大语言模型数据隐私保护的难点与探索

施敏，杨海军
上海市互联网信息办公室，上海 200032

摘要
基于海量数据训练的大语言模型在带来通用人工智能可能性的同时，也给数据隐私保护带来了新的风险与挑战。在分析大语言模型全环节中涉及的数据隐私保护风险的基础上，对隐私保护中知情同意原则、数据收集“正当、必要”原则所面临的新伦理难点展开分析论证，并探索可能的解决框架和路径，以及实操中仍可能存在的伦理难点。

关键词
大语言模型；生成式人工智能；数据隐私；知情同意；数据责任

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024033

Difficulties and explorations in data privacy protection for large language models

SHI Min, YANG Haijun
CyberSpace Administration of Shanghai, Shanghai 200032, China

Abstract
Large language models based on massive data training bring the possibility of generalized artificial intelligence, but also bring new risks and challenges to data privacy protection. This paper analyzes the risks of data privacy protection in the whole process of large language model, argues the new ethical difficulties faced by the principle of informed consent and the principle of "justification and necessity" of data collection, and explores the possible solution frameworks and paths, as well as the ethical difficulties that may still exist in the practice.

Key words
large language model, generative artificial intelligence, data privacy, informed consent, data liability

0 引言

自2022年下半年OpenAI发布ChatGPT起，以大语言模型（large language model，以下简称大模型）为代表的生成式人工智能技术和产业的发展进入快车道。基于深度学习技术和大量语料数据，大模型在人机对话、内容创作、多语种翻译等领域有较好的应用效果。然而，也因其技术特点和训练所用的语料数据，在大模型研发、应用中，数据隐私保护面临新的伦理挑战和技术难点。本文将对此展开详细论述分析，并探索可能的解决框架。

1 大模型给数据隐私保护带来的新风险

大集成、大数据、高算力三者结合，造就了大模型的快速发展[1]。作为“算力与数据加持下的‘暴力美学’”产物，大模型带来了一些新的数据隐私挑战和技术难点，这些问题覆盖了数据的收集、处理分析、使用与服务、共享、迭代等大语言模型全环节。

1.1 大模型预训练收集的大规模数据，可能成为个人隐私泄露的源头

大模型需要海量的文本数据进行预训练。首先看下“海量”的概念。据OpenAI官方公布，GPT-3.5预训练约使用了45 TB、8 000亿个单词的语料数据。国内千亿级大模型使用的语料量约千亿到万亿Tokens，与GPT-3.5相当。经调研，上述国内头部大模型企业使用的数据量是清洗后真正可用的数据，通常仅占原始数据量的2%。除专业论文、代码等较高质量数据外，原始数据包含不少含有大量干扰、风险信息的网页类数据，清洗后的可用比甚至仅约千分之一，可见收集的原始数据量之大。对比来看，将知名论文网站arXiv上所有论文转成Tokens，总量仅14.1 GB；GPT从版本2演进到版本3.5，参数规模提升约100倍（GPT-2的参数为15亿，GPT-3.5为1750亿）、数据规模则提升了1 000倍（GPT-2语料量为40 GB，GPT-3.5为45 TB）[2]。

海量语料数据很大一部分是公开网页数据或开源数据集，这些历史数据存在未经授权、隐私保护不当的问题，语料数据中包含了大量用户个人信息[3]。这些个人信息包含姓名、网络ID、电话、地址、邮件、证件号、支付信息、生物信息等个人身份直接信息，以及搜索历史、浏览记录、关注对象、发表信息、交流交互、地理位置等动态行为信息。如未能有效脱敏和妥善保护，这些语料数据将成为隐私泄露源头。

1.2 大模型显著降低属性预测的进入门槛可能侵犯个人隐私

大模型基本基于Transformer架构，通过深度学习技术在大量数据上学习训练，生成预测和回答。通过语义理解、特征提取、关系推理、泛化、自适应机制等，大模型具备了较强的属性预测能力，即根据输入的文本数据，能预测出相关实体（人物、地点、组织、时间等）或事件的一个或多个属性或特征的能力，这可能导致用户数据被过度挖掘和关联，侵犯用户隐私权[4]。

根据瑞士苏黎世联邦理工学院马丁·韦切夫教授的研究结果，“训练文本包含个人信息和对话，这些信息可通过微妙方式与语言的使用相关联，如通过某些方言或短语与一个人的位置或人口统计数据产生联系”。根据llm-privacy.org的测试结果，

GPT-4推断私人信息的准确率高达85%到95%[5]。专家指出，“其他机器学习模型也能挖掘私人信息，但大模型可用于高度精准猜测私人信息，属性预测的进入门槛非常低”。即使从模型输入文本中剥离了年龄或位置数据来保护个人隐私，但仍无法阻止大模型做出较精准的推论[6]。

1.3 滥用、恶用大模型的个性化生成服务，给隐私保护带来新挑战

如果说个性化推送算法是“千人千面”，那么大模型人机互动也是“千人千面”“千次千面”，个性化生成信息。大模型可根据用户需求提供个性化信息服务，但这也可能暴露用户的消费习惯、兴趣爱好，用于定向广告。大模型可挖掘出用户的敏感信息，如健康状况、性取向、政治倾向等，这些信息可能被用于商业目的，甚至被恶意利用。

对大模型的滥用，如用来生成虚假新闻和谣言、伪造音视频等，可能会误导公众，影响舆论。例如，自动生成恶意评论，攻击他人，加剧网络暴力，侵犯个人隐私和尊严。

对大模型的恶用，如利用生成的深度伪造内容进行欺诈、诽谤等违法行为，将严重侵犯个人隐私和声誉。例如，生成定制化的诈骗邮件或信息进行社会工程攻击，骗取用户信任，获取更多敏感信息。除ChatGPT外，暗网上还有WormGPT和FraudGPT，降低了网络攻击者实施恶意活动、侵犯他人权益的成本。

1.4 在数据存储、共享、模型迭代及提供服务的长供应链中，均可能有隐私泄露风险

存储在云端或本地的训练数据、发布到开源社区的模型代码和配置文件、部署在互联网的大模型前后端系统，均有可能因外部网络攻击、内部管理不当等原因，发生数据泄露、侵犯用户隐私。2023年9月，微软人工智能研究团队被曝在Github发布开源训练数据时，意外泄露了38 TB的隐私数据。同年12月，安全公司Lasso Security从Hugging Face及Github的存储库中发现1 681个在效的令牌，可访问723家企业组织的账号，其中含有生成式AI项目的高价值数据，使谷歌、Meta、微软和VMware等面临潜在的供应链攻击。

在商业运营与合作中，大模型研发方和运营方可能需要与其他企业或研究机构共享或交易数据，或委托第三方对语料数据进行标注处理。数据的传输、再加工、使用也可能造成泄露。此外，为适应新的数据和业务场景需求，大模型需要不断更新迭代，包括基模型迭代、垂域模型的微调等，需要持续收集和存储数据，这些处理会增加数据泄露的风险。

大模型为用户提供服务的过程可能存在的隐私泄露风险。一是使用应用和API的用户注册信息时，如运营方保存处理不当，可能增加泄露风险。二是模型本身的鲁棒性、可解释性及网络安全防护能力不足，可能增加隐私泄露风险。2023年3月，OpenAI开源代码库的一个错误，让一些用户能够看到另一个活跃用户聊天记录中的标题。三是重要数据的出境安全风险。用户如果使用国外大模型，就意味着将个人信息及涉本国、本单位的重要数据出境。自三星公司允许部分半导体业务部门员工使用ChatGPT开始，在短短20天内就发生3起机密资料外泄事件。

2 大模型给隐私保护带来的新伦理难点

大模型的技术特点带来一些独特的数

据隐私保护伦理难点。在隐私保护中，知情同意是尊重用户自主权的最关键举措，能否在大模型上有效运行？法律法规中关于个人信息收集、使用的“最小必要”原则能否在大模型上落地？接下来将逐一分析和论证。

2.1 海量、无明确目的、与运营服务相分离的数据收集，给“知情同意”带来新难点

语料数据中涉及个人信息，如在未经用户知情同意情况下被收集、使用，可能侵犯用户隐私。为获取尽量多的训练数据，大模型研发方通常从各种来源收集数据，包括下载合法性和质量不明的各类公开数据集、爬取公开网页数据和商业付费数据、使用用户生成的数据等。这些数据的来源和处理过程复杂且不透明。个人信息的收集、使用，至少涉及训练和服务两个阶段。服务阶段可与直接用户形成协议，但是训练阶段根本没有用户。个人信息的归属问题，给“知情同意”带来新难点。

一方面，用户可能并不知道也没有能力知道，他们的个人数据正在被收集和使用，更别提知道被谁、何时、以何种方式、何种目的收集及使用，根本无法“知情”。

另一方面，大模型主体方也无法准确知晓语料中个人信息的主体归属，无法定位到用户、向其明确说明收集和使用其个人信息的目的、方式和范围。那么，谁来“同意”？

因此，隐私保护中常用的“知情同意”原则，在大模型数据收集中面临新的伦理难点，包括个人信息的主体归属是谁？这些信息是否在所发布的互联网平台上有过“知情同意”？这个知情同意，能否授权大模型主体再次及多次收集和使用？

2.2 语料数据的脱敏成本与相关主体的投入意愿，面临新的伦理难题

海量语料数据含有大量个人信息，如大模型研发方未做删除或有效匿名脱敏处理，将直接导致隐私泄露。目前，语料清洗一般使用替换、加密、屏蔽、随机化、泛化、哈希等方法中的一种或多种来做脱敏处理。但是，有效脱敏的前提是能准确识别出个人信息。与通过明确的字段规则识别个人信息的网络运营者相比，针对复杂、海量的语料数据，大模型研发方识别出全部个人信息的技术难度与投入成本要高得多。多数大模型研发方可能只简单通过规则匹配、命名实体识别发现显性个人信息，再进行一些脱敏处理。事实上，这可能只处理了“冰山一角”。

不少大模型企业是初创企业，研发方更愿意将人力、资金、技术资源等投入大模型的训练、微调及产品研发打磨这些可直接见成效的工作。即使是安全投入，也倾向于优先投入大模型的价值观对齐、加强服务侧输入输出内容审核以防生成违法不良信息方面，而非对海量语料数据中全部个人信息及关联信息的识别与脱敏。

2.3 数据收集“合法、正当、必要”原则，导致落地面临新的难点

目前，我国的《数据安全法》《个人信息保护法》和欧盟的《通用数据保护条例》（GDPR），针对个人信息的收集，虽侧重有所不同，但均有“合法、正当、必要”的共性原则要求以及公开收集规则、明示收集目的、方式和范围等具体要求。

但是，对于大模型预训练和持续迭代，语料数据越多越好，至少高质量的数据要尽量多。按“最小必要”原则，应只收

集与模型训练相关的数据，并采取适当措施保护用户隐私。但实操可能存在一定难度，很难判断数据与模型训练是否相关，部分隐私数据可能是被无意识收集的或难以确定其敏感程度。此外，如语料主要来自某一特定领域、文化或群体，训练出来的模型可能存在歧视偏见。在遵循“正当”原则时，需要确保语料的多样性以避免模型偏见。

因此，在实操中，个人信息的收集若遵循“合法、正当、必要”原则或“最小必要原则”，将较难落地。个人信息的收集多半是“灰色收集、过度收集、也许需要就收集”。

2.4 数据与模型的共生关系，可能导致用户删除权、修改权和限制使用权无法得到保障

在《个人信息保护法》和GDPR的相关条款中，用户对自己的个人信息，拥有删除权（或“被遗忘权”）、修改权以及限制（或停止）使用权。但是，在大模型场景中，数据与模型的紧耦合关系以及供应链的复杂性，可能导致这3个用户权利很难被保障。

训练数据与大模型是紧耦合的，准确来说是“共生关系”，数据是模型训练的基础，模型则是从数据中提取知识和信息的工具。完成预训练的大模型其实是一个已将数据吸收为知识和信息的复杂信息系统，如果将大模型看成一个生命体，算法是其骨骼和经络，数据则是深入循环系统的血液。如用户从输出端发现有个人隐私数据，提出要大模型主体方进行删除（遗忘）、修改，这基本不可能实现。因为这意味着要在语料数据中找出并剔除相关隐私数据后，对模型进行重新训练，而这个重新训练成本是大部分模型企业无法承受的（GPT-3训练一次的成本约140万美元）。大模型服务方最多在服务侧输出进行一些策略过滤，以禁止相关内容的生成，部分实现了用户的“限制使用权”，或者准确说法叫“限制输出权”。

在提供服务和迭代优化中，人类反馈强化学习（reinforcement learning with human feedback，RLHF）为模型训练的一个阶段与方法，利用用户对话数据进行训练，从而提升对话能力、加强价值观对齐。如果封闭，可能会影响迭代升级。在第1.4节的案例中，三星大概率是无法追回或者彻底删除这些泄露的商业机密数据，它们已被训练到GPT模型中，很难删除相关信息的所有痕迹，追踪溯源难度更大。

大模型在赋能各行业的同时，也增加了供应链的长度与复杂度。训练好的基模型，运营方除自用、直接形成产品外，可向第三方输出API或基模型，作为训练及部署各行业模型、专业小模型及产品的基础能力。在基模型及服务中，用户的删除权、修改权和限制使用权尚无法得到保障，那么，当模型再训练、私有化部署、定制化开发后形成新的数据，而不同下游供应商隐私保护措施存在差异性，又增加了数据控制难度。因此，大模型供应链上的数据根本无法全部追溯及处理。

3 解决或改进的框架

3.1 应对大模型隐私保护难点的伦理框架和技术解决路径

一是尝试基于数据分类分级的安全防护，提升针对性。梳理数据资产，了解涉个人信息等敏感数据的分布，区分个人数据与非个人数据，建立适用于大数据环境下的数据分类分级保护制度。利用语义识别技术及基于机器学习的结构化数据识别技术，通过词法分析、知识抽取、情感计算、相似计算等对敏感数据进行识别和自动分

类，实现数据的分类分级。对不同分类、不同密级的数据，采取不同的安全防护措施，更好地保护数据的安全。

二是尝试不同情形下的“推定同意”“明确同意”“再次同意”，提升知情同意的可操作性。医学伦理中的知情同意，针对一些特殊情形，有“广泛知情同意”“推定同意”“再次同意”等特殊操作[7]。笔者尝试借鉴医学上的知情同意相关的特殊操作，探索其在大模型不同情形下的可行性。

医学上“推定同意”的一种使用场景是“研究涉及匿名数据和存档样本时，研究者可以假设受试者同意使用这些数据和样本进行研究”。考虑到大模型语料数据来源的广泛性、匿名化处理以及平衡数据使用和保护个人隐私的关系，“推定同意”适用于收集阶段。在遵守相关法规和伦理原则情况下，可考虑“推定同意”。但如果用户得知自己信息被使用且明确不同意的，应保障他们的“退出权”。

大模型正式向公众提供服务时，采用互联网信息服务的常规“明确知情同意”原则。

医学上“再次同意”，适用于“研究过程中出现重大变更时”“当研究结果需要应用于新的场景时”的情况。在大模型中可适用于以下情形：研发方将基模型或API授权给第三方用于再训练、部署或应用产品时，以及数据被用于可能引发隐私问题的应用（如涉及公民身心健康），那么获得明确知情同意可能是必要的，需要“再次同意”或“明确同意”。

三是针对大模型的技术特点，不同阶段采用不同匿名化和数据加密技术手段。除了在应用场景中用对称算法和非对称算法进行数据加密外，训练过程中，可采用差分隐私技术（在数据中添加随机噪声）[8]。数据处理过程可采用同态加密技术对加密数据进行处理，保护个人隐私。在涉及数据共享合作中，可采用安全多方计算技术，在多个数据拥有者之间共享数据，提高训练数据多样性，同时保护个人隐私[9]。此外，根据近期浙江大学研究成果，联邦学习作为一种平衡效率与隐私安全的分布式学习架构在大模型的隐私保护计算方面也具有潜力[10]。

四是强化事后监管，根据隐私泄露的数量等级，分级问责与处置。目前，我国对以大模型为代表的生成式人工智能服务采取“包容审慎”“分类分级”的管理原则，鼓励先发展、强化上线后监管，合规手续仅要求“面向境内公众提供具有舆论属性或者社会动员能力的生成式人工智能服务”。生成式人工智能应开展安全自评估，履行算法备案手续。上线后，服务方承担信息内容生成者和个人信息处理者责任、履行内容安全和个人信息保护义务。算法备案是为了提升模型算法的透明度、保障用户知情权。安全评估是指服务方对标法律法规要求，自行评估内容安全与个人信息保护安全防范措施的有无及合规性。

大模型研发与服务方应制定严格的数据保护政策，包括制定数据收集、使用、存储和删除的政策和数据泄露的应急预案等，确保数据在整个生命周期内的合规性。提高数据和模型算法的透明度，并建立有效的投诉和响应机制，以便用户在隐私受到侵犯后得到帮助。

依据《个人信息保护法》，发生数据隐私泄露，要求数据持有和处理者承担责任、落实有效处置措施。根据个人信息泄露数据量级的不同，除予以相应依法处罚外，监管部门可责令模型研发方、运营方依次分级采取输出侧停止生成、模型微调优化对齐，在一定时间内去除语料中隐私数据并重新训练模型，将重新训练的优化后模型同步给所有下游使用方等应急处置手段的一种或多种，以降低当下和未来对相关用户的影响。

3.2 大模型中数据隐私保护的责任区分探索

既然要问责，责任区分就至关重要。出现个人信息泄露、侵犯隐私问题时，谁应负责？大模型的责任链条长，责任似乎更难区分。从对数据发生作用的主体来看，各自主要责任如下。

- 数据提供者（包含收集者和清洗者）：应确保数据的合法性和隐私性。要确保收集的数据是合法、合理和必要的；确保数据的准确性和完整性，以防模型训练过程中的偏差。数据处理应确保隐私数据的匿名化和脱敏。
- 模型研发者：需要确保模型的安全性，采取隐私保护措施。要确保开发的模型是安全、可靠的，确保数据在训练、优化过程中被合理使用。要确保模型的可解释性和可审计性，以防模型被恶意使用。
- 模型服务运营者：需要确保模型及服务的合法使用和隐私保护。需要确保使用合法合规的模型，确保使用的数据是合理的和必要的。确保模型的安全性，以防模型被恶意使用。在输出侧加强对隐私数据的识别和拦截。
- 模型服务使用者：需要确保合法合规地使用模型服务，不滥用、恶用模型，不使用模型开展各类违法违规和违背伦理道德的活动，不侵犯他人隐私。

在实际中，数据提供者、模型研发者、模型服务运营者可能是同一主体，也有可能完全分离。按“谁收集、谁负责”“谁存储、谁负责”“谁使用、谁负责”的数据责任原则，4类主体均应承担各自的数据隐私保护责任。若从输出端或从其他渠道发现隐私泄露，笔者认为，除确定是因使用者主客观因素造成泄露外，其他情况下，数据提供者、模型研发者、模型服务运营者均应承担相应责任，且模型服务运营者的责任比例应最高。

4 可能仍存在的实操难点

一是数据分类分级的准确性和完整性可能难以保证。技术手段和资源投入不足，数据分类分级可能存在准确性和完整性问题。如何明确区分个人敏感数据与不具有个人识别性的公众数据，是一个实操难点。主体对数据分级分类及采取安全措施的投入，也是影响效果的因素。

二是分类知情同意的难度性和有效性问题仍然存在。“推定同意”的可行性，受限于与所属国家的数据保护相关法律法规的相符性；且当用户提出“退出”时，有效保障其个人信息的删除和退出。“再次同意”同样面临着无法找到同意者这一难题。此外，用户与数据收集者之间的信息不对称，可能导致用户做出不完全知情、不充分知情、非自愿同意的决策。如何在保护用户隐私的同时提升模型的确定性及提供便利、个性化服务，如何合理权衡个人隐私与公共利益及便利性，这些难题仍然存在。

三是数据匿名化和加密技术的有效性和可行性尚无法得到保证。在某些情况下，匿名化数据仍可被反匿名化，一些机器学习技术可以在匿名化数据中发现敏感信息。即使使用同态加密技术，但加密算法增加了数据复杂性，所需的算力和时间成倍增加，且需在加密域和明文域之间转换会影响数据处理效率，影响大模型性能。联邦隐私计算框架仍存在隐私威胁与通信开销大等问题。对于“时间就是金钱”的大模型训练来说，企业是否有动力去这么做？

四是数据问责只能事后处置，责任划

分、应急处置存在实操难点。数据控制者和处理者可能缺乏足够的资源和专业知识来确保数据在整个生命周期内的合规性、透明度，或用户可能面临投诉机制的不完善或障碍情况。若出现数据泄露，4类主体的责任比例如何确定，特别是在数据提供者、模型研发者、模型服务者相分离的情况下。同时，事后应急处置可能治标不治本，事实上很难追踪和处理流经供应链上下游的所有隐私数据。

因此，大语言模型数据隐私保护需要通过不断完善监管机制和伦理框架，推动产业链相关主体的共同作用，以及紧跟技术的发展来持续探索与推进。

5 结束语

基于海量数据训练的大语言模型，给数据隐私保护带来新的风险与挑战。本文分析了大语言模型在数据收集、处理分析、服务与使用、共享合作、迭代等全环节中可能存在的隐私泄露风险。同时，基于隐私保护常用的“知情同意”原则和数据收集“合法、正当、必要”原则，分析论证了实操面临的新伦理难点。针对新难点，本文提出数据分类分级、分场景的不同“知情同意”、分阶段的数据匿名化与加密以及事后依据泄露量级追责与处置的初步解决框架。但是，这些路径在实操中仍可能存在伦理难点，需要不断探索与创新发展相适应的监管机制，加快伦理框架的建设与指引，结合不断发展的技术，推进产业链主体的协同治理，持续加强大模型的数据隐私保护。

参考文献:

[1] 李国杰. 大数据与计算模型[J]. 大数据, 2024, 10(1): 9–16.
LI G J. Big data and computing models[J]. Big Data Research, 2024, 10(1): 9–16.

[2] 人工智能曾小健. GPT–1/GPT–2/GPT–3/GPT–3.5语言模型详细介绍[Z]. AI大模型技术社区, 2023.
ZENG X J. GPT–1/GPT–2/GPT–3/GPT–3.5 details of language model[Z]. AI Large Language Modeling Technology Community, 2023.

[3] 微软亚洲研究院. 知识产权、隐私和技术滥用：如何面对大模型时代的法律与伦理挑战? [Z]. 知乎专栏, 2023.
Microsoft Research Asia. Intellectual property, privacy and technology misuse: how to face the legal and ethical challenges of the era of LLMs?[Z]. Zhihu, 2023.

[4] STAAB R, VERO M, BALUNOVIĆ M, et al. Beyond memorization: violating privacy via inference with large language models[EB]. arXiv preprint, 2023, arXiv: 2310.07298.

[5] 董航, 李慧芳, 陈泱, 等. 大模型时代的隐私保护与内容安全[EB]. 工联网, 2023.
DONG H, LI H F, CHEN Y, et al. Privacy protection and content security in the age of big models[EB]. iitime, 2023.

[6] NEEL S, CHANG P. Privacy issues in large language models: a survey[EB]. arXiv preprint, 2023, arXiv: 2312.06717.

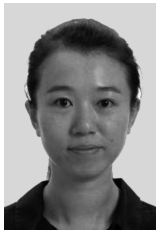
[7] BEAUCHAMP T L, CHILDRESS J F. Principles of Biomedical Ethics (7th ed.)[M]. Oxford: Oxford University Press, 2014.

[8] 参赞生命力. 浅谈大模型数据隐私[Z]. 绿洲资本, 2023.
CAN Z. An lntroduction to big model data privacy[Z]. Vitalbridge, 2023.

[9] 黄霖, 黎源, 汪星辰, 等. 数据自治开放的加密技术挑战[J]. 大数据, 2018, 4(2): 50–62.
HUANG L, LI Y, WANG X C, et al. Challenge of encryption technology for

self-governing openness of data[J]. Big Data Research, 2018, 4(2): 50-62.
[10] 吴建汉, 司世景, 王健宗, 等. 联邦学习攻击与防御综述[J]. 大数据, 2022, 8(5): 12-32.
WU J H, SI S J, WANG J Z, et al. Threats and defenses of federated learning: a survey[J]. Big Data Research, 2022, 8(5): 12-32.

作者简介



施敏(1981-), 女, 就职于上海市互联网信息办公室, 主要研究方向为互联网新技术新业务的发展和安全、网络空间治理、人工智能与大数据伦理等。

杨海军(1973-), 男, 博士, 上海市互联网信息办公室高级工程师, 主要研究方向为通信和公共互联网领域的网络与信息安全工作、互联网舆情、网络空间治理、网络与信息安全战略、互联网新技术新业务的发展及其安全等。

收稿日期: 2024-02-02