Qualcomm

AI变革正在推动终端侧推理创新

生成式模型的激增和演进，将如何改变AI格局并释放价值

2025年2月

骁龙、高通、以及其他Snapdragon与Qualcomm旗下的产品系高通技术公司和/或其子公司的产品。

目录

摘要

尖端AI推理模型DeepSeek R1一经问世，便在整个科技行业引起波澜。因其性能能够媲美甚至超越先进的同类模型，颠覆了关于AI发展的传统认知。

这一关键时刻是更广泛趋势的一部分，凸显了行业在打造高质量小语言模型和多模态推理模型方面的创新，以及这些创新正在为AI的商用应用和终端侧推理落地做好准备。这些新模型能够在终端侧运行，将加速强大边缘侧芯片的规模化扩展，并创造对此类芯片的需求。

四大趋势正在显著提高目前可在终端侧运行的AI模型的质量、性能和效率，从而推动上述变革：

- **当前先进的AI小模型已具有卓越性能。**模型蒸馏和新颖的AI网络架构等新技术能够在不影响质量的情况下简化开发流程，让新模型的表现超越一年前推出的仅能在云端运行的更大模型。
- **模型参数规模正在快速缩小。**先进的量化和剪枝技术使开发者能够在不对准确性产生实质影响的情况下，缩小模型参数规模。
- **开发者能够在边缘侧打造更丰富的应用。**高质量AI模型快速激增，意味着文本摘要、编程助手和实时翻译等特性在智能手机等终端上的普及，让AI能够支持跨边缘侧规模化部署的商用应用。
- **AI正在成为新的UI。**个性化多模态AI智能体将简化交互，高效地跨越各种应用完成任务。

高通技术公司在引领并利用从AI训练向大规模推理转型，以及AI计算处理从云端向边缘侧扩展方面具有战略优势。公司在开发定制CPU、NPU、GPU和低功耗子系统领域取得了广泛的成就。通过与模型厂商展开合作，以及面向跨不同边缘终端领域的模型部署提供工具、框架和SDK，高通技术公司赋能开发者在边缘侧加速采用AI智能体和应用。

近期对AI模型训练方式的颠覆变革和重新评估验证了AI格局即将向大规模推理转变的趋势，这将形成全新边缘侧推理计算的创新和升级周期。尽管模型训练仍将在云端进行，但推理将受益于采用高通®技术的广泛终端规模，并催生更多边缘侧AI赋能处理器的需求。

高质量AI模型目前已变得丰富且经济实惠

创新推动模型质量提升，减少开发时间和成本

AI模型训练成本的下降和开源合作相结合，让更多的开发者和组织能够进行高质量模型开发。

这种转变是由多种技术进步共同推动的。使用更长上下文文本和简化一些训练步骤，能够节省计算成本。从混合专家模型（MoE）到状态空间模型（SSM）等较新的网络架构，正在以更少的计算开销和功耗不断实现技术突破。

新一代AI模型还集成了诸如思维链推理（Chain-of-Thought Reasoning）和自我验证等先进方法，能够在数学、编码和科学推理等各种颇具挑战性的领域获得出色表现。

蒸馏（Distillation）是开发高效小模型的一项关键技术。它能够让大模型“教学”小模型，保持准确性的同时迁移知识。蒸馏技术的使用促使小型基础模型激增，包括众多面向特定任务调优的模型。

图1展示了蒸馏的强大能力。这里比较了Llama 3.3 700亿参数模型和同类DeepSeek R1蒸馏模型的LiveBench平均测试结果，显示出在相同参数规模下，蒸馏能够在推理、编程和数学任务中显著提高性能。

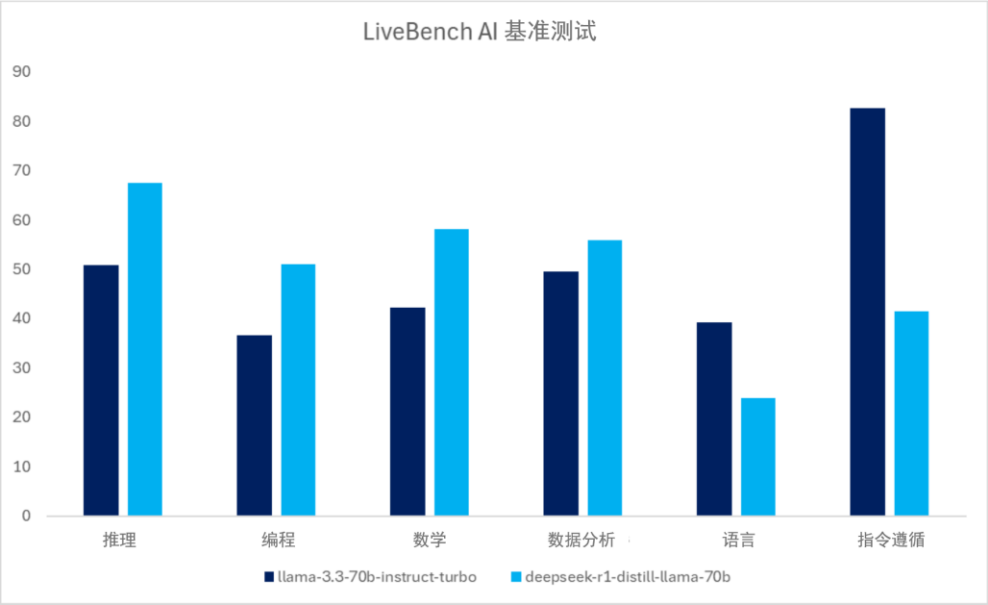

图1：Meta Llama 700亿参数模型和DeepSeek对应蒸馏模型的LiveBench AI基准测试平均结果对比。
来源：LiveBench.ai，2025年2月。

小模型在边缘侧实现了强大功能

得益于蒸馏和上述其他技术，小模型正在接近前沿大模型的质量。图2显示了DeepSeek R1蒸馏版本与其他领先模型的基准测试结果对比。基于通义千问模型和Llama模型的DeepSeek蒸馏版本展现了诸多明显优势，尤其是在GPQA基准测试中，与GPT-4o、Claude 3.5 Sonnet和GPT-o1 mini等先进模型相比，取得了相似或更高的分数。GPQA是一个关键评估指标，因其涉及解决复杂问题的深层次、多步骤的推理，这对许多模型颇具挑战性。

模型	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces 评分
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	44.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

来源：https://github.com/deepseek-ai/DeepSeek-R1

图2：数学和编程基准测试。来源：DeepSeek，2025年1月。

许多主流模型系列包括DeepSeek R1、Meta Llama、IBM Granite和Mistral Ministral都推出了小模型版本，且面向特定任务的性能和基准测试都表现出色。将大型基础模型缩减为更小、更高效的版本，不仅能实现更快的推理速度、更少的内存占用和更低的功耗，同时可以保持较高的性能水平，从而使此类模型适合在智能手机、PC和汽车等终端上部署。

量化、压缩和剪枝等进一步优化技术，有助于缩小模型规模。量化能够降低功耗，且在不明显影响准确性的情况下通过降低精度加速运算，剪枝则可以消除不必要的参数。

这些技术进步推动了高质量生成式AI模型的激增。根据Epoch AI整理的数据（图3），在2024年发布的大规模AI模型中，超过75%的模型参数在千亿规模以下。

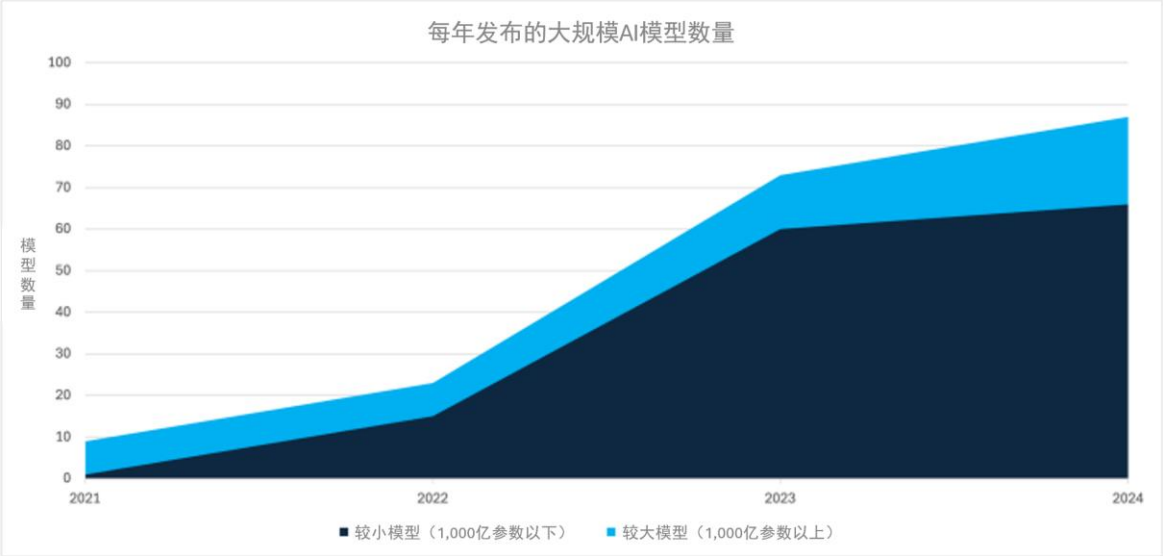

图3：每年发布的大规模AI模型数量（按参数量分类）。来源：Epoch AI，2025年1月。

AI推理创新时代已经到来

大量高质量小模型的涌现让推理工作负载再次受到关注，这是应用和服务利用模型为企业和消费者提供价值的关键环节。

高通技术公司已经优化了大量AI模型，以支持新一代面向AI的Windows 11 AI+ PC商用。同样，高通还与三星和小米等OEM厂商合作推出了众多支持丰富AI赋能特性的旗舰智能手机。

AI推理功能在终端侧的广泛普及赋能打造了丰富的生成式AI应用和助手。文档摘要、AI图像生成和编辑以及实时语言翻译现在已成为常见功能。影像方面的应用可以利用AI进行计算摄影、物体识别和实时场景优化。

这之后迎来了多模态应用的发展，这类应用结合多种数据类型（文本、视觉、音频和传感器输入），以提供更丰富、更具情境感知能力且更加个性化的体验。高通AI引擎结合了定制NPU、CPU和GPU的能力，能够在终端侧优化此类任务，使AI助手能够在不同沟通模式间切换，并生成多模态输出。

智能体AI（Agentic AI）是下一代用户交互的核心。AI系统能够通过预测用户需求，并在终端和应用内主动执行复杂工作流，进行决策和管理任务。高通技术公司注重高效、实时的AI处理，支持智能体在终端侧持续安全地运行，同时依靠个人知识图谱准确定义用户偏好和需求，无需依赖云端。随着时间推移，这些技术进步正在为AI成为主要UI奠定基础，通过自然语言和基于图像、视频与手势的交互简化人们使用技术的方式。

展望未来，高通技术公司在将AI功能融入机器人的具身AI时代也同样具有优势。利用推理优化技术专长，高通技术公司旨在支持机器人、无人机和其他自主设备（Autonomous Devices）进行实时决策，在动态的真实环境中实现精确交互。

尽管许多AI模型在云端训练，但通常蒸馏小模型在几周或几天内即可投入运营并在终端上运行。例如，在不到一周内，DeepSeek R1蒸馏模型已经能在搭载骁龙®平台的PC和智能手机上运行。

在终端内部署推理能够通过降低时延实现即时性，提高隐私性，依靠本地数据提供更多情境信息，以及实现AI特性和应用的持续运行。此外，还通过规避云推理服务相关费用，为用户和/或开发者降低了成本。这一切都将激励软件和服务提供商在边缘侧部署AI推理。

高通将成为AI推理时代引领者

作为终端侧AI的引领者，高通技术公司凭借面向边缘终端的行业领先硬件和软件解决方案，在推动AI推理时代发展上具有战略优势。这些解决方案涵盖了数十亿台智能手机、汽车、XR头显和眼镜、PC以及工业物联网终端等。

高通技术公司长期致力于开发定制CPU、NPU、GPU和低功耗子系统，同时拥有封装技术和热设计的技术专长，构成了其行业领先系统级芯片（SoC）产品的基础。这些SoC能够直接在终端侧提供高性能、高能效的AI推理。通过紧密集成这些核心组件，高通技术公司的平台可在保持电池续航和整体能效表现的同时处理复杂AI任务，这对边缘侧用例至关重要。

为了在平台上充分释放AI潜能，高通技术公司构建了强大的AI软件栈，旨在赋能软件开发者。高通AI软件栈包括库（libraries）、SDK和优化工具，可简化模型部署并提升性能。开发者可以利用这些资源，面向高通平台高效进行模型适配，缩短AI赋能应用的上市时间。高通技术公司采取开发者为中心的策略，通过简化在消费和商用产品中集成先进AI特性的过程，不断加速创新。

最后，作为高通面向各行各业规模化扩展AI战略的核心，高通与全球AI模型厂商积极合作，并提供高通AI Hub等服务[1]。在高通AI Hub上，仅需简单三步，开发者即可：1）选择模型，或引入自主模型又或基于他们的数据创建模型；2）选择任意框架和runtime，在基于云的物理设备场（cloud-based physical device farm）上撰写和测试AI应用；以及3）使用工具商业化部署其应用。高通AI Hub支持主流大语言模型和多模态大模型（LLM、LMM）系列，让开发者可在搭载高通平台的终端上部署、优化和管理推理任务。借助预优化模型库和支持定制模型优化与集成等特性，高通技术

[1] 高通AI Hub不向中华人民共和国境内公众提供生成式人工智能服务。高通AI Hub相关法律文件要求使用者在使用高通AI Hub时，遵守所有适用法律法规的规定。

公司赋能加速开发周期，同时增强了与广泛AI生态的兼容性。这种合作方式强化了高通技术公司在推动规模化、实时化AI应用方面的领导地位。

扩展覆盖所有关键边缘细分领域

得益于终端侧本地AI处理带来的增强的性能、效率、响应速度和隐私性，高通技术公司利用终端侧AI赋能众多行业、释放商业价值并支持全新用户体验。

手机

骁龙移动平台（如最新的骁龙8至尊版）通过赋能多种先进多模态生成式模型和智能体AI在智能手机上原生运行，正在推动终端侧AI功能的发展。AI在多个方面增强了智能手机功能，如通信优化、生成式图像编辑工具、个性化和无障碍功能。终端侧生成式AI正用于开发更直观、以用户为中心的特性，并在移动终端上自主执行任务。

在三星、华硕、小米、OPPO、vivo和荣耀等主要制造商基于骁龙平台推出的最新旗舰智能手机中，这种由AI驱动功能的发展趋势尤为凸显。

PC

骁龙X系列平台凭借专为实现高性能、高能效的生成式AI推理而打造的、业界领先的定制NPU核心，对定义全新AI PC品类发挥了关键作用。该NPU为Windows应用带来显著加速、增加全新特性、提升性能，并增强隐私保护和电池续航。开发者可在终端侧运行生成式AI推理，提供首次亮相在骁龙X系列PC上的Windows 11 AI+ PC先进特性。

Zoom、Affinity、Djay Pro、剪映、Moises Live和Blackmagic Design的DaVinci Resolve等流行的第三方应用，充分利用NPU在骁龙X系列平台上提供特定的AI赋能功能。

汽车

骁龙®数字底盘™解决方案在其情境感知智能座舱系统中使用终端侧AI，旨在增强汽车安全和驾驶体验。该系统利用先进摄像头、生物识别、环境传感器以及先进的多模态AI网络，提供根据驾驶员状态和环境条件而调整的实时反馈和功能。

针对自动驾驶和辅助驾驶系统，高通技术公司开发了端到端架构，利用大规模训练数据集，基于真实世界数据和AI增强数据的快速再训练、OTA更新以及包括车内多模态AI模型和因果推理在内的先进软件栈，应对现代自动驾驶和辅助驾驶的复杂性。

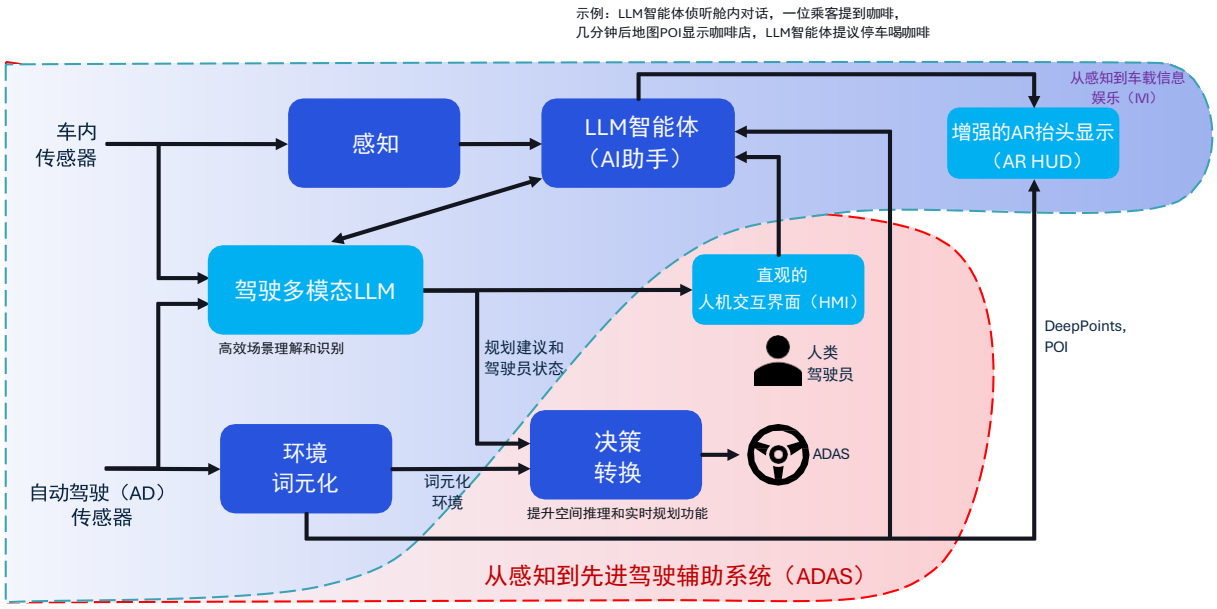

图4：简化的车内AI系统架构以支持智能座舱、自动驾驶和先进驾驶辅助。
来源：高通技术公司，2025年1月。

工业物联网

对于工业物联网和企业应用，高通技术公司近期推出Qualcomm® AI本地设备解决方案（一款可灵活摆放的本地硬件解决方案）和Qualcomm® AI推理套件（一套覆盖从近边缘到云端的AI推理软件和服务）。

边缘AI方案让敏感客户数据、调优模型和推理负载能够保留在本地，增强隐私性、可控性、能效和低时延。这对于AI赋能的业务应用至关重要，比如智能多语言搜索、定制AI助手和智能体、代码生成以及用于用户安全、安防和现场监控的计算机视觉。

网络

高通技术公司已推出AI赋能的Wi-Fi联网平台——高通®A7 Elite专业联网平台。该解决方案集成Wi-Fi 7和边缘AI，让接入点和路由器可以代表网络中的网联终端运行生成式AI推理。它支持安全、能源管理、虚拟助手和健康监测等领域的创新应用，通过在网关处理数据，从而增强隐私性和实时响应。

该联网平台有望将Wi-Fi路由器、Mesh系统、宽带网关和接入点转变为家庭和企业内部私有、本地且基于AI的小型服务器。

总结

在训练成本下降、快速推理部署和针对边缘环境的创新推动下，AI正在经历重要变革。科技行业不再仅仅聚焦于竞相构建更大的模型，而是转向如何在边缘侧实际应用中高效地部署模型。

对大型基础模型的蒸馏已催生大量更智能、更小型、更高效的模型，使各行业能够更快地规模化集成AI，特别是在终端侧加速集成。

凭借高能效芯片设计、先进AI软件栈和面向边缘应用的全面开发者支持等技术专长，高通技术公司具有引领和受益于这一变革的独特优势。

高通技术公司通过将NPU、GPU和CPU集成到终端设备中，实现了跨智能手机、PC、汽车和工业物联网领域的高性能、高能效AI推理。高通技术公司为各行业带来了高性能、经济实惠、快速响应和注重隐私的变革性AI体验。

公司的生态系统策略——包括高通AI软件栈、高通AI Hub和战略性的开发者协作——加速了自适应AI技术的部署。这些解决方案有助于满足注重实时性能、隐私和效率的相关行业需求。

随着AI创新在边缘侧爆发，高通技术公司在可扩展硬件和软件方面的投入将进一步巩固其领导力。公司正在推动一个全新时代的到来，让AI应用更加触手可及、更高效，并且融入日常生活的方方面面，推动全球多个行业的变革。