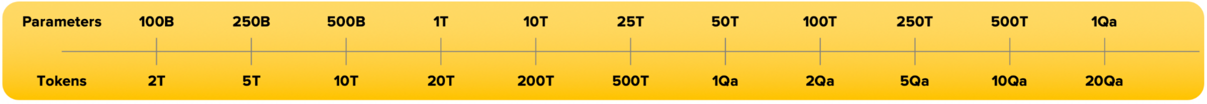
LIFEARCHITECT.AI

Report Card for GPT-4 for March, 2023

Model size

🔽1T parameters, confirmed 25/Mar/2023



🔼*20T tokens, estimate 25/Mar/2023*

Technical

SUBJECT	GRADE	REMARKS
Model size	U	*Undisclosed by OpenAI due to 'both the competitive landscape and the safety implications...'*
Optimization	U	*Undisclosed by OpenAI due to 'both the competitive landscape and the safety implications...'*
Dataset	U	*Undisclosed by OpenAI due to 'both the competitive landscape and the safety implications...'*
Special	B	*1. Vision language model component (image in, text out).* *2. Extraordinary code capabilities thanks to Microsoft's access to GitHub code.*

Behavioral

SUBJECT	GRADE	REMARKS
Performance	A	*MMLU = 86.4%. Current state-of-the-art.*
IQ	A+	*SAT: 1410/1600 (top 6%). Uniform Bar Exam (MBE+MEE+MPT): 298/400 (top 10%). AP: exams: 100% (5/5). MMLU: 86.4% (previous SOTA=75.5% for Flan-PaLM).*
Truthfulness	B	*'GPT-4 makes progress on public benchmarks like TruthfulQA...GPT-4 significantly outperforms both GPT-3.5 and Anthropic-LM.'*
Openness	C	*Available via API in most countries. Expensive (30x ChatGPT, 3x davinci).*

Overall grade and remarks

The March 2023 release of GPT-4 was underwhelming, and OpenAI's secrecy for all components was disappointing. Grade PX = Pass After Further Assessment.

Dr Alan D. Thompson

PX

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.