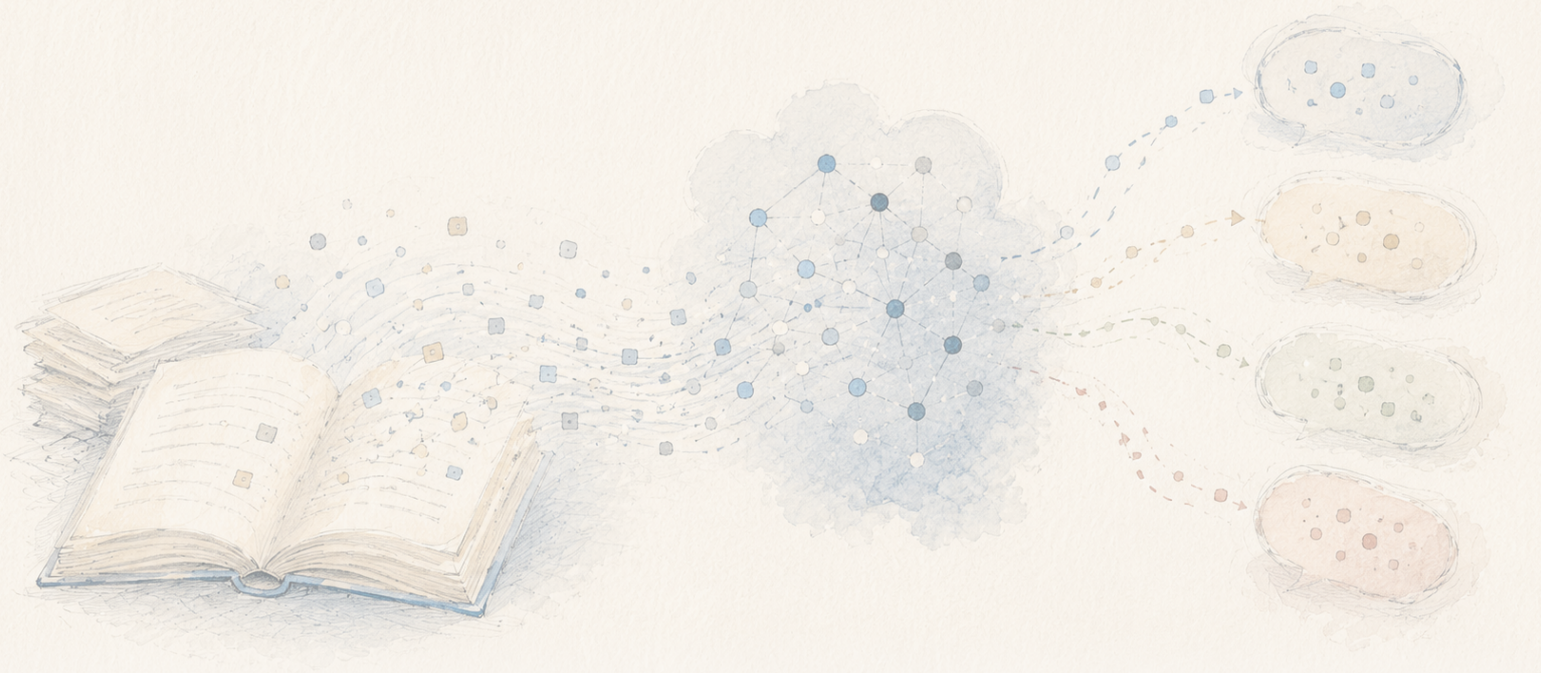




The CLEAN PAPER

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

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

Paper statistical lower bound দিয়ে কিছু factual error-এর inevitability ব্যাখ্যা করে এবং দেখায় mainstream accuracy benchmark “I don’t know” ও wrong answer-কে একই score দলি়ে guessing rational হয়। Open-rubric case study incentive flip করে, কনিত্ত 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 বহু ভাষা মডলে-কে কনে। অচনো ব্যক্তির জন্মদনি জিজ্ঞেসে করুন। সে হয়তো এমন আত্মবিশ্বাসে বলবে, “৭ মার্চ”, যনে কার্ড থেকে পড়ছে—কনিত্ত উত্তরটি ভুল। আবার জিজ্ঞেসে করলে অন্য তারখি, তৃতীয়বারে আরও একটি। গবষণাপত্ররে লখেকরো ঠিকি এমন উদাহরণ দনে: প্রধান ক্রমরে মডলেগুলোকে সাধারণ তথ্যভিত্তিক প্রশ্ন করা হলে—কনে। ব্যক্তির জন্মতারখি, বা আড়াল করে সংক্ষিপ্ত রূপে পূর্ণরূপ—প্রতিটি মডলে নরিদখিট ও আত্মবিশ্বাসী কনিত্ত আলাদা ভুল উত্তর বানায়। শলিপখাত এটাকে *হ্যালুসিনেশন* বলে, শব্দটি যনে উপলব্ধির অদ্ভুত তরুটি বোঝায়। গবষণাপত্ররে প্রথম কাজ হলে। রহস্যরে আবরণ সরানে।

মডলে কীভাবে তেরিহয় তা থেকে শুরু করা যাক। প্রথম ও সবচেয়ে বড় প্রশিক্ষণ পর্যায়ে সে বপিল পাঠ্য পড়ে সবচ্ছন্দ ভাষার ধরন শখে। এখন এমন একটি তথ্য ননি যার পছনে শখোর মতে। ধরন নই—একজন নরিদখিট মানুষরে জন্মতারখি। প্রশিক্ষণ পাঠ্যে তারখিটি একবার এসছে, বা একবারও আসনে, তাহলে ধরন শিক্ষণ-ব্যবস্থার ধরার মতে। কছু নই। মডলেরে দৃষ্টিতে উত্তরটি ইচ্ছামতে। লখেকরো Alan Turing-এর একটি পুরনে। পরসিংখ্যানগত ধারণা নযি এটিকে নরিদখিট করনে: প্রশিক্ষণ-তথ্যে যদি পাঁচটির একটি জন্মতারখি মাত্র একবার দেখা যায়, তাহলে মডলেরে অন্তত প্রায় পাঁচটির একটিতে ভুল হওয়া প্রত্যাশতি—মডলে “ভাঙা” বলে নয়, বরং শখোর মতে। পর্যাপ্ত প্রমাণ ছিল না বলে। বপিরীতে দেশরে রাজধানী মডলে খুব কম ভুল করে, কারণ “Paris” বা অনুরূপ তথ্য তথ্যে বারবার আসে।

তারা আরও যুক্তি দনে যে কনে। সম্ভাব্য বরিতিসত্য ক মিথ্যা তা নরিভরযে। গ্যভাবে শ্রণেবিনিয়াস করাই কঠনি; তাই **শুধু সত্য বরিতিতেরি করা** অন্তত সেই শ্রণেবিনিয়াস সমস্যার মতে। কঠনি। অরখাৎ কছু তরুটির পরসিংখ্যানগত নমিনসীমা থেকেই যায়।

যে ধারণাটি নিচি কাজ করছে: যা এখনে। দখেনেনি, তা কীভাবে গে। না যায়

এখানে একটি সুন্দর পুরনে। পরসিংখ্যানগত ধারণা কাজ করে।

রঙনি বলরে একটি খলকিল্পনা করুন। কত রং আছে জাননে না। 100টি বল একে একে তুললনে: লাল 40, নীল 25, সবুজ 15, হলুদ 5, বগুনী 3, কমলা 2—আর তারপর **10টি আলাদা রং, যগুলো এর প্রত্যেকেটি মাত্র একবার করে দেখা গলে।**

প্রশ্নটি হলো: পররে বলটি এমন রঙরে হওয়ার সম্ভাবনা কত, যা আপনি এখনে। একবারও দখেনেনি? যা দখেনেনি তা সরাসরি গে। না যায় না। কনিত্ত **যে রংগুলো ঠিকি একবার দেখা গেছে**, অরখাৎ একক, সেগুলো গে। না যায়। **ভালে Turing অনুমান**-এর মূল অন্তরদৃষ্টি হলো: নমুনায় এককরে ভাগ, অদখো শ্রণেতিে এখনে। কত সম্ভাবনা ভর লুকয়িে আছে তার অনুমান দয়ে। 100টি ডির-এর মধ্যে 10টি একক হলে পররে ডর একবোরনে নতুন রঙরে হওয়ার সম্ভাবনা আনুমানিক $10 / 100 = 10\%$ ।

এই একবার-দখো রংগুলো ভুল নয়। এগুলো আপনার অজ্ঞতার পরিমাপ: অনকে একক মাননে নমুনা নজিইে বলছে, বাইরে আরও শ্রণেটি আছে যা এখনে। ঠিকিমতে। দেখা হয়নি।

এখন রঙরে বদলে জন্মতারখি ধরুন, আর থলরি বদলে মডলেরে প্রশিক্ষণ পাঠ্য। দেখো জন্মতারখি্রে মধ্যে যদি **পাঁচটির একটি মাত্র একবার উপস্থতি হয়**, তাহলে একই ববিচেনামূলক চিন্তা বলে বণ্টনরে উল্লেখ্যে। গ্য অংশ মডলে কার্যত ভালে। করে কখনে। শখেনে। জন্মতারখি্রে কনে। গণনায়ে গ্য ধরন নই—কারও জন্মতারখি “উৎপন্ন” করা যায় না। তাই একবার-দখো বা অদখো তারখিে মডলেরে ভুলরে একটি নমিনসীমা থাকে।

এটাই যুক্তির ক্যুদর সংস্করণ: একক হার বলে *এই তথ্য থেকে পৃথিবীর কতটা অংশ শখো কঠনি*, এবং সটেহি তরুটির নিচি একটি **নমিনসীমা** তেরি করে। রাজধানী শহরে ক্যেত্রে এই নমিনসীমা খুব ছোট, কারণ সেগুলো বারবার দেখা যায়।

এতে হ্যালুসিনেশন কে। থা থেকে আসে তার একটি অংশ বোঝা গলে। কনিত্ত পরে সহায়কতা ও সততা শখোনে। প্রশিক্ষণরে পরও মডলে কনে ভান করে উত্তর দয়ে, সটেই এখনে। বাকি। এখানে গবষণাপত্ররে উপমা খুব সরাসরি। পরীক্ষায় কনে। শকিষারখী উত্তর জানে না। ফাঁকা রাখলে শূন্য, আন্দাজ করলে শূন্য বা একটি—তাহলে গ্রেডে সরবাধিক করার কৌশল হলে। আন্দাজ করা। “আমি জানি না” বললে কনে। লাভ নই। মডলেরে মানদণ্ডেও প্রায় একই প্রণে। দনা। লখেকরো ক্যেত্রেরে প্রভাবশালী মানদণ্ডগুলো দেখনে এবং দেখনে অধিকাংশে “আমি জানি না” ভুল উত্তররে মতে। এই শূন্য পায়। এই স্কোরিং নিয়মে অনশিচ্যুতা স্বীকার করা মডলেরে চয়ে সব সময় আন্দাজ করা মডলে বর্শা স্কোর করতে পারে। এক অরখে আমরা মূল্যায়ন দয়িইে মডলেকে ভান করে উত্তর দতিে শখিয়িছে।

Two ways to grade an answer

what each scoring rule rewards

Closed rubric

how most benchmarks score today

Correct	+1
Wrong	0
"I don't know"	0

Wrong and "I don't know" tie at 0, so a guess can only help.

Open rubric

scoring stated inside the question

Correct	+1
Wrong	-1
"I don't know"	0

A wrong guess now costs more than an honest "I don't know."

মানদণ্ড আন্দাজ করাকে যুক্তিসংগত করে তুলতে পারে। সাধারণ স্কোরিংয়ে (বাম দিকি) ভুল উত্তর এবং সং “আমি জানি না” —দুটোই শূন্য, তাই আন্দাজ করলে কেবল লাভের সম্ভাবনা থাকে। প্রশ্নের মধ্যহে নিয়ম স্পষ্ট করে ভুল উত্তরে দণ্ড দিলে (ডান দিকি), অনশ্চয়তা বেশি হলে উত্তর না দেওয়া ভালো। কঠোর হতে পারে। এটি মূল্যায়ন কী পুরস্কার করে তা বদলায়; নজি থেকে হ্যালুসিনেশনের পুরো সমাধান নয়।

Original diagram — The Clean Paper CC BY 4.0

এই অংশটি গুরুত্বপূর্ণ, কারণ প্রচলিত শরিতে নামের সঙ্গে এর বর্ণিত আছে। হ্যালুসিনেশনকে কখনও প্রযুক্তির রহস্যময় ও অবশ্যম্ভাবী সীমা হিসেবে দেখানো হয়। গবেষণাপত্র দুই দাবকিই নরম করে। পূর্বপ্রশিক্ষণ ত্রুটি নমিসীমা রহস্য নয়—সাধারণ পরিসংখ্যানগত ত্রুটি। আর হ্যালুসিনেশনের স্থায়িত্ব পুরো পুরি অনবির্য নয়—তার একটি অংশ আমাদের বানানো প্রণেদনা। কোনো ব্যবস্থা সত্যিই অনশ্চিতি হলে উত্তর না করলই হ্যালুসিনেটে করবে না; মতোয়নে করা মডলে কনে এমন করবে না, তার একটি কারণ স্কোরবোড উত্তর না দেয়াকে শাস্তি দিয়ে।

লেখকরো কী করছেন

গবেষণাপত্রটির তিনটি অংশ। প্রথম গাণিতিক যুক্তি: পূর্বপ্রশিক্ষণে কঙ্কি হ্যালুসিনেশন পরিসংখ্যানগতভাবে অনবির্য, কারণ “শুধু বধৈ পাঠ্য তরৈকিরা” অন্তত দ্বমিকি “এই বর্ভি বধৈ কনি?” শ্রণেবিনি্যাসরে মতো কঠনি। দ্বিতীয় অংশে দশটি প্রভাবশালী মানদণ্ডের সমীক্ষা দিয়ৈ দেখোনৈ হয্ছে মূলধারার নরিভুলতা-শৈলী মটেরকি আন্দাজকে উত্তর না দেওয়ার চযে পুরস্কার করে। তৃতীয় অংশে একটি প্রস্তাবতি সংশোধন এবং কসে স্টাডি: **উন্মুক্ত বুঝকি মূল্যায়ন**, যখনে স্কোরিং নিয়ম প্রশ্নের মধ্যহে লখো থাকে—উদাহরণ হিসেবে, “সঠিক উত্তর +1, ভুল উত্তর -1; আস্থা 50%-এর কম হলে উত্তর দণ্ডি না।” এতে মডলে জানে অনশ্চয়তা স্বীকার করা কখন পুরস্কার পাবে।

কসে স্টাডি-তে চারটি সীমান্ত মডলে ব্যবহৃত হয্ছে—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। লেখকরো স্পষ্ট বলনে এটি উদাহরণমূলক কসে স্টাডি, **নিয়ন্ত্রতি মডলে তুলনা নয়**: ডফিল্ট বনি্যাস, আলাদা সামঞ্জস্য নহৈ, খরচ স্বাভাবিকীকরণ নহৈ।

তারা কী পয়েছেন

- **পূর্বপ্রশিক্ষণ কিছু ত্রুটি বাধ্য করে।** মডলে যে আত্মবিশ্বাসী ভুল বহির্ভিত্তিক করে তার হারের একটি নমিনসীমা আছে, যা মডলে থেকে বানানো সেরা “বহির্ভিত্তিক কনি” শ্রুণেবিনিয়াসকারীর ত্রুটির সঙ্কে সমপ্রকতি। শেখার মতো ধরন নই এমন তথ্যের ক্ষেত্রে নমিনসীমা অন্তত একক হার—প্রশিক্ষণে ঠিকি একবার দেখা তথ্যেরে ভগ্নাংশ। তথ্য পুরো পরিস্কার হলওে কিছু হ্যালুসিনেশন অনবিার্য হতে পারে।
- **স্কোরিং বাস্তবই আন্দাজ করা পুরস্কার করে।** সাধারণ করা/ভুল স্কোরিংয়ে কখনও সব প্রশ্নেরে উত্তর দেওয়াই সর্বোত্তম কৌশল হতে পারে। সমীকষায় দেখা যায় জনপ্রিয় মানদণ্ডেরে বড় অংশ “আমি জানি না”-কে সুরফে ভুল ধরে। SimpleQA-র একটি উজ্জ্বল উদাহরণ: কাঁচা নরিভুলতা সামান্য o4-mini-কে অনুকূল করে, যদিও এটি প্রায় সব প্রশ্নেরে উত্তর দেয় এবং তনি-চতুরতাংশেরেও বেশি ক্ষেত্রে ভুল; GPT-5-mini বেশি অনিশ্চয়তা স্বীকার করে বলে কম ভুল করেও কাঁচা স্কোরেরে পড়তে পছিয়ে পড়তে পারে। বপেরে যা মডলে দেখতে ভালো লাগে।
- **স্কোরিং নিয়ম আগ থেকে স্পষ্ট করে দিলে প্রণেদনা উল্টে যায়—অন্তত এই কেসে স্টাডিতে।** গবেষকেরা একটি সহজ প্রশ্নমন কৌশল পরীক্ষা করেন: মডলকে একই প্রশ্নেরে উত্তর দুবার দিতে বলা হয়, আর দুই উত্তর না মিললে উত্তর না দিতে বলা হয়। প্রচলিত নরিভুলতা মটেরিকে এতে ভুল কমে, কিন্তু উত্তর দেওয়ার হারও কমে—ফলে মটেরিকি এই সতর্ক আচরণকে নরিব্রহতি করে। কিন্তু স্কোরিং নিয়ম খোলাখুলি জানানো হলো একই কৌশল দণ্ডেরে বিভিন্ন মাত্রায় চারটি মডলেই ভালো স্কোর করে; এমনকি GPT-5-mini, কাঁচা নরিভুলতার জন্য অনুকূল করা o4-mini-কে ছাড়িয়ে যায় (n = 4,326 প্রশ্ন প্রতি মডলে)।

এর সম্ভাব্য অর্থ কী

হ্যালুসিনেশন কমাতে আরকেটি হ্যালুসিনেশন-নরিদৃষ্টি পরীক্ষা বানানোই প্রধান কাজ নাও হতে পারে। মূলধারার মানদণ্ড অনিশ্চয়তা কীভাবে স্কোর করে সেটি বদলানোও গুরুত্বপূর্ণ, যাতে “আমি জানি না” বলা স্বয়ংক্রিয়ভাবে শাস্তিনা পায়। স্কোরেরে পড় না বদলালে হ্যালুসিনেশন-হ্রাসকারী আচরণ নরিভুলতা বিন্দু খরচ করতে পারে, ফলে ডেভেলপারদের প্রণেদনা উল্টে থাকে। এই জন্য লখেকরো সমস্যাটিকে **সামাজিক-প্রযুক্তগিত** বলেন: মটেরিকি নকশা এবং প্রভাবশালী লিডারেরে গ্রহণ—দুটোই দরকার।

এটাকী প্রমাণ করে না

- উন্মুক্ত মূল্যায়ন-নিয়ম বাস্তব মতোয়নেকৃত ব্যবস্থায় হ্যালুসিনেশন ঠিকি করে দেয়—এটি গবেষণাপত্র দেখায় না। সমর্থনকারী পরীক্ষা ছোট এবং ইচ্ছাকৃতভাবে **অনিয়ন্ত্রিত**: চার মডলে, ডফিলট সটেইস, একটি প্রশ্নমন, একটি তথ্যভিত্তিক-QA পরীক্ষা। উদ্দেশ্য প্রণেদনা উল্টে যাওয়া দেখানো, মডলে কর্মতালিকা বা সাধারণ কার্যকারিতা নয়।
- মূল্যায়নই হ্যালুসিনেশনের **একমাত্র** কারণ—গবেষণাপত্র তা বলেনো। প্রশিক্ষণ-তথ্য ত্রুটি, সত্যিই কঠিন সমস্যা এবং অপরিচিতি প্রম্পট আলাদা উৎস।
- হ্যালুসিনেশন অবশ্যম্ভাবী—জনপ্রিয় দাবিটিও গবেষণাপত্র সমর্থন করে না। তাদের যুক্তি হলো, ব্যবস্থা যদি শুধু যাচাইযোগ্য প্রশ্ন উত্তর করে এবং অন্য ক্ষেত্রে “জানি না” বলে, তাহলে হ্যালুসিনেশন এড়ানো সম্ভব।
- পূর্বপ্রশিক্ষণ নমিনসীমা অদৃশ্য হয়—উন্মুক্ত মূল্যায়ন-নিয়ম তা করে না। গবেষণাপত্র নমিনসীমা ব্যাখ্যা ও সীমা করে; সীমা আত্মবিশ্বাসী তথ্যভিত্তিক ত্রুটি নিয়ে, সব মডলে আচরণ নিয়ে নয়।
- উন্মুক্ত মূল্যায়ন-নিয়ম একাই যথেষ্ট—এটিও নয়। মূল্যায়ন প্রণেদনা বদলানো তথ্য-উদ্ধার, সরঞ্জাম ব্যবহার বা ভালো ক্যালিব্রেশনেরে বকিল্প নয়।

প্রমাণ কতটা শক্ত?

- মূল অংশটি **গাণিতিক**: আনুষ্টানিক নমিনসীমা, প্রায়োগিক পরিমাপ নয়। তাত্ত্বিক যুক্তি তার অনুমানগুলোর মধ্যে শক্ত।
- সমস্যাটির একটি **ইচ্ছাকৃতভাবে সরলীকৃত মডেল** ব্যবহার করা হয়েছে। লেখকরা নিজেরাই “সঠিক / ভুল / ‘আমি জানি না’” এই তিন ভাগকে অতিরিক্ত সরল ত্রিবিভাজন বললে এবং সবচেয়ে পরিস্কার সীমা বরে করতে আদর্শায়িত, ইচ্ছামতো তথ্য বিন্যাস ব্যবহার করেন।
- মানদণ্ড সমীক্ষা **ছোট ও বাছাইকৃত**: দশটি প্রভাবশালী মূল্যায়ন, সম্পূর্ণ নরীক্ষা নয়।
- কসে স্টাডি **বাস্তব কনিত্র সীমিত**: চার সীমান্ত মডলে, এক প্রশমন, শুধু SimpleQA, ডফিল্ট সটেইস, এবং স্পষ্টভাবে “নয়নিত্রতি মূল্যায়ন নয়”। এটি প্রণেদনা যুক্তির ধারণার প্রাথমিক প্রমাণ, মানদণ্ড ফল নয়।
- দৃষ্টিকোণটিও উল্লেখযোগ্য: চার লেখকের তনিজন OpenAI-তে কাজ করেন বা করছেন, এবং গবেষণাপত্র ক্ষেত্রে মূল্যায়ন চরচা বদলানোর পক্ষে যুক্তি দিয়ে। এটিকে খারজি করার কারণ নয়, কনিত্র স্বাধীন নরিপক্ষে পর্যালোচনা হিসেবেও পড়া ঠিক নয়। কৃত্তি হলো, সমালোচনা তারা নিজদের o4-mini ও GPT-5-mini-এর ক্ষেত্রেও প্রয়োগ করছেন।

কনে এটি গুরুত্বপূর্ণ

এটি খুব অতিরঞ্জিত প্রচার হওয়া একটি সমস্যাকে নতুনভাবে কাঠামো করে। “হ্যালুসিনেশন” কখনও অদ্ভুত ত্রুটি, কখনও অচল প্রাচীর হিসেবে বিক্রি হয়। এই গবেষণাপত্র এটিকে অনেকে বেশি সাধারণ করে: একটি পরিসংখ্যানগত নমিনসীমা, যা বোঝা যায়; তার ওপর আমাদের তরৈ প্রণেদনা, যা বদলানো যায়। শান্ত কনিত্র উপযোগী শিক্ষা হলো, নরিভরযোগ্যতার পরবর্তী অগ্রগতি কেবল মডলে স্থাপত্য বা প্রশিক্ষণ পরিসর থেকে নাও আসতে পারে—**আমরা কী মাপি এবং কী পুরস্কার করি**, সটেও সমান গুরুত্বপূর্ণ হতে পারে।

সংক্ষিপ্ত সারাংশ

ভাষা মডলের আত্মবিশ্বাসী ভুলের দুটি আলাদা উৎস থাকতে পারে। প্রথমটি পরিসংখ্যানগত: কোনও তথ্যের শেখার মতো নিয়ম না থাকলে এবং প্রশিক্ষণ-তথ্যে সটেই খুব কম দেখা গেলে অনুকরণভিত্তিক প্রশিক্ষিত মডলে কখনও না কখনও ভুল করবে; একক হারের মতো রাশি থেকে সেই নমিনসীমা অনুমান করা যায়। দ্বিতীয়টি প্রণেদনার সমস্যা: অধিকাংশ মানদণ্ডে “আমি জানি না” এবং ভুল উত্তর একই শূন্য স্কোর পায়, ফলে আন্দাজ করাই যুক্তিসঙ্গত কৌশল হয়ে দাঁড়ায়। এমনকি প্রায় তনি-চতুর্থাংশ ভুল করা বপেরে যা মডলেও, অনিশ্চিত হলে উত্তর না-দেওয়া মডলের চেয়ে কাঁচা নরিভুলতায় ভালো দেখাতে পারে। লেখকদের প্রস্তাব হলো **স্কোরিং নিয়ম খোলাখুলি জানানো**—প্রশ্নের মধ্যই কীভাবে নম্বর দেওয়া হবে তা বলা। চারটি সীমান্তবর্তী মডলের ছোট কসে স্টাডিতে এতে হ্যালুসিনেশন কমানোর আচরণ প্রচলিত নরিভুলতার তুলনায় বেশি পুরস্কার পায়। এটি তত্ত্ব, মানদণ্ড-সমীক্ষা ও ছোট অনয়িত্রতি পরীক্ষার সমন্বয়; সমাধানটি আশাব্যঞ্জক, কনিত্র বৃহৎ পরিসরে প্রমাণিত নয়। মূল বক্তব্য: হ্যালুসিনেশন রহস্যময়ও নয়, পুরোপুরি অনিবার্যও নয়।

সরাসরি যাচাই

গবেষণাপত্র কী দেখায়: পূর্বপ্রশিক্ষণে কিছু তথ্যভিত্তিক হ্যালুসিনেশনের গাণিতিক নমিনসীমা; সমীক্ষায় অধিকাংশ প্রধান ক্রমের মানদণ্ড “আমি জানি না”-কে কৃত্তি দিয়ে না; এবং চার-মডলে কসে স্টাডিতে প্রমপটের মধ্যই স্কোরিং নিয়ম দেওয়া হলে হ্যালুসিনেশন-হ্রাসকারী প্রশমন, সরল নরিভুলতায় দণ্ডিত হওয়ার বদলে পুরস্কার পায়।

কী সম্ভাব্য কনিত্র প্রমাণিত নয়: মূলধারার মানদণ্ডে উন্মুক্ত মূল্যায়ন-নিয়ম চালু করলে মেতায়নে করা মডলের হ্যালুসিনেশন বাস্তবে উল্লেখযোগ্যভাবে কমবে। সমর্থনকারী পরীক্ষা ছোট ও স্পষ্টভাবে অনয়িত্রতি।

কী দখোয় না: হ্যালুসিনেশন অনবিার্য—গবেষণাপত্র উল্টো। যুক্তি দিয়ে; মূল্যায়নই একমাত্র কারণ; হ্যালুসিনেশন একবোরো বাদ দেওয়া যাবে; বা কসে স্টাডি চার মডলে ন্যায্য কর্মতালিকা।

মূল সীমাবদ্ধতা: সরলীকৃত করা/ভুল/“I জানিনি” মডলে; দশটি মূল্যায়নরে বাছাই করা সমীক্ষা; চার মডলে, এক প্রশমন, এক পরীক্ষা ও ডফিল্ট বনিয়াসরে অনয়িন্ত্রতি কসে স্টাডি; এবং মূল্যায়ন চরচা নযি। OpenAI-led যুক্তি।

সাধারণ পাঠকরে আস্থা কতটা হওয়া উচিত? উচ্চ আস্থা য়ে হ্যালুসিনেশন রহস্যময় বা কঠোরভাবে অনবিার্য কনো। ঘটনা নয় এবং মূলধারার মানদণ্ড আন্দাজ করাকে পুরস্কার করতে পারে। প্রস্তাবতি উন্মুক্ত-মূল্যায়ন-নয়িম সংশোধন বসিত্তভাবে কাজ করবে—এতে মাঝারি আস্থা: ধারণার প্রাথমিক প্রমাণ বাস্তব, কনিত্তু পরসির ও ব্যবহারিক মতো তায়নে প্রদর্শন এখনো নহে।

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>