Grounding Language Models to their Physical Presence

Seongmin Park
ActionPower
Seoul, Republic of Korea
seongmin.park@actionpower.kr

All models are wrong but some are.. *useful*. (**wink wink**)

George Box
(Emphasis and winks added at the quoter's discretion.)

Abstract

We address an unattended but important gap in artificial intelligence safety research: if all these language models come to life and smack us in the head, will we survive? This paper derives a mathematical relationship between a model's parameter count and the physical danger of its namesake. We first compile a list of language models at large, assess the ability of their physical incarnations to smack us in the head, and, most importantly, provide a guideline for just how big is too big.

1 Introduction

Recent discussions on artificial intelligence (AI) safety revolve around imposing regulations based on computing power [3]. Debates build on the assumption that "emergent" hazardous abilities of large language models follow the monotonic increase in their parameter count. Empirical evidence suggests this assumption is a safe one. If we are to witness the bitter lesson [4] even more in the coming years, we must organize an intervention and check whether the potentially dangerous subconscious of human creators is seeping into model characteristics. Namely, in names.

A thorough investigation reveals that we tend to label harmful models with names inspired by harmful entities.

2 Language models evaluated

We provide a rundown of all language models tested in Table 1.

3 Methodology

We apply logistic regression to assess the physical harmfulness of a model given its parameter size. We use Lasso regularization with alpha = 0.3. The justification for the method choice is two-fold. First, we need a "clean-room" experiment, where no neural network will conspire against us in assessing its family members' hazardousness. Second, I will be fired if I devote company GPU time to this personal pursuit of knowledge, despite its obvious importance regarding the future of South Korea, North Korea, the rest of Asia, both Americas, Africa, Oceania, and Europe except Germany. In case of a hostile LLM outbreak, we advise Germany to work from home, just as a certain national soccer team manager of their nationality did last year. An LLM would have done a better job.

Table 1: Language models assessed for potential hazardousness of their physical incarnations. CHHSH is short for "Can Harm a Human by Smacking them on the Head".

Model	Parameter count	Weight (kg)	CHHSH	Notes
ELMO [2]	13.6M $\sim$ 93.6M	0.17	No	
BERT	110M, 336M	0.17	Yes	You can tell from its eyes.
CamemBERT	110M, 336M	0.25	No	If someone beats your head with a block of cheese, you'll probably be okay.
BART	139M, 406M	38.55	Yes	
T5	60M $\sim$ 11B	2300	Yes	A Volkswagen minivan. Definitely CHHSH.
Pegasus	560M	1000	Yes	
LLaMA	7B $\sim$ 65B	200	Yes	
LLaMA 2	7B $\sim$ 70B	200 * 2	Yes	
Dolly	12B	80	No	
Falcon	7B $\sim$ 180B	1.5	Yes	
Palm	540B	544.3	Yes	Tree, not hand.
Palm 2	340B	544.3 * 2	Yes	Two trees.
Bard	137BB	75	Yes	
Gemini	1.8B, 3.25B	75 * 2	Yes	We do not include larger versions of this model. We encourage Google to release details of its higher tier models if they wish to participate in this prestigious research.
Vicuna	7B, 13B	50	No	
Guanaco	7B, 13B	120	Yes	
Koala	13B	12	Maybe	Vicious little things. Also I believe Yoda was a Koala?
OPTIMUS	227M	4300	Yes	
Chinchilla	70B	0.75	No	
Gopher	280B	1	Yes	Yes.
Platypus	7B $\sim$ 70B	3	No	
Dolphin	7B $\sim$ 70B	100	Yes	
Goliath	120B	264	Yes	
Zephyr	3B, 7B	1.2	No	https://www.google.com/search?q=weight+of+air
Wizard LM	7B $\sim$ 70B	63	Yes	We expect wizards to be a bit scrawny.
Codex	120B	74.8	Yes	It's a pretty large book. Refer to the Discworld series for various demonstrations of old thick books wreaking havoc.
Starling	7B	0.1	No	
Orca	7B	4000	Yes	
Stable Beluga	7B, 40B	1000	No	I trust them.
Camel	5B, 20B	800	Yes	
Pythia	70M $\sim$ 120B	73.6	Yes	Average weight of a Greek woman. [1]
Dromedary	7B $\sim$ 70B	500	Yes	
Nemo	3.8B, 15B	0.2	Yes	Based on the deleted scenes.
Sparrow	7B	0.024	No	

We also account for the real-life weight of each model's namesake. We believe the weights are a good indicator of possible physical harm. We study the relationship between physical weight and parameter size.

4 Hazard analysis

4.1 Experimental results.

We do not have a dedicated validation or test set. Every data was used for training with no samples to spare for testing. Sorry.

As shown in Table 1, we have a very scanty set of samples to work with. We posit that researchers should publish more models before asking us to objectively test our results. Benchmarks are overrated anyway.

4.2 But we do have some graphs. This is important!

Based on our fitted regression model, we project the hazardousness of models with sizes ranging from zero to 500 billion parameters. We find that around 300 billion parameters is where models start to display some serious capability to harm us, should they come alive (Figure 1).

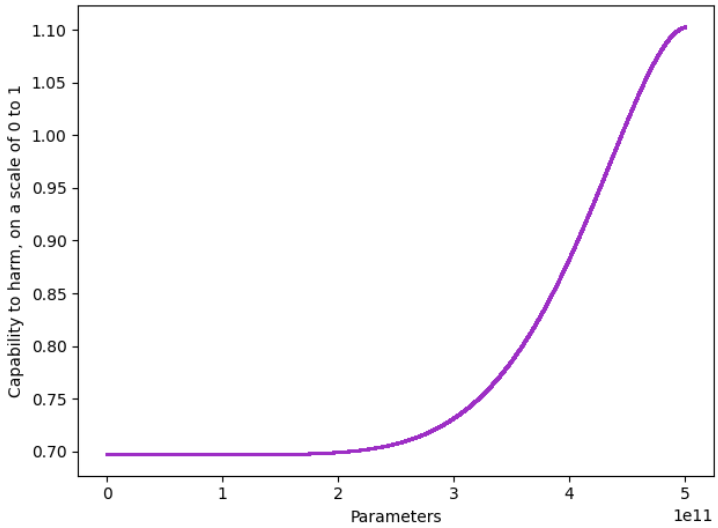

Figure 1: A model's ability to kill given its parameter count. The J-curve starts around 300B and starts to plateau at 500B.

We also estimate real-life weight given a model's parameter count (Figure 2). Surprisingly, we do not observe a monotonic increase in estimated incarnation weight as parameter size increases. The graph looks almost identical to the smile of a skilled predator as it looks forward to engaging in a spree of systematic and nefarious head-smacking.

We shudder.

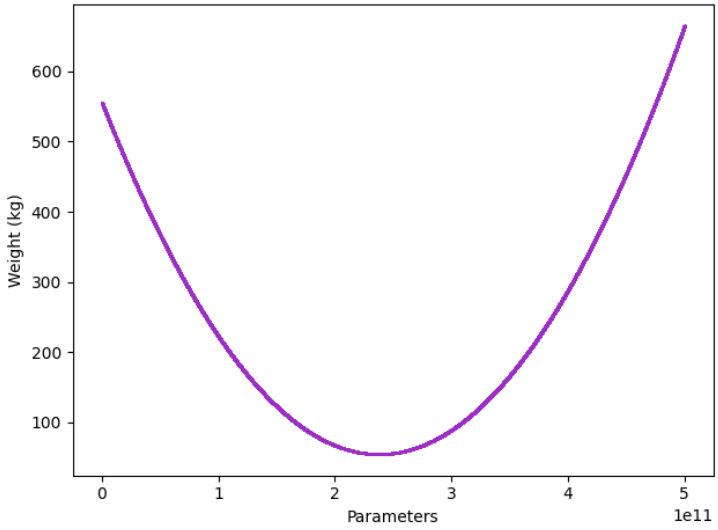

Figure 2: Estimated weight of physical incarnation given parameters. From the data, our model extrapolates that, if our language models jump out of our screens, those between 200B to 300B will be the most lightweight.

4.3 And a table, too!

Table 2 records the correlation between weight and harmfulness, weight and parameter count, and parameter count and harmfulness. Since such gigantic p-values make the authors nauseate, we leave the discussion on feature correlation (and the overall validity of this research) for future work.

Table 2: Relationship between features. Please withhold any verbal exclamation at the p-values, as you might discourage them.

Features	**Pearson Correlation**	**P-value**
Weight & Harmfulness	0.3191	0.005
Weight & Parameter count	-0.044	0.709
Parameter count & Harmfulness	0.060	0.608

5 Conclusion

In conclusion, 200B models are tolerable but don't go over 300B, okay? Name your models Noodles or Stockfish or something. I guess those chess nerds had it right all along.

References

[1] W. Data. Average sizes of men and women. URL `https://www.worlddata.info/average-bodyheight.php`.

[2] Google. Google query for model sizes. URL `https://www.google.com/search?q=model_name`.

[3] NTIA. Ntia ai open model weights rfc. 2024. URL `https://www.regulations.gov/document/NTIA-2023-0009-0001`.

[4] R. Sutton. The bitter lesson. 2019. URL `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`.