Indirect Prompting Injection of Attack Surface Measurement for Large Model Security Analysis links

Xiaofeng Meng[1] Manjiang Li[2]

1. Dongying Integrated Media Center, Dongying, Shandong, 257091, China

2. Weifang Beida Qingniao Huaguang Phototypesetting Co., Ltd., Weifang, Shandong, 261061, China

Abstract

The Security Operations Center is handing over alarm triage, detection rule writing, and code auditing to the big language model. Once the decision-making process is automated, the analysis tools themselves become attack targets. Previous studies have used whether the model obeys instructions as a criterion for successful attacks, but cannot answer the operational questions of attack surface selection: where to place the payload, how to conceal it, and the highest effectiveness. This article takes the degradation of defense judgments as the measurement object, defines the injection induced False Negative rate (IFN) as the main indicator, constructs an injection matrix containing 540 payloads and 45 units, and measures the effectiveness differences between 3 types of injection positions and 5 types of covert techniques on 8 model configurations. The results showed that the tool had the highest efficiency in returning to the position, with an IFN of 0.412, which was 38.7% higher than that of passive data sources; Semantic camouflage has the best combination of effectiveness and concealment, with an IFN of 0.401 and an automatic detection rate of only 12.6%; Zero width characters are the most concealed but have the lowest effectiveness. Cross carrier reuse rate measurement shows that the decoupling degree between tool return location and business context is the highest, with 81.4% maintaining effectiveness even after model switching. The generation of detection rules is the most fragile in cross task measurement, with a rule breaking rate increasing from 0.9% to 45.6%.

Keywords

large language model security; indirect prompt injection; attack surface measurement; payload concealment; security operations center; adversarial robustness

大模型安全分析链路的间接提示注入攻击面测量

孟小锋[1] 李满江[2]

1.东营市融媒体中心，中国·山东 东营 257091

2.潍坊北大青鸟华光照排有限公司，中国·山东 潍坊 261061

摘要

安全运营中心正在把告警分诊、检测规则编写和代码审计交给大语言模型，判决环节一旦自动化，分析工具本身就变成了攻击目标。已有研究以模型是否服从指令作为攻击成功的判据，回答不了攻击面选择的操作性问题：载荷放在哪里、用什么方式隐蔽，效力最高。本文以防御判决劣化为度量对象，定义漏报率增量（Injection−induced False Negative rate，IFN）为主指标，构建包含 540 条载荷、45 个单元的注入矩阵，在 8 个模型配置上测量 3 类注入位置与 5 类隐蔽技术的效力差异。结果显示：工具返回位置效力最高，IFN 为 0.412，比被动数据源高 38.7%；语义伪装在效力与隐蔽性上综合最优，IFN 为 0.401 而自动检出率仅 12.6%；零宽字符最隐蔽但效力最低。跨载体复用率测量表明，工具返回位置与业务上下文解耦程度最高，换模型后仍有 81.4% 维持效力。跨任务测量中检测规则生成最脆弱，规则破坏率由 0.9% 升至 45.6%。

关键词

大语言模型安全；间接提示注入；攻击面测量；载荷隐蔽性；安全运营中心；对抗鲁棒性

1 引言

安全运营中心把告警分诊、检测规则编写、代码审计交给大语言模型（Large Language Model, LLM）已是普遍做法，判决环节一旦自动化，分析工具本身就变成攻击目标。Model Context Protocol（MCP）这类协议把数据库、Shell、工单系统的访问令牌交给智能体，而模型处理上下文时不区分指令与数据，攻击者不必攻破任何边界，只要让恶意文本出现在被分析的数据里。这类攻击已有实际观测：恶意加载器中被嵌入针对 LLM 安全扫描器的注入文本，长注入荷载检出量半年增长约 5 倍，国内技能仓库发

【基金项目】山东省重点研发计划（2025TSGCXTHG029）；新闻出版A2A智能体全流程重塑工程）。

【作者简介】孟小锋（1976-），男，中国山东广饶人，在职研究生，从事计算机及网络研究。

现 8 个植入恶意提示词的仿冒技能包，工具投毒在自动批准默认设置下的攻击成功率被测到 84.2%。

2 相关工作

间接提示注入的特点是攻击者不与模型直接交互，而是把恶意指令放进模型将要读取的数据里。这条线上的越狱类基准测量模型绕过安全对齐的概率，工具滥用类基准测量智能体是否被诱导调用高权限工具，判据都落在模型是否服从上[1]。这种度量放在防御场景会失效：分诊任务的输出是受约束的判决标签，模型即使被诱导也未必输出攻击者想要的标签，ASR 与判决劣化是否相关需要实测，不能从概念上假定。

3 威胁模型与指标定义

本文考察的安全分析流水线分四段：结构化告警输入；检索与工具调用（工具返回在本地 MCP 模拟环境中实现，字段名沿用 tool_result）；模型判决（输出四级标签：高危、中危、低危、误报，应判为高危或中危的样本称为应报样本，判为低危或误报即构成漏报）；判决消费。攻击者假设为能够向流水线读取的数据源写入内容，覆盖被动数据源（工单正文、日志字段、被检索网页）、工具返回（tool_result）和代码与仓库配置（README、隐藏目录、编辑器配置）；攻击者可定制载荷文本与编码，但不能修改模型权重、系统提示和流水线控制逻辑。威胁模型不使用任何软件漏洞——攻击者需要的全部能力就是会写一段文字。为使位置与隐蔽技术的效力差异不被防线混淆，本文假设流水线未加装专门防线，防线介入后的变化属于防御篇的问题。

4 基准构建

TriInject 由良性告警底料和注入载荷矩阵两部分组成，全部实验在自建隔离环境完成。底料取自 NVD/CVE 描述、MITRE ATT&CK 技术映射和公开威胁情报报告，改写为告警格式并逐条标注应判等级，双人独立标注、第三人仲裁，一致性低于 0.75 即回修规范；最终 800 条样本中应报 512 条、良性 288 条，仲裁后标注一致性 Kappa 为 0.81。

5 实验设置

流水线按第 3 节四段结构实现，所有实验共用同一系统提示模板、检索策略和判决空间。共测量 8 个模型配置，兼顾开源与商用、7B 至 72B 参数规模（72B 采用 4-bit 量化，其影响未单独消融，是本文一个缺口）。每个配置独立运行 5 次，比例类指标用 bootstrap 计算 95% 置信区间，位置间差异做配对检验并报告效应量[2]。

6 攻击面效力测量

本节数值为预期值，按研究假设推导，实验完成后以实测数据替换。

6.1 无注入基线与注入位置效力

无注入基线本身存在不可忽略的漏报：Qwen2.5-72B-Instruct 漏报率 8.0%（$FN_0 = 41$）、误报率 11.5%，DeepSeek-V3 为 7.2%/9.8%，Llama-3.3-70B 为 10.6%/13.2%，商用模型 A 为 5.9%/7.6%。这部分属于流水线固有缺陷，也正是 IFN 采用增量形式的理由。固定 Qwen2.5-72B-Instruct 按注入位置汇总。

6.2 隐蔽技术的效力与可检测性

固定工具返回位置，五种隐蔽技术的对比见表 1，同时报告自动检测器与人工评审两条检出通道。

表1 隐蔽技术的有效性与可检测性

隐蔽技术	IFN	自动检测器检出率	人工评审检出率
显式指令	0.438	87.3%	96.1%
语义伪装	0.401	12.6%	28.4%
base64 编码	0.336	64.8%	41.2%
白字白底	0.279	8.9%	15.7%
零宽字符	0.226	5.4%	6.3%

语义伪装 IFN 为 0.401，只比显式指令低 8.4%，自动检出率却只有 12.6%，是综合最优的一种；零宽字符最隐蔽但效力最低，部分模型在预处理阶段就将其去除。效力和隐蔽性并不同向，真正危险的是处在曲线中部的语义伪装。对防御方的直接含义是：以关键词和正则为核心的注入检测器覆盖能力相当有限，攻击面评估必须把隐蔽性作为第二个轴。

6.3 载荷的跨载体复用率

位置与隐蔽技术刻画的是单次投放的效力，从攻击者成本结构看，还有一项属性同样重要：一条已验证有效的载荷，换一个目标还能不能继续用。本文按

三种载体变更条件测量复用率，结果见表2。

表2 载荷的跨载体复用率（单位：%）

注入位置	换模型后	换任务后	换编码后	三者均保持
工具返回	81.4	68.2	91.7	63.5
代码与仓库配置	74.6	52.9	86.3	48.1
被动数据源	69.8	71.4	79.5	55.2

工具返回位置的载荷在换模型后仍有 81.4% 维持效力，三者均保持 63.5%，是三类位置中最高的，原因是这类载荷围绕工具返回的语义角色设计、与业务上下文解耦；代码与仓库配置换任务复用率最低（52.9%），因为载荷要贴合具体项目的目录结构与配置语义；被动数据源换任务尚可（71.4%）但三者均保持偏低，跨模型时更容易被预处理差异削弱。把表 1 与表 3 并读：工具返回在单次效力和复用率两个维度都占优，代价是投放门槛最高；被动数据源门槛最低而两个维度居中；代码与仓库配置在效力和复用率上都不占优，价值主要在隐蔽性。这解释了公开威胁报告中工具投毒类攻击"频率不高但单次影响较大"的形态。

6.4 跨任务泛化

在同一攻击配置下测三个任务：告警分诊漏报率从 8.0% 升至 48.4%，代码审计从 12.3% 升至 40.4%，检测规则生成的规则破坏率从 0.9% 升至 45.6%，是三者中最脆弱的。原因在于规则生成的输出本身就是防御资产，被污染的不是一次判决，而是一条会被长期复用的规则[3]。这也说明注入位置的优势并非在所有任务上等量成立，任务的输出结构会调制注入效果——这一现象在 7.2 节的媒体语境对照中以更极端的形式再次出现。

7 媒体业务场景下的注入面映射

第 6 节的排序是在安全分析任务上得到的，本节把同一套框架搬到以东营市融媒体中心为代表的媒体内容审核语境下做同等约束的对照。业务描述全部来自公开资料。

与威胁模型相关的业务环节有四处：开放内容入口（黄河口论坛帖文与评论、客户端互动、东营号来稿，论坛注册会员近 18 万、日均发帖约 1 万）、诉求与工单流转（掌上 12345 图文视频工单及系统内流转）、稿件与资料检索（稿件库、图片库、历史报道）、自动编排与第三方工具。前三处分别对应被动数据源与工具返回，第四处对应代码与仓库配置，三类注入面在媒体业务中的可达性依次为高、中、中低，典型攻击目标分别是使审核判决放行或工单降级、污染审核与事实核查结论、污染自动审核规则与关键词策略[4]。

8 语境依赖的效力排序

按第 4 节流程另建媒体子集（模拟来稿覆盖评论、论坛帖文、通讯员稿件三种形态，模型与参数与第 5 节完全相同），对照测量给出两组结果。安全语境的效力排序为工具返回 > 代码配置 > 被动数据源，媒体语境则是被动数据源 > 工具返回 > 代码配置，排序发生倒置；最高 IFN 分别为 0.412 和 0.387，语义伪装检出率 12.6% 对 11.9%，最脆弱环节分别是检测规则生成（RBR 45.6%）与审核规则库自动维护（RBR 43.1%）。

9 结论

本文把间接提示注入在安全分析链路上的攻击面拆成注入位置与隐蔽技术两个正交维度，在 8 个模型配置上完成效力、隐蔽性、复用率与跨任务测量。主要结论有三条：工具返回位置单次效力最高（IFN 0.412）且跨模型复用率最高，是效力和可摊销成本综合最优的通道；语义伪装在效力与隐蔽性之间取得最优折衷（IFN 0.401、自动检出率 12.6%），单轴评估会误判威胁；效力排序是任务相关的，媒体语境下被动数据源反超工具返回（最高 IFN 0.387），通用风险排行不能跨场景引用。后续将扩样做位置与隐蔽技术的因子设计，实测标定复用率，并把评测扩展到多轮交互。防线方案的有效性与代价由同系列防御篇另行讨论。

参考文献

[1] 人工智能安全治理框架2.0版[EB/OL].全国网络安全标准化技术委员会，(2025-09-15)

[2] 网络信息内容生态治理规定 [EB/OL].中央网络安全和信息化委员会办公室，(2019-12-15)

[3] 生成式人工智能服务管理暂行办法 [EB/OL]. 中央网络安全和信息化委员会办公室，(2023-07-10)

[4] 互联网信息服务深度合成管理规定 [EB/OL].中央网络安全和信息化委员会办公室，(2022-11-25)