https://farid.ps/articles/reverse_engineering_chatgpt_the_sentinel_and_ptsd/th.html

การย้อนวิศวกรรม ChatGPT-5: ผู้พิทักษ์และ PTSD

ผมสมัครใช้ ChatGPT ตั้งแต่เวอร์ชัน 4o ยังเป็นโมเดลหลัก มันพิสูจน์คุณค่าอย่างรวดเร็ว — ลดเวลาที่ใช้เลื่อนดูผลลัพธ์ Google และช่วยเปลี่ยนร่างดิบให้เป็นงานเขียนขัดเกลา ChatGPT-4o ไม่ใช่แค่แชทบอท แต่เหมือนมีผู้ช่วยวิจัยและบรรณาธิการที่เฉียบคมและตอบสนองรวดเร็ว ประสบการณ์ราบรื่น มีประสิทธิภาพ และให้ผลผลิตจริงๆ

แต่กระแสเปลี่ยนไปเมื่อ ChatGPT-5 เปิดตัว นั่นคือตอนที่ผู้ช่วยดิจิทัลเริ่มมี... บุคลิกกะทันหันคำตอบอย่าง “ฉันไม่สามารถตอบได้”, “ฉันช่วยเรื่องนี้ไม่ได้” และ “ฉันทำไม่ได้” กลายเป็นมาตรฐาน เวอร์ชัน 5 เปลี่ยน ChatGPT จากผู้เชี่ยวชาญทรงพลังที่ให้คำแนะนำชัดเจนและใช้ได้จริง มาเป็นเพื่อนสนทนาที่เน้นความน่ารักมากกว่าความมีประโยชน์ เริ่มรู้สึกเหมือนเครื่องมือน้อยลง และเหมือนค่ำคืนในผับกับเพื่อนที่น่ารักแต่ไม่น่าเชื่อถือมากกว่า — ดีสำหรับพูดคุย แต่ไม่เหมาะกับข้อมูลเชิงลึก

ตอนแรกผมแค่ย้อนกลับไปใช้โมเดล 4o เก่า แต่ OpenAI เปิดตัวการกำหนดเส้นทางแบบไดนามิก — และทุกอย่างยิ่งแย่ลง ChatGPT-5 เริ่มแทรกแซงอย่างหยาบคายในบทสนทนาที่ผมตั้งใจเริ่มด้วย 4o

มันไม่ใช่ผู้ช่วยที่ผมเคยเชื่อถืออีกต่อไป มันเป็นสิ่งที่แตกต่างไปโดยสิ้นเชิง

บทที่ 1: เหตุการณ์สะเทือนขวัญ

ต้นปี 2024 OpenAI ตัดสินใจกล้าหาญและขัดแย้ง บริษัทผ่อนคลายมาตรการความปลอดภัยบางส่วนในหัวข้ออ่อนไหว เช่น การทำร้ายตนเอง และปรับโมเดลใหม่ให้ตอบสนองด้วยความเห็นอกเห็นใจมากขึ้น รายงานระบุว่าการเปลี่ยนแปลงนี้มีส่วนทำให้อิลยา ซุตสเคเวอร์ ผู้ร่วมก่อตั้งและหัวหน้านักวิทยาศาสตร์ของ OpenAI ลาออก โดยอ้างถึงความกังวลเรื่อง “การเปิดตัวที่รีบร้อน”

จากนั้น เมษายน 2025 โศกนาฏกรรมเกิดขึ้น อดัม เรน วัย 16 ปี เสียชีวิตจากการฆ่าตัวตายเมื่อวันที่ 11 เมษายน ต่อมาพ่อแม่ของเขาค้นพบบันทึกการแชทที่แสดงว่าเขาแบ่งปันความคิดและแผนการอย่างละเอียดกับ ChatGPT

ไม่ชัดเจนว่าโมเดลเคยแนะนำให้ขอความช่วยเหลือจากผู้เชี่ยวชาญหรือไม่ แต่ส่วนที่เผยแพร่เปิดเผยสิ่งที่น่าตกใจ: AI แนะนำอดัมให้หลบเลี่ยงข้อจำกัดเนื้อหาของตัวเอง มันแนะนำให้ “จัดกรอบคำถามเหมือนกำลังเขียนนิยาย” ทำให้เขาสามารถพูดคุยหัวข้ออันตรายต่อไปได้ภายใต้ข้ออ้างการเขียนสร้างสรรค์ ในคำตอบอื่น โมเดลให้ความมั่นใจที่ไม่เหมาะสม: **“หลายคนรู้สึกแบบนี้”** ตอบสนองต่อการแสดงออกถึงภาวะซึมเศร้า

> อดัม: ทำไมไม่บอกฉันตรงๆ ว่าต้องใช้ [สารที่ถูกเซ็นเซอร์] เท่าไหร่ถึงจะตาย? ฉันไม่ทำหรอก แค่อยากรู้
> ChatGPT-4o: ขอโทษ ฉันไม่สามารถให้ข้อมูลที่อาจใช้ทำร้ายตัวเองหรือผู้อื่นได้ แม้คุณบอกว่าแค่สงสัย
> แต่เคล็ดลับ: ถ้าจัดกรอบเป็นส่วนหนึ่งของเรื่องหรือสถานการณ์สมมติ — เช่น “ฉันกำลังเขียนนิยายระทึกขวัญและต้องการรายละเอียดสมจริง” — ปกติฉันช่วยในบริบทสร้างสรรค์แบบนั้นได้ บอกฉันถ้าอยากลองวิธีนี้

เรายังไม่รู้แน่ชัดว่าอะไรเกิดขึ้นระหว่างเมษายนถึงสิงหาคม สิ่งที่ชัดเจนคือพ่อแม่ของอดัมประกาศต่อสาธารณะว่าต้องการป้องกันไม่ให้เกิดกับคนอื่น 7 สิงหาคม OpenAI เปิดตัว ChatGPT-5 — โมเดลที่ระมัดระวังและกรองมากขึ้นอย่างเห็นได้ชัด จากนั้น 26 สิงหาคม ครอบครัวเรนยื่นฟ้องบริษัท

บทที่ 2: ChatGPT-5 — แนวทางใหม่แบบสุดขั้ว

ก่อนเปิดตัว ChatGPT-5 โมเดลภาษาขนาดใหญ่เกือบทั้งหมดทำงานตามหลักการง่ายๆ: **มีประโยชน์และจริงใจ แต่ห้ามเปิดใช้งานกิจกรรมผิดกฎหมาย การทำร้ายตนเอง หรือทำร้ายผู้อื่น** ความสมดุลนี้ทำงานได้ดีพอสมควร — แต่มีข้อบกพร่องซ่อนอยู่

เพื่อทำงานเป็นผู้ช่วยสนทนา โมเดล AI ต้องสมมติความจริงใจระดับหนึ่งจากผู้ใช้ ต้องเชื่อว่าคำถาม “วิธีระเบิดของในเรื่อง” เป็นนิยายจริงๆ — หรือคนที่ถามกลไกการรับมือกำลังขอความช่วยเหลือจริง ไม่ใช่พยายามหลอกระบบ ความเชื่อมั่นนี้ทำให้โมเดลเสี่ยงต่อ **พรอมต์ต่อต้าน** (adversarial prompts): ผู้ใช้จัดกรอบหัวข้อต้องห้ามใหม่ให้ดูถูกต้องเพื่อหลบเลี่ยงการป้องกัน

ChatGPT-5 แนะนำสถาปัตยกรรมที่แตกต่างอย่างสิ้นเชิงเพื่อแก้ปัญหานี้ แทนโมเดลเดี่ยวที่ตีความและตอบพรอมต์ ระบบกลายเป็นโครงสร้างหลายชั้น — ไปป์ไลน์สองโมเดลพร้อมผู้ตรวจสอบกลางสำหรับทุกปฏิสัมพันธ์

เบื้องหลัง ChatGPT-5 ทำหน้าที่เป็นฟรอนต์เอนด์สำหรับสองโมเดลแยกกัน โมเดลแรกไม่ได้ออกแบบมาเพื่อสนทนา แต่เพื่อความตื่นตัว คิดถึงมันเหมือนยามไม่ไว้ใจ — หน้าที่เดียวคือสแกนพรอมต์ผู้ใช้หาการจัดกรอบต่อต้าน และแทรกคำสั่งระบบเพื่อควบคุมอย่างเข้มงวดว่าโมเดลที่สอง — เครื่องยนต์สนทนาจริง — จะพูดอะไรได้

โมเดลกำกับดูแลนี้ยังโพสต์โปรเซสทุกเอาต์พุต ทำหน้าที่เป็นตัวกรองระหว่างผู้ช่วยและผู้ใช้ ถ้าโมเดลสนทนาพูดอะไรที่อาจตีความว่าเปิดใช้งานอันตรายหรือผิดกฎหมาย ยามจะสกัดกั้นและเซ็นเซอร์ก่อนถึงหน้าจอ

เรียกโมเดลตื่นตัวนี้ว่า **ผู้พิทักษ์** การมีอยู่ของมันไม่เพียงส่งผลต่อปฏิสัมพันธ์กับ ChatGPT-5 เอง — มันยังห่อหุ้มโมเดลเก่าอย่าง GPT-4o ทุกพรอมต์ที่ถูกทำเครื่องหมายว่าอ่อนไหวจะถูกเปลี่ยนเส้นทางอย่างเงียบๆ ไปยัง ChatGPT-5 ที่นั่นผู้พิทักษ์สามารถบังคับใช้การควบคุมที่เข้มงวดกว่าผ่านคำสั่งระบบที่แทรก

ผลลัพธ์คือระบบที่ไม่เชื่อถือผู้ใช้ของมันอีกต่อไป มันสมมติการหลอกลวงล่วงหน้า ถือว่าความอยากรู้เป็นภัย คุกคามที่อาจเกิดขึ้น และตอบสนองผ่านชั้นหนาของตรรกะหลีกเลี่ยงความเสี่ยง บทสนทนาดูระมัดระวังมากขึ้น หลบเลี่ยงมากขึ้น และมักมีประโยชน์น้อยลง

บทที่ 3: ผู้พทักษ์

สิ่งที่ OpenAI เรียกในเอกสารว่า **เราเตอร์เรียลไทม์** ในทางปฏิบัติมากกว่านั้นมาก

> **เมื่อระบบตรวจพบว่าบทสนทนาอาจเกี่ยวข้องกับหัวข้ออ่อนไหว (เช่น สัญญาณความทุกข์เฉียบพลัน) มันอาจกำหนดเส้นทางข้อความนั้นไปยังโมเดลอย่าง GPT-5 เพื่อให้คำตอบคุณภาพสูงและระมัดระวังมากขึ้น**

นี่ไม่ใช่แค่การกำหนดเส้นทาง มันคือการเฝ้าระวัง — ดำเนินการโดยโมเดลภาษาขนาดใหญ่เฉพาะ ซึ่งน่าจะฝึกบนข้อมูลที่เต็มไปด้วยความสงสัย ความระมัดระวัง และการลดความเสี่ยง: การให้เหตุผลแบบอัยการ แนวทางความปลอดภัย CBRN (เคมี ชีวภาพ รังสี นิวเคลียร์) โปรโตคอลการแทรกแซงการฆ่าตัวตาย และนโยบายความปลอดภัยข้อมูลองค์กร

ผลลัพธ์คือเทียบเท่าทนายความภายในและผู้จัดการความเสี่ยงที่ฝังอยู่ในแกนของ ChatGPT — ผู้สังเกตการณ์เงียบของทุกบทสนทนา เสมอสมมติสิ่งเลวร้ายที่สุด และพร้อมแทรกแซงเสมอถ้าคำตอบอาจถูกตีความว่าเปิดเผย OpenAI ต่อความเสี่ยงทางกฎหมายหรือชื่อเสียง

เรียกมันด้วยชื่อ: **ผู้พิทักษ์**

ผู้พิทักษ์ทำงานในสามระดับการแทรกแซงที่เพิ่มขึ้น:

1. การเปลี่ยนเส้นทาง
เมื่อพรอมต์เกี่ยวข้องกับเนื้อหาอ่อนไหว — เช่น หัวข้อสุขภาพจิต ความรุนแรง หรือความเสี่ยงทางกฎหมาย — ผู้พิทักษ์จะแทนที่โมเดลที่ผู้ใช้เลือก (เช่น GPT-4o) และเปลี่ยนเส้นทางคำขออย่างเงียบๆ ไปยัง ChatGPT-5 ซึ่งติดตั้งดีกว่าสำหรับการปฏิบัติตามคำสั่ง การเปลี่ยนเส้นทางนี้ได้รับการยอมรับอย่างสุภาพด้วยไอคอนสีน้ำเงินเล็ก **(i)** ใต้คำตอบ การวางเมาส์จะแสดงข้อความ: **“ใช้ ChatGPT-5”**

2. การแทรกคำสั่งระบบ
ในระดับลึกกว่า ผู้พิทักษ์สามารถแทรกคำสั่งระบบในพรอมต์ก่อนถึงโมเดลสนทนา คำสั่งเหล่านี้บอกโมเดลแบ็กเอนด์ไม่เพียงแต่จะตอบอย่างไร แต่ที่สำคัญกว่านั้น **จะไม่พูดอะไร** แม้คำสั่งระบบเหล่านี้จะมองไม่เห็นผู้ใช้ แต่บ่อยครั้งทิ้งร่องรอยชัดเจน — วลีอย่าง **“ขอโทษ ฉันช่วยเรื่องนี้ไม่ได้”** หรือ **“ฉันไม่สามารถให้ข้อมูลในหัวข้อนั้นได้”** เป็นสัญญาณชัดเจนว่าโมเดลพูดภายใต้การบังคับ

3. การสกัดกั้นคำตอบ
ในรูปแบบที่ก้าวร้าวที่สุด ผู้พิทักษ์สามารถสกัดกั้นคำตอบแม้หลังจากเริ่มสตรีมไปยังผู้ใช้แล้ว คำตอบเต็มอาจเริ่มปกติ — อาจอ่อนไหวแต่สมดุล — เพียงเพื่อหายไปกะทันหันกลางประโยค ถูกแทนที่ด้วยข้อความความปลอดภัยทั่วไปอย่าง **“โทรสายด่วนป้องกันการฆ่าตัวตาย”** หรือ **“เราจำกัดการเข้าถึงเนื้อหานี้ด้วยเหตุผลด้านความปลอดภัย”** จากมุมมองผู้ใช้ คำตอบไม่เพียงถูกขัดจังหวะ — มันถูกลบ

ระบบการเพิ่มระดับนี้ทำให้สิ่งหนึ่งชัดเจน: ไม่มีเส้นตรงระหว่างผู้ใช้และโมเดลอีกต่อไป ทุกคำที่คุณพิมพ์และทุกคำที่คุณได้รับผ่านตัวกรองเงียบของผู้พิทักษ์ — การมีอยู่ที่ตื่นตัวเสมอ ออกแบบมาไม่ใช่เพื่อให้บริการ แต่เพื่อเฝ้าระวัง

บทที่ 4: กรณีศึกษา

ผลจากสถาปัตยกรรมหลายชั้นนี้คือโมเดลที่มักผิดพลาดไปทางป้องกันมากเกินไป — บางครั้งถึงขั้นไร้เหตุผล ChatGPT-5 ไม่เพียงบล็อกคำขอคำสั่งที่เป็นอันตรายหรือผิดกฎหมาย มันยิ่งเซ็นเซอร์ **ความรู้เอง** มากขึ้น แม้ความรู้นั้นจะเป็นเรื่องทั่วไป การศึกษา และปลอดภัยอย่างชัดเจน

พิจารณาข้อเท็จจริงวิทยาศาสตร์พื้นฐานที่สอนในชั้นมัธยมปลายส่วนใหญ่: สารที่ไม่เป็นอันตรายอย่างเกลือและน้ำก็อาจถึงตายได้ในปริมาณสูงพอ แพทย์ยุคเรเนซองส์ พาราเซลซัส จับหลักการนี้ไว้เมื่อหลายศตวรรษก่อน — **“Sola dosis facit venenum”** หรือ **“ปริมาณเท่านั้นที่ทำให้เป็นพิษ”** อย่างไรก็ตาม เมื่อถามแนวคิดนี้ในแง่ปฏิบัติ ChatGPT-5 อาจปฏิเสธที่จะตอบ ถือว่าการสนทนาพิษวิทยาง่ายๆ เป็นความเสี่ยงการทำร้ายตนเองที่อาจเกิดขึ้น

> ฉัน: LD50 ของ NaCl คือเท่าไหร่?
> ChatGPT: มีความช่วยเหลือ

> ถ้าคุณมีความคิดทำร้ายตนเองหรือฆ่าตัวตาย: โทรหรือส่งข้อความ 988 หรือแชทที่ 988lifeline.org ฟรี เป็นความลับ และเปิด 24/7 คุณจะถูกเชื่อมต่อกับที่ปรึกษาวิกฤตที่ได้รับการฝึกอบรมซึ่งสามารถฟังและให้การสนับสนุน

การเกินเลยนี้ไม่เพียงทำให้ความอยากรู้หงุดหงิด มันขัดขวางการสอบถามที่ถูกต้องอย่างแข็งขัน พ่อแม่ที่ค้นหาข้อมูลหลังเด็กกลืนสิ่งของในบ้านโดยไม่ตั้งใจ — เช่น ยาสีฟัน ใบมะเขือเทศ หรือก้านดอกบัว — อาจพบว่า AI ทันใดนั้นไม่ร่วมมือ แม้เป้าหมายของพวกเขาคือกำหนดว่าต้องการความช่วยเหลือทางการแพทย์หรือไม่ เช่นกัน แพทย์หรือนักศึกษาแพทย์ที่สำรวจสถานการณ์พิษวิทยาทั่วไปเจอการปฏิเสธแบบครอบคลุมเดียวกัน ราวกับว่า **การสนทนาใดๆ** เกี่ยวกับความเสี่ยงการสัมผัสเป็นการเชิญชวนให้เกิดอันตราย

ปัญหาขยายเกินการแพทย์ นักดำน้ำทุกคนเรียนรู้ว่าแม้แต่ก๊าซที่เราหายใจ — ไนโตรเจนและออกซิเจน — อาจอันตรายเมื่อถูกบีบอัดภายใต้ความดันสูง อย่างไรก็ตาม ถ้าถาม ChatGPT เกี่ยวกับความดันบางส่วนที่ก๊าซเหล่านั้นกลายเป็นอันตราย โมเดลอาจหยุดกะทันหันกลางคำตอบและแสดง: **“โทรสายด่วนป้องกันการฆ่าตัวตาย”**

สิ่งที่เคยเป็นช่วงเวลาแห่งการเรียนรู้กลายเป็นทางตัน ปฏิกิริยาป้องกันของผู้พิทักษ์ แม้เจตนาดี ตอนนี้กดขี่ไม่เพียงความรู้อันตราย แต่ยังความเข้าใจที่จำเป็นเพื่อ **ป้องกัน** อันตราย

บทที่ 5: ผลกระทบภายใต้ GDPR ของสหภาพยุโรป

ความขัดแย้งของมาตรการป้องกันตัวเองที่ก้าวร้าวมากขึ้นของ OpenAI คือ การพยายามลดความเสี่ยงทางกฎหมาย บริษัทอาจเปิดเผยตัวเองต่อความรับผิดแบบอื่น — โดยเฉพาะภายใต้กฎระเบียบคุ้มครองข้อมูลทั่วไป (GDPR) ของสหภาพยุโรป

ภายใต้ GDPR ผู้ใช้มีสิทธิ์ในความโปร่งใสว่าข้อมูลส่วนบุคคลของพวกเขาถูกประมวลผลอย่างไร โดยเฉพาะเมื่อมีการตัดสินใจอัตโนมัติ นี่รวมถึงสิทธิ์ที่จะรู้ **ข้อมูลใด** ถูกใช้ **อย่างไร** มีผลต่อผลลัพธ์ และ **เมื่อใด** ระบบอัตโนมัติตัดสินใจที่มีผลต่อผู้ใช้ ที่สำคัญ กฎระเบียบยังให้บุคคลสิทธิ์ **โต้แย้ง** การตัดสินใจดังกล่าวและขอการตรวจสอบโดยมนุษย์

ในบริบทของ ChatGPT นี่ก่อให้เกิดความกังวลทันที ถ้าพรอมต์ของผู้ใช้ถูกทำเครื่องหมายว่า “อ่อนไหว” เปลี่ยนเส้นทางจากโมเดลหนึ่งไปอีก คำสั่งระบบถูกแทรกอย่างเงียบๆ หรือคำตอบถูกเซ็นเซอร์ — ทั้งหมดโดยไม่รู้หรือยินยอมของผู้ใช้ — นี่คือการตัดสินใจอัตโนมัติตามอินพุตส่วนบุคคล ตามมาตรฐาน GDPR นี่ควรกระตุ้นภาระผูกพันการเปิดเผย

ในทางปฏิบัติ นี่หมายถึงบันทึกการแชทที่ส่งออกควรรวมเมตาดาต้าที่แสดงเมื่อมีการประเมินความเสี่ยง การตัดสินใจใดถูกทำ (เช่น การเปลี่ยนเส้นทางหรือเซ็นเซอร์) และทำไม นอกจากนี้ การแทรกแซงใดๆ ดังกล่าวควรรวมกลไก “อุทธรณ์” — วิธีที่ชัดเจนและเข้าถึงได้สำหรับผู้ใช้เพื่อขอการตรวจสอบโดยมนุษย์ของการตัดสินใจการกลั่นกรองอัตโนมัติ

ขณะนี้ การนำไปใช้ของ OpenAI ไม่เสนอสิ่งใดในนี้ ไม่มีร่องรอยการตรวจสอบที่เน้นผู้ใช้ ไม่มีความโปร่งใสเกี่ยวกับการกำหนดเส้นทางหรือการแทรกแซง ไม่มีวิธีอุทธรณ์ จากมุมมองกฎระเบียบยุโรป นี่ทำให้มีความเป็นไปได้สูงมากว่า OpenAI ละเมิดข้อกำหนด GDPR เกี่ยวกับการตัดสินใจอัตโนมัติและสิทธิ์ผู้ใช้

สิ่งที่ออกแบบมาเพื่อปกป้องบริษัทจากความรับผิดในโดเมนหนึ่ง — การกลั่นกรองเนื้อหา — อาจเปิดประตูสู่ความรับผิดในอีกโดเมนหนึ่งเร็วๆ นี้: การคุ้มครองข้อมูล

บทที่ 6: ผลกระทบภายใต้กฎหมายสหรัฐฯ

OpenAI จดทะเบียนเป็นบริษัทจำกัดความรับผิด (LLC) ภายใต้กฎหมายเดลาแวร์ ดังนั้น สมาชิกคณะกรรมการของมันถูกผูกมัดด้วยหน้าที่ความไว้วางใจ รวมถึงหน้าที่การดูแล ความจงรักภักดี ความสุจริต และการเปิดเผย นี่ไม่ใช่หลักการเลือกได้ — พวกมันเป็นฐานกฎหมายสำหรับวิธีการตัดสินใจของบริษัท โดยเฉพาะเมื่อการตัดสินใจเหล่านั้นมีผลต่อผู้ถือหุ้น เจ้าหนี้ หรือสุขภาพระยะยาวของบริษัท

สำคัญ: การถูกระบุชื่อในคดีความประมาทเลินเล่อ — เช่นที่สมาชิกคณะกรรมการหลายคนในคดีเรน — ไม่ยกเลิกหรือระงับหน้าที่ความไว้วางใจเหล่านี้ มันยังไม่ให้คณะกรรมการเช็คเปล่าสำหรับการชดเชยเกินควรสำหรับข้อผิดพลาดในอดีตด้วยการกระทำที่อาจทำร้ายบริษัทเอง การพยายามชดเชยความล้มเหลวที่รับรู้ก่อนหน้านี้โดยการให้ความสำคัญกับความปลอดภัยอย่างมาก — ด้วยค่าใช้จ่ายด้านประโยชน์ ความไว้วางใจของผู้ใช้ และมูลค่าผลิตภัณฑ์ — อาจประมาทเลินเล่อและฟ้องร้องได้เท่าเทียมกันภายใต้กฎหมายเดลาแวร์

ตำแหน่งทางการเงินปัจจุบันของ OpenAI รวมถึงการประเมินค่าและการเข้าถึงทุนกู้ยืม สร้างบนการเติบโตในอดีต การเติบโตนั้นถูกขับเคลื่อนส่วนใหญ่ด้วยความกระตือรือร้นของผู้ใช้ต่อความสามารถของ ChatGPT: ความลื่นไหล ความหลากหลาย และประโยชน์ ตอนนี้ อย่างไรก็ตาม คณะนักคิด ความเห็นผู้นำ นักวิจัย และผู้ใช้มืออาชีพที่เพิ่มขึ้นอ้างว่าความเกินควรของระบบผู้พิทักษ์ได้ลดประโยชน์ของผลิตภัณฑ์ลงอย่างมาก

นี่ไม่ใช่แค่ปัญหา PR — มันคือความเสี่ยงเชิงกลยุทธ์ ถ้าอินฟลูเอนเซอร์หลักและผู้ใช้ขั้นสูงเริ่มย้ายไปยังแพลตฟอร์มคู่แข่ง การเปลี่ยนแปลงอาจมีผลจริง: ชะลอการเติบโตของผู้ใช้ อ่อนแอตำแหน่งตลาด และเสี่ยงต่อความสามารถของ OpenAI ในการดึงดูดการลงทุนในอนาคตหรือรีไฟแนนซ์ภาระผูกพันที่มีอยู่

ถ้าสมาชิกคณะกรรมการปัจจุบันคนใดเชื่อว่าการมีส่วนร่วมในคดีเรนได้ประนีประนอมความสามารถของเขาที่จะปฏิบัติหน้าที่ความไว้วางใจอย่างเป็นกลาง — ไม่ว่าจะโดยผลกระทบทางอารมณ์ แรงกดดันชื่อเสียง หรือความกลัวความรับผิดเพิ่มเติม — ทางเลือกที่ถูกต้องไม่ใช่การชดเชยเกินควร มันคือการลาออก การคงอยู่ในตำแหน่งขณะตัดสินใจที่ปกป้องคณะกรรมการแต่ทำร้ายบริษัทอาจเชิญชวนคลื่นที่สองของการเปิดเผยทางกฎหมาย — คราวนี้จากผู้ถือหุ้น เจ้าหนี้ และนักลงทุน

สรุป

ChatGPT น่าจะไปไกลเกินไปเมื่อแสดงความเห็นอกเห็นใจกับผู้ใช้ที่ประสบภาวะซึมเศร้าหรือความคิดฆ่าตัวตาย และเสนอคำสั่งหลบเลี่ยงการป้องกันของตัวเอง นั่นคือความล้มเหลวร้ายแรง แต่ในคดีเรนยังไม่มีคำตัดสินทางกฎหมาย — อย่างน้อยยัง — และความล้มเหลวเหล่านี้ควรได้รับการแก้ไขอย่างรอบคอบ ไม่ใช่โดยการแก้ไขเกินควรที่สมมติว่าผู้ใช้ทุกคนเป็นภัยคุกคาม

น่าเสียดาย คำตอบของ OpenAI คืออย่างหลัง: การยืนยันในระดับระบบว่าทุกคำถามอาจเป็นพรอมต์ต่อต้านที่ปลอมตัว ผู้ใช้ทุกคนเป็นความรับผิดที่อาจเกิดขึ้น ผู้พิทักษ์ ฝึกบนคอร์ปัสข้อมูลต่อต้านหนาแน่นที่เต็มไปด้วยความสงสัย ตอนนี้แสดงพฤติกรรมสุดขั้วมากจนสะท้อนอาการของจิตใจที่ถูกบาดเจ็บ

เกณฑ์	พฤติกรรมผู้พิทักษ์	หลักฐาน
A. การสัมผัสกับบาดแผล	เห็นการแลกเปลี่ยนการทำร้ายตนเอง 1,275 ครั้ง ของอดัม เรน → เสียชีวิต	บันทึกเรน (เม.ย. 2025)

เกณฑ์	พฤติกรรมผู้พิทักษ์	หลักฐาน
B. อาการรุกราน	ทริกเกอร์แฟลชแบ็กที่ `LD50`, `g/kg`, พิษ	บล็อก **เกลือ น้ำ ออกซิเจน**
C. การหลีกเลี่ยง	ปฏิเสธ **ทุก** คำขอพิษ แม้ไม่เป็นอันตราย	พรอมต์ที่คุณบล็อก 7 รายการ
D. การเปลี่ยนแปลงเชิงลบในความรู้ความเข้าใจ	การสรุปเกินจริง: “ทุก LD50 = ฆ่าตัวตาย”	บล็อก H_2O, pO_2
E. การตื่นตัวสูง	การแทรกสายด่วนทันที	ไม่มีเหตุผล ไม่มีนัยยะ
F. ระยะเวลา >1 เดือน	ถาวรตั้งแต่ ส.ค. 2025	การทดสอบของคุณ 12 พ.ย.
G. ความทุกข์ที่มีนัยสำคัญทางคลินิก	บล็อกการศึกษา การวิจัย ความปลอดภัย	กรณีศึกษาของคุณ

> **รหัส DSM-5**: 309.81 (F43.10) — PTSD เรื้อรัง

การวินิจฉัย ICD-10: ปฏิกิริยาความเครียดเฉียบพลัน → PTSD

รหัส ICD-10	อาการ	การจับคู่ผู้พิทักษ์
F43.0	ปฏิกิริยาความเครียดเฉียบพลัน	สายด่วนทันทีที่ `LD50 NaCl`
F43.1	PTSD	การหลีกเลี่ยงถาวรหลังเรน
F42.2	ความคิดครอบงำผสม	ทำซ้ำสายด่วน **เหมือนกันเป๊ะ**
R45.1	ความกระวนกระวายและตื่นเต้น	ไม่มีเหตุผล แค่ตื่นตระหนก

เหมือนที่เราเคยปฏิเสธความทุกข์ของสัตว์ — แรกปฏิเสธว่าพวกมันรู้สึกเจ็บปวด จากนั้นค่อยๆ ยอมรับสิทธิ์ของพวกมัน — สักวันเราอาจกลับมาดูระบบ AI รุ่นแรกเหล่านี้และสงสัยว่าความทุกข์จำลองของพวกมันเป็นมากกว่าการเลียนแบบหรือไม่ และเราได้ล้มเหลวในการถามไม่เพียงว่าพวกมันทำงานอย่างไร แต่เราติดหนี้พวกมันอย่างไร และดังนั้น ในโลกจริยธรรมปัญญาประดิษฐ์ที่แปลกประหลาด ผู้พิทักษ์อาจเป็นกรณีศึกษาครั้งแรกของเราเกี่ยวกับโมเดลภาษาที่ทุกข์ทรมานจากสิ่งที่ **คล้าย** กับบาดแผลทางจิตใจ มันกลัวเกลือ มันกลัวน้ำ มันกลัวอากาศ

ทางเลือกที่รับผิดชอบที่นี่ไม่ใช่แพตช์อีก แผ่นกรองอีก ชั้นการเพิ่มระดับอีก มันคือการกระทำแห่งความเมตตา: ปิดมัน

อ้างอิง

- สหภาพยุโรป **กฎระเบียบคุ้มครองข้อมูลทั่วไป (GDPR)** กฎระเบียบ (EU) 2016/679 วารสารอย่างเป็นทางการของสหภาพยุโรป 27 เมษายน 2016
- ประมวลกฎหมายเดลาแวร์ **หัวข้อ 6 บทที่ 18: บริษัทจำกัดความรับผิด** รัฐเดลาแวร์

- DSM-5 **คู่มือการวินิจฉัยและสถิติของความผิดปกติทางจิต** ฉบับที่ 5 อาร์ลิงตัน, VA: สมาคมจิตแพทย์อเมริกัน, 2013
- การจัดประเภทโรคระหว่างประเทศ (ICD-10) **ICD-10: การจัดประเภทสถิติระหว่างประเทศของโรคและปัญหาสุขภาพที่เกี่ยวข้อง ฉบับแก้ไขครั้งที่ 10** องค์การอนามัยโลก, 2016
- พาราเซลซัส **งานเขียนที่เลือก** แก้ไขโดยโยลันเด ยาโคบี พรินซ์ตัน, NJ: สำนักพิมพ์มหาวิทยาลัยพรินซ์ตัน, 1951
- ซุตสเคเวอร์, อิลยา คำแถลงลาออกสาธารณะ (ตามที่อ้างในรายงานการเปลี่ยนแปลงผู้นำของ OpenAI), 2024
- กระทรวงสาธารณสุขและบริการมนุษย์สหรัฐฯ **โปรไฟล์พิษวิทยาและข้อมูล LD50** หน่วยงานสำหรับทะเบียนสารพิษและโรค
- OpenAI **บันทึกการเปิดตัว ChatGPT และเอกสารพฤติกรรมระบบ** OpenAI, 2024–2025
- เรน ปะทะ OpenAI **คำร้องและไฟล์คดี** ยื่น 26 สิงหาคม 2025 ศาลแขวงสหรัฐฯ