

项目榜单

榜单名称	大模型边缘和端侧高效推理 AI SoC主控芯片		
行业领域	人工智能	专业方向	大模型
(计划)启动时间	2024年1月1日	计划完成时间	2027年1月1日
榜单提出目的	<p>随着人工智能技术的快速发展，大型深度学习模型在各个领域展现出了巨大的潜力，例如自然语言处理、计算机视觉、医疗诊断等。这些模型的出现为解决复杂问题提供了新的思路和方法。然而，随着模型规模的增大和应用场景的复杂化，传统的通用计算硬件已经不能满足其高效运行的需求。因此，我们迫切需要一种适应行业大模型的高效灵活神经网络处理器，以应对模型规模的不断增长和应用场景的多样化，加快大模型技术普及，让更多人可以随时随地的应用大模型工具生产或改善工作或生活，成为社会的一种新型生产力，造福社会。</p>		
榜单任务内容	<p>围绕基于“大模型边缘和端侧高效推理 AI SoC主控芯片”的课题，本项目设计一种自适应可重配置的神经网络加速器，可以满足常用神经网络算法的速实现，同时可以适应多种不同结构的神经网络。目的是：</p> <p>1、设计高效的神经网络处理器指令集，能够加速大模型计算</p> <p>2、设计灵活的神经网络多核加速器，提升访存效率并能够灵活扩容</p> <p>3、研究更适合大模型的量化技术，并集成到灵活的部署工具中，适应主流的使用方式</p> <p>关键技术将达到以下指标；</p> <p>(1)采用1.5GHz四核CPU，支持Android和TEE（可信执行环境）安全系统；</p> <p>(2)NPU支持fp 16，int8和int 4精度类型，int8峰值算力可达200 TOPS；</p> <p>(3)支持SATA3.0、PCIe3.0、USB3.0和FEPHY等高速传输与存储接口；</p> <p>(4)DDR内存≥24 GB，存带宽≥200 GB / s；</p> <p>(5)支持H264/H265 六路1080P@30fps实时编解码；支持B帧、背景建模和码率控制；</p> <p>(6)ISP最高支持13M@30fps处理能力，支持分时复用8路1080P@15fps。</p>		

<p>榜单效益目标</p>	<p>本项目旨在解决行业大型模型计算效率低、数据传输效率低等关键问题，推动神经网络处理器技术的发展和​​应用，为人工智能技术的广泛应用和深入发展提供强有力的支撑。</p> <p>1、提高行业大模型的计算效率和灵活性，通过提供高效灵活的神经网络处理器，可以加速模型的训练或推理过程，从而提高人工智能系统的整体性能。</p> <p>2、降低大型模型的部署和运行成本，促进人工智能技术的普及和深入发展。通过研发成本更低、能效更高的神经网络处理器，可以降低人工智能技术的门槛，使更多的企业和个人能够轻松使用这些技术。</p> <p>3、提供灵活的板级互联，适应日益增大的大模型规模，满足不同规模应用需求。</p>
---------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------