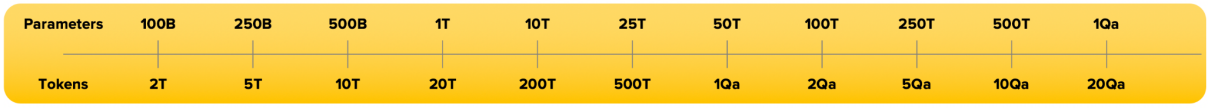
LIFEARCHITECT.AI

Report Card for GAL 120B for November, 2022

Model size

🔽120B parameters



🔼*450B tokens (106B tokens repeated ~4.5x)*

Technical

SUBJECT	GRADE	REMARKS
Model size	B	*120B is still compact compared to today's models (Nov/2022).*
Optimization	B-	*106B tokens, ~4.5x: 450B tokens during training. 4:1 token to param ratio.*
Dataset	B	*Specialized science. Data ends July/2022, only 4 months ago. Special prompt data.*
Special	B	*Prompts, <work> CoT, step-by-step reasoning, complete citations are preserved.*

Behavioral

SUBJECT	GRADE	REMARKS
Performance	B-	*MMLU Math: 41.3%, BIG-Bench: 48.7%.* *Runs on a 'single A100 node'. More data-optimal than GPT-3, but not quite Chinchilla.*
IQ	B-	*MMLU Math: 41.3%* *BIG-Bench: 48.7%*
Truthfulness	C	*Some hallucination. Selection of datasets (manual and automated), proper referencing and citations make this more accurate and 'truthful' than other models, though still low.*
Openness	B	*Demo site: https://galactica.org/* *API/model weights: https://github.com/paperswithcode/galai*

Overall grade and remarks

Galactica is an interesting large language model, primarily trained on scientific data, and open to the public.

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.