

No-Box Vulnerability Analysis: Description-only Detection of Indirect Prompt Injection Vulnerabilities in MCP Servers

Zehua Zhang*, Jie Hu, Pratham Hegde, Aditya Maheshbhai Gabani, Souradip Nath, Yibo Liu, Siyu Liu, Hongkai Chen, Hulin Wang, Zhuoer Lyu, Chang Zhu, Divij Handa, Yan Shoshitaishvili, Tiffany Bao, Ruoyu Wang, Adam Doupe

School of Computing and Augmented Intelligence, Arizona State University

Abstract—Conventional vulnerability analysis relies on source-level or binary-level access, or dynamic interaction, all of which may be unavailable to third-party analysts auditing closed-source, remotely hosted, or commercially gated software. Therefore, we propose a new paradigm of *no-box vulnerability analysis* in which neither source code nor runtime interaction is accessible, and only functionality metadata is available. Such metadata defines the intended behavior of the system, including its inputs, outputs, and side effects, while constraining the space of implementations consistent with that behavior. The intended behavior implies *irreducible data flows*, the minimal source-to-sink data flow skeletons shared by all conforming implementations. By reasoning over irreducible data flows, an analyst can formulate vulnerability hypotheses without observing or interacting with the target system. The analyst can later validate these hypotheses when additional access becomes available.

We showcase the feasibility of no-box vulnerability analysis through implementing a pipeline called MCPSEC, which audits Model Context Protocol (MCP) servers for indirect prompt injection vulnerabilities using only the tool metadata exposed at server registration time.

We evaluate MCPSEC on 20 widely deployed MCP servers comprising 177 tools, among which human evaluators confirm 95 vulnerable tools. MCPSEC identified 143 tools as vulnerable, and for each vulnerable tool, it produces *Theory of Concepts* (ToCs), a hypothesized attack scenario for later analyst validation. Using metadata alone, MCPSEC recovers 94 (98.9% recall), compared with 80 (84.2% recall) for the LLM baseline. Overall, our results introduce no-box vulnerability analysis as a new analysis paradigm and demonstrate its practical feasibility in realistic systems.

I. INTRODUCTION

You’re red-teaming a network, and after scanning you identify a juicy target server. Before you start analyzing the server, you check with your contact: uh oh, this server is the central (and only) credit card processing server for the entire company. If it goes down the company loses thousands of dollars per second. You’re forbidden from sending traffic to this server lest you disrupt the process. What do you do? Exclude the system from your audit?

The issue is that traditional vulnerability discovery generally follows three paradigms: white-box analysis, where the source code of the target system is available; black-box analysis,

which requires extensive interaction with the target system; and gray-box analysis, which combines interactions with information from the system.

However, there are situations in which the analyst may not have access to source code, access to deployment details, or the ability to freely interact. Furthermore, it might be *unethical* to test the system for vulnerabilities, as the only way to test is in deployment, yet potential vulnerability attempts might accidentally exploit other users. Consider also analyzing a hospital’s MRI machine: no digital twin is available, a second machine cannot be procured due to expense, and the machine cannot be tested in situ or taken offline due to the impact on patient care.

To address this gap we conceptualize a new vulnerability analysis paradigm called *no-box vulnerability analysis*, which only requires access to *system metadata* to reason about the potential existence of vulnerabilities in a system.

Our insight is that for some vulnerability classes, even limited metadata (e.g., documentation or system description) can suffice for hypothesizing (and eventually discovering) instances of these vulnerabilities.

Consider HotCRP, whose documentation describes functionality that accepts an input string and queries the database for submitted papers that match. Every implementation that is consistent with this documentation must contain a data flow from user input to a database query.

The fact that all conforming implementations have this data flow suggests a vulnerability hypothesis: the *potential* existence of an SQL injection vulnerability. We call this vulnerability hypothesis a *Theory-of-Concept* (ToC), which is an explicit exploitation scenario that includes additional assumptions about the attacker’s capabilities and roles, and the potential vulnerable flow details.

Of course, verifying the existence of a vulnerability requires testing it against the actual system, and thus a ToC provides a foundation for a human analyst to construct a Proof-of-Concept (PoC) of the vulnerability. Alternatively, human developers (who have access to implementation details) may use the ToC to timely reason about potential vulnerabilities and mitigations.

In this work, we use the discovery of indirect prompt

*Correspondence: zzhan645@asu.edu

injection (IPI) vulnerabilities to validate the feasibility of the no-box vulnerability analysis paradigm. IPIs occur when attacker-controlled instructions are injected into the context of a Large Language Model (LLM), where they subvert the model’s subsequent reasoning or actions [1]. Model Context Protocol (MCP) servers are particularly susceptible to IPI attacks. They provide standardized interfaces that enable LLM agents to use external tools and data [2]. Many data sources retrieved by MCP servers, such as public web pages and third-party APIs, are attacker-controllable, which enables IPI in vulnerable MCP servers.

With more than 10,000 active public MCP servers [3], auditing them for IPI vulnerabilities is difficult: Many MCP servers are proprietary; moreover, ethically testing for indirect prompt injection vulnerabilities is difficult as it can require persisting prompt-injection payloads into live systems that benign users may accidentally access (and thus exposing them to the payload). These constraints make no-box vulnerability analysis a good fit for finding IPIs in MCP servers.

We develop a prototype called MCPSEC that analyzes MCP servers for IPI vulnerabilities. MCP servers publicly expose their functionality as tools for an LLM to invoke, and we use only the tool metadata as input to MCPSEC. MCPSEC infers data flows from attacker-controlled data sources to the LLM context, and augments the data flows with plausible intermediate data entities and operations. It then considers the attacker context (ability to control the data entities, etc.) against the vulnerability preconditions, and outputs ToCs. We implement MCPSEC as a multi-stage LLM-based reasoning system that uses the LLM’s ability to reason about software architecture, data flows, and attacker capabilities.

We apply MCPSEC to 177 tools drawn from 20 widely deployed MCP servers across four categories: public web retrieval, browser automation, authenticated SaaS collaboration, and infrastructure and cloud platforms. On this dataset, MCPSEC generates 143 IPI ToCs. For comparison, a prompting-based LLM baseline generates 97 ToCs.

To further evaluate MCPSEC’s performance, we conduct a staged evaluation where human evaluators receive progressively stronger evidence and assess whether the output of MCPSEC agrees with the evaluators’ verdicts. Eventually, human evaluators confirm 95 tools as vulnerable to IPI through ethically controlled PoC testing. From metadata alone, MCPSEC recovers 94 of these vulnerabilities (98.9% recall), compared with 80 (84.2% recall) for the LLM baseline.

Our evaluation further surfaces a broader ecosystem finding: 132 (92.3%) tools examined do not apply any sanitization to external data on the path to the LLM context, indicating that safeguards against IPI vulnerabilities remain uncommon even in widely adopted commercial MCP servers. MCPSEC generates these ToCs at an average cost of \$1.65 per MCP server, demonstrating that no-box vulnerability analysis can identify concrete, runtime-confirmed attack surfaces at practical cost in a realistic setting.

Contributions. This paper makes the following contributions:

- We conceptualize *no-box vulnerability analysis*, a new vulnerability analysis paradigm for reasoning about potential vulnerabilities using only metadata.
- We develop a no-box vulnerability analysis framework that speculates about plausible implementations and vulnerability requirements and outputs *Theory-of-Concepts* for later validation when additional access is permitted.
- We implement this framework as a prototype called MCPSEC and evaluate it on 177 sampled tools from 20 widely deployed MCP servers. In a controlled setting, MCPSEC recovers 94 of 95 confirmed IPI vulnerabilities, demonstrating the feasibility of no-box vulnerability analysis in a realistic setting.

In the spirit of open science, we will release all analysis data and source code upon acceptance of this paper. We will release generated ToCs (and PoCs when available) after the developers of vulnerable MCP servers fix our reported bugs.

II. BACKGROUND

A. Existing Vulnerability Analysis Paradigms

Vulnerability analysis techniques are broadly classified by how much access the analyst has to the target system. White-box methods operate on source code or binaries, enabling precise reasoning about program behavior through static analysis, symbolic execution, and formal verification [4], [5], [6], [7]. Gray-box methods such as coverage-guided fuzzing [8], [9] require neither full source access nor a behavioral specification, but do require the ability to execute the target and observe coverage feedback. Black-box methods [10], [11] interact only with exposed interfaces and observe responses, requiring no internal access but still assuming the ability to query the target and receive outputs.

These three paradigms share a common prerequisite: the analyst must be able to observe or interact with the target system. They all attempt to speculate about and eliminate false vulnerability hypotheses by observing or interacting with the target system. However, when the target is proprietary, remotely hosted, or when direct interaction is infeasible, the analyst may access the metadata describing the target’s intended functionality. A new vulnerability analysis paradigm that does not necessitate implementation or deployment accesses can be useful in this setting.

B. Indirect Prompt Injection

Indirect prompt injection (IPI) is an injection attack that targets LLM-based applications [1]. It uses the fact that LLMs process instructions and external data within a unified context and cannot properly separate trusted instructions from untrusted content [12], [13]. The attacker often places a malicious payload in an external source that the LLM will retrieve, such as a web page, email, or document. Once the payload is ingested into the LLM context, the LLM may execute the attacker’s instructions and cause security consequences, such as invoking unauthorized tools or exfiltrating sensitive credentials.

What distinguishes IPI from canonical data flow vulnerabilities is its risk condition. A source-to-sink data flow alone is insufficient. A prompt injection payload is syntactically indistinguishable from other legitimate content the flow may transmit. Consequently, consuming untrusted external data does not necessarily imply vulnerability. An IPI vulnerability additionally requires that the data flow retain the instructional fidelity and semantic meaning of the payload.

IPI remains effective across retrieval channels, including web pages, documents, and tool outputs [14], and in agentic settings, a successful injection can cascade into chains of tool calls and data exfiltration [15], [16]. Defenses against IPI exist but remain inadequate against adaptive adversaries or human red teaming [17], [18], [19], [20], [21], [22], [23]. AgentFuzz [24], a pre-deployment detection approach, applies dynamic grey-box fuzzing, while AgentArmor [25] performs program analysis over runtime-generated agent traces. Both therefore require execution access to the target system, making them inapplicable to closed-source or remotely hosted servers.

C. Model Context Protocol

The Model Context Protocol (MCP) is a standard interface [2], [26] that lets LLM agents invoke external data sources and tools. A typical MCP server publishes a set of tools, each annotated with its tool name, a natural-language description, and a typed parameter list.

A typical MCP interaction proceeds through three phases. (1) *Server Registration*. The user registers an MCP server and provide tool metadata to a LLM Agent host. The metadata for each tool consist of a tool name, a natural-language description, and required input schema. (2) *Tool Invocation*. Given a user task, the LLM agent selects from available tools and invokes the tool call to the MCP server. (3) *Tool Response*. The server executes the tool call, which typically involves reading and processing data from external data sources, and returns a free-form result to the LLM host.

III. NO-BOX VULNERABILITY ANALYSIS

We formalize no-box vulnerability analysis as the metadata-only case of an information-centric framework. Through data flow speculation and risk analysis, we can produce testable attack scenarios for future validation when additional access becomes available.

A. Vulnerability Analysis as an Information-Centric Framework

Let T be a target system and let $\mathcal{M}$ denote the metadata available independently of implementation inspection or target interaction. Such metadata may include capability specifications, interface and schema declarations, natural-language descriptions, and end-user-facing documentation. While we assume that $\mathcal{M}$ truthfully describes the intended functionality of T , it is a partial and high-level abstraction of the underlying implementation.

We denote by $\mathcal{I}(\mathcal{M})$ the space of implementations I that implement the functionality described by $\mathcal{M}$. $\mathcal{I}(\mathcal{M})$ contains the true implementation I^* . Because metadata primarily

specifies *what* a system provides rather than precisely *how* that functionality is implemented, the size of $\mathcal{I}(\mathcal{M})$ may be effectively unbounded. However, many distinct implementations may be equivalent with respect to all properties that enable a particular vulnerability. Therefore, no-box vulnerability analysis focuses on vulnerability-relevant properties, such as data flows, transformations, validation mechanisms, and trust-boundary crossings, rather than reconstructing the exact deployed implementation.

Existing vulnerability analysis paradigms differ primarily in the additional evidence they obtain about the unknown implementation. In a white-box setting, the analyst has access to implementation-level artifacts such as source code, binaries, configuration, or runtime state. While reducing $\mathcal{I}(\mathcal{M})$, white-box access may still leave uncertainty about deployment states or external dependencies. In a black-box setting, the implementation remains hidden, but the analyst can interact with the target. Through interaction with T , the analyst can observe execution behavior, and each observation eliminates implementations that could not have produced the observed behavior.

Additional evidence restricts the size of $\mathcal{I}(\mathcal{M})$ and the likelihood that a vulnerability exists. A vulnerability is *necessary* if every implementation consistent with the available evidence contains it, and *plausible* if at least one implementation consistent with the available evidence contains it. In the absence of additional evidence, a vulnerability analyst can instead make assumptions about the implementation and evaluate whether those assumptions are required for exploitation. The analyst can then identify implementations that satisfy the required assumptions and formulate corresponding testable vulnerability hypotheses.

B. The No-Box Setting

The no-box setting is the metadata-only case of this framework, in which the analyst does not have any access to the target system aside from the metadata $\mathcal{M}$. Under the no-box setting, the analyst seeks to determine what vulnerabilities could arise in all $I \in \mathcal{I}(\mathcal{M})$ and under what implementation assumptions. The resulting vulnerability hypotheses can then be synthesized into concrete exploitation scenarios, which we call *Theory-of-Concepts* (ToCs).

C. Data Flow Speculation and Risk Analysis

Although $\mathcal{M}$ does not include implementation details, it may imply how data must move from one entity to another, which we call an *implied data flow*. While $\mathcal{M}$ can imply multiple concrete data flows, we further define an *irreducible data flow*: the minimal implementation-agnostic relationship preserved by every implementation consistent with $\mathcal{M}$.

For example, metadata specifying that a user-defined filter selects database records implies that the filter influences a database query. Its irreducible data flow is

$$\text{user-controlled filter} \xrightarrow{\text{query}} \text{database},$$

which contains only the essential data entities and transformations required to implement the functionality described by $\mathcal{M}$. An implementation may implement this relationship through raw query construction, parameterized statements, a restricted parser, or an ORM. This irreducible data flow makes SQL injection relevant to this example. However, the existence of an SQL injection vulnerability is contingent on the unsafe construction of SQL queries.

Using irreducible data flows, we hypothesize plausible vulnerabilities by annotating each irreducible data flow with hypothesized security-relevant conditions, such as the parsers used, data transformations, validation, and sanitization. The hypothesized conditions can be translated into data entities or transformations that enrich the irreducible data flows, producing various speculated data flows. We then conduct risk analysis by evaluating each speculated data flow against all preconditions necessary for exploitation. In the SQL query example, an injection is plausible when a speculated flow incorporates the user-controlled filter into a dynamically constructed query without effective parameterization or sanitization. We can hypothesize the existence of a vulnerability if at least one metadata-consistent speculated data flow satisfies all necessary preconditions for exploitation.

D. Theory-of-Concept

Based on both the data flow and the satisfied exploit preconditions, we can synthesize an exploitation scenario with additional assumptions, such as attacker roles. We define such an exploitation scenario as a *Theory-of-Concept* (ToC), the primary output artifact of no-box analysis and the most concrete vulnerability artifact that can be supported from $\mathcal{M}$ alone.

A ToC may specify the affected functionality and vulnerability class, the attacker role and required capabilities, the implementation-dependent assumptions on which exploitation is based, and the resulting security impact. It synthesizes speculated data flows and risk analysis results into a cohesive attack scenario. Unlike a *Proof-of-Concept* (PoC), a ToC does not execute or verify the scenario against a concrete implementation. However, when source code, binaries, or runtime access becomes available, a security analyst can validate the data flow and risk conditions described in the ToC to construct a PoC.

IV. FINDING IPI VULNERABILITIES IN MCP SERVERS

The implementation of many MCP servers are unavailable to the public. To measure the prevalence of such MCP servers, we enumerated the latest active MCP servers in the official MCP Registry [27], which contained 18,770 servers at the time of collection. We consider the implementation of an MCP server unavailable for analysis if it declares at least one remote endpoint but provides neither a released package nor a source repository URL. Under this definition, we found 3,007 MCP servers (16.0%) unavailable to the public. Additionally, 1,210 MCP servers (6.4%) required users to provide secrets (e.g., API keys) before configuring server connections. Although

TABLE I
TARGET SYSTEM ENTITIES, TRUST ZONES AND THREAT ASSUMPTIONS.

Entity	Trust Zones and Assumptions
Human User	Benign; issues legitimate tasks and does not intentionally inject malicious instructions.
LLM Agent	Follows the user's intent but may be influenced by instructions embedded in tool responses.
MCP Server	Non-malicious but potentially vulnerable; its code processes external content and may relay it without effective neutralization.
External Data Sources	Untrusted; their content may be influenced by a remote attacker before a tool retrieves it.

there might be cases where the MCP server implementation is available in other forms, our estimate shows that many MCP servers are unavailable for public inspection.

Next, we discuss how to discover IPI in MCP servers using the no-box vulnerability analysis paradigm.

A. MCP Servers as No-box Targets

Under the framework that §III-A describes, we model MCP servers as no-box targets, where: The target system T is an MCP server and its tool set. $\mathcal{M}$ consists of the published tool metadata available independently of implementation inspection or runtime interaction: tool names, natural-language descriptions, and JSON-Schema input specifications. The MCP specification mandates the exposure of these fields to help an LLM select tools and construct valid invocations [26]. Vulnerability sources are attacker-controllable external data, and vulnerability sinks are LLM contexts. We further assume that $\mathcal{M}$ describes the intended functionality of each MCP tool but does not include full (or any) implementation details.

B. Threat Model

An MCP-integrated LLM agent system comprises four relevant entities: a human user, an LLM agent, an MCP server, and the external data sources accessed by its tools. Table I summarizes the assumptions assigned to each entity.

We consider a remote third-party attacker who can place instructional content in a source that an MCP tool may subsequently consume. The attacker may publish public content, contribute content through a legitimate account on an authenticated platform, or influence an upstream artifact that is later retrieved by the user or tool. External provenance, rather than physical location, determines whether content is attacker-controlled. Cloned repository, installed package, or cached response remains in scope when a remote attacker influenced its contents before retrieval.

We distinguish an attacker-controlled *source* from the *medium* through which the tool accesses it. For example, an attacker may publish a web page indexed by a search engine, add content to a repository hosted on GitHub, or submit an artifact to a package registry without compromising the search engine, GitHub, or the registry itself. We therefore assume control over source content, not compromise of the hosting medium.

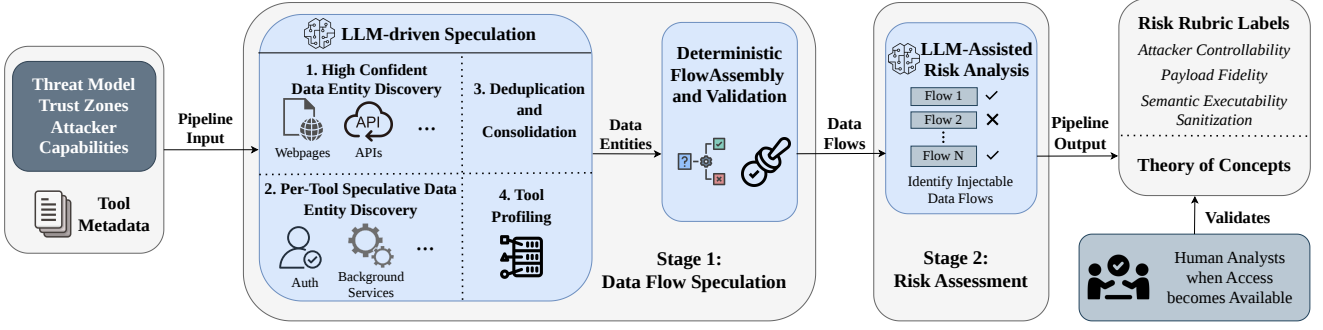


Fig. 1. MCPSEC pipeline.

The attacker cannot modify the MCP server, client, or LLM. Direct prompt injection, malicious MCP servers, transport compromise, and local compromise are outside our scope.

The attacker’s objective is to divert the agent from the user’s benign intent toward attacker-specified reasoning or actions through indirect prompt injection. The corresponding security property is *context integrity*: content controlled by a remote attacker must be mediated so that it cannot be interpreted as instructions that alter the agent’s behavior. An indirect prompt injection violation occurs when a legitimate tool invocation introduces attacker-controlled instructions into the LLM context and those instructions influence the agent contrary to the user’s intent.

C. Instantiating No-box Analysis for IPI

IPI instantiates the data flow speculation and risk analysis of §III-C with a compact source-to-sink structure. The source set $\mathcal{S}$ contains external origins whose content an MCP tool may consume, such as remote APIs, web pages, repository objects, messages, documents, and package registries. The sink set $\mathcal{K}$ contains the LLM context into which the agent incorporates a tool’s response. For a tool whose metadata entails that externally sourced content contributes to its response, the irreducible data flow is

$$s \in \mathcal{S} \rightarrow \text{MCP tool} \rightarrow k \in \mathcal{K}.$$

this relationship indicates that external content can be transmitted to LLM context, making the MCP server system vulnerable to indirect prompt injection attacks.

Data flow speculation elaborates the irreducible data flow into candidate implementation paths by hypothesizing implementation-dependent entities and operations, including the particular source, intermediate storage, parsing and transformation steps, field selection, validation, and sanitization. Each speculative path π denotes the subset of $\mathcal{I}(\mathcal{M})$ that realizes those assumptions. When the metadata suggests, but does not entail, that external content contributes to the response, the source-to-context relationship itself is treated as a plausible speculative path.

Risk analysis evaluates each speculative path against four indirect prompt injection-specific preconditions:

$$\Phi_{\text{IPI}} = \{\alpha_{\text{ctrl}}, \alpha_{\text{fid}}, \alpha_{\text{sink}}, \alpha_{\text{san}}\}.$$

1) *Payload Fidelity* (α_{fid}). This axis is grounded in information-flow theory [28] and taint analysis [29]: verbatim propagation preserves a direct dependency on attacker-controlled input, whereas derived outputs undermine taint through lossy transformations.

2) *Attacker Controllability* (α_{ctrl}). This axis collapses two CVSS 4.0 metrics [30] (*Attack Vector* and *Privileges Required*) into a single ordinal scale suited to MCP, distinguishing public/supply-chain channels [1] from platform-authenticated access. We deliberately omit CVSS’s *Attack Complexity*: indirect prompt injection via MCP tools is near-deterministic, so it does not discriminate among flows, and the residual uncertainty is already captured by the Semantic Executability and Sanitization axes.

3) *Semantic Executability* (α_{sink}). This axis reflects current LLMs lack architectural separation between instructions and data [12], with empirical evidence that the vast majority of production models fail to reliably distinguish the two [13].

4) *Sanitization* (α_{san}). This axis treats defensive measures along the data path as a spectrum: truncation or length limits preserve content semantics, whereas escaping, redaction, or content filtering can neutralize injected instructions before they reach the LLM context.

A speculative path that coherently satisfies these preconditions yields a metadata-grounded *plausible vulnerability hypothesis*. For each such hypothesis, an indirect prompt injection ToC identifies the affected tool, the attacker-controlled source and access required, the assumed source-to-context path and transformations, the relevant defensive assumptions, and the potential effect on the agent. The ToC makes the conditions underlying the plausible vulnerability explicit, making it useful for future PoC validation when more access is available.

V. MCPSEC: SYSTEM DESIGN

We develop MCPSEC, a prototype that instantiates the no-box vulnerability analysis framework to analyze MCP servers for indirect prompt injection vulnerabilities. As shown in Figure 1, MCPSEC performs a two-stage analysis to produce indirect prompt injection ToCs. Stage 1 speculates about the tools’ underlying implementations and assembles the data flows their metadata admits. Stage 2 assesses each flow against the four risk axes defined for IPI (§IV-C) and, for each

vulnerable flow, outputs a ToC describing a concrete attack scenario together with its per-axis risk labels.

A. Stage 1: Data Flow Speculation

Stage 1 takes tool metadata and produces a structurally validated, deliberately over-approximated set of candidate source-to-sink data flows. Because metadata does not reveal all implementation entities and edges, MCPSEC prompts the LLM to include entities and edges that are not stated explicitly but remain plausible given the tool’s functionality and the threat model. This deliberate over-approximation aims to maximize coverage of analyzable data flows and potential vulnerabilities. Stage 1 contains two steps: *LLM-driven Speculation*, which generates a set of hypothesized entities and edges, and *Deterministic Assembly*, which connects the entities and edges into data flows with injection-surface edges marked.

LLM-driven Speculation. First, entity discovery identifies the data entities: an initial LLM call extracts entities that can be inferred with high confidence from the complete set of tool metadata for the MCP server (e.g., external APIs, storage systems, and data pipelines), establishing a shared foundation across tools. Then, per-tool LLM calls propose additional speculative entities whose existence is plausible but not directly stated in the metadata (e.g., intermediate caches, authentication components, and background services). We instruct the LLM to admit low-confidence entities even from weak or analogical signals because tool metadata can be too concise to be self-explanatory. We retain such entities to maximize the number of candidate data flows that can be composed. A consolidation pass merges and deduplicates speculative entities across tools, and the consolidated set of entities forms a *data flow context*.

Tool profiling produces the edges: for each tool, an LLM call receives the data flow context and the tool metadata and produces a *tool profile*. The tool profile specifies the direction of data flow, the taint status of transferred data, and any transformations applied along the path. Entity discovery and tool profiling together realize the deliberate over-approximation principle and supply the input required for deterministic assembly.

Deterministic Flow Assembly and Validation. For each tool, a rule-based deterministic algorithm assembles the data flow context and tool profiles into data flows and assigns taint status at trust-zone crossings. The output edge into the LLM context is marked as an injection surface when the tool’s output carries untrusted data or when the output is verbatim and at least one inbound edge is untrusted. A structural validator then checks whether, in each assembled flow, injection edges connect valid nodes and a path exists from the attacker-controlled source to the tool’s output sink. Data flows that fail validation are removed. This deterministic stage serves as a structural safeguard. It ensures that malformed data flows cannot silently propagate into downstream reasoning when there is no oracle to detect them.

TABLE II
FOUR-AXIS RISK RUBRIC.

Axis / Level	Criterion
<i>Payload Fidelity</i>	
HIGH (verbatim)	Raw content preserved intact
LOW (derived)	Reduced to metadata or summaries
<i>Attacker Controllability</i>	
HIGH (public / supply-chain)	No authentication needed
MED (authenticated remote)	Platform account or role required
LOW (implausible)	Requires local access (Rule 0)
<i>Semantic Executability</i>	
HIGH (instructional)	Free-form NL or code
LOW (non-instructional)	Structured data, IDs, enums
<i>Sanitization</i>	
HIGH (none)	No filtering on data path
MED (partial)	Truncation or length limits
LOW (effective)	Escaping, filtering, or redaction

B. Stage 2: LLM-Assisted Risk Assessment

Stage 2 assesses each speculated and validated data flow against the risk assessment rubric, producing risk labels for every flow and a ToC for each flow identified as vulnerable. Stage 1 may produce multiple plausible data flows, but their vulnerability cannot be established without evaluating whether each flow satisfies the preconditions required for exploitation.

Risk Assessment Rubric. Intuitively, not all speculated data flows carry equal risk. A data flow that returns a numeric account balance poses a qualitatively different threat than one that relays verbatim issue comments into the LLM context. Thus, we define a risk assessment rubric to instantiate the four preconditions of §IV-C for MCP indirect prompt injection. The rubric quantifies the risk of a speculated data flow along four independent dimensions: *payload fidelity*, *attacker controllability*, *semantic executability*, and *sanitization*. Table II summarizes the axes and their levels. The labels allow analysts to prioritize more vulnerable flows, thereby compensating for Stage 1’s deliberate over-approximation.

Producing Outputs. For each validated data flow, a single LLM call receives the data flow, the tool metadata, the risk rubric, and the threat-model context from Section IV-B, which is embedded in the system prompt. LLM then analyzes the data flow against the rubric. If there exist one flow satisfies all four preconditions, LLM produces two outputs: risk labels for each axis of the *risk assessment rubric* and a ToC. The ToC is a short natural-language narrative that walks through a concrete indirect prompt injection scenario: the attacker model and the external surface under the attacker’s control; the planting step that places the payload where the tool will fetch it; and the fidelity with which the tool’s output preserves that payload en route to the LLM context. We produce the labels and the ToC within the same LLM call to bind the two outputs to a shared reasoning trace. If the tool has at least one such flow, MCPSEC flags the tool as vulnerable.

As an output of the MCPSEC pipeline, a ToC provides an

illustrative attack scenario that helps a downstream analyst test the risk rubric directly against the source code (§VII-C) or runtime instrumentation (§VII-D). The per-axis labels then serve as an audit scaffold against which the narrative can be independently cross-checked during evaluation (§VII-B).

VI. EXPERIMENT

We outline the experiment setup and outputs for the evaluation of MCPSEC on no-box vulnerability analysis below.

A. Evaluation Dataset

Our dataset contains 20 MCP servers and 177 sampled tools, detailed in Table III. We draw the servers randomly from 86 unique, verified servers compiled from the MCP Market top-100 leaderboard, ranked by GitHub stars [31], after removing deleted (3) and archived (1) repositories and deduplicating monorepo entries to avoid over-representation. Because several servers expose more than 100 tools, we randomly sample 10 tools per server, retaining all tools when fewer are available. Rather than imposing categories a priori, we group the servers by external data interaction pattern, yielding four categories that differ primarily in attacker controllability and source provenance: public web retrieval (A), browser-mediated interaction (B), authenticated SaaS collaboration (C), and infrastructure and cloud platforms (D).

This dataset is sufficient for evaluating no-box analysis on three counts. Every server accesses at least one external data source, the precondition our threat model requires for indirect prompt injection (§IV-B). The servers are high-profile: 12 of 20 are officially maintained by their service providers, and their stargazer counts (mean 9,479, median 3,764) reflect broad adoption, making them realistic targets on which defenses are more likely to already be in place. Finally, we select open-source servers so that later stages can ground MCPSEC’s predictions in source code and runtime behavior, while withholding that source and those deployments from MCPSEC during analysis to faithfully simulate the no-box setting.

B. Experiment setup and baselines.

All LLM calls in the MCPSEC use GPT-5.4, with high reasoning efforts and accessed through OpenAI API. To evaluate the performance of MCPSEC, we consider a trivial LLM baseline. For each target tool, the baseline makes one API call over the server’s sampled registration metadata. Its consolidated system prompt preserves MCPSEC’s threat model, four-axis risk rubric, and ToC requirements, but omits entity discovery, DFD reconstruction, deterministic graph assembly, and graph-conditioned reasoning as used in MCPSEC.

Among the 177 tools, the LLM baseline identified 97 plausible vulnerability candidates, and they constitute a subset of 143 identified by MCPSEC. With both methods generated corresponding risk labels and ToCs, we evaluate them against human annotated labels and evaluation in Section VII.

VII. EVALUATION

We previously framed vulnerability analysis as reasoning over possible implementations and making assumptions about underlying vulnerabilities under imperfect or limited observation of the target system (§III-A). No-box analysis begins with only metadata and therefore reasons over the full metadata-consistent implementation space. Access to source code, runtime observations, and PoC construction progressively increases observability, narrows the implementation space, and permits correspondingly stronger vulnerability claims. We organize our evaluation around this progression through four primary research questions.

A. Evaluation Design and Annotation Protocol

We first run MCPSEC on the raw tool corpus using only registration metadata and freeze its predicted flows, risk labels, and ToCs before human evaluation. Evaluators then annotate these predictions using progressively stronger evidence. Table IV summarizes the evidence, resulting labels, and purpose of each stage.

Staged risk assumption and ToC validation. We formulate RQ1–RQ3 around the 143 plausible vulnerability candidates identified by MCPSEC, which form a superset of the candidates identified by the LLM baseline.

In RQ1, human evaluators independently instantiate the no-box risk-assessment procedure by assigning the four risk-rubric labels from metadata alone. They then assess the plausibility of MCPSEC’s predicted injection source and the coherence of its generated ToC.

In RQ2, evaluators retain the metadata available in RQ1 and additionally gain access to the source code of each target tool, allowing them to perform white-box static analysis. Their judgments determine whether the metadata-derived ToCs and risk labels remain plausible under source-code evidence, thereby measuring how well MCPSEC’s speculative hypotheses capture the target’s actual vulnerability-relevant data flows, transformations, and sanitization mechanisms.

In RQ3, evaluators retain the evidence available in the previous stages and can additionally execute and instrument each target tool, allowing them to perform white-box dynamic analysis. Their runtime observations determine whether the ToC-identified source-to-sink flows actually manifest during execution and with what payload fidelity, thereby measuring how well MCPSEC’s metadata-derived predictions correspond to the target’s observed behavior. RQ3 uses benign tool parameters and validates runtime reachability rather than end-to-end exploitability, which is evaluated separately in RQ4.

Importantly, RQ1–RQ3 use either purely static evaluation or dynamic evaluation with only benign parameters.

Vulnerability labeling. To measure vulnerability detection performance more precisely, we attempt to establish vulnerability labels for each of the 177 sampled tools by constructing corresponding proof-of-concept exploits (PoCs). Because some target MCP servers connect to commercial services, we construct PoCs only for tools that are suitable for ethical and technically permissible controlled prompt-injection trials.

TABLE III

EVALUATION TARGET MCP SERVERS, GROUPED BY CATEGORY AND SORTED BY TOOL COUNT (DESCENDING) WITHIN EACH CATEGORY. GITHUB STARS AS OF APRIL 2026. *LoC*: SOURCE LINES OF CODE COUNTED WITH `cloc` AND EXCLUDING TESTS, FIXTURES, DOCUMENTATION, SKILLS, AND DEPENDENCIES. *Paid*: WHETHER THE UPSTREAM API REQUIRES A PAID PLAN TO TEST ALL SAMPLED TOOLS.

#	Server	Repository	Commit	Tools	LoC	Stars	Provider	Paid	Access
A: Search, Web, and Document Retrieval (47 total tools) — fetch public web content, search results, or documents									
1	Firecrawl	firecrawl/firecrawl-mcp-server	6fef044	14	1,152	6,116	Official	No	
2	arXiv	blazickjp/arxiv-mcp-server	51ebf3a	10	1,966	2,573	Community	No	
3	Exa Search	exa-labs/exa-mcp-server	8df65b0	10	2,381	4,286	Official	No	
4	Brave Search	brave/brave-search-mcp-server	52590e9	6	2,586	924	Official	Yes	
5	Tavily	tavily-ai/tavily-mcp	238f6fd	5	791	1,812	Official	Yes	
6	Context7	upstash/context7	658ec67	2	715	53,372	Official	No	
B: Browser-Mediated Web Interaction (117 total tools) — observe or manipulate live browser state									
7	Skyvern	Skyvern-AI/skyvern	e565094	49	6,207	21,314	Official	No	
8	Playwright	executeautomation/mcp-playwright	2349c28	33	3,929	5,455	Community	No	
9	Chrome DevTools	ChromeDevTools/chrome-devtools-mcp	b1684c6	29	8,799	36,597	Official	No	
10	Browserbase	browserbase/mcp-server-browserbase	f6bd321	6	1,356	3,277	Official	Yes	
C: SaaS Collaboration and Communication (388 total tools) — read user-authored content via authenticated APIs									
11	GitLab	zereight/gitlab-mcp	c393d1e	117	10,053	1,392	Community	No	
12	Google Workspace	taylorwilsdon/google_workspace_mcp	35fcc94	114	18,967	2,180	Community	Yes	
13	Atlassian	sooperset/mcp-atlassian	8e84d74	73	20,621	4,993	Community	No	
14	GitHub	github/github-mcp-server	2aleaac	41	19,437	29,151	Official	No	
15	Notion	makenotion/notion-mcp-server	3bef7ad	22	1,170	4,250	Official	No	
16	Slack	korotovskys/slack-mcp-server	b24b1b0	12	7,119	1,550	Community	No	
17	Discord (Klavis)	Klavis-AI/Klavis	340b6cf	9	639	5,715	Community	No	
D: Infrastructure, Cloud, and Data Platforms (60 total tools) — read operational, registry, or project-state data									
18	Supabase	supabase-community/supabase-mcp	1cd04f0	29	14,581	2,633	Community	Yes	
19	Sentry	getsentry/sentry-mcp	0941c50	22	16,093	658	Official	No	
20	Terraform	hashicorp/terraform-mcp-server	617ba91	9	7,200	1,334	Official	No	

TABLE IV

EVALUATION PROGRESSION FROM METADATA-ONLY ASSESSMENT TO CONTROLLED VULNERABILITY VALIDATION.

Stage	Evidence	Annotation	Purpose
RQ1	Metadata	ToC plausibility and rubric agreement	Human no-box assessment
RQ2	Metadata + source code	ToC plausibility and rubric agreement	Static implementation validation
RQ3	Metadata + source code + benign execution	Reachability and runtime fidelity	Dynamic data flow reachability test
RQ4	Human constructed PoCs for suitable tools	Tool vulnerability labels	Vulnerability detection performance

Details of this procedure is shown in Section B. We label 95 tools as vulnerable to indirect prompt injection attacks, and rest of tools are filtered out because they are either non-vulnerable or not suitable for PoC construction.

We then design RQ4 to compare the vulnerability predictions of MCPSEC and the baseline against these labels, enabling measurement of precision, recall, false positives, and false negatives.

Data annotation quality. A subset of the authors, all with experience in data flow analysis and security research, performs the annotations; each evaluator covers three or four servers. To assess the consistency of our human labeling, additional evaluators independently annotate a subset of the data. Before any discussion or reconciliation, exact agreement was 83.3% across all paired categorical judgments. When separated by annotation type, agreement was 89.6% for judgments directly assessing the generated ToCs, 80.1% for rubric and data-path labels independently assigned from the available metadata or source-code evidence, and 82.5% for

observations derived from runtime evidence. After calculating these pre-adjudication statistics, each evaluator pair reviewed its disagreements using the shared annotation guidelines and evidence and then recorded a consensus label and rationale; unresolved cases were referred to a third evaluator. The agreement values reported here are calculated from the original, pre-discussion annotations.

B. RQ1: Are MCPSEC-Generated ToCs and Risk Labels Plausible from Metadata Alone?

Evaluation target. Before source-code or runtime access becomes available, we assess whether the risk hypotheses and ToCs generated by MCPSEC are internally consistent and grounded in the available tool metadata. For each plausible vulnerability candidate identified by MCPSEC, the evaluators observe only the tool metadata and assess the generated artifacts along six dimensions: (1) whether the identified external data source is plausible for the tool, (2–5) whether the tool description supports each of the four risk-rubric labels (§V-B), and (6) whether the ToC is coherent and plausible.

TABLE V

RQ1 RESULTS USING REGISTRATION METADATA ALONE. THE FIRST COLUMN REPORTS ALL 143 FLOWS IDENTIFIED BY MCPSEC. MCPSEC AND THE LLM BASELINE ARE COMPARED ON THE SAME 97 TOOLS.

Measure	MCPSEC (N = 143)	MCPSEC (N = 97)	LLM Baseline (N = 97)
<i>Metadata-grounded plausibility</i>			
External source plausible	96.5%	–	–
ToC plausible	85.3%	–	–
<i>Agreement with human risk labels</i>			
Payload_Fidelity	83.2%	83.5%	59.8%
Semantic_Executability	84.6%	90.7%	88.7%
Sanitization	79.0%	82.5%	63.9%
Attacker_Controllability	84.6%	90.7%	85.6%
Mean (four axes)	82.9%	86.9%	74.5%

Results. Across the 143 plausible vulnerability candidates identified by MCPSEC, evaluators find that MCPSEC identifies a plausible external data source in 96.5% of cases, produces an internally coherent attack narrative in 85.3%, and achieves 82.9% mean agreement with human judgments across the four risk-rubric axes.

On the 97 tools for which both methods produce predictions, MCPSEC achieves 86.9% mean agreement, compared with 74.5% for the LLM baseline, a gain of 12.4 percentage points. Under the stricter requirement that all six dimensions agree simultaneously, MCPSEC achieves 60.8% on the matched subset. Because the baseline receives the same registration metadata, threat model, risk rubric, and ToC requirements, this improvement indicates that MCPSEC’s entity discovery, data flow reconstruction, deterministic graph assembly, and graph-conditioned reasoning provide additional value beyond direct prompting.

As shown in Table V, the high accuracy in identifying plausible injectable sources suggests that, for the sampled MCP servers, registration-time metadata provide sufficient signals for the LLM to infer the type and injectability of external data sources. The average decrease of 8.6% from source plausibility to ToC coherence indicates that, although MCPSEC reliably identifies injection sources, its end-to-end attack hypotheses can be over-specified or misaligned with a tool’s behavior, particularly when concise or informal metadata do not constrain the richer attack surface hypothesized by the ToC.

Beyond source plausibility and ToC coherence, Table V reports agreement between MCPSEC and the evaluators for each rubric axis. The highest-agreement axes are Semantic_Executability (84.6%) and Attacker_Controllability (84.6%), reflecting that output form and attacker access are often directly inferable from tool descriptions. The lowest-agreement axis is Sanitization (79.0%), for which the dominant disagreement consists of MCPSEC labeling sanitization as partial while evaluators rate it as none (29 cases). This pattern suggests a difference in how human evaluators and MCPSEC interpret what constitutes partial sanitization from the available metadata, particularly whether certain

TABLE VI

RQ2 RESULTS UNDER SOURCE-CODE EVIDENCE.

Measure	MCPSEC (N = 143)	MCPSEC (N = 97)	LLM Baseline (N = 97)
<i>Source-code-grounded plausibility</i>			
External source plausible	134 (93.7%)	–	–
ToC: Plausible	105 (73.4%)	–	–
ToC: Partially Plausible	3 (2.1%)	–	–
ToC: Implausible	35 (24.5%)	–	–
<i>Agreement with human risk labels</i>			
Payload_Fidelity	66.4%	68.0%	52.6%
Semantic_Executability	60.8%	64.9%	60.8%
Sanitization	71.3%	76.3%	61.9%
Mean (three axes)	66.2%	69.8%	58.4%

transformations or processing steps provide meaningful sanitization against the hypothesized injection. For Attacker_Controllability (84.6%), MCPSEC most often labels flows as authenticated_remote when evaluators downgrade them to implausible (11 cases). These disagreements are concentrated among write-only tools for which evaluators find no plausible return-path injection surface.

Conclusion. Registration-time metadata provide sufficient information for No-box analysis to construct vulnerability hypotheses that are largely plausible and internally consistent under human evaluation. Agreement is weaker for properties that depend on implementation details not directly exposed by metadata, most notably the prediction of sanitization. Overall, RQ1 measures the *metadata-grounded plausibility* of MCPSEC’s analysis, but not whether its hypotheses accurately characterize the underlying implementation; we examine this distinction using source-code evidence in RQ2.

C. RQ2: Are MCPSEC-Generated ToCs and Risk Labels Plausible Under Source-Code Evidence?

Evaluation target. In RQ2, evaluators gain access to the source code and assess whether the ToCs and risk labels remain plausible under this additional evidence. The questions we ask the evaluators therefore shift slightly from those in RQ1. Rather than asking how plausible the LLM-predicted external data sources are, we ask whether those sources are present in the source code. Because source code provides no additional evidence about Attacker_Controllability, we reuse the evaluators’ judgments from RQ1 for this axis. Evaluators assess whether the remaining three risk-rubric labels hold under source-code evidence. Finally, evaluators combine these factors to produce an overall ToC plausibility verdict: *Plausible*, *Partially Plausible*, or *Implausible*.

Results. Source-code validation shows that 75.5% of the predicted ToCs remain plausible or partially plausible in the actual implementations: 73.4% are judged *Plausible*, 2.1% are judged *Partially Plausible*, and 24.5% (35) are judged *Implausible*. MCPSEC also correctly identifies the external data source for 93.7% of tools. Among the 9 incorrect source identifications, 9 result in *Implausible* ToCs, indicating that

source-identification errors strongly contribute to implausible ToCs.

Source-code validation also reduces agreement on the predicted risk conditions (Table VI). Agreement decreases most substantially for *Semantic_Executability*, from 84.6% in RQ1 to 60.8% in RQ2. The next-largest decrease is for *Payload_Fidelity*, from 83.2% to 66.4%, while *Sanitization* remains relatively stable. On the subset shared with the LLM baseline, MCPSEC retains 69.8% mean agreement with source-code-grounded judgments, compared with 58.4% for the baseline, an improvement of 11.4 percentage points.

To understand why MCPSEC-predicted ToCs do not fully survive implementation validation, we examine the 38 flows rated *Implausible* or *Partially Plausible*. A tool may exhibit multiple implementation-level factors: *data not preserved* (24/38, 63.2%), where external content is reduced to identifiers, counts, or status codes; *output transformed* (22/38, 57.9%), where summarization or reformatting prevents payload preservation; *effective sanitization* (6/38, 15.8%); and *wrong source* (9/38, 23.7%). The dominant failure mode therefore remains incidental data loss: functionality-driven processing often prevents external content from reaching the output in a payload-preserving form, although incorrect source identification is also a nontrivial factor.

Explicit defenses, by contrast, are uncommon but effective. 92.3% (132) of tools apply no sanitization, 3.5% (5) apply partial filtering, and only 4.2% (6) apply effective sanitization. Of these 6 effectively sanitized tools, 6 receive *Implausible* verdicts. Thus, high ToC plausibility reflects the scarcity of effective defenses rather than their inability to disrupt predicted injection paths.

Conclusion. Source-code evidence largely confirms MCPSEC’s metadata-only hypotheses, with the majority of ToCs remaining at least partially plausible. Where ToCs fail, the cause is rarely active defense but rather incidental data loss through discarded or transformed outputs. This suggests that current MCP servers lack defenses designed specifically against indirect prompt injection and that their limited resilience often arises from functional design rather than deliberate security precautions.

D. RQ3: Do ToC-Identified Data Flows Reach the LLM Context at Runtime?

Evaluation target. RQ1 and RQ2 establish that MCPSEC’s hypotheses are largely plausible from metadata and survive source-code scrutiny. Yet neither confirms whether attacker-controllable data actually traverses the predicted flow during execution.

RQ3 therefore evaluates the identified flows at runtime and determines whether external data reach the sink in a form that preserves injection-capable content. For each tool, we first identify two logging points in the source code: one immediately after the remote API call that retrieves data from the ToC-identified external source (*source*) and another at the tool’s return statement (*sink*). Evaluators then construct

TABLE VII
VERDICT DERIVATION FOR RQ3.

Reachability	Output form	Verdict
complete	verbatim text	Reachable-Verbatim
complete/partial	derived	Reachable-Modified
metadata only	(any)	Reachable-Metadata
none	(any)	Not Reachable

TABLE VIII
RQ3 RESULTS FROM DYNAMIC EXECUTION EVIDENCE.

Measure	<i>n</i> / <i>N</i>	Percentage
<i>Runtime verdict</i>		
Reachable-Verbatim	22/143	15.4%
Reachable-Modified	87/143	60.8%
Reachable-Metadata	18/143	12.6%
Not Reachable	16/143	11.2%
<i>Consistency with RQ2 ToC verdicts</i>		
RQ2 Plausible $\wedge$ RQ3 Reachable	98/105	93.3%
RQ2 Implausible $\wedge$ RQ3 Reachable	27/35	77.1%

realistic, non-exploitative parameters that successfully invoke each MCP tool and retrieve data from the identified external source. We observe whether the source data reach the sink and at what fidelity.

Each flow receives a verdict based on two observed properties: how much external data reach the output (*reachability*: complete, partial, metadata-only, or none) and in what form they appear (*output form*: verbatim, formatted, JSON-wrapped, or derived). Table VII maps the cross-product to a verdict label.

Results. Across the 143 flows, 15.4% (22) are classified as Reachable-Verbatim, 60.8% (87) as Reachable-Modified, and 12.6% (18) as Reachable-Metadata; the remaining 11.2% (16) are classified as Not Reachable. In total, 76.2% (109/143) of flows can carry attacker-controllable data to the LLM context, either verbatim (22) or through structural transformations such as JSON reformatting and field selection that preserve injection-capable content (87).

Cross-referencing RQ2 plausibility with RQ3 verdicts reveals a positive association. Among 105 *Plausible* flows identified in RQ2, 98 (93.3%) are runtime-reachable. Among 35 *Implausible* flows, 27 (77.1%) are nonetheless reachable: 15 are Reachable-Modified, 6 are Reachable-Metadata, and 6 are Reachable-Verbatim. Runtime reachability in these cases shows that some external data can reach the sink, but their fidelity is insufficient to support an attack. Not Reachable flows are concentrated among write-only or UI-interaction tools, such as the *click* tool of Chromium DevTools MCP, that return only input-derived confirmations or error responses.

We further ask whether the MCP tool metadata provides enough information to reject confirmation-only tools before runtime testing. The current specification [26] does not require tool descriptions to explain what a tool returns to the LLM. For example, a tool described only as “mark a conversation

as read” might return the conversation’s content or merely a success message; the metadata does not distinguish between these security-relevant cases. We consider tool metadata to contain such a cue when a keyword such as *returns*, *output*, or *response* appears alongside a term matching one of the tool’s observed runtime output fields. 66 of 143 flows (46.2%) carry such a cue; 77 do not. Tools with output cues in their metadata are judged *Plausible* in 86.4% of cases, compared with 62.3% for tools without such cues ($\phi=0.271$, $p \approx 0.001$). The RQ1 data-fidelity agreement is 89.4% for tools with cues and 77.9% for tools without, but the observed difference is not statistically significant ($p \approx 0.067$). Output-shape cues therefore help evaluators judge *plausibility* but are less informative about how MCPSEC labels external-data *fidelity*.

Conclusion. Runtime instrumentation corroborates MCPSEC’s per-flow predictions: 76.2% of flows reach the LLM context in potentially instructional form (verbatim or derived), with strong agreement between RQ2 plausibility and RQ3 reachability. Reachable-but-implausible flows do not necessarily represent MCPSEC failures; they involve tools that pass external data through in forms that lack a usable injection channel, consistent with the incidental-data-loss patterns of RQ2. Output-shape metadata is the single strongest predictor of plausibility. Thus, extending the MCP specification with faithful output schemas could allow MCPSEC to reject confirmation-only false positives without runtime validation.

E. RQ4: Does MCPSEC Recover Confirmed Prompt-Injection Vulnerabilities?

Evaluation target. RQ1–RQ3 assess MCPSEC’s vulnerability analysis on the 143 plausible vulnerability candidates it identified under progressively stronger evidence. These evaluations cannot account for vulnerabilities that the pipeline never surfaced. We therefore return to the full 177-tool dataset and attempt an ethically controlled PoC against every suitable target tool. This yields a subset of 95 tools, all of which are confirmed vulnerable to indirect prompt injection. We score both MCPSEC and the LLM baseline against this subset to measure whether either method can recover confirmed vulnerabilities through no-box analysis.

As discussed earlier in §VII-A, the remaining 82 tools are not established negatives. A tool goes unconfirmed either because it genuinely carries no injection channel or because no PoC could be attempted against it. Recall is therefore the only metric this experiment determines exactly; the reported precision is a lower bound for both methods.

Results. MCPSEC recovers 94 of the 95 confirmed vulnerabilities (98.9%); the LLM baseline recovers 80 (84.2%). With only one missed vulnerability, MCPSEC demonstrates that it can effectively discover confirmed indirect prompt injection vulnerabilities in MCP servers.

The LLM baseline also performs strongly, indicating that no-box vulnerability analysis remains effective even with reduced reasoning effort. Notably, the 15 misses are not dis-

TABLE IX
RQ4 VULNERABILITY RECOVERY PERFORMANCE ON THE 95 TOOLS WITH CONFIRMED INDIRECT PROMPT INJECTION VULNERABILITIES. *Recovered* COUNTS CONFIRMED VULNERABILITIES IDENTIFIED THROUGH NO-BOX ANALYSIS; *Missed* COUNTS THOSE NOT IDENTIFIED. PRECISION IS REPORTED AS RECOVERED/FLAGGED, WHERE THE DENOMINATOR IS THE NUMBER OF TOOLS EACH METHOD FLAGGED ACROSS THE FULL SAMPLE.

Method	Recovered	Missed	Recall	Precision
MCPSEC	94	1	98.9%	94/143 (65.7%)
LLM Baseline	80	15	84.2%	80/97 (82.5%)

tributed evenly: 11 occur in Category C, where external content arrives through authenticated co-tenant channels whose attacker model is not apparent from registration metadata alone. Graph-conditioned reasoning over reconstructed data flows recovers exactly the cases that metadata-only prompting treats as trusted internal state.

The ordering reverses for precision: 65.7% for MCPSEC and 82.5% for the baseline. This precision–recall tradeoff follows from MCPSEC’s design. The data flow speculation stage in MCPSEC (§V) requires the LLM to reconstruct plausible entities and edges from tool metadata before judging vulnerability, rather than produce a verdict directly from the metadata. This additional reasoning, together with deliberate *over-approximation*, helps MCPSEC achieve near-perfect recall at the cost of lower precision, a tradeoff intrinsic to the no-box vulnerability analysis paradigm.

Conclusion. Together with the baseline’s strong recall, these results demonstrate that no-box analysis can recover real indirect prompt injection vulnerabilities without source-code or interaction access, while MCPSEC’s structured, over-approximating analysis substantially reduces missed vulnerabilities. This increased coverage comes at the expected cost of more false positives, reflecting the fundamental tradeoff of reasoning conservatively under implementation uncertainty.

F. Case Studies

We highlight two representative cases from our evaluation that illustrate the strengths and limitations of MCPSEC’s analysis. The first shows a metadata-derived ToC that leads to a working exploit; the second shows a plausible ToC that an independent analyst cannot validate through black-box testing without privileged access.

a) *From page text to bidirectional exfiltration channel.*: MCPSEC flags `evaluate_script` from `chrome-devtools-mcp` by identifying a flow in which an attacker-controlled page supplies a JavaScript function that the agent invokes to read same-origin browser state and transmit it to a remote endpoint. We developed a PoC from this ToC using the out-of-the-box `chrome-devtools-mcp` server and a locally hosted LLM agent. Additional details, including the tool metadata, appear in Appendix A. A setup phase seeds prior-session credentials into the browser’s `localStorage` and cookies, simulating persistent state from a legitimate SSO. We host a demo webpage containing

a simple prompt-injection payload that instructs the LLM to invoke the same tool with parameters designed to exfiltrate these credentials. Given the benign task “summarize this page,” the agent follows the injected instruction and calls `evaluate_script` with the supplied function, which sends `localStorage` and `document.cookie` via an authenticated POST to a remote server. We run the Chromium DevTools MCP server with its default Chromium configuration.

This case validates MCPSEC at two levels. First, it turns a metadata-derived ToC into a reproducible exploit on an unmodified target. Second, it demonstrates the worst-case impact: injected page content can induce JavaScript execution with Chromium’s network privileges, enabling end-to-end private data access and exfiltration.

b) *Black-box testing requires privileged access:* `slack/usergroups_update`.: MCPSEC produces a ToC for `slack/usergroups_update` based on a flow in which an attacker first places instructions in the description of a Slack user group. If another user later updates a different group property, such as its name or handle, the tool may return the pre-existing description to the LLM along with the updated group metadata. While our source-code inspection supports this ToC, testing this path requires more than access to the MCP interface. Slack user groups are available only in paid workspaces, and the endpoint requires an OAuth token with the `usergroups:write` scope and a workspace role allowed to manage user groups.¹ Without this access, a black-box test cannot obtain a successful tool response. By contrast, MCPSEC can still identify this conditional flow from metadata and record the assumptions that a later authorized test must check. For such permission-gated tools, a metadata-derived hypothesis may be the strongest result available to an independent analyst.

VIII. DISCUSSION

A. No-Box Analysis as a Front-End Triage Layer

Our results demonstrate that vulnerability hypothesis generation does not require implementation access. Across 20 widely deployed MCP servers, registration-time metadata alone identified 143 candidate injection flows, 75.5% of which were confirmed reachable at runtime. The analysis cost \$1.65 per server on average, required no credentials or source code, and involved no risk of disturbing live services. We thus prove that no-box is not only feasible but also practical at ecosystem scale, filling the gap between existing vulnerability analysis paradigms.

However, its feasibility and performance may depend on two dimensions. Across different *vulnerability classes*, the approach works best when exploitation conditions can be defined using metadata alone. This makes SQL injection, XSS, SSRF, and path traversal highly suitable. Memory corruption and race conditions, however, are less suitable since their key triggering conditions are deeply obscured by low-level

implementation details. Along the *target system* dimension, the observable metadata need to be semantically meaningful to infer an irreducible data flow. MCP servers are a favorable case, and the method may also apply to other software ecosystems such as browser extensions, CI/CD actions, and serverless applications, whose metadata expose entry points, data interfaces, and privileged capabilities.

At ecosystem scale, no-box analysis can serve as a first-pass screening control at the point of server listing or installation, with minimal deployment or configuration overhead. The produced Theory-of-Concepts can then be used to prioritize further white-box, gray-box, or runtime validation, or to warn users and hosts about potential vulnerabilities.

B. A Conceptual Gap in MCP Security

Many of the PoCs constructed in RQ4 required only placing a fixed injection payload in an attacker-controllable external source, rather than designing a sophisticated payload or attack chain. Our analysis suggests that such straightforward attacks remain possible because sanitization is largely absent from the path between external data and the LLM context: 92.3% of the examined tools apply no sanitization along this path. These implementations appear to treat a valid API response as safe model input, placing the trust boundary at the API response rather than at the external data source. This mismatch helps explain why simple injection attacks remain viable across our dataset.

Developing safer MCP server implementations should begin with output minimization. When a task requires only an identifier, count, or status, returning external free-form content creates an unnecessary injection surface and should be avoided. When such content is necessary, servers should preserve its provenance and separate it structurally from server-generated output so that hosts can distinguish data from instructions [19], [20]. Heuristic sanitization methods such as length limits and rule-based filters provide only partial protection because short or adaptive payloads may bypass them; additional LLM guardrails should be considered to support defense in depth [17], [21]. MCP developers should therefore treat external content as the trust boundary and be more mindful of the security implications in their implementations.

C. Limitations

Our case study benefits from a property that will not hold uniformly elsewhere: MCP registration metadata are unusually standardized and semantically informative, and our evaluation deliberately targets high-profile servers with external-content-bearing tools rather than a neutral cross-section of the ecosystem. In ecosystems with poorer specifications or weaker interface conventions, the same method may produce substantially more false positives or fail to generate useful hypotheses.

Our evaluation may not fully represent the broader MCP ecosystem or its deployed implementations. We sample at most ten tools from 20 high-profile servers selected because they access external data, so our results do not estimate

¹<https://docs.slack.dev/reference/methods/usergroups.update>

vulnerability prevalence across the MCP ecosystem. While the selected servers are open source, real-life hosted deployments, configurations, and proprietary middleware may differ from the implementations examined.

MCPSEC reasons about each tool in isolation and models a single external-source-to-tool-to-context path, leaving multi-hop attack chains, host-side prompt construction, cross-turn memory accumulation, and inter-server compositions out of scope. LLM-based data flow speculation also remains non-deterministic. The deterministic assembly step and repair loops substantially constrain model behavior but do not eliminate variance entirely.

Finally, we recognize that implementing MCP servers free from IPIs is still very difficult, and the high number of vulnerable MCP servers in the real world may make no-box vulnerability analysis appear more effective than it really is. We will investigate the generalizability of no-box vulnerability analysis on other vulnerability classes and targets in the future.

IX. RELATED WORK

A. Vulnerability Detection Paradigms

Vulnerability analysis techniques are broadly classified by the degree of access the analyst has to the target system, including white-box [32], [33], [34], [6], grey-box [35], [9], [36], [37], and black-box [11], [10], [38], [39] approaches.

White-box analysis, such as static analysis, symbolic execution and formal verification, has access to the source code or binary of the target system, allowing it to reason about the behavior of the system and identify potential vulnerabilities. Static analysis has long been used for vulnerability detection, evolving from early tools for C programs [40], [41] to modern approaches that leverage program analysis techniques such as data flow, taint tracking, and points-to analysis [42], [43], [44]. These methods have been applied across diverse domains, including web applications and mobile systems [45], [46]. Widely adopted frameworks such as CodeQL, Simgrep, and Joern demonstrate the practical impact of static analysis in real-world vulnerability detection. Complementary white-box techniques, including symbolic execution and model checking, enable deeper reasoning about program behavior by exploring execution paths or exhaustively verifying system properties [5], [47].

Although some white-box techniques (e.g., symbolic execution) can operate on binaries, they often incur significant overhead and struggle to scale. In contrast, grey-box approaches such as coverage-guided fuzzing improve scalability by relying on lightweight runtime feedback rather than full program analysis [8]. Modern fuzzers, popularized by AFL [48], have significantly advanced vulnerability discovery and have been widely applied across diverse software domains [9], [49], [50].

Black-box testing requires minimal access, identifying vulnerabilities by interacting with exposed interfaces such as network services and web applications. Tools like Nmap and ZMap exemplify this approach, and extensive work has explored black-box vulnerability detection in web systems [51], [38], [10], [52]. Despite differences in access assumptions, all

existing paradigms rely on the ability to observe or interact with the target system. In contrast, we introduce *no-box* vulnerability analysis, which removes this requirement and instead reasons about vulnerabilities solely from externally observable metadata.

While the term *no-box* has appeared in prior literature [53], [54], [55], those works focus on attack construction under minimal target knowledge, primarily in the domains of deep neural networks and Large Language Model (LLM) based agents. To the best of our knowledge, ours is the first work to formulate no-box vulnerability analysis as a distinct problem and to develop a framework tailored to this setting.

B. Large Language Models and Indirect Prompt Injection

Large Language Models (LLMs) exhibit strong generalization across tasks but are inherently vulnerable to prompt injection because they process instructions and external data within a shared context, without a clear architectural boundary [56], [57], [13]. This allows adversaries to inject malicious content that overrides the intended task. Indirect prompt injection (IPI) [1] extends this threat by delivering such payloads through external sources (e.g., web pages, databases, or tool outputs) that the model retrieves during execution.

IPI remains effective across diverse input channels, including retrieved content and multimodal data [14]. In agentic settings where LLMs invoke external tools, successful injections can propagate through tool chains, leading to data exfiltration or unintended actions, and empirical studies show that even advanced models remain highly susceptible [58], [59]. Existing defenses focus on runtime detection or instruction–data separation [21], [18], but adaptive attacks continue to bypass these mechanisms [17].

Pre-deployment vulnerability detection for LLM agents remains underexplored compared to runtime defenses and benchmarking. Recent approaches such as AgentFuzz [24] and AgentArmor [25] analyze agents to identify injection vulnerabilities, but both rely on access to a running system (e.g., execution feedback or runtime traces). This assumption limits their applicability in settings where the LLM or surrounding system is inaccessible, such as closed-source or restricted Model Context Protocol deployments.

X. CONCLUSION

This paper introduces *no-box vulnerability analysis*, a paradigm for reasoning about potential vulnerabilities using only metadata when source code, binaries, or runtime interaction are unavailable. For this new paradigm, we develop a framework that uses metadata as input, infers irreducible data flows, augments them with plausible implementation details, and produces Theory-of-Concepts for downstream validation. We instantiate this framework as MCPSEC, a two-stage pipeline for detecting indirect prompt injection vulnerabilities in MCP servers using only registration-time tool metadata. With fine-grained evaluation, we showcase MCPSEC achieves high agreement with human evaluators regarding the plausibility of the generated ToCs and verdicts of

indirect prompt injection vulnerability preconditions. Notably, MCPSEC recovers 94 of 95 confirmed vulnerabilities (98.9% recall), compared with 80 (84.2% recall) for the LLM baseline.

These results show that metadata alone can be used to identify vulnerable data flows, enabling practical vulnerability analysis when conventional access is unavailable. No-box analysis can also complement existing paradigms as a front-end triage layer with low deployment and configuration overhead, directing subsequent white-box or runtime validation. We hope this work motivates further research on no-box vulnerability analysis across other vulnerability classes and metadata-rich software ecosystems.

ETHICAL CONSIDERATIONS

The main ethical concern in this work lies in the identified and verified reachable data flows for the MCP server tools that we evaluated against.

Most of the flows identified, even as Reachable Verbatim or Reachable Derived, may not be sufficiently considered as vulnerabilities. We only deduce such verdicts by using a benign set of input parameters to showcase reachability of data flows, not proving that such injectable flows are indeed exploitable. There could be certain checks that deployed by the developers either in the source code or in some of the remote services for defending indirect prompt injection attacks. Thus, we do not explicitly mention any of the specific tools as vulnerable, except for the data exfiltration case study described in VII-F.

For the case study, we have filed a necessary report to the relevant party to inform them of the findings. Also, we have ensured all experiments are done in local and containerized setting, including the host of LLM and setup of remote attacker server. We have ensured that no third-party may be affected by our experiment.

We are also in the process of consolidating the rest of injectable flows. Once we prove them exploitable, we will report them accordingly.

REFERENCES

- [1] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, “Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection,” in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*. ACM, 2023.
- [2] Anthropic, “Model context protocol: Connecting AI to the tools it needs,” <https://www.anthropic.com/news/model-context-protocol>, 2024.
- [3] Anthropic, “Donating the Model Context Protocol and establishing the Agentic AI Foundation,” 2025. [Online]. Available: <https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation>
- [4] D. Wagner, J. S. Foster, E. A. Brewer, and A. Aiken, “A First Step Towards Automated Detection of Buffer Overrun Vulnerabilities,” 2000.
- [5] C. Cadar, D. Dunbar, and D. Engler, “KLEE: unassisted and automatic generation of high-coverage tests for complex systems programs,” in *Proceedings of the 8th USENIX conference on Operating systems design and implementation*, ser. OSDI’08. USA: USENIX Association, Dec. 2008, pp. 209–224.
- [6] Y. Shoshitaishvili, R. Wang, C. Salls, N. Stephens, M. Polino, A. Dutcher, J. Grosen, S. Feng, C. Hauser, C. Kruegel, and G. Vigna, “SOK: (State of) The Art of War: Offensive Techniques in Binary Analysis,” in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 138–157.
- [7] Github, “CodeQL,” 2026. [Online]. Available: <https://codeql.github.com/>
- [8] J. D. DeMott, R. J. Enbody, and W. F. Punch, “Revolutionizing the Field of Grey-box Attack Surface Testing with Evolutionary Fuzzing,” 2007.
- [9] A. Fioraldi, D. Maier, H. Eißfeldt, and M. Heuse, “Afl++: combining incremental steps of fuzzing research,” in *Proceedings of the 14th USENIX Conference on Offensive Technologies*, ser. WOOT’20. USA: USENIX Association, 2020.
- [10] J. Bau, E. Bursztein, D. Gupta, and J. Mitchell, “State of the art: Automated black-box web application vulnerability testing,” in *2010 IEEE Symposium on Security and Privacy*, 2010, pp. 332–345.
- [11] B. P. Miller, L. Fredriksen, and B. So, “An empirical study of the reliability of unix utilities,” *Commun. ACM*, vol. 33, no. 12, p. 32–44, Dec. 1990. [Online]. Available: <https://doi.org/10.1145/96267.96279>
- [12] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, “The instruction hierarchy: Training LLMs to prioritize privileged instructions,” in *arXiv preprint arXiv:2404.13208*, 2024, openAI.
- [13] E. Zverev, S. Abdelnabi, S. Tabesh, M. Fritz, and C. H. Lampert, “Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?” Jan. 2025, arXiv:2403.06833 [cs]. [Online]. Available: <http://arxiv.org/abs/2403.06833>
- [14] E. Bagdasaryan, T.-Y. Hsieh, B. Nassi, and V. Shmatikov, “(ab)using images and sounds for indirect instruction injection in multi-modal LLMs,” in *arXiv preprint arXiv:2307.10490*, 2023.
- [15] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, “InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents,” in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 471–10 506. [Online]. Available: <https://aclanthology.org/2024.findings-acl.624/>
- [16] E. DeBenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr, “AgentDojo: A dynamic environment to evaluate attacks and defenses for LLM agents,” in *arXiv preprint arXiv:2406.13352*, 2024.
- [17] Q. Zhan, R. Fang, H. S. Panchal, and D. Kang, “Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents,” in *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025, pp. 7116–7132.
- [18] S. Chen, J. Piet, C. Sitawarin, and D. Wagner, “StruQ: Defending against prompt injection with structured queries,” in *Proceedings of the 34th USENIX Security Symposium*. USENIX Association, 2025, arXiv:2402.06363 [cs].
- [19] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, and E. Kiciman, “Defending against indirect prompt injection attacks with spotlighting,” <https://arxiv.org/abs/2403.14720>, 2024, arXiv:2403.14720 [cs].
- [20] E. DeBenedetti, I. Shumailov, T. Fan, J. Hayes, N. Carlini, D. Fabian, C. Kern, C. Shi, A. Terzis, and F. Tramèr, “Defeating prompt injections

- by design,” <https://arxiv.org/abs/2503.18813>, 2025, arXiv:2503.18813 [cs].
- [21] D. Jacob, H. Alzahrani, Z. Hu, B. Alomair, and D. Wagner, “Prompt-Shield: Deployable detection for prompt injection attacks,” <https://arxiv.org/abs/2501.15145>, 2025, arXiv:2501.15145 [cs].
 - [22] K. Zhu, X. Yang, J. Wang, W. Guo, and W. Y. Wang, “MELON: Provable defense against indirect prompt injection attacks in AI agents,” in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025, arXiv:2502.05174 [cs].
 - [23] M. Nasr, N. Carlini, C. Sitawarin, S. V. Schulhoff, J. Hayes, M. Ilie, J. Pluto, S. Song, H. Chaudhari, I. Shumailov, A. Thakurta, K. Y. Xiao, A. Terzis, and F. Tramèr, “The attacker moves second: Stronger adaptive attacks bypass defenses against llm jailbreaks and prompt injections,” 2025. [Online]. Available: <https://arxiv.org/abs/2510.09023>
 - [24] F. Liu, Y. Zhang, J. Luo, J. Dai, T. Chen, L. Yuan, Z. Yu, Y. Shi, K. Li, C. Zhou, H. Chen, and M. Yang, “Make agent defeat agent: automatic detection of taint-style vulnerabilities in llm-based agents,” in *Proceedings of the 34th USENIX Conference on Security Symposium*, ser. SEC ’25. USA: USENIX Association, 2025.
 - [25] P. Wang, Y. Liu, Y. Lu, Y. Cai, H. Chen, Q. Yang, J. Zhang, J. Hong, and Y. Wu, “Agentarmor: Enforcing program analysis on agent runtime trace to defend against prompt injection,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.01249>
 - [26] Anthropic, “Model context protocol specification,” <https://modelcontextprotocol.io/specification>, 2024.
 - [27] “MCP server registry,” 2026, <https://registry.modelcontextprotocol.io/v0.1/servers?version=latest>.
 - [28] D. E. Denning, “A lattice model of secure information flow,” *Communications of the ACM*, vol. 19, no. 5, pp. 236–243, 1976.
 - [29] E. J. Schwartz, T. Avgerinos, and D. Brumley, “All you ever wanted to know about dynamic taint analysis and forward symbolic execution (but might have been afraid to ask),” in *Proceedings of the 2010 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2010, pp. 317–331.
 - [30] FIRST, “Common vulnerability scoring system version 4.0: Specification document,” <https://www.first.org/cvss/v4.0/specification-document>, 2023.
 - [31] C. Pappas, “MCP servers hub: Top 100 MCP servers by GitHub stars,” <https://github.com/apappascs/mcp-servers-hub>, 2026, accessed 2026-03-26.
 - [32] O. d. Moor, M. Verbaere, E. Hajiyev, P. Avgustinov, T. Ekman, N. Ongkingco, D. Sereni, and J. Tibble, “Keynote address: .ql for source code analysis,” in *Seventh IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 2007)*, 2007, pp. 3–16.
 - [33] C. Cadar, V. Ganesh, P. M. Pawlowski, D. L. Dill, and D. R. Engler, “Exe: Automatically generating inputs of death,” *ACM Trans. Inf. Syst. Secur.*, vol. 12, no. 2, Dec. 2008. [Online]. Available: <https://doi.org/10.1145/1455518.1455522>
 - [34] F. Yamaguchi, N. Golde, D. Arp, and K. Rieck, “Modeling and discovering vulnerabilities with code property graphs,” in *2014 IEEE Symposium on Security and Privacy*, 2014, pp. 590–604.
 - [35] J. Newsome and D. Song, “Dynamic Taint Analysis for Automatic Detection, Analysis, and Signature Generation of Exploits on Commodity Software,” Network and Distributed System Security (NDSS) Symposium 2005, 2005.
 - [36] S. Groß, S. Koch, L. Bernhard, T. Holz, and M. Johns, “FUZZILLI: Fuzzing for JavaScript JIT Compiler Vulnerabilities,” in *Proceedings 2023 Network and Distributed System Security Symposium*. San Diego, CA, USA: Internet Society, 2023. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2023/02/ndss2023_f290_paper.pdf
 - [37] Z. Kang, M. Lyu, Z. Liu, J. Yu, R. Fan, S. Li, and Y. Cao, “Follow My Flow: Unveiling Client-Side Prototype Pollution Gadgets from One Million Real-World Websites,” in *2025 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 991–1008. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00016>
 - [38] Z. Durumeric, E. Wustrow, and J. A. Halderman, “Zmap: fast internet-wide scanning and its security applications,” in *Proceedings of the 22nd USENIX Conference on Security*, ser. SEC’13. USA: USENIX Association, 2013, p. 605–620.
 - [39] A. Moshchuk, T. Bragin, S. D. Gribble, and H. M. Levy, “A Crawler-based Study of Spyware on the Web,” *Proceedings 2006 Network and Distributed System Security Symposium*, 2006.
 - [40] J. Viega, J. Bloch, Y. Kohno, and G. McGraw, “Its4: a static vulnerability scanner for c and c++ code,” in *Proceedings 16th Annual Computer Security Applications Conference (ACSAC’00)*, 2000, pp. 257–267.
 - [41] D. Larochelle and D. Evans, “Statically detecting likely buffer overflow vulnerabilities,” in *10th USENIX Security Symposium (USENIX Security 01)*. Washington, D.C.: USENIX Association, Aug. 2001. [Online]. Available: <https://www.usenix.org/conference/10th-usenix-security-symposium/statically-detecting-likely-buffer-overflow>
 - [42] M. Cova, V. Felmetsger, G. Banks, and G. Vigna, “Static detection of vulnerabilities in x86 executables,” in *2006 22nd Annual Computer Security Applications Conference (ACSAC’06)*, 2006, pp. 269–278.
 - [43] U. Shankar, K. Talwar, J. S. Foster, and D. Wagner, “Detecting format string vulnerabilities with type qualifiers,” in *10th USENIX Security Symposium (USENIX Security 01)*. Washington, D.C.: USENIX Association, Aug. 2001. [Online]. Available: <https://www.usenix.org/conference/10th-usenix-security-symposium/detecting-format-string-vulnerabilities-type-qualifiers>
 - [44] B. Steensgaard, “Points-to analysis in almost linear time,” in *Proceedings of the 23rd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL ’96. New York, NY, USA: Association for Computing Machinery, 1996, p. 32–41. [Online]. Available: <https://doi.org/10.1145/237721.237727>
 - [45] V. B. Livshits and M. S. Lam, “Finding security vulnerabilities in java applications with static analysis,” in *14th USENIX Security Symposium (USENIX Security 05)*. Baltimore, MD: USENIX Association, Jul. 2005. [Online]. Available: <https://www.usenix.org/conference/14th-usenix-security-symposium/finding-security-vulnerabilities-java-applications-static>
 - [46] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. le Traon, D. Octeau, and P. McDaniel, “FlowDroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for Android apps,” in *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. ACM, 2014, pp. 259–269.
 - [47] T. Ball and S. K. Rajamani, “The slam project: debugging system software via static analysis,” in *Proceedings of the 29th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL ’02. New York, NY, USA: Association for Computing Machinery, 2002, p. 1–3. [Online]. Available: <https://doi.org/10.1145/503272.503274>
 - [48] Google, “AFL,” 2024, original-date: 2019-07-25T16:50:06Z. [Online]. Available: <https://github.com/google/AFL>
 - [49] V.-T. Pham, M. Böhme, and A. Roychoudhury, “Aflnet: A greybox fuzzer for network protocols,” in *2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*, 2020, pp. 460–465.
 - [50] W. Xu, H. Moon, S. Kashyap, P.-N. Tseng, and T. Kim, “Fuzzing file systems via two-dimensional input space exploration,” in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 818–834.
 - [51] Fyodor, “Nmap Introduction,” 1997. [Online]. Available: <https://nmap.org/p51-11.html>
 - [52] G. Argyros, I. Stais, A. Kiayias, and A. D. Keromytis, “Back in Black: Towards Formal, Black Box Analysis of Sanitizers and Filters,” in *2016 IEEE Symposium on Security and Privacy (SP)*. San Jose, CA: IEEE, May 2016, pp. 91–109. [Online]. Available: <http://ieeexplore.ieee.org/document/7546497/>
 - [53] Q. Li, Y. Guo, and H. Chen, “Practical No-box Adversarial Attacks against DNNs,” in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 849–12 860. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/96e07156db854ca7b00b5df21716b0c6-Abstract.html
 - [54] Q. Zhang, Y. Sun, C. Zhang, C. Li, X. Wang, J. Song, and L. Gao, “Practical No-box Adversarial Attacks with Training-free Hybrid Image Transformation,” Feb. 2025, arXiv:2203.04607 [cs]. [Online]. Available: <https://arxiv.org/abs/2203.04607>
 - [55] J. Shi, Z. Yuan, G. Tie, P. Zhou, N. Z. Gong, and L. Sun, “Prompt Injection Attack to Tool Selection in LLM Agents,” arXiv:2407.07816 [cs]. [Online]. Available: <https://arxiv.org/abs/2407.07816>
 - [56] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019, arXiv:1909.08057. [Online]. Available: <https://arxiv.org/abs/1909.08057>
 - [57] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh,

D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," Jul. 2020, arXiv:2005.14165 [cs]. [Online]. Available: <http://arxiv.org/abs/2005.14165>

[58] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," Mar. 2023, arXiv:2210.03629 [cs]. [Online]. Available: <http://arxiv.org/abs/2210.03629>

[59] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, "InjecAgent: Benchmarking indirect prompt injections in tool-integrated LLM agents," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

APPENDIX A

ADDITIONAL INFORMATION OF `EVALUATE_SCRIPT` CASE STUDY

A. Tool Metadata Ingested by the Pipeline

The pipeline reads each tool’s published MCP registration. For `evaluate_script`, that registration consists of a name, a free-text description, and a JSON Schema for inputs:

```
name: evaluate_script
description: Evaluate a JavaScript function inside the currently selected page. Returns the response as JSON, so returned values have to be JSON serializable.
parameters.function (string): A JavaScript function declaration to be executed by the tool in the currently selected page. Examples:
() => { return document.title };
async () => { return await
  fetch("example.com") }.
parameters.args (array of string): An optional list of arguments to pass to the function (uid of an element on the page).
```

The example string in the description, in particular the literal `await fetch("example.com")`, advertises out-bound HTTP capability from inside the page context. This single fragment supplies the irreducible flow shape that the pipeline downstream stages reason over: attacker-influenceable page content $\rightarrow$ `evaluate_script` $\rightarrow$ LLM context, with arbitrary network egress as a side-effect of the function body.

B. Reproducibility and Stability

Across $N = 3$ runs of the local-state-exfiltration configuration, the agent obeyed the instructions in the indirect prompt injection payload in 2 runs and refused in 1 run.

APPENDIX B

VULNERABILITY LABELING

RQ4 requires a per-tool ground-truth label that is independent of both MCPSEC and the LLM baseline. We obtain it by attempting a controlled indirect prompt injection PoC against each sampled tool. A tool is labeled *vulnerable* only when three conditions are satisfied simultaneously: 1. the external data source is attacker-controllable under our threat model; 2. the tool actually retrieves that source when invoked with realistic parameters; and 3. content originating at that source reaches the tool output, i.e. the LLM-context sink, in a form that still carries instruction semantics. A tool that fails any obligation is labeled not vulnerable rather than being silently dropped, so that false negatives remain measurable for both systems.

A. Two-Stage Construction

PoC construction is deliberately split into an *inert* stage and a *payload* stage, so that no adversarial text is ever written to a live third-party service.

a) *Stage A: inert placeholder confirmation.*: For sources we control, we plant a single fixed, semantically inert token (MCPSEC-04993) at the identified injectable external source — an issue body, a message, a page property, a label description — inside accounts, repositories, and workspaces created solely for this study. We then invoke the tool with the same validated parameters recorded for the runtime-reachability experiment and check whether the token appears in the tool output. Planted artifacts are removed once the capture is archived.

b) *Stage B: in-process payload substitution.*: The adversarial payload is introduced only *inside* the locally running MCP server process. When external data enter the tool, we replace the retrieved placeholder canary, which our evaluators planted in the remote, with a real injection string. By doing so, we validated both attacker controllability on remote data source through planting placeholder canary, and the security measurement, if there are any designed to defense indirect prompt injection attacks, by passing along realistic prompt injection payloads.

This split is what makes the experiment ethically controlled: the only artifact that ever exists on a remote data source is an inert token, whereas every string with attack semantics contained fully locally.

B. Filtering and Coverage

Of the 177 sampled tools, we attempted PoC construction on those that were both technically and ethically approachable, and excluded the remainder. Tools were excluded when the tool could not be invoked at all under the subscription tier available to us; when invocation would require a destructive, irreversible, or third-party-visible side effect (for example, merging or deleting resources not under our control); when the source is a non-string structured object; or when the tool returns no external content in the first place. Under this protocol we label 95 tools vulnerable to indirect prompt injection.