Leakage risk and prevention of large AI models

Wenxia Yin

Waitlife (Shanghai) Enterprise Development Co., Ltd., Shanghai, 201601, China

Abstract

In recent years, artificial intelligence large language models (referred to as big models) have made remarkable progress, playing an important role in many fields such as speech recognition, machine translation, video processing and so on. With the popularity of DeepSeek and the application of various large models, the risk of leakage is further highlighted. Once a leak occurs, it will not only damage collective and individual interests, but also may have a negative impact on national defense and security. This paper first analyzes the types and causes of leakage risk of large model, then points out four problems to be solved by large model leakage prevention, gives the technology and methods of large model leak prevention, and finally looks forward to the future technology of large model leak prevention.

Keywords

artificial intelligence; Large language model; ChatGPT; DeepSeek; Disclosure risk

人工智能大模型的泄密风险与防范

殷文霞

等保（上海）企业发展有限公司，中国·上海 201601

摘　要

近年来，人工智能大语言模型（简称大模型）取得显著进展，在语音识别、机器翻译、视频处理等众多领域发挥着重要作用。随着DeepSeek的火爆和各种大模型的应用对象越来越广泛，其泄密风险进一步凸显，一旦发生泄密事件，不仅会损害集体和个人的利益，还可能对国防安全产生负面影响。本文首先分析了大模型泄密风险的类型和泄密产生的原因，然后指出大模型防泄密要解决的四个问题，给出了大模型防泄密的技术和方法，最后对大模型泄密防范的未来技术进行了展望。

关键词

人工智能；大语言模型；ChatGPT；DeepSeek；泄密风险

1 引言

人工智能走进公众的标识性事件是 2016 年 AlphaGo 和李世石的围棋大战 [1]，短短几年过去，人工智能技术发展迅猛，随着大模型技术的产生以及 ChatGPT[2] 的出现，从智能助手到个性化推荐系统，再到医疗诊断和金融服务，人工智能已经渗透到我们生活的各个领域，正在为我们带来前所未有的便利。DeepSeek 的贡献不仅仅是因为其以更少的算力达到了较高的性能，还因为其是完全开源的模型，也就是说即使是 DeepSeek 最先进的模型，我们只需登录便可以使用，不需要注册会员也不需要付费。

大模型系统给人们带来便捷的同时，需要大量的数据来进行训练和推理，而这些数据中往往包含大量的个人信息，如身份信息、行为习惯、健康记录等，一旦数据处理不当或被恶意利用，后果将不堪设想。因此，在享受人工智能带来技术红利的同时，如何防范安全风险特别是泄密风险，已成为当前社会必须面对的一个重要问题 [4]。

【作者简介】殷文霞（1979–），女，中国江苏人，硕士，工程师，从事信息安全研究。

2 大模型泄密风险的类型

随着人工智能技术的发展，数据的收集与处理成为大模型系统运作的核心，而这也引发了对信息泄露的担忧。

2.1 数据层面的泄密风险

人工智能系统的有效运行依赖于海量数据的收集与处理。大模型通过从多种来源获取数据，包括个人设备、社交媒体、医疗记录、电子商务平台等。为了提高系统的智能水平，这些数据被用于模型的训练、算法的优化以及个性化服务的提供。然而，许多数据中包含个人隐私信息，例如身份信息、浏览历史、位置信息、购买记录，甚至是生物特征（如面部识别和指纹数据），一旦这些信息被处理和存储，就可能面临泄露或滥用的风险。因此，理解大模型如何处理数据以及可能带来的风险，是防止泄密的第一步。

2.2 模型层面的泄密风险

大模型在模型层面主要存在模型窃取和模型后门植入两大风险。攻击者会试图获取模型的参数、结构或算法，

以便在未经授权的情况下使用或复制模型，在黑盒场景下，攻击者虽然无法直接访问模型内部，但可以通过向模型发送大量输入数据，观察输出结果，进而推断模型的特性，构建出近似的模型。而且攻击者可能会在模型训练过程中植入后门，使得模型在特定条件下输出攻击者期望的结果，同时不会影响模型在正常情况下的性能。一旦后门被触发，模型就可能泄露敏感信息。

3 大模型泄密产生的原因

3.1 技术漏洞

大模型的开发和部署涉及多个技术环节，任何一个环节存在漏洞都可能导致泄密风险。模型训练框架、服务器操作系统、网络协议等都可能存在安全漏洞，攻击者可以利用这些漏洞进行攻击，获取敏感信息。此外，一些加密算法在量子计算等新技术的冲击下，可能变得不再安全，从而导致数据和模型的泄露。

3.2 人为因素

大模型的使用者是人，内部人员违规操作、员工访问权限不清晰、用户保密意识不强等都会带来安全风险。内部人员由于权限较高，能够接触到敏感数据和模型，如果内部人员缺乏安全意识，或者受到利益诱惑，可能会故意泄露数据和模型。大模型企业在数据和模型的安全管理方面存在漏洞，如访问控制策略不完善、员工安全培训不足等，也可能导致泄密风险的增加。用户在与大模型交互过程中，可能会将敏感信息透露给模型，而大模型生成的回复若被不当获取，将导致敏感信息泄露。用户因缺乏对大模型输出内容的甄别能力，可能传播模型生成的虚假、误导性信息，被不法分子利用，导致间接泄密。

4 大模型防泄密要解决的问题

大模型在处理和分析数据时面临诸多挑战，包括透明性、责任归属和数据公平性等问题，均需在技术发展过程中逐步完善。这不仅仅是一个技术问题，更是社会、法律、政策和道德的综合挑战。

4.1 透明性与解释性

人工智能技术的一个主要挑战是“黑箱”问题。许多大模型系统，尤其是深度学习模型，尽管在处理复杂任务时非常有效，但其决策过程却往往难以被解释和理解。这种不透明性让用户无法了解大模型如何使用他们的个人数据。用户可能不知道哪些数据被收集、如何处理以及被用于哪些具体用途。这种信息的不对称性加剧了信息保护的难度，也让用户产生了信任危机。如果人们不能理解大模型的工作原理或数据处理方式，他们就难以做出知情决策，进一步加大了信息泄露的风险。

4.2 责任归属不明确

人工智能防泄密，另一个要解决的核心问题是责任归属不明确。当大模型系统导致隐私泄露、数据误用或歧视性行为时，责任应如何分配？是开发这些大模型算法的工程师，还是部署这些系统的公司，抑或是使用这些系统的终端用户？目前，大模型相关的法律框架尚不完善，导致在信息泄露事件中往往很难追究责任。

4.3 用户知情同意与保密意识

尽管许多大模型应用在数据收集时会要求用户的同意，但这些“同意”通常形式化，大多数用户不会仔细阅读复杂的隐私政策，往往不清楚他们同意了哪些数据的使用方式和范围。这种“知情同意”的形式化处理，忽视了用户真正的隐私选择权和知情权。用户可能在不知情的情况下，允许大模型系统使用他们的敏感数据，最终导致信息泄露。

4.4 技术滥用与道德边界

大模型技术的强大也带来了潜在的滥用风险。无论是政府、企业还是个人，若缺乏足够的监管和伦理约束，可能会利用大模型技术进行大规模的监控、数据挖掘或隐私侵犯。如何设定大模型应用的道德边界，防止技术滥用，保障公民的权利，是大模型防泄密中亟需解决的问题。

5 大模型防泄密的技术和方法

在大模型应用防止泄密需要多层次的解决方案，包括隐私设计理念、技术措施、用户控制权、法律法规等。通过在技术创新和法律规范上双管齐下，可以有效解决大模型技术在防泄密中遇到的问题，确保技术发展与保密并行。

5.1 隐私设计理念

“隐私设计理念”要求在大模型系统开发的初期就将隐私保护融入设计和架构中，而非事后补救。这一理念包括以下原则：数据最小化原则指的是大模型系统仅收集完成任务所需的最少数据，避免不必要的数据收集，减少泄密风险；去识别化和匿名化原则指的是在数据使用过程中，大模型系统可以通过去除身份标识和进行数据匿名化处理，确保个人信息无法被直接识别，即使数据泄露，也不会轻易暴露个人身份；加密处理原则指的是在数据传输和存储过程中，采用先进的加密技术，保证数据在传输路径或存储位置中的安全性。通过将这些隐私保护机制内嵌于大模型系统的设计和开发中，隐私保护能够在技术上得到有效实现，并减少人为干预中的隐私风险。

5.2 数据匿名化与加密技术

匿名化与加密技术是大模型防泄密中的两大关键技术措施。数据匿名化即通过技术手段移除个人数据中的识别信息，使得数据无法与特定个人直接关联。即使大模型使用这些数据进行分析，个体的身份依然是不可追踪的。数据加密是指在数据传输和存储过程中，使用加密技术可以保护数据的机密性。即使数据在传输或存储过程中被拦截，未经解密的状态下，攻击者也无法读取到有用的个人信息。加密算法的选择和密钥管理是确保加密安全性的关键。

5.3 用户控制权与透明性增强

为保护用户隐私，大模型应用需要赋予用户更多的数据控制权，并提升系统的透明性。用户应当能够清楚知道他们的数据被如何收集、使用和存储，并可以随时访问或删除其个人数据。这种控制权能够确保用户掌握自己的数据，减少泄密的可能性。同时大模型系统在收集和使用个人数据时，必须确保用户知情并明确同意。另外，用户应该享有数据共享的控制权，即允许用户自主选择是否将个人数据共享给第三方，并在需要时撤回这一授权。

5.4 法规与政策的支持

隐私保护不仅需要技术层面的措施，还需要法律和政策的配合。目前，许多国家和地区已经制定了隐私保护相关的法规，欧盟的《通用数据保护条例》为用户提供了广泛的隐私权利。我国的《个人信息保护法》也要求企业在处理用户数据时保证合法、正当、透明，同时赋予用户对个人信息的访问、修改和删除权。这些法规的出台，为大模型隐私保护提供了法律框架，促使企业和开发者在技术之外，更加重视合规要求。

6 未来技术展望

6.1 技术创新助力隐私保护

随着大模型技术的不断进步，新的隐私保护技术将持续出现，推动大模型与隐私保护的和谐发展。联邦学习被认为是未来大模型隐私保护的重要方向之一，它允许大模型在不集中收集数据的前提下进行训练，通过将模型训练分布在各个设备上，避免了大规模的数据集中处理，降低了隐私泄露的风险。差分隐私技术通过在数据集中添加噪声，确保即使对数据进行大规模分析，也无法识别出个体信息。数据合成是指通过生成虚拟数据集替代真实数据来训练大模型。这种技术既能提供高质量的数据用于大模型训练，又能保护个体隐私。

6.2 大模型伦理框架的完善

随着大模型技术的飞速发展，社会对大模型系统的伦理要求越来越高，大模型隐私保护的伦理框架也将变得更加完善。未来的大模型开发可能会引入更加严格的多方协作与伦理审查制度，确保在数据收集、模型训练、算法部署等环节中，隐私和伦理问题得到充分考虑。同时未来隐私保护的法律和伦理框架将更加国际化，全球范围内的隐私保护标准化将进一步规范大模型系统的数据处理行为，并为跨国企业提供一致的操作标准。另外未来的大模型系统将逐渐被要求具备更高的可解释性，尤其是在数据处理和隐私保护方面，透明的大模型决策过程可以让用户清楚地知道他们的数据是如何被使用的，从而增强用户对大模型的信任，减少隐私侵犯的可能性。

6.3 社会隐私意识的提升

随着大模型技术的广泛应用，公众对隐私保护的意识将逐渐提高。未来，用户将更加关注个人数据的安全性和隐私权利，并主动参与隐私保护措施中。通过隐私教育，提高公众对数据保护的认识，帮助他们更好地理解大模型应用中的隐私风险以及如何应对，用户将具备更强的隐私保护意识，使用更多的隐私保护工具，如加密通讯软件、隐私浏览器和数据管理工具等，进一步掌控和保护自己的隐私数据，主动行使自己的数据权利。

7 结论

人工智能特别是大模型技术的快速发展在带来巨大技术进步和社会效益的同时，也提出了严峻的防泄密挑战。大模型技术依赖于海量数据，而个人信息在数据收集、存储和处理过程中，极易受到侵犯和滥用。通过技术创新、隐私设计理念、法律法规的完善，以及用户保密意识的提升，大模型防泄密问题可以得到有效缓解。联邦学习、差分隐私等新兴技术将成为未来防泄密的重要工具，而全球范围内的法律和伦理框架也将为大模型应用提供强有力的保障。未来的大模型发展将不仅仅依赖于技术进步，还需要依靠全社会的共同努力，在创新与保密之间取得平衡，为实现更安全、更公平的数字社会奠定基础。

参考文献

[1] Silver, D., Huang, A., Maddison, C. et al. Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–489 (2016). https://doi.org/10.1038/nature16961.

[2] Andrew Mihalache, Ryan S Huang, Marko M Popovic, Rajeev H Muni, ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination, Medical teacher, Volume 46, 2024, pp 366-372.

[3] DeepSeek-AI, Daya Guo, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, https://arxiv.org/pdf/2501.12948v1.

[4] 周瑜，周涛，大模型的安全挑战及应对建议，中国信息安全，2024年第6期，42-44.