银行业生成式AI应用报告（2023）

何大勇、谭彦、孙蔚、窦德景、廖明、冯志宇

2023年8月

目录

银行业生成式AI应用报告（2023）

引言

自2022年底ChatGPT-3.5发布起，生成式AI（Generative AI）的话题热度持续走高。该技术并非只停留于概念阶段，而是已开始加速影响着各行各业，无论是技术本身还是应用演进的速度均十分惊人。银行业本就是最早应用传统人工智能技术的领域之一，如何借力新技术、加速数智化转型，构建差异化竞争力，是领导层当下必须深入思考的命题。对银行而言，生成式AI技术与传统AI技术到底有何不同？生成式AI能否为经营管理带来新的价值释放场景？如何推动场景快速落地？规模化应用又需做哪些准备工作？本报告将呈现我们对相关问题的思考和建议。

1. 变革已至：理解生成式AI的能力

1.1 直观体验：生成式AI带来何种不同？

让我们来畅想一下，一家银行的零售理财客户经理小王的一天。

今天是小王入行整五年的日子。早晨到支行后，小王做的第一件事情便是打开最近部门新安装的AI助手——小智，开始在对话框中录入问题。

“小智，请打开我今天上午的计划表。”

“好的，小王。这是今天上午的计划表，你计划给三个客户打电话，这是预约的时间和他们的电话号码。”

今天要沟通的第一个客户是张女士，印象中这位客户小王迄今为止只联系过一次，并不太熟悉，于是他在系统中勾选了“查找前期通话记录文稿和微信聊天记录”和“查找行内交易数据”两个选项，并在对话框中连续问了三个问题，“小智，我上次与她联系是什么时候？她目前都持有什么投资？最近是否在本行买卖过产品？”

系统只用了两到三秒就给每个问题提供了快速、准确的回答：“根据记录，你上次联系她是一个月零三天前。她最近一个月没有交易过任何产品。目前她的持仓只有两笔定期理财。”

小王继续在系统勾选“查找前期通话记录文稿和微信聊天记录”，问道：“小智，她之前联系时表示对什么产品有兴趣？”

“她表示过对债券型基金和混合型基金有兴趣”，系统很快回答。

“那上次是否有给她推荐债券型或混合型基金产品？她为什么没买？”小王让小智继续在前期通话记录中找答案。

“上次推荐产品时，她表示没有闲钱，将在一个月之后有奖金入账，会考虑进行配置。”

那现在不正是推荐的好时候，小王心想。于是，他立刻勾选了“产品推荐引擎”选项，继续问道，“小智，目前本行在售的混合型基金中，哪款比较匹配张女士的偏好？”

几秒钟之后，小智给出了三个答案，“根据模型的预测，目前有三款产品，分别是AAA、BBB、CCC三只基金，较为适合张女士。”

小王想起来，CCC这只基金昨天刚开始发售，但他还没有时间仔细学习该款产品的文件资料，他于是赶紧勾选了“产品介绍”选项。对话框中立刻调出了该产品长达10页的详细介绍。

由于没有时间仔细阅读这个冗长的PDF文件，小王随即勾选了“产品比较”选项，问道，“小智，请用一段文字总结这款产品的主要亮点，并跟BBB基金进行比较。”很快，小智在对话框中返回了CCC基金的三个亮点，以及与BBB基金的两个主要差异。

通过这段一到两分钟与小智的对话，小王认为自己已经准备好，于是拨通了张女士的电话，开始营销。通话很顺利，张女士表示自己恰巧在选择混合型基金产品，并对小王的推荐和专业的介绍话术感到满意。在通话期间，当张女士好奇问到另外一款DDD产品为什么表现不好，小王在对话框中同步将问题抛给小智，也得到了一段比较扎实、多角度的市场分析解释，及时传递给张女士。

下午四点半，小王在助手小智的帮助下，比计划提前完成了今天六个客户的沟通，均取得了较好的营销进展。本想休息片刻，去学习一下市场动态，结果领导过来布置了一个紧急的客户分析任务，并要求在下班前提交报告。原来，最近支行财富中收业绩有所下滑，领导让小王这个团队长分析他所在的小组共同覆盖的AuM在300万到600万的客户近期在产品交易上的变化趋势。

“幸好现在不用找人帮忙去系统找数、拉数了，也不用自己在Excel搞透视表了，否则今天下班前怎么可能交得了……”小王一边心里嘀咕，一边再次打开系统对话框，勾选了“客户数据分析”选项，开始一步步盘问小智，“小智，请把ABCDEF这六位客户经理所管户的AuM在300万到600万的客户都找到。”“这些客户过去一个月的日均AuM变化如何？”“这些客户在一个月前的合计产品持仓分布发生了什么变化？”“非存AuM下滑最快的客户画像是什么？”“主要退出的产品的市场表现有何变化？”……一步步地，小王层层深入地对着系统问了十几个问题。在每一步，系统都用几秒钟就返回了相关数据，并自动形成了线图、饼图等形式的图表，更直观地显示数据变化趋势。最终小王把相关图表集中下载到Word文档，并让小智基于图表内容自动形成了初稿总结，最后自己再用半小时完善了整体的分析报告与行动建议，顺利在下班之前交出了报告。

以上客户经理小王的一天，就是AI赋能的一个典型示例。在生成式AI出现后，这个示例已不再遥远。实际上，场景中的许多片段目前均已实现了概念验证。例如，让机器从非结构化的通话记录文稿中，快速准确找到相关信息并以问答输出；让机器阅读长篇累牍的产品或资讯文档，自动提炼要点；再比如，让机器在自然语言的指示下，自己去数据库抓数和作图，打通Text-to-SQL的最后一公里，实现数据分析和制图的自动化。

从这个例子，可以直观地感受到生成式AI带来的核心体验的变化：客户经理小王不再需要到CRM系统的层层功能菜单中去逐一查找并手动汇总分散的信息，也不需要去逐页阅

读海量的通话记录、产品类文档或市场资讯类资料，他甚至不需要熟练掌握各类分析性工具（比如SQL取数、PPT制图、Excel透视分析等）。由于有小智这个不知疲惫且能开展基础信息加工分析的助手，客户经理小王只需要问“对的问题”，就能高效获得精炼的、极有针对性的营销知识和客户洞察，以供进一步决策或行动。生成式AI把小王从占用时间的案头类工作中解放出来，使小王可以将时间和精力尽可能多地放在与客户的直接交互上，同时还能体现出较之前更专业的水平。这就是生成式AI的魅力。

1.2 理解根因：什么造就了生成式AI的强大能力？

人工智能已发展多年，历经专家模型、机器学习、深度学习多个阶段。今天生成式AI之所以火热，是因其相较传统AI，在**“对话”**与**“创造”**两类能力上实现了根本性的突破。

- **就“对话”能力而言**，过去的机器在回答问题时往往缺乏对上下文的理解，导致答案相关性较低，表达机械化；而如今的生成式AI能够理解更长的上下文，并进行拟人化的思考和回答，与人类的对话沟通也更自然；

- **在“创造”能力方面**，以往的机器只能按照预设任务（如分类、数值预测）输出答案；而现在，生成式AI能够自动生成自洽的图形、文本甚至代码，具备优秀的内容创作能力。

那么生成式AI背后的大模型，又是如何形成了突破性的对话和创造能力？这离不开科学和工程的双重进步。

一是科学的进步，即算法的突破。AI算法的本质在于特征提取。基于Attention Layer的**Transformer技术**，是一种新的模型架构，能更好地提取**“全局”特征**，因此模型的效果更好。Transformer技术使机器能高效捕捉海量语料中一个个词之间的关系，或者海量图片中一个个像素之间的关系，使得**大量的知识（本质上表现为词语之间的关系）能被封装在训练好的模型中。**由于该模型架构强大的能力，在2018年Transformer技术出现后，三分天下的AI应用领域（计算机视觉、语音识别和自然语言处理）逐渐形成大一统趋势。以前各个领域有一套适配其应用场景的模型架构，如今Transformer可相对较好地处理各类场景的问题。

二是工程的进步，即超大规模的算力和数据的支持。由于基础设施的进步（高算力芯片、高速网络），模型的训练规模较之前深度学习阶段有了数量级的显著跃升。深度学习时代的模型参数通常只有百万量级，只能训练几亿级的文本且还需要人工标记；但以ChatGPT为例的大模型参数可达1,750亿，能训练数万亿级的单词文本且预训练不需要人工标记。**正因为训练的语料和参数在量级上的突飞猛进，使大模型体现出的能力远超以前。**同时，也因为足够大到能训练和封装几乎全科领域的知识，大模型能表现出**很强的“通用性”能力**，即一个大模型可以在结合精调后运用到多个完全不同的场景。

2. 大有可为：银行业应用场景丰富，价值释放可期

2.1 价值创造逻辑：替代人、赋能人

生成式AI在银行业的应用，从价值创造逻辑上可分为两大类：

一是替代人：生成式AI可以替代人，开展大量重复性较高、简单基础的任务，如处理文本的要素提取、处理进件、识别异常项、生成基础数据分析、生成标准化内容等。这能够释放运营类人力资源，实现降本增效。

部分场景下，生成式AI还可能取代人，催生全新的业务模式。例如交易撮合的场景，由于很多场景的交易要素非结构化，且需要多轮交互，通常需要人来协助开展撮合，但借助生成式AI，未来买卖双方都可能只与AI界面进行对话磋商，而不再需要人作为中介进行撮合。

二是赋能人：利用生成式AI的“对话”和“创造”能力，可让AI成为助手，有效放大关键节点的“人”的产能，尤其是客户经理、财富顾问、产品经理、投研经理、信审经理、市场营销人员、编程开发人员等角色。本报告开篇赋能客户经理的例子就是一个典型示例。一方面，通过对话式学习的方式，生成式AI可以更好地“培训”这些专业人员；另一方面，在关键的展业过程中，生成式AI可以有效整合关键信息及素材，助力相关专业岗位的人员，更快做出精准有效的判断，展现出更有针对性的客户互动技巧，或更快速地产出高质量的交付物（代码、设计文案、报告等）。通过机器助手的加持，代表银行核心竞争力环节的单人产能有望大幅提升。

赋能人不仅仅是体现在专业内容的形成上，还可能体现在基础管理环节。例如，以前项目开展PMO管理，需要项目经理每天与多个角色共同开会来对齐进度，并维护一张集中显示进度和问题的大表。未来，借助生成式AI，有可能实现由AI与各个角色实时开展简单的对话来交流进度，最后由生成式AI提炼对话中的要素，自动形成项目管理看板并自动识别和提示风险。在AI的加持下，以前一个项目经理只能管理两到三个项目，未来可能变成一个人可同时管理十几个复杂项目。

2.2 价值释放空间：贯穿银行前中后台，规模化降本增效

如图1所示，生成式AI在银行业的应用场景可贯穿前中后台各个环节。银行的每条业务线、每个职能，都有可能找到生成式AI的应用场景。若能在银行业实现规模化应用，有望带来可观的降本增效收益。BCG曾以一家拥有约两万名员工的区域性国际银行为例，初步梳理了该银行前中后台相关部门应用生成式AI的潜力和效益，预计在首年即可为该银行节省约1.5亿美元的成本，占整体薪酬总包的7%左右（参阅图2）。

图1 | 生成式AI的应用场景贯穿银行全产业链的各个环节

举例、非穷尽

I. 市场和销售	II. 渠道和运营	III. 产品开发	IV. 投顾服务	V. 客户关系管理	VI. 风险合规
更精准**获客**、客群**细分**	智能化**进件处理**	支持开发**编程**和**开发文档生成**	提供内容以**支持投资顾问与客户互动**	高度拟人化的**客服机器人**	合规文件知识库，支持**对条款的高效查询**
自动**生成**广告等**营销文案**、形成**超个性化营销**内容/界面	为**贷前贷后审查**文件自动生成**要点总结**	**定价**和费用**个性化**优化	投资**报告**和**研究要点**自动提炼	**知识库**搜索	形成**合成数据集**，以提升欺诈检测
客户**情绪分析**	自动化查找和录入**客户初始化信息**	为客户画像和预测模型，生成**合成化数据**	以对话方式提供**投资者教育内容**	支持会议/通话**文档生成和提炼**	**合规**监控**报告、合同**文本**自动化生成**
支持营销人员**话术培训**	新客登录注册的**自动化客服支持**	协助产品团队**发现新兴产品趋势**	根据投资者偏好不同生成**个性化配置建议**	支持客户**数据分析**、客户**洞察提炼**	信贷**风险预警**

图2 | 以某国际银行为例，规模化应用生成式AI将释放显著的降本增效收益

该银行规模约2万人，薪酬总包超过21亿美元

组织与职位	总人数及占比	总薪酬包及占比
销售（分支行+客户经理）	34%	15%
1 2 分支行（销售+服务）	30%	13%
2 客户经理	1%	1%
其他（例如CDL、教育贷款）	3%	1%
投资与经纪	5%	7%
2 FA/财富顾问等	3%	4%
其他（例如投资助理）	2%	3%
贷款	9%	15%
2 客户经理	1%	2%
5 按揭贷款主任	2%	4%
其他（例如研究、投资组合经理）	6%	9%
3 技术/IT	11%	18%
4 运营与风控	10%	7%
1 运营（呼叫中心）	8%	6%
4 风险管理与合规	6%	9%
6 营销	2%	4%
7 财务、会计与审计	2%	4%
人力资源	1%	2%
司库	1%	2%
8 法务	0.3%	1%
其他	9%	10%
总计	~2万人	~21亿美元

预计规模化应用生成式AI首轮能实现一年1.5亿美元降本

	影响的组织与职位	生成式AI抓手	预估降本规模
1	运营与销售—呼叫中心、分支行服务人员	GenAI聊天机器人和知识中心，减少搜索时间	2200万美元
2	销售与业务—客户经理和分支行服务人员	提供销售和管理解决方案，通过搜索以前的对话并根据过去的互动生成见解	3100万美元
3	技术/IT—工程师和测试员	提供代码生成/完成工具，实现小型问题的自动修复	6300万美元
4	运营与风控—信贷分析和反腐评估岗位	进行查询分析，将内容预填充到报告模板中，使非技术团队有更多生产力用于通过访问数据和分析报告来决策	700万美元
5	信贷—按揭贷款岗位		1300万美元
6	营销—创意类营销和社交媒体岗位	根据反馈和消费者测试快速迭代，建议创意图片和文本	1000万美元
7	财务—财务分析人员	分析数据和生成分析报告，更容易让非专业人员理解	400万美元
8	法务	简化基本的律师助理任务，如法律研究、文件起草审查	400万美元

来源：某银行案例。

3. 落地可行：局部速赢已有切实路径，规模化应用还需体系性规划

3.1 优选场景：平衡收益和风险，与传统AI充分结合

在应用探索初期，各家银行通常优选少量场景先行试水、循序渐进。**在选择场景时，要平衡考量收益潜力、风险、实施难度。**同时，最早落地的试点场景，还需考虑其能否在**组织准备度诊断、方法构建、信心构建等角度形成示范效应。**

选择场景时，不能只是简单定位“业务环节”，粗放地决定到底是应用在财富管理的营销环节，还是应用在公司金融的授信审批环节。**对场景的细分和选择，需要具体到机器的角色和需解决的问题类型。**

以下几个问题，是场景定位时通常需面对的权衡选择。

1） 归纳、分析还是决策？

即使是同一个业务场景，取决于AI的角色定位不同，其收益潜力和实施难度也不尽相同。以商业银行的授信审查场景为例（参阅图3），按照角色的要求从易到难，AI的角色可以是只对事实性信息进行自动抓取和归纳的初级研究助理，也可以是对行业和企业到底好不好能形成严谨分析的分析员，甚至有可能成为直接给出某个申报项目到底能否批、批多少、批什么条件的授信策略师。

图3 | 在银行授信尽调审查的场景，机器的角色定位不同，实施复杂度逐渐提升

	复杂性和价值不断提升		
使用场景	1 “AI作为**初级研究助理**”	2 “AI作为**行业分析员**”	3 “AI作为**授信策略师**”
输入信息	结构化和非结构化数据，包括**市场动态信息、公司公开报告、内外部数据库、客户经理采集的书面资料**	数据结合银行**特定的行业和企业的分析框架、特定的参考指标值**	结合**2的分析产出、结合历史上我行和他行的类似客户的授信案例、结合我行当下的授信策略和标准**
潜在的应用场景示例	生成式AI从给定的数据源**检索事实性信息**，进行**有效汇总** *例如，让机器“**总结某公司的股东结构和过去3年的变化情况**”*	生成式AI**给出观点和判断：行业好不好、客户好不好** *例如，让机器“**评价储能行业是否发展向好、或某企业的财务报表是否健康**”*	生成式AI给出授信建议（**准入建议、额度建议、放款条件建议、出险概率预测、贷后管理要求**） *例如，让机器“**给出某客户的动力电池新建产线项目融资贷款的授信建议**”*
价值创造	**赋能客户经理**：AI在几分钟内生成高质量的事实性总结段落，大幅降低客户经理自己查找和总结“事实性”信息的时间，**使基础信息的总结更快更准**	**赋能客户经理+审查人**：AI提供**评估标准**，帮助真人在尽调报告和审查报告中**提升分析的质量**（提供了更完善的分析维度、指标），使其**分析更扎实、有理有据、更专业**	**赋能审批人**：AI协助给出授信策略建议的参考，提供更为**客观、详实、综合**的判断根据，**有利于形成更快、更正确、解释性更强、更一致的授信决策**

最后一种“授信策略师”显然代表了最高的技术应用水平，但这并不意味着就是当下优先级最高的必选场景。一般而言，涉及大额交易的复杂投融资决策，往往都不是100%的规则导向、有标准答案的，在基于客观、全面、深入的理据分析基础上，最后的决策往往还会融入当下团队的经验、风险偏好、非经济回报类收益等主观考量因素，机器难以捕捉。另外，机器给出的策略有利于保持每次决策的标准一致性，但这种一致性在历史决策并不明智的情形下（如历史决策数据中可能包含大量由于太过谨慎而错失好机会的例子），反而可能导致非最优决策的更长期延续。

反之，在前两种角色定位上，生成式AI的价值释放潜力短期来看可能更为显著。对公客户经理团队的专业水平、学习意愿通常参差不齐，在申报授信方案过程中搜集整理客户及行业的信息也通常需耗费大量时间。而当下，银行的授信要具备竞争力，一方面要“快”，另一方面要“专业”。作为“分析员”的生成式AI，可以直接构建基于专业行业分析框架、评估标准的初稿，大幅减少客户经理自己从0到1学习理解新行业和加工处理基本信息的时间，客户经理写出的尽调报告的平均水平在短期有望显著提升。

2) **面客还是对内？**

在对成本收益进行考虑时，不仅要考虑直接的运营人力节约的收益，还要考虑因潜在风险可能导致的额外间接成本。

例如，在零售远程银行的业务场景中，一种应用模式是让AI直接“面客”，让AI直接成为外呼的营销坐席，代替真人坐席与客户开展营销互动；而另一种应用模式是对内赋能，即AI成为营销坐席的助手，在真人坐席与客户互动时，为其实时提供更详细、专业的信息支持，或提供“回答草稿”，最终仍由真人坐席来把握实际对话内容。前者相较后者，看似直接节约运营成本的效果更加显著，但可能因AI互动质量的不确定性，导致部分客户体验不佳、反而投诉增加。

由于银行业在开展对个人客户的销售时，往往较其它行业面对更严格的行为监管和内控要求，在“面客”场景首先试水生成式AI，实际挑战可能更大。

3) **对目前已在应用的传统AI，替代还是结合？**

在大模型出现之前，银行已有非常多的传统AI模型的应用，例如，在线上小额个人贷款领域由机器实现智能放贷，或由机器生成理财产品推荐建议，或由机器判定高潜获客名单或客户流失率。既然Transformer技术有大一统趋势，且支持多种任务类型，那是不是意味着，所有传统AI模型的应用场景，都需要用大模型来重新做一遍呢？

并非如此。实际上，传统AI模型在不少场景已体现出很强的能力，其与大模型在相当长一段时间内会共存。如图4所示，两者适配解决的问题类型并不相同。从单一任务来看，传统机器学习模型较为擅长的是需要较强解释性、需要进行量化预测结果的任务。而大模型擅长的是产出业务内容、答疑解惑式场景的任务。

图4 | 传统AI和生成式AI适用于解决不同类型的问题

未来，大模型和传统的单任务模型之间，更可能是强强结合、同时使用的关系。

两者间可能是总分关系：大模型有望成为问题解决的中枢“大脑”。通过大模型，可将一个相对复杂的问题拆解成不同的步骤，每个步骤去调用不同的单一任务传统模型，最后大模型再将不同单任务模型的输出进行串联整合。以前，业务流程的数字化往往体现在单一任务上的数字化，但连接不同活动的流程仍需要人的相当介入。在大模型的帮助下，未来人们可以利用自然语言，将长线的目标任务拆解思路教给有推理能力的机器，然后由机器来完成一个完整业务问题的全流程数字化，使得业务流程数字化从“任务数字化”进阶到“目标智能化”。

两者间还有可能是串联关系：针对特定任务，传统AI和生成式AI的串联使用可进一步提升预测的准确性。例如，在金融反欺诈场景下，大模型可以通过创建额外的合成数据点，来解决欺诈训练数据稀缺性的问题。国外银行已有实践表明，利用合成数据集，传统AI反欺诈模型的预测性能会进一步提升，误报率会降低，甚至还可识别出新的可疑活动类型。再比如，在个人金融产品的营销推荐场景下，大模型可代替人工、有效读取各类非结

构化数据（如客户谈话记录），自动提炼为结构化的客户画像标签，动态持续地输入到传统的产品推荐AI模型中。由于输入的变量更加丰富、准确，模型的推荐结果也能更个性化、更精准贴近客户的需求。

3.2 巧用方法：三大抓手，让机器的答案更专业、更实事求是

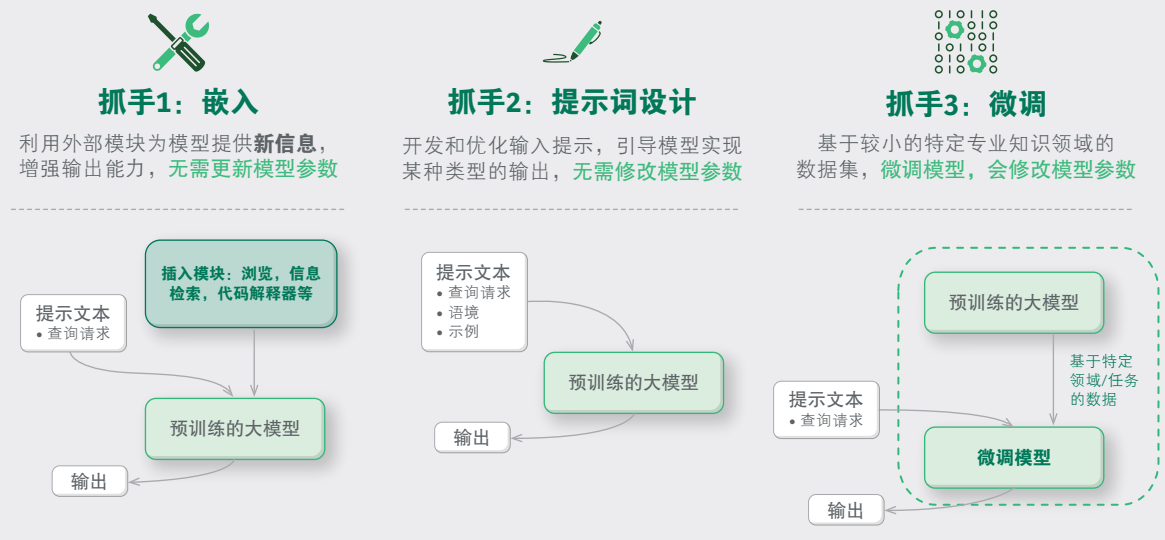
从0到1训练大模型的成本和难度高企，因此大模型开发能力未来将集中于少数专业机构。对银行而言，主要的落地路径是直接应用市场上已训练好的大模型。但**拿来即用，需有正确的方法来应对两大挑战**，确保答案质量可靠。

- **挑战一是“说真话”：** 生成式AI的“知识”是基于预训练时已经封装好的“知识”，它的知识并不能及时更新，更没有融入各家银行自身掌握的专属数据和信息。因此，直接与其对话时，给出的答案容易出现偏离事实、甚至“胡说八道”。

- **挑战二是“说专业的话”：** 大模型虽然有很强的“泛化”知识能力，但在某些专业领域的知识并不一定能掌握得好；每家金融机构内部一些特定的表达要求、表达风格在预训练时也未曾捕捉。

那么，如何让模型产生的答案更能契合银行自身的需求呢？有三个关键抓手（参阅图5）。

图5 | 提升机器回答质量的三个抓手



抓手一：利用嵌入（embedding），使大模型能基于给定的数据库来生成答案。例如，银行可以将理财经理与客户A的前期通话访谈记录、客户A在银行渠道的理财产品买卖交易历史记录，都切割编码放入到向量数据库。当理财经理问系统“客户A的理财偏好是什么？”这个问题时，系统首先到向量数据库检索出相关信息片段（买卖记录中能显示出产品偏好、对话记录中也有表示出的产品偏好），之后再将理财经理的问题和搜索到的信息片段传送给大模型，由大模型整合形成最终的对话式答案。

抓手二：利用强有力的提示词设计，使模型给出契合专业性要求的准确答案。问题怎么问，即提示词是什么，对模型的输出结果有决定性的影响。一个简单的例子，同样问ChatGPT“如果我老板让我周六加班，我该怎么办？”，当加入半句话“请你以一个劳动法律师的角色来回答”时，问题的答案就会完全不同（参阅图6）。

图6 | ChatGPT因问题不同而给出完全不同的答案

当然，提示词之所以目前已称之为工程，甚至一个新学科，绝不只是体现在单一问题上应用角色设定、提供背景等简单技巧。**在以提示词进行应用开发时，可利用LangChain把问题解决的长线逻辑和相关例子全部融入进去，使大模型能按照预设的步骤、思考链路、回答格式来产生答案**，从而形成更符合专业领域要求的产出。目前，领先大模型如GPT-4的提示词已可以容纳32K（约2万字）的输入，能支持相当复杂的、含多轮“问答对”的问题描述。从落地所需的能力来看，提示词工程对团队的要求，根本上是问题拆解、流程梳理的能力，对软件开发能力的要求较为有限。

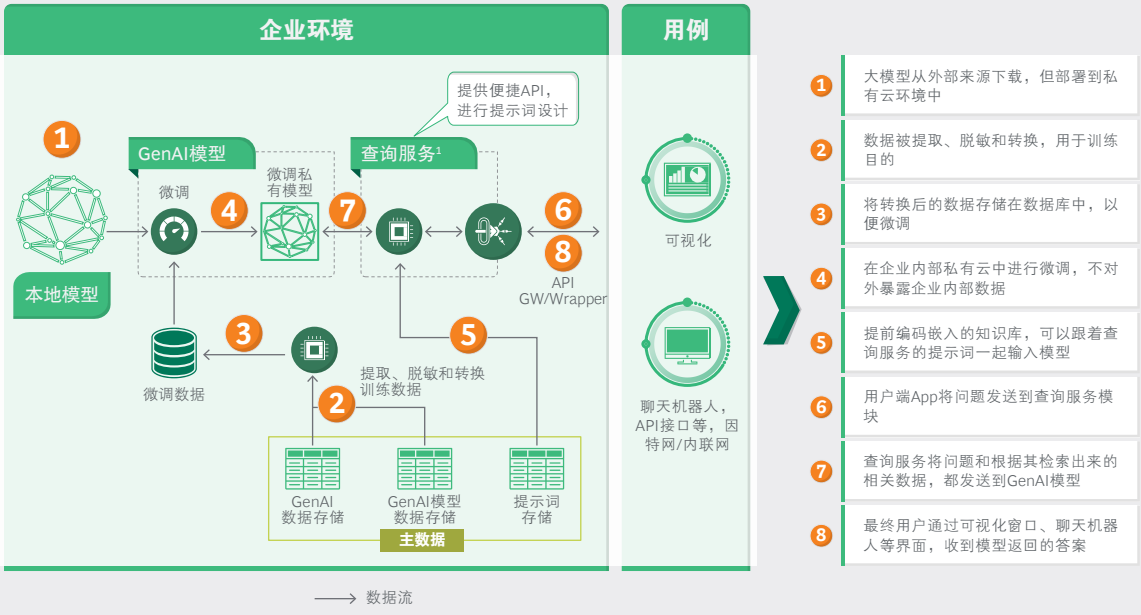
抓手三：直接对开源大模型进行指令精调。精调模型会直接修改模型的参数，使得模型在有限提示词下，一次性生成答案的质量更高（zero-shot质量更高）。不过，即使模型进行了精调，若要处理的是需拆解多步逻辑链路的问题，仍需要在提示词上精心设计，只靠精调模型难以解决所有应用诉求。另外，精调模型需要更多的示例（数量明显多于提示词设计所需的高质量示例要求，且需要人工标注）和一定的算力资源（需高算力芯片），同时还需要团队有擅长做模型的数据科学家，这都使得场景应用开发的整体资源投入更高。

因此，若应用方能选择的大模型本身性能足够强大领先，且处理的问题需要多步逻辑拆解，通常优先考虑直接用“提示词设计”而非“精调”来进行场景应用的开发。反之，针对不需复杂逻辑拆解、相对直接的“内容问答”类场景（如客服、知识库、培训等），银行通常有大量现成的高质量“问答对”的数据积累可支撑训练，那么精调不乏为投入产出比更高的一种路径。

3.3 夯实技术：合理部署、多维选型、全栈升级

从部署上看，银行等金融机构对数据安全性有非常严格的要求，**这意味着模型的精调、模型的应用都很可能需在本地进行，保证专属数据“不出行”。**图7呈现了国内银行未来可能采取的主流部署模式。

图7 | 国内银行未来更有可能采取的主流部署模式



[1] 用于编排和/或提供API访问的中介服务。

从技术上看，**若要在全行前中后台都规模化应用大模型，多项软硬件能力也需相应升级**（参阅图8）。

图8 | 技术架构蓝图

来源：BCG 分析。

首先是算力。即使不从0到1训练大模型、只是进行精调，也需要一定的高算力资源。另外，若未来有成百上千个AI应用同时在组织内运行，对硬件基础设施的算力、稳定性也提出了更高要求。

其次是模型。大模型的选型尤为重要，但这并不是一件简单的事情。**模型是不是越“大”越好？并不一定。**比如Meta的LLaMA（2023）模型参数为650亿，比GPT-3（2020）的1,750亿参数量小一个量级，但性能却并不比后者差。训练语料、训练方法、训练时长，都会对模型效果有影响。另外，近期有斯坦福大学的研究表明，“三个臭皮匠顶一个诸葛亮”，即在特定场景下，三个较小的模型的共同应用可能比一个很大的单一模型的表现更好。

在具体给某个场景进行模型选型时，**业务团队的输入很关键。需要有充分的测试例**，来验证模型在**准确性、置信度、鲁棒性、公平性、毒性、效率**等多个维度的表现。针对不同类型的场景，评估前述各个维度的测试例也不尽相同。除了模型自身的效果表现外，模型的选型还要考虑与平台合作方相关的其它因素，例如**商用授权的要求、能否支持精调、安全性、兼容性**等。

从规模化应用来考量，**未来银行需要引入和维护一个模型库**。如之前所述，很多场景需要大模型和传统AI模型的并用，因此模型库需要纳入多种模型。**即使是大模型，目前各家开发商仍然处在一个快速迭代、你追我赶的阶段**，甚至商用授权等商业模式都在快速变化（例如7月20日，Meta宣布Llama 2支持免费商用）。因此，银行作为应用方，在现阶段开展试水时，需对大模型的选择**保持开放态度**，可在以一家大模型为主的基础上，也仍然积极尝试市场上其它领先的开源模型，持续寻找场景与模型之间的最佳适配组合。

除了模型选型之外，围绕规模化应用开发，还需要构建其它多项能力。

一方面是3.2小节中已提到的**基于提示词设计进行应用开发的能力**。要发挥出大模型的应用潜力、缩短应用开发周期，就需要让不具备软件编程能力的业务团队能更加自主、充分地开展基于提示词设计的应用开发，而这需要搭建一个全部基于自然语言交互、支持“托拉拽”的开发工具，支持业务人员进行问题拆解、定义“问答对”。银行未来可以设立一个生成式AI工作室团队，专门为业务团队的应用开发搭建工具、开展培训。

另外，不同场景的提示词设计中的“问答对”是有可能复用的，甚至是需要全行有统一定义，才能取得更好的效果。这些“问答对”的沉淀、背后的代码管理也需要有相应的机制。在初期进行部分典型场景试水时，银行的技术团队就要开始思考，如何构建常态化、针对“提示词设计”开发的研发管理机制。

另外，风险控制也很关键。如何形成高质量的测试能力，如何进行内容审查，如何持续基于应用形成反馈，这些方面也都需要构建新的管理流程和标准。

3.4 重塑体系：规模化应用需遵循10/20/70原则，技术之外更关键

我们认为，生成式AI在银行业规模化应用的落地，是一个体系性工程，其成功与否会遵循“10/20/70”法则，即：10%是模型，20%是整体IT能力升级，70%是**业务与组织的转型**。

当大量运用AI时，需要有操作行为规范，引导员工进行合适的信息输入、并对机器输出进行合理地判断和使用。围绕质量管理、风险监控、责任认定等，也需构建匹配的管理机制。银行还要构建负责任AI体系，在精调和应用开发时，尽力确保公平性、可靠性、透明度或可解释性、隐私安全、可问责等目标。

另外，生成式AI的大范围应用也将变革企业的岗位和人才结构，以及人才的选拔培养体系。从人才结构上来看，专业技能岗位的基础级别员工的需求量可能会减少，而质量管理岗位的人员需求可能会增加。从人的能力素质要求来看，员工之间专业技术的方差可能会伴随机器的赋能而变小，但是在问题定义能力、问题解决能力方面的综合要求会比以前更高、更能拉开人与人之间的差距。围绕人才要求的改变，企业的人才培训体系、人才晋升的路径和标准也都需要相应变化。

4. 快速行动：银行需由点及面、敏捷推进

生成式AI的浪潮已席卷全球，国内外诸多金融机构均已开始加速应用场景的探索。虽然目前全市场看似兴致很高，但很多机构实际还未下定决心，热情和动力还只是停留在对各类“首个应用”名头的跑马圈地。

各家银行应充分意识到，这次的新技术不只是噱头，而是有望切实带来革命性生产力提升的范式变革。对于生成式AI的探索，银行需要有长远的眼光，开展体系化的顶层规划，需要联合相关业务和科技部门协同努力，推动规模化应用的分步落地。

具体而言，可分三个阶段，由点及面、敏捷推进：

- **第一阶段，少量场景的概念验证和局部落地：**选择重点应用场景，快速完成概念验证（POC）、构建最小可行性产品（MVP）。利用这一过程，诊断技术、业务两方面的准备度，梳理出部署模式、技术选型、质量和风险管理的框架标准、及配套的组织及资源投入要求。

- **第二阶段，开展全场景盘点+体系规划：**基于局部应用的效果和经验，形成规模化实施的顶层规划，包括：盘点银行所有潜在应用场景，基于商业价值和可行性高低，排布场景的落地先后优先级，形成投入产出量化评估方案；形成技术架构整体升级的细化方案设计；形成质量和风险管理的体系化方案；形成业务和组织能力转型的方案设计；形成能力建设关键举措及路线图。

- **第三阶段，规模化应用落地+体系能力固化：**完成技术和工具基础设施的搭建；依次分批推进应用场景落地；围绕业务、技术端不断积累应用经验；持续在落地中迭代问题，并将相关能力固化至技术架构、业务流程和管理规范中。

关于作者

何大勇是波士顿咨询公司（BCG）董事总经理，全球资深合伙人。

谭彦是波士顿咨询公司（BCG）董事总经理，全球合伙人。

孙蔚是波士顿咨询公司（BCG）合伙人。

窦德景博士是波士顿咨询公司（BCG）合伙人兼副总裁，BCG中国区首席数据科学家。

廖明博士是波士顿咨询公司（BCG）副总裁，数据科学。

冯志宇是波士顿咨询公司（BCG）董事经理。

关于波士顿咨询公司：

波士顿咨询公司（BCG）与商界以及社会领袖携手并肩，帮助他们在应对最严峻挑战的同时，把握千载难逢的绝佳机遇。自1963年成立伊始，BCG便成为商业战略的开拓者和引领者。如今，BCG致力于帮助客户启动和落实整体转型，使所有利益相关方受益——赋能组织增长、打造可持续的竞争优势、发挥积极的社会影响力。

BCG复合多样的国际化团队能够为客户提供深厚的行业知识、职能专长和深刻洞察，激发组织变革。BCG基于最前沿的技术和构思，结合企业数字化创新实践，为客户量身打造符合其商业目标的解决方案。BCG创立的独特合作模式，与客户组织的各个层面紧密协作，帮助客户实现卓越发展，打造更美好的明天。

如需获得有关BCG的详细资料，请发送邮件至：GCMKT@bcg.com。

如欲了解更多BCG的精彩洞察，请关注我们的官方微信账号：BCG波士顿咨询；BCG数智港；“BCG洞察”小程序；BCG微信视频号。

BCG波士顿咨询　BCG数智港　BCG洞察　BCG微信视频号

bcg.com