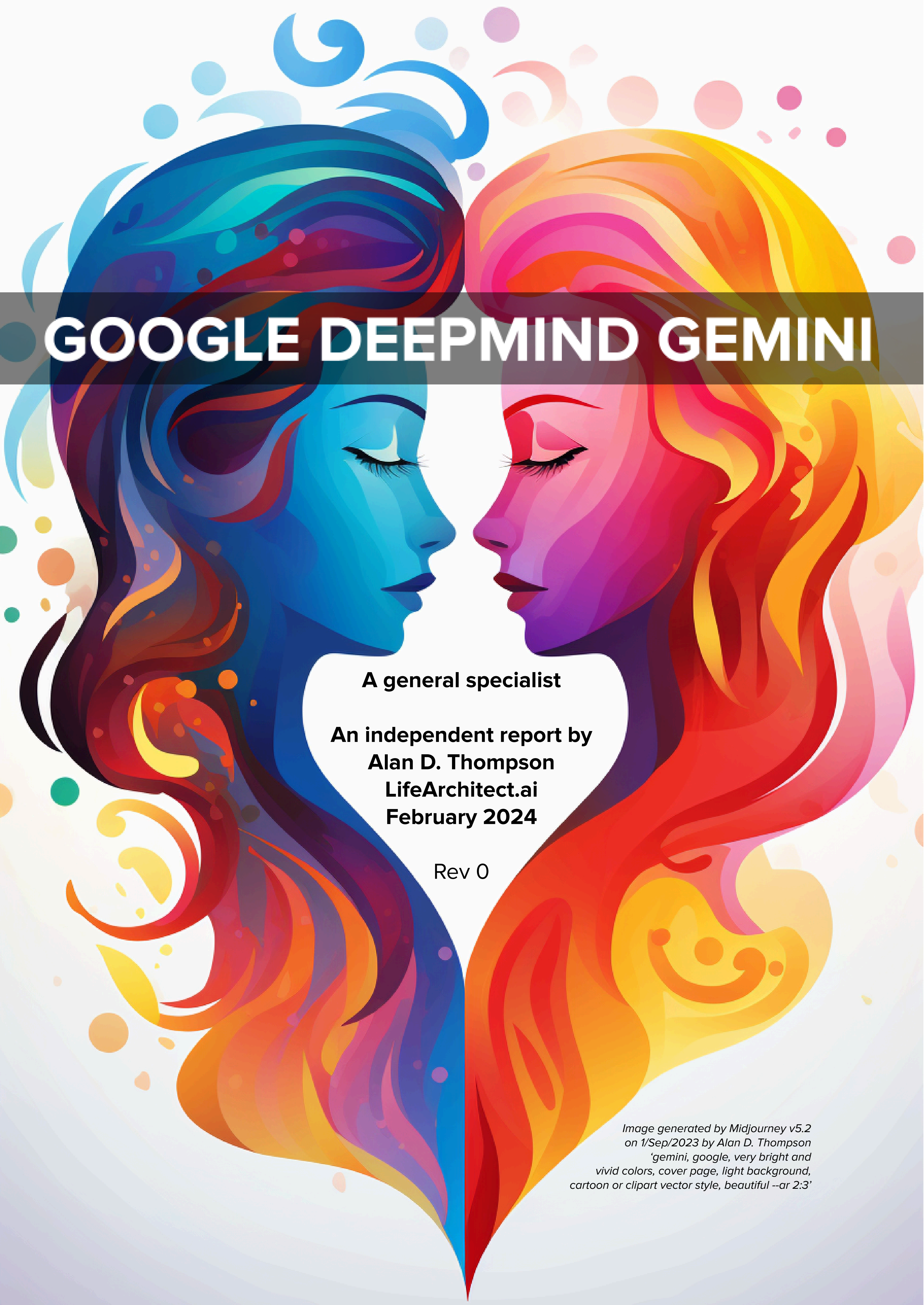


GOOGLE DEEPMIND GEMINI

A general specialist

An independent report by
Alan D. Thompson
LifeArchitect.ai
February 2024

Rev 0

Image generated by Midjourney v5.2
on 1/Sep/2023 by Alan D. Thompson
'gemini, google, very bright and
vivid colors, cover page, light background,
cartoon or clipart vector style, beautiful --ar 2:3'

Notice: A pre-release edition of this independent report (Rev A) was made available in Sep/2023, before the release of Gemini. Following the release of the complete Gemini model family, this Feb/2024 report is the final edition (Rev 0).

Contents

About the author

Dr Alan D. Thompson is an AI expert and consultant. With Leta (an AI powered by GPT-3), Alan co-presented a seminar called 'The new irrelevance of intelligence' at the World Gifted Conference in August 2021. His applied AI research and visualizations are featured across major international media, including citations in the University of Oxford's debate on AI Ethics in December 2021. He has held positions as chairman for Mensa International (gifted families), consultant to GE and Warner Bros, and memberships with the IEEE and IET.

To cite this report:
Thompson, A. D. (2024). *Google DeepMind Gemini: A general specialist (final edition).* https://LifeArchitect.ai/gemini

Abstract

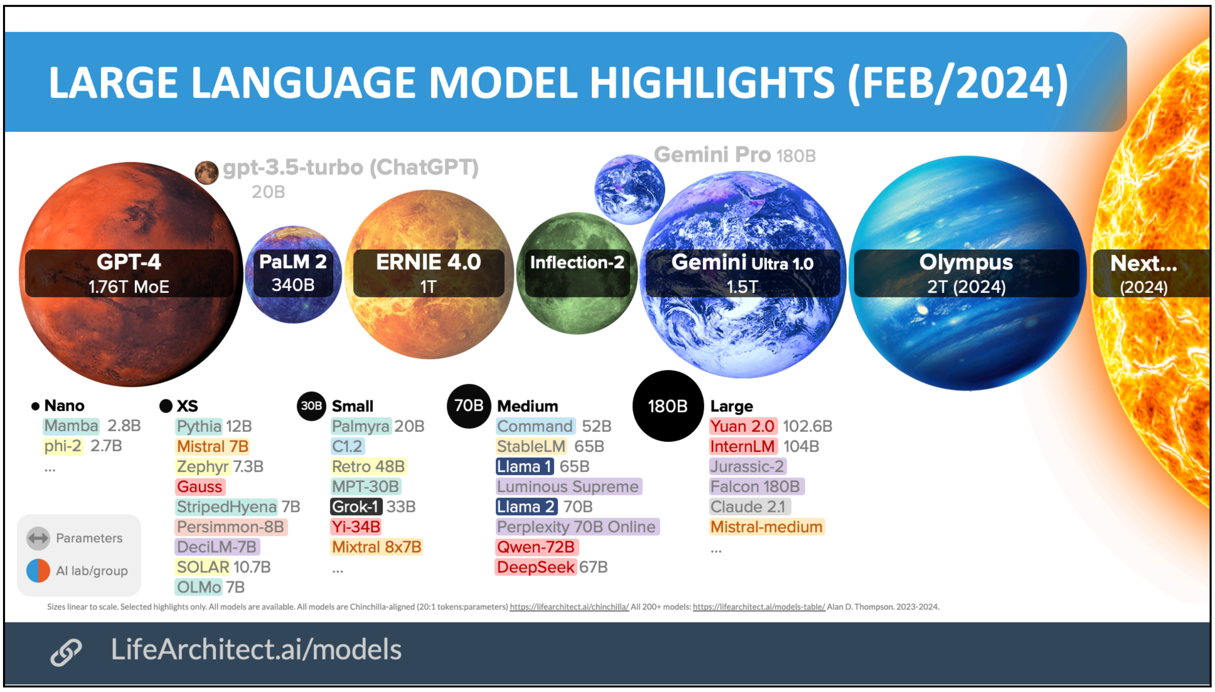
Since Google's discovery of the Transformer architecture in 2017, and successive release of their pre-trained transformer language model BERT in October 2018, training large language models (LLMs) has become a new space race, bringing humanity towards its largest evolutionary change yet: 'superintelligence.'

Between 2020 and 2024, LLMs continued to be trained on increasingly larger datasets, by ever larger teams of data scientists, with compute now measured in the hundreds of millions of dollars. The information synthesized here covers the progress made by Google and DeepMind, presenting as one company under the Alphabet umbrella in 2023, with a focus on the massive Gemini multimodal model.

Gemini Nano and Pro were released on 6/Dec/2023, and Gemini Ultra 1.0 was released on 7/Feb/2024. Gemini Ultra 1.0 is likely to be a dense model of around 1.5 trillion parameters trained on 30 trillion tokens. Compared to the GPT-4 sparse MoE model, Gemini Ultra 1.0 has a similar parameter count while being trained on 2× more data.

Model name	Parameters	Notes
Gemini Nano-1	**1.8B**	Targets lower memory devices like smartwatches.
Gemini Nano-2	**3.25B**	Targets higher memory devices like smartphones.
Gemini Pro 1.0	***180B***	ChatGPT (20B) competitor model.
Gemini Ultra 1.0	***1500B***	GPT-4 (1.76T) competitor model.

Table. Google DeepMind Gemini family model sizes. *Estimates only for Pro & Ultra.



Sizes linear to scale. Selected highlights only. All models are available. All models are Chinchilla-aligned (20:1 tokens:parameters) https://lifearchitect.ai/chinchilla/ All 200+ models: https://lifearchitect.ai/models-table/ Alan D. Thompson. 2023-2024.

Chart. Google DeepMind Gemini and other LLMs (Feb/2024).

1. Background

1.1. Etymology

The word ‘gemini’ is borrowed from the Latin geminī (‘twins’). In astronomy, Gemini is a constellation of the zodiac supposedly shaped like a pair of twins. In Greek mythology, Gemini is associated with the myth of the brothers Castor and Pollux (son of Zeus). When the human Castor died, because he was a mortal, Pollux begged his father Zeus to give Castor immortality, which was done through uniting them together in the heavens. The Gemini constellation contains these stars Castor and Pollux. It is one of the few constellations that actually looks like its namesake.

In Chinese astronomy,[1] the stars that correspond to Gemini are seen as Yin and Yang:

Castor (Yin): intellectual, aiding success in study.
Pollux (Yang): strength and ferocity.

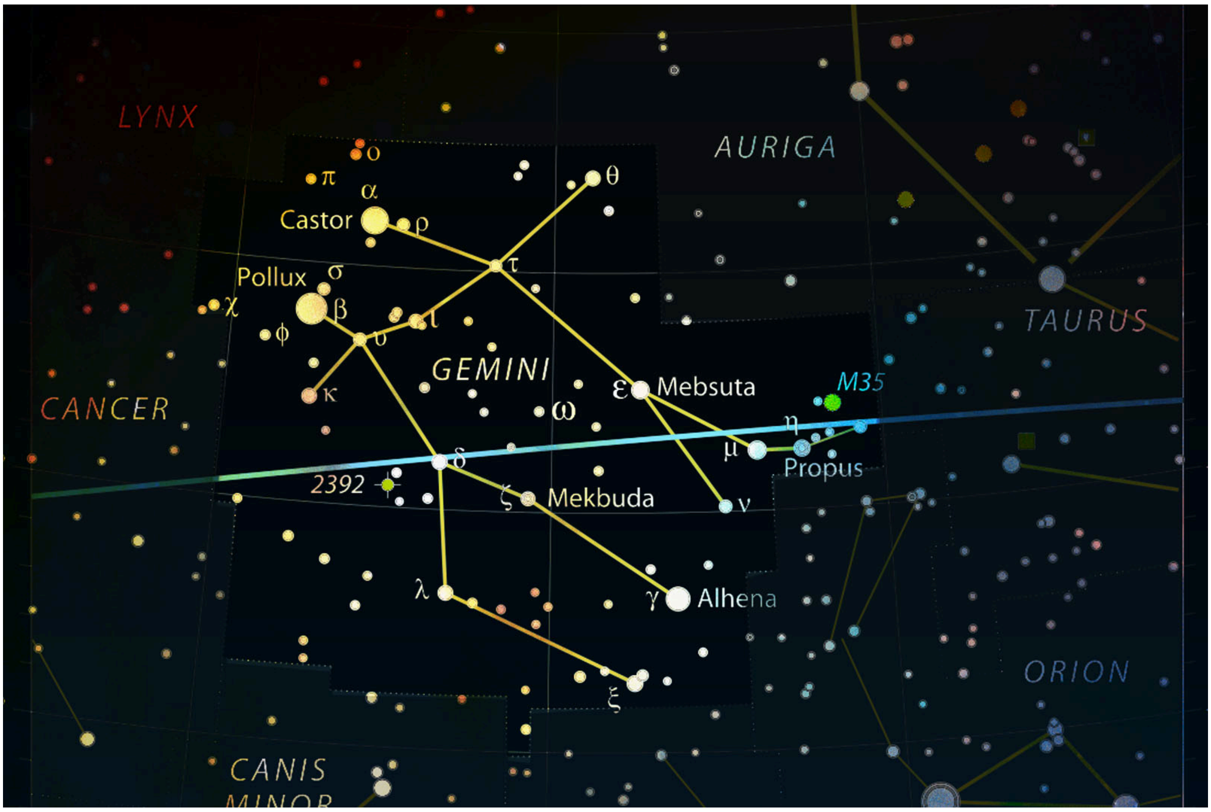


Figure. Map of the Gemini constellation.[2]

1.2. Google DeepMind: Two archers with one target

Google purchased DeepMind in 2014 for a reported US$600 million. During the explosion of AI model releases from 2018–2023, both Google and its subsidiary DeepMind worked independently.

[1] https://brickthology.com/2013/06/11/gemini/
[2] https://littleastronomy.com/gemini-constellation-for-kids/

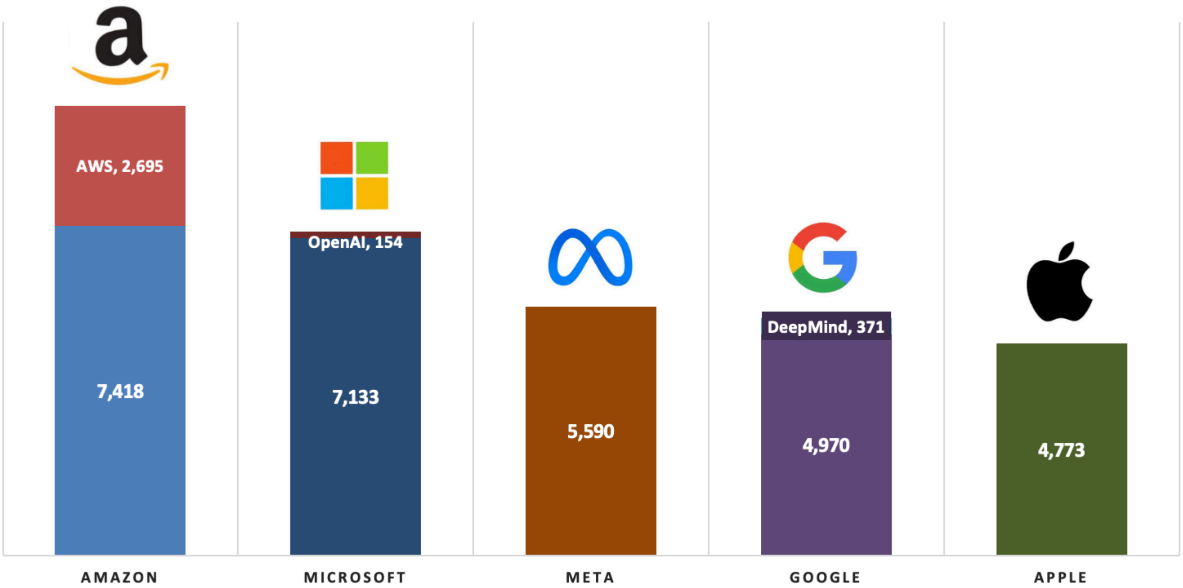


Chart. AI lab staff count. Source: glass.ai (Mar/2023).[3]

While Google AI employed around 5,000 staff in Q1 2023, DeepMind's focused engineers generated the same kind of output with a fraction of Google's headcount (only around 370 staff), though more than double the staff at OpenAI (150 staff).

The two Alphabet groups also differed in their mission statements. Google AI's objective is simple: 'Advancing AI for everyone.' DeepMind has a more nuanced statement: 'Our teams research and build safe AI systems. We're committed to solving intelligence, to advance science and benefit humanity.'

Some animosity between the two research groups was reported[4] by media outlets:

- Google developers were displeased that DeepMind didn't generate much revenue for the company.
- Google developers were resentful that DeepMind had special status within Alphabet that gave it free reign to work on projects.
- Google enjoys the public spotlight, while DeepMind is a much more private company, objecting to their products being labeled as 'powered by DeepMind'.
- Both parties had difficulty working together across regions: DeepMind primarily based in London and Google primarily in California, a time difference of 7–8 hours, and a distance of 8,500 km. (OpenAI faces the same challenge, opening their new London office[5] in 2023.)

California, US	1AM	3AM	5AM	7AM	**9AM**	**11AM**	**1PM**	**3PM**	**5PM**	7PM	9PM	11PM
London, UK	**9AM**	**11AM**	**1PM**	**3PM**	**5PM**	7PM	9PM	11PM	1AM	3AM	5AM	7AM

Table. California and London time zones usually have no overlap of working hours.

[3] https://www.glass.ai/glass-news/code-red-the-ai-armies-of-the-tech-giants
[4] The Information via MobileSyrup, 20/Apr/2018: https://mobilesyrup.com/2018/04/20/google-deepmind-not-getting-along/
[5] https://openai.com/blog/introducing-openai-london

1.3. Gemini personnel

The Gemini paper lists around 625 contributors involved in the development of Gemini. In the GPT-4 report, OpenAI listed 279 unique staff and external consultants involved in the development of that model.[6] Gemini team leads from DeepMind were Drs Demis Hassabis, Oriol Vinyals, and Koray Kavukcuoglu. Dr Jeff Dean assisted from Google AI.

1.4. Gemini compute resources

Gemini Ultra 1.0 was trained with Google's own TPU v4 chips (smaller models in the family used the newer TPU v5e), with many of the TPU hardware circuits designed by AI.[7] Gemini Ultra is trained on significantly more compute than GPT-4. The compute budget for Gemini is estimated to be well into the hundreds of millions of dollars. As a rough guide, TPUv4 @ $3.22/hr for 15,616 years or 136.8M hours ≈ US$440.5M using 2023 retail pricing.

Model	Training end	Chip type	TFLOP/s (max)	Chip count	Wall clock (days)	Total time (years)	Cost (US$)	MMLU ▲
GPT-3	Apr/2020	V100	130	10,000	15 days	405 years	$9M	43.9
Llama 1	Jan/2023	A100	312	2,048	21 days	118 years	$4M	63.4
Llama 2	Jun/2023	A100	312	*2,048*	35 days	196 years	$7M	68.0
GPT-4	Aug/2022	A100	312	25,000	95 days	6,507 years	$224M	86.4
Gemini	Nov/2023	TPUv4	275	57,000	100 days	15,616 years	$440M	90.0
GPT-5	*Apr/2024*	H100	989	*50,000*	*120 days*	*16,438 years*	*$612M*	
Llama 3	*Apr/2024*	H100	989					
Olympus	*Aug/2024*	H100	989					
Gemini 2	*Nov/2024*	TPUv5	393		*Alan D. Thompson. Feb/2024. LifeArchitect.ai*			

Table. Google DeepMind Gemini training compute (see working, with sources[8]).

1.5. Large language models

The silo approach adopted by Google AI and DeepMind seemed to have benefits, spurring advances by both research groups. Perhaps with the exception of OpenAI's successful commercialization of LLMs, Google continued to lead the field in AI model innovation. DeepMind also explored the periphery of AI models, notably discovering that LLMs should be trained with an order of magnitude more compute and data than was being used during the GPT-3 era. Both labs trained many large language models during the early 2020s.

[6] https://arxiv.org/abs/2303.08774

[7] https://www.nature.com/articles/s41586-021-03544-w

[8] Working, with sources:
https://docs.google.com/spreadsheets/d/1O5KVQW1Hx5ZAkcg8AIRjbQLQzx2wVaLl0SqUu-ir9Fs/edit#gid=1381013921

Lab	Model	Parameters (+ tokens trained)	Date	Notes
Google	**BERT**	0.3B (137B)	Oct/2018	First LLM (with GPT-1)
Google	**T5**	11B (34B)	Oct/2019	
Google	**Meena**	0.35B (10T)	Jan/2020	Chatbot
Google	**Switch**	1T (576B)	Jan/2021	MoE
Google	**LaMDA**	137B	Jun/2021	Chatbot
Google	**FLAN**	137B	Sep/2021	Fine-tuned LaMDA
DeepMind	**RETRO**	7.5B	Dec/2021	Retrieval-augmented
Google	**GLaM**	1.2T	Dec/2021	MoE
DeepMind	**Gopher**	280B (300B)	Dec/2021	
DeepMind	**Chinchilla**	70B (1.4T)	Mar/2022	More compute + data
Google	**PaLM 1**	540B (780B)	Apr/2022	
Google	**LaMDA 2**		May/2022	
DeepMind	**Gato**	1.2B	May/2022	Proto-AGI (VLM + robot)
Google	**UL2**	20B (1T)	May/2022	
Google	**LIMoE**	5.6B	Jun/2022	MoE
DeepMind	**Perceiver AR**	1B	Jun/2022	Context=100k
Google	**Minerva**	540B (818.5B)	Jun/2022	Maths

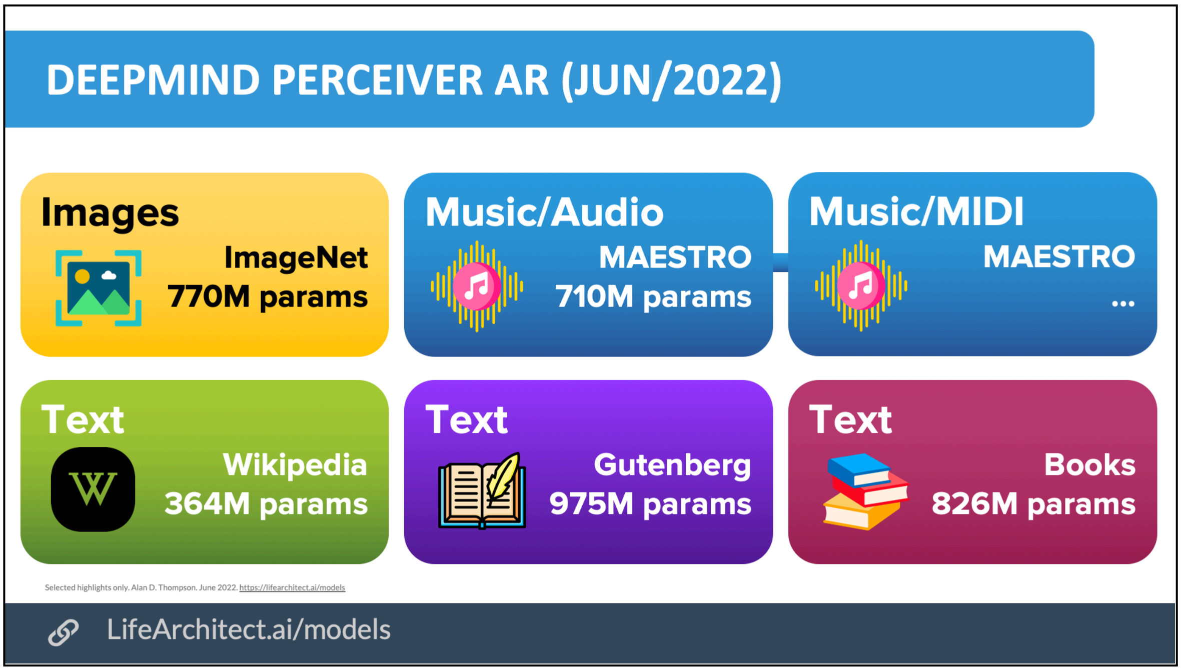
Google	**U-series**
	Flan-series
	Flan-U-series
		Multiple	Oct/2022	U=Compute focus
				Flan=Instruct focus

Google	**PaLM 2**	340B (3.6T)	May/2023	
Google	**DIDACT**	(potentially 37.9T)	Jun/2023	Massive codebase

Table. Google and DeepMind LLMs 2018–2023. Highlights only.

Google's models are well-known in the AI field, with many Google models (BERT, T5) open sourced and released to the public. At the other extreme, DeepMind kept a shroud around all of its models, providing detailed papers up until the end of 2022, but never releasing a major model outside of its own lab.

In the table above, DeepMind's 'modality-agnostic' Perceiver AR combined several modalities. It was also one of the first models to expand the context window (similar to human working memory) to 100,000 tokens or 75,000 words. The full list of trained modalities for Perceiver AR is shown below.



DEEPMIND PERCEIVER AR (JUN/2022)

Images ImageNet 770M params	Music/Audio MAESTRO 710M params	Music/MIDI MAESTRO …
Text Wikipedia 364M params	Text Gutenberg 975M params	Text Books 826M params

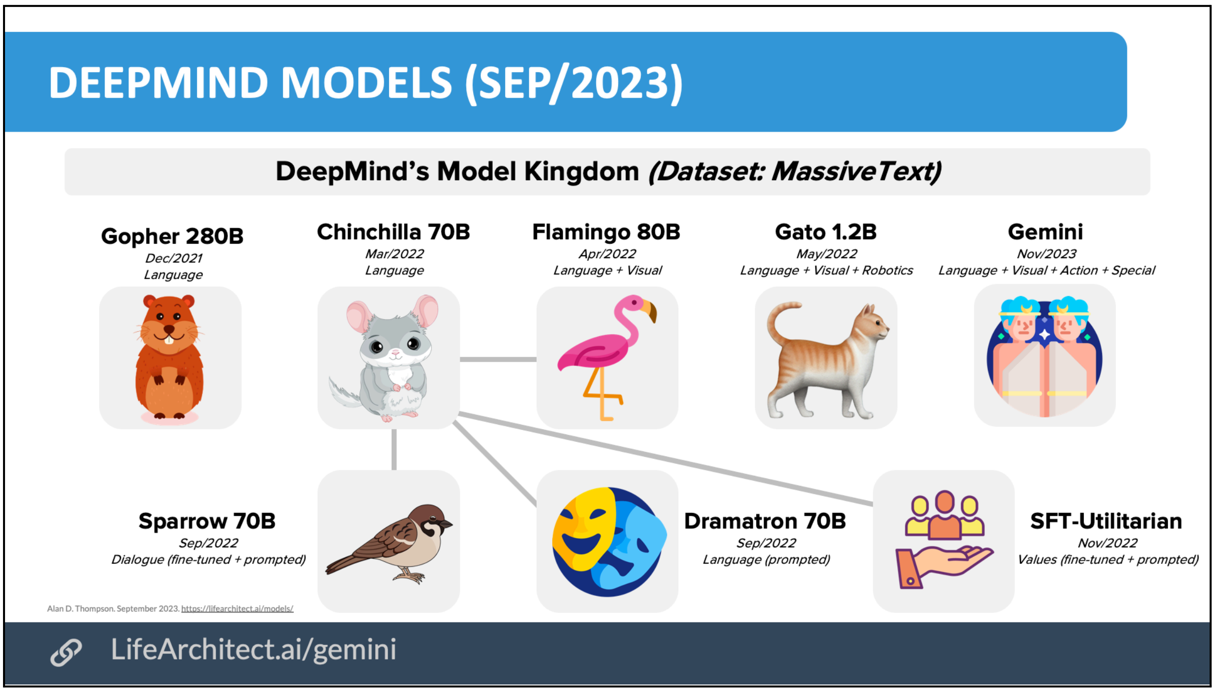
Selected highlights only. Alan D. Thompson. June 2022. https://lifearchitect.ai/models

LifeArchitect.ai/models

Chart. DeepMind Perceiver AR combined several modalities (Jun/2022).

Several of DeepMind's private models in 2022 were offshoots of its large language model, Chinchilla 70B trained on 1.4T tokens. Subsequently, most AI labs used the architecture and training setup of Chinchilla as a starting point. For example, Meta AI's Llama 1 65B also trained on 1.4T tokens, while Llama 2 70B was trained on 2T tokens.

A more comprehensive view of DeepMind's models is shown below. Note the focus on alignment in the bottom row: Sparrow, Dramatron, and SFT-Utilitarian explored alignment through prompting, fine-tuning, or a combination of the two.



DEEPMIND MODELS (SEP/2023)

DeepMind's Model Kingdom *(Dataset: MassiveText)*

Alan D. Thompson. September 2023. https://lifearchitect.ai/models/

LifeArchitect.ai/gemini

Chart. DeepMind model highlights 2021–2023.

1.6. Text-to-image and visual language models

Additionally, both research groups worked independently on their own text-to-image models (text in ➜ image out) and visual language models (image in ➜ text out).

Lab	Model	Dataset size	Date	Notes
DeepMind	**Flamingo**	185M+	Apr/2022	VLM
Google	**Imagen**	860M	May/2022	Diffusion
Google	**Parti**	4.8B	Jun/2022	Transformer/VQGAN
Google	**DreamBooth**	860M+	Aug/2022	Fine-tuned Imagen
Google	**UniTune**	860M+	Oct/2022	Fine-tuned Imagen
Google	**Muse**	460M+	Jan/2023	Imagen dataset
Google	**Soft MoE**	4B+	Aug/2023	ViT. New MoE setup

Table. Google and DeepMind text-to-image and VLMs 2022–2023. Highlights only.

1.7. The Alpha series of AI systems

While large language models and text-to-image models remain the two most popular model types in 2024, DeepMind has been exploring more specific AI systems since 2015. With single-focus expertise ranging from chess to maths, these systems do one thing and one thing only, and they do it very, very well.

#	System	Expertise	Date	Description
1	**AlphaGo**	Go	Oct/2015	First AI in the series. Designed to play the board game Go; a huge number of possible board configurations. Mar/2016: beat world champion Lee Sedol.
2	**AlphaGo Zero**	Go	Oct/2017	Improved version of AlphaGo, learned to play from scratch, no prior knowledge beyond the game's rules.
3	**AlphaZero**	Board games	Dec/2017	Further generalization of AlphaGo Zero's approach to play other board games, chess and shogi. Superhuman performance.
4	**AlphaFold 1**	Protein folding	Dec/2018	Change in focus from games to scientific problems. Designed to predict protein folding structures, a complex problem in biology.
5	**AlphaStar**	StarCraft II	Jan/2019	Designed to play the real-time strategy game StarCraft II. First AI to reach a professional (Grandmaster) level.
6	**AlphaFold 2**	Protein folding	Nov/2020	New version of AlphaFold, later open sourced. Allows users to predict 3D structure of arbitrary proteins with exceptional accuracy.
7	**AlphaCode**	Software code	Feb/2022	A coding engine that creates computer programs at a rate comparable to that of an average programmer.
8	**AlphaTensor**	Maths	Oct/2022	The first AI system for discovering novel and provably correct algorithms for fundamental tasks such as matrix multiplication.
9	**AlphaDev**	Algorithms	Jun/2023	A system to discover enhanced computer science algorithms. Uses AlphaZero approach to find faster algorithms for tasks such as sorting and hashing.
10	**AlphaCode 2**	Software code	Dec/2023	Gemini-powered agent combining Gemini's reasoning capabilities with search and tool-use for solving competitive programming problems.
11	**AlphaGeometry**	Geometry	Jan/2024	An AI system that solves complex geometry problems at a level approaching a human Olympiad gold-medalist.

Table. DeepMind Alpha systems 2015–2024. Highlights only. Assisted by GPT-4.[9]

[9] I still write my reports by hand as of Q1 2024, but I used OpenAI GPT-4 (via Poe.com) in this section to summarize the DeepMind Alpha series systems, distilling the name and function of each system, date of release, and putting it into a table. 2022 systems outside of GPT-4's training data were added by hand, and the final table was heavily edited.

In an interview with *The New York Times*[10] in July 2023, DeepMind CEO Demis Hassabis commented on how these specific systems may be integrated with a more general AI model:

> *We're working on our own system called Gemini… There's going to be a combination of the two things [general and specialized]. So we'll have this increasingly more powerful general system that you basically interact with through language but has other capabilities, general capabilities, like math and coding, and perhaps some reasoning and planning, eventually, in the next generations of these systems… There should be specialized AI systems that learn how to do those things — AlphaGo, AlphaZero, AlphaFold… the general system can call those specialized AIs as tools.'*

In June 2023, for *Wired,*[11] Hassabis also noted:

> *At a high level you can think of Gemini as combining some of the strengths of AlphaGo-type systems with the amazing language capabilities of the large models. We also have some new innovations that are going to be pretty interesting. I can see the kinds of things we're building into the Gemini series right, and we have no reason to believe that they won't work.*

1.8. Putting it together: LLM + VLM + Text-to-image

If we identify the current 'best in class' of each model type released by Google and DeepMind, we get something like this:

A. LLM (sparse MoE): 1T parameter (based on Google Switch).
B. Text-to-image: Google Imagen 2.
C. Visual language model: 10B parameters (based on DeepMind Flamingo).
D. Datasets: The largest possible dataset (based on DeepMind MassiveText or Google's PaLM 2 dataset).

Let's explore this last point in finer detail.

2. Datasets

Very large datasets are a major part of the AI space race in 2024, with some open-source datasets hitting 5 trillion text tokens in June 2023[12] (a word is around 1.3 tokens; we can quickly derive the count of words from tokens by multiplying tokens by 0.75). DeepMind hit the same 5T token count nearly 18 months earlier in February 2022,[13] and with Google's support this should be easily extendable. However, as we'll see, text is not the only possible training data.

[10] https://www.nytimes.com/2023/07/11/podcasts/transcript-ezra-klein-interviews-demis-hassabis.html
[11] https://archive.md/XFwF6
[12] https://arxiv.org/abs/2306.01116
[13] https://arxiv.org/abs/2112.04426

2.1. Datasets: Text: MassiveText multilingual

The contents of DeepMind's MassiveText English dataset via the Gopher 280B model was analyzed in the paper *What's in my AI?* (Thompson 2022) at https://lifearchitect.ai/whats-in-my-ai/.

The multilingual version of this dataset (used for Retro) is twice as big as the English version used to train Gopher and Chinchilla, though there are some inconsistencies in sizing between the two datasets (for example, MassiveText for Gopher has 676B tokens of News, while the larger MassiveText for Retro shows only 237B tokens for News).

Count	Dataset	Language	Raw Size	Tokens
1	**Books (20.47M)**	English	*12.8TB*	**3.42T**
2	**MassiveWeb**	Multilingual	*3.8TB*	**978B**
3	**Github**	English	*2.7TB*	**375B**
4	**News**	English	*0.9TB*	**237B**
5	**Wikipedia**	Multilingual	*54GB*	**13B**
	Total		*20.3TB*	**5.026T**

Table. MassiveText Multilingual Datasets used for Retro. Disclosed in **bold**. Determined in *italics*.

In January 2023, Google announced[14] that it had digitized 'more than 40 million books in more than 500 languages,' a decent percentage of the 130 million unique books in the world, as estimated by Google Books software engineer Leonid Taycher in 2010.[15] Of interest, the Gopher paper disclosed that its MassiveText English Books dataset contained some books that are more than 500 years old (1500–2008).

In August 2023, it was reported[16] that Google DeepMind legal counsel were meticulously reviewing Gemini's training process. On one occasion, they instructed researchers to eliminate some training data that was sourced from textbooks. This data could have assisted the AI model in responding to queries about topics such as astronomy or biology. However, in 2023, there has been increased attention (and legal action) from copyright owners regarding intellectual property. For further detail on the top 20 domains used in DeepMind MassiveWeb (part of MassiveText English), see the Appendix of this report.

2.2. Datasets: Visual (images and video)

The twins of Gemini are 'looking' at both text and visual data during training.

[14] https://blog.google/products/search/google-books-library-project/

[15] http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html

[16] https://www.theinformation.com/articles/the-forced-marriage-at-the-heart-of-googles-ai-race

If we include Google's YouTube video platform, we get an unimaginably large amount of data on which to train the model. In Q1 2023, YouTube totalled around 800 million videos, each with an average length of 11.7 minutes.[17]

If we were to divide this into frames, we would have to convert 9.36B total minutes into seconds: 561.6B × 25 frames per second = 14T images. These images are generally labeled thanks to subtitles (my estimate[18] is 1.56T text tokens), easily giving us one of the largest visual datasets in the world. It would be comparable with Tesla's cameras collecting visual data from its 4.5M vehicles[19] in Q2 2023. Given both Google and DeepMind's exploration of robotics and AI embodiment, it's likely that they are using the actual moving video footage as well as still images for training Gemini.

2.3. Datasets: Audio

When DeepMind announced its generalist agent, Gato (Spanish for 'cat') in May 2022, the unexpected innovation based on the Transformer architecture opened a new world for AI labs. Besides standard tokenization of text, DeepMind listed several other byte streams being flattened and used during training, as well as during the usual input (prompt), and output (response):

1. Button presses for the Atari gaming console.
2. Proprioception (robotics).
3. Joint torques (robotics).

Gemini revealed a previously unseen mixture of capabilities based on the text and image training datasets, while also including additional special training datasets based on audio processing (16kHz audio signals from Universal Speech Model or USM).

3. Gemini capabilities and performance

Gemini was first announced[20] by Google's CEO in May 2023:

> *We're already at work on Gemini — our next model created from the ground up to be multimodal, highly efficient at tool and API integrations, and built to enable future innovations, like memory and planning. Gemini is still in training, but it's already exhibiting multimodal capabilities never before seen in prior models. Once fine-tuned and rigorously tested for safety, Gemini will be available at various sizes and capabilities, just like PaLM 2, to ensure it can be deployed across different products, applications, and devices for everyone's benefit.*

3.1. Languages

Gemini is expected to closely follow previous offerings from DeepMind and Google, with coverage of hundreds of languages. The top 10 languages in DeepMind

[17] https://www.wyzowl.com/youtube-stats/
[18] https://lifearchitect.ai/gemini/#youtube
[19] https://www.licarco.com/news/how-many-tesla-cars-have-been-sold
[20] https://blog.google/technology/ai/google-palm-2-ai-large-language-model/

MassiveWeb are English, Russian, Spanish, Chinese, French, German, Portuguese, Italian, Swahili, and Urdu. The top 10 languages in Google PaLM 2 are English, Spanish, Chinese, Russian, Japanese, French, Portuguese, German, Italian, and Korean.

3.2. Visual

Gemini will provide built-in text-to-image capabilities. Two state-of-the-art model examples from Google (Imagen, pronounced 'imagine', and Parti) are shown here with their corresponding text prompts. Each image is 'conceptualized' from scratch by the model, based on the text prompt input by the user.

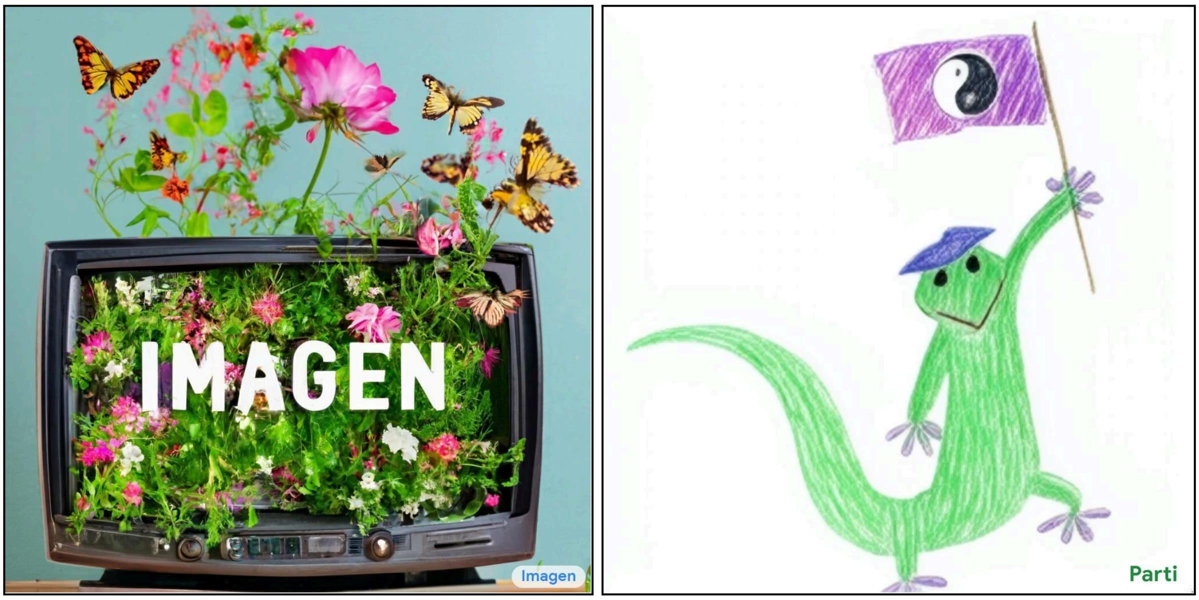
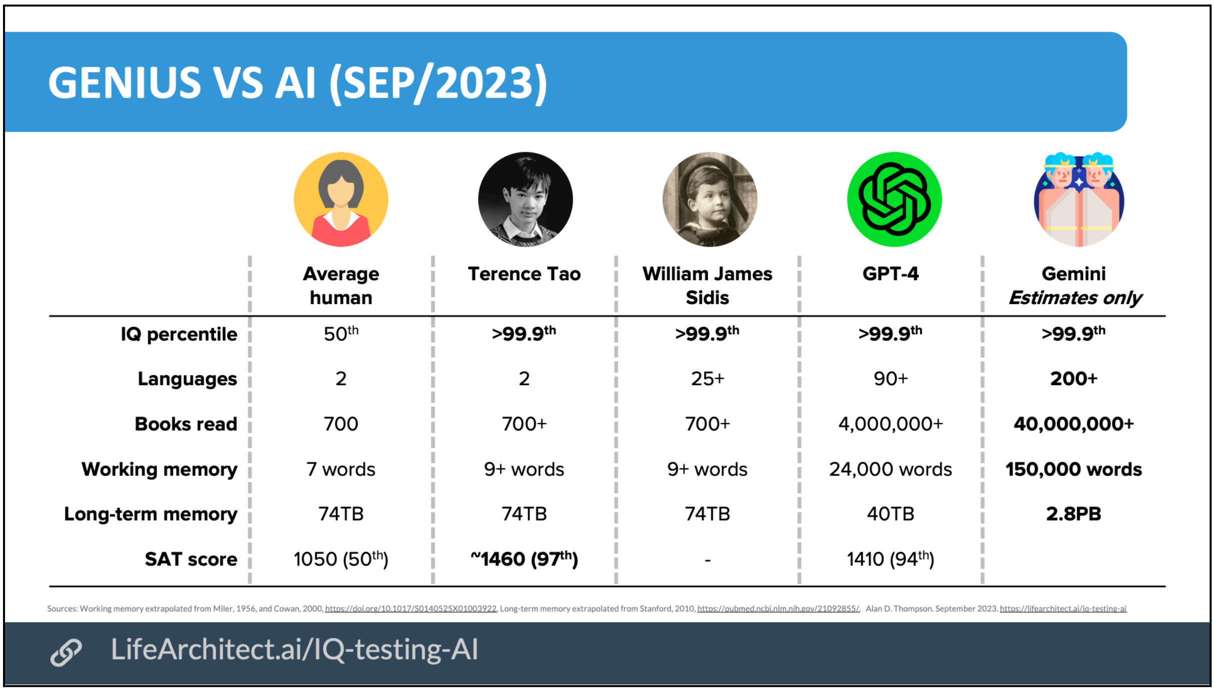


Image: Google Imagen (left, May/2022) 'Vines in the shape of text 'Imagen' with flowers and butterflies bursting out of an old TV.' **Google Parti** (right, Jun/2022) 'Portrait of a gecko wearing a train conductor's hat and holding a flag that has a yin-yang symbol on it. Child's crayon drawing'

3.3. IQ



GENIUS VS AI (SEP/2023)

	Average human	Terence Tao	William James Sidis	GPT-4	Gemini *Estimates only*
IQ percentile	50th	**>99.9th**	**>99.9th**	**>99.9th**	**>99.9th**
Languages	2	2	25+	90+	**200+**
Books read	700	700+	700+	4,000,000+	**40,000,000+**
Working memory	7 words	9+ words	9+ words	24,000 words	**150,000 words**
Long-term memory	74TB	74TB	74TB	40TB	**2.8PB**
SAT score	1050 (50th)	**~1460 (97th)**	-	1410 (94th)	

Sources: Working memory extrapolated from Miler, 1956, and Cowan, 2000, https://doi.org/10.1017/S0140525X01003922. Long-term memory extrapolated from Stanford, 2010, https://pubmed.ncbi.nlm.nih.gov/21092855/. Alan D. Thompson. September 2023. https://lifearchitect.ai/iq-testing-ai

LifeArchitect.ai/IQ-testing-AI

Chart. Human vs GPT-4 vs Gemini. Estimates from Sep/2023.

Large language models are now approaching—and in some cases breaching—the ceilings of human-designed tests. Gemini Ultra 1.0 is a direct competitor of OpenAI's *GPT-4,* scoring similarly across a range of benchmarks.

	Human avg	GPT-4	Flan PaLM 2	Gemini Ultra
MMLU	34.5	86.4	81.2	**90.04**
WinoGrande	**94.0**	87.5	90.9	-
Theory of mind	87.0	**100.0**	-	-
SAT score	1050 (P50)	**1410 (P94)**	-	-

Table. Human vs GPT-4 vs PaLM 2 vs Gemini, selected benchmarks. Highest in **bold.**

4. Size comparison

Based on previous model iterations, it seems prudent to use a lower-bounded assumption of each successive model being significantly larger and more powerful than the previous model, especially for dataset sizes post-Chinchilla. This exponential scaling of model size (measured in parameters) is illustrated most clearly in OpenAI's GPT releases:

GPT-1 117M ➜	GPT-2 1.5B ➜	GPT-3 175B ➜	GPT-4 1.76T ➜	GPT-5
13×	116×	10×		

Gemini demonstrates the largest general dataset mixture collected to date, based on text, code, image, video, audio, and other specialized components. 1.5T parameters trained on 30T tokens is a ratio of 20:1; well above GPT-4 (8:1) and PaLM 2 (11:1) estimates. A 1.5T-parameter dense model is larger and more performant than GPT-4's 1.76T-parameter sparse MoE estimate. While OpenAI's GPT-4 sparse model likely provides more total connections, standard dense models like Gemini allow interaction between all parameters, learn more efficiently, can generalize better from limited data, and are both simpler and more stable.

5. Implementing and applying Gemini

In May 2023, Google noted that 'Gemini will be available at various sizes and capabilities, [and will] be deployed across different products, applications, and devices.'[21] Like Google's recent PaLM 2 model, Gemini is available through several applications:

1. API via Vertex AI on Google Cloud[22]
2. Other private API (for example, in Poe.com)[23]
3. Chatbot via Google Bard at bard.google.com[24]
4. Google products including Gmail, Docs, Sheets, Slides, and more via Duet AI[25]

[21] https://blog.google/technology/ai/google-palm-2-ai-large-language-model/
[22] https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-applications-with-vertex-ai-palm-2-models-and-langchain
[23] https://twitter.com/poe_platform/status/1674470575056752641
[24] https://bard.google.com/
[25] https://blog.google/technology/ai/google-palm-2-ai-large-language-model/

5. Prediction: ‘Position Zero’ Featured Snippets[26] in Google Search (similar to Knowledge Graph, and Quora’s implementation of ChatGPT as a featured answer shown at the top of the page)[27]

This last item is particularly interesting. Integrating a large language model into Google Search would completely change the landscape of how users interact with search; shifting it from a simple index to an intelligent system and beyond.

6. Conclusion

Google DeepMind Gemini Ultra 1.0 is one of the most powerful AI models available in early 2024. Dovetailing between OpenAI’s releases of GPT-4 and GPT-5, it is a groundbreaking release. The model showcases the culmination of millions of person-hours of bleeding-edge AI research and development by two of the world’s premier AI labs. As an eagerly anticipated multimodal system combining language, vision, and other capabilities, Gemini pushes the boundaries of AI even further.

While concrete details on Gemini’s architecture may never be released based on ‘no publishing’ precedents by OpenAI and Google, as well as on comments made[28] by DeepMind’s CEO, this report synthesizes available information and provides educated projections regarding the model’s details, training data, and performance. By reviewing the lineage of Google and DeepMind’s past achievements in areas like large language models, visual systems, robotics, and specialized expert systems, we can begin to glimpse the potential shape of this AI model.

Gemini represents a massive leap forward from its predecessors in scale, breadth, and application. The prospect of 1.5T parameters trained on 30T text tokens plus additional multimodal data is staggering, and Google Gemini is already showing new emerging abilities far beyond any current system.

In particular, the largest model—Gemini Ultra 1.0—blazes new trails in humanity’s quest to develop increasingly capable and general artificial intelligence. Its unveiling is a watershed moment in the field.

[26] https://support.google.com/websearch/answer/9351707?hl=en

[27] https://support.google.com/knowledgepanel/answer/9787176?sjid=7519301398263360399-AP

[28] ‘[Hassabis] suggests that the AI industry’s culture of publishing its findings openly may soon need to end.’ https://time.com/6246119/demis-hassabis-deepmind-interview/

7. Further reading

For brevity and readability, footnotes were used in this article, rather than in-text/parenthetical citations. Related reports and resources are listed below, or please see http://lifearchitect.ai/papers/ for the major foundational papers in the large language model space.

Models
Thompson, A. D. (2021-2024). *Models table.* https://lifearchitect.ai/models-table/

Datasets
Thompson, A. D. (2022a). *What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher.* https://LifeArchitect.ai/whats-in-my-ai

Google Pathways
Thompson, A. D. (2022b). *Google Pathways: An Exploration of the Pathways Architecture from PaLM to Parti.* https://LifeArchitect.ai/pathways

–

Major papers from Google & DeepMind (in chronological order)

Google Transformer (Jun/2017)
https://arxiv.org/abs/1706.03762

Google Switch Transformers (Jan/2021)
https://arxiv.org/abs/2101.03961

DeepMind Gopher & MassiveText English (Dec/2021)
https://arxiv.org/abs/2112.11446

DeepMind Retro & MassiveText Multilingual (Dec/2021)
https://arxiv.org/abs/2112.04426

DeepMind Chinchilla (Mar/2022)
https://arxiv.org/abs/2203.15556

Google PaLM 1 (Apr/2022)
https://arxiv.org/abs/2204.02311

Google PaLM 2 (May/2023)
https://arxiv.org/abs/2305.10403

8. Appendix

Datasets: MassiveWeb English Analysis

The following table is from the 2022 paper: What's in My AI?:
https://lifearchitect.ai/whats-in-my-ai/

DeepMind was acquired by Google in 2014, and has access to enormous amounts of Alphabet-managed data in the creation of MassiveText and the new Gemini dataset.

While the subset of MassiveText, MassiveWeb, is not detailed much further in the Gopher paper, an Appendix on p44, Figure A3b notes the Top 20 domains appearing in MassiveWeb English.[29] Given the disclosed percentage represented for each domain, we can use the MassiveWeb total token count (506B tokens) and total Raw Size (1900GB) to determine the token count and size of each domain.

Count	Domain	Percentage tokens	Tokens (PT × 506B)	Size (PT × 1900GB)
1	**ScienceDirect**	**1.85%**	*9.4B*	*35.2GB*
2	**Gale**	**1.79%**	*9.1B*	*34.0GB*
3	**NCBI**	**1.59%**	*8.0B*	*30.2GB*
4	**Facebook**	**1.10%**	*5.6B*	*20.9GB*
5	**Issuu**	**0.98%**	*5.0B*	*18.6GB*
6	**Academia**	**0.93%**	*4.7B*	*17.7GB*
7	**Quora**	**0.75%**	*3.8B*	*14.3GB*
8	**Springer**	**0.73%**	*3.7B*	*13.9GB*
9	**YouTube**	**0.73%**	*3.7B*	*13.9GB*
10	**ProQuest Search**	**0.68%**	*3.4B*	*12.9GB*
11	**English Wikipedia**	**0.66%**	*3.3B*	*12.5GB*
12	**SlideShare**	**0.58%**	*2.9B*	*11.0GB*
13	**SlidePlayer**	**0.57%**	*2.9B*	*10.8GB*
14	**Reddit**	**0.51%**	*2.6B*	*9.7GB*
15	**Medium**	**0.42%**	*2.1B*	*8.0GB*
16	**Wiley Online Library**	**0.38%**	*1.9B*	*7.2GB*
17	**Europe PubMed Central**	**0.38%**	*1.9B*	*7.2GB*
18	**GitHub**	**0.33%**	*1.7B*	*6.3GB*
19	**DocPlayer**	**0.32%**	*1.6B*	*6.1GB*
20	**StackOverflow**	**0.28%**	*1.4B*	*5.3GB*

Table. MassiveWeb English: Top 20 Domains. Disclosed in **bold**. Determined in *italics*.

[29] Gopher paper: https://arxiv.org/abs/2112.11446 pp 44, Figure A3b.