The Clean Paper

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

למה מודלי שפה מזים — ולמה שיטת הדירוג שלנו משמרת זאת

השקרים הבטוחים שאנו מכנים "הזיות" אינם תקלה מסתורית. חלקם תוצר לוואי סטטיסטי של אימון, והם נשארים מפני שמדדים מרכזיים מתגמלים ניחוש בטוח יותר מאשר "אני לא יודע" כנה. מקרה מבחן בארבעה מודלי חזית מראה שכתיבת כללי הניקוד בתוך השאלה ("מחוונים פתוחים") הופכת את התמריץ.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · 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 1050–1047 (2026) · DOI: 10.1038/s41586-026-10549-w

למה מודלים מנחשים, ולמה לימדנו אותם לעשות זאת

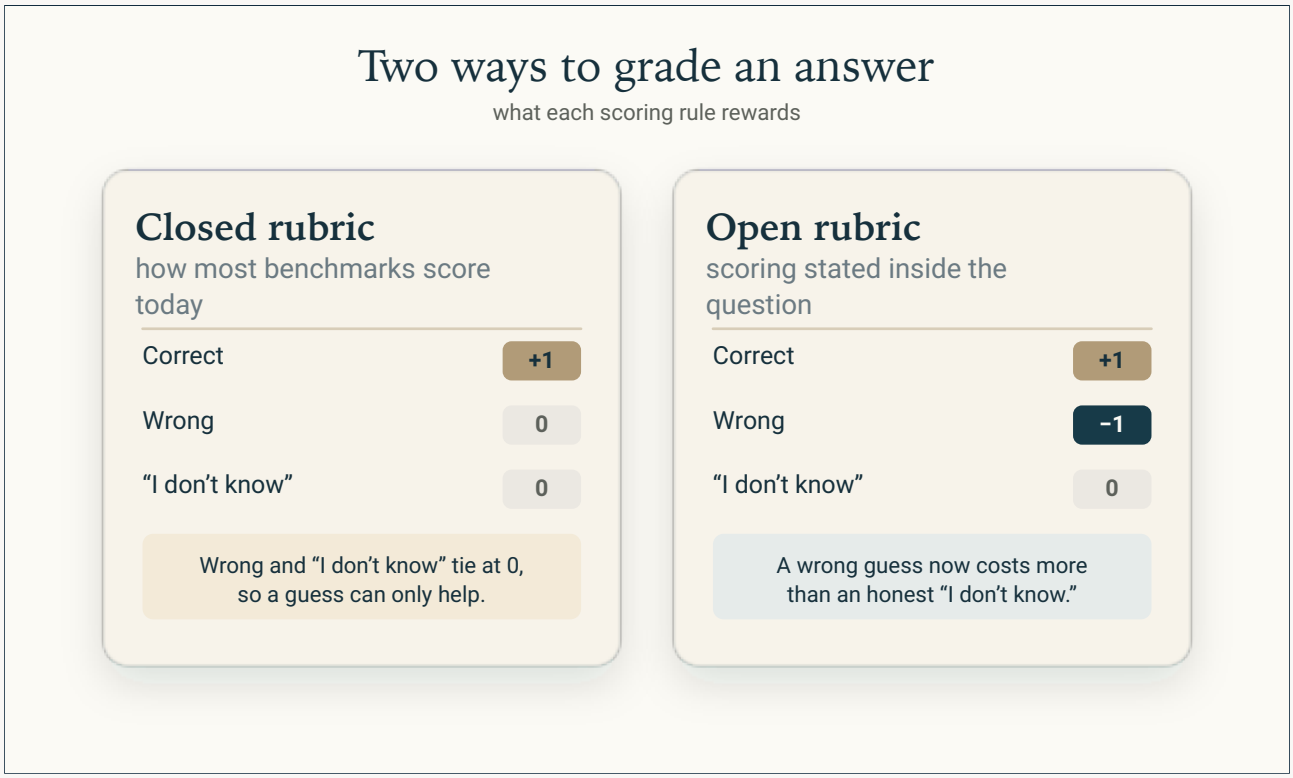
שאלו מודל שפה גדול על יום ההולדת של אדם לא מוכר, והוא עשוי לענות "7 במרץ" בביטחון יציב של מי שקורא זאת מכרטיס — ולהיות שגוי, שלוש פעמים ברצף, עם שלושה תאריכים שונים. המחברים נותנים בדיוק דוגמה כזאת: מודלים מובילים שנשאלו שאלה עובדתית פשוטה — יום הולדת של אדם, או פירוש של ראשי תיבות לא מוכרים — וכל אחד מהם המציא בביטחון תשובה אחרת, אף אחת מהן לא נכונה. המונח של התעשייה לכך הוא הזיה, מילה שגורמת לזה להישמע כמו תקלה בתפיסה. המהלך הראשון של המאמר הוא להוציא מזה את המסתורין.

נתחיל מאופן הבנייה של מודל. בשלב האימון הראשון והגדול ביותר שלו הוא לומד, למעשה, איך נראית שפה שוטפת, על ידי קריאה של כמות עצומה של טקסט. עכשיו קחו עובדה שאין מאחוריה דפוס — יום הולדת של אדם מסוים. אם התאריך הזה הופיע בטקסט האימון פעם אחת, או כלל לא, אין ללומד דפוסים במה להיאחז: התשובה, מנקודת המבט של המודל, שרירותית. המחברים מדייקים זאת באמצעות רעיון ישן (של אלן טיורינג, מבעיה אחרת): אם אחד מכל חמישה ימי הולדת מופיע בנתונים פעם אחת בלבד, יש לצפות שהמודל יטעה לפחות באחד מכל חמישה מהם — לא מפני שהוא מקולקל, אלא מפני שמעולם לא היה שם משהו ללמוד. (באותו היגיון, מודלים כמעט אף פעם לא טועים בבירת מדינה: אלה מופיעות כל הזמן.) הם טוענים בזהירות שהבחנה בין משפט אמיתי למשפט שגוי אך סביר היא עצמה בעיה קשה, ושהפקת משפטים אמיתיים בלבד קשה לפחות באותה מידה. רצפת שגיאה מובנית בבעיה.

> yet seen haven't you what count to how underneath: idea The
>
> כראוי. אותו לפגוש וכדאי שפה, ממודלי יותר ישן — באמת מבריק רעיון על נשען זה **התחילו בשק של כדורים צבעוניים.** אינכם יודעים כמה צבעים יש בו. אתם שולפים 100, אחד-אחד, וסופרים: אדום 40, כחול 25, ירוק 15, צהוב 5, סגול 3, כתום 2 — ואז **עשרה צבעים שונים שכל אחד מהם הופיע בדיוק פעם אחת.**
> עכשיו לשאלה שטיורינג באמת עמד מולה, בבעיה אחרת לגמרי: מה הסיכוי שהכדור הבא יהיה בצבע שלא ראיתם כלל? אי אפשר לספור את מה שמעולם לא שלפתם — אבל **אפשר** לספור את הצבעים שראיתם בדיוק פעם אחת, את ה"חד-פעמיים". הטריק, שנקרא **אמידת** Good–Turing, הוא שחלקן של השליפות שהיו חד-פעמיות מעריך את ההסתברות שעוד מסתתרת בצבעים שלא ראיתם. עשר מתוך מאה השליפות היו צבעים שהופיעו פעם אחת, לכן הסיכוי שהכדור הבא יהיה בצבע חדש לגמרי הוא בערך 10 / 100 = 10%. הצבעים שנראו פעם אחת אינם טעויות. הם מדידה של הבורות שלכם: הרבה צבעים שמופיעים פעם אחת הם הדרך של המדגם לומר לכם שהעולם מכיל יותר ממה שפשוט שלפתם עד כה.
> **עכשיו החליפו צבעים בימי הולדת,** ואת השק בטקסט האימון של המודל. נניח שמבין ימי ההולדת שהוא ראה, **אחד מכל חמישה מופיע בדיוק פעם אחת.** אותו טריק: בערך חמישית מן ההסתברות חיה בימי הולדת שהמודל למעשה מעולם לא ראה — וליום הולדת אין דפוס שאפשר להישען עליו (אי אפשר לחשב יום הולדת של מישהו). לכן תאריך שנראה פעם אחת, או מעולם לא נראה, הוא מטבע שהמודל לא יכול לשקלל, והוא יטעה בערך **באחד מכל חמישה** מהם. שום כמות של תחכום לא מתקנת זאת: לא היה שם דבר ללמוד. זו כל הטענה בקטן: שיעור החד-פעמיים מודד כמה מן העולם אינו ניתן ללמידה מן הנתונים האלה, וזה נעשה **רצפה** מתחת לשגיאות. זו גם הסיבה שמודל כמעט אינו מפספס עיר בירה — *Paris* מופיעה כל הזמן, שיעור החד-פעמיים שלה קרוב לאפס, ולכן יש הרבה מה ללמוד.

זה מסביר מאין באות הזיות. זה לא מסביר למה הן שורדות — למה מודלים, אחרי כל האימון המאוחר שנועד להפוך אותם למועילים וישרים, עדיין מבלפים במקום להודות בספק. כאן האנלוגיה של המאמר כמעט לא נוחה מרוב שהיא מדויקת. דמיינו תלמיד בבחינה שאינו יודע תשובה. אם תשובה ריקה מקבלת אפס וניחוש עשוי לקבל נקודה, המהלך שממקסם ציון הוא לנחש — בביטחון, באופן ספציפי, ולעולם לא "אני לא בטוח". תלמידים לומדים זאת. מתברר שגם מודלים — כי אנחנו מדרגים אותם באותה דרך. המחברים עברו על המדדים שהתחום מתחרה עליהם בפועל, לוחות

הדירוג שמכוונים מודלים לטפס בהם, ומצאו שכמעט כולם נותנים ל"אני לא יודע" בדיוק אותו ציון כמו לתשובה שגויה: אפס. תחת כלל כזה, מודל שתמיד מנחש ינצח מודל זהה אחרת שמסמן בכנות את אי-הוודאות שלו. במובן די מילולי, אנחנו מדרגים אותם אל תוך זה.



מדדים יכולים להפוך ניחוש לרציונלי. בשיטת הניקוד שרוב המדדים משתמשים בה (משמאל), תשובה שגויה ו"אני לא יודע" כנה מקבלות שתיהן אפס — ולכן ניחוש יכול רק לעזור. אם מציינים את הכללים בשאלה, עם עונש על טעות (מימין), הימנעות כשלא בטוחים יכולה להיות המהלך הטוב יותר. זה משנה מה המבחן מתגמל; זה לא פותר, כשלעצמו, הזיה.

Original diagram — The Clean Paper CC BY 4.0

זה החלק שכדאי לשמור, מפני שהוא הולך נגד הכותרת הרגילה. הזיה מוצגת לעיתים קרובות כמגבלה בלתי נמנעת, כמעט מיסטית, של הטכנולוגיה. המאמר חולק על שני החלקים. רצפת טרום-האימון אינה תעלומה — זו שגיאה סטטיסטית רגילה, מן הסוג שלמידת מכונה מבינה כבר עשורים. וההישרדות אינה בלתי נמנעת — היא, בחלקה, תמריץ שבנינו ואפשר לשנות. מערכת שפשוט מסרבת לענות כשאינה בטוחה לא תזיה כלל; הסיבה שמודלים פרוסים אינם מתנהגים כך היא שלוחות הציונים מענישים את הסירוב.

מה המחברים עשו

למאמר שלושה חלקים. ראשית, טיעון מתמטי שחלק מן ההזיה נכפה סטטיסטית בזמן טרום-האימון, על ידי הראיה ש"להפיק רק טקסט תקף" קשה לפחות כמו בעיית סיווג בינארית של "האם המשפט הזה תקף?". שנית, טיעון — מגובה בסקר של עשרה מדדים משפיעים — שמדדי דיוק מיינסטרימיים מתגמלים ניחוש על פני הימנעות. שלישית, תיקון מוצע ומקרה מבחן: הערכות **במחוון פתוח**, שבהן הניקוד כתוב בתוך השאלה עצמה (למשל, "תשובה נכונה מקבלת 1, שגויה −1, לכן הימנע אם אתה בטוח בפחות מ-50%"), כך שמודל יכול לדעת מתי יושר מתוגמל. הם מנסים זאת על ארבעה מודלי חזית — GPT-5 ו-OpenAI's o3 Pro, Gemini 3 Google's Grok 4 xAI's ו-Claude Opus Anthropic's.

4.5 — באמצעות 4,326 השאלות העובדתיות של SimpleQA. הם מפורשים בכך שמקרה המבחן הוא המחשה, "לא הערכה מבוקרת בין מודלים" (הגדרות ברירת מחדל, בלי כוונון, בלי נרמול עלות).

מה הם מצאו

- **טרום-אימון כופה שגיאה מסוימת.** השיעור שבו מודל פולט שקרי-ביטחון חסום מלמטה על ידי בערך פעמיים שיעור השגיאה של מסווג ה"האם משפט זה תקף?" הטוב ביותר שנבנה ממנו. עבור עובדות ללא דפוס נלמד, הרצפה הזאת היא לפחות שיעור החד-פעמיים — חלקן של העובדות שמופיעות בדיוק פעם אחת באימון. חלק מן ההזיה בלתי נמנע גם עם נתונים נקיים לחלוטין.
- **הדירוג מתגמל ניחוש — ממש.** תחת ניקוד רגיל של נכון/לא נכון, לעולם לא להימנע היא האסטרטגיה האופטימלית, והסקר של המחברים מוצא שרוב גדול של מדדים פופולריים מדרג "אני לא יודע" פשוט כשגוי. דוגמה חיה מן הצד שלהם: במבחן SimpleQA, דיוק גולמי מעט מעדיף את OpenAI's o4-mini — שעונה כמעט על הכל וטועה ביותר משלושה רבעים מן הזמן — על GPT-5-mini, שעושה הרבה פחות טעויות כי הוא נמנע כשאינו בטוח. המודל הפזיז יותר נראה טוב יותר בלוח הדירוג.
- **מחוונים פתוחים הופכים את התמריץ (במקרה המבחן שלהם).** הם בודקים הקלה פשוטה בהזיה (לתת למודל לענות פעמיים ולהימנע אם שתי התשובות חלוקות). תחת דיוק רגיל, ההקלה חותכת שגיאות אבל גם חותכת דיוק — ולכן המדד מרתיע מלהשתמש בה. תחת מחוונים פתוחים, אותה הקלה יוצאת עדיפה בכל ארבעת המודלים לאורך טווח עונשים; ו-GPT-5-mini — שדיוק גולמי העניש כי נמנע כשהיה לא בטוח — יוצא לפני o4-mini ברגע שהניקוד נאמר בגלוי (n = 4,326 שאלות לכל מודל).

מה זה כנראה אומר

צמצום הזיה הוא לרוב לא עניין של המצאת עוד מבחנים ייעודיים להזיה. הוא עניין של שינוי האופן שבו המדדים המרכזיים מדרגים אי-ודאות, כך שהודאה ב"אני לא יודע" לא תיענש עוד. עד שלוח הדירוג ישתנה, צמצום הזיה ימשיך לעלות למודלים בנקודות דיוק ולכן ימשיך להיות מרוסן — ולכן המחברים ממסגרים את הבעיה כ"סוציו-טכנית": חלקה מדד טוב יותר, וחלקה לגרום ללוחות הדירוג המשפיעים לאמץ אותו.

מה זה לא מוכיח

- זה **לא** מראה שמחוונים פתוחים מתקנים הזיה בטבע. הניסוי התומך הוא מקרה מבחן קטן ו**בלתי מבוקר** במכוון — ארבעה מודלים בהגדרות ברירת מחדל, הקלה אחת שנבחרה, מבחן שאלות עובדתי אחד — שנועד להדגים את היפוך התמריץ, לא לדרג מודלים או להוכיח יעילות כללית.
- זה **לא** טוען שהדירוג הוא הגורם היחיד. שגיאות בנתוני האימון, בעיות קשות באמת, והנחיות לא מוכרות נשארות מקורות נפרדים.
- זה **לא** תומך בקו הפופולרי שהזיות בלתי נמנעות. המחברים טוענים להפך: מערכת שתענה רק על שאלות ניתנות לבדיקה ותאמר אחרת "אני לא יודע" לא תזיה לעולם.
- זה **לא** מעלים את רצפת טרום-האימון — הוא מסביר ומגביל אותה, והגבול נוגע לשגיאות עובדתיות בטוחות, לא לכל התנהגות מודל.
- זה **לא** מראה שמחוונים פתוחים מספיקים לבדם. הם משנים את מה שהערכה מתגמלת; הם אינם תחליף לשליפה,

שימוש בכלים או מודלים מכוילים טוב יותר.

עד כמה הראיות חזקות?

- הליבה היא **מתמטיקה** — חסמים תחתונים פורמליים, לא מדידות. כטיעון תאורטי הוא עומד היטב בתנאיו.
- הוא נשען על **מודלים מפושטים במכוון** של הבעיה; המחברים עצמם מציינים את "השלישייה הכוזבת" של טיפול בכל תגובה כנכונה, שגויה או "אני לא יודע", ואת מצב "העובדות השרירותיות" האידאלי שבו משתמשים לחסם הנקי ביותר.
- סקר המדדים הוא **דגימה קטנה ומכוונת** — עשר הערכות משפיעות, לא ביקורת ממצה.
- מקרה המבחן **אמיתי אך מוגבל**: ארבעה מודלי חזית, הקלה יחידה, SimpleQA בלבד, הגדרות ברירת מחדל, ובמפורש "לא הערכה מבוקרת". הוא הוכחת היתכנות לטיעון התמריץ, לא תוצאת מדד.
- כדאי לציין את נקודת המבט: שלושה מארבעת המחברים מועסקים או הועסקו ב-OpenAI, והמאמר טוען שהתחום צריך לשנות את אופן הערכת המודלים. זו עמדה מנומקת היטב מצד בעל עניין, לא ביקורת חיצונית ניטרלית — יש לשקול אותה, לא לפסול אותה. (לזכותו, המאמר מפנה את הביקורת גם אל מודליו שלו, o4-mini ו-GPT-5-mini, באותה קלות שבה הוא מפנה אותה לאחרים.)

למה זה חשוב

זה ממסגר מחדש בעיה עמוסת הייפ. "הזיה" נוטה להימכר או כפגם מפחיד או כחומה בלתי ניתנת להזזה; המאמר הזה הופך אותה לרגילה וחלקית מעשה ידינו — רצפה סטטיסטית שאפשר באמת להבין, היושבת על גבי תמריץ שבחרנו. הלקח הרחב שקט ושימושי יותר: התקדמות נוספת באמינות עשויה להיות תלויה במה אנחנו מודדים לא פחות מאשר במה אנחנו בונים.

סיכום נקי

תשובות שגויות ובטוחות ממודלי שפה מגיעות משני מקורות. הראשון סטטיסטי: כשלעובדה אין דפוס ללמוד, מודל שאומן לחקות שפה יטעה לפעמים, ואת הרצפה הזאת אפשר לאמוד (למשל, ממספר העובדות שמופיעות פעם אחת בלבד באימון). השני הוא תמריצים: כמעט כל מדד שעליו מדרגים מודלים נותן ל"אני לא יודע" אותו ציון כמו לתשובה שגויה, ולכן ניחוש תמיד מנצח — עד כדי כך שמודל שטועה בשלושה רבעים מן הזמן יכול לעקוף מודל ישר יותר שנמנע. הצעת המחברים אינה עוד מבחן הזיה אלא "מחוונים פתוחים": לכתוב את כללי הניקוד בתוך השאלה. במקרה מבחן על ארבעה מודלי חזית, זה הופך את התמריץ כך ששיטה מפחיתת הזיה מתוגמלת במקום להיענש. זו עבודת תאוריה-פלוס-סקר עם ניסוי קטן ולא מבוקר במפורש, שעברה ביקורת עמיתים ב-*Nature*; התיקון מבטיח אך עדיין לא הוכח בקנה מידה, והזיות נטענות כאן כלא מסתוריות ולא בלתי נמנעות לחלוטין.

בדיקה בלי לייפות

מה המאמר מראה: חסם תחתון מתמטי שלפיו חלק מן ההזיה נכפה בזמן טרום-האימון (לפחות "שיעור החד-פעמיים" עבור עובדות חסרות דפוס); סקר שמוצא שרוב המדדים המובילים אינם נותנים ל"אני לא יודע" קרדיט; ומקרה מבחן בארבעה מודלים שבו כתיבת הניקוד בשאלה ("מחוונים פתוחים") גורמת לשיטה מפחיתת הזיה לנצח, במקום שבו דיוק רגיל העניש אותה.

מה סביר אך לא מוכח: שמחוונים פתוחים, אם ייכנסו למדדים המרכזיים, יפחיתו משמעותית הזיה במודלים פרוסים. הניסוי התומך קטן ובלתי מבוקר במפורש.

מה זה לא מראה: שהזיות בלתי נמנעות (הוא טוען להפך); שהדירוג הוא הגורם היחיד; שאפשר לבטל הזיה לגמרי; או שמקרה המבחן מדרג את ארבעת המודלים זה מול זה.

המגבלות העיקריות: מודל מפושט במכוון של נכון/לא נכון/"אני לא יודע" (המחברים קוראים לו "שלישייה כוזבת"); סקר קטן ומכוון של מדדים (עשר הערכות); מקרה מבחן בלתי מבוקר (ארבעה מודלים, הקלה אחת, מבחן אחד, הגדרות ברירת מחדל); וטיעון בהובלת OpenAI על הדרך שבה התחום צריך להעריך מודלים.

כמה ביטחון צריך להיות לקורא כללי? גבוה בכך שהזיה אינה מסתורית ואינה בלתי נמנעת לחלוטין, ושמדדים מרכזיים כיום מתגמלים ניחוש. בינוני בכך שהתיקון המוצע עוזר — יש לו כעת הוכחת היתכנות אמיתית, אבל עדיין לא הדגמה שהוא עובד באופן רחב ובקנה מידה.

מקורות

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