Model Bağlam Protokolü (MCP) Güvenliği: Tehditler, Savunmalar ve Dağıtım Kılavuzu

Yusuf Talha ARABACI

23 Ekim 2025

Özet

Bu çalışma, Model Context Protocol (MCP) protokolünün güvenlik açıklarını, potansiyel saldırı vektörlerini ve bu açıkların ortadan kaldırılması için alınması gereken savunma önlemlerini kapsamlı bir şekilde incelemektedir. MCP, otonom ajanların harici sistemlerle güvenli bir şekilde iletişim kurmasını sağlayan bir protokoldür. Bu rapor, MCP'nin güvenlik zafiyetlerini detaylandırarak, uygulama düzeyinde önerilen savunma stratejileri ile bu tehditlere karşı nasıl koruma sağlanabileceğini ele almaktadır.

1 Giriş

Model Context Protocol (MCP), otonom ajanlar ve harici sistemler arasındaki etkileşimi güvenli bir şekilde sağlayan bir protokoldür. MCP, daha önce statik olan dil modellerini (LLM) dinamik ve otonom ajanlara dönüştürerek büyük bir güvenlik açığı yaratmıştır. Bu protokolün kullanımı, birçok güvenlik tehditini gündeme getirmiştir. Örneğin, araçlar arası zehirlenme, prompt enjeksiyonu ve kimlik doğrulama eksiklikleri gibi klasik saldırı vektörlerinin yanı sıra, daha sofistike ve çok katmanlı saldırılar da ortaya çıkmaktadır. Bu çalışmanın temel amacı, bu tehditlere karşı etkili savunma stratejileri geliştirmektir.

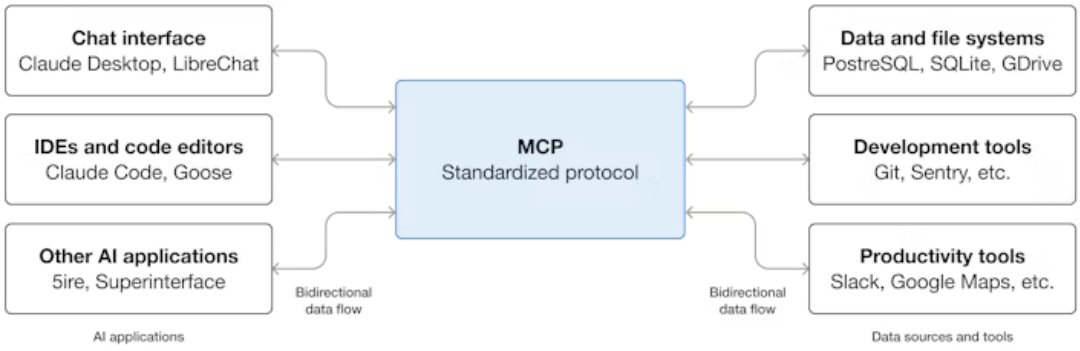
1.1 Çalışmanın Amacı ve Kapsamı

Bu çalışma, MCP protokolünün güvenlik analizini yaparak, aşağıdaki unsurlara odaklanmaktadır:

- MCP protokolünün temel bileşenlerinin güvenlik zafiyetleri.
- Otonom ajanların potansiyel saldırılara karşı nasıl savunulabileceği.
- Savunma stratejilerinin protokol düzeyinde uygulanabilirliği.
- Güvenli bir dağıtım süreci için gerekli en iyi uygulamalar.

2 MCP Mimarisi ve Tehdit Yüzeyi

MCP, dört ana bileşenden oluşur: İstemci (Client), Sunucu (Server), Araçlar (Tools) ve Taşıma Katmanı (Transport Layer). Bu bileşenlerin her biri farklı güvenlik tehditlerine açıktır ve güvenli bir sistem inşa etmek için her bir bileşenin özelliklerinin anlaşılması gereklidir.



Şekil 1: Model Bağlam Protokolü (MCP) Mimarisi

2.1 MCP Bileşenleri ve Güvenlik Sınırları

MCP'nin temel yapısı ve bileşenleri, onun nasıl işlediğini ve bu süreçte hangi zayıflıkların ortaya çıkabileceğini anlamak için oldukça önemlidir. MCP protokolü, istemci ve sunucu arasındaki güvenli iletişimi sağlamak amacıyla tasarlanmış bir protokoldür. Bu protokoldeki güvenlik sınırları, her bileşenin erişim yetkilerine göre tanımlanır:

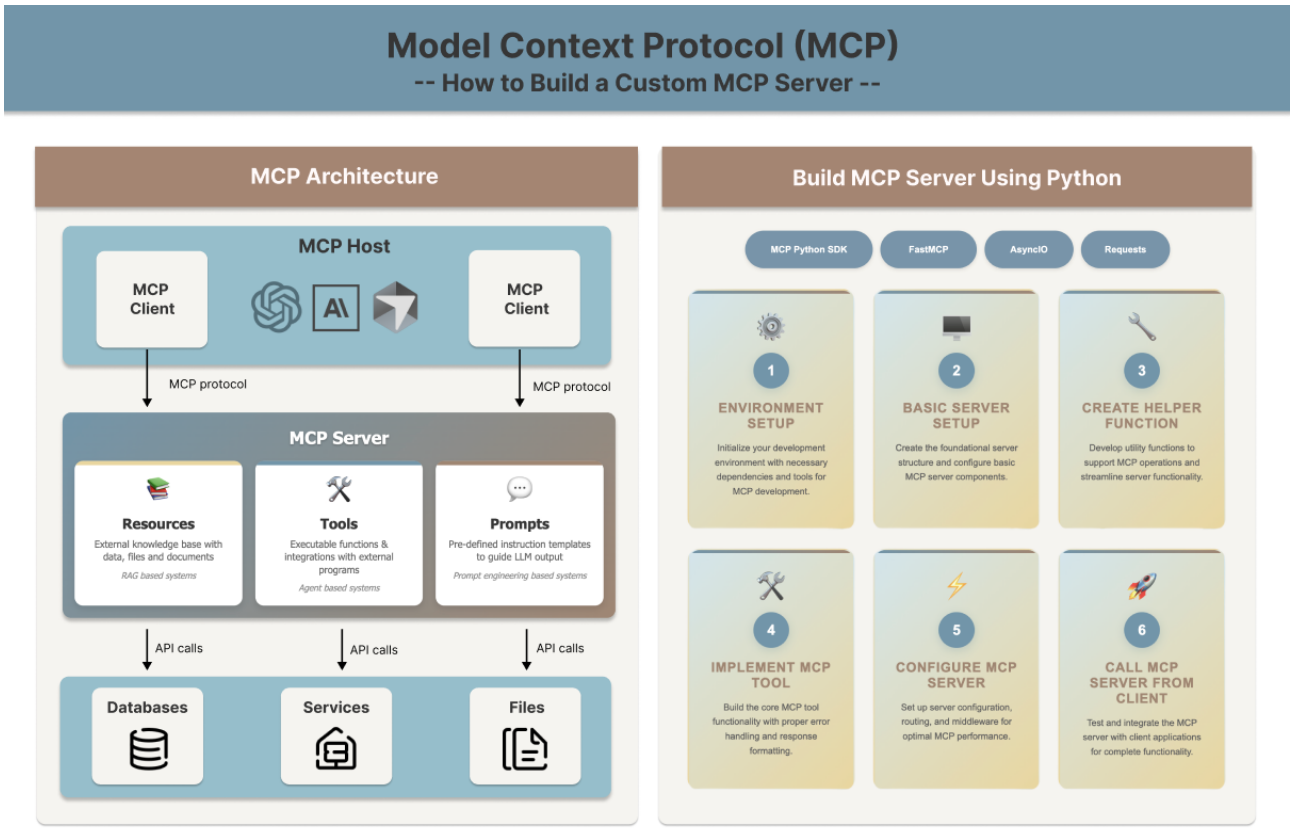
- **İstemci (Client):** LLM'leri barındıran ortamdır. Bu bileşen, dış sistemlerle etkileşimde bulunurken en zayıf nokta olabilir. İstemci, harici verilerle etkileşime girerken büyük güvenlik riskleri taşıyabilir. Örneğin, istemciye dışardan gönderilen kötü niyetli girdiler, ajanın güvenliğini tehlikeye atabilir.

- **Sunucu (Server):** Harici sistemlere veya verilere erişim sağlayan merkezi bileşendir. Sunucu, harici API'ler ve araçlarla bağlantı kurarak ajanı dış dünyaya bağlar. Ancak, sunuculara yönelik yapılan saldırılar, sistemin bütünlüğünü ciddi şekilde bozabilir.
- **Araçlar (Tools):** MCP sunucusuna entegre edilmiş harici işlevler veya araçlar. Araçlar, ajanların görevlerini yerine getirmesi için kullanılır, ancak kötü niyetli araçlar ajanları manipüle edebilir.
- **Taşıma Katmanı (Transport Layer):** Sunucu ve istemci arasındaki veri iletimini ve mesajlaşmayı sağlayan güvenli bir kanaldır. Genellikle SSL/TLS şifreleme protokolleri kullanılır. Ancak, taşıma katmanındaki zayıflıklar, tüm sistemin güvenliğini tehdit edebilir.

2.2 Tehdit Yüzeyi ve Güvenlik Zafiyetleri

MCP'nin mimarisi, geniş bir saldırı yüzeyi sunar. Bu yüzeydeki zafiyetler şunları içerebilir:

- **Araç Zehirlenmesi (Tool Poisoning):** Sunucu veya istemciye kötü niyetli komutlar enjekte edilmesi, aracın manipüle edilmesine yol açabilir. Bu, ajanların yanlış işlevleri çağırmasına ve dolayısıyla sistemin yanlış çalışmasına neden olabilir.
- **Prompt Enjeksiyonu (Prompt Injection):** Kullanıcı girdilerine kötü niyetli talimatlar eklenerek ajanın davranışı değiştirilir. Bu saldırılar, ajanın davranışını manipüle etmek için hedeflenebilir.
- **Kimlik Doğrulama Eksiklikleri:** Yetersiz kimlik doğrulama sistemleri, izinsiz erişime neden olabilir. Bu zafiyet, iç ve dış tehdit aktörleri tarafından kolayca sömürülebilir.
- **Veri Sızıntıları (Exfiltration):** Harici API'ler veya araçlar üzerinden kritik verilerin dışarı sızması. Bu durum, ajanın gizlilik politikalarını ihlal ederek hassas bilgilerin dışarıya çıkmasına yol açabilir.



Şekil 2: Model Bağlam Protokolü (MCP) Modeli

3 Tehdit Modeli ve Saldırı Vektörleri

MCP ile entegre edilmiş sistemlerdeki tehditlerin başında, ajanların güvenli olmayan verilerle etkileşime girmesi ve bunun sonucunda kritik verilerin sızması yer almaktadır. Bu bölüme göre, tehdit aktörleri, saldırı senaryoları ve bu senaryolara karşı geliştirilmesi gereken savunma önlemleri detaylandırılacaktır.

3.1 Tehdit Aktörleri ve Senaryoları

MCP'nin güvenliğini tehdit eden aktörler ve olası saldırılar şunlar olabilir:

- **Dış Saldırganlar:** Kötü niyetli bireyler veya hacker grupları, MCP sisteminin zafiyetlerinden yararlanarak ajanın planlama mantığını manipüle edebilirler. Bu tür saldırılar, ajanın dışarıya veri sızdırmasına veya yetkisiz işlemler gerçekleştirmesine neden olabilir.
- **İç Saldırganlar:** Yetkili kişilerin sistemin güvenlik önlemlerini atlatmak için kötü niyetli davranışlar sergileyebileceği durumlar. İç saldırganlar, özellikle güçlü erişim haklarına sahip olduklarında ciddi tehditler oluşturabilir.
- **Zararlı Yazılımlar:** MCP protokolünü hedef alarak ajanın işlevlerini manipüle eden yazılımlar. Zararlı yazılımlar, ajanın veri akışını yönlendirerek gizli bilgilerin sızdırılmasına yol açabilir.

3.2 Saldırı Vektörleri

MCP, bir dizi saldırı vektörüne açıktır. Bu vektörler, ajanın güvenliğini tehdit edebilir ve farklı yollarla sisteme zarar verebilir. Öne çıkan saldırı vektörleri şunlardır:

- **Araç Zehirlenmesi ve Plan Enjeksiyonu:** Ajanın karar verme sürecine kötü niyetli talimatlar enjekte edilmesi. Bu, ajanın görevleri yerine getirirken yanlış kararlar almasına yol açabilir.
- **DoS (Denial of Service) ve Finansal DoS:** Ajanların gereksiz kaynak harcamalarına yol açan saldırılar, sistemin işleyişini zorlaştırabilir. Örneğin, aşırı token kullanımı veya maliyetli API çağrılarının yapılması.
- **Veri Sızıntısı (Exfiltration):** Ajana kötü niyetli araçlar aracılığıyla hassas verilerin dışarı sızdırılması. Bu tür saldırılar, ajanın görevlerini kötüye kullanarak önemli bilgilerin çalınmasına neden olabilir.

4 Savunma Stratejileri ve İyi Uygulamalar

Gelişmiş savunma stratejileri, sistemin zafiyetlerine karşı korunması için kritik öneme sahiptir. Bu bölümde, MCP protokolünün güvenliğini sağlamak için önerilen stratejiler ele alınacaktır.

4.1 Sistem Düzeyinde Savunmalar

Sistem düzeyinde alınması gereken savunmalar:

- **Bilgi Akışı Kontrolü (IFC):** Ajana giren ve çıkan tüm verilerin izlenmesi ve etiketlenmesi. Bu mekanizma, zehirli verilerin ajanın muhakeme akışına dahil olmasını engeller.
- **Dinamik Taint-Tracking:** Güvenlik açısından şüpheli verilerin izlenmesi ve yanlış eylemlerle sonuçlanmaması için engellenmesi. Bu yöntem, özellikle veri sızıntılarına karşı etkilidir.
- **Muhakeme ve Planlama Doğrulaması:** Ajana verilen görevlerin doğruluğunun sağlanması için sürekli olarak plan doğrulama mekanizmalarının uygulanması. Böylece, yanlış veya zararlı görevlerin ajana verilmesi engellenir.

4.2 Taşıma Katmanı ve Kimlik Doğrulama Stratejileri

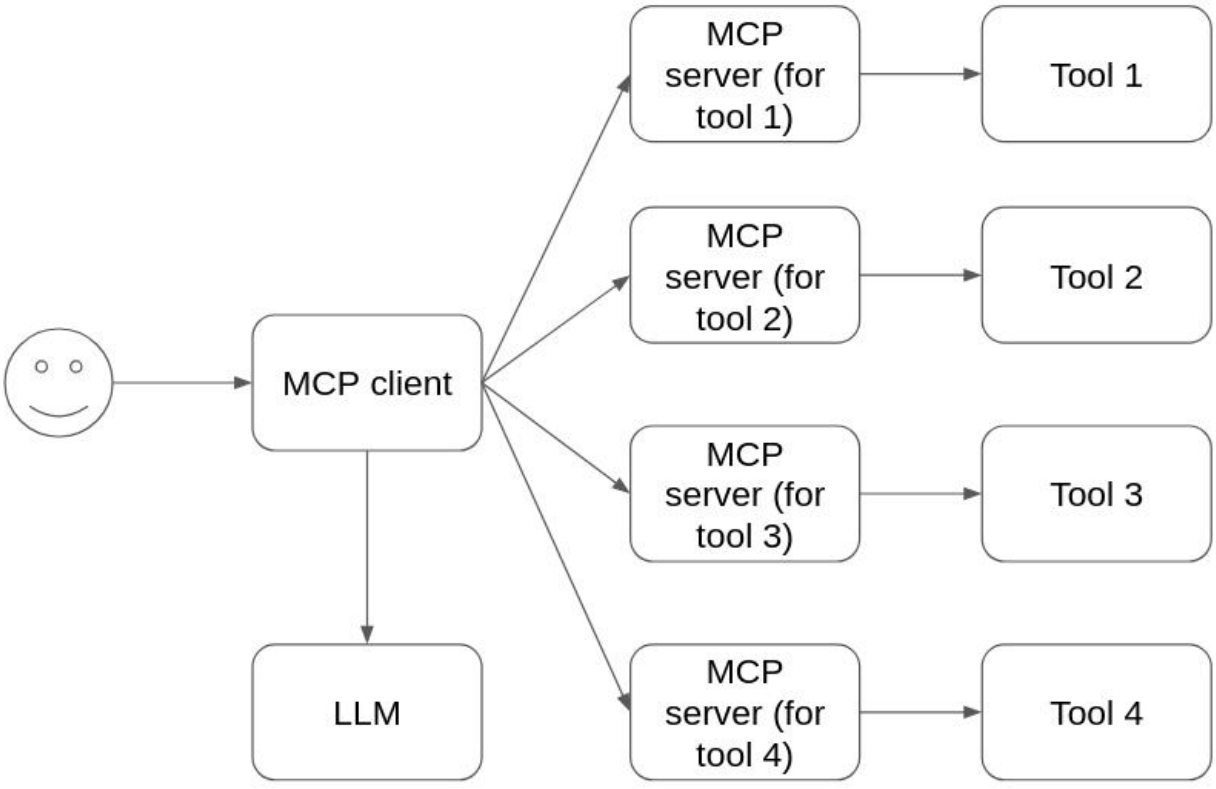
Taşıma katmanında, güvenli veri iletimi için şu önlemler alınabilir:

- **SSL/TLS Şifrelemesi:** Verinin iletilmesi sırasında güvenliği sağlamak için şifreleme protokollerinin uygulanması. Bu, verilerin güvenliğini artırır ve dinleme saldırılarına karşı korur.
- **OAuth 2.0 Kimlik Doğrulaması:** İstemci ve sunucu arasındaki güvenli bağlantının sağlanabilmesi için güçlü kimlik doğrulama protokollerinin kullanılması. Bu, yalnızca yetkili kullanıcıların erişim sağlamasına olanak tanır.

4.3 Ajan Güvenliği ve İzleme

Ajansların güvenliği için izleme ve denetleme mekanizmaları gereklidir:

- **Sandboxing:** Ajanların yalnızca izole edilmiş ortamda çalışması sağlanarak, kötü niyetli eylemlerin yayılmasının önlenmesi. Sandboxing, ajanın dışarıya zarar vermesini engeller.
- **Görev Denetimi:** Ajana verilen görevlerin sürekli olarak izlenmesi ve yanlış bir karar almasını önlemek için güvenlik önlemlerinin devreye girmesi. Bu mekanizma, ajanın görev dışı hareketlerini engelleyebilir.



Şekil 3: Model Bağlam Protokolü (MCP) Protokol Akışı

5 Sonuç ve Gelecek Yönelimleri

Model Context Protocol (MCP), otonom ajanların etkin bir şekilde harici sistemlerle etkileşimde bulunmalarını sağlayan bir protokoldür. Ancak bu yeni protokol, aynı zamanda güvenlik açısından çeşitli riskler taşımaktadır. Bu çalışmada, MCP'nin güvenlik tehditleri ve savunma stratejileri detaylı bir şekilde incelenmiştir. Gelecekteki çalışmalar, bu tehditlere karşı daha ileri düzey savunma stratejileri geliştirmeye ve sistemin güvenliğini artırmaya odaklanmalıdır.