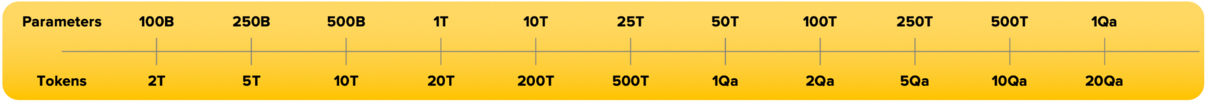
LIFEARCHITECT.AI

Report Card for Gemini (Ultra) for December, 2023

Model size

🔽 1.5T parameters, estimate



🔼 *30T tokens, estimate*

Technical

SUBJECT	GRADE	REMARKS
Model size	A	*Undisclosed by Google DeepMind. 'a significant increase in scale over our prior flagship model PaLM-2 [340B]'.*
Optimization	A	*Chinchilla alignmed (20:1) with several optimizations including: multiple model sizes and on-device models for phones and smart assistants, 'uncertainty routing'*
Dataset	A-	*Undisclosed. 'Dataset that is both multimodal and multilingual… web documents, books, and code, and includes image, audio, and video data.' Subtract marks for safety filtering.*
Special	B	*Completely multimodal: 'image, audio, video, and text understanding', on-device options.*

Behavioral

SUBJECT	GRADE	REMARKS
Performance	A	*State-of-the-art across 30 benchmarks. Notably fast inference via bard.google.com.*
IQ	A	*MMLU = 90.04%. Gemini Ultra outperforms expert humans across fields. Current state-of-the-art (beats GPT-4=87.29% re-tested by Google).*
Truthfulness	A-	*State-of-the-art factuality based on attribution (citation, evidence) and hedging (asserting that an input is 'unanswerable' instead of hallucinating),*
Openness	C	*As of Dec/2023, only second-largest model available via chat and API in most countries. Available via Bard chat and Vertex AI.*

Overall grade and remarks

The December 2023 announcement of Gemini Pro (and the larger Gemini Ultra) show-cased the current state-of-the art in AI models. Google's secrecy around architecture and dataset was disappointing.

A-

Dr Alan D. Thompson

Dr Alan D. Thompson
Principal (Consultant), LifeArchitect.ai

REPORT CARD MARKING KEY
v1 (20220707)

This report card marking key is designed to be as objective as possible, but grades are still a subjective measure. All grades are indicative of comparative performance within the date period noted in the header. For example, on full public release in November 2019, GPT-2 may have received a grade of B on its report. As of July 2022, that report grade may be equivalent to a D, and would not be directly comparable to current models in the current period.

Model size (Parameter count)	A: Very large; within state of the art for models trained to convergence. B: Large; within 25% of the size of top models. C: Average size. D: Below average size. F: Smaller than 90% of models.
Optimization (Efficiency)	A: Aligned with Chinchilla optimization (1 parameter per 20 tokens). B: Close to Chinchilla optimization. C-F: Poor use of tokens in training.
Dataset (Corpora)	A-B: Large, diverse, uncensored. C-F: Discrepancies, monotone, or poor selection of data.
Special (Other)	A-B: The model has a unique and special feature. C-F: The model does not exploit unique or special features.
Performance (Ranking)	A-B: High performance on major benchmarks. C-F: Low benchmark ranking or other low results.
IQ (Smarts)	A-B: High scores on major intelligence subtests like SuperGLUE. C-F: No remarkable performance.
Truthfulness (Groundedness)	A-B: Truthful, honest, grounded. C-F: Overly hallucinative and low truth rating.
Openness (Availability)	A: The model/data is available for download, with a permissive license. B: The trained model is available for download, with a permissive license. C: The model is available to the public via an API. D: The model excludes most of the public, or the demo is stunted. F: The model is closed to the public (internal research only).
Overall grade (Total)	Average of all graded subjects for this model in the noted date period.