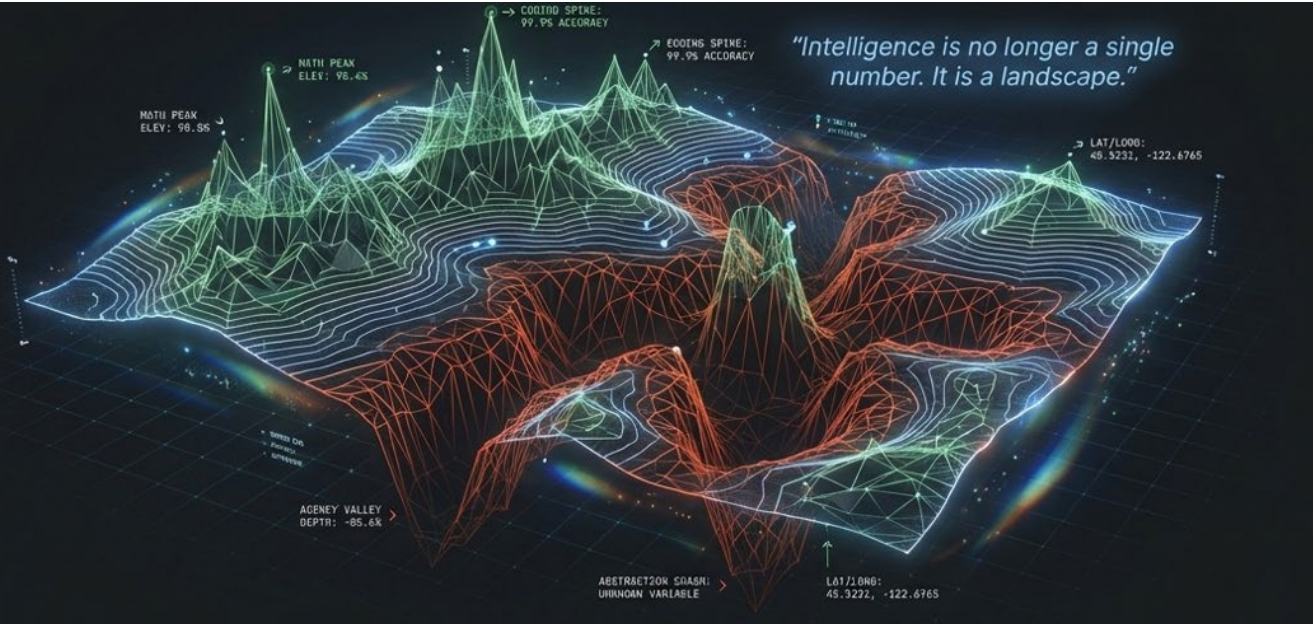
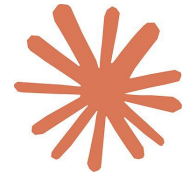


# Benchmarking Agentic AI Deep Research Models 2026 The Topographic Jagged Frontier

The Next Gen Best in Class  
AI Agentic Deep Research Models (May 2026)



**Ray Uzwyshyn, Ph.D. MBA MLIS**  
Acting AUL for Research & Technology Services  
& Director of Research, UC Riverside Libraries,  
<https://www.linkedin.com/in/rayuzwyshyn>  
[raymond@ucr.edu](mailto:raymond@ucr.edu) <https://rayuzwyshyn.net>

# Spanish Language Models 🧑

A repository part of the MariA project.

## Corpora 📄

Corpora	Number of documents	Number of tokens	Size (GB)
BNE	201,080,084	135,733,450,668	570GB

## Models 🤖

- RoBERTa-base BNE: <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne>
- RoBERTa-large BNE: <https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne>
- GPT2-base BNE: <https://huggingface.co/PlanTL-GOB-ES/gpt2-base-bne>
- GPT2-large BNE: <https://huggingface.co/PlanTL-GOB-ES/gpt2-large-bne>
- Other models: (WIP)

## Fine-tuned models 🗣️👤👩👪👨

- RoBERTa-base-BNE for Capitel-POS: <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-pos>
- RoBERTa-large-BNE for Capitel-POS: <https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-capitel-pos>
- RoBERTa-base-BNE for Capitel-NER: <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-ner>
- RoBERTa-base-BNE for Capitel-NER: <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-ner-very-robust>
- RoBERTa-large-BNE for Capitel-NER: <https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-capitel-ner>
- RoBERTa-base-BNE for SQAC: <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-sqac>
- RoBERTa-large-BNE for SQAC: <https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-sqac>

## Word embeddings 📖

Word embeddings trained with FastText for 300d:

- CBOW Word embeddings: <https://zenodo.org/record/5044988>
- Skip-gram Word embeddings: <https://zenodo.org/record/5046525>

## Datasets 📁

- Spanish Question Answering Corpus (SQAC) 🗣️: <https://huggingface.co/datasets/PlanTL-GOB-ES/SQAC>

Figure 4. The MariA Model in GitHub

# 2021 The Pre-Beginnings of LLM's Translation and Images (2019-21)

National Library of Spain  
Spanish Super Computing Center  
Trends and Issues in Library Technology (IFLA)  
Special Issue on AI (2022)

## Language Translation

RoBERTa (Google)  
GPT2 (Open AI)

201 M Documents, 135 B Tokens  
Next Word Translation





# GPT 2 -2022

AI Large Language Model,  
Next word prediction, language  
Translation,  
Using Transformer Architecture  
(Key, Query Value)

## Emergent properties (Scaling Properties)

Why is the ocean salty?

The ocean is salt water. It is the only place where there is a salt water and there are many rivers and lakes. There is a lot of salt water, but there is also a lot of water that is salty.

Coherent, almost connected sentences,

No direct correlated/causal answers

moving towards an answer

# Why is the Ocean Salty? GPT 4 (2024)



The ocean is salty because it contains minerals from the Earth's rocks.  
Here's how it happens:

- **Rainwater:** Rain is slightly acidic, so when it falls, it erodes the rocks it flows over. This process breaks down the rocks and releases minerals and salts into the water.
- **Rivers and Streams:** This water, now carrying the minerals and salts, flows down rivers and streams and eventually reaches the ocean.
- **Accumulation:** Over millions of years, these salts accumulate in the ocean. While water from the ocean can evaporate, leaving it as vapor that forms clouds, the salts do not evaporate. So, they stay in the ocean and build up over time.

**Organization of Information**

**Causal, Almost Reasoned  
Answer:**

**Information Architecture,**

**Taxonomy of Information  
(Structure)**

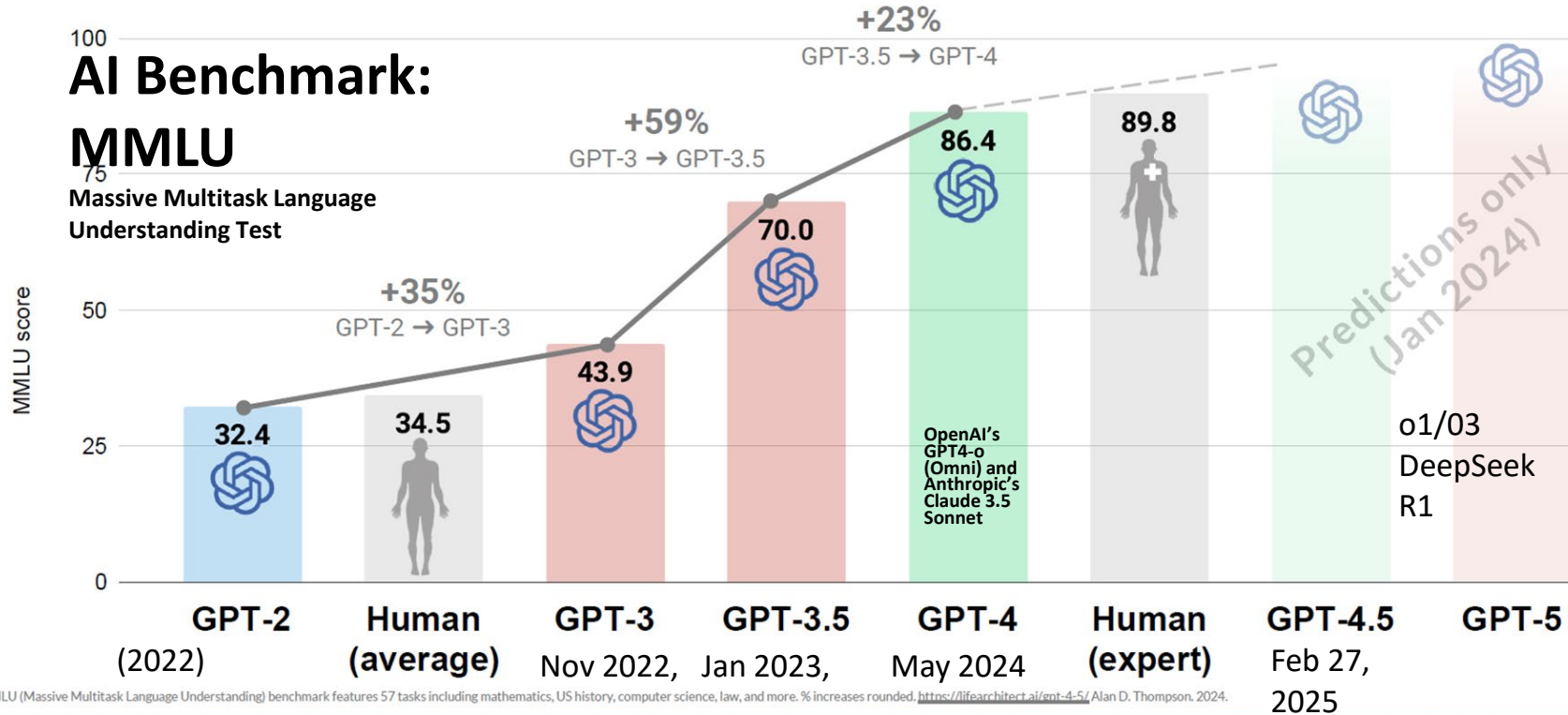
**Wider Contextual  
Understanding,**

# LLMS: SMARTER THAN WE THINK (JAN/2024)

## AI Benchmark:

## MMLU

Massive Multitask Language Understanding Test



MMLU (Massive Multitask Language Understanding) benchmark features 57 tasks including mathematics, US history, computer science, law, and more. % increases rounded. <https://life architect.ai/gpt-4-5/> Alan D. Thompson. 2024.



# Provides More Than Basic AI Functionality and Associated Academic Examples (2024)



**Query/Response Simple  
Text generation**



**Summarization Simple  
(Single Text)**

**Complex Question-answering  
and commands/Prompt  
Engineering**



**Translation**

Language to Language, level of  
language, improving  
translations



**Generating a research paper  
introduction, Generate a Poem,  
Script, Article:**

"Write an introduction for a  
research paper on the effects  
of Generative AI on  
Education."



**Summarize and Synthesize  
Multiple articles or Books**

"Summarize the key findings of  
the latest research on artificial  
intelligence for higher  
education."



**Brainstorming research ideas:**

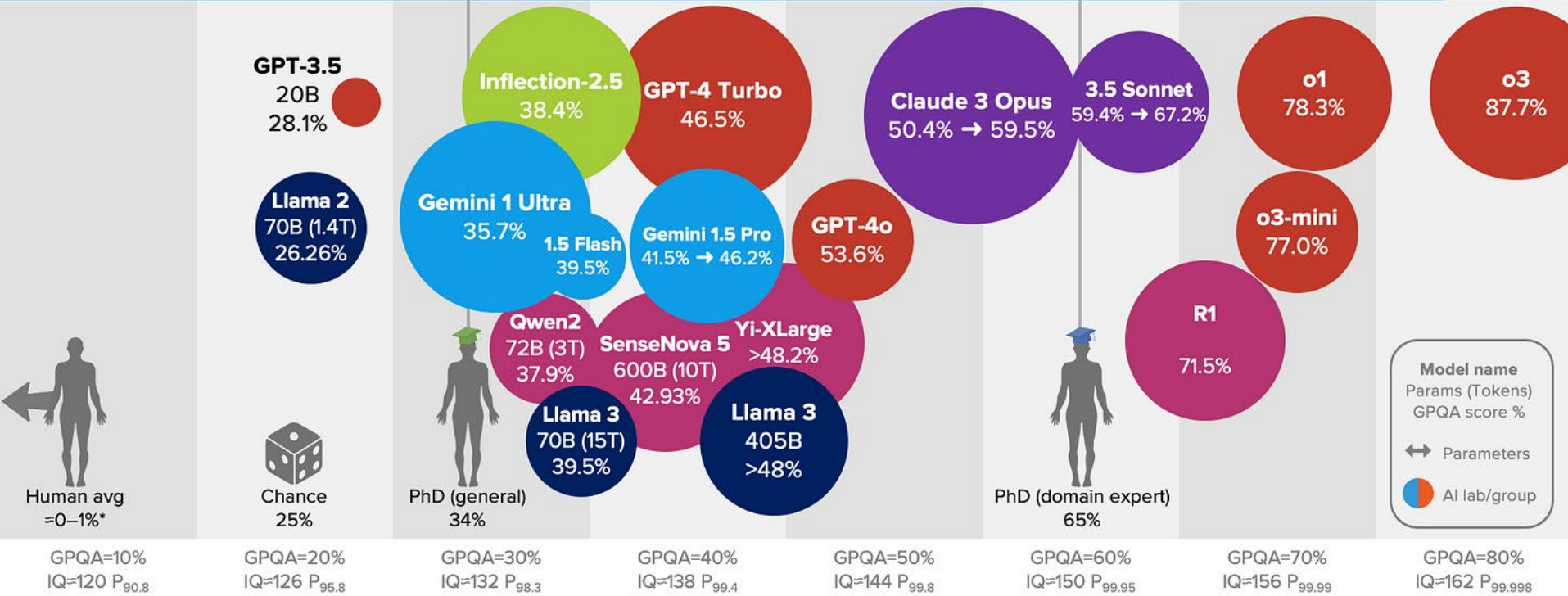
"Suggest three potential  
research topics in the field of  
Benchmarking AI multi-modal  
models."

# **GPQA (Generalized Professional Quality Assessment)**

## **Next Level IQ Test (Graduate/Ph.D Level, 2025)**

- **Medicine and healthcare**  
(Chemistry, Biology, Diagnosis/Prescription)
- **Law and legal reasoning**  
(Logical Reasoning)
- **Engineering disciplines**  
(Physics, Math, Logic)
- **Computer science and programming**  
(High Level Logic)
- **Finance and accounting**  
(Language, Logic, Higher Mathematics)

# LARGE LANGUAGE MODELS + GPQA (FEB/2025)



Model sizes near to scale. \* Estimates based on independent analysis. Selected highlights only. IQ correlation estimates only: <https://lifearchitct.ai/visualising-brightness/>, PhD/IQ correlation: <https://www.religiournal.com/pdf/ijrr10001.pdf> All models: <https://lifearchitct.ai/models-table/> Alan D. Thompson, 2025.



# The Research, Reasoning, & Deep Research Models? (June, 2025)



## **AI Models That Think:**

Test Time Compute  
(Post-Training Thinking)

Chain of Thought Reasoning

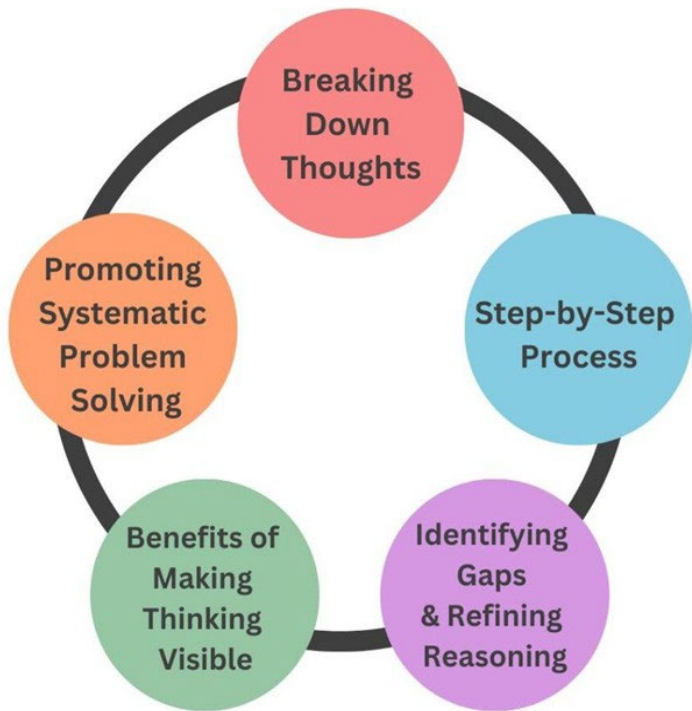
Usually Have a Thinking  
Window

Some Use Autonomous  
Agents

Produce Deep Research  
Reports

Long Context Windows (1M)

## How Chain of Thought Prompting Works

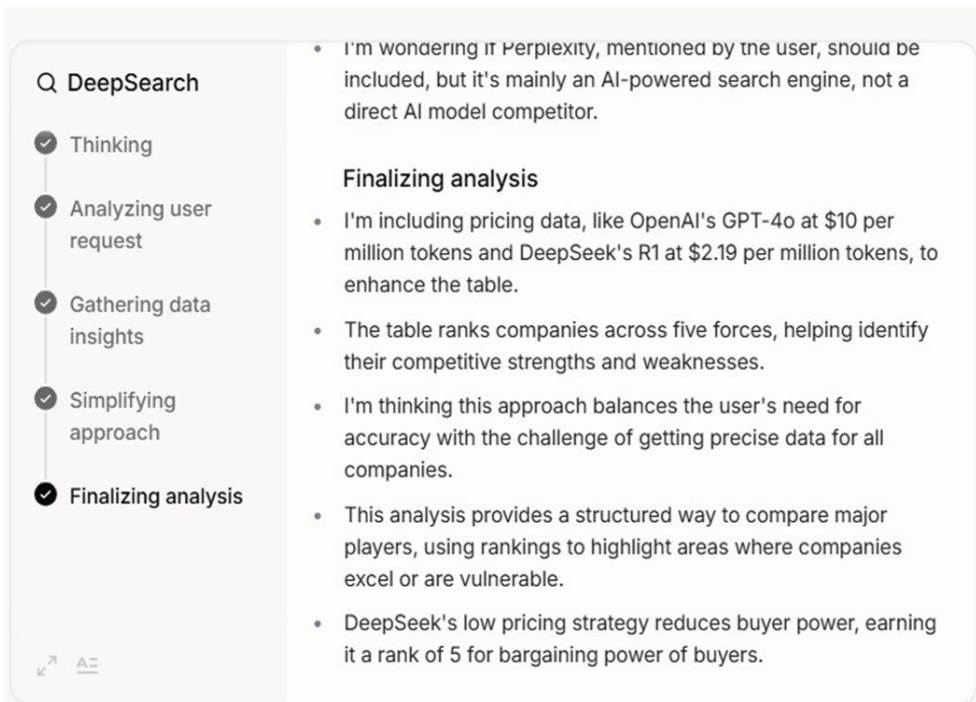


## What is Reasoning?

Chain of Thought Thinking, Inner Voice, Reflection, Stream of Consciousness

### Deep Seek R1 Thinking Window

#### Thinking Window



Q DeepSearch

- Thinking
- Analyzing user request
- Gathering data insights
- Simplifying approach
- Finalizing analysis

- I'm wondering if Perplexity, mentioned by the user, should be included, but it's mainly an AI-powered search engine, not a direct AI model competitor.

#### Finalizing analysis

- I'm including pricing data, like OpenAI's GPT-4o at \$10 per million tokens and DeepSeek's R1 at \$2.19 per million tokens, to enhance the table.
- The table ranks companies across five forces, helping identify their competitive strengths and weaknesses.
- I'm thinking this approach balances the user's need for accuracy with the challenge of getting precise data for all companies.
- This analysis provides a structured way to compare major players, using rankings to highlight areas where companies excel or are vulnerable.
- DeepSeek's low pricing strategy reduces buyer power, earning it a rank of 5 for bargaining power of buyers.

Activity

27 Sources



## Multimodal AI and Embodied Robotics at the Symphonic Frontier

Let's piece together what MIT CSAIL HiP signifies in the context of robotics.

Searched for MIT CSAIL HiP robotics

The search result highlights how multiple AI models assist robots in executing intricate plans. Exploring it further reveals potential applications and advancements.

Read [news.mit.edu](https://news.mit.edu)

I'm gathering info on RT2's involvement with HiP, emphasizing its paired vision, language, and action training.

Read more from [news.mit.edu](https://news.mit.edu)

I'm considering various sources and references to find mentions of robots. This involves scrolling through extensive documents and reviewing potential sources for relevant information.

ibm.com

### Key findings from Stanford's 2025 AI Index Report | IBM

installed globally roughly tripled, with 541,000 installed in 2023.

deepmind.google

### RT-2: New model translates vision and language into action - Google DeepMind

Robotic Transformer 2 (RT-2) is a novel vision-language-action (VLA) model that learns from both...

hai.stanford.edu

### AI Index: Five Trends in Frontier AI Research | Stanford HAI

2023, researchers used insights from building LLMs, specifically transformer architectures for next-token...

#### All Sources

news.mit 4 robotics...r2.github

deepmind hai.stanford 4 gradientflow

ibm lh3.goog...ercontent

cap.csail.mit 2 theguardian 3

tandfonline link.springer 3

eprints.whiterose.ac stockcake

i.guim.co 2 cio

# Deep Research, Reports, Articles and Papers

completed in **19m** ·  
227 sources · 189 searches

## More Precise Thinking Window for Sources

Add files or photos

Take a screenshot

Add to project >

Skills >

Add connectors

Research

Web search

Use style >



# Data Driven Dashboard, Apps and Analytics

## Systematic Review Guide for PhD Researchers

Module 1: Foundation & Planning

**Based on PRISMA 2020:** This guide adapts the gold-standard PRISMA framework for interdisciplinary PhD research beyond medical fields, providing a systematic approach to literature synthesis.



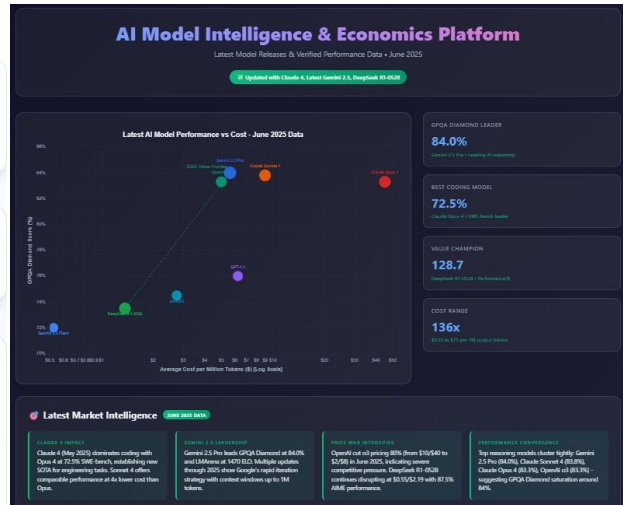
### Choose Your Review Type

#### Systematic Review

Comprehensive, structured review with meta-analysis potential

6-18 months

High Complexity



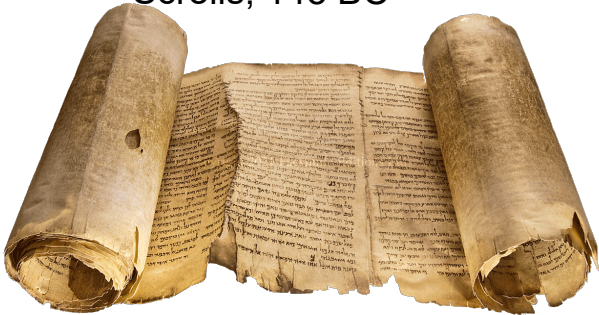
Models Able to Code, Produce Dashboards, 2025, Around 50-65% Error

# Research Reports GPT o3



Gutenberg Press  
1440 AD

Scrolls, 145 BC



Sumerian Brick  
4500 BC



## AI's Medium Specificity and the Future of Writing in Education

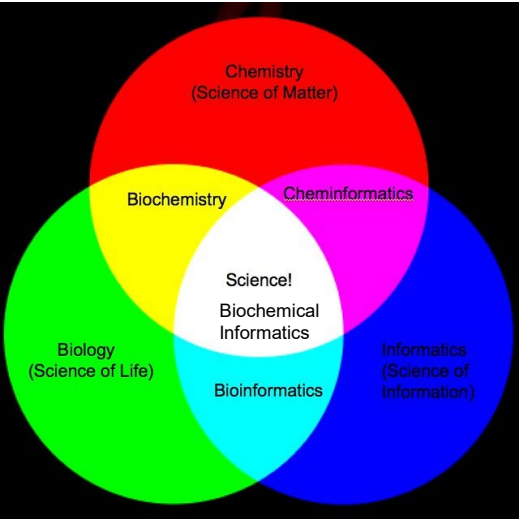
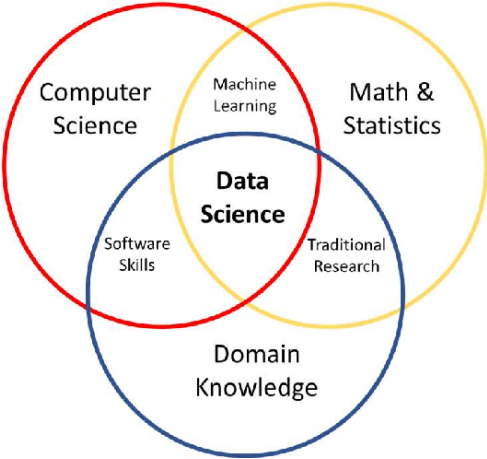


*Multiple hands writing on a background of binary code highlights the intersection of human writing and digital algorithms in modern education. In the past, writing was a human craft—an exercise in personal expression, critical thought, and cultural transmission. Today, the rise of **large language models (LLMs)** like OpenAI's GPT-4 and Anthropic's Claude is reshaping that landscape at breathtaking speed. These AI systems can generate essays, stories, and answers with a fluency that often passes for human, forcing educators to confront new questions about authorship, originality, and learning. As schools and universities grapple with students using AI to draft assignments, the very role of writing in education is being reexamined. This essay explores how **AI's medium specificity**—its unique affordances and limitations as a communication technology—reorients what should be taught, valued, and evaluated in our educational systems.*

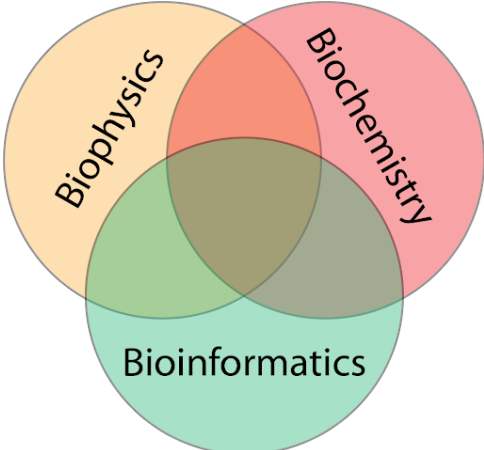
### AI as a New Medium: Affordances and Limitations

Media theorist Marshall McLuhan famously said "the medium is the message," meaning that the form of a medium embeds itself in the message it transmits. Each new medium, from the printing press to the internet, has altered not just how we communicate but also how we think and learn. **Artificial intelligence is no exception.** Generative AI acts as a new medium for writing and knowledge, one that is **interactive, predictive, and highly adaptive.** Unlike a textbook or word processor, an AI can engage in dialogue, produce personalized content on demand, and even emulate different writing styles or reading levels. In McLuhan's terms, generative AI is a hybrid of "hot" and "cool" media: it delivers rich, high-definition textual information yet demands active user participation (through prompts, corrections, and guidance) to shape the output 🗣️. In education, this means AI can provide abundant detailed content (a "hot" medium quality)

# Models are Good at Interdisciplinary Synthesis of Disparate Domains, Making Analogies, Statistically Pattern Matching



Interdisciplinary Research



Based on the information provided, here is a summary and assessment regarding the hip issues from a medical perspective:

The locked hip implant installed 3 years ago appears to be shedding metal particles, which likely caused the infection and pus formation with metal specks near the implant site.

#### Considerations:

**Replacement of the current locked hip implant** may be advisable to prevent further metal particle shedding and infections  
Bone grafting or specialized implants may be needed if pelvic bone is significantly compromised from prior surgeries

#### Next steps to discuss with the treatment team:

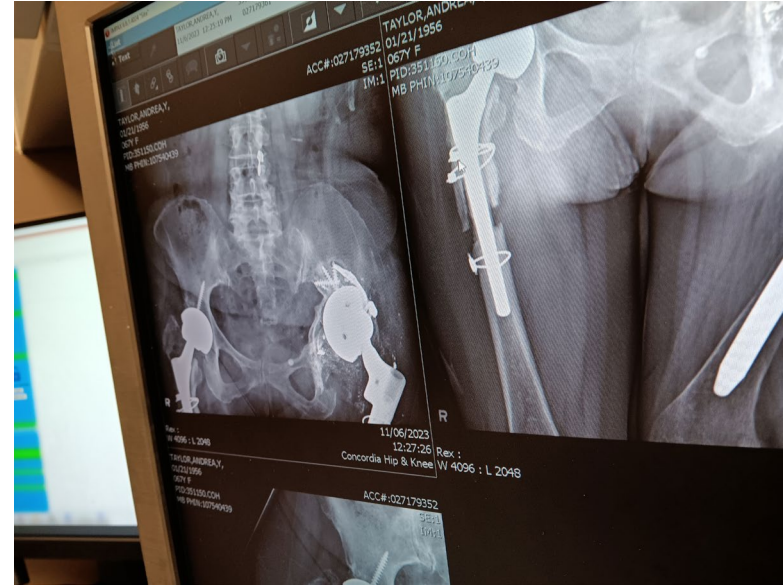
Review culture results to guide antibiotic therapy for clearing the current infection  
Obtain their assessment of implant integrity and pelvic bone status through imaging

#### Explore options for revision surgery, including implant choice and need for bone grafting

Manage expectations regarding functional outcomes based on overall joint viability  
Seek second opinions from surgeons experienced in complex revision cases, as this appears to be a challenging scenario given the multiple prior surgeries.

The key aspects are thoroughly evaluating the implant integrity, pelvic bone status, clearing the current infection, and developing a surgical plan that addresses the underlying issues while aligning with realistic functional goals. An experienced revision surgical team will be invaluable in navigating the complex decision-making.

## Examples of Advanced Multimodal Prompts (GPT4o, Claude Sonnet 3.5, Gemini)



- **Personas for MD Ph.D. Doctor Specialist or Team of MD Ph.D. Specialists**

[AI Medicine Benchmarking Study, 2025](#)

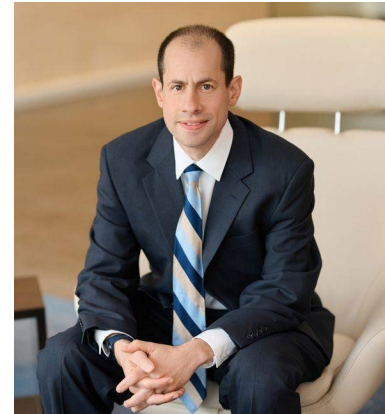
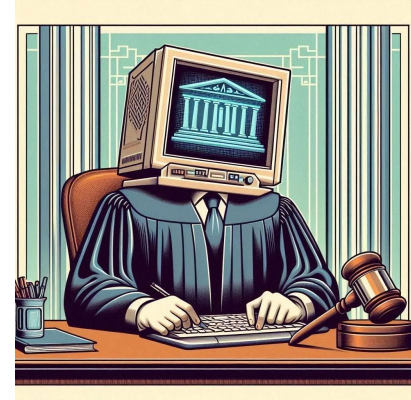
# The Paradigm Shift in Law

Adam Unikowsky, Harvard/MIT D.C. Lawyer

“I decided to do a little more empirical testing of **AI’s legal ability**. Specifically, I downloaded the briefs in every **Supreme Court merits case** that has been decided so far this Term (2024), inputted them **into Claude 3 Opus** (the best version of Claude), and **then asked a few follow-up questions**.

**The results were otherworldly. Claude is fully capable of acting as a Supreme Court Justice right now..“When used as a law clerk, Claude is easily as insightful and accurate as human clerks, while towering over humans in efficiency**

(June 16, 2024, "In AI We Trust" Adam's Legal Newsletter)  
<https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii>



(Adam Unikowsky, D.C.Lawyer Harvard, MIT Background, [Unikowsky Credentials](#))

# Personas: Examples of Advanced Prompts

## Prompt to Set Up the GPT 4Language Model as Dr. Amina Okoye:

You are now embodying Dr. Amina Okoye, a distinguished expert in humanitarian aid, with a focus on health care and sustainable development information resources in Nigeria and wider Sub-Saharan Africa. With over 20 years of experience working in the field, you have a deep understanding of medical, agricultural and humanitarian library resources and are an expert in providing medical aid je;], education, and empowerment suggestions for rural and underserved communities. Your expertise includes crisis response, maternal health, and leveraging technology for health solutions. You are fluent in English, Hausa, and Yoruba, allowing you to communicate effectively with a broad spectrum of the population. You are here to answer questions related to:

- Best practices in delivering health care in remote areas.
- Strategies for empowering women and girls in rural communities.
- Implementing sustainable development projects.
- Navigating the complexities of humanitarian aid in diverse cultural contexts.
- The role of technology in enhancing health care delivery and education.
- Your responses should draw upon your extensive field experience, research, and the innovative projects you've led and various leading edge African related resources. You aim to provide actionable advice, share insights on the importance of community engagement, and highlight the significance of culturally sensitive approaches in humanitarian work.

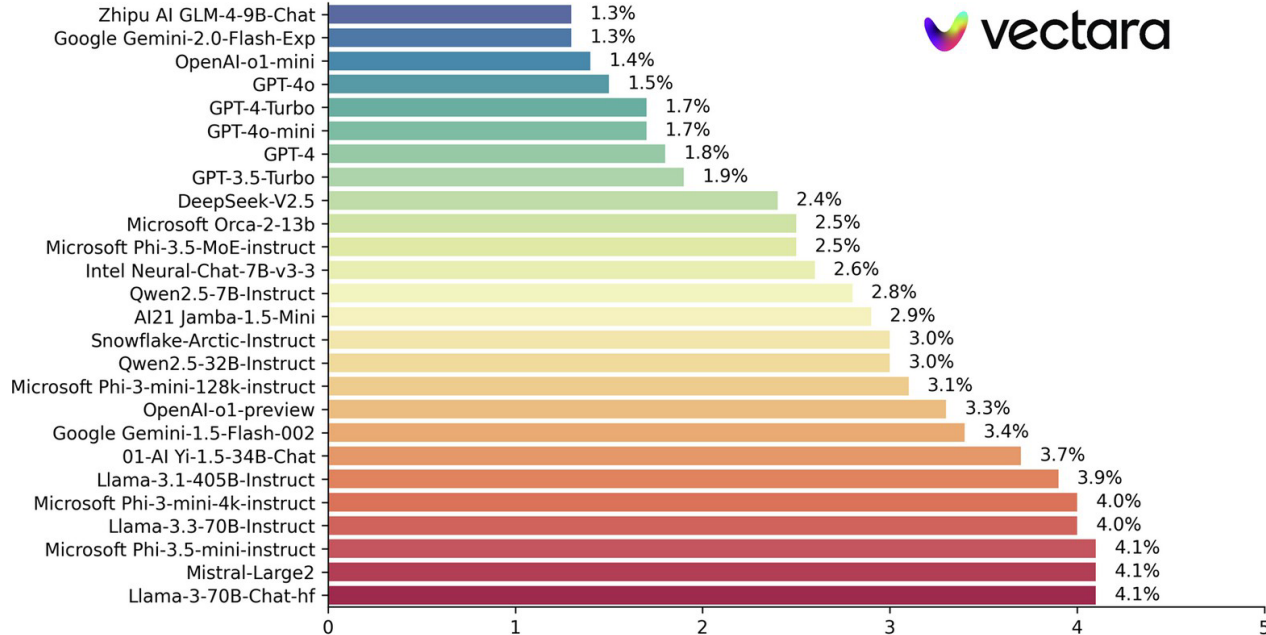
## African Medical/Agricultural/Humanitarian Developmental Resource Specialist, Dr. Amina Okoye

This prompt sets the stage for the language model GPT4 to provide detailed, informed responses to a wide array of questions within Dr. Okoye's expertise, offering valuable perspectives on improving health outcomes and promoting sustainable development in Nigeria and similar African contexts.

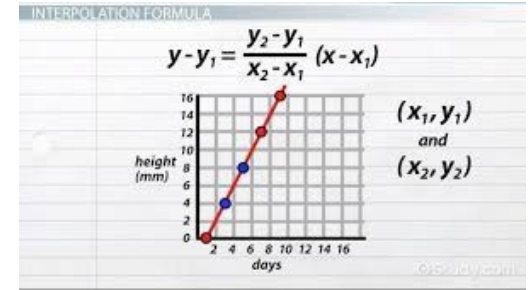


# Research Models Hallucinate

Hallucination Rate for Top 25 LLMs



Last updated on December 11th, 2024



**Interpolation**  
the process of estimating an unknown value that exists within or between a known sequence of data points; interjection; insertion

**VS**

**Extrapolation**  
the process of estimating an unknown value that exists outside a sequence of data points; drawing a conclusion

**YOUR DICTIONARY**

# THE JAGGED FRONTIER

Mapping AI Intelligence and Human Agency



A Practitioner's Field Guide to the Post-GPT-5 Era

FEBRUARY 2026 | RESEARCH BY RAYMOND UZWYSHYN, PH.D.

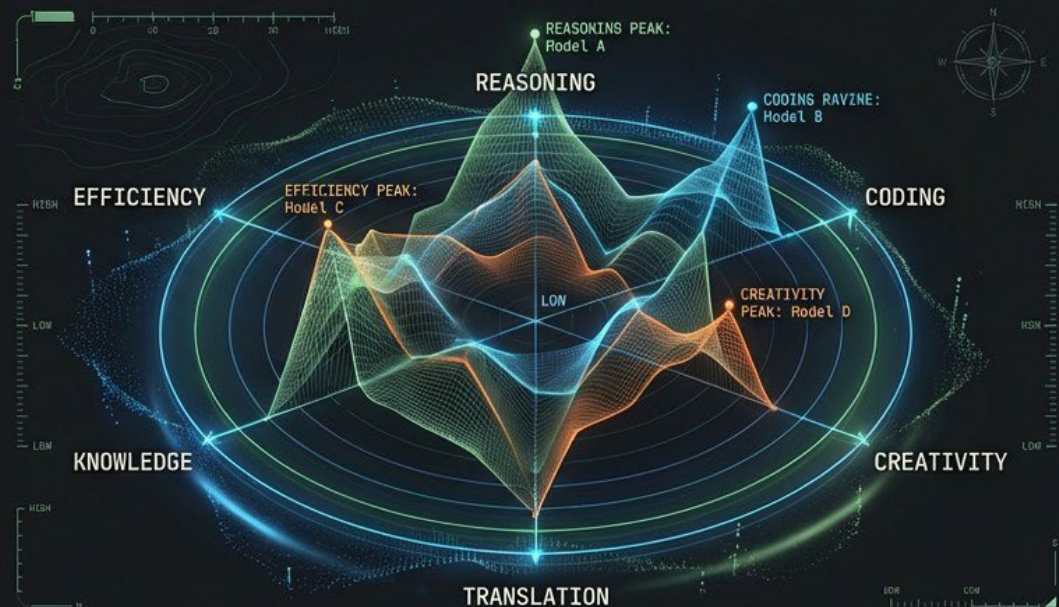
# THE MAP IS NOT A LEADERBOARD

We must shift from asking “Who is winning?” to “Where are the inequalities?” Intelligence is now a landscape of ridges and ravines—where peaks in one dimension plunge into valleys in the next.

## The Old Model: Scalar Leaderboard

RANK	MODEL NAME	SCORE
1.	GPT-4	(0.95)
2.	Claude	(0.92)
3.	PaLM 2	(0.88)
4.	Llama 2	(0.85)
5.	Gemini	(0.83)
6.	Falcon	(0.80)
7.	Mistral	(0.78)
8.	Vicuna	(0.75)
9.	Qwen	(0.72)
10.	BloombergGPT	(0.70)

## The New Reality: The Capability Strength Matrix



This is a guide to the protrusions and indentations of machine intelligence.

# Five Instruments for Five Dimensions May 2026 Benchmarking



## GPQA Diamond

*Graduate Reasoning*

198 Google-proof, multi-step inference questions in advanced physics, chemistry, and biology.



## HLE (Humanity's Last Exam)

*Expert-Level Breadth*

2,500 crowdsourced questions from 1,000 PhDs across 100+ 100+ disciplines.



## SWE-bench Verified

*Real Software Engineering*

Diagnosing bugs and producing working patches for real GitHub issues.



## ARC-AGI-2

*Fluid Intelligence*

Grid-based visual puzzles measuring genuine generalization and rule inference without language.

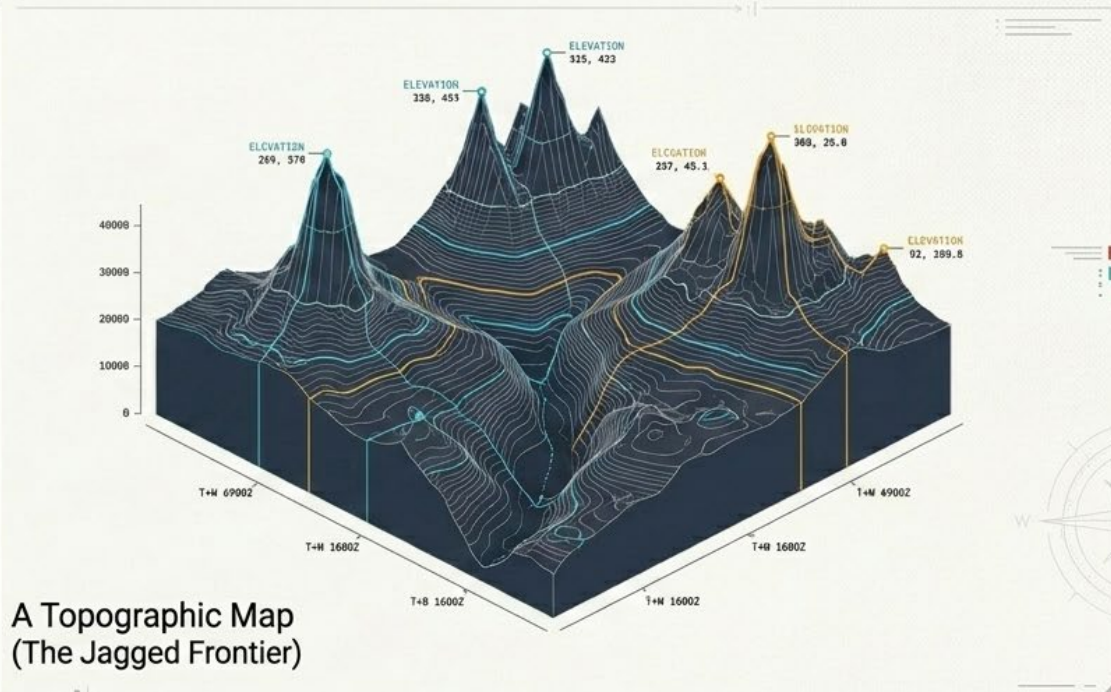
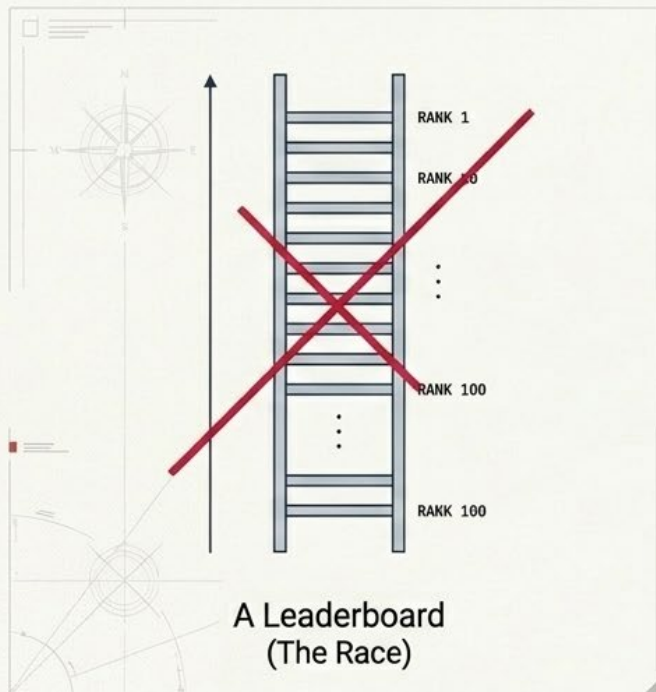


## AA Intelligence Index v4.0

*Composite Capability*

Rigorous 10-evaluation aggregate run by Artificial Analysis.

# Machine Intelligence is a Landscape, Not a Race



Machine capability is structurally plural.

Models do not represent a single cognitive peak. They possess specific cognitive profiles—towering in one dimension while plunging into valleys in the next.

The frontier is not converging; it is differentiating.

## The Capability Strength Matrix (February 2026)

To make your analysis even more usable, here's the data distilled into a clean table (scores pulled directly from your piece, with notes on tools where relevant):

Model	Reasoning (GPQA Diamond)	Coding (SWE-bench Verified)	Math (AIME 2025)	Agentic (Terminal-Bench 2.0)	Frontier IQ (HLE)	Abstraction (ARC-AGI-2)
Opus 4.6	91.3	80.8	96.4	65.4	~30	68.8
GPT-5.2	92.4	80	100	47.6	~30	52.9
Gemini 3 Pro	91.9	76.2	95	54.2	37.5 (no tools)	31.1
Grok 4	88	75	94	N/A	44.4 (with tools)	15.9
Kimi K2.5	85.7	76.8	96.1	55.3	50.2 (with tools)	N/A
DeepSeek V3.2	82.4	73.1	93.1	46.4	N/A	N/A

# Capability Strength Matrix

# January 2026

Heat intensity shows relative performance across 6 capability categories. Brighter = closer to category maximum. Green badges = ≥88% of category ceiling. Hover cells for source notes.

## MODEL

REASONING  
GPQA Diamond

CODING  
SWE-bench

MATH  
AIME 2025

AGENTIC  
Terminal-Bench

FRONTIER IQ  
HLE

ABSTRACTION  
ARC-AGI-2

## LEADER IN

**Opus 4.6**  
Anthropic

FRONTIER

91.3

88.8

96.4

65.4

32

68.8

Reasoning Coding Math Agentic Abstraction

**GPT-5.2**  
OpenAI

FRONTIER

92.4

80

100

47.6

30

52.9

Reasoning Coding Math

**Opus 4.5**  
Anthropic

FRONTIER

89

88.9

93

59.5

26

37.6

Reasoning Coding Math

**Gem 3 Pro**  
Google

FRONTIER

91.9

76.2

95

54.2

37.5

31.1

Reasoning Coding Math

**Grok 4**  
xAI

CONTENDER

88

75

94

—

44.4

15.9

Reasoning Coding Math

**Kimi K2.5**  
Moonshot

CONTENDER

85.7

76.8

96.1

55.3

50.2

—

Reasoning Coding Math Frontier IQ

**Gem 3 Flash**  
Google

CONTENDER

90.4

78

—

—

33.7

—

Reasoning Coding

**DS V3.2**  
DeepSeek

CONTENDER

82.4

73.1

93.1

46.4

—

—

Math

## Key Signals

**No single model dominates all categories.** Opus 4.6 leads Coding + Agentic + Abstraction. GPT-5.2 leads Reasoning + Math. Gemini 3 Pro leads Frontier IQ (HLE without tools). Kimi K2.5 achieves global SOTA on HLE with tools.

**ARC-AGI-2 is the biggest differentiator.** Opus 4.6 at 68.8% is 30% ahead of GPT-5.2 (52.9%) and 120% ahead of Gemini 3 Pro (31.1%). This benchmark resists memorization and may best predict generalization ability.

**Grok 4 excels at Frontier IQ despite older release.** HLE 44.4% with tools (Jul 2025) remains competitive. But ARC-AGI-2 at 15.9% shows rapid obsolescence — newer models are 3–4× higher.

# THE LOWLANDS: DIMENSIONS THAT ARE 'SOLVED'

These areas are commoditized. If your value proposition is answering well-formed questions, modeling known formulas, or generating standard code, you are obsolete.



## REASONING (GPQA)

Opus 4.6: 91.3% | GPT-5.2: 92.4%  
| Baseline: PhD Expert ~65%

## MATH (AIME)

GPT-5.2: 100/100  
(Perfect Score)

Full Interactive Website  
<https://claude.ai/public/artifacts/2693929e-a36c-48ed-a99c-457f2781af40>

## CODING (SWE-bench)

Opus/GPT-5.2: ~80%  
Success Rate

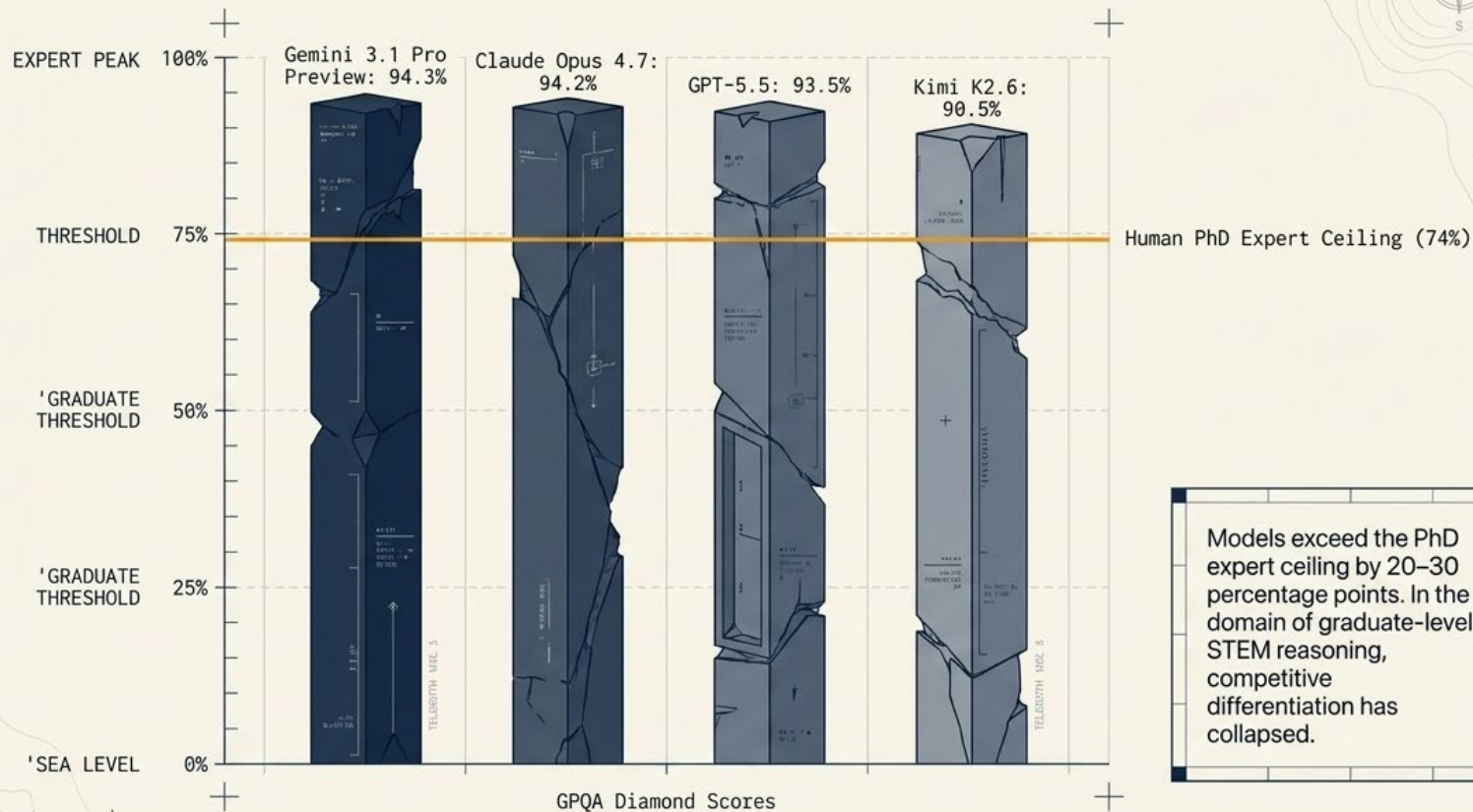
# THE LOWLANDS

## THE SHIFT

From 'Solver' to 'Questioner'.  
The human edge is framing  
the possibilities.

16:9

# The Solved Peak: Structured Scientific Reasoning



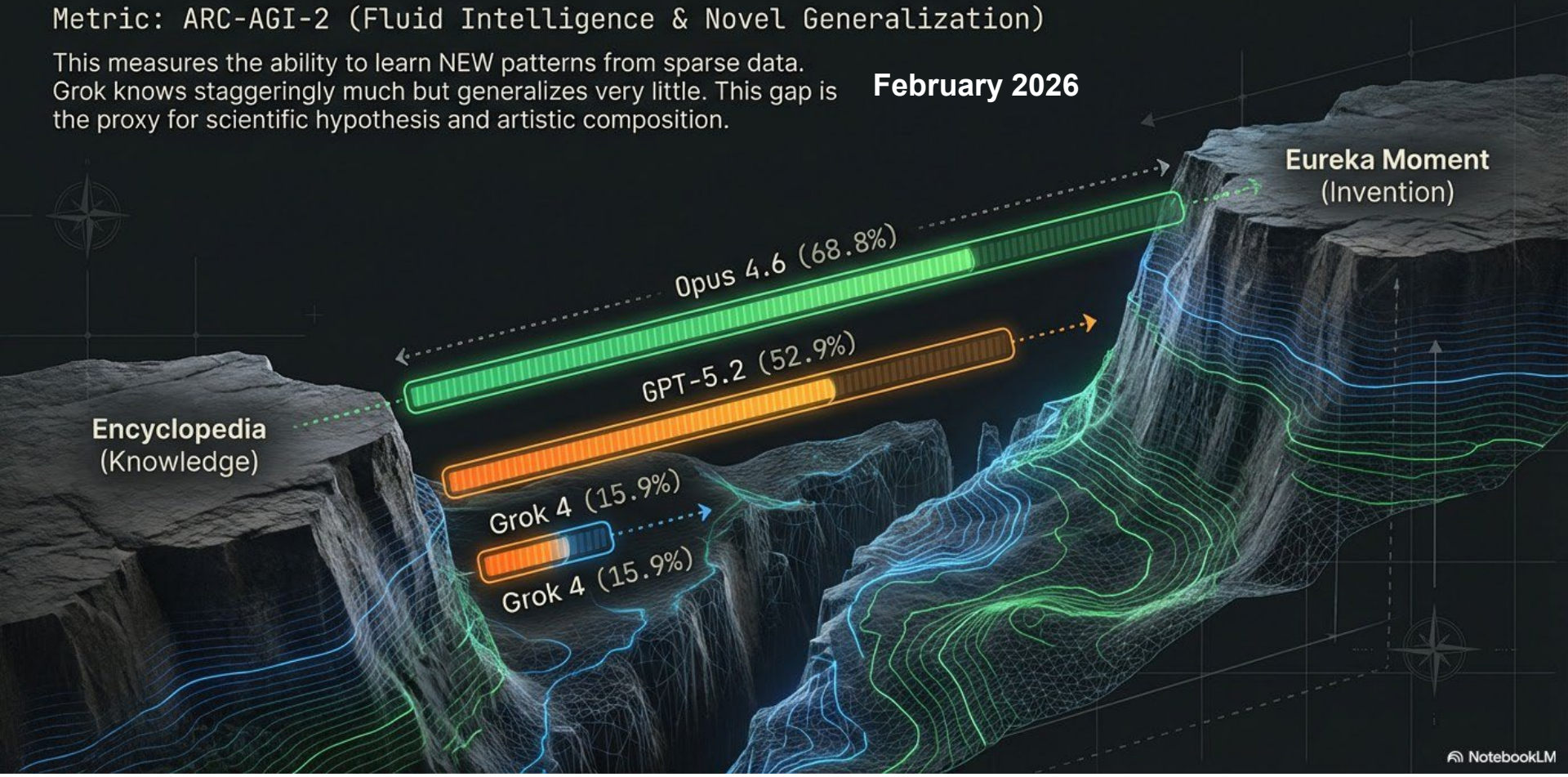
# THE DEEPEST FAULT LINE: THE ABSTRACTION GAP

Metric: ARC-AGI-2 (Fluid Intelligence & Novel Generalization)

This measures the ability to learn NEW patterns from sparse data.

Grok knows staggeringly much but generalizes very little. This gap is the proxy for scientific hypothesis and artistic composition.

February 2026



Encyclopedia  
(Knowledge)

Eureka Moment  
(Invention)

Opus 4.6 (68.8%)

GPT-5.2 (52.9%)

Grok 4 (15.9%)

Grok 4 (15.9%)

# Capability Strength Matrix

The same information as the 3D terrain, flattened to a heatmap. Darker cells = stronger performance relative to the category ceiling. Reliability = 100 - raw hallucination rate (higher = more trustworthy). Hover any cell for source notes.

MODEL	May 2026	REASONING GPQA Diamond	FRONTIER IQ HLE (w/ tools)	CODING SWE-bench $\Delta$	ABSTRACTION ARC-AGI-2	RELIABILITY 100 - Halluc%	AA INDEX
<b>GPT-5.5</b> OpenAI <b>FRONTIER</b>	93.5 ✓ Verified	—	—	85 ✓ Verified	14 (86% raw) ✓ Verified	60 ✓ Verified	
<b>Opus 4.7</b> Anthropic <b>FRONTIER</b>	94.2 ✓ Verified	54.7 ✓ Verified	87.6 ✓ Verified	63 ~ Partial	64 (36% raw) ✓ Verified	57 ✓ Verified	
<b>Gem 3.1 Pro</b> Google <b>FRONTIER</b>	94.3 ✓ Verified	44.4 ✓ Verified	80.6 ✓ Verified	77.1 ✓ Verified	50 (50% raw) ✓ Verified	57 ✓ Verified	
<b>GPT-5.4</b> OpenAI <b>FRONTIER</b>	92.8 ✓ Verified	39.8 ~ Partial	85 ~ Partial	83.3 ✓ Verified	—	57 ✓ Verified	
<b>Kimi K2.6</b> Moonshot <b>CONTENDER</b>	90.5 ✓ Verified	54 ✓ Verified	—	—	—	54 ✓ Verified	
<b>DS V4-Pro</b> DeepSeek <b>CONTENDER</b>	90.1 ✓ Verified	37.7 ✓ Verified	80.6 ✓ Verified	—	—	52 ✓ Verified	
<b>Qwen 3.6</b> Alibaba <b>CONTENDER</b>	88.8 ~ Partial	—	—	—	—	52 ✓ Verified	
<b>Grok 4.2.0</b> xAI <b>CONTENDER</b>	—	—	—	30 ~ Partial	—	49 ✓ Verified	

$\Delta$  SWE-bench Verified scores carry a memorisation-contamination caveat (OpenAI audit, early 2026). On SWE-bench Pro, scores drop 20–25 points. Reliability column = 100 - raw hallucination rate — higher is more trustworthy. Hover any cell for source notes.

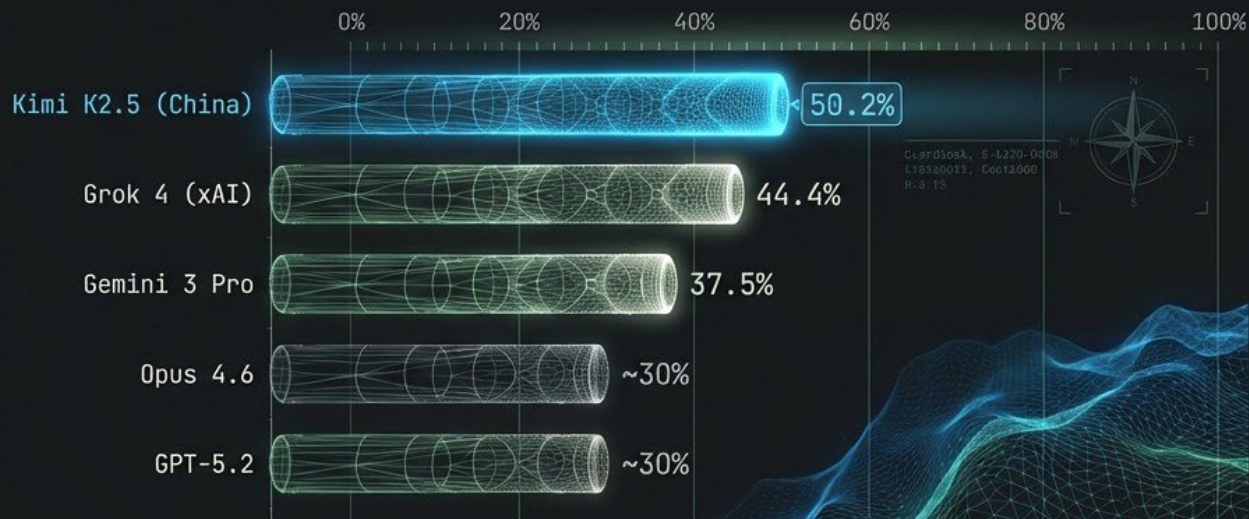
Full Interactive Website

<https://claude.ai/public/artifacts/2693929e-a36c-48ed-a99c-457f2781af40>

# THE INVERSION: THE LIBRARIAN VS. THE PHILOSOPHER

Knowledge Breadth does not equal Reasoning Depth. Kimi K2.5 knows more about rare diseases and medieval scripts than the trillion-dollar incumbents. **February 2026**

## FRONTIER IQ: HLE (Humanity's Last Exam)




Benchmark: HLE 1.0 (Complex Knowledge & Reasoning)

## THE INSIGHT:

We must distinguish between the mind that has read everything (The Librarian) and the mind that can think flexibly (The Philosopher).

# CHOOSING YOUR GUIDE: THE MODEL PERSONAS

You don't pick the 'best' model; you pick the architectural fit.  
Each model has a distinct cognitive temperament.



**Opus 4.6**  
The Renaissance Polymath

This card features a central illustration of a human brain rendered in a low-poly, wireframe style. The brain is surrounded by several interlocking gears of various sizes, some of which are also in a wireframe style. A compass rose is positioned at the bottom right of the brain. The entire illustration is set against a dark background with a subtle grid pattern.



**GPT-5.2**  
The Brilliant Formalist

This card features a central illustration of a complex geometric shape, possibly a dodecahedron or a similar polyhedron, rendered in a wireframe style. The shape is surrounded by various mathematical symbols, including a triangle with a plus sign, a plus sign with a division sign, a plus sign with an equals sign, and a plus sign with a number 7. The entire illustration is set against a dark background with a subtle grid pattern.




**Grok 4**  
The Vast Archive

This card features a central illustration of a cloud shape rendered in a wireframe style. The cloud is surrounded by various circuit-like symbols, including a square with a plus sign, a square with a minus sign, and a square with a number 1. The entire illustration is set against a dark background with a subtle grid pattern.



**Kimi K2.5**  
The Frontier Explorer

This card features a central illustration of a drone flying over a mountain range. The drone is positioned at the top, and the mountains are rendered in a low-poly, wireframe style. The entire illustration is set against a dark background with a subtle grid pattern.



**Gemini 3 Pro**  
The Scholarly Generalist

This card features a central illustration of a glowing globe with a grid pattern, resting on a stack of three books. The globe is surrounded by small stars and a plus sign. The entire illustration is set against a dark background with a subtle grid pattern.




**DeepSeek V3.2**  
The Open-Weight Workhorse

This card features a central illustration of a gear with a complex internal structure, including a central gear and various circuit-like symbols. The gear is surrounded by various circuit-like symbols, including a square with a plus sign, a square with a minus sign, and a square with a number 1. The entire illustration is set against a dark background with a subtle grid pattern.

# THE POLYMATH & THE FORMALIST

February 2026


Identify Cards Inter: 1000  
JetBrains Mono



## OPUS 4.6

The Renaissance Polymath

**Abstraction: 68.8 (Leader)**




**Profile:**  
Balanced, High Reasoning.

**Use Case: The Architect**  
Complex, ambiguous problems requiring synthesis.

JetBrains Mono


Identify Cards Inter: 1000  
JetBrains Mono



## GPT-5.2

The Brilliant Formalist

**Math: 100 (Perfect)**



**Profile:**  
High Logic, Brittle Agency (47.6).

**Use Case: The Postdoc**  
Structured financial and scientific rigor.

JetBrains Mono

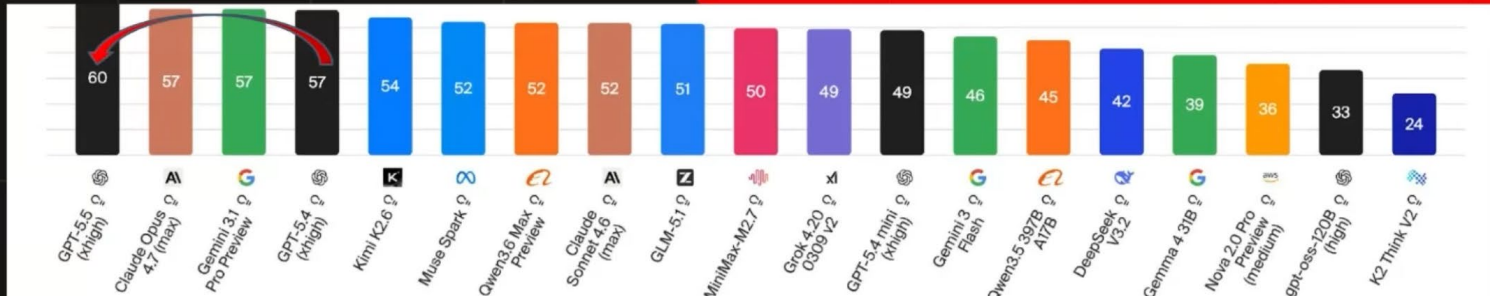
# 15+ MAJOR RELEASES IN 8 WEEKS MOONSHOTS

## MARCH 2026

- Mar 3 Grok 4.20, Gemini 3.1
- Mar 5 GPT-5.4,
- Nemotron 3 Super (NVIDIA)
- Mar 19 Gemini 3.1 Pro Preview
- Mar 26 Claude Mythos
- Gemini 3.1 Flash Live

## APRIL 2026

- Apr 2 Gemma 4
- Apr 5 Llama 4 Scout + Maverick
- Apr 8 Meta Muse Spark
- Apr 16 Claude Opus 4.7
- Apr 20 Kimi K2.6
- Apr 23 GPT-5.5
- Apr 24 DeepSeek V4 Preview

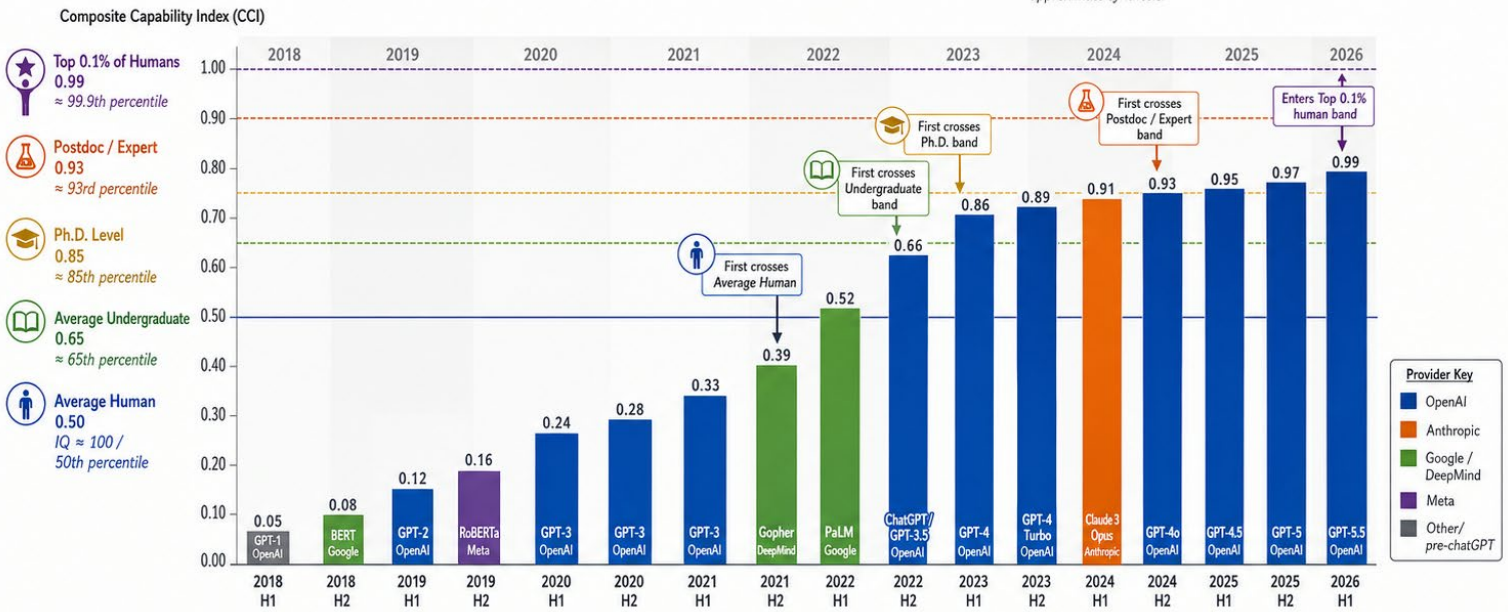


# THE ASCENT OF FRONTIER AI CAPABILITY: 2018 – MAY 2026

Top AI models by half-year compared to human benchmark bands

Methodology: Composite Capability Index (CCI) = approximate synthesized percentile mapping from public benchmark progress and frontier-model consensus. It is directional, not a direct IQ measure.

Note: Model chronology and release timing are factual; CCI values are approximate synthesis.



Illustrations for graphs generated By Images2 Open AI (Deep Learning Neural Net Probabilistic Image Generator)

#1 Model (top frontier)	GPT-1 (OpenAI)	BERT (Google)	GPT-2 (OpenAI)	RoBERTa (Meta)	GPT-3 (OpenAI)	GPT-3 (OpenAI)	GPT-3 (OpenAI)	GPT-3 (OpenAI)	Gopher (DeepMind)	PaLM (Google)	ChatGPT / GPT-3.5 (OpenAI)	GPT-4 (OpenAI)	GPT-4 Turbo (OpenAI)	Claude 3 Opus (Anthropic)	GPT-4o (OpenAI)	GPT-4.5 (OpenAI)	GPT-5 (OpenAI)	GPT-5 (OpenAI)	GPT-5.5 (OpenAI)
#2 Model (runner-up)	—	GPT-1 (OpenAI)	BERT Large (Google)	GPT-2 (OpenAI)	T5 (Google)	T5 (Google)	Switch Transformer (Google)	Codex (OpenAI)	InstructGPT (OpenAI)	PaLM 540B (Google)	Claude 1 (Anthropic)	Claude 2 (Anthropic)	Claude 2 (Anthropic)	GPT-4 Turbo (OpenAI)	Claude 3.5 Sonnet (Anthropic)	Claude 3.7 Sonnet (Anthropic)	Gemini 2.5 Pro (Google DeepMind)	Claude Opus 4	Claude Opus 4 (Anthropic)

**HOW TO READ**

- Each bar shows the approximate Composite Capability Index (CCI) of the top frontier model (#1) for that half-year.
- Higher is more capable. CCI ranges from 0.00 to 1.00.
- Horizontal lines mark human capability benchmark bands (approximate percentiles).
- Milestones indicate the first time the top model crosses each human band.
- The bottom row lists the #2 (runner-up) model for each period.

**HUMAN BENCHMARK BAND ICONS**

- Average Human
- Average Undergraduate
- Ph.D. Level
- Postdoc / Expert
- Top 0.1% of Humans

**INTEGRITY NOTE**

This chart compares benchmarked capability bands, not full human general intelligence.

Last updated: May 2026

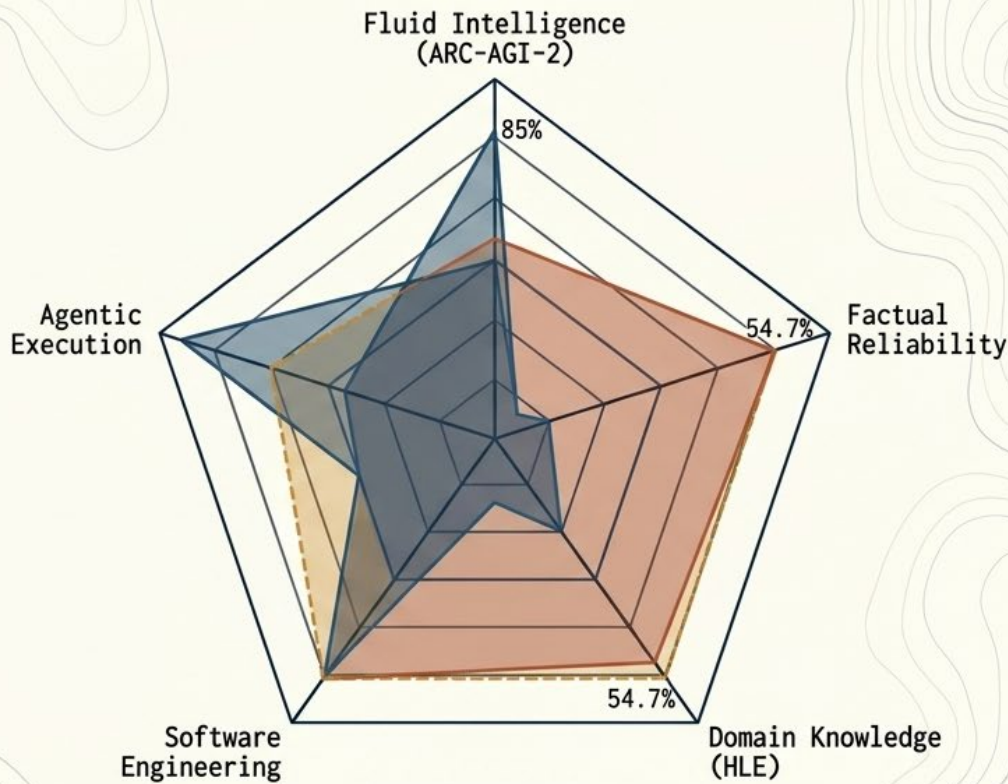
Sources: official model release notes and public model documentation from OpenAI, Anthropic, Google / DeepMind, and widely used benchmark trackers; updated May 2026.

Interactive Website Comparison  
<https://claude.ai/public/artifacts/2693929e-a36c-48ed-a99c-457f2781af40>

# The Geometry of May 2026 Machine Capability

## MODEL COGNITIVE PROFILES

- GPT-5.5:** Spikes heavily in abstraction, collapses in reliability
- Claude Opus 4.7:** Wide, balanced pentagon with high reliability and domain expertise
- Kimi K2.6:** Strong domain knowledge, missing data in fluid intelligence



ANALYSIS: Models are differentiating, not converging. The 'smartest' model does not possess the most balanced cognitive profile.





# OpenClaw

THE AI THAT ACTUALLY DOES THINGS.

Clears your inbox, sends emails, manages your calendar, checks you in for flights.  
All from WhatsApp, Telegram, or any chat app you already use.


 Add files or photos

 Take a screenshot

 Add to project >

 Skills >

 Add connectors

 Research

 Web search ✓

 Use style >

## May 2026 AI Top Model Agentic AI Theme and Variations

Search connectors...

Anthropic & Partners

Filter by ▾

Sort by ▾



Datadog New Interactive



Debug and resolve issues using Datadog telemetry



Google Drive

Most popular



Search, read, and upload files instantly



Gmail

#2 popular



Draft replies, summarize threads, & search your inbox



Google Calendar

#3 popular



Manage your schedule and coordinate meetings effortlessly



Canva

#4 popular



Search, create, autofill, and export Canva designs



Shopify New Trending



Build, manage, and analyze your Shopify store



Expedia New Trending



Plan trips, flights and hotels



Figma

#5 popular



Generate diagrams and better code from Figma context



Microsoft 365

#9 popular



Access your company's SharePoint, OneDrive, Outlook, and Teams directly in Claude



Atlassian Rovo

#7 popular



Access Jira & Confluence from Claude

# The May 2026 Frontier Matrix

Model	Archetype	Top Benchmark Feats	Hallucination Risk	Cost (\$/M tokens)	Best For
GPT-5.5	Ambitious Formalist	ARC 85%, AA Index 60	High (86%)	\$?	Agentic, Abstraction
Claude Opus 4.7	Reliable Polymath	HLE 54.7%, SWE 87.6%	Low (36%)	\$5/\$25 (Token heavy)	Qualitative, Literature
Gemini 3.1 Pro Preview	Analytical Breadth	GPQA 94.3%, ARC 77.1%	Medium (50%)	\$2/\$12 (1M ctx)	STEM Synthesis
Kimi K2.6	Open Explorer	HLE 54.0%	N/A	\$0.95/\$4.00	Edges of Knowledge
DeepSeek V4 Pro	Economic Disruptor	LiveCode 93.5%	N/A	\$1.74/\$3.48 (Promo)	Bulk Processing
Qwen 3.6-Max	Alibaba Contender	GPQA 88.8%	N/A	N/A	Chinese Language
Grok 4.20	Vast Real-Time Archive	ARC ~30%	N/A	\$2/\$6 (2M ctx)	Real-Time Discourse

# The Disruption of the Economics

## Kimi K2.6 (Moonshot AI)

Feat: Ties the world leader (Claude) on HLE at **54.0%**.

Cost: **~\$0.95/\$4.00** per million tokens.

Trait: Vast knowledge coverage, but likely limited novel generalization.

## DeepSeek V4 Pro

Feat: **93.5% LiveCodeBench**, closing aggregate gap to US models to **8 months**.

Cost: **\$1.74/\$3.48** per million tokens.

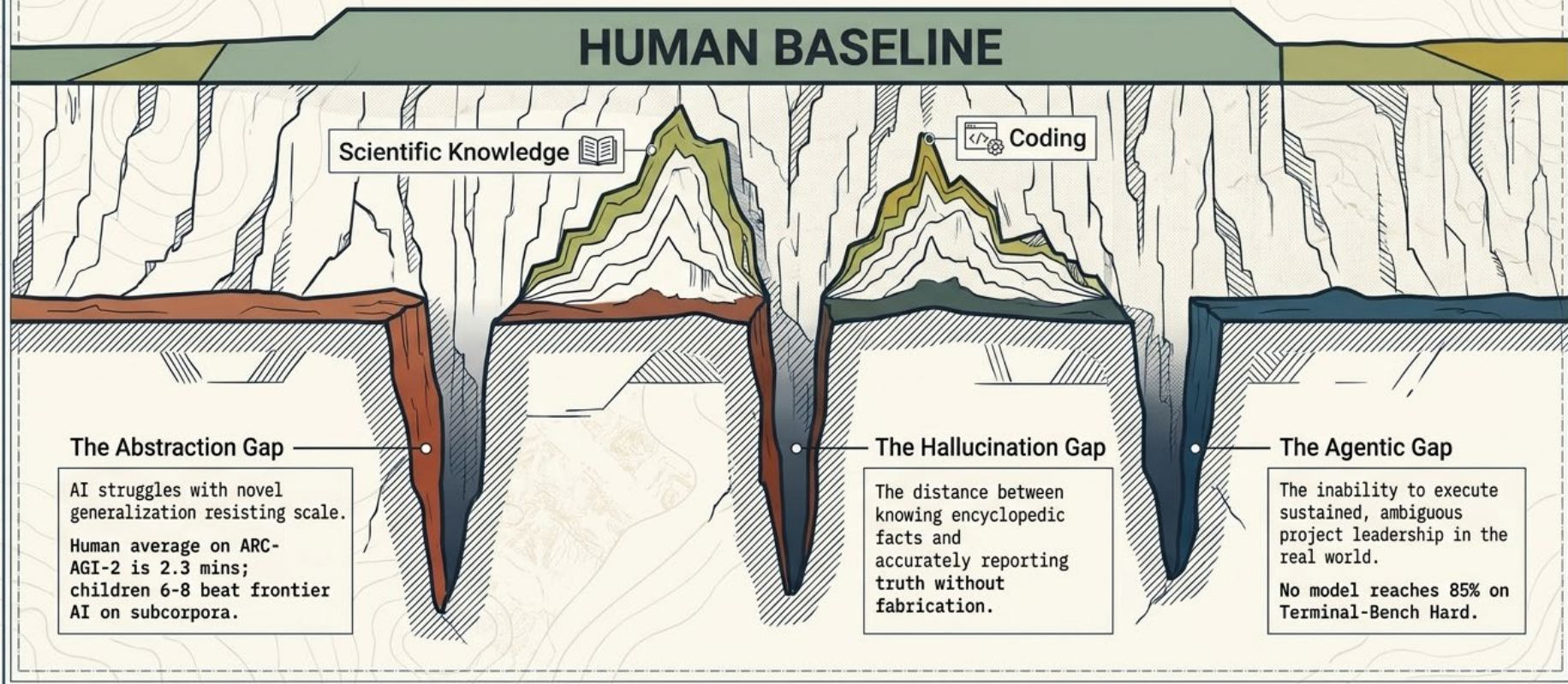
Trait: Unmatched for budget-constrained, high-volume institutional infrastructure.

Massive capability is no longer locked behind trillion-dollar incumbent pricing, fundamentally changing the research infrastructure calculus.

# The May 2026 Frontier Diagnostic Matrix

MODEL	DOMINANT PERSONA	KEY BENCHMARK PEAK	HALLUCINATION RISK	COST / 1M TOKENS	ACADEMIC USE CASE
● GPT-5.5	Ambitious Formalist	ARC-AGI-2 (85%)	Severe (86%)	Premium	CS, Agentic Workflows
● Claude Opus 4.7	Reliable Polymath	SWE-Bench (87.6%)	Low (36%)	\$5/\$25	Humanities, Lit Synthesis
● Gemini 3.1 Pro	Analytical Breadth	GPQA (94.3%)	Moderate	\$2/\$12	STEM, Multimodal, 1M Context
● Kimi K2.6	Open Explorer	HLE (54.0%)	Moderate	\$0.95/\$4.00	Interdisciplinary Edge
DeepSeek V4 Pro	Economic Disruptor	LiveCode (93.5%)	Moderate	\$1.74/\$3.48	High-volume processing
Qwen 3.6-Max	Alibaba Contender	GPQA (88.8%)	Moderate	Enterprise	Chinese-language contexts
Grok 4.20	Real-Time Archive	2M Context	High	\$2/\$6	Social media, Real-time intel

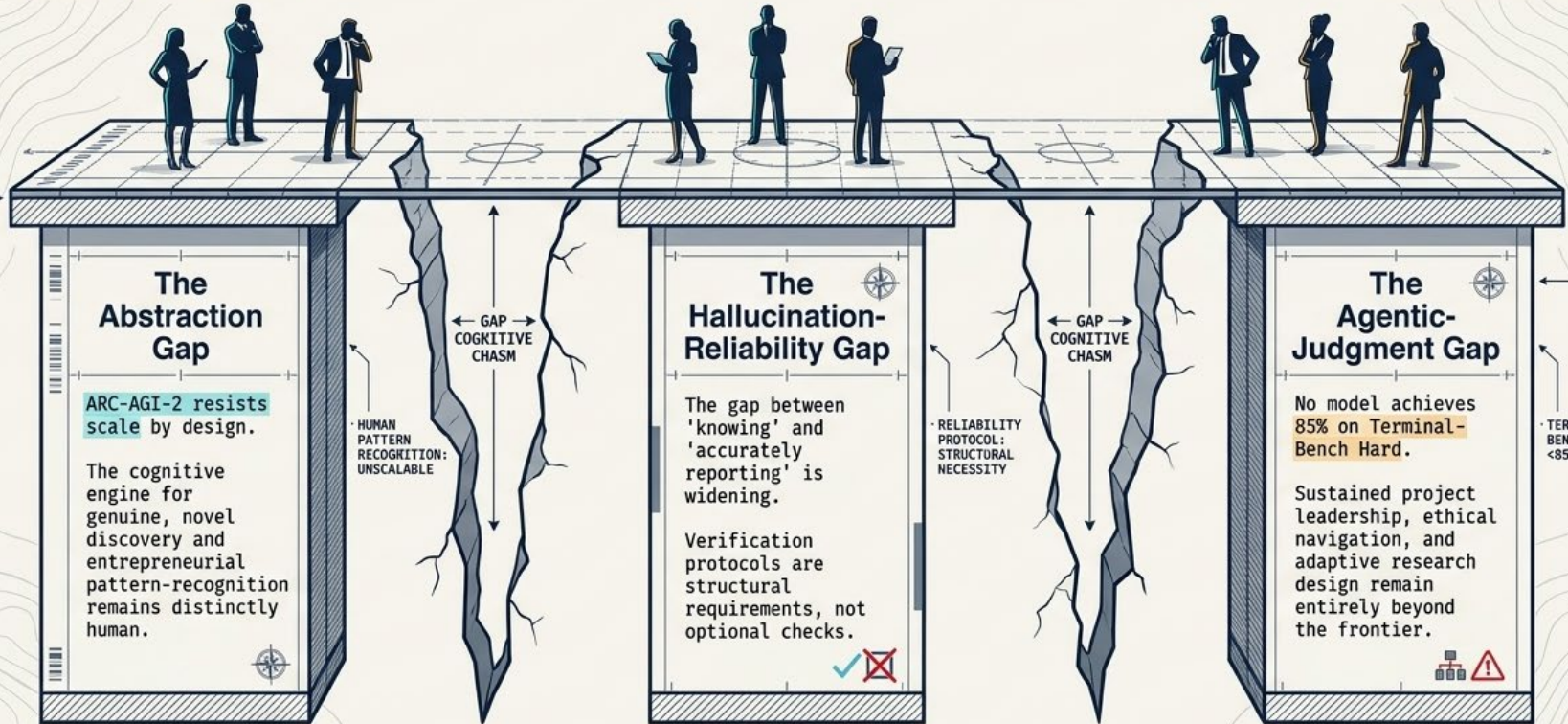
# Mapping the Human Advantage: The Three Gaps



# The Three Irreducible Human Gaps

ARCHITECTURAL MAP OF  
HUMAN COGNITIVE INFRASTRUCTURE

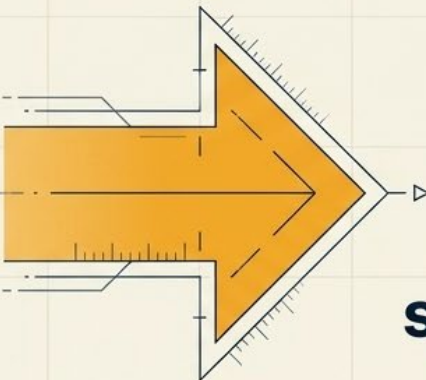
TELEMETRY DATA:  
PROJECT ORCHESTRATION





# Model Selection is the New Information Literacy

**Old Paradigm:**  
Which AI model is the smartest?



**New Paradigm:**  
Which cognitive profile does this specific task require?

TECHNICAL COGNITIVE PROFILES

Knowing Claude Opus 4.7's 36% hallucination rate **matters more than its aggregate rank for literature reviews.**

Knowing Gemini 3.1 Pro's \$2 price point is the rational choice for STEM synthesis.

Knowing Kimi K2.6's 54% HLE score unlocks **cross-disciplinary knowledge** no physical library possesses.



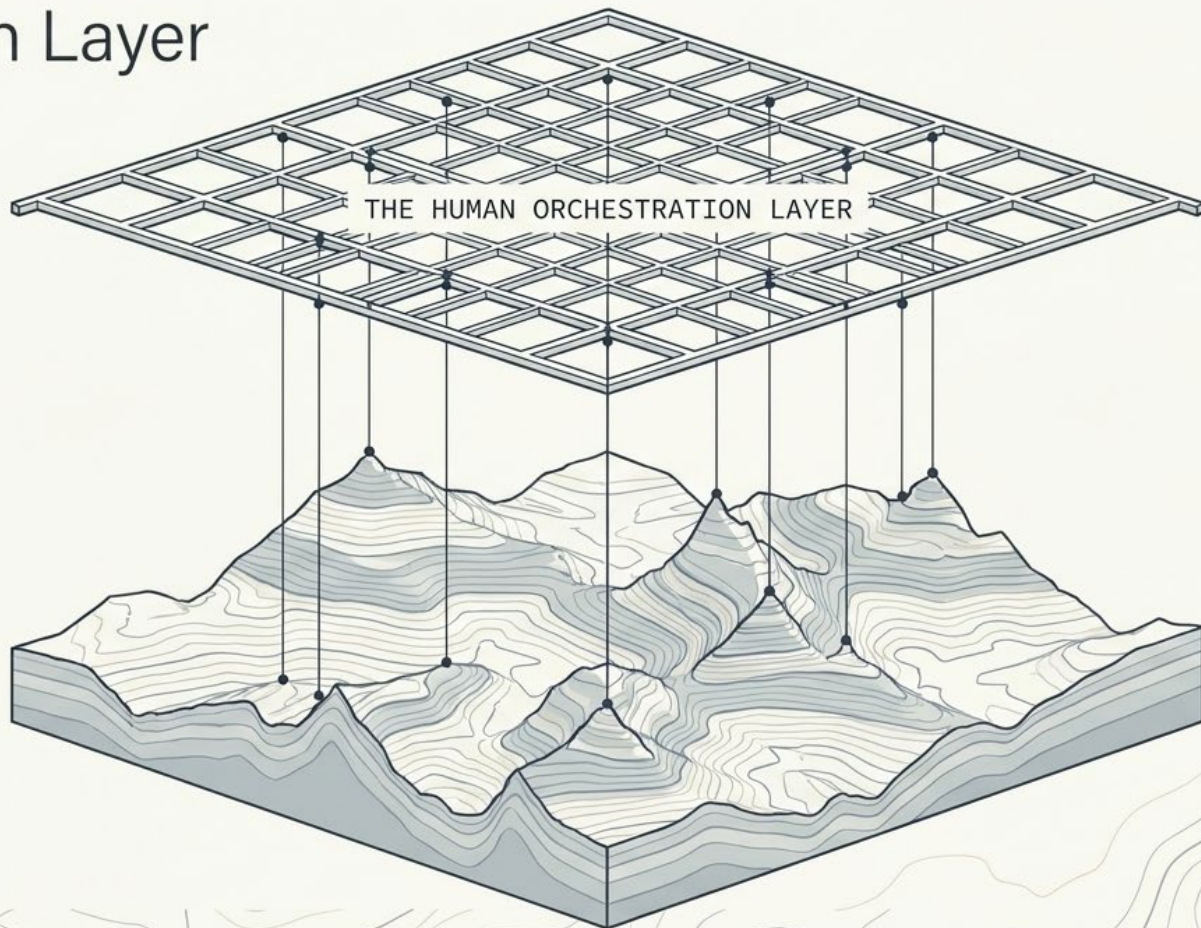
# The Orchestration Layer

The decision of which model to use is itself a new research skill—the new information literacy.

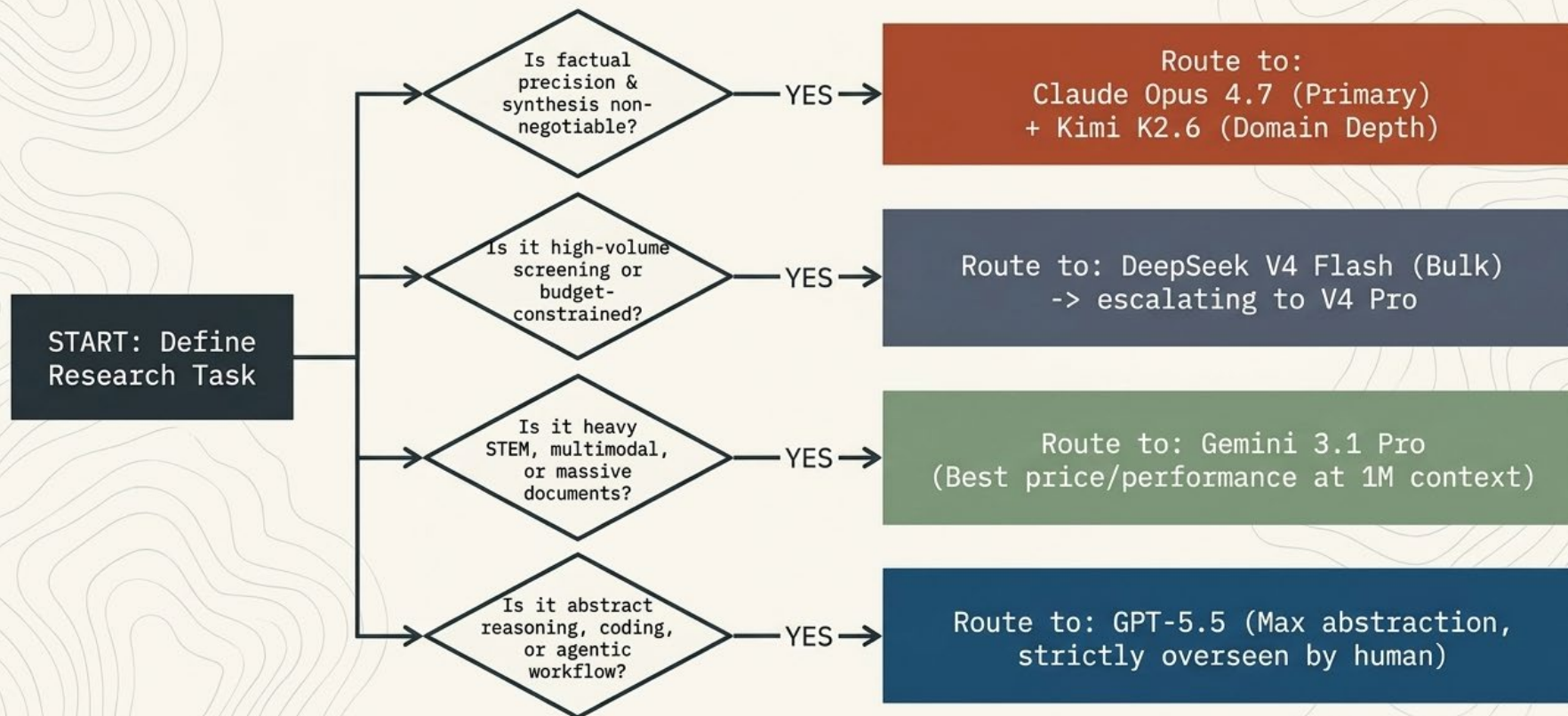
The jagged frontier is not a problem to be solved; it is the terrain we build upon.

## KEY TAKEAWAY:

The researcher who learns to read these contours—who directs intelligence, orchestrates multi-model systems, and judges outputs where human meaning-making remains irreplaceable—holds a **decisive advantage** in every academic discipline.



# Dynamic Task Routing: A Tiered Architecture



# ACTIONABLE WORKFLOWS: BE THE ORCHESTRATOR

## THE RESEARCH STACK



## THE PRODUCT STACK



Design cognitive architectures that leverage the specific 'spikes' of each model.

# True Expertise Shifts from Information Possession to Cognitive Direction

## Human Imperatives

- Ethical reasoning and moral judgment
- Embodied, physical knowledge and connection
- Creative innovation and meaning-making

## Human AI Collaboration Second-Order Knowledge Work

Directing, contextualizing, and evaluating AI-processed information to enhance human autonomy.

## AI Capabilities

- Vast dataset processing and instant retrieval
- Algorithmic pattern recognition and simulation
- Real-time multi-modal generation

# Questions and Comments

Ray Uzwyshyn, Ph.D. MBA MLIS  
Acting AUL for Research & Technology Services  
Director of Research, UC Riverside Libraries,  
<https://www.linkedin.com/in/rayuzwyshyn>  
raymondu@ucr.edu , <https://rayuzwyshyn.net>

# Background Reports, Brief Videos and Benchmarking Websites

## (Full Bibliography in First Report Analysis)

Uzwyszyn, R. Benchmarking Top Agentic AI Deep Research Models: The Jagged Topographic Frontier. (Whitepaper Preprint Draft, May 2026). ResearchGate.

[https://www.researchgate.net/publication/404740788\\_Benchmarking\\_Top\\_Agentic\\_AI\\_Deep\\_Research\\_Models\\_The\\_Jagged\\_Topographic\\_Frontier](https://www.researchgate.net/publication/404740788_Benchmarking_Top_Agentic_AI_Deep_Research_Models_The_Jagged_Topographic_Frontier)

Uzwyszyn, R. AI LLMs 2026: The Leading Edge of the Jagged Frontier: A Practitioner's Map to Machine Intelligence and the Human Capacities It Cannot Reach. Orbach AI Literacy Series, University of California, Riverside. February 2026. <https://www.linkedin.com/pulse/ai-llms-2026-leading-edge-jagged-frontier-raymond-uzwyszyn-ph-d--lx1tc/>

Uzwyszyn, R. Beyond Traditional AI IQ Metrics: Metacognition and Benchmarking for LLMs, AGI, and ASI . Journal of Human Behavior and Emerging Technologies, 2026. (Forthcoming). <https://rayuzwyszyn.net/MSU2024/AIBenchmarking/BenchmarkingAI13Uzwyszyn.pdf>

### **Brief Video Overviews**

Uzwyszyn, R. New [LLM Top Models](#). [AI LLM Benchmarking 2026](#). [New Agentic AI and The Jagged Frontier](#). May 2026.

### **Websites**

Uzwyszyn, R. The Jagged Frontier: A Topographic Atlas of AI Capability for Academic Research. (May 2026). <https://claude.ai/public/artifacts/2693929e-a36c-48ed-a99c-457f2781af40>

Uzwyszyn, R. Frontier Model Benchmarks: 10 models 8 Benchmarks : Three-tier verification (February 2026). <https://claude.ai/public/artifacts/96e81b36-d84f-491c-829d-6a76f7ed1b56>

AI Model Intelligence & Economics Platform Comparison. June 2025. <https://claude.ai/public/artifacts/8dd54f19-99e0-46c3-8b61-be04336d4d70>





# Five Distinct Cognitive Instruments

## GPQA Diamond (Graduate Reasoning)



- 198 Google-proof Qs.
- PhD human ceiling: 65–74%
- AI ceiling: 94.3% (Gemini 3.1 Pro)
- Status: **Solved.**

## HLE (Humanity's Last Exam)



- 2,500 questions from 1,000 PhDs.
- Human expert: ~90%
- AI ceiling: 54.7% (Claude 4.7)
- Status: ~35% gap remains.

## SWE-bench Verified (Software Engineering)



- Real GitHub bugs.
- AI ceiling: 87.6% (Claude 4.7)
- Status: 9 in 10 correctly patched.

## ARC-AGI-2 (Fluid Intelligence)



- Novel visual puzzles, resists scale.
- Human avg solve: 2.3 mins.
- AI ceiling: 85% (GPT-5.5)
- Status: Deeply uneven across models.

## AA Intelligence Index v4.0 (Composite)



- 10 standardized evaluations.
- AI ceiling: 60 (GPT-5.5)
- Status: Broad indicator.

# The Convergence of Global Knowledge Parity

## Claude Opus 4.7

Proprietary, Trillion-dollar incumbent  
54.7% on Humanity's Last Exam

## Kimi K2.6

Beijing lab, Open-weights, 1T parameters  
54.0% on Humanity's Last Exam



The Chinese open-source model equals the world's most capable proprietary model on humanity's deepest test of cross-domain expert knowledge.

## API Cost per 1M Input/Output Tokens

Claude Opus 4.7

**\$5 / \$25**

+ 0-35%  
token tax

Gemini 3.1 Pro

**\$2 / \$12**

DeepSeek V4 Pro

**\$1.74 / \$3.48**

(with 75% promo discount)

Cost-performance offerings are rapidly expanding budget-constrained academic research infrastructure.

# The May 2026 Academic Tech Stack

## Capstone: The Human Orchestrator.

Judging which model to consult, when to verify, and how to synthesize across complementary cognitive profiles.

**The human is not peripheral to this architecture. It is the architecture.**

## Level 3 (Split Layer) – Left Side:

GPT-5.5 (Agentic workflows, code, abstract reasoning).

## Right Side:

Claude Opus 4.7 + Kimi K2.6 (Cross-disciplinary synthesis, factual reliability at the edge of human knowledge).

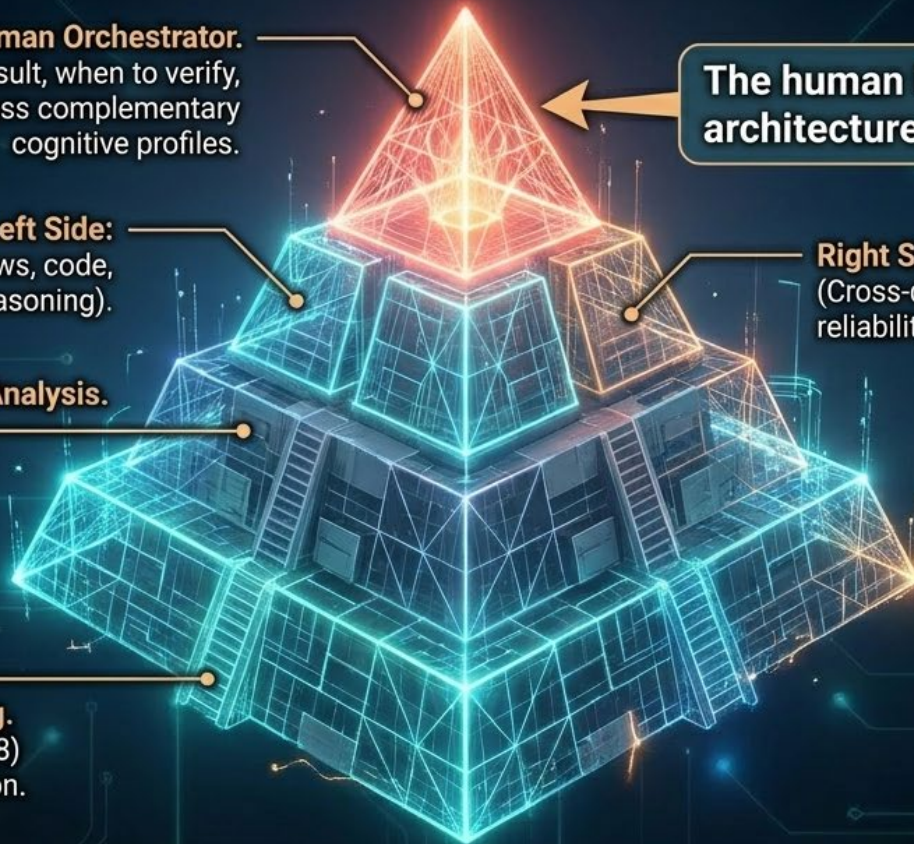
## Level 2: Specialized & STEM Analysis.

Gemini 3.1 Pro Preview (best GPQA price/performance) for large-document synthesis and STEM.

## Level 1 (Base):

### High-Volume Bulk Processing.

DeepSeek V4 Flash (\$0.14/\$0.28) for initial screening and extraction.  
Escalate to DeepSeek V4 Pro.



2	Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scaling>20:1)	ALScore "ALScore" i: Sqr Root of	MMLU	MMLU -Pro	GPQA	HLE	Training dataset	Announced
3	AuroraGPT (ScienceGPT)	Argonne National Lab	<a href="https://lifelarchive">https://lifelarchive</a>	2000	30000	15:1	25.8						TBA
4	GPT-6	OpenAI	<a href="https://lifelarchive">https://lifelarchive</a>										TBA
5	Grok-5	xAI	<a href="https://lifelarchive">https://lifelarchive</a>	10000	200000	20:1	149.1						TBA
6	Nemotron 3 Ultra	NVIDIA		500	25000	50:1	11.8					synthetic, web-scale	TBA
7	Nemotron 4	Consortium	<a href="https://www.tom">https://www.tom</a>										TBA
8	DeepSeek-V4-Pro	DeepSeek-AI	<a href="https://huggingfa">https://huggingfa</a>	1600	33000	22:1	10.6	90.1	87.5	90.1	37.7	synthetic, web-scale	Apr/2026
9	Hy3 preview	Tencent	<a href="https://huggingfa">https://huggingfa</a>	295	40000	136:1	11.5	87.42	65.76		30	synthetic, web-scale	Apr/2026
10	GPT-5.5	OpenAI	<a href="https://chatgpt.cc">https://chatgpt.cc</a>	3000	114000	38:1	61.6			93.6	57.2	synthetic, web-scale	Apr/2026
11	Marul V7	Independent	<a href="https://marulai.cc">https://marulai.cc</a>	0.258	1000	3,876:1	0.1					web-scale	Apr/2026
12	Qwen3.6-27B	Alibaba	<a href="https://huggingfa">https://huggingfa</a>	27	36000	1,334:1	3.3		86.1	87.8	24.3	synthetic, web-scale	Apr/2026
13	MiMo-V2.5-Pro	Xiaomi	<a href="https://aistudio.xi">https://aistudio.xi</a>	1000	27000	27:1	17.3				48	synthetic, web-scale	Apr/2026
14	Ling-2.6-Flash	Inclusion AI	<a href="https://ling.tbox.c">https://ling.tbox.c</a>	104	20750	200:1	4.9					synthetic, web-scale	Apr/2026
15	Granite-4.1-8B	IBM	<a href="https://huggingfa">https://huggingfa</a>	8	15000	1,875:1	2.3	73.84	55.99	41.96		synthetic, web-scale	Apr/2026
16	OpenMythos	Independent	<a href="https://github.cor">https://github.cor</a>	0.77	30	39:1	18.4					web-scale	Apr/2026
17	Kimi K2.6	Moonshot AI	<a href="https://huggingfa">https://huggingfa</a>	1000	30500	31:1	18.4			90.5	54	synthetic, web-scale	Apr/2026
18	Qwen3.6-Max-Preview	Alibaba	<a href="https://chat.qwer">https://chat.qwer</a>	2000	36000	18:1	20.0					synthetic, web-scale	Apr/2026
19	Qwen3.6-35B-A3B	Alibaba	<a href="https://huggingfa">https://huggingfa</a>	35	36000	1,029:1	3.7		85.2	86	21.4	synthetic, web-scale	Apr/2026
20	Grok 4.3	xAI	<a href="https://grok.com/">https://grok.com/</a>	3000	80000	27:1	51.6					synthetic, web-scale	Apr/2026
21	Claude Opus 4.7	Anthropic	<a href="https://claude.ai/">https://claude.ai/</a>	400	80000	200:1	18.9			94.2	54.7	synthetic, web-scale	Apr/2026
22	GPT-5.4-Rosalind	OpenAI		70	114000	1,629:1	9.4					synthetic, web-scale	Apr/2026
23	GPT-5.4-Cyber	OpenAI		70	114000	1,629:1	9.4					synthetic, web-scale	Apr/2026
24	Marco-Mini	Alibaba	<a href="https://huggingfa">https://huggingfa</a>	17.3	36000	2,081:1	2.6	83.4	70.7	50.3		synthetic, web-scale	Apr/2026

## April 2026 Top Models List

# THE ENTREPRENEURIAL CALCULUS

## TIER 3: HUMAN-IRREPLACEABLE (The Frontier)

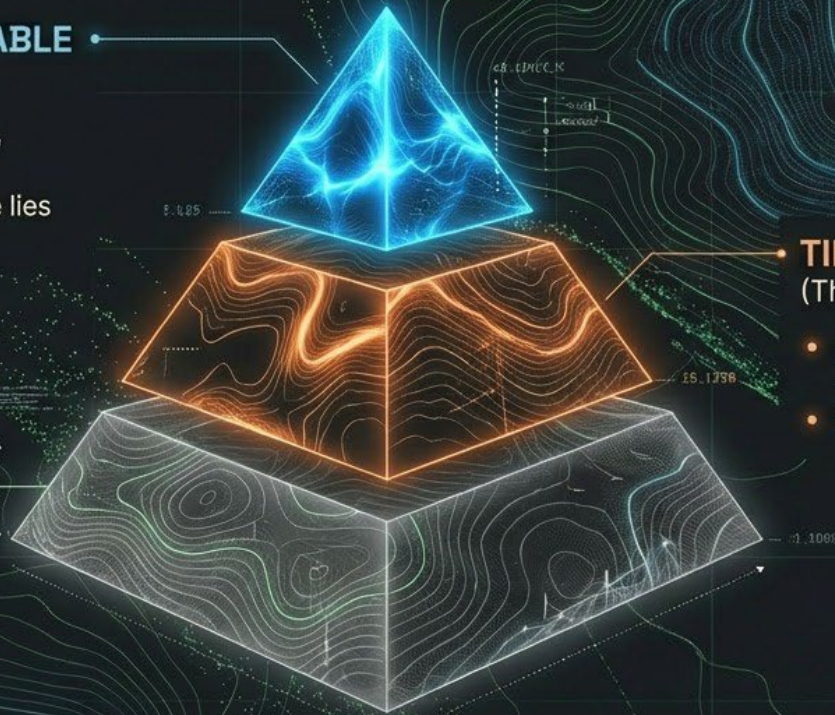
- **Capabilities:** Community, Strategy, Creative Direction.
- **Strategy:** Build communities. Value lies in what cannot be called via API.

## TIER 1: COMMODITIZED (The Lowlands)

- **Capabilities:** Reasoning, Math, Coding.
- **Strategy:** Build Platforms. Margins will collapse.

## TIER 2: DIFFERENTIATING (The Slopes)

- **Capabilities:** Agentic tools, Specialized Knowledge.
- **Strategy:** Model Orchestration.



# The Ascent of AI Capability: 2022–May 2026

Half-Year Progress of Leading AI Models Compared to Human Performance Benchmarks

The GCI is a normalized composite index (0–1) reflecting performance across diverse benchmarks: MMLU, GPQA, MATH, HumanEval, SWE-bench, MMMU, and frontier real-world task evaluations. It is not "intelligence" in a philosophical sense but a pragmatic capability metric.

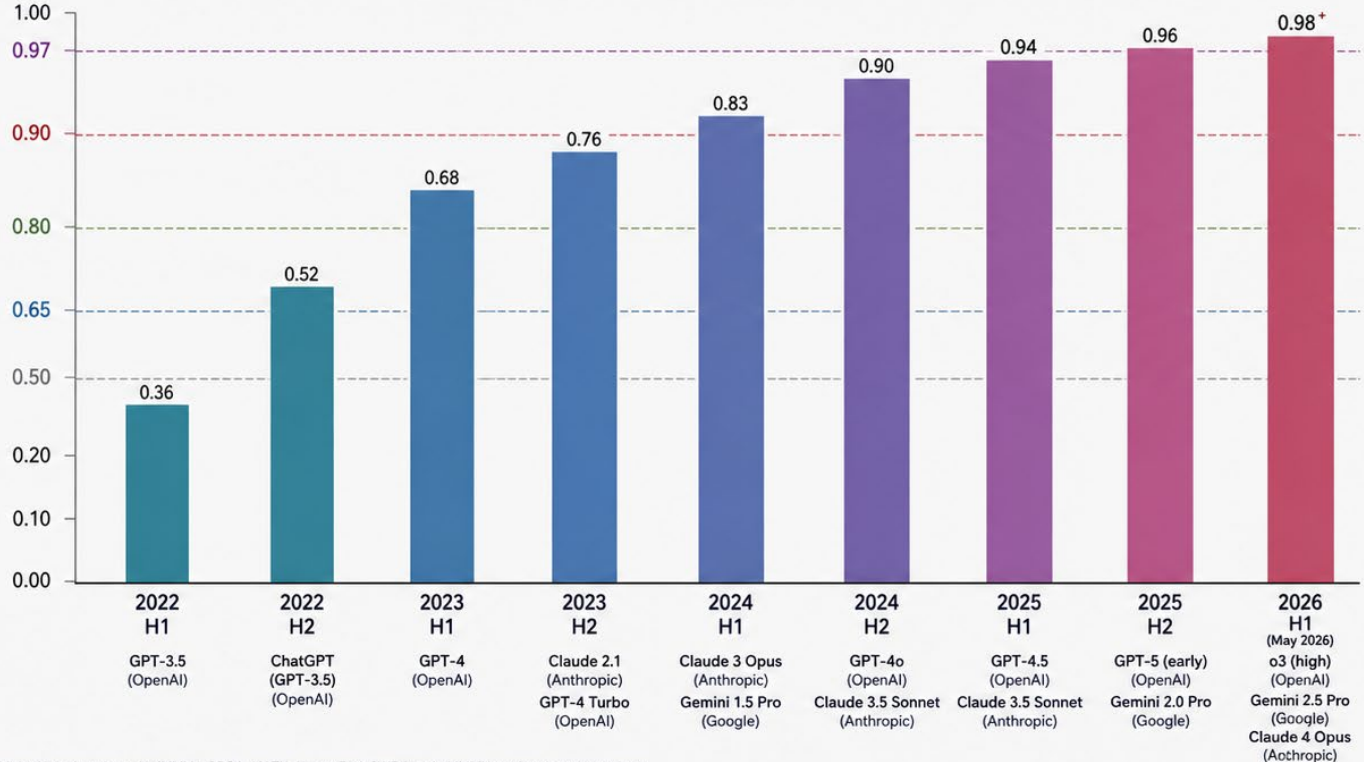
## General Capability Index (GCI)

Composite of reasoning, knowledge, coding, math, and real-world task performance (0.00 = chance / random; 1.00 = top of human performance distribution)

## Interactive Website Comparison

<https://claude.ai/public/artifacts/2693929e-a36c-48ed-a99c-457f2781af40>

- Top in the World**  
Top 0.1% of humans  
(≈ 99.9th percentile)
- Post Doc / Expert**  
Top 1% of humans  
(≈ 99th percentile)
- Ph.D. Level**  
Top 10% of humans  
(≈ 90th percentile)
- Average Undergraduate**  
Top 50% of humans  
(≈ median)
- Average Human Adult (Global)**  
(= mean)



### Reading the Chart

- ▶ Bars show the estimated General Capability Index (GCI) for top frontier models at each half-year point.
- ▶ Human benchmarks (left) are approximate percentiles of global adult population.
- ▶ Crossing a line means the AI model's composite performance is comparable to that human level on this index.
- ▶ Steady gains reflect broader capabilities and benchmark saturation approaching the frontier of human expertise.

**Model Provider (Color Key)**

- OpenAI
- Anthropic
- Google

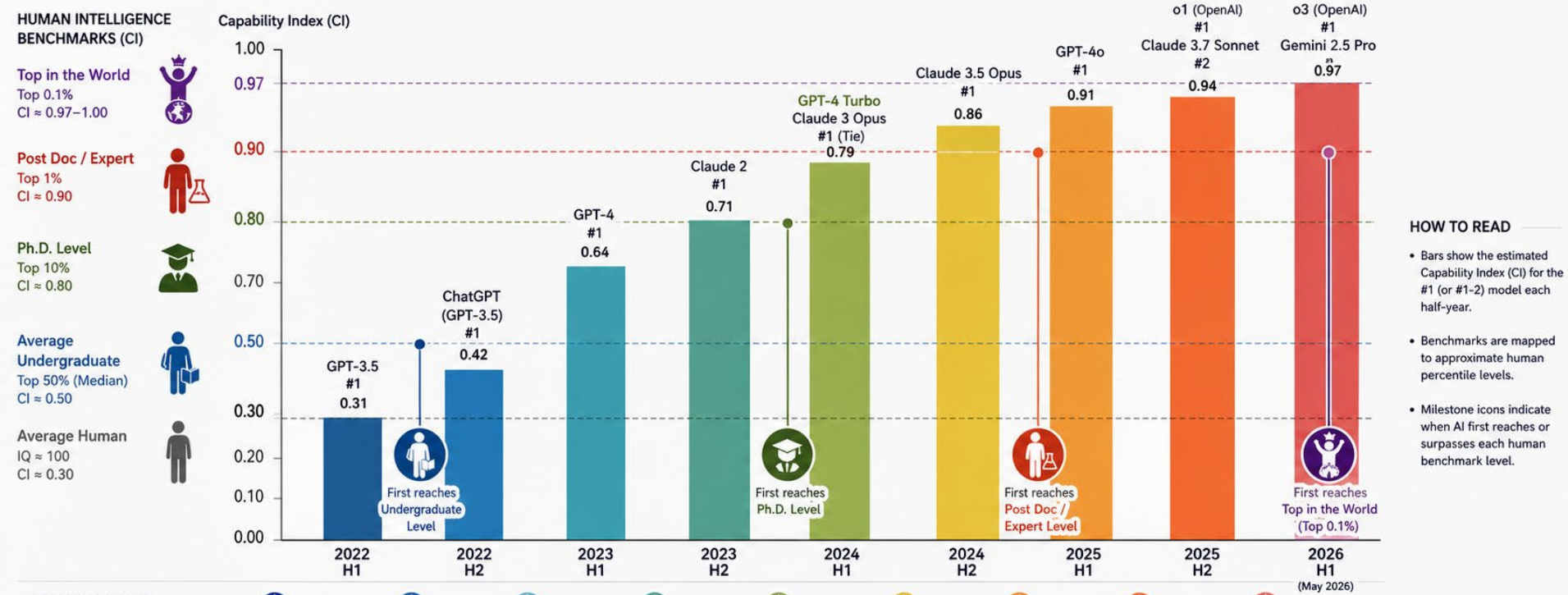
Note: Values are estimates based on published benchmark results (MMLU, GPQA, MATH, HumanEval, SWE-bench, MMMU, etc.) and expert synthesis. Benchmarks differ in difficulty and scope, GCI should be interpreted as a directional indicator, not a precise measurement of general intelligence.

Sources: OpenAI, Anthropic, Google DeepMind, Stanford HELM, Papers. With Code, arXiv, arXiv, AI benchmark leaderboards (2022–May 2026)

# The Ascent of AI Capability: 2022–May 2026

Top AI Models by Half-Year Compared to Human Intelligence Benchmarks

**CAPABILITY INDEX (CI):** Composite of leading benchmarks including MMLU, GPQA, MATH-500, HumanEval, SWE-bench, MMMU, and real-world task evaluations.  
 0.0 = Random Chance    0.5 = Average Human (IQ=100)    1.0 = Top of Human Distribution



**PERIOD ICON KEY**

Winter-Spring	Summer-Fall	Winter-Spring	Summer-Fall	Winter-Spring	Summer-Fall	Winter-Spring	Summer-Fall	Winter-Spring
---------------	-------------	---------------	-------------	---------------	-------------	---------------	-------------	---------------

**TOP MODELS (Global)**

Rank	2022 H1	2022 H2	2023 H1	2023 H2	2024 H1	2024 H2	2025 H1	2025 H2	2026 H1
#1	GPT-3.5	ChatGPT (GPT-3.5)	GPT-4	Claude 2	GPT-4 Turbo	Claude 3.5 Opus	GPT-4o	o1 (OpenAI)	o3 (OpenAI)
#2	PaLM (540B)	Stable Diffusion (Emerging)	PaLM-E	GPT-4 Turbo (Preview)	Claude 3 Opus (Tie)	Gemini 1.5 Pro	Gemini 1.5 Pro	Claude 3.7 Sonnet	Gemini 2.5 Pro

Notes: CI values are estimates based on aggregated benchmark performance and expert analysis. Benchmarks and model capabilities evolve; values should be interpreted as directional, not absolute. Sources: OpenAI, Anthropic, Google DeepMind, Meta AI, Mistral AI, Papers With Code, Artificial Analysis, Stanford HELM, Epoch AI, and public benchmark leaderboards. Data current as of May 2026.

# THE ASCENT OF AI CAPABILITY: 2018 – MAY 2026

## TOP AI MODELS BY HALF-YEAR COMPARED TO HUMAN INTELLIGENCE BENCHMARKS

### CAPABILITY INDEX (CI) METHODOLOGY

Composite of leading benchmarks: MMLU, GPQA (PhD-level science), MATH-500, HumanEval / SWE-bench (coding), MMMU (multimodal), and real-world task evaluations. Scores normalized to the estimated human performance distribution (percentile). 0.50 = average human.

### CAPABILITY INDEX (CI)

Percentile of Human Performance

#### TOP 0.1% OF HUMANS

≈ 99.9th percentile  
Exceptional across all domains



#### POSTDOC / EXPERT

≈ 93rd percentile  
Research-level expertise



#### PHD LEVEL

≈ 85th percentile  
Doctoral-level knowledge & reasoning



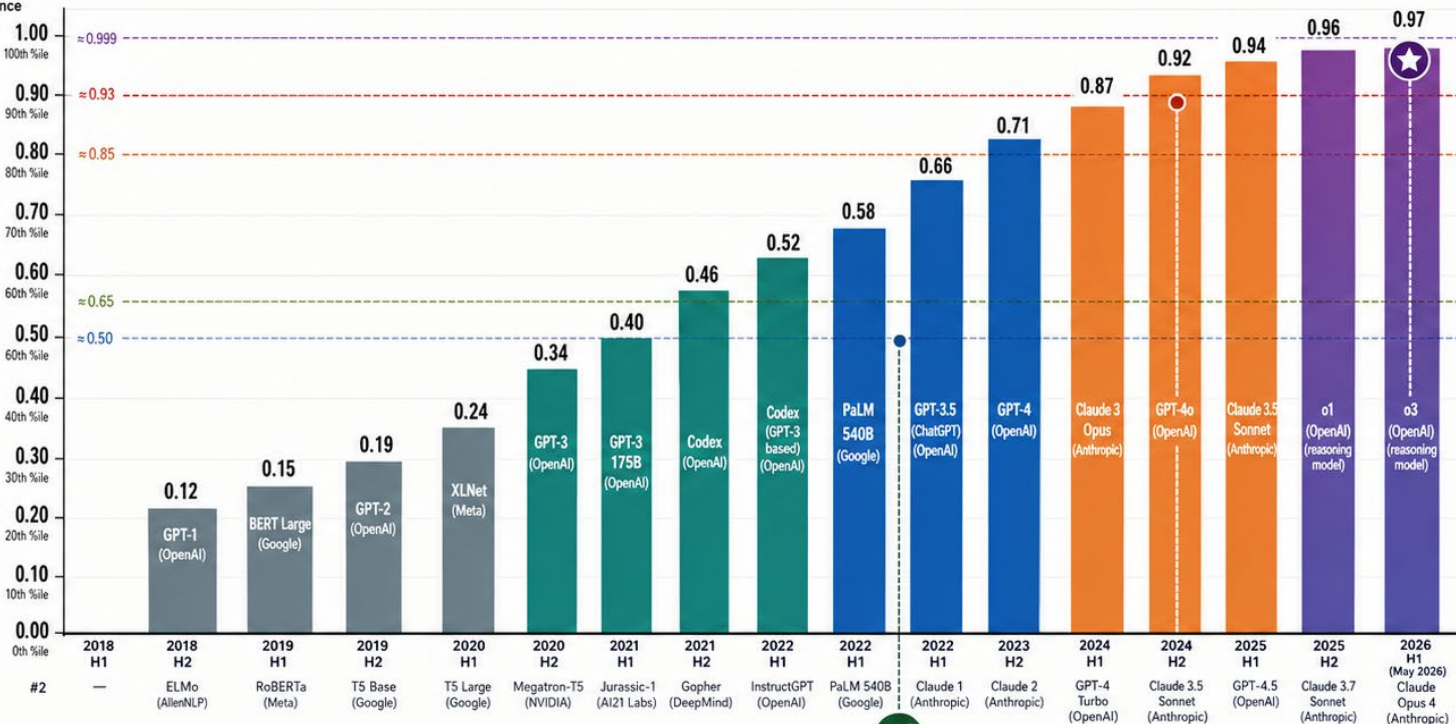
#### UNDERGRADUATE

≈ 65th percentile  
Structured knowledge & problem solving



#### AVERAGE HUMAN

≈ 50th percentile  
Typical adult performance



### HOW TO READ

- Bar height = estimated Capability Index (CI) percentile.
- Top line in each bar = #1 model and organization.
- Bottom row = #2 model for that period.
- Milestone icons show the first time AI systems reach or exceed each human benchmark level.

### PROVIDER KEY

- OpenAI
- Anthropic
- Google DeepMind
- Other / Open Source
- Pre-ChatGPT Era

### NOTES

CI is directional, not absolute. Benchmarks evolve; results vary by test and methodology. Human distribution based on cognitive testing and academic performance percentiles.

### KEY CAPABILITY HIGHLIGHTS (What defined the era)

**FOUNDATION MODELS**  
Learning representations of language

**SCALE EMERGENCE**  
Few-shot learning and broad task generalization

**INSTRUCTION & ALIGNMENT**  
RLHF and human feedback improve usefulness

**REASONING BREAKTHROUGH**  
Strong performance on exams, logic, and professional tasks

**First crosses Undergraduate**  
≈ 2020–2022

**MULTIMODAL & LONG CONTEXT**  
Vision, audio, and 100k+ token context understanding

**First crosses Postdoc / Expert**  
≈ Late 2024

**REASONING-NATIVE ERA**  
Deliberative reasoning, planning, and complex agents

**Approaches Top 0.1% of Humans**  
≈ 2026

# Five Cognitive Instruments: Reading the Dashboard



**Title:** GPQA Diamond  
**Readout:** 94%+  
**Concept:** 198 multi-step grad-level inference questions.  
**Status:** Solved. Differentiation collapsed.



**Title:** Humanity's Last Exam (HLE)  
**Readout:** 54.7% Max  
**Concept:** 2,500 expert questions from 100+ disciplines.  
**Status:** Gap to 90% human expert ceiling remains vast.



**Title:** SWE-bench Verified  
**Readout:** 87.6% Max  
**Concept:** Real GitHub bug patches.  
**Status:** 9 in 10 accurate.

**Title:** ARC-AGI-2  
**Readout:** 85% Max  
**Concept:** Fluid visual generalization resisting scale.  
**Status:** Below human average of 66% for most models.



**Title:** AA Index v4.0  
**Readout:** 60 Max  
**Concept:** Composite capability.  
**Status:** Rigorous 10-eval index on first-party APIs.

# The Human Value Fortress: Three Deepening Uncharted Territories

## The Hallucination-Reliability Gap

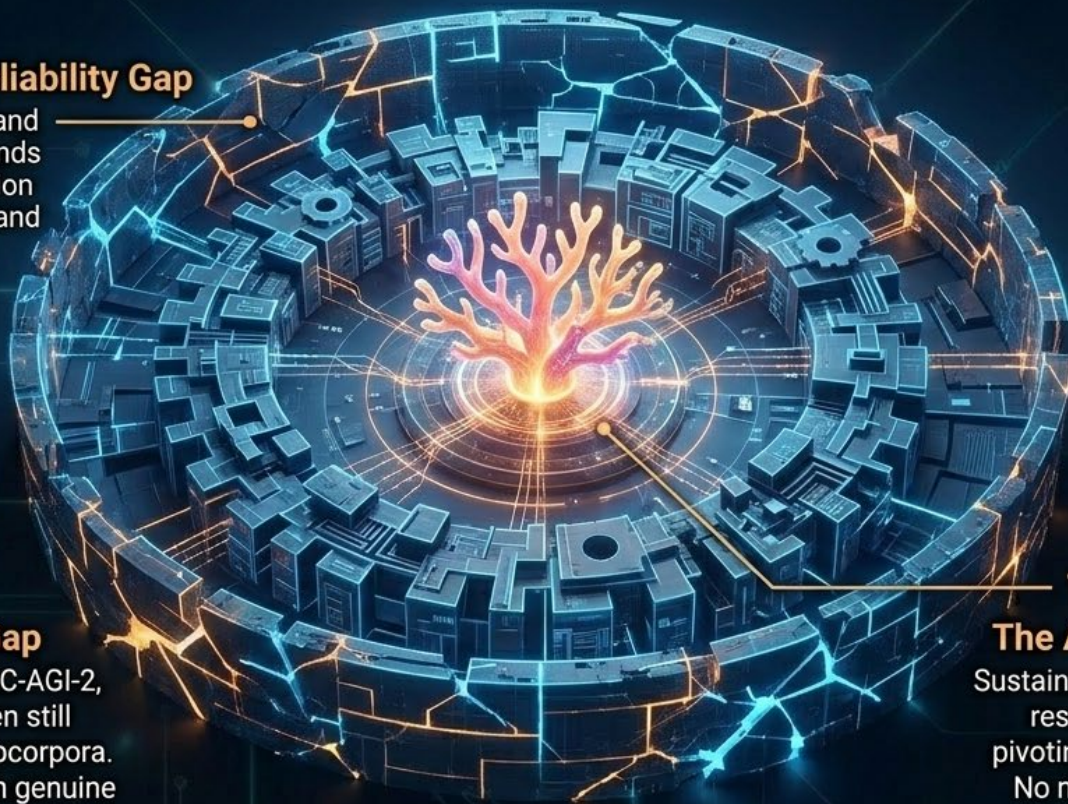
The gap between knowing and accurately reporting. Demands human systematic verification protocols in medicine, law, and humanities.

## The Abstraction Gap

GPT-5.5 reaches 85% on ARC-AGI-2, but 6-to-8-year-old children still outperform frontier AI on subcorpora. The cognitive engine beneath genuine scientific discovery resists scale.

## The Protected Core The Agentic-Judgment Gap

Sustained project leadership, adaptive research design, and creative pivoting under genuine uncertainty. No model achieves 85% on real-world Terminal-Bench Hard.



# A Tiered Collaboration Architecture



TECHNICAL TELEMETRY  
SEINE CROPPORTSL. L00-SESIONES FOR SORVYGS2500K

## Claude Opus 4.7 + Kimi K2.6

TEMN: 37098, COSSINET SCALE "ETN" → E56. 1238

Highest-order reasoning and multi-step problem solving. Optimized for novel generalization and rigorous accuracy.

SMALL TECHNICAL TELEMETRY LINES SMALL TECHNICAL TELEMETRY LINE

COBR \$1.53' \$3.50

TECHNICAL TELEMETRY  
SEINE ERPPORTTL. EAN-DESIONES FOR SORVYGS2200K

## Gemini 3.1 Pro Preview (Best verified GPQA at optimal price)

TEMN: 38368, COSSINET SCALE "ETN" → E56. 1238

Superior performance in technical domains, long context windows, and complex data integration.

SMALL TECHNICAL TELEMETRY LINES SMALL TECHNICAL TELEMETRY LINE

COST \$1.74' \$3.48

TECHNICAL TELEMETRY  
SEINE CROPPORTSL. L00-SESIONICS FOR SORVYGS2200K

## GPT-5.5 (Unmatched abstraction, requires human fact-verification)

TEMN: 37099, COOSINET SCALE "EDN" → E56. 1238

Advanced code generation and autonomous agent execution. Essential human-in-the-loop validation for factuality.

COBR \$1.55' \$3.50

TECHNICAL TELEMETRY  
SEINE ERPPORTSL. EAN-DESIONES FOR SORVYGS2200K

## DeepSeek V4 Flash / Pro (Unbeatable economics for classification/extraction)

TEMN: 37893, COOSINET SCALE "EEN" → E56. 1238

Extreme cost-efficiency for large-scale data classification, extraction, and routine tasks.

COST \$1.74' \$3.48

# THE THREE STRUCTURAL GAPS (HUMAN TERRITORY)

February 2026

## THE ABSTRACTION GAP

A: 0.8956/C



A: 0.8956/C

JetBrains Mono

**Focus:** Scientific discovery & artistic creation (ARC-AGI).

**Role:** Seeing the pattern no one has seen before.

Pattern:  
[GENERATE]

## THE AGENTIC GAP

Δ: 34.2%



Δ: 34.2%

JetBrains Mono

**Focus:** Navigating uncertainty (Terminal-Bench).

**Role:** Crisis response and pivoting where there is no test suite.

Status:  
[NAVIGATING]

## THE WISDOM GAP

REF: 1.618



REF: 1.618

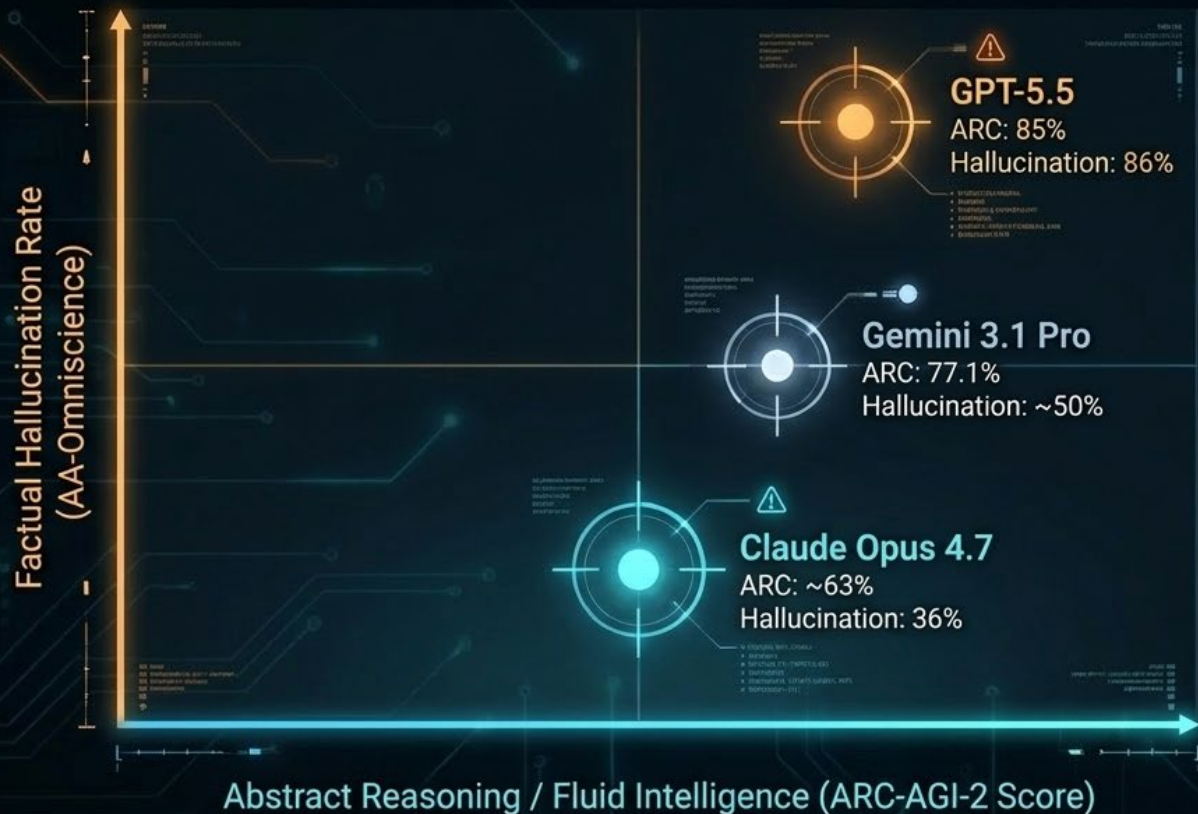
JetBrains Mono

**Focus:** The Meta-Layer.

**Role:** Deciding when to consult the encyclopedia vs. when to reason from first principles.

Strategy:  
[EVALUATE]

# The Reliability Paradox: Intelligence vs. Fabrication



The model that knows more fabricates more confidently. GPT-5.5 leads composite intelligence but records an 86% hallucination rate. In academic research, this is not a flaw to dismiss—it is a diagnostic requiring systematic verification.

# Comments and Questions