人工智能原生云构建与加速核心能力指南

目录

背景

从AI云到原生AI云：云服务的比较分析

平台能力需求

面对原生AI云的挑战

AI加速而生：腾讯云（AI原生）全景分析

云平台架构能力

- **基础设施层**

  加速计算

  网络和边缘加速

  存储加速

- **模型库**

- **工程工具层**

  部署和微调加速

  内容质量管理

  数据处理效率提升

  发展提升

- **应用层**

- **全栈式安全解决方案**

结论

关于腾讯云

参考材料

背景

在人工智能时代，我们正处于一场非凡的技术革命之中。这一深刻的转变，以其动态的势头和重大影响，正在重塑全球商业格局和社会进步。随着人工智能创新的快速发展，它们正在渗透到决策、创新和价值创造的各个领域，成为社会进步的关键驱动力。跨国公司、初创企业、成熟的研究中心和个别先驱者 alike 都不可避免地被这场由人工智能驱动的变革浪潮所席卷。

全球科技格局随着在大型语言模型、语音模型和视频模型等领域的突破而跳动着创新的脉搏，这些突破不断推动人工智能技术的边界。这种竞争动力不仅激发出新想法，还推动着各行各业持续的能量波。人工智能技术的演化和集成使得技术提供商能够完善和提升他们的产品，加强他们的竞争地位。例如，独立软件供应商（ISVs）越来越多地将人工智能功能融入到他们的产品中，以在这个变革时代确立自己的地位。正如Gartner在《CTO的生成式人工智能技术景观指南》中预测的那样："到2026年，超过70%的独立软件供应商（ISVs）将在其企业应用中嵌入生成式人工智能（GenAI）功能，这比目前的不到1%有大幅增加。"

生成式人工智能已成为商业领域的颠覆性力量，吸引了全球各行各业高管们的关注。其关键优势在于自动化决策过程和内容创作，通过提高效率和增加价值来革新商业运营。正如Gartner在《CTO的生成式人工智能技术景观指南》中所预测："到2026年，超过80%的企业将使用生成式AI API、模型，或在其生产环境中部署了生成式AI赋能的应用程序，这一比例将从今天的不到5%显著增加。"

所有这些变化标志着人工智能原生时代的强劲开端。在这个阶段，大型语言模型（LLMs）将作为基础技术，推动现有应用的显著转型，并催生全新的应用类别。Copilot和AI Agent的引入是这一转变的强烈证据。

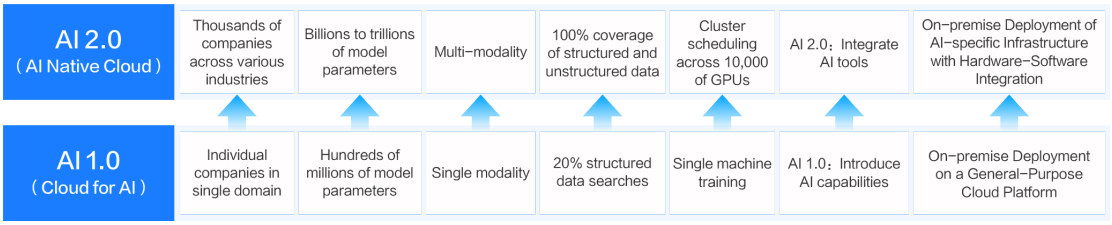
在不久的将来，托管在云上的大多数应用程序将平稳过渡到AI原生应用。

AI原生应用依赖于AI原生云，这促使云服务持续进化以满足AI原生时代的需求。下一代AI原生云解决方案将赋予IT领导者精细调整在这个转型时代价值、成本和风险之间微妙平衡的能力，使他们能够自信地与当前趋势保持一致并引领行业进步。

从人工智能云到AI原生云：云平台能力要求比较分析

我们激动地见证了从AI云到AI原生云的转变，这标志着云计算功能的重大飞跃。新兴的AI原生云不仅仅是一个技术进步，它还成为了推动用户业务转型和创造力的核心力量。AI原生云无缝地将AI技术嵌入到云计算服务的各个方面，为用户提供更智能和自动化的服务。此外，AI原生云优先考虑开放性和生态系统培养。通过提供强大的API和SDK，它为第三方开发者营造了一个动态的环境，以促进尖端AI原生服务和应用的创造。

图1：从云AI到AI原生云的转变



以下概述了增强能力的必要性：

1. 涵盖性以实现更广泛的参与：在过去，人工智能技术仅限于特定的行业和用户群体。然而，随着AI原生技术的出现，通过大型语言模型（LLMs）解锁了广泛的技能，使不同行业中的用户广泛参与成为可能。云平台必须具备可扩展的架构，以有效地适应这一多元化的参与者格局。建立一个稳健的框架对于满足各种规模和需求的用户至关重要，确保云生态系统中的每位用户都能获得定制的解决方案。

2. 精通训练大型语言模型：此前，人工智能模型的参数量通常在数亿范围内。随着AI原生时代的到来，模型参数量发生了巨量增长，达到了数十亿、数百亿，甚至万亿级别。云平台必须具备管理这种前所未有的训练任务规模的能力。对于它们来说，不断优化计算资源以满足日益增长的计算需求，同时保持最佳效率和成本效益，至关重要。

3. 多模态支持。过去，用户主要利用单一模态模型，如语言模型、语音模型和视频模型。在AI原生时代，整合并理解多种数据模态的多模态模型将增强功能并应用于更广泛的应用场景。

4. 加强多模态检索功能：在过去，用户主要依赖于结构化数据。然而，随着嵌入技术、多模态特性和AI原生时代向量化的出现，文档、音频、图像和视频等非结构化数据现在正被有效利用。云平台必须提供强大的跨模态检索能力，以支持这一数据利用方式的转变。

5. 简化集群调度：在AI原生时代，随着模型参数激增，单机训练显得不足。云平台必须支持集群调度，以容纳数千甚至数万个GPU来满足大型语言模型（LLM）训练的基本需求。集群调度系统应展现出智能和效率，自动优化资源分配，减少等待时间，并提高整体训练效率。

6. 授权发展提升：在AI 1.0阶段，开发者需要具备对云服务和AI的坚实理解，才能有效构建、训练和部署AI应用，通常需要整合各种工具和平台。然而，在AI原生时代，AI开发的门槛显著降低。开发者现在可以利用简洁、高质量的代码和工具迅速创建和部署AI应用。此外，云服务提供商提供了多样化的预训练模型和可定制模型选择，进一步简化了AI应用的开发流程。这不仅极大地提升了开发效率，还极大地减少了代码量，使得开发者能够更多地专注于软件产品设计。

7. 在本地部署中的适应性：在目前的AI原生环境中，用户对安全和数据隐私的关注程度达到前所未有的水平。随着AI基础设施相关成本的不断上升，企业越来越重视内部基础设施提供的规模经济和运营效率。因此，将AI解决方案本地部署以促进LLMs的培训和推理已成为越来越多的用户的优选方法。尽管传统用户通常会管理计算资源并部署AI应用程序。

在通用云计算平台上，集群配置和网络等关键领域对可扩展性和性能的要求在AI原生时代已经变得极其严格。这要求专为AI场景设计的专业化计算、网络和存储基础设施，需要硬件和软件紧密合作，以实现满足用户对高性能和可靠性需求的本地部署。

8. 确保内容质量和安全：在人工智能原生时代，云平台不仅要满足数据、应用程序和网络安全的传统标准，还要应对内容质量保证和安全性的不断变化挑战。面对可能含有敏感数据或侵犯知识产权的内容，用户依赖云平台提供强大的“基线策略”来防范此类风险。

在上述观点的基础上，在原生AI时代，一个具备端到端能力的强大云平台应运而生，该平台涵盖了基础设施、模型、工程工具、应用程序以及AI信任和安全五大关键组件，成为用户熟练驾驭这一变革时代技术格局的首选。

面对原生AI云的挑战

随着人工智能原生时代的全面展开，云计算技术的发展充满挑战。新兴的人工智能原生云必须持续创新，以应对以下七个关键技术挑战：

1. 自动部署：随着AI原生生态系统的不断发展，云平台必须优先考虑出色的用户友好性，以降低入门门槛。克服关键技术挑战，如简化标准和AI环境的部署流程，以及通过简单点击实现无缝GPU驱动程序安装，是平台必须克服的关键任务。

2. 自动化操作：用户努力优化GPU计算能力，尤其是在由数千甚至数万台GPU组成的集群中，效率至关重要。下一代云平台必须具备智能自动化功能，以实现集群操作的流畅性，包括高效的调度和容错机制，以保持稳定性并最大化大规模部署的计算能力。

3. 提升集群性能：在不断发展的人工智能原生云环境中，提高存储和网络传输速度，同时最小化计算过程中的闲置时间至关重要。云服务必须提升存储和网络传输速率，以最小化

计算延迟显著。努力实现零数据包丢失、提升承载能力，并达到2Tbps的集群吞吐量，代表着性能优化的顶峰。

向量化技术：向量化技术将非结构化和结构化数据转换为向量表示，促进有效的相似度检索，并为训练大型语言模型（LLMs）提供稳健的数据基础。该技术巧妙地存储和检索以向量格式表示的模型训练参数，加强大规模并行计算，并加速模型训练流程。此外，向量化技术成为用户解决通过RAG部署生成式人工智能过程时遇到的诸如幻觉、知识停滞和数据安全问题等挑战的关键解决方案。

5. 模型微调：在AI原生云的范畴内，LLM（大型语言模型）的微调对于最大化其可用性至关重要。考虑通过模型优化进行微调的人工智能编程助手，它能够生成符合用户偏好的优质代码。然而，模型微调是一项复杂的任务，需要仔细关注数据质量、审慎选择LLM、保持泛化能力，以及始终坚持伦理和监管标准。

6. 兼容性与可扩展性：即将推出的AI原生云，旨在适应涵盖公共云和本地部署的多元化部署策略，将面临一系列技术挑战。首先，该平台必须解决兼容性问题，确保在多样化的部署环境中实现一致的性能和用户体验。可扩展性成为多中心部署的另一个关键方面。随着用户企业的发展和市场需求的变化，AI原生云应具备无缝扩展的灵活性，以满足不断变化的计算需求。

7. 减少幻觉：在AI原生时代，生成式AI输出的精确性至关重要。下一代云必须通过利用高级调校工具链和综合RAG解决方案，巧妙地减轻广泛的LLM对话中的幻觉，从而保障生成内容的准确性和可靠性。

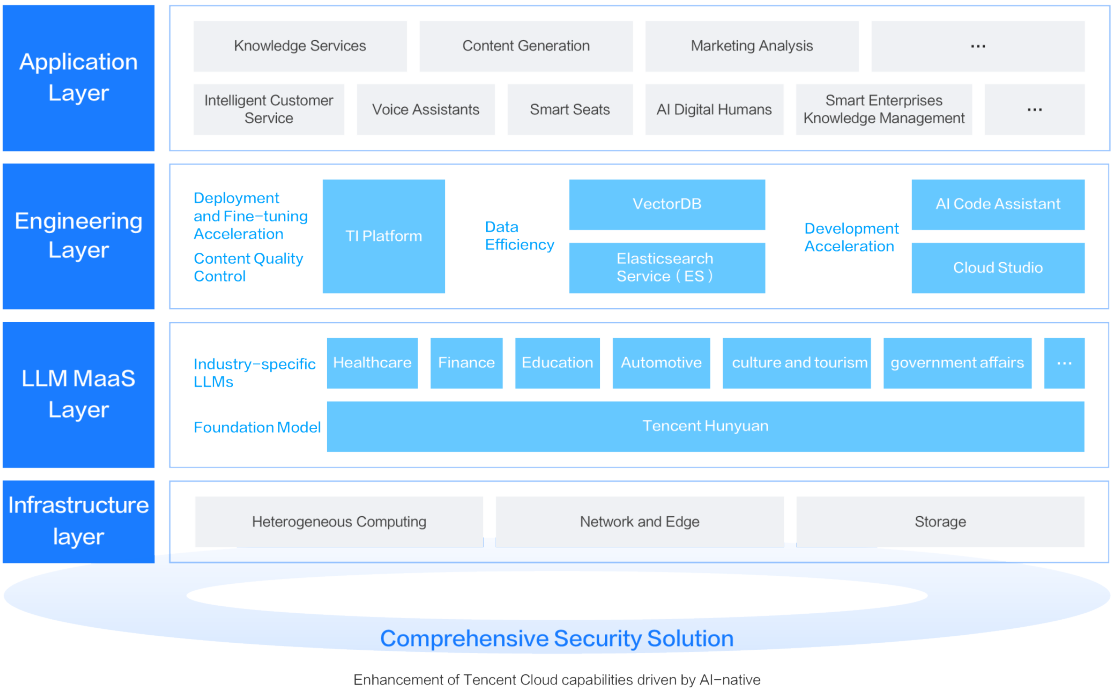
面对新时代的挑战，AI原生云的出现不仅象征着技术升级，也展现了创新思维和对卓越的追求。唯有持续推动技术边界，云平台才能在AI原生趋势中保持领先地位，为用户提供强有力的支持和丰富的增长机遇。

图2：AI 2.0的进步

人工智能加速而生：腾讯云（AI原生云）平台架构能力全景分析

腾讯云向用户提供由生成式AI驱动的先进云架构，涵盖五大关键能力：AI基础设施、模型与框架、AI工程、AI应用和安全。下一代云计算助力全面加速大规模语言模型（LLM）的训练、推理和应用部署，释放各个行业中多样化MaaS的效率，并加速AI原生应用程序的采用。

图 3：腾讯云（AI原生云）平台架构能力



基础设施层

基础设施层包含三个关键能力：计算加速、网络和边缘加速以及存储加速。这些特性旨在弥补计算、存储和网络性能上的不足，共同构建起一个坚固可靠的平台基础。

加速计算

人工智能已经进入了一个快速发展的阶段，尤其是在竞争激烈的生成式AI领域。为了抓住机遇，用户需要迅速迭代他们的语言模型（LLMs）。这推动了大规模、高性能异构计算能力的需求激增，将焦点从以CPU为中心的计算范式转变为以GPU为中心的计算范式。

此外，在生成式AI时代，模型参数已达到万亿级别，这在训练和推理过程中都带来了前所未有的可扩展性、性能、容错性和成本效率方面的挑战。

首先，用户对计算错误有极低的容忍度，因为训练中断需要从头开始。计算和时间成本的双重压力使得任何延迟都无法接受。

其次，每个阶段——预训练、后训练、微调和推理——都要求

GPU计算能力，为用户创造了一个持续且具有挑战性的环境。在维护模型性能和产品体验的同时，用户还必须确保自身的可持续性。寻找成本效益高的异构计算能力以最大化投资回报率是他们面临的关键问题。

腾讯云异构计算

用户痛点

- 培训

效率： 在快速演变的生成式人工智能领域中，效率和时间对商业成功至关重要。客户需要大幅减少大型语言模型（LLM）的训练时间，从而产生了对高性能计算指数级的需求。

稳定性： 培训必须不间断；任何中断都意味着从头开始，这是不可接受且无法容忍的。

- 推断

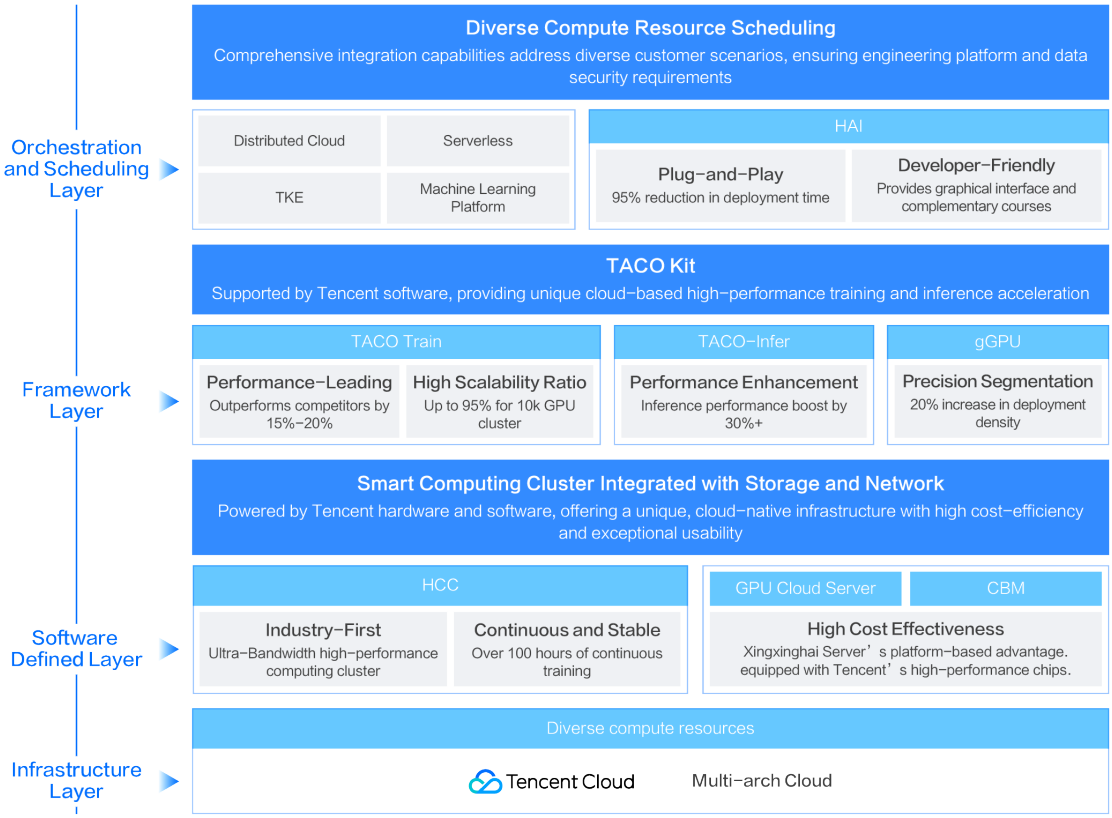
延迟： 人工智能推理需要用户请求的前向传播，这需要提供高吞吐量和低延迟的计算支持以维持无缝的用户体验。

费用： 与训练需求大规模GPU算力不同，推理需要成本效益的计算解决方案。

产品解决方案

在基础设施层，腾讯云异构计算作为一个关键工具，通过其多架构云平台提供强大的AI支持。通过无缝整合和优化软硬件，它使得并行计算能力更强大，从而显著提升大型语言模型（LLMs）的训练和推理过程。腾讯云提供包括高性能计算集群（HCC 2.0）、云裸金属（CBM）、云服务器、HAI、容器和云函数等多种实例选择，提供了一个多样化且全球领先的实例选项系列。

图4：腾讯云的产品解决方案



优势

领先规模 腾讯云管理着超过1.5亿个计算核心，提供行业领先的16 EFLOPS（1.6 x 10^18 FLOPS）基于人工智能的计算能力。

卓越表现： 在硬件方面，它配备了最先进的GPU芯片和腾讯独有的、行业专属的3.2T RDMA StarPulse网络，适用于集群。在软件方面，腾讯的TACO加速框架，包括TACO Train和TACO Infer，针对训练和推理加速进行了优化。这一集成的软硬件解决方案提供了卓越的性能，将万亿参数模型的训练时间缩短了80%。

超级稳定性： 腾讯云提供行业领先的99.9%服务等级协议。统一的TACO软件界面抽象底层硬件差异，确保稳定平台。

易用性： 该平台具有一键式GPU驱动安装、基本环境的自动化部署、批量任务管理和资源管理工具。它支持快速部署各种AI环境，包括ChatGLM-6B和StableDiffusion。

成本效率： 腾讯云支持按需使用vGPU和qGPU容器级别的资源分区，粒度可达5%，提供精确的GPU弹性服务。其云原生网络架构允许混合训练和推理工作负载，最大程度地为用户节省成本。

成功案例

腾讯云异构计算服务覆盖全国90%以上的LLM客户，巩固了其在市场上作为首选和可靠的AI基础设施的地位。它已经交付了

大规模异构智能计算服务，为众多行业客户提供支持，例如百川智能、MiniMax、Moonshot AI、快手、小红书（RED）、XtalPi 和什么值得买。

图5：腾讯云的成功故事

部署选项：公有云、专用云和本地部署。

腾讯云智能计算基础设施

用户痛点

本地部署：用户要求因自用或隐私问题，将计算能力、数据和整个AI解决方案托管在本地。

综合人工智能解决方案：为搭建本地人工智能中心，用户需要一套涵盖计算能力、网络、存储、加速框架等多个方面的全面解决方案。

多架构云 用户寻求在各个软件和硬件架构上的全面支持。

开放性和兼容性： 用户需要智能计算解决方案足够开放，以便与现有硬件设备集成。

效率与稳定性： 用户需要顶级推理和训练性能以提高效率，同时确保系统保持稳定和健壮。

操作能力： 鉴于运营和维护团队规模有限，人工智能解决方案必须集成全面的运营和维护功能，以确保系统稳定运行。

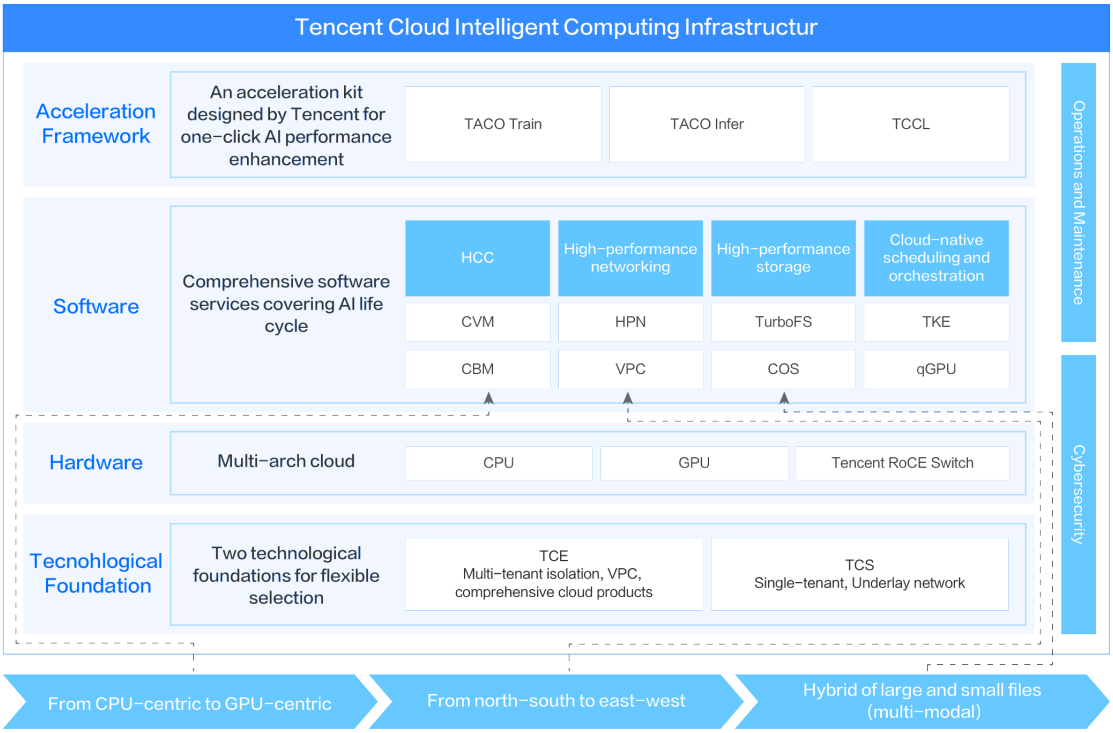
产品解决方案

腾讯云智能计算基础设施旨在为AI应用场景量身打造的计算基础设施。它涵盖了如HCC、TurboFS和IHN等产品，提供全面的高性能计算、存储和网络能力，以支持上层应用平台的人工智能解决方案。

腾讯云智能计算基础设施可在TCE技术基础上实现，具备多租户隔离和全面云产品特性。

或在TCS技术架构上，它提供灵活、轻量级和经典的底层网络。这种灵活性使其能够轻松满足各种AI场景的需求。

图6：腾讯云的产品解决方案



优势

综合人工智能解决方案： 腾讯云提供了一套全面的智能计算产品，包括计算、网络、存储和数据库。这些组件可以根据具体需求集成到一个统一系统中，或单独选择使用。

杰出表现： 通过腾讯的公共云和其超大规模智能计算中心展示，该套件提供了卓越的性能和稳定性。
自力更生： 凭借软件和硬件的自主研发，腾讯对其核心技术的完全自给自足得到保证。

安全和合规： 提供全面的云安全服务，符合分级保护、可信云、关键基础设施和国家级加密要求的标准。

开放兼容性： 腾讯的软件包括标准接口和成熟的协议，能够与第三方硬件产品集成。

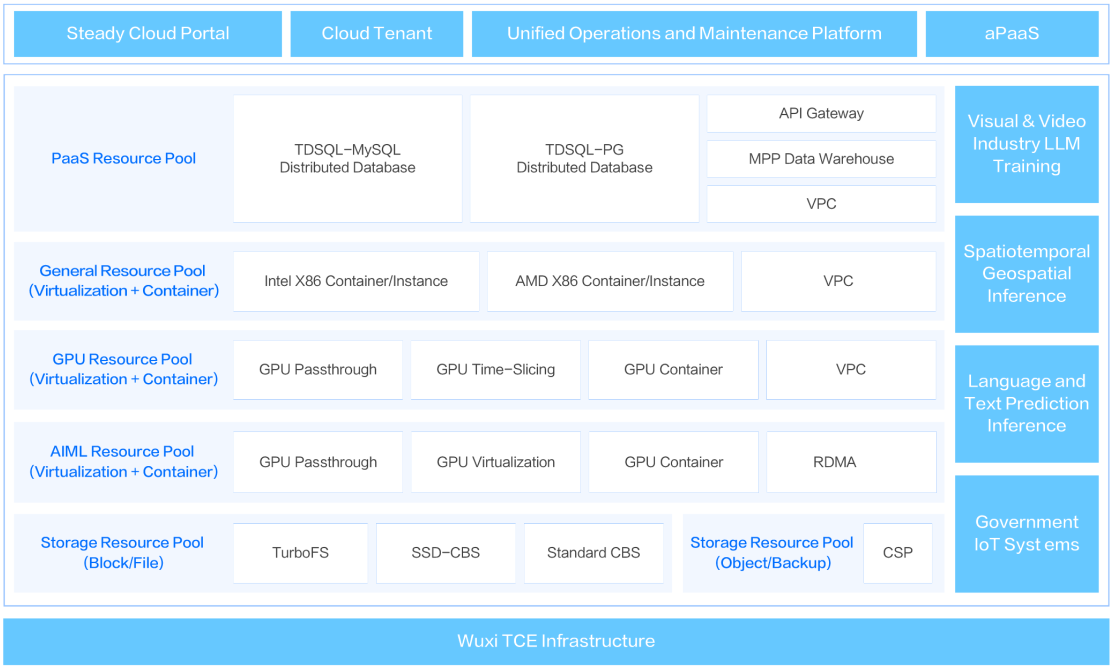
多架构云： 与主流硬件供应商兼容，并支持混合部署。

成功案例

图 7：腾讯云的成功故事之一

基于广州，Steady Technology是一家国内领先的数据中心（IDC）和IT基础设施服务提供商，为包括金融、制造和零售在内的多个行业提供IDC解决方案、数据中心服务及相关网络服务。Steady Cloud以其核心基础设施——腾讯云智能计算基础设施为基础，支持异构计算、通用计算、分布式存储、云原生和安全服务。它提供一站式自助云服务、弹性云、大规模并行计算和在线LLM训练服务，配备高性能计算集群和全闪存分布式存储，为人工智能（AI）和机器学习（ML）场景提供全面解决方案。

图8：解决方案架构图



部署选项：

私人部署

网络和边缘加速

尽管大多数用户专注于提升计算能力，但他们往往忽略了生成式AI对网络和边缘计算的挑战。随着快速

AGI 的发展使得基于 LLM 推荐算法或智能生成能力的软件和硬件工具将成为下一代超级应用浪潮的驱动力。无论其形式如何，这股由 LLM 和生成式 AI 主导的新一波 AI 原生应用，将对底层计算网络和应用分发网络提出全新的要求。

用户依赖前一代计算网络可能会在模型训练中遇到负面影响。例如，计算集群可能难以管理大规模的训练数据集，缓慢的网络传输速度会延长训练周期，多个训练任务之间缺乏带宽隔离也可能导致干扰。

此外，用户需要解决生成式AI发展给应用分发网络带来的新要求。首先，降低延迟：如智能客服和翻译等应用需要实时响应，更高的延迟会削弱生成式AI的有效性。其次，增强安全性：AI模型和数据面临如DDoS勒索攻击和非法数据抓取等威胁。鉴于大多数用户的收入和估值与大型语言模型（LLMs）和超级应用（Super Apps）紧密相关，加强安全措施是紧迫的。

腾讯云连接网络

用户痛点：

对高带宽的需求：高性能计算网络需要达到Tb级别的单服务器访问能力，例如主流的单个网络接口卡（NIC）2x100G，总计每台服务器1.6T或3.2T。整个网络应容纳数万台GPU。对低延迟的需求：一个稳定高效的计算网络必须将数据中心网络的延迟从毫秒级降低到恒定的10微秒。

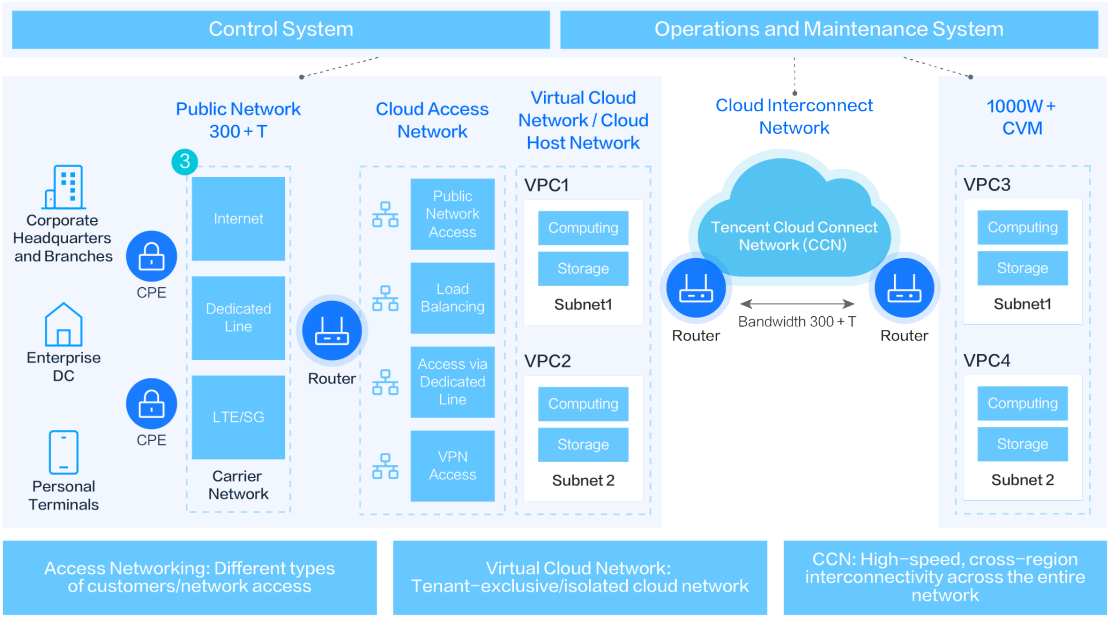
对广泛覆盖的需求：随着LLM推理和生成能力的快速增长，上层应用场景也将迅速扩展。因此，应用分布网络需要延伸到边缘，覆盖越来越多的用户。

需求对最小抖动：在高速度数据传输网络连接的集群中，即使是0.1%的数据包丢失也可能导致50%的计算能力损失。因此，如主动拥塞控制和TCP流级确定性链路负载平衡等功能对于构建端到端无损耗网络是必需的。

产品特性

腾讯云助力繁荣的AI时代，提供行业领先的网络基础设施、网络架构、公共网络接入、云互连以及高性能计算网络解决方案和能力。

图9：腾讯云的产品功能



Tencent Cloud Connect Network Product Architecture Capabilities

优势

大规模计算网络和高性能超级计算网络：
腾讯的3.2T RDMA网络，被称为StarPulse，具有专有的TiTa和TCCL协议。它提供单个VPC，拥有300万个网络元素节点和单节点100G物理网络接入。该配置支持超高性能计算任务，如广告和视频服务、人工智能大型语言模型（AI LLMs）以及超过10万张卡的GPU集群。它实现了卓越的集群性能和智能操作，网络延迟在10μs到40μs之间，接近零包丢失，负载率超过90%。
超宽带宽，实现跨区域高速、稳定的连接： 腾讯

云连接网络（CCN）提供高达300T的带宽，支持超过20个全球区域。它提供各种网络类型，如专用线路、VPN和VPC，为超过1000万云虚拟机（CVM）实现高速、安全、稳定的互联。

大规模用户接入和灵活扩展：该大规模、低延迟接入网络支持灵活扩展，覆盖全球用户。它为超过13亿用户提供接入，为超过100万企业客户提供高可用性、高质量和高并发接入。在亚洲质量排名首位，拥有65毫秒的云接入时间。

成功案例

图10：腾讯云的成功案例

部署选项：公有云、专用云和本地部署。

腾讯云 EdgeOne

用户痛点：

用户面临高延迟、弱网络、断连和拥塞等问题，这些问题影响了性能。网络攻击，如DDoS勒索、非法数据抓取和机器人攻击普遍存在。传统解决方案往往无法同时保障“加速”和“安全”，而安全产品价格昂贵，通常在攻击发生后才被被动购买。传统计费模式（基本+可选）可能导致意外成本。复杂或个性化的场景需要灵活的服务支持，随着企业的发展，新的需求需要高度可扩展的技术架构。个人开发者在使用高性能边缘计算能力时面临高昂成本。配置和管理这些资源需要深厚的专业知识，从而形成重大的财务和技术障碍。

产品特性： 腾讯云EdgeOne是中国首创的全功能安全加速边缘平台。它具备全球部署的统一多软件架构，提供用户就近服务。EdgeOne提供包括域名解析、动态和静态智能加速、TCP/UDP第四层加速、DDoS/CC/Web/Bot防护、边缘函数计算以及边缘媒体处理（转码）在内的集成边缘服务，助力用户提升速度和效率。

优势：

性能加速：

支持静态内容分发和动态加速。

提供对L3/4/7网络协议和业务数据缓存加速的支持。

核心节点通过专用线路相互连接，实现了端到端的低延迟。

小于50毫秒。

凭借有利的亚太节点，它成为了中国公司海外扩张和外国公司进入亚太地区的首选CDN服务提供商。

增强安全性：

提供对DDoS和CC保护的支撑，网页保护和机器人管理。利用Anycast网络架构进行近源净化。提供超过15 Tbps的保护能力。能够在平均3秒内识别并减轻大多数DDoS攻击。

一款综合性的全功能安全加速产品，确保在无序列影响业务性能的情况下实现安全和加速。该产品将安全定位为一种经济实惠、易于获取的产品，而非仅限于专家使用。该解决方案为商业构建了一个“免疫系统”，提供无形且无缝的安全保护。

灵活性与弹性：

仅对保护后的干净交通收费，避免意外账单。

引入了在架构层面的服务链概念，使得在相同的边缘节点上集成各种软件服务成为可能，从而在边缘解锁无限可能。边缘函数和规则引擎等特性允许实现个性化的定制化业务操作，灵活适应不同的业务场景。

技术开放性：

开放边缘节点功能，包括边缘函数、可编程特性和边缘基础设施计算能力。

提供了一个开放的协作平台OpenEdge，为开发者支持下一代无服务器边缘AI应用的开发，并在边缘解锁无限可能。

成功案例： 腾讯云EdgeOne为中国前十大收入生成游戏公司的超过70%，海外前十大社交音频和短视频应用超过80%，以及主要国有银行的33%提供服务。

图 11：腾讯云的成功案例

部署选项：公有云

存储加速

在传统的人工智能时代，用户的存储基础设施常常无法满足生成性人工智能的性能需求。这项技术需要巨大的数据存储以及快速、安全的数据处理和检索以完成复杂任务。

在训练过程中，必须快速将大量数据集加载到GPU中，以避免计算资源的闲置，这会延长训练时间并浪费计算能力。在推理场景中，高读写速度对于模型快速访问和存储数据至关重要。此外，安全性问题，尤其是内容安全性，至关重要，因为缺乏审查机制的快速交互可能对拥有大量用户的超级应用造成重大风险。

腾讯云对象存储（COS）

用户痛点：

性能：在预处理过程中，生成式人工智能需要从存储集群中快速检索数据，这要求高元数据操作性能、低文件访问延迟和高存储集群吞吐量。多协议支持：存储系统必须支持各种协议，以处理数据收集、预处理、训练、推理和应用过程中的混合工作负载。

阶段化，通过打破不同数据基础之间的壁垒，实现无缝的数据管理和流动。弹性存储：对于原始和加工后的数据，需要大量的存储空间，存储系统需要具备弹性可伸缩性，以适应模型和数据集的增长。数据安全：生成式AI数据的快速增长因快速交互而带来高度的内容安全风险，追求数据来源的成本高昂。数据管理：管理非结构化数据的高成本、低元数据利用率以及提高生成式AI应用性能的有效数据管理方法的需求。

产品特性

腾讯云对象存储（COS）作为集中的数据存储仓库，提供快速公网访问、传输能力以及庞大的存储容量。利用GooseFS缓存加速技术，COS实现亚毫秒延迟、百万级IOPS和Tbps吞吐量，从而优化存储性能。云无限（CI）提升了生成式AI在数据预处理、模型训练和安全治理等流程中的效率。

数据本地化：GooseFS可以将数据调度到本地GPU节点磁盘，缩短文件I/O路径，增强数据局部性，支持亚毫秒级延迟、百万IOPS和Tbps吞吐量。统一存储语义：支持COS、Hadoop、S3和FUSE等多种存储语义，使其适用于各种计算生态系统和应用场景。统一命名空间：通过统一的文件系统命名空间管理不同的远程存储服务，如COS、TStor和Cloud HDFS，提供丰富的智能数据流策略。集成内容安全：存储回调自动触发审查，提供快速审计响应和毫秒级结果输出，并为AIGC提供自定义审计策略模型。智能检索元洞察：根据数据内容创建特征索引，为数据清洗提供多模式数据检索和分析能力，以及数据存储期间模型训练和分类管理。统一腾讯云生态系统服务：与COS操作集成，包括日志记录、身份验证和监控。

CFS Turbo是一种高性能的文件存储解决方案，提供高达TiB/s的吞吐量和数千万IOPS。它支持在检查点读写、样本数据检索以及训练和推理过程中的模型分发等高并发场景下的生成式AI，最大化GPU利用率并加速AIGC业务增长。

高性能：采用全闪存阵列、RDMA和专有客户端，实现从客户端到服务器的无缝端到端并行处理，极大地提升了整体性能。数据分层：利用人工智能将数据根据访问频率分为热层和冷层。

频率、文件大小和路径配置，有效降低大规模数据环境中的存储成本。

混合云：采用分布式云解决方案在用户数据中心部署服务，同时将控制平面与公共云集成，将公共云的能力扩展到本地环境。

腾讯云统一服务：集成了各种云服务，包括日志记录、身份验证、警报以及PaaS平台（例如，机器学习平台）和容器服务，以满足多样化的用户需求。

优势

集群吞吐量：通过计算亲和性任务调度增强数据局部性并提高整体集群带宽。丰富的数据缓存策略优化数据流管理，提升访问加速。在PB级全闪存集群中提供高达2Tbps的集群吞吐量。

访问延迟：通过客户端缓存短路读取、数据预取和并行I/O功能减少GPU文件访问延迟，通过RDMA优化网络开销，实现亚毫秒级文件访问延迟。

元数据规模：支持通过高并发KV数据库实现元数据节点的并行扩展和分层管理。通过数据库和表分区技术增强单节点元数据规模。通过线程模型优化和KV数据结构增强优化元数据访问性能和效率。实现集群元数据规模高达数十亿，元数据OPS高达数百万。

全面数据加工：通过低延迟、高准确度的云存储原生内容审查，解决生成性数据内容安全问题。通过丰富数据标签快速识别和分类生成性数据。使用先进AV1数据压缩技术显著降低数据存储成本。

成功故事：

图12：腾讯云的成功案例

百川智能，迷你最大，元石，右脑智能

部署选项：

公共云与私有部署

云对象存储（COS）：用户可以通过控制台、API、SDK和工具等多种方式轻松与公有云集成，确保快速访问。TStor：一种即插即用的私有云部署解决方案，与公有云存储无缝集成。它支持数据互操作性，使应用程序能够按需、随时、随地访问相关数据。数据加速器Goose FileSystem（GooseFS）：用户可以部署和管理GooseFS

通过腾讯云控制台、API和TCCLI在CVM、TKE、GPU和EMR等不同计算服务中。云无限（CI）：用户可以通过腾讯云控制台、API和SDK激活和配置CI，以集成必要的数据处理能力。云文件存储（CFS）Turbo：用户可以通过控制台、API或其他方法创建存储实例，通过云服务器、容器和PaaS平台实现并行文件存储的快速挂载或利用。

模型库

基础模型

除了 Hunyuan，腾讯云已完全集成并对超过20个主流模型进行了兼容，包括 Llama 2、Falcon、Dolly、Vicuna、Bloom 和 Alpaca。这些主流模型支持直接部署和调用，特色包括简化应用程序流程和低代码操作。

行业特定且专属的大规模模型

同时，腾讯云希望“大型模型”能够对行业有更深入的理解，并更易于实施。他们旨在针对用户的痛点和需求量身定制解决方案，帮助用户创建独特的行业专属大型模型。

基于此，腾讯云还创建了一个针对行业大型模型的精选商店，其中包括涵盖金融、媒体、文化及旅游、政务、教育等各个场景的多个高质量行业特定大型模型。它支持多种模型训练任务，并允许用户按需使用。

这样，用户不仅可以快速融合独特场景和数据来微调和生成专属模型；他们还可以根据其业务场景的需求，定制具有不同参数和规格的模型服务。有关模型库的集成和支持功能，请参阅由腾讯云发布的另一份报告中提供的全面描述。

图13：相关报告的二维码

工程工具层

AI原生云需要云基础设施与各种AI平台和工具紧密集成。在模型部署和基于模型库的微调，以及基于此开发智能应用方面，必须提及腾讯云工程工具层的强大和丰富功能。

部署和微调加速

腾讯云TI平台

腾讯云TI平台是基于腾讯先进的人工智能能力和多年技术专长构建的全栈式AI开发服务平台。该平台专为开发者、政府和企业用户设计，简化了行业特定AI部署的整个流程链，包括数据采集、处理、算法开发、模型训练、评估、部署以及AI应用开发。该平台帮助用户快速创建和部署AI应用，管理端到端AI解决方案，并加速数字化转型，同时促进AI产业生态系统的发展。

部署选项：腾讯云TI平台产品支持公有云访问、本地部署和专用云部署。

关于TI平台核心能力的详细信息，请参考腾讯云发布的另一份报告，题目为《生成式AI行业发展路径研究》。

图14：相关报告的二维码

内容质量管理

内容质量管理涉及用户评估模型输出的准确性和合规性。在部署生成式人工智能应用时，用户必须解决模型幻想问题以确保输出准确性。此外，由大型语言模型生成的内容应避免法律合规问题，例如处理敏感信息和尊重知识产权。

降低幻觉对于致力于推进生成式AI的用户至关重要。生成式AI的意外影响引起了像CEO和COO这样的非技术性高管的注意，他们通常在接受生成式AI项目成果中扮演着角色。大型语言模型（LLM）幻觉的低准确性可能会侵蚀管理层对生成式AI的信任，阻碍其对价值的认可。如果没有高级管理层的支持，用户可能会在后续的生成式AI项目中遇到困难。

大型模型可能会输出敏感信息，例如色情或暴力内容，或生成受知识产权保护的内容。这些问题比生成结果的准确性更重要。无论生成结果的准确性如何，用户都需要一个“回退策略”。

用户痛点

生成式人工智能促进了丰富的创造力，但也带来了与数据真实性、内容合规、用户隐私和身份以及伦理考量相关的史无前例的挑战。从监管的角度来看，国家已在三个关键领域确立了AIGC的合规要求：数据、内容和算法。

数据合规：利用特定大型语言模型的AIGC提供商负责确保预训练数据的合法性，并优化生成人工智能产品的训练数据来源。

内容合规：利用生成式人工智能产品提供聊天、文本、图像和语音生成服务等服务的组织和个人应承担该产品的内容生产者的责任。提供商必须按照法律规定对生成的图像、视频和其他内容进行标识，并履行信息内容责任。

管理和提升该平台信息和内容的治理。算法合规：算法推荐服务提供商必须遵守算法注册、算法评估、相对算法透明度以及建立强大的用户权利保障机制等要求。

产品特性

满足合规要求

为了提高知识获取的准确性，我们考虑上下文因素，对当前查询进行精炼，并利用向量数据库来识别相关信息。将大型语言模型（LLMs）与搜索引擎集成，使我们的解决方案能够高效地从互联网上的官方网站和行业来源检索大量数据，生成类似人类的答案。这种方法增强了答案的多样性和相关性，有效地减少了答案的幻觉。

此外，为了解决与LLMs相关的幻觉问题，我们在腾讯云LLM知识引擎平台层提供了一套先进的调优工具链。此工具链帮助用户预先生成和验证问题和答案，促进对话测试、答案优化和部署激活等过程。在整个问答提取过程中，平台支持包括文本、图形和表格在内的多种输出格式。验证功能包括突出显示和跟踪原始文本片段，显著提高验证效率。为了增强运营，我们提供了强大的反馈机制和片段修正工具，有效地降低对话中的幻觉。

图15：腾讯云的产品特色

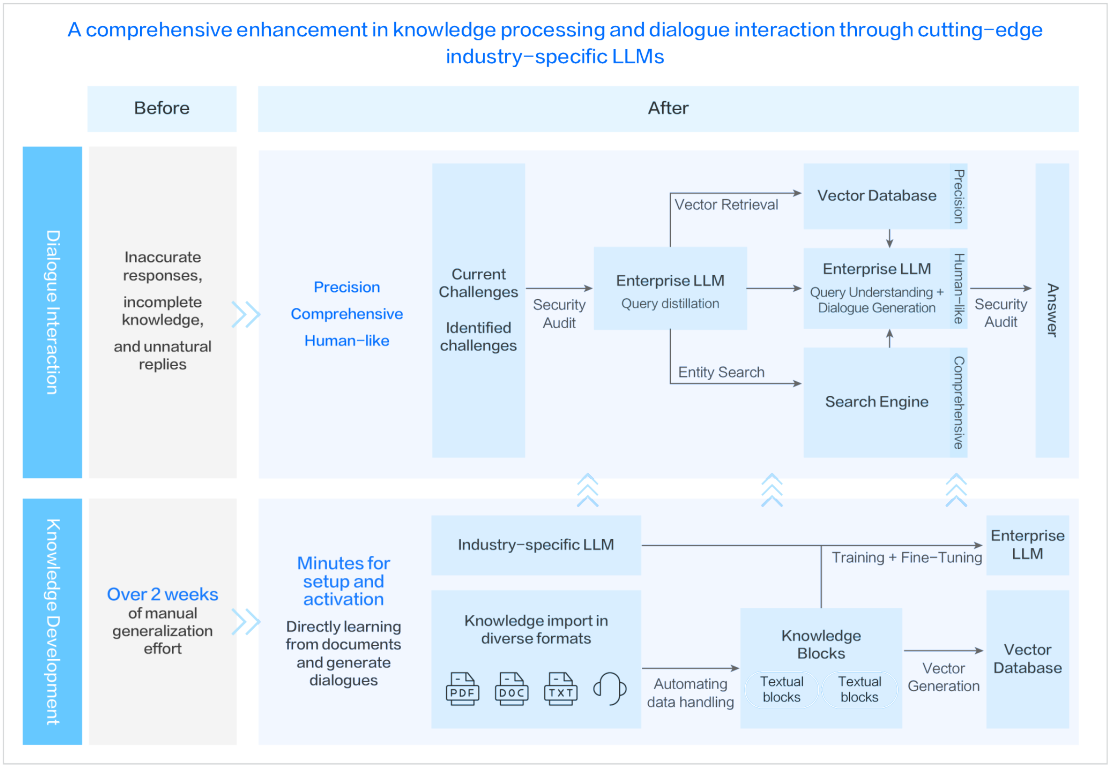


图16：腾讯云的产品特性

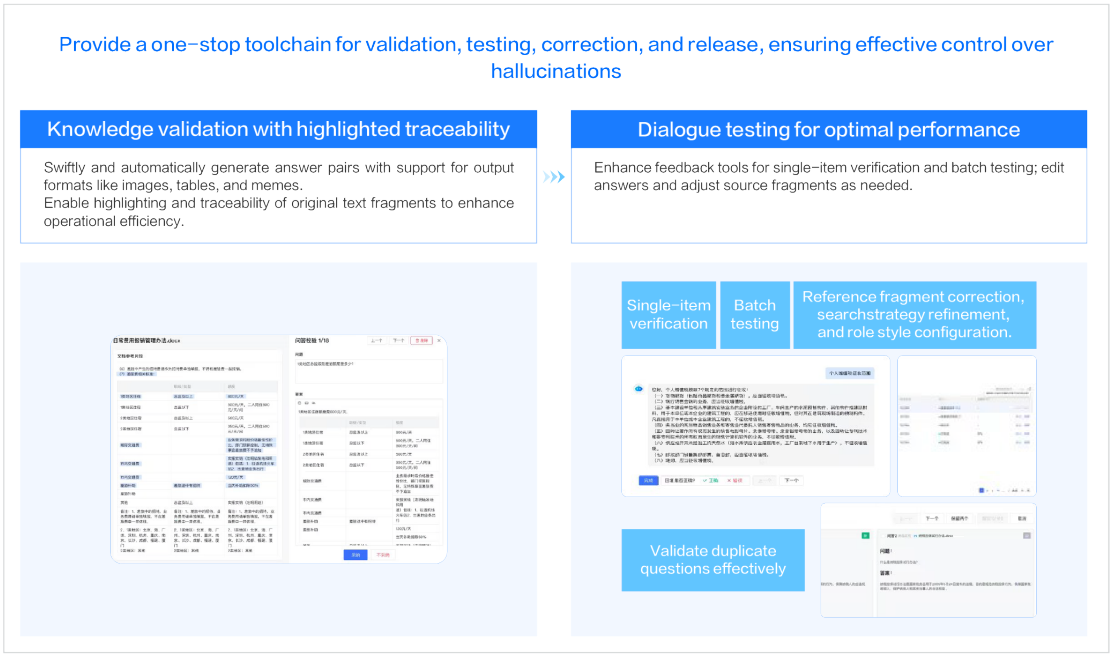
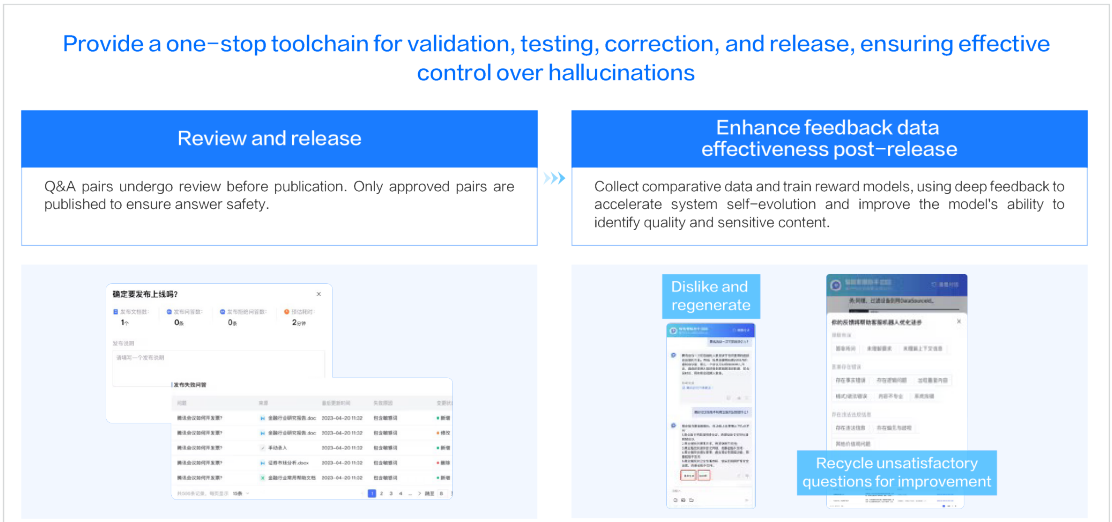


图 17：腾讯云的产品特性



满足合规要求

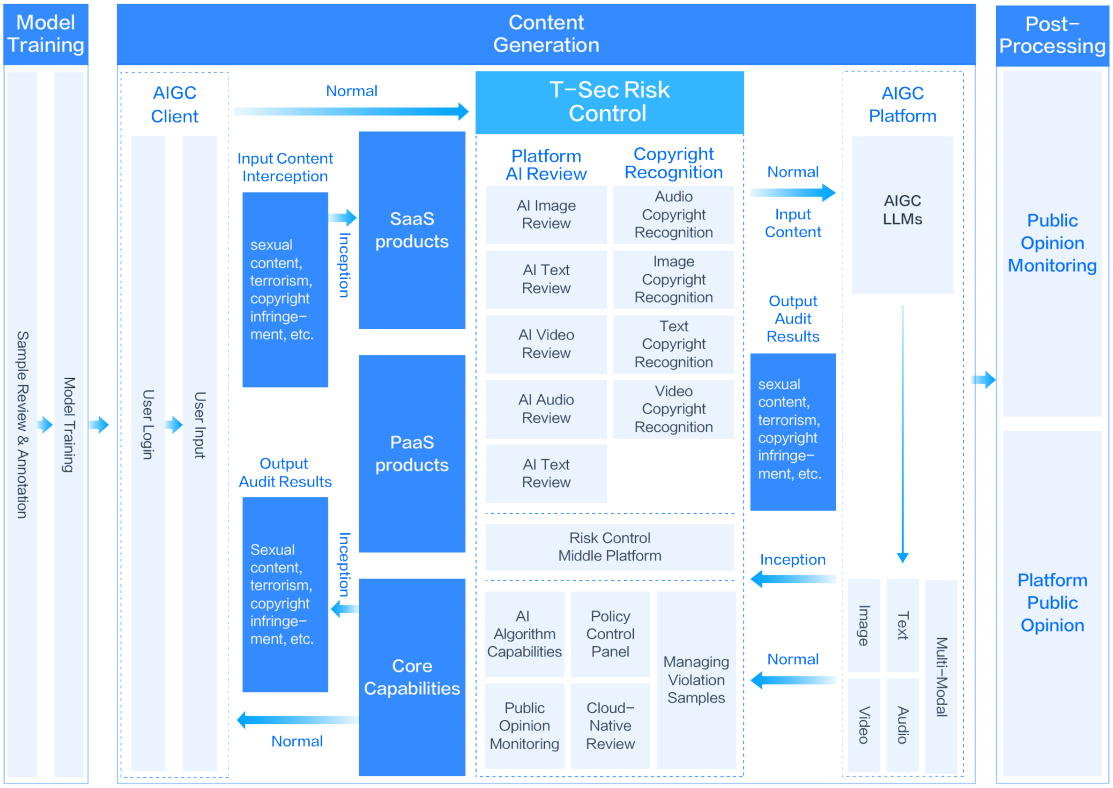
AIGC平台生成大规模数据。腾讯T-Sec为用户提供涵盖六个维度的全面内容安全解决方案：界面输入、内容预处理、模型识别、政策协助、平台调度分析和手动标注操作。通过整合机器审核、安全专家、编辑审查和版权服务，AIGC用户可以一次完成整个审查流程，在无缝操作中有效解决AIGC不同阶段的内核痛点。

腾讯T-Sec风险控制为用户提供跨六个维度的全面内容安全解决方案：界面输入、内容预处理、模型识别、策略辅助、

平台调度分析，以及手动标注操作。与腾讯云的对象存储、实时直播、点播视频、实时音视频等功能无缝集成，它允许AIGC用户通过单一集成在云中进行全面的内容安全审计，显著提高效率。

丰富经验：超过20年的运营经验，加上处理数十亿案例的违规行为经验。高精度：机器准确率达到99.99%。一键集成：与腾讯云五大云组件的无缝交互。定制解决方案：提供十多种定制识别服务。敏捷响应：针对服务的小时级响应。

图18：腾讯T-Sec风险控制解决方案



数据处理效率提升

在从模型到应用的旅程中，软件工程工具扮演着至关重要的角色，其中数据分析工具发挥着关键作用。高效的数据管理涉及简化数据处理工作流程、提升数据质量，并部署先进技术以实现稳健的数据分析。这种方法确保了数据的最优利用，加速了决策过程，提升了运营效率，并加强了市场竞争力。

在人工智能原生时代，两个重要的技术挑战在数据之路上显得尤为突出。

效率提升，为用户设置了巨大的障碍。

首先，数据向量化是提高数据效率的初始挑战。在AI原生时代，数据检索的新需求出现，需要通过向量化技术高效检索非结构化数据。此外，对于希望通过RAG实现生成式AI的用户，他们面临着如幻觉、知识停滞和数据安全问题，而数据向量化成为最优解决方案。

其次，数据协同驾驶员在提升数据效率的旅程中带来了终极挑战。数据协同驾驶员的引入将数据分析的效率提升到了前所未有的水平。它使数据分析民主化，使用户能够直接与数据互动，并通过智能对话获得洞察。这种自动化分析的实现不仅显著简化了数据分析过程，还消除了分析请求排队瓶颈，赋予用户数据驱动的决策能力。

腾讯云VectorDB

用户痛点

对于LLMs，训练需要大量高质量的数据集，导致在模型训练期间产生大量时间成本。VectorDB在训练阶段高效地以向量形式存储参数，便于高效地清洗和检索对应数据。此外，它支持大规模并行计算，加速模型训练过程。

关于RAG领域，幻视、知识停滞和数据安全障碍等问题阻碍了用户基于大型语言模型（LLMs）部署各种智能问答应用。腾讯云向量数据库提供了一站式RAG检索解决方案，作为LLMs的“外部知识库”，协助客户构建高质量的RAG应用。

产品特性

腾讯云VectorDB是一种完全托管的企业级分布式数据库服务，由公司内部开发。它支持高达4096维度的向量数据以及各种索引类型和相似度计算方法。单个索引可支持数十亿个向量规模，能够处理数百万的QPS，并提供毫秒级的查询延迟。

除了存储和检索向量数据之外，腾讯云VectorDB提供嵌入功能以及一站式RAG检索解决方案，能够快速构建高质量的外部知识库供LLM使用。此外，它还可广泛应用于推荐系统、自然语言处理以及其他AI领域。

优势

低成本与高性能： 首先通过由中国信息通信研究院（CAICT）进行的标准化测试和性能可伸缩性测试。单个实例可以存储数十亿个向量数据，实现每秒五百万次查询（QPS），并提供毫秒级的响应延迟，性能超过行业平均水平1.5倍以上，同时单次QPS成本降低75%。

低门槛： 一站式RAG检索解决方案。

函数如文档拆分、嵌入和检索排名已集成到VectorDB中，从而简化了RAG中的数据处理工作流程。这种集成增强了数据检索的召回率，并将数据访问效率提高了十倍，与传统解决方案相比。

使用便捷 提供了Python、Java和Golang等语言的支持SDK，以及LangChain和Llamaindex框架的支持，根据实际需求快速集成腾讯云VectorDB，并提供了相应的工具，便于开发高价值生成式AI应用。

成功案例

图19：腾讯云的成功案例

博世、百川智能、XVERSE、新东方、猿辅导、粉笔、猿辅导

部署选项：公有云

腾讯DB for DBbrain

腾讯云DBbrain（DBbrain）是由腾讯云提供的自助管理云服务，为用户提供数据库性能优化、安全和管理等功能。DBbrain运用大数据技术和专家经验引擎，快速复制资深数据库管理员的经验，自动化大量传统手动数据库运维任务。它适用于云和非云用户，有效保障数据库服务的安全性、稳定性和高效运行。

用户痛点：

数据库根本原因分析和优化解决方案：当数据库运行不优化或异常时，就像人生病一样，表现出各种症状。通过这些表现来识别根本原因需要经验丰富的DBA进行仔细调查。用户迫切需要一种能够自动诊断数据库异常根本原因并提供优化建议或解决方案的服务，减少对人员专业知识的依赖。

实时数据库诊断与优化建议：对数据库进行手动检查或使用传统工具分析可能无法及时检测问题或提供快速解决方案，因此缺乏时效性，并可能错失对问题的早期响应，这些问题可能会升级到危急状态。用户需要一个能够全天候持续诊断数据库、自动通知管理员已诊断的问题，并快速提供解决方案的服务。

对数据库链的全面洞察：用户需要全面了解数据库中每个SQL语句、会话和事务的详细操作信息和统计数据。他们需要知道哪些业务SQL语句正在数据库的每个工作节点上运行。他们应能够从数据库的角度快速观察业务健康状态并追溯到业务请求的源头。

数据库健康检查报告：用户不仅需要评分大量数据库的健康状态以快速识别需要关注的实例，而且还需要对每个实例进行全面的详细健康检查。这些报告需要归档以进行健康跟踪和问题分析。

产品特性

实时根本原因分析和优化建议：利用结合大数据、规则引擎、机器学习和LLM技术的综合解决方案。实现数据库问题的24/7连续检测、发现和诊断，迅速识别数据库风险为用户，并提供建议的优化解决方案和改进建议。

端到端分析与洞察：建立每个SQL语句与业务的关联，监控每个SQL语句在数据库各个节点的运行细节，并实现任何时间、任何节点、任何业务的SQL、会话和事务的回溯分析。同时，它还可以提供高级统计分析，以便快速定位核心问题。

数据库检查：利用实时诊断结果对数据库进行评分，通过仪表盘展示风险排名，并为数据库实例提供详细的每日健康检查，生成健康检查报告。

优势：

通过结合规则引擎、机器学习和LLM技术的综合解决方案，有效地诊断数据库问题并提供优化建议。这种方法满足客户对实时性的期望，并克服了单一模型解决方案的限制。端到端分析将业务流程与SQL和数据库节点相连接，为用户提供行业领先的洞察力。这一全面的信息还作为学习资源，持续增强实时诊断的丰富性和深度。详细的检查报告为高级用户提供对数据库操作的全面视图，帮助他们识别和解决问题，而不会忽略任何问题。

部署选项

公共云，私有云

腾讯云 Elasticsearch 服务 (ES)

用户痛点：

随着人工智能的兴起，尽管语言模型在回答一般性问题时表现出色，但当应用于企业服务时，它们面临着重大挑战。

知识截止点：这些模型是在至特定日期的数据上训练的，这限制了它们对近期发展和更新的了解。

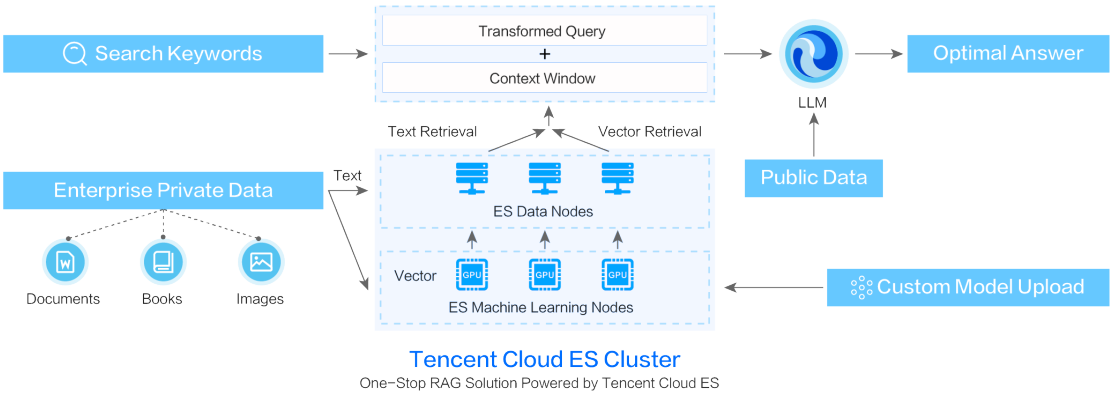
缺乏特定领域知识：他们的知识来源于公开的互联网数据，往往缺少企业所需的专门见解。

为应对这些挑战，一种新的技术方法应运而生：检索增强生成（RAG）。RAG通过向量和信息检索技术将大型语言模型与特定业务知识库无缝集成，以提高答案的准确性。实施RAG复杂，需要专业技术、对业务环境深入理解、大规模数据处理和算法优化的结合，以达到最佳效果。

产品特性

腾讯云ES提供了强大的基于云的AI增强功能，特点是一个功能全面且包含实施RAG所需所有功能的引擎。它支持在一个统一的技术堆栈中对文本和向量的混合搜索，并将自然语言处理与大型语言模型集成。这使用户能够利用由AI驱动的先进搜索功能，为搜索和分析提供新的体验。

图20：腾讯云的产品特性



优势：

低入门门槛：一体化RAG解决方案

腾讯云ES为模型上传、矢量生成、存储、检索以及与LLMs的集成提供一站式解决方案。这种全面的方法超越了传统的单点解决方案，满足了用户在构建RAG应用时对全面需求。

高性能：处理数百万QPS并扩展至数百亿个向量

它支持高达4096个维度，处理数十亿个向量的操作，平均向量检索响应延迟控制在毫秒以内，并允许与腾讯的专有软件和硬件技术集成。在类似规格的对比下，腾讯云ES通过优化的独立查询链路和改进的分布式锁，展现了显著的性能提升和成本优势。

更高精度：矢量和文本的独特混合搜索能力

ES的混合搜索利用向量检索的灵活性，提供多样化的结果，满足不同用户的查询需求和偏好。它集成了关键词搜索逻辑、排序、筛选和其他功能，以有效应对复杂的查询需求，从而提高搜索结果的准确性和可解释性。

增强智能：完美整合 使用大型语言模型实现轻松的AI问答应用

ES与LangChain集成，以帮助构建复杂的数据管道和生成式AI应用。它通过API无缝集成LLMs，简化了AI驱动智能问答应用的构建。

专用机器学习节点

腾讯云Elasticsearch服务支持向量化模型的上传、管理和部署，高效完成向量生成并有效提升向量推理能力，同时与数据节点隔离以确保在线搜索服务的稳定性。

成功故事：

自2018年正式发布以来，腾讯云ES不断优化围绕最关键的用户痛点：成本、性能和稳定性，赢得了众多客户的信任。腾讯云ES已完全管理云端的Elasticsearch技术栈，持续加强其开发能力，并将ES与更多云原生特性相结合，如独特的数据路径、自主索引和存储计算分离技术。它也是国内首家推出商业级ES Serverless服务的提供商，提供按需使用和完全无需维护的功能。随着AI浪潮的到来，ES迅速实现了AI集成解决方案，以满足新用户的发展需求。

图21：腾讯云的成功案例

部署选项：公有云

腾讯云商业智能

用户痛点：

降低数据分析的门槛：使非专业人士能够通过自然语言访问数据。I 提高效率：允许即兴分析，无需正式报告；用户通过提问获取数据。I 增加便利性：随时随地启用对话式分析，甚至在出差或数据查看不切实际的情况下。

产品特性：

腾讯云商业智能提供了一套全面的企业智能（BI）功能，包括数据源集成、数据建模以及数据可视化与分析。它使业务操作员能够迅速获取可操作的见解，以支持明智的决策制定。该系统采用敏捷的、自助服务的方法设计，通过直观的拖放功能，使用户能够轻松地开发复杂的报告。此外，它还支持报告共享、自动分发以及其他企业协作场景。

优势：

腾讯云商业智能利用一个专门的100亿级别语言大模型进行数据分析，该模型能准确解释自然语言查询的语义。它支持上下文连贯的多轮对话和智能后续问题，以澄清用户意图，使新手用户也能通过简单提问进行数据分析。腾讯云商业智能支持多个平台，包括小程序、H5和PC，使用户能够随时随地开展数据分析。该产品具有强大的业务学习功能，使用户能够创建和整合针对其特定需求的行业知识库。

成功案例

猫眼娱乐

腾讯云商业智能为猫眼提供了数据治理、仓储和BI分析的综合解决方案。借助腾讯云商业智能，猫眼迅速建立了覆盖电影/票房销售、流量监控、运营评估和自动化数据检索等关键场景的企业分析平台。

王府井全球购物

作为王府井集团旗下的旗舰跨境电子商务项目，王府井全球购物迅速开发了一个基于腾讯云商业智能的云数据分析和业务报告展示平台。该平台支持跨

核心业务领域，如库存、销售分析和分销。

高（中国）

高丝（中国）面临着来自各个系统数据碎片化的挑战，缺乏统一的数据管理平台。腾讯云商业智能使高丝能够从不同的系统中整合数据，消除数据孤岛，并快速构建一个统一用户数据平台。

XVERSE

腾讯云商业智能满足XVERSE每日用户行为数据分析需求，支持在用户行为、开发、测试和反馈等多维度的快速数据可视化。该方案满足他们的业务分析需求并支持明智的决策。

部署选项：公有云

发展提升

生成式AI凭借其革命性的能力正在推动软件开发边界的扩展。这一转变不仅引入了软件创建的新范式，而且解锁了提升开发效率和质量的机遇。

在其核心，生成式AI通过产生针对项目需求定制的多样化软件元素（如代码、图像和文本）在创造力方面表现出色。这种能力通过在每个步骤中提高效率，简化了软件开发的所有阶段——从编码、测试到代码审查。

此外，生成式人工智能在提升开发标准方面展现出巨大的潜力。通过深入理解编程语言的上下文，它能够自主生成高质量的代码，有效减少开发生命周期中的缺陷和错误。AI生成的代码不仅符合行业标准最佳实践，还通过其清晰、精炼的风格和便于开发的特性，提升了整体产品质量。

然而，这种人工智能驱动的开发方法给传统的集成开发环境（IDE）带来了挑战。这些IDE最初设计时没有充分集成人工智能，导致在计算资源配置、工具集成和运营管理方面存在限制。这些缺陷不仅会影响开发者的用户体验，还可能造成计算能力利用的效率低下。

此外，为用户选择合适的AI代码助手至关重要。虽然AI助手

在代码生成方面表现出色，其有效性取决于对用户特定编码风格和标准的理解。如果不能掌握这些细微差别，可能会导致人工智能生成的代码与实际开发需求不完全一致，从而影响其可用性。

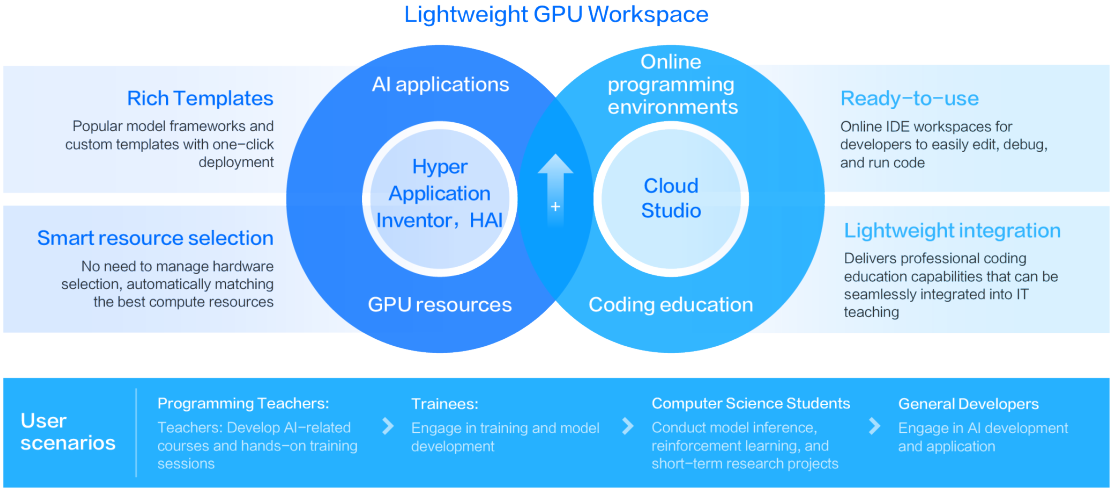
腾讯云开发者平台

用户痛点：

缺乏提供最佳用户体验的在线集成开发环境，低集成成本和最低的维护开销。缺乏一致的编码运行时环境。

产品特性与优势

图22：腾讯云的产品特色与优势



部署选项：

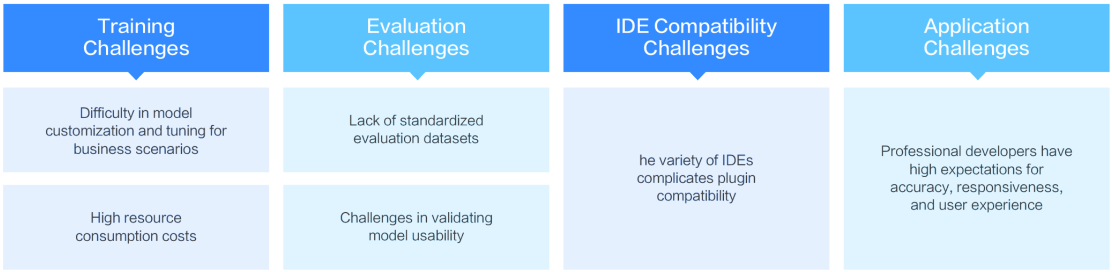
公有云

腾讯云AI代码助手

用户痛点：

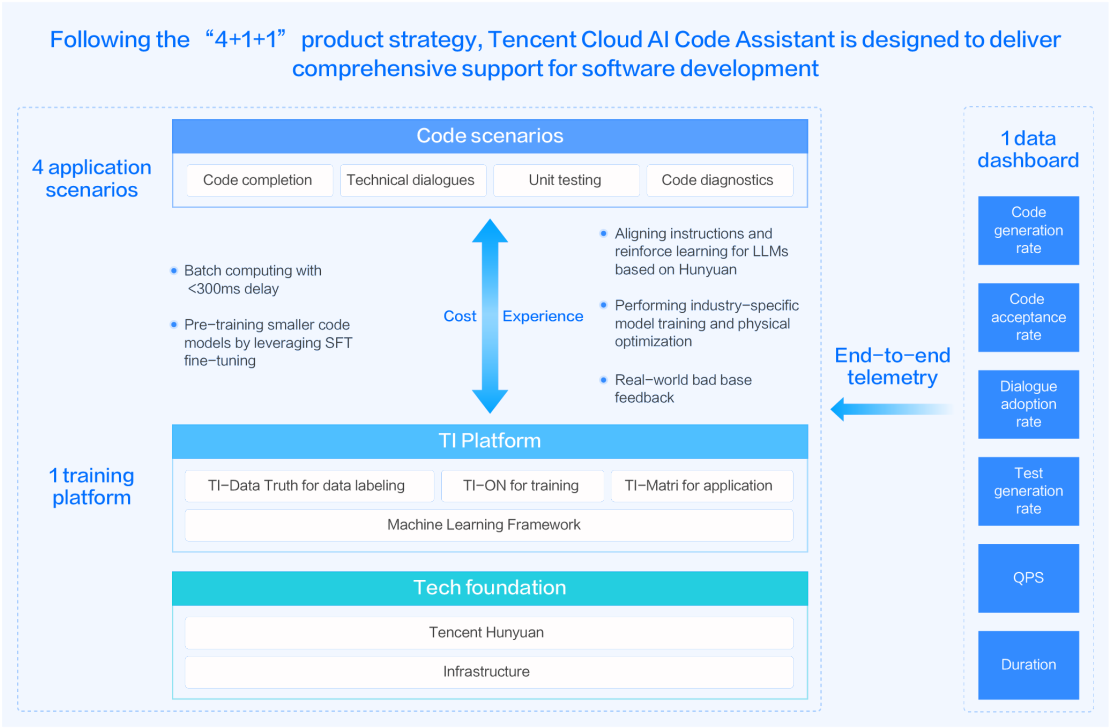
随着人工智能的快速发展，软件工程正朝着以智能驱动的范式演进。AI代码助手，作为一种常用且价值高的应用，正引领这一变革。在技术探索过程中，出现了四种不同的挑战：

图23：用户痛点



产品特性

图24：腾讯云的产品特性



优势

1. 多模式能力：

交付针对企业客户定制的预训练、成本低效的专业模型，支持主流GPU。

通过利用专有企业数据集在本地环境进行微调和迭代，腾讯云AI代码助手在零延迟条件下实现了相比同类开源模型3到5倍的处理性能。

2. 绩效评估：

与企业客户协作，通过高级微调优化模型。利用公开代码资源和企业特定样本，精确调整模型以适应特定行业场景。

在部署模型之前，根据客户的业务场景定制已验证且有效的评估标准以进行细微调整。确保评估结果符合公司的编码标准和协议，保证代码精确且详尽。

发电量

3. IDE 兼容性：

整合腾讯云AI代码助手，作为插件在企业客户端内部系统中后部署，确保不同集成开发环境之间的兼容性。

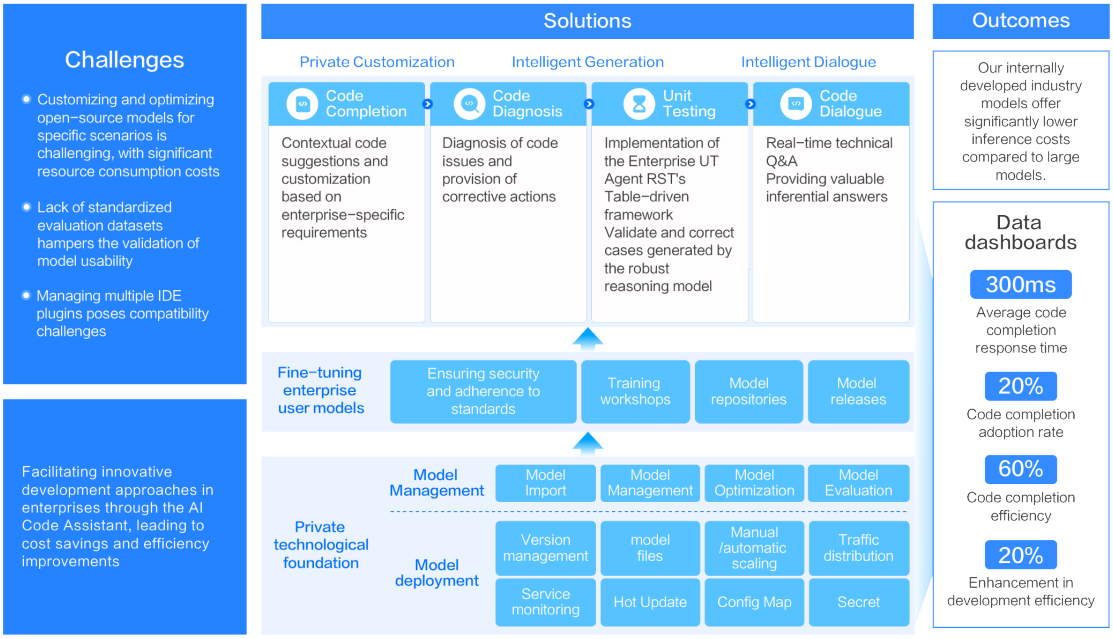
4. 发展提升：

提供全面解决方案和简化的应用程序体验，尤其针对企业开发者在代码补全场景中的需求。

实现了上下文感知的自动代码补全功能。提供包括行内和侧边栏对话在内的智能辅助功能。支持跨多个文件的复杂代码补全，并整合了提示扩展接口以增强用户交互。

成功案例：一家领先的金融机构。

图25：腾讯云的成功故事之一



部署选项：公有云，本地部署

腾讯云媒体处理服务协作助手

媒体处理能力丰富，包括转码、音频和视频增强、智能审查、智能分析、质量检验、截图捕获、录制、评估等功能，以及数百个配置参数。客户需要承担一定的学习成本，才能理解和熟练使用这些功能。

在使用增值功能之前，客户需要了解产品功能、效果、示例、计费方法和集成过程，以便做出明智的决策。此前，这需要在整个集成过程中与各种角色如客户、产品经理和研发人员等进行沟通，以收集所有必要的信息，这在人力和时间成本上都是昂贵的。

产品特性：

腾讯云媒体处理服务Copilot通过人工智能将广泛的媒体处理功能简化为会话式界面，在媒体处理场景中创建了一个真正的AI助手，该助手可以替代实际用户操作。它能够理解用户意图、调用接口、生成媒体处理MPS模板，降低用户理解和使用成本，并增强媒体处理MPS的多种原子能力和模板参数调整的可用性。此外，它还具有实时处理音频和视频文件的能力。

优势

敏捷性与柔性

腾讯云媒体处理服务Copilot可以降低使用媒体处理服务（MPS）的门槛，通过自然语言交互提供编辑MPS各种功能的 capability。支持模板参数的生成和修改。使任务安排和政策设置得以调整。促进任务创建和启动。

高可用性

允许用户在对话框中直接发送指令并接收反馈。即时显示处理结果并提供用于对比的工具。

广泛资源支持

Copilot可以回答与音频和视频相关的专业问题。

此产品可协助腾讯云媒体处理技术的集成。

它包括全面的媒体处理计费模型和定价信息。

服务。

部署选项：公有云、专用云、混合云、SDK等。

应用层

腾讯云提供一系列智能应用，如腾讯会议、腾讯文档、企业微信（微信办公）、腾讯乐相和腾讯起点，同时通过人工智能数字人、知识引擎等高级解决方案丰富软件功能。

智能客服。除此之外，腾讯云还提供了从这些智能应用中衍生出的多样化的技术产品和功能，包括生成式AI助手和引擎，旨在显著提升用户应用智能。

为了深入了解我们的应用层能力，请参阅腾讯云发布的另一份报告，标题为《生成式AI产业实施路径研究报告》。

图26：相关报告的二维码

全栈式安全解决方案

腾讯云架构由腾讯云安全加固，提供全面的安全措施以维护AI信任。

为了对我们的核心安全能力进行彻底探索，请参阅腾讯云发布的另一份报告，标题为“生成式人工智能产业发展路径研究报告”。

图27：相关报告的二维码

结论

随着我们步入人工智能原生的时代，我们见证了商业领域的深刻变革。

并且推动社会进步。正如之前所强调的，人工智能技术的应用范围持续扩展，大型语言模型（LLMs）作为新的基石，引领着前所未有的应用。云作为人工智能技术的关键基础设施，在提升能力如LLM训练、高效的集群调度和全面的数据检索等方面发挥着至关重要的作用，涉及众多利益相关者。

腾讯云将自己定位为AI原生时代的领导者，通过以生成式AI为中心的全面升级。它不仅为用户提供了一个强大的平台，以快速采用通用人工智能（AGI），而且加速了他们走向商业成功的道路。这一进步不仅标志着技术进步，也标志着对未来商业潜力的深刻探索和实现。

展望未来，大规模模型的应用范围将持续扩大，从边缘场景向核心应用发展，影响力不断增强。然而，企业在生产环境中部署大规模模型时面临着一系列挑战。其中，“对模型参数的盲目崇拜”和“对数据准备的误解”是两家企业需要防范的最关键的挑战。

首先，企业必须揭开大规模模型参数规模的神秘面纱，避免盲目追求规模。通过精确的模型评估，企业可以选择最符合自身需求的模型，以防止因规模过大而导致的成本上升和性能下降。此外，数据准备策略需要重新评估。传统的数据准备流程通常排除异常值，但在大规模模型训练中，这些异常值和意外数据可能成为提升模型泛化能力的关键。

《AI原生云构建与加速核心能力指南》不仅是一本技术指导书，也提供了一个可执行的蓝图。它引导企业在AI原生时代寻求价值、成本和风险之间的最佳平衡。

让我们携手前进，在人工智能原生时代探索无限可能——这不仅是对技术创新的证明，也是对人类独创性和创造力的肯定。未来已经到来；让我们拥抱人工智能原生时代的曙光，开启这一新篇章。

关于腾讯云

腾讯云，腾讯集团旗下的领先云计算部门，

中国服务提供商，向全球提供先进的技术解决方案。其能力包括针对企业、组织、机构和个人开发者定制的云计算、人工智能和大数据服务。目前，腾讯云的基础设施遍布21个区域，在58个区域运行，拥有超过一百万台全球服务器和3200多个全球加速节点，此外还有200T全球带宽储备和400多个权威认证。

凭借其卓越的技术实力，腾讯云开发了一系列针对特定行业的综合解决方案，培育了一个开放、协作的云生态系统，并赋能各行各业在数字化转型之旅中取得进展。

参考材料

1. Gartner, Inc.,《首席技术官指南：生成式人工智能技术图谱》，G00793970