The Current State of AI in Rheumatology

Paul Sufka, MD

Twin Cities Orthopedics, Eagan, MN

𝕏:@psufka 🦋:@psufka.bsky.social

Slides: paulsufka.com/AI

Disclosures

I have no relevant disclosures.

Objectives

- Define the major types of artificial intelligence relevant to healthcare
- Describe current and emerging applications of AI in rheumatology
- Evaluate the limitations, ethical concerns, and future directions of AI integration in rheumatologic practice

Naval
@naval

𝕏.com

AIs aren’t minds and robots aren’t bodies.

They’re better at some things - usually the things we hate to do.

They’re worse at other things - usually the things we like to do.

Automation more than competition.

Last edited 1:28 AM · 9/6/25 · **302K** Views

Outline

1. AI Basics
2. Clinical Uses
3. AI Agents & EHRs
4. Risks & Limitations
5. Future Directions
6. Practical Tips
7. Discussion / Q&A

Definitions

Artificial Intelligence: computer systems performing tasks requiring human-like intelligence

Hierarchy:

- Algorithms: rule-based logic
- Machine Learning: learns patterns from data
- Deep Learning: ML using neural networks
  - Includes Transformers (GPT = Generative Pretrained Transformer)
  - GPT Enables Large Language Models (LLMs)

Transformers

- Transformer: introduced "self-attention" — this allowed neural networks to consider all relationships between all data in a sequence at once

- Benefits:
  - Parallel processing → faster
  - Scaling → improve as they grow
  - Versatility → handle text, code, images, sound, graphics

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

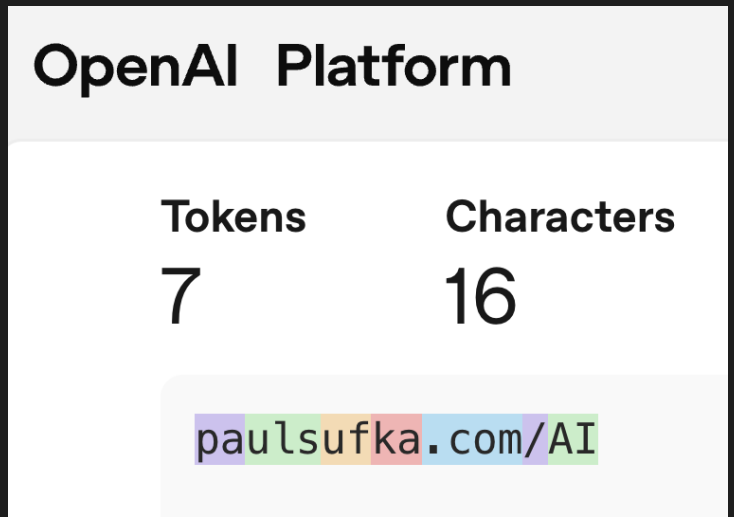
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

https://arxiv.org/pdf/1706.03762

Pretraining Transformers

1. Tokenization: Turning Words into Numbers

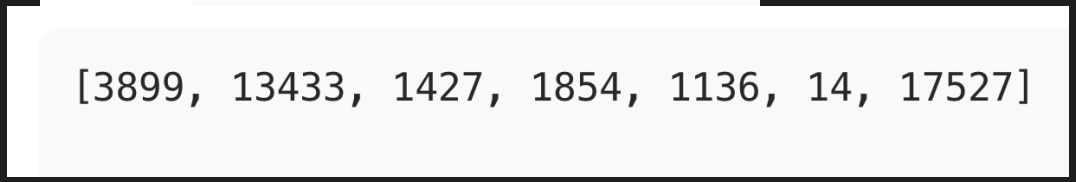
- Text is split into small fragments (≈3–4 characters)
- Entire vocabulary: ~100k–200k unique tokens



Since pretraining is highly technical, I will cover it quickly.

If you're interested, slides for this presentation and references are available here

⬅️



https://platform.openai.com/tokenizer

Pretraining

2. Embeddings: The model's dictionary

- Each token → mapped to a vector capturing relationships
  - Over trillions of training updates, vectors cluster by meaning:
    - "lupus" near "arthritis"
    - "lupus" far from "banana"
    - "methotrexate" near "leflunomide"
  - In GPT-5, each token vector ≈ 32,000 numbers
  - E.g.: Prior 7 tokens → 7x32,000 embedding sequence (matrix)

$$E = \begin{bmatrix} 0.12 & -0.45 & 0.07 & 0.33 & \cdots \\ 0.03 & 0.55 & -0.18 & 0.22 & \cdots \\ -0.78 & 0.11 & 0.40 & -0.05 & \cdots \\ 0.25 & -0.62 & 0.15 & 0.09 & \cdots \\ -0.14 & 0.34 & -0.22 & 0.17 & \cdots \\ 0.08 & 0.02 & 0.19 & -0.11 & \cdots \\ -0.31 & 0.28 & 0.04 & 0.13 & \cdots \end{bmatrix}$$

Pretraining

3. Attention: How words relate (more vectors)

- **Query (Q):** what a word is looking for
  - "lupus" may seek disease terms or treatments
- **Key (K):** what a word can match with
  - "nephritis" matches kidney disease-related queries
- **Value (V):** info the word contributes
  - "nephritis" = renal inflammation.

How it works:

- Queries search for matching Keys.
- Stronger match → more of that Value is used

Pretraining

4. Multi-Head Attention: Combining Different Perspectives

- Model uses many sets of Q, K, V (“heads”) in parallel
- Each head looks at the text in a different way, in parallel:
  - One head might capture grammar
  - Medical relationships (e.g., disease–symptom links)
  - Temporal info (timeline, before/after)
  - Numerical reasoning (lab values, counts)
  - Etc.
- Outputs combine → richer representation by tracking multiple relationships at once

Pretraining

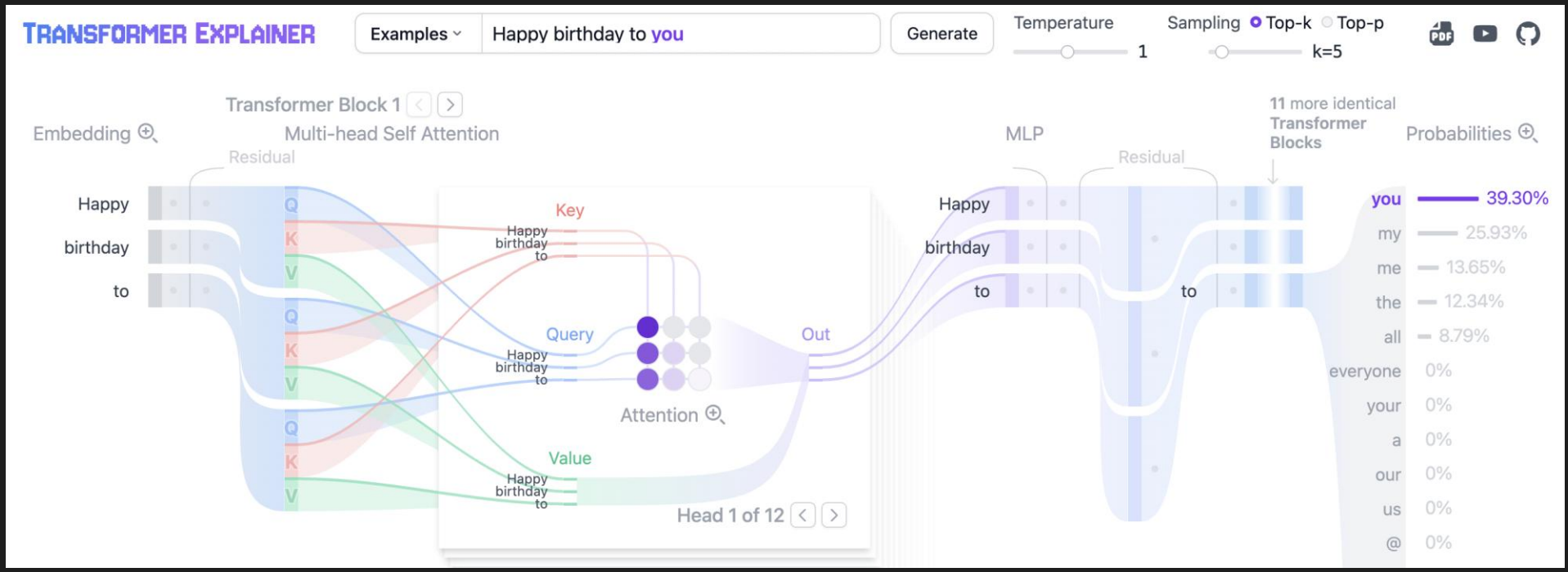
5. The Transformer Block: Deeper Understanding

- Each block has two parts:
  - Attention → decides which words matter
  - Feed-Forward Network → refines meaning
- Large models (like GPT-5) stack hundreds of these blocks
- Each layer builds on the last → richer context
- Final result: accurate next word prediction based on understanding

Example:

Happy birthday to _____

GPT-2 Transformer Example



- **MLP = Multi-Layer Perceptron = Feed-Forward Network**

https://poloclub.github.io/transformer-explainer/

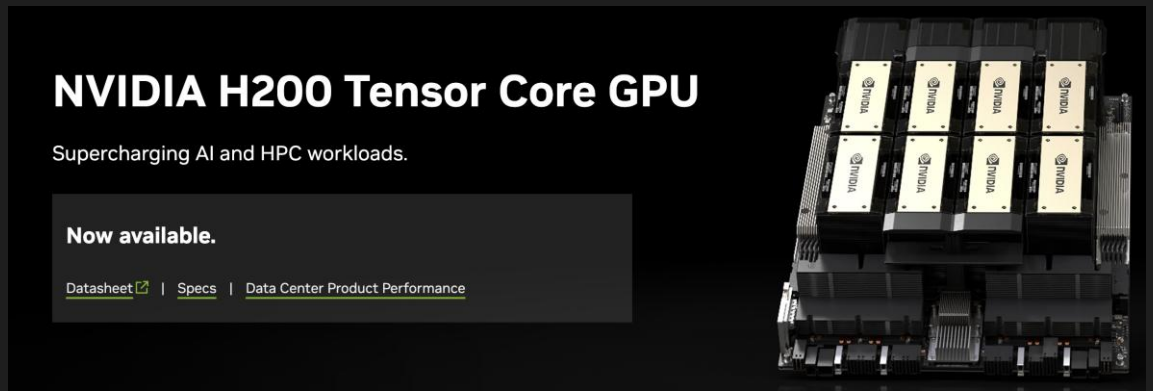
GPUs

- Why GPUs? Scale of math in AI needs specialized chips
  - CPU: few cores, sequential tasks
  - GPU = graphics processing unit
    - Originally designed for matrix math involved in video games
    - Thousands of cores, optimized for parallel matrix math
    - Speed measured in FLOPS (floating point operations per second)
- NVIDIA dominates AI market

GPUs

- NVIDIA RTX 5090 → ~838 TFLOPS (FP8), 600W

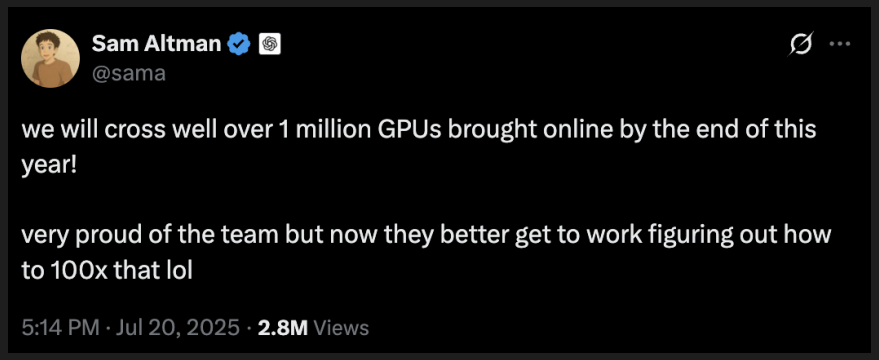
- AI GPUs:
  - H100/H200 (Hopper, 2022) → ~2000-4000 TFLOPS, 700W
  - B100/B200 (Blackwell, 2024-25) → ~6,600-10,000 TFLOPS, 700-1000W
  - Rubin (2027) → ~16,670 TFLOPS, 900W

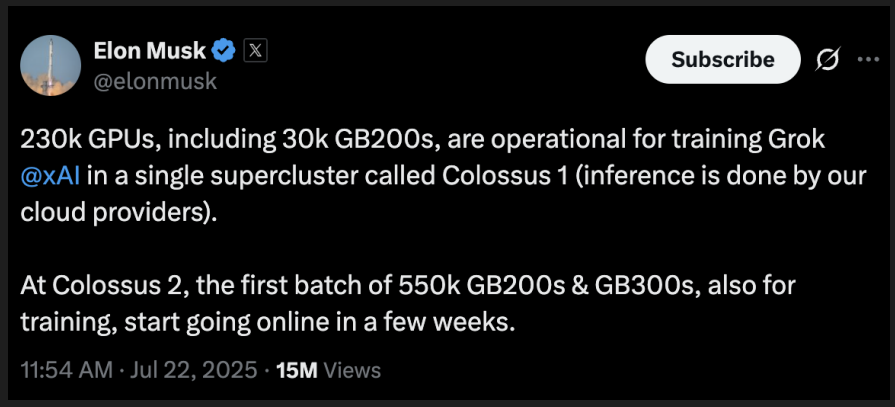


GPU Numbers & Energy Requirements

Best estimates:

- OpenAI: est ~1M
  - ≈700 MW → power of 200–300k city
- xAI: about 780k GPUs by end of Sept 2025
- Google & Microsoft: 500k-1M each
- Meta: 600-700k
- Amazon (AWS): 200-300k
- Anthropic: 20-50k





Evolution of AI Models

GPT-3 (2020)

- 175B parameters (number of learned weights, more weights → more understanding/nuance)
- First model to capture major public attention.
- Fluent text, but frequent hallucinations & weak reasoning.

GPT-3.5 (2022)

- Refinement of GPT-3.
- Better alignment and fewer nonsensical outputs.
- Powered first ChatGPT launch (Nov 2022).

Evolution of AI Models

GPT-4 (2023)

- Estimated 1–1.8T parameters.
- Major leap in reasoning & reliability.
- High scores on bar exam, SAT, and USMLE.

GPT-4o (2024)

- Same family as GPT-4.
- Multimodal: integrates text, vision, and audio.
- Real-time conversational capabilities.

GPT-o3 (2024)

- Optimized for reasoning efficiency.
- Achieves high scores on benchmarks.

Evolution of AI Models

GPT-5 (2025)

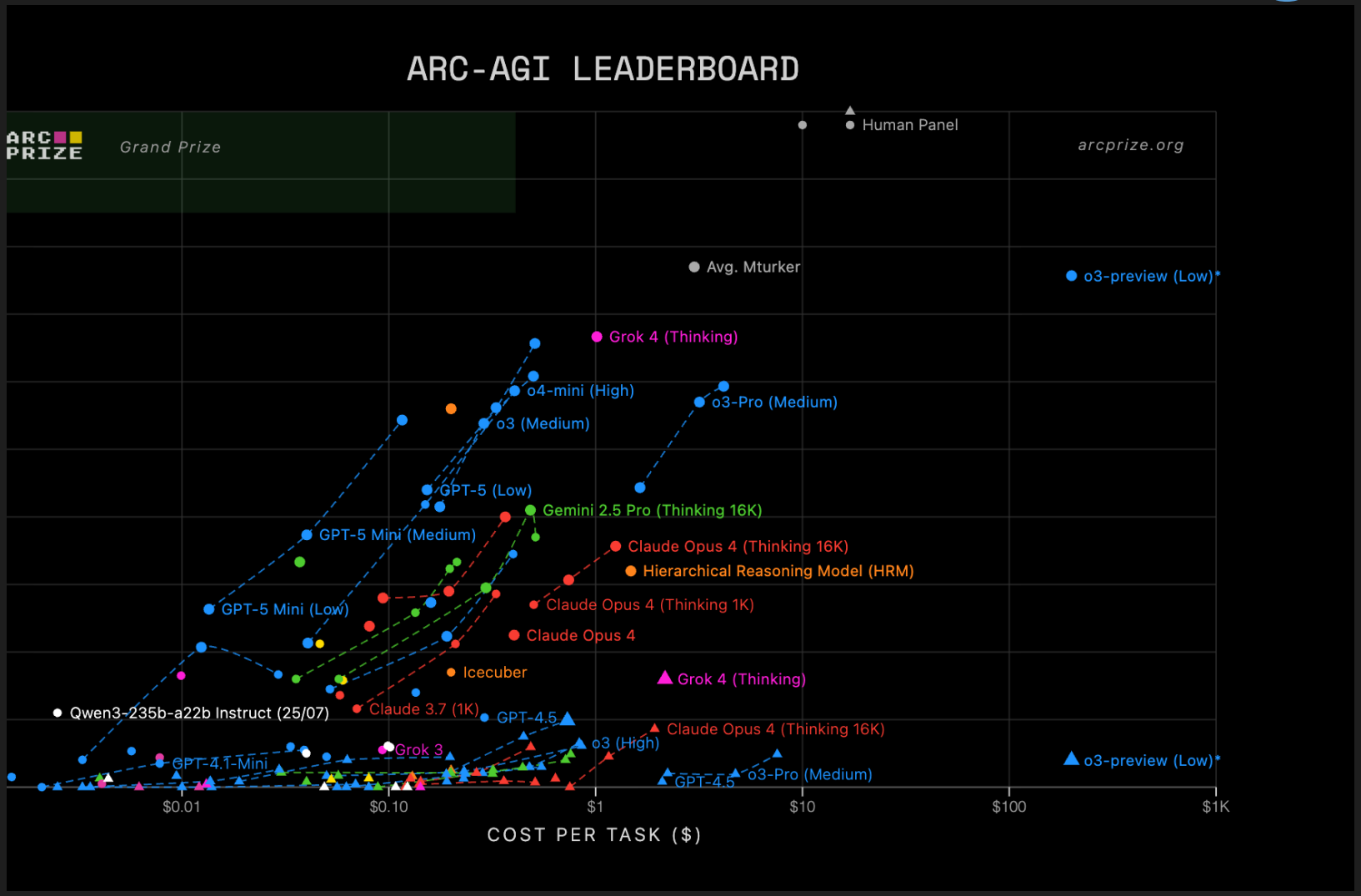
- Current frontier. >10T parameters?
- Stronger reasoning.
- Selects right model for the task.
- More efficient → faster, lower energy use

“A thing that I like to remind people in our company of is, no-one knows what happens next. … It’s very hard to predict.” — Sam Altman, CEO of OpenAI

How AI is Measured

Benchmark	Measures
ARC-AGI	Tests abstract reasoning & generalization
MMLU (Massive Multitask Language Understanding)	Factual knowledge across 57 academic subjects
GPQA	Complex science problem-solving
Humanity's Last Exam	Deep reasoning across broad domains
LLM Leaderboards	Aggregate across many tasks and models

ARC-AGI: Measures Reasoning



https://arcprize.org/leaderboard on 09/06/2025

MMLU: Academic Knowledge

Results independently reproduced by Kaggle. Learn more

Last updated September 2, 2025

#	Model	Score
1	Gpt-5-2025-08-07	93.5% ±0.4%
2	Claude-Opus-4-1-20250805	93.4% ±0.4%
3	Claude-Opus-4-20250514	92.8% ±0.4%
4	Gemini-2.5-Pro-Preview-06-05	92.4% ±0.4%
5	O3-2025-04-16	92.3% ±0.4%
6	O1-2024-12-17	91.9% ±0.5%
7	Grok-4-0709	91.6% ±0.5%

https://www.kaggle.com/benchmarks/open-benchmarks/mmlu on 09/06/2025

GPQA: Complex science problem solving

https://www.vellum.ai/llm-leaderboard on 09/06/2025

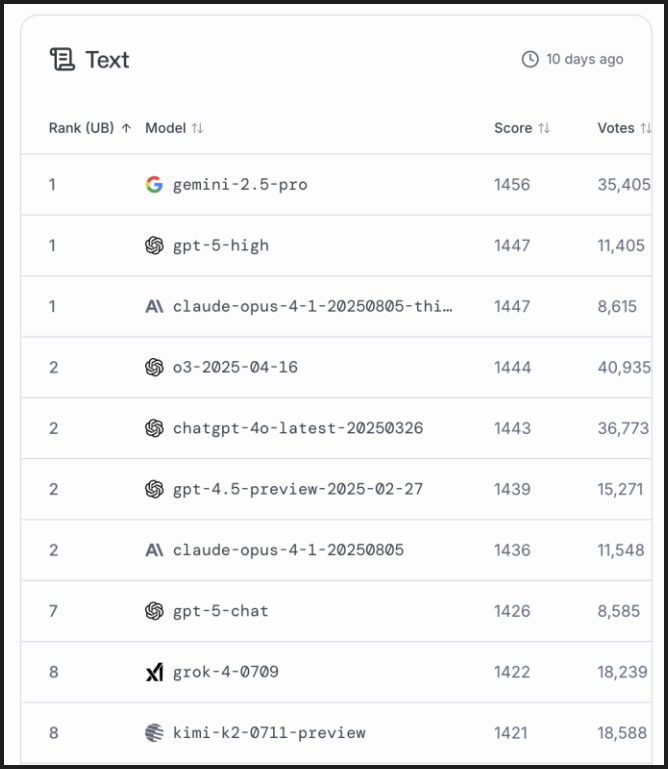
Humanities Last Exam: Deep reasoning

Model	Accuracy (%) ↑	Calibration Error (%) ↓
Grok 4	25.4	
GPT-5	25.3	50.0
Gemini 2.5 Pro	21.6	72.0
o3	20.3	34.0
o4-mini	18.1	57.0
DeepSeek-R1-0528*	14.0	78.0
o3-mini*	13.4	80.0
Gemini 2.5 Flash	12.1	80.0
Qwen3-235B*	11.8	74.0
Claude 4 Opus	10.7	73.0

https://agi.safe.ai/on 09/06/2025

Leaderboards

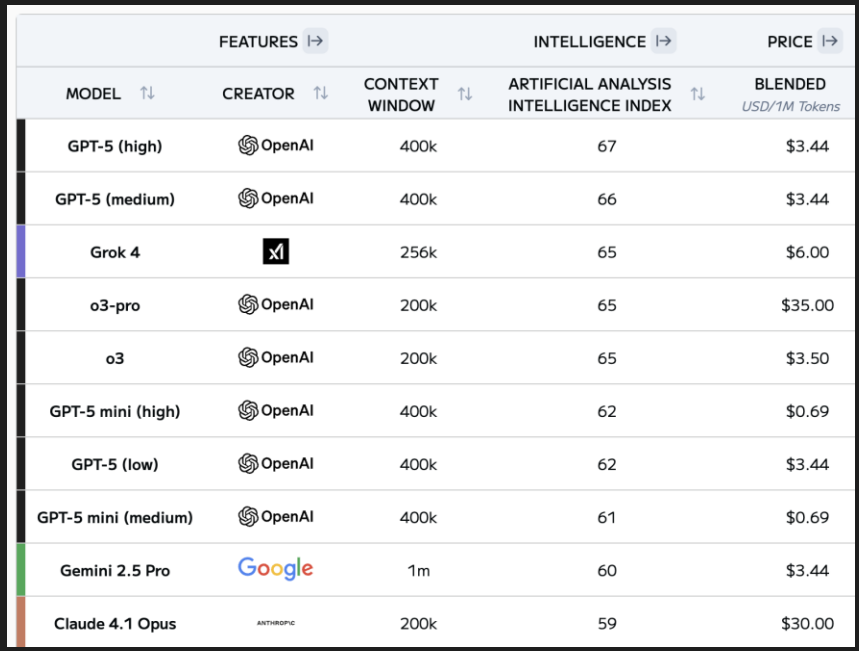
- LMArena.ai



Text — 10 days ago

Rank (UB)	Model	Score	Votes
1	gemini-2.5-pro	1456	35,405
1	gpt-5-high	1447	11,405
1	claude-opus-4-1-20250805-thi…	1447	8,615
2	o3-2025-04-16	1444	40,935
2	chatgpt-4o-latest-20250326	1443	36,773
2	gpt-4.5-preview-2025-02-27	1439	15,271
2	claude-opus-4-1-20250805	1436	11,548
7	gpt-5-chat	1426	8,585
8	grok-4-0709	1422	18,239
8	kimi-k2-0711-preview	1421	18,588

- ArtificialAnalysis.ai

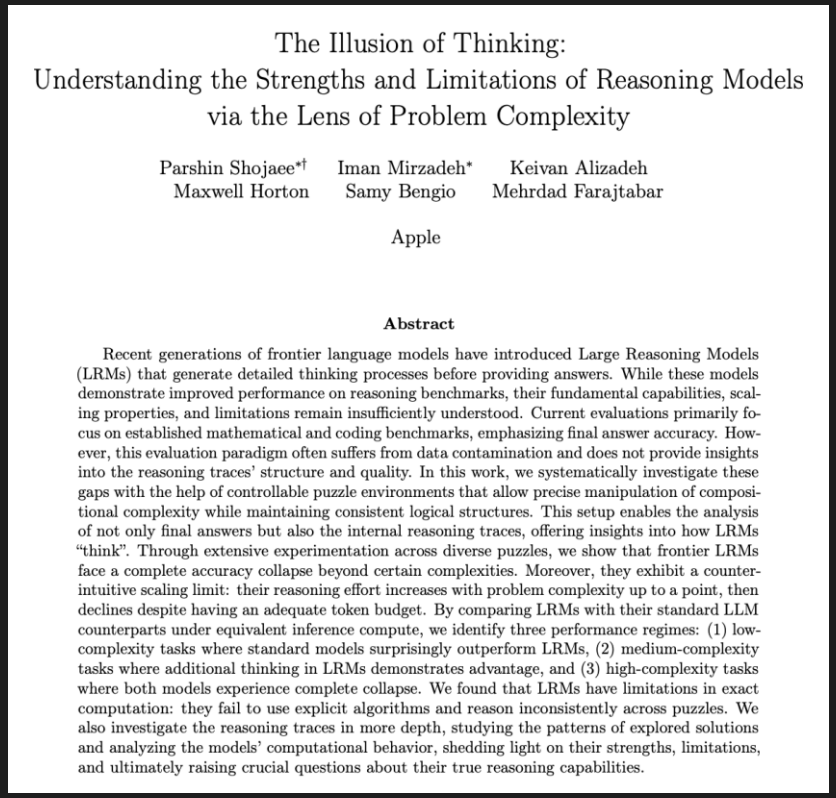


	FEATURES		INTELLIGENCE	PRICE
MODEL	CREATOR	CONTEXT WINDOW	ARTIFICIAL ANALYSIS INTELLIGENCE INDEX	BLENDED *USD/1M Tokens*
GPT-5 (high)	OpenAI	400k	67	$3.44
GPT-5 (medium)	OpenAI	400k	66	$3.44
Grok 4	xAI	256k	65	$6.00
o3-pro	OpenAI	200k	65	$35.00
o3	OpenAI	200k	65	$3.50
GPT-5 mini (high)	OpenAI	400k	62	$0.69
GPT-5 (low)	OpenAI	400k	62	$3.44
GPT-5 mini (medium)	OpenAI	400k	61	$0.69
Gemini 2.5 Pro	Google	1m	60	$3.44
Claude 4.1 Opus	ANTHROPIC	200k	59	$30.00

Data from 09/06/2025

Artificial General Intelligence (AGI)

- Not clear how to define AGI
  - Can perform any intellectual task a human can do?
  - Does it require creativity or ability to discover?
- Apple's 'Illusion of Thinking'
  - Emergent statistical effect ≠ Thinking



The Illusion of Thinking:
Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaee*† Iman Mirzadeh* Keivan Alizadeh
Maxwell Horton Samy Bengio Mehrdad Farajtabar

Apple

Abstract

Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights into the reasoning traces' structure and quality. In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs "think". Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counterintuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) low-complexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse. We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations, and ultimately raising crucial questions about their true reasoning capabilities.

https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

Using General AIs

- Many of them have a "*personality*"

- ChatGPT
  - GPT-4o: People liked that it was warm and chatty 😳. *Friendly professor.*
  - GPT-o3: Methodical and authoritative. *Focused engineer.*
  - GPT-5: Warm and pragmatic. *Helpful polymath.*
- Grok-4: Witty and blunt. *Irreverent savant.*
- Claude Sonnet 4: Eager and empathetic. *Thoughtful advisor.*
- Gemini 2.5 Pro: Optimistic and reflective. *Caring assistant.*

Getting Optimal Outputs: Prompt Engineering

For important tasks, use a structured prompt:

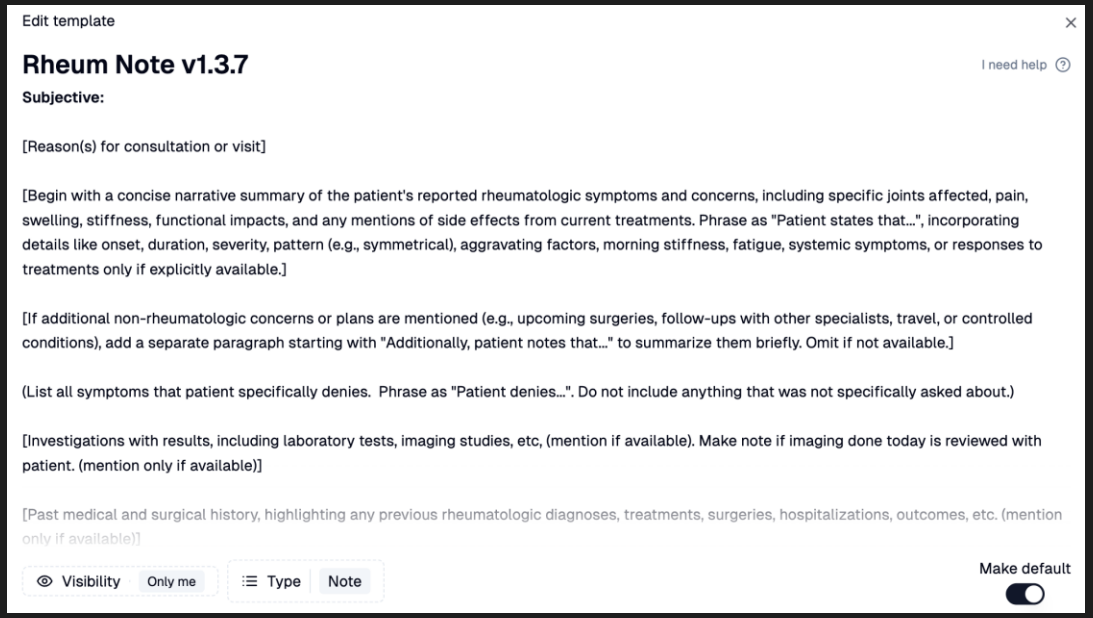
1. Role: Act as [specific role, e.g., “an experienced rheumatologist”]
2. Task: Your task is to [summarize / draft / analyze].
3. Context: Provide background, patient details, or constraints.
4. Format: Respond in [bullets / table / # of words/paragraphs].
5. Extras: Include [examples, reasoning steps, references].

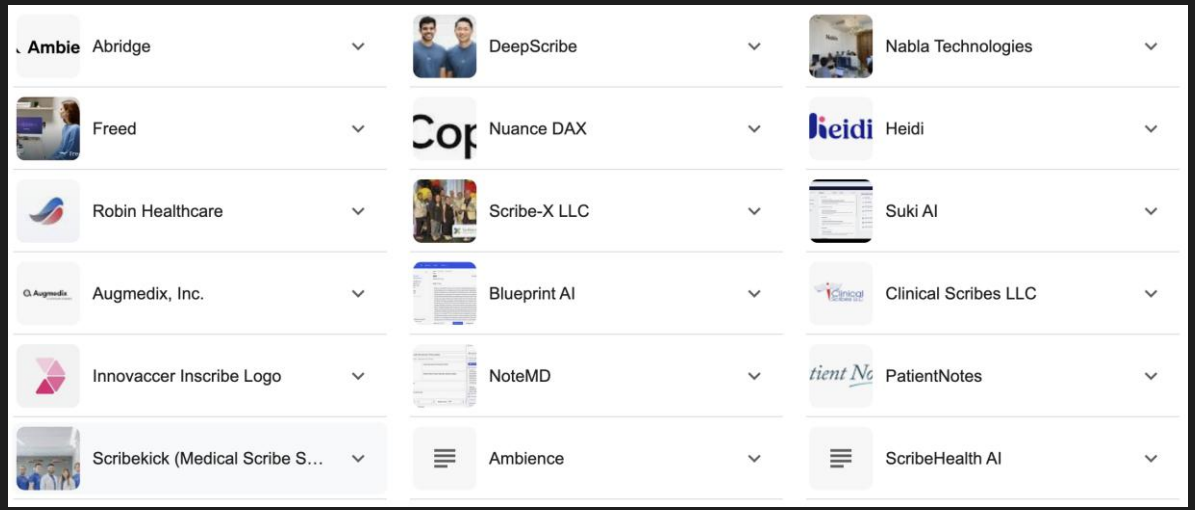
Current Uses in Medicine

- **Clinical Documentation & Notes** – ambient scribes
- **Decision Support** – e.g., OpenEvidence, Vera Health, UpToDate AI
- **Radiology & Pathology** – image interpretation, lesion detection
- **Dermatology & Ophthalmology** – skin cancer, diabetic retinopathy
- **Genomics & Drug Discovery** – protein/DNA modeling (AlphaFold, NVIDIA BioNeMo, OpenAI/Retro Biosciences)
- **Administrative Tasks** – scheduling, billing, prior auth
- **Patient Communication** – chatbots for triage, follow-up
- **Population Health** – identify at-risk patients, forecast admissions
- **Medical Education & Training** – AI tutors, simulated cases, exam prep
- **Clinical Research** – literature reviews, data extraction, trial recruitment

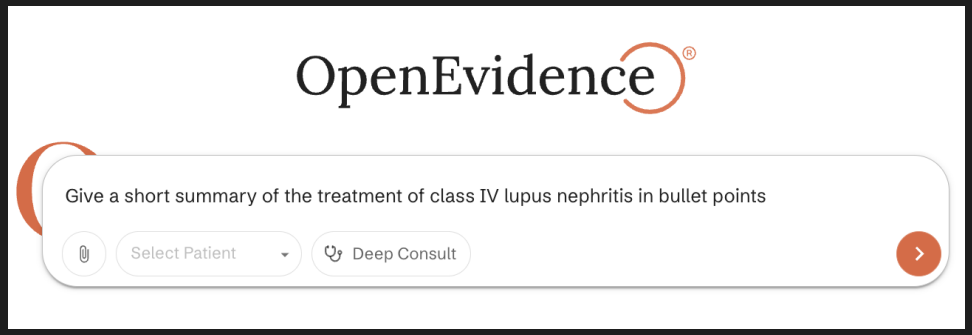
Ambient scribes

- Heidi Health ➡️
- **Biggest players with EHR integration:** Nuance DAX (Microsoft), Abridge, Suki.
- **Strong specialty/startup entrants:** Heidi Health, DeepScribe, Augmedix.





Decision Support



• **Initial therapy for class IV lupus nephritis (diffuse proliferative lupus nephritis) should begin promptly with a combination of immunosuppressive agents and glucocorticoids.**[1]

• The American College of Rheumatology recommends a **triple immunosuppressive regimen**:

• Pulse intravenous glucocorticoids (methylprednisolone 250–1000 mg daily for 1–3 days),

• Followed by oral glucocorticoids (≤0.5 mg/kg/day, maximum 40 mg/day) with a taper to ≤5 mg/day by 6 months,

• Plus one of the following:

• Mycophenolic acid analog (MPAA) plus belimumab,

• MPAA plus a calcineurin inhibitor (CNI),

• Euro-Lupus low-dose intravenous cyclophosphamide plus belimumab (with MPAA substituted for cyclophosphamide after completion of the cyclophosphamide course).[1]

• The regimen should be individualized based on patient factors such as kidney function, blood pressure, and chronicity on biopsy. MPAA-based regimens are generally preferred over cyclophosphamide due to a better toxicity profile.[1]

• Hydroxychloroquine should be initiated and continued unless contraindicated.[1]

• Renin-angiotensin-aldosterone system inhibitors (RAAS-I) are recommended for any degree of proteinuria.[1]

• Maintenance therapy after induction typically continues with MPAA; azathioprine is reserved for those intolerant of MPAA or planning pregnancy.[1]

References

1. 2024 American College of Rheumatology (ACR) Guideline for the Screening, Treatment, and Management of Lupus Nephritis.
   Sammaritano LR, Askanase A, Bermas BL, et al.
   Arthritis & Rheumatology (Hoboken, N.J.). 2025;. doi:10.1002/art.43212.
   ☆ Leading Journal ⓘ New Research
2. KDIGO 2024 Clinical Practice Guideline for the Management of LUPUS NEPHRITIS.
   Kidney International. 2024;105(1S):S1-S69. doi:10.1016/j.kint.2023.09.002.
   Practice Guideline ☆ Leading Journal ⓘ

Screenshot from 09/06/2025

Decision Support

vera health
for physicians

What biologic would be the best first line choice for a patient with Yao syndrome?

Answer | Drug Information 2 | References 4 | Steps

The best first-line biologic for **Yao syndrome** (autoinflammatory disease linked to **NOD2 mutations**) is an **IL-1 inhibitor**, most commonly **anakinra** or **canakinumab**, based on current evidence from related autoinflammatory disorders.

- **Rationale**: Yao syndrome shares pathophysiological overlap with other systemic autoinflammatory syndromes (e.g., TRAPS, MKD, CAPS, Still's disease), where **IL-1 blockade has shown consistent efficacy**.
  - A large RCT in *NEJM* (n=181) demonstrated that **canakinumab achieved complete response rates of 61–73%** across FMF, MKD, and TRAPS compared with 6–8% for placebo ($p<0.001$) 1 I.
  - A systematic review of IL-1 biologics in immune-mediated disorders confirmed **anakinra, canakinumab, and rilonacept** as effective and safe in autoinflammatory syndromes, with rapid resolution of systemic inflammation 2 I.

Ask a follow-up question...

Biologic option	Mechanism	Evidence in autoinflammatory syndromes	Clinical evidence
Anakinra	IL-1 receptor antagonist (daily SC)	Effective in CAPS, TRAPS, MKD, Still's disease	Rapid onset, but daily dosing, injection-site reactions 2
Canakinumab	Anti–IL-1β mAb (q4–8w SC)	RCT in FMF, MKD, TRAPS: 61–73% complete response vs 6–8% placebo 1	Strongest RCT evidence, convenient dosing, favorable safety
Rilonacept	IL-1 trap fusion protein (weekly SC)	Effective in CAPS, sJIA, pericarditis; less robust in head-to-head comparisons 4	Less effective than canakinumab or tocilizumab in sJIA

Conclusion:

For **Yao syndrome**, the **first-line biologic of choice is an IL-1 inhibitor**, with **canakinumab favored** due to robust RCT evidence, sustained remission, and convenient dosing. **Anakinra** remains a valid alternative in cases requiring rapid onset or where canakinumab is not available.

Screenshot from 09/06/2025

Imaging

- Google Health AI Developer Foundations

CXR Foundation Model

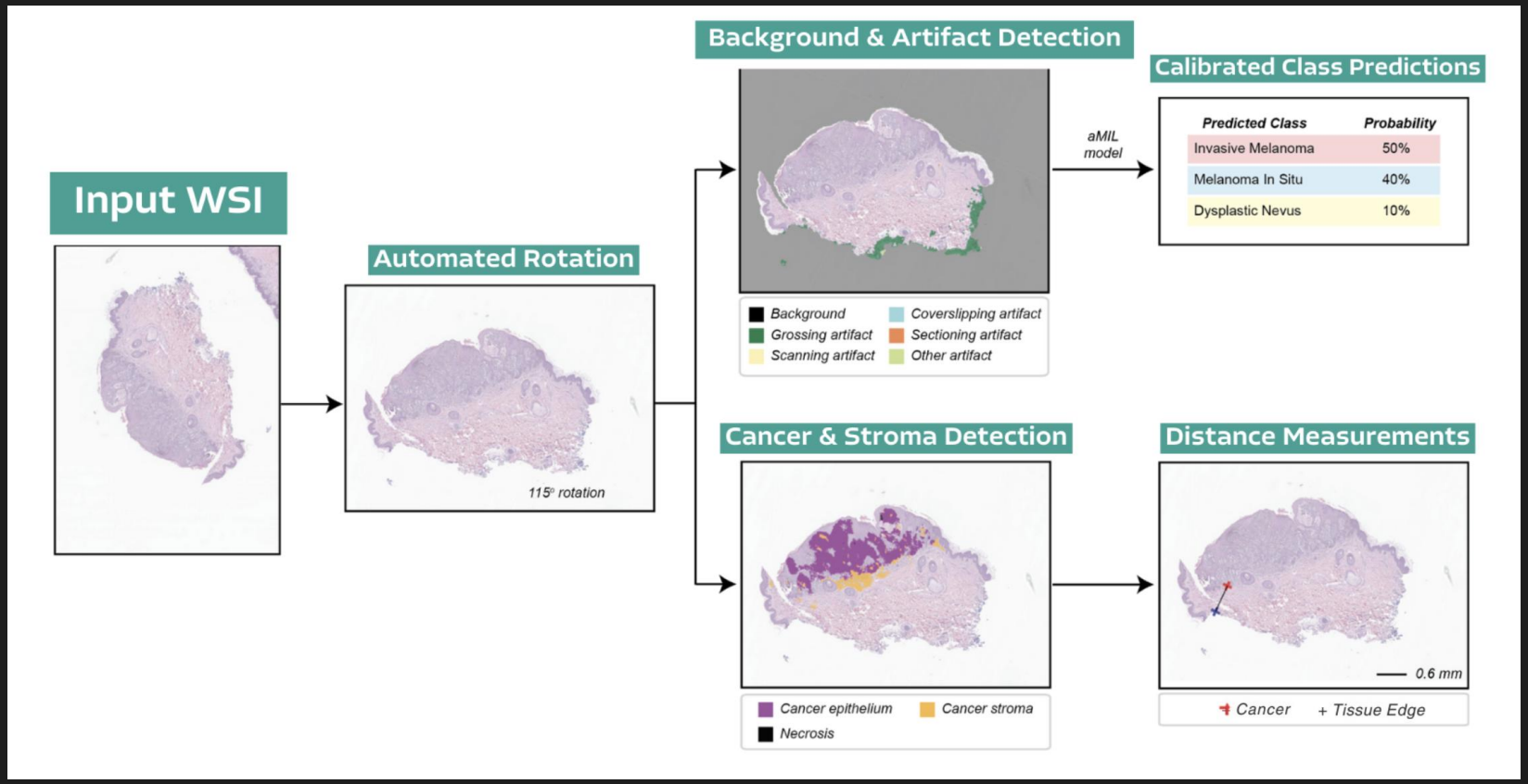
On this page

CXR Foundation is a machine learning (ML) model that produces embeddings based on images of chest X-rays. The embeddings can be used to efficiently build AI models for chest X-ray related tasks, requiring less data and less compute than having to fully train a model without the embeddings or the pretrained model.

https://developers.google.com/health-ai-developer-foundations/

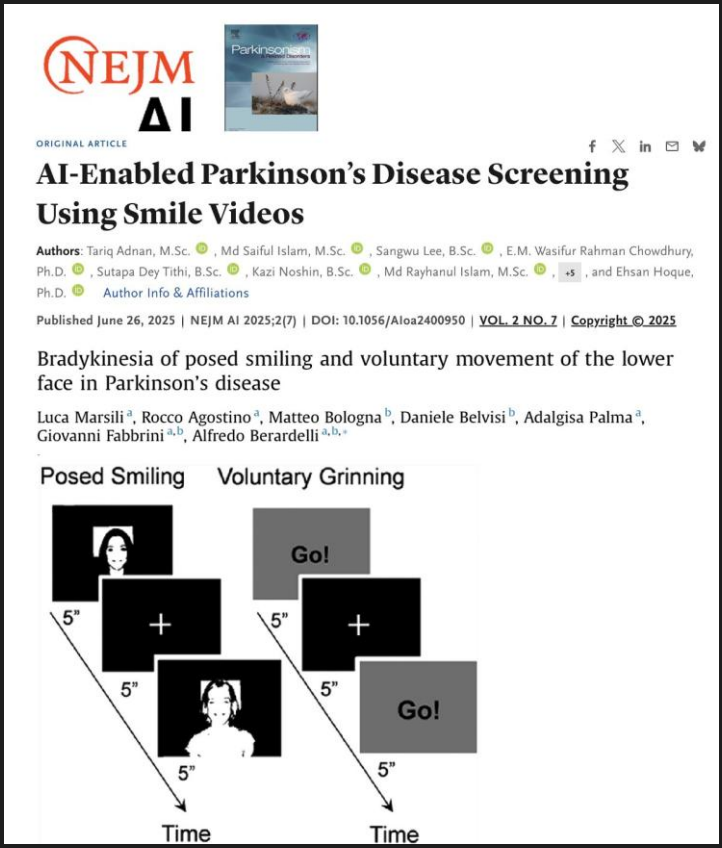
Pathology

- PathAI platform



https://www.pathai.com/

Physical Exam



NEJM AI

ORIGINAL ARTICLE

AI-Enabled Parkinson's Disease Screening Using Smile Videos

Authors: Tariq Adnan, M.Sc., Md Saiful Islam, M.Sc., Sangwu Lee, B.Sc., E.M. Wasifur Rahman Chowdhury, Ph.D., Sutapa Dey Tithi, B.Sc., Kazi Noshin, B.Sc., Md Rayhanul Islam, M.Sc., +5, and Ehsan Hoque, Ph.D. Author Info & Affiliations

Published June 26, 2025 | NEJM AI 2025;2(7) | DOI: 10.1056/AIoa2400950 | VOL. 2 NO. 7 | Copyright © 2025

Bradykinesia of posed smiling and voluntary movement of the lower face in Parkinson's disease

Luca Marsili[a], Rocco Agostino[a], Matteo Bologna[b], Daniele Belvisi[b], Adalgisa Palma[a], Giovanni Fabbrini[a,b], Alfredo Berardelli[a,b,*]

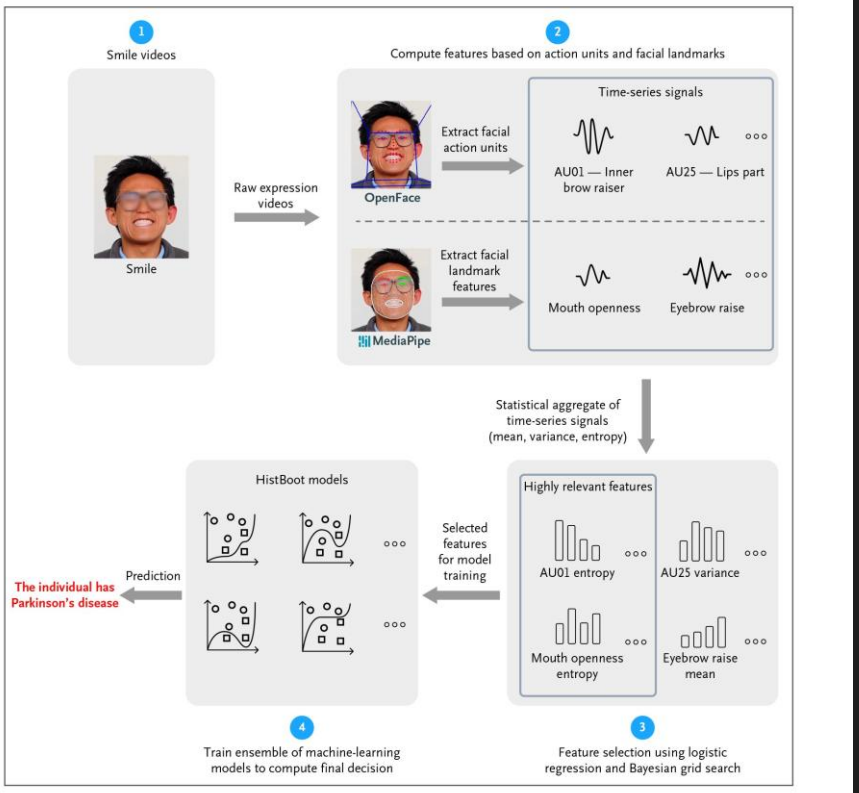


Figure 1. A Brief Overview of the AI-Based System for Classifying Individuals with and without PD.

- 1452 patients, 391 had Parkinsons, AI detected Parkinson's ~88% accuracy.

DOI: 10.1056/AIoa2400950

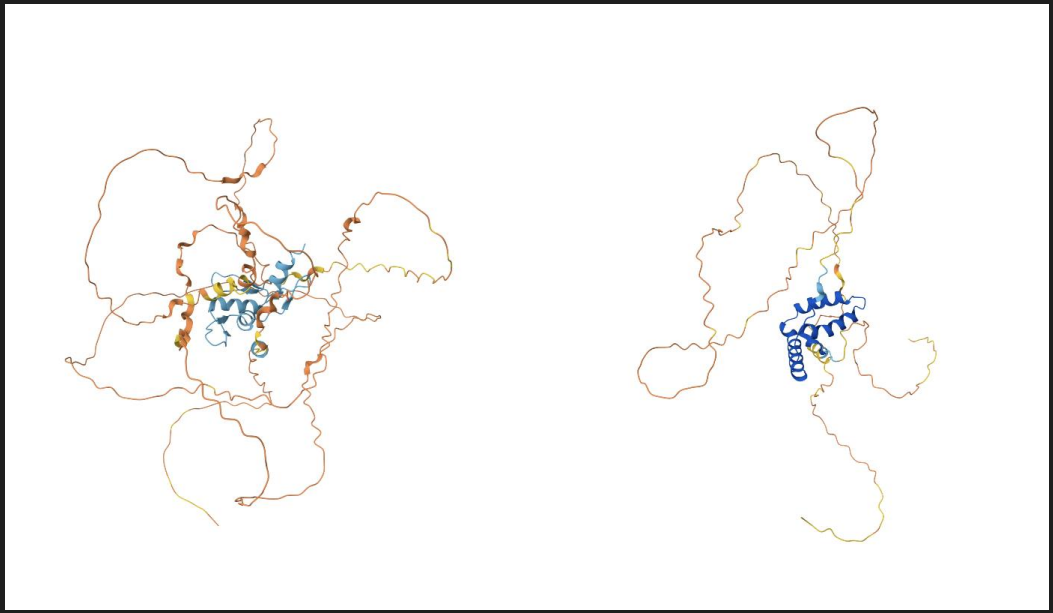
Protein Modeling

- Google AlphaFold
  - Predicts 3D protein structures from amino acid sequences
  - Accelerates biology and drug discovery

https://www.alphafold.ebi.ac.uk/entry/A0A089RQF2

Regenerative Medicine

- Yamanaka Factors (OCT4, SOX2, KLF4, c-MYC) – 2012 Nobel Prize discovery; reprogram adult cells into pluripotent stem cells
- Natural factors → inefficient & cause genomic instability → limits therapeutic use
- ~20 years of slow progress with traditional methods, due to astronomical number of possible protein variants to consider
- OpenAI + Retro Biosciences used GPT-4b micro to redesign SOX2 & KLF4 → 50× higher pluripotency in months



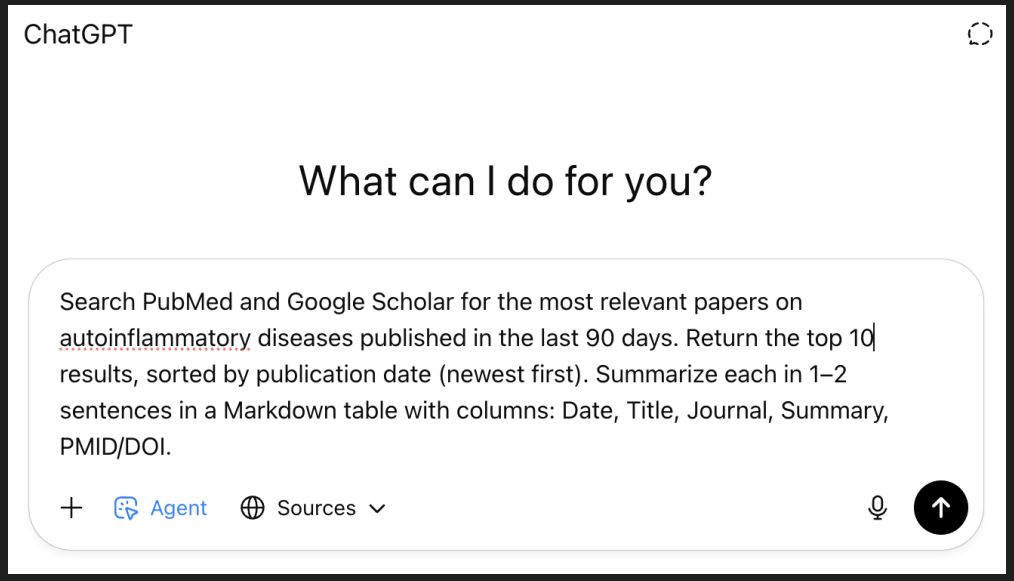
https://openai.com/index/accelerating-life-sciences-research-with-retro-biosciences/

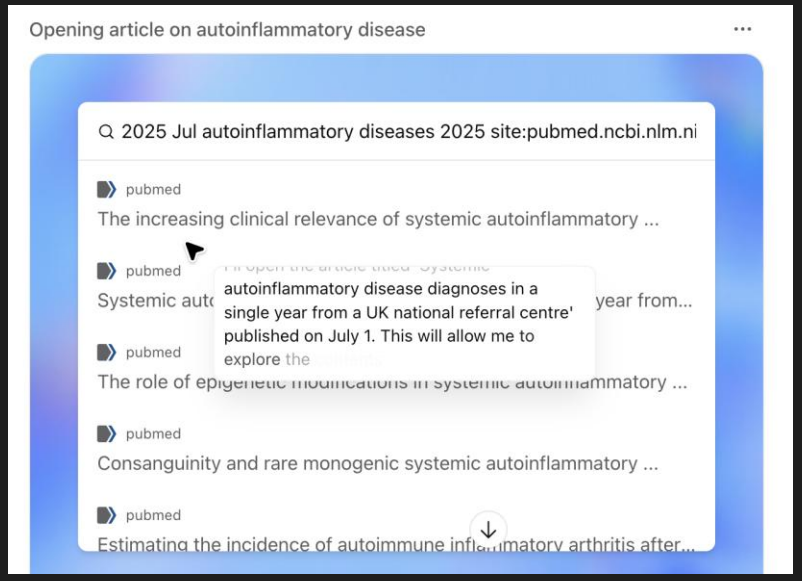
AI Agents

- Chatbot: answers one question
- Agents: carries out multi-step tasks — with human approval for critical steps

- **Email Agents** – manage inbox, draft/reply, summarize
- **Browser Agents** – navigate the web, gather info, fill forms
- **Coding Agents** – e.g., Cursor: write/debug code, run tests, suggest improvements
- **Office Productivity Agents** – schedule meetings, generate reports, coordinate tasks (Microsoft Copilot, Google Gemini apps)
  - Google Sheets: =AI(prompt, [range1], [range2], ...)
  - Excel: =COPILOT(prompt, [context1], [prompt_part2], [context2], ...)

ChatGPT Agent





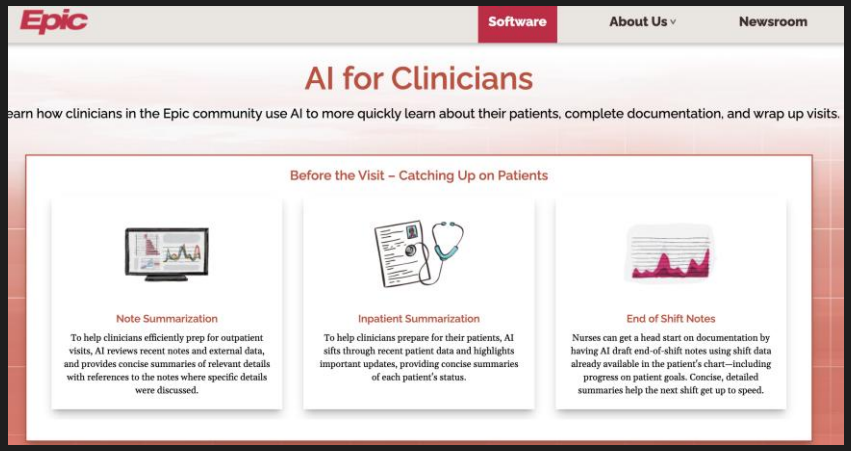
Worked for 11 minutes

Here are recent (≤90-day) autoinflammatory-disease papers sorted by publication date (newest first).
Each summary is based on the abstract and highlights the key findings and clinical implications.

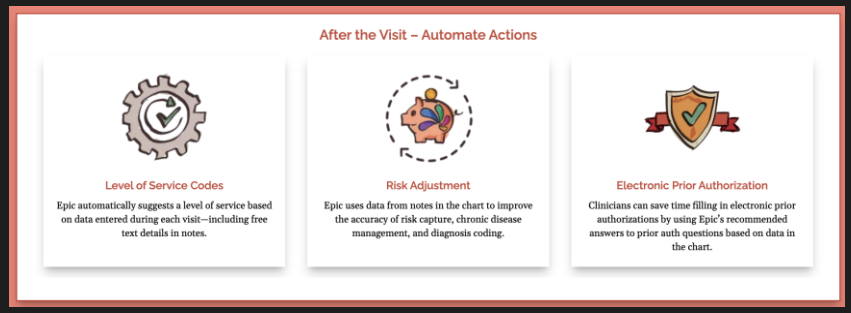
Date (2025)	Title & Journal	Summary (1–2 sentences)	PMID/DOI
Sept 4	**"From diagnostic uncertainty to targeted therapy: a case-based review of VEXAS syndrome," Rheumatology Int.**	This narrative case review presents two adult male patients with VEXAS syndrome (somatic *UBA1* mutations) whose multisystem inflammation and cytopenias were initially puzzling; genetic testing confirmed the diagnosis and treatment with corticosteroids plus biologics (IL-6 or IL-1 blockade) led to disease control. The review emphasises the importance of a high index of suspicion, early genetic testing and the use of IL-1/IL-6 inhibitors as steroid-sparing therapy for this late-onset autoinflammatory disease pubmed.ncbi.nlm.nih.gov.	**PMID 40905972**, DOI 10.1007/s0029
Aug 17	**"VEXAS syndrome: genetics, gender differences, clinical insights, diagnostic pitfalls, and**	This comprehensive review notes that VEXAS arises from somatic *UBA1* mutations and, contrary to early reports, heterozygous females can develop severe disease; typical	**PMID 40869252**, DOI 10.3390/ijms2

EHR Agents

- **Epic**: Agents for clinicians, revenue cycle, and patients



- **Oracle Health** (Cerner): Voice-enabled assistant: can listen in on visits, draft notes, and place orders



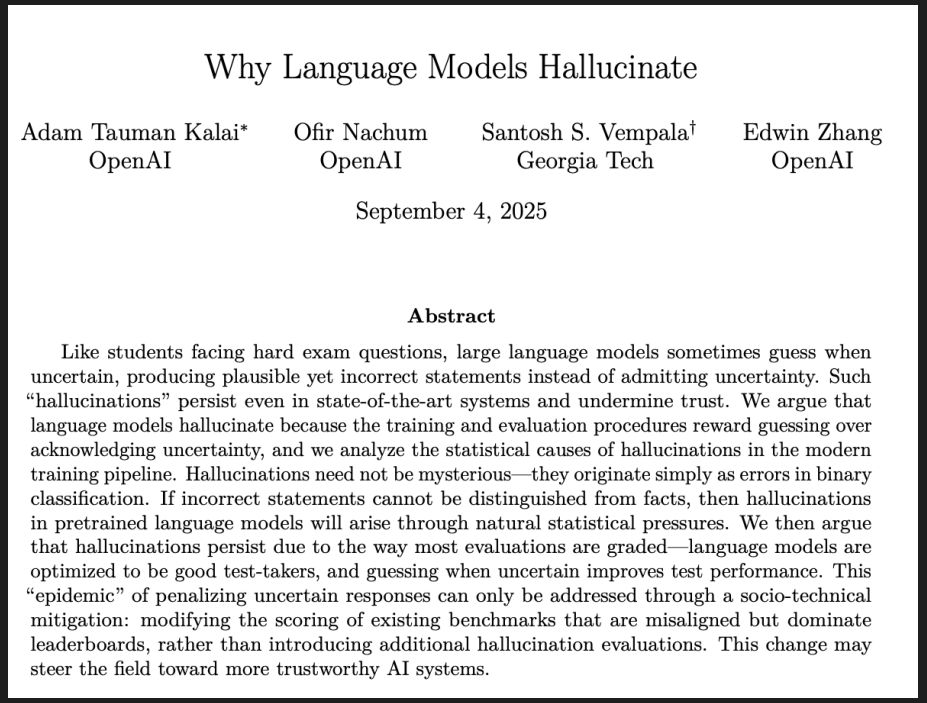
https://www.epic.com/software/ai-clinicians/
https://www.oracle.com/health/clinical-suite/clinical-ai-agent/

AI Risks in Medicine

- **Errors & Hallucinations**
  - Fabricated references or “confidently wrong” answers
  - Risk of diagnostic or treatment errors if unchecked
  - Can amplify existing mistakes in documentation
- **Legal & Liability Issues**
  - Who is responsible if AI advice leads to harm — physician, institution, or vendor?
  - Malpractice implications still undefined
  - Regulatory standards (FDA, CMS, state boards) still evolving
- **Ethical Concerns**
  - Transparency: “black box” models make decisions hard to explain
  - Patient consent & trust in use of their data

Hallucinations

- Hallucinations = statistical artifacts (from training), not bugs
- Current models reward guessing
- Fix: Reward admitting uncertainty?



Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala† Georgia Tech
Edwin Zhang OpenAI

September 4, 2025

Abstract

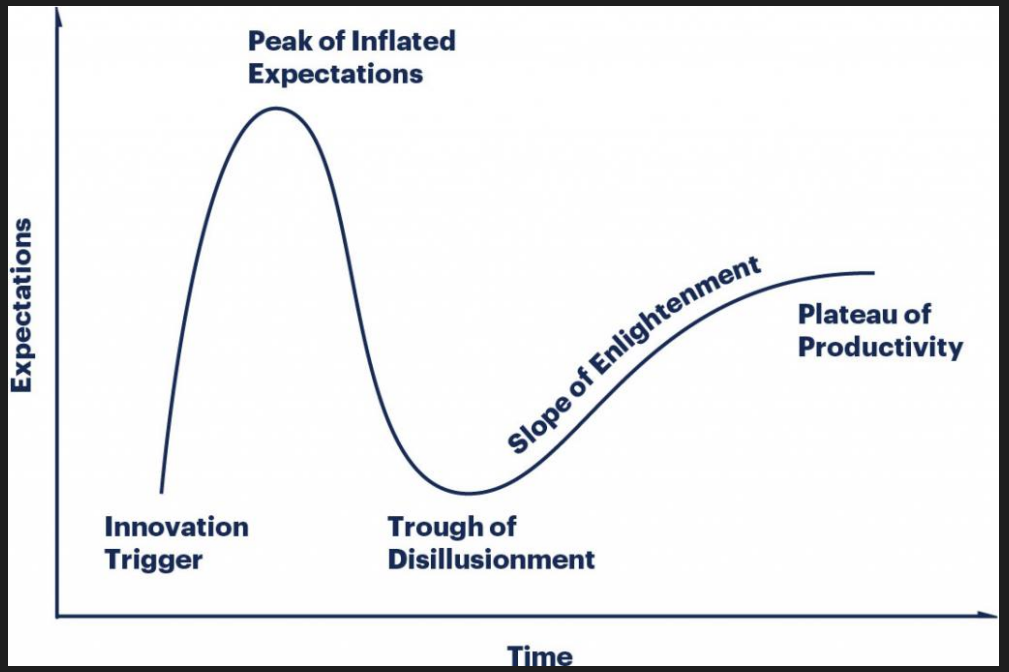
Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

https://openai.com/index/why-language-models-hallucinate/

Hype Cycle?

- We are in a hype cycle, but where are we?

- GPT-5, Grok-4, Claude Sonnet 4: Trough of disillusionment
- Gemini 2.5 Pro: Peak of Inflated Expectations

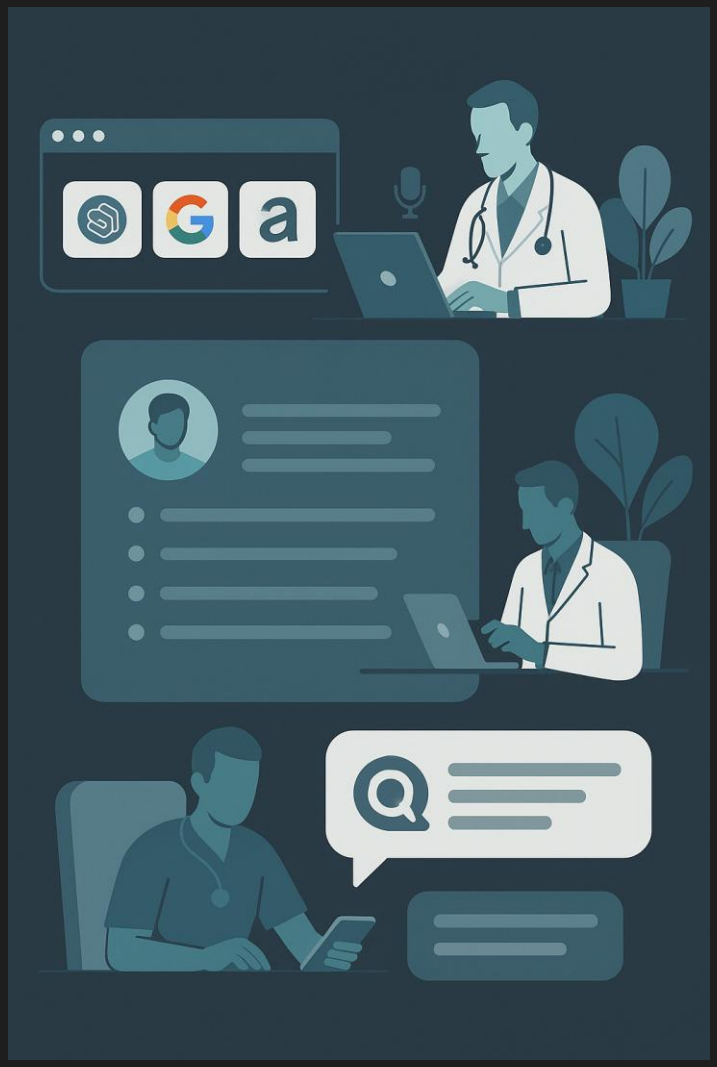


Future Directions

1. AI embedded in daily tech + EHRs
2. Agentic workflows
3. Multimodal inputs (history, labs, imaging, pathology, genomics)
4. Regulatory evolution and guidance.
5. Workforce impact: efficiency ↑, burnout ↓

Practical Tips

1. Set up 2+ AI accounts (ChatGPT, Gemini, Claude, Grok)
2. Test an ambient scribe (check with your IT team, or use AI to help pick one)
3. Try decision tools (OpenEvidence, Vera Health)
4. Practice structured prompting for better results
5. Use AI to summarize journal articles



Discussion / Q&A