

MCP-Diag: A Deterministic, Protocol-Driven Architecture for AI-Native Network Diagnostics

Devansh Lodha, Mohit Panchal and Sameer G. Kulkarni
Indian Institute of Technology Gandhinagar

Abstract—The integration of Large Language Models (LLMs) into network operations (AIOps) is hindered by two fundamental challenges: the *stochastic grounding problem*, where LLMs struggle to reliably parse unstructured, vendor-specific CLI output, and the *security gap* of granting autonomous agents shell access. This paper introduces MCP-Diag, a hybrid neuro-symbolic architecture built upon the Model Context Protocol (MCP). We propose a deterministic translation layer that converts raw `stdout` from canonical utilities (`dig`, `ping`, `traceroute`) into rigorous JSON schemas before AI ingestion. We further introduce a mandatory “Elicitation Loop” that enforces Human-in-the-Loop (HITL) authorization at the protocol level. Our preliminary evaluation demonstrate that MCP-Diag achieving 100% entity extraction accuracy with less than 0.9% execution latency overhead and 3.7 \times increase in context token usage.

I. INTRODUCTION

The escalating complexity of modern network infrastructure, characterized by distributed microservices and dynamic, software-defined topologies, demands a paradigm shift in network management [1]. Traditional manual approaches where operators type commands into SSH terminals are no longer scalable and are prone to human error [2], [3]. This has given rise to AIOps, utilizing Large Language Models (LLMs) to automate diagnostics and operational workflows [4], [5]. However, integrating LLMs with essential command-line utilities is impeded by two fundamental architectural barriers: the “Translation Gap” and the “Governance Gap.”

The ‘Translation Gap’ arises because utilities such as `ping`, `traceroute`, and `dig` were designed for human interpretation, not machine consumption. Their output is unstructured and highly variable; for instance, the textual output of `ping` differs subtly between `iputils` (Linux) and BSD-derived systems (macOS), and handles timeout representations differently [6]. Relying on an LLM to “screen-scrape” this data is inherently probabilistic [7]. An agent might correctly identify latency in one session but hallucinate an IP address in the next.

The ‘Governance Gap’ is even more critical. Granting an LLM direct shell access via standard subprocess libraries poses an immense security risk. Malicious actors could exploit prompt injection attacks to trick the AI into executing arbitrary commands [8], [9]. Even without malicious intent, a stochastic agent stuck in a reasoning loop could “hallucinate” a destructive command (e.g., `rm -rf /`) or accidentally trigger a Denial of Service (DoS) by spawning thousands of concurrent probes. A production-grade framework must enforce the principle of least privilege at the protocol level, not

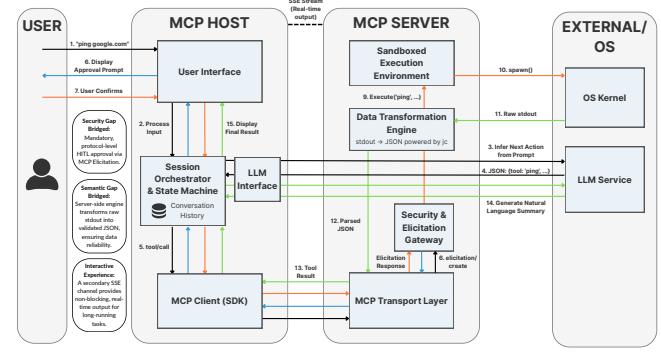


Fig. 1: MCP-Diag Architecture: The MCP Host (left) orchestrates the flow between the user, the LLM, and the MCP Server (right). Critically, all tool execution is centralized on the Server, which enforces governance via a mandatory MCP Elicitation loop (blue arrows) back to the Host, and guarantees reliable data by transforming raw `stdout` into validated JSON.

just via “system prompts” that can be bypassed. We propose a solution built upon the Model Context Protocol (MCP) [10]. We introduce **MCP-Diag**, a functional implementation and architectural validation of MCP for network diagnostics. Our system bridges these gaps via three key mechanisms:

- **Deterministic Grounding:** We inject a middleware layer (powered by the `jc` library [6]) that intercepts raw `stdout` and serializes it into rigorous, type-safe JSON schemas. This ensures the LLM reasons over structured data objects, eliminating parsing hallucinations.
- **Protocol-Enforced Governance:** We utilize MCP’s elicitation primitive to make tool execution impossible without an explicit, cryptographic-like handshake from the human operator.
- **Hybrid Transport:** We introduce a dual-channel communication model, using standard MCP transports for control logic and Server-Sent Events (SSE) for asynchronous data streaming, preventing agent timeouts during long-running convergence tests.

II. RELATED WORK

The landscape of LLM-integrated operations tools (summarized in Table I) is characterized by a trade-off between generality and reliability. General agentic frameworks such as LangChain [11] and Open Interpreter [12] offer high flexibility but suffer from high hallucination risks due to probabilistic

parsing of raw text. AI Terminals like Shell-GPT [13] and Gorilla CLI [14] introduce “User-Gated” safety, where commands require manual execution, yet they lack the stateful session orchestration necessary for autonomous remedial workflows (e.g., “Trace the path only if latency exceeds 50ms”).

Early implementations in the Model Context Protocol ecosystem, such as mcp-nettools [15], act as simple wrappers that pass raw binary output directly to the LLM. MCP-Diag is the first architecture to implement a *Secure Translation Layer*. Unlike existing solutions, we utilize *Deterministic Output Parsing* (via `jc` schemas) to ensure the LLM reasons over verified data objects. Furthermore, our custom client extends the protocol to solve the “Blocking Wait” problem inherent in standard MCP tool calls by utilizing a side-channel SSE stream for real-time diagnostic feedback.

III. SYSTEM ARCHITECTURE

The MCP-Diag system uses decoupled client-server model and adheres to the open Model Context Protocol standard. The design follows the “Sidecar Pattern,” where the diagnostic logic is isolated from the inference engine. The architecture comprises two primary components: the MCP Server, which acts as a deterministic execution gateway, and the MCP Host, which functions as the stateful session orchestrator.

A. Design and Execution Model

The control flow, as visualized in Fig. 1, proceeds through a “U-shaped” lifecycle is described below.

1. AI-Driven Plan Generation (Steps 1-5): The lifecycle begins when the user provides a natural language query (Step 1). The Session Orchestrator appends this to the history (Step 2) and queries the LLM (Step 3). The LLM returns a structured tool call plan (Step 4), which the Host sends to the Server as a formal `tool/call` request (Step 5).

2. Secure, Server-Side Execution (Steps 6-12): The Security Gateway of MCP Server initiates a mandatory Elicitation loop (Steps 6-8). On receiving user confirmation, it dispatches the command (Steps 9-10). The resulting raw `stdout` is piped to the Data Transformation Engine (Step 11), which uses the `jc` utility to parse it into a structured JSON object (Step 12).

3. Synthesis and Response (Steps 13-15): The Server returns the clean JSON to the Host (Step 13). The LLM synthesizes this structured data into a human-readable summary (Step 14), completing the loop (Step 15).

B. Secure Data Transformation Pipeline

To resolve the “Translation Gap,” the Server implements a strict input-output sanitization pipeline. Unlike traditional agent frameworks that expose a raw pseudo-terminal (PTY) to the LLM, MCP-Diag enforces a rigid data contract. The execution flow is formalized in Algorithm 1.

When a tool call is received, the system spawns two coupled processes: the diagnostic tool (e.g., `ping`) and a serialization engine (the `jc` library). The standard output (`stdout`) of the diagnostic tool is piped directly to the standard input (`stdin`) of the serializer. This ensures that the LLM never interacts with

Algorithm 1 Subprocess IPC and Schema Enforcement

```

Require: ToolRequest  $R = \langle \text{Cmd } C, \text{Args } A \rangle$ 
Ensure: Validated JSON Object  $J \in \text{Schema}(C)$ 
1: procedure HANDLETOOLCALL( $R$ )
2:    $\mathcal{M} \leftarrow \text{Payload}(\text{“Authorize execution of: ” } +C)$ 
3:    $\tau_{\text{auth}} \leftarrow \text{ELICIT\_HANDSHAKE}(\mathcal{M})$   $\triangleright$  Protocol-level HITL
4:   if  $\tau_{\text{auth}}.\text{action} \neq \text{ACCEPT}$  then
5:     return Err(AuthDenied)
6:   end if
7:    $P_{\text{diag}} \leftarrow \text{Spawn}(C, A)$   $\triangleright$  Process 1: Diagnostic Tool
8:    $P_{\text{parse}} \leftarrow \text{Spawn}(``\text{jc}``, [\`--\` +C])$   $\triangleright$  Process 2: Serializer
9:    $\Phi(P_{\text{diag}}.\text{stdout} \rightarrow P_{\text{parse}}.\text{stdin})$   $\triangleright$  Atomic Stdio Pipe
10:   $\Lambda(P_{\text{diag}}.\text{stdout} \rightarrow \text{SSE\_Buffer})$   $\triangleright$  Async Client Feed
11:   $\text{Wait}(P_{\text{diag}}, P_{\text{parse}})$ 
12:   $\varepsilon \leftarrow \text{ExitCode}(P_{\text{diag}})$ 
13:  if  $\varepsilon \in \text{SuccessSet}(C)$  then  $\triangleright$  Tool-specific exit validation
14:     $J \leftarrow \text{Deserialize}(P_{\text{parse}}.\text{stdout})$ 
15:    return  $J$ 
16:  else
17:    return Err(RuntimeError,  $P_{\text{diag}}.\text{stderr}$ )
18:  end if
19: end procedure

```

unstructured text; it only receives verified, schema-compliant JSON objects.

C. Protocol-Enforced Governance

To address the “Security Gap,” MCP-Diag rejects the industry-standard approach of client-side confirmation dialogs. Instead, we implement governance at the protocol level. The Server’s state machine defaults to a `BLOCK` state for all sensitive network operations. Upon receiving a request, the Server suspends the execution thread and emits an `elicitation` message. The operation remains atomically locked until a valid signed confirmation token is received from the Host.

Security Inheritance: It is crucial to note that by architecting strictly atop the Model Context Protocol, MCP-Diag inherits the security posture of the standard itself. This shared security responsibility model significantly reduces the attack surface compared to bespoke agent frameworks.

D. Hybrid Transport Layer

Network diagnostics introduce a unique concurrency challenge: tools like `ping` are long-running processes that can exceed standard RPC timeouts. We resolve this by implementing a Hybrid Transport Architecture:

- **Control Plane (Synchronous):** All atomic operations—handshakes, capability negotiation, and elicitation requests—occur over the standard MCP transport (Stdio/HTTP).
- **Data Plane (Asynchronous):** For streaming tools, the Server opens a secondary side channel using Server-Sent Events (SSE). The Host subscribes to this stream, allowing the Server to push real-time `stdout` frames immediately as they occur.

IV. SYSTEM VALIDATION

We validated the protocol adherence of the MCP-Diag server using a standard, third-party MCP host (VS Code MCP Host). As illustrated in Fig. 2, the server strictly enforces

TABLE I: Architectural Comparison of LLM-Driven Operational Tools

System	Category	Output Parsing	Hallucination Risk	Execution Safety	Human-in-the-Loop	State Awareness	Streaming
mcp-nettools [15]	MCP Wrapper	Probabilistic (Raw Text)	High	Host-Dependent	None (Pass-through)	Stateless (One-off calls)	Blocking (Wait for Exit)
LangChain [11]	Toolkit	Probabilistic (Regex/Raw)	High	Client-Side (Optional “y/n” Prompt)	None (Pass-through)	Window-Based (Token limit)	Blocking (Wait for Exit)
Open Interpreter [12]	General Agent	Probabilistic (Raw Text)	High	Isolation Only (Net. Exposed)	Chat-Based (Can Bypass)	Stateful Loop (Preserves CWD)	Buffered (Piped Output)
Shell-GPT [13]	CLI Assistant	Probabilistic (Raw Text)	Medium (Suggestion)	User-Gated (Flag Prompt)	Enforced (User Execs)	Session-Based (Cache File)	Real-Time
Warp AI [16]	AI Terminal	Probabilistic (Raw Text)	Medium (Suggestion)	User-Gated (CMD+Enter)	Enforced (User Execs)	Terminal Context (Shell History)	Real-Time (Native UI)
Gorilla CLI [14]	Cmd Generator	Probabilistic (Raw Text)	Low (Finetuned Model)	User-Gated (User Execs)	Enforced (User Execs)	Stateless (One-off Cmds)	Real-Time
Anthropic Comp. Use [17]	UI Agent	Probabilistic (Visual OCR)	High (Visual Errors)	Isolation Only (Net. Exposed)	High Friction / None (Observation)	Stateful (Visual History)	Real-Time (Visual Proc.)
MCP-Diag (This Work)	MCP Tool	Deterministic (jc Schema)	Very Low (JSON, Validated)	Protocol-Level (Mandatory Elicit.)	Enforced (Blocking State)	Stateful Session (Orchestrated)	Real-Time (Hybrid SSE)

the mandatory elicitation handshake for sensitive operations, ensuring compliance with the JSON-RPC 2.0 based MCP specification.

This functional test confirms that the architecture is compatible with any compliant MCP host. While the core MCP transport logic is strictly maintained, our custom client implements an optional SSE side-channel extension for real-time, line-by-line streaming. This extension enables low-latency feedback for long-running diagnostic tasks without violating the request-response cycle of the primary MCP transport.



Fig. 2: MCP-Diag (server) running in VS Code (Host).

V. PERFORMANCE EVALUATION

The complete source code for the MCP-Diag server implementation, the custom client, and the automated benchmark suite has been made publicly available [18].

To quantify the architectural trade-offs of deterministic grounding, we conducted a large-scale benchmark ($N = 500$ trials) targeting the global Top 500 domains [19]. The benchmark task required the agent to execute a traceroute to a target domain and extract the third-hop IP address.

A. Benchmark Configuration and Methodology

We compared two distinct architectural patterns to isolate the impact of protocol-driven diagnostics:

- **Baseline:** A standard agentic workflow where the prompt is sent to the LLM, which generates a raw CLI command string. The system parses this string using regex, executes it via a direct subprocess call, and returns the unstructured stdout to the LLM for probabilistic entity extraction.
- **MCP-Diag (Benchmark Mode):** The system executes tools via the MCP tool-call primitive. For this study, the server was configured in a “Headless” mode with the SSE side-channel and the mandatory Elicitation loop disabled. This configuration isolates the computational latency of serialization, transport, and schema validation from human reaction time.

All trials utilized the Gemma-3-27b-it model accessed via the Google AI API. The client-side orchestration and MCP server were executed within a local Bun/V8 JavaScript runtime. To ensure statistical validity, we enforced a JIT Warmup (preliminary execution to stabilize the runtime) and Path Priming (unmeasured traceroute before every trial to populate intermediate hop ARP tables and routing caches).

B. Grounding Fidelity and Reliability

Experimental data (Fig. 3C) indicates that the *ShellTool* baseline is inherently unstable, exhibiting stochastic failures in 0.4% of trials. These errors occurred when non-standard vendor stdout formatting (e.g., unexpected ICMP error codes or null-hop asterisks) exceeded the regex-based parser’s capabilities. By enforcing a rigid, type-safe JSON contract, MCP-Diag achieved 100% extraction accuracy across the entire dataset.

C. Architectural Tax: Latency and Token Cost

We quantify the “Architectural Tax” as the overhead introduced by the protocol layer:

- 1) **Latency Overhead:** As shown in Fig. 3A, the MCP layer introduces a mean overhead of 311ms. In the context of the traceroute task (avg. 34s), this tax is < 0.9%, resulting in a negligible user experience impact of +0.2% (Fig. 3B).

MCP-Diag vs Baseline Performance Evaluation

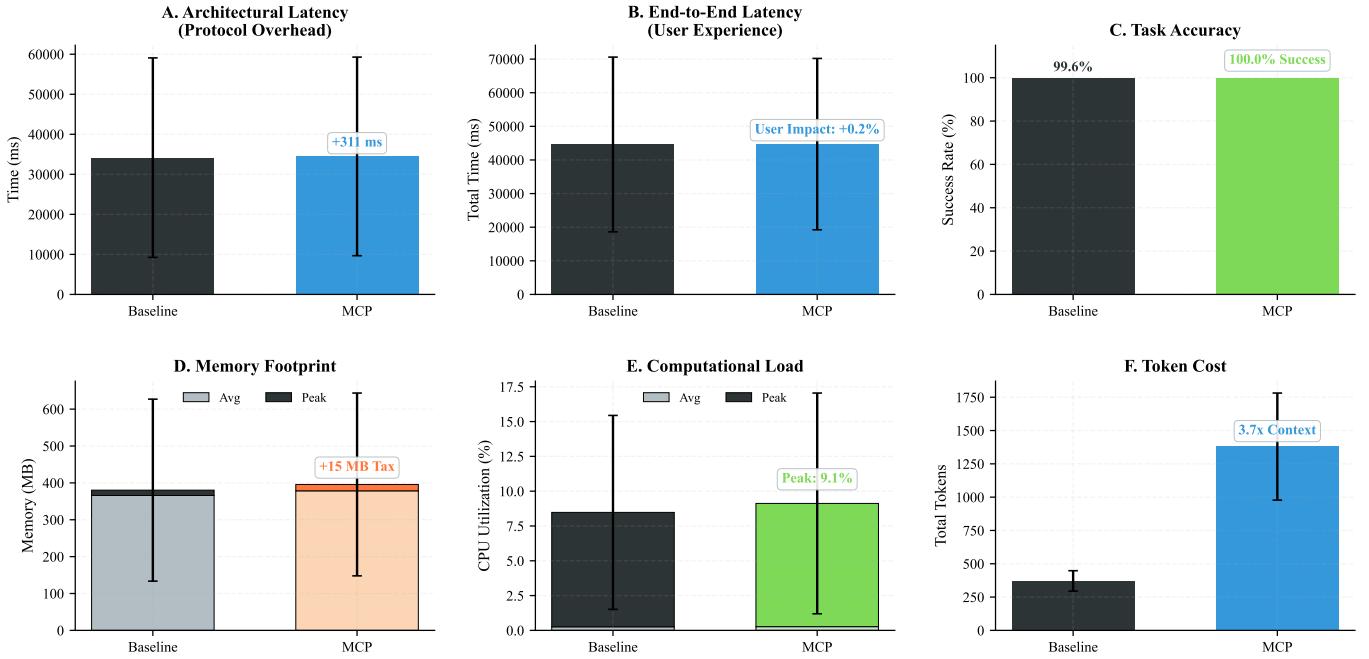


Fig. 3: Comparison of MCP-Diag and baseline ShellTool for A) Protocol overhead B) Latency overhead C) Task Accuracy D) Memory overhead E) CPU overhead and F) Token Overhead over 500 trials.

- 2) **Token Consumption:** Enforcing JSON schemas results in a $3.7\times$ increase in context token usage per turn (mean 1100 vs. 300, Fig. 3F). Because the LLM processes input tokens in parallel, this increase does not significantly degrade time-to-first-token (TTFT) latency.

D. Resource Efficiency

Fig. 3D, E show that the persistent MCP sidecar imposes a memory tax of only +15 MB and a peak CPU increase of +1.1%. This demonstrates that the benefits of deterministic grounding can be achieved with a minimal footprint suitable for resource-constrained edge deployments.

VI. CONCLUSION AND FUTURE WORK

This paper introduced MCP-Diag, a novel architecture that provides a principled, security-first solution to the critical “Security Gap” and “Translation Gap” hindering the application of LLMs to network diagnostics. By synthesizing mandatory protocol-level elicitation, deterministic server-side data transformation, and a hybrid communication model, our work establishes a practical, production-ready blueprint for the future of AIOps. The architectural foundation of MCP-Diag enables several ambitious avenues for future research:

1. Complexity : As diagnostic tools evolve to capture deeper system states (e.g., full packet captures via tshark), the richness and depth of the structured JSON output will naturally increase. We anticipate that rapid advancements in LLM context windows and the development of more intelligent,

adaptive parsers will seamlessly accommodate this growth, allowing for increasingly granular diagnostics without compromising the semantic rigor that our architecture guarantees.

2. Autonomous Diagnostic Agents: We plan to utilize MCP’s Prompt primitive to create complex, multi-step reasoning chains. With this approach, agent can autonomously diagnose a high-level symptom by sequencing ping, traceroute, and dnsLookup calls, while adhering to mandatory user approval at critical state transitions.

3. Scalability and Concurrency: Our current study focuses on the fundamental interaction strictly between a single MCP Host and Server. In future, we will explore the dimensions of concurrent execution and scalability. We will also investigate a distributed orchestration layer capable of managing thousands of concurrent MCP-Diag instances across a network fabric, transforming the architecture from a local diagnostic utility into a global sensor network.

ACKNOWLEDGMENT

We thank the TPC Chairs and anonymous reviewers for their thoughtful feedback. This work is supported by grants from The Govt. of India, Department of science and technology, SERB Core Research Grant: CRG/2023/008021, MeitY Grant: 4(3)/2024-ITEA and Department of Telecommunications (DoT) 5G Use Case Lab, IIT Gandhinagar.

REFERENCES

- [1] C. Alice and H. Bukola, "Overcoming software-defined networking (SDN) challenges: Emerging solutions and future research directions," *EasyChair Preprint*, no. 14687, 2024, easyChair Preprint no. 14687. [Online]. Available: <https://easychair.org/publications/preprint/pMvmj>
- [2] J. Edelman, S. S. Lowe, and M. Oswalt, *Network Programmability and Automation: Skills for the Next-Generation Network Engineer*. Sebastopol, CA: O'Reilly Media, 2018.
- [3] J. Schoenwaelder, "Overview of the 2002 IAB network management workshop," Internet Requests for Comments, RFC Editor, RFC 3535, May 2003. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc3535.txt>
- [4] L. Zhang, T. Jia, Z. Liu, Y. Feng, W. Zeng, and Y. Yang, "A survey of aiops in the era of large language models," *ACM Computing Surveys*, 2025, just Accepted.
- [5] C. Liu, X. Xie, X. Zhang, and Y. Cui, "Large language models for networking: Workflow, advances and challenges," 2024. [Online]. Available: <https://arxiv.org/abs/2404.12901>
- [6] K. Brazil, "jc: CLI tool and python library that converts the output of popular command-line tools to JSON," 2024, GitHub Repository. Accessed: Dec. 20, 2025. [Online]. Available: <https://github.com/kellyjonbrazil/jc>
- [7] J. Hong, N. V. Tu, and J. W.-K. Hong, "A comprehensive survey on LLM-based network management and operations," *International Journal of Network Management*, vol. 35, no. 6, 2025. [Online]. Available: <https://doi.org/10.1002/nem.70029>
- [8] OWASP Foundation. (2025) OWASP Top 10 for Large Language Model Applications (v2025). Specifies Vulnerability LLM06: Excessive Agency. Accessed: Dec. 20, 2025. [Online]. Available: <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
- [9] National Vulnerability Database. (2025, May) CVE-2025-1753: Arbitrary Shell Execution via Argument Injection. Accessed: Dec. 20, 2025. [Online]. Available: <https://nvd.nist.gov/vuln/detail/CVE-2025-1753>
- [10] Agentic AI Foundation. (2025) Model Context Protocol Specification. V2025.11.25. Accessed: Dec. 20, 2025. [Online]. Available: <https://modelcontextprotocol.io/specification>
- [11] H. Chase, "LangChain: Building applications with LLMs through composability," 2025, software Version 0.3. [Online]. Available: <https://github.com/langchain-ai/langchain>
- [12] K. Lucas, "Open Interpreter: A natural language interface for computers," 2023, GitHub Repository. [Online]. Available: <https://github.com/OpenInterpreter/open-interpreter>
- [13] E. (TheR1D), "Shell-GPT: A command-line productivity tool powered by AI," 2023, GitHub Repository. [Online]. Available: https://github.com/TheR1D/shell_gpt
- [14] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large Language Model Connected with Massive APIs," 2023. [Online]. Available: <https://arxiv.org/abs/2305.15334>
- [15] D. Kruijt, "mcp-nettools: A Model Context Protocol implementation providing network diagnostics," 2025, GitHub Repository. [Online]. Available: <https://github.com/dkruijt/mcp-nettools>
- [16] Warp Technologies, Inc. (2025) Warp: The terminal for the 21st century. [Online]. Available: <https://www.warp.dev>
- [17] Anthropic. (2024, Oct.) Developing a computer use model. [Online]. Available: <https://www.anthropic.com/news/developing-computer-use>
- [18] D. Lodha, M. Panchal, and S. G. Kulkarni, "MCP-Diag: A Deterministic, Protocol-Driven Architecture for AI-Native Network Diagnostics," <https://github.com/devansh-lodha/mcp-diag>, 2025.
- [19] Moz. (2025) The Moz Top 500 Websites. [Online]. Available: <https://moz.com/top500>