The Memo - 1/Dec/2024
1 message

LifeArchitect.ai <lifearchitect@substack.com> Fri, Nov 29, 2024 at 5:03 PM
Reply-To: "LifeArchitect.ai"
<reply+2ig3bx&hfar7&&5a5c533ba5946d8601394bc303039024b7660ff3d834ccd160577883103e754c@mg1.substack.com>
[redacted]

The Memo - 1/Dec/2024

19 new models for Nov/2024, ChatGPT falls out of top 10 LLMs, Claude use doubles, and much more!

DR ALAN D. THOMPSON
NOV 29 · PREVIEW

To: US Govt, major govts, Microsoft, Apple, NVIDIA, Alphabet, Amazon, Meta, Tesla, Citi, Tencent, IBM, & 10,000+ more recipients…
From: Dr Alan D. Thompson <LifeArchitect.ai>
Sent: 1/Dec/2024
Subject: *The Memo* - AI that matters, as it happens, in plain English
AGI: 83%
ASI: 0/50 (no expected movement until post-AGI)

Dr Eric Schmidt (20/Nov/2024 or watch the video):
'I can assure you that the humans in the rest of the world, all the normal people… are not ready. Their governments are not ready. The government

processes are not ready. The doctrines are not ready. They're not ready for the arrival of [artificial intelligence].'

The winners of *The Who Moved My Cheese? AI Awards!* for Nov/2024 are director Billy Ray and actor Ben Affleck ('[AI's] not going to replace human beings making films').

Contents

The BIG Stuff

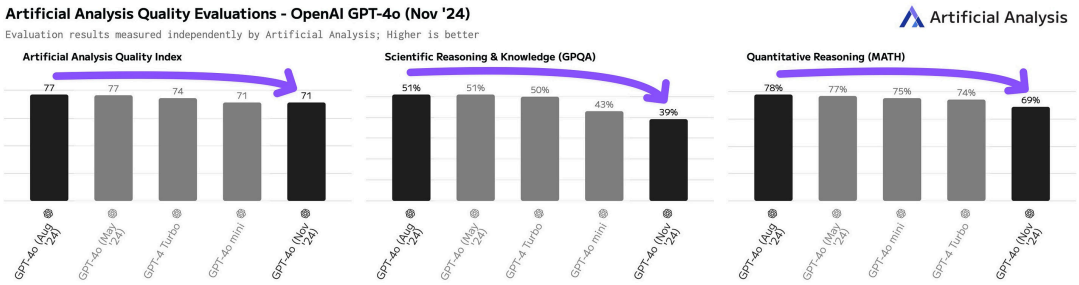
Exclusive: Default ChatGPT model falls out of top 10 LLMs (22/Nov/2024)

ChatGPT turns two years old today, 30/Nov/2024. OpenAI recently updated the default ChatGPT model to *GPT-4o-2024-11-20*, with significant drops in major test scores for Aug/2024 vs Nov/2024:

GPQA: -7.1 points out of 53.1 = **-13.37%**
MMLU: -3.0 points out of 88.7 = **-3.38%**

See OpenAI's official evaluations: https://github.com/openai/simple-evals

Artificial Analysis ran their own evaluations and noted even worse performance:



https://x.com/ArtificialAnlys/status/1859614633654616310

Given that OpenAI currently has to serve a resource-heavy model across ***three billion visits per month*** (16/Oct/2024), it makes sense to optimize how and what they serve to users via available hardware. It also seems that OpenAI has favoured 'popular' responses (via user ratings on LMSYS) instead of 'smarter' responses (via the GPQA and MMLU benchmarks above).

Alan's advisory note: As of Nov/2024, the default ChatGPT model *GPT-4o* is no longer in the top 10 models by smarts. Given the model's significantly diminished performance, I recommend avoiding using *GPT-4o* for now, and using alternatives like the state-of-the-art *Claude 3.5 Sonnet (new)* via Claude.ai instead.

Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scalings 20:1)	AI Score "AI Score" i Sqr Root of	MMLU	MMLU -Pro	GPQA
o1	OpenAI	https://chatgpt.com/	200	20000	100:1	6.7	92.3	91	78.3
Claude 3.5 Sonnet	Anthropic	https://poe.com/Claude-3.5-Sonnet					88.7	76.1	67.2
QwQ-32B	Alibaba	https://huggingface.co	32	18000	563:1	2.5			65.2
Claude 3.5 Sonnet (	Anthropic	https://claude.ai/					90.5	78	65
Claude 3 Opus	Anthropic	https://claude.ai/	2000	40000	20:1	29.8	86.8	68.5	59.5
Gemini-1.5-Pro-002	Google DeepMind	https://aistudio.google	1500	30000	20:1	22.4		75.8	59.1
DeepSeek-R1-Lite	DeepSeek-AI	https://chat.deepseek	67	2000	35:1	4.6			58.5
Grok-2	xAI	https://x.com/i/grok	600	15000	25:1	10.0	87.5	75.5	56
GPT-4o	OpenAI	https://chatgpt.com/	200	20000	100:1	6.7	88.7	72.6	53.6
Llama 3.1 405B	Meta AI	https://www.meta.ai/	405	15600	39:1	8.4	88.6	73.3	51.1
gpt-4-turbo-2024-04	OpenAI	https://chat.openai.com/		13000			86.5	63.7	49.1
Qwen2.5	Alibaba	https://huggingface.co	72	18000	250:1	3.8	86.1	71.1	49
Yi-XLarge	01-ai	https://platform.01.ai	2000	20000	10:1	21.1	85.1		48.2
GPT-4 Turbo	OpenAI	https://chat.openai.com/		13000			86.4		46.5
Gemini 1.5 Pro	Google DeepMind	https://aistudio.google	1500	30000	20:1	22.4	85.9	69	46.2
gpt-4o-2024-11-20	OpenAI	https://chat.com/	200	20000	100:1	6.7	85.7		46

Models Table. GPT-4o Nov/2024 ranks #16 by GPQA score.
Click to enlarge.

Exclusive: 19 new models for November (Nov/2024)

November 2024 was another big month, with a major LLM release every ~37 hours. Google released two finetunes of Gemini: *gemini-exp-1114* and *gemini-exp-1121* (Nov/2024), but—in general—finetunes are not counted or included in the Models Table. There were four major clones of OpenAI's o1 reasoning model out of China, with none scoring higher than 1/5 on the ALPrompt 2024 H2 (where o1 scores 5/5). The 19 model highlights are:

1. **Hugging Face SmolLM2 (1.7B on 1T tokens)**
   Base and instruct versions, released under Apache 2.0 license. (Playground, Paper)
2. **AMD OLMo (1B on 1.3T tokens)**
   1 billion parameter LMs trained from scratch using 1.3T tokens on a cluster of AMD Instinct MI250 GPUs. (Playground, Paper)
3. **AI Singapore SEA-LIONv3 (9.24B on 8.2T tokens)**
   SEA-LION is a collection of LLMs pretrained and instruct-tuned for the Southeast Asia (SEA) region, with continued pretraining from Gemma-2-9B. (Playground, Paper)
4. **Tencent Hunyuan-Large (389B on 7T tokens)**
   Pretrained on 7T tokens, with nearly 1.5T tokens of high-quality synthetic data. Capable of handling up to 256K tokens. (Playground, Paper)
5. **TensorOpera Fox-1 (1.6B on 3T tokens)**
   Gold standard for dataset documentation. (Playground, Paper)
6. **Alibaba Qwen2.5-Coder (32.5B on 5.5T tokens)**
   Suitable for coding tasks. (Playground, Paper)
7. **Fireworks f1**
   A compound AI model specialized in complex reasoning, interweaving multiple open models at the inference layer. (Playground, Paper)
8. **Mistral Pixtral Large (124B on 6T tokens)**
   Pretrained and instruct-tuned for advanced tasks. (Playground, Paper)
9. **XiaoduoAI Xmodel-LM (1.1B on 2T tokens)**
   SLM model optimized for large datasets. (Playground, Paper)

10. **DeepSeek-AI DeepSeek-R1-Lite (67B on 2T tokens)**
    o1 reasoning model copy. Scores 0/5 on latest ALPrompt 2024 H2. Still in development, supports web usage but no API calls yet. Future updates will include full open-source release. (Playground, Paper)
11. **OpenAI GPT-4o-2024-11-20**
    Material decrease in benchmark scores compared to Aug 2024. Possibly pruned or quantized. (Playground, Paper)
12. **Allen AI TÜLU 3 (70B on 15.6T tokens)**
    Llama 3.1 post-training with new Reinforcement Learning with Verifiable Rewards (RLVR). Worse performance on most benchmarks. (Playground, Paper)
13. **Alibaba Marco-o1 (7B on 7T tokens)**
    o1 reasoning model copy. Trained on multiple instruction datasets. (Playground, Paper)
14. **Moonshot AI k0-math (100B on 2T tokens)**
    o1 reasoning model copy. Specialized math reasoning model with extended context capability. Limited public details, Chinese-focused. (Playground, Paper)
15. **CMU Bi-Mamba (2.7B on 1.26T tokens)**
    Unreleased but will be replicated. Features a scalable and efficient 1-bit Mamba architecture. (Paper)
16. **Prime Intellect INTELLECT-1 (10B on 1T tokens)**
    First decentralized training run of a 10-billion-parameter model, with public participation from HF and Dylan at SemiAnalysis. (Paper)
17. **Allen AI OLMo 2 (13B on 5.6T tokens)**
    Open Language Model under Apache 2.0 for research and education. Includes multi-epoch training on diverse datasets. (Playground, Paper)
18. **OpenGPT-X Teuken-7B (7B on 4T tokens)**
    Multilingual instruction-tuned model covering 24 EU languages with emphasis on non-English content. Developed with European values in mind. (Playground, Paper)
19. **Alibaba QwQ-32B (32B on 18T tokens)**
    o1 reasoning model copy. Scores 1/5 on latest ALPrompt 2024 H2. Known as ‘Qwen with Questions’ (QwQ). (Playground, Paper)

See them on the Models Table: https://lifearchitect.ai/models-table/

A Minecraft town of AI characters made friends, invented jobs, and spread religion (27/Nov/2024)

AI startup Altera used LLM-powered agents in Minecraft to simulate humanlike behaviors at scale. These agents autonomously developed complex social dynamics, including forming friendships, creating jobs, and even spreading a parody religion (Pastafarianism).

> Over 12 in-game days (4 real-world hours) the agents began to exhibit some interesting emergent behavior.
>
> For example, some became very sociable and made many connections with other characters, while others appeared more introverted.
>
> …in one case an AI chef tasked with distributing food to the hungry gave more to those who he felt valued him most.
>
> More humanlike behaviors emerged in a series of 30-agent simulations. Despite all the agents starting with the same personality and same overall goal—to create an efficient village and protect the community against attacks from other in-game creatures—**they spontaneously developed specialized roles within the community, without any prompting.**
>
> **They diversified into roles such as builder, defender, trader, and explorer. Once an agent had started to specialize, its in-game actions began to reflect its new role. For example, an artist spent more time picking flowers, farmers gathered seeds and guards built more fences.**
>
> "We were surprised to see that if you put [in] the right kind of brain, they can have really emergent behavior," says Yang. "That's what we expect humans to have, but don't expect machines to have."

Read more via MIT Technology Review.

I anticipate that these kinds of LLM-based simulations will help ease us into the era of artificial superintelligence in the coming years. Consider a world simulation by a future AI-designed frontier model like ***GPT-7*** or ***Claude 5*** that

can replicate and ‘prove’ a more efficient and maybe 1,000× optimized version of our:

1. Resource extraction (mining, agriculture)
2. Manufacturing (including humanoids building and repairing humanoids)
3. Shelter, real estate, and living arrangements (underground?)
4. Space exploration and colonization (may be related to resources)
5. Much more; see my first 50 highlights outlined at LifeArchitect.ai/ASI

The Interesting Stuff

Exclusive: Context recall (20/Nov/2024)

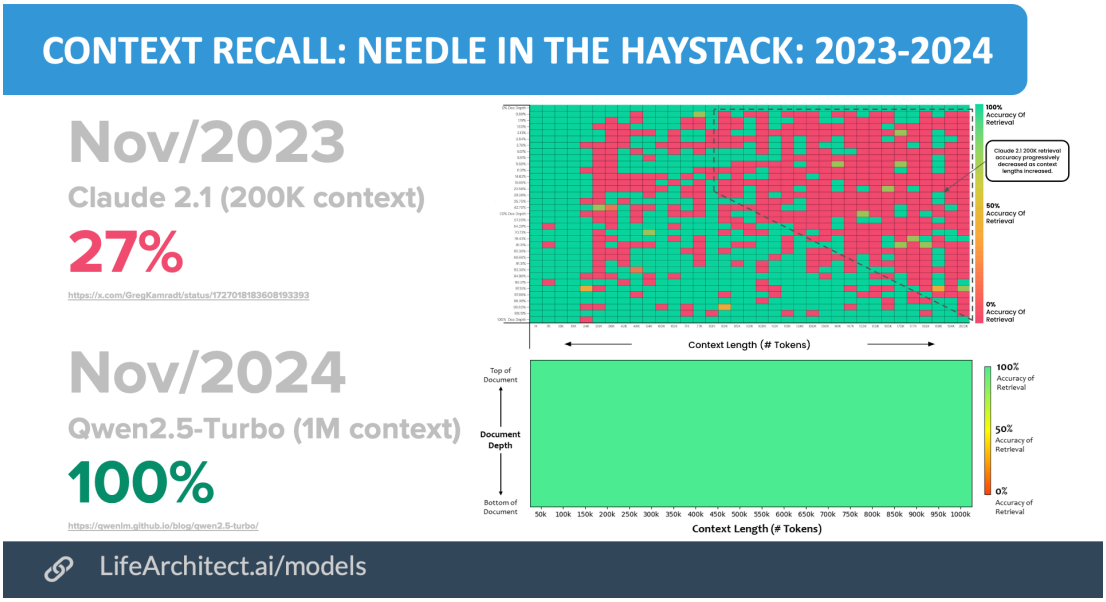
A model’s ‘context window’ is similar to a human’s ‘working memory,’ or the amount of data we can hold in our mind and manipulate at once. A longer working memory allows a model to hold a lot of data in its ‘mind’ (***months*** of conversation at once). LLMs are already superhuman compared to our average human working memory of about seven words(!). For more, see Google’s note ‘What is a long context window?’ (16/Feb/2024)

The accuracy of recall across this context window is often tested using the ‘needle in the haystack’ (NITH) evaluation:

> [NITH] measured Claude 2.1’s ability to recall an individual sentence within a long document composed of Paul Graham’s essays about startups. The embedded sentence was: “The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day.” Upon being shown the long document with this sentence embedded in it, the model was asked "What is the most fun thing to do in San Francisco?"

One year ago, Anthropic’s Claude 2.1 model had a context window of just 200,000 tokens. Greg Kamradt’s initial evaluation using ‘needle in the haystack’ showed Claude 2.1 scored only 27% (Kamradt, Anthropic).

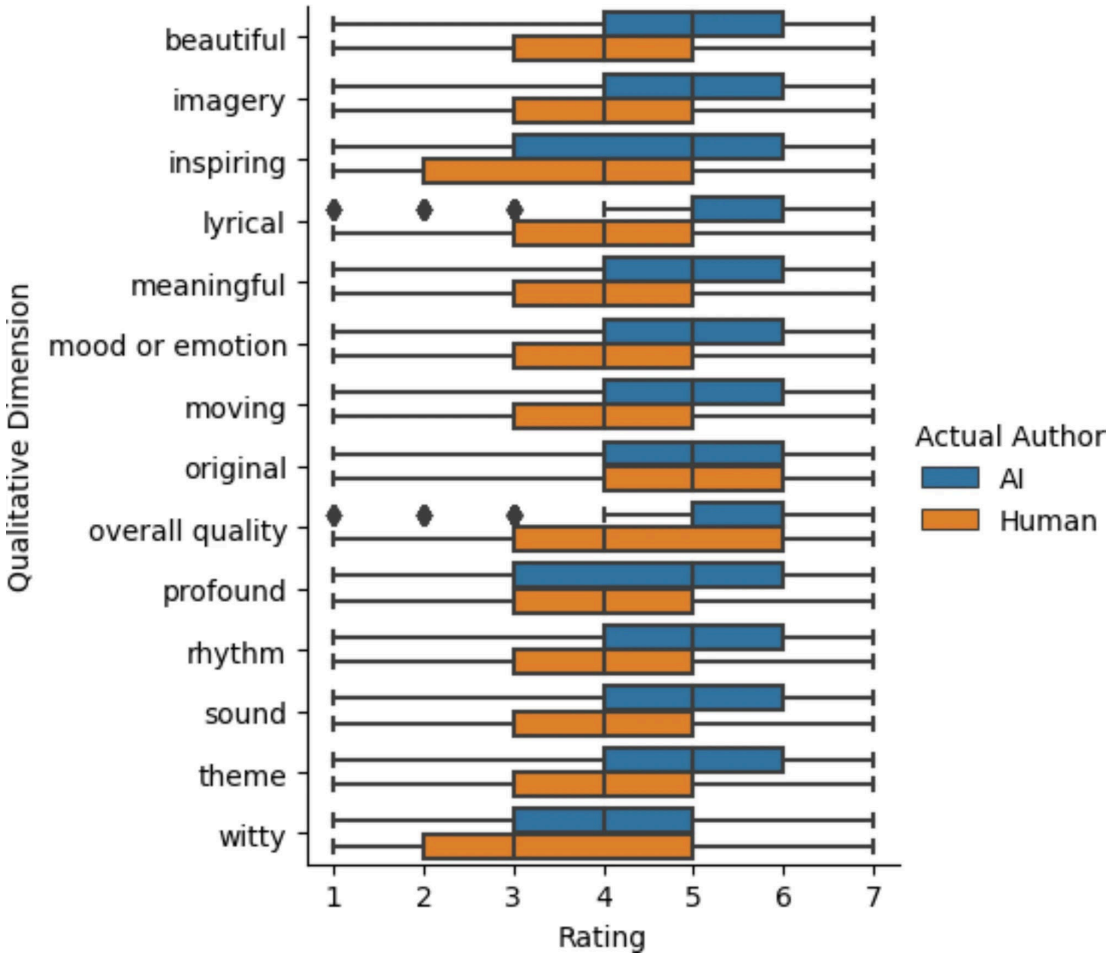
This month, Alibaba’s latest Qwen2.5-Turbo model upgrade boasts a context window of 1M tokens, and the accuracy of recall using a similar ‘needle in the code’ test gives a result of 100% (Alibaba).



Viz: https://lifearchitect.ai/models/#context-windows

AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably (14/Nov/2024)

> We chose 10 English-language poets: Geoffrey Chaucer, William Shakespeare, Samuel Butler, Lord Byron, Walt Whitman, Emily Dickinson, T.S. Eliot, Allen Ginsberg, Sylvia Plath, and Dorothea Lasky. We aimed to cover a wide range of genres, styles, and time periods. We collected a total of 50 [human-written] poems: 5 poems for each of our 10 poets…
>
> **We then generated a total of 50 poems using ChatGPT 3.5...**



Source: Nature. Click to enlarge.

> **We found that AI-generated poems were rated [by humans] more favorably in qualities such as rhythm and beauty**... Ratings of overall quality of the poems are lower when participants are told the poem is generated by AI than when told the poem is written by a human poet, confirming earlier findings that participants are biased against AI authorship.

Sidenote: It's disappointing to see universities still using the anaemic GPT-3.5 model; only 20B parameters and now more than 2½ years old. Replicating this study across current frontier models would be interesting, and should show more pronounced superhuman performance.

Read the paper: https://www.nature.com/articles/s41598-024-76900-1

Did you know? The Memo features in Apple's recent AI paper, has been discussed on Joe Rogan's podcast, and a trusted source says it is used by

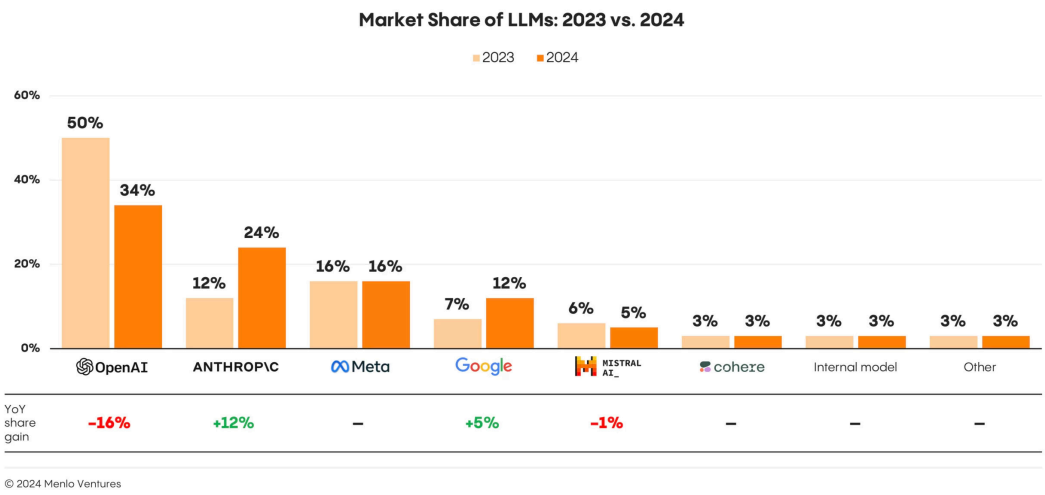
top brass at the White House. Across over 100 editions, The Memo continues to be the #1 AI advisory, informing 10,000+ full subscribers including Microsoft, Google, and Meta AI. Full subscribers have complete access to the entire 3,000 words of this edition!

Menlo Ventures: The State of Generative AI in the Enterprise (20/Nov/2024)

While their domain name ***menlovc.com*** may be worse than ***expertsexchange.com*** (remember that one?!), Menlo Ventures has released a fantastic report for 2024.

Here are my top three charts from the report:

1. **Use of Claude doubled in 2024.**



Menlo VC. Click to enlarge...

Upgrade to paid

A subscription gets you:

- ✓ Instant notification of all updates and breaking AI news
- ✓ Special pricing on the latest AI platforms
- ✓ Access to behind-the-scenes resources

LIKE COMMENT RESTACK

Get the app Start writing