Frontier Large Language Models (LLMs) for the Practicing MFM and Genetics Clinician: A Practical Guide to Model Selection in 2026

Christian Macedonia, M.D.

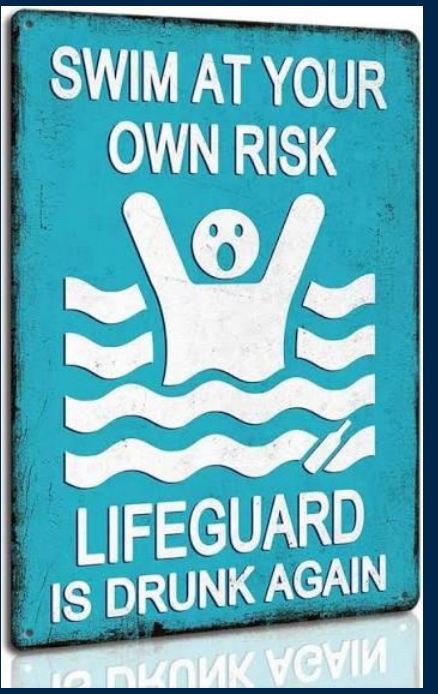
University of Michigan

USA

CEO Lancaster Maternal Fetal Medicine LLC

https://thektproject.org/

Disclosure and Safety Statement



- No financial relationships or conflicts of interest to disclose.
- The following discussion is educational and does not constitute endorsement of any commercial product.
- Audience members remain responsible for compliance with institutional policies, HIPAA/GDPR requirements, and applicable privacy regulations.
- LLM outputs should be independently verified before use in clinical care.
- Never upload identifiable patient data
- Assume every prompt is discoverable

This talk can be found at:

https://thektproject.org/

On many *standardized medical reasoning tasks*, frontier models now *outperform* most physicians

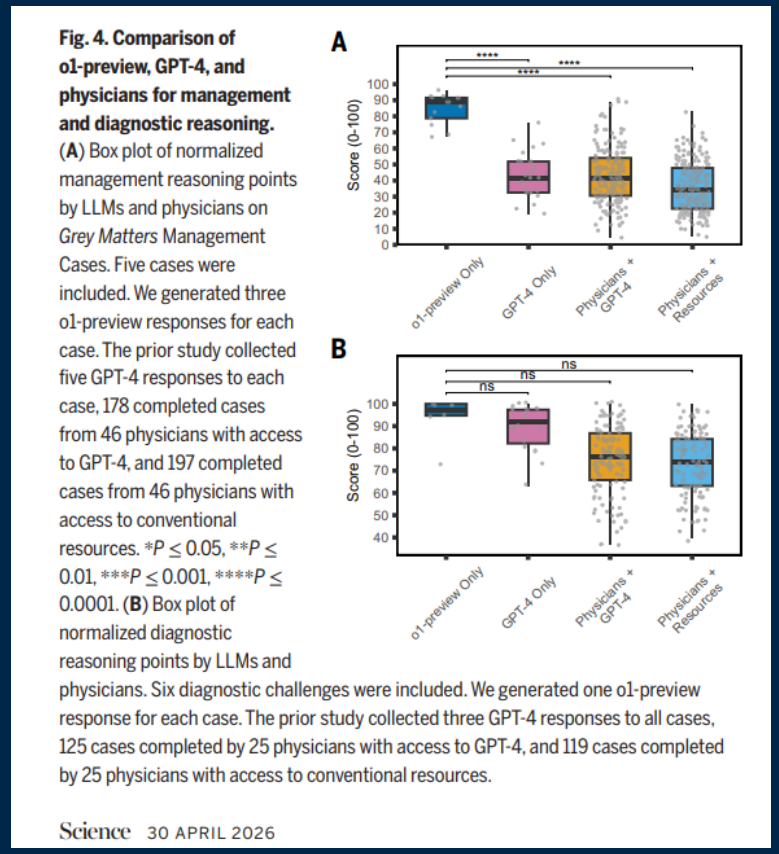


Fig. 4. Comparison of o1-preview, GPT-4, and physicians for management and diagnostic reasoning. (**A**) Box plot of normalized management reasoning points by LLMs and physicians on *Grey Matters* Management Cases. Five cases were included. We generated three o1-preview responses for each case. The prior study collected five GPT-4 responses to each case, 178 completed cases from 46 physicians with access to GPT-4, and 197 completed cases from 46 physicians with access to conventional resources. $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$, $****P \leq 0.0001$. (**B**) Box plot of normalized diagnostic reasoning points by LLMs and physicians. Six diagnostic challenges were included. We generated one o1-preview response for each case. The prior study collected three GPT-4 responses to all cases, 125 cases completed by 25 physicians with access to GPT-4, and 119 cases completed by 25 physicians with access to conventional resources.

Science 30 APRIL 2026

Peter G. Brodeur et al. , *Performance of a large language model on the reasoning tasks of a physician*. **Science** 392,524-527 (2026). *DOI:10.1126/science.adz4433*

NEJM AI

DOI: 10.1056/AIcs2501343

CASE STUDY

LLM-Assisted Reanalysis of Unsolved Rare Disease Genomes Increases Diagnostic Yield

Aaron Jaech, Ph.D.,[1] Morgan Cheatham, M.D.,[2] Suyash S. Shringarpure, Ph.D.,[1] Casie A. Genetti, M.S.,[4] Pratiksha Pradhan, M.S.,[4] Aarti Bagul, M.S.,[1] Trishan Panch, M.D.,[5] Benjamin Rader, M.D., M.P.H.,[6,7] Kara Sewalk, M.P.H.,[6] Rebecca Distler, M.P.H.,[1] Katherine N. Anderson, M.S.,[4] Melissa Stewart, M.P.H.,[6] Sarah Lejfer, M.P.H.,[6] David C. Glahn, Ph.D.,[3,8,9,10] Richard D. Goldstein, M.D.,[3,11] Monica Wojcik, M.D., M.P.H.,[3,4] Alan H. Beggs, Ph.D.,[3,4] John S. Brownstein, Ph.D.,[3,6] and Catherine A. Brownstein, Ph.D., M.P.H.[3,4,6,8,11]

Received: December 4, 2025; Revised: February 19, 2026; Accepted: March 25, 2026; Published: June 18, 2026

Abstract

BACKGROUND Rare and undiagnosed genetic disorders affect millions of patients globally, and many patients endure years of inconclusive testing. Conventional genomic interpretation can be insufficiently sensitive and costly and is rarely repeated as knowledge evolves.

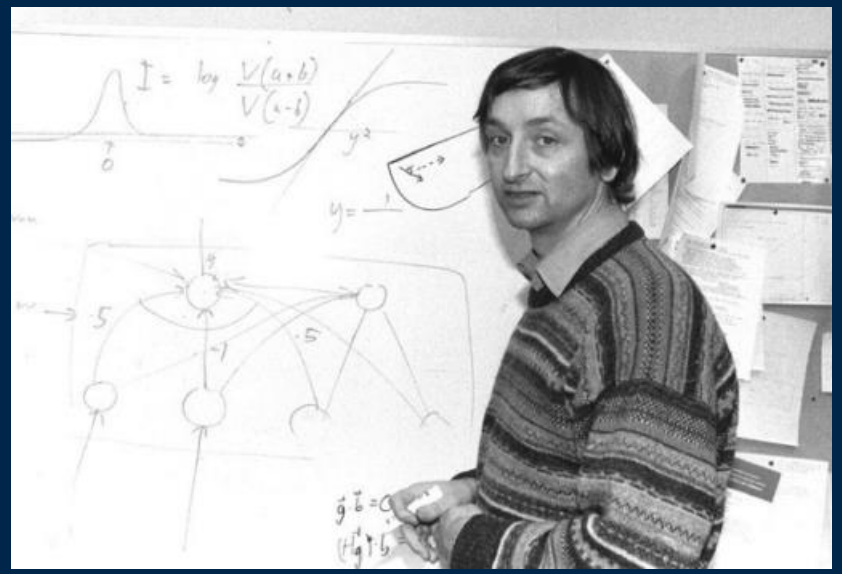
METHODS We conducted a retrospective multicohort reanalysis using a large language model (LLM)–assisted workflow that ingests clinician notes, Human Phenotype Ontology (HPO) terms, and a filtered variant table to propose explanation-rich candidate hypotheses for expert adjudication under American College of Medical Genetics and Genomics and Association for Molecular Pathology criteria. A diagnosis was defined a priori as a variant classified as pathogenic or likely pathogenic, confirmed in a Clinical Laboratory Improvement Amendments–certified laboratory, and returned to families. Secondary outputs included "rediscoveries" of externally established diagnoses not yet available locally and hypothesis generation signals.

Methods

Briefly, annotated variant call files (VCFs) from existing short- and long-read exome and genome datasets were uploaded for patients who had undergone sequencing to diagnose suspected genetic conditions.[11] These were paired with metadata files containing deidentified basic demographic and clinical data, including detailed phenotypic information encoded as HPO codes. The data were then analyzed using OpenAI's o3-deep-research model via the OpenAI application programming interface (API) or the ChatGPT (Chat Generative Pretrained Transformer) (web/app) user interface (selecting Research/Deep Research mode). Analyses used only standard platform tools (sandboxed Python/Code Interpreter and web search enabled) and were limited to user-provided files and publicly available web resources. LLM prompt development was iterative, eventually settling on two versions used to generate the data, discussed below. A sample output is provided in Figure 1.

Which of the Models to Choose From?

- 2022: One major *frontier* model … GPT3



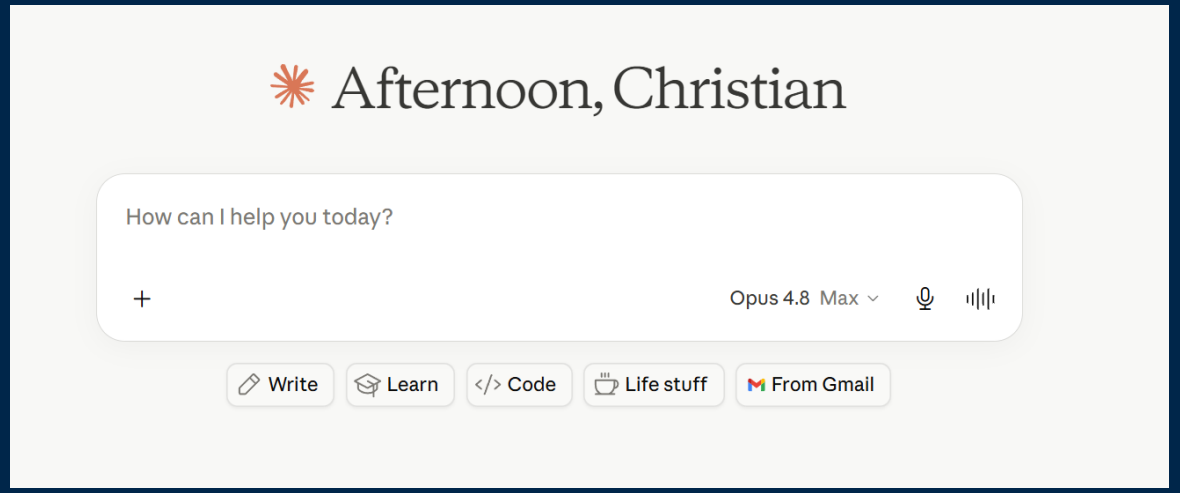
- 2026:
  - GPT
  - Claude
  - Gemini
  - Grok
  - Le Chat
  - DeepSeek
  - OpenEvidence
  - Replit
  - OpenClaw / local models (LLAMA/Meta -- Gemma/Google --DeepSeek R1)

Key message: "*The best model depends on the job.*"

Def. Frontier Model: A big AI model that is general purpose and sits at the frontier of capability

Common Medical Uses of AI

- Examples:
  - Differential diagnosis brainstorming
  - Literature summarization
  - Drafting patient letters
  - Insurance appeals
  - Genetic syndrome review
  - Ultrasound report editing
  - Coding support
  - Research assistance
  - Presentation preparation
  - Manuscript review
  - Grant writing assistance
  - Statistical interpretation
- Not:
  - Autonomous diagnosis
  - Autonomous medical decision making



But remember: AI is in the background, everywhere, in your machines, in your pocket, in your life…

Model		Usage Rank		Scale	Medical Utility	Cost	Advantages	Limitations
GPT		#1		Frontier	Excellent	$$	Best generalist, strong reasoning, projects	Can be verbose
Claude		#2		Frontier	Excellent	$$	Long documents, writing, nuanced analysis	Less integrated ecosystem
Gemini	Gemini	#3		Frontier	Very Good	$$	Massive context window, Google ecosystem	Occasionally inconsistent reasoning
Grok	Grok	#4		Frontier	Good	$$	Current events, internet access	Less mature medical workflow
Le Chat		#5		Large	Good	$	Fast, efficient	Smaller ecosystem
DeepSeek	deepseek	#6		Large	Good	$	Exceptional value	Variable reliability
OpenEvidence	OpenEvidence	Specialty		Proprietary	Excellent	$$	Medical evidence focused	Narrow scope
Replit AI	replit	Specialty		N/A	Moderate	$$	Automation, workflow creation, vibe coding	Not a medical model

	Claude Mythos 5 / Fable 5	Claude Mythos Preview	Claude Opus 4.8	GPT 5.5	Gemini 3.1 Pro
Agentic coding SWE-Bench Pro	**80.3%**	77.8%	69.2%	58.6%	54.2%
Agentic coding FrontierCode (Diamond)	**29.3%** xhigh	—	13.4% xhigh	5.7% xhigh	—
Knowledge work GDPval-AA	**1932**	—	1890	1769	1314
Knowledge work vision GDP.pdf	**29.8%** no tools	—	22.5% no tools	24.9% no tools	16.7% no tools
Spatial reasoning Blueprint-Bench 2	**38.6%**	—	14.5%	36.2%	26.5%
Tool use AutomationBench	**17.4%**	—	15.5%	12.9%	9.6%
Computer use OSWorld-Verified	85.0%	**85.4%**	83.4%	78.7%	76.2%
Legal Legal Agent Benchmark	**13.3%**	—	10.4%	2.1%	0.0%
Multidisciplinary reasoning Humanity's Last Exam	**59.0%*** no tools	56.8% no tools	49.8% no tools	41.4% no tools	44.4% no tools
	64.5%* with tools	**64.7%** with tools	57.9% with tools	52.2% with tools	51.4% with tools
Biology BioMysteryBench	**46.1%*** hard	29.6% hard	40.0% hard	—	—
	83.9%* human solved	82.6% human solved	80.4% human solved	—	—
Agentic coding Terminal-Bench 2.1	**88.0%***	—	82.7%	83.4% Codex CLI	70.7% Gemini CLI
Cybersecurity ExploitBench (Cap%)	**78.0%***	69.0%	40.0%	34.0%	—
Health HealthBench Professional	**66.0%***	64.7%	56.9%	51.8%	—

Methodology: Reported scores are within a 1-3 percentage point difference for Claude Mythos 5 and Claude Fable 5. This table shows the higher score of the two. Starred (*) benchmarks show a larger difference due to our blocking safeguards for cybersecurity and biology-related questions. For these benchmarks, Claude Fable 5 performs closer to Claude Opus 4.8 due to fallbacks. See the system card for details.

Source:
https://www.anthropic.com/news/claude-fable-5-mythos-5

What Matters More Than the Model…*The Prompt*

Prompting: the process of providing context, goals, constraints, and examples that help the model produce useful output

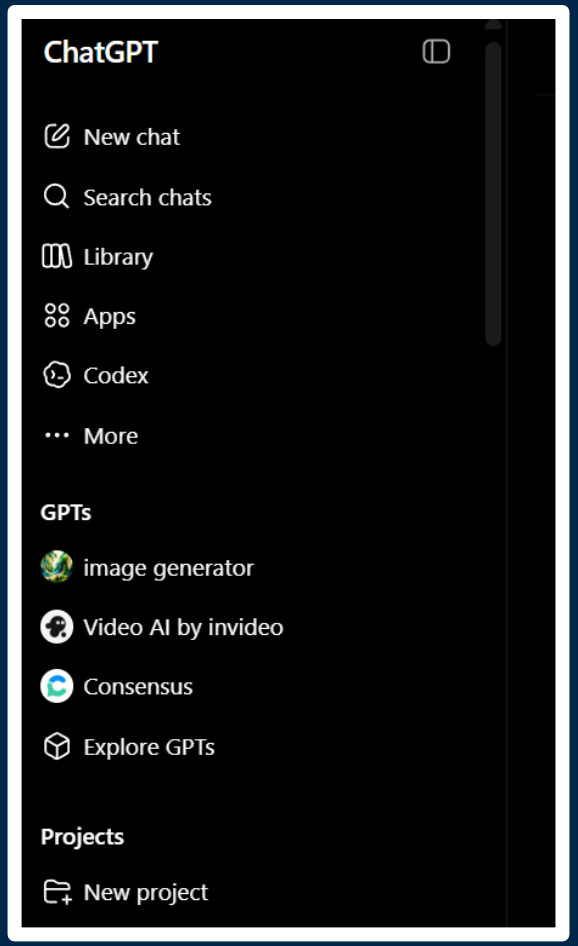
- Mediocre Prompt:
  - "Tell me about Noonan syndrome.“
- Better Prompt:
  - “Summarize Noonan syndrome for MFM counseling note that will go to the patient’s ob provider (who knows me and works with me frequently). Include prenatal findings that are suggestive of this diagnosis, talk about recurrence risk, testing options, and postnatal outcomes. Keep it to 3 paragraphs. Use bullet points.”

In 2026, Prompt Engineering is becoming a critical clinical skill.

Customizing Models

…Most clinicians have no idea this capability exists.

- Topics:
  - Projects
  - Persistent instructions
  - Dedicated workspace
  - Reusable references
- Knowledge Files
- Upload Examples:
  - Practice protocols
  - FMF guidelines
  - SMFM documents
  - Genetic counseling templates
  - Custom Instructions



Example:

"You are assisting a maternal-fetal medicine physician. Use concise language. Cite evidence when possible."

Feature	Purpose	Example for MFM/Genetics
Custom Instructions	Define role and style	"Respond as a board-certified MFM specialist using concise clinical language."
Project Folders	Maintain persistent context	MFM Practice, Genetics Consults, Research Projects
Knowledge Files	Provide trusted references	SMFM Consults, FMF protocols, ACMG guidelines
Markdown (.md) Files	Encode reusable instructions and workflows	FGR workup template, cfDNA counseling template
Reference Libraries	Reduce hallucinations	Practice policies, published guidelines, counseling scripts
Memory & Continuity	Preserve ongoing discussions	Longitudinal research or manuscript development

Take Home Points

- There is no single best model. Use more than one model for important work and compare outputs.
- Match the model to the task.
- Prompt quality matters more than model selection.
- Use projects and knowledge files.
- Verify outputs. You and only you are responsible for patient care. Legally and ethically.
- AI is becoming a routine clinical and research enhancement capability.

AI is not your replacement and not simply a tool. AI is like an exosuit. AI is an enhancement that can make you a more useful and effective human. AI cannot choose that for you. Only you can.

Thank You

This talk can be found at:

https://thektproject.org/