LIFEARCHITECT.AI

Report Card for <u>PaLM</u>   for <u>July, 2022</u>

Model size

▼540B parameters

Parameters	100B	250B	500B	1T	10T	25T	50T	100T	250T	500T	1Qa
Tokens	2T	5T	10T	20T	200T	500T	1Qa	2Qa	5Qa	10Qa	20Qa

▲*780B tokens*

Technical

SUBJECT	GRADE	REMARKS
Model size	A+	*At 540B parameters, PaLM is the largest dense large language model trained to convergence.*
Optimization	D	*Released pre-Chinchilla, using older scaling laws ('beliefs'). Google can learn from peers at DeepMind to ensure models are trained both efficiently & effectively.*
Dataset	B	*The text dataset covered a broad range of literature. Apparent censorship was low. It would be simple to increase visibility of the Common Crawl (web) in the next release.*
Special	B	*The Pathways architecture and roadmap is especially interesting, designed as discrete and interconnected modules that may help emulate the outputs of the human brain.*

Behavioral

SUBJECT	GRADE	REMARKS
Performance	A+	*PaLM outperforms nearly all current models, and some human performance, across a broad range of benchmarks..*
IQ	A+	*PaLM hits top 5 scores on a range of benchmarks including SuperGLUE and others: https://super.gluebenchmark.com/leaderboard & https://lifearchitect.ai/iq-testing-ai/*
Truthfulness	B	*PaLM seems to have solved many of the challenges associated with hallucinations and low truth scores, but the paper is untestable.*
Openness	F	*Google PaLM is not currently available to the public, and it seems that there are no plans to make it available in the short term.*

Overall grade and remarks

The April 2022 Google Pathways language model (PaLM) is one of the largest and most complete models announced. The Pathways architecture and family (including Jul/2022 text-to-image model, Parti), offer a glimpse into the future of AI.

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.