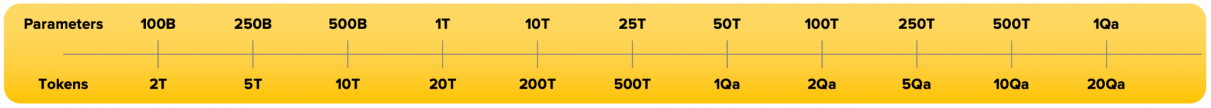
LIFEARCHITECT.AI

Report Card for LLaMA-65B for February, 2023

Model size

🔽65B parameters



🔼*1,400B tokens*

Technical

SUBJECT	GRADE	REMARKS
Model size	A	*Current state-of-the-art based on tokens trained, similar to DeepMind Chinchilla.*
Optimization	A	*1.4T tokens (1,400B). 22:1 tokens:parameters. Chinchilla was 20:1. GPT-3 was only 1.7:1.*
Dataset	B-	*Common crawl and other standard 'public data'.*
Special	C	*Instruction-tuned version called LLaMA-I. Interesting results.*

Behavioral

SUBJECT	GRADE	REMARKS
Performance	A	*MMLU = 63.4%. Outperforms Google PaLM Minerva 62B on GSM8k, despite not being fine-tuned on maths data. Outperforms Gopher 280B and Chinchilla 70B on TriviaQA.*
IQ	A	*MMLU = 63.4%.*
Truthfulness	A	*Better than GPT-3 (0.57 vs GPT-3 = 0.28)*
Openness	F	*Available to researchers only, non-commercial licence only.*

Overall grade and remarks

Fantastic addition to the open-ish models like OPT-175B and BLOOM.

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.