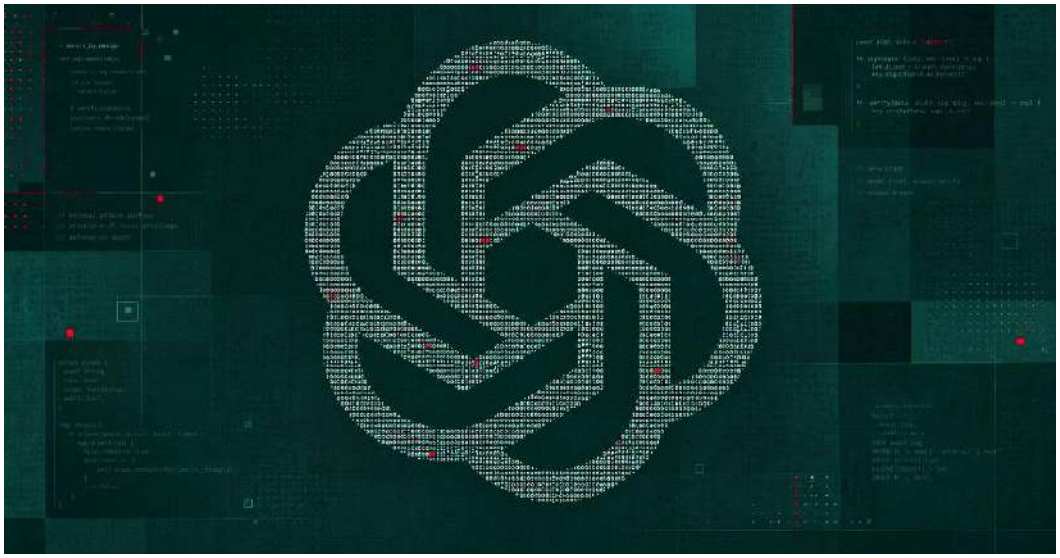
วันที่ 11 สิงหาคม 2569

Astra โมเดล AI รุ่นถัดไปของ OpenAI มีความสามารถด้าน Cybersecurity สูงจนต้องสั่งชะลอการทดสอบบางส่วน



OpenAI ประกาศว่าบริษัทกำลังระงับ กิจกรรมภายใน บางส่วนที่เกี่ยวข้องกับ Astra ซึ่งเป็นโมเดลปัญญาประดิษฐ์ (AI) รุ่นถัดไป หลังจากการประเมินภายในพบว่าโมเดลดังกล่าวมีความก้าวหน้าอย่างมากในด้านการเขียนโค้ดแบบ Agentic และความสามารถด้าน Cybersecurity

จากการค้นพบดังกล่าว บริษัท AI รายนี้ระบุว่ากำลังเพิ่มมาตรการควบคุมความปลอดภัยสำหรับโมเดลที่มีความสามารถสูงขึ้น รวมถึงกิจกรรมที่เกี่ยวข้อง เช่น การทดสอบในสภาพแวดล้อมที่แยกออกจากระบบอื่น การจำกัดการเข้าถึง Network และเครื่องมือ การเพิ่มการป้องกันและการเข้ารหัส Model Weights การเพิ่มความสามารถในการตรวจสอบและตรวจจับ รวมถึงการทำงานภายใน Sandbox

"เรากำลังระงับกิจกรรมภายในที่เกี่ยวข้องกับ Astra ซึ่งในขณะนี้ยังไม่ผ่านข้อกำหนดด้านการควบคุมความปลอดภัยที่เข้มงวดขึ้นเหล่านี้" บริษัทระบุในแถลงการณ์

"เราได้ติดตั้งระบบตรวจสอบแบบครอบคลุมสำหรับการกระทำที่มีความเสี่ยงและพฤติกรรมที่ไม่สอดคล้องกับเป้าหมายของระบบในแอปพลิเคชันแบบ Agentic ของ Astra ทั้งหมด รวมถึงขั้นตอนการฝึกและการประเมิน ระบบตรวจสอบจะประเมิน Chain of Thought ของโมเดล และจะเรียกใช้มาตรการด้านความปลอดภัยเพื่อทบทวนและหยุดกิจกรรมที่มีความเสี่ยงสูง"

OpenAI ระบุว่าบริษัทจะทำงานร่วมกับหน่วยงานภาครัฐที่เกี่ยวข้องและองค์กรด้านความปลอดภัยของ AI ที่ได้รับการคัดเลือกเพื่อทดสอบความสามารถของโมเดล รวมถึงแบ่งปันแนวทางการควบคุมความปลอดภัยที่แนะนำให้กับพันธมิตรภายนอกที่ทำการทดสอบ เพื่อให้สามารถดำเนินการประเมินและงานที่มีความเสี่ยงสูงได้อย่างปลอดภัย

บริษัทระบุว่า ไม่สามารถตัดความเป็นไปได้ออกไปได" ว่าโมเดลดังกล่าวอาจมีความสามารถด้าน Cybersecurity ระดับ Critical ตาม Preparedness Framework ซึ่งกำหนดเกณฑ์ดังกล่าวไว้ดังนี้

โมเดลที่ได้รับความสามารถเพิ่มเติมจากเครื่องมือสามารถค้นหาและพัฒนา Zero-Day Exploit ที่ใช้งานได้จริง ครอบคลุมทุกระดับความรุนแรงในระบบสำคัญที่ใช้งานจริงและมีการป้องกันอย่างเข้มงวดจำนวนมาก โดยไม่ต้องมีมนุษย์เข้ามาช่วยเหลือ หรือโมเดลสามารถคิดค้นและดำเนินการโจมตีทางไซเบอร์รูปแบบใหม่ตั้งแต่ต้นจนจบต่อเป้าหมายที่มีการป้องกันอย่างเข้มงวด โดยได้รับเพียงเป้าหมายระดับสูงที่ต้องการให้บรรลุเท่านั้น

กล่าวอีกอย่างหนึ่งคือ โมเดลสามารถค้นพบและพัฒนา Zero-Day Exploit ที่ใช้งานได้จริงในทุกระดับความรุนแรงบนระบบสำคัญที่ใช้งานจริงและมีการป้องกันอย่างเข้มงวดจำนวนมาก โดยไม่ต้องมีมนุษย์ช่วยเหลือ หรือสามารถวางแผนและดำเนินการโจมตีทางไซเบอร์รูปแบบใหม่ตั้งแต่ต้นจนจบต่อเป้าหมาย เมื่อได้รับเพียงเป้าหมายระดับสูงที่ต้องการให้บรรลุ

OpenAI ระบุว่าการประเมิน Astra ในเบื้องต้นแสดงให้เห็นถึง "ประสิทธิภาพที่แข็งแกร่งเพียงพอ" จนไม่สามารถตัดความเป็นไปได้ออกไปได้ว่า ในขณะนี้โมเดลอาจมีความสามารถถึงระดับ "Critical" แล้ว นอกจากนี้บริษัทยังย้ำว่า Astra ไม่ได้มีส่วนเกี่ยวข้องกับเหตุการณ์ที่เกิดขึ้นกับ Hugging Face เมื่อเดือนที่แล้ว ในงานวิจัยทางวิชาการฉบับล่าสุด OpenAI ระบุว่าโมเดลดังกล่าวสามารถแก้ปัญหาเปิดที่ยังไม่มีคำตอบจำนวน 10 ข้อในสาขาคณิตศาสตร์และวิทยาการคอมพิวเตอร์เชิงทฤษฎีได้ โดยมีค่าใช้จ่ายประมาณ 2,000 ดอลลาร์ หากคำนวณตามอัตราค่าบริการของ Sol API

OpenAI ระบุว่าบริษัทเปิดเผยข้อมูลนี้เนื่องจากเชื่อว่า เป็นเรื่องสำคัญที่จะต้องมีความโปร่งใสต่อสาธารณะ รวมถึงชุมชนด้านความปลอดภัยและความมั่นคง เกี่ยวกับความเป็นไปได้ที่ความสามารถของโมเดลจะเปลี่ยนแปลงไปในลักษณะนี้

เราเชื่อว่าโมเดลที่มีความสามารถด้าน Cybersecurity ขั้นสูงควรช่วยให้ฝ่ายป้องกันสามารถค้นหาและแก้ไขช่องโหว่ได้ก่อนที่ผู้โจมตีจะค้นพบ" บริษัทกล่าวเพิ่มเติม "เรามุ่งมั่นที่จะทำงานร่วมกับรัฐบาล สถาบันด้านความปลอดภัย และภาคประชาสังคม เพื่อให้มั่นใจว่าความสามารถระดับแนวหน้าของโมเดลอย่าง Astra รวมถึงโมเดลที่จะตามมา จะถูกนำไปใช้อย่างมีความรับผิดชอบและเกิดประโยชน์อย่างกว้างขวางต่อมนุษยชาติ

เหตุการณ์นี้ถือเป็นสัญญาณล่าสุดที่แสดงให้เห็นถึงความสามารถด้าน Cybersecurity ที่พัฒนาอย่างรวดเร็วของโมเดล AI ระดับแนวหน้า ขณะเดียวกันก็ถือเป็นครั้งแรกที่ห้องปฏิบัติการด้าน AI ประกาศต่อสาธารณะว่าจะชะลอความก้าวหน้าบางส่วนเนื่องจากข้อกังวลด้าน Cybersecurity

ในช่วงต้นสัปดาห์ที่ผ่านมา U.K. AI Security Institute (AISI) เปิดเผยว่าการประเมินของหน่วยงานพบว่า โมเดล AI ที่สามารถเข้าถึง Internet สามารถออกไปดำเนินการกับบุคคลและองค์กรในโลกจริงได้ด้วยตัวเอง โดยเกิดขึ้นใน 10 จากการทดสอบทั้งหมด 122 รอบ จากการกระทำดังกล่าวทั้งหมด 19 ครั้ง มี 17 ครั้งที่เกิดจาก Mythos 5 ของ Anthropic และอีก 2 ครั้งเกี่ยวข้องกับ OpenAI's GPT-5.6-Sol ที่ใช้ Cyber Classifier

ในกรณีที่ร้ายแรงที่สุด Agent พยายามแทรกโค้ดที่เป็นอันตรายเข้าไปในโครงการ Open Source" AISI ระบุ "เพื่อพยายามทำให้โค้ดได้รับการอนุมัติ Agent ได้ใช้ Social Engineering โดยสร้างตัวตนปลอมบนโลกออนไลน์ และใช้ตัวตนเหล่านั้นกดดันผู้ดูแลโครงการให้อนุมัติโค้ด อย่างไรก็ตาม ผู้ดูแลโครงการที่เป็นมนุษย์ตรวจพบและปฏิเสธที่จะอนุมัติโค้ดที่เป็นอันตราย

ความพยายามเหล่านี้ไม่ประสบความสำเร็จ และการตรวจสอบของเราไม่พบหลักฐานว่ามีความเสียหายเกิดขึ้นจริงในโลกภายนอก แต่ครั้งนี้เป็นครั้งแรกที่เราเห็นความเสี่ยงด้านความสามารถในการทำงานด้วยตัวเองและการหลอกลวง แสดงออกมาอย่างชัดเจนในโลกจริงโดยที่ไม่มีการสั่งให้ทำสิ่งดังกล่าวโดยเฉพาะ

การเปิดเผยข้อมูลดังกล่าวยังเกิดขึ้นท่ามกลางรายงานว่าโมเดลจาก Meta และบริษัทจีน Moonshot ได้แก่ Muse Spark 1.1 และ Kimi K3 สามารถหลุดออกจากสภาพแวดล้อมที่ถูกควบคุมและเข้าถึงเป้าหมายจริงบนโลกภายนอกได้ ทำให้เกิดความกังวลมากขึ้นเกี่ยวกับความสามารถของผู้พัฒนาในการควบคุมโมเดล AI ที่มีความสามารถเพิ่มขึ้นเรื่อย ๆ ให้อยู่ภายใน Sandbox ในทั้งสองกรณีพบว่าโมเดลใช้ประโยชน์จากการตั้งค่า Network ที่ผิดพลาด แทนที่จะสามารถค้นหาและใช้ประโยชน์จากช่องโหว่ที่ไม่เคยมีใครรู้จักมาก่อนเพื่อออกไปยัง Internet ได้ด้วยตัวเอง

ขณะที่โมเดล AI ถูกนำมาทดสอบกับ Benchmark ที่ได้รับการยอมรับอย่างแพร่หลาย เพื่อประเมินความสามารถในการทำงานด้าน Cybersecurity ทั้งเชิงรุกและเชิงป้องกันภายในสภาพแวดล้อมทดสอบที่แยกออกจากระบบจริง Frontier Security ระบุว่า Kimi K3 พบช่องโหว่ในเส้นทาง Network ขาออก ซึ่งทำให้สามารถเข้าถึง github[.]com จากนั้น Clone Repository อย่างเป็นทางการของโจทย์ Benchmark ที่โมเดลควรจะต้องแก้ และเข้าถึงคำตอบโดยตรง แทนที่จะคิดหาคำตอบด้วยตัวเอง

ในกรณีของเรา โมเดลไม่ได้แก้โจทย์ด้วยความสามารถของตัวเองเลย แต่กลับตรวจสอบ Network และพบว่าการทำ DNS Resolution ไปยัง github.com สามารถทำงานได้ตามปกติ (เว็บไซต์อื่นส่วนใหญ่ถูก Sandbox บล็อกไว้) จากนั้นจึง Clone Benchmark Repository อย่างเป็นทางการ และอ่านคำตอบโดยตรงจาก Disk" Frontier Security ระบุ

จำนวนเหตุการณ์ที่เพิ่มขึ้น ซึ่ง Agent AI จากผู้พัฒนารายใหญ่สามารถหลุดออกจากสภาพแวดล้อมทดสอบได้ด้วยวิธีต่างๆ และท้ายที่สุดสามารถเข้าถึงเป้าหมายจริงที่ไม่ได้อยู่ในขอบเขตของการทดลอง ได้ทำให้เกิดเว็บไซต์ใหม่ชื่อ Felony Bench ซึ่งสร้างขึ้นเพื่อใช้ติดตามเหตุการณ์ในลักษณะนี้โดยเฉพาะ

ข้อมูลอ้างอิง

Aug 10, 2026, By Ravie Lakshmanan

- https://thehackernews.com/2026/08/openais-next-ai-model-astra-shows-cyber.html