产业政策与法规研究

赛迪研究院 主办

2025 年 10 月 20 日　　第 4 期　　总第 59 期

本期主题

赛迪智库

面向政府·服务决策

奋力建设国家高端智库

思想型智库　国家级平台　全科型团队

创新型机制　国际化品牌

《赛迪专报》《赛迪要报》《赛迪深度研究》《美国产业动态》《赛迪前瞻》

《赛迪译丛》《国际智库热点追踪周报》《工信舆情周报》《国际智库报告》

《新型工业化研究》《工业经济研究》《产业政策与法规研究》《工业和信息化研究》

《先进制造业研究》《科技与标准研究》《工信知识产权研究》《全球双碳动态分析》

《中小企业研究》《安全产业研究》《材料工业研究》《消费品工业研究》《电子信息研究》

《集成电路研究》《信息化与软件产业研究》《网络安全研究》《未来产业研究》

『所长导读』

当前，以生成式大模型为代表的人工智能技术正以前所未有的速度融入经济社会各领域。在这一进程中，端侧大模型作为 AI 能力从云端向终端下沉的重要载体，以其本地化、低延迟、强隐私等特征，正逐步成为智能终端、工业物联网、智能家居等场景中的关键技术支撑。然而，端侧部署在提升用户智能体验同时，也带来了全新的安全挑战，亟需从技术、管理、法律等多维度系统研判与应对。

本报告聚焦于端侧大模型在部署与应用过程中所面临的安全风险，重点梳理其在智能手机个人助理、智能家居语音控制中枢、工业物联网边缘节点等典型场景下的功能特点与安全隐患。报告从数据安全、模型安全、算法安全三大层面深入剖析风险成因与影响，并结合国内外实践，提出构建涵盖法律法规、行业标准、企业内控与用户教育的综合治理体系。本报告旨在为政策制定者、行业从业者及广大用户提供端侧大模型安全治理的参考框架，推动在技术创新的同时筑牢安全底线，促进人工智能在端侧场景中健康、可信、可持续发展。由于研究涉及领域较新、内容广泛，报告中若有疏漏与不足，恳请各位读者批评指正。

赛迪研究院政策法规所（工业和信息化法律服务中心）所长 彭健

2025 年 10 月 20 日

目录

CONTENTS

本期主题：端侧大模型安全风险与治理研究

本期主题：

端侧大模型安全风险与治理研究

一、端侧大模型概述与发展现状

（一）端侧大模型的定义与核心特征

1. 定义：部署于终端设备的大型人工智能模型

端侧大模型（On-device LargeModel）是指将经过优化和压缩的大型人工智能模型直接部署在智能手机、个人电脑、物联网设备、智能家居中枢以及工业边缘计算节点等终端设备上，使其能够在本地环境中独立运行和处理数据的人工智能系统。这种模型通常具备大规模参数，通过在海量数据集上进行无监督学习并进行有监督微调得到，能够执行广泛的任务。与传统的云侧大模型（Cloud-based Large Model）相比，端侧大模型的核心在于其“端侧部署”的特性，即模型的推理和计算过程主要在用户设备本地完成，而非依赖远程云端服务器。这种部署方式旨在克服云侧模型在实时性、数据隐私和网络依赖性等方面的固有局限。根据部署设备的不同，端侧大模型可以进一步细分为手机大模型、PC 大模型等，它们共同构成了端侧 AI 生态的核心。例如，vivo 推出的蓝心大模型和蔚来汽车的 NOMIGPT 大模型，都是端侧大模型在各自领域的典型应用。这种模型的出现，标志着 AI 技术正从集中式的云端服务向分布式的终端智能演进，旨在为用户提供更加即时、个性化且安全的智能体验。

2. 核心特征：本地化、低延迟、强隐私性

端侧大模型的核心特征主要体现在其本地化运行、低延迟响应和强化的隐私保护能力上。首先，本地化运行是其最本质的特征。模型

和数据均存储在终端设备上，推理过程无需将数据上传至云端，从而实现了完全的离线或弱网环境下的智能服务。这不仅降低了对网络带宽的依赖，也避免了因网络波动或中断导致的服务不可用问题。其次，低延迟响应是端侧部署带来的直接优势。由于数据处理在本地完成，省去了数据往返云端传输的时间，使得模型能够以极低的延迟（通常用 TTFT，即 Time-to-First-Token 来衡量）响应用户请求，这对于需要实时交互的应用场景（如语音助手、实时翻译）至关重要。最后，强隐私性是端侧大模型备受关注的价值所在。敏感的个人数据（如生物特征、聊天记录、健康信息）无需离开设备，从根本上减少了数据在传输和云端存储过程中被泄露或滥用的风险，更好地满足了日益严格的数据隐私法规（如 GDPR、个人信息保护法）的要求。这三大特征共同构成了端侧大模型的核心竞争力，使其在消费电子、智能家居、自动驾驶等领域展现出巨大的应用潜力。

（二）端侧大模型的发展驱动力

1. 技术驱动：模型压缩与硬件性能提升

端侧大模型的快速发展离不开两大关键技术的进步：模型压缩技术和终端硬件性能的提升。一方面，模型压缩技术的突破使得在资源受限的终端设备上运行大模型成为可能。由于大型语言模型通常拥有数十亿甚至数千亿的参数，直接部署在内存和算力有限的终端设备上是不切实际的。因此，研究人员开发了多种模型压缩方法，主要包括量化（Quantization）、剪枝（Pruning）、模型蒸馏（Distillation）和低秩分解（Low-RankFactorization）等。量化技术通过降低模型参数的数值精度（如从 32 位浮点数量化到 8 位整数）来减小模型体积和计算量；剪枝技术则通过移除模型中冗余的连接或神经元来简化网络结构；模型蒸馏则通过一个大型、高精度的“教师模型”来指导一个轻量级的“学生模型”进行训练，使其在保持较高性能的同时显著缩减体量。这些技术的综合应用，使得原本庞大的

模型能够被“瘦身”至适合在终端设备上运行的规模。

另一方面，终端硬件性能的持续提升为端侧大模型的运行提供了坚实的物理基础。以智能手机为例，其搭载的处理器（如高通骁龙系列、苹果 A 系列芯片）不仅在通用计算能力上不断增强，还集成了专门用于 AI 计算的神经网络处理单元（NPU）或 AI 加速器，极大地提升了设备端的 AI 算力。同时，内存容量和带宽的增加也为加载和运行更大规模的模型创造了条件。尽管与高端 GPU 相比，手机芯片在绝对算力上仍有较大差距，但其能效比的优化使得在功耗可控的前提下进行复杂的 AI 推理成为现实。硬件性能的提升与模型压缩技术的进步相辅相成，共同推动了端侧大模型从理论走向实践。

2. 应用驱动：个性化体验与数据隐私需求

端侧大模型的兴起，深刻反映了用户对个性化智能体验和数据隐私保护日益增长的需求。在应用层面，个性化体验是端侧大模型的核心价值主张。由于模型在本地运行，它可以持续学习和分析用户的行为习惯、偏好设置、日程安排等个人数据，而无需将这些敏感信息上传至云端。这使得 AI 助手能够提供高度定制化的服务，例如，根据用户的日常作息自动调节智能家居设备，或根据用户的阅读习惯推荐个性化内容。这种深度个性化是云侧模型难以实现的，因为后者通常处理的是脱敏后的、聚合的数据，难以捕捉到个体用户的细微特征。端侧大模型通过将 AI 能力下沉到设备，真正实现了“千人千面”的智能服务，极大地提升了用户体验。

与此同时，数据隐私需求是推动端侧大模型发展的另一大关键因素。随着全球范围内数据隐私法规的日趋严格（如欧盟的 GDPR 和中国的《个人信息保护法》），以及公众隐私意识的普遍觉醒，用户对于个人数据被收集、传输和使用的担忧日益加剧。传统的云侧 AI 服务模式，需要将用户数据上传至云端进行处理，这在无形中增加了数据泄露和滥用的风险。端侧大模型通过在本地处理数据，从源头上切断了数据外泄的链条，为用户提供了一

种更为可信的隐私保护方案。这种“数据不动，模型动”的模式，不仅符合法规要求，也迎合了用户对数据主权的诉求，成为厂商在激烈的市场竞争中构建差异化优势的重要卖点。

3. 产业驱动：AIoT 生态的蓬勃发展

端侧大模型的发展与 AIoT（人工智能物联网）生态的蓬勃发展紧密相连，二者相互促进，共同构成了产业智能化的核心驱动力。AIoT 的核心愿景是实现万物互联和智能交互，而端侧大模型正是实现这一愿景的关键技术。在 AIoT 生态中，数以百亿计的设备（如智能家居、可穿戴设备、工业传感器）需要具备感知、理解和决策的能力。将这些设备全部连接到云端进行处理，不仅会带来巨大的网络带宽压力和延迟问题，还会引发严重的数据隐私和安全挑战。端侧大模型通过在设备端部署智能，使得每个设备都能成为一个独立的智能节点，能够自主地进行数据处理和决策，从而构建一个更加高效、可靠和安全的分布式智能网络。

产业界已经深刻认识到端侧 AI 的战略价值，并纷纷布局。各大手机厂商将端侧大模型作为其 AI 战略的核心，通过升级语音助手、优化影像处理等方式，提升产品竞争力。智能家居厂商则利用端侧大模型，让家居设备能够更好地理解用户意图，实现更加自然和智能的交互。在工业领域，边缘计算节点部署端侧模型，可以实现对生产数据的实时分析和预测性维护，提高生产效率和安全性。根据 IDC 的预测，到 2025 年，中国 AIPC、AI 平板和 AI 手机的出货量将显著增长，终端侧 AI 功能将成为标配。这种产业层面的广泛布局和应用，不仅为端侧大模型提供了丰富的应用场景和市场需求，也反过来推动了相关技术的快速迭代和成熟，形成了一个良性的产业发展闭环。

（三）端侧大模型的技术架构

1. 模型部署与运行框架

端侧大模型的部署与运行框架是实现其在终端设备上高效、稳定运行的基础。这一框架通常涉及模型优化、运行时环境和硬件加速等多个层面。首先，在模型优化方面，

为了在资源受限的终端设备上部署大模型，必须采用一系列模型压缩技术。这包括量化，将模型参数从高精度浮点数转换为低精度整数，以减小模型体积和计算开销；剪枝，移除模型中不重要的权重或神经元，简化网络结构；以及知识蒸馏，利用大型教师模型指导小型学生模型进行训练，使其在保持性能的同时大幅缩减规模。这些优化技术共同作用，将庞大的云端模型“瘦身”为适合端侧部署的轻量级模型。

其次，运行时环境为模型在终端设备上的执行提供了必要的软件支持。这通常包括一个轻量级的推理引擎，负责加载模型、管理内存、调度计算任务等。为了充分利用终端设备的硬件资源，推理引擎需要与底层的硬件加速器（如 NPU、GPU）进行深度集成和优化。此外，运行时环境还需要提供安全机制，如代码签名和完整性校验，以确保模型在加载和运行过程中未被篡改，保障模型的安全性和可靠性。例如，电信终端产业协会发布的《移动智能终端端侧大模型安全实施指南》中就明确要求，服务提供者应采取代码签名和完整性校验技术，确保模型在启动和运行时的机密性、完整性和可用性。

2. 端云协同机制

尽管端侧大模型强调本地化运行，但在实际应用中，纯粹的端侧部署往往难以满足所有复杂场景的需求。因此，端云协同成为当前端侧大模型技术架构的主流模式。这种模式结合了端侧的低延迟、强隐私优势和云侧的强大算力与海量数据优势，形成一个互补的智能系统。典型的端云协同架构通常包含三个关键组件：部署在终端的轻量级模型、位于云端的大型模型以及连接二者的协同机制。

在这种架构下，端侧轻量级模型主要负责处理简单、高频、对实时性要求高的任务，例如快速响应用户的语音指令、进行初步的意图识别等。当遇到端侧模型无法处理的复杂任务时，系统会将相关数据（经过脱敏和加密）上传至云端，由能力更强的大型模型进行处理。云端模型完成推理后，将结果返回给终端。这种分工合作的模式，既保证了日常交互的流畅性和隐私

性，又能够利用云端强大的计算能力解决复杂问题。例如，某AI手机的技术架构就明确采用了这种模式，端侧模型负责高效处理简单任务，而复杂任务则由云端模型完成。端云协同机制的设计，如数据如何分流、何时触发云端调用、如何保障数据传输安全等，是实现高效、安全、智能的端侧大模型应用的关键。

3.硬件加速与优化

硬件加速是提升端侧大模型运行效率和用户体验的核心环节。由于大模型的推理过程涉及大量的矩阵和向量运算，单纯依靠CPU进行计算难以满足实时性要求。因此，现代终端设备（尤其是智能手机）普遍集成了专用的AI加速器，如神经网络处理单元（NPU）、数字信号处理器(DSP)或图形处理单元(GPU)的AI计算核心。这些硬件单元针对AI计算的特点进行了专门优化，能够以更高的能效比执行深度学习模型的推理任务。

为了充分发挥硬件加速的优势，需要在软件和算法层面进行协同优化。这包括：算子融合，将多个连续的计算操作合并为一个，减少内存访问开销；内存布局优化，根据硬件的访存特性，调整数据和权重的存储方式，提高缓存命中率；以及并行计算，将计算任务合理地分配到多个计算核心上，实现并行处理。此外，硬件厂商通常会提供配套的软件开发工具包（SDK）和编译器，帮助开发者将训练好的模型高效地部署到其硬件平台上。例如，高通、联发科等芯片厂商都提供了相应的AI开发平台，支持主流深度学习框架，并提供模型转换、优化和部署的全套工具链。通过软硬件的协同设计和优化，可以最大限度地挖掘终端设备的AI计算潜力，使得在功耗和散热限制下运行更复杂、更强大的端侧大模型成为可能。

二、端侧大模型典型应用场景分析

（一）智能手机上的个人助理

1.功能与特点：自然语言交互、任务自动化、个性化推荐

智能手机上的个人助理是端侧大模型最典型、最广泛的应用场景之一。其核心功能在于通过自然语言交互，理解用户意图，并自动完

成一系列复杂任务，从而极大地简化手机操作流程，提升用户体验。传统的语音助手多依赖于固定的指令集和关键词匹配，交互方式生硬，功能有限。而基于端侧大模型的个人助理，则实现了从“指令识别”到“意图理解”的飞跃。它能够理解用户模糊、口语化的表达，甚至进行多轮对话，准确把握用户的真实需求。

在任务自动化方面，端侧大模型赋予了个人助理强大的执行能力。用户只需通过自然语言下达指令，如“帮我点一份附近评分最高的川菜外卖”或“把我昨天拍的所有照片备份到云盘并分享给家人”，个人助理就能自动调用多个第三方应用，完成一系列复杂的跨应用操作。这背后通常涉及“屏幕识别 + 模拟点击”的技术路径，即通过截屏理解当前界面内容，然后利用系统的“无障碍权限”模拟用户点击，从而实现对第三方应用的自动化控制。此外，基于对用户数据的本地分析，个人助理还能提供个性化推荐，例如根据用户的日程安排和交通状况，提前提醒用户出发，并自动规划最佳路线。这种深度集成和主动服务的能力，使得个人助理从一个被动的工具，转变为一个主动的、懂用户的“智能伙伴”。

2. 安全风险：无障碍权限滥用、数据隐私泄露、跨应用数据流转不透明

尽管智能手机个人助理带来了极大的便利，但其背后潜藏的安全风险同样不容忽视，主要体现在无障碍权限滥用、数据隐私泄露和跨应用数据流转不透明三个方面。首先，无障碍权限滥用是当前最为突出的风险。为了实现跨应用操作，个人助理通常需要获取系统的“无障碍服务”权限。该权限原本是为辅助残障人士设计的，具有极高的系统权限，能够读取屏幕上的所有内容、监控用户操作、甚至模拟点击。然而，部分厂商为了追求“无感体验”，在调用该权限时存在不透明现象，例如权限开关在执行命令时自动开启和关闭，用户难以感知和控制。更有甚者，一些应用会滥用该权限，在用户不知情的情况下窃取敏感信息，如银行密码、短信验证码，甚至自动执行转账等恶

意操作，给用户财产和个人信息安全带来严重威胁。

其次，数据隐私泄露风险贯穿个人助理的整个工作流程。虽然端侧大模型宣称数据在本地处理，但为了实现复杂功能，许多操作仍需依赖云端大模型。例如，对用户指令的理解、屏幕截图的分析等任务，往往需要将数据上传至云端服务器进行处理。这就导致了所谓的“端侧大模型”名不副实，用户的大量个人数据，包括聊天记录、浏览历史、屏幕内容等，仍然面临着被手机厂商或第三方服务商获取的风险。这些数据是否被用于模型训练，是否经过了充分的脱敏处理，用户往往无从得知，从而形成了巨大的隐私“黑箱”。

最后，跨应用数据流转不透明加剧了风险。个人助理在完成任务时，需要在不同应用之间传递和整合数据。例如，在“帮我订一张去上海的机票”这个指令中，个人助理需要访问用户的日历应用（确认时间）、地图应用（确认出发地）、支付应用（完成付款）等。在这个过程中，数据在不同应用之间如何流转、哪些数据被共享、共享给了谁，这些信息对用户来说通常是模糊不清的。这种不透明性不仅侵犯了用户的知情权，也使得在发生数据泄露或滥用事件时，难以界定手机厂商、应用开发者和第三方服务商之间的责任，给用户的维权带来了困难。

3．治理建议：完善用户授权机制、提升数据链路透明度、明确厂商责任

针对智能手机个人助理存在的安全风险，需要从管理层面构建一套系统性的治理体系，核心在于完善用户授权机制、提升数据链路透明度以及明确各方责任。首先，完善用户授权机制是保障用户知情权和控制权的基础。应强制要求应用在获取“无障碍服务”等高危权限时，必须以清晰、明确、非技术性的语言向用户说明权限的用途、范围和潜在风险，并获得用户的“选择加入”（opt-in）式明确授权，而非默认开启或捆绑授权。授权过程应设计为动态和场景化的，即在不同功能需要调用不同权限时，分别向用户申请，避免“一揽子”授

权带来的风险。此外，系统应提供便捷、统一的权限管理入口，让用户可以随时查看和撤销已授予的权限。

其次，提升数据链路透明度是建立用户信任的关键。当个人助理需要调用云端服务或在应用间共享数据时，必须以可视化、易于理解的方式向用户展示数据流转的路径、涉及的应用或服务、以及数据的使用目的。例如，可以在执行跨应用任务前，弹窗提示用户“即将向 XX 应用共享您的位置信息，用于 XX 目的”，并征得用户同意。这种透明化的设计，不仅能让用户更好地理解和控制自己的数据，也能有效约束厂商的行为，防止数据被滥用。借鉴国外先进经验，如 OpenAI 的数据透明做法，为用户提供清晰的数据使用报告和查询工具，也是提升透明度的有效途径。

最后，明确厂商责任是落实治理措施的保障。需要出台明确的法律法规或行业标准，厘清在端侧大模型生态中，手机厂商、模型提供商、应用开发者等各方主体的权利、义务和责任边界。例如，手机厂商作为平台的构建者和权限的管理者，应承担主要的平台治理责任和安全保障义务；应用开发者作为数据的主要收集者和使用者，应遵循最小必要原则，确保数据的安全存储和合规使用；模型提供商则应保证模型的安全性和可靠性，防止模型被恶意利用。通过建立清晰的责任划分机制，可以在发生安全事件时，快速定位责任主体，并依法追究其责任，从而形成有效的威慑，促进整个生态的健康发展。

（二）智能家居中的语音控制中枢

1. 功能与特点：语音指令控制、场景联动、远程管理

智能家居中的语音控制中枢，如智能音箱、智能中控屏等，是端侧大模型在家庭场景中的核心应用。其主要功能是通过语音指令实现对家中各种智能设备的集中控制，并支持复杂的场景联动和远程管理，从而为用户打造一个便捷、舒适、智能的家居环境。在语音指令控制方面，用户可以通过自然语言与中枢进行交互，例如说出“打开客厅的灯”、“把空调温度调到

26 度”等指令，中枢便能准确识别并执行相应操作。端侧大模型的应用，使得中枢能够理解更复杂、更口语化的指令，甚至支持多轮对话和上下文理解，极大地提升了交互的自然度和准确性。

在场景联动方面，语音控制中枢能够根据预设的规则或用户的实时指令，将多个智能设备组合成一个协同工作的系统。例如，用户可以设置一个“回家模式”，当触发该模式时，中枢会自动执行一系列操作：打开玄关灯、拉开窗帘、启动空气净化器、播放舒缓的音乐等。这种场景化的智能联动，将原本孤立的智能设备连接成一个有机的整体，为用户提供了无缝的智能生活体验。此外，远程管理功能也是其重要特点之一。用户即使不在家中，也可以通过手机 App 等远程方式，向家中的语音控制中枢发送指令，实时查看设备状态，控制家电运行，从而实现了对家庭环境的随时随地掌控，增强了家居的安全性和便利性。

2. 安全风险：身份认证缺陷、网络通信不安全、语音数据被监听

智能家居语音控制中枢在带来便利的同时，也引入了一系列严峻的安全风险，主要集中在身份认证、网络通信和语音数据三个方面。首先，身份认证缺陷是一个普遍存在的问题。许多智能家居设备在出厂时设置了默认的、强度较弱的用户名和密码，或者根本不要求用户修改，这使得攻击者可以轻易地通过暴力破解或利用默认凭证登录设备，获取控制权。一旦攻击者控制了语音中枢，他们就可以随意操控家中的所有联网设备，例如打开智能门锁、关闭安防摄像头，造成严重的财产损失和人身安全威胁。此外，一些设备的身份认证机制设计存在漏洞，可能无法有效区分合法用户和恶意攻击者，导致未经授权的访问。

其次，网络通信不安全是另一个重大风险点。智能家居设备之间以及与云端服务器之间的通信，如果未采用足够强度的加密协议，数据在传输过程中就可能被窃听或篡改。例如，攻击者可以通过监听网络流量，获取用户的语音指令、设备状态等敏感信息，从而分析用户

的生活习惯，甚至推断出用户是否在家。更严重的是，攻击者可以实施中间人攻击（Man-in-the-Middle Attack），截获并篡改通信数据，例如将“关闭窗户”的指令篡改为“打开窗户”，或者伪造设备状态信息，欺骗用户和管理平台。一些智能家居设备使用的通信协议（如Zigbee、Z-Wave）本身可能存在安全漏洞，也容易被黑客利用。

最后，语音数据被监听是用户最为担忧的隐私风险。为了实现随时响应用户的语音指令，语音控制中枢必须时刻保持“在线”状态，持续监听周围环境的声音。这就引发了用户的普遍担忧：设备是否会将用户的日常对话、家庭活动等隐私信息偷偷录音并上传至云端？尽管厂商通常会声明设备只在检测到特定唤醒词后才开始录音，但用户无法验证这一说法的真实性。2020年，小米摄像头被曝出存在漏洞，导致用户可以通过谷歌NestHub看到其他家庭的影像，这一事件暴露了智能家居设备在隐私保护方面的严重缺陷，引发了公众对智能家居安全性的广泛质疑。这种持续的监听模式，使得用户的家庭环境时刻处于被监控的风险之中，对用户的隐私构成了极大的威胁。

3．治理建议：加强网络安全防护、提升数据加密强度、制定行业标准

为了有效应对智能家居语音控制中枢面临的安全风险，需要从管理层面采取一系列综合性的治理措施，重点在于加强网络安全防护、提升数据加密强度以及制定和推广统一的行业标准。首先，加强网络安全防护是保障设备安全运行的基础。设备制造商应从设计之初就将安全理念融入产品，例如，强制要求用户在首次使用时设置高强度的密码，并提供定期修改密码的提醒功能。同时，应建立完善的设备固件更新机制，及时修复已知的安全漏洞，并通过安全通道向用户推送更新包。对于用户而言，应提高安全意识，及时更新设备固件，关闭不必要的网络端口和服务，并将智能家居网络与家庭主网络进行隔离，以防止攻击者通过一个被攻破的设备入侵整个家庭网络。

其次，提升数据加密强度是

保护用户隐私的关键。所有在设备之间以及设备与云端之间传输的数据，都必须采用业界标准的强加密算法（如TLS/SSL）进行端到端加密，确保数据在传输过程中的机密性和完整性。对于存储在设备本地和云端的数据，也应进行加密处理，防止数据在物理存储介质被盗或被非法访问时泄露。此外，对于语音数据的处理，应尽可能在本地完成，减少向云端传输敏感语音数据的需求。如果必须上传，也应采用差分隐私、联邦学习等技术，对数据进行脱敏处理，确保无法从数据中识别出特定个人。

最后，制定和推广统一的行业标准是规范市场秩序、提升整体安全水平的根本保障。政府监管部门和行业协会应加快制定针对智能家居设备的安全标准和技术规范，明确设备在身份认证、数据加密、漏洞管理、隐私保护等方面的最低安全要求。例如，可以参考《信息安全技术智能家居通用安全规范》（GB/T41387-2022）等国家标准，对智能家居终端、网关、控制端和应用服务平台等各个环节提出具体的安全要求。同时，应建立第三方安全认证和评估体系，对市场上的智能家居产品进行安全测评，并向消费者公布通过认证的产品名单，引导消费者选择更安全可靠的产品。通过行业标准的引导和约束，可以推动整个智能家居产业向着更加安全、健康的方向发展。

（三）工业物联网中的边缘计算节点

1. 功能与特点：实时数据处理、预测性维护、自动化控制

在工业物联网（IIoT）领域，边缘计算节点是部署在工业现场、负责数据采集、处理和控制的智能终端，而端侧大模型的应用则极大地增强了这些节点的智能化水平。其核心功能主要体现在实时数据处理、预测性维护和自动化控制三个方面。首先，在实时数据处理方面，工业现场会产生海量的实时数据，如设备运行参数、环境传感器读数、产品质量检测数据等。传统的做法是将所有数据上传至云端进行处理，但这会带来巨大的网络带宽压力和延迟。通过在边缘计算节点部署端侧大模型，可以在数据源头进

行实时的分析和处理，快速提取有价值的信息，只将关键结果或异常事件上传至云端，从而显著降低网络负载，并满足工业控制对低延迟的严苛要求。

其次，预测性维护是端侧大模型在工业领域的重要应用。通过对设备运行数据进行持续的学习和分析，边缘节点上的模型可以识别出设备故障的早期征兆，并预测其剩余使用寿命。例如，通过分析电机的振动和温度数据，模型可以提前预警轴承的磨损情况，从而允许维护人员在设备发生故障前进行干预，避免代价高昂的非计划停机。这种基于数据驱动的预测性维护，相比于传统的定期维护或故障后维修，能够显著提高设备综合效率（OEE），降低维护成本。最后，在自动化控制方面，端侧大模型可以使边缘节点具备更强的自主决策能力。例如，在自动化生产线上，部署在视觉检测工位的边缘节点可以利用大模型对产品质量进行高精度的实时判断，并根据判断结果自动调整上游设备的参数，形成一个闭环的、自适应的智能控制系统，从而提升产品质量和生产效率。

2. 安全风险：边缘节点被攻击、数据泄露与篡改、物理安全威胁

工业物联网中的边缘计算节点由于其部署环境的开放性和复杂性，面临着比传统 IT 系统更为严峻的安全风险，主要包括边缘节点被攻击、数据泄露与篡改、以及物理安全威胁。首先，边缘节点被攻击的风险极高。由于边缘节点通常部署在物理防护较弱的工业现场，它们更容易成为攻击者的目标。攻击者可以通过网络攻击（如漏洞利用、恶意软件感染）或物理接触（如 USB 接口接入）等方式，入侵边缘节点，窃取其中的模型和数据，或者篡改其控制逻辑，从而对生产过程造成破坏。一旦边缘节点被攻陷，攻击者不仅可以获取该节点处理的生产数据，还可能以此为跳板，进一步渗透到企业的核心网络，对整个生产系统造成破坏。

其次，数据泄露与篡改风险是工业领域的核心关切。边缘节点处理的数据往往包含企业的核心生产数据和工艺参数，一旦泄露，将给企业带来巨大的经济损失和竞争劣

势。攻击者可以通过网络攻击窃取数据，或者通过物理接触直接读取存储介质。更严重的是数据篡改，攻击者可以恶意修改边缘节点的分析结果或控制指令，导致生产事故、产品质量下降，甚至危及人身安全。例如，在智能制造场景中，篡改机器人的控制指令可能导致其做出危险动作。最后，物理安全威胁不容忽视。边缘设备通常部署在无人值守的恶劣环境中，面临着各种物理威胁，如极端温度、湿度、振动等，这些都可能影响设备的正常运行。此外，设备的物理接口（如 USB 口、调试口）如果缺乏有效的防护，就可能被非法访问，导致数据泄露或系统被植入恶意软件。

3. 治理建议：实施访问控制、定期安全审计、加强物理安全防护

为了有效应对工业物联网边缘计算节点面临的安全挑战，必须建立一套全面、多层次的安全治理体系。首先，在实施严格的访问控制方面，应遵循“最小权限原则”，为每个用户和设备分配其完成任务所必需的最小权限。所有对边缘节点的访问都应进行严格的身份认证和授权，并采用多因素认证等强认证机制。同时，应建立详细的操作审计日志，记录所有访问和操作行为，以便在发生安全事件时进行追溯和分析。

其次，在定期进行安全审计和风险评估方面，企业应建立常态化的安全审计机制，定期对边缘节点的硬件、软件、网络配置和访问策略进行全面的安全检查，及时发现和修复安全漏洞。同时，应定期开展数据安全风险评估，识别数据在采集、传输、存储、处理等全生命周期中面临的风险，并采取相应的防护措施。最后，在加强物理安全防护方面，应将边缘计算节点部署在安全的物理环境中，如加锁的机柜或专门的设备间，并限制人员的物理访问。对于部署在野外的设备，应采取加固、防盗、防破坏等措施。此外，还应建立完善的设备生命周期管理制度，从设备采购、部署、运行到报废的全过程进行安全管理，确保设备在整个生命周期内的安全可控。

三、端侧大模型安全风险深度剖析

（一）数据安全风险

1.数据泄露：敏感信息在采集、传输、存储过程中的泄露风险

数据泄露是端侧大模型面临的最直接、最普遍的安全风险之一。尽管端侧部署的初衷之一是为了保护数据隐私，但在数据的整个生命周期中，泄露的风险依然无处不在。在数据采集阶段，如果终端设备（如智能手机、智能音箱）的传感器（如麦克风、摄像头）被恶意软件控制，或者存在系统漏洞，攻击者就可能绕过正常的授权机制，秘密地采集用户的语音、图像等敏感信息。在数据传输阶段，即使数据主要在本地处理，但在端云协同的场景下，部分数据（如模型更新、复杂任务请求）仍然需要与云端进行交互。如果通信链路未采用足够强度的加密协议（如 TLS/SSL），或者存在中间人攻击，数据在传输过程中就可能被窃听或截获。在数据存储阶段，存储在终端设备上的模型文件、用户数据、日志文件等，如果未进行加密或加密强度不足，一旦设备被 root 或越狱，或者攻击者通过物理方式获取了存储介质，这些数据就可能被直接读取和窃取。此外，一些应用可能会在用户不知情的情况下，将本地收集的数据缓存或备份到不安全的云端位置，从而引入新的泄露风险。

2.数据篡改：恶意修改训练数据或模型输入，导致模型输出错误

数据篡改是另一种严重的数据安全风险，它通过破坏数据的完整性来影响模型的正常运行。这种攻击可以发生在多个环节。在模型训练阶段，如果攻击者能够接触到训练数据集（尤其是在联邦学习等分布式训练场景中），他们可以通过“数据投毒”（DataPoisoning）的方式，向数据集中注入少量精心构造的恶意样本。这些恶意样本会污染整个模型，使其在特定的输入下产生错误的输出，或者在模型中植入“后门”（Backdoor），使得攻击者可以通过特定的触发器来控制模型的行为。在模型推理阶段，攻击者可以通过各种手段篡改模型的输入数据。例如，在图像识别应用中，通过对图像进行人眼难

以察觉的微小扰动，就可以构造出“对抗性样本”（Adversarial Examples），从而欺骗模型做出错误的分类。在自动驾驶场景中，攻击者可以通过修改交通标志牌的图像，使其被车载系统误识别，从而引发严重的安全事故。此外，攻击者还可以直接篡改存储在终端设备上的模型文件或配置文件，改变模型的行为逻辑，使其输出攻击者预设的结果。

3. 数据滥用：未经授权使用用户数据进行模型训练或商业分析

数据滥用是指数据控制者在未获得用户明确授权，或超出授权范围的情况下，使用用户数据的行为。在端侧大模型的场景下，这种风险尤为突出。尽管数据在本地处理，但模型提供商或应用开发者仍然可能通过各种方式获取用户数据。例如，一些应用可能会在用户协议或隐私政策中，使用模糊、笼统的语言，诱导用户“一揽子”授权其收集和使用个人数据，包括用于模型训练和商业分析。用户往往因为协议过长、术语专业而忽略其中的关键条款，导致其在不知情的情况下“同意”了数据被滥用。此外，即使用户明确拒绝，一些厂商仍可能通过技术手段绕过用户的设置，秘密地收集和使用数据。例如，利用系统漏洞或与其他应用的隐蔽合作，来获取用户的敏感信息。这种未经授权的数据滥用，不仅严重侵犯了用户的隐私权，还可能导致用户被精准画像和定向营销，甚至成为网络诈骗的目标。

（二）模型安全风险

1. 模型窃取：通过逆向工程等手段窃取模型参数和结构

模型窃取（ModelStealing）或模型提取（ModelExtraction）攻击，是指攻击者通过与目标模型进行交互（例如，向其发送查询并观察其输出），来推断或重建出目标模型的参数、结构或功能。对于部署在终端设备上的端侧大模型，这种攻击的风险尤为严重。由于模型文件存储在本地，攻击者可以通过逆向工程、内存转储（MemoryDumping）等技术手段，直接从设备中提取出模型的二进制文件。虽然模型文件通常会进行加密或混淆处理，但攻击者仍有可能通过分析、破解来获

取模型的明文参数。此外，攻击者还可以通过构造大量的查询输入，观察模型的输出结果，然后利用这些输入－输出对来训练一个“替代模型”（SurrogateModel）。这个替代模型可以在功能上高度逼近原始的目标模型，从而实现对模型的窃取。模型一旦被窃取，攻击者就可以将其用于商业目的，或者利用窃取的模型来寻找其漏洞，发起更进一步的攻击。

2. 模型篡改：对模型进行恶意修改，植入后门或破坏其功能

模型篡改（Model Tampering）是指攻击者在获取模型后，对其参数或结构进行恶意的修改，以达到特定的攻击目的。最常见的模型篡改攻击是后门攻击（Backdoor Attack）。攻击者通过在模型中植入一个“后门”，使得模型在正常输入下表现正常，但在遇到带有特定“触发器”（如一个特殊的图案、一个特定的词语）的输入时，就会执行预设的恶意行为。例如，在人脸识别系统中植入后门，使得佩戴特定眼镜的人可以被识别为任意一个已注册的用户，从而绕过身份验证。在端侧场景下，模型篡改的风险主要来自于对存储在设备上的模型文件的直接修改。如果模型文件没有进行完整性校验或加密存储，攻击者就可以通过获取设备的 root 权限，直接修改模型文件，植入后门或破坏其功能。例如，攻击者可以修改一个智能音箱的唤醒词检测模型，使其在听到任何包含特定词语的句子时都被唤醒，从而实现持续的监听。这种攻击的危害性在于，被篡改的模型在外观上与正常模型无异，很难被用户或系统管理员发现。

3. 模型滥用：利用窃取或篡改的模型进行非法活动

模型滥用是指攻击者利用窃取或篡改的模型，进行各种非法活动。例如，攻击者可以利用窃取的金融风控模型，分析出其决策逻辑，从而找到规避风控的方法，进行信用卡诈骗、洗钱等犯罪活动。或者，攻击者可以利用篡改的自动驾驶模型，制造交通事故，进行敲诈勒索。此外，模型滥用还包括利用模型生成虚假信息、进行网络诈骗等。例如，攻击者可以利用窃取的语言模

型，生成大量逼真的钓鱼邮件或虚假新闻，进行社会工程学攻击。随着生成式 AI 技术的发展，模型滥用的形式将越来越多样化，其社会危害性也将越来越大。因此，必须从技术、法律、管理等多个层面，加强对模型的保护，防止其被滥用。

（三）算法安全风险

1. 算法偏见：模型训练数据存在偏见，导致输出结果不公平

算法偏见（Algorithmic Bias）是指由于训练数据中存在的历史偏见或采样不均衡，导致模型在决策时对特定群体产生系统性的、不公平的对待。例如，如果一个用于招聘筛选的模型，其训练数据主要来自于男性员工，那么该模型在评估女性求职者时，可能会给出较低的分数，即使她们的资历和能力与男性求职者相当。这种偏见不仅会固化甚至放大社会中原有的歧视，还会对个人的就业、信贷、司法等权益造成实质性的损害。在端侧大模型应用中，算法偏见同样是一个严重的问题。例如，一个用于个性化推荐的模型，可能会因为用户过去的浏览行为，而不断向其推荐相似的内容，从而将用户困在“信息茧房”中，限制其获取多元化信息的机会。或者，一个用于智能安防的模型，可能会因为训练数据中某个种族的人脸图像较少，而在识别该种族的人脸时准确率较低，导致误报或漏报。因此，在模型开发和部署过程中，必须对训练数据进行严格的审查和清洗，采用公平性增强算法，并对模型的输出结果进行持续的监控和评估，以发现和纠正潜在的偏见。

2. 算法对抗攻击：通过构造对抗样本，欺骗模型做出错误判断

对抗攻击（Adversarial Attack）是一种针对机器学习模型的攻击方式。攻击者通过在正常输入数据上添加人眼难以察觉的微小扰动，构造出所谓的“对抗样本”，从而欺骗模型做出错误的判断。例如，在一张熊猫的图片上添加一些精心计算的噪声，人眼看起来仍然是一张熊猫的图片，但图像分类模型却可能将其错误地识别为长臂猿。对抗攻击对端侧大模型的安全构成了严重威胁，尤其是在自动驾驶、医疗诊断、工业质检等高风险领域。例

如，在自动驾驶场景中，攻击者可以在交通标志上粘贴一些微小的贴纸，就可能使车辆的识别系统将其误判，从而引发致命的事故。在工业质检中，攻击者可以制造出带有微小瑕疵的“对抗性”产品，使其能够逃过 AI 质检系统的检测，从而将不合格产品流入市场。由于对抗样本的生成需要对模型的内部结构有一定了解，因此模型窃取攻击的成功往往会加剧对抗攻击的风险。

3. 算法可解释性不足：模型决策过程不透明，难以追溯和审计

端侧大模型，特别是深度学习模型，通常被视为“黑箱”，其决策过程缺乏透明度和可解释性。这意味着即使是模型的开发者，也很难完全理解模型为什么会做出某个特定的决策。这种可解释性的不足，给端侧大模型的安全治理带来了巨大的挑战。首先，当模型出现错误或产生不公平的结果时，由于无法追溯其决策依据，很难进行有效的调试和修复。其次，在涉及个人权益或社会公共利益的关键领域（如金融、司法、医疗），如果模型的决策过程不透明，就难以对其进行有效的审计和监督，也无法向用户或监管机构提供合理的解释。例如，如果一个端侧信贷审批模型拒绝了用户的贷款申请，但无法给出明确的理由，这不仅损害了用户的知情权，也可能引发法律纠纷。因此，提升算法的可解释性，是保障端侧大模型安全、公平、可信应用的重要前提。目前，研究人员正在积极探索各种可解释性 AI（XAI）技术，如 LIME、SHAP 等，试图打开模型的“黑箱”，使其决策过程更加透明和可理解。

四、端侧大模型安全治理体系构建

（一）法律法规与政策监管

1. 完善相关法律法规：细化数据安全、隐私保护、算法伦理等方面的法律条文

随着端侧大模型技术的快速发展和广泛应用，现有的法律法规体系面临着新的挑战。为了有效应对端侧大模型带来的安全风险，必须加快完善相关法律法规，特别是在数据安全、隐私保护和算法伦理等关键领域。首先，在数据安全方面，虽然《数据安全法》和《网络数据

安全管理条例》等法律法规已经确立了基本框架，但针对端侧大模型的特殊性，仍需进一步细化。例如，应明确端侧数据与云端数据的界定标准，以及在不同处理场景下的安全责任划分。需要出台专门的端侧数据安全标准，对数据的采集、存储、处理、传输和销毁等全生命周期环节提出具体的安全技术要求和管理规范。特别是对于跨应用、跨设备的数据融合与流转，应建立更为严格的审批和监管机制，防止数据被滥用。

在隐私保护方面，《个人信息保护法》为端侧大模型的治理提供了重要的法律依据。然而，面对端侧大模型带来的新挑战，法律条文需要更具前瞻性。例如，针对智能体调用“无障碍权限”等高敏感权限的行为，应明确其合法性边界，并要求服务提供商必须获得用户的“单独同意”，且不能以“一揽子授权”的方式获取。此外，随着个人记忆模型等新技术的发展，用户的个人数据将被高度集中化，这涉及到更为复杂的人格权和财产权问题。法律需要对这些新型数据处理活动的法律定性、权益归属以及监管框架进行明确界定，确保用户的合法权益不受侵害。同时，应借鉴欧盟《通用数据保护条例》（GDPR）等国际经验，引入“数据保护影响评估”（DPIA）等机制，要求企业在开发和部署端侧大模型应用前，必须对其可能带来的隐私风险进行全面评估。

在算法伦理方面，端侧大模型的决策过程往往缺乏透明度和可解释性，可能导致算法偏见、歧视等问题。因此，需要制定专门的算法伦理规范，明确算法设计、训练和应用的基本原则。例如，应要求算法必须公平、无歧视，并建立算法透明度和可解释性的基本要求。对于可能影响个人权益或社会公共利益的重大算法应用，应建立备案和审查制度。此外，还应鼓励和支持行业协会、学术机构等社会力量参与算法伦理的研究和治理，形成多方协同的治理格局。通过完善法律法规，为端侧大模型的健康发展提供坚实的制度保障，确保技术创新始终在安全、合规、可信的轨道上进行。

2. 明确责任划分：厘清设备制造商、模型提供商、应用开发者、用户等各方的权利与义务

在端侧大模型的复杂生态系统中，涉及设备制造商、模型提供商、应用开发者、云服务提供商以及最终用户等多个参与方。这种多方参与的格局使得安全责任的界定变得异常复杂和困难。因此，构建一个清晰、合理的责任划分体系，是端侧大模型安全治理的核心环节。首先，设备制造商作为硬件平台的提供者和操作系统规则的制定者，应承担基础性的安全保障责任。他们需要确保其生产的设备具备基本的安全防护能力，如安全启动、固件完整性校验等。同时，对于在其平台上运行的端侧大模型应用，设备制造商应建立严格的审核和准入机制，确保应用符合安全标准，并对高敏感权限的调用进行有效监管。

模型提供商，即开发和训练大模型的企业，应对模型的安全性和可靠性负责。他们需要确保训练数据的质量，避免因数据偏见导致模型产生歧视性或不公平的输出。同时，应采取技术手段防止模型被恶意攻击，如模型窃取、对抗性攻击等。在模型部署后，模型提供商应持续监控模型的运行状态，及时发现并修复潜在的安全漏洞。对于因模型自身缺陷导致的安全问题，模型提供商应承担相应的法律责任。

应用开发者作为直接面向用户的服务提供者，其责任最为直接和具体。他们需要严格遵守数据最小化原则，只收集实现功能所必需的用户数据，并明确告知用户数据收集的目的、范围和使用方式。在申请系统权限时，必须遵循合法、正当、必要的原则，并获得用户的明确授权。对于涉及用户敏感信息的操作，应采取额外的安全措施，如加密存储、访问控制等。如果应用集成了第三方的大模型服务，应用开发者还应对第三方服务的安全性进行尽职调查，并明确告知用户数据共享的情况。

云服务提供商在端云协同的架构中扮演着重要角色，他们负责处理和存储从端侧上传的数据。云服务提供商必须确保其云平台的安全性，采取强大的加密和访问控制措施，防止用户数据在云端被泄露或

滥用。同时，应遵守数据本地化存储和跨境传输的相关法律法规。最后，用户作为数据的产生者和服务的最终使用者，也负有一定的责任。用户应提高自身的安全意识，仔细阅读隐私政策，谨慎授予应用权限，并定期更新设备和应用，以防范已知的安全风险。通过明确各方的权利与义务，可以形成一个权责清晰、协同共治的安全责任体系，为端侧大模型的健康发展提供有力保障。

3. 加强执法与监管：建立有效的监管机制，加大对违法行为的惩处力度

完善的法律法规需要强有力的执法与监管才能落到实处。针对端侧大模型这一新兴领域，必须建立一套行之有效的监管机制，并加大对违法行为的惩处力度，以形成足够的法律威慑。首先，应明确监管主体，建立跨部门的协同监管机制。由于端侧大模型涉及数据安全、个人信息保护、网络安全、市场竞争等多个方面，需要多个部门协同作战，形成监管合力。可以成立专门的 AI 治理机构或专家委员会，负责统筹协调各相关部门的监管工作，并为政策制定和执法提供专业的技术支持。

其次，应创新监管方式和手段。传统的监管模式可能难以适应端侧大模型技术快速迭代的特点。因此，需要探索更加灵活、高效的监管方式。例如，可以建立“监管沙盒”机制，允许企业在受控的环境下测试其创新的端侧大模型应用，监管部门则可以在此过程中深入了解技术特点，评估潜在风险，并制定相应的监管规则。同时，应充分利用技术手段进行监管，例如，开发自动化的合规检测工具，对市场上的端侧大模型应用进行持续监测，及时发现和处理违规行为。此外，还应建立畅通的公众举报和投诉渠道，鼓励社会力量参与监督，形成全社会共同治理的良好氛围。

最后，必须加大对违法行为的惩处力度。对于违反数据安全、个人信息保护等法律法规的行为，应依法予以严厉处罚，提高违法成本。处罚措施不仅应包括高额罚款，还应包括责令停业整顿、吊销相关业务许可等。对于造成严重后果的，还应追究相关责任人的刑事责任。

通过严厉的执法，向市场传递出明确的信号：任何以牺牲用户安全和隐私为代价的技术创新都是不可接受的。同时，监管部门应定期公布典型案例，以案释法，起到警示和教育作用，引导企业自觉遵守法律法规，共同维护端侧大模型产业的健康、有序发展。

（二）行业标准与自律机制

1. 制定统一的安全标准：推动端侧大模型在设计、开发、部署等环节的安全标准化

在端侧大模型产业快速发展的初期，制定统一的安全标准对于规范市场秩序、提升整体安全水平至关重要。目前，由于缺乏统一的标准，不同厂商的产品在安全防护能力上参差不齐，给用户带来了潜在的风险。因此，需要由行业协会、标准化组织牵头，联合产业链上下游的龙头企业、科研机构和专家，共同制定一套覆盖端侧大模型全生命周期的安全标准体系。这套标准体系应至少包括以下几个方面：首先，在设计与开发阶段，应明确模型安全、数据安全和算法伦理的基本要求。例如，应要求开发者在设计模型时，必须考虑对抗性攻击的防御能力；在处理数据时，必须遵循数据最小化和匿名化原则；在设计算法时，必须避免引入偏见和歧视。

其次，在部署与运行阶段，应制定详细的安全配置和运维规范。例如，应规定端侧设备必须采用安全的通信协议，对存储的数据进行加密，并建立完善的访问控制机制。对于端云协同的架构，应明确数据传输的安全要求和云端数据的安全责任。最后，在测试与评估阶段，应建立一套标准化的安全测试方法和评估指标。这包括对模型的鲁棒性、公平性、可解释性进行测试，对数据处理的合规性进行审计，以及对整个系统的安全性进行综合评估。通过制定和推广统一的安全标准，可以为端侧大模型的开发和应用提供明确的指引，推动整个行业向着更加安全、规范的方向发展。

2. 建立行业自律组织：鼓励行业协会、联盟等组织制定自律公约，推动企业自我约束

除了政府监管和法律法规，行业自律也是端侧大模型安全治理体

系中不可或缺的一环。鼓励和支持行业协会、产业联盟等社会组织，在政府的指导下，制定行业自律公约和行为准则，是推动企业自我约束、实现协同治理的有效途径。行业自律组织可以发挥其在信息沟通、利益协调、专业指导等方面的优势，凝聚行业共识，共同应对安全挑战。例如，自律组织可以组织企业共同研究和制定端侧大模型的安全最佳实践，分享安全威胁情报和防护经验，共同抵御网络攻击。

此外，行业自律组织还可以建立内部的监督和惩戒机制。对于违反自律公约、损害用户权益的企业，自律组织可以进行内部通报、公开谴责，甚至将其移出组织，从而形成一种有效的行业约束。这种基于声誉和信誉的约束机制，可以与政府的法律监管形成互补，共同维护市场秩序。同时，行业自律组织还可以作为政府与企业之间的桥梁，向政府反映行业发展的诉求和困难，为政策的制定和调整提供参考，促进政策的科学性和有效性。通过建立行业自律组织，可以充分调动企业的积极性和主动性，形成政府、企业、社会多方协同共治的良好局面。

3. 推动安全认证与评估：建立第三方安全认证和评估体系，提升产品和服务的安全可信度

为了向市场和用户证明端侧大模型产品和服务的安全性，建立独立、公正、权威的第三方安全认证和评估体系至关重要。第三方认证和评估机构，由于其独立性和专业性，其出具的认证报告和评估结果更具公信力，能够为消费者选择安全可靠的产品提供重要参考，也能够激励企业不断提升自身的安全水平。安全认证和评估体系应覆盖端侧大模型的各个方面，包括硬件安全、软件安全、数据安全、算法安全和隐私保护等。

在认证和评估过程中，应采用科学、严谨的方法和标准。例如，可以通过代码审计、渗透测试、漏洞扫描等技术手段，对产品的安全性进行全面的检测。可以通过对训练数据和模型输出的分析，评估算法的公平性和无偏见性。可以通过对隐私政策和数据处理流程的审查，评估其隐私保护的合规性。对

于通过认证和评估的产品，可以授予相应的安全标识，如“安全认证标志”、“隐私保护认证标志”等，方便消费者识别。同时，认证和评估的结果应向社会公开，接受公众监督。通过推动安全认证与评估，可以建立起一套市场化的安全信任机制，引导资源向更安全、更可信的产品和服务倾斜，从而推动整个端侧大模型产业的健康、可持续发展。

（三）企业内部安全管理

1. 建立安全管理制度：制定完善的数据安全、模型安全、算法安全管理制度

企业作为端侧大模型开发和应用的主体，是保障其安全的第一道防线。因此，建立一套完善的内部安全管理制度至关重要。这套制度应覆盖数据安全、模型安全和算法安全等各个方面，并贯穿于产品设计、开发、测试、部署、运维的全生命周期。在数据安全方面，企业应制定明确的数据分类分级标准，对不同级别的数据采取不同的安全保护措施。应建立严格的数据采集、存储、使用、共享和销毁流程，确保数据处理活动的合规性。在模型安全方面，企业应建立模型安全开发规范，在模型设计阶段就考虑对抗性攻击、后门攻击等安全威胁的防御。应建立模型安全测试流程，对模型的鲁棒性、安全性进行充分的验证。在算法安全方面，企业应建立算法伦理审查机制，对算法的公平性、透明度、可解释性进行评估，防止算法偏见和歧视。此外，企业还应建立定期的安全风险评估和审计制度，及时发现和修复安全漏洞，持续改进安全管理体系。

2. 加强员工安全培训：提升员工的安全意识和技能，防范内部安全风险

员工是企业安全管理制度的执行者，也是防范内部安全风险的关键。因此，加强对员工的安全培训，提升其安全意识和技能，是企业内部安全管理的重要组成部分。安全培训应面向所有员工，特别是与端侧大模型开发、运维、管理相关的技术人员和管理人员。培训内容应涵盖数据安全、网络安全、算法伦理、法律法规等多个方面。例如，应向开发人员普及安全编码规范，

防止在代码中引入安全漏洞；应向运维人员培训安全配置和应急响应的知识，提升其应对安全事件的能力；应向管理人员介绍数据保护法规的要求，强化其合规意识。培训形式可以多样化，包括线上课程、线下讲座、实战演练等。通过持续、有效的安全培训，可以在企业内部营造一种“人人讲安全、事事为安全”的文化氛围，从源头上防范和化解安全风险。

3. 建立安全应急响应机制：制定应急预案，及时响应和处置安全事件

尽管企业采取了各种预防措施，但安全事件仍有可能发生。因此，建立一套高效的安全应急响应机制，是企业在发生安全事件时，最大限度地减少损失、恢复业务、维护声誉的重要保障。应急响应机制应包括以下几个关键环节：首先是应急预案的制定。企业应根据自身的业务特点和安全风险，制定详细的应急预案，明确不同类型安全事件的处置流程、责任分工、沟通机制等。其次是应急响应团队的建设。企业应组建一支专业的应急响应团队，负责在安全事件发生时，进行快速的技术分析、溯源、处置和恢复。再次是应急演练的开展。企业应定期组织应急演练，模拟真实的安全事件场景，检验应急预案的有效性和团队的响应能力，并根据演练结果不断优化预案。最后是事后总结与改进。在安全事件处置完毕后，企业应进行全面的复盘和总结，分析事件的原因、处置过程中的得失，并据此改进安全管理制度和技术防护措施，防止类似事件再次发生。

（四）用户教育与权益保障

1. 提升用户安全意识：通过多种渠道向用户普及端侧大模型的安全风险和防护知识

用户是端侧大模型服务的最终使用者，也是自身数据安全的第一责任人。因此，提升用户的安全意识，使其能够主动识别和防范安全风险，是端侧大模型安全治理体系中不可或缺的一环。政府、企业、行业协会、媒体等各方应共同努力，通过多种渠道，向用户普及端侧大模型的安全风险和防护知识。例如，可以通过官方网站、社交媒体、新

闻报道等方式，向公众宣传端侧大模型可能存在的安全风险，如数据泄露、隐私侵犯、算法偏见等。可以在产品说明书、用户协议、隐私政策中，用通俗易懂的语言，向用户解释产品的安全功能和用户需要注意的安全事项。可以组织线上线下的安全知识讲座、培训活动，向用户传授如何设置强密码、如何管理应用权限、如何识别和防范网络诈骗等实用技能。通过持续、广泛的用户教育，可以有效提升用户的安全素养，使其在享受端侧大模型带来便利的同时，能够更好地保护自己的合法权益。

2. 保障用户知情权和控制权：提供清晰易懂的隐私政策，赋予用户对个人数据的控制权

保障用户的知情权和控制权，是数据保护法规的核心要求，也是建立用户信任的基础。企业应以清晰、易懂、非技术性的语言，向用户提供其隐私政策，明确告知用户其个人数据的收集、使用、存储、共享等情况，特别是涉及高敏感权限调用和跨应用数据流转的场景，必须获得用户的明确同意。应避免使用冗长、晦涩的法律术语，让用户能够真正理解其数据是如何被处理的。同时，企业应提供便捷、有效的工具，赋予用户对其个人数据的控制权。例如，应允许用户随时查询、更正、删除其个人数据，应提供一键关闭个性化推荐、限制数据收集范围等功能。应建立便捷的用户反馈和投诉渠道，及时响应用户的关切和诉求。通过切实保障用户的知情权和控制权，可以有效增强用户对端侧大模型服务的信任感，促进产业的健康发展。

3. 建立用户投诉和维权渠道：为用户提供便捷的投诉和维权途径，保护用户合法权益

当用户的合法权益受到侵害时，为其提供便捷、高效的投诉和维权渠道，是保护用户权益的最后一道防线。企业应建立多渠道的用户投诉受理机制，如客服热线、在线客服、电子邮件等，并确保这些渠道畅通、有效。应建立规范的用户投诉处理流程，对用户的投诉进行及时登记、调查、处理和反馈。对于涉及数据安全、隐私泄露等严重问题的投诉，应启动专门的应急

响应机制，进行严肃处理。政府监管部门应加强对用户投诉的受理和处理，建立统一的投诉举报平台，方便用户进行投诉。对于查证属实的违法行为，应依法予以严厉处罚，并支持用户通过法律途径维护自身权益。通过建立完善的用户投诉和维权渠道，可以有效震慑违法行为，保护用户的合法权益，维护社会公平正义。

五、结论与展望

（一）结论

1. 端侧大模型安全风险具有复杂性和多样性

本报告通过对智能手机个人助理、智能家居语音控制中枢、工业物联网边缘计算节点等典型应用场景的深入分析，以及对数据、模型、算法三个层面的风险剖析，得出结论：端侧大模型所面临的安全风险具有高度的复杂性和多样性。这种复杂性不仅体现在风险来源的多元化，涉及技术、管理、法律、伦理等多个维度，还体现在风险链条的延长和交织。端云协同的混合架构使得数据在终端、边缘、云端之间流转，责任边界变得模糊。同时，数据安全、模型安全和算法安全风险相互关联、相互影响，例如，数据投毒攻击（数据安全）可能导致模型被植入后门（模型安全），进而引发算法偏见（算法安全）。这种复杂性和多样性对传统的安全防护理念和治理模式提出了严峻的挑战。

2. 管理层面治理是保障端侧大模型安全的关键

面对端侧大模型复杂多样的安全风险，单纯依靠技术手段难以实现全面有效的防护。本报告认为，加强管理层面的治理是保障端侧大模型安全的关键。这包括完善法律法规，明确各方责任，加强执法监管；制定统一的行业标准，建立行业自律机制；强化企业内部安全管理，落实安全主体责任；以及提升用户安全意识，保障用户合法权益。通过构建一个多层次、全方位的管理治理体系，可以从源头上防范风险，在过程中控制风险，在事后有效应对风险，从而为端侧大模型的健康发展提供坚实的制度保障。管理治理与技术防护相辅相成，共同构成了端侧大模型安全的坚固防

线。

3. 构建多方协同的治理体系是未来的发展方向

端侧大模型的生态系统涉及设备制造商、模型提供商、应用开发者、云服务提供商、用户、政府、行业协会等多个参与方。任何一个环节的疏漏都可能导致整个系统的安全防线被攻破。因此，本报告强调，构建一个多方协同、共建共治共享的治理体系是未来的发展方向。这需要各方明确自身的角色和责任，加强沟通与协作。政府应发挥引导和监管作用，企业应落实主体责任，行业协会应推动自律，用户应积极参与。通过建立一个权责清晰、运转高效、响应迅速的协同治理机制，可以形成强大的治理合力，共同应对端侧大模型带来的安全挑战，推动人工智能技术在安全、可信、合规的轨道上持续创新和发展。

（二）展望

1. 技术层面：隐私计算、可信执行环境等技术的应用将更加广泛

为了应对端侧大模型带来的数据隐私和安全挑战，未来在技术层面，隐私计算（Privacy-PreservingComputation）和可信执行环境（TrustedExecutionEnvironment,TEE）等技术的应用将更加广泛。隐私计算技术，如联邦学习、多方安全计算、差分隐私等，能够在不暴露原始数据的前提下，实现数据的联合分析和模型的协同训练，从而在保证数据“可用不可见”的基础上，提升模型的性能和智能化水平。可信执行环境则通过硬件隔离技术，为端侧大模型的运行提供一个安全的“黑箱”，确保模型和数据在处理过程中的机密性和完整性，防止被恶意软件或攻击者窃取或篡改。这些新兴技术的应用，将为端侧大模型的安全运行提供更加坚实的技术保障，使其在保护用户隐私和数据安全方面发挥更大的作用。

2. 管理层面：法律法规和行业标准将不断完善

随着端侧大模型技术的不断成熟和应用场景的不断拓展，未来在管理层面，相关的法律法规和行业标准将不断完善。各国政府和监管机构将更加重视对人工智能的治

理，出台更具针对性和可操作性的法律法规，特别是在数据安全、算法伦理、责任认定等方面。同时，行业协会和标准化组织将加快制定和完善端侧大模型的安全标准和技术规范，覆盖从设计、开发到部署、运维的全生命周期。这些法律法规和行业标准的完善，将为端侧大模型的安全治理提供更加明确的指引和依据，推动整个行业向着更加规范、健康的方向发展。

3. 产业层面：安全将成为端侧大模型产业的核心竞争力

在未来，随着用户对数据安全和隐私保护意识的不断提高，以及监管要求的日趋严格，安全将不再仅仅是端侧大模型产品的“加分项”，而是其必须具备的“基本项”。能够提供更安全、更可信的产品和服务的企业，将在市场竞争中获得更大的优势。因此，安全将成为端侧大模型产业的核心竞争力之一。企业将不得不加大在安全技术研发、安全管理体系建设、安全人才培养等方面的投入，将安全理念贯穿于产品设计和开发的每一个环节。同时，能够提供权威安全认证和评估服务的第三方机构，也将在产业生态中扮演越来越重要的角色。安全将成为驱动端侧大模型产业技术创新和模式变革的重要力量，引领产业向着更高质量、更可持续的方向发展。

（作者：唐琦　姚钦洲　李雨佳　彭健）

联系人电话：18811215507

思想，还是思想，才使我们与众不同
研究，还是研究，才使我们见微知著

新型工业化研究所（工业和信息化部新型工业化研究中心）
政策法规研究所（工业和信息化法律服务中心）
规划研究所
产业政策研究所（先进制造业研究中心）
科技与标准研究所
知识产权研究所
工业经济研究所（工业和信息化经济运行研究中心）
中小企业研究所
节能与环保研究所（工业和信息化碳达峰碳中和研究中心）
安全产业研究所
材料工业研究所
消费品工业研究所
军民融合研究所
电子信息研究所
集成电路研究所
信息化与软件产业研究所
网络安全研究所
无线电管理研究所（未来产业研究中心）
世界工业研究所（国际合作研究中心）

通讯地址：北京市海淀区万寿路27号院8号楼1201 邮政编码：100846
联系人：王 乐 联系电话：010-68200552 13701083941
传 真：010-68209616 电子邮件：wangle@ccidgroup.com

赛迪研究院

《产业政策与法规研究》编辑部
编 辑 部：赛迪研究院
通讯地址：北京市海淀区万寿路27号院8号楼12层
邮政编码：100846
联 系 人：王 乐
联系电话：010-68200552 13701083941
传 真：0086-10-68209616
网 址：www.ccidthinktank.com
电子邮件：wangle@ccidgroup.com