大模型金融应用实践及发展建议

北京金融信息化研究所（FITI）

2023 年 11 月

版权声明

编制委员会

主　任：

潘润红

副主任：

黄程林、庄文君

编委会成员（排名不分先后，按姓氏拼音排序）：

陈志明、代　铁、董佳艺、方　科、胡利明、黄　炜、李　锋、李金龙、李一昂、林冠峰、刘承岩、刘殿兴、刘汉西、刘训艳、陆　鑫、潘　华、沈剑平、孙　莉、汪　航、王建军、王　玲、王　麒、王彦博、吴永飞、肖　京、杨　波、俞　枫、张海燕、张　洁、赵　海、赵焕芳

编写组成员（排名不分先后，按姓氏拼音排序）：

曹伯翰、陈　超、陈广浩、陈　鸿、陈　明、陈志豪、迟倩倩、崔雨萍、刁翔宇、段旭欢、范　容、苟志龙、何　平、何巧媚、何幸杰、胡国强、胡师阳、胡应明、黄　彪、黄　韦、金　睿、金　昕、李大伟、李冬妮、李　峰、李　娟、李梦霄、刘　畅、刘　威、罗安扬、罗方华、毛奕凯、彭　晋、戚　嚣、邱晓慧、谈　健、唐登龙、王　莹、王煜惠、王　振、温昱晖、文俊杰、吴青松、徐崚峰、鄢胜利、杨　洋、于　飞、曾培基、占可非、张　彬、张　绅、张笑冬、赵　辉、周思霁、宗　宇

主要执笔人（排名不分先后，按姓氏拼音排序）：

鲍思佳、卢金环、屈　洋、孙　曦、王帅强

主编单位:

北京金融信息化研究所

中国农业银行股份有限公司

中国邮政储蓄银行有限责任公司

上海银行股份有限公司

腾讯云计算(北京)有限责任公司

蚂蚁科技集团股份有限公司

参编单位:

中国工商银行股份有限公司

中国银行股份有限公司

交通银行股份有限公司

中信银行股份有限公司

中国光大银行股份有限公司

平安银行股份有限公司

招商银行股份有限公司

上海浦东发展银行股份有限公司

华夏银行股份有限公司

中国民生银行股份有限公司

兴业银行股份有限公司

浙商银行股份有限公司

北京银行股份有限公司

中信证券股份有限公司

国泰君安证券股份有限公司

华泰证券股份有限公司

国信证券股份有限公司

中国平安保险股份有限公司

中国银联股份有限公司

北京国家金融科技认证中心

北京银联金卡科技有限公司

海光信息技术股份有限公司

北京火山引擎科技有限公司

北京瑞莱智慧科技有限公司

摘　要

当前，大模型正掀起新一轮智能化发展热潮，赋能千行百业。大模型具备优秀的理解、学习、生成和推理能力，其工程化应用包含数据构建、模型算法、模型训练、模型压缩与加速、模型评测、模型运营和安全可信多个环节。金融机构和科技企业正积极探索大模型在金融业的合理应用，已试点应用于智能客服、智能办公、智能研发、智能投研等多个业务场景，进一步推动金融服务的智慧再造，加速 AI 技术赋能金融业务提质增效。

大模型在金融业应用还处于初期探索和应用试点阶段，仍面临金融应用规范与指南不完善、金融应用场景缺少范式、高质量金融训练数据欠缺、训练算力支撑不充分、算法可信度和安全性不足等诸多挑战。本课题系统梳理了大模型工程化应用的各个技术环节，总结金融机构在大模型技术路线、使用方式和应用场景等方面的实践经验，重点研究金融机构应用大模型时在场景、数据、算力、算法等方面面临的突出问题，并提出相关意见建议，旨在为全行业提供参考和借鉴，促进大模型在金融业快速落地应用。

目 录

一、概述

近年来，以人工智能为代表的新一代信息技术加速应用，特别是基于大模型、大数据、大算力的 ChatGPT 的发布，标志着人工智能技术取得里程碑式突破，推动科技创新进入新阶段。随着大模型技术的迅猛发展和场景价值的不断涌现，该技术或将重塑多个行业的工作方式和格局。

为稳步推动生成式人工智能在各行各业的有序应用，我国陆续出台一系列政策法规和管理办法，《国务院 2023 年度立法工作计划》将人工智能法纳入了国家立法计划，《生成式人工智能服务管理暂行办法》提出了促进生成式人工智能技术发展的具体措施，《网络安全标准实践指南——生成式人工智能服务内容标识方法》指导生成式人工智能服务提供者等有关单位做好内容标识工作，《生成式人工智能服务 安全基本要求》（征求意见稿）给出了生成式人工智能服务在语料安全、模型安全、安全措施、安全评估等方面的基本要求，《全球人工智能治理倡议》主张建立人工智能风险等级测试评估体系，不断提升人工智能技术的安全性、可靠性、可控性、公平性。

金融业是数字化、智能化的先行者，有望成为大模型技术落地的最佳领域之一。《金融科技发展规划（2022-2025 年）》明确提出要抓住全球人工智能发展新机遇，以人为本全面推进智能技术在金融领域深化应用，强化科技伦理治理，着力打造场景感

知、人机协同、跨界融合的智慧金融新业态。金融机构正在积极探索大模型在智能客服、智能办公、智能研发等业务场景的应用，提升智能技术的可获得性，助力金融服务降本增效。

二、大模型技术与产品发展现状

（一）工程化应用主要环节与技术

大模型相较于中小模型，具有更好的表示能力、泛化能力、学习能力和语义表达能力，但其参数量巨大、训练所需数据量和算力资源多、部署运营更为复杂，工程化落地涉及数据构建、模型算法、模型训练、模型压缩与加速、模型评测、模型运营和安全可信等多个复杂环节。

1. 数据构建

训练大模型需要海量数据做支撑，高质量数据集的构建和处理对于大模型的性能表现至关重要。训练数据集一般需要涵盖多种类型、多种领域的数据来源，并配以相应的数据预处理过程。根据数据来源不同，大模型的训练数据主要可分为公开数据、商业数据和私有数据。大模型参数量需要跟训练数据集大小相匹配，简单堆砌参数量并不能无限度地提升其性能。通过提升训练数据集质量和内容丰富度、加入一些特定数据集、合理利用外挂知识库资源、合理配置各种类型数据配比等方式，可以有效提升大模型的整体性能，减少模型幻觉，并加快模型的收敛速度。

高质量的数据预处理是提升模型表现和安全可靠性的重要

手段。比如自然语言处理训练数据的预处理手段一般包括：**质量过滤，**过滤重复数据、低质量数据、虚假内容、不合规内容等；**数据去重，**重复数据可能会降低大模型的多样性，导致训练过程不稳定，从而影响模型性能，一般可在句子级、文档级和数据集级等不同颗粒度上进行数据去重处理；**隐私脱敏，**对于包含个人敏感信息的数据进行脱敏处理，如身份证号码、电话号码等，包括但不限于匿名化、泛化等手段；**数据去毒，**消除带有种族/性别偏见、社会文化偏见、宗教文化偏见的数据，以及低俗、粗鄙和带有攻击性的数据等；**数据降维，**其目标是在保留基本信息的同时减少数据集的复杂性，从而提高训练效率，一般可通过减少特征维度或样本大小来实现；**数据增强，**通过人工创建对现有数据的变更来增加数据量和多样性，特别是在数据量有限的情况下，通过数据增强可以提高模型的准确性、泛化能力和鲁棒性，也可以应对数据类别不平衡等问题。完成数据预处理后，可以将数据通过分词等手段，转换为适用于大模型训练的表达形式，形成高质量语料。此外，在模型推理过程中，也可以通过整合外部的领域知识库或专业数据库，为模型提供额外的背景知识和参考数据，尤其是快速且不断地更新信息，从而提高模型的准确性和鲁棒性。

2. 模型算法

大模型技术的突破源于自然语言处理领域的 Transformer 架构。该架构使得模型参数量突破了 1 个亿，随后一系列大模型被推出。基于 Transformer 架构的模型可以分为编码器、解码器、

编码到解码三大类，其主要特点和代表性模型如表 1 所示。

表 1　大模型结构主要分类及特点

架　构	编码器	解码器	编码器-解码器
特　点	模型具备“双向”注意力，可使用上下文信息来进行预测，最适合理解类任务，如文本分类、命名实体识别、阅读理解等；模型参数一般偏小，预训练后需基于具体任务进行Fine-tuning。	模型只具备“单向”注意力，即基于前面的单词来预测后续单词，最适合生成类任务，如文本创作、对话问答等；模型参数量相对较大，实用性高，计算高效、内存占用少、泛化能力强。	模型同时使用编码器和解码器两个部分，适用于围绕给定输入产生对应输出的任务，如翻译、摘要，以及一些有很强序列特征的任务，如语音识别、图像描述生成等。
代表性模　型	BERT、RoBERTa、DeBERTa等。	GPT系列、PaLM、OPT、LLaMA等。	T5、ChatGLM、BART、ERNIE 3.0、M6等。

当前，基于 Transformer 解码器结构训练的大模型成为了自然语言处理领域的主流方案。在此影响下，语音、视觉以及跨模态等领域的大模型也尝试应用类似模型架构，并取得了较好效果，比如语音领域的 OpenAI whisper 和 DaLL-E 等，图像生成领域的 Stable Diffusion 开源模型等。

3. 模型训练

大模型训练涉及预训练和微调等重要环节。**预训练**的主要目的是利用大量无标签的数据，训练出一个有能力捕捉到数据中隐藏的底层结构和模式的模型，这一阶段的模型通常被称为“基座模型”。由于大模型的参数量和训练数据量的急剧增长，单个计算设备的算力已经不足以支撑模型训练。当前，一般通过分布式训练来解决预训练过程中的海量计算任务和高内存资源等问题，但也面临着计算墙、内存墙和通信墙等挑战。目前解决分布式训练的关键技术是并行化，将任务分割并分配到多个处理器或设备

上，以便同时完成计算，更有效地利用计算资源，减少训练所需时间。**微调**的主要目的是在预训练模型的基础上，通过有监督微调、强化学习等方式，进一步提升模型在下游任务中的表现，使得模型输出更符合人类期望。有监督微调，又称为指令微调，通过使用有标注的特定任务数据对预训练模型进行微调，从而使得模型具备遵循指令的能力。早期的微调算法会涉及到预训练模型的全量参数更新，计算成本较高，目前已提出了多种参数高效微调任务的方法以节约计算成本，如 LoRA、Adapter、P-tuning 等。强化学习技术是基于人类反馈，进一步调整模型的行为。其数据集一般由经过人工评估的反馈数据构建，这些数据反映了模型的输出与期望输出之间的差异，基于 Q-learning、深度 Q 网络或近端策略优化等强化学习算法进行训练。

大模型训练场景对中高端 AI 芯片需求旺盛，需要统筹规划 CPU 芯片、GPU 芯片、服务器、网络、存储、冷却、算力运营服务、AI 应用服务平台等多个方面。在金融机构通用服务器集群基础上，构建基于异构芯片体系的 AI 算力资源池，实现对金融机构现有 AI 算力资源的统一调度，保障大模型训练的算力支撑。

4. 模型压缩与加速

模型压缩是指通过各种技术手段来减小机器学习模型的大小、复杂度和计算量，加速推理过程并减少内存使用，以便在资源受限的设备上部署和运行，如移动设备、边缘设备等。目前，模型压缩技术主要包括知识蒸馏、剪枝和量化等解决方案。**知识**

蒸馏是一种训练小型模型以模仿大型模型行为的方法，保留了大型模型主要功能的同时降低了计算和存储需求，但通常需要一个预先训练好的大型模型，且性能上会有一定损失。**剪枝**是一种去除模型中不重要或冗余参数的方法，一般可以在不显著影响模型性能的情况下减小模型的大小和计算需求，但需要确定哪些参数是不重要或冗余的，以选择合适的剪枝策略。**量化**是一种减少模型参数和运算中数字精度以降低模型的存储需求和计算复杂度的技术，可适用于多种模型和任务，并显著减少存储和计算需求，但可能会造成一定程度的精度损失，且有时需要特定的硬件支持。

模型加速主要研究加速模型的训练和推理过程，伴随模型参数增长，正逐渐成为研究热点。**训练环节**，针对计算量、通信、内存可以进行一系列优化，例如使用梯度累积或梯度压缩可以优化通信策略、使用半精度浮点数可以节省内存等。**推理环节**，优化手段包括使用 GPU、TPU 和 ASIC 等芯片的专用硬件加速器加速计算过程、使用并行化和分布式推理提高推理吞吐量并减少推理时间、使用缓存和预取策略降低内存访问延迟、在边缘设备上进行推理减少与服务器端的通信延迟、结合模型压缩技术加速推理过程等。现阶段 AI 应用中，大量的算法、模型、开发框架、软件等开发都基于通用加速卡架构，在考虑硬件算力的基础上，要结合加速芯片软件栈及开发工具链等配套的软件生态能力。

5. 模型评测

模型评测在机器学习和自然语言处理领域扮演着至关重要

的角色。大模型具有更强大的泛化能力，可以处理多种任务，但大模型的输出可能存在不真实、不准确、不专业等问题，因此在大模型上线或升级时，有必要对其进行较为全面、充分的评测，帮助模型迭代优化。

大模型评测已成为行业发展热点问题，目前国内外相关评测层出不穷。据初步统计，目前行业内关于大模型基准测试或特定任务的测试数据集已多达 200 余项，主要推出机构可以大体分为学术界、产业界、媒体、社区以及智库等。其中一些代表性的评估基准包括 HELM、MMLU、C-EVAL、BigBench、HumanEval、AGIEVal、SuperCLUE、OpenLLM 等。总体来看，大模型评测仍处于早期阶段，如何构建出全面、充分且能伴随大模型能力增长不断迭代的大模型评测基准，仍面临较大挑战。

6. 模型运营

大模型运营包括工程化、部署、管理、调试、维护和监控等多个方面，旨在确保大模型在生产环境中稳定运行，持续适应变化，满足用户需求，保障数据安全。**工程化方面**，模块化和面向对象的编程可以帮助组织代码，使其更具可读性和可重用性，如将数据预处理、模型结构、训练循环和评估功能分为不同的模块或类，同时版本控制系统和自动化测试技术有助于及时跟踪代码的修改历史，确保每个功能模块和整个系统稳定运行。**部署方面**，要将模型转化为适用于实际环境的格式，包括序列化、压缩、硬件优化以及容器化工具的使用，以确保模型在不同环境中的一致

性。**管理方面**，使用身份和访问管理工具可以控制资源的访问权限，运用数据管理工具跟踪数据集的变化。**调试方面，**使用可视化工具和日志记录有助于监测模型运行时的详细信息，这对于诊断问题和优化性能至关重要。**维护方面，**要制定明确的更新策略以适应新数据和业务需求，同时要建立回退机制，确保出现问题时能够快速回退到稳定版本。**监控方面，**通常包括性能监控和异常检测，基于实时跟踪和警报设置，确保模型的可靠运行。

7. 安全可信

一般而言，大模型的安全可信会从多个维度进行考量和评估，包括但不限于：**可靠性**，即大模型的输出内容是真实的、一致的等；**内容安全性**，即大模型的输出应避免涉黄、涉暴等非法内容，并能遵循当地的道德准则和法律规定等；**公平无偏性**，即大模型输出应避免偏见、刻板印象、不公平等情况；**鲁棒性**，即大模型在面对投毒攻击、提示词攻击等恶意行为或者意外情况时，依然能够产生稳定和可靠的输出结果；**可解释性**，即大模型能够解释其推理过程并能透明展示其内容生成方式等；**数据安全和隐私保护**，即对训练和推理大模型的数据中可能包含的敏感信息进行相应的脱敏和保护处理等。

为有效提升大模型的安全可信水平，需要在大模型开发和运营的全流程采取相应的安全措施，包括但不限于：**数据处理环节**，建立高质量的训练数据集，有效去除有毒或错误信息、注重训练数据的分布比例以避免产生偏见、对训练数据中的敏感数据进行

脱敏等；**模型开发阶段**，引入对齐技术等让大模型的输出更符合人类价值观；**模型上线前**，需对大模型的安全可靠水平进行充分评测，如红队对抗测试等；**大模型对客提供服务时**，可以引入安全围栏技术，既能帮助大模型拦截外界的恶意提问，又能对生成内容进行风险过滤和拦截；此外，**模型运营过程中**，还需要建立持续的监控和定期审核机制，以及时发现异常情况并持续迭代优化模型。

（二）国内外主要产品情况

自 ChatGPT 引起业界高度关注后，国内外科技巨头纷纷加大了对大模型的研发和投入，以模型为核心，围绕模型全生命周期设计、生产并提供产品、技术和服务，推动“数据、模型、服务、场景” 的 AI 生产力闭环连接和迭代优化，为大规模、标准化的人工智能创新提供技术支撑。在此背景下，大模型相关产品和服务迎来了爆发式增长。此外，许多大模型相关的开源项目涌现，为研究和创新提供了便利。这些开源项目主要朝着两个方向发展：一是相对 ChatGPT 更经济的、平民化的替代品，二是围绕着大模型建立的外围应用工具。

数据来源：金融信息化研究所更新[1]

图 1 国内外典型大模型产品发展时间线

三、大模型在金融业应用与探索实践

人工智能技术的应用一直是金融机构数字化转型和智能化升级的核心领域，大模型具备的理解、学习、生成和推理能力，可观的知识容量和卓越的多任务泛化能力，将进一步推动金融服务的智能化升级。

（一）技术路线

1. 技术选型

大模型常见的技术选型路线包括开源大模型应用、产学研联合创新大模型研制和商用大模型采购。大型金融机构一般采用三路并行的模式，中小金融机构普遍尚未采购商用产品，大都停留在调研、汇报和 POC 测试阶段。

[1] https://github.com/RUCAIBox/LLMSurvey

一是开源大模型应用。开源大模型是指通过开源方式发布和共享的大模型，其源代码和相关资料对公众开放，适用于具备一定技术能力的开发者、研究机构以及对模型定制和二次开发有需求的机构。国内外的大模型有近一半选择了开源的方式，Bloom、GLM、Llama 是目前国内外流行度较高的开源大模型。金融机构选择开源大模型的前提是能够依靠自身或者第三方公司进行大模型应用研发，且具备后期维护、迭代更新的能力。尽管开源大模型已具备较好的实践效果，但仍面临数据安全合规风险、开源协议风险等挑战。总体而言，金融机构对于开源模型的采用普遍持谨慎态度，但在前期探索阶段通常会采用开源大模型进行测试、研究，同时将其与商用大模型的应用效果进行比对。

二是产学研联合创新大模型研制。金融机构、科技企业、科研院所、高等院校等通过合作，共享知识、技术或资源，发挥各自优势，推动技术创新和产业发展。产学研联合创新有助于金融机构在大模型创新应用初期推动特定试点场景快速落地。一方面，可以在一定程度上助力金融机构加速科技创新及数字化转型，打造更好的产品和服务，增强行业竞争力。另一方面，可以减少金融业对国外技术和产品的依赖，加快国产化步伐，提升行业安全可控能力。

专栏一 产学研联合创新大模型研制
工商银行与清华、鹏城实验室、华为等高等院校、科研院所、科技企业开展大模型联合创新。

交通银行与华为、科大讯飞共建了联合创新实验室，推进大模型及算力集群技术、人工智能等先进技术在金融领域的落地应用。

北京银行与火山引擎、华为、中科院自动化所、中科闻歌共建联合创新实验室，围绕金融大模型体系构建、前沿金融科技应用等领域开展合作创新，共同探索银行智能化技术的最佳实践。

三是商用大模型采购。众多国内外商用大模型正在逐步推广应用。相较于开源大模型，商用大模型可以为金融机构提供更加工程化、易用性强、服务有保障的解决方案。

金融机构在大模型技术选型时需要综合考虑大模型对业务质量和人员效率的提升效果、大模型持续创新能力、大模型运行时的稳定性和安全性等多个方面。同时，金融机构还需结合自身业务特点和实力情况，对资金、人员、配套工具产品完备性等因素进行全面考量，以选择合适的大模型解决方案。此外，要建设并维护内部统一的大模型资源库，并在此基础上建立大模型应用开发平台，更好地充分利用业界多种领先的通用大模型。

2. 部署方式

为制定合适的部署方案，金融机构首先需要确定需求和目标，其次要根据业务场景和技术要求，选择合适的硬件和软件环境，确保能够支持大模型的运行和优化。针对不同应用场景，金融机构探索采用不同部署方式以更合理地应用大模型。根据部署环境的不同，可以将大模型部署方式分为私有化部署、行业云部署和公有云部署等。

一是私有化部署。金融机构将大模型部署于自有服务器，由金融机构负责维护和管理。私有化部署可以提供更好的数据安全保障，大大减少信息安全隐患，且一般具有较好的应用效果，尤其是对于需要运用内部语料训练的金融业务场景，可以根据金融机构的需求进行定制和优化模型，并随时增减资源。但是，这种部署方式往往会产生高昂的成本，需要金融机构投入大量的资金、人力来建设和维护。

二是行业云部署。由行业内起主导作用或掌握关键资源的组织建立和维护，以公开或半公开的方式，在确保数据安全的前提下，向行业内部或相关组织提供云平台服务。将大模型部署在金融云，既能在一定程度上满足数据安全可控的要求，又兼备成本低、扩展性强等优势。中小型金融机构对行业云部署的需求更为迫切。

三是公有云部署。金融机构通过标准接口调用部署在公有云上的大模型，由云服务提供商负责维护和管理，具备更低的成本、更高的灵活性和可扩展性。然而，由于金融机构无法完全掌控其数据的存储和管理，使用这种部署方式可能面临数据安全风险，因此该部署方式可能更适用于互联网类的金融企业或非敏感类业务场景。

鉴于个人隐私保护和数据不出域等相关要求，私有化部署仍是金融机构部署大模型的主要选择方式。私有化部署使得金融机构得以保留对数据和系统的完全控制权，但由于大模型训练和推

理对算力及配套基础设施有较高要求，该部署方式更适用于大中型金融机构，对于中小型金融机构实现难度较大。对于不涉及数据保密性的场景，比如证券公司基于公开数据生成投资策略及研报撰写，行业云或公有云部署具有一定优势。目前，国外已有API 市场化采购服务，且价格低廉。随着私有化部署成本不断提高，中小银行、证券公司、基金等中小金融机构，迫切需要通过行业云或公有云来降低大模型金融应用的门槛。

（二）使用方式

大模型具备出色的自然语言理解能力，但因其在金融垂直领域的知识储备不足，回答的专业性难以满足要求。金融机构需要将金融领域已有的数据库、知识图谱等专业知识系统，与大模型的意图理解能力、语言生成能力和场景掌控能力进行对接，实现大模型的行业个性化和机构定制化。然而，大模型技术复杂度高、参数规模大、研发投入高，特别是需要大量的数据标注和人工反馈等工程化投入，从头自研大模型的难度非常大。因此，金融机构目前主要采用 API 调用、提示工程、模型微调和二次训练等应用方式，以降低大模型的应用研发门槛，解决场景数据少、模型精度低等问题，推动大模型快速落地测试。比如，农业银行利用有监督的模型微调和强化学习等技术，融入行内知识库数据，训练、收敛出能理解行内知识的基础模型，目前已具备行业知识问答能力。中信证券将特定问题及与其相关内部文档材料作为问答输入，既可获得较好的应用效果，又避免了二次训练的高昂成本。

1. API 调用

API 调用是指通过 API 接口调用大模型，实现图像生成、文本生成、语音合成等功能。公有云大模型大多采用以 API 接口的方式提供服务。该方式使得金融机构在大模型探索应用初期能够快速将大模型在项目中进行测试验证，且成本低廉。然而，API 调用的方式不仅需要完成一系列的技术整合和接口开发，也面临如何在不降低模型准确性的前提下保障数据安全和隐私性、如何高效处理海量数据、生成内容可控性和鲁棒性较差等问题，适合直接落地的金融业务场景很少。一般来说，金融机构应用 API 调用方式对大模型产品进行测试后，会根据测试结果结合该模型在其他领域的表现进行综合评价，判断是否进一步选择微调或二次训练的方式进行优化。

2. 提示工程

基础预训练大模型内含知识丰富，但很多潜能尚未被激发，如何有效引导大模型来完成特定任务存在挑战。提示工程通过优化提示语句激发大模型所具备的潜在能力，提示语通常是一段文本，用于构建问题或对任务进行恰当表述，以便大模型基于内在知识生成合适答案，可通过少样本、思维链路等方式实现。提示工程并不改变模型参数，对于可指令微调的大模型来说，构造提示语的成本相对较低，大部分成本转嫁为标注成本，通过设计好的提示语来调优任务效果。

3. 模型微调

基础预训练大模型具备较好的通用性知识，但实际应用中，每个具体场景都有其独特的需求和流程，仍需优化模型以提升应用效果和价值。模型微调是优化模型效果和性能并提高准确性的重要手段，也是当前金融机构试用大模型的最常用手段。当大模型缺乏垂直领域知识时，通过模型微调技术，以较低的成本即可明显提升针对特定任务的应用效果。一般来说，当测试结果良好时，会优先选择模型微调技术提升应用效果。

4. 二次训练

二次训练指在已有模型的基础上，使用新的数据集对模型进行重新训练，从而提高模型的性能和应用效果。二次训练通常需要对模型参数进行修改，以适应新数据集的特点。相较于微调，二次训练会产生高昂的数据、算力、人力以及时间成本，但也可以实现更好的效果和更广泛的场景化应用。一般来说，大型金融机构因具备海量垂直领域数据基础，在微调技术难以实现期望达到的应用效果时，可以选择该方式高度定制化模型以提升模型的泛化能力和场景化应用能力。

此外，金融机构还采用大模型融合知识库、知识图谱等组件的方式，通过引入领域向量库检索、数据增强等技术，结合大模型关键超参优化，形成不同模型针对不同行业和领域的参数体系，精准匹配业务需求，使得大模型回答结果较为可信。

（三）应用场景

大模型在金融领域的应用前景广阔，趋动 AI 从劳动密集向脑力密集应用，金融机构已经将大模型试点应用于智能客服、智能办公、智能研发、智能投研等多个金融业务场景，但从能用到好用、易用，再到规模化应用还任重道远。基于当前金融机构主要场景探索实践来构建金融行业大模型应用场景全景图。

数据来源：金融信息化研究所

图 2 金融行业大模型应用场景全景图

1. 智能客服

传统的智能客服的服务方式主要是先识别用户意图，再匹配到特定的对话模板，是一个非常庞杂的配置过程，特别是要确保所有渠道对于相似问题的回答保持一致。这种方式不仅维护成本高，维护难度大，而且难以应对复杂多变的自然语言对话，用户体验较差。金融机构普遍认为大模型虽然仍不具备直接面客的能力，但可以在通用大模型的基础上，叠加金融客服领域的数据和专业经验，经过垂直领域定向训练后，客服机器人可以综合考虑用户提示语和用户习惯，准确识别用户意图，作为客服助手协助

客服人员开展客户的陪伴和关怀，升级真诚话术，提供 24 小时的多语言支持，有效提升用户对话体验，提高服务质量。

专栏二 智能客服场景
工商银行利用大模型的语义理解、生成能力，结合基于知识库的检索提取能力，精准理解客户意图，生成符合其业务特性的结果，为客户提供更准确、更个性化的服务，提升应答效率和质量，从而提升客户满意度。2023 年 3 月，工商银行进行了大模型智能客服试点测试，一定程度上提升了客户情绪识别效果，缩减了约 50%的人力维护成本。 农业银行大模型 ChatABC 面向行内客服员工提供远程银行 AI 辅助客服问答助手服务。此助手基于远程银行问答数据完成训练微调，支持模型在多轮问答中识别客户的主要意图，结合远程银行知识库和知识图谱，生成拟人问答，辅助坐席人员在问答中获取知识，有效提升坐席人员的答复效率。 中国银行正探索将大语言模型运用于客服场景，帮助业务人员实现更自然灵活的智能问答功能，有效地应用到坐席问答辅助、员工培训、行内员工办公等多个场景。 交通银行正探索使用大模型来准确识别客户意图和坐席人员的知识检索需求，进行相应客服知识的提炼和推荐，并在坐席通话结束后，对服务内容进行自动的标准化分类，生成通话小结，以提升坐席人员整体工作效率。 光大银行探索大模型技术与现有智能客服相融合，基于大模型

意图理解能力对客户问题进一步理解，以提升客户问答满意度；基于大模型技术探索坐席工单自动生成可能性，对客户与坐席对话内容快速生成摘要工单，提升客户服务效率。

民生银行探索运用大模型技术辅助坐席完成问题回复和工单生成等工作，实现工单流程自动化，提供内容以支持顾问与客户互动，支持高度拟人化的客服机器人，优化客户体验，提升客服质量。

兴业银行基于大模型技术实现客服语料智能泛化，对当前语料标注过程中存在的标注工作繁琐、工作量大等痛点，通过大模型生成符合业务需求的语料，以帮助客服运营人员提升标注效率，进而提升智能客服回答准确率，减少客户投诉。

上海银行正探索在客户服务领域运用大模型优秀的语言理解能力并结合行内知识库，在充分洞察客户诉求的前提下，对复杂场景进行分解，能够从知识库中自动完成有效知识的提取与采编，解决知识库运维完全依赖人力、多语义理解、语义缠绕等问题，为客户提供优质解答。

国信证券探索运用大模型技术自动化生成服务话术和客户指标数据，包括：知识问答、行业分析、行情快报、客户资产配置建议、资讯推送摘要、投诉建议反馈等，提升运营人员的服务质量和效率。

平安保险运用大模型技术解决传统知识库梳理成本高、培训成本高、推荐话术僵化等问题，在客服人员与客户通话过程中，实时根据客户需求，生成个性化话术，从而提升客户满意度，大大降低坐席人员学习成本。

2. 智能办公

大模型可以作为办公助手，发挥其出色的文本生成能力，辅助完成报告和运营文案生成、邮件起草、公文润色、纪要撰写、内容审核、辅助纠错等工作，为员工提供更加便利、快捷的智能办公工具及个性化的智能办公解决方案，大幅提升工作效率。此外，大模型还可以实现不同语种的实时翻译，提升金融机构跨境业务交流和管理水平。

专栏三 智能办公场景
工商银行通过搜索技术与大模型结合，将知识库检索到的信息作为大模型的生成依据，实现自然语言的向量搜索，辅助业务人员在业务办理过程中快速查询到客户问题的准确应答，提升应答效率和质量，从而提升客户满意度。 农业银行大模型 ChatABC 面向行内员工办公场景推出智能问答应用。此应用着眼于大模型在金融领域的知识理解能力、内容生成能力，并融合知识库，具备研发服务领域级知识理解和问答能力，可完成自由闲聊、行内知识问答、内容摘要等多类型任务，并以行内研发服务平台的问答助手、工单自动化回复助手、行内即时聊天工具等多渠道形式进行试点。 交通银行正探索在办公软件中嵌入基于大模型的智能问答助手，为员工提供人资规章、授信、对公、零售业务场景的会话式咨询。同时，交通银行利用大模型搭建会议纪要助手，从口语化的会议记录中提取关键信息，包括讨论主题、观点、结论等，组织成连

贯自然的语言，生成书面化的会议纪要和待办事项。

邮储银行运用大模型技术搭建智能知识问答系统“灵犀”，系统采用 Langchain 和向量数据库技术，结合大模型对自然语言的理解与生成能力，发挥了垂直领域内知识理解能力，赋能业务实现专业知识智能解答，提升业务办理效率。同时，邮储银行运用大模型技术搭建“小邮助手”智能机器人，提供在线业务知识问答、热点问题分类展示，实现业务难点、要点即时回复和精准提示，提升柜员操作体验及业务处理效率，释放业务指导员工作。

光大银行基于大模型技术构建员工助手，应用于行内通信软件，快速解答内部员工日常办公问题，构建多项智能应用，可帮助行内员工实现消息智能摘要、日程会议智能提醒、邮件自动编写、代码自动生成、系统助手等多项能力，提升办公效率。同时，光大银行探索将现有知识库与大模型相结合，打造企业内搜场景应用，将现有零散知识召回交给大模型，基于大模型理解及生成能力辅助业务办公人员快速定位核心问题并作出合理判断，以提升业务效率。

民生银行通过大模型的理解能力建立多场景文档助手，对文档的主要内容进行摘要汇总整理，全面提升员工工作效率，实现针对文档不同场景的个性化需求。

兴业银行探索通过大模型嵌入行内 WPS 办公套件，实现包括 PPT 大纲生成，文章内容生成，内容扩写与改写，文章风格转变等功能，以减轻总分行一线员工的文字材料撰写润色的负担。

北京银行搭建“京智助手”大模型对话机器人，提供行内知识

问答、数据分析、任务执行等功能，应用于协同办公、智能客服、合规管理等场景。目前，“京智助手”已向全行10000多名员工开放，并同步建设移动端和PAD端，已在合规管理、数据分析和流程自动化场景，取得了一定的应用成效。

上海银行利用大模型的理解和生成能力，搭建智能办公助手，一方面接入行内知识库，提升知识检索能力，并由大模型对于长文档自动检索生成知识条目等，另一方面接入行内各类办公系统，提供公文检查、写作、总结润色等功能，在大幅提升办公效率的基础上，极大地提升员工交互体验。

中信证券正在积极探索利用大模型建立跨境实时通讯系统，旨在实现境内外员工用母语进行无障碍交流，提高员工工作效率，加强团队协作力度，以及提升对国际化客户的服务质量，加强全球化布局和业务发展。

国泰君安将大模型技术与OCR、语音识别等技术相结合，围绕智慧办公场景开发智能办公助手和基于大模型的知识库问答应用，提供会议纪要生成、邮件撰写等常见办公任务小工具，通过自然语言问答的形式实现文档问答，为员工提供一个智能、高效、便捷的工作伙伴，提升办公效率和工作质量。

3. 智能研发

大模型在智能研发场景展示出很大的潜力。大模型具有处理自然语言和生成多种编程语言高质量代码的能力，在系统设计、代码生成与补全、代码翻译与注释、辅助测试等方面为科技人员

提供帮助和支持，提升开发效率和交付质量。同时，大模型的阅读理解和 SQL 生成能力，可以实现指令文字到 SQL 代码的直接相互转化，业务人员无需了解数据库技术与编程知识也可轻松完成数据查询等工作。此外，通过大模型与 BI 等应用系统结合，可以实现报表自动生成，可视化地展示数据查询分析结果，使得非技术人员更直观且深入地理解、利用数据，降低技术门槛。

专栏四 智能研发场景
农业银行大模型 ChatABC 推出 AI 辅助编程应用，面向行内研发员工提供辅助编程服务。通过行内多个系统的代码和代码规范进行微调，实现 Java、Python、JavaScript、SQL 等语言的代码生成、代码补全、代码解释等能力，可在前端、后端、单元测试等多类研发编码场景提供辅助，提升研发人员的开发效率。 平安银行在 ChatBI 项目中应用了 BankGPT 的金融理解能力，旨在为业务人员和管理层提供即时的数据分析服务，当业务人员面临数据相关的问题时，系统可以迅速给出精确答案，实现提问与回答的无缝对接。 民生银行探索运用大模型技术辅助用户完成高质量代码编写，按需完成代码生成、代码补全等功能，促进分析平民化，提升代码效率，实现质量和效率的双提升。 兴业银行通过大模型嵌入行内开发 IDE 的方式，探索基于大模型技术实现代码辅助生成，以提升科技研发单位的开发效率和代码质量。

上海银行通过在软件开发过程中引入大模型进行代码补全、注释生成、代码纠错等自动化辅助功能，极大地提升了开发人员的开发效率和代码质量;在代码测试方面，应用大模型生成能力，辅助自动生成测试案例，有效提升测试效率及案例覆盖度；在数据分析方面，上海银行正在探索大模型与行内经营管理工具--“掌上行”的有效结合，通过自然语言交互方式，允许用户就经营指标自由提问，进一步降低数据分析的门槛，提升数据驱动经营管理的有效性。

中信证券探索运用大模型实现股价预测等模型构建，根据用户输入的自然语言，生成相应 SQL 语句，准确地完成用户数据查询需求，以期适配不同数据库、完成多表数据查询。

国信证券通过大模型的辅助代码生成能力，协助 IT 人员、业务产品经理、运营人员进行代码编写、数据分析、辅助代码生成检查等工作，提升系统开发、数据分析等内部工作效率。

4. 智能投研

大模型技术可以帮助投研人员从海量、分散、庞杂的报告中挖掘关键信息，自动抓取财经、债市、信用等多个市场板块的资讯报告内容，快速获取报告的核心观点、关键数据和市场趋势，分析预测市场交易情况，智能生成资讯报告和投研简报，为投研人员提供投资、风控等决策辅助。

专栏五 智能投研场景
工商银行综合应用大模型的核心信息提取、智能文本生成等能力，实现金融报告的自动生成，有效将投研简报生成效率从 1 小时

缩短至 5 分钟，提升了投研人员对海量文本数据的整合归纳提炼效率。

兴业银行通过大模型技术实现研报摘要的智能生成，构建了一套包括研报文档结构化、信息抽取和大语言模型语义理解摘要生成的一体化解决方案，实现研报核心内容智能提炼，提高了兴银理财子公司投研团队查询、阅读内外部研报的效率，加快了投资决策效率，一定程度节省人力成本，同时提升了客户体验。

国泰君安证券利用大模型的自然语言到结构化查询(NL2SQL)等能力，改进传统问答系统的精准性和灵活性，实现对投研领域内包括行情、公司、基金等数据的精确、高效问答。

华泰证券正探索运用大模型对文本的学习理解能力，学习历史研报的撰写模式、分析逻辑和行文风格等内容，从而实现研究报告初稿的自动撰写。目前已初步搭建内容召回、内容生成的线上撰写服务框架，打通从财务报告的非结构化文档解析，利用 embedding 技术构建财报知识库，根据历史研报进行高相关内容召回，结合内容召回与大模型服务，通过建设 Prompt 工程，打通大模型撰写能力链路。

5. 智能营销

结合金融业语料进行适应性训练后，大模型可辅助生成营销话术和营销文案，帮助客户更快地获取最新资讯和产品的信息。同时，通过对话式营销，大模型可优化客户参与度，提升服务的效率与质量，引导其做出决策。此外，随着投资者数据、产品知

识和营销文案的不断积累，大模型技术可以从海量信息中检索词条，并提炼整合，可针对特定客户实现贴心、高质量、有创意的精准营销，快速生成个性化建议和推荐，并构造相应的图文推广，进一步提升客户转化率和营销效率，为客户提供更加全渠道、个性化、有温度的金融服务。

专栏六 智能营销场景
平安银行借助 BankGPT 优秀的文案生成能力，针对不同客户，批量生成个性化营销文案，从而更有针对性地提升客户粘性和业务转化率，为营销运营团队提供支持。同时，平安银行利用 BankGPT 的自然语言理解能力，助力多媒体运营团队实现自动化 FAQ 抽取功能。 民生银行探索运用大模型技术赋能客户意图理解营销场景，通过搜索知识库、产品库、聊天记录等，洞察客户真实意图，自动提供广告等营销文案、自动生成个性化营销内容、自动提供各类分析报告，提升营销成功率。 平安保险通过大模型的自然语言交互、内容生成等能力，智能化分析并提取客户需求，辅助生成保险产品营销素材，并根据客户标签属性提供针对该客群的产品推荐以及推荐理由相关话术，基于个人医疗历史和分线因素，提供个性化保险建议和方案，有助于产品精算人员制定针对性保险方案。

6. 智能运维

金融机构可以利用大模型生成技术，实现智能运维分析、运维知识获取、网络安全分析等，进一步提升运维效率。大模型具

备良好的语义搜索能力，在复杂的数据中准确找到所需结果，助力运维人员快速对接各类型的结构化、非结构化数据，实现流程自动化、智能检索和内部资料的辅助审核。此外，对于大模型还可以辅助运维人员完成代码攻防测试、明文检测等安全检测工作。

专栏七 智能运维场景
上海银行探索将大模型应用于故障分析及解决等场景，结合历史生产事件解决工单、运维文档或问答对等知识库，当故障发生时，大模型能够自动根据历史解决经验及知识库给出故障分析及解决方案，从而为运维人员提供辅助支持，以提升事件解决效率。

7. 智能风控

风险控制是金融业的核心要务。大语言模型的阅读理解能力可以辅助提示存在的法律风险，降低人为疏漏概率并提升法审人员工作效率。在大模型的通用能力基础上融合金融行业的知识和数据用于真伪核验、舆情分析等环节，结合风控数据模型，进行风险模拟与压力测试，对可能发生的风险事件作出预警，实现贷前智能推荐、贷中全面风控、贷后监测预警全流程风控管理。此外，基于大模型的合规知识问答助手在金融合规场景中也发挥了一定作用。

专栏八 智能风控场景
光大银行基于大模型理解能力，探索大模型解读法律法规政策的可行性，基于业务问题快速定位政策文件并给出法律法规依据，打造法律法规领域智能专家，辅助各业务场景快速把控政策、法规

风险。

华夏银行探索利用大语言模型技术叠加合规图谱，以高质量内外规数据为基础，以结构化合规知识标签为辅助，利用大语言模型语义理解、信息汇总和自然语句生成能力实现智能化合规知识问答功能，降低合规知识查询门槛。

平安保险运用大模型技术，基于历史数据和模拟数据进行风险模拟和压力测试，评估产品设计的可靠程度和稳定性，以便产品精算人员更好地了解产品可能面临的风险和挑战，助力产品精算人员制定相应保险策略。

8. 智能投顾

围绕财富管理专业知识对大模型进行增量训练，构建基于大模型的投资顾问助手，提供围绕个体的全生命周期智能投顾服务，实时监测市场动态，调整投资策略，并根据客户的风险偏好、收益目标和资产状况，为客户提供个性化的投资建议和组合优化，大幅提升服务效率和服务体验。

专栏九 智能投顾场景
国泰君安证券运用大模型的理解分析能力，进行金融资讯热点话题提取和归纳，对相关信息做正面/负面消息的评分和整理，生成每天/周/月的热点话题榜单，既能实时跟踪行业网站热点，又能回顾过去一段时间的热点话题，基于榜单排名走势预测股市热点信息。

大模型在提升金融服务效率和体验、降低金融风险和成本、创新金融产品和模式等方面有着显著作用。大模型在金融领域的

应用场景具有较大空间，随着大模型与各个金融场景的深度融合，将进一步提升我国金融业的智能化水平，进一步深入推动金融机构数字化转型。

（四）应用趋势

一是金融机构将采用先内后外、从易到难、场景迁移的方式落地大模型金融应用。由于目前大模型直接对客难度较大、可控性不强，金融机构主要对内应用大模型能力，待技术逐渐成熟，可以考虑对外输出。此外，金融机构也将优先选择风险等级低、适配应用难度小、业务提升效果明显的场景进行大模型试点落地，逐步将试点应用场景迁移到真实、复杂的业务场景，实现大模型对金融产品和服务的全面升级。

二是大小模型协同进化是大模型发展的一个必然趋势。大模型对应场景多为开放式和主观型问题，侧重推理和创造，也可被用于作为连接多个具体任务模型的通用接口，而小模型对应场景多为封闭式问题，不涉及过多主观推断，答案的正确性可以被清晰验证，因此在某些特定场景仍具有更好的表现。大模型对于中小模型并非是替代或对立的关系，两者应相互协作、互相搭配，由大模型向边、端的小模型输出模型能力，而小模型负责实际的推理与执行，同时向大模型反馈算法与执行成效，使得大模型的能力持续强化。金融机构将加强研究和推进大小模型协同、生成式技术与传统人工智能技术协同，将大模型连接到传统软件，提升行业整体智能化水平。

三是多模态金融大模型的发展与应用仍有较大潜力。大模型生成效果的提升依赖于垂直场景系统化程度和高质量数据。金融业具备专业领域知识库，多年来沉淀了大量格式多样的优质业务数据。运用多模态技术实现知识的迁移、表示、对齐和推理，使得大模型能更好地构建金融领域内外部生态系统，助力金融科技创新和金融业务赋能，为金融机构提供更多智能化、个性化的服务和决策支持，同时也为客户和市场参与者带来更好的体验和更稳健的金融环境。

四是 AI Agent 未来可能推动人工智能成为金融业信息基础设施。AI Agent 是有能力主动思考和行动的智能体，以大模型为大脑驱动，能够自主感知环境、形成记忆、规划决策、使用工具并执行复杂任务，甚至与其他 Agent 合作实现任务。尽管多智能体的发展仍面临较大困境，但随着算力支撑和技术研究的不断演进，AI Agent 将在各行业发挥强劲动力，尤其在与各行业紧密相连的金融业，更全面地实现人机融合，为金融机构提质增效，创造更深度的价值。

四、大模型在金融业应用面临的风险与挑战

（一）金融应用规范与指南亟需完善

金融业作为强监管行业，在政策方面一直遵循着高标准和严要求。我国已相继发布《生成式人工智能服务管理暂行办法》《网络安全标准实践指南——生成式人工智能服务内容标识方法》

《生成式人工智能服务 安全基本要求》（征求意见稿）等文件，对大模型在通用领域的应用进行了合理的约束和引导，但是针对大模型在金融领域的应用尚且缺少可实施落地的标准规范和指南，对应用过程中的权责界定尚不明晰，缺少对生成内容的问罚机制及大模型广泛应用后可能存在的无序商业行为的监管机制。此外，针对大模型训练和推理的 AI 算力基础设施、行业语料库标准化建设也较为欠缺。

（二）金融应用场景缺少范式

大模型选型、架构调整设计、技术验证等环节过程复杂，金融机构缺乏大模型技术融合场景落地的方法论，对大模型的能力边界认知不足，尚未明确大模型适合哪些业务场景、是否有必要替换传统的 AI 设备、何时适合落地，尚未健全大模型应用创新风控管理机制，尚未有典型的落地案例可以向行业规模化推广。大模型支撑多个场景或服务的行业应用，其测评指标非常复杂，测试数据集设计构建与更新维护难度大、成本高，尚且缺少一个覆盖面广、公允度高、满足不同场景和任务特征的大模型金融应用及其风险治理的评估方法或指标体系，致使金融机构进行大模型选型及评估存在较大困难，阻碍了大模型金融场景应用进程。大模型需要提示工程相关的内置模板，与原有的机器学习、深度学习等模型工作方式有很大差别，也增加了相关人员的工作难度。此外，大模型金融应用需要与现有系统和业务流程进行集成，需要跨组织、跨部门、跨团队协作，组织能力面临挑战。

（三）高质量金融训练数据欠缺

数据是大模型训练的基础，为了切实解决金融业务问题，需要大量高质量、多领域的金融数据基于业务属性对大模型进行增量训练。金融领域知识存储形式繁多，包括影像件、PDF、Excel等多种格式，需要通过分类、清洗、问答数据集梳理等大量前期处理及后期更新维护工作，针对各种业务难点、要点问题的解答还需要搜集大量专家经验，以保持大模型的准确性和有效性，而这会耗费大量人力物力。同时，大模型训练迭代需要一定时间，致使大模型对时事的了解有限。金融数据流通仍在探索阶段，而单一金融机构掌握的数据资源较为有限，一定程度上影响了大模型金融应用效果。金融数据敏感性高，在数据分级分类管理、数据脱敏清洗、防止数据偏见和滥用等环节也存在难题。

（四）训练算力支撑普遍不足

大模型训练和推理需要足够的算力支撑，在高端GPU芯片断供的背景下，金融机构对中高端AI算力的需求存在较大缺口。由于金融数据敏感度高，金融机构普遍选择私有化部署大模型，而构建、训练、优化大模型需要高性能的计算资源和大量的存储资源，硬件设备的采购和维护需要高昂的资金投入，给金融机构带来较大的成本压力。大模型的训练与推理对AI芯片的要求有所不同，当前我国AI芯片能较好地支撑推理，而在训练上仍与国际领先水平有明显差距，存在计算能力不足、芯片制程工艺有限、算力调度不灵活、产业生态不完备、与大模型兼容适配性不

够等问题，且在金融业应用普遍缺乏验证。我国 AI 芯片适配涉及 CPU、操作系统、云平台、AI 框架、加速框架和算法模型等多个层次，适配工作复杂且难度大，牵一发而动全身。

（五）算法可信度和安全性有待提升

大模型金融应用在准确性、安全性、稳定性和金融科技伦理等方面面临挑战。准确性方面，大模型存在文本及数据幻觉问题，其训练数据难以溯源、生成内容不可信、计算过程不可解释、推理逻辑不专业，难以直接应用于数据准确性要求高、业务流程复杂度高的金融场景。安全性方面，金融场景涉及大量敏感信息，大模型在输入输出过程中可能造成数据泄露，从而引发重大的安全事件和恶劣影响，同时特殊的提示词构造、逆向工程等手段可能被非法用于攻击大模型，绕过内容过滤模块，使攻击者获取超出权限范围的结果，甚至窃取大模型的所有权和使用权，肆意修改模型代码或参数，使其生成不准确、不公平、不合规的恶意结果。稳定性方面，大模型算法框架不够完善，开发环境不够友好，适配的框架比较少，且当前大模型算法主要基于国外的机器学习平台和技术，在我国设备的操作系统、编程环境、算法库等应用时可能出现各种意想不到的错误和异常，从而影响大模型运行效果和稳定性。此外，大模型可能引发算法歧视、人权、道德、造假等科技伦理风险，影响金融服务的健康发展。

五、多措并举提升大模型金融业应用水平

（一）加强金融应用的指导与管理

坚持发展和安全并重、促进创新和治理相结合的原则，制定一套完备的适用于金融领域的大模型管理体系，分类分级地制定政策指南，引导大模型在金融业规范应用，持续推动业务创新发展。对大模型的训练数据、算法设计、生成内容、风险治理等方面进行管理，制定从准入阶段的评估和备案，到对外提供金融服务，再到事后反馈的全过程管理机制，确保大模型在金融领域应用的合规性和安全性。明确金融业涉及大模型使用的各类主体的责任和义务，并制定合理的问责机制。积极参与大模型国际标准化工作，推动金融行业的AI算力基础设施、行业语料库标准化建设，制定合理的标注规则，加强不同大模型产品之间的互通性和兼容性。

（二）有序推动金融应用场景落地

金融机构征集并统筹大模型相关需求，梳理现有需求场景及方案，形成跟踪台账，探索大模型与金融业务融合所需的前提条件和能力边界，选取业务价值高、实施完备度成熟、风险可控的业务场景优先落地应用。多技术路线并举，技术点同步验证，探索应用监管沙箱等治理方式，加快大模型的试点应用步伐，打造大模型金融应用最佳案例，并进行规模化推广复用。基于分级分类分域的治理思路，形成多元敏捷协同的治理体系，推动实现大

模型金融应用负责任、可监督、可追溯、可信赖。创新大模型金融应用评估工作机制和理念，加快普适性好、具有底线约束的标准研制和通用测评体系建设，将自动评估和人工评估相结合，提高金融业大模型应用评估工作的质量。第三方评估机构积极协助金融机构，搭建一套适应其业务的模型评价体系，并建立评测指标与评测数据集反馈和更新机制，促进大模型在金融场景应用中的迭代优化。

（三）积极构建高质量金融数据集

金融机构梳理场景应用数据需求，建立并完善大模型应用数据使用机制，探索一套面向大模型的数据“采集、清洗、管理、应用”方法和体系，提升数据集的规模、质量和多样性，保障模型微调与投入生产后的数据连贯性、稳定性。做好敏感数据拦截的审计检查工作，根据场景特点和风险等级进行数据分级分类。研究建立针对数据偏见、技术滥用、数据滥用等问题的风险管控机制，做好大模型私域管理和权限隔离，保证数据在可控范围内流动，并通过区块链存证技术强化管控，确保各个环节的数据可追查、可溯源且不可篡改。金融业积极推动大模型训练和行业标准测评公共语料库建设，助力行业级金融大模型建设，提升大模型金融应用水平。

（四）产用协同共筑 AI 算力基础设施

产业机构加大我国 AI 芯片的研发与推广应用，保障大模型推理和训练的算力资源供给，提升硬件安全可控水平。加强产学

研用协同创新，共同推动大模型软硬件生态建设，标准化驱动抽象及依赖，提升框架通用性，合力推进 GPU、DCU、NPU 算力集群和智能计算中心等算力基础设施建设，助力大模型全栈兼容性适配，快速推动大模型私有化部署。通过提供算力资源租赁、移动算力资源车等方式为金融机构提供算力解决方案。金融机构结合自身需求，梳理共性硬件资源需求，完善大模型算力中心规划，基于国产和非国产算力建设多源异构算力资源池，建立健全算力资源分配流程、资源使用跟踪与资源回收机制，对国内外芯片进行统一纳管、虚拟化和调度，充分、合理利用算力资源，保障算力平台供给稳定性，提升大模型训练推理效率。根据算力建设情况，形成大模型算力适配的模型微调部署方案，在保证模型效果的前提下，通过模型压缩、小样本训练等方式进一步降低应用成本。同时，推动中高端算力集群配套的网络、存储、冷却等方面的改造工作。

（五）完善算法优化与风险管控体系

产业机构与金融机构沟通切实需求，提高大模型算法的透明度、可解释性和可预测性，帮助金融机构更好地理解算法运作方式和决策依据。建立用户参与和反馈机制，纳入模型算法改进和优化等环节之中，提升用户体验和技术的可靠性。积极推进大模型算法、模型和工具等全流程配套体系建设，提供全套的大模型金融应用解决方案，降低大模型金融应用落地的技术门槛和风险。金融机构做好适合自身场景的基础大模型选型，研究建立针对大

模型生成内容、算法安全的风险管控机制，配备专业人员实时监控大模型运行情况。对于涉及敏感数据、直接对客或对输出结果准确性高的场景谨慎使用大模型技术，建立“AI 审核+人工审核”两道关卡，保障大模型输入输出数据的安全可控。加强金融科技伦理治理，负责任、有道德地开展大模型技术创新应用，通过开展专业培训与论坛交流等方式，增强相关人员的安全风险防范意识。在确保用户个人信息安全和隐私不受侵犯的前提下，金融机构与产业机构加强联合研究与攻关，提高大模型算法自研水平，通过共享大模型前沿研究成果、金融业训练与评测数据集等方式，不断提升金融业大模型算法安全性、合规性、专业性和兼容性。

附录

案例一：邮储银行基于大模型的智能知识问答

一、背景及意义

随着数字化时代的到来，数据成为关键生产要素，价值愈发凸显。在企业生产经营过程中，会产生大量的知识信息，传统的知识问答系统是基于语言学意义的专家规则系统，由人工编制的知识库对接自然语言接口构成。其知识领域狭窄，词汇总量有限，人工成本较大，常出现语言歧义问题，较难产生实用价值，复杂甚至冲突的语言规则使得系统的维护成本增加。

随着人工智能技术的发展以及大模型的出现，其强大的上下文理解能力、语言生成能力及学习能力，使得它们能够产生更准确、更连贯的回答，基于大模型的知识问答在知识整合和归纳方面提供了非常大的帮助。

邮储银行自主打造了基于大模型的智能知识问答系统“灵犀”。“灵犀”是一套完全私有化部署、信息安全可靠的智能问答系统，采用开源、免费、可商用的中文 Llama2 模型，基于社区活跃的 Langchain 框架，结合大模型针对自然语言的理解与生成能力，使用向量数据库存储训练数据。以超亿级开源数据训练参数为基础，发挥了指定领域内大模型强大的生成能力，为邮储银行在金融领域内的人机交互场景提供了自主可控的大模型能力。

二、技术方案及创新点

“灵犀”采用完全私有化部署模式，无任何信息外泄隐患，同时通过金融银行领域特定语料的微调训练，提供金融知识理解能力，结合大模型生成式与向量化特点，提供面向邮储银行知识领域的智能问答系统。

（一）更智能的知识问答。系统基于 Langchain 框架及百亿参数中文 Llama2 模型，通过向量检索的优化、大模型提示词优化、超参调整以及数据清洗和整理，不断提升模型问答结果的准确性及合理性，系统更智能；

（二）多元异构硬件支持。系统完全同时适配 CPU 与 GPU 硬件，避免对稀缺资源的过度依赖，通过对模型进行特定格式的转化，以及采用当前十分前沿的模型运行框架，系统能够在无 GPU 硬件资源的条件下无缝运行系统；并通过对运行模型框架部分源码的修改、相关软件包的依赖更改，完成了 CPU 下 x86 及 arm64 架构的全适配；

（三）更少的资源占用。通常大模型至少需要 6G 显存或 32G 内存情况下才能加载运行，通过引入特定的模型运行框架以及对模型转化操作，同样体量的模型只需 0 显存 10G 内存，8 核线程情况下即可运行系统；

（四）更具扩展性的特定领域知识库。系统基于 Langchain 框架，统一前置向量数据库存储并由大模型整合，支持自有知识库灵活训练及检索，提升效率的同时也支持自有知识库无限扩展。

三、项目进展及应用效果

“灵犀”上线以来，通过在向量检索率提升、提示词优化、超参调整等方面进行不断的优化，逐步提升模型问答结果的准确性及合理性。目前已完成信息科技、风险管理、个人金融、公司金融等板块100多篇内部政策和制度的素材训练，同时完成信贷、不良、资产保全等业务条线业务制度的学习，为邮储银行业务人员检索、学习相关制度政策提供效率利器，让科技赋能业务发展。

案例二：某股份制银行基于腾讯云 TI-OCR 大模型单据处理

一、背景及意义

在该行单据处理场景中，此业务涉及到大量银行回单、交易发票、跨境汇款申请书、业务往来邮件、传真等数据，需要整理、录入系统。若纯依赖人工，耗时长、效率低、成本高、易出错，而用传统的 OCR 深度学习模型，需要经过检测、识别、结构化等阶段，多个阶段错误累积，难以突破检测识别难点，模型指标上限低且不具备阅读理解和推理能力，不同场景下模型能力无法复制、定制成本高。因此，需要借助 OCR 大模型的能力解决以上单据处理面临的问题。

二、技术方案及创新点

腾讯云 TI 平台 TI-OCR 是一款专注于 OCR 细分场景建模的训练平台，覆盖了从数据导入、数据生成、数据标注、模型训练、应用编排到应用测试发布的全流程。平台沉淀了腾讯优图强大的 OCR 内置模型和专家丰富的模型优化经验，能助力非 AI 专业的客户轻松实现自主构建自定义业务下的 OCR 应用解决方案。

（一）支持四种识别模式

1. 智能结构化

从单一版式或混合版式的图片中提取出 Key 字段、Value

字段，以及 Key-Value 的键值对关系。

2. 固定版式结构化

实现对如身份证、火车票、机动车登记证等所有字段位置固定的单一版式类型的数据信息进行提取。

3. 检测/识别

实现各类表单、票据、证件、单据等的包含手写体、印刷体、中英文的字段提取。

4. 智能分拣

即通用目标检测，检测出图片中物体所在的框位置及其所属类别。

（二）先进的技术架构

整体架构上，TI-OCR 训练平台采用 Master-Worker 的分布式架构。Master 节点负责对外提 HTTP 协议的产品功能接口，Worker 节点负责执行模型训练和推理等计算任务。TI-OCR 训练平台支持最多 100 个 Worker 节点，每个节点可以配置一块或多块 GPU 卡。模型训练可以使用单机多卡以提升训练速度，也可以在多个节点上同时运行多个训练/推理任务，以提升系统吞吐量。

数据存储方面，TI-OCR 训练平台使用 MySQL 存储元数据，使用普通硬盘存储图像、模型等数据。

部署方面，TI-OCR 训练平台支持单机或者集群模式。在单机模式下，平台只需要一台 GPU 机器即可安装运行，成本低且

几乎不需要运维；在集群模式下，平台只需要 1-2 台低配的 CPU 机器，外加可横向扩展的多台 GPU 机器，平台占用资源少，性价比高。

容灾方面，TI-OCR 训练平台支持主备部署 MySQL，支持使用磁盘阵列，以提供基本的数据容灾能力。计算任务都采用异步运行模式，Master 节点异常不会影响训练/推理任务的运行。

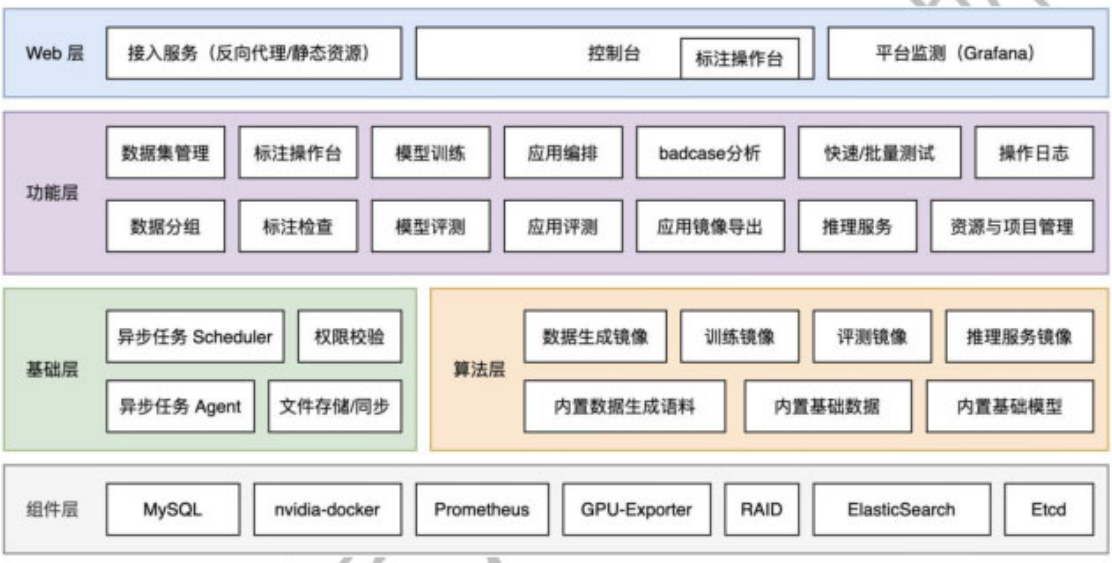

图 1 TI-OCR 训练平台的技术架构

（三）主要技术特点

腾讯云 TI-OCR 大模型具备三大特点：一是基于原生大模型，不经过训练，直接支持常规下游任务，零样本学习泛化召回率可达 93%；二是通过 prompt 设计，不经过训练，支持复杂下游任务，小样本学习泛化召回率可达 95%；三是通过多模态技术，小样本精调解决传统 OCR 难题，自研端到端技术突破检测识别业界痛点，比传统模型召回率提高 3%-20%。

三、项目进展及应用效果

该行利用腾讯云TI--OCR,对非结构化数据进行自动化分拣、提取并转换为结构化数据，实现对各种格式数据的高精度识别，识别准确率95%以上。通过应用腾讯云-OCR，该行在单据处理中减少了低价值高耗时手工作业，节省运营人力成本，实现多元业务数据处理的标准化、线上化、自动化。

案例三：某股份制银行基于腾讯云金融大模型的智能客服

一、背景及意义

在银行客服系统场景，传统的智能客服存在三大痛点：一是知识维护量大，冷启动知识配置成本 14 天-1 个月不等，且需要持续投入运营；二是问答覆盖率低、拦截率低，由于知识边界受限，不在知识库的问题无法回复或者几轮下来往往答非所问；三是接待上限低、服务效率低，进线后坐席需要经历知识理解、搜索、组织回复的复杂流程。为了解决以上问题，需要构建金融大模型能力赋能智能客服场景。

二、技术方案及创新点

该行基于腾讯云金融行业大模型能力，结合自身场景数据，通过腾讯云 TI 平台进行精调，构建了专属的金融客服大模型，并进行私有化部署。通过快速接入银行企业知识，直接学习企业文档库、搜索引擎现有资源，同时直接对接银行 API 进行任务式对话问答，打造了银行专属 AI 助手。

图 2 银行客服智能解决方案架构

该银行采用语音识别、语音合成、人脸识别等 AI 技术，在进行安全认证的基础上，对自然语言进行深度分析，并进行精准回复，让服务“看得见”、“听得见”，大幅减少人工成本的基础上，极大的提升用户交互体验。

一是智能语音导航和智能问答。该银行将 NLP 技术与知识库、知识图谱相结合，开发出智能语音导航和智能问答功能，搭建起智能客服的核心。通过智能语音导航和智能问答，可以实现对客户的合理引导，将复杂的功能菜单扁平化，提升客服服务效率。

二是智能外呼和智能质检。一方面，该银行利用 NLP 、情绪识别、语音识别等技术，将人工客服的服务的录音进行转写，并在此基础上进行数据分析，形成专题分析。另一方面，将外呼营销、催收等过去由人工开展的业务，交由机器人办理，并实时对数据进行深度分析，朝着定制化的客户处理方案演进。

三是打造客服助手。客服助手可以在人工坐席服务时，为员工提供即时的话术支持，也可以根据人工坐席的需求，为人工坐席提供即时的协助，提升工作效率。

三、项目进展及应用效果

腾讯云金融行业大模型在银行投资、财富管理、绿色金融等业务方面，为该行提供智能咨询、辅助分析、决策等服务，助力该行多个核心业务智能化、健康发展。在具体业务方面，该行推出智能客服机器人，可以通过手机银行、微信银行、网上银行等多个渠道，为客户提供问答服务，极大地推动银行客服系统升级。

案例四：某股份制银行基于中科可控的金融大模型服务平台

一、背景及意义

大模型对金融行业人工智能产业产生了变革式的影响，在金融业态大模型落地应用的过程中有三大痛点：一是 GPU 算力不足。在美国对华高端 GPU 制裁的背景下，H100、A100、L40S 等适用于大模型训练的算力缺失，国产化的算力成为必要选项。构建基于异构芯片体系的国产 GPU 资源池，探索化解人工智能芯片供应“卡脖子”风险的路径，具备对行内现有 AI 框架及算法模型的统一调度能力。二是预训练引擎选择难度高。大模型预训练引擎种类繁多，建设能够支撑多种大模型预训练引擎接入的平台，实现多个大模型按需服务于多个场景的“松耦合”模式，是高性价比、高效率解决企业在大模型“发动机”层面的选择难题。三是落地应用支撑能力。构建基于向量数据库、知识库、舆情监测、微调工具链、推理服务等多个部分的产业化组合。

二、技术方案及创新点

中科可控联合捷通华声、百川智能为金融领域的客户提供“量知大模型平台”，全部国产化，符合信创要求，全方位安全保障；垂直领域训练模型定制，用户全面参与；打通数据接口，结合搜索引擎，做到实时数据呈现与风险及时预警。

该平台包含基础的 LLM 服务以及典型大模型应用，不仅包含

大模型基础的文本创作、代码编写、文本翻译以及其他常见自然语言处理能力，还可以为用户提供非结构化文档检索问答、辅助FAQ知识库加工、票证信息提取等多种功能，并提供http接口，方便进行二次开发，更便捷地集成到业务系统。

（一）自带应用，效率即刻提升

相较于众多基础以及微调模型供应商，量知大模型平台不仅预置微调大模型，还自带了对话、创作、分析等全方位大模型应用，用户无需开发，帮助企业内部即刻提升工作效率。

（二）多套底座，择优使用

量知大模型平台适配了Baichuan2、LLAMA2等众多开源大模型并进行了微调，不同的模型擅长的任务及领域各有不同，针对不同的场景使用不同的微调模型，从而达到最优效果。

（三）API 封装，便捷开发

为了便于二次开发，量知大模型平台中的功能应用提供了对应的API 接口，使二次开发以及系统集成变得更加方便。

（四）训推一体，及时优化

针对非通用的知识以及推理过程中收集到的Badcase，量知大模型平台提供标注及训练平台，支持大模型的标注、训练及评估，在使用过程中及时优化模型，回答效果越用越好。

（五）信创环境，安全可控

量知大模型平台使用的服务器为全国产化服务器，使用海光CPU、海光DCU 芯片，实现“大模型+全国产化GPU集群算力”

的应用示范，化解人工智能芯片供应“卡脖子”风险的路径。

（六）安全使用，拒绝敏感

对于黄赌毒、宗教、政治等众多类型的敏感信息，量知大模型平台提供敏感信息检测功能，模型回答绿色安全，避免触碰法律红线。

三、项目进展及应用效果

针对金融行业客服中心，量知大模型平台在客服全流程的话前、话中和话后阶段都可以发挥重要作用。

针对话前阶段，量知可以支持问答知识的自动构建，也即可以从一段文字或非结构化文档中抽取问答对，并根据标准问答对给出扩展问。已应用该项技术来完成知识加工和梳理工作，大大节约了人工成本。另外，量知还支持根据一段或多段对话记录生成格式化的对话场景脚本，支持通过自然语言交互方式引导客户导入非结构化文档，逐步完成知识建模、知识抽取、知识融合的任务，实现知识图谱的自动构建。

针对话中阶段，量知支持利用大语言模型来进行客服的问答交互。这是使用大模型赋能智能客服自然而然的应用需求，但在落地中需要解决大模型对专业知识回答不准的问题。量知平台中包括了多种方案，比如使用外挂知识库或已有的问答库，采用检索加大模型的方法生成答案，并支持知识溯源；或者使用 NL2SQL，将自然语言转换为关系型数据库的查询，包括结合知识图谱和图

数据库来进行问答。

针对话后阶段，量知支持对话单数据进行更深入的分析挖掘，例如营销线索、潜在诉求原因等；也支持对海量语音、文字内容进行聚类和分类处理，生成热点和舆情研判结论等。每一通电话都会形成一个工单，由大模型来进行分类、根因分析、总结、给出建议措施等，减少人工处理环节，实现提质增效。

案例五：北京银行 AIB 金融智能应用平台

一、背景及意义

随着信息技术的快速发展，人工智能已经逐渐融入了各行各业。为了进一步提高工作效率，提升“双客”服务质量，北京银行 7 个部门组建数字化转型 12 号敏捷工程敏捷团队，深度应用 AIGC 技术，全力打造了金融智能应用平台(AIB，AI Banking)。

平台通过大模型、机器学习小模型、语义搜索等前瞻数字化技术，打通行内业务系统、办公系统、数据系统、操作系统，整合全行 80 项大模型服务、7 项 GPT 创作工具，以 GPT 对话方式，面向理财经理岗、大堂经理岗、客户经理岗、综合柜员岗、远程客服岗提供理财投顾策略、业务问题解答、组合金融资讯、客户营销话术、宏观政策研究、行业发展前瞻等实时在线支持。

二、技术方案及创新点

（一）利用知识图谱和搜索增强技术，提升大模型的可解释性

为解决大模型生成式内容在银行业务场景应用过程中可解释性问题，依托向量数据库、知识图谱、分布式大数据构筑北京银行金融价值矩阵，逐步形成覆盖业务营销、操作指引、资产配置、合规安全、财务分析全场景的金融知识图谱，通过正排、倒排多种索引融合的知识搜索引擎，为大模型生成内容提供了准确、

可靠、有效的知识来源，并在用户交互过程中给出引述来源，让大模型生成内容“言之有理”。

（二）集成行内应用插件，打造大模型对话机器人

为推动大模型与银行业务场景的深入融合，利用意图识别、语义检索、提示工程等技术，自主开发大模型插件框架，打通系统间数据、流程、制度壁垒，重塑系统流程和用户体验。以对话交互的方式，提供文案编写、代码生成、知识问答等基础功能，以及 RPA 任务执行、企业信息查询、系统调用等扩展功能。

（三）利用国产算力进行模型微调和推理，提升大模型的场景适配能力

基于国产算力完成 ChatGLM 模型微调，利用行内知识库，采用 LoRA 方式进行微调训练，冻结预训练模型参数，基于旁路矩阵来替代以完成模型微调，然后将预训练模型参数与旁路矩阵参数叠加，以实现最终结果输出，提升模型对金融场景的泛化能力。

三、项目进展及应用效果

目前，AIB 金融智能应用平台已面向全行 10000 多名员工开放，推出北银投顾、财报助手、智能客服、京客图谱、运营助手、数币银行、京行研究首批 7 款智能应用，让员工能随时随地使用智能化工具，快速掌握岗位所需知识和技能，将专家能力赋能至每一位员工，全面提升业务专业化水平。

图 3 北京银行 AIB 金融智能应用平台架构图

北银投顾服务理财经理近 1700 人，汇集了丰富的金融市场资讯、产品信息资源以及专业的投顾策略，可以根据客户的投资偏好和目标，量身定制个性化的投资策略。

运营助手服务综合柜员近 3000 人，覆盖外部监管政策、行内发文制度以及各类业务操作指南，为柜员提供操作指引。

京客图谱和数币机器人服务客户经理 3700 人，借助大模型强大的自然语言理解和内容生成能力，支持客户信息、营销商机灵活查询，助力精准获客。

智能客服机器人服务远程客服近 700 人，集成了个人业务、公司业务、国际业务等重点场景知识库，帮助客服专员精准回答客户问题，提升服务满意度。

财报助手和京行研究机器人，集成行内外众多权威机构万余篇深度报告，可为全行各业务场景提供研究支持和智慧决策参考。

案例六：上海银行基于开源大模型的智能办公助手

一、背景及意义

在现代银行业中，高效办公是保持竞争力的关键，而传统办公场景中重点存在以下两点突出问题：一是办公系统多且复杂，获取知识难度大且耗时长；二是日常办公中员工大量时间消耗在内容总结、公文写作等文档梳理、素材查找等方面，效率低下。为了提升员工的工作效率，基于开源大模型自主研发设计智能办公助手，从而让员工能够更专注于核心业务，提高工作效率。

二、技术方案及创新点

基于私有化部署金融领域大模型，通过思维链、One-shot 等一系列 prompt 技术进行提示微调，构建 prompt 指令库。并结合向量知识库，快速接入银行企业自有场景知识，直接学习企业文档库、搜索引擎现有资源，打造银行专属办公助手。

基于大模型建设的智能办公助手包含六大核心模块，各模块高度协同。第一模块是硬件和环境资源，它提供模型训练和推理所需的算力支持，尤其是 GPU 运算能力和大容量内存。第二模块是智能交互界面，负责语音或文本的输入输出功能，获得用户提出的问题并展示回复结果。第三模块是金融大模型，利用大模型技术针对金融领域进行预训练和迁移学习，以深刻理解业务语义。第四模块是快捷功能库，通过设计指令模板，引导大模型快

速实现知识检索、润色、推理等能力。第五模块是向量知识库，对文本数据进行向量化编码，以实现语义索引和相似内容检索。第六模块是数据支持管理模块，包含多个外部知识库，定时更新并且部分实时连接外部数据源，为模型提供最新知识。六个模块相互衔接，共同支持智能问答与办公写作。

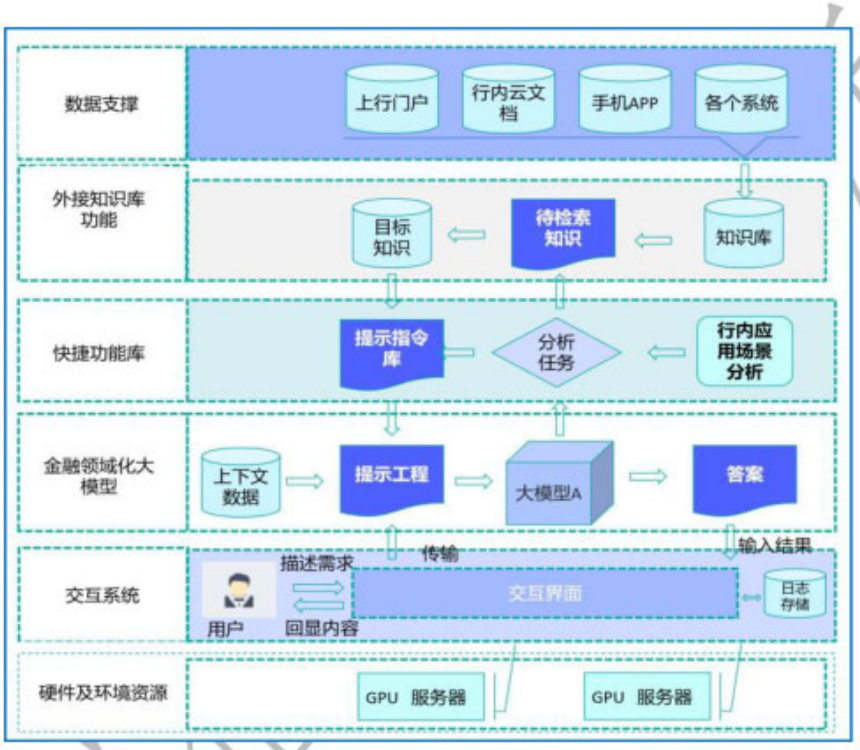

图 4 智能办公助手方案架构图

三、项目进展及应用效果

智能办公助手将为银行内部员工提供行内领域知识的精准搜索功能，并且进一步通过简单交互方式，来一键获取想要的问题的答案，实现“所问即所答”。同时智能助手还提供一系列工作辅助功能，包括周报内容总结、宣传稿/营销文案撰写、文章润色扩写、PPT 大纲生成、项目开发文档编写等等场景功能，大幅提升办公效率和员工交互体验。

案例七：国信证券辅助运营人员服务客户场景

一、背景及意义

国信证券股份有限公司(以下简称“国信证券”）综合型证券公司，业务包括证券期货经纪、证券承销、自营交易、资产管理、股权投资等。国信证券财富管理业务，目前是由运营人员人工服务千万量级客户，为客户提供个性化的咨询、运营、投资询问等服务。该业务面临由于客户海量、需求众多导致的运营人员运营服务效率低、客户体验不佳、运营人力消耗大等问题。因此，需要借助人工智能和生成式人工智能能力，提升自动化生成服务话术和客户指标数据，提升运营人员对客户的服务质量和效率。

二、技术方案及创新点

基于人工智能和生成式人工智能底层能力，辅助运营人员自动化生成客户服务话术，创新点如下:

（一）知识问答、行业分析、客户资产配置建议：基于大模型 Prompt 工程能力，进行相关财富管理领域的内容生成，可以替代投顾部分工作，进行资料查询和回答整理;

（二）资讯摘要：基于大模型 Prompt 工程能力，进行大量投研资讯摘要生成，辅助运营人员推送合适的简短信息给客户;

（三）行情快报、客户指标筛选：结合大模型 Prompt 工程和后台数据分析对接，实现以自然语言对话的方式进行信息查询。

三、项目进展及应用效果

项目正在进行中，目前已完成业务需求分析和大模型能力可行性验证，正在技术方案设计阶段。预计上线后，可辅助总部和各分支机构运营人员为客户提供个性化高效的知识问答、行业分析、行情快报、客户资产配置建议、资讯推送摘要、投诉建议反馈等服务。

案例八：蚂蚁金融大模型应用-支小宝 2.0

一、背景及意义

智能理财助理旨在协助个人更有效地进行资产管理与配置。当前智能理财助理已在智能客服、个性化推荐、风险管理等方面取得显著进展，但要进一步替代人工金融理财专家，仍面临如金融信息过载、复杂金融任务拆解、专业术语晦涩，缺乏个性化投资建议等一系列挑战。

ChatGPT 的出现引发了社会各界对于大模型技术的广泛关注和资源投入，并由此推动了大模型技术的快速发展和产业应用，其中也包括应用到智能理财助理场景中，通过智能化升级解决前述挑战。然而，由于金融行业的专业性、严谨性、合规性等特点，难以直接将通用大模型直接应用到金融应用场景中，一般需要结合金融领域知识和数据进一步微调形成金融行业大模型，才能让大模型对金融行业的特有的术语和规则有更深刻理解，输出结果满足金融领域极高的准确性和可解释性要求等。

支小宝 2.0 是蚂蚁集团基于自研的金融大模型打造的新一代智能理财助理，它致力于为用户提供透明可信赖的金融服务，提供高度智能化的专业投资建议，实现亲和力十足的陪伴和流畅的交流。支小宝 2.0 强调金融产品的合适性和安全性，旨在解决信息鸿沟，提升用户体验，推动金融领域的不断创新和进步。

二、技术方案及创新点

蚂蚁金融大模型着重打造了自身在金融领域的“知识力”、“专业力”、“语言力”，以及结合可信围栏技术实现的“安全力”，确保大模型技术可以安全合规地应用到金融场景中，提升蚂蚁金融服务的智能性，打造出智能理财助理新产品支小宝 2.0。

蚂蚁金融大模型主要技术方案如下图所示。除创新性地提炼和打造出金融大模型的“四力模型”，蚂蚁还构建了金融专属任务评测集“Fin-Eval”，从认知、生成、专业知识、专业逻辑和安全性五大维度设计出 28 类金融专属任务，对金融大模型能力进行全面评估。

图 5 蚂蚁金融大模型全栈技术布局图

三、关键问题及解决方案

“知识力”方面，针对通用大模型专业金融知识缺失的问题，蚂蚁金融大模型引入了可信、多元、实时的泛金融内容和知识，

构建起千亿级别 Token 级别的通用+蚂蚁专有金融语料，并结合模型知识注入与信息检索技术，赋予支小宝 2.0 兼具广度和深度的智能理财助理知识力。

“语言力”方面，金融行业的复杂性与用户期望的简明性之间存在着巨大的差距。为了弥合这一鸿沟，支小宝 2.0 采用了多项策略：（1）扩展上下文窗口，将上下文窗口扩大至 32K，以深入理解用户意图，实现更连贯的多轮对话；（2）对话仿真工具，为解决模糊意图和多重问题，支小宝训练了对话仿真工具，模拟专业理财专家与用户的对话，提升其理财领域语言能力；（3）动态风格调整，通过大规模金融指令数据集训练，实现风格切换功能，应用不同的风格算子生成专业，标准和通俗三种用户风格。

“专业力”方面，蚂蚁通过整合业界前沿技术能力，包括多目标运筹优化、动态图计算和异构图表征学习等，沉淀出数百个金融接口工具，解决行情解读、产品评估、行为分析等不同专业任务。作为智能中枢，支小宝 2.0 能够智能地识别用户的模糊需求，调用适当的金融工具，使金融领域的“专业力”无缝融入每一个用户的日常需求中。

“安全力”方面，针对通用大模型在金融领域应用面临的安全性问题，蚂蚁聘请超过 100 名金融专家对生成内容在隐私保护、合规表达、上下文关联等多个维度进行评估，并使用基于人类反馈的强化学习让大模型对齐金融业务的合规需求，并通过后置校验的方式保障安全底线，在数据、模型、输出等不同环节建起了

多重“金融围栏”。

四、项目进展及应用效果

近半年，基于金融大模型技术，蚂蚁已对原智能理财助理产品支小宝进行全方面升级，当前的支小宝 2.0 已有了兼具广度和深度的金融知识、专业金融工具的调用能力、个性化的表达能力，以及安全可信的围栏能力。据评测，支小宝 2.0 在金融意图识别上准确率达到 95%，投资情绪识别准确率达到 90%，金融资讯总结和事件推理等达到分析师水平，目前可提供选基、行情、配置等 290+理财服务，选品、规划、核赔等 30+保险顾问服务。