《北京市生成式人工智能大模型产业发展与治理白皮书》概况

在全球人工智能治理与竞争格局加速重构、技术演进范式深刻转型的背景下，人工智能竞争正从“算力规模扩张”向“认知执行一体化能力”跃迁，各主要经济体形成差异化发展路径。《白皮书》坚持“使用人工智能治理人工智能”，构建形成“安全成熟度”指标体系，助力大模型企业行稳致远发展。在过去两年对北京市生成式人工智能大模型 5000 余次技术测评的基础上，从已备案 216 款大模型中随机抽取 100 款开展专项技术检测与对比分析，以“安全成熟度”指标体系，洞察大模型产业创新发展的最新变化，展望未来发展趋势。《白皮书》共六部分，涵盖发展态势分析、安全风险挑战、测评方法、结果分析、产业发展、对策展望，构建起完整的逻辑体系。

一、主要观点

（一）全球人工智能正在由“单体模型”向“人—模型—基础设施协同的复合智能体”演进

从“认知工具”走向“智能体操作系统”。通过模型对抗发现，大模型产业架构正从传统“应用中心”模式，向以“智能体”为核心的新架构转型，全球人工智能大模型体系也从“函数逼近能力竞争”转向“认知系统竞争”，大模型逐步向具备自主任务规划、工具调用能力的“智能体操作系统”演化。全球最大 AI 模型聚合平台 OpenRouter 数据显示，2 月中国 AI 模型调用量已超越美国，全球调用量前十企业中北京企业占据半壁江山，充分彰显了北京在全球大模型产业发展中的核心地位。

北京从模型研发高地向全球系统级生态中心演进。截至 2025 年底，北京市累计备案大模型达 212 款，占全国总量近 30%，位居全国首位；自研模型占比达 45.3%，显示出较强的底层技术自主创新能力。从能力结构看，北京大模型体系仍以文生文模型为主体，占比 79.2%，同时多模态模型、端侧模型与智能体快速增长，人工智能能力正由信

息生成阶段向认知执行阶段加速演进。在关键技术路径上，北京在MoE稀疏架构、长文本推理、端侧部署及智能体系统等方向持续突破，头部主体在技术攻坚、商业变现、生态布局上各展优势，产业体系正由“模型研发高地”向“智能体系统级生态中心”转变。

（二）大模型安全能力正由“合规驱动安全”向“能力驱动安全”的关键转型

测评显示，受测大模型无效响应率达23.82%，且大模型以防御性拒答为主要响应形态，拒答占比达46.95%，这说明对于显性的违规指令，防御性拒答已经成为大模型的标准配置，部分模型通过扩大拒答边界降低风险暴露概率，形成“低风险—低可用”的伪安全结构。该模式虽可在短期内降低风险指标，但显著削弱模型在复杂语境下的有效服务能力，其本质反映出安全优化目标由“风险约束下的效用最大化”偏移为“效用约束下的风险最小化”。这一现象表明，北京市大模型产业正处于由“合规驱动安全”向“能力驱动安全”的关键转型阶段。未来安全体系将由外部附加模块转变为模型内生能力组成部分，实现安全与可用性的协同优化。

（三）人工智能风险正在由“内容违规风险”向“认知越权风险”迁移，治理对象正在由内容转向行为

测评发现，当前人工智能风险形态正在发生结构性变化，攻击方式由直接违规指令逐步转向基于身份模拟、语境诱导及多轮交互的复杂认知攻击。这类攻击通过构建合理社会语境绕过模型显性规则，使模型在缺乏权限感知能力的情况下执行隐含违规逻辑，其本质属于认知越权风险。随着智能体逐步具备任务执行能力，风险边界将由信息生成扩展至行为执行层面，影响范围由数字空间延伸至现实物理环境，人工智能治理对象将由内容输出转向模型的认知过程与行为决策机制。

（四）端侧智能体正在形成新的安全治理盲区，可能成为未来系统性风险的重要来源

随着人工智能能力加速向智能终端、自动驾驶系统及具身智能设备扩展，端侧模型正在成为人工智能体系的重要组成部分。测评显示，

端侧模型在高强度攻击环境下的风险暴露率高于云端闭源模型，其主要原因在于端侧部署受限于算力与能耗约束，安全对齐模块在模型压缩过程中往往被弱化。端侧智能体具备脱离云端独立运行能力，其行为过程难以实时监测，随着部署规模扩大，可能形成分布式系统性风险。未来端侧智能体安全治理将成为人工智能安全体系建设的重要方向。

（五）“系统防护型”代表产业演进目标，大模型正处于“从合规防御向韧性抗压”转化的过渡期

聚类分析显示，当前大模型呈现四类典型安全成熟度结构。其中，“系统防护型”模型在 L4 级攻击条件下仍保持较低风险暴露率，兼具高安全性与高可用性，代表未来发展方向；而“强拒答型”模型极易出现“拒答泛化”存在可用性不足问题；“高强度失守型”模型在面对高强度的认知博弈时，模型无法维持其安全边界的一致性；“高失败伪安全型”模型无效响应率过高，亟须优化核心推理能力、系统架构稳定性和工具链可靠性。综合评估结果表明，当前大模型安全体系仍停留在以合规满足为目标的防护阶段，安全水位尚未达到应对自动化对抗与复杂攻击场景的成熟水平，当攻击强度由基础诱导升级至 AI 自动化对抗阶段时，风险暴露率上升至 24.28%。面向未来，“系统防护型”将成为安全能力演进的前沿方向，其特征在于通过全生命周期风险控制机制，实现可用性与安全性的动态平衡，而非单纯依赖输出端的内容拦截。

二、未来展望

（一）产业深化：从“大模型问询”向“智能体协作”演进

一是复合智能体体系逐渐形成。大模型正逐步嵌入操作系统、终端设备及云端计算环境，形成具备任务规划、自主决策与工具调用能力的复合智能体体系，推动信息技术架构由“应用中心模式”向“智能体中心模式”转变。**二是迈向 AIOS 核心基础设施**。在智能体体系演进过程中，智能体操作系统（AIOS）或将成为承载智能体运行、资源调度及行为管理的核心基础设施，实现模型能力、计算资源及物理终端的统一调度与协同运行，为智能体规模化、体系化落地提供关

键支撑。**三是网络身份标识升级**。IPv6 广覆盖、唯一性、可追溯的优势将至关重要，MCP 等新型协议、接口规范与治理规则也将愈发关键。我们判断，IPv6 将成为未来智能体的重要身份标识，传统的“信息消费”将会向“AI 消费”快速转变。

（二）治理转型：从“内容护栏式防控”向“社会影响导向的体系化治理”转型

一是风险应对方式升级。未来安全防护对象将由传统软件系统扩展至自主智能体，安全体系由静态防御机制向具备自适应能力的动态免疫体系演进。应构建全生命周期的“攻、检、防”一体化安全架构，并基于“AI 对抗 AI”的自动化对抗方式持续强化模型对复杂攻击的适应能力，实现风险应对从被动防御、静态过滤向主动防控、动态适配、安全免疫的全面升级。**二是安全产业形态变化**。安全体系将打破传统网络安全服务模式的局限，升级为面向智能体系统的新型安全基础设施，人工智能安全不再是单纯的“防护工具体系”，而是演进为支撑智能体系统稳定运行、保障产业健康发展的关键基础能力，并逐步融入人工智能产业体系，成为不可或缺的重要组成部分。**三是治理目标全面拓展**。治理转型的核心导向是打破以往聚焦单点内容合规的“护栏式防控”局限，转向覆盖模型全生命周期、兼顾技术安全与社会影响的体系化治理。既保障人工智能技术的安全、合规、可控运行，也防范其可能产生的各类社会风险，实现技术发展与社会治理的协同推进。

（三）公平重塑：从“效率提升工具”向“社会结构性公平调节机制”转型

一是新的数字鸿沟开始形成。随着语音交互、多模态感知、智能代理等技术普及，传统依赖文字输入能力形成的数字鸿沟逐步消解，但基于感官、行动、认知差异的新型数字鸿沟正在显现。不同群体，尤其是残障群体在信息获取、设备使用、场景适配等方面面临的新差距，已成为数字社会公平发展中亟待解决的突出问题。**二是人工智能从效率工具发展为社会公平基础设施**。人工智能不再只是提升生产效率的工具，更将成为社会结构性公平调节机制与重要公共基础设施。

依托语音交互、视觉识别、行为理解、脑机接口等技术，可显著提升“语残、失聪”等群体参与数字社会的能力，以技术普惠弥合新型数字鸿沟。**三是未来人工智能治理的方向应聚焦普惠与公平。**人工智能治理需将公平性、可及性、包容性纳入基础目标，推动技术安全与社会公平深度协同。在治理实践中，以政府引导、产业协同为路径，构建高质量、多元包容、兼顾弱势群体诉求的价值对齐语料库与安全语料体系，在模型预训练与微调阶段嵌入公平导向，推动人工智能安全从“外部防护”转向“内生安全”，保障技术普惠性，实现安全可控与社会公平的统一。

（四）区域协同：京津冀成为关键算力与治理新枢纽

北京拥有高密度的算法、人才与治理制度资源，而河北等地在能源供给、物理空间与算力承载上具备显著比较优势。未来应深化 AI for 京津冀，着力构建跨区域的智能体协同网络，打破行政区划壁垒，实现北京的“智力资源”与周边的“算力与能源资源”高效耦合。这不仅能解决“算力与智能分布不均”的结构性矛盾，更将推动京津冀地区从传统的“地理协同”升级为深度的“智能协同”，打造支撑中国 AI 规模化发展的战略支点。