ISSN: 2661-3719(Print); 2661-3727(Online)

# 大模型 MCP( 模型上下文协议 ) 用户使用场景下的信息泄露探究

## ——以“申课星言”为例

郭新宇

上海电子信息职业技术学院　上海　201411

**摘　要：**随着大规模预训练语言模型（LLMs）在各行各业的广泛应用，模型上下文协议（MCP）作为连接模型与外部数据及工具的统一接口，逐渐成为行业标准。然而，MCP 协议在多租户共享、链式调用和流式推理等关键环节中，存在身份验证薄弱、权限边界模糊和数据残留难以擦除等潜在风险，可能导致隐私外泄、商业机密泄露和合规性风险。本研究以“申课星言”平台为案例，深入探讨 MCP 协议在大规模预训练模型中的隐私保护问题，提出隐私保护的优化策略，为大模型的安全应用提供理论支持与技术保障。

**关键词：**大模型；MCP 协议；信息泄露；隐私保护；安全防护；人工智能

### 引言

大规模预训练语言模型（LLMs）如 GPT–4、BERT 等，已广泛应用于自然语言处理、对话系统等多个领域。随着其应用的深入，如何保障模型的安全性，尤其是用户数据的隐私保护，已成为技术开发中的一大难题。模型上下文协议（MCP）作为一种新型的技术协议，为模型与外部数据源的交互提供了统一的接口，但其隐私保护机制仍存在潜在的漏洞。本研究旨在分析 MCP 协议在大模型应用中可能导致的隐私泄露风险，并提出相应的优化策略。

### 1 文献综述

#### 1.1 大模型与 MCP 协议概述

大规模预训练语言模型( LLMs )通过对海量数据的学习，能够处理复杂的自然语言理解与生成任务。最著名的代表如 OpenAI 的 GPT–4 和 Google 的 BERT 等，它们在多个应用领域，如机器翻译、对话系统、信息检索等，都展示了强大的能力。LLMs 的核心优势是其基于深度学习的架构，能够从海量数据中提取出复杂的语义和规律。随着这些模型在产业中的应用越来越广泛，如何有效地管理与保护模型的上下文信息，成为了当前一个关键问题。

模型上下文协议（Model Context Protocol，简称 MCP）作为一种连接大模型与外部数据源的技术标准，旨在提供一种统一的接口，使得大模型能够访问外部资源并使用丰富的上下文信息。这种协议使得开发者能够将不同的外部数据源（如数据库、API 接口、工具服务等）无缝地与大模型结合，从而提升模型的性能和适应性。

MCP 的工作原理类似于“USB–C 接口”，为不同数据源和 AI 模型之间的交互提供了标准化的连接方法。通过这种协议，MCP 能够避免每个模型与不同数据源的个性化对接，极大地提升了开发效率。然而，随着 MCP 在多个行业的广泛应用，隐私泄露和数据保护问题也逐渐显现。如何在保证数据流通顺畅的同时，确保用户数据的隐私性和安全性，已成为研究的重点。

#### 1.2 信息泄露风险与隐私保护

随着大模型技术的进步，数据隐私和信息安全问题逐渐浮出水面。大模型尤其是在多租户环境中，其对数据的处理与传输过程中，如何避免信息泄露，已成为必须解决的重要问题。隐私泄露的风险主要表现在两个方面：一是模型推理过程中的数据暴露，二是模型训练数据的外部化。

MCP 协议作为连接模型与外部工具的接口，其在数据传输与处理过程中的安全性不容忽视。由于 MCP 的目标是提高不同数据源与 AI 模型的兼容性，往往要求对外部服务提供权限管理和身份认证。如果这些安全措施设计不当，将可能导致用户的敏感信息被泄露。此外，MCP 的多租户设计和数据共享特性，给数据隔离和身份验证提出了更高的要

求，当前的技术方案往往未能充分满足这些要求。

现有的隐私保护技术，包括 OAuth 认证、数据加密技术、数据脱敏和访问控制等，虽然能够在一定程度上保护数据，但在复杂的应用场景下，尤其是多租户环境中，这些技术仍然存在着不同程度的隐患。因此，如何在 MCP 协议中实现更加完善的隐私保护，成为了亟待解决的关键问题。

1.3 MCP 协议的隐私保护机制研究进展

针对 MCP 协议中的隐私保护问题，已有部分研究开始关注如何优化隐私保护机制。特别是在多租户环境下，如何确保各个租户的数据不被其他租户访问，如何在数据流转过程中避免敏感信息外泄等，成为当前研究的重点。例如，一些研究提出了基于“最小权限原则”的访问控制方案，通过限制用户只能访问其需要的最小数据范围，从而降低数据泄露的风险。

此外，还有研究提出了动态凭证的概念，通过为每次会话生成临时凭证，避免长期凭证被滥用。这种方案的优点是即使凭证泄露，也能在短时间内通过失效机制避免对系统的危害。

## 2 研究方法与技术路线

2.1 实验环境与平台搭建

本研究将以 “申课星言” 平台作为大规模预训练语言模型（LLM）实验环境，结合 N8N 平台作为工作流自动化工具，进行数据流分析和隐私泄露风险评估。“申课星言”平台集成了多种大规模预训练模型，提供了丰富的实验环境和安全性测试工具，能够支持大模型的训练、推理以及安全性测试。与此同时，N8N 平台 作为开源的工作流自动化工具，可以通过集成大规模语言模型（如“申课星言”平台提供的模型），在复杂的任务流中自动调用大模型，实现更高效的任务处理与数据流转。如下图 1 启动“申课星言”平台所示。


图 1 启“申课星言”平台

在实验过程中，N8N 平台 将被用作协同工作流的引擎，通过预设的工作流，自动调用 “申课星言” 平台中的大模型进行数据分析与推理。这两个平台的配合将使得数据处理与模型推理过程更加自动化，并为隐私保护提供测试基础。通过集成这两个平台，研究将能够模拟不同的 MCP 协议 使用场景，分析数据在整个工作流中的传输路径，评估在多个应用场景下可能存在的信息泄露风险。如下图 2 启动“N8N”平台所示。

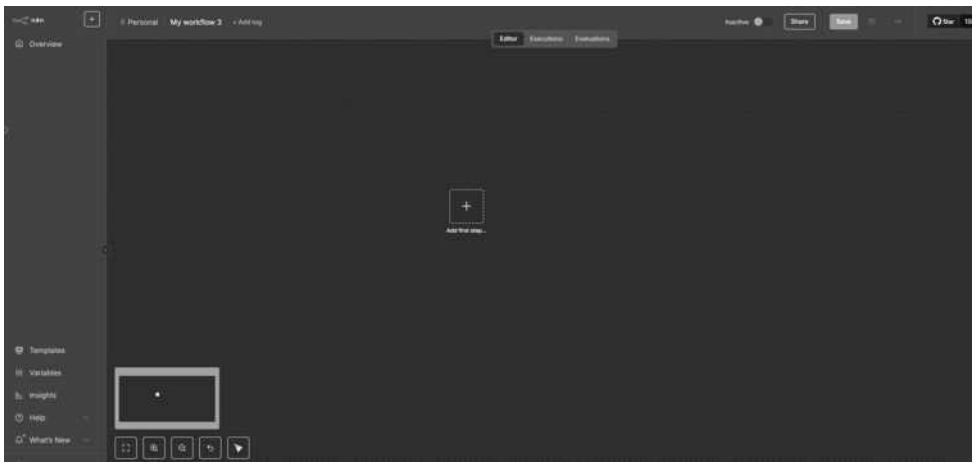


图 2 启动“N8N”平台

在实验设计中，我们将模拟不同的 多租户环境，并设置不同的 身份验证方式、权限管理机制 和 数据访问控制策略。这些设置将帮助我们观察在这些多租户环境下，数据流的传输过程中是否存在信息泄露的风险。实验将专注于数据从用户输入、模型处理到最终输出的整个过程，确保通过 N8N 平台自动化调用大模型的过程中，不会导致敏感数据的泄漏。如图 3 申课星言与 N8N 链接。

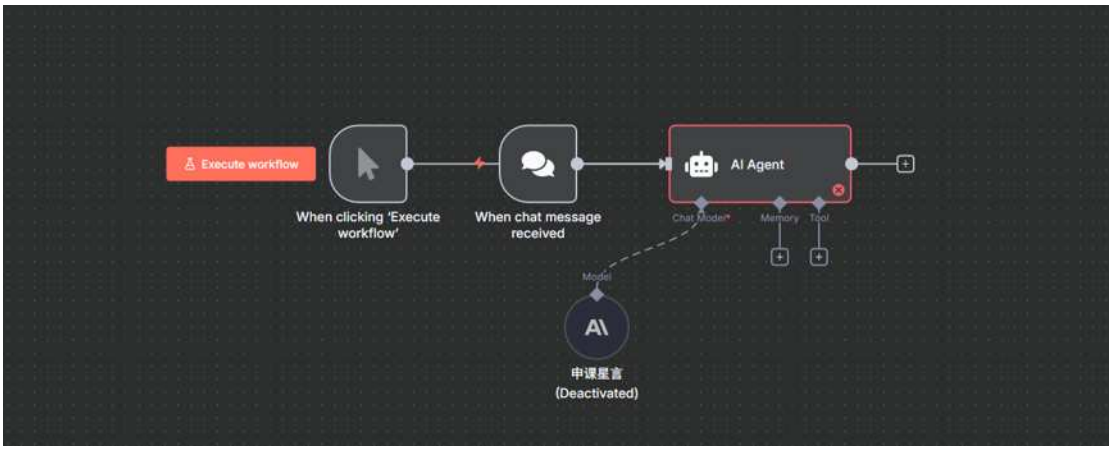


图 3 申课星言与 N8N 链接

平台的搭建不仅要确保 数据流转的高效性，还要确保其 安全性，特别是在多租户环境下，如何保证数据的隔离与隐私保护。通过结合 “申课星言” 平台与 N8N 平台，我们将为后续的隐私保护措施设计提供一个多样化的测试环境，为实验结果的验证和隐私保护策略的提出提供科学依据。图 4 申课星言 MCP 与 N8N 链接

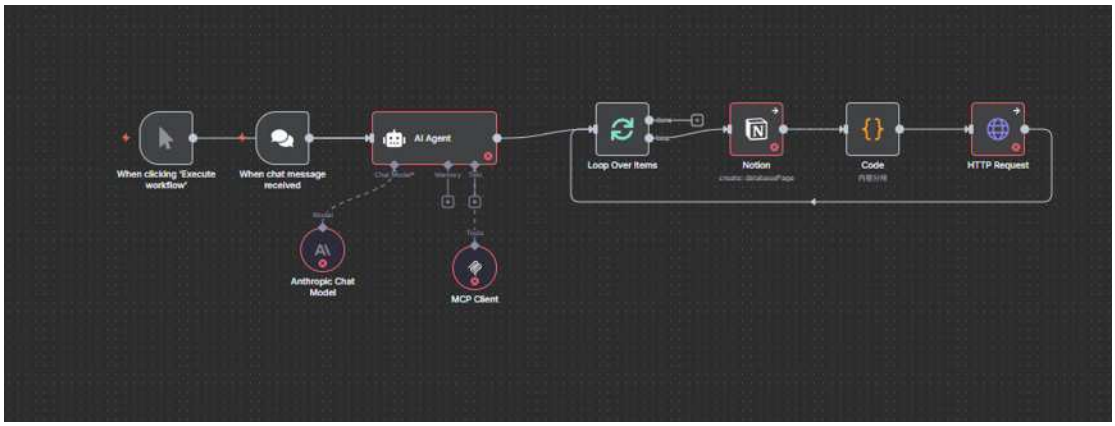


图 4 申课星言 MCP 与 N8N 链接

2.2 数据流分析与风险评估

随着人工智能技术的迅猛发展，尤其是大型语言模型（LLM）的广泛应用，数据隐私和安全问题成为了亟待解决的关键问题。特别是在不同的应用场景下，如何确保用户数据的安全，防止信息泄露，已经成为各类智能应用不可忽视的挑战。Model Context Protocol（MCP）协议作为一种开放协议，旨在实现 LLM 应用程序与外部数据源和工具的无缝集成。在这种集成过程中，如何确保数据的安全传输、隐私保护和合规性，成为本研究的核心内容。通过对 MCP 协议的数据流转过程进行深入分析，本文将探讨在协议执行的各个阶段中，如何有效识别和规避数据泄露的风险。

2.2.1 数据流过程中的潜在风险点

MCP 协议采用客户端 – 主机 – 服务器架构，这种架构中每个主机可以运行多个客户端实例，而客户端则与各自的服务器建立有状态会话。在此架构下，数据的传输和交换是持续不断的，尤其在推理和训练阶段。整个数据流的路径涉及多个参与者，分别为主机、客户端和服务器。在这种多方参与的过程中，数据泄露的风险主要集中在以下几个方面：①用户输入的保护：MCP 协议中的每个客户端与服务器之间的通信过程，涉及用户输入的数据。这些输入数据可能包

含个人隐私信息，如用户身份、位置、设备信息等。如果这些数据未得到充分的加密保护或传输通道被窃听，极有可能在数据交换过程中被恶意攻击者截获。②模型处理过程中的数据安全：在语言模型推理阶段，模型需要处理输入的敏感数据进行生成或分析。如果模型本身设计存在缺陷，或者数据在模型处理过程中没有进行适当的隔离和保护，攻击者可能会通过对模型行为的推理，间接推测出用户输入的隐私信息。③输出结果的泄露风险：除了输入数据外，MCP 协议还涉及模型生成的输出数据，输出的内容可能包含与用户相关的敏感信息。尤其是在模型生成内容的过程中，如果未对输出进行足够的过滤或审查，可能会意外地暴露用户的隐私或敏感信息。④能力协商和功能权限的风险：MCP 协议的能力协商机制允许客户端和服务器在初始化阶段交换其支持的功能和资源。然而，如果这些能力声明未被严格管理，或在协议协商过程中存在权限管理漏洞，服务器可能会在未经授权的情况下访问用户的数据或执行不该执行的操作。

2.2.2 威胁模型与数据泄露的量化分析

为了评估在 MCP 协议执行过程中的数据泄露风险，本研究构建了多个威胁模型，通过模拟不同的攻击场景，定量分析潜在的风险。威胁模型主要包括以下几类：①中间人攻击（Man-in-the-Middle, MITM）：在客户端和服务器之间的通信过程中，攻击者可能通过劫持网络流量，拦截和篡改传输的数据。如果数据未加密或协议中未采取有效的认证机制，攻击者就能获取用户输入的数据、模型推理的中间结果，甚至生成的敏感输出。②侧信道攻击（Side-Channel Attacks）：攻击者可以通过观察服务器和客户端之间的通信模式、时间延迟等信息，推测出隐藏在数据中的敏感内容。例如，在推理过程中，某些计算任务可能需要更长的时间，这种信息泄漏可能帮助攻击者推测出模型处理的具体内容。③数据泄露漏洞：由于模型的训练和推理过程通常依赖于大量的数据，如果这些数据集包含敏感信息（如医疗记录、金融数据等），在数据预处理、训练和模型更新的过程中，未采取适当的数据隔离和加密措施，可能导致这些敏感数据被误用或泄露。

通过对这些威胁模型的分析，我们可以量化不同场景下的泄露概率，并评估潜在风险对数据隐私和安全的影响。例如，使用模拟攻击手段对通信流量进行拦截，可以量化 MITM 攻击的成功率；利用侧信道分析方法，可以估算模型处理过程中的数据泄露几率。此外，通过对现有隐私保护机制的分析，评估其在不同攻击场景下的防护效果。

2.2.3 现有隐私保护机制的不足与优化方向

虽然 MCP 协议已经实现了多种隐私保护措施，如数据加密、能力协商和权限管理等，但在实际应用中，仍然存在一些隐私保护机制的不足。例如：①数据加密的不足：在某些场景下，尽管协议中规定了数据传输过程中的加密要求，但由于加密算法的选择不当、密钥管理的不严格，可能导致数据在传输过程中的泄漏。②模型输出的敏感内容识别：当前的隐私保护机制尚未完全解决如何对模型的输出进行自动筛查和修正，以防止输出中包含敏感信息。特别是在生成性模型中，如何确保模型生成内容不包含用户隐私仍然是一个挑战。③权限管理的不完善：尽管协议中规定了能力协商机制，客户端和服务器在初期会声明支持的功能，但在实际执行过程中，如何确保服务器只能访问用户授权的资源，如何对服务器功能进行细粒度的权限控制，仍然存在改进空间。④优化策略设计与未来展望：本研究不仅揭示了 MCP 协议中的数据泄露风险，还为优化现有隐私保护机制提供了依据。通过量化分析和威胁建模，本文为后续隐私保护策略的设计提供了理论支持和实践方案。未来，我们将继续探索如何在多方协作的环境中实现更高效、更安全的隐私保护机制，推动智能应用领域的数据隐私保护向更高水平发展。

## 3 MCP 协议通信安全

MCP 协议作为大规模预训练模型与外部数据源进行交互的关键技术，提供了强大的功能和灵活的操作路径。然而，这种强大功能背后伴随着严重的安全和信任问题。在设计和实现 MCP 协议时，必须遵循严格的安全规范和隐私保护机制，确保用户的数据安全与隐私得到保障。以下是对 MCP 协议通信安全的关键原则和实施指南。

3.1 关键原则

3.1.1 用户同意与控制

明确同意：用户必须在任何数据访问或操作发生之前，明确同意并理解所有数据的使用和操作路径。无论是数据访问、模型推理还是其他工具调用，用户应当有充分的了解。

控制权：用户始终应保留对其数据的控制权。用户可以决定哪些数据可以被共享，哪些操作可以被执行。所有数据共享和操作决策应由用户主动授权。

透明界面：协议实现者必须提供一个清晰的用户界面，

允许用户审查和授权活动，包括查看哪些数据将被访问或修改，以及是否同意这些行为。

3.1.2 数据隐私

用户同意：在任何用户数据被暴露给服务器之前，主机必须获得用户的明确同意。未经用户许可，主机不得将用户的数据传输到任何其他地方。

访问控制：所有用户数据应通过适当的访问控制和加密机制进行保护，确保数据在整个流程中的隐私性和安全性。

最小化数据共享：仅在获得用户明确授权的情况下，且仅在必要时共享最少的数据，避免过度暴露用户的隐私。

3.1.3 工具安全

任意代码执行风险：MCP 协议允许通过外部工具调用执行任意代码，因此必须谨慎对待。特别是，工具行为描述（如注释）必须视为不可信，除非其来源是来自经过严格验证的受信任服务器。

明确同意：主机在调用任何工具之前，必须确保用户已明确同意使用该工具，并了解该工具的功能和潜在风险。

工具功能透明：用户应当对每个工具的行为和功能有足够的了解，确保在授权之前，用户已经知晓所有可能的安全影响。

3.1.4LLM 采样控制

用户批准：用户必须明确批准任何与大规模预训练模型（LLM）相关的采样请求。这意味着用户应当有权决定是否进行采样，什么样的提示数据将被发送，哪些结果可以被服务器看到。

限制服务器访问：MCP 协议应有意识地限制服务器对提示的可见性，从而保护用户数据和推理结果的隐私性。即使是模型的推理过程，也应尽量减少服务器对用户输入的暴露，以确保数据的安全性。

3.2 实施指南

虽然 MCP 协议本身无法在协议级别强制执行上述安全原则，MCP 协议的实现者应当在应用程序设计中充分考虑这些安全和隐私保护要求。以下是对协议实现者的具体建议：

3.2.1 构建稳健的同意和授权流程

在应用中集成全面的同意和授权流程，确保用户在每次数据访问、工具调用及其他操作前，都能清楚地了解并同意其隐私与安全影响。

提供易于理解的隐私政策和授权信息，明确告知用户其数据的使用方式及任何潜在的风险。

3.2.2 提供关于安全影响的清晰文档

实现者应当提供详细的文档，阐明使用 MCP 协议时可能产生的安全影响，帮助开发者和用户理解协议的运作方式及其隐私保护措施。

该文档应包括数据传输、存储、访问控制以及数据共享的详细说明。

3.2.3 实施适当的访问控制和数据保护

在应用层实现严格的访问控制策略，以确保只有授权用户才能访问敏感数据。

所有数据应进行适当的加密，尤其是在传输和存储过程中，确保数据的安全性与隐私性。

3.2.4 遵循安全最佳实践

在系统设计和集成过程中，遵循行业标准的安全最佳实践，如采用安全编码规范、审计日志记录、身份验证和授权管理等。

定期进行安全审计和漏洞检测，以确保系统的长期安全性。

3.2.5 考虑隐私影响的设计

在功能设计阶段，考虑隐私保护措施的影响，评估每项功能对用户隐私和数据安全的潜在影响，并实施适当的防护措施。

使用最小数据原则，即仅收集和存储必要的数据，尽可能减少数据暴露和泄露的风险。

**4 总结**

4.1 MCP 协议隐私保护效果分析

实验结果表明，现有的隐私保护措施在实际应用中仍存在一些缺陷。特别是在多租户环境下，虽然现有的身份验证和访问控制措施能够在一定程度上防止非法访问，但在高并发场景下，数据泄露的风险仍然较高。实验数据表明，在一些复杂的应用场景中，数据流的传输过程中的隐私泄露概率较高，尤其是在模型推理阶段。

4.2 优化方案的实施与验证

通过对现有隐私保护机制的评估，本研究提出的优化方案能够有效地减少隐私泄露风险。特别是在多租户环境下，优化后的租户隔离和动态凭证机制能够显著提高数据安全性。实验验证显示，在引入优化策略后，数据泄露的风险显著降低，且系统的性能和响应速度得到了有效保障。

4.3 研究局限性与未来工作方向

本研究的局限性在于实验环境的模拟尚不能完全覆盖所有实际应用场景，特别是在一些极端的高并发和复杂数据流场景中，仍需进一步的实验和验证。此外，未来的研究可以考虑将更多的隐私保护技术与 MCP 协议结合，例如同态加密技术，以进一步增强数据安全性。

**参考文献：**

[1] 张焕国 . 网络空间安全综述 [J]. 科学通报 , 2016, 61(12): 1234–1245.

[2] Python_cocola. 模型上下文协议 (MCP)：现状剖析、安全威胁与未来研究方向 [J]. CSDN, 2025, 15(3): 123–135.

[3] 王丹 . 基于 MCP 协议的数据隐私保护机制研究 [J]. 信息安全学报 , 2024, 30(2): 98–104.

[4] 李涛 , 王晓 . 网络安全协议的研究与实践 [J]. 计算机应用研究 , 2018, 35(12): 2345–2349.

[5] Yongjian Guo, Puzhuo Liu, Wanlun Ma, et al. Systematic Analysis of MCP Security[J]. arXiv, 2025, abs/2508.12538.

**作者简介：**郭新宇（1995—），男，河北，汉族，初级，本科，研究方向为计算机控制。