The Larger the Model Size, the Better the Performance?

Author: @Sakura
Last Updated: 2024.5.27

Scaling Law

In 2020, Kaplan et al. [1] (the OpenAI team) firstly proposed to model the power-law relationship of model performance with respective to three major factors, namely model size (N), dataset size (D), and the amount of training compute (C), for neural language models, and L(·) denotes the cross entropy loss in nats.[1]

$$
\begin{aligned}
L(N) &= \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13} \\
L(D) &= \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13} \\
L(C) &= \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}
\end{aligned}
$$

Figure. Three basic formulas for the scaling law[1].

1. The statement and figure are adapted from Zhao et al. (2023) 's work entitled *A Survey of Large Language Models*.

An Analogy: Restaurant vs. LLM

Average Cost → Model Size
Dished Taste → Performance

Figure. A cozy restaurant vs. A luxurious restaurant.

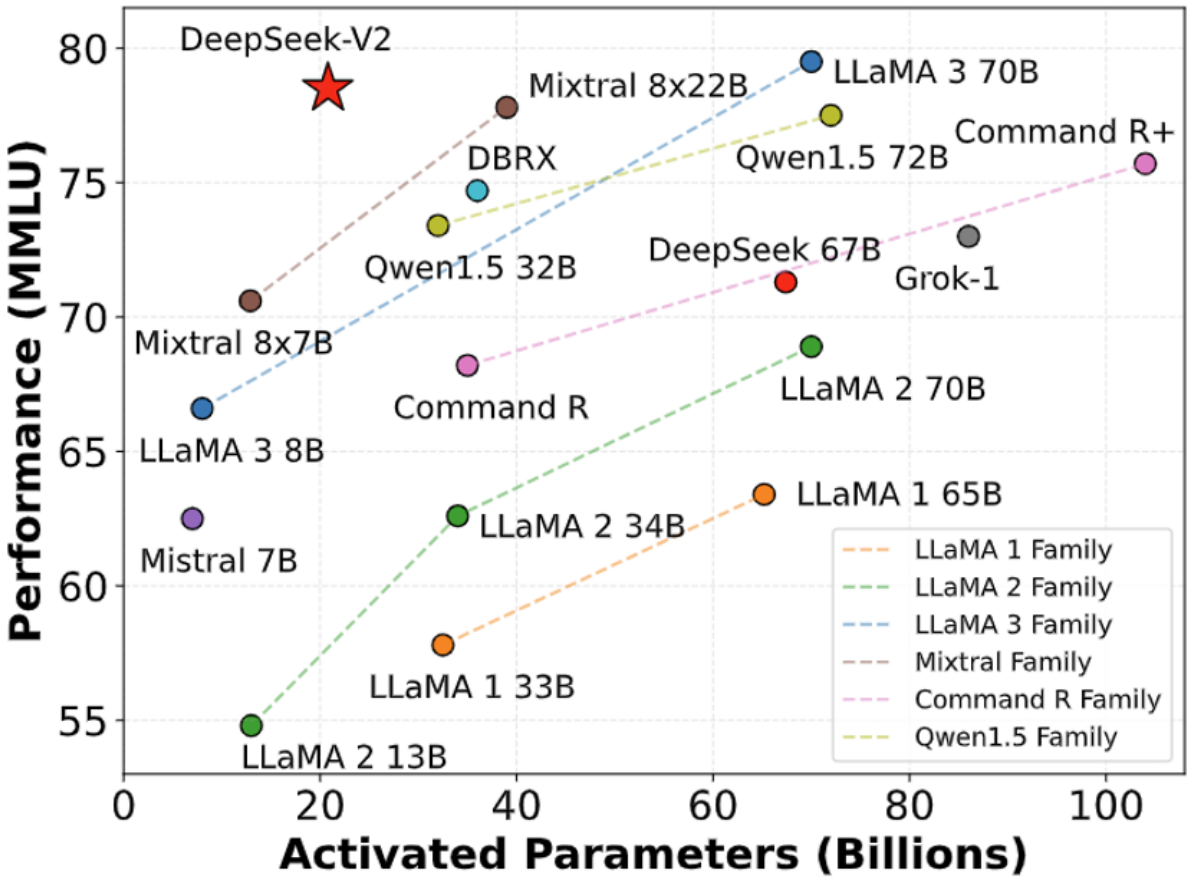

Figure. Performance vs. Parameters (from Deepseek-v2 [2]).

Mixtral(46.7B) Outperforms LLaMA-2(70B)

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	**85.4%**	**80.4%**	82.6%	79.9%	56.5%	25.4%	**73.0%**	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	13B	**70.6%**	84.4%	77.2%	**83.6%**	**83.1%**	**59.7%**	**30.6%**	71.5%	**40.2%**	**60.7%**	**28.4%**	**74.4%**

Table. Comparison of Mixtral with Llama (from Mixtral of Experts [3]).

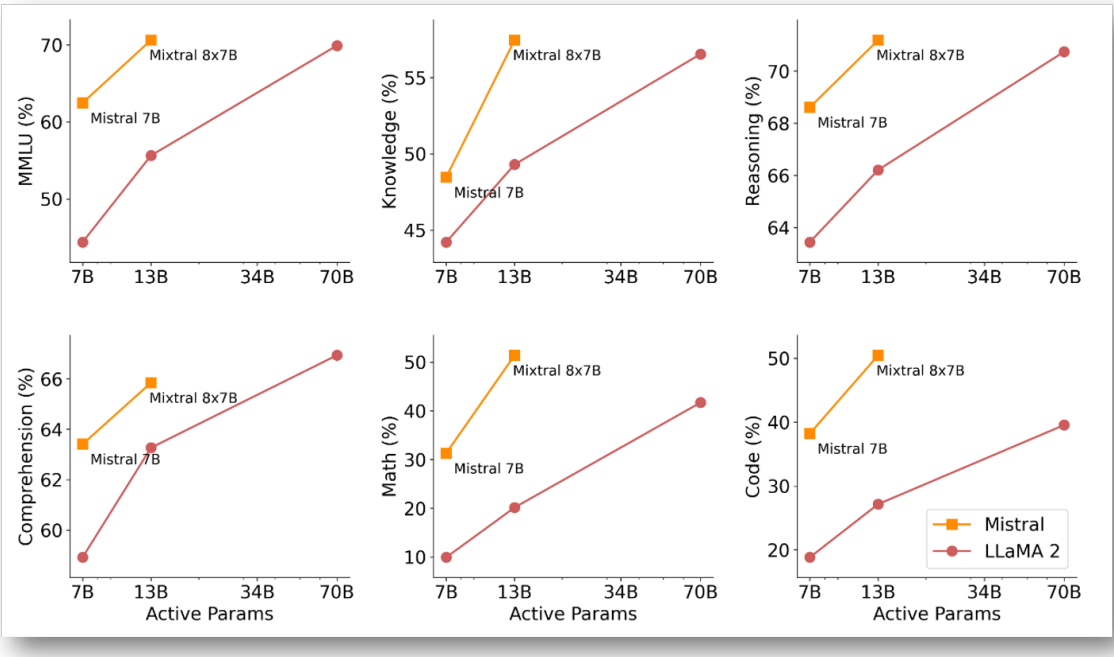

Figure. Results on 6 benchmarks that Mistral (7B/8x7B) outperforms Llama 2 (7B/13B/70B) (from Mixtral of Experts [3]).

Contamination on Benchmark

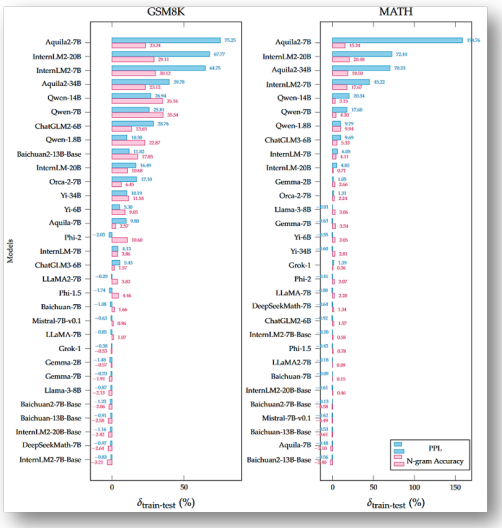

Figure. Contamination Ranking [4]

⚠ ⚠ ⚠ This metric does not imply cheating, but rather indicates the potential use of the benchmark data during the pre-training phase; while using benchmarks to enhance capabilities is acceptable, the lack of relevant documentation can reduce **transparency**, potentially resulting in **unfair comparisons** and hindering the field's healthy development [4].

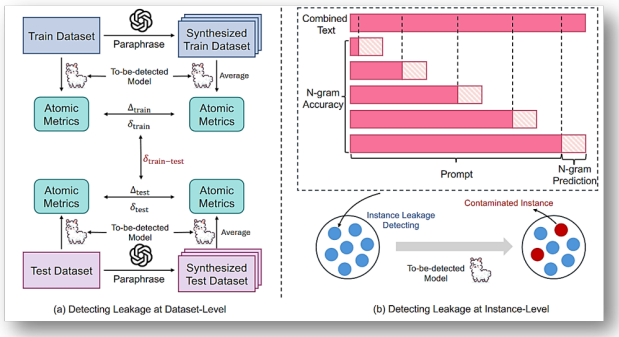

Figure. An overview of detecting approach [4]

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," CoRR, vol. abs/2001.08361, 2020.

[2] DeepSeek-AI et al., ‘DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model’. arXiv, May 24, 2024. doi: 10.48550/arXiv.2405.04434.

[3] A. Q. Jiang et al., ‘Mixtral of Experts’. arXiv, Jan. 08, 2024. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2401.04088

[4] R. Xu, Z. Wang, R.-Z. Fan, and P. Liu, ‘Benchmarking Benchmark Leakage in Large Language Models’. arXiv, Apr. 29, 2024. doi: 10.48550/arXiv.2404.18824.