CAICT 中国信通院

集智 专题报告

数据治理研究报告

——端侧大模型数据治理法律要点研究

（2025 年）

中国信息通信研究院政策与经济研究所

中国信息通信研究院互联网法律研究中心

2025年12月

版权声明

前　言

当前，数据作为新型生产要素，是数字化、网络化、智能化的基础，已快速融入生产、分配、流通、消费和社会服务管理等各个环节，深刻改变着生产方式、生活方式和社会治理方式。党中央高度重视通过数据基础制度建设释放数据要素价值。2025 年 10 月 23 日，党的第二十届中央委员会第四次全体会议通过的《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》中明确提出，“健全数据要素基础制度，建设开放共享安全的全国一体化数据市场，深化数据资源开发利用。促进实体经济和数字经济深度融合，实施工业互联网创新发展工程。加快人工智能等数智技术创新，突破基础理论和核心技术，强化算力、算法、数据等高效供给”。

今年以来，随着人工智能技术加速向智能终端设备深度渗透，端侧大模型正以“算力前移、数据本地、场景深耕”的颠覆性变革重塑智能产业格局，形成覆盖智能手机、智能汽车、可穿戴设备等万亿级智能终端市场的产业生态。这场技术革命不仅催生出个性化语音助手、实时健康监测、本地化决策支持等创新应用，更通过“数据优先本地处理”的特性重构了隐私计算范式。然而，技术双刃剑效应在端侧场景有所表现。在数据保护层面，端侧大模型采集的生物特征、行为轨迹等敏感数据面临多重风险：一方面，模型逆向攻击技术可通过分析输出结果反推原始数据，芝加哥大学研究团队已成功从端侧语音合成模型的输出中还原出 87%的原始声纹特征；另一方面，设备丢失或侧信道攻击可能导致本地模型参数泄露，进而暴露用户行为模式。更严

峻的是，端侧大模型降低了虚假信息生成的技术门槛，攻击者仅需单台终端即可伪造逼真的深度伪造内容。

自 2020 年起，中国信通院互联网法律研究中心已持续发布五本数据治理领域的研究报告，形成系列化成果。值此之际，该中心紧密结合端侧大模型技术的前沿进展与应用实践，精心编纂了第六本数据治理专题报告——《数据治理研究报告-端侧大模型数据治理法律要点研究（2025 年）》。本报告立足部署在智能终端设备上的大模型所具有的独特技术架构与运行机理，系统剖析其在数据采集、存储、处理、传输及销毁等全生命周期环节中潜藏的风险特性，通过全面梳理国内外相关立法动态与实践经验，创新性地提出了一套兼具前瞻性与实操性的综合治理策略，旨在为打造安全可靠、合规有序的端侧智能生态环境提供有力的理论支撑与实践指引。

目 录

一、端侧大模型发展概况与技术特点

端侧大模型是指通过量化、蒸馏、裁剪等轻量化技术，将大模型核心推理能力下沉至手机、平板、车载设备、智能家居等终端硬件，依托本地算力完成数据处理、智能交互与任务执行，无需依赖云端持续支撑即可实现离线运行的轻量化 AI 模型体系，其核心价值在于平衡隐私保护、低延迟响应与终端算力适配需求。随着人工智能技术的快速发展，端侧大模型逐渐成为实现智能设备本地计算能力的关键技术，可在终端直接实现数据闭环与实时决策，具有本地化、轻量化、个性化与协同化等特征。

（一）本地化：终端自治与数据保护升级

端侧大模型通过隐私增强型计算架构和离线能力支撑两方面技术设计，实现了终端自治的功能，构建了完整的“数据-模型-服务”的数据安全闭环体系。**一方面，**端侧大模型能够突破网络依赖限制，在弱网或无网环境下提供稳定性服务，直接在终端实现功能支持。以北科瑞声手机端侧大模型为例[1]，其支持离线会议纪要生成与多语种实时翻译功能。端侧大模型这种本地化能力不仅优化了用户体验，还通过减少云端算力消耗降低了企业的运营成本。**另一方面，**端侧大模型通过隐私增强型计算架构可保障数据安全。如，部分智能手表类端侧大模型在实时监测心电数据时，功能分析完全部署在设备端，原始生理信号无需上传云端，可以在一定程度上避免数据泄露。

[1] 《离线可以实时翻译、转写、字幕自动生成！高效开会神器打造 AI PC 最佳使用场景》https://m.jiemian.com/article/12041252.html

（二）轻量化：资源约束下的高效能突破

端侧大模型通过算法优化与硬件协同实现了性能提升，以突破终端设备在算力、内存和功耗方面的严格限制。**一方面**，通过量化、剪枝和知识蒸馏等核心模型压缩技术，在保证模型核心性能（精度、推理效果）能够有效发挥的前提下，降低模型的参数量、计算量与存储开销，实现模型的轻量化部署与高效推理，适配移动端、边缘设备等资源受限场景。**另一方面**，利用资源动态分配机制来优化资源使用效率。端侧设备的资源有限且异构，不同的设备在算力、内存、功耗等方面存在差异。资源动态分配机制可以根据设备的实时状态和任务需求，灵活地分配计算、存储和通信等资源，以提高资源利用率和模型运行效率。这种资源动态分配机制使端侧大模型在资源受限环境下仍能有效平衡性能与能耗。

（三）个性化：用户行为驱动的场景智能

端侧大模型通过分析用户的输入习惯、运动轨迹、生物特征等本地数据，在设备端动态更新知识库与服务策略，实现基于本地数据的深度个性化适配。**一方面**，端侧大模型依托本地部署低延迟的优势，深度挖掘具体场景下的实时交互数据、用户行为偏好及环境特征，通过持续迭代优化模型推理逻辑与适配策略，精准匹配场景核心需求，实现服务精准度、响应效率及适配性的全方位提升。以输入法场景为例，讯飞输入法的端侧大模型通过分析用户历史输入记录，在设备端构建个性化词库与预测模型，使候选词准确率较通用模型显著提升。在健康管理领域，华为 WATCH 5 设备搭载的“玄玑感知系统”整合

了多种生理数据，可实时分析用户健康状况并提供个性化建议。**另一方面，**端侧大模型凭借强大的多模态情境感知能力，可深度融合文本、语音、图像、设备状态及用户行为轨迹等多维度实时数据，精准捕捉场景核心需求与潜在偏好，不仅进一步拓宽了个性化服务的覆盖边界与体验深度，更推动服务逻辑从“用户主动触发的被动响应”，升级为“基于实时情境研判的主动预判与前置服务”，让个性化体验更具时效性、适配性与前瞻性。如，部分智能汽车的语音助理能根据用户日程推荐餐厅并提前预约充电桩，可分析驾驶者的日常路线、音乐偏好、座椅调节习惯等，生成定制化建议，还能根据天气或交通状况调整建议等；部分智能家居系统的端侧大模型可根据用户的作息规律、室内光照与温湿度自动调节设备运行状态，实现“无感化”的智能体验。

（四）协同化：端云融合与生态互联

端侧大模型依托精细化分层架构设计，实现了“端-边-云”三级资源的动态优化配置与高效调度，不仅最大化发挥各层级硬件资源的性能优势，更构建起各终端、节点间互联互通、协同运作的生态体系。**一方面，**通过对实时性任务与复杂任务的智能分流实现全域计算协同，其中，对时延敏感的实时性任务（如语音唤醒、障碍物检测、即时交互响应等）由端侧本地直接处理，可显著降低传输时延、保障服务流畅性；而算力需求高的复杂任务（如多模态语义理解、大规模数据训练迭代、高精度图像解析等）则智能分流至边缘节点或云端，借助更强算力支撑完成高效处理，形成“端侧响应快、边云算力足”的协同

闭环。**另一方面，**依托“端-边-云”分层架构搭建的互联体系，端侧大模型通过多设备智能互动进一步强化全域协同化效能。不同终端设备基于统一的协同协议与数据交互标准，可实现感知数据、任务结果、服务状态的实时共享与联动调度，例如智能家居场景中，端侧设备可同步环境感知数据，边缘节点统筹分配响应任务，云端优化全局服务策略，形成“设备联动-数据互通-任务协同”的闭环。这种多设备互动模式，既弥补了单一终端的算力、感知局限，又让各设备的功能优势形成互补，使“端-边-云”协同从“任务分流”升级为“全域联动”，大幅提升整体服务的响应效率、覆盖范围与适配精度。

二、端侧大模型数据治理法律要点分析

端侧大模型一定程度上是在基于受信任的隔离环境下进行部署和运行的，既能避免终端上收集和产生的数据过多与外部环境进行交互，又能在长期使用和学习的过程中，利用端侧数据将通用模型训练为高度个性化的模型。近年来，各类大模型密集部署，开始规模化应用于智能手机、个人电脑、智能汽车等智能终端设备，进一步强化了设备在图文等领域的处理能力，拓展了人机交互场景，能够更灵活地响应用户需求。但端侧大模型的数据处理方式相比传统云端模型更具隐蔽性、碎片化、场景化特征，且依托设备本地的多传感器、系统权限、离线交互场景，形成了“本地采集+轻量上传+隐性关联”的复杂数据处理体系，因此也引起了针对部分数据治理法律要点的讨论。

（一）数据处理前：“告知-同意”规则的适配

“告知-同意”是个人信息保护立法中确立的个人信息处理核心

规则。一般情况下，个人信息处理者要向个人信息主体告知处理活动的具体情况并在获得个人同意后，才能处理个人信息。“告知-同意”是保障个人对其个人信息处理知情权和决定权的重要手段。但随着端侧大模型日渐与端侧设备融合，深入日常生活，对数据的处理更加频繁且实时进行，传统立法中的“告知-同意”如何适配成为难题。

1.端侧模型通过海量公开数据进行训练，其中可能包含的个人数据难以履行用户同意

端侧大模型内置于终端设备进行应用前，已经基于互联网数百亿语料进行训练，这些语料库与模型的任务相关联，比如文本数据、图像数据、音频数据、交易数据等，其中也不乏有大量个人数据。**一方面，**模型开发者训练基础模型获取数据的重要途径之一是网络爬取，在这一过程中，虽然模型开发者会通过“数据清洗”的方式去除个人信息及其他数据，但由于“数据清洗”通常是自动进行的，无法百分百保证个人数据不再出现在清理后的数据集中。在“Reddit 起诉 Anthropic 抓取公开帖子构成违约、侵占动产、不正当竞争案”中，Reddit 在提交给旧金山高等法院的起诉状中就强调了“删除帖文仍被训练”的风险。[2]在上述情况下，模型开发者获取个人数据主体的同意几乎不大可能。荷兰数据保护机构在其发布的《GDPR 适用条件：AIGC 合规路径》中表示，“在收集过程中处理个人数据，即使在数据集清洗后，这些数据集仍可能包含个人数据。”[3]德国汉堡数据保

[2] 旧金山高等法院官网文件：https://webapps.sftc.org/ci/CaseInfo.dll?CaseNum=CGC25625892&SessionID=7ED3E27E25FF2D2DF9DAEFD37A4BF30FA2441144

[3] 荷兰数据保护机构《GDPR 适用条件：AIGC 合规路径》（全文翻译）:https://mp.weixin.qq.com/s/jtWME0Fh1wP3MFMo0p1A_A

护和自由信息委员会发布的《讨论文件：大语言模型和个人资料》报告中也表示，“针对特定任务进行优化的模型微调，可以在某些情况下重现训练数据，包括个人数据。”[4]另一方面，虽然模型部署到终端设备时，并未存储数据，但是其按照用户指令输出的内容中也有可能会出现个人数据的痕迹。在此情况下，模型部署后对其个人数据产生的数据处理活动，以及应用场景、应用目的等具体内容，无法通知到此前的训练数据主体，更无法获取数据主体的同意。EDPB 发布的《人工智能系统和数据保护的安全基础培训》报告中表示，AI“工作流程中的标注人员的反馈可能会无意中泄露个人数据”[5]。2021 年，已有人通过技术手段成功从 GPT-3 的输出中提取个人信息，以及利用 ChatGPT 中的多步骤提示来诱导个人信息的生成。又如，根据我国《个人信息保护法》第十三条[6]关于个人信息处理合法性依据的规定，除需要获取个人知情同意外，也可以在履行合同必需、履行法定职责或者法定义务必需、应对突发公共卫生事件等特殊情况下，不经个人同意即可处理个人信息。假如大模型在使用个人数据进行训练时符合上述特殊情形不必经过个人同意，后续再将大模型应用于个人终端设备提供商业服务，也无法再获取训练数据主体的同意。尤为值得注意的是，端侧大模型以提供精准化、个性化服务为核心目标，其训练与运行需调用大量包含用户特征的个人数据，这也使得上述数据安全隐患在端侧大模型的应用场景中表现得更为突出。

[4] https://www.bfdi.bund.de/SiteGlobals/Forms/Suche/Expertensuche_Schnelleinstieg_Formular.html?nn=252136&resourceId=270942&input_=252136&pageLocale=de&cl2Categories_Themen=&cl2Categories_Themen.GROUP=1&templateQueryString=llm&submit.x=0&submit.y=0

[5] https://www.edpb.europa.eu/edpb_en

[6] http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html

2.端侧模型通过一揽子授权获取数据具有现实需求，但特殊场景单独授权要求难保障

终端设备变身“贴心助理”的前提是“更懂用户”。融合了大模型的终端设备需要实时、快速对用户的要求做出反应，对收集和处理数据的效率也提出较高要求，这意味着用户需要让渡更多的个人隐私才能换取便捷的服务。如果每次处理数据、响应用户需求都要进行告知和索取权限，反应速度将受到较大影响。端侧模型的本地化迭代依赖数据的持续投喂，量化、蒸馏后的轻量化模型需通过用户交互数据不断优化参数，而分步授权易导致数据采集碎片化，无法形成完整的用户行为画像与模型优化样本，削弱端侧模型的个性化能力与迭代效率。如果通过一揽子授权模式一次性开放核心权限，即可保障数据采集的完整性与连续性，完美适配端侧模型“实时处理、持续迭代”的技术需求。但在这种情况下，《个人信息保护法》中规定的敏感个人信息、向第三方提供个人信息、向境外传输个人信息等特殊场景需要单独同意的要求则难以对应落实。

3.端侧模型大幅度扩展设备功能范围，用户难以精准行使知情权

端侧大模型通过本地化部署将 AI 推理能力深度融入终端设备，大幅扩展了设备功能边界，从手机的离线文案生成、语音实时交互，到车载设备的智能导航与场景化控制，再到智能家居的跨设备联动响应，设备不再是单一工具，而是具备自主感知、决策与服务能力的智能终端。但这种功能的跨越式扩展，也让用户难以精准行使知情权：

一方面，端侧模型的功能实现依赖多源数据的实时采集与联动处理，涵盖传感器数据、本地缓存、设备状态等多维度信息，用户无法清晰知晓设备具体采集了哪些数据、用于支撑哪项扩展功能；**另一方面，**扩展功能与端侧模型的推理逻辑深度绑定，部分功能的触发场景、数据使用范围具有动态性与隐蔽性，用户既难以预判设备会通过模型衍生出哪些附加功能，也无法准确追溯每项功能背后的数据处理链路，进一步弱化了用户对功能与数据关联的认知。

4. 特殊群体保护缺乏有效判断机制

端侧大模型凭借本地化部署实现全场景智能交互，大幅渗透至老人、未成年人、残障人士等特殊群体的日常生活，但其技术架构与应用模式的特殊性，导致针对特殊群体的保护缺乏有效判断机制，难以精准识别群体身份、匹配保护需求，最终使特殊群体在数据安全、内容适配等方面的权益得不到针对性保障。端侧模型依赖本地设备数据独立完成推理决策，缺乏云端集中式的身份核验与群体标签体系，且特殊群体的行为特征易与普通用户混淆，如未成年人使用成人设备、老年人操作习惯接近普通用户，端侧模型无法通过单一交互数据精准判定使用者是否为特殊群体，更难以区分不同特殊群体的差异化保护需求（如未成年人需内容过滤、残障人士需权限适配）。而根据我国法律规定，在处理不满 14 周岁儿童个人信息时，应当取得其父母或监护人的同意，保障儿童及其父母或者其他监护人的合法权益。同时，不满 14 周岁儿童的所有个人信息均被个保法纳入敏感个人信息进行保护，因此，按照法律规定，针对儿童个人信息的任何处理活动都需

要获取父母或者监护人的同意。

（二）数据处理中：数据收集存储规则的效力

1. 最小必要原则的适用性挑战

最小必要原则是国内外个人信息保护立法中被普遍采纳的个人信息处理原则，其源自传统的比例原则。当前，我国《个人信息保护法》[7]明确规定，“收集个人信息，应当限于实现处理目的的最小范围，不得过度收集个人信息。”最小必要原则的最关键前提是需要有非常明确、清晰的数据处理目的，然后才能够确定处理个人信息的最小范围和类别。但是就很多端侧大模型来说，需要为用户提供更加智能化、精准化的服务，如果严格遵守最小必要原则的话，服务的智能化、精准化程度就会受到影响。如，在使用本地会议纪要快速记录和整理功能时，原则上仅需通过录音功能记录会议内容即可，但为了精准将内容定位到每一个人，就需要识别所有发言人的语音，为此搜集的个人信息还可能属于敏感个人信息。以“提升体验”为名的这类数据处理行为是否还应当适用最小必要的原则呢？

2. 端云协同导致数据存储的期限管理无法有效落实

目前，大多数端侧大模型尚不能实现完全的数据本地化处理，因此，端侧大模型主要采用“终端+云端”的混合技术路线，无法真正实现用户终端离线运行。在终端设备部署的大模型可在本地完成对绝大部分数据的处理与分析，并结合用户输入、传感器数据等信息，持续进行初步的微调或更新（例如记录用户的语音特征和使用习惯等）。

[7] http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html

之后再**由端侧设备上传本地参数至云端，构建出更精准的个性化模型。而个人终端上的参数并不能够绝对剔除所有的个人信息，或完全不能够从中提取出个人信息，**因此，上传到云端的参数被储存到云端，可能会导致个人信息上传至云端。《个人信息保护法》第十九条规定“除法律、行政法规另有规定外，个人信息的保存期限应当为实现处理目的所必要的最短时间”。被上传到云端的个人信息可能无法在处理目的实现后及时删除，从而超出针对个人信息存储的期限管理要求。

（三）数据处理后：用户数据权利的保护

1. 多主体多链条下的数据泄露风险

端侧大模型的技术落地与商业化生态，决定了其数据处理行为必然呈现多主体参与、多链条联动的核心特征，这种跨主体、跨环节的交互体系，打破了传统数据“单一主体闭环管理”的模式，形成多主体多链条数据的交互模式。**一方面，主体之间具有关联性。**数据在多主体间的流转多以“模型优化”“功能适配”“生态联动”为由，通过系统后台、设备缓存、离线同步等渠道完成，使得端侧大模型的数据安全风险具备了强传导性——任一主体的安全漏洞不再是“单点风险”，而是会通过数据交互链路，快速传导至其他所有参与主体，形成“一处失守，全域沦陷”的局面，这也是端侧大模型泄露风险相比单一主体应用显著升高的关键原因。**另一方面，数据之间具有关联性。**端侧大模型的多链条联动，使得数据并非孤立存在，而是跨链条、跨场景的关联整合：采集环节的多模态数据（语音、图像、轨迹）与推理环节的用户指令数据关联，存储环节的缓存数据与传输环节的端云

同步数据关联，复用环节的生态数据与迭代环节的模型优化数据关联，最终形成完整的用户精准画像。一旦某一链条发生数据泄露，并非单一类型的数据泄露，而是关联的多维度、高敏感数据的泄露，其危害会通过链条联动快速扩散，从“单一数据泄露”升级为“用户全维度隐私泄露”，甚至引发财产损失、人身安全威胁等次生危害。

2. 用户行使个人权利需要更多保障机制

端侧大模型对个人数据的处理兼具实时性与复杂性，涉及的参与主体及数据处理链条繁多，个人用户根据《个人信息保护法》针对端侧大模型行使自身合法权利，存在一定现实挑战。**第一，**更正删除权的行使存在障碍。与传统计算机将特定信息存储于单一位置不同，端侧大模型的人工智能系统中，个人数据并非集中存储，而是分散于神经网络的数十亿模型参数之中。同时，大模型具有较强的数据记忆顽固性，个人很难精准知晓自身哪些数据被模型采集和使用，更无法从海量分散的数据中精准行使权利。以搜索引擎的数据擦除为例，其可通过两种明确方式实现：一是从源头删除包含个人信息的源网页，二是在搜索引擎索引中删除与个人数据相关链接的特定条目。但这两种方法均无法适用于端侧大模型，即便能从模型的训练数据集中删除个人数据，也无法直接改变已完成训练的模型参数，相关操作仅能在模型后续的迭代训练中逐步生效，更正与删除的效果存在显著滞后性。**第二，**限制及拒绝处理权的行使存在现实障碍。行使限制或拒绝权的核心前提，是个人对自身信息的处理主体、处理方式、处理范围等情况具备充分知情权。而在端侧大模型的应用场景下，数据处理的多主

体、多链条特征，易导致用户对自身信息的实际处理情况不知情；即便用户知晓相关处理行为，因数据处理环节涉及多方主体且权责边界模糊，用户也难以直接向相关责任方（如云端模型服务提供方）有效主张权利。**第三，**个人信息可携权存在技术壁垒。端侧大模型中，个人信息以模型训练权重的形式存在，该存储格式本身不具备可移植性；即便实现技术层面的移植，相关参数也仅能在特定的人工智能系统中适配发挥作用，无法在不同平台、不同模型间自由迁移，直接导致个人信息可携权难以落地。

3. 多主体数据处理引发责任认定困境

在端云协同的技术路线下，数据流转涉及终端厂商、应用开发者、第三方服务提供商等多个主体，多种场景，任何单一主体都难以完全掌握和控制整个服务链条上的数据处理情况。一旦出现数据安全事件，各方的权责不清，个人用户进行维权面临较大挑战。

一方面，模型提供者与模型部署者之间的责任划分界限不明确。第三方模型提供者通过本地化部署向用户提供服务，但模型内部运行机制（如训练数据来源、算法逻辑）对部署者而言属于“黑箱”，第三方模型提供者在开发大模型时存在非法行为，是否会影响部署应用阶段的合法性问题不清晰。在内容生成场景下，如果大模型使用未经授权爬取的版权文本进行训练，导致部署该模型的企业生成内容涉及侵权，模型提供者与模型部署者的责任如何划分目前并未有明确的界定。在端侧场景中，两大主体在各维度的行为更是深度绑定、不可分割。提供者完成端侧轻量化模型的量化、蒸馏，但需针对部署者的硬

件算力（如手机 NPU、车机芯片）做定制化适配，部署者也会基于端侧场景需求，对模型权重做二次调优，二者均参与模型“落地版”的开发，责任无法单一定界。

另一方面，端云结合场景下的数据安全责任不易区分。端云结合是大模型落地的主流模式，其核心特征是将模型推理、数据采集的轻量化环节下放至端侧，模型训练、迭代优化、数据整合的核心环节保留在云端，形成“端侧采集-本地推理-端云同步-云端优化-模型下推”的闭环链路。这一模式虽兼顾了端侧的低延迟、高隐私优势与云端的强算力、强迭代能力，却让数据处理行为跨越端云双域，涉及模型提供者、端侧部署者、云端服务商、硬件厂商等多主体的交叉参与，且数据流转、行为管控、技术实现均呈现端云耦合特征。这种跨域性、耦合性与多主体性相互叠加，使得数据安全责任的认定失去了单一域、单一主体下的清晰边界，最终导致端云结合场景下的大模型数据安全责任极易产生交叉，成为该模式下数据安全治理的核心痛点。

（四）处理全链条：数据污染与内容输出的偏差

端侧大模型通过本地化部署实现了隐私保护与低延迟响应，但可能因数据污染、模型幻觉或对抗攻击生成误导性内容，对个人决策、社会信任及公共安全构成威胁。

1. 本地化训练的数据污染影响模型结果

训练数据的质量与多样性直接影响模型输出。端侧大模型的本地化迭代依赖终端设备采集的零散数据，这类数据缺乏云端的统一清洗与校验机制，易混入错误信息、主观偏见、恶意内容。**一方面，**个人

终端的交互数据本身具有碎片化、个性化特征，若用户长期输入片面观点、虚假信息，模型会在本地迭代中强化这类偏差，形成“偏见固化”；**另一方面，**恶意主体可通过诱导用户输入、伪装合法应用植入数据等方式，污染端侧模型的本地训练样本，使其生成预设的误导性内容。相较于云端模型的集中数据管控，端侧数据的分散性让污染行为难以被及时发现与溯源。

2. 有限算力下的模型幻觉带来认知偏差

模型出现“幻觉”，是指模型生成的内容与现实世界事实或用户输入不一致的现象。端侧大模型为适配终端算力，需通过量化、蒸馏等技术进行轻量化压缩，这一过程会不可避免地损失部分模型精度与逻辑推理能力，导致模型在处理模糊指令、冷门知识时，易生成“看似合理、实则错误”的内容。同时，端侧模型的迭代速度远慢于云端——云端可依托海量数据与强算力快速修正幻觉问题，而端侧模型多依赖用户手动更新或厂商推送补丁，使得幻觉缺陷在终端设备中持续存在，误导用户决策。

3. 本地环境潜在的对抗攻击导致安全盲区

端侧设备的安全防护能力普遍弱于云端服务器，恶意主体可通过输入对抗样本、篡改模型参数、利用硬件接口漏洞等方式，对端侧大模型实施攻击。例如，通过添加微小扰动的语音指令、文本内容，诱导模型误判语义，生成与用户需求相悖的误导性信息；或直接篡改端侧模型的本地权重，使其在特定场景下固定输出有害内容。由于对抗

攻击的隐蔽性强，且端侧模型缺乏云端的实时安全监测机制，攻击行为往往在造成危害后才被察觉。

三、全球端侧大模型数据治理法律制度比较研究

在数据治理方面，欧美模式一直以来都是全球的两种典型代表模式。面对人工智能的飞速发展，欧盟目前仍坚持在 GDPR 的基础上推进个人数据保护，但同时欧盟数据保护委员会（EDPB）及成员国机构也通过发布指南、细则等文件的形式，进一步细化端侧大模型链条上的各方责任。美国则采取了分行业、分场景的个人信息保护策略，更强调通过市场机制和技术标准实现治理目标。在欧美之外，我国则形成了层次分明、相互协同的人工智能法治基础格局，既有顶层立法，也有专门针对人工智能的配套规定，对端侧大模型的个人信息保护提出系统性要求。

（一）欧盟在严格保护个人数据基础上进行灵活处理

欧盟 GDPR 构建了极为严格的个人数据保护框架，为端侧大模型应用场景下的个人数据处理提供了基础性的规则框架。但有关人工智能大模型相关场景如何适用 GDPR 的问题，在欧盟范围内遭遇了广泛争议。德国、荷兰等成员国数据保护机构在执法案例中，对如何适用 GDPR 下的相关义务作出了不同解答。欧洲数据保护委员会（EDPB）发布的《关于人工智能模型中个人数据处理的特定数据保护问题的 28/2024 号意见》对人工智能大模型中的个人数据处理合规问题作出了具有约束性的答复，正面回应了各成员国数据保护机构的争议焦点。此外，EDPB 还推出相关指南为人工智能企业如何开展数

据合规提供详细指引。

1. 欧盟 GDPR 建立“权利优先”的数据保护框架

欧盟 GDPR 构建了极为严格的个人数据保护框架，增强个人对数据流转全环节的控制权，GDPR 的相关要求在端侧大模型的场景下依然适用。**一是设计和默认的数据保护原则**（Data Protection by Design and by Default，DPbDD）。EDPB 在《关于 GDPR 第 25 条设计和默认的数据保护指南》中指出，履行 DPbDD 义务时，应重点考虑保护措施的先进性、实施成本、数据处理性质及可能对个人数据权利产生影响等因素。例如，数据保护措施是否处于行业的最新水平，在数据处理活动全流程均需评估其先进性等要求。**二是以“告知—同意”为核心的个人数据保护规则。**作为数据处理的合法性基础之一，最基本的要求是告知用户数据处理的目的、数据类型、存储及共享范围等事项，在此基础上取得数据主体的明确同意。尽管存在适用困难，告知—同意规则仍在人工智能场景中发挥着重要作用。**三是以“同等保护”为核心的数据跨境流动规则。**在“端云结合”的技术业态中，手机终端厂商通常与第三方云端模型达成合作，只要其中一方位于境外，则可能存在调用境外算力或数据境外存储的情况。根据 GDPR 的规定，需要通过充分性决定、标准合同或公司约束性规则来证明其数据跨境活动的合规性。

此外，欧盟《人工智能法》也与 GDPR 紧密衔接、协同运作，对人工智能中的数据治理提出一定要求。《人工智能法》基于风险将人工智能分为不可接受风险、高风险、有限风险和最小风险四类，不

同风险的人工智能系统承担的数据治理义务不同。例如，高风险人工智能必须确保用于训练模型的数据集在相关性、代表性、准确性和完整性方面达到标准，以尽量减少歧视性风险。同时高风险人工智能还必须设计允许人类监督员干预或停止系统的功能，为基于个人数据的自动化决策机制提供了重要的制衡机制。

2. 各成员国数据保护机构为人工智能数据治理提供范本

欧盟各成员国数据保护机构发布文件对人工智能场景中的个人数据保护问题作出说明，如人工智能大模型的匿名性、第三方模型合法性对后续部署使用的影响以及合法利益作为数据处理合法性基础等问题，对端侧大模型场景中的个人数据保护具有一定的借鉴意义。

（1）德国在数据保护原则基础上明确特殊场景下 LLM 的个人数据保护要求

德国联邦和各州独立数据保护监督机构会议（DSK）在《人工智能和数据保护指南》中明确，LLM 的部署必须遵循数据保护原则，尤其强调“谨慎处理输入与输出中包含个人数据的情形”。根据该指南，公共部门和企业在使用 LLM 时应优先选择不涉及个人数据的应用场景，若无法避免，则必须取得处理数据的合法性基础，并采取封闭系统架构限制数据外流风险。对涉及敏感的情形，DSK 特别指出需警惕“间接识别性数据”的泄露，并要求在模型训练阶段就完成脱敏处理。

汉堡州数据保护机构发布的《讨论文件：大语言模型和个人数据》

[8]对 LLM 的技术原理进行了剖析，得出了 LLM 并不存储个人数据的关键结论，并在此基础上明确相关实践的合规要求。**第一，**某公司使用了第三方开发的 LLM，后来发现第三方在模型训练过程中使用了个人数据，且未取得合法依据。在这一情形下，主要由第三方模型开发者承担法律责任。如果该公司参与了训练过程或对非法训练行为知情，则也需要承担相应责任。**第二，**有关数据主体如何行使个人权利。由于 LLM 并不存储个人数据，因此无法对 LLM 主张数据权利。但针对输出结果或其他可能存储个人信息的数据库，仍可主张相关权利。**第三，**某公司希望使用自己的训练数据对第三方 LLM 进行微调。如果该公司提供的训练数据中包含了个人数据，则应取得将该数据用于模型训练的合法性，并满足数据主体的权利主张。**第四，**某公司将第三方 LLM 部署在本地服务器上。部署行为本身并不涉及对个人数据的处理，但该公司一方面应确保人工智能系统输入和输出的合规性要求，另一方面应采取措施防止数据提取。**第五，**通过应用程序接口（API）等方式使用第三方 LLM，同样需要满足输入和输出端的合规性，同时在选择第三方服务提供商时，应确保其具备防止数据提取的安全防护能力，并在 LLM 投入使用前，通过合同形式明确双方的责任分配。

德国巴登—符腾堡州的数据保护机构发表文章[9]，对汉堡州数据保护机构采取的立场进行了点评。文章指出，LLM 和人工智能系统应被视为一个整体，虽然 LLM 在训练后未直接存储原始个人数据，

[8] *Discussion Paper: Large Language Models and Personal Data*, at https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf

[9] *Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz*, at https://www.baden-wuerttemberg.datenschutz.de/rechtsgrundlagen-datenschutz-ki/

但其输入和输出的交互机制可能成为识别自然人的关键。例如，用户通过姓名、出生日期、地理位置等提示词，生成与可识别个人相关的信息。此时模型输出的结果与用户输入的上下文结合，便可能构成 GDPR 意义上的个人数据。

（2）荷兰基于模型训练不足的合法性基础明确不同阶段的个人数据保护要求

荷兰数据保护机构在《生成式人工智能适用 GDPR 的条件》[10]报告中明确，**几乎所有现成的生成式人工智能在合法性方面都存在不足。**其主要理由是：生成式人工智能依赖所谓的“基础模型”，人工智能企业主要通过爬取的互联网数据训练模型。这些训练数据中很可能包含了未经同意收集的个人数据，尽管这只占训练数据很小的一部分，但无法改变数据非法获得的事实。

该报告还进一步指出了用于训练和微调模型的数据合法性问题。荷兰数据保护机构指出，爬取数据的数据主体与生成式人工智能的开发者通常不存在直接关系，目前只能以 GDPR 第 6 款 1 项 f 条“合法利益”作为收集训练数据的合法性基础。是否能够适用“合法利益”条款可依据欧洲数据保护委员会（EDPB）发布的《基于 GDPR 第 6 款 1 项 f 条处理个人数据的指南》进行判断，即遵循“三步走”的判断标准：首先，数据控制者必须存在合法利益，这些利益必须是合法、明确和现实存在的；其次，对个人数据进行处理对于满足合法利益是必要的，不存在对数据主体影响更小的手段；最后，需要在数据主体

[10] *GDPR preconditions for generative AI*, at https://www.autoriteitpersoonsgegevens.nl/en/documents/gdpr-preconditions-for-generative-ai?sessionid=

基本权利和合法利益之间进行权衡，确保不会对其基本权利和自由产生不合理的影响。

（3）爱尔兰明确了模型训练中的个人公开信息使用问题

爱尔兰数据保护委员会（DPC）发布关于 Meta 利用公开个人信息训练 AI 的声明[11]，从 DPC 与 Meta 展开的多次交锋出发，总结了大模型训练中的个人公开信息使用的问题。DPC 在声明中明确，Meta 为响应 DPC 的要求做出了多项重大措施和改进，包括：对“异议表格”的内容进行更新，用户可通过此种方式阻止自己的数据被 Meta 用于大模型训练；确保“异议表格”适用于欧洲所有司法管辖区，并提供更长的时间访问权限；更新去标识化、数据集过滤及输出过滤等数据保护措施等。DPC 将持续监督“异议表格”的落实情况，确保用户知晓人工智能训练如何影响其个人数据权利，以及可以采取哪些措施防止个人数据被用于此目的，确保所有用户都有机会提出异议。同时，DPC 还提醒所有使用社交媒体和互联网平台的用户定期审查其隐私设置和控制选项。

3. EDPB 回应人工智能场景中个人数据保护的关键争议

为回应各成员国数据保护机构之间的争议、提高 GDPR 执法的统一性，EDPB 发布具有约束性的意见，对部分重点问题作出解释和界定，并发布相关指南帮助人工智能企业提升合规性。

（1）《关于人工智能模型中个人数据处理的特定数据保护问题

[11] *DPC statement on Meta AI*, at https://www.dataprotection.ie/en/news-media/latest-news/dpc-statement-meta-ai

的 28/2024 号意见》

为回应爱尔兰 DPC 提出的解释请求、解决各成员国在人工智能模型中个人数据保护方面存在的争议，EDPB 发布了《关于人工智能模型中个人数据处理的特定数据保护问题的 28/2024 号意见》[12]（以下简称“《意见》”），**对人工智能模型开发及部署阶段个人数据保护问题提出了具有约束性的意见。**《意见》主要回应了三方面的争议：**一是**人工智能模型应在何时被认定为“匿名化”；**二是**在人工智能的开发和部署阶段，个人数据控制者如何证明存在“合法利益”，并作为数据处理的合法性基础；**三是**人工智能模型在开发阶段非法处理数据是否会对该人工智能模型后续的使用产生影响。

一是人工智能模型的匿名性应由数据保护主管机关进行逐案评估。EDPB 认为，根据个人数据训练的人工智能模型不能在所有情况下都视为匿名化。要使人工智能模型视为匿名化，则须满足：直接从人工智能模型中提取个人数据、有意或无意地从查询中获得个人数据的可能性，“对于任何数据主体来说应该是微不足道的”。数据保护主管机关需要对识别出个人数据的可能性进行彻底的评估，并以此判断人工智能模型的匿名性。

二是明确了人工智能模型中个人数据处理如何适用合法利益条款的问题。第一，人工智能模型的数据控制者或第三方处理数据的合法利益限于特定情形。《意见》明确合法利益应当逐案评估，并列举

[12] EDPB, *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*, at https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

开发对话体服务协助用户、开发人工智能系统检测欺诈性内容或行为、改进系统的威胁检测功能三类情形。**第二，**对于合法利益的追求是合法且必要的。根据欧盟法院判例及 EDPB 此前的指导意见，应考量数据主体的基本权利和自由，并结合 GDPR 第 5 条 1 款 c 项的数据最小化原则，对合法利益的必要性进行审查。在人工智能模型开发的过程中，应特别注意处理的个人数据数量，是否与追求合法利益成比例，同时也要考虑数据最小化原则。此外，数据控制者与数据主体之间是否存在直接关系（第一手数据）、数据控制者是否采取保障个人数据的技术措施等因素也会影响必要性。**第三，**逐案权衡合法利益与数据主体的基本权利和自由。《意见》指出，人工智能模型的开发和部署可能对受《欧盟基本权利宪章》保护的私人和家庭生活权、个人数据受保护权等权利带来严重风险，如开发阶段在违背数据主体意愿或在其不知情的情况下对个人数据进行爬取，部署阶段通过数据提取或模型反转推断出学习阶段的个人数据。数据控制者应从处理数据的性质、数据处理的场景、数据处理可能进一步产生的后果进行判断。另外，数据主体的合理预期，即在数据收集时能否预期可能会为该目的进行处理，在平衡测试中也发挥着重要作用。

三是澄清开发阶段数据处理的非法性是否会对部署应用等后续阶段产生影响。《意见》划分了三种情形：**第一种**情况是控制者非法处理个人数据以开发模型，个人数据保留在模型中，部署阶段由同一控制者进行处理。《意见》明确此种情况应根据案件的具体情况逐一评估，如后续处理活动是以 GDPR 第 6 条 1 款 f 项作为合法性基础，

则初始活动处理的非法性会显著影响合法利益条款的适用。**第二种**情况是控制者非法处理个人数据以开发模型，个人数据被保留在模型中，但部署阶段由另一控制者处理。数据保护主管机构应充分考虑部署阶段的控制者是否开展了详尽的评估，以确定人工智能模型不是通过非法处理个人数据开发的，如控制者应考虑数据是否来源于个人数据泄露，或数据已经受到了主管机构或法院的侵权认定。主管机构对于后续数据处理的判断可能因控制者的评估结果及监管期待而有所不同。**第三种**情况是在部署阶段对个人数据进行另一次处理之前，对模型进行了匿名化处理。如果可以证明人工智能模型的后续操作不涉及个人数据的处理，则不适用 GDPR。《意见》明确，只要人工智能模型保持足够的匿名性，则部署阶段处理的合法性不受初始阶段处理合法性的影响。

（2）《人工智能安全与数据保护的法律合规培训课程》

为帮助企业内数据保护官等专业人员开展人工智能安全及数据保护领域的合规工作，EDPB 发布了《人工智能安全与数据保护的法律合规培训课程》[13]（以下简称“《课程》”），全面梳理了人工智能领域的数据合规要求。

一是对数据主体权利的行使提出指导性建议。《课程》明确了在人工智能大模型语境下行使限制和反对处理权、更正删除权、可携权等数据主体权利时可能遇到的困难，并提出了具有针对性的解决方案。**在限制和反对处理权方面，**首要的问题是确保数据主体知晓自己的数

[13] EDPB, *Law & Compliance in AI Security & Data Protection*, at https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/law-compliance-ai-security-data_en

据被用于人工智能训练。当数据主体无法直接联系到第三方模型开发者时，可向人工智能系统的使用或部署者行使权利。《课程》以 InnovaHospital 为例，若患者知晓医院使用第三方开发的模型进行诊断后，明确反对将自己的数据用于第三方模型训练，则可以向医院提出要求。医院不得将该数据用于训练（或微调模型），且必须确保开发者也不得使用该数据进行模型训练。**在更正删除权方面，**人工智能系统不像传统计算机那样将特定信息存储于单一位置，个人数据通常分散于神经网络内的数十亿参数当中。因此，更正删除权利的行使在人工智能大模型的场景中面临困难。在数据主体提出要求时，《课程》明确数据控制者必须响应其要求。作为最终的解决方案，数据保护机构可以在必要时采取“算法驱逐”（algorithmic disgorgement）的措施，即强制删除不合规的大模型。数据保护从业者需要和软件开发人员合作，寻找模型删除之外的替代方案，如使用 CPR 技术使从模型中删除数据可行。**可携权方面，**针对输入和输出阶段包含个人数据的情形，可携权的行使并不存在任何争议。但如上所述，由于人工智能模型的特殊性，难以将训练权重与特定自然人相联系。即便能够识别出特定个人，训练权重也难以直接移植到其他系统中。因此，数据保护从业者应向技术团队确认模型中嵌入的个人数据是否是可移植的格式。

二是明确训练阶段数据处理的合规性要求。《课程》指出，由于 GDPR 第 6 条 1 款（b）至（e）项中的“必要性”要求，其不太可能作为训练阶段数据处理的合法性基础，（a）项的“同意”和（f）项

的“合法利益”条款则更有可能适用。**针对同意，**在人工智能训练阶段可能需要处理成千上万的个人数据，使得取得同意存在困难。且数据主体与大型人工智能企业可能存在不对等的权力关系、难以就人工智能如何使用个人数据提供具体信息等因素，会使得同意条款的适用变得更加复杂。但适用同意条款必须解决这些困难，否则可能因不符合 GDPR 第 7 条规定而被认定为无效。**针对合法利益，**EDPB 意见已经对该条款的适用给出明确意见，《课程》进一步指出，该条款不适用于数据主体基本权利更加显著或涉及特殊类型个人数据的情形。正当利益条款可能为人工智能开发者提供更大的灵活性，但人工智能企业需要在平衡测试中考量人工智能所引发的特定风险。**针对数据再使用，**数据控制者必须观察新的处理目的是否与原目的兼容，GDPR 第 6 条 4 款列举了在此情况下必须考虑的因素。例如，原来的处理目的与预期的处理目的之间的关联。《课程》明确了在人工智能语境下必须考虑的三个特定因素：一是在考虑处理目的之间的关系时，必须考虑人工智能模型的训练目的；二是评估受训练人工智能可能对数据主体产生的影响，及训练行为本身可能引发的数据泄露等风险；三是考虑为应对各类风险所采取的技术和组织措施。

三是明确 GDPR 数据处理原则在人工智能语境下的适用。合法、公平和透明原则方面，合法性是指任何数据处理行为都必须为法律所允许，且满足法律要求。公平性不仅意味着处理训练数据时的公平，也意味着必须应对已完成人工智能系统中可能出现的算法偏见。透明原则要求开发者关注自身及人工智能系统数据处理的透明性。**数据最**

小化方面，《课程》明确了充分性、相关性和必要性三要素。充分性要求开发者确保所用数据质量能够满足当前的任务需求，可参考《人工智能法》第 10 条有关高风险人工智能数据质量的要求。相关性要求开发者能够说明训练数据与其处理目的的相关性，如为了实现对学生课程表现的预测，收集其社交媒体上的信息不具有相关性。必要性要求开发者在不同的数据处理方案中权衡，选择需要处理更少数据的方案。**准确性方面，**不仅包括训练阶段数据使用的准确性，也包括人工智能应用阶段生成内容的准确性。**存储限制方面，**训练数据、测试数据和验证数据以及输入输出阶段的数据，都必须满足 GDPR 的存储时间要求。**完整性和保密性方面，**要求开发者关注数据投毒、后门和环境攻击、对抗性攻击、模型提取和反转攻击等特定类型安全风险，采取相应对策防止个人数据泄露等安全事件发生。

（二）美国采取“场景化保护”灵活路径

与欧盟不同，美国采取了分行业、分场景的端侧大模型数据保护策略，从联邦举措到各州立法，再到司法实践引领规则，以及行业自我监管，形成了多层次的监管框架，更强调通过市场机制和技术标准实现治理目标。整体来看，美国监管框架总体上试图在促进 AI 创新和保护消费者权益之间取得平衡，避免过度监管阻碍技术发展。

1. 形成联邦层面轻监管、各州分别重点治理的模式

一是对 AI 行业监管进行整体松绑。2025 年 1 月 23 日，特朗普 2.0 政府颁布了《消除美国人工智能领导障碍行政命令》[14]（Removing

[14] Removing Barriers to American Leadership in Artificial Intelligence，https://www.whitehouse.gov/preside

Barriers to American Leadership in Artificial Intelligence，第 14179 号行政命令），主要目标是“消除美国人工智能领导障碍”。第 14179 号行政命令弱化了此前拜登行政令中[15]关于数据安全和隐私保护的严格要求，转而强调促进技术创新和减轻企业合规负担。这一政策转向反映了美国联邦政府在 AI 监管上的态度变化，即从强调风险防范转向鼓励创新发展。

二是多州积极开展实践，形成各具特色的监管模式。各州则通过多样化的立法填补空白，形成了各具特色的监管模式。**加州**在人工智能监管领域采取与欧洲较为接近的强监管思路，通过《AB 2013 法案》，以州立法的形式制定了全美最严格的隐私保护和透明度标准，要求生成式 AI 企业披露训练数据的来源、类型和处理方式，以提升数据可追溯性；禁止通过诱导性设计等方式诱骗用户同意数据收集；允许用户拒绝 AI 自动化决策（如端侧推荐算法），并要求企业披露数据用途；规定生物识别数据（如端侧人脸识别）需额外保护等。**马里兰州**通过《在线数据隐私法案》，首次引入“实质性数据最小化（substantive data minimization）”的概念，对数据的收集和使用进行严格限制，要求仅在提供消费者所需的特定服务或产品时才可使用相关数据，避免过度采集。罗拉多州针对重点领域明确要求 AI 招聘工具进行偏见检测，并向用户提供算法决策的解释权。得克萨斯州拟议的《负责任

ntial-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

[15] 拜登于 2023 年签署的《关于安全、可靠、可信地开发和使用人工智能的行政命令》（第 14110 号行政命令）是美国联邦政府对于 AI 行业进行综合性监管的最主要尝试，要求 AI（包括端侧模型）开发者确保数据隐私，并遵循美国国家标准与技术研究院（NIST）的《人工智能风险管理框架》，在涉及国家安全、经济安全或公共健康的高风险模型训练前向政府报备，并共享安全测试结果。然而，该行政命令的执行面临特朗普 2.0 政府挑战。

的 AI 治理法案》草案借鉴欧盟《人工智能法案》，将 AI 系统分为不同风险等级，要求高风险系统（如医疗、就业领域的 AI）进行严格的公平性测试和隐私影响评估，并设立监管机构监督执行等。

2. 针对重点技术、行业、场景明确严格的数据治理要求

一是明确部分人工智能技术的个人信息保护底线。如，《删除法案》[16]为防止 AI 生成内容侵犯个人隐私提供了法律依据。2025 年 5 月 19 日，美国总统特朗普正式签署《删除法案》（Take It Down Act），对 AI 深度伪造技术的滥用进行监管，明确禁止在未经授权的情况下创建、发布或传播包含他人肖像、声音或其他可识别特征的 AI 生成隐私内容，并设定了具体的责任、处罚和执法机制；建立了针对此类违法行为的民事和刑事责任框架；规定了内容发布平台的责任和义务，要求其建立有效的内容报告和删除机制；明确美国联邦贸易委员会（FTC）为执法机构，负责监督法案的执行和处理相关投诉等事项。

二是明确特定行业数据治理的具体要求。美国尚未出台联邦层面的全面隐私保护法律，而是依靠《健康保险可携性和责任法案》[17](the Health Insurance Portability and Accountability Act，HIPAA)、《儿童在线隐私保护法案》[18](Children’s Online Privacy Protection Act，COPPA)、《电子通信隐私法》[19]（Electronic Communications Privacy

[16] Take It Down Act，https://www.congress.gov/bill/119th-congress/senate-bill/146

[17] the Health Insurance Portability and Accountability Act，1996 年 8 月 21 日正式成为法律，https://www.congress.gov/bill/104th-congress/house-bill/3103/text

[18] Children's Online Privacy Protection Act ，该法案于 1998 年 10 月 21 日颁布，2000 年 4 月 21 日生效。2025 年 4 月 22 日，美国联邦贸易委员会（FTC）修订《儿童在线隐私保护法》（COPPA），以加强儿童信息保护，赋予家长更多控制权，该修正案于 2025 年 6 月 23 日生效。https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa

[19] Electronic Communications Privacy Act，https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285

Act，ECPA）等行业特定法规进行规制。例如，医疗健康领域的 AI（如端侧诊断模型）适用 HIPAA，要求医疗数据匿名化处理；COPPA 限制收集 13 岁以下儿童数据，端侧 AI 玩具、教育应用等的数据采集需遵守相关规定等。另外，《联邦贸易委员会法》（FTC Act, 15 U.S.C. §45）规定联邦贸易委员会有权处罚“不公平或欺骗性商业行为”，包括 AI 系统滥用用户数据等隐私保护违规行为。例如，联邦贸易委员会（FTC）依据《联邦贸易委员会法》第 5 节，对 AI 系统的隐私违规行为（如数据泄露、欺骗性声明）进行执法。FTC 因亚马逊 Alexa 默认保存用户语音数据且对删除权限进行误导，对其提起诉讼。

三是针对版权保护明确具体数据治理要求。美国版权局于 2025 年 5 月 9 日公布《版权与人工智能》报告第三部分，讨论生成式人工智能（Gen AI）系统开发时使用他人著作当作训练数据，是否造成侵权及可否构成合理使用问题。该报告认为，Gen AI 的系统开发与训练确实会造成侵害版权的高度风险，至于是否可以主张合理使用，主要看最后利用的方式和个案判断。**部分地区判例法在 AI 训练数据版权、合理使用等方面为行业提供指导。**2025 年 6 月 23 日，美国加州北区地方法院对 Bartz v. Anthropic PBC 案作出判决，在该案中，Anthropic 从盗版网站免费下载了超过七百万本受版权保护的电子书，还购买了数百万本纸质书籍并进行扫描数字化。法院认为大模型公司使用特定图书数据集训练大模型，以及将购买所得的图书转换为数字格式的行为属于合理使用；但下载盗版图书用以搭建数据库的做法不属于合理使用，在数据采集阶段必须确保来源合法。法院指出，版权法并不赋

予作者阻止他人使用其作品进行训练或学习的权利。这一判决为 AI 训练数据的版权问题提供了重要指导，明确了合理使用原则在 AI 领域的适用，同时强调了数据来源合法性的重要性。

3. 统一的端侧大模型数据保护要求在持续探索中

美国各界对端侧大模型，乃至整个人工智能领域的数据保护规定仍有迫切的发展需求，一些机构和部门已经在重要的报告中提出了人工智能时代数据保护方面的具体要求。如，美国国家标准与技术研究院 NIST 发布《人工智能风险管理框架》草案，要求组织对 AI 系统中的数据处理活动进行全面的隐私风险评估，识别可能的隐私风险；AI 系统应仅收集和使用实现其功能所必需的最少数据；AI 系统的设计、开发和部署应遵循隐私设计原则，如匿名性、保密性和用户控制等；鼓励开发隐私增强技术，如去标识化、差分隐私或联邦学习等，以在保护隐私的同时保持数据的可用性。NIST 还和国际标准化组织（ISO）合作推动全球 AI 标准的互认，例如《ISO/IEC 42001:2023 人工智能系统可信性要求》要求端侧大模型在设计阶段融入隐私保护机制，并提供第三方审计报告。美国国家安全局（NSA）发布《AI 数据安全：用于训练和操作 AI 系统的数据安全最佳实践》指南，强调在 AI 系统的开发、测试和运行过程中保护数据的重要性；IEEE 发布的《隐私机器学习标准》针对边缘计算环境提出了具体的安全要求。

4. 行业自律推动企业合规

美国的重点通过引导行业自律和相关实践，主导端侧大模型隐私保护技术创新。FTC 通过“安全港”计划鼓励企业制定隐私保护政策。

2025 年 1 月，FTC 强调 AI 产品默认需要隐私保护和安全措施，要求企业在部署前后评估风险，防止欺诈和隐私侵犯，这对端侧大模型的数据处理和用户通知有直接指导意义。各科技企业对于 AI 使用风险都积极进行制度建设，例如微软的《AI 负责任透明度报告》指出，微软在开发 AI 软件的过程中致力于遵循负责任 AI、隐私合规及数据合规等原则，会根据有关标准对 AI 模型的功能和局限性进行评估，同时会协同产品的利益相关方确认 AI 软件的风险，并确保向用户公开必要的材料以确保透明度；苹果等企业采用“本地优先”策略，通过硬件加密（如 AES-256）、安全隔区（Secure Enclave）、端到端加密，以及内存内处理（PIM）和近内存处理（PNM）等技术将敏感数据保留在设备本地，在设备端完成数据推理，减少云端传输风险；谷歌等企业通过隐私标签、本地化训练等措施提升用户信任。此外，各大企业之间对于 AI 应用也达成部分共识，2023 年谷歌、微软、OpenAI 等 15 家企业签署“AI 安全与隐私承诺”，承诺在发布 AI 产品前完成充分的内部和外部测试，且在整个行业、政府、民间社会和学术界共享有关 AI 产品风险的信息。这些行业自律措施为 AI 系统的开发和使用提供了额外的保障，补充了政府监管的不足。

（三）中国在现有法律框架下开展实践探索

我国人工智能立法稳步推进，形成了层次分明、相互协同的人工智能法治基础格局。《网络安全法》《数据安全法》《个人信息保护法》从不同侧重点对人工智能进行规范，《网络数据安全管理条例》加强对训练数据处理活动的安全管理，《生成式人工智能服务管理暂

行办法》《人工智能生成合成内容标识办法》等进一步细化上位法要求。同时，不断出现的司法执法实践也为端侧大模型的数据治理提供了一定的参考和指导。

1. 现有立法在端侧大模型的适用

第一，明确划分不同主体的不同数据治理责任。

端侧大模型在使用过程中涉及终端厂商、模型服务提供者、APP 服务提供者等多方主体。在这一生态中，终端厂商作为硬件和操控系统提供者，具有获取基础用户数据的先天优势，能够收集用户的操作系统版本、设备型号及位置信息等，APP 服务提供者则关注用户的浏览历史、点赞评论等行为数据，模型服务提供者可能大量接触用户语音数据，将其拆解转化为任务流程，实现用户提出的操控要求。由于数据在不同主体之间流动，导致在各业务环节责任难以划分。

主体责任划分的关键在于区分终端厂商、第三方模型服务提供者、APP 服务提供者之间是否构成共同处理或委托处理个人信息的法律关系。我国《个人信息保护法》第 20 条、第 21 条对共同处理、委托处理情形作出规定，明确相关责任承担要求。**对于终端厂商，**其面向用户提供智能服务时，明确知晓收集的个人信息种类与处理目的，符合《个人信息保护法》中对于个人信息处理者的定义。**对于 APP 服务提供者，**在端侧大模型的场景下，实际上是人工智能模拟用户在 APP 中进行读屏理解、操作，在未达成与终端厂商的合作时，APP 对于人工智能的控制能力有限。即使 APP 服务提供者履行了其自身义务，也无法对可能的不当操作做出及时地响应或干预。在这种情形

中，APP 服务提供者与终端厂商不构成“共同处理”。**对于模型服务提供者，**有关责任划定的情形则更加复杂。在用户直接调用第三方模型服务提供者的情况中，如用户使用识图功能，将圈选或滚动截屏的内容发送给模型服务提供者，第三方模型服务提供者直接收集个人信息并决定处理方式，可能构成单独的个人信息处理。如果终端厂商对个人信息进行初步处理，再借助第三方模型进行深入分析，两者围绕数据处理的决策是不可分离且对数据处理的目的和方式均产生了实际性影响，则可能构成“共同处理”。此外，如果终端厂商将所收集的个人信息加密后以自身名义上传至第三方模型服务提供者，并以自身名义输出结果，则可能被认定为“委托处理”。

第二，重点确立“告知-同意”规则的适用要求。

使用端侧大模型提供服务的动态数据处理的模式需要更加灵活适用《个人信息保护法》《网络数据安全管理条例》（简称“《条例》”）等规定中的“告知-同意”规则，以减少端侧大模型业务场景中的个人信息泄露及权益侵害风险。

《条例》对“告知-同意”规则作了细化要求，明确告知的方式、展示要求和具体内容，如个人信息处理的规则应当集中公开展示、易于访问并置于醒目位置，内容明确具体、清晰易懂。《条例》还明确即便获得个人同意也必须遵循个人信息处理的必要性要求，同时不得超出个人同意的目的、方式、种类、保存期限等处理个人信息。为满足相关要求，终端厂商通常会采取动态授权的机制，在需要截屏获取信息时，通过实时弹窗说明用途，说明“当前屏幕内容将用于优化对

话模型调整”，并允许分项选择同意。

在敏感个人信息方面，我国《个人信息保护法》明确只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，个人信息处理者方可处理敏感个人信息，且处理敏感个人信息需要取得个人的单独同意。由于端侧大模型业务场景下的个人信息收集范围扩大，厂商可能通过录屏等方式获得更多敏感信息，因此数据处理者应根据用户使用场景预设数据收集和使用范围，在涉及敏感信息收集时，须在告知中详细说明涉及收集敏感个人信息的情形，告知该行为的必要性以及可能对用户权益产生的影响。并且，随着场景的变换，数据处理者需要在新的信息收集场景中重新履行告知义务和获得用户的单独同意。另外，《个人信息保护法》明确不满十四周岁未成年人信息属于敏感个人信息，涉及收集此类信息，应当取得未成年人的父母或其他监护人的同意。实践中，已有应用服务提供者通过人脸识别方式验证监护人身份并获得同意。

2. 司法实践提供具体指引

区分不同阶段确认不同主体的责任。杭州“AI 奥特曼”案件中明确大模型服务提供者对于训练阶段的数据原则上不承担责任。杭州互联网法院认为，生成式人工智能通常具有 4 个重要阶段，分别是数据输入、数据训练、内容输出、内容使用，主张对不同阶段区别处理。本案中，杭州互联网法院认定，被告虽然提供生成式人工智能服务，但数据输入端侵权训练素材由用户上传，在内容输出阶段侵权模型和侵权图片生成后亦由用户决定发布或分享，无证据证明被告与用户共

同提供侵权作品，被告未直接实施受信息网络传播权控制的行为，不构成直接侵权。杭州中院同样分阶段分析认为，从数据输入和数据训练阶段看，生成式人工智能的数据来源不仅有平台自己输入的训练数据库，还有平台在服务用户过程中，由用户输入的数据。当服务提供者向公众提供由用户参与训练的模型服务时，有来自全球各地的海量用户对模型进行数据“投喂”，这些数据的合法性和版权状态可能各不相同。在此情况下，若严格要求服务提供者对用户输入端的每一份数据进行逐一审查和验证，既不具有可行性，也与其法律属性不相适应，无疑会加重开发监管负担，势必阻碍生成式人工智能的发展，因此服务提供者的注意义务应当与其身份及信息管理能力相适应。

3. 技术标准持续探索

技术标准在探索端侧大模型治理路径、细化落实法律要求方面发挥重要作用，中国信息通信研究院在端侧大模型风险治理方面开展了两大探索实践。**一是**端云协同场景下的“云上用模”研究。中国信通院已启动《云上人工智能安全发展研究报告（2025 年）》编制工作，聚焦云上人工智能安全发展痛点，深入分析人工智能安全发展趋势和技术演进路径。同时，“云上大模型安全”系列标准的研制工作也已正式启动，覆盖基础设施安全、模型安全、数据安全、工具安全与内容安全等全生命周期能力要求，适配一体机、私有化部署、MaaS 等多种应用场景，实现全方位、全场景的安全标准覆盖。在数据安全方面，针对多元数据汇集引发的权责问题、安全暴露面增多及过度收集分析数据风险，提出了事前规范、事中验证相结合的解决思路。**二是**

Agent 安全评测体系构建。整体安全评测集成了智能终端控制、高精度文本体系、多模态内容审核引擎及规则驱动的行为判别模块。同时，对于 Agent 内置语言模型的安全评测，构建了专门的评测框架，通过精准提示词测试智能体如何转化指令为执行命令，并监测全流程交互数据。

四、展望与建议

面对端侧大模型潜在的对数据法律制度的冲击影响，本报告认为应当构建“发展-安全”的平衡框架——既要通过技术创新和制度完善为技术创新留足空间，释放其赋能千行百业的潜力，也要以制度规范与技术工具为筑牢数据安全屏障，让用户真正享受“智能在端，安全在握”的数字化生活，这不仅是保障个人权益的基本要求，更是维护数字时代社会稳定与发展的关键所在。**一方面，**端侧大模型并非完全“裸奔”，现有法律法规，如《网络安全法》《数据安全法》《个人信息保护法》和《网络数据安全管理条例》等，均已为包括端侧大模型在内的数据保护提供了一定的法律依据，可以在现有法律框架下进行适应性调整；**另一方面，**从精准性和长远发展来看，我们还是应当坚持在平衡安全和发展的前提下，进一步聚焦端侧大模型的数据治理法律要点，通过完善法律规则、补充监管手段、创新技术工具、加强部门合作、深化国际合作等方式，形成“政府主导、企业主责、行业自律”的多方共治生态，实现治理效能的最大化与可持续性。

（一）完善法律规则，强化制度可操作性

1. 针对处理行为的风险等级构建差别告知机制。将端侧

大模型的个人信息处理行为划分风险等级。**对于风险等级较低的处理行为，**个人信息处理者承担相对宽松的风险告知义务。在此类场景下，告知内容可侧重于信息处理的基本目的、大致范围等基础信息，告知方式也可更为简洁灵活，以保障信息处理流程的高效运转。**对于风险等级较高的处理行为，**个人信息处理者则需履行更为严格的风险告知义务。告知内容不仅要涵盖信息处理的详细目的、具体方式、可能涉及的信息类型，还需明确说明潜在风险以及可能对信息主体权益造成的影响。同时，要及时通过简洁弹窗、语音提示等轻量化方式实时告知用户，替代传统静态隐私政策，解决用户知情权模糊问题。

2. 研究探索训练数据同意例外情形。在《个人信息保护法》框架下，研究细化“为公共利益或模型优化必要”的同意豁免条件，要求模型开发者对未获得同意的个人信息进行严格脱敏处理，并留存数据清洗记录，同时赋予数据主体事后异议权。

3. 建立适配最小必要原则的弹性适用机制。针对端侧模型个性化服务需求，明确“最小必要”的判断标准应结合技术可行性与服务必要性，允许在“提升用户体验且无替代方案”的情形下适度扩大数据收集范围，但需通过独立评估论证，并向用户说明扩大收集的合理性。

4. 规范端云协同的数据存储期限。要求端侧模型明确区分本地存储数据与云端同步数据，本地数据遵循“实现目的后即时删除”的原则，云端同步数据需设置自动清理机制，且存储期限不得超过模型迭代周期，同时建立数据存储期限公示制度。

5. 强化特殊群体保护规则。将端侧智能设备纳入未成年人个人信息保护的重点监管场景，对疑似未成年人用户自动触发监护人同意流程；针对老年人、残障人士等群体，制定权限适配标准，明确简化操作、隐私保护强化等强制性要求。

（二）明确多主体责任划分，落实责任追溯机制

1. 细化前端数据分类分级标准。针对端侧大模型处理的数据类型，制定《端侧大模型数据分类分级指南》，明确不同敏感等级数据的处理规则。例如，生物特征数据需强制采用硬件级加密存储，禁止任何形式的云端传输。同时，建立“数据处理必要性评估”制度，防止以“技术需求”为名过度采集数据。

2. 完善端侧侵权认定规则。明确“端侧数据滥用”的司法解释。明确终端厂商通过系统更新偷偷激活端侧模型的语音监控功能等行为，构成非法收集个人信息。引入“举证责任倒置”原则，在用户主张数据泄露时，要求企业承担举证责任（如提供数据加密日志、访问记录等），降低用户维权成本。

3. 明确多主体责任边界。端侧大模型生态涉及设备厂商、模型开发者、APP 提供者等主体，建议根据主观状态和数据调用形式，通过制定三方主体责任清单，明确数据泄露时的追责链条，确保端侧模型发生个人信息权益侵害案例时，可以追溯至个人信息处理的各环节，确定责任主体。主观状态区分主要适用于 APP 与终端之间的责任划分，针对 APP 对操作不知情的终端的一些行为，而且 APP 已经履行了自身义务的，可以规定 APP 不承担终端产生的风险；针对 APP

对操作知情的终端的一些不当操作，而且没有采取必要措施的，可以考虑通过连带责任的方式处理。调用形式区分则可以适用于多种场景，例如，针对用户将截屏上传到第三方 AI 进行处理的情形，第三方 AI 是直接的数据处理者，履行数据处理的所有义务；针对端侧 AI 对用户信息进行了初步处理，再借助云端 AI 作进一步分析的情况，可以考虑二者可能构成“共同处理”，应当依法承担连带责任；对于终端厂商将用户信息加密后以自己的名义上传到第三方 AI 进行处理的情形，可以考虑判断构成“委托处理”，相关权利义务应当通过合同约定，委托方还要对受托方进行监督。

（三）补充监管手段，应对信息内容安全风险

1. 构建公开、可信的端侧大模型语料库。端侧大模型语料库的搭建质量，对算法程序运算生成内容的品质与可信度起着决定性作用，明确信息收集来源并打造公开、可信的语料库，是构建虚假信息生成风险全流程应对机制的关键前提。为显著提升模型对虚假信息的免疫力，鼓励探索基于可信机构（如高校、研究机构）或区块链技术的数据托管平台，支持数据主体上传、共享语料资产，建议以权威数据为核心，通过自动化加工与人工校验结合保障质量，依托安全机制防范风险，并借助行业协作推动标准化，为端侧人工智能应用提供可靠的数据基础。

2. 打造透明和可解释的端侧大模型算法模型。鼓励行业组织制定相关端侧大模型运行指南，针对算法模型的设计原则、运行流程、数据使用规范、可解释性要求等方面内容，为技术服务提供者提

供更加明确的参考和指导。鼓励技术服务提供者通过公开披露算法程序的输入端、输出端等运行过程，让用户了解算法处理数据的过程。同时，为用户提供算法程序运行的解释说明，以通俗易懂的语言解释算法的决策逻辑和依据，保障算法程序的准确性，确保生成内容真实、客观。

3. 完善信息内容审查过滤制度。建立“机器初筛+人工复核+专家终审”的三级审查机制，前端由 AI 模型对海量信息进行实时过滤，标记高风险内容；中端由专业审核团队对机器判定结果进行人工复核，结合上下文语境和来源可信度综合判断；后端邀请领域专家（如医学、法律专家）对争议性信息进行终审，确保审查准确性。同时，制定分级处置策略，对轻微违规内容采取限流、标注警示，对严重虚假信息直接删除并追溯传播链条。

（四）创新技术工具，提升以技管技能力

1. 部署端侧监管沙盒。建立“监管沙盒”环境，允许企业在隔离系统中测试新技术。同时，在监管沙盒中，采用高强度的加密算法对数据进行加密处理，确保测试过程符合数据安全规范；建立单独的应急预警机制，在不侵犯企业商业秘密、个人信息权益的前提下，针对问题频发、潜在风险严重的场景，严密监管端侧大模型的数据调用行为。例如，若检测到模型在非用户交互时段频繁访问传感器，立即触发预警并记录操作日志。

2. 优化数据清洗与脱敏技术。开发适配端侧模型的轻量化脱敏工具，对训练数据中的个人信息进行“不可逆脱敏”（如面部特征

模糊化、身份信息加密处理），同时引入人工复核机制，降低脱敏遗漏风险；针对端侧采集数据，采用“实时脱敏+本地缓存加密”技术，确保原始数据不被非法获取。

3. 构建端云协同的安全存储架构。采用“本地加密存储+云端加密传输”模式，端侧数据使用硬件加密芯片存储，传输至云端时采用端到端加密技术，防止数据在传输过程中泄露；云端建立数据隔离机制，不同主体的数据分区存储，避免交叉泄露风险。

4. 构建个人信息处理风险评估机制。构建完善的个人信息处理风险评估机制，从多个方面综合考量个人信息风险。评估手段的适当性。明确个人信息处理的技术手段必须严格限于国家实定法规预先设定的目的。评估手段的必要性。所采用的技术手段对个人合法权益的侵害必须是最小的。在端侧大模型训练过程中，应优先选择对用户隐私影响较小的数据收集和处理方式。评估手段的均衡性。从“成本—收益”框架进行分析，采用该技术手段实现预设目的所增进的社会公共利益与对个体权益的减损之间不能相差悬殊，权衡技术创新带来的社会效益和用户个人权益保护之间的关系。

5. 开发轻量化隐私计算技术。引入联邦学习、差分隐私等技术，实现端侧数据“可用不可见”，模型迭代仅使用数据特征而非原始数据，既保障模型优化需求，又避免个人信息泄露；针对端侧算力有限的问题，优化隐私计算算法，降低资源消耗。

（五）建立协同机制，提升治理能力

1. 构建“政府监管+行业自律+技术监测”三位一体监管

模式。完善端侧模型备案制度，在模型上线前报备数据处理方案、安全评估报告；行业协会制定自律标准，明确技术规范、隐私保护、数据安全流转、风险事件等方面的要求；引入第三方技术监测机构，对端侧模型进行常态化安全检测，检测结果向社会公示。

2. 建立动态评估机制。要求端侧模型上线后定期进行数据安全评估，若发生重大版本更新或安全事件，需立即开展专项评估，并将评估结果报送有关部门；对评估不合格的模型，责令限期整改，逾期未整改的暂停服务。

3. 强化部门间信息共享。定期召开跨部门联席会议，共享端侧设备安全漏洞情报，协同制定处置方案。建设端侧大模型监管数据库，整合企业备案信息、违规记录、用户投诉等数据，为跨区域执法提供支撑。

4. 建立行业间技术共享平台。由行业协会牵头，联合企业、科研机构共建端侧模型安全技术共享库，共享脱敏工具、安全算法、漏洞信息等资源，推动行业整体技术水平提升；定期组织技术交流活动，解决共性技术难题。

5. 建立便捷维权渠道。设立端侧模型数据权益保护专门平台，便利用户通过平台快速提交更正、删除、投诉等诉求，平台对接模型提供者、部署者等责任主体；对复杂诉求，引入公益诉讼支持机制，为用户提供法律援助。

（六）加强国际合作，共同应对全球性挑战

1. 推动制定端侧数据安全国际准则，明确数据收集、使用、存储的规范。同时，建立跨境数据安全监管协作机制，当发生数据安全事件时，各国监管机构能迅速协同调查，追究责任，形成有效震慑。

2. 充分发挥各国在人工智能、密码学等领域优势，通过建立联合研发项目，共享研究成果，加速端侧数据安全防护技术的突破。例如，共同研发更强大的加密算法，保障端侧数据在传输和存储中的安全性；开发先进的隐私计算技术，使数据在“可用不可见”的前提下实现价值挖掘，防止数据泄露与滥用。

3. 加强国际交流与培训，通过举办国际研讨会、培训课程等活动，促进各国在端侧数据安全领域的经验交流与知识共享。提升全球范围内对端侧大模型数据安全风险的认知和应对能力，培养专业人才队伍。

中国信息通信研究院 政策与经济研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62302973

传真：010-62302476

网址：www.caict.ac.cn