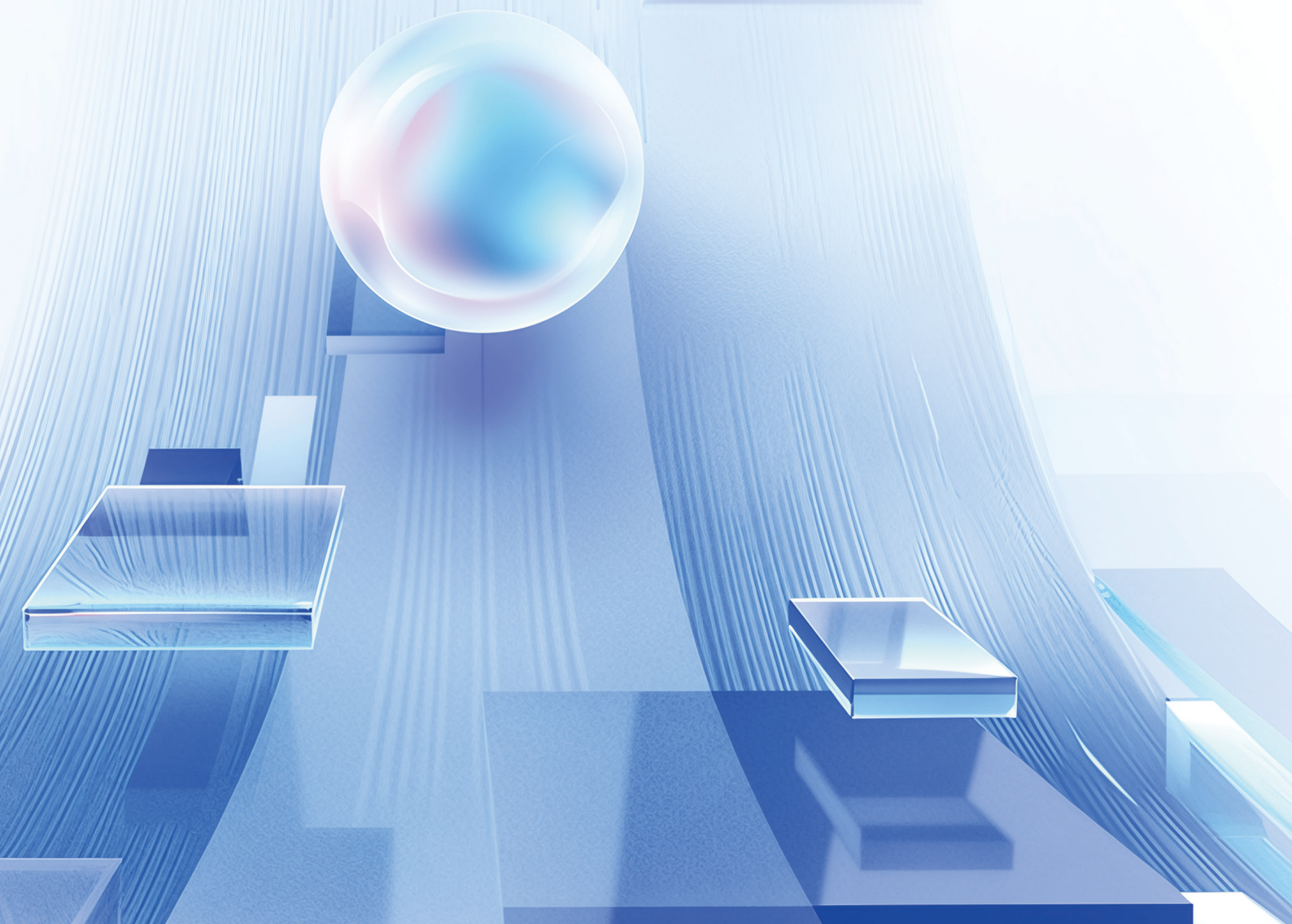
众安信科 × 众安金融科技研究院

迈向智能驱动新纪元

大语言模型赋能金融保险行业的应用纵览与趋势展望



执行摘要

2025 年初，国产大语言模型在推理性能、购置与运维成本等关键领域实现了显著优化，推动各个行业大语言模型的应用加速。与传统 AI 算法通常依赖人工定义规则和浅层神经网络不同，大语言模型在复杂语义理解、上下文推理、多任务处理和非结构化数据分析等方面展现出更高的通用性。这些进展使得大语言模型能够在多个业务领域中实现更高效的自动化处理，大语言模型正逐步推动各个行业向更加智能化的方向发展。

从当前行业实践来看，大语言模型在金融保险领域的应用已完成初步的技术验证与试点落地，正处于由试点探索向系统化集成的过渡阶段。在部署初期，行业一般优先选择在容错成本较低、决策干预门槛较小的辅助性业务场景展开，例如智能客服、智能质检、营销助手、智能快赔、Chat BI、AI 审计内控等业务领域，通过低敏感度流程的反复试点，建立模型调试与反馈机制，为后续向高复杂度核心业务的拓展奠定实践基础。以上审慎的推进方式，既体现出金融保险业对 AI“冷启动”阶段数据与经验积累不足、专业人才及硬件储备有限等客观条件的现实考量，也反映出大语言模型应用本身所需的调试周期和迭代规律。在这一背景下，金融保险机构还通过在智能试点场景构建起涵盖模型适配、业务协同与流程重构的跨部门机制，积累了系统性落地所需的组织与治理能力，这将为未来向高精度要求的业务领域拓展奠定坚实的基础与信心。

值得关注的是，大语言模型在金融保险领域的应用，不仅意味着流程效率的提升，更推动了行业信息处理范式与决策逻辑的深层次转型。这一转变并非仅是技术的替代，其本质逻辑源于机构由结构化、静态数据向多源、动态信息系统演进所引发的能力重构。过去主要依靠结构化数据与人工经验进行判断，而大语言模型的引入，使社交媒体交互、图像、语音、用户行为轨迹等非结构化信息得以系统化建模与高效分析，显著提升了风险识别的广度与响应的及时性。这一能力不仅增强了金融风险建模的灵活性，也为多源数据驱动的动态预测机制提供了算法支撑，特别是在应对突发性风险事件时，机构能够融合实时信息动态调整风险评估，提前部署资源，从而提升整体应急处置能力与运营韧性。

进一步来看，大语言模型的应用已推动银行、保险、券商等金融机构经营理念、业务逻辑与价值创造模式的整体重塑，并催生出两大显著趋势：一是金融服务的精准化跃迁，例如银行利用实时企业经营数据与社交媒体动态信息优化信贷评估体系；券商则依托知识图谱、产业链网络进行更精准的市场预测与资产配置。二是基于业务场景的跨行业生态化协同，如保险机构与医疗健康平台合作开发基于实时健康数据的预防型保险，银行与汽车、智能家居等行业共同搭建实时风险预警与主动干预的信贷服务体系。可以说，从深层次上看，大语言模型的兴起正推动金融机构构建全新的能力驱动模式。相较于以往依赖资本规模和渠道扩张的发展路径，如今的智能化转型更依赖于数据资产的有效管理、算法能力的持续演进，以及算力资源的动态协同配置。这一能力体系不仅重构了金融机构的信息处理机制与决策逻辑，也使其具备了在高度不确定环境中实现敏捷响应、精细化运营和智能资源配置的基础能力，为行业走向更高质量、更可持续的智能运营形态提供了底层支撑。

综上所述，大语言模型对金融保险行业的影响，已不再局限于技术层面的升级迭代，而正在引发一场从“经验驱动”向“数据智能驱动”的深层次认知转型。金融保险企业不仅要主动适应这一变革，更要解放固有认知模式，积极进行跨界合作、生态协同与共创共享，才能在未来智能决策时代占据主动。那些率先深入融合大语言模型技术的机构将获得因先发优势带来的技术红利，并引领行业转向数据资产、算法优化与算力生态的新型经营范式。相信随着技术成熟度曲线逐步跨越临界点，金融行业的价值创造逻辑也将随之优化和升级，这不仅体现为效率的全面提升，更是对商业创新空间与潜力的深度释放。

未来已来，唯变不变！

目录

01

大模型精进降本提速，重构价值创造底层逻辑


AI

当前全球大模型技术的发展格局逐渐呈现多元化趋势，各国在路径选择上展现出不同的技术侧重与生态布局。美国企业 OpenAI 持续通过闭源 API 服务加速市场转化，而 Meta 则以 LLaMA 系列推进开源生态，探索开放与协作并行的路径。欧洲企业如 Mistral AI 则采纳“部分开源 + 商业授权”的中间模式，在提升模型透明度的同时兼顾经济收益与技术主权诉求。在我国，形成了闭源与开源并行的发展体系，不同企业根据自身能力与场景定位采取差异化策略。其中，DeepSeek 和通义千问等模型则代表了开源技术路线，在工程效率与社区协同方面持续探索。腾讯混元则面向 B 端行业应用，强调模型的可控性与私域适配能力；字节跳动的豆包模型聚焦轻量部署与用户触达，已在多款 C 端产品中实现落地应用；百度的文心一言以闭源方式深度绑定搜索、知识图谱等业务系统，强调自有生态闭环；这些路径背后，体现了各国和企业在技术自主、市场策略与生态治理上的差异性权衡。

从技术突破路径来看，各地模型研发重心呈现分化。美国主流团队以 Scaling Law 为基础，通过扩大参数规模与优化训练机制提升性能，典型如 GPT-4 在稀疏注意力机制与强化学习反馈上的优化，使其在万亿级参数下依然具备较高推理效率。相比之下，中国团队更注重底层算法及工程层面的资源优化与实用性设计等系统性优化。这一趋势表明，技术演进正逐步从单一规模扩展向多维度协同优化转变，有助于模型在资源受限环境中的实用性扩展。

DeepSeek

DeepSeek 通过混合专家（MoE）动态路由技术将 6710 亿参数的活跃计算量压缩至 37 亿，结合自研负载均衡策略使专家模块利用率提升 24%；依托自强化学习框架（Self-Reinforcement Learning）实现无需人工反馈的思维链优化，训练效率提升 3.5 倍，并通过 DualPipe 算法在 NVIDIA RTX 4080 Super 显卡集群中达成 95% 硬件利用率，大幅降低分布式训练损耗。其基于多个知名开源大模型的深度蒸馏技术，使 32B 轻量版本在数学推理与代码生成任务中达到 GPT-4 约 80% 的基准水平。据披露，DeepSeek 的效能优势瞩目——仅为其他同规模模型训练成本（9240 万美元）的 6%。

通义千问大模型

通义千问模型家族基于大规模参数架构构建了从百亿到千亿级的完整体系，技术层面深度融合预训练基础模型与垂直领域优化能力，在对话交互、代码生成、数学推理等场景形成专项突破，并通过量化压缩、注意力机制加速等轻量化技术显著降低计算资源需求；同时积极拓展多模态理解能力，实现文本与视觉信息的协同处理。通义有望通过跨模态技术融合构建更全面的 AI 能力生态，最终形成兼顾通用化能力与行业深度应用的智能基础设施。

腾讯混元大模型

腾讯推出自研深度推理模型混元 T1，基于 Hybrid-Mamba-Transformer 创新架构，显著降低计算与内存消耗，支持超长文本高效处理（解码速度提升 2 倍）。模型通过专项优化在 MMLU-PRO（87.2 分）、CEval 等中英文推理基准中领先，适配对齐任务、指令跟随及工具调用场景，现已上线腾讯云，定价输入 1 元 / 百万 tokens、输出 4 元 / 百万 tokens，开放官网体验及企业 API 试用。

依据行业跟踪来看，随着大模型能力的持续提升，部署与场景适配问题逐渐成为模型实际价值转化的重要衡量维度。相较于早期关注模型参数规模与训练性能的阶段，当前的关注点正转向如何在多元环境中实现模型与算力资源、业务流程及系统接口的有效对接。这一变化反映出模型开发正从纯粹的算法突破，延伸至工程体系与生态协同的系统能力建设。

在已有实践中，部分团队通过结构设计的调整增强模型的可部署性。以 DeepSeek 为例，其采用稀疏门控的混合专家模型（MoE）架构，在推理过程中按需激活子模块，并配合流水线并行调度技术，提升算力利用效率并控制推理资源消耗。这类架构选择有助于模型在多卡集群下实现较高的资源适配度，从而为本地部署、专有环境运行等应用形式提供更多可能性。另一类策略则体现为模型结构与场景输入的深度耦合，例如豆包大模型聚焦于内容生成等高频轻量场景，通过对任务语义边界的精细控制实现快速响应与资源稳定性，适用于特定 C 端产品链路下的实际使用需求。

在企业级部署中，也有团队将模型能力与既有平台资源整合，形成较为紧密的生态应用路径。文心一言通过与百度搜索、知识图谱等业务模块集成，构建了一种以平台为基础的模型嵌套体系，适用于数据结构清晰、业务流程较为稳定的应用场景。腾讯混元则强调模型在政务、金融、制造等垂类领域的私域部署能力，并探索将企业内部知识系统与语言模型结构协同优化的路径。上述做法体现了从“模型通用能力”向“场景精度适配”的渐进转变。

尽管当前部署路径呈现多元发展态势，模型在实际落地过程中仍面临一些工程侧与生态侧的适配挑战。例如，在算力环境方面，大多数模型当前仍以 CUDA 体系为主进行推理加速，而国产编译器和执行框架的适配正在持续推进；在跨行业部署中，不同平台之间的接口规范、数据表达方式和微调流程尚未完全统一，这在一定程度上对模型迁移效率提出新的要求；此外，部分业务场景反馈周期较长，模型微调与能力更新的节奏需进一步优化，以增强其持续适配能力。这些现象反映出部署能力不仅是单一技术点的突破，更依赖于模型、系统与场景之间的协同演进。

整体而言，大模型的部署能力正在从“可用”向“可适配”迈进。从已有进展可以看出，未来的竞争焦点或将聚焦于如何构建跨架构、跨场景的柔性部署机制，以及如何通过生态联动提升模型在异构环境中的运行效率与反馈响应能力。技术演进的方向，正从中心化、静态部署逐步拓展至弹性化、协同式部署能力的构建过程之中。

> 真正的模型优势，不止于性能，更在于它如何嵌入现实系统，成为业务流程的一部分而非附加品。

一、大模型技术路线持续优化，金融保险业迎智能新机遇

随着大模型技术的快速发展，全球大模型的技术演进长期遵循“参数规模决定论”的底层逻辑。最初，通用大模型通过不断堆砌千亿级参数和海量算力来覆盖长尾场景，借助巨大的计算能力和数据量提升模型的性能。然而，这种技术范式逐渐暴露出了一些问题，尤其是在边际成本不经济方面。举例来说，GPT-4 级别的大模型单周期训练成本已接近 5000 万美元，其日均亿级请求的推理能耗估算相当于 1.2 万户美国家庭一年的用电量（Hugging Face, 2023）。这表明，虽然大规模堆砌参数能够有效提升模型的性能，但其带来的高昂成本和不可持续的能耗，促使了对更加经济且高效技术路径的探索。

这种挑战催生了垂类大模型的发展。与传统通用大模型不同，垂类大模型通过定制化训练，结合行业特定的数据集和业务流程，能够在减少计算成本的同时提供更加精准的行业解决方案。垂类大模型的一个核心优势在于它们能够深入理解特定行业的需求，结合行业知识图谱和数据结构优化决策过程，从而克服通用模型在处理特定业务时的灵活性不足和精准性问题。随着垂类大模型技术的不断进步，尤其是在计算资源优化和数据定制化训练方面的突破，它们逐渐成为解决行业痛点的关键工具。例如，在金融行业，垂类大模型能够通过对市场数据、交易记录的深度学习，提升风控管理和合规审查的效率；这些模型通过将行业数据与业务规则深度结合，能够提供更高效、精准的智能化服务，从而加速行业的数字化转型。

从产业层面来看，垂类大模型的发展不仅依赖于技术本身的进步，还与全球技术生态的深度合作密切相关。以华为云与 DeepSeek 的合作为例，推动了大模型在不同行业中的快速落地和应用。与此同时，通义千问通过开源大模型的方式，降低了中小企业接入这些先进技术的门槛，进一步推动了技术的普及化和应用场景的创新。这一开放生态使得越来越多的中小型企业能够快速构建行业定制化解决方案，推动了技术创新的扩展。

展望未来，垂类大模型的潜力将不仅限于单一行业的智能化升级。随着多模态学习和边缘计算等新技术的加速发展，垂类大模型将能够结合更多的数据源，如图像、音频、传感器数据等，提升在更加复杂、多变的业务场景中的适应能力。这一趋势为各行各业提供了更多的商业机会，尤其是在金融保险行业，垂类大模型的应用将推动业务决策、风险评估和客户服务等领域的智能化，进一步优化行业价值链。

1. 国内外大模型技术路线演进框架

(1) 国际大模型范式的起源

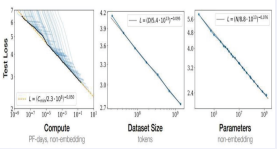
大语言模型的技术起源与发展，既是技术创新的结果，也是计算能力、行业需求与数据资源相互作用的产物。从最初的规则推理与知识图谱，到深度神经网络和大规模训练模型的引入，人工智能逐步突破了传统框架。随着 Scaling Law（规模法则）的诞生，深度学习模型的规模扩展带来了性能的指数级增长。这一法则表明，随着模型参数的增加，性能将呈现正相关，推动了大模型技术的飞速发展，为后续的大模型应用提供了坚实的理论基础。

图 1：Scaling Law 的演化：从 Pre-Training 到 Post-Training

Scaling Law 的定义

Scaling Law（扩展定律）是人工智能领域的核心理论框架，揭示了模型性能与模型规模之间的幂律关系，即通过系统性地增加模型参数量、训练数据量和计算资源，可显著提升模型的性能表现。

OpenAI 2020 年论文首次系统性验证了这一规律，奠定了大模型规模扩展的理论基础。



Scaling Law 从 Pre- Training 阶段到 Post- Training 阶段的演化

背景：

GPT-2/3/4、Gemini、Claude、Llama、Qwen 等系列模型的成功，验证了 Scaling Law 的暴力美学（即通过扩大规模提升性能）。然而，随着预训练规模持续增长，行业面临瓶颈。其核心矛盾：数据质量与数量的限制导致 Scaling Law 的收益增速放缓（社区称为“Scaling Law 撞墙”）

里程碑事件：

2024 年 9 月 12 日：OpenAI 发布 o1 模型，提出推理阶段的 Scaling 新范式：
- Test-Time Scaling：通过增加算力与响应长度（输出 token 数），模型性能持续提升。
- 对于业绩的启示：OpenAI 驳斥“撞墙论”，强调其通过双模型体系（a 系列与 GPT 系列）延续 Scaling Law 的有效性。

行业现状：

工业界与学术界开始复现 o1 的推理扩展能力，推动大模型研发方向从“规模优先”转向深度思考模型（Deep Reasoning Models），涌现出一大批 o1 模型，诸如：Qwen-QwQ、Gemini 2.0 Flash Thinking、DeepSeek R1、Kimi K1.5、GLM-Zero、Skywork-o1 等等。
O1 本质是让大模型学会自动寻找从问题到正确答案的详细中间思考步骤（如思维链 CoT），以此解决复杂问题。DeepSeek R1 和 Kimi K1.5 论文中 RL Scaling Law 得到进一步验证。

数据来源：公开资料，众安金融科技研究院

然而，尽管技术上取得了显著突破，通用大模型在特定行业中的应用仍面临挑战。通用模型在处理多种任务时表现出色，但它们的能力在应对行业专属任务时显得力不从心。例如，金融行业需要高精度的风险管理与合规检测，医疗行业则要求疾病诊断和治疗方案的精准性，这些行业的复杂性要求模型具备行业特定的专业知识，通用模型往往无法提供足够的支持。因此，垂类大模型应运而生，成为弥补这一空缺的关键。

垂类大模型（Vertical Large Models）通过专门化的训练，能够针对行业特有的数据集与业务规则进行优化，提升模型在特定领域中的精准度和效率。与通用大模型不同，垂类大模型不仅处理行业特有的数据，还能深入理解行业的专业术语和决策流程。例如，在金融领域，垂类大模型能够通过训练交易记录、信用评估、市场数据等信息，为风险评估、合规检查等任务提供精准支持。而在医疗领域，垂类大模型通过学习患者病历、医学影像、临床数据等，为医生提供智能化的诊断与个性化治疗方案，提升了医疗决策的精准性和效率。

随着技术的发展，垂类大模型的应用已逐步扩展至金融、医疗、法律、零售等多个行业。通过行业专用数据集的训练，垂类大模型不仅提升了任务处理的精度，也显著优化了业务流程。例如，在金融行业，垂类大模型能够通过对客户行为数据和交易历史的分析，自动化完成风控模型和合规审查，提升了金融机构的决策效率；在医疗领域，垂类大模型通过分析不同病历数据，生成个性化的治疗建议，帮助医生提高诊断的精度。垂类大模型不仅能够优化行业工作流，还能加速行业的智能化转型，进而提升企业的核心竞争力。

垂类大模型的成功不仅依赖于技术本身，更得益于技术生态的深度融合。以 OpenAI 与微软为例，微软通过其强大的 Azure 云平台为 OpenAI 提供了必要的计算资源支持，并将大模型技术推广至全球市场。通过 Azure OpenAI Service，微软不仅推动了大模型的普及，还加速了跨行业应用的落地。与此同时，像 Grok 这样的新兴技术公司，通过开源大模型的方式，降低了技术的进入门槛，促进了中小企业接入先进技术，推动了行业技术的普及化。这种技术生态的变革，使得大模型的技术能力不再局限于大型企业，更多的中小企业也开始受益。

随着多模态学习和边缘计算等技术的快速发展，垂类大模型的能力将进一步增强。在未来，垂类大模型将能够结合更多的数据类型（如图像、音频、视频等）和更为复杂的业务场景。通过集成来自不同领域的数据，垂类大模型将能够提供更加灵活和智能的决策支持，从而推动更多行业的智能化转型。在这个过程中，如何平衡技术创新与数据隐私、合规性等监管问题，依然是技术发展的关键挑战。

总体来看，垂类大模型的崛起，不仅推动了各行业的数字化转型，还为企业带来了前所未有的商业机会。随着技术不断发展与商业应用逐步扩展，垂类大模型将在推动行业智能化浪潮的过程中扮演更为重要的角色。它们不仅代表了技术应用的突破，更是推动行业创新和差异化竞争的重要工具。未来的商业机会将集中于如何有效整合行业知识、数据能力与推理机制，借助垂类大模型实现行业的定制化转型。企业若能深度融合这一技术，必将抢占行业智能化转型的先机，从而在未来的竞争中占据有利地位。

(2) 国产大模型技术的突破

自 2023 年以来，中国在大模型技术的自主可控发展上取得了实质性进展，特别是在算力自主化和数据优化方面。以华为昇腾 910 和寒武纪思元系列为代表的国产 AI 芯片，已逐步承担起大模型训练与推理的关键算力支撑，提升了本土 AI 基础设施的独立性。然而，算力国产化并不仅仅是硬件的替换，更标志着软硬件协同体系的重构。这一体系的有效运转，需依托国产模型的深度适配与优化。DeepSeek、通义千问等模型在架构层面对国产芯片进行了有针对性的适配与压测验证，从而提升了国产算力的利用效率，也为大模型生态的闭环建设奠定了基础。

在技术能力方面，DeepSeek-R1 代表了国产大模型在推理性能与训练效率上的突破。通过引入混合专家（MoE）架构与 GRPO 强化学习策略等多方面创新，DeepSeek-R1 在数学推理、代码生成等复杂任务中的表现已达到甚至超越 GPT-4.5。这一技术突破标志着中国大模型在“性能”与“成本”之间实现了理想的平衡，特别适合用于私有化部署和垂直行业应用，推动了国产模型的广泛应用。如图所示，自 2024 年 11 月起，我国大模型厂商开始了发布具备推理能力的大模型，并于 2025 年 1 月形成集中释放态势，呈现出从预研试水到规模化落地的演进规律。

图 2：类 o1 深度思考模型发布时间线

数据来源：公开资料，众安金融科技研究院

从技术演进路径来看，国产大模型生态逐步呈现功能分化与专业化趋势。语言生成模型（如 DeepSeek V3、通义千问 Qwen2.5）专注于深化语义理解与对话生成能力，而以 DeepSeek-R1、文心 X1 为代表的推理模型，则在逻辑决策与结构化任务中不断突破，逐步在金融、科研、教育等高认知要求的行业场景中形成优势。此外，国内厂商在多模态和垂直场景的拓展方面也取得了积极进展。通义千问等系列模型，已经在医疗问诊、教育、科研建模等领域开始试点应用，展现了高于通用模型的专业适配能力。腾讯“混元”与百度“文心”系列则在工业制造、智能教学等领域实现了产业应用，推动大模型从通用生成工具向行业认知平台转型。

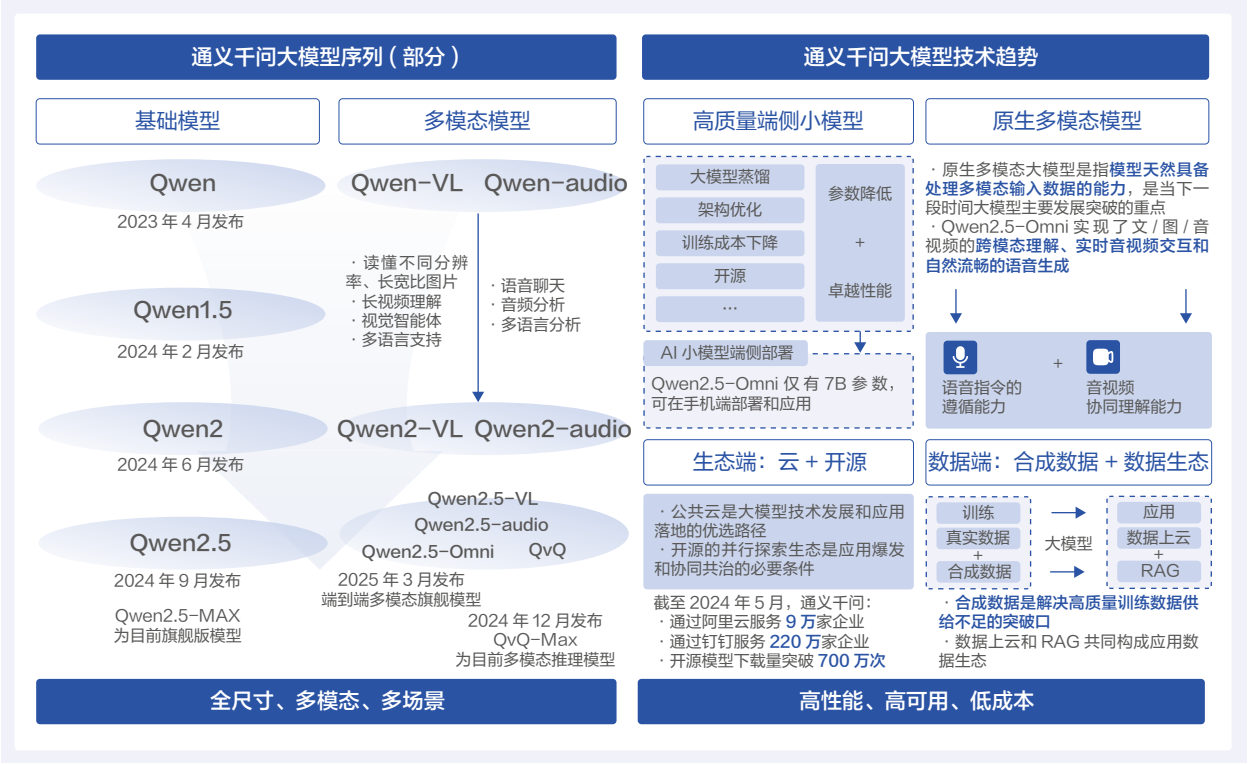
在生态建设层面，国产大模型厂商普遍采取“开源协同 + 轻量化部署”的双轮策略，推动 AI 能力向更多场景渗透，尤其是中小企业和边缘计算领域。DeepSeek 通过开放模型权重、API 接口及文档，提升了开发者社区的活跃度，加速了中小企业的应用落地。同时，蒸馏技术的应用使得大模型能力得以迁移到轻量模型（如 Qwen-32B、MiniGLM），显著降低了终端部署的成本，并保持了关键任务性能。在多个应用场景中，轻量化模型已经实现了对 OpenAI-o1-mini 的局部超越，成为终端设备部署中的可行路径。这一进展加速了大模型的商业普及速率，并推动国产大模型在“高性能、低成本、强适配”的目标上迈出了坚实步伐。

2. 中国大模型崛起筑基金融业升维

(1) 中国多模态大模型最新发展现状跟踪

近年来，多模态大模型技术成为人工智能领域的核心研究方向之一。与传统的单模态模型相比，多模态模型通过融合文本、图像、音频、视频等多种信息，显著提升了机器对复杂场景的理解与生成能力。海外技术巨头如谷歌、OpenAI 率先推出原生多模态模型 Gemini 和 GPT-4o，通过统一架构实现跨模态信息的深度融合。而国内以阿里通义千问为代表的多模态技术也取得显著突破，尤其在中文场景优化和开源生态建设方面表现突出。具体而言，海外企业在多模态大模型领域起步较早，技术路线以“原生多模态架构”为核心。例如，OpenAI 的 GPT-4o 进一步强化了多模态推理能力，支持用户通过自然语言指令实现跨模态任务调度，可以根据图像生成描述性文本或基于文本生成图像草图。这类模型的共性在于通过大规模跨模态预训练，构建了模态间语义关联的深度理解能力，为后续任务适配提供了扎实基础。然而，其技术细节与训练数据仍高度封闭，限制了行业生态的开放性发展。

图 3：通义千问技术新趋势



数据来源：公开资料，众安金融科技研究院

在我国，通义千问通过模型的不断更新迭代，在多模态和模型性能优化等方面实现了明显提升，结合生态端的开源实践和数据端的生态打造，通义千问实现了”全尺寸、多模态、多场景“+”高性能、高可用、低成本“的多模型应用与生态，重塑了传统的人机交互方式，有望带动商业模式的进一步突破与更新。其中 2025 年 2 月发布的 Qwen2.5-Omni 模型，通过“双核架构 + 流式处理”技术实现了全模态实时交互。具体而言，该模型将视觉、文本等不同模态的处理模块解耦为独立核心，同时通过流式数据管道实现模态间信息的动态融合。例如，在处理“根据图片生成故事”任务时，视觉核心首先提取图像语义特征，文本核心则基于特征生成连贯文本，二者通过共享注意力机制实现高效协同。这一设计不仅降低了多模态任务的响应延迟，还提升了模型在复杂场景下的稳定性。在性能表现上，通义千问的轻量化多模态模型 QvQ 刷新了多项任务纪录。例如，其视觉推理模块通过引入“链式思维（CoT）增强方法”，将数学问题分解为多个中间推理步骤，并结合视觉特征进行分阶段求解。此外，通义千问的多模态开源生态覆盖文本、图像、音频等全场景需求，已开源的 Qwen-VL 和 Qwen-Audio 模型支持开发者快速构建定制化应用。这种“全模态、全尺寸”的开源策略，显著降低了多模态技术的应用门槛，推动了中文大模型生态的繁荣。

推动多模态大模型快速发展的背后，是算力基础设施、数据资源整合能力与算法创新的共同驱动。

在算力层面

中国近年来持续加码智能计算中心和高性能计算平台建设，支撑大规模预训练与推理推送的算力资源需求，保障模型训练的效率和稳定性。

在数据层面

多模态模型对海量异构数据提出了高质量、高多样性、高对齐度的标准，促进数据治理、数据标注、数据融合等产业链环节的升级优化。

在算法层面

越来越多的研究聚焦于统一模态表示学习、跨模态注意机制、视觉与语言协同建模等领域，不断提升模型对现实世界语义关系的建模与泛化能力。

在模型能力提升的同时，落地场景也将日益丰富。多模态大模型已被应用于智能客服、媒体内容生成、教育辅助、工业质检、自动驾驶等多个领域，其中一些初步商用化的成果已显现出良好的经济与社会价值。例如，电商平台通过多模态模型实现商品图文生成、搜索优化与用户画像重建，极大提升了用户体验与平台转化率；从国际比较视角来看，中国多模态大模型的发展路径在一定程度上体现了“技术自研 + 产业导向”的组合策略。与 OpenAI、Google DeepMind 等企业以算法驱动为核心的路径不同，中国科技企业往往采取场景牵引式发展模式，在特定任务上快速打磨模型性能，通过真实业务推动模型反馈优化，形成更具产业落地能力的技术形态。这种“任务驱动式”发展策略不仅缩短了模型从实验室到应用的转化周期，也帮助企业在市场中形成了特有的应用优势。

中国多模态大模型的发展也面临若干挑战：

数据

高质量跨模态数据的采集与标注成本极高，尤其在视频、语音等模态中仍存在数据稀缺、样本偏差等问题，影响模型泛化能力

能力

模型在语义一致性、事实准确性推理逻辑性方面仍需进一步提升，当前一些模型在生成内容时仍存在 " 幻觉”现象，限制其在高风险业务领域的广泛应用

资源

模型规模不断扩大带来的算力成本和能耗问题，也引发了关于绿色 AI 和低碳训练的思考

尽管挑战犹存，但随着政策引导、产业协同与技术演进的持续深入，中国多模态大模型有望在未来实现新一轮的突破。

政策引导：加强基础设施与制度保障

政府层面正在推动算力资源国家级统筹与跨区域调度机制，加快建设国家大模型开源社区和标准体系，支持关键技术联合攻关。

产业协同：构建行业专家模型与智能体系统

企业层面，头部科技公司持续深化行业合作，通过垂直场景优化模型微调方式，逐步构建起一批“行业专家型”多模态模型，并推动其向具备商业闭环能力的智能体系统演进。

科研突破：攻坚关键理论瓶颈

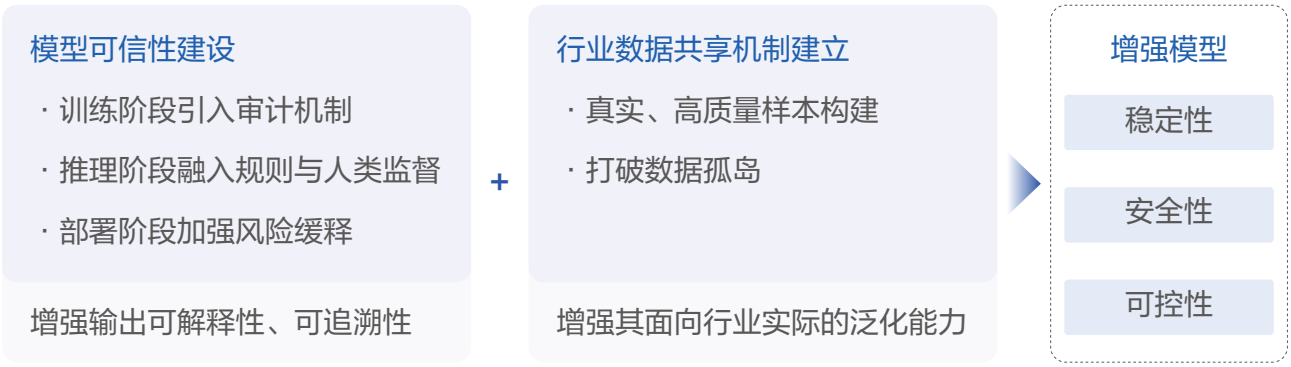
研究层面，学界正加大对模型可解释性、可控性和安全性的投入，推动跨模态推理、语义一致性约束等基础课题研究，试图解决模型“能看能说但不能理解”的关键痛点。

(2) 大模型时代金融保险业智能化转型迎来新机遇

在此发展背景下，多模态大模型为金融保险行业的智能化转型带来诸多启示。金融与保险作为高度依赖数据与模型支撑的行业，具备天然的数据优势和应用迫切性。传统的金融风控、保险核赔、客户服务等环节多依赖规则驱动与结构化数据分析，难以处理非结构化信息所蕴含的巨大价值。而多模态大模型所具备的图文、语音、视频等多源数据理解能力，正好填补了这一能力空白，为建立面向未来的“多模态金融认知系统”提供了基础技术支撑。

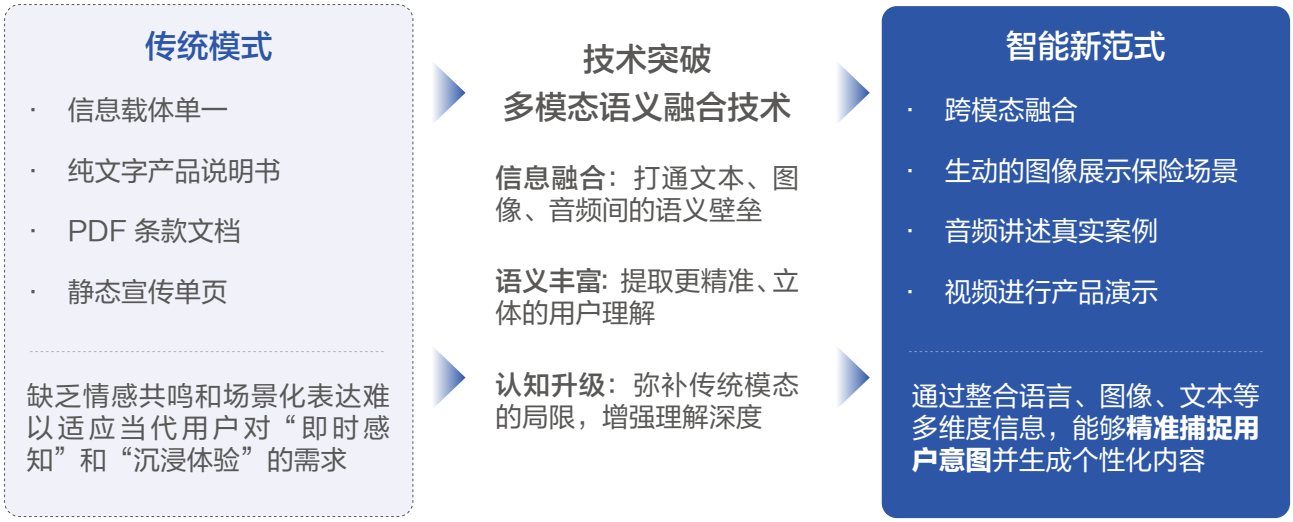
具体而言，多模态模型在智能客服与销售辅助场景中对提升用户体验有所助力。通过融合客户语音交流、社交媒体文本、身份图像等信息，模型可构建出更加真实全面的客户画像，实现对用户意图的精准识别与产品匹配。在当前，多模态大模型一般应用在理赔审核与欺诈识别等环节，通过图像与语音协同分析理赔材料与事故现场信息，辅助判断理赔的合理性并提供潜在欺诈风险的识别结果，提升审核准确率与风控能力。此外，对于保险中介机构或销售渠道，通过文生图、图生视频等多模态模型能力还被用于短视频、图文、音频等营销内容的批量生产，为获客与渠道运营效率的赋能。

从中长期视角出发，多模态大模型有望成为金融保险企业构建“数字智能中台”的核心引擎。当前，金融保险行业普遍面临数据分散、系统割裂、模型难以迁移等痛点，制约了智能化战略的纵深推进。而多模态大模型所具备的统一语义建模能力，能够打通语言、图像、表格、结构化数据之间的语义屏障，为企业构建具备语义理解、逻辑推理、知识生成与任务执行能力的智能服务基础设施。依托这一基础平台，保险企业可快速实现产品问答、营销话术生成、风险洞察、运营优化等多种能力的模块化调用，让实现“以模型为中心”的敏捷业务创新成为可能。



此外，随着多模态大模型与业务融合的不断深入，金融保险行业有望探索出更加智能化的运营与服务模式。通过 AI agent 进行自主分析问题、制定应对策略、与人类协作决策的能力，从而在客户服务、风险管控、营销策划、流程自动化等方面实现实质性提升。这种“AI+ 人机协同”的业务新范式，有望重塑保险从销售到理赔、从风控到再保的全链条流程逻辑，推动行业迈向以用户为中心、以认知为核心的智能化新时代。

当然，要实现上述智能化范式的落地，关键之一在于多模态大模型底层技术的不断进步与升级。要求多模态大模型的技术演进从单一模态融合向跨模态统一表征学习深度推进，此过程的核心在于构建共享语义空间以消除文本、图像、音频等异构数据的鸿沟。同时，还需要轻量化部署技术如稀疏激活架构（MoE）与动态剪枝进一步降低推理成本，结合边缘计算实现设备端实时响应，使得金融保险行业得大模型在低算力环境下仍能保持高效性能。



综上所述，中国多模态大模型在基础设施、算法能力与应用场景等方面正呈现出系统化发展趋势，正在逐步构建起具有中国特色的多模态智能生态体系。对于金融保险行业而言，这一技术演进不仅提供了强大的赋能路径，也促使企业重新思考未来的组织能力结构、客户服务策略与数据治理机制。多模态大模型不只是一个工具或平台，它还将深刻地改变金融认知与决策方式，为保险行业打开一条通向“语义智能驱动”的高质量发展新路径。在全球化竞争中，开源社区与国际标准输出将是重要的一环，本土技术通过中文能力优势与场景适配性向国际市场延伸，而算力国产化与软件生态的协同突破将决定长期竞争力。未来，谁能率先构建起可控、可信、可持续的多模态智能中台，谁就会有更大的可能在金融科技的下一轮竞赛中掌握主导权。

3. 大模型前沿范式演进及行业启示

(1) 大模型技术的用户交互范式重构

当前，大模型应用正在推动用户交互范式发生根本性重构，其核心矛盾从“如何精准描述需求”转向“如何高效传递意图”。传统对话式 AI 要求用户将模糊的需求逐步转化为清晰的结构化语言，这种由人类思维转译为机器语言的“翻译损耗”，在复杂场景尤其明显。这一问题的本质在于交互界面未能有效贴合人类认知直觉，迫使用户学习并适应机器的表达规则，显著提高交互成本。展望未来，新一代交互技术逐渐接纳非结构化的表达方式，例如通过大模型的上下文补全、语义推理与场景建模能力，实现碎片化意图到明确任务链的快速转化。**这体现了底层交互逻辑的根本性变革：交互主导权从人类主动转译逐渐转向 AI 主动解析。**

随之而来的变革是大模型在服务形态上的进化，由单纯的信息检索工具转型为多模态服务中枢。当用户提出宽泛请求时，AI 可主动协调跨模态数据（如文本到图表再到代码）、动态调用 API（本地资源与云端协同分配），并根据设备感知系统自动优化服务组合，实现全面综合的解决方案。这种技术变革打破了传统工具链的壁垒，将 AI 升级为企业与用户高效协作的“生产力中台”。

进一步而言，技术融合的趋势进一步推动了生态系统的深度整合，交互界面从单一信息入口逐渐转型为多场景智能服务的中枢节点。横向资源整合与纵向深度集成，使得单一接口能够有效联动多种服务，例如支付或医疗、保险等领域的综合服务场景。

交互界面变革的最终趋势，是界面与设备操作系统及硬件资源深度融合，形成“系统级全场景 AI”能力，包括内存优化管理、语义搜索嵌入及芯片架构优化等，全面提升用户交互体验。这种交互模式进一步从被动响应转型为主动预测，通过动态用户画像与场景预加载技术，实现服务的实时认知预测与高度个性化推荐。

目前交互模式痛点

01
意图识别精确性不够完善

02
长尾需求容易被误解

03
跨平台任务协同可能出现服务延迟或断层

04
个性化推荐与隐私保护之间的平衡问题仍未有效解决

针对以上问题，技术发展正在模型架构、计算范式、增强人机信任三个层面走向融合发展。

三个层面的融合

模型架构层面	将通用大模型与行业专有知识相结合，建立针对性更强的“通才 + 专才”混合模型
计算范式层面	采用边缘设备轻量化模型与云端重型模型的协同计算模式，优化服务响应效率
增强人机信任方面	引入因果推理机制，提升 AI 决策透明度和可解释性

(2) 金融领域多智能体系统应用与演进前景

人工智能技术加速演进，正在推动金融领域的智能化进程从单点突破走向多维系统扩展。多智能体系以集群智能、快速适应和高效决策能力为核心优势，成为海内外金融机构在数据分析、风险管理与智能决策领域的当前热门的探索方向之一。

当前阶段，多智能体系概念仍处于初步探索期，应用实践普遍以单 agent 部署为起点。单智能体阶段的主要任务，是在现有模型能力框架内，积累场景理解、任务执行与协同运作的经验。随着应用实践不断深入，单智能体在覆盖复杂业务需求时，逐步显现出对推理深度与知识广度的更高要求，促使行业同步关注底层大模型能力的结构特性。强化学习在这一体系中，主要发挥基于奖励机制的路径优化作用，能够有效提升决策成功率，但不直接赋能新的推理能力或知识体系。因此，模型的实际渗透能力最终取决于基底架构设计、token 量规模及训练数据的多样性与泛化水平。这些基础特性，成为多智能体系早期探索阶段演进节奏的重要决定因素。

基于对当前大模型的基础能力边界的认知，行业实践逐步将视角延伸至应用端的局部增强策略。金融机构开始探索通过引入高质量的领域“小数据”，结合模型蒸馏、数据蒸馏和模型微调等，以支撑特定业务场景下智能体的精细化适应与全流程优化。这种以垂直领域为导向的策略，使细分领域应用成为当前智能化演进的重要支撑方向。尽管小规模域性模型在总体泛化能力上与全尺度基底模型存在差异，但凭借对专业场景的高适配性，在特定业务模块、细分产品线和专业服务场景中展现出更高的智能适配性与部署可行性，推动金融行业智能应用从局部优化向体系化深化演进。这一局部增强路径，不仅提升了智能系统在垂直细分市场中的响应效率，也为后续多智能体体系在复杂业务结构中的协同演进奠定了扎实基础。

在体系构建方面，多智能体系统由多个独立决策的 agent 组成，依托协同互动、动态竞争与自适应机制，共同完成复杂环境下的智能决策任务。金融机构通常从单 agent 场景试点出发，逐步向多 agent 协同演进，覆盖数据分析、风险管控、智能交易与智能服务等核心领域，持续积累技术成熟度与运营协作能力，推动智能化水平稳步升级。面向未来，多智能体系统的发展趋势将包括跨系统数据协同能力的提升、分工与资源配置的智能优化，以及连续性微调与动态协作机制的深化。这一过程中，需同步打破数据孤岛，贯通流程链路，完善通信协同与负载均衡机制，同时在经营机制与合规管理层面推进行业级系统优化，以确保智能体系的稳定演进与可持续增长。

二、新技术驱动成本急速下探，垂类大模型助推流程再造

在过去的大模型发展周期中，高昂的训练成本、复杂的运维要求以及推理阶段对算力资源的持续消耗，曾是制约大模型规模化商业落地的主要挑战。然而，随着模型结构创新、算法优化与系统调度技术的持续演进，围绕训练、运维与推理环节的成本正在实现系统性下探。以专家模型架构（MoE）、混合精度训练、模型蒸馏与并行调度框架为代表的新技术组合，正在显著提升资源利用率与模型推理效率，为大模型的商业化应用构建起更具可行性的成本结构基础。与此同时，面向具体行业场景的垂直大模型正逐步补充并优化通用模型的能力边界，推动金融、保险、零售等关键领域在流程结构与服务链条上实现深度重构，并在效率提升的同时降低部署成本与算力资源消耗。通过构建更具行业适配性的模型体系，企业得以在核心业务环节中实现更精细的智能能力嵌入。

1. 大模型产业化三阶成本曲线下探

在金融企业推动基础大模型落地应用的过程中，成本管理始终是决策者最为关注的核心问题。尤其是当大模型技术应用能力成为行业智能化竞争的关键变量时，如何以最低的资源代价换取最优的业务收益，就成为企业决策者必须解答的命题。从行业实践来看，当前金融企业在基础模型应用上所面临的成本结构主要由三部分构成：购置成本、推理成本以及持续运维成本，三者共同决定了技术部署的门槛与可持续性。

(1) 购置成本：筑牢智能底座的前期投入

购置成本是金融企业应用基础模型的第一道门槛，也是资本投入最为集中的环节。这一阶段的投入往往具有一次性、结构化、不可逆的特点，涵盖硬件设施的部署、模型本体的获取、以及面向具体业务场景的系统开发。对于本地化部署模式而言，企业需要采购高性能服务器、GPU/TPU 计算集群，同时承担数据中心的建设或租赁费用，尤其在对合规性与数据安全性有高度要求的银行、证券、保险等机构中，自建 AI 平台往往是默认选项。在模型采购方面，企业可能面临两类选择：

01

API 调用模式

直接通过 API 调用商业化模型，如 DeepSeek 等，虽无需复杂部署，但需要支付持续的调用费用和服务订阅费

02

本地化部署模式

基于开源模型进行本地化适配与二次开发，如通义千问等开源基础模型，企业可自由下载使用，但需投入大量人力物力完成数据适配、功能开发与业务集成，这也构成了另一个隐性成本

从开发角度看，购置成本还包括围绕模型构建的一系列应用系统的搭建，例如智能客服系统、风控建模平台、合规审计助手等。这类应用往往需要定制化开发、跨部门协同，甚至需要引入外部技术服务商完成集成测试。以某头部保险机构为例，在构建面向理赔流程的智能问答系统时，虽选择了开源模型作为底层语言引擎，但在医疗术语本地化、数据加密协议适配、与原有理赔系统对接等环节上仍投入了超过数百万的定制化开发费用。值得注意的是，虽然开源模型在表面上显得“零成本”，但其背后所依赖的工程开发与模型调优投入往往远高于商业 API 的“即插即用”，尤其是在金融高度结构化和监管约束密集的场景下，这一差距更为显著。

(2) 推理成本：运行业务量的弹性支出项

而在基础大模型完成部署进入实际运行阶段后，推理成本迅速成为决定模型可用性与业务可扩展能力的第二大核心支出项。所谓推理成本，指的是模型在运行过程中所产生的实时算力需求及其衍生的带宽、电力、冷却、系统资源等全套运行费用总和。

与一次性投入为主的购置成本不同，推理成本呈现出明显的长期性、业务波动敏感性与资源负荷传导性，其管理效率将直接影响模型整体投入产出比。

在中国金融行业中，许多中大型银行、保险机构和证券机构出于数据安全、系统可控、合规审计等考虑，普遍选择本地化部署模式运行大模型。这意味着企业需完全自担算力运行所带来的各类成本压力。其中，GPU 集群运行的电力消耗成为推理阶段最直观的直接成本，而更大的隐性成本则来自于机房的冷却系统、电源冗余配置、系统运维人力、软件环境升级、网络带宽保障、负载均衡与容灾架构等多项配套系统支出。

相比之下，采用云平台提供的模型即服务（MaaS）或软件即服务（SaaS）模式，在推理成本初期具备明显的成本弹性优势。多数国内主流大模型厂商如 DeepSeek、阿里“通义千问”API 服务、腾讯“混元模型”等，均已推出基于 Token 计价的商业化推理服务。按公开定价标准，例如 DeepSeek-R1 的输出费用 16 元 / 百万 tokens（含思考过程 + 答案），企业可按调用量自由扩缩，无需提前部署硬件或预留算力资源。

然而，这种云端服务模型在中国金融应用中也存在一定边界，特别是在核心高频场景或计算强依赖任务中，推理成本往往会随着调用量急剧上升而快速积累，形成指数级的“Token 成本爆炸”效应。具体而言，在如个人信用评估、反洗钱交易路径建模、服务电话等金融机构日常高密度任务中，模型需连续处理复杂的逻辑链条、跨数据源语义融合与长文档抽取请求，导致单次任务调用 Token 数远超普通客服场景的平均水平。

更重要的是，考虑到国内多数云厂商在金融级模型服务中尚处于探索阶段，调用定价机制尚未充分市场化，金融机构在大规模调用时常面临成本上限不明、资源调度不确定、服务等级协议（SLA）弹性有限等运营不确定性。因此，越来越多的机构开始采取“本地 + 云”的混合部署架构：将合规与频繁调用场景转向本地化运行，以保障安全与成本可控；而将外围场景或试验性功能留在云端，获取快速上线与灵活调用优势。难度，企业可将更多精力集中在业务创新和客户服务上。

综上所述，中国情境下的推理成本结构具有明显的“部署模式分化”特征。在政策合规、安全可控与成本弹性之间取得平衡，已成为当前金融企业智能化部署中的核心设计命题。通过精细化测算调用结构、分场景拆分运行策略并构建成本监测机制，金融机构有望在保障业务连续性的同时，建立更加可控、高效的模型推理体系。

(3) 运维成本：保障模型演进的持续投入

在金融机构推动大模型应用的全过程中，持续运维成本作为继购置成本与推理成本之后的第三类核心支出，正逐步成为 AI 投入是否具备长期可持续性的重要衡量指标。不同于一次性投入的硬件采买或按需增长的模型调用费用，持续运维成本主要由模型本体的持续优化、业务系统的长期打磨以及数据治理与合规的日常管理所组成，其支出模式具有结构复杂、更新频繁、职责交叉等特征，极易被金融企业在初期规划中低估，进而形成模型部署“短期有用、长期失控”的隐性挑战。

模型微调

从模型自身角度来看，持续微调已成为提升模型在垂直行业保持长期表现能力的刚性要求。金融行业数据更新周期快、行业术语更迭频繁、政策逻辑高度动态化，大模型若长期不做优化，极易出现“知识滞后”“语义漂移”与“规则脱节”等问题，影响其对新语境的理解与应对。这类微调工作往往涉及数据工程、提示词优化、模型安全性回归测试等多个环节，需周期性投入工程与专业资源。

应用打磨

应用系统的持续演进也构成了运维成本的另一重要来源。在模型部署初期，大多企业聚焦于核心模型能力的验证与输出准确率，但随着模型逐步嵌入金融业务流程中，用户体验、前端交互、反馈机制、功能迭代等系统外围环节的重要性迅速提升。例如，某大型保险机构将大模型用于构建理赔问答助手，引入更符合客户话术的响应模板，并将高频纠错信息用于训练集反向修正。整个优化流程周期近多个月，涉及产品设计、NLP 工程、客户关系三方部门协作，虽不属于模型本体投入，但对最终业务价值实现具有决定性意义。

合规治理

数据治理与合规成本则是持续运维中最具复杂性与不可压缩性的支出项。金融行业对客户数据、交易数据、医疗信息等均存在高度敏感性要求，大模型在处理上述数据时，必须具备完善的数据脱敏、日志审计、访问控制与加密存储机制。此外，随着监管对 AI 技术使用的透明化、审计化提出更高要求，模型调用行为必须具备可回溯性与合规性，输出结果需经由人机协同审阅，特别是在风控、授信、反洗钱等关键场景。该类系统的建设成本可能会超过初始模型的调用费用，成为 AI 系统典型的需要治理补强的场景。

为提升运维效率与分摊长期成本，已有金融机构正逐步探索与 AI 厂商、科研机构构建联合实验室或持续共建平台，通过行业知识标签体系共建、调优数据资源共享、模型应用层组件协同开发等方式，既提升了模型对金融语境的深度适应性，也显著降低了单家企业独立承担全部运维任务的边际压力。

运维成本的三大构成与应对策略

模型微调	应用打磨	合规治理
· 金融行业语境高度动态（术语、政策、数据更新快）	· 模型嵌入流程后，前端交互与反馈机制成为关键	· 必须建立脱敏、日志审计、权限控制、加密系统
· 大模型需定期微调，避免“知识滞后”“规则脱节”	· 示例：保险理赔问答助手引入高频纠错优化模板	· 场景如风控、授信、反洗钱，需人机协同审阅
· 需投入：数据工程、提示词设计、安全性回归测试	· 多部门协作（产品/NLP/客服），周期长，工种多	· 成本或高于初期预算模型调用，需要审慎规划
应对知识更新与语义演化	支撑用户体验与业务适配	确保系统安全与监管一致性

为提升运维效率与分摊长期成本，已有金融机构正逐步探索与 AI 厂商、科研机构构建联合实验室或持续共建平台，通过行业知识标签体系共建、调优数据资源共享、模型应用层组件协同开发等方式，既提升了模型对金融语境的深度适应性，也显著降低了单家企业独立承担全部运维任务的边际压力。

值得注意的是，持续运维不仅是技术工作，更是一种组织能力体现。国内部分金融机构正尝试将模型运维能力纳入信息科技部或数据治理部的职责体系中，配套设立“模型使用登记制度”“模型表现评估机制”与“调用行为责任制”，推动从“技术驱动”向“制度保障”过渡。这种方式虽然初期建设工作量大，但有助于解决运维成本归属不清、调用行为不规范、模型责任不明等长期治理痛点。

综上所述，持续运维成本的本质是一种与模型生命周期并行的能力支出结构，决定了金融机构大模型从可部署走向可用、可控与可持续的进阶能力。随着模型能力平台化、部署规范化与治理机制制度化的推进，金融企业应当在战略层面明确持续运维作为一种“长期投入机制”进行资源配置与组织安排，确保 AI 投资能够从“试点试用”过渡到“全面落地”，从而实现模型驱动下的业务稳态提升与智能化演进。

在基础模型的技术周期中，金融企业所扮演的角色远不止“使用者”这么简单。事实上，越来越多的机构在技术与业务高度融合的趋势下，正在逐步演化出“双重身份”——既是模型服务的使用者（User），也是基础能力的开发者（Developer）。这种角色的切换并非仅仅是一种组织功能的变化，更是企业资源配置方式、能力建设路径和长期价值观的体现。

当金融企业以“用户”的视角参与基础模型部署时，其核心诉求往往聚焦于成本控制与业务敏捷性。在此阶段，企业通常尚未具备深度的 AI 研发能力，更倾向于通过外包、平台服务或开源集成等方式快速落地应用场景。这一类企业的典型做法是以“SaaS+ 轻定制”形式完成模型部署，优先选择云服务平台的成熟产品，避免前期高昂的硬件采购和系统搭建成本。

这类“用户型”金融企业所面临的最大挑战是场景适配性。一方面，基础模型往往源于通用语料训练，其输出结果容易“贴标签”但难以“理解业务”，这使得其在精细化金融任务中表现不佳；另一方面，在云服务架构下，企业往往无法触及模型底层结构，因此缺乏对异常行为的解释能力与优化手段。为缓解这一问题，许多机构逐步尝试引入提示工程（Prompt Engineering）作为“业务适配器”，通过对提示词结构的精细调控，引导模型输出更贴合金融语境的答案。然而，当企业不断积累模型调用经验、掌握关键接口与流程后，部分组织将开始转向“开发者”角色。这种转变的核心特征是内部构建 AI 平台、自主管理模型生命周期，并围绕业务痛点开发定制能力。开发者型金融企业往往具备技术团队、数据处理能力和业务抽象能力，能够通过构建“模型中台”“模型仓库”等体系结构，实现模型资产的沉淀与复用。从策略角度看，开发者型金融机构更强调生态整合与场景重构。

此外，由于其具备基础能力构建能力，往往会主动对接金融行业的专属数据工具与系统接口，例如与某金融机构的风控系统联动构建支付异常识别机制，甚至引入政策法规数据库对监管文件进行自动标注与解读。以此为基础，这类企业在模型功能设计上更强调“解释性”与“合规性”，通过数据注释、调用日志、行为溯源等机制建立“AI 治理中台”，确保技术部署不仅能跑、还能管。

值得注意的是，企业从“用户”向“开发者”的过渡并非一蹴而就，也不一定是所有机构的最优路径。对于资源有限、业务模式偏传统的中小机构而言，全流程自研反而可能带来冗余投入与管理复杂性上升的问题。因此，更多企业选择在“双重角色”之间维持一种“策略平衡”：即在非核心场景中扮演用户角色，以平台即服务（PaaS）方式快速落地模型功能；而在核心能力建设上，如反欺诈、信用评估、智能投研等方面，则倾向于自建能力，以形成差异化竞争壁垒。这种混合模式也促进了产业链上下游的“协作—共建”机制。例如，保险机构在智能核保环节中采用商用模型作为语义分析基础，但在保单解释、医疗数据脱敏等模块上依然选择自研方式，以规避数据外泄与模型误判风险。

更进一步来看，双重角色策略不仅是模型部署的成本平衡工具，更是金融企业塑造“数字能力资产”的关键路径。通过持续积累业务场景的模型调用数据、微调日志与反馈闭环，企业能够形成面向不同场景的“模型能力谱系”，从而在未来 AIGC 时代来临时实现快速适配与灵活组合。比如，某大型金融集团通过在多个业务条线中部署文本生成与理解模型，最终构建起一个可统一调用的金融语言模型中枢，既支持人力资源的招聘评估，也支持财务部门的合同审阅，展现出“从技术到平台”的系统跃迁路径。

由此可见，企业越早认识到这一角色演化的路径，越能在成本结构的设计中实现“有的放矢”的部署策略，确保技术投入与业务成果之间形成稳定正向反馈循环。随着越来越多中国金融机构正将基础模型能力建设纳入战略发展规划，购置成本不仅体现企业对新一代技术趋势的响应速度，更代表其在数据能力、算力能力和系统工程能力上的前瞻性布局。

2. 垂类大模型推进行业智能化转型

(1) 垂类大模型破解现实落地之困

金融大模型的产业化进程本质上是技术适配性与行业特殊性相互作用的结果。从底层逻辑看，金融业务对模型的精准性、实时性、合规性要求远高于通用场景，这决定了技术路径必须从“通用能力外溢”转向“场景深度定制”。

早期探索阶段，机构普遍尝试将通用大模型直接套用至金融场景，但很快发现金融数据的异构性与模型输入格式的冲突，决策可解释性需求与“黑箱”特性的矛盾，合规刚性约束与模型迭代灵活性的冲突三重核心矛盾。这种矛盾推动国际厂商与国内金融保险机构技术路径分化。

通用大模型在金融场景落地过程中的三大核心矛盾以及技术路径

三重核心矛盾

1. **金融数据的异构性 VS 模型输入格式的冲突**

金融数据的异构性
- 结构化交易数据
- 非结构化文档
- 实时市场信号

2. **决策可解释性需求 VS “黑箱”特性的矛盾**

3. **合规刚性约束 VS 模型迭代灵活性的冲突**

推动

技术路径分化

国际厂商：倾向于通过金融数据的预训练和垂直领域知识注入来提升模型的专业性。

国内金融保险机构：依托业务场景流量形成数据闭环，以“通用底座 + 场景插件”模式不断优化。

构成国内金融保险大模型生态格局的三类核心参与者——科技巨头、专业厂商与金融机构，正在基于自身禀赋走向各自差异化的发展路径。

金融行业大模型应用生态图谱

这种竞争格局揭示出行业底层逻辑——场景理解深度与生态资源整合能力正在超越参数规模，成为新的竞争维度。

行业专用模型的商业化落地面临结构性挑战，暴露出技术理想与商业现实的鸿沟。数据治理层面，合规要求下的信息脱敏处理与模型性能需求形成根本矛盾。例如，信贷风控模型需解析用户通话录音中的情绪特征，但匿名化处理可能削弱语义理解精度，迫使机构在合规与效率间寻找平衡。技术 – 商业的价值闭环同样执行挑战，模型部署带来的算力成本、运维投入与场景收益难以精确匹配，部分机构陷入“技术超前于业务”的困境。这些挑战折射出行业核心命题——大模型的商业化必须突破“技术可行”到“商业必要”的鸿沟。

(2) 场景牵引下的垂类模型新生态

未来竞争的核心将聚焦于技术架构的深度适配能力与生态系统协同效能。需要特别指出的是，技术发展的战略重心应从追求完全自动化转向系统鲁棒性验证、长期运维成本优化，以及通过渐进式技术升级构建可持续的竞争优势。

多模态技术通过整合文本、图表、语音等异构数据，正在重构金融分析范式：基础数据处理由模型完成，人类专家聚焦战略决策。边缘计算与云端协同的架构在隐私保护与实时性间取得平衡，但复杂场景仍需云端算力支持。监管科技（RegTech）的内生化趋势催生新型技术栈，规则引擎与语义理解的结合虽提升合规效率，但完全自动化仍需长期探索，并且考虑到系统稳定性及投入成本，完全自动化不应是唯一目标。产业分工的重构催生“技术层 – 中间层 – 应用层”新型生态：底层厂商提供算力与基础模型，中间层聚焦场景适配工具开发，应用层聚合垂直领域微调模型。这种解耦推动商业模式从硬件采购转向服务订阅[1]，但对机构的组织适配能力提出更高要求。

从技术架构演进视角看，金融大模型的核心竞争维度已从单一模型性能比拼，升级为数据资产化运营[2]、分布式算力网络[3]与动态合规嵌入三位一体的能力体系。多模态融合改变信息处理逻辑，端云协同重构基础设施，联邦学习突破数据孤岛限制。这些技术演变标志着金融业进入新阶段，大模型不再局限于效率优化工具的定位，而是通过重塑数据处理逻辑、风险控制体系与服务交互模式，成为驱动金融基础设施系统性变革的核心引擎。场景价值的深度挖掘成为商业成功关键，智能投研需打通“数据 – 模型 – 交易”闭环，保险科技依赖跨领域数据整合，支付清算需解决多模态信息处理难题。这些场景的共性在于，模型价值与业务流程重构深度绑定，脱离场景逻辑的技术堆砌难以持续。此外，监管适配与技术发展的动态平衡将长期影响行业进程。长期来看，只有在技术突破、场景深耕、监管适配与组织进化间找到平衡的企业，才能在智能化浪潮中占据主导地位。

1. 这种分层模式推动商业模式从CAPEX主导的硬件采购（一次性购买服务器、软件许可）转向OPEX主导的服务订阅（按调用量付费、年度订阅合约），机构可动态调整资源投入，避免技术迭代的沉没成本

2. 主流金融场景（如风控、投研、核保等）的算法框架（Transformer）已相对成熟，竞争差异化能力在于高质量数据集，算法精准度会因数据集的喂养而得以提升

3. 因为金融业务对低延迟、高并发、强稳定的要求更高，算力部署应从“资源堆砌”升级为“网络化能力”

3. 智能运营从效率提升到流程重构

人工智能自主智能体[4]与工作流[5]技术的发展经历了从人工编排到智能化、自主化的演变。早期，企业主要依赖人工方式设计和管理工作流程，虽然在流程明确性和管理结构上有优势，但在应对快速变化的市场需求时显得灵活性不足。随着大语言模型（LLM）的崛起，AI 增强的工作流工具开始出现，提升了任务处理的智能化水平。然而，这些系统的编排仍需人工干预，未能完全实现自主决策。近年来，第三代自主系统逐步登场，具备动态任务分解、实时路径优化等能力，展现出更高的智能性和商业潜力。

人工智能技术的演进路径正从单点模型优化转向系统化架构迭代，这一转型的底层逻辑源于大型语言模型在产业实践中暴露的核心矛盾：大语言模型基于静态语料训练的知识固化特性，与商业场景对动态实时决策的需求形成。

当金融风险预测需整合突发市场信号，或医疗诊断需同步最新临床指南时，传统大语言模型的响应机制往往会面临可靠性衰减，而检索增强生成（RAG）通过外部知识库的实时调用，本质上是在 AI 系统中重建动态知识更新机制。进一步地，AI Agent 作为协调大语言模型、RAG 及其他模块的操作系统，通过任务分解与多工具调度，实现了从内容生成到问题解决的能力拓展。

在产业落地层面，技术协同正在引发运营流程的渐进式重构。金融领域的实践验证了这一路径：投研类 AI Agent 通过 RAG 抓取实时市场数据（如大宗商品价格波动、地缘政治事件），经 LLM 生成风险信号后，驱动自动化交易系统调整持仓权重。这一闭环将传统投研中“信息采集 – 分析 – 执行”的线性流程加速为近实时响应，其核心价值在于通过动态知识迭代降低模型误判概率，而非单纯压缩时间成本。但是，当前，AI Agent 在各行业的应用日益广泛，但在实际落地过程中仍面临多重挑战。

4. AI Agent

5. Workflow

执行效果的评估与保障

首先，执行效果的评估与保障是一大难题。AI Agent 在封闭环境中测试表现优异，但在复杂多变的实际场景中，执行可靠性存在不确定性。

高风险决策的危险行为管理

其次，高风险决策的危险行为管理也亟待解决。在处理涉及资金转账等高风险操作时，频繁的人工审批可能导致审批疲劳，增加安全隐患。

其他重点关注的问题

此外，默认行为的安全性与用户体验之间的平衡、推理透明性、监控成本控制、责任追溯机制以及紧急停止机制等方面，都是企业在应用 AI Agent 时需要重点关注的问题。

针对上述挑战，业界提出了多种解决方案。例如，通过引入模型推理和评估工作流，结合监督微调（SFT）和低秩适应（LoRA）等技术，以及可视化的多维模型评估工具，企业可以在部署前充分验证 AI Agent 的性能，确保其在复杂多变的实际场景中保持高可靠性。在高风险决策管理方面，利用工作流工具，将复杂任务拆解为多个子任务，明确每个环节的权限和审批流程，减少人工审批疲劳，平衡操作效率与安全性。此外，通过智能体协作，设计合理的交互标准，确保在保障安全的同时提升用户体验，避免因频繁确认导致的服务效率下降。采用可视化的模型评估和提示工程优化工具，展示 AI Agent 的决策逻辑，增强用户对 AI 决策过程的理解和信任，特别是在医疗诊断等敏感领域。灵活配置监控策略，避免过度监控导致的成本增加和误判问题，实现精准且经济的监控。在专有网络中完成模型定制和应用程序开发，结合可定制的内容治理规范和人工干预工具，确保数据安全，便于责任追溯，强化风险管控。设计全局熔断机制，确保在系统故障时能够迅速停止相关操作，防止损失扩大。

展望未来，AI Agent 的发展将会愈加注重高质量数据集的构建和专有模型的微调，以确保其在实际应用中的精确性和适应性。通过结合大模型和特定场景的小模型，AI Agent 能够在执行时保持较高的可控性和准确率。这种集成大规模的基础模型与场景化微调的方式，将使得 AI Agent 在特定行业和任务中的表现更为精确，能够在更加复杂和动态的环境中实现高效决策。同时，基于具体场景的专有模型微调能够使 AI Agent 充分发挥其在特定领域的优势，从而提高行业应用的效率。

随着 AI Agent 在多样化场景中的应用，执行风险的管理和策略也将变得更加重要。随着应用场景的复杂性增加，AI Agent 在处理实时信息、复杂决策和高动态的环境时，仍需具备强大的适应能力和容错机制。在此过程中，技术架构将向更加模块化、灵活的方向发展，以确保能够实时根变化的环境进行优化和调整。

AI Agent 在复杂场景下的能力清进与技术适配路径

原因现状	多样化落地场景 + 场景复杂性提升 + 高动态环境挑战
↓ 要求	
AI Agent	高适应性 + 高容错需求
↓ 推动	
技术架构方向	模块化 + 灵活化架构 + 实时优化适配能力提升

尤其是在金融保险行业等需要快速响应和处理复杂风险场景的领域，AI Agent 将扮演越来越重要的辅助决策角色。例如，金融交易和应急管理等场景中的复杂决策需求，促使了 AI Agent 在知识检索精度与响应速度上的进一步优化，这一进程将推动 RAG（检索增强生成）技术不断突破其现有局限，实现知识获取和处理的时效性和准确性的平衡，满足行业对实时决策支持的高度要求。

案例	金融交易 + 应急管理典型场景
↓ 推动	
AI Agent	优化响应速度 + 检索精度
↓ 进一步推动	
RAG	RAG 技术突破现有局限
↓ 实现	
效果	“实时 + 准确”知识获取和处理，满足行业对实时决策支持的高度要求

这一发展趋势同样促进了向量数据库与边缘计算技术的协同发展。通过将数据处理任务向网络边缘迁移，AI Agent 能够在更接近数据源的位置进行实时分析，从而减少响应时间并降低数据传输的延迟。这种协同不仅提高了处理速度，还在一定程度上降低了对中心化计算资源的依赖，提高了系统的可扩展性和容错能力。同时，随着 AI Agent 技术的不断成熟，LLM（大语言模型）的角色也会发生变化。从最初的知识存储和信息获取，逐渐转向以语义理解和逻辑推理为核心的决策支持角色。这样，AI Agent 将不再仅仅依赖庞大的数据存储，而是通过智能推理为行业决策提供支持，推动了行业决策的自动化和智能化。

与此同时，随着 AI Agent 多模态交互能力的不断增强，其功能边界也有望得到扩展。通过结合文本、语音、图像等多种交互方式，AI Agent 能够更全面地理解用户需求和场景背景，从而提升其处理复杂任务的能力。这种多模态能力的提升，将使得 AI Agent 不仅能处理更多元化的任务，还能够在复杂决策场景中表现得更为灵活和高效。随着这些新功能的不断加入，AI Agent 的应用场景也将不断扩展，不仅仅局限于单一领域或任务，而是能够在更加广泛的行业中提供决策支持，推动行业全方位的智能化转型。

然而，技术融合的进程也带来了新的挑战和潜在风险：

这些挑战的根源，实际上来源于技术组件协同中的“能力代差”。不同技术模块之间的能力差异，可能会导致协作时的效率低下，甚至发生功能上的冲突。为应对这一挑战，企业需要通过动态权限控制、增量学习机制以及强化学习等方案来进行缓解。动态权限控制机制能够确保在数据访问、任务执行等过程中，只有具备相应授权的模块和人员才能参与，从而保证系统的稳定性和合规性。增量学习机制则可以帮助 AI Agent 在长期运行中不断自我优化，逐步提升其应对复杂情境的能力。这些技术手段将为 AI Agent 的落地提供有力保障，并确保其在快速发展的同时，能够保持可控性、准确性和合规性。

总体而言，AI Agent 技术的未来发展将紧密围绕高质量数据集和专有模型微调、技术融合、执行风险管理等多个维度展开。企业在应用 AI Agent 时，必须充分认识到技术的挑战和风险，通过科学的策略和技术手段，推动 AI Agent 的智能化转型，确保其在实际应用中的可持续发展。随着这些技术突破的实现，AI Agent 将不仅成为决策支持的工具，更将成为推动行业创新和发展的核心力量。

02

大模型赋能保险全链，落地有赖行业深度洞察

随着以 ChatGPT 为代表的大语言模型从实验室探索阶段逐步迈向商业应用，金融保险行业对其商业潜力的关注持续升温。然而，截至 2025 年初，尽管行业对其寄予厚望，但要真正实现规模化落地，依然面临一系列关键挑战。

2025 年以前大模型在金融保险行业落地挑战

01

数据安全及合规

我国金融保险机构在应用此类模型时需要严格遵循本土数据安全及合规

02

高成本

传统高性能大语言模型普遍依赖海外高成本 GPU 算力，给中小保险机构造成显著经济压力，且增加了供应链的不确定性风险

03

场景适配

通用型大语言模型缺乏与保险行业场景深度结合的适配性，难以精准满足业务需求，从而限制了技术落地的真正商业价值

2025 年初以来，随着 DeepSeek 和通义千问等国产大语言模型在成本控制与推理性能上的持续突破，大模型的大规模商用逐渐成为现实。这一趋势正促使金融保险机构重新评估数字化转型路径，并探索更广泛、更深入的大语言模型应用场景。

DeepSeek 系列模型和通义千问系列模型

本土化

满足严格的本地化数据安全的需求，实现数据处理与算力资源的自主掌控

低成本

适配国产算力，实现了数据处理与算力资源的自主掌控

高性能

在多项性能评测中表现出色，并在部分指标上接近国际一线品牌的商用版本

基于国产大模型的技术突破，各大保险机构纷纷迅速接入低成本大模型，在前期大语言模型应用试点的基础上，各大保险机构对大语言模型的商用实践也从观望期待变得更加理性和务实，更加倾向于从业务环节中的细节入手进行精准迭代。**目前，大语言模型赋能范围已覆盖保险业务价值全链条及中后台管理的各个环节**，已经解锁且应用落地较为成熟的场景集中在销售辅助、智能商业分析、智能质检、代码运维等多个领域，在提升运营效率的基础上，实现了业务流程的智能升级。

然而，要使大语言模型在金融保险行业真正发挥效能，需要对业务流程进行深入细致的拆解，**这需要同时具备对业务逻辑的深刻理解和对不同大语言模型的技术特性的准确把握**，通过精准定位痛点环节，在不同的场景中进行深入细微环节的大语言模型的适配及替代，实现技术与业务的无缝衔接，从而在特定场景中实现人机协作或智能化替代，提升整体业务效能。在大语言模型实际部署过程中，仍需采取小步试点的方式，确保落地的稳健性。另外，在大语言模型应用落地的过程中，需要谨慎关注并有效防范数据安全、AI幻觉、监管合规等问题，建立健全的治理体系及机制，只有确保安全、合规、稳定运行，保险行业的智能化转型才能真正行稳致远。

一、国产低成本大模型的突破，使其大规模商用成为可能

2022 年 11 月 30 日，随着 GPT-3.5 的爆发以及随后一系列大语言模型的快速涌现，使保险行业在业务运营和办公场景中应用此类技术做好了充分的理论准备，并获得了初步的实践积累。2025 年 1 月，中国人工智能公司深度求索（DeepSeek）开源发布其多种参数模型，以“小模型实现大智慧”的技术以及本地化安全部署等特点引爆行业，解决了大语言模型商业落地中的算力成本痛点，并在一定程度上为数据安全增加了保障，使得中小保险机构本地化部署大语言模型应用成为可能，让大语言模型真正具备规模化商用能力。随后，阿里巴巴推出了通义千问轻量级系列模型，以 QwQ-32B 推理模型为例，该模型以 32 亿参数的精简架构，实现了与 DeepSeek-R1 相当的性能表现，尤其在数学推理、代码生成和问题解决等领域表现优异。通义千问系列模型的设计注重降低部署成本，支持在本地环境中运行，适用于消费级显卡，进一步降低了中小型企业引入人工智能技术的门槛。低成本、高性能的国产大语言模型相继推出，为金融保险机构探索大语言模型的应用价值提供了更优的技术选择。大型金融保险机构进一步深化应用场景的布局，而越来越多的中小型保险机构也开始部署大语言模型，推动其在行业内的广泛普及。

2025 年初以前，国内外主流大语言模型（如 OpenAI 的 GPT 系列、Google DeepMind 系列、文心一言、通义千问的通用大语言模型等）在大规模商用时面临算力成本高、数据安全以及行业适配性不足等诸多挑战。

算力成本高

以 GPT-4 为例，其参数规模可达数千亿乃至万亿级别，在模型的训练和推理过程中需消耗大量高性能 GPU 算力，导致企业应用成本居高不下。

数据安全风险亦构成显著合规隐患

保险行业的数据涉及敏感的个人信息和财务记录，如采用云端的大语言模型服务，则需将数据上传至第三方服务器，存在数据安全及合规性风险。例如，OpenAI 的隐私政策指出，用户输入的数据可能被用于进一步训练和优化模型，若未明确选择退出，则默认视为同意。这对于强调数据保密性和安全性的保险机构而言难以接受。

行业适配不足

通用大语言模型普遍缺乏保险行业的专业知识及术语训练，直接应用时难以精准处理复杂的保单条款及理赔规则。同时，当时主流模型的训练数据多以英文语料为主，对中文应用场景存在一定局限性。尤其是，闭源模型通常限制企业基于自身专有数据进行微调和定制开发，进一步限制了保险机构在模型适配方面的灵活性。

2025 年初，国产大语言模型在低成本高效能上的突破，为上述问题带来新的解决思路。由杭州深度求索人工智能基础技术研究有限公司（DeepSeek Inc.）研发的 DeepSeek 系列模型，以及阿里巴巴发布的通义千问系列模型，凭借开源开放、低成本高性能等优势，为保险机构，尤其是中小型保险机构实现低成本技术升级提供了重要选项。与传统的通用大语言模型相比较，DeepSeek 系列模型以及通义千问系列模型在部署成本和计算资源消耗方面具有显著优势，极大地降低了保险企业应用人工智能技术的门槛，使得包括中小型保险机构在内的各类企业，都能够以可承受的成本引入先进的人工智能技术，推动其业务流程的智能化转型。

1. 国产算力适配确保数据安全合规

DeepSeek 系列模型及通义千问系列模型本地部署与国产算力适配的突破，正在加速金融保险行业对大语言模型的本地化应用，使其在数据安全、算力成本和供应链稳定性上更具可行性。

(1) 完全本地部署保障数据安全及适用性

DeepSeek 系列模型及通义千问系列模型均支持本地部署，这意味着企业可以在自己的服务器或私有云环境中运行模型，从而更好地控制数据安全和模型性能，最大限度地降低数据泄露风险，确保数据隐私和安全符合严格的法规标准。

此外，这些模型还支持高度定制化部署。企业可以根据自身数据特征和业务需求，对模型进行定制训练和优化，从而提升在特定业务场景下的准确性和适用性。同时，借助本地部署方案，企业可以灵活调整计算资源，降低运维成本，进一步提升大语言模型的整体效能。

(2) 适配国产算力降本提升供应链稳定性

在人工智能技术的部署过程中，算力是一个不可忽视的因素。传统的大语言模型通常依赖于高性能的 GPU 服务器，不仅部署和运营成本高昂，而且在国内环境中，还存在供应链不确定性及对外部技术依赖的潜在风险。

DeepSeek 通过自主研发的动态稀疏训练架构和多模态算法，大幅降低了训练和推理能耗。在保证推理精度的基础上，DeepSeek 系列模型将训练能耗降低至同类模型的十分之一，并首次实现了千亿级大语言模型低成本商业化的可行性。此外，阿里巴巴的通义千问系列中多个中小参数量模型在设计之初即考虑到注重推理效率与资源适配，具有较低的显存需求，从而进一步降低了中小型保险企业应用先进大语言模型技术的硬件成本。

随着国产人工智能芯片制造商的快速发展，DeepSeek 和通义千问系列模型等国产大语言模型的算力需求正获得更多本土企业的支持。海光信息、华为昇腾等国产芯片厂商的最新技术已能高效适配这些模型，显著降低对进口高端 GPU 的依赖。国产芯片的引入不仅缓解了算力供应问题，也提升了大语言模型技术的自主可控性，同时降低保险机构应用人工智能的门槛，为行业提供更稳健、可持续的技术支持。

2. 低成本高性能破解行业成本难题

DeepSeek 系列模型及通义千问系列模型国产大语言模型的高性能与低成本突破，使得大语言模型大规模商用成为可能。过去，特别是对于中小型保险机构而言，高昂的算力及部署成本限制了行业智能化发展的广度与深度。然而，DeepSeek-R1 和通义千问系列模型（以 QwQ-32B 模型为例）的推出，凭借媲美国际主流模型的推理能力，以及大幅降低的计算资源消耗，使大模型在保险行业的大规模商用变得可行。同时，借助轻量级架构、国产算力适配及本地化部署方案，这些模型不仅降低了企业的初始投资与运维成本，还赋能保险机构更积极低成本的场景应用探索，从而推动更快速的大语言模型的落地应用。

图 4：DeepSeek 模型性能展示

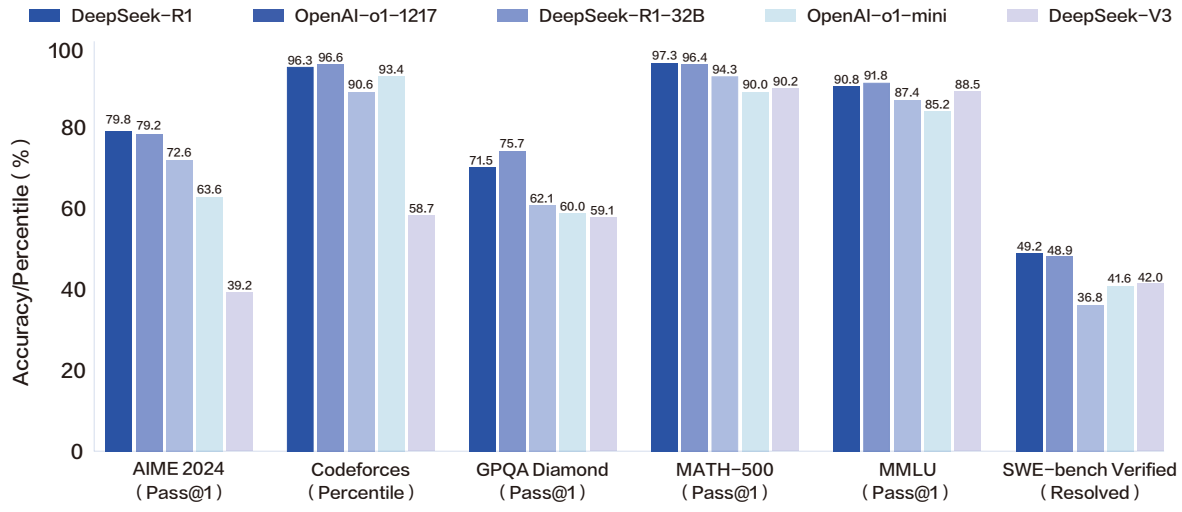


Figure 1 | Benchmark performance of DeepSeek-R1

数据来源：DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [6]

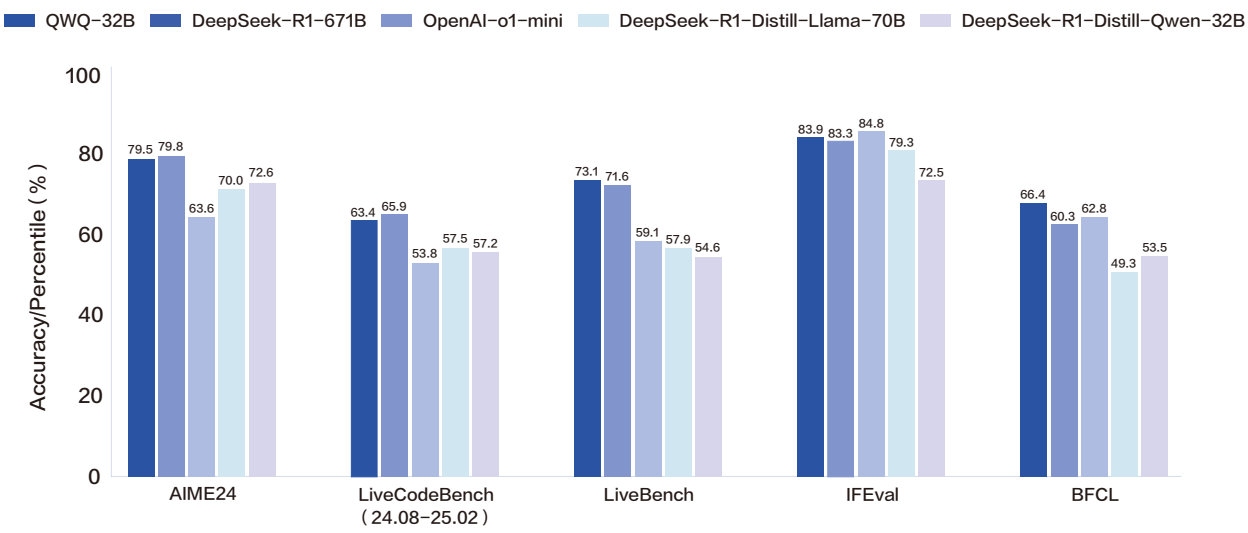
6. https://github.com/DeepSeek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf

(1) 国产模型多项指标比肩国际一线品牌

DeepSeek-R1 在数学、代码、自然语言推理等任务上，性能比肩 OpenAIo1 正式版[7]，而 QwQ-32B 模型在数学、代码及自然语言推理任务中实现了媲美 DeepSeek-R1 满血版的性能。

DeepSeek-R1 在数学、代码和自然语言推理任务中的表现非常出色，其在 AIME2024 测试中 Pass@1 得分达到 79.8%，在 MATH-500 测试中取得 97.3% 的成绩，并在编程任务中以 2029 的 CodeforcesElo 评分远超 96.3% 的参赛模型。此外，在 MMLU、MMLU-Pro 和 GPQADiamond 等知识测评中，DeepSeek-R1 分别获得 90.8%、84.0% 和 71.5% 的分数，创意写作、一般问答、编辑、总结等任务中也有优异表现，整体性能与 OpenAI-o1 正式版不相上下，从而大幅降低了中小型保险企业在技术实现上的成本。

图 5：QwQ-32B 模型性能展示



数据来源：通义官方博客 QwQ-32B:Embracing the Power of Reinforcement Learning[8]

7.https://github.com/DeepSeek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf

8.https://qwenlm.github.io/blog/qwq-32b/

而阿里云通义千问团队推出的QwQ-32B 模型则是一款仅具320亿参数的推理模型，在数学推理、代码生成和通用问答等方面实现了“质的飞跃”，在 AIME24、LiveCodeBench 等权威评测中，其成绩已接近满血版 DeepSeek-R1。此外，通义千问系列模型还集成了智能体（Agent）能力，能够在使用工具时对推理过程进行动态调整，从而显著提升了模型在复杂任务中的灵活性和适应性。

(2) 低成本接入引发险企更积极模型试点

DeepSeek 与通义千问系列模型通过采用创新的架构设计和训练策略（包括强化学习和模型蒸馏），大语言模型部署所需的计算资源、硬件成本与运营费用，为中小型保险企业以较低投资启动大语言模型应用提供了切实可行的路径；同时，这些技术进一步提升了模型的推理能力和业务效率，有力推动了保险行业的智能化转型。

大幅降低硬件部署及使用成本

过去两年，大语言模型的购买和部署成本普遍以数百万元甚至上千万元计，主要原因在于大语言模型对算力和存储资源的高要求，只能由实力雄厚的超大型企业或付费能力强的机构承担。相比之下，DeepSeek 和通义千问系列模型在架构和训练策略上进行了创新优化，大幅降低了算力消耗和使用成本，为保险行业规模化部署大语言模型打开了经济可行性。

在硬件部署的节省方面

DeepSeek 系列模型及通义千问系列模型提供 7B、32B 等轻量级模型选择，及对中低端硬件的适配，大幅降低了硬件部署成本，使中小保险机构也能以相对合理的成本启动大语言模型应用进程。

在使用成本节省方面

DeepSeek 系列模型及通义千问系列模型通过其各自的技术突破，大幅压降了使用成本。其中，DeepSeek-R1 模型 API 调用价格仅为前期同类模型的约 1/10，大幅降低了开发者的使用门槛，北京大学发布的白皮书指出 DeepSeek-R1 每百万 Token 的推理成本低至 0.14 美元，不足行业典型水平的十分之一。而阿里通义团队的两阶段大规模强化学习优化，仅用约 1/20 的参数量，就达到了满血版 DeepSeek-R1 相当的推理水平，推理成本仅为 DeepSeek-R1 的 1/10。

引发更积极的大模型应用尝试

对金融保险行业而言，DeepSeek 以及通义千问系列模型的低成本不仅降低了部署和运维开支，同时也降低了保险机构探索智能化升级的门槛。

更快速的业务创新周期

低成本让中小企业也能有底气进行大语言模型应用的“快速试错”。在保险的新产品开发、精准营销策略或反欺诈检测等高频场景中，企业可以借助 DeepSeek 与以及通义千问系列模型在小范围内尝试不同策略，通过模型迭代来快速发现最优方案，而不必担心“烧钱”或算力不足的问题。

更广泛的应用场景覆盖

部署与算力成本降低后，大语言模型应用可以从原本局限于“大型客户服务系统”或“高端金融风险评估”的高门槛领域，扩展至更多细分环节。例如，自动化的保单信息匹配、理赔文件审核、保险产品的交叉销售策略制定等，这些原本较琐碎且需要人工干预的业务，如今都可能因低成本大语言模型技术获得升级。

更灵活的维护及升级

DeepSeek 和通义千问系列模型本地化部署与蒸馏模型策略令后续的维护更加简单。保险机构可以在内部服务器或国产卡上平稳运行模型，遇到模型更新或业务需求变动时，也可快速完成升级和再训练，无须依赖第三方云服务或海外算力资源。这样有助于企业从长期维度上实现更可控、更经济的大语言模型能力拓展。

DeepSeek 和通义千问系列模型的应用正在降低保险机构引入大语言模型的成本门槛，并优化成本结构。一方面，训练与推理的能耗和硬件开支大幅降低；另一方面，得益于开源和本地部署模式，软件许可与数据使用成本趋近于零。这使保险机构能够将节省的预算投入业务创新，而不再受高昂算力成本的制约。随着低成本大语言模型的持续成熟，保险行业对大模型的投资回报率（ROI）将显著提升，加速全行业的智能化升级进程。

3. 中文语义优化适配保险多种场景

作为高度数据驱动的行业，金融保险领域涉及海量的非结构化数据，包括文本记录、语音交互、影像资料及复杂合同文件等，核心业务环节如承保、理赔、风控及客户服务，都高度依赖数据分析、模式识别和智能决策。非结构化数据的解析是大语言模型的强项，因为这些模型通过在大规模、多样化的文本数据上进行训练，能够捕捉语言的复杂模式和语义关系，从而有效处理和理解非结构化数据，如文本记录、语音交互和影像资料等。然而，要使大语言模型在金融保险等特定行业中成功应用，必须具备高度的行业适配能力。这是因为不同的大语言模型在架构设计、训练数据和优化目标上各有侧重，导致其在处理特定任务时表现出不同的优势和局限性。

因此，选择和应用大语言模型时，必须根据具体业务需求，评估模型的特性与行业需求的匹配程度，以确保其在实际应用中发挥最大效能。

DeepSeek 系列模型与通义千问系列模型在保险机构行业适配性上各具优势，为保险行业提供了更加灵活的智能化解决方案。其中，DeepSeek 系列模型在中文语义理解方面进行了针对性优化，显著减少歧义，特别是在合规审查、理赔风控等对精准度要求极高的场景下，展现出强大的业务适配能力。此外，DeepSeek 具备蒸馏小模型的能力，能够针对不同的业务需求，压缩模型规模，使其更适用于特定场景，提高运行效率的同时降低算力消耗。这使得保险机构可以灵活选择适合自身业务的模型版本，实现高效能与低成本的平衡。而通义千问系列模型则在本地化部署方面具备明显优势，能够无缝嵌入保险机构原有的 IT 系统，降低企业的技术改造成本，确保现有业务流程与 AI 技术的顺畅融合。此外，通义千问系列模型具备智能体（Agent）能力，能够根据实时业务需求动态调整推理过程，在复杂任务处理上展现出更强的适应性。无论是在欺诈检测、风险评估，还是智能理赔、客户服务等高频场景，通义千问系列模型都能通过强化学习优化策略，提高模型在长期业务应用中的稳定性与决策精准度。

当前，保险行业对 DeepSeek 系列模型与通义千问系列模型的应用已呈现出细分化趋势。根据官方发布的参数对比，DeepSeek-V3 定位为通用型大语言模型，旨在满足广泛的商业和研究需求。它在自然语言处理、多语言处理和常规自然语言理解任务中表现出色，追求高性价比。DeepSeek-V3 更适合处理通用的客户服务、市场营销和理赔审核等任务，适用于日常运营中需要高效和准确处理大量客户数据的场景。而 DeepSeek-R1 及通义千问系列模型则定位于推理大模型，更适合处理复杂的推理任务，如风险评估、欺诈检测及复杂案件的决策支持。

二、保险机构快速接入大模型，当前应用聚焦于内部提效

DeepSeek 是国内首个开源的轻量级国产大语言模型，基于其“开源、灵活、低成本、高效能”的技术优势迅速引起广大消费者及各行各业的广泛关注。DeepSeek 开源以来，仅用了 7 天即达成 1 亿用户，创下史上最快纪录，远超其他科技产品。而 ChatGPT 用时 2 个月才达成 1 亿用户。截至 2025 年 2 月 1 日 DeepSeek APP 日活用户突破 3000 万，超过 ChatGPT 目前日活（以 5200 万计算）的一半。据多方舆情分析显示，截至 2025 年 2 月 24 日，阿里云、华为云、腾讯云以及三大运营商等巨头先后接入 DeepSeek，部分地方政府也接入 DeepSeek，拉开本地配置浪潮。随着大语言模型技术的快速发展，保险行业也在积极接入 DeepSeek 系列模型，并进行落地应用尝试，以提升运营效率和智能化水平。

1. 险企加速 AI 中台升级或模型启航

自 2024 年 12 月份 DeepSeek-V3 模型发布，尤其是 2025 年 1 月份 DeepSeek-R1 模型发布以来，保险机构接入进程迅速加快，尤其是 2 月份进入集中接入期。截至 3 月 9 日，据不完全统计，已有 30 家保险机构接入，涵盖各类规模的保险机构，以及集团、财险、寿险、再保险、资管等各个类型的保险机构。整体来看，大型保险机构较早布局，而部分中小型保险机构也迅速积极接入，另有部分中小保险机构密集开展技术调研，确定试点应用场景，并陆续进入招投标采购环节。

保险机构在接入 DeepSeek 时呈现出两类主要模式。部分中大型公司将 DeepSeek 接入原有大语言模型底座，根据 DeepSeek 的优势，通过测试后将其应用于更加适配的生产场景，实现智能化升级；部分中小型保险机构则利用 DeepSeek 低成本、高性能的特性，首次引入大语言模型，在已经相对成熟的应用场景进行尝试，另外，对于新应用场景，也有中小保险机构的策略是先接入，再在实践中探索和推进应用。

2. 提效场景先行客户交互谨慎探索

当前，保险行业对 DeepSeek 的应用仍处于早期落地阶段，保险机构主要将其应用于内部提效场景，而在直接面向客户的交互领域仍保持谨慎探索。

根据公开信息据不完全统计，从整体来看，目前接入 DeepSeek 的保险机构主要将 DeepSeek 应用于内部办公、营销领域以及运营领域，其中内部办公应用场景最多，其次为营销场景，再次为运营场景。在内勤办公场景中，目前已接入的部分公司中，应用主要涉及 8 个细分场景，其中最主要的应用场景为内勤问答，另外也有应用于流程优化、代码运维、文档编辑、数据解析、AI 生成等环节，优化日常工作效率。在营销客服场景中，目前已接入的部分公司中，应用主要涉及 8 个细分场景，其中应用最多的场景销售辅助、客户服务，也有保险机构将 DeepSeek 应用于保单双录等场景，运用 R1 推理模型，为用户提供保险双录操作建议，或者将其集成至 APP，以优化客户使用体验。在运营领域的应用相对较少，主要为 4 个场景，主要应用方向为智能核保、理赔陪练、理赔质

图 6：险企现阶段 DeepSeek 应用盘点

注：颜色深浅代表应用次数，颜色越深应用次数越多

数据来源：公开资料，众安金融科技研究院，数据统计截至 2025 年 3 月上旬

检、产品开发领域。其中，在核保环节，当前使用方法为，基于客户医疗报告的多维度解读，利用 DeepSeek 快速分析个体风险特征，输出核保建议，显著缩短核保周期；在产品开发领域，当前主要用于条款审核，利用大模型重构保险条款，对歧义、缺失条款进行消解或补全；助力提升保险业务的自动化程度。

资管公司也积极接入 DeepSeek 模型，并展开应用探索。其中，大型保险资管公司对 DeepSeek 的应用相对较深入，并已拓展至投研领域。例如，某大型国有保险资管公司已将 DeepSeek 接入多个数字员工场景，包括路演会议、人力资源、公司治理、另类投资等领域，并推出集模型体验与数字员工孵化于一体的智能工坊，构建全方位智能生态体系；中小型保险资管公司仍在初期探索阶段，主要用于智能问答、要素提取、智能检索、数据提取等任务领域，辅助内部员工提高数据分析和投资研究的智能化水平。

业务内容决定了大语言模型尝试的切入方向的差异，保险业务侧重内部运营和销售支持，而保险资管公司则更注重数据挖掘与智能投研。目前，保险机构已经将 DeepSeek 应用于极为细分具体场景，并产生了较好的效果，未来仍有广阔的场景供保险机构进行探索，进一步提升智能化水平。

三、大模型持续迭代细微环节，降本增效实并现智能升级

大语言模型的应用，本质上是企业在深刻理解原有业务领域的基础上，将传统业务流程进行更精细的拆解，并定义其中高度标准化或重复性的场景，通过精准定位痛点环节，引入大模型及其他智能化技术进行自动化替代，不断深入迭代优化流程。一方面，它通过智能化手段提升现有业务流程的效率，使传统流程更加精准、高效；另一方面，它在应用过程中也往往带来意想不到的增量价值。

例如，BI 可视化能力的增强，使数据分析和业务洞察更加直观，进一步贴合实际业务需求，推动企业决策方式的升级。同时，在营销领域，通过大语言模型的赋能，代理人能够随时随地生成个性化营销素材，显著增强营销触达能力。

当前，大语言模型在保险行业应用仍处于初步探索阶段，通过场景化验证、知识库搭建、系统集成的三阶段路径，实现更好的提升营收、降本增效、优化客户体验和强化风险管理的效果。

从大语言模型的场景落地探索路径上来看，保险机构优先选择在容错成本较低、决策干预门槛较小的辅助性业务场景展开，通过低敏感度流程的反复试点，建立模型调试与反馈机制，为后续向高复杂度核心业务的拓展奠定实践基础。

经过对大模型目前在保险领域的典型应用的盘点，其赋能范围已覆盖保险业务价值全链条及中后台管理的各个环节，并正逐步深入更多细分场景，优化各个关键节点。

从保险价值链来看，大语言模型典型的应用场景包括，保险销售、核保出单，客户服务以及理赔办理各个环节，在产品开发领域，大语言模型也有望做出知识库，替代其中部分简单重复性工作；从内部办公研运来看，大语言模型可以赋能决策支持、审计内控、IT 研发及运维管理，办公助手等领域。

图 7：AI 智变重构运营新范式

数据来源：公开资料，众安金融科技研究院

1. 现阶段大模型典型业务应用场景

从保险价值链来看，大语言模型典型的应用场景包括产品开发定价、营销辅助、承保自动化、客户服务、理赔自动化、智能质检等；其中重点应用场景包括智能客服、智能质检、营销助手、智能快赔等。

(1) 产品碎片化保险动态定价及条款审核

目前，大语言模型在产品精算领域的国内应用尚处于探索阶段。已有部分保险机构尝试将其用于产品开发和定价环节，例如在产品规则检验、责任拆分与落库、理赔数据结构化、碎片化产品的动态定价等场景中进行实践探索。但整体来看，其在精算环节的落地仍处于初期试点阶段，未来具备广阔的应用拓展空间。

国外在该领域的实践也值得关注。根据美国精算师协会与英国精算师协会的调研，部分英国与美国的保险机构及研究机构已开始应用大语言模型优化精算流程，提升风险评估、产品定价与市场预测的准确性，从而扩展可承保人群和业务边界。其中，财产保险在应用人工智能技术方面的普及程度显著高于其他险种。

依托大语言模型对非结构数据的分析能力以及推理能力，精算师可以更加便利的分析多模态的客户数据，并根据历史数据进行智能化的情景分析与压力测试，从而更好的研发适合更匹配市场需求的产品，更好的模拟极端情境下的长期风险评估，为保费定价、资本管理提供科学支持。

更好探测客户需求，开发更多需求适配的产品。保险产品同质化问题，不仅困扰着中国保险行业，也让全球保险行业感到力不从心。随着大语言模型的广泛应用，这一局面未来有望迎来转机。借助大语言模型技术，大语言模型可帮助保险机构分析大量非结构化数据，如医疗记录和客户反馈，从而有望提升探测客户的偏好、需求和痛点的效率，为打造更加个性化、差异化、丰富和市场需求的产品奠定了基础。这为设计全新的保险产品或优化现有产品的保障范围、条款内容及费率结构提供了更多可能性与灵活性。

智能化的情景分析与压力测试，为保险定价等提供科学支持。此外，大语言模型还能模拟不同的风险场景，为精算师提供更全面的风险评估支持，并优化定价策略。在保险产品开发过程中，精算师需要评估不同宏观经济条件、市场波动及政策变化对产品的影响。传统上，精算师依赖大量的历史数据进行风险评估和费率设定。然而，当面对那些带有独特风险特征的新型保险产品时，这种方式却常常陷入“无米之炊”的困境。大语言模型通过训练，根据历史数据模拟大量潜在场景，生成反

映现实世界场景的数据，为保费定价、资本管理提供科学支持。例如，美国精算师已经在利用人工智能模拟气候变化的长期影响，而未来的应用可能会更进一步，将更多社会和治理因素纳入风险评估中[9]。

(2) 营销场景提升触达效率精准匹配需求

在营销领域，大语言模型能够分别在线下代理人展业场景以及线上营销场景中，帮助代理人及公司公司精准推荐客户、制定个性化的营销策略，还能优化客户管理流程，从而吸引更多客户，提高成交率，带来保费收入的稳步增长。

代理人展业场景：营销辅助及培训辅助

在代理人展业场景中，大语言模型不仅帮助销售团队从海量客户数据中洞察需求，提供个性化营销内容，还通过模拟真实对话场景提升销售人员的专业技能。此外，借助自然语言理解、多模态生成和计算机视觉等先进技术，大语言模型能够快速生成高质量的营销素材及代理人形象设计。这种智能化解决方案不仅节省了时间和成本，还增强了客户的接受度和信任感，为企业创造了更大的市场价值。

营销辅助：联结客户的“智慧桥梁”

在营销辅助方面，大语言模型可以帮助销售团队从数据中洞察客户需求，并提供创意化的营销支持。它通过分析海量客户数据，包括在线行为、历史交易记录以及兴趣偏好，构建出全面立体的用户画像。这些画像不仅帮助保险机构精准识别目标客户，还能根据历史数据预测客户最可能感兴趣的产品，从而为销售人员提供精准的决策支持。当客户提出疑问时，销售人员可以快速依托大语言模型智能查询相关的保险条款、政策解读或经典案例，并给予快速、清晰的答复。这样的即时响应不仅展现了销售员的专业水平，也增强了客户的信任感，同时大幅提高了营销效率。此外，大语言模型通过分析海量市场数据，优化客户接触策略，帮助保险机构更高效地分配销售资源，让每一份营销预算方案都更有价值。在营销辅助方面，大语言模型可以帮助销售团队从数据中洞察客户需求，并提供创意化的营销支持。它通过分析海量客户数据，包括在线行为、历史交易记录以及兴趣偏好，构建出全面立体的用户画像。这些画像不仅帮助保险机构精准识别目标客户，还能根据历史数据预测客户最可能感兴趣的产品，从而为销售人员提供精准的决策支持。当客户提出疑问时，销售人员可以快速依托大语言模型智能查询相关的保险条款、政策解读或经典案例，并给予快速、清晰的答复。这样的即时响应不仅展现了销售员的专业水平，也增强了客户的信任感，同时大幅提高了营销效率。此

9.https://contingencies.org/generative-ai-applications-for-actuaries/

外，大语言模型通过分析海量市场数据，优化客户接触策略，帮助保险机构更高效地分配销售资源，让每一份营销预算方案都更有价值。

培训辅助：量身定制的“销售训练营”

过去，保险机构的培训模式像是在搭建“一刀切”的流水线，依托既定架构对销售人员进行分级分类，再给各个类别的销售人员培训不同的内容。然而，随着销售队伍日益年轻化，这种“传统大锅饭”式的培训正逐渐失去吸引力。年轻销售员需要的并非单一的理论，而是更贴合实际、更有针对性的技能提升。比如，新招募的业务员常常面临困惑：我到底需要学习什么？公司提供的内容是否对我的发展有帮助？大语言模型等大语言模型通过定制化设计培训课程，精准识别每位业务员的薄弱环节。借助大语言模型，销售人员能够模拟与客户的真实对话场景，例如：面对挑剔客户如何化解疑虑、面对犹豫不决的客户如何推动成交。通过这些互动，业务员不仅能提前练习应答策略，还能更深入地掌握沟通技巧和说服能力，从而在实际销售中如虎添翼。

素材生成：智能生成海报及代理人形象

在传统营销海报制作中，文案策划、设计排版和视觉调整往往需要多个团队协作，流程冗长且效率受限。大语言模型结合自然语言理解、和多模态生成技术，使营销人员仅需输入一句话和一张图，即可快速生成高质量海报，大幅降低创作成本并提升内容多样性。此外，它还能借助人脸识别、姿态估计及风格迁移技术，确保品牌人物形象在不同营销场景中保持一致，增强视觉连贯性。例如，自动生成人物“儿童－青年－老年”不同阶段的形象，使品牌宣传更具沉浸感。相比传统方法需要数天才能完成少量设计，大语言模型可在一天内高效产出 1000+ 张风格统一的营销形象，显著提升生产效率。同时，大语言模型还能自动识别手绘草图，并结合计算机视觉和深度神经网络优化图像结构，仅需 2 分钟即可完成手绘到海报的转换，而传统方法通常耗时 0.5 至 1 个人工日。这一能力提升了创意产出的灵活性，使营销团队能够更快测试不同设计方案，加速市场推广节奏。

另外，借助大语言模型，代理人随时随地拍张照片，即可有效生成专业代理人形象包装。在金融和保险行业，代理人的专业形象直接影响客户的信任感。大语言模型结合计算机视觉和深度学习，能够自动调整代理人的表情、服饰，并快速生成符合市场需求的展业形象，助力品牌打造更专业的营销形象。借助大语言模型，在 5 分钟内，大语言模型可生成一张符合市场需求的专业形象照，而传统人工方式可能需要 2–3 个工作日，成本较高且一致性难以保证。

大语言模型技术，为营销与展业领域带来了全新的生产方式。通过智能海报生成以及智能化代理人形象塑造，极大地提升了内容创作的效率和一致性。未来，随着多模态生成和个性化推荐技术的不断进步，企业的营销策略将更加智能化、精准化，为品牌传播创造更大价值。

私域运营场景：自动化营销线索激活及个性化回复

目前，在私域运营场景中，大语言模型凭借其出色的自然语言处理能力，结合对客户多模态情绪的分析，能够准确理解并快速回应客户需求。

在销售线索激活的环节中，大语言模型也能技术发挥着重要作用。当业务人员通过扫码登录微信或企业微信并导入联系人后，大语言模型会根据预设策略自动加好友。这一过程中，大语言模型通过自然语言处理和机器学习算法，理解并模拟人类的社交行为，使得机器人能够像真人一样与潜在客户建立联系。

同时，大语言模型能够自动与潜在客户展开对话。在互动中，大语言模型可以运用深度学习算法分析对话内容，生成自然且智能的回复，迅速拉近与客户的距离。同时，机器人会记录下每一条有价值的信息，进行线索清洗及会话总结以及学习优化，从而提升线上营销效率，为后续销售跟进提供可靠依据。此外，大语言模型机器人还能保持“永不疲倦”的状态，可以做到 7*24 小时在线服务，秒级回复客户消息。

然后，大语言模型等大语言模型技术拥有实时学习和进化能力。通过不断地与人类互动、收集数据和优化算法，大语言模型能够快速掌握销售技巧，提高获客效率。从自动化内容生成到智能决策，再到自适应优化，大语言模型有望深度整合自动化销售流程，为保险行业开启了全新的营销篇章。

此外，在应对坐席团队长期存在的痛点方面，大语言模型同样展现出价值。传统坐席团队面临人工坐席流动率高、培训周期长等难题，导致坐席人员的专业积累不足，服务水平不稳定。随着大语言模型的应用，以及领域知识不断沉淀，这些问题将会得到有效缓解。

(3) 承保环节承保提速及时风险识别预警

在承保核保出单环节，大语言模型通过其强大的智能分析与深度学习能力，为保险机构提供全方位的辅助核保支持。首先，大语言模型能够自动识别并处理来自多种来源的非结构化信息，如医疗报告、体检记录、客户过往理赔记录等，有效提升数据的整合效率和准确性。随后，大语言模型会根据事先制定的核保规则，将收集到的客户信息与核保要求进行自动匹配，对潜在风险进行初步判断。最后，大语言模型可结合 RPA（机器人流程自动化）技术，快速生成核保建议和出单方案，为人工核保提供精确而高效的决策支持，显著缩短核保流程，提高作业质量与服务效率。例如，车险领域承保环节，对于低风险客户，通过解析图片分析，结合大语言模型知识库，实现自动报价及出单。

(4) 服务领域从被动应答到情绪价值提供

大语言模型在客户服务的能力应用与营销环节基本相同，通过整合语音识别、自然语言处理等前沿技术，为保险机构打造智能客服体系。一方面，针对语音渠道，大语言模型的大语言模型智能语音客服能够在第一时间接听客户来电，识别并理解客户诉求，提供实时高效的解决方案；另一方面，在文字渠道，大语言模型智能文字客服可以通过智能对话引擎，及时为客户解答常见问题，降低人工坐席负担。同时，大语言模型还能针对复杂场景提供大语言模型坐席辅助功能，为客服人员或销售团队提供实时建议与知识库检索，确保回复准确性与专业性。另外，结合人工智能合规检测能力，大语言模型能够在对外沟通的各个环节进行用语规范与合规性监测，有效控制合规风险、提升客户服务质量与效率。

人工客服领域

借助大语言模型的自然语言理解能力，结合深度融合的保险行业知识库、产品知识库及高频 FAQ 数据库，保险企业能够构建高效智能的客服体系，在产品推荐的精准度、复杂场景的处理能力及客户服务体验方面实现全面升级。通过智能客服系统的意图识别与深度理解，企业得以在客户旅程的各个环节提供个性化、精准化的智能支持，使服务从传统的被动响应模式，转变为主动洞察与高效互动。

售前环节

智能客服能够基于多轮会话理解客户需求，通过语义解析、上下文分析及行为数据建模，精准识别用户意图，并自动推送匹配的视频、图文、FAQ 解答及个性化推荐链接。这不仅能够优化客户信息获取的便捷度，还能有效缩短决策路径，提升产品推荐的精准度和用户体验。基于客户画像和实时数据分析，AI 系统能够预测客户可能感兴趣的保险产品或服务，并提供有针对性的营销引导，从而提高投保转化率。

售中环节

AI 系统能够结合实时交互数据快速刻画用户画像，智能预判客户的核心关注点，并主动推送“猜你想问”模块，提供高度契合的个性化服务。与此同时，智能客服能够基于用户行为模式识别不同客户群体的特征，实现差异化服务策略。例如，对高净值客户提供 VIP 专属咨询服务，对高频问询用户优化问题响应机制，确保每位客户都能获得最适合的交互体验。这种精准、高效的服务模式，不仅能提高客户满意度，还能有效增强品牌粘性和用户忠诚度。

售后服务环节

智能客服能够自动关联用户保单，实时判断服务进度，并提供智能化的流程引导。通过 AI 驱动的智能分流系统，客户的咨询能够被精准分类，既能高效处理可自助解决的常见问题，又能确保复杂需求快速转交至人工坐席，减少服务等待时间。同时，用户可以通过智能客服一键查询保单详情、理赔状态及增值权益，实现自助化、无缝化的售后体验。这一能力的增强，使保险服务从“以产品为中心”向“以客户为中心”转型，进一步提升客户的服务体验和企业的运营效率。

在面向人工坐席的管理端，智能客服体系同样发挥着重要作用。系统支持智能会话分区与状态提醒，使人工坐席能够更高效地管理多线程对话，确保高意向客户能够得到及时响应。同时，开放式智能应用接口为坐席人员提供实时 AI 辅助，包括智能话术推荐、最佳应答策略及高效知识检索，显著降低人工操作的复杂性，提升整体服务效率。基于大语言模型的语义分析能力，AI 还可帮助坐席人员动态优化客户沟通策略，使客户交互更加精准高效。

得益于这一智能化升级，某保险机构客服体系实现了全天候、稳定、高效的运营模式。在实际应用中，智能客服不仅显著提升了客户体验，还推动了业务核心指标的优化。7×24 小时的智能服务覆盖，使高意向客户的 7 日内投保率较行业平均水平提升 286%。AI 机器人在自动问题解决方面的表现同样卓越，其问题解决率高达 90%，显著降低了人工客服的工作负担。同时，基于千万级语料库优化的智能客服模型，其准确率达到 95%，确保了智能客服的交互质量。在人机协同模式的支持下，人工坐席的接待访客量大幅提升，整体服务效率得到了显著增强。

未来，随着大语言模型的技术演进，智能客服系统将进一步向深度个性化、语境自适应及多模态交互发展。文本、语音、视频等多种交互模式的融合，将使客服系统更加智能化、人性化。保险企业需要积极拥抱智能化升级趋势，通过不断优化客户服务体验、提升运营效率，构建以客户为中心的智能服务生态，推动行业数字化迈向新的高度。

语音客服领域

电话 AI 语音客服机器人通过主动服务交互，实现精准化的客户关怀，并在智能策略、语音交互、意向判定及可视化经营分析等多个维度深度优化，以提升客户体验和业务价值。在智能策略层面，该系统融合人工经验与 AI 深度学习能力，构建动态调整的策略体系，支持服务关怀、风险减损、产品续保及服务引导等多种策略，以适应不同客户需求及业务场景。

在智能语音交互方面

AI 语音客服依托先进的自然语言处理和语音识别技术，能够精准识别客户意图，并进行多轮动态交互。同时，系统结合多场景人声适配技术，使交互更加自然、高效。在核心技术层面，大语言模型负责构建对话流，包括自动化决策流程及基于 RAG[10] 的知识增强问答机制，使系统能够实时调用企业知识库，为客户提供高准确度的解答和个性化推荐。与此同时，TTS[11] 语音合成技术的深度应用，使 AI 客服能够通过多种风格的音色模拟，增强客户体验。例如，在生日祝福场景下，采用甜美风格的语音增强情感连接；在客户回访时，御姐音色能够提升专业感和信任度；而在灾害提醒场景中，系统选择严肃男声，使信息传达更加权威，增强客户的响应意愿。

在意向判定方面

AI 系统通过语义分析、情绪识别及多轮交互数据，构建客户意图智能判定模型。系统能够对客户交互数据进行深度解析，并结合交互轮次、语气变化等因素，对客户意向进行精准分级，包括高意向度、中意向度、低意向度、拒绝、投诉 / 辱骂及未接通等类别。基于不同意向等级，AI 客服能够智能匹配个性化的后续服务策略。例如，对于高意向客户，可通过人工坐席跟进提升转化率，而对于低意向或拒绝客户，则可调整服务节奏或进入长期培养路径。此能力确保资源配置最优，最大化客户价值。

在可视化经营分析方面

AI 客服系统通过数据驱动的方式，全面监控服务运营情况。系统实时收集并分析呼叫结果、交互数据及用户行为特征，并通过运营监控大盘直观展现关键业务指标，使企业能够精准评估智能客服的运行效果，并基于数据反馈优化服务策略。此外，借助大语言模型的深度赋能，AI 语音客服能够在关键时间节点提供精准服务。例如，在生日、节假日等特殊时刻，通过智能语音进行个性化关怀，以提升客户满意度；在台风、洪灾等自然灾害发生时，向受影响区域客户提供实时提醒，助力风险减损；在增值服务即将到期、理赔结案等关键业务节点，通过智能通知提醒客户，从而提升续保率、增强客户粘性。

10.RAG（Retrieval-Augmented Generation，检索增强生成）是一种结合信息检索（Retrieval）和文本生成（Generation）的自然语言处理（NLP）架构。它通过先从外部知识库或文档集合中检索相关信息，再利用生成式模型（如大语言模型）结合检索到的内容进行回答，从而提升生成文本的准确性、可靠性和知识覆盖范围。RAG 能够弥补传统生成式模型（如 GPT）因训练数据局限性导致的信息缺失问题，使模型具备实时知识获取能力，广泛应用于智能问答、文档摘要、对话系统等场景

11.TTS（Text-to-Speech，文本转语音）是一种语音合成技术，能够将书面文本转换为自然流畅的语音输出。TTS 系统通常包括文本分析、语言处理、语音合成等多个模块，其中涉及音素分割、韵律建模、波形合成等关键技术，以确保生成的语音在语调、节奏和音色上尽可能接近真人发音。现代 TTS 技术广泛采用深度学习方法，如基于神经网络的 WaveNet、Tacotron 2 等模型，以提升语音的自然度和表达能力。TTS 广泛应用于智能语音助手、客服机器人、无障碍辅助（如视障人士语音阅读）、智能导航等领域

整体而言，这一整合性使保险企业不仅能够显著提升客户体验，还能有效降低运营成本，推动服务模式向智能化、个性化和精细化方向加速演进。

(5) 理赔领域实现自动理赔及智能反欺诈

理赔环节的效率和准确性不仅直接关系到客户的满意度，更是企业运营成本控制的重点所在。通过引入大语言模型，保险机构能够在理赔流程的各个环节实现智能化升级，大幅提升处理效率和风险控制能力，从而为客户提供更快捷、更可靠的服务体验。

材料解析环节

材料解析环节不仅是理赔流程中最为耗时的部分，也是最容易出错的一环，其效率和准确性直接影响客户体验和企业成本。在理赔过程中，保险机构需要处理大量的非结构化数据，如医疗费用清单、发票和理赔申请表等。传统手工处理方式耗时且易出错。大语言模型具备强大的自然语言处理能力和多模态理解能力，能够高效处理非结构化数据，通过将光学字符识别技术与自然语言处理能力相结合，大语言模型可以从扫描件、图片或视频中提取关键信息，并将其转化为结构化数据。此外，大语言模型通过对历史数据的学习，可以不断优化对不同模板和格式的适应性，有效提高费用清单处理的准确性和效率。这直接减少了人工干预的需求，降低了错误率和审计成本。例如，某上市保险机构将大语言模型应用于医疗费用清单的解析，成功应对了格式差异和模板多样化的问题，使解析有效性从原来的 24% 提升至 65%，大大提高了费用清单处理的准确性和效率。

理算环节

理算环节作为理赔流程中的重要步骤，不仅决定了赔付金额的准确性，更是考验着保险机构自动化水平的关键指标。理算涉及对医疗费用、赔付比例等复杂因素的计算。大语言模型通过深度学习算法和知识图谱技术，能够自动进行疾病分类和费用核算，减少了对人工操作的依赖。同时，大语言模型基于大规模医疗知识库和疾病命名规范，能够精准识别和分类各种疾病名称，通过语义匹配和上下文理解，模型可以有效解决疾病库不全和名称不规范的问题，进一步提高了自动理算率，加快了理赔处理速度，确保了理算结果的准确性。例如，某上市保险机构利用大语言模型技术自动进行疾病分类，解决了疾病库不全和名称不规范的问题，使疾病分类的有效性达到了 93%，进一步提升了自动理算率，显著提高了理赔处理效率。

风控环节

在风控环节中，如何高效识别潜在的欺诈行为并降低运营风险，是每一个保险机构都在努力攻克的难题。大语言模型通过整合历史理赔数据、医疗知识库和外部公开信息，可以构建一个以疾病为中心的海量知识网络，能够有效识别异常行为模式，发现隐藏的造假骗保行为，并对不合理的医疗风险提供实时预警，这为理赔人员提供了高效的判责依据，提升了运营决策的效率。借助大语言模型技术，理赔人员可以快速判责，降低了欺诈风险和运营成本。例如，某保险机构利用大语言模型技术加强了对医疗风险的识别能力，助力理赔人员高效判责，实现高效率的运营决策。

(6) 质检构建全业务场景的动态风险屏障

在保险行业面临日益复杂的市场环境和客户需求的背景下，传统质检方式在合规管控中暴露出一系列难以忽视的问题：抽检覆盖不足导致风险漏检频发，规则僵化难以有效识别复杂的业务语义场景，以及对新业务规则响应迟缓，无法及时满足监管要求。为有效解决这些痛点，某上市保险机构开发了AI 质检平台，通过创新性地融合大语言模型、ASR 语音识别[12]和智能规则引擎等前沿技术，构建出了一套覆盖全渠道、全场景的智能化质检解决方案。

技术实现层面

AI 质检平台依托高并发、高可用的技术架构，成功支持日均百万级请求量的稳定运行。同时，平台具备强大的多模态处理能力，能够同时处理多种音频与文本数据，实现电话、企业微信、在线客服等多渠道的全量智能质检，真正做到风险的全面覆盖。通过 AI 规则与关键词规则的深度融合，平台不仅能够精准捕捉明确的违规关键词，更能够深层理解复杂的语义表述，从而实现质检精准度与效率的协同提升。此外，通过多租户架构设计，平台能灵活适配不同业务线的质检需求，实现一套系统在多个场景中的高效复用。统一的规则模板管理则进一步确保了合规标准的一致性和管理的高效性。配套完善的可视化分析体系，通过多维度报表与实时监控清晰展现质检结果与趋势变化，为风险管控与业务决策提供精准的数据支撑。

业务应用层面

AI 质检平台通过引入大语言模型显著降低了规则设定的技术门槛，业务人员可直接使用自然语言描述质检规则，无需漫长的技术开发与模型训练周期，即可快速上线实施，大幅提升了对业务变化的

12. ASR（Automatic Speech Recognition）自动语音识别是一项技术，旨在将人类的语音信号转换为计算机可读的文本。这项技术的核心目标是使计算机能够理解和处理人类语言，从而实现人机交互。ASR 技术广泛应用于语音拨号、语音导航、室内设备控制、语音文档检索和听写数据录入等领域

响应速度。同时，大语言模型强大的语义理解能力，使 AI 质检的准确性与人工质检接近，显著提高了隐性风险识别的精准度。

AI 质检体系覆盖保险服务的全流程，实现“检前预防、检中管控、检后追踪”的一体化风险闭环管理。在检前环节，通过敏感词实时拦截与监控，提前预防风险；在检中环节，平台提供覆盖健康险、意外险、旅行险、家财险等超过 200 个险种的智能质检模型，离线 ASR 模型则基于千万级保险行业语料深度优化，全文语音识别准确率达到 98% 以上，关键字识别准确率超过 99%。结合人工质检的抽检模式与标准化评分机制，实现质检过程的精准化与一致化；在检后环节，平台通过实时预警机制，使监管人员能够迅速在企业微信端处理违规事件，并结合私域运营和全流程数据跟踪，实现问题的快速发现与闭环解决。此外，平台还提供趋势分析、交叉分析及原因挖掘等多维智能分析能力，有效赋能业务持续优化。

实践层面

AI 质检平台展现出显著的业务价值

100% 质检覆盖
有效消除风险盲区

45% 人均产能提高
AI 辅助人工的模式

62% 提升
风险检出率

质检带来的监管威慑效应

16% 降低
坐席差错率

32% 下降
客户投诉率

在风险管控方面

100% 覆盖
健康险高风险会话

85% 降低
极高风险差错

100% 覆盖率
私域渠道的高风险规则

16% 降低
总体差错率

92% 覆盖率
客服业务规则

但与此同时，AI 质检平台仍需持续应对技术与业务层面的挑战。例如，在技术方面，如何不断提高大语言模型在保险场景下的精准性和稳定性，以及持续优化平台在高并发情况下的性能与成本之间的平衡；在业务方面，如何实现规则标准化与业务个性化需求的平衡，加强质检结果与实际业务运营的有效闭环管理，仍是团队持续探索的方向。目前，通过深入业务场景分析和技术架构优化，团队正积极应对这些挑战，推动平台能力的持续迭代。

展望未来，AI 质检平台仍具有广阔的发展前景。一方面，平台将向实时质检和预警方向进一步延伸，提升风险防范的及时性；另一方面，将深入挖掘质检数据价值，为保险业务提供更具前瞻性的洞察与决策依据。随着大语言模型和 AI 技术的持续升级，质检领域还将涌现更多创新应用场景，推动保险质检向更高效、更精准、更智能的方向发展。AI 质检平台作为技术与业务深度融合的典型应用，将持续为保险行业的数字化转型与智能化风控提供坚实的保障与动力。

2. 当前大模型典型中后台应用场景

在中后台管理领域，大语言模型已赋能决策支持、审计内控、日常办公以及 IT 研发运维等场景，通过提供智能化、自动化的解决方案，大幅提高工作效率及智能化水平。

(1) 智能数据分析助力高效业务决策

在当今数据驱动的商业环境中，企业对数据决策的依赖程度日益加深，但传统的 BI（商业智能）系统在实际应用中仍面临着操作难度较高、口径繁多等诸多挑战。大语言模型技术的引入，可以优化传统数据分析模式下的诸多痛点。

痛点解决

传统 BI 系统的主要痛点

高操作门槛

- 数据提取及分析依赖人工操作
- 使用者需具备深厚的数据分析经验和行业背景知识

手动归因分析

- 操作人员耗费大量时间进行手动分析以查找问题根因
- 分析效率低，难以实时响应业务变化

大语言模型的优化方案

自然语言交互，降低使用门槛

- 数据查询更加直观易用
- 使用者无需具备专业的数据分析技能，即可获取关键信息

AI 自动归因分析，提升效率

- 利用大语言模型的自动化归因能力，快速定位业务异常的核心影响因素
- 显著缩短从问题发现到原因分析的时间

智能升级

增强数据可视化能力

- 自动生成清晰易读的分析报告与可视化图表
- 应用于市场趋势分析、保单销售情况、理赔风险评估等多种业务场景
- 帮助管理层快速掌握业务动态，制定更精准的战略决策

自学习与优化能力

- 基于用户的查询与反馈，不断优化自身模型
- 提高数据决策的精准度，长期贴合企业需求

AI 自动预警与预测分析

- 提供实时业务监控与异常波动预警
- 提出相调整建议，帮助企业更敏捷地应对市场变化，降低运营风险

优化电脑端 BI 系统，提升智能水平及人群普适度

在电脑端，借助大语言模型自然语言处理以及智能分析能力，使传统 BI 系统的查询、分析和数据呈现方式得到了的优化，让企业能够更加方便的地进行业务诊断、市场研判和战略调整。

凭借大语言模型具备自然语言理解能力，使用无需学习各类终身数据口径知识以及掌握复杂的 BI 工具，只需通过文本或语音方式进行数据查询，实现实时数据提取与决策建议，极大地降低了 BI 工具的使用门槛，提高数据的应用效率。例如，业务人员可以直接输入“最近三个月某事业部的 ROI 是多少？”系统会自动从数据库提取数据，并提供直观的可视化分析。

在数据可视化方面，BI 系统依托大语言模型实现了更直观、更智能的数据呈现。企业管理层可以一键生成分析报告，查看业务增长趋势、销售转化率、客户行为分析等关键指标，并以折线图、柱状图、热力图等形式呈现，使数据驱动的决策更加精准高效。同时，系统还能自动生成智能报告，为高层管理者提供更深入的业务洞察，支持更加科学的战略规划。

大语言模型还增强了 BI 系统的情景分析和预测能力。在数据异常波动或市场趋势变化的情况下，系统能够识别异常，自动发出预警，并结合历史数据进行智能预测。例如，当某产品的市场份额下降时，BI 系统能够自动评估其对整体业务的潜在影响，并建议管理层采取相应的市场应对策略，以降低风险，提高业务稳定性。

移动端集成 ChatBI，助力用户随时随地快速决策

目前，已有保险机构针对在移动端（如钉钉、企业微信）内集成 ChatBI 助手，使用人员可直接通过聊天对话的方式获取日常关键数及分析结果。基于大语言模型的强大自然语言处理能力，移动端 ChatBI 能够通过轻量化交互方式，从海量数据中提取有价值的信息。与传统 BI 系统相比，ChatBI 分析助手的核心价值在于随时随地提供数据支持，ChatBI 不再局限于复杂的 BI 界面，而是通过智能对话交互，让使用在日常工作场景中就能获取数据分析结果，使其能够快速响应业务变化，构建更加灵活的数据决策体系。

从用户需求来看

ChatBI 主要面向管理团队、业务团队和数据团队，满足他们在不同层级上的数据分析需求。

从管理团队来看

从 ChatBI 能够实时提供市场趋势分析、业务增长预测、销售渠道表现评估等核心数据，帮助决策层快速了解整体业务动态。例如，管理人员可以在钉钉、企业微信等移动端工具上直接输入“2 月新单业绩是多少？”或“某渠道的转化率为何下降？”ChatBI 能够迅速解析查询意图，自动检索系统数据，并生成可视化报告，为业务决策提供有力支持。

从业务团队来看

ChatBI 能够提供销售数据、客户行为分析、市场需求预测，使销售人员能够精准调整策略，提高转化率。

从数据团队来看

ChatBI 则提供更灵活的数据查询能力，帮助分析师在日常工作中更高效地完成数据整理和业务洞察。

ChatBI 的智能化数据分析能力，使其成为企业数字化转型的重要工具，助力企业从传统的数据分析模式向智能化决策升级。未来，随着人工智能技术应用的不断加深，ChatBI 将在保险行业的精准营销、智能核保、风险管理等多个领域发挥更大价值，帮助企业建立更加高效的数据驱动型管理体系。

(2) 审计内控场景实现智能流程优化

在审计内控环节，大语言模型的智能化能力能够帮助保险机构大幅提升合规审查与风险管理的效率与准确性。

01

首先，大语言模型可针对内部制度合规性进行批量审查，自动识别并标注可能存在违反合规要求的问题点，为审计人员和内控团队提供精确的风险排查指引

02

其次，大语言模型具备对产品条款与相关材料进行深度解析的能力，能够快速完成条款审核并生成审计报告，让审计工作更加高效和透明

03

最后，通过对政策文件的深入理解与解读，大语言模型可以生成审计建议，帮助企业及时优化内部流程与管理制度，有效降低合规风险

这样一来，大语言模型不仅为审计内控带来了流程上的自动化与智能化，也为企业的合规运营和风险防控提供了更有力的技术支撑。

(3) 办公场景智能升级驱动效率提升

在日常办公领域，大语言模型已实现较为成熟的落地应用。将会有越来越多的公司将大语言模型嵌入公司内部系统，提供给公司内部员工适用，提升各部门的日常办公效率及智能化水平。其中，在办公提效与自动化文档处理方面，员工依托该智能中台的文案助手，帮助自动生成图文，大大缩短文案撰写时间；还可以利用个人大语言模型助理进行日程管理、会议纪要整理、文档自动生成等办公任务；在审计工作中，通过嵌入处罚知识库，内勤可以快速准确查找相关处罚信息，助力公司各项工作开展的合规性。

除了辅助员工日常办公外，还可以通过数字员工的方式，让大语言模型协助完成特定的办公任务。数字员工是通过大语言模型技术驱动的虚拟员工，专注于处理重复性和流程化的工作，执行任务高效、精准，同时，它借助大语言模型技术，突破预设规则限制，拥有更丰富的知识和深度分析能力。

(4) 研发运维场景优化软开流程效率

在研发与运维场景中，大语言模型在优化软件开发流程和提升开发效率方面，堪称“效率革命”的典范。代码助手工具可以实时解决复杂问题、生成代码片段，还能进行代码调试与优化，极大地减轻了开发人员的工作负担。

3. 小步试点借力经验实现稳健落地

尽管在保险行业部署大语言模型，可以大幅提升业务效率与客户体验。但在由于落地过程有一定的困难及复杂度，需要在谨慎的基础上小步试点，不断迭代，或者借鉴先行公司成熟的经验，确保不出现系统性风险。同时，为确保大模型应用的安全可控，还需建立完善的治理规则体系，有效防范潜在的安全风险和幻觉问题，确保模型的稳定性和可靠性。

(1) 深谙业务逻辑精准破解场景痛点

在大语言模型场景落地应用方面，如何使模型更加贴合企业特定领域的数据和业务场景，也是在模型接入阶段需着重解决的问题。由于大语言模型需要在深刻理解原有业务领域的基础上，将传统业务流程进行更精细的拆解，并定义其中高度标准化或重复性的场景，找出关键痛点，引入大模型及

其他智能化技术进行自动化替代，因此，大模型的场景应用需要同时基于对业务逻辑的深刻理解以及对不同大语言模型的的特点的深刻理解。在这种背景下，只有采取 " 技术赋能 + 业务深耕 " 的双轮驱动模式，业务部门牵头制定统一规则，确保特定细分流程标准一致性，技术团队持续优化大模型应用效果、提升识别准确率、增强规则引擎性能，才有可能将大语言模型与企业已有业务场景有效融合，以及实现对现有业务系统、业务流程进行赋能。此外，企业还需考虑大语言模型的使用限制与风险，以确保在实际场景融合研发过程中的稳定性和可行性。这一系列的挑战，都要求企业在实践中进行深入研究、精确分析和创新应用，以实现大语言模型技术在各个层面上的最优整合和应用，推动企业业务的持续创新和发展。

鉴于大语言模型技术在保险业务场景融合应用方面的多重挑战，如何确保技术的稳健落地并实现持续优化成为关键。在此背景下，采用分阶段的实施策略至关重要。通过“小步试点”策略、人工复核、数据本地化部署以及持续优化等措施，确保大语言模型能够在保障合规性、安全性和业务效率的同时，顺利推进其在保险领域的全面应用。

小步试点，逐步推进，稳健落地

在正式大规模应用大语言模型前，选取典型的业务场景中的可以嵌入大语言模型能力的具体细分环节试点，并事前在较为可控且风险容忍度相对较高的场景中验证技术与流程，可快速迭代、不断优化技术模型和业务流程。同时，在试点过程中，建立实时的监测机制，跟踪大语言模型对业务的影响和潜在风险，并在必要时进行及时调整，确保风险敞口可控。

大语言模型初筛与人工复核并行，确保合法合规

在业务环节中先由大语言模型进行初步筛查或决策（如核保打分、理赔风险识别、营销推荐），快速处理大量重复性任务，提高效率。对大语言模型输出的关键结果，如核保建议、理赔审核、营销目标客户等，由专业人员进行二次审查，结合经验判断与合规要求做必要修正。同时，人工审核也能为模型提供持续的反馈数据，帮助模型自我优化。

关键数据本地化部署

关键业务数据需要本地化存储，并建立多重备份与容灾机制，预防服务器故障或网络攻击导致的数据丢失。定期组织网络安全和系统安全审计，关注系统漏洞、访问权限、日志记录等，确保对潜在威胁做到及时预警和处理。

持续迭代与优化

大语言模型并非“一劳永逸”。在部署后，需要在业务环境中持续收集新数据，不断进行训练、校准与迭代。同时，还要结合市场变化、监管政策更新，定期对模型算法进行检视和合规性验证。

完善应急预案

对大语言模型系统故障、错误输出或违规事件，需制定明确的应急响应机制，包括问题定位、系统回退、数据修复、用户告知、责任划分及损失赔付等内容，确保在事发后能够迅速应对。

随着对大语言模型的不断升级迭代，以及保险机构对大语言模型应用的安全有序探索，未来将会有更多的险企接入大语言模型，并有望解锁不同的业务场景及办公链条中各环节的更多的小模块，从而赋能保险价值链各链条中更多场景以及内部办公研运的各个环节，持续强化智能化能力。

(2) 完善治理机制构建可控模型体系

虽然借助大语言模型强大的自然语言处理、自动化分析能力和智能决策支持能力，在科学妥善使用的情况下，可以极大地提升业务效率和客户体验。然而，任何技术的变革都伴随着风险与挑战，大语言模型也不例外。在保险行业，尤其是涉及数据敏感度高、决策影响深远的业务环节，大语言模型的应用需要谨慎对待。其中，数据安全与合规治理以及 AI 幻觉是两大核心挑战。大语言模型需要处理大量的用户信息、交易数据和企业机密数据，而在数据存储、传输及使用过程中，可能面临数据泄露、未经授权访问、模型篡改等安全风险。此外，由于训练数据的固有局限性和模型生成方式的不可控性，大语言模型可能会生成虚假信息、逻辑自相矛盾的回答或事实错误的结论，这在保险核保、理赔决策、风险评估等环节可能带来重大合规和运营风险。应对这些挑战，需要从技术、管理等多维度入手。在数据安全方面，企业需要加强数据加密、访问控制、模型防篡改、算法公平性审查等措施，以保障数据安全性和合规性。在 AI 幻觉治理方面，业界正在探索外部知识检索（例如 RAG 等）、事实一致性验证、优化生成策略等手段，以提升大语言模型的可靠性。此外，金融机构可采用分级管理方式，根据不同应用场景的风险级别，采取相应的管控措施，确保高风险场景中的模型应用具备充分的透明度、可解释性和人工监督机制。

数据安全与合规的治理

大语言模型是一种基于人工智能技术的深度学习搜索引擎或数据挖掘工具。它可以通过深度学习和自然语言处理等技术，帮助用户进行信息检索、数据分析和预测等操作。然而，像所有涉及敏感数据处理和人工智能技术的应用一样，大语言模型也面临一定的数据安全隐患。

数据泄露和隐私侵犯

大语言模型需要处理大量的用户数据和敏感信息，如搜索历史、个人偏好、甚至在某些情况下，可能涉及公司机密数据。若这些数据存储、传输或处理不当，可能会导致数据泄露，尤其是如果没有足够的加密措施，敏感信息可能会被黑客窃取。

未经授权的访问和数据滥用

深度学习模型可能存储和处理大量的私人数据。如果没有严格的身份验证和访问控制机制，黑客或恶意用户可能会获得对数据的访问权限，滥用这些信息用于诈骗、网络攻击或非法行为。

数据篡改

如果大语言模型的算法和模型受到攻击（例如，通过输入伪造的数据进行训练），攻击者可能操控数据检索的结果或模型预测，导致不准确的信息传播，甚至引发大规模的误导或诈骗行为。

模型反向工程与数据泄漏

深度学习模型本身具有较高的复杂性，反向工程攻击者可能通过分析模型输出的行为来推断训练数据，进而获得未经授权的数据。

算法偏见和不公平性

深度学习模型可能因训练数据的偏见或不全面性，而导致生成不公正的搜索结果或数据分析。比如，模型可能强化某些群体的观点而忽略其他群体的需求，这可能引发歧视和不公平现象。

虚假信息生成

大语言模型等系统有时可以生成相关内容推荐和搜索结果，若不加以控制，恶意用户可以利用这些推荐系统传播虚假信息或进行广告刷量等操控行为。

为应对以上的数据安全问题，可采取数据加密，安全审计建立模型防篡改机制，进行算法透明性与公平性审查，进行反作弊与内容验证，应对举措：

数据加密与访问管控

通过数据加密与隐私保护，对所有敏感数据进行加密存储和传输，确保数据在传输过程中不被窃取。同时，加强数据访问控制，使用多层次的身份验证机制确保只有授权用户可以访问敏感数据

审计监控双保障

定期对系统进行安全审计，确保没有漏洞或不合规的行为。建立实时监控机制，及时检测和应对潜在的安全威胁，如未经授权的访问尝试或异常数据请求

建立模型防篡改机制

采用模型加密和防篡改技术，确保模型的完整性。可以使用数字签名和校验机制来确认模型和数据没有被篡改

进行算法透明性与公平性审查

引入透明的算法审查和测试机制，定期评估模型的公平性和偏见问题。对训练数据进行去偏见处理，并尽可能使用多样化的训练数据，以确保生成的结果公正无偏

进行反作弊与内容验证

为搜索结果和数据分析结果建立反作弊系统，利用大语言模型检测并防止虚假信息的传播。可以通过技术手段识别和屏蔽虚假的内容、数据源或广告

搭建法律合规与伦理框架

确保大语言模型应用符合数据保护和隐私法规（如 GDPR、CCPA 等），制定严格的伦理标准，确保模型和数据的使用不会侵犯个人隐私或产生社会不公平

通过这些应对举措，大语言模型可以有效降低数据安全隐患，提升系统的可靠性和用户的信任度。

对 AI 幻觉的应对治理

AI 幻觉是指人工智能模型，特别是大语言模型，在没有足够或正确的训练数据支持的情况下，生成看似合理但实际错误的内容。一般而言，大语言模型幻觉可以分为三种类型：

输入冲突幻觉

输入冲突幻觉是指生成的内容与用户提供的输入不一致

上下文冲突幻觉

上下文冲突幻觉则是指生成的内容与之前生成的信息相互矛盾

事实冲突幻觉

事实冲突幻觉则发生在生成的内容与已知的世界知识不符时

这三种幻觉类型体现了大语言模型生成内容时可能存在的不同层次的错误。而理解这一现象至关重要，倘若大语言模型被广泛应用于数据分析、决策支持、客户服务等领域，金融机构需要对 AI 幻觉做出相应治理。

从幻觉的来源来看，一般分为数据、训练以及生成及推理：

幻觉来自训练数据

大语言模型的知识和能力主要来源于预训练数据，如果使用了不完整或过期的数据，就容易导致知识错误，从而引发幻觉现象。此外，大语言模型往往会捕捉到虚假的相关性，尤其在回忆长尾信息和进行复杂推理时表现出困难，这进一步加剧了幻觉的出现。

幻觉来自训练

在预训练阶段，大语言模型通过学习通用表征并捕捉广泛的知识，通常采用基于 transformer 的架构，在庞大的语料库中进行因果语言建模。然而，固有的架构设计和研究人员所采用的特定训练策略，可能会引发与幻觉相关的问题。在对齐阶段，通常通过监督微调和从人类反馈中强化学习来提升模型表现。尽管对齐过程显著提高了大语言模型的响应质量，但也带来了幻觉产生的风险，主要表现为能力不对齐和信念不对齐。

幻觉来自生成或推理

经过预训练和对齐后，解码在体现大语言模型能力方面发挥着重要作用。然而，解码策略的某些缺陷可能导致大语言模型出现幻觉。解码过程中的潜在原因可以从两个关键因素深入探讨。首先，解码策略固有的随机性，例如采用采样生成策略（如 top-p 和 top-k）所引入的随机性，可能会导致幻觉的产生。其次，不完善的解码表示也是一个重要原因。在解码阶段，大语言模型使用顶层表示法预测下一个标记，但顶层表示法存在局限性，主要体现在上下文关注不足和 Softmax 瓶颈这两个方面。

为此，检测大语言模型中的幻觉对于确保生成内容的可靠性和可信度至关重要。传统的衡量标准主要依赖于词语重叠，无法区分可信内容和幻觉内容之间的细微差别。这样的挑战凸显了为大语言模型幻觉量身定制更复杂的检测方法的必要性。

目前监测方法主要分为“检索外部事实”和“不确定性估计”两种策略。

检索外部事实

该方法是一种直观的策略，旨在将模型生成的内容与可靠的知识来源进行比较，从而有效地指出大语言模型输出中的事实不准确之处。尽管许多幻觉检测方法依赖外部知识源进行事实检查，但也有一些方法可以在零资源环境下解决这一问题，无需检索。这些策略的基本前提是，大语言模型幻觉的起源本质上与模型的不确定性有关。

不确定性估计

不确定性估计的方法大致可分为两种：基于内部状态和大语言模型行为。前者的前提是可以访问模型的内部状态，而后者适用于更受限制的环境，仅利用模型的可观测行为来推断其潜在的不确定性。

忠实性幻觉[13]的检测方法：该方法主要侧重于确保生成的内容与给定上下文保持一致，从而避免无关或矛盾输出的潜在隐患。

13. 忠实性幻觉（faithfulness illusion）指大语言模型在回答或生成文本时，表面上看起来逻辑清晰、语言流畅，但却与用户的指令或上下文不一致的一种现象

常见的检测方法包括基于事实度量、基于分类器的度量、基于 QA 的度量方法、不确定性估计以及基于 prompt 的度量方法。

基于事实度量

基于事实度量通过检测生成内容与源内容之间的事实重叠度来评估忠实度

基于分类器的度量

基于分类器的度量则利用经过训练的分类器来区分生成内容与源内容之间的关联程度

基于 QA 的度量

基于 QA 的度量方法通过问题解答系统验证源内容与生成内容之间的信息一致性

不确定性估计

不确定性估计则通过测量模型对其生成输出的置信度来评估忠实度

基于 prompt 的度量

基于 prompt 的度量方法则让大语言模型充当评估者，通过特定的 prompt 策略来评估生成内容的忠实度

检测大模型幻觉为后续解决方案打下了基础，当前，对于解决大语言模型幻觉已有多种策略来提高生成模型在自然语言处理任务中的准确性和可靠性。一种常见方法是结合检索和生成，使用像 RAG（Retrieval-Augmented Generation）这样的模型，在生成过程中结合检索相关文档或信息，从而减少幻觉现象。此外，模型可解释性和后处理也是重要手段，通过分析生成模型的输出并使用可解释性工具、规则引擎或其他后处理方法来识别和纠正潜在的幻觉问题。模型融合与集成则通过将多个生成模型的输出进行融合或集成，提升生成文本的准确性，常见的方法包括投票、加权平均等。优化生成策略，如使用束搜索（beamsearch）、拓扑抽样（top-ksampling）或核心抽样（nucleussampling），有助于平衡生成文本的多样性和准确性。在预训练和微调过程中，使用更高质量和更具代表性的数据集，以及更多有标签的数据或强化学习方法来提高模型性能，微调阶段则需使用与目标任务更相关的数据集，以便模型更好地适应特定场景。提示词工程也可以通过使用更好的提示词进行正确性引导，而增加多样性则能通过引入随机性或多样性，避免模型过度依赖特定信息，从而提升生成内容的质量。

除了在工程上策略应对之外，对于金融机构的应用场景进行分级管理也是一种降低大语言模型应用风险的一种方式。例如：高风险大语言模型应用，诸如用于人寿、健康保险的核保与定价系统，可能对个人的财务和社会公平性产生较大影响，因此需要实施严格的风险管理措施，如确保数据质量、降

低算法偏见、提供透明的决策依据、以及建立人工监督机制。此外，系统还需要具备防止数据操纵和保证准确性的安全性要求，以确保决策过程的可信度和公正性。

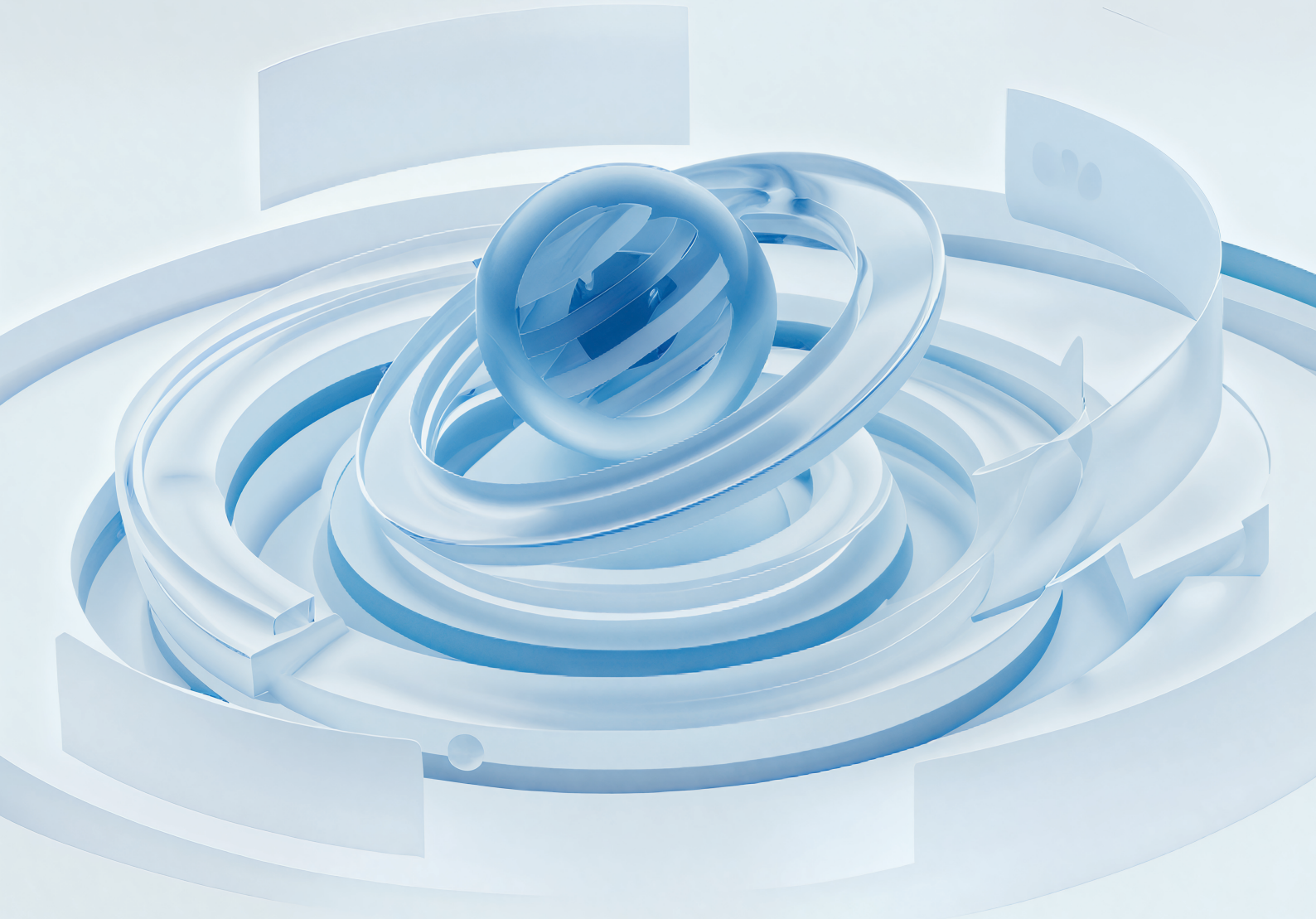
对于有限风险大语言模型应用，例如客服聊天机器人或部分自动化理赔系统，虽然这些系统对用户权益的影响相对较小，但仍需确保一定程度的透明度和可解释性。用户应当明确知道自己正在与大语言模型互动，同时大语言模型的决策逻辑应当简洁明了，用户能理解拒赔等关键决策的依据。这一类别的应用，虽然风险较低，但依然要求提高决策过程的透明度，以降低金融消费者对大语言模型的疑虑。

最低风险的大语言模型应用则主要是一些对用户权益影响极小的工具，如保险机构内部的数据分析或文档分类系统。对于这一类应用，主要鼓励企业根据自身情况制定伦理准则，并提升员工的大语言模型应用素养，以保证大语言模型技术的合理使用和管理。对这些应用的管理更多的是基于行业自律和企业责任。

总的来说，分级管理的核心理念是根据不同大语言模型应用的风险程度，采取差异化的应对及治理措施，既保障了消费者的权益，又为大语言模型技术的发展提供了必要的空间。

03

合作范式的系统演变，从单边集成到机制协同



当前，大模型在金融保险行业的商业化路径正逐步围绕“精准度门槛”这一核心临界点展开: 只有当模型在关键业务任务中的预测误差率降至业务可接受水平，模型输出才具备复用性与部署价值。精准度不仅代表模型是否“有效”，更决定其是否“可靠”，从而成为技术可用性向商业可行性转化的拐点。支撑这一跃迁的，是“算力—算法—数据”三位一体的支撑体系：算力与算法为能力生成提供边界框架，而真正决定大模型应用的商业表现的，是高质量、合规且具备业务语义深度的数据资产。

> 特别在当前算法开源化、算力商品化趋势加剧的背景下，高质量数据集的构建能力，已成为金融机构能否真正掌握智能能力的核心分水岭。

具体而言，在算力层面，为了满足金融保险行业对算力可及性、性价比与合规性的多重要求，部分云厂商正与大模型厂商展开协作，推出包括按量计费、包月套餐、推理实例预留等多样化的算力服务形式；在算法层面，当前主流模型结构与微调路径在开源生态的推动下已趋于透明，技术共享与复制门槛相对降低；与算法相比，数据作为智能能力构建的根基，具有获取门槛高、合规约束严、工程落地复杂等多重挑战。在金融保险这类高度合规、精度要求高的行业，数据壁垒不仅决定了大模型微调的起点，更决定了大模型应用可否突破“商业部署到规模应用”的能力门槛。

一、数据要素价值加速显性化，倒逼从技术到系统化重构

当前金融保险行业在构建高质量数据集进程中，正经历从局部效率优化到系统性协同的范式跃迁。**这一转型的核心矛盾在于这一资产化过程并非简单的数据积累或技术升级，而是对数据治理体系、组织能力结构与跨机构协作机制的系统重构。**中小机构因缺乏稳定的数据获取通道和高质量标注体系，在短期内形成具有迁移价值的数据资产存在挑战，而头部机构虽具备平台能力，却在独立解决数据碎片化、高频低值与语义漂移等结构性问题上并不经济。这一“局部能力冗余与系统协同错配”并存的格局，反映出行业在高质量数据集构建过程中面临的核心矛盾已从“算力瓶颈”转向“数据要素资源分配不均”。正是在这一背景下，在我国政府的国家政策及制度指引之下，金融保险行业正逐步探索出三类具有现实可行性与制度嵌入性的协同机制路径，通过重塑数据生产与流通关系，突破智能化转型的关键约束。

横向协同机制

横向协同机制以跨平台、跨生态的拼图式连接为核心，特别适用于客户认知建模、风险信号补全与跨场景推荐等需要多维输入的大模型应用场景。典型路径包括保险机构与电商、出行、社交平台的深度集成，企业通过标准化接口接入经用户授权、并经合法处理的业务相关特征信息，用于构建场景语义感知模型、用户偏好预测模型及多轮交互式智能客服系统。以某头部保险科技平台为例，其在合规数据接口框架下训练了适用于碎片化场景的反欺诈推理引擎和用户语义识别模型，实现了对大模型通用能力的场景化重构。该模式的价值在于构建了以用户为中心、以任务为驱动的动态数据拼图体系，降低了模型开发门槛，提升了智能部署效率。

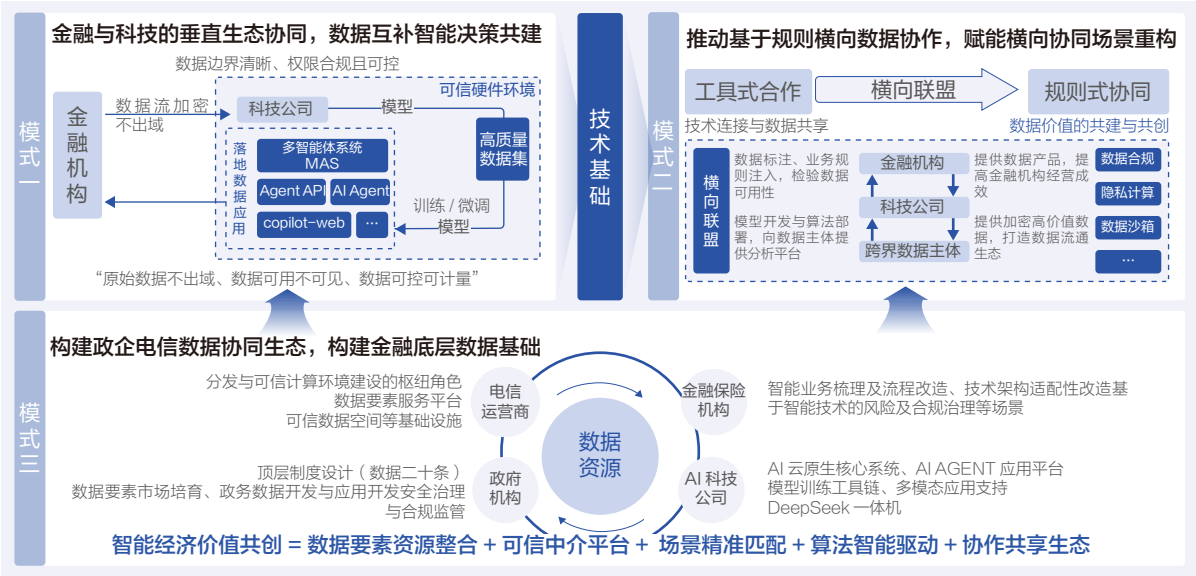
垂直协同机制

垂直协同机制则以内生系统的一体化协同为特征，主要服务于集团企业、产业链条或业务纵深场景下的大模型落地需求。通过打通内部数据中台，实现多业务板块的标签体系统一与语义标准对齐，企业不仅能够形成以客户为主线的统一数据视图，还可据此训练具备跨任务泛化能力的垂类模型。如某金融集团通过统一客户标识体系整合信贷、理财、保障等多个产品模块的数据资产，形成用于风险预测、资产推荐与服务分层的大模型微调语料集，显著提升了模型复用效率与客户运营精度。此类机制的本质，是通过垂向整合打破“数据孤岛”，实现数据资产向模型资源的高效转化。

政企协同机制

政企协同机制则在高合规性与公共数据价值释放之间建立制度桥梁，成为金融保险机构构建可信数据基础设施的重要抓手。通过对接医疗医保系统、征信体系、交通监管与司法平台，企业可在合规授权下获得结构化、权威性强的数据资源，用于增强大模型在理赔、风控、授信等高责任场景中的决策鲁棒性。通过制度化对接公共信用、医疗支付、交通监管与司法等领域的数据资源，企业在合规授权的框架下，得以构建结构化、权威性强的外部数据输入体系，提升大模型在高责任业务场景中的应用稳健性与风险判断能力。此类机制依托可信数据中介平台、标准化数据授权协议与可验证接口设计，为 AI 能力嵌入关键金融基础设施提供了制度保障。

图 8：金融保险业智能化路径：垂直、横向与生态合作的协同模式



数据来源：公开资料，众安金融科技研究院

从横向协同的生态嵌入，到垂直协同的数据贯通，再到政企协同的制度植入，三类机制共同构成了智能化系统中数据要素获取、治理与应用的多维协同体系。它们不仅回应了当前企业在大模型部署中面临的数据边界问题，更引导企业从“自有数据管理”走向“数据协作共建”的新范式。对于金融保险机构而言，这一协同体系不仅是模型效果优化的能力支点，更是构建数据壁垒、形成生态优势、提升行业话语权的战略基础。

二、垂直横向及生态数据协同，构建全行业共享智能底座

如上所述，在当前金融保险行业智能化转型持续深化、数据要素价值加速显性化的背景下，行业智能化演进已不再是局部技术能力的叠加过程，而是关系到资源重组、组织结构与制度安排的系统性重构。特别是在数据资产化路径上，单一机构孤立应对数据构建、治理与应用问题的边际收益正在递减，行业逐步共识的方向是——通过生态化方式推动数据协同体系建设。

在这一趋势下，智能技术商业生态构建不仅是一种技术架构选择，更成为推动行业转型、重塑合作范式与增强系统韧性的战略路径。回顾行业内已有实践，可以发现生态构建主要沿三条路径展开：其一是面向制度共建与能力共享的赋能共创，政府、金融机构、大模型厂商与电信运营商共同打造数据驱动的基础设施体系；其二是以业务与算法深度融合为特征的垂直整合，重塑机构内部数据与技术的联动结构；其三则是以多主体联动为基础的横向扩展，解决跨机构数据要素高效流通与规则协同的问题。这三种路径协同作用，共同构成了中国金融保险行业智能生态系统的战略雏形。

1. 政企协同：推动数据要素流通新路径

自“数据二十条”政策出台以来，我国在数据要素流通和交易方面的顶层设计逐步清晰，明确提出“要构建数据基础制度，健全数据、流通交易、收益分配和安全治理机制”，并通过国家数据局的设立，统筹推进全国范围内的数据资源整合、标准制定与跨部门、跨行业的数据要素市场建设。其根本目的在于打破外部“数据孤岛”、激发数据资源价值潜能，并将数据要素从“沉睡资产”转化为“生产引擎”，带动社会治理效能提升和产业升级。

依据产业实践跟踪，在这一政策架构下，电信运营商因其在网络基础设施和用户数据积累方面的天然优势，成为数据要素市场的关键支撑方。与以往局限于通信行业内部的数据应用不同，近年来，三大运营商积极响应国家数据治理战略，尤其在与地方政府、国家数据局等机构的合作中扮演了“公共数据服务提供商”与“基础数据整合运营商”的双重角色。具体而言，运营商通过接入政府公共管理数据、构建人口热力图、出行行为画像、城市通勤模型等数据产品，为地方政府在城市治理、交通调度、风险预警等方面提供数据支撑。例如，在某南方沿海城市的数据要素流通试点中，运营商与市大数据局合作建立了“城市时空大脑”，基于数据处理引擎将多个政务部门的脱敏数据进行融合治理，并以API 方式服务于智慧交通、应急响应等场景，实现了政府治理数据从“部门自用”向“多方协同”的跃升。

在数据要素市场化和公共数据治理体系逐步完善的背景下，金融保险机构正逐步突破“数据使用方”的角色限制，系统性地嵌入由政府数据资源与电信运营商构建的外部数据生态，推动形成具有中国特色的“数据 – 平台 – 场景”三元协同模型。依据行业跟踪来看，**金融保险机构正与电信运营商及国家数据管理机构共同构建起以高质量数据资产、大模型能力和多元业务场景为核心的新型协同机制。**这一合作模式突破了传统的数据调用或供应关系，转向基于数据全生命周期价值共创的生态共建逻辑。

图 9：数据要素驱动的协同生态

数据来源：公开资料，众安金融科技研究院

在数据侧

电信运营商凭借其覆盖全国的网络基础设施和高频用户行为数据积累，构成了具备高密度、高实时性和高交叉价值的数据底座。通过与保险机构在客户画像、风险数据、赔案信息等内部数据的融合，三方借助隐私计算、多方安全计算（MPC）等前沿数据协同技术，实现联合建模与跨域知识迁移，显著增强风控模型的泛化能力与实用价值。

在平台侧

电信运营商在取得客户授权或符合法定情形下，根据数据产品化和平台化战略，将脱敏后的数据封装为标准化的风险图谱、行为洞察、区域活跃度等数据产品。保险机构在合法合规的前提下与具备合法数据资质的数据提供商合作，通过“数据脱敏 + 分场景授权”的模式快速对接区域化定价、精准营销和灾害预警等业务场景，显著提升产品的灵活配置与市场响应速度。此外，伴随数据资产入表、估值与质押机制逐步建立，部分保险机构开始探索通过数据资产参与资本市场的可能性，试图

构建“数据 + 金融 + 技术”三位一体的新型资产组合，推动数据从要素向资本的跃迁。

在场景侧

更具前瞻性的合作已进入智能场景的深度融合阶段，特别是在以物联网、AI 和边缘算力为基础的感知层技术不断成熟的背景下，保险机构正在实现从“事后承保”向“事中感知、事前干预”的战略跃迁。例如，某财险与某电信运营商在工业厂房保险、公共安全责任险等领域的合作，通过部署 IoT 传感器采集实时风险信号（如温湿度、电力异常、烟雾浓度等），嵌入承保和理赔模型，形成实时风险监测和自动理赔触发机制，显著提升运营敏捷度与服务体验。同样在车险领域，基于 5G 边缘计算环境构建的 AI 图像识别能力，已广泛应用于远程定损、理赔自动审核等流程。

> 在制度与基础设施层面，电信运营商与国家数据局在区域数据平台建设、数据标准制定、数据资产估值等关键环节中逐步形成主导力，保险机构则通过嵌入式参与，融入数据生态的运行与协同体系。

在多个城市级治理项目中，保险机构已不再只是服务接入者，而是成为“标准共建者”与“场景共创者”，例如在智慧城市应急响应平台中，保险产品被作为风险缓释机制原生嵌入至城市治理流程，实现服务与政策的协同一体化，扩大了保险机构在社会系统中的功能边界。

综上所述，政府机构、电信运营商与金融保险机构构建的“三元协同”合作机制，正推动金融保险业从传统的风险对价逻辑走向基于数据要素、智能系统与生态共治的复合型价值创造体系。其深层价值不仅在于模型精度和运营效率的提升，更在于帮助保险机构建立横跨数据流通、智能建模、场景服务与资本运营的全栈式能力体系，构筑面向未来的差异化竞争壁垒与持续性增长引擎。在三元协同机制构建过程中，应坚持合法合规、数据安全两大前提，切实保障消费者权益。

2. 垂直整合：构建企业级智能协同底座

近年来，随着金融保险行业对智能化能力的系统性要求不断提升，“垂直整合”这一理念正在发生变化。它已不再仅仅意味着在公司内部实现不同业务系统之间的数据打通，而是在数据、算法、算力与组织资源之间，构建起具备稳定运行机制与动态调度能力的智能能力基础。其关键在于，企业能否通过整合内部能力与外部资源，打通从数据采集、模型调用到业务执行的全流程路径，推动智能化从技术辅助走向生产核心。

在企业引入智能技术的早期阶段，一般金融保险机构对垂直整合的理解多集中于打通保单、理赔等关键系统，实现基础信息互通。但随着大语言模型、语义识别、实时推理等能力逐步进入业务主流程，传统的“点对点对接”策略已显局限。智能化真正的挑战在于，如何将数据、模型与业务场景实现结构性融合，使 AI 能力不再停留在“外部调用”或“辅助判断”的工具层，而成为业务流程本身的一部分。这一过程的核心在于三点逻辑推演：首先，企业需建立面向 AI 的数据预处理与标准化机制，使原本零散、异构的数据具备被模型高效理解与调用的前置条件；其次，模型能力本身需通过 API 或平台组件化形式嵌入业务流程之中，实现按需触发与过程联动，而非事后分析或批量处理；最后，所有模型判断结果需具备业务可接受性与系统可执行性，即其输出可直接驱动业务动作或被纳入决策路径，完成模型输出与流程执行之间的耦合闭环。唯有完成这三重结构适配，AI 能力才能从“附加组件”转变为“内生模块”，真正融入企业的日常运营逻辑。这一趋势迫使企业将垂直整合上升为能力构建机制，推进从局部项目向系统工程的转型。

从当前行业跟踪来看，多数金融保险机构即便拥有多年的业务积累和相当规模的数据资产，但在支撑智能系统运行的真正“可用数据体系”仍存在优化空间。问题的本质原因在于多数机构的数据系统是围绕职能部门建设的，而智能化场景往往跨越业务条线，需调用多个系统的数据资源。以理赔审核为例，一个智能模型可能涉及调用 CRM 数据、历史赔案与风险标签等信息，但这些数据分别归属在运管、风控、理赔等不同部门，权限体系与目标导向各异，导致数据流贯通存在难度。并且，仅靠局部技术打通无法解决根源问题，为此，金融更保险机构必须在组织、权限与流程层面重构数据调用机制，才能为智能化流程提供稳定支撑。

与此同时，模型嵌入也对业务运行机制提出更高要求。传统系统的部署可割裂运转，但智能模型依赖连续反馈机制进行动态优化。这一过程涵盖模型推理、业务响应、结果回流与策略更新，链条长、环节多，横跨多个系统与组织边界。任何一个环节响应滞后，都大概率造成模型能力在系统中的“干扰”甚至“停滞”。因此，企业不仅要接入模型，更要为其设计可持续运行的反馈路径，这要求组织具备构建智能化闭环的能力架构，而不再依赖于分散的系统维护逻辑。此外，企业还面临能力复用效率低、建设重复投入高等问题。一套可泛化的语义标签体系，原本可以支撑营销、风控、客服等多个

场景，但由于缺乏平台化能力承载与统一治理机制，往往在不同部门中被重复构建，形成资源浪费。要破解“能力碎片化”这一转型挑战，企业必须从项目交付逻辑转向平台服务逻辑，推动能力标准化、接口化、组件化，真正实现数据与模型的统一承载与复用。

归根结底，垂直整合的重点并不仅仅是构建一个“大中台”或“新平台”，而在于建立一套可以持续调度智能资源、动态支持业务演进的能力框架。这种体系要求企业打破职能本位的组织设计逻辑，将流程中的关键节点转化为可组合、可编排的能力模块，并通过统一的调度平台实现跨系统、跨数据、跨模型的智能资源统筹。其本质是让AI从边缘插件演进为生产环节中的内嵌角色，具备面向未来迭代、自我优化和持续扩展的运行机制。从战略角度看，垂直整合的成功标志在于企业能否将智能能力从“单点试点”拉升为“系统运营”，进而获得组织层面的敏捷性与结构性升级。

随着这一趋势推进，一类新兴角色正成为垂直整合中的关键枢纽——以智能平台建设与能力部署为导向的技术合作方。这类合作方涵盖云厂商、大模型厂商、行业集成商、行业解决方案提供商等。在具体实践中，这些技术方不仅提供模型 API 和算力资源，更围绕应用软件开发与底层平台国产化需求，构建软硬一体化的部署体系，并结合行业知识和工程经验，参与 AI 场景设计、模型测试与工程集成等全流程工作。部分技术厂商正在以 MaaS（模型即服务）平台为基础，构建适配特定行业的 AIGC 能力集。通过在保险、银行等行业场景中积累模型调优经验与业务规则参数，他们不仅提供 API 级调用接口，还结合数据标注、任务编排、规则策略等模块，形成可快速部署的智能流程组件。这些企业还与金融机构共建 AI 应用实验室，围绕金融保险的核心流程测试场景模型适配性，发布联合解决方案，并基于工程化积累，提供从模型选型、集成部署到工具平台封装的全流程交付服务，实现在场景中的深度嵌入与可持续优化。其部署路径已从传统项目制转向模块共创与运营共担的生态协作模式，是金融保险行业推进智能能力构建的关键协作框架。

面向大模型在金融场景的落地实践，不同类型的金融机构因其发展阶段、数据治理能力和技术架构基础的差异，逐步形成了多元化的协作路径。

当前主流合作模式大致可分为三类：

01

模型底座合作模式

即技术企业提供基础大模型能力，金融机构结合自身行业语料、业务场景与标签体系，参与联合微调，提升模型对本行业语义和逻辑的适应性

02

训练平台共建模式

双方基于联邦学习或私有化训练平台开展合作，实现“数据不出域、算法就近训练”，在保障数据安全与监管合规前提下，增强模型能力与业务场景的持续匹配

03

应用场景嵌入模式

由技术企业输出标准化、模块化的模型服务组件，以轻量化方式对接金融机构既有系统，显著降低集成成本并加快部署节奏

三类模式分别对应金融机构在技术投入策略、模型控制力与部署敏捷性等方面的不同诉求。

模型底座合作模式，适用于头部金融机构

主要适用于已建立数据治理体系、具备内部算法团队的头部金融机构，如国有大型银行、全国性保险集团及大型资产管理机构。这类机构重视模型定制化程度与长期知识资产积累，典型应用包括智能投研分析引擎、多语言舆情解读系统、面向专业服务的知识问答模型等，强调高适配度与知识表达能力的持续演进。

训练平台共建模式，契合专业型金融机构

更契合监管要求严格、对模型训练过程可控性要求较高的专业型金融机构，如政策性银行、区域性商业银行、信用保险机构及券商风控部门等。其核心诉求在于在确保数据主权与算法透明度的基础上，实现模型的本地化适配与敏捷更新。应用场景包括文本合规审查系统、信贷风险预警引擎、智能化合规抽查助手等，强调可解释性、训练自主性与合规一致性。

应用场景嵌入模式，适用于中小金融机构

广泛适用于资源配置有限、数字化程度相对较低且更关注研发的投入产出比的中小金融机构，如区域性保险机构、农信机构、基金销售平台等。这类机构更关注智能能力的接入门槛与投入回报效率，通常通过 API 化或低代码模型组件快速集成，满足客服辅助、问答生成、智能外呼、业务流程引导等轻量级任务需求。该模式部署周期短、改造压力低，是当前大模型在中低复杂度金融场景中落地的重要路径之一。

总体来看，三类协作模式映射出金融机构在智能化能力建设路径上的分层演进策略。大型机构倾向于深度定制与模型能力沉淀，适合构建“专属型智能中台”；合规压力大、技术基础中等的机构，更偏好兼顾安全与灵活性的训练平台共建路径；而中小型机构则以“轻量部署、快速见效”为目标，优先采用场景嵌入策略。上述模式构成了金融行业内部大模型能力部署的分级结构，为进一步衡量训练成本、部署效率与投资回报率（ROI）等核心变量提供了实践基础，也为技术企业提供了切入金融智能化生态的路径指引。

3. 横向协同：拓展跨场景智能联动边界

在智能化系统逐步由“点状接入”走向“结构性嵌入”的演进过程中，金融保险企业对数据能力的依赖也正经历深刻转变。随着我国经济结构的系统性重构，金融保险机构所面临的挑战与业务机遇也同步发生转化：一方面，保险责任向复合型场景延展，客户行为日益呈现跨平台、多设备、非线性等特征，传统基于静态字段与规则配置的模型体系在识别能力和响应速度上已无法胜任。另一方面，新兴业态在机器人运营、数字健康、自动化出行等领域持续涌现，衍生出大量脱离传统范式的金融保险需求。企业要在这一新环境下构建可持续的智能运营能力，显然无法仅依赖自身系统中沉淀的历史数据。

正是在上述背景下，金融保险机构与其他行业基于业务场景的横向协同由此成为破题的关键。它不再是简单的数据获取或平台对接行为，而是作为企业跨越数据壁垒、能力边界与场景感知盲区的协同路径重新进入决策核心。其逻辑起点并非“数据在哪里”，而是“业务问题如何被更智能地解决”。企业需要以真实业务场景为牵引，反向推导所需数据结构、模型形式与反馈路径，并在这一过程中构建起与外部生态的动态协同机制。横向协同的价值不止于扩大数据来源，更在于通过跨主体共建实现业务判断力的结构性增强。但理想与现实之间仍存在显著落差。在实践中，多数协同项目仍停留在接口连接与资源置换阶段，缺乏围绕场景构建能力闭环的系统设计。接口不通、数据结构不一、模型不迁移、规则不互认，这些技术与机制层面的瓶颈，使协同难以超越“项目合作”的范畴。其本质问题在于，横向协同若脱离具体场景驱动、缺乏共同收益机制与演化路径，便无法在组织内部沉淀能力，也无法在生态层面构建粘性。

为此，企业需要将横向协同从“理念设想”转化为面向业务的路径设计，不再停留在宏观叙述层面。围绕场景重构协同路径的过程中，企业不再停留于策略层面，而是进入机制设计与资源匹配的实际部署阶段。其中，协同机制是否具备落地价值，关键在于其嵌入场景的适配程度，以及是否能形成数据与反馈的正向闭环。换句话说，协同策略的有效性，不取决于“协同能不能做”，而取决于“在哪些任务形态中协同最具成本效益与能力沉淀价值”。

因此，在横向协同机制推进初期，企业往往聚焦于高频、高损与高潜力三类关键场景作为优先启动区域。这些场景在业务规模、风险权重及生态扩展性方面具备显著的协同驱动力，适合作为机制验证与能力沉淀的首选路径。高频场景的流程耦合性与触点密度高，适宜作为协同机制的稳定入口；高损场景的判断难度与风险外溢性强，亟需横向补强以提升识别精度与防控能力；高潜力场景则通常位于生态延伸或新兴服务方向，通过前置协同机制部署，有助于企业抢占数据接口与模型标准的主导权。三类启动场景的共通特征在于——协同成本可控、数据接口明确、机制复用性强，既具备快速成效的可行性，也为后续机制标准化与能力模块化奠定基础。

在可行性优先的原则下，企业应重点聚焦三类协同任务结构：一是决策增强，解决判断能力的瓶颈；二是流程优化，提升运营效率与响应速度；三是客户洞察，实现对用户动态认知的持续深化。它们分别对应不同的数据类型、技术手段与治理要求，构成横向协同在金融保险领域落地的三大基础单元。

01

高频场景

高频场景指的是在业务流程中重复出现、对运营稳定性产生持续影响的环节。这些节点虽单次价值不高，但由于出现密度大，协同带来的效率提升能够在累计中释放显著运营红利

02

高损场景

高损场景对应风险敞口大、处理成本高的关键节点，在这些场景中构建协同能力不仅可以提高识别精度，还能有效控制损失边界

03

高潜力场景

高潜力场景关注的是智慧出行、数字健康等新兴领域，这些领域内的标准尚未固化。率先进行协同工作形成先发优势

正是基于以上“启动场景”所揭示的能力空白与协同潜力，经验表明，横向协同的价值落点主要集中在三类关键能力维度：其一，判断力的提升，即在复杂、不确定环境中增强风险识别与响应的准确性；其二，执行效率的优化，通过多系统联动提升流程响应的稳定性与及时性；其三，客户理解的深化，以动态、语义化画像机制增强对用户行为、动机与场景的感知能力。这三类能力维度不仅构成了协同机制的功能边界，更成为企业构建智能能力体系的结构基础。围绕上述能力诉求，企业在实际部署中逐步形成了三种典型的横向协同形态，分别对应不同场景任务、数据来源与价值机制。它们既是能力目标的映射单元，也是协同机制设计的路径入口。

第一类为决策增强型协同，主要面向反欺诈、动态定价、智能核保等高敏感度场景。此类场景具有预测性要求高、可变性强的特征，单靠内部历史数据往往难以形成前瞻性判断。因此，企业需要通过接入外部信用记录、行为序列、风险信号等变量，构建多源评分模型，拓展模型视野。这种协同强调的是“认知外延”——即如何通过跨域数据丰富模型输入，提高决策链条对未来态势的感知能力。

第二类为流程优化型协同，聚焦于身份认证、理赔审核、交易验证等标准化流程。该类型的核心目标并不在于复杂判断，而在于通过系统间的实时互联与数据即时注入，减少人工干预、降低处理延迟。企业一般通过标准化 API 对接构建端到端的闭环处理能力。其机制设计重心在于“工程可行性”与“调用低延迟”，强调的是如何将外部系统的能力嵌入业务流程，实现柔性连接与精准触发。

第三类为客户洞察型协同，体现出更强的战略属性。其目标在于突破静态标签体系对用户认知的限制，转向动态、语义驱动的画像机制。在此模式下，企业通过与电商、社交、出行等多元平台的数据拼接，构建出分层、多维、情境感知型的用户画像体系，从而支撑更高阶的产品定制、客户运营与生命周期管理策略。该类型的协同强调“深度理解”而非表层识别，目标是赋予企业对用户行为的预测性与可解释性，使用户管理从“事后运营”走向“预判干预”。

这三类协同形态虽在结构上各自独立，承担的任务目标也各不相同，但在能力建设层面却构成了一个互为支撑的闭环体系。其中，决策增强型协同为企业提供前瞻性的风险识别与策略判断能力，流程优化型协同提升了关键业务节点的执行效率与流程稳定性，而客户洞察型协同则拓展了企业对用户价值的理解维度，支撑个性化运营与持续服务优化。正是这种协同形态与能力诉求之间的映射关系，使横向协同机制的构建逻辑逐步从“资源驱动”向“能力导向”转变。相比于以往以数据覆盖面为目标的被动整合，当前更强调以具体能力缺口为起点，反向识别所需的外部数据要素、系统接口与反馈机制，从而推动协同设计的主动性与系统性提升。从演进角度看，企业可视业务阶段与能力成熟度，选择不同类型协同作为启动点：以流程优化型作为低风险入口、逐步过渡到决策增强型的高价值协同，最终通过客户洞察型完成用户认知系统的系统性跃迁。这种渐进式协同策略，将协同机制从单点工具升维为企业能力重塑与生态融合的核心支撑架构。