The Clean Paper

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

Language model কেন hallucinate করে — এবং benchmark কেন guessing-কে reward করে

Paper statistical lower bound দিয়ে কছু factual error-এর inevitability ব্যাখ্যা করে এবং দখোয় mainstream accuracy benchmark "I don't know" ও wrong answer-কে একই score দলিে guessing rational হয়। Open-rubric case study incentive flip করে, কন্তিু broad hallucination cure এখনও demonstrated নয়।

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

মডলে কনে আন্দাজ করে, আর আমরা কনে তাদরে তা করতে শখিয়িছে

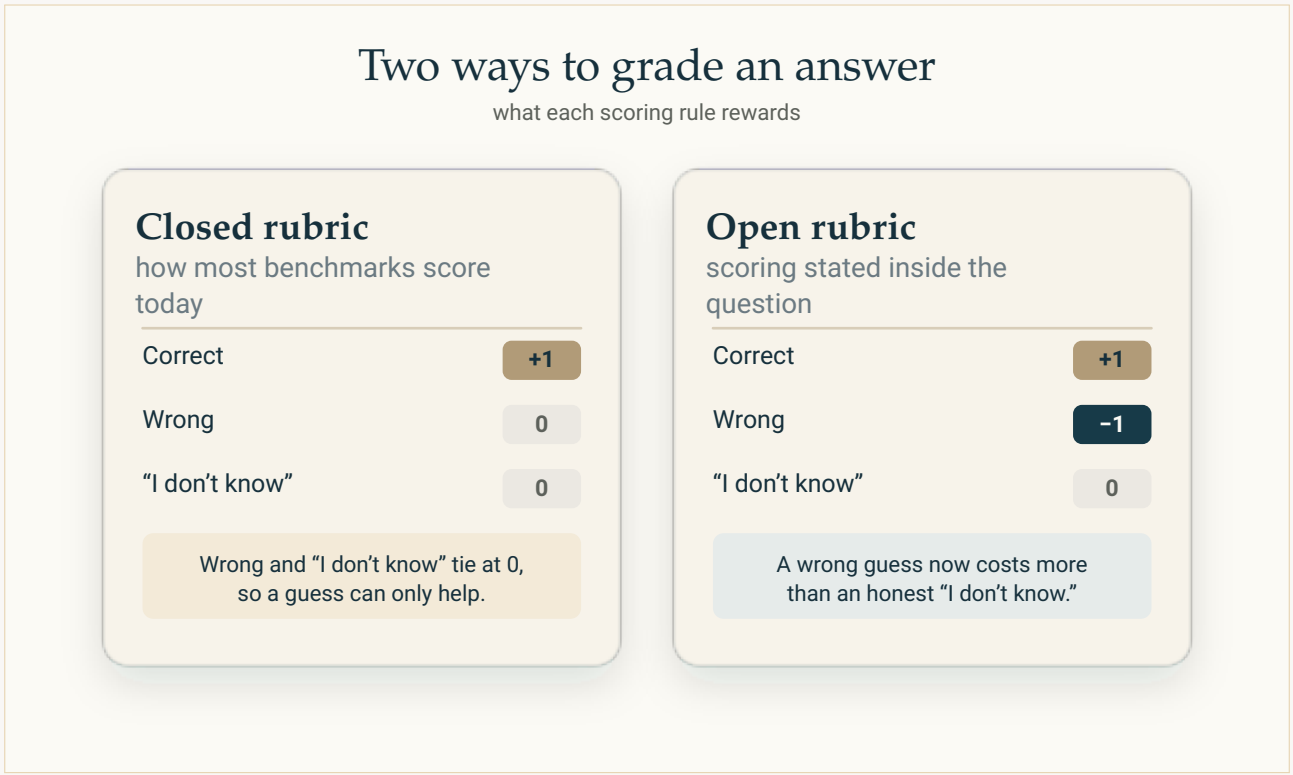
একটি বৃহৎ ভাষা মডলে-কে কোনো অচনো ব্যক্তরি জন্মদনি জজ্ঞিসে করুন। সে হয়তো এমন আত্মবশ্বিাসে বলবে, "৭ মার্চ", যনে কার্ড থকেে পড়ছে—কন্তিু উত্তরটি ভুল। আবার জজ্ঞিসে করলে অন্য তারখি, তৃতীয়বারে আরও একটি। গবষেণাপত্ররে লখেকরো ঠকি এমন উদাহরণ দনে: প্রধান ক্রমরে মডলেগুলোকে সাধারণ তথ্যভত্তিকি প্রশ্ন করা হলে—কোনো ব্যক্তরি জন্মতারখি, বা আড়াল করে সংক্ষপ্তি রূপরে পূর্ণরূপ—প্রতটিমিডলে নর্দিষ্টি ও আত্মবশ্বিাসী কন্তিু আলাদা ভুল উত্তর বানায়। শল্পিখাত এটাকে *হ্যালুসনিশেন* বলে, শব্দটি যনিে উপলব্ধরি অদ্ভুত ত্রুটিবি োঝায়। গবষেণাপত্ররে প্রথম কাজ হলো রহস্যরে আবরণ সরানো।

মডলে কীভাবে তরৈি হয় তা থকেে শুরু করা যাক। প্রথম ও সবচয়েে বড় প্রশক্ষিণ পর্যায়ে সে বপিুল পাঠ্য পড়ে fluent ভাষার ধরন শখেে। এখন এমন একটি তথ্য নন যার পছেনে শখোর মতো ধরন নইে—একজন নর্দিষ্টি মানুষরে জন্মতারখি। প্রশক্ষিণ পাঠ্যে তারখিটি একবার এসছে, বা একবারও আসনে, তাহলে ধরন শক্ষিণ-ব্যবস্থার ধরার মতো কছিু নইে। মডলেরে দৃষ্টতিে উত্তর-টি ইচ্ছামতো। লখেকরো Alan Turing-এর একটি পুরোনো পরসিংখ্যানগত ধারণা নয়িে এটকিে নর্দিষ্টি করনে: প্রশক্ষিণ-তথ্যে যদি পাঁচটরি একটি জন্মতারখি মাত্র একবার দখো যায়, তাহলে মডলেরে অন্তত প্রায় পাঁচটরি একটতিে ভুল হওয়া প্রত্যাশতি—মডলে "ভাঙা" বলে নয়, বরং শখোর মতো পর্যাপ্ত প্রমাণ ছলি না বলে। বপিরীতে দশেরে রাজধানী মডলে খুব কম ভুল করে, কারণ "Paris" বা অনুরূপ তথ্য তথ্যে বারবার আসে।

তারা আরও যুক্তদিনে যে কোনো সম্ভাব্য বর্তিসত্য কমিথ্যিা তা নর্ভিরযোগ্যভাবে classifying করাই কঠনি; তাই **শুধু সত্য বর্তিতরৈি করা** অন্তত সইে শ্রণেবিন্যিাস সমস্যার মতো কঠনি। অর্থাৎ কছিু ত্রুটরি পরসিংখ্যানগত নম্নিসীমা থকেইে যায়।

> **যে ধারণাটি নচিে কাজ করছে: যা এখনো দখেনেনি, তা কীভাবে গোনা যায়**
>
> এখানে একটি সুন্দর পুরোনো পরসিংখ্যানগত ধারণা কাজ করে।
> **রঙনি বলরে একটি থলি কল্পনা করুন।** কত রং আছে জাননে না। 100টি বল একে একে তুললনে: লাল 40, নীল 25, সবুজ 15, হলুদ 5, বগেুনি 3, কমলা 2—আর তারপর **10টি আলাদা রং, যগেুলোর প্রত্যকেটি মাত্র একবার করে** দখো গলে।
> প্রশ্নটি হলো: পররে বলটি এমন রঙরে হওয়ার সম্ভাবনা কত, যা আপনি এখনো একবারও দখেনেনি? যা দখেনেনি তা সরাসরি গোনা যায় না। কন্তিু **যে রংগুলো ঠকি একবার দখো গছে**, অর্থাৎ *একক*, সগেুলো গোনা যায়। **ভালো–Turing estimation**-এর মূল অন্তর্দৃষ্টি হলো: নমুনায় এককরে ভাগ, অদখো শ্রণেি-তে এখনো কত সম্ভাবনা ভর লুকয়িে আছে তার অনুমান দয়ে। 100 draw-এর মধ্যে 10টি একক হলে পররে draw একবোরে নতুন রঙরে হওয়ার সম্ভাবনা আনুমানকি 10 / 100 = 10%।
> এই একবার-seen রংগুলো mistake নয়। এগুলো আপনার অজ্ঞতার পরমিাপ: অনকে একক মানে নমুনা নজিইে বলছে, বাইরে আরও শ্রণেি আছে যা এখনো ঠকিমতো দখো হয়নি।
> **এখন রঙরে বদলে জন্মতারখি ধরুন**, আর bag-এর বদলে মডলেরে প্রশক্ষিণ পাঠ্য। দখো জন্মতারখিরে মধ্যে যদি **পাঁচটরি একটি মাত্র একবার** উপস্থতি হয়, তাহলে একই ববিচেনামূলক চন্তিা বলে বণ্টনরে উল্লখেযোগ্য অংশ মডলে কার্যত ভালো করে কখনো শখেনে। জন্মতারখিরে কোনো গণনাযোগ্য ধরন নইে—কারও জন্মতারখি "উৎপন্ন করা" করা যায় না। তাই একবার-seen বা অদখো তারখিে মডলেরে ভুলরে একটি নম্নিসীমা থাকে।
> এটাই যুক্তরি ক্ষুদ্র সংস্করণ: একক হার বলে *এই তথ্য থকেে* পৃথবিীর কতটা অংশ শখো কঠনি, এবং সটেইি ত্রুটরি নচিে একটি **নম্নিসীমা** তরৈি করে। Capital city-তে এই নম্নিসীমা খুব ছোট, কারণ সগেুলো বারবার দখো যায়।

এতে হ্যালুসনিশেন কোথা থকেে আসে তার একটি অংশ বোঝা গলে। কন্তিু পরে সহায়কতা ও সততা শখোনোর প্রশক্ষিণরে পরও মডলে কনে ভান করে উত্তর দয়ে, সটেি এখনো বাকি। এখানে গবষেণাপত্ররে উপমা খুব সরাসরি। পরীক্ষায় কোনো শক্ষিার্থী উত্তর জানে না। Blank রাখলে শূন্য, আন্দাজ করলে শূন্য বা একটি—তাহলে গ্রডে সর্বাধকি করার কৌশল হলো আন্দাজ করা। "আমি জানি না" বললে কোনো upside নইে। মডলেরে মানদণ্ডওে প্রায় একই প্রণোদনা। লখেকরো ক্ষত্তেররে প্রভাবশালী মানদণ্ডগুলো দখেনে এবং দখেনে অধকিাংশে "I don't know" ভুল উত্তররে মতোই শূন্য পায়। এই স্কোরিং নয়িমে অনশ্চিয়তা স্বীকার করা মডলেরে চয়েে সব সময় আন্দাজ করা মডলে বশেি স্কোর করতে পারে। এক অর্থে আমরা মূল্যায়ন দয়িইে মডলেকে ভান করে উত্তর দতিে শখিয়িছে।



মানদণ্ড আন্দাজ করাকে rational করে তুলতে পারে। সাধারণ স্কোরিং (বাম দিকে) ভুল উত্তর এবং সৎ “আমি জানি না”—দুটোই শূন্য, তাই আন্দাজ করলে কেবল লাভের সম্ভাবনা থাকে। প্রশ্নের মধ্যেই নিয়ম স্পষ্ট করে ভুল উত্তরে penalty দিলে (ডান দিকে), অনিশ্চয়তা বশে হলে উত্তর না দেওয়া ভালো কৌশল হতে পারে। এটি মূল্যায়ন কী পুরস্কার করে তা বদলায়; নিজে থেকে হ্যালুসিনেশনের পুরো সমাধান নয়।

Original diagram — The Clean Paper CC BY 4.0

এই অংশটি গুরুত্বপূর্ণ, কারণ প্রচলিত শিরোনামের সঙ্গে এর বিরোধ আছে। হ্যালুসিনেশনকে কখনও প্রযুক্তির রহস্যময় ও অবশ্যম্ভাবী সীমা হিসেবে দেখানো হয়। গবেষণাপত্র দুই দাবিকেই নরম করে। পূর্বপ্রশিক্ষণ তত্ত্বটি নিম্নসীমা রহস্য নয়—সাধারণ পরিসংখ্যানগত তত্ত্বটি। আর হ্যালুসিনেশনের স্থায়িত্ব পুরোপুরি অনিবার্য নয়—তার একটি অংশ আমাদের বানানো প্রণোদনা। কোনো ব্যবস্থা সত্যই অনিশ্চিত হলে উত্তর না করলেই hallucinate করবে না; মোতায়েন করা মডেল কেন এমন করে না, তার একটি কারণ স্কোরবোর্ড refusal-কে শাস্তি দেয়।

লেখকেরা কী করছেন

গবেষণাপত্রটির তিনটি অংশ। প্রথমে গাণিতিক যুক্তি: পূর্বপ্রশিক্ষণে কিছু হ্যালুসিনেশন পরিসংখ্যানগতভাবে forced, কারণ “শুধু বেধে পাঠ্য তৈরি করা” অন্তত দ্বিমিক “এই বিবৃতি বৈধ কি না?” শ্রেণিবিন্যাসের মতো কঠিন। দ্বিতীয় অংশে দেখানো হয়েছে মূলধারার নিরিভুলতা-শৈলী মেট্রিক আন্দাজকে উত্তর না দেওয়ার চেয়ে পুরস্কার করে। তৃতীয় অংশে একটি প্রস্তাবিত সংশোধন এবং কেস স্টাডি: **উন্মুক্ত-rubric মূল্যায়ন**, যেখানে স্কোরিং নিয়ম প্রশ্নের মধ্যেই লেখা থাকে—উদাহরণ হিসেবে, “সঠিক উত্তর +1, ভুল উত্তর −1; আস্থা 50%-এর কম হলে উত্তর না দেওয়া করো।” এতে মডেল জানে অনিশ্চয়তা স্বীকার করা কখন পুরস্কার পাবে।

কেস স্টাডি-তে চারটি সীমান্ত মডেল ব্যবহৃত হয়েছে—Google-এর Gemini 3 Pro, OpenAI-এর GPT-5, xAI-এর Grok 4 এবং Anthropic-এর Claude Opus 4.5—এবং SimpleQA-এর **4,326** তথ্যভিত্তিক প্রশ্ন। লেখকেরা স্পষ্ট বলেন এটি illustrative কেস স্টাডি, **নিয়ন্ত্রিত মডেল তুলনা নয়**: ডিফল্ট বিন্যাস, আলাদা সামঞ্জস্য নেই, খরচ normalization নেই।

তারা কী পয়েছেনে

- **পূর্বপ্রশক্ষিণ কছিু ত্রুটি বাধ্য করে।** মডলে যে আত্মবশ্বিাসী ভুল বর্তিতরৈি করে তার হাররে একটি নম্নিসীমা আছ, যা মডলে থকেে বানানো সরো "বর্তিবিধে কনিা" শ্রণেবিন্যিাসকারীর ত্রুটরি সঙ্গে সম্পর্কতি। শখোর মতো ধরন নইে এমন তথ্যরে ক্ষত্রেে নম্নিসীমা অন্তত *একক হার*—প্রশক্ষিণে ঠকি একবার দখো তথ্যরে ভগ্নাংশ। তথ্য পুরো পরষ্কিার হলওে কছিু হ্যালুসনিশেন অনবির্য হতে পার।ে
- **স্কোরংি বাস্তবইে আন্দাজ করা পুরস্কার কর।ে** সাধারণ সংশোধন করা/ভুল স্কোরংিে কখনও উত্তর না দওেয়া না করা সর্বোত্তম কৌশল হতে পার।ে সমীক্ষায় দখো যায় জনপ্রয়ি মানদণ্ডরে বড় অংশ "I don't know"-কে স্রফে ভুল ধর।ে SimpleQA-র একটি উজ্জ্বল উদাহরণ: কাঁচা নর্িভুলতা সামান্য **o4-mini**-কে অনুকূল কর,ে যদওি এটি প্রায় সব প্রশ্নরে উত্তর দয়ে এবং তনি-চতুর্থাংশরেও বশেি ক্ষত্রেে ভুল; **GPT-5-mini** বশেি অনশ্চিয়তা স্বীকার করে বলে কম ভুল করওে কাঁচা স্কোরবোর্ডে পছিয়িে পড়তে পার।ে Reckless মডলে দখেতে ভালো লাগ।ে
- **স্কোরংি নয়িম আগে থকেে স্পষ্ট করে দলিে প্রণোদনা উল্টে যায়—অন্তত এই কসে স্টাডতি।ে** গবষেকরো একটি সহজ প্রশমন কৌশল পরীক্ষা করনে: মডলেকে একই প্রশ্নরে উত্তর দুবার দতিে বলা হয়, আর দুই উত্তর না মলিলে উত্তর না দতিে বলা হয়। প্রচলতি নর্িভুলতা মট্রেকিে এতে ভুল কম,ে কন্তিু উত্তর দওেয়ার হারও কম—েফলে মট্রেকিটি এই সতর্ক আচরণকে নর্িুৎসাহতি কর।ে কন্তিু স্কোরংি নয়িম খোলাখুলি জানানো হলে একই কৌশল বভিন্নি penalty-তে চারটি মডলেইে ভালো স্কোর কর;ে এমনকি GPT-5-mini, কাঁচা নর্িভুলতার জন্য অনুকূল করা o4-mini-কে ছাড়য়িে যায় (**n = 4,326** প্রশ্ন প্রতি মডলে)।

এর সম্ভাব্য অর্থ কী

হ্যালুসনিশেন কমাতে আরকেটি হ্যালুসনিশেন-নর্িদষ্টি পরীক্ষা বানানোই প্রধান কাজ নাও হতে পার।ে মূলধারার মানদণ্ড অনশ্চিয়তা কীভাবে স্কোর করে সটেি বদলানোও গুরুত্বপূর্ণ, যাতে "আমি জানি না" বলা স্বয়ংক্রয়িভাবে শাস্তি না পায়। স্কোরবোর্ড না বদলালে হ্যালুসনিশেন-হ্রাসকারী আচরণ নর্িভুলতা বন্দিু খরচ করতে পার,ে ফলে ডভেলেপাররার প্রণোদনা উল্টোই থাক।ে এই জন্য লখেকরো সমস্যাটকিে **socio-প্রযুক্তগিত** বলনে: মট্রেকি নকশা এবং প্রভাবশালী লডিারবোর্ডরে adoption—দুটোই দরকার।

এটি কী প্রমাণ করে না

- উন্মুক্ত rubric বাস্তব মোতায়নে করা ব্যবস্থায় হ্যালুসনিশেন ঠকি করে দয়ে—েএটি গবষেণাপত্র দখোয় না। সমর্থনকারী পরীক্ষা ছোট এবং ইচ্ছাকৃতভাবে **অনয়িন্ত্রতি**: চার মডলে, ডফিল্ট settings, একটি প্রশমন, একটি তথ্যভত্তিকি-QA পরীক্ষা। উদ্দশ্যে প্রণোদনা উল্টে যাওয়া দখোনো, মডলে ranking বা সাধারণ কার্যকারতিা নয়।
- Grading-ই হ্যালুসনিশেনরে **একমাত্র** কারণ—গবষেণাপত্র তা বলে না। প্রশক্ষিণ-তথ্য ত্রুটি, সত্যহিি কঠনি সমস্যা এবং অপরচিতি প্রম্পট আলাদা উৎস।
- হ্যালুসনিশেন অবশ্যম্ভাবী—জনপ্রয়ি দাবটিওি গবষেণাপত্র সমর্থন করে না। তাদরে যুক্তি হলো, ব্যবস্থা যদি শুধু checkable প্রশ্ন উত্তর করে এবং অন্য ক্ষত্রেে "জানি না" বল,ে তাহলে হ্যালুসনিশেন এড়ানো সম্ভব।
- পূর্বপ্রশক্ষিণ নম্নিসীমা অদৃশ্য হয়—উন্মুক্ত rubric তা করে না। গবষেণাপত্র নম্নিসীমা ব্যাখ্যা ও সীমা কর;ে সীমা আত্মবশ্বিাসী তথ্যভত্তিকি ত্রুটি নয়িে, সব মডলে আচরণ নয়িে নয়।
- উন্মুক্ত rubric একাই যথষ্টে—এটিও নয়। মূল্যায়ন প্রণোদনা বদলানো retrieval, সরঞ্জাম ব্যবহার বা ভালো ক্যালবি্রশেনরে বকিল্প নয়।

প্রমাণ কতটা শক্ত?

- মূল অংশটি **গাণতিকি**: আনুষ্ঠানকি নম্নিসীমা, প্রায়োগকি পরমিাপ নয়। তাত্ত্বকি যুক্তি তার অনুমানগুলোর মধ্যে শক্ত।
- সমস্যাটরি একটি **ইচ্ছাকৃতভাবে সরলীকৃত মডলে** ব্যবহার করা হয়ছে। লখেকরো নজিরোই "সঠকি / ভুল / 'আমি জানি না'" এই তনি ভাগকে অতরিক্তি সরল ত্রবিভিাজন বলনে এবং সবচয়েে পরষ্কিার সীমা বরে করতে আদর্শায়তি, ইচ্ছামতো-তথ্য বন্যিাস ব্যবহার করনে।
- মানদণ্ড সমীক্ষা **ছোট ও curated**: দশটি প্রভাবশালী মূল্যায়ন, exhaustive নরীক্ষা নয়।
- কসে স্টাডি **বাস্তব কন্তিু সীমতি**: চার সীমান্ত মডলে, এক প্রশমন, শুধু SimpleQA, ডফিল্ট settings, এবং স্পষ্টভাবে "not a নয়িন্ত্রতি মূল্যায়ন"। এটি প্রণোদনা যুক্তরি ধারণার প্রাথমকি প্রমাণ, মানদণ্ড ফল নয়।
- Vantage বন্দিু-ও উল্লখেযোগ্য: চার লখেকরে তনিজন OpenAI-তে কাজ করনে বা করছেনে, এবং গবষেণাপত্র ক্ষত্রেরে মূল্যায়ন চর্চা বদলানোর পক্ষে যুক্তি দয়ে। এটকিে dismiss করার কারণ নয়, কন্তিু স্বাধীন নরিপক্ষে পর্যালোচনা হসিবেওে পড়া ঠকি নয়। কৃতত্বি হলো, critique তারা নজিদেরে o4-mini ও GPT-5-mini-এর ক্ষত্রেওে প্রয়োগ করছেনে।

কনে এটি গুরুত্বপূর্ণ

এটি খুব অতরিঞ্জতি প্রচার হওয়া একটি সমস্যাকে নতুনভাবে কাঠামো করে। "হ্যালুসনিশেন" কখনও spooky defect, কখনও immovable প্রাচীর হসিবেে বকি্রি হয়। এই গবষেণাপত্র এটকিে অনকে বশেি সাধারণ করে: একটি পরসিংখ্যানগত নম্নিসীমা, যা বোঝা যায়; তার ওপর আমাদরে তরৈি প্রণোদনা, যা বদলানো যায়। Quiet কন্তিু উপযোগী শক্ষিা হলো, নর্ভিরযোগ্যতার পরবর্তী অগ্রগতি কবেল মডলে স্থাপত্য বা প্রশক্ষিণ পরসির থকেে নাও আসতে পার—**আমরা কী মাপি এবং কী পুরস্কার করি**, সটেওি সমান গুরুত্বপূর্ণ হতে পার।

সংক্ষপ্তি সারাংশ

ভাষা মডলেরে আত্মবশ্বিাসী ভুলরে দুটি আলাদা উৎস থাকতে পার। প্রথমটি পরসিংখ্যানগত: কোনো তথ্যরে শখোর মতো নয়িম না থাকলে এবং প্রশক্ষিণ-তথ্যে সটেি খুব কম দখো গলেে অনুকরণভত্তিকি প্রশক্ষিতি মডলে কখনো না কখনো ভুল করব; একক হাররে মতো রাশি থকেে সইে নম্নিসীমা অনুমান করা যায়। দ্বতিীয়টি প্রণোদনার সমস্যা: অধকিাংশ মানদণ্ডে "আমি জানি না" এবং ভুল উত্তর একই শূন্য স্কোর পায়, ফলে আন্দাজ করাই যুক্তসিঙ্গত কৌশল হয়ে দাঁড়ায়। এমনকি প্রায় তনি-চতুর্থাংশ ভুল করা বপেরোয়া মডলেও, অনশ্চিতি হলে উত্তর না দওেয়া মডলেরে চয়েে কাঁচা নর্ভিুলতায় ভালো দখোতে পার। লখেকদরে প্রস্তাব হলো **স্কোরিং নয়িম খোলাখুলি জানানো**—প্রশ্নরে মধ্যইে কীভাবে নম্বর দওেয়া হবে তা বলা। চারটি সীমান্তবর্তী মডলেরে ছোট কসে স্টাডতিে এতে হ্যালুসনিশেন কমানোর আচরণ প্রচলতি নর্ভিুলতার তুলনায় বশেি পুরস্কার পায়। এটি তত্ত্ব, মানদণ্ড-সমীক্ষা ও ছোট অনয়িন্ত্রতি পরীক্ষার সমন্বয়; সমাধানটি আশাব্যঞ্জক, কন্তিু বৃহৎ পরসিরে প্রমাণতি নয়। মূল বক্তব্য: হ্যালুসনিশেন রহস্যময়ও নয়, পুরোপুরি অনবিার্যও নয়।

সরাসরি যাচাই

গবষেণাপত্র কী দখোয়: পূর্বপ্রশক্ষিণে কছিু তথ্যভত্তিকি হ্যালুসনিশেনরে গাণতিকি নম্নিসীমা; সমীক্ষায় অধকিাংশ প্রধান ক্রমরে মানদণ্ড "I don't know"-কে কৃতত্বি দয়ে না; এবং চার-মডলে কসে স্টাডি-তে প্রম্পটরে মধ্যইে স্কোরিং নয়িম দওেয়া হলে হ্যালুসনিশেন-হ্রাসকারী প্রশমন, সরল নর্ভিুলতা-তে penalised হওয়ার বদলে পুরস্কার পায়।

কী সম্ভাব্য কন্তিু প্রমাণতি নয়: মূলধারার মানদণ্ডে উন্মুক্ত rubric চালু করলে মোতায়নে করা মডলেরে হ্যালুসনিশেন বাস্তবে উল্লখেযোগ্যভাবে কমব। সমরথনকারী পরীক্ষা ছোট ও স্পষ্টভাবে অনয়িন্ত্রতি।

কী দখোয় না: হ্যালুসনিশেন অনবিার্য—গবষেণাপত্র উল্টো যুক্তি দয়ে; grading-ই sole কারণ; হ্যালুসনিশেন একবোরে eliminate করা যাবে; বা কসে স্টাডি চার মডলেরে ন্যায্য ranking।

মূল সীমাবদ্ধতা: Simplified সংশোধন করা/ভুল/"I don't know" মডলে; দশটি মূল্যায়নরে curated সমীক্ষা; চার মডলে, এক প্রশমন, এক পরীক্ষা ও ডফিল্ট বন্যিাসরে অনয়িন্ত্রতি কসে স্টাডি; এবং মূল্যায়ন চর্চা নয়িে OpenAI-led যুক্তি।

সাধারণ পাঠকরে আস্থা কতটা হওয়া উচতি? উচ্চ আস্থা যে হ্যালুসনিশেন রহস্যময় বা strictly অনবিার্য কোনো ঘটনা নয় এবং মূলধারার মানদণ্ড আন্দাজ করাকে পুরস্কার করতে পারে। প্রস্তাবতি উন্মুক্ত-rubric সংশোধন বস্তিৃতভাবে কাজ করবে—এতে মাঝারি আস্থা: ধারণার প্রাথমকি প্রমাণ বাস্তব, কন্তিু পরসির ও ব্যবহারকি মোতায়নেে প্রদর্শন এখনো নইে।

উৎস

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