

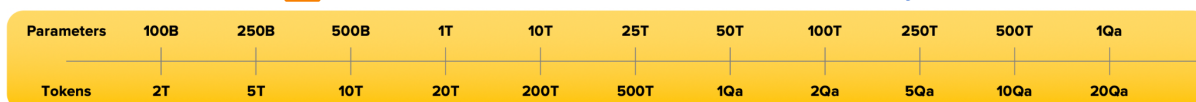
LIFEARCHITECT.AI

Report Card for PaLM 2

for May, 2023

Model size

▼ 340B parameters, CNBC confirmed [17/May/2023](#)



▲ 3.6T tokens, CNBC confirmed [17/May/2023](#)

Technical

SUBJECT	GRADE	REMARKS
Model size	A	Undisclosed by Google. 'significantly smaller than the largest PaLM model [540B] but uses more training compute'. CNBC confirmed 340B parameters.
Optimization	U	Undisclosed by Google.
Dataset	U	Undisclosed.. 'The pre-training corpus is significantly larger than the corpus used to train PaLM'. Multilingual. Team of 18. With DeepMind partnership, may use MassiveText .
Special	B	1. Multilingual: 'larger models can handle more disparate non-English datasets [helps] English language understanding performance'. 2. Compute-optimal.

Behavioral

SUBJECT	GRADE	REMARKS
Performance	A	WinoGrande = 90.9%. Current state-of-the-art (beats GPT-4=87.5%).
IQ	A	MMLU: 78.3/81.2% (SOTA=GPT-4=86.4%). PaLM 2 did very badly on SuperGLUE=86.4%, only slightly higher than GPT-3, RoBERTa, and worse than PaLM 540B, T5, ERNIE 3.0.
Truthfulness	U	Not addressed directly.
Openness	B	Available via API in most countries. Available via Bard chat.

Overall grade and remarks

The May 2023 release of PaLM 2 was an interesting and rapid development. Google's secrecy for all components was disappointing. Grade PX = Pass After Further Assessment.

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.

