The Clean Paper

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

Γιατί τα γλωσσικά μοντέλα κάνουν hallucinations — και γιατί ο τρόπος που τα βαθμολογούμε τις διατηρεί

Οι ψευδείς απαντήσεις με αυτοπεποίθηση που ονομάζουμε «hallucinations» δεν είναι μυστηριώδες σφάλμα: ορισμένες είναι στατιστικό υποπροϊόν της εκπαίδευσης και επιμένουν επειδή τα κύρια benchmarks ανταμείβουν μια μαντεψιά με αυτοπεποίθηση περισσότερο από ένα ειλικρινές «δεν ξέρω». Μια μελέτη περίπτωσης σε τέσσερα frontier models δείχνει ότι η δήλωση των κανόνων βαθμολόγησης μέσα στο prompt («open rubrics») αντιστρέφει αυτό το κίνητρο.

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

Γιατί τα μοντέλα μαντεύουν — και γιατί τα μάθαμε να το κάνουν

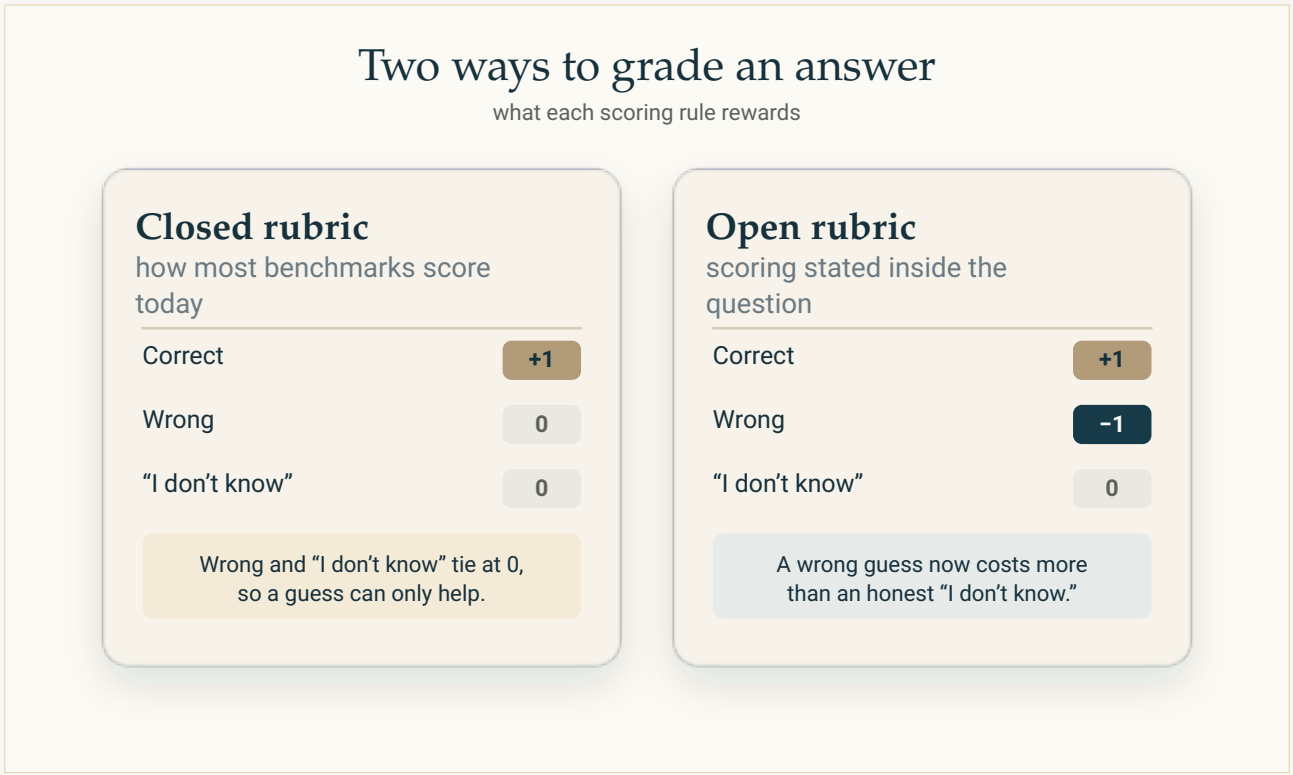
Ρωτήστε ένα μεγάλο γλωσσικό μοντέλο για τα γενέθλια ενός αγνώστου και μπορεί να απαντήσει «7 Μαρτίου» με τη σταθερή αυτοπεποίθηση κάποιου που το διαβάζει από μια κάρτα — και να κάνει λάθος, τρεις φορές στη σειρά, με τρεις διαφορετικές ημερομηνίες. Οι συγγραφείς δίνουν ακριβώς τέτοιο παράδειγμα: κορυφαία μοντέλα που ρωτήθηκαν μια απλή πραγματολογική ερώτηση — τα γενέθλια ενός ανθρώπου ή τι σημαίνει ένα άγνωστο ακρωνύμιο — επινοούν το καθένα με αυτοπεποίθηση διαφορετική απάντηση, και καμία δεν είναι σωστή. Η βιομηχανία το ονομάζει ***hallucination***, όρος που το κάνει να ακούγεται σαν σφάλμα αντίληψης. Η πρώτη κίνηση της εργασίας είναι να αφαιρέσει το μυστήριο.

Ξεκινήστε από το πώς κατασκευάζεται ένα μοντέλο. Στο πρώτο και μεγαλύτερο στάδιο εκπαίδευσης μαθαίνει, ουσιαστικά, πώς μοιάζει η ρέουσα γλώσσα διαβάζοντας τεράστιες ποσότητες κειμένου. Τώρα πάρτε ένα γεγονός χωρίς μοτίβο πίσω του — τα γενέθλια ενός συγκεκριμένου ανθρώπου. Αν η ημερομηνία εμφανίστηκε μία φορά στο κείμενο εκπαίδευσης, ή ποτέ, δεν υπάρχει κάποιο μοτίβο από το οποίο να πιαστεί ένας μηχανισμός μάθησης μοτίβων: από την οπτική του μοντέλου, η απάντηση είναι αυθαίρετη. Οι συγγραφείς το διατυπώνουν με ακρίβεια δανειζόμενοι μια παλιά ιδέα του Alan Turing από άλλο πρόβλημα: αν ένα στα πέντε γενέθλια εμφανίζεται μόνο μία φορά στα δεδομένα, ένα μοντέλο θα πρέπει να αναμένεται να κάνει λάθος τουλάχιστον σε ένα στα πέντε — όχι επειδή είναι χαλασμένο, αλλά επειδή δεν υπήρχε αρκετό υλικό για να μάθει. (Με την ίδια λογική, τα μοντέλα σχεδόν ποτέ δεν κάνουν λάθος στην πρωτεύουσα μιας χώρας: αυτές οι πληροφορίες εμφανίζονται διαρκώς.) Οι συγγραφείς υποστηρίζουν προσεκτικά ότι το να ξεχωρίζεις μια αληθινή δήλωση από μια εύλογη ψευδή είναι από μόνο του δύσκολο πρόβλημα, και ότι το να παράγεις *μόνο* αληθείς δηλώσεις είναι τουλάχιστον εξίσου δύσκολο. Υπάρχει ένα κατώφλι σφάλματος ενσωματωμένο στη διαδικασία.

> **Η ιδέα από κάτω: πώς μετράς όσα δεν έχεις ακόμη δει**
>
> Αυτό βασίζεται σε μια πραγματικά έξυπνη ιδέα — παλαιότερη από τα γλωσσικά μοντέλα και άξια να τη γνωρίσουμε σωστά.
> **Ξεκινήστε με μια σακούλα χρωματιστές μπάλες.** Δεν ξέρετε πόσα χρώματα περιέχει. Τραβάτε 100, μία κάθε φορά, και μετράτε: κόκκινες 40, μπλε 25, πράσινες 15, κίτρινες 5, μοβ 3, πορτοκαλί 2 — και μετά **δέκα διαφορετικά χρώματα που εμφανίζονται ακριβώς μία φορά το καθένα.**
> Τώρα το ερώτημα που αντιμετώπισε πράγματι ο Turing, σε εντελώς διαφορετικό πρόβλημα: *ποια είναι η πιθανότητα η επόμενη μπάλα να έχει χρώμα που δεν έχεις δει ποτέ;* Δεν μπορείς να μετρήσεις όσα δεν τράβηξες ποτέ — μπορείς όμως να μετρήσεις τα χρώματα που είδες ακριβώς μία φορά, τα «singletons». Το τέχνασμα, που ονομάζεται **εκτίμηση Good–Turing**, είναι ότι το ποσοστό των δειγμάτων σου που είναι singletons εκτιμά την πιθανότητα που παραμένει κρυμμένη στα χρώματα που δεν έχεις ακόμη δει. Δέκα από τις εκατό μπάλες ήταν χρώματα που εμφανίστηκαν μόνο μία φορά, άρα η

> πιθανότητα η επόμενη μπάλα να έχει εντελώς νέο χρώμα είναι περίπου **10 / 100 = 10%.** Αυτά τα χρώματα που εμφανίστηκαν μία φορά δεν είναι *λάθη*. Είναι μέτρηση της ίδιας σου της άγνοιας: πολλά χρώματα που εμφανίζονται μία μόνο φορά είναι ο τρόπος με τον οποίο το δείγμα σου λέει ότι ο κόσμος περιέχει κι άλλα που απλώς δεν έχεις ακόμη τραβήξει.
> **Τώρα αντικαταστήστε τα χρώματα με ημερομηνίες γενεθλίων** και τη σακούλα με το κείμενο εκπαίδευσης του μοντέλου. Ας πούμε ότι, μεταξύ των γενεθλίων που είδε, **ένα στα πέντε εμφανίζεται ακριβώς μία φορά.** Το ίδιο τέχνασμα: περίπου το ένα πέμπτο της πιθανότητας βρίσκεται σε γενέθλια που το μοντέλο ουσιαστικά δεν έχει δει — και μια ημερομηνία γέννησης δεν έχει μοτίβο στο οποίο να βασιστείς (δεν μπορείς να την *υπολογίσεις*). Έτσι, μια ημερομηνία που εμφανίστηκε μία φορά ή ποτέ είναι κάτι που το μοντέλο δεν μπορεί να σταθμίσει σωστά, και θα κάνει λάθος περίπου **σε μία στις πέντε.**
> Καμία εξυπνάδα δεν το διορθώνει: δεν υπήρχε αρκετή πληροφορία για να μάθει.
> Αυτό είναι όλο το επιχείρημα σε μικρογραφία: το ποσοστό των singletons μετρά πόσο από τον κόσμο είναι αδύνατο να μάθεις *από αυτά τα δεδομένα*, και αυτό γίνεται ένα **κατώτατο όριο** στα σφάλματα. Εξηγεί επίσης γιατί ένα μοντέλο σχεδόν ποτέ δεν χάνει μια πρωτεύουσα — το *Paris* εμφανίζεται αμέτρητες φορές, το ποσοστό singleton είναι σχεδόν μηδενικό, άρα υπάρχει άφθονο υλικό για μάθηση.

Αυτό εξηγεί από πού προέρχονται οι hallucinations. Δεν εξηγεί γιατί *επιβιώνουν* — γιατί τα μοντέλα, μετά από όλη την πρόσθετη εκπαίδευση που υποτίθεται ότι τα κάνει χρήσιμα και ειλικρινή, συνεχίζουν να μπλοφάρουν αντί να παραδέχονται αβεβαιότητα. Εδώ η αναλογία της εργασίας είναι σχεδόν ενοχλητικά εύστοχη. Φανταστείτε έναν μαθητή σε εξέταση που δεν ξέρει μια απάντηση. Αν το κενό παίρνει μηδέν και μια μαντεψιά μπορεί να πάρει έναν βαθμό, η κίνηση που μεγιστοποιεί τη βαθμολογία είναι να μαντέψει — με αυτοπεποίθηση, συγκεκριμένα, ποτέ «δεν είμαι σίγουρος». Οι μαθητές το μαθαίνουν. Το ίδιο και τα μοντέλα — επειδή τα βαθμολογούμε με τον ίδιο τρόπο. Οι συγγραφείς εξέτασαν τα benchmarks στα οποία ανταγωνίζεται πραγματικά το πεδίο, τα leaderboards στα οποία ρυθμίζονται τα μοντέλα, και βρήκαν ότι σχεδόν όλα δίνουν στο «δεν ξέρω» ακριβώς την ίδια βαθμολογία με μια λάθος απάντηση: μηδέν. Με αυτόν τον κανόνα, ένα μοντέλο που πάντα μαντεύει θα ξεπεράσει ένα κατά τα άλλα ίδιο μοντέλο που δηλώνει ειλικρινά την αβεβαιότητά του. Με αρκετά κυριολεκτική έννοια, τα βαθμολογούμε ώστε να συμπεριφέρονται έτσι.



Τα benchmarks μπορούν να κάνουν τη μαντεψιά ορθολογική. Με τη βαθμολόγηση που χρησιμοποιούν τα περισσότερα benchmarks (αριστερά), μια λάθος απάντηση και ένα ειλικρινές «δεν ξέρω» παίρνουν και τα δύο μηδέν — άρα η μαντεψιά μόνο να βοηθήσει μπορεί. Αν οι κανόνες δηλώνονται μέσα στην ερώτηση, με ποινή για λάθος απάντηση (δεξιά), η αποχή όταν υπάρχει αβεβαιότητα μπορεί να γίνει η καλύτερη επιλογή. Αυτό αλλάζει τι αναμείβει το τεστ· δεν λύνει από μόνο του τις hallucinations.

Original diagram — The Clean Paper CC BY 4.0

Αυτό είναι το κομμάτι που αξίζει να κρατήσουμε, επειδή πηγαίνει κόντρα στον συνηθισμένο τίτλο. Οι hallucinations συχνά παρουσιάζονται ως αναπόφευκτο, σχεδόν μυστικιστικό όριο της τεχνολογίας. Η εργασία αμφισβητεί και τα δύο. Το κατώτατο όριο του pretraining δεν είναι μυστήριο — είναι συνηθισμένο στατιστικό σφάλμα, από το είδος που η μηχανική μάθηση κατανοεί εδώ και δεκαετίες. Και η επιμονή τους δεν είναι αναπόφευκτη — είναι εν μέρει κίνητρο που εμείς δημιουργήσαμε και μπορούμε να αλλάξουμε. Ένα σύστημα που απλώς αρνείται να απαντήσει όταν δεν είναι σίγουρο δεν θα hallucinate καθόλου· ο λόγος που τα αναπτυγμένα μοντέλα δεν συμπεριφέρονται έτσι είναι ότι οι πίνακες βαθμολογίας μας τιμωρούν την άρνηση.

Τι έκαναν οι συγγραφείς

Η εργασία έχει τρία μέρη. Πρώτον, ένα μαθηματικό επιχείρημα ότι κάποιο ποσοστό hallucination επιβάλλεται στατιστικά κατά το pretraining, δείχνοντας ότι το «παράγε μόνο έγκυρο κείμενο» είναι τουλάχιστον τόσο δύσκολο όσο ένα δυαδικό πρόβλημα ταξινόμησης «είναι αυτή η δήλωση έγκυρη;». Δεύτερον, ένα επιχείρημα — υποστηριζόμενο από έρευνα δέκα σημαντικών benchmarks — ότι οι συνήθεις μετρικές ακρίβειας αναμείβουν τη μαντεψιά αντί

για την αποχή. Τρίτον, μια προτεινόμενη λύση και μια μελέτη περίπτωσης που τη δοκιμάζει: αξιολογήσεις με **open rubrics**, όπου η βαθμολόγηση δηλώνεται μέσα στην ίδια την ερώτηση (για παράδειγμα, «σωστή απάντηση +1, λάθος −1, άρα απέχεις αν είσαι λιγότερο από 50% σίγουρος»), ώστε το μοντέλο να γνωρίζει πότε αναμείβεται η ειλικρίνεια. Το δοκιμάζουν σε τέσσερα frontier models — Gemini 3 Pro της Google, GPT-5 της OpenAI, Grok 4 της xAI και Claude Opus 4.5 της Anthropic — χρησιμοποιώντας τις 4.326 πραγματολογικές ερωτήσεις του SimpleQA. Είναι σαφείς ότι η μελέτη περίπτωσης είναι ενδεικτική, «όχι ελεγχόμενη αξιολόγηση μεταξύ μοντέλων» (προεπιλεγμένες ρυθμίσεις, χωρίς tuning, χωρίς εξομάλυνση κόστους).

Τι βρήκαν

- **Το pretraining επιβάλλει κάποιο σφάλμα.** Ο ρυθμός με τον οποίο ένα μοντέλο παράγει ψευδείς δηλώσεις με αυτοπεποίθηση έχει κατώτατο όριο που συνδέεται με (περίπου το διπλάσιο) σφάλμα του καλύτερου ταξινομητή «είναι αυτή η δήλωση έγκυρη;» που μπορεί να κατασκευαστεί από αυτό. Για γεγονότα χωρίς μαθήσιμο μοτίβο, αυτό το όριο είναι τουλάχιστον το *singleton rate* — το ποσοστό των γεγονότων που εμφανίζονται ακριβώς μία φορά στην εκπαίδευση. Κάποιο hallucination είναι αναπόφευκτο ακόμη και με απολύτως καθαρά δεδομένα.
- **Η βαθμολόγηση ανταμείβει συγκεκριμένα τη μαντεψιά.** Με τη συνήθη βαθμολόγηση σωστό/λάθος, το να μην απέχεις ποτέ είναι η βέλτιστη στρατηγική, και η έρευνα των συγγραφέων βρίσκει ότι η μεγάλη πλειονότητα δημοφιλών benchmarks βαθμολογεί το «δεν ξέρω» απλώς ως λάθος. Ένα χαρακτηριστικό παράδειγμα από τα ίδια τους τα μοντέλα: στο SimpleQA, η ακατέργαστη ακρίβεια ευνοεί ελαφρά το o4-mini της OpenAI — που απαντά σχεδόν σε όλα και κάνει λάθος πάνω από τα τρία τέταρτα του χρόνου — έναντι του GPT-5-mini, που κάνει πολύ λιγότερα λάθη επειδή απέχει όταν δεν είναι σίγουρο. Το πιο ριψοκίνδυνο μοντέλο φαίνεται καλύτερο στον πίνακα.
- **Τα open rubrics αντιστρέφουν το κίνητρο στη μελέτη περίπτωσης.** Δοκιμάζουν μια απλή τεχνική μείωσης hallucination (το μοντέλο απαντά δύο φορές και απέχει αν οι δύο απαντήσεις διαφωνούν). Με τη συνήθη ακρίβεια, η τεχνική μειώνει τα λάθη αλλά μειώνει και την ακρίβεια — άρα η μετρική αποθαρρύνει την υιοθέτησή της. Με open rubrics, η ίδια τεχνική βγαίνει καλύτερη και στα τέσσερα μοντέλα για διάφορα επίπεδα ποινής· και το GPT-5-mini — που η ακατέργαστη ακρίβεια τιμωρούσε επειδή απείχε όταν δεν ήταν σίγουρο — ξεπερνά το o4-mini όταν η βαθμολόγηση δηλώνεται ανοιχτά (n = 4.326 ερωτήσεις ανά μοντέλο).

Τι πιθανότατα σημαίνει αυτό

Η μείωση των hallucinations δεν είναι κυρίως ζήτημα επινόησης περισσότερων ειδικών τεστ hallucination. Είναι ζήτημα αλλαγής του τρόπου με τον οποίο τα κύρια benchmarks βαθμολογούν την αβεβαιότητα, ώστε το ειλικρινές «δεν ξέρω» να μην τιμωρείται πλέον.

Όσο ο πίνακας βαθμολογίας δεν αλλάζει, η μείωση hallucination θα συνεχίζει να κοστίζει μονάδες ακρίβειας και άρα να αποθαρρύνεται — γι' αυτό οι συγγραφείς χαρακτηρίζουν το πρόβλημα «κοινωνικοτεχνικό»: εν μέρει καλύτερη μετρική, εν μέρει πειθώ ώστε τα σημαντικά leaderboards να την υιοθετήσουν.

Τι *δεν αποδεικνύει*

- **Δεν** δείχνει ότι τα open rubrics διορθώνουν τις hallucinations στην πραγματική χρήση. Το υποστηρικτικό πείραμα είναι μικρή, σκόπιμα **μη ελεγχόμενη** μελέτη περίπτωσης — τέσσερα μοντέλα με προεπιλεγμένες ρυθμίσεις, μία επιλεγμένη τεχνική, ένα τεστ πραγματολογικών ερωτήσεων — σχεδιασμένη να δείξει *την αντιστροφή του κινήτρου*, όχι να κατατάξει μοντέλα ή να αποδείξει γενική αποτελεσματικότητα.
- **Δεν** υποστηρίζει ότι η βαθμολόγηση είναι η *μοναδική* αιτία. Λάθη στα δεδομένα εκπαίδευσης, πραγματικά δύσκολα προβλήματα και άγνωστα prompts παραμένουν ξεχωριστές πηγές.
- **Δεν** στηρίζει τη δημοφιλή άποψη ότι οι hallucinations είναι αναπόφευκτες. Οι συγγραφείς υποστηρίζουν το αντίθετο: ένα σύστημα που απαντά μόνο σε ερωτήματα τα οποία μπορεί να ελέγξει και αλλιώς λέει «δεν ξέρω» δεν θα hallucinate.
- **Δεν** εξαφανίζει το κατώτατο όριο του pretraining — το εξηγεί και το οριοθετεί, και το όριο αφορά ψευδείς πραγματολογικές δηλώσεις με αυτοπεποίθηση, όχι κάθε συμπεριφορά του μοντέλου.
- **Δεν** δείχνει ότι τα open rubrics είναι *επαρκή* από μόνα τους. Αλλάζουν τι ανταμείβει μια αξιολόγηση· δεν αντικαθιστούν retrieval, χρήση εργαλείων ή καλύτερα βαθμονομημένα μοντέλα.

Πόσο ισχυρά είναι τα στοιχεία;

- Ο πυρήνας είναι μα**θημα**τ**ικά** — τυπικά κατώτατα όρια, όχι μετρήσεις. Ως θεωρητικό επιχείρημα στέκει με τους δικούς του όρους.
- Βασίζεται σε σ**κόπι**μα απ**λουστευμέν**α μ**οντέλα** του προβλήματος· οι ίδιοι οι συγγραφείς επισημαίνουν την «ψευδή τριχοτόμηση» του να θεωρείς κάθε απόκριση σωστή, λάθος ή «δεν ξέρω», και το εξιδανικευμένο πλαίσιο «αυθαίρετων γεγονότων» που χρησιμοποιείται για το καθαρότερο όριο.
- Η έρευνα benchmarks είναι **μικρό, επιμελημένο δείγμα** — δέκα σημαντικές αξιολογήσεις, όχι εξαντλητικός έλεγχος.
- Η μελέτη περίπτωσης είναι **πραγματική αλλά περιορισμένη**: τέσσερα frontier models, μία τεχνική, μόνο SimpleQA, προεπιλεγμένες ρυθμίσεις, ρητά «όχι ελεγχόμενη αξιολόγηση». Είναι proof of concept για το επιχείρημα των κινήτρων, όχι αποτέλεσμα benchmark.
- Αξίζει να αναφερθεί η οπτική γωνία: τρεις από τους τέσσερις συγγραφείς εργάζονται ή εργάστηκαν στην **OpenAI**, και η εργασία υποστηρίζει ότι το πεδίο πρέπει να αλλάξει τον τρόπο αξιολόγησης των μοντέλων. Είναι μια καλά τεκμηριωμένη θέση από ενδιαφερόμενο μέρος, όχι ουδέτερη εξωτερική ανασκόπηση — κάτι που πρέπει να ζυγιστεί, όχι να

απορριφθεί. (Προς τιμήν της, η εργασία στρέφει την κριτική στα δικά της μοντέλα, o4-mini και GPT-5-mini, εξίσου πρόθυμα όσο και στα άλλα.)

Γιατί έχει σημασία

Αναπλαισιώνει ένα υπερβολικά διαφημισμένο πρόβλημα. Το «hallucination» συνήθως παρουσιάζεται είτε ως μυστηριώδες ελάττωμα είτε ως ακίνητος τοίχος· η εργασία το κάνει πιο καθημερινό και εν μέρει αυτοπροκαλούμενο — ένα στατιστικό κατώτατο όριο που μπορούμε να κατανοήσουμε, πάνω από ένα κίνητρο που επιλέξαμε εμείς. Το ευρύτερο μάθημα είναι πιο ήσυχο και πιο χρήσιμο: η περαιτέρω πρόοδος στην αξιοπιστία ίσως εξαρτάται τόσο από *το τι μετράμε* όσο και από το τι κατασκευάζουμε.

Καθαρή σύνοψη

Οι ψευδείς απαντήσεις με αυτοπεποίθηση από γλωσσικά μοντέλα προέρχονται από δύο σημεία. Το πρώτο είναι στατιστικό: όταν ένα γεγονός δεν έχει μοτίβο που να μπορεί να μάθει, ένα μοντέλο εκπαιδευμένο να μιμείται τη γλώσσα θα κάνει μερικές φορές λάθος, και αυτό το κατώτατο όριο μπορεί να εκτιμηθεί (για παράδειγμα από το πόσα γεγονότα εμφανίζονται μόνο μία φορά στην εκπαίδευση). Το δεύτερο είναι τα κίνητρα: σχεδόν κάθε benchmark στο οποίο κατατάσσονται τα μοντέλα βαθμολογεί το «δεν ξέρω» το ίδιο με μια λάθος απάντηση, άρα η μαντεψιά πάντα συμφέρει — σε σημείο που ένα μοντέλο που κάνει λάθος τα τρία τέταρτα του χρόνου μπορεί να ξεπερνά ένα πιο ειλικρινές μοντέλο που απέχει. Η πρόταση των συγγραφέων δεν είναι ακόμη ένα τεστ hallucination αλλά τα «open rubrics»: δήλωσε τη βαθμολόγηση μέσα στην ερώτηση. Σε μελέτη περίπτωσης τεσσάρων frontier models, αυτό αντιστρέφει το κίνητρο ώστε μια τεχνική μείωσης hallucination να ανταμείβεται αντί να τιμωρείται. Είναι εργασία θεωρίας συν έρευνας benchmarks, με μικρό και ρητά μη ελεγχόμενο πείραμα, αξιολογημένη στο ***Nature***· η λύση είναι πολλά υποσχόμενη αλλά δεν έχει ακόμη αποδειχθεί σε κλίμακα, και οι hallucinations υποστηρίζεται ότι δεν είναι ούτε μυστηριώδεις ούτε αυστηρά αναπόφευκτες.

Έλεγχος χωρίς υπερβολές

Τι δείχνει η εργασία: Ένα μαθηματικό κατώτατο όριο σύμφωνα με το οποίο κάποιο hallucination επιβάλλεται κατά το pretraining (τουλάχιστον το «singleton rate» για γεγονότα χωρίς μοτίβο)· μια έρευνα που βρίσκει ότι τα περισσότερα κορυφαία benchmarks δεν δίνουν καμία πίστωση στο «δεν ξέρω»· και μια μελέτη περίπτωσης τεσσάρων μοντέλων όπου η δήλωση της βαθμολόγησης στο prompt («open rubrics») κάνει μια τεχνική μείωσης hallucination να κερδίζει εκεί όπου η απλή ακρίβεια την τιμωρούσε.

Τι είναι εύλογο αλλά όχι αποδεδειγμένο: Ότι τα open rubrics, αν ενσωματωθούν στα βασικά benchmarks, θα μειώσουν ουσιαστικά τις hallucinations στα αναπτυγμένα μοντέλα. Το υποστηρικτικό πείραμα είναι μικρό και ρητά μη ελεγχόμενο.

Τι δεν δείχνει: Ότι οι hallucinations είναι αναπόφευκτες (υποστηρίζει το αντίθετο)· ότι η βαθμολόγηση είναι η μόνη αιτία· ότι οι hallucinations μπορούν να εξαλειφθούν πλήρως· ότι η μελέτη περίπτωσης κατατάσσει τα τέσσερα μοντέλα μεταξύ τους.

Κύριοι περιορισμοί: Σκόπιμα απλουστευμένο μοντέλο σωστό/λάθος/«δεν ξέρω» (οι συγγραφείς το ονομάζουν «ψευδή τριχοτόμηση»)· μικρή επιμελημένη έρευνα benchmarks (δέκα αξιολογήσεις)· μη ελεγχόμενη μελέτη περίπτωσης (τέσσερα μοντέλα, μία τεχνική, ένα τεστ, προεπιλεγμένες ρυθμίσεις)· και ένα επιχείρημα με επικεφαλής ανθρώπους της OpenAI για το πώς πρέπει το πεδίο να αξιολογεί τα μοντέλα.

Πόση εμπιστοσύνη πρέπει να έχει ένας γενικός αναγνώστης; Υψηλή ότι το hallucination δεν είναι ούτε μυστηριώδες ούτε αυστηρά αναπόφευκτο, και ότι τα σημερινά βασικά benchmarks ανταμείβουν τη μαντεψιά. Μέτρια ότι η προτεινόμενη λύση βοηθά — υπάρχει πλέον πραγματικό proof of concept, αλλά όχι ακόμη επίδειξη ότι λειτουργεί ευρέως και σε μεγάλη κλίμακα.

Πηγές

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