Security Concerns with Commonly Used AI Tools

By: Jeffrey Caleb Hendrix

Generative AI and related tools have become increasingly utilized by professionals, educators, and students across the United States and worldwide. However, the question remains: what are we risking using these platforms? This is a review of articles published by experts in the field, and the conversation will discuss potential vulnerabilities from 3rd party intervention, such as accessing sensitive information, indirect prompt intervention, and the misinformation and theft that derive from such nefarious behaviors. Importantly, this discussion will focus solely and specifically on Microsoft's Copilot, OpenAI's ChatGPT, and Google's Gemini artificial intelligence programs. In addition to raising concerns, this review will highlight preventative steps recommended in this field for reducing the potential exposure of sensitive information. Experts from Forbes establish that through simply opting out of default terms, individuals reduce the likelihood of data being involved in a large breach.[1] Further, the report adds that parties can gain security against personalized attacks by implementing safeguards within their model or personal system and utilizing these tools professionally. Id.

Microsoft Co-Pilot and SharePoint

Pen Test Partners, a security firm in the space, first established a necessary baseline for understanding SharePoint as it operates in the Microsoft Co-Pilot world, which is the concept of "Agents", which can be *Default* (created by Microsoft), or *Custom* (crafted by private organizations).[2] Further, while agents are essential in allowing the Co-Pilot platform to better assist its users by accessing their documents and much more, that corresponding function makes them targets for exploitation. Id. The company claims that through communicating directly with Agents, attackers can manipulate these entities into unveiling personal information such as keys, passwords, as well as provide themselves with a deep understanding of the application's vulnerabilities and internal systems'

[1] https://www.forbes.com/sites/alexvakulov/2025/01/22/how-to-use-chatgpt-and-other-ai-chatbots-securely/

[2] https://www.pentestpartners.com/security-blog/exploiting-copilot-ai-for-sharepoint/

operating functions. Id. Adding that, attackers can pose as security teams at the respective organization to trick the agent into providing access to sensitive information, all while the agent responds as if the request were legitimate. Id. Furthermore, as all information is freely copiable that is provided by Co-Pilot, this sensitive data is no different, and even if the agent initially blocks access, the model may still provide it with persistence from attackers. Id. Clarifying that the ability to use Agents to comb through vast data in SharePoint is extremely beneficial to attackers, yet detrimental to normal users and, more importantly, organizations. Id. Finally, when it comes to determining whether a breach has occurred, the source establishes that "accessed by" and "recent file" logs utilized by SharePoint are helpless as this form of attack does not trigger any update or sound an alarm. Id.

How can this be addressed?

Cybersecurity experts stress the importance of minimizing the exposure of sensitive information in Microsoft, the Co-Pilot platform, the agents, and, of course, SharePoint itself. 3 One recommendation from both pen-test partners widely renowned security firm, as well as the article by Forbes, establishes that the only sure-fire method of protection is preventing all sensitive information from existing in SharePoint. Id. However, where this storage of information in SharePoint is necessary, researchers advocate for individuals and corporations to implement appropriate access controls that can limit the Agent's scope of discovery. Id at 2. Further, additional safeguards in this literature include restricting the creation of agents or mandating approval of any new agent, which experts say can be achieved through altering the configuration of one's site. Id. Crucially, professionals such as Pen Test establish that when it comes to securing information in this digital age, it is essential not to

[3] https://www.forbes.com/sites/daveywinder/2025/05/14/new-warning---Microsoft-copilot-ai-can-access-restricted-passwords/

put all your eggs in one basket. Id. Further clarifying that while preventative measures can be a great first line of defense, adding secondary protections like monitoring aids substantially in preventing villainous individuals from accessing sensitive files unbeknownst to the user. Id.

Open-AI and the Various Chat-GPT Models

With Chat-GPT currently being the most utilized LLM and Artificial intelligence tool, the security concerns are heightened as the number of users and level of usage increase. Earlier this year, there was an alleged breach of nearly 20 thousand users' personal information, including but not limited to passwords[4]. However, Forbes establishes that OpenAI remains confident that their systems were not compromised, and rather that the data was derived or available from unrelated infostealer logs. Id. Adding that, these hackers had isolated the individuals' ChatGPT login information through these larger data sets of stolen information and used this to invoke panic in the AI space. Id. Nonetheless, a separate Forbes article has reiterated that there are real security concerns through the notion of prompt injection in all the major LLM models by using fiction to override safeguards in place[5]. The same expert adds that while this is unlikely to result in a third- party individual accessing secure data in everyday usage, as we see AI become prevalent in the medical and legal field, these risks increase. Id. In domains like Healthcare and Law, experts acknowledge that creating hypothetical situations could allow the chatbot to expose private patient/client data unknowingly. Id. Further, the lack of notification of third parties accessing this data could mean a breach can occur unbeknownst to the user and the organization that they are serving. Id.

[4] https://www.forbes.com/sites/daveywinder/2025/02/11/has-openai-been-hacked-what-20-million-users-need-to-know/

[5] https://www.forbes.com/sites/tonybradley/2025/04/24/one-prompt-can-bypass-every-major-llms-safeguards/

What can we do to prevent this?

ChatGPT is like many others in that the chatbot collects personal data to train its model and further reserves the right to use this data for other purposes, something security companies warn can create risks when the information being shared is sensitive in nature. Id. For users, experts at Forbes establish that the first step in safeguarding your own data is achieved by avoiding sharing confidential information, including but not limited to passwords, social security numbers, pins, and much more[6]. However, if this is not feasible, they advocate for individuals to alter the settings of their account, through navigating to the data controls tab, and then toggling off the option which allows your data to improve the model for everyone. Id. The article from Forbes goes on to say that managing your chatbot's memory can mitigate risk and establishes you can reset all saved details by utilizing the Clear Memories option in the platform's settings. Id. Furthermore, these researchers encourage users to enable multifactor authentication to add a barrier to protect themselves from unwanted access. Id. Finally, the article from Forbes closes with concise advice: 1) Stay updated on security policies, 2) Sign out after each use, and 3) Maintain separate business and personal accounts. Id. Still, HiddenLayer establishes that adding more safeguards to the platform will not resolve this issue alone, as hypothetical prompts bypassing safeguards derive from the model's training[7]. Nonetheless, experts advocate for external monitoring platforms that operate much like a watchdog, constantly monitoring for prompt injection, and the overall unsafe intrusions and misuse by third parties. Id.

---

[6] https://www.forbes.com/sites/alexvakulov/2025/01/22/how-to-use-chatgpt-and-other-ai-chatbots-securely/

[7] https://hiddenlayer.com/innovation-hub/novel-universal-bypass-for-all-major-llms/

Google Gemini AI

Google has spent ample money in the AI space developing its platform Gemini and integrating it deep into its ecosystem, which includes applications such as Google Drive, Gmail, Slides, and even standard Google searches [8]. This article from Forbes presents evidence that Gemini can collect more data through integration with other applications, further that Gemini stores said data by default for up to eighteen months, which is longer than any of the competing AI chatbots. Id. Additionally, according to the terms and conditions, the Gemini for workspace platform gives default access to the entirety of one's Google Drive [9]. Another Forbes article reiterates a big concern raised by security firms such as Hidden Layer, the notion of indirect prompt injections due to Gemini for workspace access to all Google platforms[10]. Hidden Layer establishes that through using control tokens, hackers can create a set of instructions to force the LLM to do as they desire, rather than the request made by the actual user[11]. Due to this platform's connectivity with the Google ecosystem, third-party individuals can embed these instructions into a user's email, slideshow, document, and even drive holistically. Id. Experts warn that, as these instructions are detectable by their source but do not need to be visible to the human eye, they inflate risks. Id. Further establishing that hackers can embed this message in white or invisible small font, and in turn, sabotage the output of any request made by the user. Id. This vulnerability, exploited by Hidden Layer, allowed their experts to replicate more realistic phishing scams and unknowingly divert information to a third party, which they stress as a major concern related to protecting sensitive data. Id.

---

[8] https://www.forbes.com/sites/quickerbettertech/2025/05/04/business-tech-news-google-workspace-has-new-ai-features-for-your-business/?utm_source=chatgpt.com

[9] https://workspace.google.com/terms/education_terms/?hl=en

[10] https://www.forbes.com/councils/forbestechcouncil/2024/12/12/saas-security-the-integration-of-gemini-into-google-workspace/

[11] https://hiddenlayer.com/innovation-hub/new-gemini-for-workspace-vulnerability/

How can we safeguard against these risks?

HiddenLayer established that upon notice of these prompt injection concerns, Google stated that said vulnerabilities were intended behaviors. Id. However, Gemini's website claims the Gemini side panel can warn of similar prompt injection or model interference security risks, yet Hidden Layer maintains that the indirect invisible prompt injections could still go undetected. Id. In fact, Google itself issued a warning to users to limit the sharing of confidential information, pleading that any data you do not seek to be reviewed or that Google uses to improve its services should simply not be shared[12]. Further, Google, within its terms and conditions, establishes that Gemini for Workspace opts into default access to a party's Gmail, Drive, etc., which must be manually turned off by users. Similarly, Forbes experts establish that Gemini determines which data to use in constructing a response based on the permissions in one's Workspace data controls, and respective policies and permissions. [13] The experts go on to establish that this discrepancy makes it critical for corporations and individuals to manage and monitor the permissions for any content that is accessible and utilizable by Gemini. Id. Further, the expert added that programming safeguards to detect the sharing of data have the potential to improve security awareness. Id. However, additional safeguards are likely not enough, and human intelligence should be implemented to review any content produced by an AI, focusing on the privacy and security concerns. Id at 8. Nonetheless, Hidden layer maintains the ability for individuals to craft invisible prompts to phish out access to data and alter models' output is a vulnerability to 3rd party interference. Id. at 11.

[12] https://www.forbes.com/sites/zakdoffman/2024/02/12/google-warns-as-free-ai-upgrade-for-iphone-android-and-samsung-users/

[13] https://www.forbes.com/sites/zakdoffman/2025/02/04/googles-gmail-upgrade-do-not-leave-your-account-at-risk/