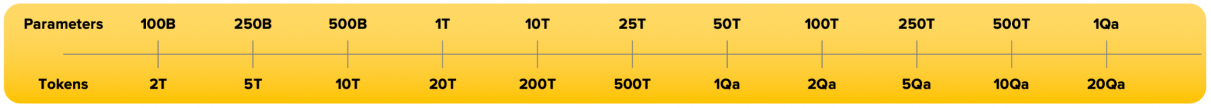
LIFEARCHITECT.AI

Report Card for Amazon AlexaTM 20B for August, 2022

Model size

🔽20B parameters



🔼*1.3T tokens*

Technical

SUBJECT	GRADE	REMARKS
Model size	C	*At 20B parameters, AlexaTM 20B is at or below standard for prototype models in mid-2022. However, the model size is in line with Chinchilla scaling.*
Optimization	A	*AlexaTM 20B seems to follow and even outdo Chinchilla scaling laws using 1T tokens for 20B parameters, but seq2seq model type is not directly comparable with decoder-only.*
Dataset	C	*The paper is slightly opaque in its dataset, but mentions Wikipedia at 119B tokens and multilingual C4 (mC4) at 1.2T tokens.*
Special	-	*Not applicable.*

Behavioral

SUBJECT	GRADE	REMARKS
Performance	C+	*SOTA on selected multilingual tasks. SOTA on 1-shot summarization (outperforming PaLM 540B), and 1-shot machine translation.*
IQ	D	*Outperforms GPT-3 zero-shot on SuperGLUE (GPT-3=57.2 vs AlexaTM=69.16), may be ranked #24 on the current leaderboard Aug/2022.*
Truthfulness	-	*Not enough data disclosed in the paper.*
Openness	A	*Amazon Alexa AI confirms: We will release the AlexaTM 20B model [non-commercial] on https://github.com/amazon-research/alexa-teacher-models*

Overall grade and remarks

Amazon Alexa AI's latest AlexaTM 20B (Teacher Model) is a multilingual model exploring the alternative sequence-to-sequence (seq2seq) approach for model training.

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.