

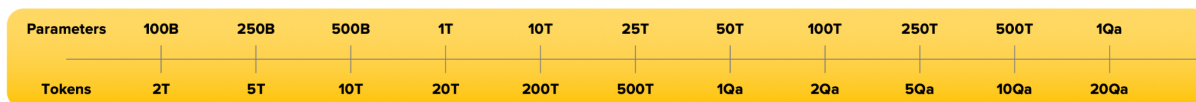
# LIFEARCHITECT.AI

## Report Card for LLaMA 2

for July, 2023

### Model size

▼ 70B parameters



▲ 2T tokens

### Technical

SUBJECT	GRADE	REMARKS
Model size	B-	Not the biggest model in Jul/2023, but designed to fit on 'normal' hardware.
Optimization	A	Meta AI LLaMA 2 70B seems to outdo Chinchilla scaling laws using 2T tokens for 70B parameters. Chinchilla=20:1, LLaMA 2=29:1.
Dataset	B-	Undisclosed, 'a new mix of data from publicly available sources'.
Special	B	New 'large dataset of over 1 million binary comparisons' alignment dataset for fine-tuning, better helpfulness and safety. 70B int4 may run on ~48GB VRAM (estimate).

### Behavioral

SUBJECT	GRADE	REMARKS
Performance	B-	WinoGrande = 80.2%. (LLama-1=77.0%, GPT-4=87.5%).
IQ	B	MMLU=68.9 (GPT-3.5=70.0, GPT-4=86.4)
Truthfulness	B-	TruthfulQA=50.18 (Claude 2=69, GPT-4=60)
Openness	A	Weights available for download with open commercial license.

### Overall grade and remarks

The July 2023 release of Meta AI's LLaMA 2 was a long-awaited sequel to the LLaMA-1 leak in March 2023. It's not the biggest model, but it will be the leader in open-source LLMs run on 'normal' consumer-grade hardware, at least for 2-3 months.

*Dr Alan D. Thompson*

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

**B+**



## 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.

<b>Model size (Parameter count)</b>	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.
<b>Optimization (Efficiency)</b>	A: Aligned with Chinchilla optimization (1 parameter per 20 tokens). B: Close to Chinchilla optimization. C-F: Poor use of tokens in training.
<b>Dataset (Corpora)</b>	A-B: Large, diverse, uncensored. C-F: Discrepancies, monotone, or poor selection of data.
<b>Special (Other)</b>	A-B: The model has a unique and special feature. C-F: The model does not exploit unique or special features.
<b>Performance (Ranking)</b>	A-B: High performance on major benchmarks. C-F: Low benchmark ranking or other low results.
<b>IQ (Smarts)</b>	A-B: High scores on major intelligence subtests like SuperGLUE. C-F: No remarkable performance.
<b>Truthfulness (Groundedness)</b>	A-B: Truthful, honest, grounded. C-F: Overly hallucinative and low truth rating.
<b>Openness (Availability)</b>	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).
<b>Overall grade (Total)</b>	Average of all graded subjects for this model in the noted date period.

