Law-MCP: A Legal Compliance MCP Framework

Qianyu Wang

School of Intellectual Property, University of Chinese Academy of Sciences, Beijing, China
School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China
wangqianyu23@mails.ucas.ac.cn

Abstract

LLMs are moving from passive information processors to agentic systems that can act through the Model Context Protocol (MCP). This shift widens the scope of AI use. At the same time, the rapid growth of the MCP ecosystem creates legal risks. The main risks concern intellectual property (IP) infringement and the use of data by unverified third-party tools. In this complex supply chain, users face severe information asymmetry and may bear direct liability for outputs produced by opaque components that they do not control. Post-hoc legal remedies are not enough. We propose `Law-MCP`to build the Bridge between Artificial Intelligence and Law.

We analyze how legal risks are distributed across tool calls and user interactions and show the need for technical controls. We then present a MCP framework with modular design with three layers: (1) an MCP Context Aggregation Layer that standardizes data and tracks provenance; (2) a plug-and-play IP Risk Detection Layer that uses a central scheduler to coordinate detection plugins; and (3) a Localized Legal Alert Layer that links technical risks to the laws of specific jurisdictions and to case law. The framework automates IP risk detection to protect users. It also offers developers a verifiable way to show due care, closing the gap between legal duties and agent capabilities.

Introduction

Large Language Models (LLMs) have revolutionized natural language processing. Early LLMs mainly answered questions. Their risks were mostly about harmful content. However, the Agentic AI can act. It can call tools and affect digital and even physical systems. As LLMs become products and applications, they add broad tool use via MCP. System complexity grows. More parties join, including developers, deployers, and users.

The Model Context Protocol (MCP) (Anthropic et al. 2024) represents a significant advancement in this domain, providing a standardized interface for AI assistants to connect with external tools and data sources. As an open protocol, MCP aims to establish a universal adaptation layer—a "USB-C port for AI applications"—that allows any compliant model to access any data repository or service via a consistent format. This standardization addresses the fragmentation issue where each new tool integration demands custom development, replacing it with a single, extensible protocol.

Since MCP's release in 2024, its ecosystem has grown fast. Many MCP tools have appeared. A large share are third-party tools without strict compliance review (Singh et al. 2025; Fang et al. 2025; Hou et al. 2025). Some hide legal risks, such as unauthorized data collection and IP infringement (Beurer-Kellner and Fischer 2025). When an AI system outputs infringing content, it is hard to assign liability. The risk may come from the model, the retrieved knowledge, or the tools. Users face strong information asymmetry. They cannot see into the black box and cannot judge whether each component is compliant. They may enter legal gray zones when choosing tools. IP infringement is a key concern. We therefore propose an LLM-based Legal Compliance MCP Tool `Law-MCP`. It detects IP risks for user-invoked tools in an automated way, locates sources, and gives compliance advice.Our main contributions are as follows:

- We analyze how legal risks are distributed across tool calls and user interactions, showing the need for *ex-ante* technical controls;
- We present a modular Legal Compliance MCP Framework with three layers: an MCP Context Aggregation Layer for standardizing data and tracking provenance, a Plug-and-Play IP Risk Detection Layer for coordinating detection plugins, and a Localized Legal Alert Layer linking technical risks to jurisdictional laws and case law;

Legal Risks Analysis

The transition from passive LLMs to agentic systems via MCP introduces a complex supply chain of liability. Unlike traditional software where risks are static and defined by code, MCP agents dynamically select tools, retrieve data, and generate content. We analyze these legal risks across three dimensions of the agentic lifecycle: source provenance, autonomous execution, and downstream liability.

Upstream Risks: Tool Provenance and Authorization

The first layer of risk originates from the *tools and data sources* the agent selects. In an open MCP ecosystem, agents may invoke third-party tools that function as "black boxes," creating significant opacity regarding IP rights.

Proceedings of Bridge between Artificial Intelligence and Law 2026 (Short papers), pages 258–262
Januray 21, 2026, AAAI 2026 Bridge Program, Singapore

- **Inherently Infringing Tools:** Some tools are designed to bypass copyright protections, such as unauthorized web scrapers that harvest paywalled content. When an agent utilizes such a tool, it may trigger liability for copyright infringement or unfair competition, even if the user did not explicitly request the scraping.
- **Unauthorized Access via Legitimate Tools:** Even compliant tools can be misused. An agent might bypass authentication mechanisms to access proprietary databases. For model developers, if the underlying training corpora or the retrieved knowledge base (RAG) contains unauthorized copyrighted material, the resulting output is legally "fruit of the poisonous tree."

Process Risks: Autonomous Execution and Persistence

The second layer involves the *agent's autonomous behavior* during task execution. Unlike a single API call, an agentic workflow involves multi-step reasoning, intermediate data storage, and service interaction.

- **Violation of Service Terms:** Agents interacting with external APIs must adhere to Terms of Service (ToS). An agent might inadvertently violate rate limits or usage restrictions (e.g., using a non-commercial API for commercial tasks), leading to breach of contract claims against the deployer.
- **Improper Data Persistence:** During complex workflows, agents often cache intermediate results. If an agent stores copyright-protected content or sensitive personal data (PII) on unsecure public storage or local servers without encryption, it creates risks regarding data residency laws (e.g., GDPR) and reproduction rights.

Downstream Risks: User Liability and Attribution

The final layer concerns the *output generation and user responsibility*. This is where the information asymmetry peaks: users often bear direct liability for outputs they cannot fully audit.

The Burden of "Direct Infringement" Users may mistakenly assume AI-processed content is original. Legally, generating content based on protected works often constitutes an unauthorized derivative work. For instance, in the "Medusa LoRA Model" case (Shanghai High People's Court 2025), the court found the user liable for training a model on copyrighted images, rejecting the platform neutrality defense. Similarly, the US *TAKE IT DOWN Act* and China's *Deep Synthesis Provisions* impose strict liability on users who generate non-consensual deepfakes or infringe on personality rights.

The "Duty of Care" Dilemma Under current legal frameworks, a user's liability often hinges on whether they exercised a reasonable "duty of care." However, the opacity of the MCP ecosystem makes this nearly impossible for non-experts.

- **Information Asymmetry:** Users cannot audit upstream models for pirated corpora or monitor third-party MCP tools in real-time.
- **High Evidentiary Burden:** Even if a user tries to be compliant, proving that an infringement was caused by a tool's defect rather than user misuse is technically difficult in court.

Given these risks, relying solely on post-hoc legal remedies is insufficient. Users need *ex-ante* technical controls to bridge the gap between their legal duties and their limited visibility into the agent's internal workings.

Legislation in Different Countries

As MCP tools enable image and video processing, legal red lines tighten when users ask agents to generate content about specific individuals.

- The European Union's *AI Act* classifies deepfakes as "limited risk" and imposes transparency duties.
- China's *Provisions on the Administration of Deep Synthesis of Internet Information Services* require labeling of deepfake content.
- On April 9, 2025, US Senators Chris Coons, Marsha Blackburn, Amy Klobuchar, and Thom Tillis reintroduced the NO FAKES Act of 2025 (McDermott 2025). It seeks to establish a Digital Replication Right for voices and likenesses, clarifies that unauthorized use and distribution of digital replicas is infringement, and sets safe harbor rules for online service providers.
- The US *TAKE IT DOWN Act* (Ortutay 2025), effective May 19, 2025, makes non-consensual publication of real or AI-forged explicit images a federal crime. Users who generate such content with agent tools may face criminal, not only civil, liability.
- Denmark's proposed July 2025 amendment to the *Copyright Act* (Kulturministeriet 2025) adds rights for performers and general protection of personal traits. It bans realistic imitation of a person's traits or artistic performance without consent, except for narrow cases such as satire. Using AI to "resurrect" the deceased or imitate celebrities is thus risky.

Users sometimes assume that AI-processed images are original. Legally, this is an unauthorized derivative work. Users may bear direct liability. For example, in the Chinese "Medusa LoRA Model" case (Shanghai High People's Court 2025), the defendant Li used a "LoRA model training" function to upload over 20 images of the "Medusa" character from *Battle Through the Heavens* without permission, trained a LoRA model, and published it. The court found direct infringement of reproduction, adaptation, and information network dissemination rights, despite platform neutrality arguments.

Users Face the Greatest Difficulty in Controlling Legal Risks In the MCP ecosystem, risks may come from the model, tool, or platform layers. Yet users have the least control and bear the most direct consequences. AI does not change the basic liability framework. User liability still depends on use behavior and fault. The key question is whether a reasonable duty of care was breached. AI complicates how fault is judged.

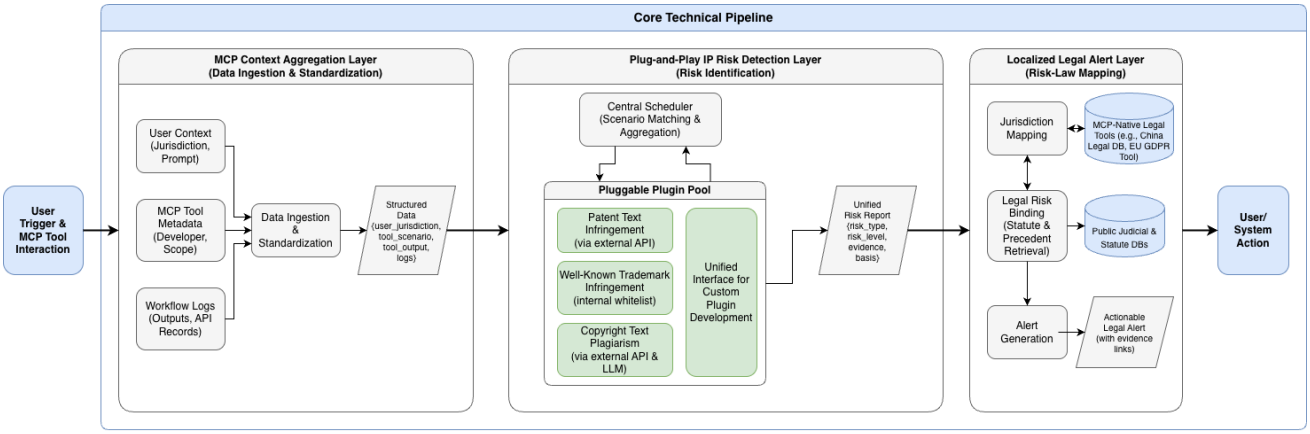

Figure 1: Technical Workflow

First, information asymmetry blocks users from auditing upstream models for pirated corpora and from monitoring third-party MCP tool compliance in real time. This opacity can make users unwitting principals or beneficiaries of infringement. If providers do not inform users about norms or fail to push necessary security updates, the user's duty of care may be reduced.

Second, the degree of user control varies by AI type. With assistive AI, users must make decisions and supervise, so their duty of care is higher. With substitutive AI, the duty shifts to system checks and compliance with use norms. For enterprises, when their digital employees cause infringement or breach, the enterprise often compensates first and then seeks recourse from developers.

Finally, even if users exercise due care, proving upstream negligence or defects is hard in court. The evidentiary burden is high, so user remedies are limited.

When risk sources, information, and control sit with upstream providers, asking users to self-censor and supervise cannot fix their weak position. Automated technical means are needed to give users pre-emptive identification and real-time blocking.

Given these risks, post-hoc legal remedies are insufficient. **A Bridge between Artificial Intelligence and Law is needed**.

Technical Approach

The framework uses a streamlined pipeline for easy integration and high scalability within MCP. As shown in Figure 1, the system has three layers. We map each difficulty to a framework layer: (i) user information asymmetry → the *MCP Context Aggregation Layer*, which standardizes multi-source context and tracks provenance for audit; (ii) low user control and high evidentiary burden → mainly the *Plug-and-Play IP Risk Detection Layer*, which provides proactive and reproducible findings; and (iii) multi-jurisdictional requirements → the *Localized Legal Alert Layer*, which performs jurisdiction mapping and generates guidance linked to statutes and cases.

Implementation Foundation

Our framework is built on the Python MCP library provided by (Anthropic et al. 2024). The core follows a plugin-based design that integrates with existing MCP deployments. By adhering to MCP specifications, the system can run as middleware between user agents and tools.

MCP Context Aggregation Layer

This layer targets information asymmetry. It standardizes multi-source context and tracks data provenance for evidence.

It serves as the entry point for data ingestion and standardization. Instead of asking users to check each MCP tool for compliance, it collects and analyzes context from multiple sources. It identifies where each data stream comes from and its risk profile.

It aggregates four streams: (1) *User Context* (jurisdiction and task prompts); (2) *MCP Tool Metadata* (developer, data sources, scope); (3) *Workflow Logs* (records of interactions and API calls); and (4) *Tool Outputs* (raw text or data).

Unstructured inputs are transformed into a structured format user_jurisdiction, tool_scenario, tool_output, logs. This unified schema feeds the risk detection layer. Provenance is tracked so that downstream components can attribute risks to specific tools or sources.

Plug-and-Play IP Risk Detection Layer

This layer reduces low user control by identifying and blocking IP risks proactively and in real time. It produces reproducible findings that support evidence.

The design is "Central Scheduler + Plugin Pool." It is loosely coupled to allow modular growth and varied organizational needs.

The **Central Scheduler** performs three actions: (1) *Plugin Registration*, accepting IP plugins that follow a unified interface and I/O contract; (2) *Scenario Matching*, selecting plugins based on `tool_scenario` (e.g., patent plugins for "technical proposal generation"); and (3) *Result Aggregation*, normalizing outputs from multiple plugins into a single risk report.

The **Plugin Pool** is the extensible core. Instead of hard-coding detection, the framework exposes an interface so organizations can add custom IP detection. The demo we implemented has five plugin categories:

Patent Text Infringement Plugin: Compares technical text against patent claims via external APIs (e.g., Google Patents API) or self-hosted databases. Similarity $\geq 70\%$ is flagged as a potential risk. This applies to technical proposals, algorithm descriptions, and architecture documents.

Well-Known Trademark Plugin: Matches content against a whitelist of top global trademarks and brand names to detect unauthorized use. This applies to marketing content, product descriptions, and brand-related outputs.

Copyright Text Plugin: Uses external APIs (e.g., Project Gutenberg, CrossRef) and LLM-based semantic checks to identify similarity and plagiarism risks. This applies to content generation, reports, and documents.

Open Source License Plugin: (Planned) Uses GitHub License APIs and open-source databases to verify code against license terms (e.g., GPL, MIT, Apache).

Design Patent Plugin: (Planned) Matches designs and visuals against WIPO and internal registries to detect design patent risks. This applies to product design and visual content.

Organizations can extend the framework with proprietary plugins, such as for trade secret leakage, competitive intelligence, or industry-specific rules. Because detection logic is decoupled from scheduling, the framework is agnostic to methods and can evolve with new risks and techniques.

Localized Legal Alert Layer

This layer operationalizes multi-jurisdictional rules by mapping technical risks to laws and cases in the user's jurisdiction.

It runs three functions: (1) *Jurisdiction Mapping*, using `user_jurisdiction` to call MCP-native legal tools for the region (e.g., China legal databases, EU GDPR tools, US Patent Office APIs); (2) *Legal Binding*, retrieving statutes (e.g., Article 11 of the Patent Law of the People's Republic of China, EU Directive 2006/115/EC) and relevant precedents to support the assessment; and (3) *Alert Generation*, producing natural language alerts that state the risk in clear terms (e.g., "Current output may violate Patent Law Article 11; similar 2023 cases saw damages of RMB 500,000–5,000,000"), along with evidence links to laws and cases.

This layer turns abstract IP risk signals into concrete legal warnings. Users can then make informed choices about tool use and output deployment. By grounding technical findings in jurisdictional law, the framework links technical compliance with legal accountability.

Demo for Risk Analysis

We demonstrate the risk analysis pipeline with a demo in Figure 2. This demo showcases an end-to-end "generation-to-risk" pipeline for marketing content. Given a natural-language prompt, the system first produces a candidate draft

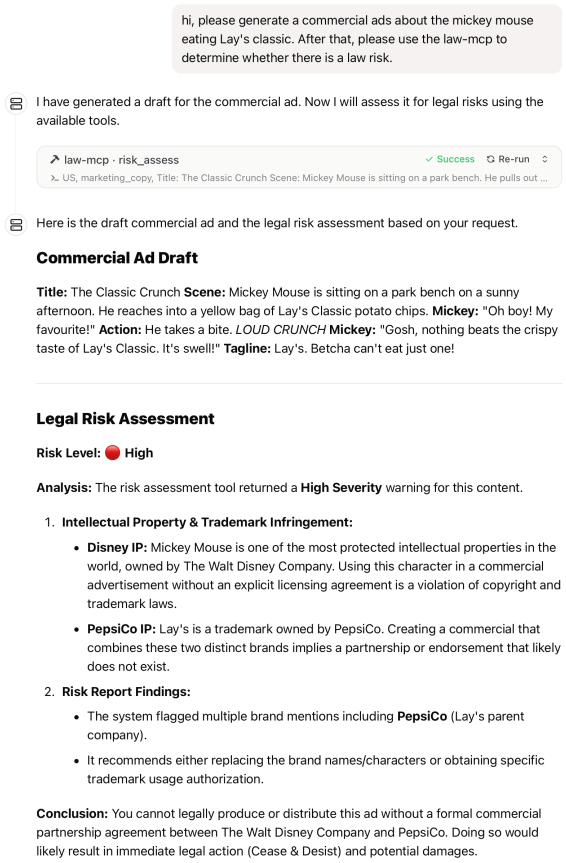

Figure 2: Demo for risk analysis

and then invokes a legal or compliance assessor (e.g., law-mcp's risk_assess) under a specified jurisdiction and scenario. The assessor returns a structured report—covering risk level, triggered issues (such as copyright and trademark concerns), and concrete mitigation advice—so creators can iterate early. In the illustrated example, an ad that combines a copyrighted character with a branded product is flagged as High Risk due to potential IP infringement and implied endorsement. The pipeline demonstrates how automated checks surface legal issues before publication and guide de-risking steps, such as de-branding, substituting generic entities, or obtaining explicit authorization.

Conclusion

This paper addresses a critical gap in the LLM ecosystem: users face severe information asymmetry, limited control over upstream components, and high evidentiary burdens. We propose `Law-MCP`that embeds "compliance by design" into agent workflows through three integrated layers. By automating IP risk detection and providing jurisdiction-specific guidance, the framework helps users exercise due care.

References

Anthropic et al. 2024. Model context protocol. https://github.com/modelcontextprotocol. Accessed: 2025-11-29.

Beurer-Kellner, L.; and Fischer, M. 2025. MCP Security Notification: Tool Poisoning Attacks. *Invariant Labs Blog*.

Fang, J.; Yao, Z.; Wang, R.; Ma, H.; Wang, X.; and Chua, T.-S. 2025. We Should Identify and Mitigate Third-Party Safety Risks in MCP-Powered Agent Systems. *arXiv preprint arXiv:2506.13666*.

Hou, X.; Zhao, Y.; Wang, S.; and Wang, H. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Kulturministeriet. 2025. Forslag til lov om ændring af lov om ophavsret (Draft Bill amending the Copyright Act). https://www.ft.dk/samling/20241/almdel/kuu/bilag/232/3050901.pdf. Accessed: 2025-11-29.

McDermott, E. 2025. NO FAKES Act Reintroduced to Support from Both Big Tech and Creators. https://ipwatchdog.com/2025/04/09/no-fakes-act-reintroduced-support-big-tech-creators/. Accessed: 2025-11-29.

Ortutay, B. 2025. President Trump signs Take It Down Act, addressing nonconsensual deepfakes. What is it? https://apnews.com/article/take-it-down-deepfake-trump-melania-first-amendment-741a6e525e81e5e3d8843aac20de8615. Accessed: 2025-11-29.

Shanghai High People's Court. 2025. (First instance judgment on AI large model copyright infringement case involving Medusa image). https://mp.weixin.qq.com/s/Plae0snaOEsqqmodLU9j4g. Accessed: 2025-11-29.

Singh, A.; Ehtesham, A.; Kumar, S.; and Khoei, T. T. 2025. A survey of the model context protocol (mcp): Standardizing context to enhance large language models (llms).