Volume 39 | Issue 9 Article 16

9-20-2024

Development trends of large models and Tencent's independent innovation practice

Jason Si
Tencent Research Institute, Shenzhen 518054, China

Development trends of large models and Tencent’s independent innovation practice

Abstract

The article discusses the emerging trends and application prospects of current large models, using Tencent’s Hunyuan large model as an example. It focuses mainly on innovations and implementations of large models in China. Companies like Google, Meta, and OpenAI have launched powerful models such as Google’s Gemini and Meta’s Llama 3, which have made significant progress in multi-modal applications and reasoning capabilities. China’s large models have significantly improved performance and efficiency by adopting the MoE (Mixture of Experts) architecture. Specifically, with its self-developed MoE trillion-parameter large model and deep learning framework, Tencent has made breakthrough advancements in large model technology and achieved exceptional performance in muti-modal applications. Moreover, Tencent has launched a one-stop AI agent creation and distribution platform. Tencent understands that industry-wide, large models are key to implementing AI and strategies, and is actively supporting the application of large models across sectors such as retail, education, finance, healthcare, media, transportation, and government, helping these industries enhance quality and efficiency.

Keywords

multimodal; Hunyuan large model; deep learning framework

引用格式：司晓．大模型发展趋势及腾讯公司自主创新实践．中国科学院院刊，2024, 39(9): 1631-1638, doi: 10.16418/j.issn.1000-3045.20240814004.
SI J. Development trends of large models and Tencent's independent innovation practice. Bulletin of Chinese Academy of Sciences, 2024, 39(9): 1631-1638, doi: 10.16418/j.issn.1000-3045.20240814004. (in Chinese)

大模型发展趋势及腾讯公司自主创新实践

司 晓

腾讯研究院　深圳　518054

摘要　文章探讨了大模型发展趋势和应用前景，以腾讯混元大模型为例，重点剖析了国内大模型的自主创新和落地实践。当前，国际上谷歌、Meta、OpenAI等公司纷纷推出更强大的模型，如美国谷歌公司的Gemini和美国Meta公司的Llama 3，这些模型在多模态应用和推理能力上取得了显著进展。国内大模型通过采用MoE（混合专家模型）架构，显著提升了模型性能和效率。腾讯公司在大模型技术上取得了突破性进展，其自研的MoE万亿参数大模型、Angel机器学习平台等，推动腾讯混元在多模态应用中表现优异，并推出了一站式AI智能体创作与分发平台腾讯元器。行业大模型成为人工智能+战略落地的关键，腾讯公司积极助力大模型在零售、教育、金融、医疗、传媒、交通、政务等领域的应用，推动各行业提质增效。

关键词　多模态，腾讯混元大模型，深度学习框架

DOI　10.16418/j.issn.1000-3045.20240814004

CSTR　32128.14.CASbulletin.20240814004

随着人工智能技术的飞速发展，大模型时代的到来正引领全球科技创新的新潮流。当前，全球人工智能技术呈现出快速迭代、多模态融合，以及大规模应用的趋势。各大科技公司纷纷推出新一代大模型产品，竞相争夺技术制高点。谷歌、Meta、OpenAI等公司，作为国际科技巨头不断突破技术瓶颈，推出更强大、更高效的大模型，推动人工智能在各个领域的深度应用。

在此背景下，作为中国领先的互联网企业，腾讯公司通过自主研发混元大模型，在自然语言理解和生成、文生图、视频生成等方面取得了显著突破。腾讯公司不仅在技术上实现了与国际先进水平的对标，更

修改稿收到日期：2024年8月30日；预出版日期：2024年9月5日

通过一系列创新应用，推动了人工智能在产业互联网、消费互联网和科学研究等领域的落地和普及。

本文详细探讨腾讯在大模型与新技术领域的自主创新成果，及其在各行各业的应用实例，分析腾讯如何通过技术创新和战略布局，引领人工智能新时代的发展方向，为行业内外提供有价值的参考和借鉴。

1 值得关注的大模型新趋势

1.1 国外大模型加速迭代，第一梯队竞争更为激烈

谷歌公司发布的Gemini，是最早主打原生大模型的厂商，无缝跨文本、图像、视频、音频和代码，在MMLU（大规模多任务语言理解）方面优于人类专家的模型，准确率达到90%（人类专家为89.8%）。Anthropic公司的Claude 3在一些基准测试中，超越了OpenAI公司的GPT-4和谷歌的Gemini 1.0 Ultra。Meta公司发布的Llama 3，基于超过15万亿token训练，具有更强的推理和代码能力，其更强能力的4 000亿参数模型蓄势待发。OpenAI公司2024年初发布的文生视频模型Sora，可以直接输出长达60秒的高质量视频，将多模态大模型推进了一大步。GPT-4o带来了全新的大模型交互方式，可以实时对音频、视觉和文本进行推理，新模型支持50种不同的语言，处理速度提升200%，价格下降了50%。

近日，谷歌公司发布Gemini 1.5 Pro，将上下文窗口扩展到200万，意味着可以给模型输入2小时视频、22小时音频、超过6万行代码，或者140多万单词。谷歌公司还发布了AI Agent智能体Astra，不仅可以实时视频交互，还可以搭配未来谷歌的AR原型眼镜使用，为未来基于新硬件的新交互带来了新的想象空间。苹果公司则首次发布其人工智能战略，即将推出的新Siri应用程序（APP）定位为一个陪伴助手，将带来全新的交互体验。Siri能更自然地回答问题、读懂模糊表达、理解多轮对话，还将能够识别屏幕、总结信息、个性化定制、跨应用操作、在App中执行数百项操作，有望形成跨越众多APP的新应用生态。

1.2 MoE（混合专家模型）成为国内模型发力追赶的有效路径，模型做小、做专成重要趋势

当前流行的Transformer架构，在推理时需要把参数全部过一遍，非常耗费计算资源，被行业普遍认为是一种不经济的底层架构。OpenAI在GPT-4的研发中，探索成功了一条新的路径，即依靠混合专家模型（MoE）来提升模型能力，将原来的稠密模型改为稀疏模型。其原理是每次训练和推理只激活与问题对应的一部分专家，使其能够在保持高参数规模的同时，降低实际计算量，提高模型的效率和速度，达到“术业有专攻”的效果。

GPT-4有1.8万亿参数，使用了16位专家，每个专家都有大约1 110亿参数。国内大模型研发紧随国际步伐，已有腾讯、Minimax等企业，积极转向MoE架构，取得了很好的效果，部分能力已经追平甚至赶超GPT-4。

1.3 行业大模型成为人工智能+战略落地的最后一公里

从国际上看，头部应用ChatGPT的大众用户数已经出现见顶态势（网站全球访问量2023年6月首次出现环比负增长），同时企业用户数开始快速增长（从2024年1月的15万增长到4月的超过60万），反映出大模型需求增量正在从大众市场转向企业市场。

国内来看，在国家政策指引下，各地已经陆续出台了一系列鼓励措施，积极推动大模型在各类场景的应用，包括北京、上海、杭州、广东、重庆等10多个省市，涉及金融、医疗、教育、制造、城市治理、交通、农业农村等多个行业领域。行业头部企业也在积极推进行业应用创新，如腾讯、阿里巴巴等模型厂商推出的基于云计算的MaaS（模型即服务）模式，让各行业根据需要灵活定制，提高行业大模型应用的创新效率。

2 腾讯自主创新突破大模型关键技术

2.1 自研MoE万亿参数大模型，能力比肩GPT-4

腾讯混元大模型在多模态上取得积极进展。

（1）文生文。处于国内领先水平，在中文语境的全景测试中，超过GPT-4.0 Turbo。腾讯混元支持32 k与256 k长文模型，其中256 k具有强大的超长文本处理能力。支持单文档最长1 000万字的超长文处理，能够一次性解析最多50个文件，在长文档的阅读理解和大规模数据分析方面展现出强大性能。

（2）文生图。全面升级至中文原生DiT（Diffusion Transformer）架构，具备多轮绘图能力，测评结果国内领先，比肩国外Midjourney模型和Stable Diffusion模型，现已全面对外开源。评测数据显示，腾讯混元文生图模型是目前效果最好的开源文生图模型，整体能力达国际领先水平，被广泛用于素材创作、商品合成、游戏出圈等多项业务及场景中，被《央视新闻》《新华日报》等20余家媒体在实际的新闻内容生产中采用。

（3）图生视频。腾讯和清华大学、香港科技大学联合推出全新图生视频模型“Follow-Your-Click”。图生视频模型主要功能包括局部动画生成和多对象动画，支持多种动作表达，如头部调整、翅膀拍动等。在生成3D层面，腾讯混元支持多种内容形态生成3D效果，单图仅需30秒即可生成3D模型。

（4）文生视频。采用跟Sora相同的DiT技术，可以生成高质量的16秒视频。腾讯混元大模型积极助力新华社、《人民日报》、中央电视台等主流媒体，生成视频、海报等内容素材，如《人民日报》科幻大片《珍AI地球》等。

2024年5月，腾讯混元大模型再次全面升级，大模型APP版“腾讯元宝”正式上线，致力于成为每个用户生活与工作的人工智能（AI）助手。同时，还上线了一站式AI智能体创作与分发平台“腾讯元器”。用户可以在元器平台上直接创建智能体，并将这些智能体一键分发到QQ、微信客服、腾讯云等渠道，并支持分发到微信公众号和小程序。如，腾讯元宝与热门电视剧《长相思》联动的AI Agent角色聊天，互动总热度已超过2.6亿。当前，腾讯混元模型资源对智能体创作全部免费，帮助广大开发者和用户更好地获得人工智能加持的创新能力。

2024年8月2日，中文多模态大模型SuperCLUE-V基准8月榜单发布，腾讯混元大模型凭借其在多模态理解方面的卓越表现，在众多参评模型中脱颖而出，斩获国内大模型排名第1位，稳居卓越领导者象限。榜单中，腾讯混元大模型总成绩仅略低于GPT-4o，表现好于Claude3.5-Sonnet和Gemini-1.5-Pro，获得总分71.95的高分。显示出在技术和应用层的综合优势。

2.2 自主研发业界领先的机器学习平台

机器学习平台是当前我国在大模型领域需要重点突破的核心技术之一。腾讯自2015年起，自研了Angel大规模机器学习平台，从硬件到软件实现多方面突破。

（1）Angel平台有效破解万亿参数模型训练推理难题。目前，Angel机器学习平台支持单集群多达10万张GPU的组网规模。随着模型训练所需要GPU数量的增加，如何高效互联万卡甚至十万级卡片变得至关重要。英伟达、微软、谷歌、亚马逊等国外科技公司在网络互联方面持续投入。据传，GPT-4的训练使用了约2.5万张A100 GPU，而GPT-5的训练预计使用约5万张H100 GPU（1张H100的算力约等于3张A100）；Meta已经构建了2个包含2.4万张H100的算力集群，用于训练Llama 3大模型，以追赶GPT-5；美国企业家马斯克旗下人工智能公司xAI已经建成一个包含10万张H100的超级算力集群，来支持其Grok大模型的迭代。

（2）腾讯通过自主研发的Angel PTM和Angel HCF

框架，专注于大模型训练和推理，支持单任务万卡级别的超大规模训练和大规模推理服务部署。该平台大模型训练效率提升至主流开源框架的2.6倍，千亿级大模型训练可节省50%的算力成本。在推理方面，Angel平台的推理速度提高了1.3倍，在腾讯混元大模型的文生图应用中，推理耗时从原本的10秒缩短到了3—4秒。

（3）腾讯自研的深度学习框架实现网络、存储、软硬协同3个方面的技术突破。① 自研的星脉RDMA高速网络提升了30%的通信性能，支持单集群12.8万张GPU和单节点3.2 T带宽。通过自研的TCCL通信库等软件，可以高效互联大规模异构计算资源，成本比国外主流的InfiniBand网络下降了70%。② 通过显存+主存统一存储管理技术，可以高效利用低端GPU，并通过模型算子之间的显存共享和优化，使端到端推理性能提升至业界平均水平的2.3倍。③ 采用GPU网络软硬件协同，通过全方位软硬件监控，实现了大模型任务99.5%的稳定性。中国电子学会科技成果鉴定委员会认为，Angel平台整体技术达到国际先进水平，其中面向all-to-all通信的高效缓存调度与管理技术、自适应预采样与图结构搜索技术达到国际领先水平，授予Angel机器学习平台中国电子学会2024年度科技进步奖一等奖。

（4）腾讯Angel平台通过开源开放为产业创新提供支撑，推动行业进步及人才培养。腾讯于2018年将Angel开源给Linux基金会旗下的LF AI基金会，是国内首个LF AI基金会的顶级项目，并获得了2019年度最受欢迎中国开源软件。同时，Angel相关成果已集成到腾讯云一站式机器学习平台，支持中国高校计算机大赛-微信大数据挑战赛、“觅影”医学人工智能算法大赛等各项国内赛事，推动科技人才培养。

3 融入千行百业，打造新质生产力

作为腾讯全链路自研的实用级大模型，混元大模型在内部业务领域的应用不断深化，目前内部近700项业务已经接入腾讯混元大模型，包括腾讯云、腾讯广告、腾讯会议、腾讯文档、微信读书、微信搜一搜等多个应用场景，单日调用次数超过千亿。同时，混元也通过腾讯云服务教育、公司外部零售、金融、医疗、传媒、交通、政务等多个行业客户。

3.1 助力消费互联网智能化升级

（1）腾讯会议。其是最先通过微调接入混元大模型的产品，可提供会中问答、会议总结、会议待办项整理等能力。其中，基于混元推出的AI小助手，可即时回答会议内外问题。例如，走神的时候，问一下小助手：刚刚说了什么？就会获得准确的总结。过去4个月，腾讯会议AI小助手的每日调用量增长了20倍。

（2）腾讯文档。为方便日常办公协作，腾讯文档接入混元大模型，可提供文档辅助创作、文本润色、文本校阅、表格公式和图表生成、PPT生成等能力，有效提高创作的效率和体验。

（3）微信读书。拥有庞大的书库资源，涵盖了各行业广泛的知识。基于混元大模型微信读书推出了AI问书、AI大纲等新功能，通过对现有图书内容的智能提炼和分析，生成准确和高质量的回答，方便用户查阅专业知识，并可以按图索骥，拓展学习地图。

（4）微信输入法。正式上线“一键AI问答”功能，用户只需要在微信内聊天框中输入内容后加一个符号“=”，即可获取人工智能生成的回答内容，非常简便易用。

（5）腾讯内部设计。大模型还可显著提升素材创作生产效率，当前腾讯内部各类设计需求中，90%的角色、场景、地图等原画创作都可由腾讯混元辅助生成。同时，基于腾讯混元的能力，超过99%的标识（Logo）可以通过AI设计。

（6）个体创作。随着大模型与人机协作的深入，个体创作的门槛进一步降低，越来越多的个体借助大模型外脑成为“斜杠青年”“超级生产者”，甚至开启

自己的“一人企业”。

3.2 为各行业提质增效注入新动能

（1）产业服务。在产业服务领域，腾讯推出行业大模型为企业客户提供涵盖模型预训练、模型精调、智能应用开发等一站式行业大模型解决方案。以高浓度的行业数据，加强模型对行业专业知识的理解；结合搜索增强与实时查询能力，提升模型解决产业问题的实时性、准确度、安全性等能力。目前已与20多个行业结合，提供超50个行业大模型解决方案，覆盖零售、教育、金融、医疗、传媒、交通、政务等主要行业，已在智能问答、内容创作、数据分析、代码助手等多个场景开展应用。例如，腾讯云医学行业大模型，已经达到大模型SOTA（最优效果）的水平。60亿参数的小模型就能支持好“问医问药”，也能辅助医生撰写专业文书，如电子病历、出院小结等。目前，该模型已经在上海瑞金医院超过10个头部临床科室为医生和患者提供导辅诊、随访建议、生成电子病历等能力。以体检报告生成为例，平均每5秒即可自动生成一份总结报告，每天自动生成超过500份，报告采纳率达到96%以上。

（2）广告创意。广告是大模型率先落地的行业，通过腾讯混元“文生图”，可以高效创作广告素材，让创作效率提升10倍以上，在提供更多创意素材的同时降低设计成本。基于大模型图像创作引擎，用户可以很方便地使用线稿生图功能，上传产品线稿设计图后，通过提示词和参数设定，快速生成实物设计图，大幅缩短创作与生产周期。腾讯还推出一站式AI广告创意平台“妙思”，提供AI多模态生成能力，提升营销内容创作工作效率，助力提升广告生产及投放效率，其中图生图平均点击通过率提升15%。

（3）客服。客服是最符合大模型知识交互的场景之一，腾讯的企点客服大模型文本机器人接入了大模型知识引擎执行账单查询、退换货类的查询任务，部署成本降低50%。

（4）知识引擎。除了强大的基础大模型，低门槛的开发工具也让大模型更“好上手”。针对知识管理场景，以RAG（检索增强生成）技术架构为基础，整合了OCR文档解析、向量检索、大语言模型、多模态大模型等技术，腾讯推出大模型知识引擎，让AI不仅懂“产业”，更懂“企业”和“产品”。其中，腾讯云向量数据库每日支撑超过3 700亿次向量检索请求，可支持千亿级向量规模存储，百万级QPS及毫秒级查询延迟，适用于大模型的训练推理、RAG场景、AI应用及搜索推荐服务，实现企业数据接入AI的效率比传统方案提升10倍。目前，已在政务、金融、教育、出行、零售等多个行业落地。例如，在人才培训场景，知识引擎可以结合腾讯乐享知识学习平台，把员工的智慧聚集成企业知识库，促进内部知识分享和传播；再如，在客服场景，知识引擎可以融入到客服系统，让客服人员更准确、更高效率地解答客户的问题。除了助力知识管理，腾讯也发布了“大模型图像创作引擎”和“大模型视频创作引擎”，帮企业在图像和视频创作上提质提效。这些大模型原生工具链有效解决了大模型“上手难”的问题，与大模型产品应用一起让“开箱即用”的AI加速落地产业，大幅提升AI普惠水平。

目前，腾讯围绕大模型已经构建起全链路的产品矩阵，包括底层基础设施、自研大模型、模型开发平台、智能体开发平台和面向场景的多元智能应用等，帮助企业客户将大模型快速落地到场景中去（图1）。

3.3 AI4S（人工智能驱动科学研究）加速科研突破

大模型促进AI for science（AI4S）的兴起，腾讯积极响应提供科研用高性能算力，加速推动大模型等AI技术在科学计算领域的应用，探索“科技向善”的更多可能，例如以下2个方面。

（1）天文方面。腾讯、国家天文台、复旦大学计算机科学技术学院联合启动“探星计划”，用云+AI帮助中国天眼FAST加快处理每天庞大的数据量；并通

图 1　腾讯全链路自研大模型助力各行业提质增效

Figure1　Tencent's self-developed full-chain large models help various industries to improve quality and efficiency

过视觉 AI 分析，更高效找到快速射电暴、脉冲星线索。截至目前，已发现 30 颗脉冲星。在比脉冲星探索精度要求更高更快的快速射电暴领域，腾讯还设计了一套全新端到端的 AI 算法，实现同等算力下推动信号处理效率提速 1 800 倍，1 年内就发现 2 颗快速射电暴。

（2）文化方面。腾讯创新性地将文字检测、摹本生成、字形匹配等 AI 算法，综合应用于甲骨文研究。在“殷契文渊”甲骨文数据标注和处理基础上，通过定制化算法，不断丰富完善甲骨文模型库，截至目前已建立覆盖 143 万字的全球最大甲骨文单字数据库，提升甲骨文识别与考释、甲骨论著内容提取等的效率。通过探索甲骨文研究的人机协同新模式，云+AI 将进一步焕活汉字源头，向更多人展现甲骨文的无限魅力。

4　相关建议

（1）激发大模型企业的自主创新活力。支持头部大模型企业构建具有国际竞争力的通用大模型，优先在数据集和算力资源方面保障头部企业需求，鼓励企业进行关键核心技术攻关和原始创新，力争在全球大模型技术加速迭代演进中跻身前列。推出面向中小企业的人工智能培育计划，为本地中小企业和科研机构提供先进云算力、数据集、模型、软件和培训等一站式资源包，扶持中小企业和开发者敏捷创新。

（2）优化大模型发展的数据、人才和算力要素支持。推进政府公共数据开放、版权数据资源整合和行业数据共享。落实国家“数据要素×”政策，设立高质量数据专项工程，加大政府公共数据开放力度，以机器可读方式推进公共数据开源开放，探索建立国家级的公共语料库。制定国家层面的科技人才引进政策，不拘一格吸引全球顶尖 AI 人才，探索通过“科技绿卡”等政策吸引海外 AI 专家，构建“留学—工作—永居”无缝衔接的引才路径。

（3）以行业大模型应用为主推方向，加快人工智能+落地。推动千行百业加快行业大模型的实用落地，充分发挥我国市场庞大、企业众多、应用场景丰富的特点，激发各行业、企业的应用需求，以需求牵引应用创新和算力建设。鼓励产学研用协同创新，采取大模型应用“揭榜挂帅”等方式，促进产学研合作攻关，对行业大模型或应用创新创业企业给予政策支持。

（4）激励新技术和产业发展，将 AI 立法定位为“发展法”。建议当前人工智能立法聚焦支持突破核心技术，鼓励业界尽快推出并应用产品，摆脱“卡脖子”困境。在安全方面，监管规则按照已有规定的内容、数据安全等现行立法执行即可，不宜另行增设新的行业门槛。立法应注重与行业充分沟通，提供稳定的市场预期。

（5）用治理创新开辟新空间，鼓励负责任的大模型应用。着眼当下，促进 AI 发展应用，坚持“科技向善”，增进社会福祉，避免盲目放大未来可能出现的风险。探索建立支持大模型价值对齐的技术和管理措施，推动形成相关的政策指南、标准、技术规范等，以促进负责任的、安全可信的大模型的推广使用。

参考文献

1 Naveed H, Khan A U, Qiu S, et al. A comprehensive overview of large language models. ArXiv, 2023, doi: 10.48550/arXiv.2307.06435.

2 Makridakis S, Petropoulos F, Kang Y F. Large language models: Success and impact. forecasting, 2023, 5(3): 536-549.

3 The 13 biggest AI stories of 2023. HAI Stanford. (2023-12-14)[2024-09-02]. https://hai.stanford.edu/news/13-biggest-ai-stories-2023.

4 Zhang Z, Liu Y, et al. A bibliometric review of large language model research from 2017 to 2023. arXiv, 2023, doi: 10.48550/arXiv.2304.02020.

5 Taulli T. Large language models. (2023-06-03)[2024-09-02]. https://link. springer. com/chapter/10.1007/978-1-4842-9367-6_5.

6 Anil R, Borgeaud S, Alayrac J B, et al. Gemini: A family of highly capable multimodal models. arXiv, 2023, doi: 10.48550/arXiv.2312.11805.

7 Georgiev P, Lei V I, Burnell R, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv, 2024, doi: 10.48550/arXiv.2403.05530.

8 Akter S N , Yu Z C, Muhamed A, et al. An in-depth look at Gemini's language abilities. arXiv, 2023, doi: 10.48550/arXiv.2312.11444.

9 TeamGemini, Google. Gemini ultra benchmark comparison. (2023-12-06)[2024-09-02]. https://assets.bwbx.io/documents/users/iqjWHBFdfxIU/r7G7RrtT6rnM/v0.

10 Google's Gemini: Setting new benchmarks in language models. (2023-08-15) [2024-09-02]. https://www.superannotate. com/blog/googles-gemini-setting-new-benchmarks-in-language-models.

11 赵鑫. 大语言模型(LLM)综述与实用指南. (2023-04-20)[2024-09-02]. https://arthurchiao.art/posts/2023/llm-guide/.
Zhao X. A comprehensive review and practical guide to large language models (LLMs). (in Chinese)

12 以ChatGPT为代表的大型语言模型研究进展. (2023-05-15)[2024-09-02]. https://www.nsfc.gov.cn/csc/20345/20348/pdf/2023/202305-714-723.pdf.
Research progress of large-scale language models represented by ChatGPT. (2023-05-15)[2024-09-02]. https://www. nsfc. gov. cn/csc/20345/20348/pdf/2023/202305-714-723.pdf. (in Chinese)

Development trends of large models and Tencent's independent innovation practice

SI Jason

（Tencent Research Institute, Shenzhen 518054, China）

Abstract The article discusses the emerging trends and application prospects of current large models, using Tencent's Hunyuan large model as an example. It focuses mainly on innovations and implementations of large models in China. Companies like Google, Meta, and OpenAI have launched powerful models such as Google's Gemini and Meta's Llama 3, which have made significant progress in multi-modal applications and reasoning capabilities. China's large models have significantly improved performance and efficiency by adopting the MoE (Mixture of Experts) architecture. Specifically, with its self-developed MoE trillion-parameter large model and deep learning framework, Tencent has made breakthrough advancements in large model technology and achieved exceptional performance in muti-modal applications. Moreover, Tencent has launched a one-stop AI agent creation and distribution platform. Tencent understands that industry-wide, large models are key to implementing AI and strategies, and is actively supporting the application of large models across sectors such as retail, education, finance, healthcare, media, transportation, and government, helping these industries enhance quality and efficiency.

Keywords multimodal, Hunyuan large model, deep learning framework

司　晓　腾讯集团副总裁，腾讯研究院院长。中南财经政法大学法学博士，斯坦福大学访问学者。兼任中国法学会理事、中国版权协会常务理事、深圳市版权协会会长。长期从事互联网产业、法律、经济、政策等领域的实践和学术研究工作。
E-mail: jasonsi@tencent.com

SI Jason　Vice President of Tencent and Dean of Tencent Research Institute. He was a visiting scholar of Stanford Law School and was officially invited to be a postgraduate supervisor of Peking University Law School since 2017. He is currently President of the Shenzhen Copyright Association, Council Member of China Law Society, Executive Director of Copyright Society of China, and Deputy Director of China Industrial Internet Development Alliance. His scholarship addresses legal and economy public issues associated with the Chinese Internet industry, and provides a unique combination of both academic and practical experiences.
E-mail: jasonsi@tencent.com

■责任编辑：文彦杰