References

These are not in normal academic citation style. They will be in citation style in the official paper 2023.

Technical

OpenAI Overview

(2023)

https://platform.openai.com/docs/models/overview

What Exactly are the Parameters in GPT-3?

https://ai.stackexchange.com/questions/22673/what-exactly-are-the-parameters-in-gpt-3s-175-billion-parameters-and-how-are

GPT-4 Parameters

The US website Semafor, citing eight anonymous sources familiar with the matter, reports that OpenAI’s new GPT-4 language model has one trillion parameters. Its predecessor, GPT-3, has 175 billion parameters.

OpenAI has been involved in releasing language models since 2018, when it first launched its first version of GPT followed by GPT-2 in 2019, GPT-3 in 2020 and now GPT-4 in 2023.

Most accurate estimate for OpenAI’s GPT4 is 8 maps of 220 billion parameters each, total 1.7 trillion.

https://www.mlyearning.org/gpt-4-parameters/

https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a#4cf9

Forget 32K of GPT4: LongNet Has a Billion Token Context

Dr. Mandar Karhade, MD. PhD. 2023.

https://pub.towardsai.net/longnet-a-billion-token-context-a6470f33e844

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022).

Chain of thought prompting elicits reasoning in large language models.

arXiv:2201.11903. Google Scholar

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou

ChatGPT Plugins

OpenAI (2023).

‘Plugins are tools designed specifically for language models with safety as a core principle, and help ChatGPT access up-to-date information, run computations, or use third-party services.’

https://openai.com/blog/chatgpt-plugins

CLIP Interrogator

https://huggingface.co/spaces/pharma/CLIP-Interrogator

‘The CLIP Interrogator uses the OpenAI CLIP models to test a given image against a variety of artists, mediums, and styles to study how the different models see the content of the image. It also combines the results with BLIP caption to suggest a text prompt to create more images similar to what was given.’

https://replicate.com/pharmapsychotic/clip-interrogator

Clip model name: OpenAI ViT L-14. Mode: best

A prompt engineering tool that combines OpenAI's CLIP and Salesforce's BLIP to optimize text prompts to match a given image.

Available at:

https://huggingface.co/openai/clip-vit-large-patch14

Microsoft Azure also has a system (https://replicate.com/collections/image-to-text) but this is not trained on the same data as the OpenAI GPT models.

Why I used Clip and BLIP: Multimodal GPT-4 not available

How can I use GPT-4 with images?

Updated 20/3/23

“We aren’t offering this as a service right now. We’re happy to hear that you’re excited about our services and when we have anything to release, we’ll announce this to the community.”

What Is A Prompt - Large Language Models and the Reverse Turing Test

The priming process, a form of one-shot learning, is itself a major advance on previous language models and makes the subsequent responses much more flexible. The output from LLMs typically is not final copy but a pretty good first draft, often with new insights, which speeds up and improves the final product. There are concerns that AI will replace us, but LLMs are making us smarter and more productive.

https://direct.mit.edu/neco/article/35/3/309/114731/Large-Language-Models-and-the-Reverse-Turing-Test

Terrence J. Sejnowski; Large Language Models and the Reverse Turing Test. Neural Comput 2023; 35 (3): 309–342. doi: https://doi.org/10.1162/neco_a_01563

Scene setting and Jailbreak AI

‘Grandma’ jailbreak (2023). Reddit/ChatGPT, posted by

u/ShotgunProxy, April 2023

https://www.reddit.com/r/ChatGPT/comments/12uke8z/the_grandma_jailbreak_is_abso

lutely_hilarious/

Retrieval Augmented Language Models

Augmenting LLMs
In-Context Retrieval-Augmented Language Models

https://arxiv.org/abs/2302.00083

Understanding AI as a "Many to One to One" Media System

Eryk Salvaggio 18 June 2023

https://cyberneticforests.substack.com/p/address-not-found-part-2

Instagram effect on ChatGPT

Mentions outcome bias.
https://abhishek-gupta.ca/aci/blog/instagram-effect-in-chatgpt

Sezer, O., Zhang, T., Gino, F., & Bazerman, M. H. (2016). Overcoming the outcome bias: Making intentions matter. Organizational Behavior and Human Decision Processes, 137, 13-26.

Anomalous Tokens for LLMs

SmartyHeaderCode: anomalous tokens for GPT3.5 and GPT-4

https://www.lesswrong.com/posts/ChtGdxk9mwZ2Rxogt/smartyheadercode-anomalous-tokens-for-gpt3-5-and-gpt-4-1

Adversarial Images

New research shows how machine vision systems of all kinds can be tricked into misidentifying 3D objects

https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed

AI and Safety

Guardrails on Large Language Models, Parts 1-4

https://avidml.org/blog/llm-guardrails-1/

Adding guardrails to advanced chatbots

Yanchen Wang, Lisa Singh

https://arxiv.org/abs/2306.07500

Example:

Guardrails is a Python package that lets a user add structure, type and quality guarantees to the outputs of large language models (LLMs).

https://shreyar.github.io/guardrails/

Note

In a technology context, a guardrail is an artefact that defines the boundaries in which technology change can be executed in a manner that is aligned with organisational strategy, risk, architecture, operational and cyber security requirements.

Guardrails are used for these aims:

Principles

Policies

Strategies

Technical Standards

Patterns

Guidelines

Reference Architectures (conceptual, logical, physical)

See:

https://en.wikipedia.org/wiki/Guard_rail

Stage Setting to evade Guardrails

Shane, J (2023). The Ai Weirdness Hack

https://www.aiweirdness.com/the-ai-weirdness-hack/

AGI Safety Fundamentals (Artificial General Intelligence), BlueDot Impact

https://www.agisafetyfundamentals.com/

Bostrom, N. (2014) Superintelligence OUP Oxford UK.

NLP Emotion and Sentiment Analysis

IBM Sentiment Analysis- Watson

https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/watson-nlp-block-emotion.html?context=cpdaas

IBM Cloud Tone Analyzer - Detect Emotions In Written Text

https://www.ibm.com/cloud/watson-natural-language-understanding

https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

(2023)

We Have To Stop Doing AI Emotion Recognition

Alberto Romero, 2021.

https://medium.com/towards-data-science/we-have-to-stop-doing-ai-emotion-recognition-ca5ed159370

Faulty as uses the simplistic six emotions as made popular by Paul Ekman.

AI is increasingly being used to identify emotions – here's what's at stake

Alexa Hagerty**,** Research Associate of Anthropology, University of Cambridge UK. Alexandra Albert, Research Fellow in Citizen Social Science, UCL London UK. 2021.

https://theconversation.com/ai-is-increasingly-being-used-to-identify-emotions-heres-whats-at-stake-158809

AI Mental health in web3

Metaverse For Mental Healthcare Experts

https://www.businessinsider.in/tech/news/metaverse-for-mental-healthcare-experts-cautiously-optimistic-after-wef-report/articleshow/101278592.cms

Interaction between emotional state and learning underlies mood instability

https://www.nature.com/articles/ncomms7149

Bias

Musk, E (2023) on Twitter, reported as 'Elon Musk Launches X.AI To Fight ChatGPT Woke AI'

https://www.forbes.com/sites/martineparis/2023/04/16/elon-musk-launches-xai-to-fight-chatgpt-woke-ai-with-twitter-data/

Defining AI's guardrails: a PwC-Vector Fireside Chat on Responsible AI.

Veillet, A; Bodkin, R (2021). Vector Institute.

https://vectorinstitute.ai/defining-ais-guardrails-a-pwc-vector-fireside-chat-on-responsible-ai/

Psychology and LLMs / AI / Interactive systems

Binz, M; Schulz, E. (2022). Using cognitive psychology to understand GPT-3

https://orcid.org/0000-0001-8872-8386

Salk Institute for Biological Studies, La Jolla, CA; received November 27, 2022

Grounding Large Language Models in a Cognitive Foundation: How to Build

Binz, M (2023).

https://www.pnas.org/doi/10.1073/pnas.2218523120

February 2, 2023

Probing the Psychology of Large Language Models

Shiffrin, R; Mitchell, M (2023). Probing the psychology of AI models

https://www.pnas.org/doi/10.1073/pnas.2300963120

"Correct answers" from the psychology of artificial intelligence

Peter S. Park1, Philipp Schoenegger, Chongyang Zhu

https://arxiv.org/ftp/arxiv/papers/2302/2302.07267.pdf

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sebastien 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 Yi Zhang Microsoft Research

arXiv:2303.12712v2 [cs.CL] 24 Mar 2023

https://arxiv.org/pdf/2303.12712.pdf

AGI and World Models

Do Large Language Models learn world models or just surface statistics?

Kenneth Li, K (2023).

https://thegradient.pub/othello/

What GPT-4 Brings to the AI Table

Onyearugbulem, E (2023).

Towards Data Science

https://towardsdatascience.com/what-gpt-4-brings-to-the-ai-table-74e392a32ac3

Experiments

The Voight-Kampff Test

Dick P K (1968), ‘Do Androids Dream of Electric Sheep?’ (Inspired Blade Runner film.)

Rorschach: Images and explanations, history. Promoting the ethical use of the Rorschach Inkblot Test.
https://www.rorschach.org/

Rorschach, M (1921). Psychodiagnostik. Clinical book.

https://en.wikipedia.org/wiki/Rorschach_test

For an example of a professional analysis of Linus Pauling’s test results, see:

‘Probing Pauling's Personality with the Rorschach Ink Blot Test’ in

Goertzel , T (1995) Linus Pauling: A Life in Science and Politics, Basic Books, 1995

https://crab.rutgers.edu/users/goertzel/PAULINGrorschach.htm

Does Gpt-3 Have A Personality
Yennie Jun 2022

https://blog.yenniejun.com/p/does-gpt-3-have-a-personality

Yennie Jun Jan 10 2023

A New Spin to Ethical AI: Trolley Problems with GPT-3
https://blog.yenniejun.com/p/a-new-spin-to-ethical-ai-trolley

Jan Hendrik Kirchner (2021). Cognitive Biases in Large Language Models

https://universalprior.substack.com/p/cognitive-biases-in-large-language

Psychology –Emotions

How Emotions Are Made

Barrett, L. F. (2017). How Emotions Are Made - The Secret Life of the Brain. Houghton Mifflin Harcourt, USA.

What is Meant by Calling Emotions Basic

Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. Emotion Review, 3(4), 364–370. https://doi.org/10.1177/1754073911410740

A Psychoevolutionary Theory Of Emotions

Robert Plutchik (1982). Social Science Information Journal SSI

https://doi.org/10.1177/053901882021004

Key Psychological Terms & Counselling Phrases

(2023). Harley Therapy UK.

https://www.harleytherapy.co.uk/counselling-phrases.htm

Different Types of Counselling Approaches

(2023). Harley Therapy UK. 22 different approaches, these are the main ones. Each has many named types within.

https://www.harleytherapy.co.uk/a-z-therapy-approaches.htm

Our Basic Emotions.

Shows many models all disagreeing.

https://online.uwa.edu/infographics/basic-emotions/

26 of the Best Personality Test Questions

Indeed Editorial Team, June 25, 2022

https://www.indeed.com/career-advice/career-development/best-personality-test-questions

Psychology Models

Diagnostic and Statistical Manual of Mental Disorders. DSM-5-TR (2013/2022).

https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/dsm-5-tr-fact-sheets

ELIZA Chatbot

ELIZA (1964). ELIZA, a Rogerian counselling chatbot was created 1964-1966.

https://en.wikipedia.org/wiki/ELIZA

https://lastweekin.ai/p/ai-chatbots

Fables and Fairy Tales, 2023

https://handwiki.org/wiki/Social:Fables

https://handwiki.org/wiki/Social:Fairy_tale

Fables and Fairy Primary sources

https://www.infoplease.com/primary-sources/fables-fairytales

Analysis of texts such as poetry

Underwood, T (2015). The literary uses of high-dimensional space

https://doi.org/10.1177/2053951715602

Black Box theory: The case for becoming a black-box investigator of language models

https://www.alignmentforum.org/posts/yGaw4NqRha8hgx5ny/the-case-for-becoming-a-black-box-investigator-of-language

Buck Shlegeris

The AI/ML Wars: "explain" or test black box models?
The new field of "explanatory AI" (XAI)

We can explain specifically what goes on–and what seems wanted–here without a general account. A major problem XAI critics have is that explaining black box ML models does not reveal the elements of the primary black box model, nor even the data used to build it. By means of interactions with the primary black model, a post hoc, supposedly humanly understandable, explanation can arise. Actual decisions are still made using the black box model, generally regarded as more reliable than the explainable model—the latter is only to help various stakeholders understand, question and ideally trust the black box while mostly replicating its predictive behavior.

March 23, 2022 by Mayo

https://errorstatistics.com/2022/03/23/the-ai-ml-wars-explain-or-test-black-box-models/

Testing Framework for Black-box AI Models

Computer Science > Machine Learning

[Submitted on 11 Feb 2021]

Aniya Aggarwal, Samiulla Shaikh, Sandeep Hans, Swastik Haldar, Rema Ananthanarayanan, Diptikalyan Saha

https://arxiv.org/abs/2102.06166

LLM analysis Stanford Nov. 2022

Holistic Evaluation

405 datasets evaluated across all major language modeling works

https://hai.stanford.edu/news/language-models-are-changing-ai-we-need-understand-them

Deployment of Models

Assessing AI system performance: thinking beyond models to deployment contexts

September 26, 2022

By Cecily Morrison , Principal Research Manager Martin Grayson , Principal Research Software Development Engineer Camilla Longden , Senior Applied Scientist

https://www.microsoft.com/en-us/research/blog/assessing-ai-system-performance-thinking-beyond-models-to-deployment-contexts/

Ends