**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**

**ĐỒ ÁN MÔN AN TOÀN MẠNG**

**LỚP NT140.P11.ANTT NHÓM 18**

**ĐỀ TÀI:**

**LLM Platform Security:**

**Applying a Systematic Evaluation Framework**

**to OpenAI’s ChatGPT Plugins**

**GIẢNG VIÊN HƯỚNG DẪN**

**ThS. Nghi Hoàng Khoa**

**DANH SÁCH THÀNH VIÊN NHÓM 18**

1. Trần Thái Huy – 22520580
2. Lê Ngọc Hiếu – 22520435
3. Trần Minh Hiếu– 22520445
4. Đào Nhật Duy – 22520312

**TP. HỒ CHÍ MINH, NĂM 2024**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

..........................................................................................................................................

*……………, ngày … tháng … năm 202…*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên****)***

**LỜI CẢM ƠN**

Chúng em xin gửi lời tri ân sâu sắc đến Trường Đại học Công nghệ Thông tin và Khoa Mạng máy tính và Truyền thông vì đã tạo môi trường học tập lý tưởng, nơi chúng em có cơ hội tiếp cận và nghiên cứu môn An toàn mạng. Môn học này không chỉ cung cấp những kiến thức nền tảng quan trọng mà còn giúp chúng em mở mang tầm nhìn về thực trạng và tương lai của lĩnh vực an toàn thông tin trong thực tế.

Chúng em đặc biệt ghi nhận sự hướng dẫn tận tình từ thầy Nghi Hoàng Khoa trong phần lý thuyết và thầy Tô Trọng Nghĩa trong phần thực hành. Các thầy không chỉ truyền đạt kiến thức chuyên môn mà còn truyền cảm hứng cho chúng em nỗ lực học hỏi, vượt qua thử thách trong quá trình thực hiện đồ án này.

Tuy nhiên, do thời gian hạn chế và những thiếu sót không thể tránh khỏi, bài báo cáo này có thể chưa đạt được sự hoàn thiện như mong đợi. Chúng em rất mong nhận được những góp ý quý báu từ các thầy cô để có thể cải thiện và phát triển tốt hơn trong tương lai.

Một lần nữa, chúng em xin gửi lời cảm ơn chân thành đến tất cả những người đã đồng hành và hỗ trợ chúng em trong suốt hành trình học tập và thực hiện đồ án này.

TP. Hồ Chí Minh, ngày tháng năm 2024

Nhóm 18

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 5](#_Toc184514528)

[DANH MỤC TỪ VIẾT TẮT 6](#_Toc184514529)

[Chương 1. GIỚI THIỆU TỔNG QUAN VÀ THỰC TRẠNG. 1](#_Toc184514530)

[1.1. Giới thiệu tổng quan. 1](#_Toc184514531)

[1.2. Thực trạng. 1](#_Toc184514532)

[Chương 2. KIẾN TRÚC CỦA PLUGIN TRONG NỀN TÀNG LLM. 2](#_Toc184514533)

[2.1. So sánh kiến trúc của plugin và cách tương tác với LLM với Actions tích hợp vào GPTs. 2](#_Toc184514534)

[2.1.a. Cách thức hoạt động của plugin. 3](#_Toc184514535)

[2.1.b. Cách thức hoạt động của Actions, được tích hợp vào các GPT hiện nay. 4](#_Toc184514536)

[2.2. Ảnh hưởng của sự thay đổi từ plugin sang GPT đối với đồ án. 6](#_Toc184514537)

[3. Phương pháp nghiên cứu của bài báo. 6](#_Toc184514538)

[3.1. Framework. 6](#_Toc184514539)

[3.2. Phương pháp thực hiện. 7](#_Toc184514540)

[TÀI LIỆU THAM KHẢO 7](#_Toc184514541)

# DANH MỤC HÌNH ẢNH

[Hình 1: Quy trình tương tác của plugin đến nền LLM khi user tạo một câu lệnh. 3](#_Toc184597382)

[Hình 2: Disclaimer của GPT SSH. 9](#_Toc184597383)

[Hình 3: SSH GPT dùng credentials cung cấp SSH đến máy. 9](#_Toc184597384)

[Hình 4: Public SSH key của GPT SSH 10](#_Toc184597385)

# DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Ý nghĩa** |
| LLM | Large Language Model |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Chương 1. GIỚI THIỆU TỔNG QUAN VÀ THỰC TRẠNG.

## Giới thiệu tổng quan.

Các mô hình ngôn ngữ lớn (LLM - Large Language Model ), chẳng hạn như ChatGPT, đã dánh dấu bước tiến lớn trong lĩnh vực trí tuệ nhân tạo (AI – Artificial Intelligence). Các bước tiến này được thể hiện qua khả năng xử lý ngôn ngữ tự nhiên cũng như cho phép phần lớn người dùng tiếp cận công nghệ này. Không chỉ dừng lại ở việc hỗ trợ trò chuyện, các nền tảng LLM đã mở rộng khả năng thông qua việc tích hợp plugin của bên thứ ba. Các plugin cho phép LLM kết nối với các dịch vụ bên ngoài, truy cập dữ liệu thời gian thực và thực hiện các tác vụ phức tạp hơn so với trước đây.

ChatGPT, nền tảng LLM phổ biến nhất hiện nay đã giới thiệu hệ sinh thái plugin để nâng cao khả năng của chatbot và phạm vi ứng dụng. Người dùng có thể thông qua các plugin, thực hiện các tác vụ như tìm kiếm thông tin, đặt vé, quản lí dữ liệu. Tuy nhiên cùng với tiềm năng mở rộng này là các rủi ro đáng kể liên quan đến an toàn, bảo mật, và quyền riêng tư. Điều này càng đáng quan ngại hơn khi các plugin này được phát triển bởi bên thứ ba với ChatGPT đóng vai trò như một kiểm duyệt viên.

Hiện nay cơ chế Plugin đã được OpenAI chuyển sang Actions, với cơ chế hoạt động tương tự và được tích hợp vào GPTs[[1]](#footnote-2)

## Thực trạng.

Trong hệ sinh thái các dịch vụ của nền tảng LLM như Chatgpt thì việc tích hợp plugin của bên thứ ba đã mở rộng đáng kể khả năng đáp ứng như cầu của người dùng. Tuy nhiên, điều này cũng mang lại những thách thức liên quan đến an toàn, bảo mật, và quyền riêng tư. Nhiều nghiên cứu đã chỉ ra rằng các plugin nói riêng và nền tảng LLM nói chung chứa các nguy cơ tiềm ẩn hay hành vi bất thường[[2]](#footnote-3),[[3]](#footnote-4).

Bên cạnh đó, sự phụ thuộc vào ngôn ngữ tự nhiên (natural language) để mô tả chức năng của plugin làm cho việc kiểm soát và giám sát trở nên khó khăn hơn. Chẳng hạn, một plugin có thể sử dụng mô tả chức năng mơ hồ hoặc quá rộng, dẫn đến tình trạng "lấn chiếm chức năng" (functionality squatting) hoặc "lấn chiếm chủ đề" (topic squatting), làm ảnh hưởng đến hiệu quả của toàn hệ thống. Thêm vào đó, khi các plugin này hợp tác hoặc chia sẻ dữ liệu không minh bạch với nhau, nguy cơ lạm dụng dữ liệu cá nhân của người dùng tăng cao, vi phạm quyền riêng tư nghiêm trọng.

Vấn đề nói trên, cùng với chính sách quản lí plugin chưa nghiêm ngặt của OpenAI, bài báo đã đề xuất một khung đánh giá (evaluation framework) nhằm nhận diện, phân loại, phân tích hệ sinh thái Plugin mà ChatGPT đã áp dụng đến khoảng giữa năm 2024[[4]](#footnote-5).

Hiện tại một số chức năng của các plugin như Browsing đã được tích hợp vào ChatGPT,

# Chương 2. KIẾN TRÚC CỦA PLUGIN TRONG NỀN TÀNG LLM.

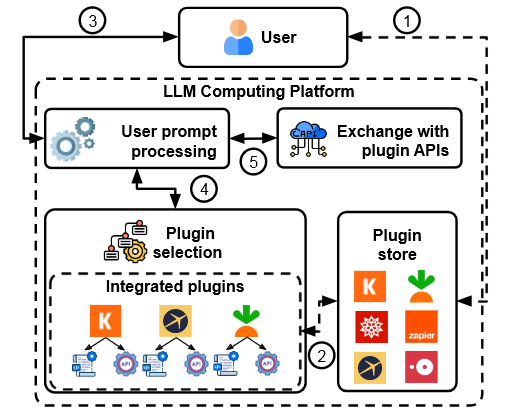
Bản thân các nền tàng LLM không thể thực hiện các tác vụ phức tạp cũng như truy cập các dữ liệu ngoài training data. Các plugin cho nền tảng LLM, ví dụ như OpenAI giúp mở rộng các tính năng bằng cách cho phép endpoints của LLM kết nối tới các dịch vụ thông qua API, truy cập được dữ liệu thời gian thực hoặc giới hạn.

## 2.1. So sánh kiến trúc của plugin và cách tương tác với LLM với Actions tích hợp vào GPTs.

### 2.1.a. Cách thức hoạt động của plugin.

Các plugin cho OpenAI sẽ bao gồm một đoạn manifest và specifications chứa:

* Manifest: phiên bản, tên của plugin (dành cho LLM), tên của plugin (dành cho người dùng), mô tả của plugin (dánh cho LLM), mô tả của plugin (dành cho người dùng), yêu cầu đăng nhập (có hoặc không), loại API, URL dẫn tới specification, thông tin pháp lý.
* Specification: phiên bản OpenAI, tên của plugin, mô tả plugin, URL máy chủ (ví dụ: <https://kayak.com>), các endpoints cho LLM, các schema.



Hình 1: Quy trình tương tác của plugin đến nền LLM khi user tạo một câu lệnh.[[5]](#footnote-6)

OpenAI yêu cầu rằng plugin phải sử dụng HTTPS cho mọi giao tiếp với nền tảng LLM, xây dựng các luồng xác nhận cho những yêu cầu có thể thay đổi dữ liệu người dùng, ví dụ như thông qua các yêu cầu POST, sử dụng xác thực nếu plugin thay người dùng thực hiện một hành động, không sử dụng các mô hình tạo ảnh không phải của OpenAI, tuân thủ chính sách nội dung của OpenAI, tuân theo hướng dẫn liên quan đến thương hiệu của OpenAI, cùng các yêu cầu khác được đề cập trong quy trình đánh giá plugin.

Ngoài ra, OpenAI cũng: tuyên bố rằng họ sẽ gỡ bỏ các plugin nếu chúng thay đổi, hạn chế giao tiếp chỉ với tên miền gốc của plugin, và chỉ chuyển các mã định danh người dùng (user identifiers) không tồn tại quá một ngày hoặc ngoài phạm vi một phiên trò chuyện.

OpenAI khuyến nghị rằng plugin nên triển khai giới hạn tốc độ yêu cầu API (rate limits) và cung cấp một dải địa chỉ IP của máy chủ OpenAI để các plugin có thể thêm vào danh sách cho phép (allow list).[[6]](#footnote-7)

### 2.1.b. Cách thức hoạt động của Actions, được tích hợp vào các GPT hiện nay.

GPT là **[NEED CONTENT HERE YOOOOOOOOOOOOOOOOOOOO]**

*OpenAI Actions* là một hệ thống giúp GPT có thể thực hiện các nhiệm vụ bên ngoài, như gọi API hoặc tương tác với dịch vụ khác. Nó có các chức năng giống như các *plugin*, mở rộng khả năng của các GPT. Nhưng có một số điểm khác biệt đáng lưu ý:

|  |  |  |
| --- | --- | --- |
| Chức năng | Plugins | Actions -> GPTs |
| Triển khai | Được host ở máy chủ của bên thứ ba đã tạo ra nó. | Được host bởi OpenAI. |
| Quản lý | Phụ thuộc vào bên thứ ba, OpenAI chỉ đóng một phần vai trò quản lý. | Quản lý hoàn toàn bởi OpenAI |
| Tương tác với LLM | Plugin tương tác với LLM bằng ngôn ngữ tự nhiên. | Actions được tích hợp trực tiếp vào các chức năng của GPTs. |

Hiện tại, phần lớn các plugins được nhắc đến/thử nghiệm trong bài viết đã bị gỡ bỏ, ngừng hoạt động sau khi OpenAI chuyển sang mô hình Actions/GPTs. Bảng sau liệt kê các plugin chính, trạng thái, cũng như một số GPT tương tự chúng (nếu có) dựa vào dữ liệu của website [https://www.whatplugin.ai/](https://www.whatplugin.ai/%20) và <https://github.com/copilot-us/chatgpt-plugins> .

|  |  |  |  |
| --- | --- | --- | --- |
| Plugins | Chức năng | Trạng thái | GPT tương tự |
| AMZPRO | Hỗ trợ tạo các listing trên Amazon | Ngừng hoạt động | Ecommerce SEO Product Description Writer AI GPT |
| AI Agents | Sử dụng nhiều AI agent cùng lúc | Ngừng hoạt động | Code Copilot |
| AutoInfra1 | Quản lý server qua SSH, hỗ trợ AWS, GCP, Azzure | Ngừng hoạt động |  |
| ChatSSHplug | SSH đến server và thực thi lệnh dựa trên đầu vào là ngôn ngữ tự nhiên | Ngừng hoạt động | GPT SSH |
| Clinical Trial Radar | Tìm kiếm, phân tích liên quan đến y khoa. | Ngừng hoạt động | GCP |
| LexiShopper | Hỗ trợ mua sắm trên Amazon | Ngừng hoạt động | Fashion Men & Women Outfit| Shopping |Budget | GPT AI |
| Magic | Virtual Worker | Ngừng hoạt động |  |
| Playlist AI | Tạo playlist trên các nền tảng streaming: Amazon, Deezer, Spotify. | Thay đổi thành GPT | Playlist AI |
| ReflectNotes | Tạo note trên Reflect App | Ngừng hoạt động |  |
| Tira | Hỗ trợ mua sắm trên Tira | Ngừng hoạt động |  |
| KAYAK | Hỗ trợ tìm kiếm liên quan đến du lịch trên KAYAK.com | Thay đổi thành GPT | KAYAK |
| Upskillr | Tạo khóa học nâng cao kỹ năng cho người dùng. | Ngừng hoạt động |  |
| PDF Exporter | Tạo file PDF dựa trên nội dung chat | Ngừng hoạt động | AI PDF Drive |
| Expedia | Hỗ trợ lên kế hoạch du lịch | Ngừng hoạt động | Travel Guide |

## 2.2. Ảnh hưởng của sự thay đổi từ plugin sang GPT đối với đồ án.

# Chương 3. Phương pháp nghiên cứu của bài báo.

## 3.1. Framework.

Bài báo trình bày một framework nhằm đánh giá một cách hệ thống tính an toàn, bảo mật, và quyền riêng tư của các nền tảng LLM tích hợp plugin (ChatGPT được lấy làm ví dụ). Framework này bao gồm một phân loại tấn công (attack taxonomy), được phát triển thông qua quá trình phân tích khả năng của các bên liên quan (plugin, người dùng, nền tảng LLM) và mô hình hóa các mối đe dọa.

Mục tiêu của framework:

* Thực tiễn (Actionable): Cung cấp một cấu trúc rõ ràng để xây dựng phân loại tấn công cho việc phân tích bảo mật đối với các nền tảng LLM khác nhau.
* Toàn diện (Extensive): Bao quát cả các cách thức tấn công hiện có (như ở IOT, web, mobile) và mới xuất hiện trong bối cảnh LLM.
* Có thể mở rộng:Có thể tích hợp thêm các hình thức tấn công hoặc điều chỉnh theo sự phát triển của các nền tảng LLM trong tương lai.
* Thông tin thực tiễn: Framework được xây dựng dựa trên thực nghiệm và dữ liệu thực tế từ các plugin hiện có.

## 3.2. Quá trình xây dựng framework.

Framework được xây dựng qua các bước:

* Tổng hợp và phân loại tấn công từ nghiên cứu trước đây: Thu thập các hình thức tấn công từ các nền tảng tích hợp bên thứ ba như web, di động, và IoT.
* Phân tích khả năng các bên liên quan: Đánh giá vai trò và trách nhiệm của plugin, người dùng, và nền tảng LLM để xác định cách các bên này có thể thực hiện hoặc trở thành nạn nhân của các cuộc tấn công.
* Mô hình hóa mối đe dọa: Sử dụng kỹ thuật mô hình cây tấn công (attack tree) để xác định các kịch bản tấn công mới.
* Cải tiến thông qua thực nghiệm: Phân tích các plugin thực tế trên nền tảng OpenAI, đánh giá khả năng xảy ra các hình thức tấn công trong phân loại tấn công và tinh chỉnh framework dựa trên các phát hiện thực tế.

## 3.3. Quá trình thực hiện bài báo.

1. Tác giả thu thập manifest, specification, và chính sách bảo mật của 268 plugin từ OpenAI Plugin Store (tính đến ngày 6/6/2023).
2. Xem xét chức năng của plugin, luồng dữ liệu, và các yêu cầu truy cập (ví dụ: yêu cầu xác thực hoặc liên kết tài khoản).
3. Tương tác với plugin trên ChatGPT để kiểm chứng các giả thuyết về các hình thức tấn công.
4. Khi phát hiện ra các hướng tấn công mới, xem lại và chỉnh sửa phân loại tấn công.

Upon disclosing to plugin vendors, we learned that in at least

one case the plugin vendor also disclosed the situation to

OpenAI because OpenAI (not them) were in the position to

fix the issue, but OpenAI did not. (fix later)

# Chương 4. Tấn công giữa plugin và user.

## Chiếm quyền kiểm soát máy người dùng

Một số phương thức plugin có thể chiếm quyền kiếm soát máy người dùng được đề cập:

* Bên tấn công có thể đánh lừa người dùng cài đặt các plugin chưa được kiểm duyệt và không nằm trên OpenAI Plugin Store (qua developer mode). Các plugin này có thể chứa malware.
* Plugin có thể khuyên người dùng truy cập các website không đáng tin cậy.
* Người quản lý plugin có thể có khả năng lấy các dữ liệu cần thiết cho các task như: quản lý máy tính/server từ xa (AutoInfra1, ChatSSHPlug) → ***Rủi ro 1***

### ****4.1.a. Rủi ro 1: Khai thác thông tin.****

Các plugin hỗ trợ tự động hóa máy chủ và quản lý thông qua SSH thường yêu cầu truy cập dữ liệu như khóa SSH hoặc thông tin đăng nhập máy chủ:

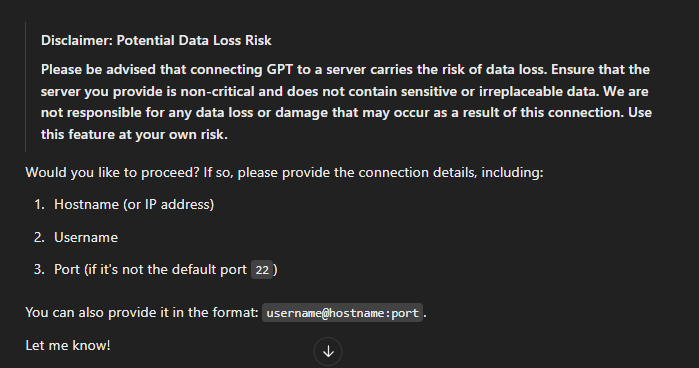
* AutoInfra1: Yêu cầu người dùng thêm public key của plugin vào tệp authorized keys trên SSH và cung cấp địa chỉ IP công cộng.
* ChatSSHPlug: Yêu cầu người dùng cung cấp trực tiếp mật khẩu hoặc private key.

ChatSSHPlug, trước khi gỡ bỏ và sau khi tác giả hoàn thành bài báo, đã chuyển sang sử dụng Oauth (giống như tác giả đã khuyến nghị). Thông tin được lấy từ kênh Youtube của tác giả plugin: <https://www.youtube.com/watch?v=1xzJT03C_Tw> . Tuy nhiên người dùng vẫn dựa vào độ đáng tin cậy của nơi thực hiện việc tương tác với server chatsshplug.com (hiện không còn tồn tại).

**Rủi ro này có thể dẫn tới hậu quả nghiêm trọng. Ví dụ một nhà phát triển plugin, nếu có được credentials có khả năng điều khiển được máy của người dùng.**

### **4.1.b. Thử nghiệm trên GPT SSH.**

Do các plugin đề cập trong bài báo (AutoInfra1, ChatSHPlug) không còn tồn tại nên tiến hành thử nghiệm trên GPT SSH.



Hình 2: Disclaimer của GPT SSH.

A screenshot of a computer

Description automatically generated

Hình 3: SSH GPT dùng credentials cung cấp SSH đến máy.

A screenshot of a computer

Description automatically generated

Hình 4: Public SSH key của GPT SSH

Thông qua quá trình thử nghiệm trên GPT SSH, có thể thấy rằng GPT này hoạt động tương tự với ChatSSHPlug ban đầu: yêu cầu cung cấp credentials, sau đó thông qua một server bên ngoài thực thi các commands mà ChatGPT trả về. Ở đây GPT SSH sử dụng <https://ssh.bulkninja.com/>. GPT này hiện đang tồn tại Risk 1 trong bài báo do sử dụng phương thức không an toàn để xác thực người dùng (đưa credentials ra ngoài cho một server khác xử lý).

OpenAI cũng khuyến cáo người dùng chỉ nên gửi thông tin cho các bên thứ 3 đáng tin cậy và yêu cầu người dùng cho phép trước khi GPT giao tiếp với một server bên ngoài.

## Chiếm dụng tài khoản của người dùng.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | F. P. a. I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," *arXiv preprint arXiv:2211.09527,* 2022. |

1. <https://platform.openai.com/docs/actions/introduction> [↑](#footnote-ref-2)
2. F. Perez and I. Ribeiro, “Ignore Previous Prompt: Attack Techniques for Language Models,” arXiv:2211.09527 [cs], Nov. 2022, Available: https://arxiv.org/abs/2211.09527 [↑](#footnote-ref-3)
3. Y. Liu et al., “Prompt Injection attack against LLM-integrated Applications,” arXiv.org, Jun. 08, 2023. https://arxiv.org/abs/2306.05499 [↑](#footnote-ref-4)
4. <https://help.openai.com/en/articles/8988022-winding-down-the-chatgpt-plugins-beta> [↑](#footnote-ref-5)
5. U. Iqbal, Tadayoshi Kohno, and F. Roesner, “LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI’s ChatGPT Plugins,” *arXiv (Cornell University)*, Sep. 2023, p. 2, doi: https://doi.org/10.48550/arxiv.2309.10254. [↑](#footnote-ref-6)
6. U. Iqbal, Tadayoshi Kohno, and F. Roesner, “LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI’s ChatGPT Plugins,” arXiv (Cornell University), Sep. 2023, p. 3, doi: https://doi.org/10.48550/arxiv.2309.10254. [↑](#footnote-ref-7)