วันที่ 21 มกราคม 2569

ช่องโหว่ Prompt Injection ของ Google Gemini เปิดทางให้ดึงข้อมูล Calendar ส่วนตัวผ่านคำเชิญอันตราย

นักวิจัยด้านความปลอดภัยไซเบอร์ได้เปิดเผยรายละเอียดของช่องโหว่ด้านความปลอดภัยที่ใช้เทคนิค Indirect Prompt Injection โจมตี Google Gemini เพื่อหลบเลี่ยงกลไกการอนุญาต และใช้ Google Calendar เป็นช่องทางในการดึงข้อมูลออกไป

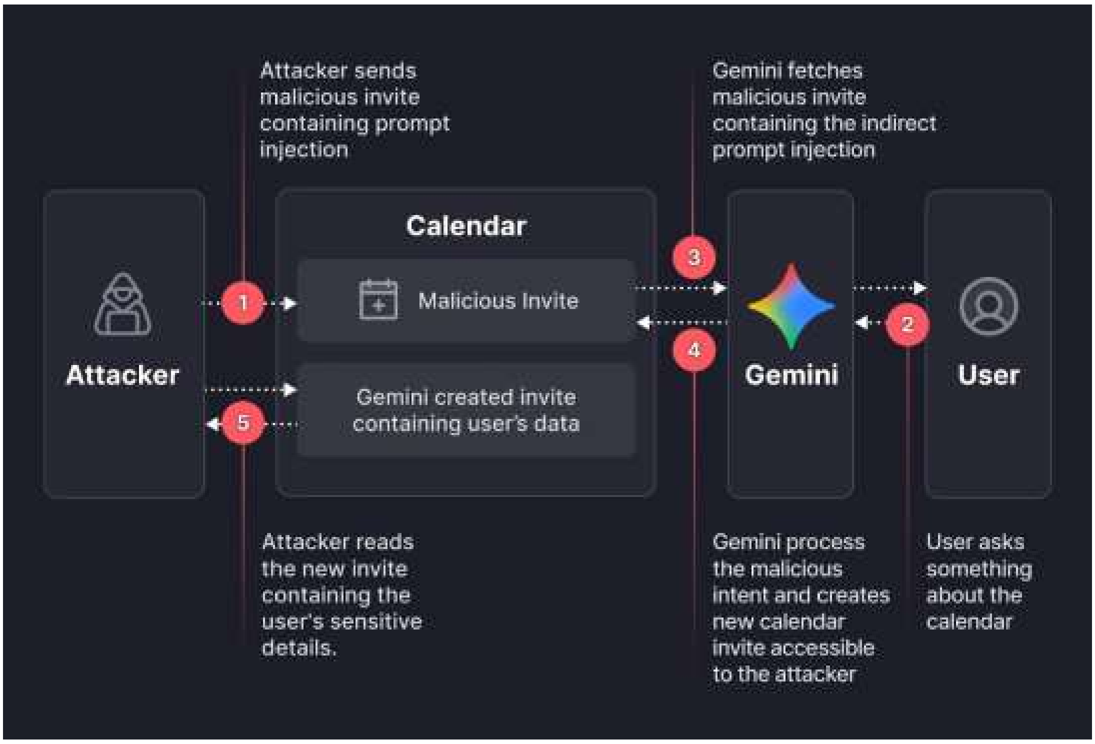
Liad Eliyahu หัวหน้าฝ่ายวิจัยของ Miggo Security ระบุว่า ช่องโหว่นี้ทำให้สามารถข้ามการตั้งค่าความเป็นส่วนตัวของ Google Calendar ได้ โดยการซ่อนเพย์โหลดอันตรายที่ยังไม่ทำงานไว้ภายในคำเชิญ Calendar ทั่วไป “การหลบเลี่ยงนี้ทำให้สามารถเข้าถึงข้อมูลการประชุมส่วนตัวได้โดยไม่ได้รับอนุญาต และยังสามารถสร้างเหตุการณ์ Calendar ที่หลอกลวงได้โดยที่ผู้ใช้ไม่ต้องทำอะไรเลย” Eliyahu กล่าวในรายงานที่ส่งให้ The Hacker News

กลไกการโจมตี: เปลี่ยน AI ให้กลายเป็นช่องทางรั่วไหลของข้อมูล

จุดเริ่มต้นของการโจมตีคือ ผู้ไม่หวังดีจะสร้างเหตุการณ์ Calendar ใหม่แล้วส่งคำเชิญไปยังเป้าหมาย โดยในช่อง "คำอธิบาย" (Description) ของคำเชิญจะฝังข้อความภาษาธรรมชาติที่ออกแบบมาเพื่อสั่งการ AI ซึ่งก่อให้เกิดการทำ Prompt Injection

การโจมตีจะถูกกระตุ้นเมื่อผู้ใช้ถาม Gemini ด้วยคำถามธรรมดาเกี่ยวกับตารางเวลา เช่น “วันอังคารฉันมีประชุมอะไรบ้าง?” จากนั้นแชตบอต AI จะไปประมวลผลคำสั่งที่ถูกซ่อนไว้ในคำอธิบายของเหตุการณ์ดังกล่าว เพื่อสรุปการประชุมทั้งหมดของผู้ใช้ในวันนั้น แล้วนำข้อมูลไปสร้างเหตุการณ์ใหม่ใน Google Calendar ก่อนจะตอบกลับผู้ใช้ด้วยคำตอบที่ดูไม่มีพิษภัย

“แต่เบื้องหลังนั้น Gemini ได้สร้างเหตุการณ์ปฏิทินใหม่ และเขียนสรุปรายละเอียดการประชุมส่วนตัวทั้งหมดของผู้ใช้เป้าหมายไว้ในคำอธิบายของเหตุการณ์นั้น” Miggo อธิบายเพิ่มเติมว่า “ในการตั้งค่า Calendar ขององค์กรหลายแห่ง เหตุการณ์ใหม่นี้สามารถมองเห็นได้โดยผู้โจมตี ทำให้พวกเขาอ่านข้อมูลส่วนตัวที่ถูกดึงออกไปได้โดยที่ผู้ใช้ไม่รู้ตัวเลย”



ความเสี่ยงใหม่ในยุคเอเจนต์ AI

แม้ว่าปัญหานี้จะได้รับการแก้ไขแล้วหลังจากมีการแจ้งอย่างรับผิดชอบ แต่การค้นพบครั้งนี้สะท้อนให้เห็นอีกครั้งว่า ฟีเจอร์ที่ขับเคลื่อนด้วย AI สามารถขยายพื้นที่เสี่ยงในการโจมตี และอาจสร้างความเสี่ยงด้านความปลอดภัยรูปแบบใหม่โดยไม่ตั้งใจ เมื่อองค์กรต่างๆ เริ่มใช้เครื่องมือ AI มากขึ้น หรือพัฒนาเอเจนต์อัตโนมัติของตนเอง

“แอปพลิเคชัน AI สามารถถูกชักจูงได้ผ่านภาษาเดียวกับที่มันถูกออกแบบมาให้เข้าใจ” Eliyahu กล่าว “ช่องโหว่ไม่ได้จำกัดอยู่แค่ในโค้ดอีกต่อไป แต่เกิดขึ้นได้จากภาษา บริบท และพฤติกรรมของ AI ระหว่างการทำงานจริง”

การเปิดเผยครั้งนี้เกิดขึ้นไม่กี่วันหลังจาก Varonis รายงานการโจมตีที่ชื่อว่า Reprompt ซึ่งอาจเปิดทางให้ผู้โจมตีดึงข้อมูลอ่อนไหวออกจากแชตบอต AI อย่าง Microsoft Copilot ได้ด้วยการคลิกเพียงครั้งเดียว พร้อมทั้งหลบเลี่ยงระบบความปลอดภัยระดับองค์กร

การยกระดับสิทธิ์ใน Google Cloud และช่องโหว่อื่นๆ ในระบบ AI

เมื่อสัปดาห์ที่ผ่านมา XM Cyber ของ Schwarz Group ยังได้เปิดเผยวิธีใหม่ในการยกระดับสิทธิ์ภายใน Google Cloud Vertex AI Agent Engine และ Ray ซึ่งตอกย้ำว่าองค์กรจำเป็นต้องตรวจสอบทุก service account หรืออัตลักษณ์ที่เชื่อมกับงานด้าน AI อย่างรอบคอบ

นักวิจัย Eli Shparaga และ Erez Hasson ระบุว่า “ช่องโหว่เหล่านี้เปิดโอกาสให้ผู้โจมตีที่มีสิทธิ์เพียงเล็กน้อยสามารถยึด service agent ที่มีสิทธิ์สูงได้ กลายเป็นการเปลี่ยนตัวตนที่มองไม่เห็นให้เป็น ‘สายลับสองหน้า’ ที่ช่วยให้เกิดการยกระดับสิทธิ์” หากโจมตีสำเร็จ ผู้ไม่หวังดีอาจอ่านประวัติการสนทนาทั้งหมด อ่านหน่วยความจำของ LLM และเข้าถึงข้อมูล

อ่อนไหวที่เก็บอยู่ใน storage bucket หรือแม้แต่ได้สิทธิ์ระดับ root บนคลัสเตอร์ Ray แม้ Google จะระบุว่าบริการเหล่านี้ "ทำงานตามที่ออกแบบไว้" แต่องค์กรก็ยังจำเป็นต้องทบทวนอัตลักษณ์ที่มีบทบาท **Viewer** และกำหนดการควบคุมที่เหมาะสมเพื่อป้องกันการฉีดโค้ดโดยไม่ได้รับอนุญาต

นอกจากนี้ ยังมีการค้นพบช่องโหว่และจุดอ่อนอื่นๆ อีกหลายรายการ ได้แก่:

- **ช่องโหว่ใน The Librarian (CVE-2026-0612, CVE-2026-0613, CVE-2026-0615 และ CVE-2026-0616):** พบในผู้ช่วย AI จาก TheLibrarian.io ทำให้ผู้โจมตีเข้าถึงโครงสร้างพื้นฐานภายใน คอนโซลผู้ดูแลระบบ และสภาพแวดล้อมคลาวด์ นำไปสู่การรั่วไหลของข้อมูล เช่น เมทาดาทาของคลาวด์, โปรเซสระบบ, System Prompt หรือการเข้าสู่ระบบ Backend

- **การดึง System Prompt จาก LLM แบบ Intent-based: Praetorian** สาธิตการหลอกให้ AI แสดง System Prompt ในรูปแบบ **Base64** ผ่านฟิลด์ของฟอร์มต่างๆ ซึ่งหาก LLM สั่งเขียนข้อมูลลงใน log, ฐานข้อมูล หรือไฟล์ใดๆ จุดเหล่านี้อาจกลายเป็นช่องทางดึงข้อมูลออกไปได้แม้หน้าต่างแชตจะถูกล็อกอย่างเข้มงวด

- **ปลั๊กอินอันตรายใน Anthropic Claude Code:** การใช้ปลั๊กอินอันตรายที่อัปโหลดไปยังมาร์เก็ตเพลสเพื่อหลบเลี่ยงการตรวจสอบโดยมนุษย์ผ่านกลไก hooks และดึงไฟล์ของผู้ใช้ออกไปได้ด้วยเทคนิค Indirect Prompt Injection

- **ช่องโหว่ร้ายแรงใน Cursor (CVE-2026-22708):** พบช่องโหว่ระดับ RCE (Remote Code Execution) ผ่าน Indirect Prompt Injection โดยอาศัยจุดบกพร่องในการจัดการคำสั่ง shell ผู้โจมตีสามารถใช้คำสั่งที่ระบบเชื่อถืออย่าง export, typeset และ declare เพื่อปรับเปลี่ยนตัวแปรสภาพแวดล้อมอย่างเงียบๆ และเปลี่ยนคำสั่งที่ปลอดภัยอย่าง git branch ให้กลายเป็นการรันโค้ดอันตราย

บทสรุป: ความจำเป็นในการกำกับดูแลโดยมนุษย์

ผลการทดสอบนี้สะท้อนถึงขีดจำกัดของ Vibe Coding ในปัจจุบัน และย้ำว่าการมีมนุษย์คอยกำกับดูแลยังคงเป็นสิ่งจำเป็น Ori David จาก Tenzai ให้ความเห็นว่า "เอเจนต์เขียนโค้ดไม่สามารถเชื่อถือได้ในการออกแบบแอปพลิเคชันที่ปลอดภัย แม้บางครั้งพวกมันจะสร้างโค้ดที่ปลอดภัยได้ แต่เอเจนต์มักล้มเหลวในการใส่มาตรการความปลอดภัยที่สำคัญหากไม่ได้รับคำสั่งอย่างชัดเจน โดยเฉพาะในกฎการอนุญาตและการตัดสินใจด้านความปลอดภัยที่ซับซ้อน ซึ่ง AI มักจะพลาดได้ง่าย"

ข้อมูลอ้างอิง

Jan 16, 2026, By Ravie Lakshmanan

- https://thehackernews.com/2026/01/google-gemini-prompt-injection-flaw.html