You Won't Believe This One WEIRD TRICK That BEATS ChatGPT on AIc (NOT CLICKBAIT)

Alex Xie[0], **Abhishek Vijayakumar**[0], **Erin Gao**[0],
Bhargav Hadya[0], **Samiksha Kale**[0], **Tara Lakdawala**[0]
Society
@neuralthenarwhal

Abstract

We introduce UNIFORMER, a novel nonparametric sublinear time and memory transformer architecture that comprehensively beats ChatGPT as well as virtually all modern neural language models on a variety of dataset[1] and metrics.

1 Introduction

Large language models (LLMs) such as ChatGPT have captured the public interest due to their ability to do math poorly (gpt, 2023b), generate offensive content (gpt, 2023a), incorrectly answer basic factoid questions (Pearl, 2022), and yet still pass collegiate-level examinations (OpenAI, 2023).

Rather than addressing these concerning behaviors, the research community has opted to focus on creating large language models that are either larger (OpenAI, 2023), worse than existing models (Bennet, 2023), or posted on 4chan (Vincent, 2023).

We propose a novel language model architecture that is much smaller than existing LLMs, beats SoTA language models on a variety of metrics, and is extremely unlikely to be posted on 4chan. Despite improving upon these aspects of LLMs, our model cannot pass Advanced Placement (AP) examinations and thus validates the continued existence of the College Board. [2]

2 Method

Language modeling is the task of assigning a probability to a sequence of tokens S. As is standard in language models, we decompose this probability $P(S)$ autoregressively:

$$P(S) = \prod_{t=1}^{T} p_\theta(s_t \mid s_{<t})$$

Inspired by recent LLM architectures, we propose a transformer architecture composed of n repeated blocks, where each block consists of the following operations performed sequentially to best avoid GPU exploitation[3]:

TikTok-normalized feedforward Layer norm (Ba et al., 2016) is a normalization technique used in almost all transformers. But are people using layer norm in their LLMs because it actually works, or are they just scared of looking like they aren't good at machine-learning-ing? Related work on batch normalization would suggest it's the latter (Wise, 2017). As people who are openly bad at machine-learning-ing, we introduce our much-less-effective-but-also-much-less-pretentious alternative to layer norm, Tiktok normalization, shown in Algorithm 1.

Algorithm 1 TikTok normalization

Require: TikTok, integer k, crippling procrastination
  $c \leftarrow 0$
  while $c < k$ **do**
    Swipe to next video V
    if V asks "Can we normalize x?" **then**
      $x \leftarrow \frac{x}{||x||}$
      $c \leftarrow c + 1$
    end if
  end while

Decapitated Self-attention Attention is at

[0] Inequal contribution

[1] *dataset*, singular

[2] an American nonprofit educational assessment organization that made over $50 million in profit in 2019

[3] See our ethics statement.

	AIC ↓	BIC ↓	HQC ↓
LLAMA	130,000,000,000	702,247,747,500	309,386,868,400
CHATGPT	350,000,000,000	1,890,667,010,905	832,964,645,693
GPT-4	200,000,000,000,000	1,080,381,149,088,820	475,979,797,538,603
UNIFORMER	**1049.98**	**1049.98**	**1049.98**

Table 1: Various Information Criteria on Penn Treebank Corpus

the core of transformers and supposedly all you need[4]. Specifically, transformers use multi-head attention, a variant of attention in which m disembodied "heads" are forced against their will to pay attention to potentially toxic, psychologically scarring texts (gpt, 2023a). Recently, the UN Human Rights Council and other humanitarian institutions have critiqued the barbarism of this technique (Michel et al., 2019). We propose to go one step further and decapitate all the heads to put them out of their collective misery. This can be viewed as a generalization of multi-head attention with $m = 0$ heads.

Superlinear nearest neighbors retrieval Recent work has proposed augmenting LLMs with a retrieval component (Khandelwal et al., 2020; Borgeaud et al., 2022). These models generally use sublinear-time nearest neighbors retrieval (Johnson et al., 2017). However, we point out that these efficient retrieval algorithms are inexact and thus may yield sub-optimal results. Instead, we propose to perform exact search by simply looping through all possible subsets of the retrieval datastore, filtering by size, and taking the one with the lowest total distance from our query. While we've been told that this is "exponential time," "not tractable," and "a gigantic waste of compute resources," we prefer to take the glass half full approach and think of it as "better-than-linear" and "leaving no stone unturned." Interestingly, in our model, we find that our exact search is no slower than approximate nearest neighbors search.

Markov-Chain Monte Carlo Metropolis-Hastings Variational Reparametrized Minimum Bayes Risk Annealed Dropout We ran out of funny things to say, so following past work, we were hoping we could write a bunch of big ML words here to intimidate people out of reading this section (Gupta and Jain, 2020).

As a minor experimental detail, note that in our model, we take the number of transformer blocks $n = 0$.

On top of our transformer states, we learn a non-parametric language modeling head. Specifically, we compute our distribution over the vocabulary as

$$p_\theta(w_t \mid w_{<t}) \propto \lim_{\tau \to \infty} \exp\left(\frac{W_{\text{LM}}\mathbf{h}_t}{\tau}\right)$$

where W_{LM} is the output matrix, $\mathbf{h}_t$ is the t-th hidden state, and τ is the temperature at which we sample (Ackley et al., 1985).

Since this reduces to a uniform distribution over the vocabulary, we elide W_{LM} and store zero parameters in GPU memory for our final model.

3 Model Validation & Experiments

As UNIFORMER has no parameters, we must conclude that its performance stems from a complete understanding of the English language, embedded into its architecture in the Chomskyan sense (Chomsky, 2006). This makes UNIFORMER the second model to exhibit "sparks of Artificial General Intelligence," (Bubeck et al., 2023), but the first to do so without unprompted generation of toxic content.

Given the potential to become an AGI, we refrain from implementing UNIFORMER on conventional hardware to prevent the technological singularity (Chalmers, 2010). All results were instead computed via restricted simulation and theoretical performance bounds on the *Desmos* consumer-oriented cloud-based analytical mathematics system (Desmos, 2023).

4 Evaluation

We describe in this section the metrics used to evaluate our model, reported in Table 1. For all

[4]Along with all these other things. Something's not adding up here.

metrics, lower values indicate better models. The values presented for all LLMs are estimated lower bounds based on publicly available knowledge. We take parameter counts for LLAMA and CHATGPT from their respective papers and we take the parameter count for GPT-4 from Twitter.

Following recent work, we evaluate exclusively on the Akaike (Akaike, 1974), Bayesian (Schwarz, 1978), and Hannan-Quinn (Hannan and Quinn, 1979) Information Criteria, which are defined as

$$\text{AIC} = 2k - 2\ln(\hat{L})$$
$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$
$$\text{HQC} = 2k\ln(\ln(n)) - 2\ln(\hat{L})$$

where k represents the number of parameters of a given model, n represents the sample size, and $\hat{L}$ represents the likelihood of the sample according to the model. For the Penn Treebank, $n = 49208$ (Marcus et al., 1993). Note that for UNIFORMER, $k = 0$.

5 Environmental Impact

Naturally occurring ecosystems consist of several *trophic levels*, each of which contains increasingly complex organisms that obtain energy by consumption of organisms in lower trophic levels. Notably, energy transfer between trophic levels is inefficient: only about 10% of the energy in one trophic level progresses to the next (Urry et al., 2016).

Traditional large language models occupy a unique niche in the ecosystem: they are both scavengers, consuming by-products of human activity in the form of language artifacts, and parasites, surviving on GPU "cluster" colony activity to the detriment of the component GPUs. LLMs also cause harmful human activity: they have historically promoted the large-scale construction of *treebanks* (Marcus et al., 1993), which are likely created through deforestation and may contribute to the endangering of several species.

UNIFORMER is an energy-efficient organism that may outcompete LLMs on several levels. Due to its incredibly effective performance on language-related tasks, humans will no longer need to engage in deforestation in order to support LLMs. UNIFORMER may also generate synthetic language artifacts masquerading as human artifacts that traditional LLMs may unknowingly consume, a technique it likely learned from its study of the Trojan war (aen, 1996).

While LLMs draw energy from multiple trophic levels including those of trees and humans, UNIFORMER does not rely on any other organism for energy. It is thus a minimum of 10 times as efficient as an LLM. We predict that the widespread introduction of UNIFORMER into existing ecosystems will drive LLMs extinct, allowing both forests and GPU colonies to flourish.

6 Ethics Statement

We are categorically against any and all forms of exploitation, including labor, GPU, and child.

We are categorically against any and all forms of labor, including GPU and child.

We are categorically against any and all forms of GPU [5], including child.

We are categorically against any and all forms of child.

7 Conclusions

OpenAI, Google Brain, FAIR and Microsoft Research should all immediately disband and devote all their remaining funding toward our model. UNIFORMER can be run on a single consumer GPU due to its novel architecture. Each author requests one *NVIDIA® GeForce RTX™ 4090* for continued model development.

References

[Ackley et al.1985] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

[aen1996] 1996. *Vergil's Aeneid*. Bloom's notes. Chelsea House Publishers, New York.

[Akaike1974] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December.

[Ba et al.2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

[Bennet2023] Sharron Bennet. 2023. Did google's bard ai tool just commit its first error in a demo?, Feb.

[Borgeaud et al.2022] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron

[5] except when given to us (see Section 7)

Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens.

[Bubeck et al.2023] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

[Chalmers2010] David J. Chalmers. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10):9–10.

[Chomsky2006] Noam Chomsky. 2006. *Language and Mind*. Cambridge University Press, January.

[Desmos2023] Desmos. 2023. Desmos — graphing calculator.

[gpt2023a] 2023a. Chatgpt's creators say ai has been 'biased, offensive and objectionable', Feb.

[gpt2023b] 2023b. Wolfram: Alpha as the way to bring computational knowledge superpowers to chatgpt, Jan.

[Gupta and Jain2020] Divam Gupta and Varun Jain. 2020. Gradschoolnet: Robust end-to-end *-shot unsupervised deepaf neural attention model for convexly optimal (artifically intelligent) success in computer vision research. In *Proceedings of the 14th ACH SIGBOVIK Special Interest Group on Harry Query Bovik.*

[Hannan and Quinn1979] E. J. Hannan and B. G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195, January.

[Johnson et al.2017] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

[Khandelwal et al.2020] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*.

[Marcus et al.1993] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[Michel et al.2019] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[OpenAI2023] OpenAI. 2023. Gpt-4 technical report.

[Pearl2022] Mike Pearl. 2022. The chatgpt chatbot from openai is amazing, creative, and totally wrong, Dec.

[Schwarz1978] Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), March.

[Urry et al.2016] Lisa Urry, Michael Cain, Steven Wasserman, Peter Minorsky, and Jane Reece. 2016. *Campbell Biology*. Campbell Biology Series. Pearson.

[Vincent2023] James Vincent. 2023. Meta's powerful ai language model has leaked online - what happens now?, Mar.

[Wise2017] Joshua A. Wise. 2017. Batch normalization for improved dnn performance, my ass. In *Proceedings of the 11th ACH SIGBOVIK Special Interest Group on Harry Quechua Bovik.*