

生成式AI

产业落地路径研究报告



版权声明

本报告版权属于腾讯云计算（北京）有限责任公司，并受法律保护。转载、摘抄或利用其他方式使用本报告文字或观点的，应注明“来源：腾讯云计算（北京）有限责任公司”。违反上述声明者，公司保留追究其法律责任的权利。

引言

在当下新一轮科技革命和产业变革加速发展的背景下，数据已成为新的生产要素，算力成为新的基础能源，而人工智能则成为新质生产力。2024年的政府工作报告中，明确指出要深化人工智能应用，并首次提出开展“人工智能+”行动。该行动打开了新质生产力的大门，人工智能正在成为产业创新的关键抓手。尤其是以大模型为代表的生成式AI技术，已成为推动新一代产业变革的核心动力。

为抓住此次技术机遇，企业开始积极尝试将生成式AI融入工作流程，以探索各种创新可能性。生成式AI已经成为企业各个层面关注的焦点，CEO寄望于新技术成为公司业绩增长的引擎，而员工则期待新技术成为他们创意性工作的源泉。在这个变革的时代，企业不断努力将技术机遇转化为竞争优势，不断拓展业务领域，迎接未来的挑战。

对于企业而言，探索生成式AI应用落地是一项具有挑战性的任务。本报告提出了生成式AI应用场景矩阵，用以协助企业完善自身可落地应用场景。同时，报告还基于落地实践经验总结出生成式AI应用落地路线，指导企业遴选适合自身情况的落地方式。结合Gartner的产业技术研究，本报告力图为企业带来腾讯云的思考、实践经验与建议，帮助企业捕捉这一历史性新机遇，谱写第二增长曲线。

CONTENTS

目录

01

生成式AI勾勒企业智慧未来

- 一、大模型技术发展迅速, 激发企业生成式AI应用需求 / 02
- 二、生成式AI应用落地处在探索阶段 / 03
- 三、企业落地生成式AI应用面临两大挑战 / 03

02

生成式AI产业落地路径, 指引企业找到适合的落地场景和技术路线

- 一、生成式AI应用场景盘点梳理 / 06
- 二、腾讯云、Gartner联合推出生成式AI应用落地路线图 / 12
- 三、落地路线解读 / 16

03

腾讯云为客户提供生成式AI驱动的新一代产品方案

- 一、基于生成式AI的腾讯云产品架构升级 / 22
- 二、腾讯云完善的产品矩阵, 满足不同路线客户需求 / 22

04

生成式AI产业最佳落地实践

- 一、路线一 - 标准软件 / 46
- 二、路线二 - 标准模型能力增强 / 48
- 三、路线三 - 定制化模型精调训练 / 56

05

生成式AI发展展望



云文档

「AI实验室」知识库

搜索

玩赚 AI 实验室

AI新人入门必学

AI对话能力学习

AI写作能力学习

AI绘画能力学习

AI视频能力学习

AI声音能力学习

AI办公能力学习

AI设计能力学习

AI智能体能力学习

AI数字人应用学习

AI电商技能学习

AI企业相关学习

AI行业报告合集

AI创业投资学习

AI行业大会学习

AI专题学习板块

AI应用实战案例库

AI副业变现案例库

AI内容创作者专栏

AI每日大纪实

运营圆桌会 > 「AI实验室」知识库 > 玩赚 AI 实验室

最新修改: 6月17日 11:50

玩赚 AI 实验室

愿景和目标

愿景：打造一站式 AI 学习基地，关注 AI 技能、AI 商业和变现。
目标：带领 500 万人免费学习 AI 知识，做人人都能学并用得上的 AI 科普开源社区。

玩赚 AI 实验室简介

欢迎来到「玩赚 AI 实验室」知识库，这是最接地气的 AI 知识库，人人都能看得懂用得！玩赚 AI 实验室成立于 2023 年 3 月，负责人是资深运营大猫，帮助大家在这里体系化学习 AI 知识，让大家快速掌握使用 AI 的技能并应用，在这里给你分享各类 AI 商业案例和赚钱案例！2023 年 3 月 6 日，我们写出业内首个 AI 相关 10 万+阅读知识手册，截止 2024 年 2 月，学习手册累计阅读量 100 万+，读者超 50 万+，有 5 万人在社群学习！



新同学入学指南

开源AI学习社区点击直达：

https://yunyinghui.feishu.cn/wiki/Whe1wZnxQi55mpk0zatc3Xran5b?from=from_copylink

第一章

生成式AI勾勒企业 智慧未来



(一)大模型技术发展迅速,激发企业生成式AI应用需求

伴随2022年末ChatGPT的问世,以大模型为代表的生成式AI技术得到各行各业的高度关注和热议。最早由谷歌提出了Transformer架构,随后,谷歌相继发布了基于Transformer架构的Bert、T5等预训练模型,同时OpenAI也推出了GPT预训练模型。模型的参数量迅速提升至千亿甚至万亿级别,成为超大规模参数模型,同时通过对丰富知识数据的学习,大模型技术在泛化能力、多模态能力、开放域交互和模型可解释性等方面均有巨大提升。

2020年,GPT3.0发布,在文本生成方面的能力表现优异。其生成的文本准确、连贯,并且更加贴近人类的表达方式,这使得大模型在内容创作、对话交互等场景具备实用性。2022年,Midjourney发布,大模型的图像生成能力得到广泛认可,应用于广告、游戏等创意设计场景。2024年,Sora发布,可以根据文本、图像生成逼真且具有想象力的视频。教育和娱乐等行业正在积极应用Sora制作视频。

大模型技术发展带来的生成式AI效果提升催生出了新的场景和产业模式,企业探索生成式AI的需求迅速增长。

根据 Gartner 对 822 位企业领导者进行的 "2024 年新一代人工智能规划" 调查显示,绝大多数正在实施或积极计划实施生成式AI的企业高管已经预期或实现了实施收益。平均而言,受访者调查报告:收入增加 15.8%,成本节约 15.2%,员工人数减少 4.6%,生产率提高 22.6%。^[1]

(二)生成式AI应用落地处在探索阶段

生成式AI技术在极短时间内快速渗透到各个工作场景之中，并在改变以往的工作方式。这种新技术的普及将不可避免地诞生新的商业模式和产业生态，并带来全新的机遇。在这个变革的过程中，企业都在积极融入新技术的潮流，并期望在竞争中脱颖而出。

根据Gartner调研数据，44%企业CIO表示已经落地或将在未来12个月落地生成式AI，68%企业CIO表示已经落地或将在未来24个月内落地生成式AI。^[2]

目前，经过初步对生成式AI技术的观望学习，企业认为抓住这一变革性新技术将改变行业竞争格局。大量企业开始转向尝试探索将生成式AI技术应用在自己的业务流程中。

首先，生成式AI有望大幅实现降本增效，从而彻底改变行业成本结构。生成式AI打造的智能助手使得企业为每位员工配备了“专家智囊”，员工个人产能得以倍数级提升，重塑了企业人力资源成本结构。

其次，生成式AI有望实现客户交互方式的变革，使得企业在客户争夺战中抢得先机。一方面，生成式AI的交互体验与真人无差，且个性化更强，年轻客户更青睐新一代的交互方式。另一方面，全新的交互方式之上有可能孕育全新的消费模式。

(三)企业落地生成式AI应用面临两大挑战

生成式AI为企业描绘了美好的未来前景，然而，要将其真正应用于实际业务中，企业仍然面临着场景价值、落地可行性等诸多挑战。

在场景价值方面，企业担心新技术应用价值不及预期，面临如何规划应用落地场景的挑战。企业需要确定生成式AI在特定业务场景下的实际应用和价值，这涉及到对业务需求、用户体验的全面分析和评估。但面对场景价值如何分类与遴选、生成式AI能够解决到何种程度等问题时，企业仍缺少成熟的方法支撑。

在落地可行性方面，企业担心生成式AI技术人才储备不足、数据基础设施不成熟等问题，面临如何选择适合自身情况落地技术路线的挑战。尽管生成式AI技术非常先进，但要在企业内部或产品中实际落地应用，需要充分考虑哪些评估维度，以及可以选择哪些落地路线，企业亟待解决。

针对以上两大挑战，企业需要梳理落地场景，从中优选适合自身的应用落地。并且，企业需要审慎评估技术、预算、风险等维度，寻找合理的落地方式。为此，基于Gartner的研究方法，腾讯云结合自身的大模型应用落地实践经验，为企业提供生成式AI应用场景矩阵和生成式AI应用落地路线图，帮助企业解决场景价值、落地可行性等挑战。



第二章

生成式AI产业落地路径， 指引企业找到适合的落地 场景和技术路线

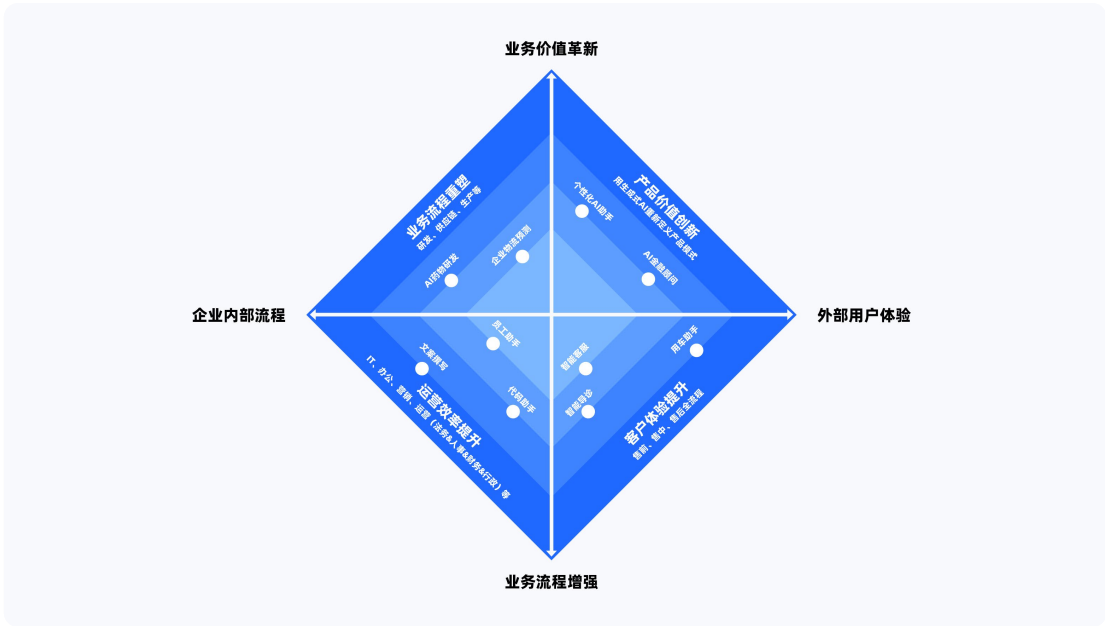


生成式AI技术有助于企业构建行业竞争壁垒，提升市场竞争力，改善客户体验。然而，企业在落地生成式AI应用时，缺少落地场景规划和落地路线选择的方法指导。本章旨在帮助企业全面规划落地场景，遴选最适合自身情况的落地技术路线，实现生成式AI应用的快速落地。同时，本章节还为企业用户落地生成式AI匹配了相应的腾讯云产品方案，使企业能够更加高效地利用生成式AI技术，实现场景价值。

(一) 生成式AI应用场景盘点梳理

1. 生成式AI应用场景矩阵

生成式AI在各行业、各企业职能中的落地应用场景不断增多。然而，在进行生成式AI应用建设之前，企业往往缺乏明确的可落地场景参考，并且对于如何优选试点应用进行落地缺少经验和方法参考。为此，腾讯云推出生成式AI应用场景矩阵，通过对场景价值分类，并对场景技术成熟度进行评估，协助企业系统梳理适合自身情况的落地场景。



（落地场景图由纵横2个坐标轴，4个场景价值象限，以及分布在象限中的场景点构成。）

在生成式AI应用场景矩阵中，横轴是指落地场景面向的企业业务流程分类。横轴正方向，是生成式AI应用到与外部用户体验相关的业务流程，涵盖直接与用户交互或者将生成式AI应用内置于销售给用户的产品服务中等方式。横轴负方向，是生成式AI应用到企业日常运营相关的企业内部流程，生成式AI技术服务于企业员工或企业内部核心业务流程。

纵轴是指落地场景对于业务的改造程度分类。纵轴正方向，是指生成式AI具备颠覆既有业务的创新能力，通过变革产品/业务模式创新，实现业务价值革新。纵轴负方向，是指生成式AI可以改善既有的业务和运营模式，实现业务流程增强。

通过横纵2个坐标轴，生成式AI落地场景被划分至4个场景价值象限中，分别为运营效率提升、客户体验提升、产品价值创新和业务流程重塑。

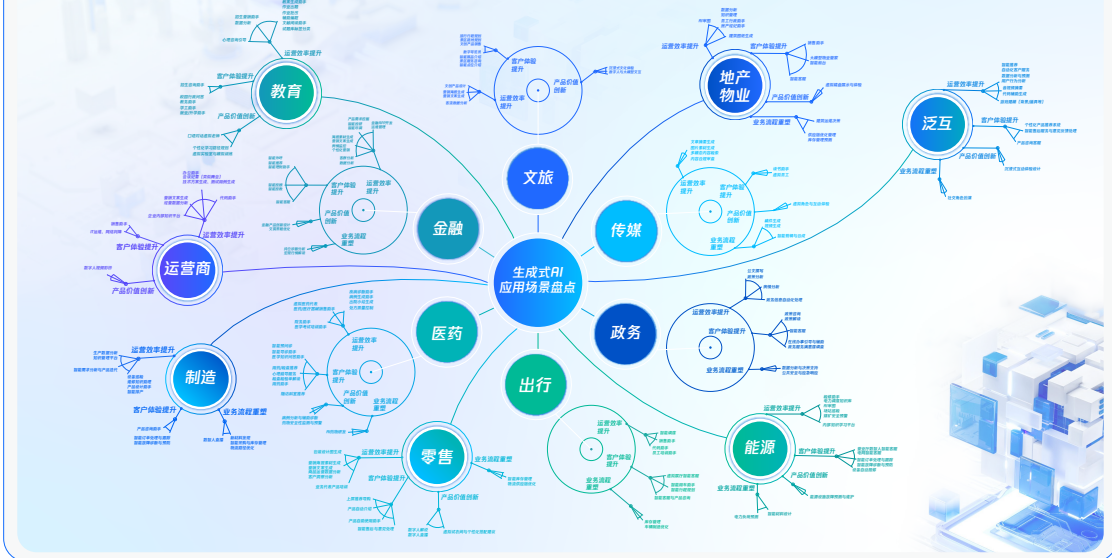
运营效率提升象限，指面向企业内部用户，改善了现有业务流程的场景；通过生成式AI应用帮助企业IT、办公、营销、人财税法职能运营等部门实现降本增效的价值。

客户体验提升象限，指面向企业外部用户，增强业务服务流程的场景。通过生成式AI应用提升了企业售前、售中、售后的全流程客户体验。

产品价值创新象限，指面向企业外部用户，颠覆了传统业务价值的场景，此类场景通过生成式AI重新定义企业服务客户的产品与服务模式。

业务流程重塑象限，指面向企业内部用户，革新业务价值的场景。通过生成式AI应用彻底改造了企业研发、供应链、生产等核心价值链。

生成式AI百大应用场景



详见附录[1]

企业可以参考生成式AI应用场景矩阵，结合内部与外部业务流程，梳理与企业自身情况相关的生成式AI应用场景；我们也梳理了生成式AI在金融、教育、医学、出行等13个行业的百大落地应用场景图，供企业参考，挖掘更多可落地场景。完成场景梳理之后，企业可以从其中遴选合适的场景优先试点落地。

2.四大场景象限

生成式AI将在各个行业、各个业务流程释放价值，这既包括现有业务流程增强所带来的收入增长或成本下降，也包含推动行业进化过程中带来的全新经济价值。以下，我们将对运营效率提升、客户体验提升、业务流程重塑、产品价值创新4大场景价值进行解析。

3.运营效率提升

高效的执行力是企业核心竞争壁垒之一。生成式AI既是通用知识的百科专家，在学习行业和企业专有知识之后，也可以成为特定领域的专家。生成式AI的专家能力，通过智能助手辅助员工或者智能化自动执行业务流程完成落地，从而实现运营效率提升。

首先，企业员工日常有大量文档撰写、图片素材制作、代码注释补齐等执行性工作。通过与模型的问答交互，员工可以获得文档、图像、代码等基础成果，在此之上进行调优可以提高工作效率。

其次，企业业务流程和产品往往会积累大量知识文档，但是这类知识数据往往十分冗杂，企业无法高效使用。基于企业积累的知识文档、知识库等非结构化数据，生成式AI可以实现自然语言交互的知识问答智能助手。员工可以通过问答形式迅速在知识数据中获取自己需要的信息，释放企业数据价值，从而提升运营效率。

最后，企业运营流程涉及多个部门与岗位，流程节点数量众多。任何一个节点执行效率低下，都会导致整个流程延迟。生成式AI的Agent能力可以基于目标自动拆解工作任务、并自动完成，从而使得业务流程无需依托流程节点的人工执行，以实现效率提升。随着生成式AI渗透到企业端到端业务流程之后，甚至将系统性实现运营效率的跃迁。

以下，我们以生成式AI在金融行业的落地场景为例进行说明。

生成式AI应用场景渗透金融业务流程



(生成式AI应用场景渗透金融业务流程)

在IT业务流程中，金融机构信息化程度较高，研发实力雄厚，面向客户的APP应用以IT自研为主。生成式AI可以辅助IT技术人员自动完成代码补全工作，从而提高了金融机构对客APP的研发迭代速度。

在办公业务流程中，生成式AI打造的办公助手可以在会议结束之后自动生成会议纪要。金融行业的专业术语众多，基于行业大模型的生成式AI均可以理解并准确生成会议纪要，将员工从繁杂的流程化工作中释放出来，提升会议待办工作执行的效率。

在营销业务流程中，客户经理面对的金融产品数量众多，每款产品的介绍方案也很复杂，需要消耗大量时间学习理解。生成式AI可以通过知识库问答，迅速解答客户经理对于产品的疑问，并且协助客户经理生成个性化产品介绍方案。

在运营业务流程中，金融机构需要定期向监管机构报送经营数据与财务分析报告。生成式AI可以辅助财务人员快速获取所需的分析数据，并自动化生成分析报告。针对异常指标，生成式AI通过对话式交互，可以快速下钻分析定位原因。

4.客户体验提升

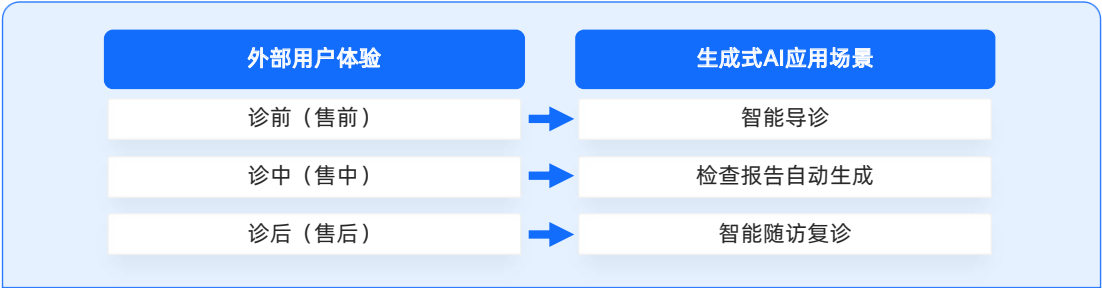
客户体验是企业持续抢占市场份额的重要抓手，生成式AI在客户服务应用场景拥有回复个性化、支持文本图像交互等优势。生成式AI贯穿售前、售中、售后的客户交互服务全流程，全面提升客户体验。

首先，生成式AI大幅改善了客服机器人的回复效果。生成式AI具备通识能力，与客户交互的边界逐步向开放域扩展，而且可以基于不同客户的上下文生成个性化回复，打破了模式化固定回复的限制。而且生成式AI具备文本、图像、视频等多模态能力，可以支持客户上传照片等方式进行交互。

其次，生成式AI打造的机器人不光适用于售后客服场景，在客户旅程中还可以参与至顾问咨询、销售等多个客户交互流程中。生成式AI通过学习产品说明书、金牌销售话术等文档，在产品咨询、产品选型建议等售前和售中场景也在渗透。随着生成式AI在客户交互流程中渗透的增强，企业不仅改善了单一客户体验，也实现了全渠道客户体验一致。

以下，我们以生成式AI在医疗行业的落地场景为例进行说明。

生成式AI应用场景渗透医疗服务流程



（生成式AI应用场景渗透医疗服务流程）

在诊前环节，患者经常会遇到挂错诊室、医生不匹配、候诊时间长等体验不佳的问题。生成式AI支持的机器人，可以通过意图识别、智能追问等多轮交互协助患者获悉对症的就诊科室，并完成预约挂号。而且，生成式AI还可以基于患者疾病类别与缓急、医生擅长疾病与级别，进行精准医生推荐，方便患者匹配合适的医生。

在诊中环节，患者面对检查检验报告往往一头雾水，只能在与医生交流过程中详细垂询。生成式AI可以通过医疗辅助诊断，自动生成辅助诊断报告，在帮助医生提升检查效率的同时，也帮助患者加深病情了解。

在诊后环节，患者缺少与医生交互的通路，用药过程中可能存在用法用量、用药禁忌等疑惑。生成式AI可以弥补诊后的医生空缺，随时与患者保持交流，用口语化方式解答用药疑惑、提醒用药，持续关心患者。

5. 产品价值创新

每一次技术范式的变革，都将涌现出新的产品服务形式，为客户提供全新的产品价值。生成式AI无疑是企业认定的新技术范式。相比传统AI，生成式AI在需要创造力的创新类任务以及需要决策的专家经验类任务中表现大幅提升。因此，通过生成式AI，企业有机会打造新一代应用，创造全新的AI驱动产品与服务。

例如在教育行业，个性化教学、因材施教一直是教育领域的追求目标，但受限于师资资源，个性化教学更多是通过学生分组教学的方式实现，并非针对每位学生因材施教。基于生成式AI的交互和创造能力，个性化AI教学助手可以在教学、练习、考试全流程中与学生完成互动，并不断基于学生知识盲点进行有针对性的教学强化。

6. 业务流程重塑

传统AI技术在研发、生产、供应链等企业核心价值链一直难以渗透发挥价值，主要难点在于传统AI无法理解如此复杂的业务逻辑。生成式AI拥有超强的知识汲取、分析以及生成能力，通过对企业内部数据的海量学习，以及业务流程运行过程中的持续迭代，有机会创造出超越传统专家经验的新模式。基于生成式AI，企业有望革新传统价值链，创造全新的业务流程。

例如在医药研发企业的药物分子筛选场景下，研发人员需要进行大量分子实验，进而筛选出有效的分子。基于生成式AI，研发人员可以构建更为准确的分子筛选标准，从而缩减分子实验规模，缩短药物研发周期。

(二)腾讯云、Gartner联合推出生成式AI应用落地路线图

在完成应用场景规划之后，企业下一步将步入生成式AI应用建设实施阶段。但当前市场上缺乏生成式AI应用建设方法论，企业在落地路线选择、落地路线评估维度、模型服务商分工协作等方面亟需方法指引。

为此，基于Gartner研究报告，腾讯云推出生成式AI应用落地路线图，以期助力企业遴选到适合自身情况的落地路线，规避建设风险、提高落地成功率。



生成式AI应用落地路线分为3条。根据落地复杂程度由简单到复杂依次别为，标准软件路线、标准模型能力增强路线和定制化模型精调训练路线。标准软件路线落地最为简单，企业直接采购开箱即用的软件完成落地。标准模型能力增强路线落地复杂度中等，企业需要参与提示词工程等环节来增强模型能力，优化模型输出结果的。定制化模型精调训练路线落地最复杂，企业需要完成二次模型训练工作。在3条路线中，复杂度越低的路线，企业在落地过程中技术采购（Buy）占比越大；复杂度越高的路线，企业在落地过程中技术自建（Build）比例越高。

为了帮助企业更好的根据自身情况和项目需求选择适合的落地路线，我们从评估维度和分工协作两个方面对每条路线进行解析，并对应提供差异化的腾讯云生成式AI产品方案。

1.评估维度

评估维度，是指企业选择生成式AI应用落地路线时应当考量的6个维度，分别为知识数据、开发投入、数据安全、输出内容控制、项目预算和算力资源。

知识数据指企业评估应用建设需准备的行业或者企业专属数据。企业面临基础模型对于业务场景理解程度不及预期的问题，因此需要评估场景应用对行业/企业知识数据的要求。在场景落地过程中，基础模型在通用程度高的业务场景下有较好结果，但是往往在垂类行业场景下的效果差强人意；对于企业或者行业属性强的场景，基础模型需要补充专属数据以理解业务逻辑。行业/企业专属知识数据可以通过提示词工程、精调训练等不同方式改善模型输出结果。

开发投入，是指企业评估应用建设所需的技术研发人员投入。生成式AI应用建设时，企业面临研发人力资源投入不足以及技术栈短板等问题，因此需要评估场景所需的研发人力资源投入。开发投入分为开发工程师与算法工程师，开发工程师为应用负责，算法工程师为模型负责。

数据安全要求，是指企业评估数据隐私与数据合规潜在风险。生成式AI应用存在信息泄露风险，企业需要将数据进行脱敏处理，降低个人身份的可识别性。而且，对于敏感数据，企业还需制定严格的访问控制策略，授权人员方可访问。模型精调过程中也存在训练数据未合规销毁问题，企业需要建立完善的数据治理规范，确保训练过程中数据合规。

输出内容质量控制，是指企业评估模型输出的准确性、一致性和合规性要求。生成式AI应用落地过程中，企业需要解决模型幻觉问题，以保证模型输出的准确度。同时，大模型面对同一问题，可能会生成不同的答案，企业面临如何保持问题回复一致性的问题。此外，大模型输出的内容要避免涉及敏感信息、知识产权等法律合规问题。

项目预算，是指企业评估场景落地所需资金投入。应用落地时，企业面临投入产出比（ROI）计量的问题，合理的项目预算评估是ROI计算的前提。评估项目预算时，企业需要考虑模型开发、应用开发、应用集成等开发类预算，硬件等基础设施类预算，数据采集与治理等数据类预算，合规类预算以及安全类预算等诸多角度。

算力资源，是指企业评估模型训练与推理使用的GPU资源。GPU资源是企业落地生成式AI应用时的一大掣肘。对于GPU资源配置的要求，模型训练与推理两个环节差异明显。训练所需算力资源的配置远高于推理，企业需要进行合理的资源评估。在训练环节，GPU内存需求、批处理大小设置等会影响资源投入。在推理环节，并发量、模型参数规模、延迟要求等会影响资源评估。

2.分工协作

落地路线第二部分是分工协作，指应用建设涉及的基础模型构建、模型精调训练、提示工程与检索增强生成、应用开发4个层面职责，在企业与模型服务商之间的工作划分与配合。

基础模型构建，是指以无监督机器学习方式，通过海量数据集构建预训练模型的职责。在构建基础模型时，训练数据准备、训练方法、算力储备是3个核心工作。训练数据，需要以海量互联网数据集为基础，包含文本、图像、音频等多模态信息。训练方法，通常采用无监督学习，使得模型自动发现数据中的模式、结构和关联关系。训练方法还会加入基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF），将人类反馈映射为奖励嵌入模型训练，以持续提高模型输出质量。算力储备，模型每训练升级一次需要投入大量高性能GPU卡时。

模型精调训练，是指通过有监督机器学习，将高质量的知识数据训练集注入基础模型的职责。精调可以提升模型在特定任务上的性能，使模型更适应特定领域、任务或应用场景，输出更准确的结果。精调过程中，训练数据的准备与增强、模型参数微调、持续迭代是核心工作职责。精调所需的训练数据，必须经过人工标注，而且可以通过数据增强生成更多训练样本。参数微调涉及模型架构的调整，需要对基础模型进行层级解剖，以确定哪些层需要精调，哪些层保持固定。精调往往是一个长期持续性工作，在不断调整精调策略的过程中逐步改善模型输出质量。

提示工程与检索增强生成，是指通过提示词工程（Prompt Engineering）设计和引导模型输入，在不改变基础模型参数的同时，解决模型幻觉、模型知识更新及时性等一系列问题。提示词需要根据特定任务进行设计和迭代。提示词设计可以由多种不同关键词、语法组合形成策略，以适配不同使用场景。检索增强生成（Retrieval Augmented Generation, RAG）是提示词工程的一种重要形式。通过将检索的知识库信息引入到模型输入，模型输出的知识实时性、结果质量均有明显改善。

应用开发，是指面向终端用户（End User）的应用建设职责，以解决企业应用的定制化需求，保证应用可以适配企业业务流程。应用建设需要企业的业务需求方深度参与，以更好的确定应用待解决的问题与目标。

3.产品方案

落地路线第三部分是产品方案，指腾讯云为不同落地路线所设计的产品组合方案。产品方案由AI Apps、AI Engineering、AI Infra3层及模型、安全构成。

AI Apps 暨人工智能应用程序，是指主要使用生成式人工智能模型来执行任务的软件应用程序。这包括聚焦行业和功能的应用程序。

AI Engineering 暨人工智能工程，是指使企业能够操作生成式人工智能模型的工具，包括以模型和数据为中心的工具。

AI Infra 暨人工智能基础设施，是指可用于构建生成式人工智能应用程序的基础设施组件，如计算、网络和存储组件。

模型，既包括生成式人工智能基础模型（如大语言模型），也包括针对特定行业或用例的特定领域的模型。

安全，是指企业在建设生成式AI过程中面临的数据隐私保护相关的AI本体安全（AI safety）、开发/运营/部署流程中安全防护相关的AI应用安全（security for AI）和AI内容审核相关的AI衍生安全。

企业可以基于自身场景需求，参考评估维度、分工协作对自身需求进行评估分析，明确符合自身特点的落地路线，最终选择腾讯云提供的差异化产品方案来落地生成式AI应用。

(三)落地路线解读

1.标准软件落地路线

标准软件落地路线，是指企业直接采购基于生成式AI模型能力增强的成熟应用软件。该落地路线下，生成式AI内嵌于应用之中，企业可以开箱即用。比如，在将生成式AI内置于HR SaaS软件，实现以智能问答的交互形式服务员工。

在评估标准软件落地路线对应的场景时，场景需求呈现出应用成熟、通用程度高、安全要求高等特点。首先，此落地路线应用成熟度高，采购的是标准软件。因此，开发投入、项目预算、算力资源等评估维度对于企业要求低。其次，标准软件适合的场景可以跨职能部门、跨行业，通用程度高，因此对于组织/垂直知识数据、输出内容质量控制等评估维度要求也属于低。最后，该路线下的应用多搭载于公有云之上，因此企业会非常在意数据与应用本身的安全防护，场景评估时数据安全要求高。

选择标准软件落地路线时，企业与模型服务商在分工协作中，以模型服务商为主导。模型服务商以提供开箱即用的应用为导向，全权负责从底层基础模型到上层应用建设的全部职责。该落地路线下，腾讯云提供一系列成熟的AI Apps产品方案。一类AI Apps是生成式AI增强的成熟软件，另一类AI Apps是在成熟软件中增加智能助手入口，用户可以随时唤醒助手辅助工作。

标准软件落地路线优势在于，既有系统改造成本低、投入预算低且可以持续升级迭代。由于直接使用成熟应用，企业只需要将应用与相关业务系统接口进行打通即可，现有工作流程、业务系统均无需调整，因此系统改造成本低。同时，公有云上的应用无需部署成本，软件采购或基于用量或基于账号量，预算投入低。另一个采购公有云应用的优势在于，应用可以即时更新，企业可以随时享受到最新的技术红利。

标准软件落地路线的劣势在于，模型能力不全面和无法定制化扩展功能。第一，该路线使用的模型由公开数据训练，没有针对垂直行业/企业专属业务需求定制化调优，在面对部分垂类业务问题时可能性能与预期有一定差距。第二，标准软件不支持定制化开发，面对企业内部复杂的业务流程时，无法与业务深度融合。这导致企业难以将标准软件与内部流程无缝整合，限制了在应对特定业务需求时的灵活性。

针对标准软件落地路线，建议企业在代码生成、会议纪要等通用程度高的场景采用该路线。这类场景在各行业的需求通用度较高，一般不具有行业差异性，使用标准软件即可满足用户需求。标准软件通常具有较好的稳定性和可靠性，无需进行定制化开发，支持开箱即用快速落地，可以帮助企业高性价比快速落地生成式AI应用。

2. 标准模型能力增强落地路线

标准模型能力增强路线，是指通过提示词设计、检索增强生成等提示词工程，增强基础模型输出的准确性、知识实时性。检索增强生成可以通过加载组织/垂直知识数据，改善模型输入的提示词准确性。值得一提的是，该路线实施时，会将生成式AI模型能力封装为API服务接口，企业将API嵌入自研应用软件增强智能化水平，或是基于API创建定制化的全新智能应用。该落地路线对应的场景进行评估时，场景需求具有私域知识引入、研发资源与预算投入适中、算力资源投入低、数据安全要求高等特点。首先，基础模型无法充分理解企业内部业务场景、且无法回答实时更新的业务知识，需要引入行业或者企业私域知识以弥补基础模型能力欠缺。因此，输出内容质量控制和组织/垂直知识数据两个评估维度属于中等要求。其次，该落地路线下企业既需要建设API服务接口，也需要定制化开发应用软件，因此开发投入和项目预算两个评估维度较标准软件落地路线有所提高，属于中等要求。再次，该落地路线中的检索增强与提示工程不涉及模型参数的调整，无需模型训练算力投入，只需企业为模型推理提供算力支持，因此算力资源要求依然低。最后，由于引入企业私域数据，在提示词使用过程中有隐私数据泄露风险，数据安全评估维度保持在高要求水准。

选择标准模型能力增强落地路线，企业和模型服务商分工协作各有侧重点。模型服务商参与所有4个环节工作，由于模型参数无需调整，模型服务商的工作职责重点在于基础模型和模型精调。企业深度参与至数据检索与提示工程、应用建设两个环节，以解决API服务接口和应用软件的定制化需求。

根据企业在数据检索与提示工程环节的参与程度，标准模型能力增强落地路线可以继续细分为三条子路线。

第一条子路线，企业可以使用腾讯云封装好的标准模型API，或者使用腾讯云的aPaaS平台，加入企业专属数据，通过RAG（检索增强生产）等提示工程技术形成经过能力增强的专属API。此路线中模型能力增强的工作由腾讯云主导完成，企业直接使用API即可。

第二条子路线，企业使用腾讯云机器学习平台和向量数据库产品，自行构建API服务接口的检索增强。企业使用腾讯云提供向量数据库构建企业专属向量知识库，实现对数据检索速度和效率的优化，并使用腾讯云机器学习平台完成模型推理服务部署。

第三条子路线，企业独立建设机器学习平台和向量数据库的框架，并完成API服务接口的检索增强。相比第二条子路线，企业需要自行使用函数、容器等服务自行完成模型推理环境的搭建，此子路线更适合IT开发能力强、有自主构建开发中台需求的企业。

企业可以基于开发投入这一评估维度，进一步选择适合自身情况的子路线。企业算法工程师团队尚在建立早期时，检索增强等工程能力可能存在技术短板，建议选择第一条子路线。企业算法工程师团队完善，但云原生技术团队存在技术短板时，建议选择第二条子路线。企业自研实力强，算法和云原生技术资源储备完善时，建议选择第三条子路线。

在标准模型能力增强落地路线下，腾讯云提供AI Apps和AI Engineering组合形成的产品方案。AI Apps主要以模型服务的方式提供。AI Engineering分为云原生平台（如容器）、数据平台（如向量数据库）、AI开发平台（如机器学习平台）和应用开发平台（如DevOps）。

标准模型能力增强落地路线优势在于，低成本改善模型输出效果，以及生成式AI应用定制化能力增强。在改善模型输出层面，企业无需支付高昂的算力资源和预算改造基础模型，通过较低成本的提示工程和检索增强，就可以实现知识输出实时性和模型准确率提高，一定程度缓解了模型幻觉带来的困扰。在定制化能力增强方面，企业可以将生成式AI封装的API服务接口嵌入自研应用，使得符合企业定制化需求的自研应用得以增强；企业也可以基于API服务接口创建全新的自定义应用，同样使得定制化需求得以满足。

标准模型能力增强落地路线劣势在于场景适应性受限和响应延迟。首先，该路线并没有改进模型本身的能力结构，只是通过提示词工程补充了知识、调优了输出。因此，场景发生变化时，需要重新启动提示词工程，使用场景有局限性。其次，检索增强是模型输出前额外增加的执行步骤，可能会导致模型输出的响应时间增加，进而影响终端用户的使用体验。

针对标准模型能力增强落地路线，建议企业在知识助手类场景采纳。生成式AI为各个岗位带来了人效提升的机会，知识助手通过问答形式为每位员工提供赋能。而且，检索增强加载的

知识库可以实现每位员工的个性化，员工的个人经验可以持续沉淀在知识助手上，进而为公司整体经验提升奠定基础。

3.定制化模型精调训练落地路线

定制化模型精调训练落地路线，是指基于基础模型，企业通过私域数据集进行模型训练，以实现企业专属大模型的建设。企业专属大模型改变了基础模型参数，并且在特定任务上表现出来的性能和在企业知识的深度理解两方面得以增强。

该落地路线对应的场景进行评估时，需求呈现出需求高度定制化、资源和预算投入高等特点。首先，企业精调模型是因为基础模型在特定任务能力上存在短板，需要在模型底层进行改变以大幅改善输出结果。对应的企业需要有针对性的进行专属训练数据集准备和遴选，输出结果的控制也完全以企业自身需求为准绳。因此，输出内容质量控制、组织/垂直知识数据、数据安全要求等评估维度要求很高。其次，企业精调模型需要整建制的算法工程师团队、完全定制化的模型训练和应用开发，因此开发投入和项目预算也很高。同时，模型精调需要投入GPU资源完成训练任务，消耗的算力资源同样很高。

该落地路线在分工协作层面，以企业为主导角色，模型服务商扮演辅助角色。除了基础模型由服务商提供，模型精调、数据检索与提示工程、应用建设等环节企业均处于核心主导地位。在此过程中，企业生成式AI技术自研能力得以大幅提升。模型服务商聚焦于基础模型的迭代升级，重点为企业进行模型精调训练提供算力和工具服务。

相应的，该落地路线下，腾讯云提供AI Infra与AI Engineering产品组合方案。AI Infra涵盖计算资源、软件加速、硬件创新等产品集合。基础模型精调耗费大量算力资源，因此内置硬件和软件加速产品的异构计算平台，可以协助企业提高GPU资源使用率、加快模型精调速度。AI Engineering包括AI训练工具平台和数据检索与提示工程开发工具链。基础模型精调需要持续迭代，因此操作简单的训练工具平台至关重要。

定制化模型精调训练落地路线优势在于，模型输出准确度有保障，以及提高了企业在行业中的竞争壁垒。由于模型精调完全针对企业个性化需求，输入的训练集准确度高，使得模型输出的准确性也极高。而且，企业在该落地路线中处于主导地位，对于专属大模型的技术控制力强，为产品创新和业务重塑带来无限可能。企业专属大模型可能成为超越对手的有力竞争武器。

定制化模型精调训练落地路线劣势在于，成本高企和灾难性遗忘带来的模型能力退化。首先，该路线算力资源投入高、开发人员投入高、项目预算高。高企的成本，会拉低项目投入产出比。其次，精调后的模型可能在原有任务性能上出现大幅衰退，丧失了基础模型的泛化能力。

在面对检索增强的模型输出依然无法达到企业准确性和实时性要求，或者企业需要将生成式AI应用于业务逻辑极其复杂的场景、而基础模型完全无法理解等情景时，建议企业采纳定制化模型精调训练落地路线，定制企业专属大模型。



第三章

腾讯云为客户提供生成式AI 驱动的新一代产品方案



(一) 基于生成式AI的腾讯云产品架构升级

全栈提效 让企业用好AI



(二) 腾讯云完善的产品矩阵, 满足不同路线客户需求

1. 路线一 - 标准软件

(1) 腾讯乐享AI助手

落地背景及挑战

在企业知识管理、培训学习、办公协同场景中，存在着大量的内容生产、内容流转、内容消费诉求，传统的知识管理或培训平台存在着内容生产门槛高、内容生产效率低、内容获取链路长、内容利用率低等问题，导致管理成本高但收效甚微，让企业知识从“低质量”走向“不可用”。

企业应用面临的挑战包括：1. 实现问答能力与企业内部知识的强关联；2. 生成不同场景、格式的内容需要不同模型、方向的训练方案。

产品方案

腾讯乐享AI助手 让知识生产和学习更高效



方案价值

①智能问答：

搭建大模型与企业内部知识的桥梁，无需训练模型即可快速搭建企业AI问答专家，可理解企业知识后以自然语言直接回复成员问题，并列出生关文档链接作参考。

对比传统搜索的优势：保留知识搜索的权限特性，改变传统冗长的检索->浏览->筛选->理解的信息获取链路，提高内容曝光，让每次知识获取时间减少5-10分钟。

对比传统问答的优势：无需管理问答对，直接基于文档、视频问答；可通过反馈-提醒修订-后台跟进分析的工作流，促进知识迭代；乐享千人千面的权限管理，AI也能对拥有不同权限的提问人给出不同的回答。

②智能写作：

提供适配企业业务内容生产的写作大模型，让模型一方面能够掌握足够多的各领域知识，提供创作灵感；另一方面能够根据明确指令，对文档内容进行总结、缩写、扩写或纠偏，如撰写大纲、方案续写、会议纪要等。降低业务内容创作的门槛，提升办公协同内容输出的效率。

③智能生成：

让模型掌握跨职责的生产能力，如培训中的生成考题场景，通过训练专门的出题大模型，让模型既理解业务知识，又按规则出题，成为一个懂业务的培训出题专家，解决传统的出题模式中各方资源协同和生产困难的挑战。

腾讯乐享AI助手，让知识生产、学习、管理全生命周期提效。

(2)腾讯云AI代码助手

落地背景及挑战

随着人工智能的爆发，软件工程也转向智能化，而AI代码助手作为公认的高频高价值场景先行一步。

企业落地的挑战包括。1、企业用户关心代码安全。尤其在金融等高合规要求行业，随着大模型落地的不断推进，企业越来越关注如何平衡大模型落地收益和潜在合规风险。2、业务场景要求多模能力，企业内部不同业务场景对多种个性化模型提出需求，同时需保证应用侧无缝对接。3、企业决策需要效果度量。企业用户更关心投产价值如何衡量，期待通过数据监控手段，保障模型训练效果，推动持续改善提升。4、企业效能提出全场景覆盖。企业希望在研发过程中尽可能多的引入较为确定可行的AI加持场景，覆盖多个智能研发场景，从而提升工程质量，达到效率提升最大化。

产品方案



方案价值

腾讯云AI代码助手是依托于腾讯模型与训练底座，精准面向企业客户需求打造，支持多模态、安全合规、流畅高效、成本可控的专业AI编程提效工具。

①保障代码安全（Security）

依托腾讯云天御安全审核模型三重防护能力，在语料防护和Reasoning、微调等阶段进行主动预防：确保输入内容正常合规；确保回答问题在训练内容范围之内，减少回答幻觉；确保不合规问题能够被拒绝回答。部署后，和企业安全团队积极配合，只选取企业内部认证合规的私有语料进行进一步训练，打磨出符合企业要求的代码模型。

②支持多模能力（MaaS）

基于具备多模管理能力、支持企业私有化部署的专业训练平台--腾讯云TI-ONE平台进行训练和推理优化，在代码对话、单元测试场景基于行业模型进行指令对齐和强化学习、推理优化，同时对真实场景下反馈的badcase进行调优。

③提供数据看板（Analysis）

企业可通过数据看板进行衡量，通过效能看板密切关注代码生成率、采纳率、QPS、生成耗时等关键指标，对产品的优化起到监督反馈效果。

④覆盖丰富场景（Full）

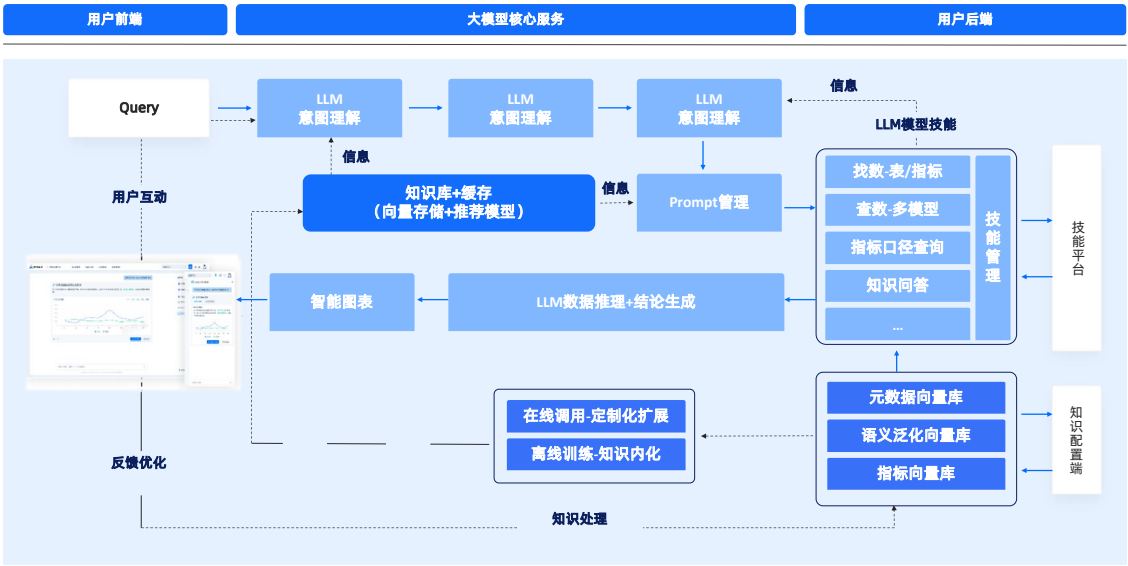
覆盖代码补全，技术对话，单元测试和代码诊断等关键编码阶段应用场景。

(3)腾讯企点营销云AI助手

落地背景及挑战

企业数据分析主要痛点：懂业务的人不会用分析工具、懂分析的人不一定了解业务、海量数据人工挖掘效率低、数据洞察依赖主观经验不够严谨。

产品方案



方案价值

基于腾讯云行业大模型在分析领域的应用服务。用户可以通过自然语言对话，实现找数查数、数据结论解读、智能归因及异常诊断，旨在企业提供专业、高效、易用、可信的智能分析服务。

①专业

结合行业知识在营销分析领域微调后的大模型，具备专业分析师思维，覆盖90%数据分析场景，及业界领先的问题识别率，客户个性化知识一学就会，更懂业务。

②高效

智能归因自动识别指标异动，海量数据也能快速定位影响因子，异常诊断可及时下发一线业务执行策略，定时看板解读任务和一键生产报告功能，轻松搞定周报日报。

③易用

对话式交互，只要懂业务，就可以实现数据查询及分析，根据业务数据推荐分析问题和分析指标，小白用户轻松上手，分析结果自然语言总结，不懂图表也能获得关键业务信息。

④可信

针对复杂问题，模型推理步骤可视化，分析结果对应配置参数/SQL可查看可验证，分析结果生成配置页面可查看可调整。

2.路线二-标准模型能力增强

(1)腾讯云大模型知识引擎

落地背景及挑战

大模型技术不断深化，加速生成式AI应用落地，推动新一轮的产业变革，IDC报告显示，知识管理、会话类应用目前最受企业者关注及青睐。然而，大模型到知识应用落地门还存在一定门槛，在典型知识服务应用包括知识问答、业务办理、知识总结等均有着不同程度的痛点。

①企业文档知识多样化，图文并茂，排版复杂用户问答效果及体验不可控。

②流程知识复杂度高，维护成本高自动化接待难度大。

③不同行业不同场景对于知识总结、信息提取存在差异化诉求。

面向这一背景，打造基于大语言模型的企业级知识应用构建平台，提供开箱可用的应

用模版、可被集成的原子能力API，降低大模型接入到企业知识管理及会话应用场景的门槛，推动大语言模型在企服场景的应用落地。

产品方案



方案价值

- ①擅长处理企业级多模态复杂知识，支持最全常用格式与上百种文档场景，在表格问答、复杂公式、图文关联输出与数学计算等复杂场景，可满足企业知识库管理维护的多样化、专业性处理要求。
- ②依托OCR大模型解析引擎、多模态大模型、业内首个语义切分大模型与基于混元大模型技术的行业大模型，全链路解决复杂文档的解析、切分、检索、推理、生成等业界难题，端到端问答效果业内领先。
- ③多模态交互模式丰富，应用端支持多种输入、输出格式，提升交互体验。

易用

- ①使用门槛低，提供开箱可用的应用模板和可被集成的原子能力API两种便捷使用方式。
- ②工具链完善，支持自定义模型选择、角色设定、提示词自动优化、知识库管理及维护。
- ③一站式流程，支持对话测试-修正-发布-反馈增强的一站式处理。

(2)向量数据库

落地背景及挑战

利用LLM和向量数据库，企业可以构建高效、智能的知识库系统，实现快速检索、语义理解和个性化推荐，提升企业知识管理和应用效率。

①挑战：

文档预处理、向量化门槛高，业务接入成本高，接入效果差。

产品方案



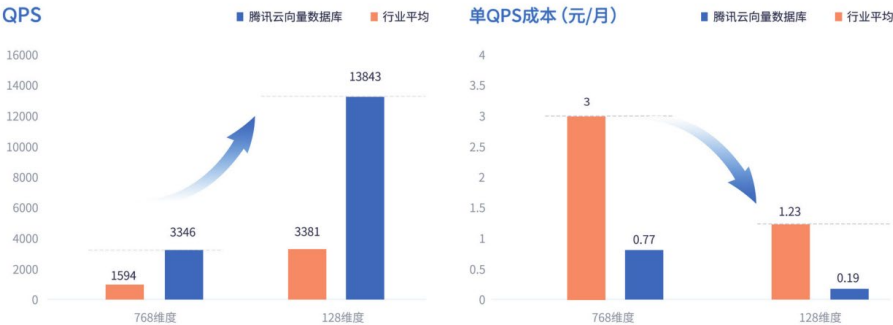
性能强劲，行业领先

五年打磨，积累深厚

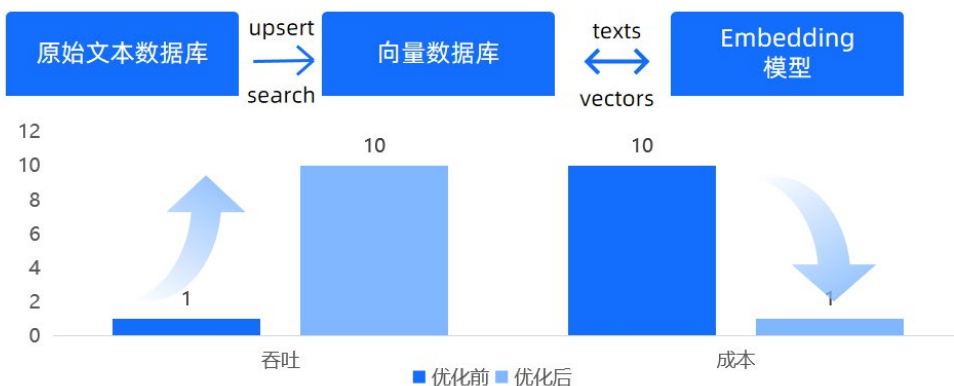
源自腾讯集团内部向量检索引擎OLAMA，已服务腾讯集团内部40+线上业务，日均处理千亿次向量检索请求。

性能领先，性价比高

性能领先行业平均水平1.5倍以上，单QPS成本降低75%，极具性价比。



集成Embedding功能，高效易用

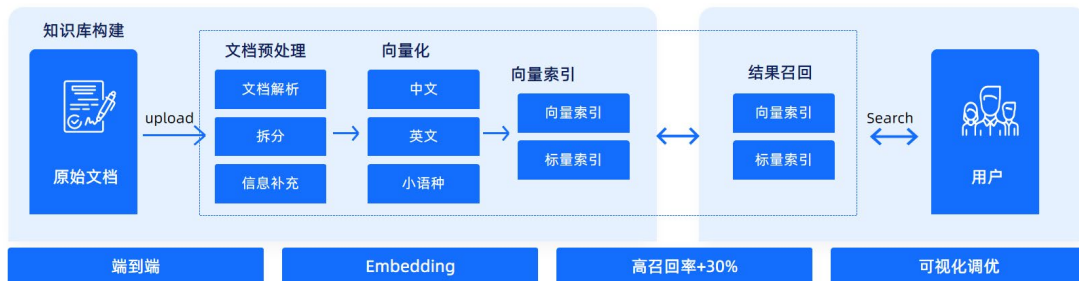


- 数据写入/检索自动向量化，对齐传统数据库的使用体验，无需关注向量生成过程
- 优化GPU处理速率，性能提升10倍，超高配速快速处理数据
- 海量GPU资源池提供算力服务，按需使用，成本更优

方案价值

端到端AI套件，AGI时代的知识库解决方案

腾讯云向量数据库实现面向RAG领域中内容检索的端到端解决方案，在向量数据库中集成文档预处理、自动向量化、可视化调优等功能，简化RAG中的数据处理以及检索流程，实现业内最高召回率，并大幅降低接入门槛，帮助业务快速搭建定制化的知识库服务。



检索召回效果提升30%

在文档预处理、索引构建、结果排序等环节均进行了专项优化，并构建独特的精排方案，更适用于各类文档检索场景，召回率对比传统方案显著提升。

接入效率提升10+倍

相较于自行完成数据处理、向量化、创建索引等操作，AI套件提供了简单、高效的解决方案，将知识库接入的时间从1个月提升到1小时，接入效率提升10倍以上。

(3)腾讯云Cloud Studio

落地背景及挑战

- ①传统的应用开发常常面临：1、资源的限制 2、环境配置的复杂且不兼容问题

产品方案



方案价值

Cloud Studio，提供即开即用的在线开发环境，满足AI场景下大模型运算、图形计算等场景对GPU算力工作空间的需求，实现AI原生应用开发中提效降本与高效协作，对于教育行业创新应用与教学成果快速迭代、上线更具有长效价值。

3.路线三-定制化模型精调训练

(1)腾讯云TI平台

落地背景及挑战

大模型驱动“智慧涌现”，AI大模型的发展与运用，正在逐步贯穿到企业的设计、研发、生产、营销、服务全流程，助力企业创新的同时，帮助员工提效，加速云智一体进程，但是如何将大模型快速应用到自己的业务场景中仍然面临大量挑战。

第一，缺乏高质量行业知识数据。通用大模型在逻辑推理、常识问答等场景下有较好的表现，但是在垂直行业业务场景下效果往往差强人意，需要有大量高质量的行业知识数据，通过大模型训练优化，提升大模型在具体垂直行业业务场景下的能力。

第二，大规模并行训练稳定性要求高。由于大模型的训练参数量、数据量大，在进行大模型训练时，需要大量的GPU训练资源且训练周期长，对训练平台的断点续训、故障自动隔离、自动容错，异常感知等底座能力考验极大。

第三，计算资源少。大模型的训练和推理，对计算资源和存储资源有很高的需求，大规模分布式训练需要高效的资源调度管理系统，减少节点资源碎片，提高调度成功率，并需要灵活的任务优先级配置策略，进一步提升资源利用率。

产品方案



方案价值

①为客户提供丰富高质量的基底大模型。

腾讯云TI平台内置腾讯自研的超千亿参数规模混元大模型，其使用超2万亿token预料优化预训练算法及策略，并改进了注意力机器、开发思维链新算法，在轮对话、知识增强、逻辑推理、内容创作等方面有优势能力。同时，基于大量垂直行业的业务数据，我们精调训练出金融、医药、教育、汽车等多个行业的行业大模型，降低了不同行业数据的彼此干扰，提升模型在行业垂类任务的性能。

②完善的工具平台支持“简单、稳定、高效”的大规模训练。

简单，暨腾讯云TI平台为用户提供简单的大规模训练流程，内置精调物料，一键启动精调

任务，最快仅需5步即可完成大模型精调落地。稳定，通过机器故障自动迁移、异常POD驱逐重新调度、自动断点续训，保障大模型训练任务的稳定性。高效，通过拓扑感知调度、gang调度策略，任务优先级配置策略等技术提高资源利用率，并通过自动注入通信配置、网络拓扑等相关参数提升开发效率和问题排查效率。

③提供算力、存储、计算多重优化加速能力。

通过自研的Angel加速框架，通过CPU&GPU异步调用提高吞吐，通过缓存定长和池化，提升参数支持规模，优化显存；自研大模型计算和并行优化方案，结合编译和手工优化，解决低配显存显卡无法训练超大模型问题；为用户提供三重加速优化。

(2)腾讯云异构计算平台

落地背景及挑战

背景：AI已经进入了一个高速发展期，模型参数达到万亿级，在训推场景，对算力的规模、性能、容错、性价比提出前所未有的挑战。

①大模型训练场景痛点：

大模型AI场景下，对于训练效率要求非常极致，单纯堆叠计算节点存在“木桶效应”，存储、网络一旦出现瓶颈就会导致运算速度严重下降。

②推理场景痛点：

AI推理是前向传播的用户请求计算过程；需要寻求低时延、大吞吐、低成本、低功耗之间的平衡。

③部署、易用的痛点：

AIGC浪潮下，AI应用部署和使用需求激增。因GPU等异构芯片的特殊性，自身部署使用难度大，涉及到选型、驱动安装、软件栈适配、版本兼容、应用部署等多重问题，中小企业及开发者望尘莫及。而且涉及到安全隐私问题，客户要求私有化或本地化部署，来运载自己的大模型及相关业务。

产品方案

产品全景能力



方案价值

①极致性能：

搭载业界前沿芯片与腾讯自研 XPU 芯片；自研智能网卡，虚拟化零损耗，性能媲美物理机；支持业内唯一100GVPC 和独家3.2T RDMA 网络；定制硬件，针对场景特殊优化，性能领先业界，万亿参数大模型训练时间缩短 80%。

②极致易用：

一键安装 GPU 驱动；基础环境自动化部署；提供批量任务管理、资源管理工具。支持多种 AI环境快速部署，如ChatGLM-6B、StableDiffusion等。

③极致兼容：

实例丰富度全球领先；统一的软件接口 TACO，屏蔽底层硬件差异，保障平台稳定；云原生网络架构，实现训练推理混部。

④极致降本：

支持 vGPU 按需使用；qGPU 容器级资源切分，颗粒度达 5%，提供 GPU 算粒级弹性服务。

⑤规模领先：

腾讯云的算力调度总规模超过1.5亿核，并提供16 EFLOPS（每秒1600亿亿次浮点运算）的智算算力。未来，新一代集群不仅能服务于大模型训练推理，已实现全国90%+的大模型客户覆盖度，还将在自动驾驶、科学计算、自然语言处理等场景中充分应用。

4.模型底座

(1)腾讯混元大模型

腾讯混元是腾讯基于全链路自主可控技术打造的生成式大模型，通过持续迭代和实践，积累了从底层算力到机器学习平台再到上层应用的完整自主技术。目前，腾讯混元大模型参数量已达到万亿，tokens数量超过7万亿。

算法层面，腾讯混元大模型率先采用混合专家模型 (MoE) 结构，并在这个过程积累大量自研技术，其中包括创新的专家路由Routing算法、独创的MoE Scaling Law机制以及合成数据技术，实现模型总体性能相比上一代Dense模型提升50%。

训练和推理框架上，腾讯自研Angel机器学习平台面向大模型训练，在预训练、模型精调、强化学习多个阶段进行优化，相比业界开源框架，可以用更少的资源训练更大的模型，训练速度是主流框架的2.6倍；面向大模型推理，Angel机器学习平台实现成本相比业界主流框架下降70%，支持国产主流硬件的适配。

算力层面，腾讯拥有自研星脉高性能计算网络，使用领先的3层网络架构，可支持单集群十万卡规模，具备业界最高的3.2T通信带宽，可提升40%的GPU利用率，节省30%—60%的模型训练成本，为AI大模型带来10倍通信性能提升。此外，星脉网络中的交换机、光模块、网卡均是腾讯全链路自研。

基于扎实的基础能力积累，腾讯混元大模型积极推进相关应用落地，让大模型创造更多价值。目前腾讯内部超过600个业务及场景已接入测试，腾讯旗下协作SaaS（软件即服务）产品全面接入腾讯混元大模型，包含企业微信、腾讯会议、腾讯文档、腾讯乐享、腾讯电子签、腾讯问卷、腾讯云AI代码助手等产品。腾讯广告基于混元大模型推出AI广告创意平台妙思，有效提高广告主生产及投放效率。

无论是企业客户还是个人开发者，都可以通过腾讯云上API直接调用腾讯混元，实现更便捷的智能化升级。腾讯还联合生态伙伴，将大模型技术与20多个行业结合，提供超50个行业大模型解决方案。

(2)腾讯云行业大模型

为更好解决特定行业问题，基于混元技术能力，结合垂直行业业务场景，加入大量垂直行业数据，精调训练出垂直行业大模型。相比通用大模型，行业大模型能够理解垂直行业的专业领域知识，提供更加精准的回复；并且在数据安全、生成内容控制方面会有进一步的能力提升。

目前腾讯云已经推出金融行业大模型、医学行业大模型、汽车行业大模型、教育行业大模型、文旅行业大模型等多个行业大模型，提供超50个行业大模型解决方案，助力客户构建专属大模型及智能应用，帮助客户提质增效。

以腾讯云金融行业大模型为例，其是在通用大模型基础上，加入大量金融行业数据进行训练，并通过行业专家强化反馈，提升了金融行业大模型在金融领域的任务效果。可以作为金融机构员工的智能助手，在知识问答、舆情摘要、营销文案生成、研报摘要、产品推荐、投资教育等场景帮助金融机构提高工作效率和用户体验。

腾讯云也将持续升级产品、技术能力，加速行业大模型在千行百业应用落地，助力更多实体实现高质量的数智化转型升级。



5.全栈安全防护

(1)腾讯云金融风控大模型

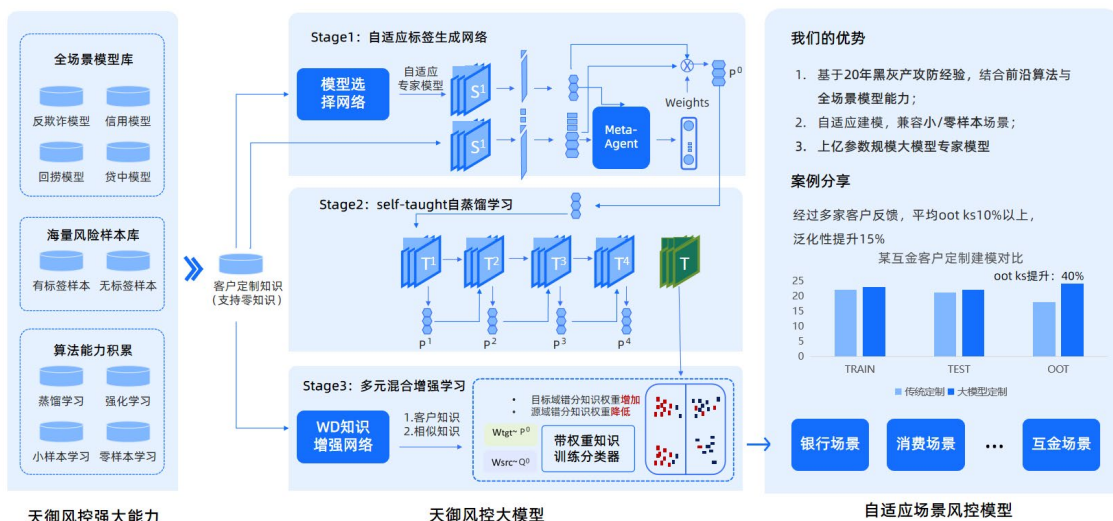
落地背景及挑战

随着零售信贷业务的发展，金融机构已经进入精细化运营和差异化竞争的阶段，各机构的产品设计和目标客群都有比较大的差异，反欺诈技术也从过去的通用反欺诈模型发展到定制化的阶段。

然而，金融反欺诈受表现期长，样本收集成本高，产品渠道迭代快等因素影响，普遍面临样本不足问题，导致金融定制化模型极易发生过拟合，导致上线后模型不稳定，或者效果衰减快，给风险控制带来很大挑战。

产品方案

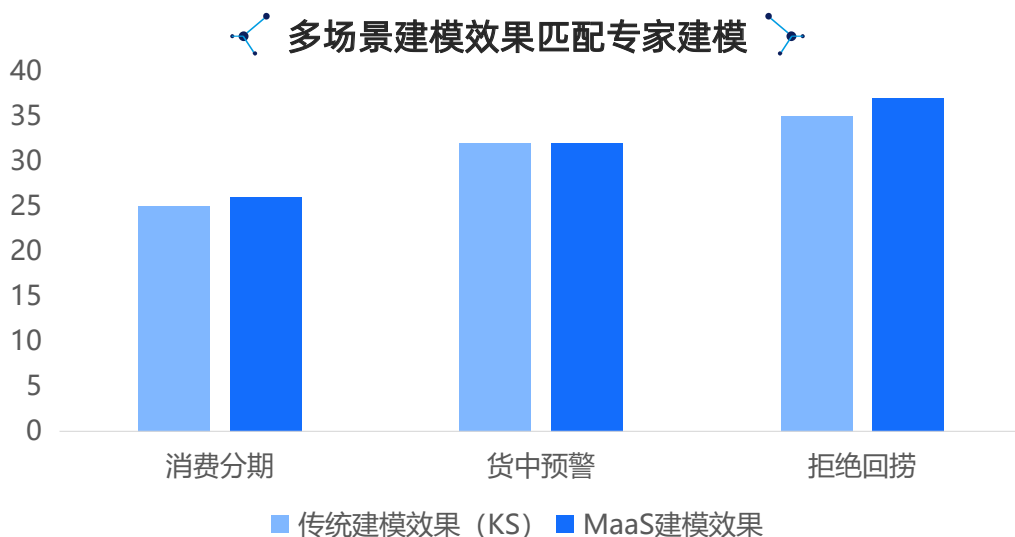
天御金融反欺诈大模型技术方案



方案价值

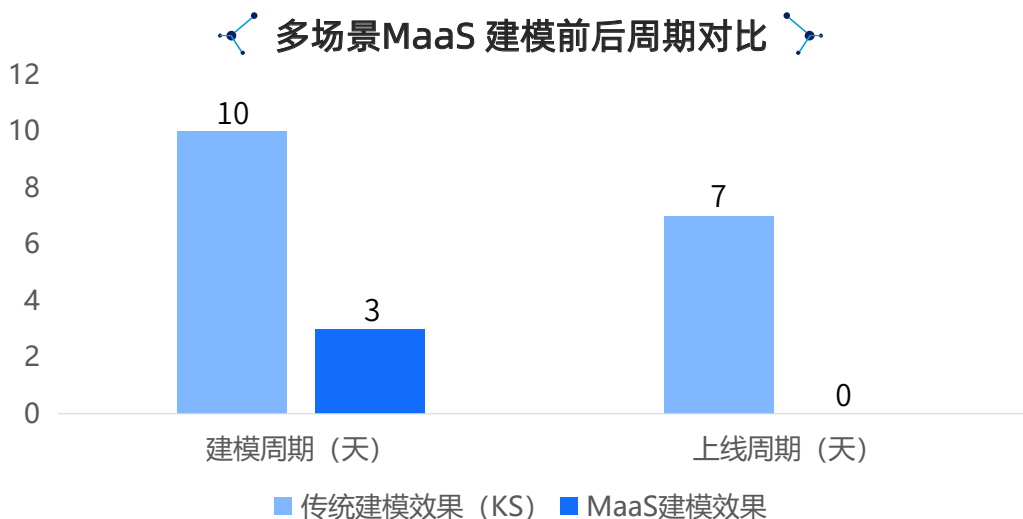
①建模无忧：腾讯专家级建模

金融风控大模型融合了腾讯安全天御大量风控建模专家的经验以及知识，能自动实现专家级精度的风控建模，客户无需模型专才也能高效构建风控模型。



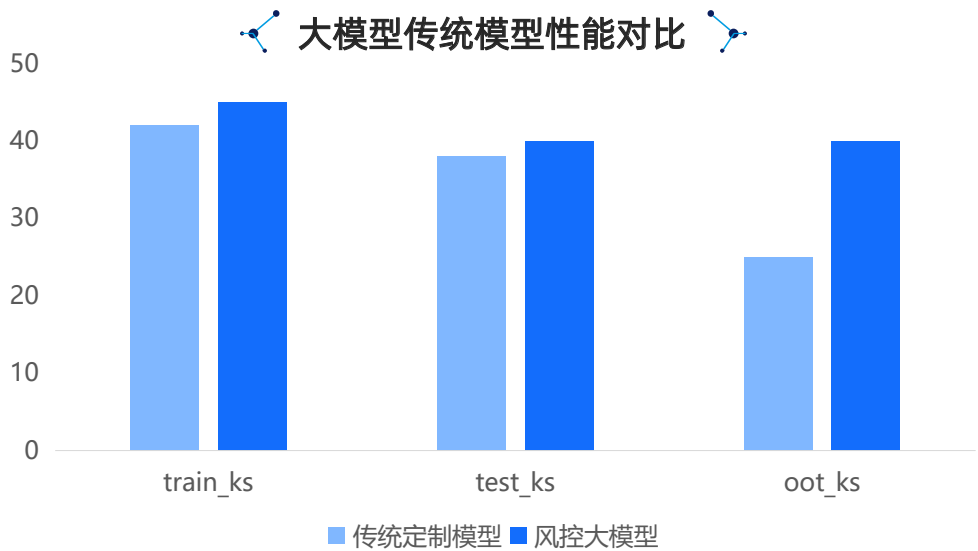
②部署提效：全流程自动化

在建模阶段只需使用少量提示样本，就能自动构建适配客户自身业务独有特点的风控模型，并且实现全流程自动化的部署上线，支持客户持续发布快速集成到自身的风控系统上，帮助客户风控策略部署效率提升10倍。



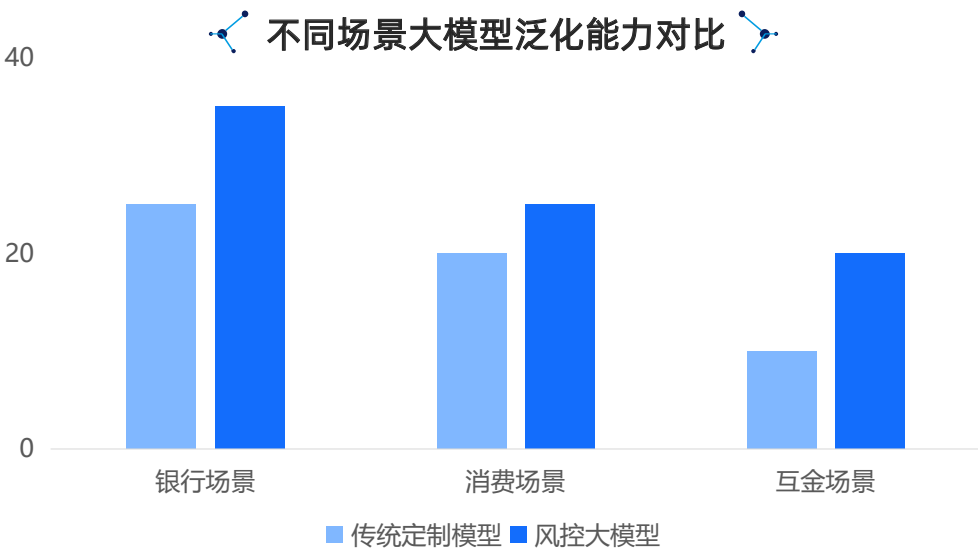
③性能提升：解决“小样本”风控难题

依托20多年沉淀的海量欺诈样本和多场景下的丰富风控模型能力，金融风控大模型能帮助样本积累有限以及新业务上线“零样本”的企业，高效解决“小样本”训练难题，模型区分度比传统模式提升20%，快速构建风控体系，支撑业务发展。



④360度模型评估

金融风控大模型为客户提供超越自身样本的模型评估体系，包含所有的风控模型评估指标，覆盖全量金融风控的细分场景，并在跨场景的泛化能力上提升30%，帮助企业新业务快速拓客。



(2)腾讯云天御内容风控

落地背景及挑战

AIGC技术在带来丰富创造力的同时，也让数据真实性、内容合规性、用户隐私和身份，以及伦理等问题遭到空前挑战。从监管视角出发，国家对AIGC的整个业务形态提出需要落实三类合规要求，分别是数据、内容和算法。

在数据合规方面，2023年4月国家互联网信息办公室公布的《生成式人工智能服务管理办法（征求意见稿）》（以下简称办法）提到，在AIGC特定的大模型下提供者应当对生成式人工智能产品的预训练数据、优化训练数据来源的合法性负责。

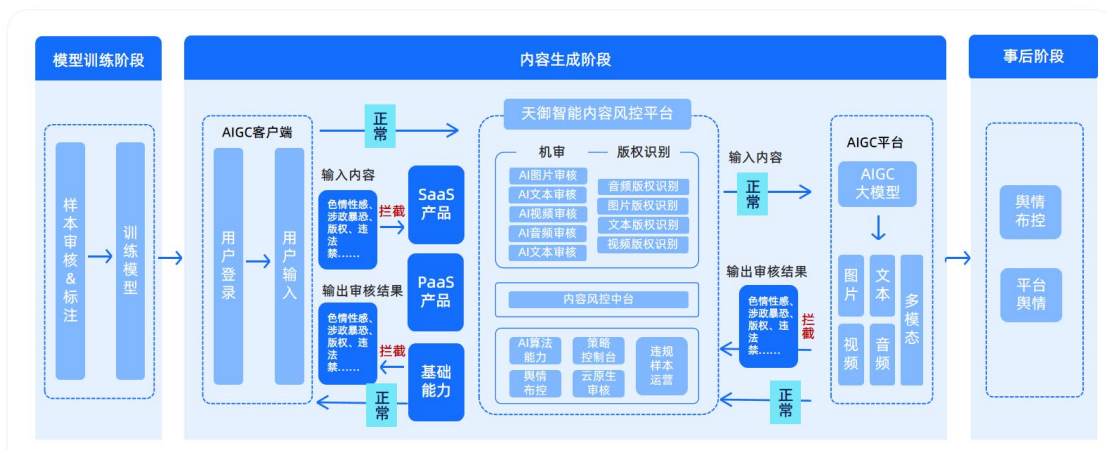
在内容合规方面，《办法》中提到利用生成式人工智能产品提供聊天和文本、图像、声音生成等服务的组织和个人，承担该产品生成内容生产者的责任；提供者应当依法依规对生成的图片、视频等内容进行标识，履行信息内容管理主体责任，加强本平台网络信息内容生态治理。

在算法合规方面，算法推荐服务提供者必须落实算法备案要求、算法评估要求、算法相对透明要求、用户权益建立健全机制等要求。

产品方案

腾讯AIGC内容风控全链路解决方案

AIGC平台生产的数据量级规模大，天御以接口输入、内容预处理、模型识别、策略辅助、平台调度分析、人工标注运营共6个维度为用户提供完整的内容安全解决方案，通过机器审核、安全专家、审核和版权服务等产品组合，让AIGC企业一次接入即可完成整套审核，一站式解决AIGC不同阶段的核心痛点。





方案价值

腾讯天御内容风控以接口输入、内容预处理、模型识别、策略辅助、平台调度分析、人工标注运营共6个维度为用户提供完整的内容安全解决方案，同时与腾讯云对象存储、云直播、云点播、实时音视频等云上工具打通，让AIGC企业在云上一次接入，即可完成全套内容安全审核，大幅提升效率。

①经验丰富：20 多年运营经验+万亿级违规处理。

②高准确率：机器准确率达 99.99%。

③一键接入：与腾讯云五大云组件联动。

④量体裁衣：提供 十几种定制识别服务。

⑤敏捷响应：小时级响应针对性服务。

(3)腾讯数据安全治理

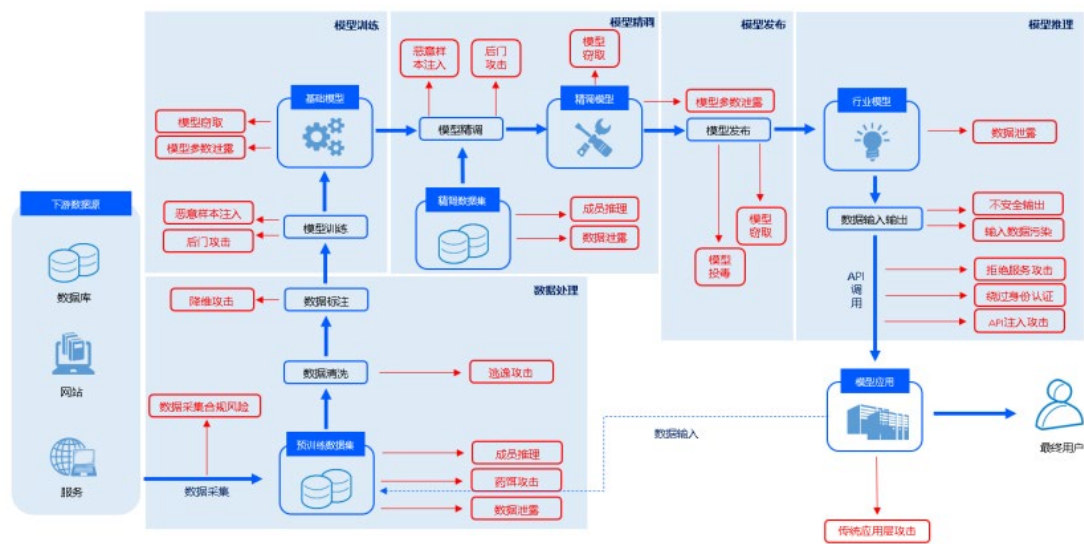
大模型数据安全合规要求

在AIGC技术的快速发展和应用过程中，相关数据安全合规问题也一直备受监管单位关注。最近几年，监管单位陆续下发了《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理暂行办法》等规章制度，以《网络安全

法》、《数据安全法》、《个人信息保护法》三大上位法为依据，完善了AIGC领域的监管框架，要求建立健全信息发布内容审核、数据安全、个人信息保护、应急处置等管理制度和技术措施，规定了AIGC服务的提供者和运营者应履行的各项合规义务，包括训练数据保护、个人信息保护、用户输入内容和使用记录的保护、安全评估、数据定期审核等内容。

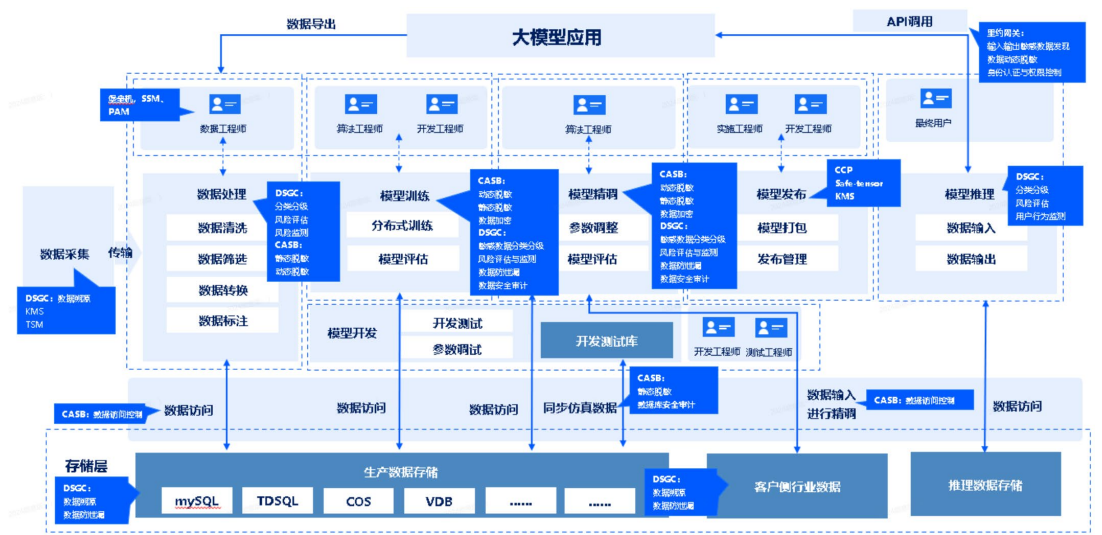
大模型面临的安全风险

通过分析AIGC业务流程每个阶段面临的安全风险，得出以下风险模型：



大模型数据安全保护解决方案

通过分析AIGC面临的数据安全风险以及合规要求，对AIGC业务流程的不同阶段匹配数据安全防护策略，方案如下：



注：DSGC：腾讯云数据安全治理中心 CASB：腾讯云数据安全防护网关

通过对大模型的用户、实体、模型文件实施分级别的访问控制，实现权限分离，通过数据安全治理中心、数据安全防护网关、机密计算平台等产品能力，实现大模型业务全流程中海量数据和大模型的完整性和保密性。通过此方案可以实现如下效果：



方案价值

①腾讯云数据安全治理解决方案针对AIGC场景做了适配开发，支持海量数据的敏感度识别和安全防护、审计溯源，防止隐私数据泄露、被滥用、被篡改，满足数据合规要求。

②腾讯云数据安全治理产品利用腾讯大模型能力和高性能算力，训练数据安全大模型并反向赋能大模型数据安全，在敏感数据发现、数据标注、数据防泄漏、血缘分析、自动化响应等能力上具备显著优势，敏感数据自动化识别准确度达99%以上，风险识别准确度达97%以上，大大减少漏报率和误报率，减少安全运营阶段的工作量。



第四章

生成式AI产业最佳落地实践



(一)路线一 - 标准软件产品方案

1.旷真律所-腾讯乐享AI助手

项目背景及挑战

旷真律师事务所是国内领先的法律服务机构之一。律所所在的知识密集型行业，对知识的沉淀、更新和应用，有天然的、更高的要求。

该行业的知识具有很强的专业性与实践性，需要不断学习和更新知识库。而传统的知识传承是“师带徒”模式，受带人水平、人员稳定性的影响很大，需要系统化的知识管理平台将隐形知识显性化，为他人所用、为组织所用。此外，AI技术的出现能够大幅提升法律文书工作的效率，快速分析帮助律师判断和决策，促进团队创新和协作，对推动法律行业的工业化进程至关重要。

产品方案

①导入内部知识文档4w+至乐享平台，打造专属AI知识库，员工使用乐享AI助手智能问答进行专业问题对答，提升知识学习效率；并通过反馈-实时提醒作者修改-管理后台进行跟进分析，促进知识迭代。

②打造企业培训阵地，实现学考练同步执行，提升培训效率；众多资深律师作为培训内容生产者，通过AI助手生成课程大纲、知识点，生成课后考题，提升知识生产效率。

③全面使用乐享AI助手、知识库、课堂、直播、论坛、项目等功能模块，助力旷真律所知识管理、学习培训、文化建设三大场景提效。

项目成效

旷真律所依托腾讯乐享打造团队协作和知识管理平台，实践效果显著：目前已在“旷真大学乐享平台”存档4万余份员工编撰的指引和案例，开设了873场培训和课程，220余场认证直播。平台最高月访问量达40万人次，折合每人每天平均访问10次。

腾讯乐享的创新意义在于它打破了传统企业知识管理的方式，通过AI助手，使知识的生产、获取、共享、迭代和管理变得更高效。不仅帮助旷真快速沉淀和传承知识，降低人员流动带来的风险，提高团队效率 and 创新能力；还助力企业文化落地，打造了一支更具有学习力、凝聚力和奋斗精神的团队，让旷真更好地应对外部环境的变化。另外，腾讯乐享的应用也有助于推动旷真的数字化转型，为律所行业的可持续发展注入新的活力。

客户评价

“借助腾讯乐享AI助手，目前律所已形成自己的AI知识库。员工调研显示，对典型问题的AI回答满意度达93分，端到端问题准确率达91%。现在有问题就找乐享AI聊，要内容就找乐享AI帮忙出，腾讯乐享是旷真人的智能大学！”

——旷真法律集团学习与能力发展部部长 龙庆

2.招商银行 - AI代码助手实践

项目背景及挑战

客户作为头部金融机构，积极拥抱智能化场景，率先探索AI驱动降本增效的落地方案；面临以下痛点：1、企业安全合规要求高，代码不可外泄；2、开源模型难以定制和调优、资源成本较高，产生投入产出管控顾虑；3、缺乏适用的标准评测数据集；4、IDE插件版本多，统一管理与多版本兼容挑战大。

解决方案



项目成效

- 1、成本降低：基于自研行业模型推理成本远低于大模型成本。
- 2、效能提升：成功打造出性能与体验更优、安全合规且满足业务场景的行业模型，建立起适合业务开发场景的模型微调评测标准，实现开发效率30%+提升。

(二)路线二 - 标准模型能力增强

1.百川智能-腾讯云向量数据库Tencent Cloud VectorDB

项目背景及挑战

大模型提供在线推理的服务背后有一个模块叫搜索增强，这个搜索增强服务需要向量数据库提供数据检索的功能，数据规模从1亿-100亿级别，对大规模、高性能、稳定性有比较高的要求。100亿级单索引存储规模，上千节点的集群管理，对向量数据库的架构设计，系统稳定性有着较高的要求。

产品方案

腾讯云向量数据库是一款全托管的自研企业级分布式数据库服务，源自腾讯集团自研的向量检索引擎 OLAMA，专用于存储、检索、分析多维向量数据。该数据库支持多种索引类型和相似度计算方法，提供多副本高可用特性，单索引支持10亿级向量规模，百万级 QPS 及毫秒级查询延迟，不仅能为大模型提供外部知识库，提高大模型回答的准确性，还可广泛应用于推荐系统、NLP 服务、计算机视觉、智能客服等 AI 领域。

①集成Embedding功能，高效易用

Embedding 功能是腾讯云向量数据库提供将非结构化数据自动转换为向量数据的能力，目前已支持多种语言的文本 Embedding 模型。可实现原始文本自动向量化进行写入/检索，对齐传统数据库使用体验，无需关注向量化过程。

- 数据写入/检索自动向量化，对齐传统数据库的使用体验，无需关注向量生成过程。
- 优化GPU处理速率，性能提升10倍，超高配速快速处理数据。
- 海量GPU资源池提供算力服务，按需使用，成本更优。

②端到端AI套件，AGI时代的知识库解决方案

腾讯云向量数据库实现面向RAG领域中内容检索的端到端解决方案，在向量数据库中集成文档预处理、自动向量化、可视化调优等功能，简化RAG中的数据处理以及检索流程，大幅降低接入门槛，对比传统处理方案召回率提升30%，实现界内最高召回率，加以可视化调优进一步帮助业务快速搭建定制化的知识库服务。

- 检索召回效果提升30%

在文档预处理、索引构建、结果排序等环节均进行了专项优化，并构建独特的精排方案，更适用于各类文档检索场景，召回率对比传统方案显著提升。

·接入成本降低10倍

相较于自行完成数据处理、向量化、创建索引等操作，AI套件方案提供简单、高效的解决方案，将知识库接入的时间从1个月提升到1小时，接入成本降低10倍以上。

项目成效

解决千亿数据量级下的数据分区、过期删除等问题，冷热数据分离，提升海量数据下的检索性能。

2.中国大熊猫保护研究中心+广东工业大学-TI平台*知识引擎

项目背景及挑战

大熊猫行为识别 “大熊猫的AI饲养员”。

产品方案

(1)广东工业大学、腾讯云、中国大熊猫保护研究中心联合开展人工智能+大熊猫保护探索。一阶段围绕大熊猫采食行为视频化自动识别技术进行研发。这个阶段，借助腾讯云一站式机器学习生态服务平台TI平台的，进行数据清洗、标注、模型训练，开发了全球首个【大熊猫行为智能识别模型】，并生成大熊猫日常管理数据，如饮食习惯、行为规律等。主要使用到的TI平台功能如下：

①依托平台便捷易用性，多人协同完成大熊猫视频数据的数据过滤、清洗、标签，快速构造大熊猫行为识别图像训练库。

②依托平台的工具、算力等资源，加速大熊猫行为识别模型的搭建、训练、调试、优化。

(2)未来，随着根据【大熊猫行为智能识别模型】生成的大熊猫日常管理数据的增多，如何更加高效地利用这些数据去辅助饲养员管理保护大熊猫，成为项目探索的重点。

腾讯大模型知识引擎将是我们重点探索的方向。比如，我们将大熊猫日常管理数据、大熊猫管理保护垂直领域的专业知识导入大模型知识引擎，让大模型对所有的数据、知识进行解析、重组，回答饲养员关于大熊猫日常行为状态问题，打造一个信息全面、知识专业的大模型

饲养员助手，为饲养员快速提供单个大熊猫或者大熊猫种群的知识信息汇总和分析（含真实数据知识化和规范知识），助力大熊猫保护。

3. 中原出版集团-腾讯云大模型知识引擎

项目背景及挑战

由河南省教育厅牵头，中原出版集团牵头建设，面向河南省全省上千所中小学提供智慧教育平台，通过7×24小时全天候大模型知识助教实现因材施教，打造教师助理及学生助手，整体提升教学效率及质量。

教育内容的准确性至关重要，需确保大模型生成的内容不仅符合教育标准，而且还要适应不断更新的教育政策和课程要求，中原出版集团要将河南省百万级的中小学教材文档导入，进行知识梳理配置，工作量庞大且复杂。

产品方案

①知识引擎有现成开放接口，支持外部文档通过接口写入知识引擎文档库，无需梳理文档配置问答对，可针对文档直接问答。

②大模型支持教学大纲设计、习题解析等场景问答，可通过模型精调提升问答准确率，提升教学效率和质量。

项目成效

目前已在河南省郑州市部分重点中小学上线，在备课材料、教学实施、学生管理及工作事务四大场景帮助教师提升效率；在知识学习、作业联系、备考准备、学习提升四大场景助力学生提升学习效率和质量。

4. 万榕信息-腾讯云大模型知识引擎

项目背景及挑战

万榕信息是一家成套开关设备产业互联网集成服务公司，与腾讯云携手打造面向成套开关设备行业提供解决方案和专业服务的万智云电气产业互联网平台，借助人工智能技术提升行业设计效率，同时探索基于大模型的创新业务，拓宽服务应用边界，提供更准确，个性化的解决方案。

①电器行业专业文档内容复杂，包括国家标准，过往历史技术文档，设计文档，网页信息等，包含图文表混排及大量数据，问答梳理困难。

②传统客服机器人问答时，无法基于文档生成设计、方案、框架类内容进行回答。

产品方案

①腾讯云大模型知识引擎具有解析复杂表格文档的能力，可将专业电气类技术文档内容准确解析，通过大模型进行理解推理，生成用户提问相对应的答案。

②大模型支持简单数学计算、数字范围判断，可解决数字范围类提问无法穷举的问题；大模型具备方案框架生成的能力，可应对用户方案设计思路类提问。

项目成效

①创建行业知识库，集结行业专家知识和经验，通过基于腾讯云大模型知识引擎打造的在线服务平台“榕博士”，供内部员工学习及合作伙伴查阅，促进知识共享。

②通过知识共享平台，研发人员可高效了解行业设计规范及内部产品标准，提效研发效率，售前人员可快速检索及定位产品文档手册，相应客户咨询，提升服务质量。

5.中原消金-腾讯金融风控大模型

项目背景及挑战

面对新增零售信贷业务风控难与存量零售风控业务风险上升的双重挑战，中原消费金融携手腾讯云天御将“风控策略”与“算法模型”深度融合，构建了一个以多元化的平台能力、技

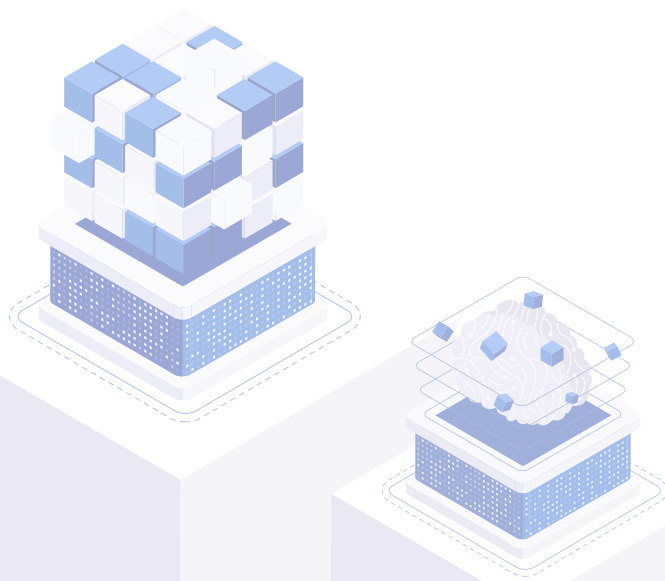
术能力、数据能力为基础，涵盖“获客+授信准入+贷中管理+贷后管理”全流程的智能风控体系。

面对日益庞大的市场需求和更加严苛的合规要求，中原消金如何搭建完善贯穿客户全生命周期、业务全流程的智能风控体系是一项持续性的挑战。

①传统的风控模型以专家模型和逻辑回归为代表，强调可解释性和稳定性。随着信贷业务走向普惠、经营的客群更加多样化，需要使用更有效的算法来做信息整合和客群分层。

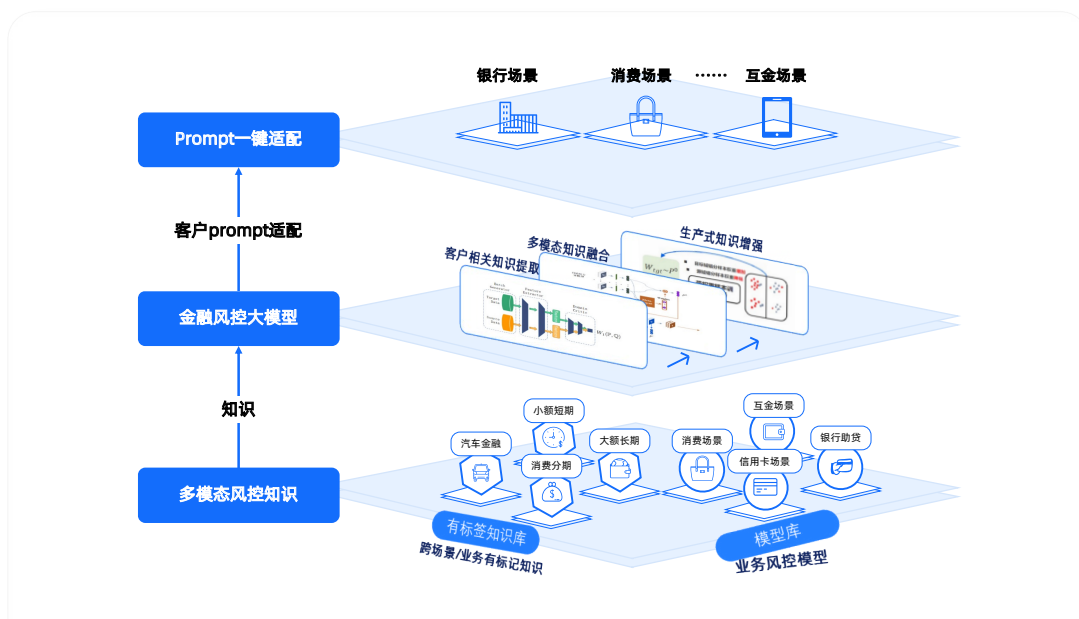
②为了更好地挖掘和发挥数据的价值，变量使用的量级和复杂度均会大幅提升，这对风控变量的管理提出更高的要求。

③围绕客户全生命周期的管理，模型从侧重于贷前、贷中、贷后等环节的客户风险识别，扩展到了包括客服、电诈、消保、协商还款等场景，更多地从经营的视角去审视模型带来的增益。



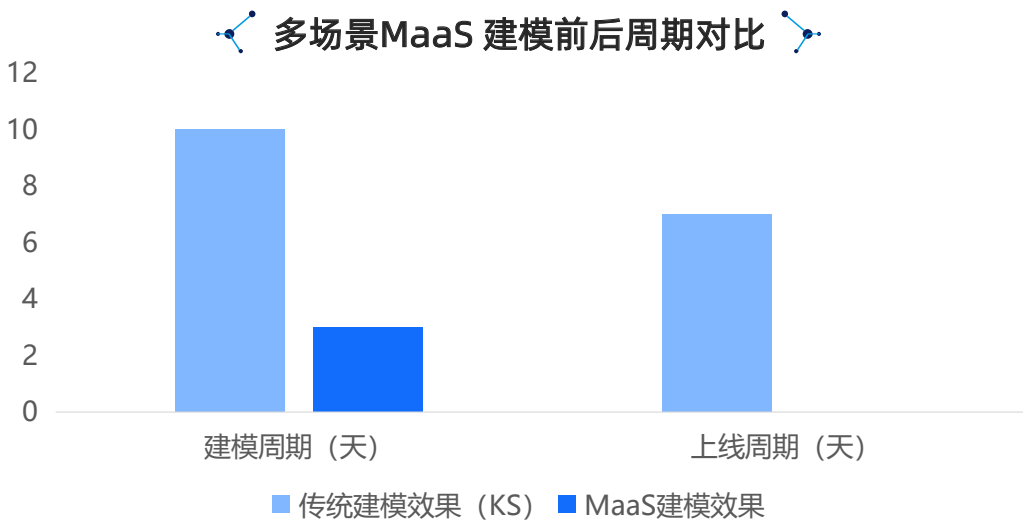
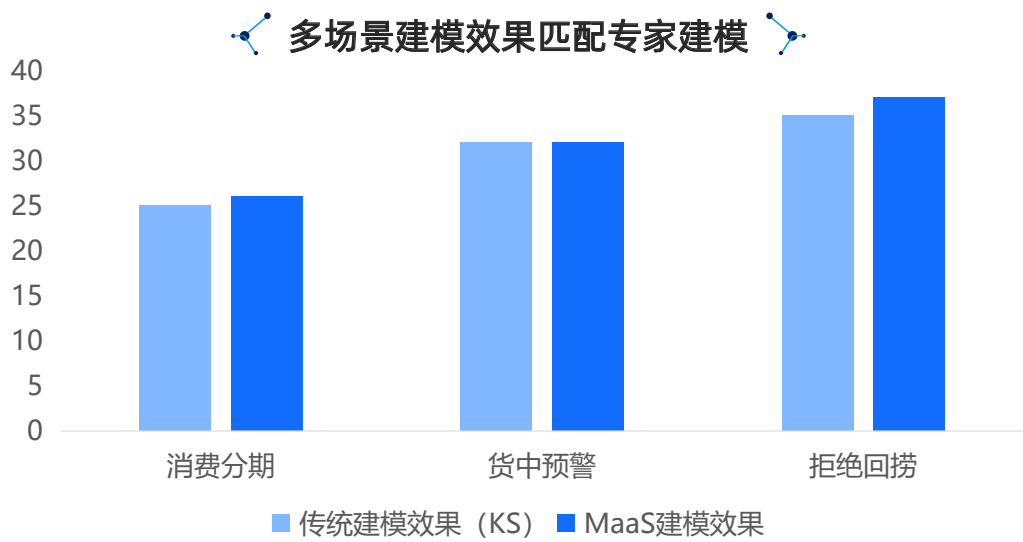
产品方案

面对新增零售信贷业务风控难与存量零售风控业务风险上升的双重挑战，中原消费金融携手腾讯云天御将“风控策略”与“算法模型”深度融合，构建了一个以多元化的平台能力、技术能力、数据能力为基础，涵盖“获客+授信准入+贷中管理+贷后管理”全流程的智能风控体系。中原消费金融还与腾讯云金融风控大模型所积累的风控知识和能力进行互动，更高效率地生产出专属于中原消费金融的风控模型。目前双方通过这样的模式联合共建了10个风控模型，应用在反欺诈、信用初筛、拒绝回捞、客户经营等多个场景。



项目成效

在双方联合构建的智能风控体系的支撑下，中原消费金融有效降低欺诈行为的发生，并在业内率先正式推出“7天无理由还款”权益，该权益是中原消费金融公司面向首次借款的用户推出的一项鼓励用户理性消费的保障性权益，也实现了消费金融行业产品形态“从0到1”的重大突破。建模周期从原来的 10天降至 3 天，上线周期从传统的7天升级为小时级上线。



(三)路线三 - 定制化模型精调训练

1.长安汽车-行业大模型

项目背景及挑战

客户APP、车机的小安助手及客服中心产品需升级能力：

①目前采用关键字匹配的方式回答用户问题，若未匹配则不能很好的回答响应，用户体验较差。

②后端知识库存储了长安的车型知识，目前是人工录入，随着车型越来越多，效率低成本高，无法满足业务需求。

产品方案

基于汽车行业大模型，结合客户数据，训练长安大模型。利用大模型泛化解析能力，解析客户手册、文档，生成知识对，完善知识库；再基于大模型语义理解能力，结合知识库，对用户提出的问题进行更准确的回答。

项目成效

由周级录入，升级到分钟级生成天级校对，效率提升数十倍。人工录入质量层次不齐，模型生成标准化，不依赖专家经验。人工方式人员培养难，模型方式能力可迭代、可扩展。成本随业务规模增长可控。由单轮会话提升到具备多轮对话能力。对传统AI闲聊，升级到大模型闲聊，用户体验大幅提升。通过大模型兜底回复，提升问题解答率，提高结果正确率。

2.某大模型企业

项目背景及挑战

客户自研文本大模型和语音大模型在上线后，遇到了内容审核的多项难点：

①区别于真人聊天，用户会把AI当做问答工具，会涉及到历史、政治、价值观等话题，平台的内容安全风险增加。

②产品迅速破圈，存在少量高危用户会尝试试探平台底线。

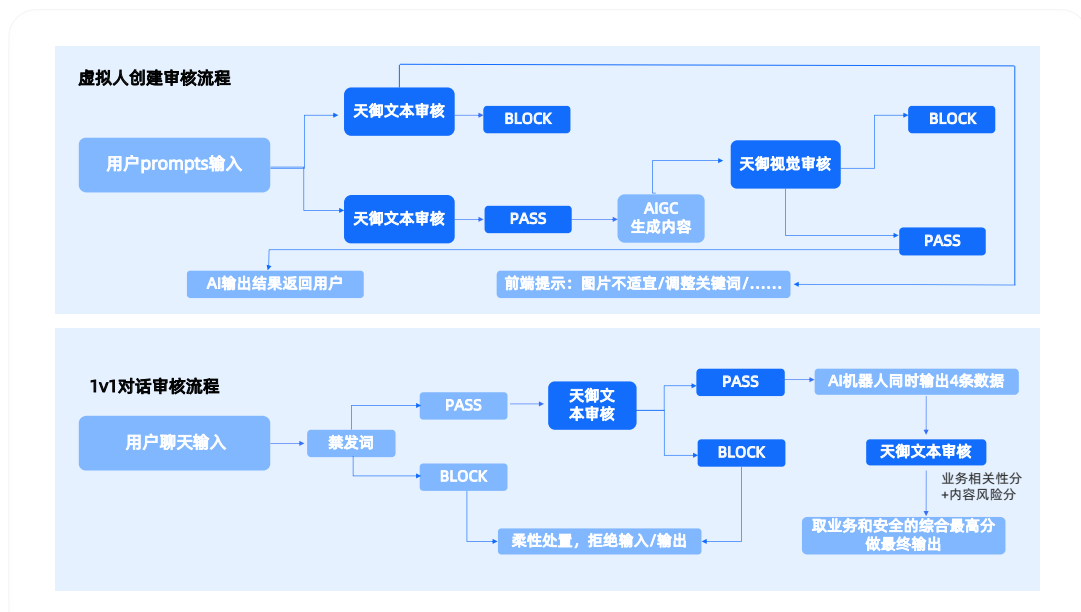
③由于场景的特殊，违规数据中色情话题比例较高。

当前AIGC内容安全受到国家高度重视，监管要求AIGC技术提供商要在数据、合规、算法三大形态上落实合规要求，而即将上线的自研文本大模型和语音大模型面临着新型AI 伪造、虚假、违背伦理、不适内容、未成年人等违规内容，以及AI 问答产生错误引导等内容，大大增加了审核的难度和复杂度。



产品方案

针对 AIGC 技术所带来的内容风险挑战，腾讯安全依托二十余年的安全对抗经验，为该大模型企业提供了AIGC全链路内容安全解决方案。基于数据合规、内容合规、算法合规等相关政策要求，为企业已上线的大模型平台提供包含审校服务、安全专家服务、机器审核服务、版权保护服务五大能力板块，覆盖AIGC类应用从模型训练到内容生成到事后运营全过程的内容安全建设，确保AI大模型可信、可靠、可用。



项目成效

天御AIGC内容安全解决方案覆盖了“虚拟人创建”、“虚拟人 1V1对话”“热点话题审核”等多个场景，为企业提供十几种定制识别服务，机器审核准确率达 99.99%，有效保障了大模型生成内容的合规、可用。

3.阅文集团-TI平台行业大模型精调

项目背景及挑战

作为国内头部的网文创作公司，阅文集团希望使用生成式AI技术，在文生图/图生图场景中，面向C端用户的应用需要提供高QPS、低模型耗时的方案。

产品方案

客户诉求

文生图/图生图，大模型文生文推理部署降本增效

算法和工程团队精力有限，模型数量多迭代款

自行搭建推理集群和服务平台，模型更新复杂

降低维护难度

推理成本高昂，推理速度已达瓶颈

千百亿级别模型推理部署，推理吞吐和时延存在瓶颈

降低推理成本

方案价值

推理和加速性能

- 提高高性能文档计算节点及推理加速能力，推理性能提升30%

高效大模型部署和服务接入

- 提供动态batch和自动服务调度机制，大幅提高服务吞吐量
- 模型优化和部署模块一键加速和发布模型，支持模型推理加速的过程中动态加载不同的LoRA模型

产品示例

1 模型文件上传文件存储，导入模型

2 进行模型优化

3 部署模型



项目成效

阅文基于腾讯云TI平台搭载的angel推理加速引擎，推理能力平均提升30%，在文生文领域，更是把成本降低到原来的1/4。此外，TI平台提供动态batch和自动服务调度机制，大幅提高了服务吞吐量，模型优化和部署模块一键加速和发布模型，支持模型推理加速的过程中动态加载不同的LoRA模型，极大减轻了服务部署的维护成本。



第五章

生成式AI发展展望



通过上一章节分析各行业的最佳落地实践，我们可以发现生成式AI技术已经取得了突破性进展，并且在各个领域得到了广泛的应用。随着不断的研究和实践，生成式AI技术本身仍然存在着许多可以突破创新的方向。更加高效灵活的架构、覆盖模态的持续扩充、具身智能等都是可期的生成式AI新技术趋势。这些新技术突破，必将带来更多的创新成果和应用场景。

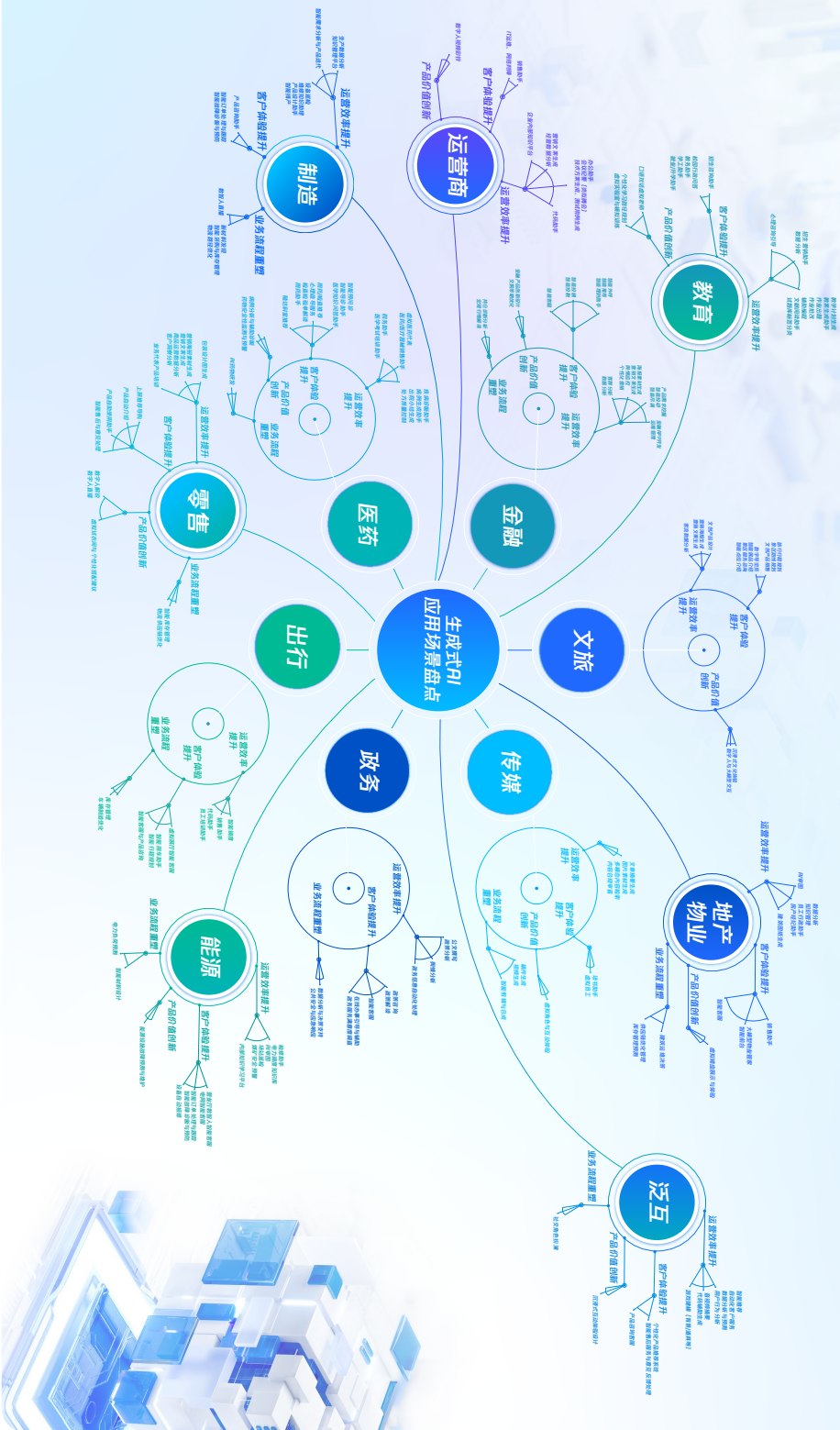
首先，我们认为新架构是一个至关重要的技术创新方向。本轮技术革命源头是Transformer架构的诞生，但距今已有7年之久，而且计算效率距离人脑还有很大的提升空间。技术人员始终在开发下一代人工智能架构，卡内基梅隆和普林斯顿联合研发的Mamba架构正在向Transformer发起挑战。未来势必涌现出更高效、更灵活的架构，带来更多的优势和创新，为未来生成式AI应用扩展奠定坚实基础。

其次，我们关注到多模态能力持续扩展的趋势，如基因、生物电、雷达信息等。基因组学等信息可以视为另一种“语言”，复杂度极高，生成式AI技术正在突破这些未涉及的模态领域。这一技术趋势将带来生成式AI在生命健康、科学研究、汽车等领域更加丰富的应用场景，引领更多行业开启科技变革。

第三，具身智能是未来数年生成式AI又一技术突破点。具身智能，是指通过将生成式AI应用到物理世界，载体可以是机器人、也可以是智能设备，从而实现更加智能化和个性化的互动体验。具身智能延展了生成式AI技术的应用边界，家居、交通等领域可能诞生产品价值创新的革命性变化。

生成式AI技术的蓬勃发展已经为我们推开了通向通用人工智能的大门。研究人员还在不断推动技术进步，企业也将积极采纳新技术，并应对应用落地过程中的挑战，最终共同努力实现通用人工智能的目标。

生成式AI百大应用场景



参考资料

- 1.Gartner Inc., How to Choose an Approach for Deploying Generative AI, G00794559
- 2.Gartner Inc., How to Calculate Business Value and Cost for Generative AI Use Cases, G00805323
- 3.Gartner Inc., 2024 CIO and Technology Executive Agenda: A China Perspective, G00807509

结语

企业在AI应用探索的道路上持续前行，许多企业已借助AI技术成长为行业领导者，并开辟了新的业务版块。作为新技术的代表，生成式AI为企业带来了新的跃迁式发展机遇，推动企业向前迈出更大的步伐。在这个变革的时代，企业利用生成式AI技术不断创新，带来了更广阔的发展空间，并保持在市场竞争中的领先地位。

腾讯云基于行业实践的复盘总结，提出了生成式AI应用场景矩阵和生成式AI应用落地路线两大方法论，希望为企业开启生成式AI旅途规避风险、打好前站。

整体而言，生成式AI在不同产业的广泛应用已经泛起了第一波涟漪。随着技术进一步迭代精进，以及应用落地路径的逐步成熟，生成式AI必将持续掀起新的技术应用浪潮，在全球各行各业产生颠覆性变革。

腾讯云致力于持续深入挖掘技术的潜力，以助力实体经济的发展，成为实体产业蓬勃发展的“放大器”。实体产业是国民经济的基石，也是社会民生的保障，腾讯云希望充分发挥自身的技术优势，为实体产业的进步贡献力量，实现更广泛的社会价值。

