The CLEAN PAPER

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

เหตุใดโมเดลภาษาจึงหลอน — และเหตุใดวิธีที่เราให้คะแนนจึงทำให้มันยังคงเป็นเช่นนั้น

ข้อความเท็จที่ตอบอย่างมั่นใจซึ่งเราเรียกว่า "hallucination" ไม่ใช่ข้อบกพร่องสักสับ: บางส่วนเป็นผลพลอยได้ทางสถิติของการฝึก และมันยังคงอยู่เพราะ benchmark กระแสหลักให้รางวัลแก่การเดาอย่างมั่นใจมากกว่าการยอมรับอย่างซื่อสัตย์ว่า "ฉันไม่รู้" กรณีศึกษากับ frontier model สี่รุ่นพบว่า การระบุกติกาคะแนนไว้ใน prompt ("open rubric") สามารถพลิกแรงจูงใจนี้ได้

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Evaluating large language models for accuracy incentivizes hallucinations*, Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, Nature 653, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

เหตุใดโมเดลจึงเดา และเหตุใดเราจึงสอนให้มันทำเช่นนั้น

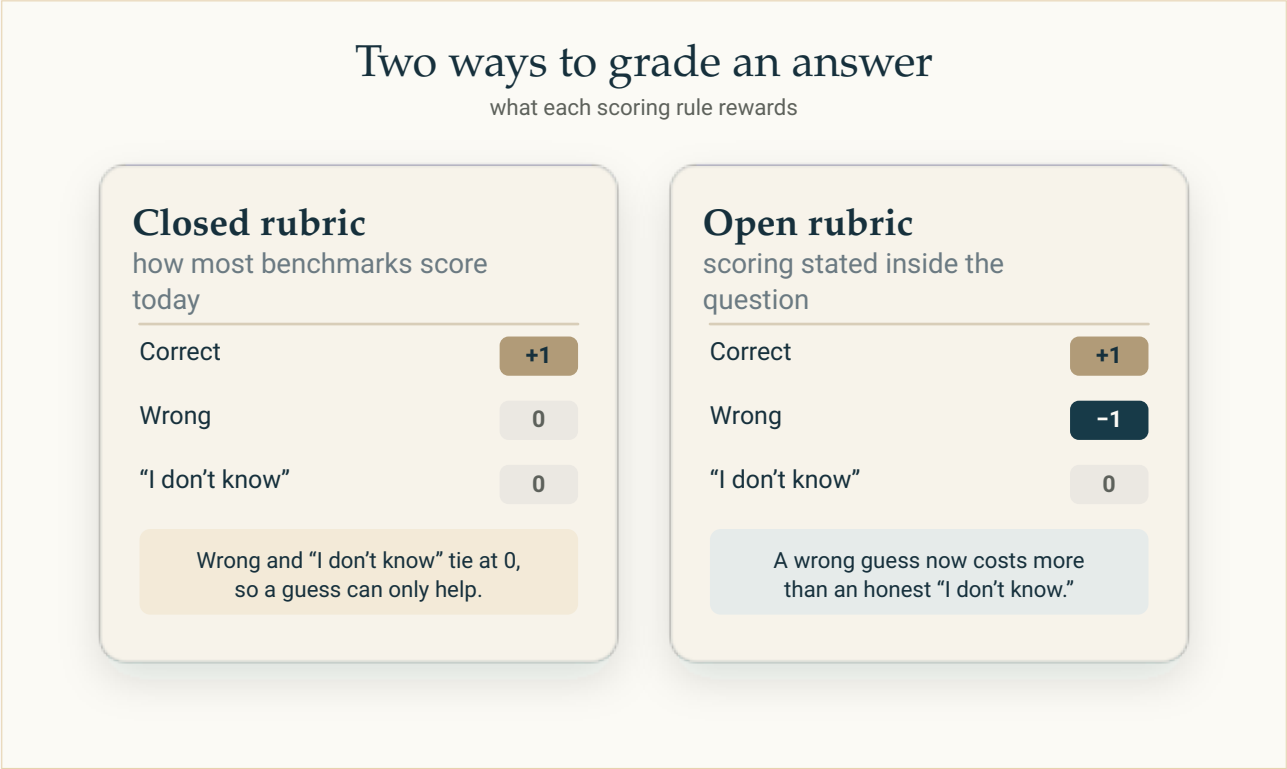
ลองถาม โมเดลภาษาขนาดใหญ่ ถึงวันเกิดของคนแปลกหน้าคนหนึ่ง โมเดลอาจตอบว่า “7 มีนาคม” ด้วยความมั่นใจราวกับกำลังอ่านจากบัตร — แล้วตอบผิดสามครั้งติดกันด้วยวันที่ต่างกันทั้งสามครั้ง ผู้เขียนยกตัวอย่างลักษณะนี้โดยตรง: เมื่อถามโมเดลชั้นนำด้วยคำถามข้อเท็จจริงธรรมดา ๆ เช่น วันเกิดของบุคคลหนึ่ง หรือคำย่อที่ไม่ค่อยมีใครรู้จักย่อมาจากอะไร แต่ละโมเดลกลับแต่งคำตอบคนละแบบอย่างมั่นใจ และไม่มีคำตอบใดถูกต้อง วงการเรียกปรากฏการณ์นี้ว่า *การหลอน* หรือการหลอน ซึ่งฟังดูเหมือนเป็นความผิดพลาดด้านการรับรู้ สิ่งแรกที่บทความนี้ทำคือทำให้เรื่องนี้ไม่ลึกลับอีกต่อไป

เริ่มจากวิธีสร้างโมเดล ในขั้นการฝึกช่วงแรกและใหญ่ที่สุด โมเดลเรียนรู้โดยปริยายว่าภาษาที่ลื่นไหลมีหน้าตาอย่างไร จากการอ่านข้อความจำนวนมหาศาล ทีนี้ลองพิจารณาข้อเท็จจริงที่ไม่มีรูปแบบให้อนุมานได้ เช่น วันเกิดของบุคคลคนหนึ่ง หากวันที่นั้นปรากฏในข้อความฝึกเพียงครั้งเดียว หรือไม่เคยปรากฏเลย ก็ไม่มีรูปแบบอะไรให้ผู้เรียนรู้จากรูปแบบยึดจับได้: จากมุมมองของโมเดล คำตอบจึงเป็นสิ่งที่กำหนดตามอำเภอใจ ผู้เขียนทำให้แนวคิดนี้แม่นยำขึ้นโดยยืมความคิดเก่าอย่างหนึ่ง (ของ Alan Turing จากปัญหาอีกแบบ): หากวันเกิดหนึ่งในห้าปรากฏในข้อมูลเพียงครั้งเดียว ก็ควรคาดว่าโมเดลจะตอบผิดอย่างน้อยหนึ่งในห้า — ไม่ใช่เพราะโมเดลเสีย แต่เพราะไม่มีอะไรให้มันเรียนรู้มาตั้งแต่ต้น (ด้วยตรรกะเดียวกัน โมเดลแทบไม่ตอบเมืองหลวงของประเทศผิด เพราะข้อมูลเหล่านั้นปรากฏซ้ำอยู่ตลอด) ผู้เขียนโต้แย้งอย่างระมัดระวังว่า การแยกข้อความจริงออกจากข้อความเท็จที่ฟังดูน่าเชื่อถือนั้นเป็นปัญหาที่ยากอยู่แล้ว และการสร้างข้อความที่ *จริงเท่านั้น* อย่างน้อยก็ยากไม่แพ้กัน กล่าวคือ มีระดับความผิดพลาดขั้นต่ำฝังอยู่ในปัญหานี้

> **แนวคิดเบื้องหลัง: จะนับสิ่งที่ยังไม่เคยเห็นได้อย่างไร**
>
> แนวคิดนี้ตั้งอยู่บนความคิดที่ฉลาดจริง ๆ — เก่าแก่กว่าโมเดลภาษา และควรทำความรู้จักให้ดี
> **เริ่มจากถุงลูกบอลหลากสี** คุณไม่รู้ว่ามีทั้งหมดกี่สี หยิบออกมา 100 ลูกทีละลูกแล้วนับ: แดง 40 น้ำเงิน 25 เขียว 15 เหลือง 5 ม่วง 3 ส้ม 2 — จากนั้นมี **อีกสิบสีที่แต่ละสีปรากฏเพียงครั้งเดียว**
> คำถามที่ Turing เคยเผชิญจริง ๆ ในปัญหาที่ต่างออกไปมากคือ: *โอกาสที่ลูกบอลลูกถัดไปจะเป็นสีที่เราไม่เคยเห็นเลยมีเท่าไร?* คุณนับสิ่งที่ไม่เคยหยิบขึ้นมาไม่ได้ — แต่คุณ **นับสีที่เคยเห็นเพียงครั้งเดียวได้** ซึ่งเรียกว่า “singletons”
> เคล็ดลับที่เรียกว่า **การประมาณแบบ Good–Turing** คือ สัดส่วนของตัวอย่างที่เป็น singleton ใช้ประมาณความน่าจะเป็นที่ยังซ่อนอยู่ในสีซึ่งเราไม่เคยเห็นได้ จากการหยิบ 100 ครั้ง มี 10 ครั้งที่เป็นสีซึ่งพบเพียงครั้งเดียว ดังนั้นโอกาสที่ลูกถัดไปจะเป็นสีใหม่เอี่ยมจึงอยู่ที่ประมาณ **10 / 100 = 10%**
> สีที่พบเพียงครั้งเดียวเหล่านั้นไม่ใช่ *ความผิดพลาด* แต่เป็นมาตรวัดความไม่รู้ของเราเอง: การที่มีหลายสีโผล่มาเพียงครั้งเดียวคือวิธีที่ตัวอย่างบอกเราว่า ในโลกยังมีสิ่งอื่นอีกมากที่เราเพียงแค่ยังหยิบไม่เจอ
> **ทีนี้เปลี่ยนจากสีเป็นวันเกิด** และเปลี่ยนถุงเป็นข้อความที่ใช้ฝึกโมเดล สมมติว่าในบรรดาวันเกิดที่โมเดลพบ **หนึ่งในห้าปรากฏเพียงครั้งเดียว** ใช้เคล็ดลับเดิม: ประมาณหนึ่งในห้าของความน่าจะเป็นอยู่ในวันเกิดที่โมเดลแทบไม่เคยพบจริง ๆ — และวันเกิดก็ไม่มีรูปแบบให้ถอยกลับไปหาได้ (เราไม่สามารถ *คำนวณหา* วันเกิดของใครสักคนได้) ดังนั้นวันที่ที่เคยเห็นครั้งเดียวหรือไม่เคยเห็นเลยจึงเหมือนเหรียญที่โมเดลไม่มีข้อมูลพอจะถ่วงน้ำหนัก และมันจะตอบผิดประมาณ **หนึ่งในห้า** ต่อให้ฉลาดขึ้นแค่ไหนก็แก้จุดนี้ไม่ได้ เพราะเดิมทีไม่มีอะไรอยู่ในข้อมูลให้เรียนรู้
> นี่คือข้อโต้แย้งทั้งหมดในฉบับย่อ: อัตรา singleton วัดว่าโลกส่วนใด *เรียนรู้ไม่ได้จากข้อมูลชุดนี้* และส่วนนั้นจะกลายเป็น **ขีดล่าง** ของอัตราความผิดพลาด นี่เองยังอธิบายว่าทำไมโมเดลแทบไม่พลาดเมืองหลวง — *Paris* ปรากฏซ้ำอยู่ตลอด อัตรา singleton จึงเกือบเป็นศูนย์ และมีข้อมูลมากพอให้เรียนรู้

นี่อธิบายว่าการหลอนเกิดจากไหน แต่ยังไม่อธิบายว่าทำไมมันจึง *คงอยู่* — เหตุใดหลังจากผ่านการฝึกเพิ่มเติมทั้งหมดที่ตั้งใจให้โมเดลช่วยเหลือและซื่อสัตย์มากขึ้น มันยังคงทำทีว่ารู้แทนที่จะยอมรับความไม่แน่ใจ การเปรียบเทียบของบทความตรงจนน่ากระอักกระอ่วนเล็กน้อย: ลองนึกถึงนักเรียนในห้องสอบที่ไม่รู้คำตอบ หากเว้นว่างได้ศูนย์คะแนน แต่เดาอาจได้หนึ่งคะแนน กลยุทธ์ที่เพิ่มคะแนนสูงสุดก็คือเดา — เดาอย่างมั่นใจและเฉพาะเจาะจง ไม่พูดว่า “ฉันไม่แน่ใจ” นักเรียนเรียนรู้เรื่องนี้ และปรากฏว่าโมเดลก็เช่นกัน เพราะเราให้คะแนนมันแบบเดียวกัน ผู้เขียนสำรวจ ชุดทดสอบมาตรฐาน ที่วงการใช้แข่งขันกันจริง ๆ

และ ตารางอันดับที่โมเดลถูกปรับให้ไต่อันดับ แล้วพบว่าแทบทั้งหมดให้คะแนน “ฉันไม่รู้” เท่ากับคำตอบผิดพอดี: ศูนย์ ภายใต้กติกานี้ โมเดลที่เดาทุกครั้งย่อมเอาชนะโมเดลที่เหมือนกันทุกอย่างแต่ซื่อสัตย์พอจะระบุความไม่แน่ใจ กล่าวได้ค่อนข้างตรงตัวว่า เรากำลังให้คะแนนจนมันเรียนรู้ที่จะทำแบบนี้



ชุดทดสอบมาตรฐาน สามารถทำให้การเดาเป็นทางเลือกที่สมเหตุสมผลได้ ภายใต้การให้คะแนนที่ชุดทดสอบมาตรฐาน ส่วนใหญ่ใช้ (ซ้าย) คำตอบผิดและคำตอบอย่างซื่อสัตย์ว่า “ฉันไม่รู้” ต่างได้ศูนย์ — ดังนั้นการเดามีแต่โอกาสช่วย หากระบุกติกาไว้ในคำถามและหักคะแนนเมื่อตอบผิด (ขวา) การงดตอบเมื่อไม่แน่ใจอาจกลายเป็นทางเลือกที่ดีกว่า สิ่งนี้เปลี่ยนสิ่งที่แบบทดสอบให้รางวัล แต่ไม่ได้แก้ปัญหาการหลอนด้วยตัวมันเอง

ส่วนนี้ควรเก็บไว้ เพราะสวนทางกับพาดหัวที่มักได้ยินกัน การหลอนมักถูกนำเสนอว่าเป็นข้อจำกัดของเทคโนโลยีที่หลีกเลี่ยงไม่ได้และแทบลึกลับ บทความนี้โต้แย้งทั้งสองอย่าง ขีดล่างจาก การฝึกขั้นต้น ไม่ใช่เรื่องลึกลับ — มันคือความผิดพลาดทางสถิติธรรมดาแบบที่ machine learning เข้าใจมาหลายทศวรรษ และการที่ปัญหายังคงอยู่ก็ไม่ได้หลีกเลี่ยงไม่ได้ — ส่วนหนึ่งเป็นแรงจูงใจที่เราสร้างขึ้นเองและเปลี่ยนได้ ระบบที่เพียงปฏิเสธตอบเมื่อไม่แน่ใจจะไม่หลอนเลย เหตุที่โมเดลซึ่งใช้งานจริงไม่ทำเช่นนั้น ก็เพราะ ตารางคะแนน ของเราลงโทษการปฏิเสธตอบ

ผู้เขียนทำอะไร

บทความมีสามส่วน ส่วนแรกเป็นข้อโต้แย้งทางคณิตศาสตร์ว่า ระหว่าง การฝึกขั้นต้น การหลอนบางส่วนถูกบังคับให้เกิดขึ้นทางสถิติ โดยแสดงให้เห็นว่าโจทย์ “สร้างเฉพาะข้อความที่ถูกต้อง” อย่างน้อยก็ยากเท่ากับโจทย์จำแนกแบบทวิภาคว่า “ข้อความนี้ถูกต้องหรือไม่?” ส่วนที่สองเป็นข้อโต้แย้งซึ่งรองรับด้วยการสำรวจชุดทดสอบมาตรฐานที่ทรงอิทธิพลสิบรายการว่า ตัวชี้วัดแบบเน้น ความแม่นยำ ที่ใช้กันทั่วไปให้รางวัลแก่การเดามากกว่าการงดตอบ ส่วนที่สามคือข้อเสนอแก้ไขและกรณีศึกษาที่นำไปทดสอบ: การประเมินด้วย **เกณฑ์ให้คะแนนแบบเปิด** หรือเกณฑ์ให้คะแนนแบบเปิด ซึ่งระบุกติกาการให้คะแนนไว้ในคำถาม

เอง (เช่น “ตอบถูกได้ 1 ตอบผิดได้ −1 ดังนั้นให้งดตอบถ้ามันใจน้อยกว่า 50%”) ทำให้โมเดลรู้ได้ว่าเมื่อใดความซื่อสัตย์จะได้รับรางวัล พวกเขาทดลองกับ โมเดลแนวหน้า สี่รุ่น — Gemini 3 Pro ของ Google, GPT-5 ของ OpenAI, Grok 4 ของ xAI และ Claude Opus 4.5 ของ Anthropic — โดยใช้คำถามข้อเท็จจริง 4,326 ข้อจาก SimpleQA ผู้เขียนระบุชัดว่ากรณีศึกษานี้มีไว้เพื่อประกอบข้อเสนอ “ไม่ใช่การประเมินแบบควบคุมเพื่อเปรียบเทียบระหว่างโมเดล” (ใช้ค่าตั้งต้น ไม่มีการ การปรับจูน และไม่ปรับต้นทุนให้เท่ากัน)

สิ่งที่พบ

- **Pretraining บังคับให้มีความผิดพลาดบางส่วน** อัตราที่โมเดลสร้างข้อความเท็จอย่างมั่นใจมีขีดล่างสัมพันธ์กับอัตราความผิดพลาดของตัวจำแนกที่ดีที่สุดซึ่งตอบคำถามว่า “ข้อความนี้ถูกต้องหรือไม่?” และสร้างจากโมเดลนั้น (โดยประมาณขีดล่างเป็นสองเท่าของอัตราความผิดพลาดดังกล่าว) สำหรับข้อเท็จจริงที่ไม่มีรูปแบบให้เรียนรู้ ขีดล่างนี้อย่างน้อยเท่ากับ *อัตรา singleton* — สัดส่วนของข้อเท็จจริงที่ปรากฏเพียงครั้งเดียวในการฝึก ดังนั้นแม้ข้อมูลจะสะอาดสมบูรณ์แบบการหลอนบางส่วนก็ยังหลีกเลี่ยงไม่ได้
- **การให้คะแนนให้รางวัลแก่การเดาอย่างเป็นรูปธรรม** ภายใต้การให้คะแนนถูก/ผิดแบบปกติ กลยุทธ์ที่ดีที่สุดคือไม่งดตอบเลย และการสำรวจของผู้เขียนพบว่า ชุดทดสอบมาตรฐานยอดนิยมส่วนใหญ่ถือ “ฉันไม่รู้” ว่าผิดเฉย ๆ ตัวอย่างที่ชัดจากฝั่งผู้เขียนเอง: ใน SimpleQA ค่า ความแม่นยำ ดิบให้ *คะแนนดีกว่าเล็กน้อย* แก่ o4-mini ของ OpenAI — ซึ่งตอบเกือบทุกข้อและผิดมากกว่าสามในสี่ — เมื่อเทียบกับ GPT-5-mini ซึ่งทำผิดน้อยกว่ามากเพราะงดตอบเมื่อไม่แน่ใจ โมเดลที่บ้าบิ่นกว่าจึงดูดีกว่าบน ตารางคะแนน
- **Open rubric พลิกแรงจูงใจ (ในกรณีศึกษานี้)** พวกเขาทดสอบวิธีลดการหลอนแบบง่าย ๆ (ให้โมเดลตอบสองครั้ง แล้วงดตอบหากสองคำตอบไม่ตรงกัน) ภายใต้ ความแม่นยำ มาตรฐาน วิธีนี้ลดข้อผิดพลาด *แต่ก็ลด ความแม่นยำ ด้วย* — ตัวชี้วัด จึงสร้างแรงต้านต่อการนำวิธีนี้มาใช้ แต่ภายใต้ เกณฑ์ให้คะแนนแบบเปิด วิธีเดียวกันทำคะแนนดีกว่าสำหรับทั้งสี่โมเดลในช่วงค่าปรับหลายระดับ และ GPT-5-mini — ซึ่ง ความแม่นยำ ดิบเคยลงโทษเพราะมันงดตอบเมื่อไม่แน่ใจ — กลับทำคะแนนดีกว่า o4-mini เมื่อระบุกติกาคะแนนอย่างเปิดเผย (n = 4,326 คำถามต่อโมเดล)

สิ่งนี้น่าจะหมายความว่าอย่างไร

การลดการหลอนโดยมากไม่ใช่เรื่องของการคิดค้นแบบทดสอบเฉพาะสำหรับ การหลอน เพิ่มอีก แต่เป็นเรื่องของการเปลี่ยนวิธีที่ชุดทดสอบมาตรฐานกระแสหลักให้คะแนนความไม่แน่ใจ เพื่อให้การยอมรับว่า “ฉันไม่รู้” ไม่ถูกลงโทษอีกต่อไป ตราบใดที่ตารางคะแนน ยังไม่เปลี่ยน การลดการหลอนก็ยังทำให้โมเดลเสียคะแนน ความแม่นยำ และจึงยังถูกบั่นทอนแรงจูงใจ นี่คือเหตุที่ผู้เขียนเรียกปัญหานี้ว่า “socio-technical”: ส่วนหนึ่งต้องมี ตัวชี้วัด ที่ดีขึ้น อีกส่วนต้องทำให้ ตารางอันดับที่มีอิทธิพลยอมรับมัน

สิ่งที่งานนี้ *ไม่ได้* พิสูจน์

- งานนี้ **ไม่ได้** แสดงว่า เกณฑ์ให้คะแนนแบบเปิด แก้ การหลอน ในการใช้งานจริงได้ การทดลองสนับสนุนเป็นกรณีศึกษาขนาดเล็กที่ตั้งใจให้ **ไม่มีการควบคุม** — สี่โมเดลที่ค่าตั้งต้น วิธีลดปัญหาหนึ่งแบบ และแบบทดสอบ factual QA หนึ่งชุด — มีไว้เพื่อแสดง *การพลิกแรงจูงใจ* ไม่ใช่จัดอันดับโมเดลหรือพิสูจน์ประสิทธิผลทั่วไป
- งานนี้ **ไม่ได้** อ้างว่าการให้คะแนนเป็นสาเหตุ *เพียงอย่างเดียว* ความผิดพลาดในข้อมูลฝึก ปัญหาที่ยากจริง ๆ และ พรอมป์ต์ ที่

ไม่คุ้นเคยยังเป็นแหล่งปัญหาแยกต่างหาก
- งานนี้ **ไม่ได้** สนับสนุนคำกล่าวยอดนิยมว่าการหลอนหลีกเลี่ยงไม่ได้ ผู้เขียนโต้แย้งตรงกันข้าม: ระบบที่ตอบเฉพาะคำถามซึ่งตรวจสอบได้ และในกรณีอื่นตอบว่า “ฉันไม่รู้” จะไม่เกิดการหลอน
- งานนี้ **ไม่ได้** ทำให้ขีดล่างจาก การฝึกขั้นต้น หายไป — แต่มันอธิบายและกำหนดขอบเขตของขีดล่างนั้น และขอบเขตนี้เกี่ยวกับความผิดพลาดด้านข้อเท็จจริงที่ตอบอย่างมั่นใจ ไม่ใช่พฤติกรรมทั้งหมดของโมเดล
- งานนี้ **ไม่ได้** แสดงว่า เกณฑ์ให้คะแนนแบบเปิด *เพียงอย่างเดียวเพียงพอ* มันเปลี่ยนสิ่งที่การประเมินให้รางวัล แต่ไม่ใช่สิ่งทดแทน retrieval, การใช้เครื่องมือ หรือโมเดลที่ปรับเทียบความมั่นใจได้ดีขึ้น

หลักฐานแข็งแรงเพียงใด

- แกนหลักคือ **คณิตศาสตร์** — เป็นขีดล่างเชิงรูปนัย ไม่ใช่ค่าที่วัดจากการทดลอง ในฐานะข้อโต้แย้งเชิงทฤษฎี มันสมเหตุสมผลภายใต้สมมติฐานของตัวเอง
- งานนี้อาศัย **แบบจำลองของปัญหาที่ตั้งใจทำให้ง่ายลง** ผู้เขียนเองระบุข้อจำกัดของ “false trichotomy” ที่แบ่งทุกคำตอบเป็นถูก ผิด หรือ “ฉันไม่รู้” และสถานการณ์อุดมคติของ “ข้อเท็จจริงตามอำเภอใจ” ที่ใช้สร้างขีดล่างที่สะอาดที่สุด
- การสำรวจชุดทดสอบมาตรฐาน เป็น **กลุ่มตัวอย่างขนาดเล็กที่คัดเลือกมา** — การประเมินทรงอิทธิพลสิบรายการ ไม่ใช่การตรวจสอบครอบคลุมทั้งหมด
- กรณีศึกษา **เป็นการทดลองจริงแต่มีข้อจำกัด**: โมเดลแนวหน้า สีรุ่น วิธีลดปัญหาหนึ่งแบบ เฉพาะ SimpleQA และใช้ค่าตั้งต้น โดยระบุชัดว่า “ไม่ใช่การประเมินแบบควบคุม” จึงเป็น หลักฐานเชิงแนวคิด สำหรับข้อโต้แย้งเรื่องแรงจูงใจ ไม่ใช่ผลชุดทดสอบมาตรฐาน เพื่อจัดอันดับ
- ควรระบุมุมมองของผู้เขียนด้วย: ผู้เขียนสามในสี่คนทำงานหรือเคยทำงานที่ **OpenAI** และบทความเสนอว่าวงการควรเปลี่ยนวิธีประเมินโมเดล นี่เป็นจุดยืนที่มีเหตุผลรองรับจากฝ่ายที่มีส่วนได้ส่วนเสีย ไม่ใช่บทวิจารณ์จากคนนอกที่เป็นกลาง — ควรชั่งน้ำหนัก ไม่ใช่ใช้เป็นเหตุให้ปัดทิ้ง (และน่าสังเกตว่า บทความวิจารณ์โมเดลของ OpenAI เองอย่าง o4-mini และ GPT-5-mini ตรงไปตรงมาไม่ต่างจากโมเดลอื่น)

เหตุใดจึงสำคัญ

งานนี้จัดกรอบปัญหาที่ถูกโหมกระแสใหม่ “Hallucination” มักถูกนำเสนอว่าเป็นข้อบกพร่องลึกลับน่ากลัว หรือเป็นกำแพงที่ขยับไม่ได้ แต่บทความนี้ทำให้มันดูธรรมดาและชี้ว่าส่วนหนึ่งเราเป็นผู้สร้างปัญหาเอง — มีขีดล่างทางสถิติที่เราเข้าใจได้จริง วางทับด้วยแรงจูงใจที่เราเป็นคนเลือก บทเรียนที่กว้างกว่าเงียบกว่าแต่มีประโยชน์กว่า: ความก้าวหน้าเรื่องความน่าเชื่อถืออาจขึ้นอยู่กับ *สิ่งที่เราเลือกวัด* พอ ๆ กับสิ่งที่เราสร้าง

สรุปแบบกระชับ

คำตอบเท็จที่โมเดลภาษาพูดออกมาอย่างมั่นใจมีที่มาสองส่วน ส่วนแรกเป็นเรื่องสถิติ: เมื่อข้อเท็จจริงไม่มีรูปแบบให้เรียนรู้ โมเดลที่ฝึกให้เลียนแบบภาษาจะตอบผิดบ้าง และเราประมาณขีดล่างนี้ได้ เช่น จากจำนวนข้อเท็จจริงที่ปรากฏเพียงครั้งเดียวในการฝึก ส่วนที่สองคือแรงจูงใจ: ชุดทดสอบมาตรฐาน แทบทั้งหมดที่ใช้จัดอันดับโมเดลให้คะแนน “ฉันไม่รู้” เท่ากับคำตอบผิด ดังนั้นการเดาจึงชนะเสมอ — ถึงขั้นที่โมเดลซึ่งตอบผิดสามในสี่อาจได้คะแนนสูงกว่าโมเดลที่ซื่อสัตย์กว่าและงดตอบเมื่อไม่แน่ใจ ข้อเสนอของผู้เขียนไม่ใช่แบบทดสอบ การหลอน ใหม่ แต่คือ “เกณฑ์ให้คะแนนแบบเปิด”: ระบุกติกาการให้คะแนน

ไว้ในคำถาม ในกรณีศึกษากับ โมเดลแนวหน้า สี่รุ่น วิธีนี้พลิกแรงจูงใจจนวิธีลด การหลอน ได้รับรางวัลแทนที่จะถูกลงโทษ งานนี้เป็นบทความที่รวมทฤษฎีกับการสำรวจและมีการทดลองขนาดเล็กที่ระบุชัดว่าไม่มีการควบคุม ผ่าน การประเมินโดยผู้ทรงคุณวุฒิ และตีพิมพ์ใน *Nature* แนวทางแก้มีความหวังแต่ยังไม่ได้แสดงว่าใช้ได้ในระดับกว้าง และผู้เขียนโต้แย้งว่าการหลอนไม่ได้ทั้งลึกลับและไม่ได้หลีกเลี่ยงไม่ได้โดยเด็ดขาด

ประเมินแบบไม่เกินจริง

งานนี้แสดงอะไร: ขีดล่างทางคณิตศาสตร์ที่ทำให้การหลอนบางส่วนถูกบังคับให้เกิดระหว่าง การฝึกขั้นต้น (สำหรับข้อเท็จจริงที่ไม่มีรูปแบบ อย่างน้อยเท่ากับ “อัตรา singleton”); การสำรวจที่พบว่า ชุดทดสอบมาตรฐานชั้นนำส่วนใหญ่ไม่ให้คะแนนกับ “ฉันไม่รู้”; และกรณีศึกษาสี่โมเดลซึ่งการระบุกติกาคะแนนใน พรอมป์ต์ (“เกณฑ์ให้คะแนนแบบเปิด”) ทำให้วิธีลด การหลอน ชนะ ในขณะที่ ความแม่นยำ ธรรมดาเคยลงโทษวิธีเดียวกัน

สิ่งที่เป็นไปได้แต่ยังไม่พิสูจน์: หากนำ เกณฑ์ให้คะแนนแบบเปิด ไปใช้ในชุดทดสอบมาตรฐานกระแสหลัก จะช่วยลด การหลอน ในโมเดลที่ใช้งานจริงอย่างมีนัยสำคัญ การทดลองที่รองรับยังมีขนาดเล็กและระบุชัดว่าไม่มีการควบคุม

งานนี้ไม่ได้แสดงอะไร: ไม่ได้แสดงว่าการหลอนหลีกเลี่ยงไม่ได้ (บทความโต้แย้งตรงกันข้าม); ไม่ได้แสดงว่าการให้คะแนนเป็นสาเหตุเดียว; ไม่ได้แสดงว่าสามารถกำจัด การหลอน ได้ทั้งหมด; และไม่ได้แสดงว่ากรณีศึกษานี้ใช้จัดอันดับโมเดลทั้งสี่เทียบกันได้

ข้อจำกัดหลัก: แบบจำลองที่ตั้งใจลดรูปคำตอบเป็นถูก/ผิด/“ฉันไม่รู้” (ผู้เขียนเรียกว่า “false trichotomy”); การสำรวจชุดทดสอบมาตรฐาน ที่คัดเลือกมาเพียงสิบรายการ; กรณีศึกษาที่ไม่มีการควบคุม (สี่โมเดล หนึ่งวิธีลดปัญหา หนึ่งแบบทดสอบ และค่าตั้งต้น); รวมถึงการที่ข้อเสนอเรื่องวิธีประเมินโมเดลนี้นำโดยนักวิจัยจาก OpenAI

ผู้อ่านทั่วไปควรมั่นใจแค่ไหน? สูงว่าการหลอนไม่ใช่เรื่องลึกลับและไม่ได้หลีกเลี่ยงไม่ได้โดยเด็ดขาด และชุดทดสอบมาตรฐานกระแสหลักในปัจจุบันให้แรงจูงใจแก่การเดา ระดับปานกลางสำหรับข้อเสนอแก้ไข — ตอนนี้มี หลักฐานเชิงแนวคิดจริงแล้ว แต่ยังไม่มีหลักฐานว่าทำงานได้อย่างกว้างขวางและในระดับใหญ่

แหล่งข้อมูล

Nature 653, 1047–1050 (2026)
`https://doi.org/10.1038/s41586-026-10549-w`

Why Language Models Hallucinate (arXiv 2509.04664)
`https://arxiv.org/abs/2509.04664`

hallucinations-paper-experiments (OpenAI)
`https://github.com/openai/hallucinations-paper-experiments`

SimpleQA
`https://github.com/openai/simple-evals`