WIKIPEDIA
The Free Encyclopedia

Large language model

A **large language model** (**LLM**) is a type of machine learning model designed for natural language processing tasks such as language generation. LLMs are language models with many parameters, and are trained with self-supervised learning on a vast amount of text.
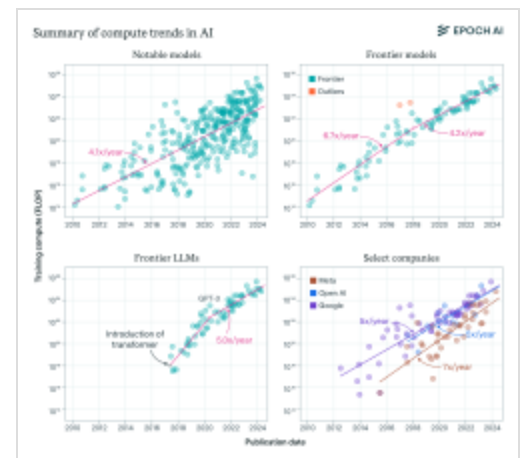
The largest and most capable LLMs are generative pretrained transformers (GPTs). Modern models can be fine-tuned for specific tasks or guided by prompt engineering.[1] These models acquire predictive power regarding syntax, semantics, and ontologies[2] inherent in human language corpora, but they also inherit inaccuracies and biases present in the data they are trained in.[3]

History

Before 2017, there were a few language models that were large as compared to capacities then available. In the 1990s, the IBM alignment models pioneered statistical language modelling. A smoothed n-gram model in 2001 trained on 0.3 billion words achieved state-of-the-art perplexity at the time.[4] In the 2000s, as Internet use became prevalent, some researchers constructed Internet-scale language datasets ("web as corpus"[5]), upon which they trained statistical language models.[6][7] In 2009, in most language processing tasks, statistical language models dominated over symbolic language models because they can usefully ingest large datasets.[8]

After neural networks became dominant in image processing around 2012,[9] they were applied to language modelling as well. Google converted its translation service to Neural Machine Translation in 2016. Because it preceded the existence of transformers, it was done by seq2seq deep LSTM networks.

The training compute of notable large models in FLOPs vs publication date over the period 2010-2024. For overall notable models (top left), frontier models (top right), top language models (bottom left) and top models within leading companies (bottom right). The majority of these models are language models.

At the 2017 NeurIPS conference, Google researchers introduced the transformer architecture in their landmark paper "Attention Is All You Need". This paper's goal was to improve upon 2014 seq2seq technology,[10] and was based mainly on the attention mechanism developed by Bahdanau et al. in 2014.[11] The following year in 2018, BERT was introduced and quickly became "ubiquitous".[12] Though the original transformer has both encoder and decoder blocks, BERT is an encoder-only model. Academic and research usage of BERT began to decline in 2023, following rapid improvements in the abilities of decoder-only models (such as GPT) to solve tasks via prompting.[13]

Although decoder-only GPT-1 was introduced in 2018, it was GPT-2 in 2019 that caught widespread attention because OpenAI at first deemed it too powerful to release publicly, out of fear of malicious use.[14] GPT-3 in 2020 went a step further and as of 2024 is available only via API with no offering of downloading the model to execute locally. But it was the 2022 consumer-facing browser-based ChatGPT that captured the imaginations of the general population and caused some media hype and online buzz.[15] The 2023 GPT-4 was praised for its increased accuracy and as a "holy grail" for its multimodal capabilities.[16] OpenAI did not reveal the high-level architecture and the number of parameters of GPT-4. The release of ChatGPT led to an uptick in LLM usage across several research subfields of computer science, including robotics, software engineering, and societal impact work.[13] In 2024 OpenAI released the reasoning model OpenAI o1, which generates long chains of thought before returning a final answer.

The training compute of notable large AI models in FLOPs vs publication date over the period 2017-2024. The majority of large models are language models or multimodal models with language capacity.

Competing language models have for the most part been attempting to equal the GPT series, at least in terms of number of parameters.[17]

Since 2022, source-available models have been gaining popularity, especially at first with BLOOM and LLaMA, though both have restrictions on the field of use. Mistral AI's models Mistral 7B and Mixtral 8x7b have the more permissive Apache License. In January 2025, DeepSeek released DeepSeek R1, a 671-billion-parameter open-weight model that performs comparably to OpenAI o1 but at a much lower cost.[18]

Since 2023, many LLMs have been trained to be multimodal, having the ability to also process or generate other types of data, such as images or audio. These LLMs are also called large multimodal models (LMMs).[19]

An illustration of main components of the transformer model from the original paper, where layers were normalized after (instead of before) multiheaded attention

As of 2024, the largest and most capable models are all based on the transformer architecture. Some recent implementations are based on other architectures, such as recurrent neural network variants and Mamba (a state space model).[20][21][22]

Dataset preprocessing

Tokenization

As machine learning algorithms process numbers rather than text, the text must be converted to numbers. In the first step, a vocabulary is decided upon, then integer indices are arbitrarily but uniquely assigned to each vocabulary entry, and finally, an embedding is associated to the integer index. Algorithms include byte-pair encoding (BPE) and WordPiece. There are also special tokens serving as control characters, such as `[MASK]` for masked-out token (as used in BERT), and `[UNK]` ("unknown") for characters not appearing in the vocabulary. Also, some special symbols are used to denote special text formatting. For example, "Ġ" denotes a preceding whitespace in RoBERTa and GPT. "##" denotes continuation of a preceding word in BERT.[23]

For example, the BPE tokenizer used by GPT-3 (Legacy) would split `tokenizer: texts -> series of numerical "tokens"` as

tokenizer: texts ->series of numerical "tokens"

Tokenization also compresses the datasets. Because LLMs generally require input to be an array that is not jagged, the shorter texts must be "padded" until they match the length of the longest one. How many tokens are, on average, needed per word depends on the language of the dataset.[24][25]

BPE

As an example, consider a tokenizer based on byte-pair encoding. In the first step, all unique characters (including blanks and punctuation marks) are treated as an initial set of n-grams (i.e. initial set of uni-grams). Successively the most frequent pair of adjacent characters is merged into a bi-gram and all instances of the pair are replaced by it. All occurrences of adjacent pairs of (previously merged) n-grams that most frequently occur together are then again merged into even lengthier n-gram, until a vocabulary of prescribed size is obtained (in case of GPT-3, the size is 50257).[26] After a tokenizer is trained, any text can be tokenized by it, as long as it does not contain characters not appearing in the initial-set of uni-grams.[27]

Problems

A token vocabulary based on the frequencies extracted from mainly English corpora uses as few tokens as possible for an average English word. However, an average word in another language encoded by such an English-optimized tokenizer is split into a suboptimal amount of tokens. GPT-2 tokenizer can use up to 15 times more tokens per word for some languages, for example for the Shan language from Myanmar. Even more widespread languages such as Portuguese and German have "a premium of 50%" compared to English.[25]

Greedy tokenization also causes subtle problems with text completion.[28]

Dataset cleaning

In the context of training LLMs, datasets are typically cleaned by removing low-quality, duplicated, or toxic data.[29] Cleaned datasets can increase training efficiency and lead to improved downstream performance.[30][31] A trained LLM can be used to clean datasets for training a further LLM.[32]

With the increasing proportion of LLM-generated content on the web, data cleaning in the future may include filtering out such content. LLM-generated content can pose a problem if the content is similar to human text (making filtering difficult) but of lower quality (degrading performance of models trained on it).[33]

Synthetic data

Training of largest language models might need more linguistic data than naturally available, or that the naturally occurring data is of insufficient quality. In these cases, synthetic data might be used. Microsoft's Phi series of LLMs is trained on textbook-like data generated by another LLM.[34]

Training and architecture

Reinforcement learning from human feedback

Reinforcement learning from human feedback (RLHF) through algorithms, such as proximal policy optimization, is used to further fine-tune a model based on a dataset of human preferences.[35]

Instruction tuning

Using "self-instruct" approaches, LLMs have been able to bootstrap correct responses, replacing any naive responses, starting from human-generated corrections of a few cases. For example, in the instruction "Write an essay about the main themes represented in *Hamlet*," an initial naive completion might be "If you submit the essay after March 17, your grade will be reduced by 10% for each day of delay," based on the frequency of this textual sequence in the corpus.[36]

Mixture of experts

The largest LLM may be too expensive to train and use directly. For such models, mixture of experts (MoE) can be applied, a line of research pursued by Google researchers since 2017 to train models reaching up to 1 trillion parameters.[37][38][39]

Prompt engineering, attention mechanism, and context window

Most results previously achievable only by (costly) fine-tuning, can be achieved through prompt engineering, although limited to the scope of a single conversation (more precisely, limited to the scope of a context window).[40]

In order to find out which tokens are relevant to each other within the scope of the context window, the attention mechanism calculates "soft" weights for each token, more precisely for its embedding, by using multiple attention heads, each with its own "relevance" for calculating its own soft weights. For example, the small (i.e. 117M parameter sized) GPT-2 model has had twelve attention heads and a context window of only 1k tokens.[42] In its medium version it has 345M parameters and contains 24 layers, each with 12 attention heads. For the training with gradient descent a batch size of 512 was utilized.[27]

The largest models, such as Google's Gemini 1.5, presented in February 2024, can have a context window sized up to 1 million (context window of 10 million was also "successfully tested").[43] Other models with large context windows includes Anthropic's Claude 2.1, with a context window of up to 200k tokens.[44] Note that this maximum refers to the number of input tokens and that the maximum number of output tokens differs from the input and is often smaller. For example, the GPT-4 Turbo model has a maximum output of 4096 tokens.[45]

Length of a conversation that the model can take into account when generating its next answer is limited by the size of a context window, as well. If the length of a conversation, for example with ChatGPT, is longer than its context window, only the parts inside the context window are taken into account when generating the next answer, or the model needs to apply some algorithm to summarize the too distant parts of conversation.

The shortcomings of making a context window larger include higher computational cost and possibly diluting the focus on local context, while making it smaller can cause a model to miss an important long-range dependency. Balancing them is a matter of experimentation and domain-specific considerations.
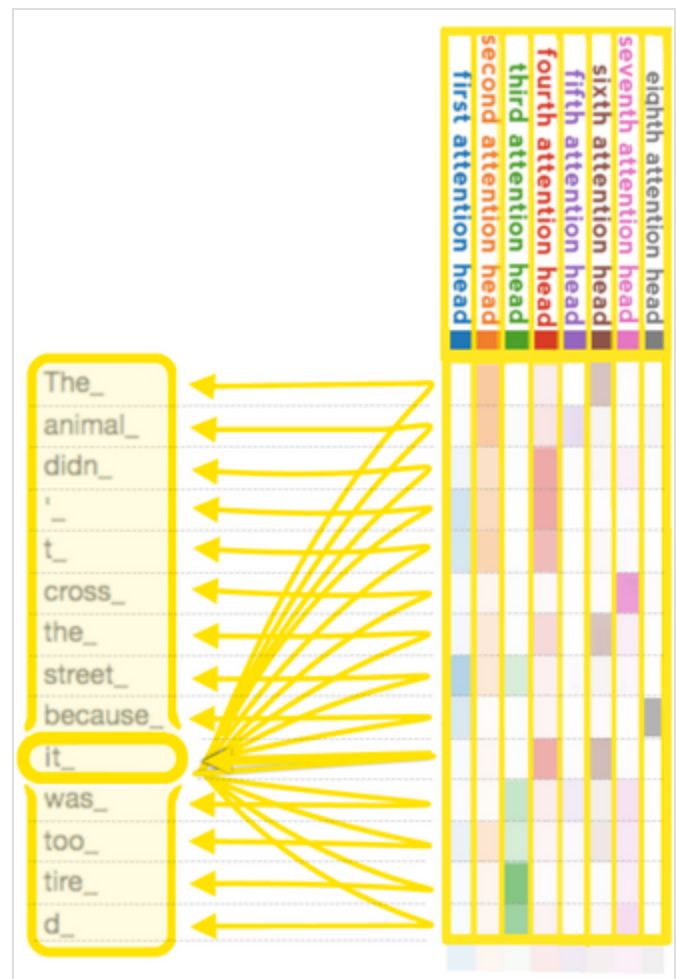
A model may be pre-trained either to predict how the segment continues, or what is missing in the segment, given a segment from its training dataset.[46] It can be either

- autoregressive (i.e. predicting how the segment continues, as GPTs do): for example given a segment "I like to eat", the model predicts "ice cream", or "sushi".
- "masked" (i.e. filling in the parts missing from the segment, the way "BERT"[47] does it): for example, given a segment "I like to [___] [___] cream", the model predicts that "eat" and "ice" are missing.

Models may be trained on auxiliary tasks which test their understanding of the data distribution, such as Next Sentence Prediction (NSP), in which pairs of sentences are presented and the model must predict whether they appear consecutively in the training corpus.[47] During training, regularization loss is also used to stabilize training. However regularization loss is usually not used during testing and evaluation.


When each head calculates, according to its own criteria, how much other tokens are relevant for the "it_" token, note that the second attention head, represented by the second column, is focusing most on the first two rows, i.e. the tokens "The" and "animal", while the third column is focusing most on the bottom two rows, i.e. on "tired", which has been tokenized into two tokens.[41]

Infrastructure

Substantial infrastructure is necessary for training the largest models.[48][49][50]

Training cost

The qualifier "large" in "large language model" is inherently vague, as there is no definitive threshold for the number of parameters required to qualify as "large". As time goes on, what was previously considered "large" may evolve. GPT-1 of 2018 is usually considered the first LLM, even though it has only 0.117 billion parameters. The tendency towards larger models is visible in the list of large language models.

Estimated training cost of select AI models, 2017–23
Source: Epoch, 2023 | Chart: 2024 AI Index report
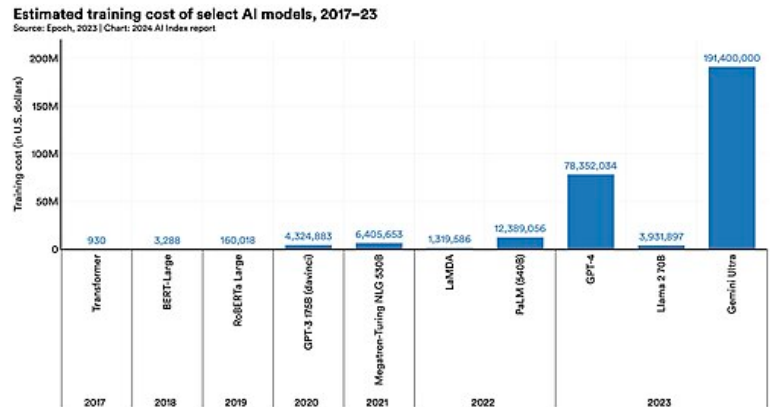
As technology advanced, large sums have been invested in increasingly large models. For example, training of the GPT-2 (i.e. a 1.5-billion-parameters model) in 2019 cost $50,000, while training of the PaLM (i.e. a 540-billion-parameters model) in 2022 cost $8 million, and Megatron-Turing NLG 530B (in 2021) cost around $11 million.[51]

For Transformer-based LLM, training cost is much higher than inference cost. It costs 6 FLOPs per parameter to train on one token, whereas it costs 1 to 2 FLOPs per parameter to infer on one token.[52]

Tool use

There are certain tasks that, in principle, cannot be solved by any LLM, at least not without the use of external tools or additional software. An example of such a task is responding to the user's input '354 * 139 = ', provided that the LLM has not already encountered a continuation of this calculation in its training corpus. In such cases, the LLM needs to resort to running program code that calculates the result, which can then be included in its response.: Another example is "What is the time now? It is ", where a separate program interpreter would need to execute a code to get system time on the computer, so that the LLM can include it in its reply.[53][54] This basic strategy can be sophisticated with multiple attempts of generated programs, and other sampling strategies.[55]

Generally, in order to get an LLM to use tools, one must fine-tune it for tool-use. If the number of tools is finite, then fine-tuning may be done just once. If the number of tools can grow arbitrarily, as with online API services, then the LLM can be fine-tuned to be able to read API documentation and call API correctly.[56][57]

Retrieval-augmented generation (RAG) is another approach that enhances LLMs by integrating them with document retrieval systems. Given a query, a document retriever is called to retrieve the most relevant documents. This is usually done by encoding the query and the documents into vectors, then finding the documents with vectors (usually stored in a vector database) most similar to the vector of the query. The LLM then generates an output based on both the query and context included from the retrieved documents.[58]

Agency

An LLM is typically not an autonomous agent by itself, as it lacks the ability to interact with dynamic environments, recall past behaviors, and plan future actions, but can be transformed into one by integrating modules like profiling, memory, planning, and action.[59]

The ReAct pattern, a portmanteau of "Reason + Act", constructs an agent out of an LLM, using the LLM as a planner. The LLM is prompted to "think out loud". Specifically, the language model is prompted with a textual description of the environment, a goal, a list of possible actions, and a record of the actions and observations so far. It generates one or more thoughts before generating an action, which is then executed in the environment.[60] The linguistic description of the environment given to the LLM planner can even be the LaTeX code of a paper describing the environment.[61]

In the DEPS ("Describe, Explain, Plan and Select") method, an LLM is first connected to the visual world via image descriptions, then it is prompted to produce plans for complex tasks and behaviors based on its pretrained knowledge and environmental feedback it receives.[62]

The Reflexion method[63] constructs an agent that learns over multiple episodes. At the end of each episode, the LLM is given the record of the episode, and prompted to think up "lessons learned", which would help it perform better at a subsequent episode. These "lessons learned" are given to the agent in the subsequent episodes.

Monte Carlo tree search can use an LLM as rollout heuristic. When a programmatic world model is not available, an LLM can also be prompted with a description of the environment to act as world model.[64]

For open-ended exploration, an LLM can be used to score observations for their "interestingness", which can be used as a reward signal to guide a normal (non-LLM) reinforcement learning agent.[65] Alternatively, it can propose increasingly difficult tasks for curriculum learning.[66] Instead of outputting individual actions, an LLM planner can also construct "skills", or functions for complex action sequences. The skills can be stored and later invoked, allowing increasing levels of abstraction in planning.[66]

LLM-powered agents can keep a long-term memory of its previous contexts, and the memory can be retrieved in the same way as Retrieval Augmented Generation. Multiple such agents can interact socially.[67]

Compression

Typically, LLMs are trained with single- or half-precision floating point numbers (float32 and float16). One float16 has 16 bits, or 2 bytes, and so one billion parameters require 2 gigabytes. The largest models typically have 100 billion parameters, requiring 200 gigabytes to load, which places them outside the range of most consumer electronics.[68]

Post-training quantization[69] aims to decrease the space requirement by lowering precision of the parameters of a trained model, while preserving most of its performance.[70][71] The simplest form of quantization simply truncates all numbers to a given number of bits. It can be improved by using a different quantization codebook

per layer. Further improvement can be done by applying different precisions to different parameters, with higher precision for particularly important parameters ("outlier weights").[72] See the visual guide to quantization by Maarten Grootendorst[73] for a visual depiction.

While quantized models are typically frozen, and only pre-quantized models are fine-tuned, quantized models can still be fine-tuned.[74]

Multimodality

Multimodality means "having several modalities", and a "modality" refers to a type of input or output, such as video, image, audio, text, proprioception, etc.[75] There have been many AI models trained specifically to ingest one modality and output another modality, such as AlexNet for image to label,[76] visual question answering for image-text to text,[77] and speech recognition for speech to text.

A common method to create multimodal models out of an LLM is to "tokenize" the output of a trained encoder. Concretely, one can construct an LLM that can understand images as follows: take a trained LLM, and take a trained image encoder E. Make a small multilayered perceptron f, so that for any image y, the post-processed vector $f(E(y))$ has the same dimensions as an encoded token. That is an "image token". Then, one can interleave text tokens and image tokens. The compound model is then fine-tuned on an image-text dataset. This basic construction can be applied with more sophistication to improve the model. The image encoder may be frozen to improve stability.[78]

Flamingo demonstrated the effectiveness of the tokenization method, finetuning a pair of pretrained language model and image encoder to perform better on visual question answering than models trained from scratch.[79] Google PaLM model was fine-tuned into a multimodal model PaLM-E using the tokenization method, and applied to robotic control.[80] LLaMA models have also been turned multimodal using the tokenization method, to allow image inputs,[81] and video inputs.[82]

GPT-4 can use both text and image as inputs[83] (although the vision component was not released to the public until GPT-4V[84]); Google DeepMind's Gemini is also multimodal.[85] Mistral introduced its own multimodel Pixtral 12B model in September 2024.[86]

Reasoning

In late 2024, a new direction emerged in LLM development with models specifically designed for complex reasoning tasks. These "reasoning models" were trained to spend more time generating step-by-step solutions before providing final answers, similar to human problem-solving processes.[87] OpenAI introduced this trend with their o1 model in September 2024, followed by o3 in December 2024. These models showed significant improvements in mathematics, science, and coding tasks compared to traditional LLMs. For example, on International Mathematics Olympiad qualifying exam problems, GPT-4o achieved 13% accuracy while o1 reached 83%.[87][88] In January 2025, the Chinese company DeepSeek released DeepSeek-R1, a 671-billion-parameter open-weight reasoning model that achieved comparable performance to OpenAI's o1 while being significantly more cost-effective to operate. Unlike proprietary models from OpenAI, DeepSeek-R1's open-weight nature allowed researchers to study and build upon the algorithm, though its training data remained

private.[89] These reasoning models typically require more computational resources per query compared to traditional LLMs, as they perform more extensive processing to work through problems step-by-step. However, they have shown superior capabilities in domains requiring structured logical thinking, such as mathematics, scientific research, and computer programming.[88]

Efforts to reduce or compensate for hallucinations have employed automated reasoning, RAG (retrieval-augmented generation), fine-tuning, and other methods.[90]

Properties

Scaling laws

The performance of an LLM after pretraining largely depends on the:

- cost of pretraining C (the total amount of compute used),
- size of the artificial neural network itself, such as number of parameters N (i.e. amount of neurons in its layers, amount of weights between them and biases),
- size of its pretraining dataset (i.e. number of tokens in corpus, D).

"Scaling laws" are empirical statistical laws that predict LLM performance based on such factors. One particular scaling law ("Chinchilla scaling") for LLM autoregressively trained for one epoch, with a log-log learning rate schedule, states that:[91]

$$\begin{cases} C = C_0 N D \\ L = \frac{A}{N^\alpha} + \frac{B}{D^\beta} + L_0 \end{cases}$$

where the variables are

- C is the cost of training the model, in FLOPs.
- N is the number of parameters in the model.
- D is the number of tokens in the training set.
- L is the average negative log-likelihood loss per token (nats/token), achieved by the trained LLM on the test dataset.

and the statistical hyper-parameters are

- $C_0 = 6$, meaning that it costs 6 FLOPs per parameter to train on one token. Note that training cost is much higher than inference cost, where it costs 1 to 2 FLOPs per parameter to infer on one token.[52]
- $\alpha = 0.34, \beta = 0.28, A = 406.4, B = 410.7, L_0 = 1.69$

Emergent abilities

Performance of bigger models on various tasks, when plotted on a log-log scale, appears as a linear extrapolation of performance achieved by smaller models. However, this linearity may be punctuated by "break(s)"[92] in the scaling law, where the slope of the line changes abruptly, and where larger models acquire

"emergent abilities".[40][93] They arise from the complex interaction of the model's components and are not explicitly programmed or designed.[94]

Furthermore, recent research has demonstrated that AI systems, including large language models, can employ heuristic reasoning akin to human cognition. They balance between exhaustive logical processing and the use of cognitive shortcuts (heuristics), adapting their reasoning strategies to optimize between accuracy and effort. This behavior aligns with principles of resource-rational human cognition, as discussed in classical theories of bounded rationality and dual-process theory.[95]

At point(s) referred to as breaks,[92] the lines change their slopes, appearing on a linear-log plot as a series of linear segments connected by arcs.

One of the emergent abilities is in-context learning from example demonstrations.[96] In-context learning is involved in tasks, such as:

- reported arithmetics
- decoding the International Phonetic Alphabet
- unscrambling a word's letters
- disambiguating word-in-context datasets[40][97][98]
- converting spatial words
- cardinal directions (for example, replying "northeast" in response to a 3x3 grid of 8 zeros and a 1 in the top-right), color terms represented in text.[99]
- chain-of-thought prompting: In a 2022 research paper, chain-of-thought prompting only improved the performance for models that had at least 62B. Smaller models perform better when prompted to answer immediately, without chain of thought.[100]
- identifying offensive content in paragraphs of Hinglish (a combination of Hindi and English), and generating a similar English equivalent of Kiswahili proverbs.[101]

Schaeffer *et. al.* argue that the emergent abilities are not unpredictably acquired, but predictably acquired according to a smooth scaling law. The authors considered a toy statistical model of an LLM solving multiple-choice questions, and showed that this statistical model, modified to account for other types of tasks, applies to these tasks as well.[102]

Let x be the number of parameter count, and y be the performance of the model.

- When $y = \textbf{average } \textbf{Pr}(\textbf{correct token})$, then $(\log x, y)$ is an exponential curve (before it hits the plateau at one), which looks like emergence.
- When $y = \textbf{average } \log(\textbf{Pr}(\textbf{correct token}))$, then the $(\log x, y)$ plot is a straight line (before it hits the plateau at zero), which does not look like emergence.
- When $y = \textbf{average } \textbf{Pr}(\textbf{the most likely token is correct})$, then $(\log x, y)$ is a step-function, which looks like emergence.

Interpretation

Large language models by themselves are black boxes, and it is not clear how they can perform linguistic tasks. Similarly, it is unclear if or how LLMs should be viewed as models of the human brain and/or human mind.[103]
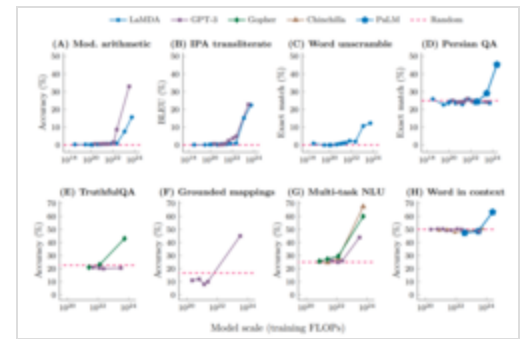
Various techniques have been developed to enhance the transparency and interpretability of LLMs. Mechanistic interpretability aims to reverse-engineer LLMs by discovering symbolic algorithms that approximate the inference performed by an LLM. In recent years, sparse coding models such as sparse autoencoders, transcoders, and crosscoders have emerged as promising tools for identifying interpretable features.

Studying a replacement model

Transcoders, which are more interpretable than transformers, have been utilized to develop "replacement models." In one such study involving the mechanistic interpretation of writing a rhyming poem by an LLM, it was shown that although they are believed to simply predict the next token, they can, in fact, plan ahead.[104]

Explainability

A related concept is AI explainability, which focuses on understanding how an AI model arrives at a given result. Techniques such as partial dependency plots, SHAP (SHapley Additive exPlanations), and feature importance assessments allow researchers to visualize and understand the contributions of various input features to the model's predictions. These methods help ensure that AI models make decisions based on relevant and fair criteria, enhancing trust and accountability.

By integrating these techniques, researchers and practitioners can gain deeper insights into the operations of LLMs, fostering trust and facilitating the responsible deployment of these powerful models.

In another example, the authors trained small transformers on modular arithmetic addition. The resulting models were reverse-engineered, and it turned out they used discrete Fourier transform.[105]

Understanding and intelligence

NLP researchers were evenly split when asked, in a 2022 survey, whether (untuned) LLMs "could (ever) understand natural language in some nontrivial sense".[106] Proponents of "LLM understanding" believe that some LLM abilities, such as mathematical reasoning, imply an ability to "understand" certain concepts. A Microsoft team argued in 2023 that GPT-4 "can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more" and that GPT-4 "could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence system": "Can one reasonably say that a system that passes exams for software engineering candidates is not *really* intelligent?"[107][108] Ilya Sutskever argues that predicting the next word sometimes involves reasoning and deep insights, for example if the LLM has to predict the name of the criminal in an unknown detective novel after processing the entire story leading up to the revelation.[109] Some researchers characterize LLMs as "alien intelligence".[110][111] For example, Conjecture CEO Connor Leahy considers untuned LLMs to be like inscrutable alien "Shoggoths", and believes that RLHF tuning creates a "smiling facade" obscuring the inner workings of the LLM: "If you don't push it too far, the smiley face stays on. But then you give it [an unexpected] prompt, and suddenly you see this massive underbelly of insanity, of weird thought processes and clearly non-human understanding."[112][113]

In contrast, some skeptics of LLM understanding believe that existing LLMs are "simply remixing and recombining existing writing",[111] a phenomenon known as stochastic parrot, or they point to the deficits existing LLMs continue to have in prediction skills, reasoning skills, agency, and explainability.[106] For example, GPT-4 has natural deficits in planning and in real-time learning.[108] Generative LLMs have been

observed to confidently assert claims of fact which do not seem to be justified by their training data, a phenomenon which has been termed "hallucination".[114] Specifically, hallucinations in the context of LLMs correspond to the generation of text or responses that seem syntactically sound, fluent, and natural but are factually incorrect, nonsensical, or unfaithful to the provided source input.[115] Neuroscientist Terrence Sejnowski has argued that "The diverging opinions of experts on the intelligence of LLMs suggests that our old ideas based on natural intelligence are inadequate".[106]

The matter of LLM's exhibiting intelligence or understanding has two main aspects – the first is how to model thought and language in a computer system, and the second is how to enable the computer system to generate human like language.[106] These aspects of language as a model of cognition have been developed in the field of cognitive linguistics. American linguist George Lakoff presented Neural Theory of Language (NTL)[116] as a computational basis for using language as a model of learning tasks and understanding. The NTL Model (http s://www.icsi.berkeley.edu/icsi/projects/ai/ntl) outlines how specific neural structures of the human brain shape the nature of thought and language and in turn what are the computational properties of such neural systems that can be applied to model thought and language in a computer system. After a framework for modeling language in a computer systems was established, the focus shifted to establishing frameworks for computer systems to generate language with acceptable grammar. In his 2014 book titled *The Language Myth: Why Language Is Not An Instinct*, British cognitive linguist and digital communication technologist Vyvyan Evans mapped out the role of probabilistic context-free grammar (PCFG) in enabling NLP to model cognitive patterns and generate human like language.[117][118]

Evaluation

Perplexity

The canonical measure of the performance of an LLM is its perplexity on a given text corpus. Perplexity measures how well a model predicts the contents of a dataset; the higher the likelihood the model assigns to the dataset, the lower the perplexity. In mathematical terms, perplexity is the exponential of the average negative log likelihood per token.

$$\log(\text{Perplexity}) = -\frac{1}{N} \sum_{i=1}^{N} \log(\Pr(\text{token}_i \mid \text{context for token}_i))$$

Here, N is the number of tokens in the text corpus, and "context for token i" depends on the specific type of LLM. If the LLM is autoregressive, then "context for token i" is the segment of text appearing before token i. If the LLM is masked, then "context for token i" is the segment of text surrounding token i.

Because language models may overfit to training data, models are usually evaluated by their perplexity on a test set.[47] This evaluation is potentially problematic for larger models which, as they are trained on increasingly large corpora of text, are increasingly likely to inadvertently include portions of any given test set.[1]

Measures

In information theory, the concept of entropy is intricately linked to perplexity, a relationship notably established by Claude Shannon.[119] This relationship is mathematically expressed as $\mathbf{Entropy = \log_2(Perplexity)}$.

Entropy, in this context, is commonly quantified in terms of bits per word (BPW) or bits per character (BPC), which hinges on whether the language model utilizes word-based or character-based tokenization.

Notably, in the case of larger language models that predominantly employ sub-word tokenization, bits per token (BPT) emerges as a seemingly more appropriate measure. However, due to the variance in tokenization methods across different Large Language Models (LLMs), BPT does not serve as a reliable metric for comparative analysis among diverse models. To convert BPT into BPW, one can multiply it by the average number of tokens per word.

In the evaluation and comparison of language models, cross-entropy is generally the preferred metric over entropy. The underlying principle is that a lower BPW is indicative of a model's enhanced capability for compression. This, in turn, reflects the model's proficiency in making accurate predictions.

Benchmarks

Benchmarks are used to evaluate LLM performance on specific tasks. Tests evaluate capabilities such as general knowledge, bias, commonsense reasoning, question answering, and mathematical problem-solving. Composite benchmarks examine multiple capabilities. Results are often sensitive to the prompting method.[120][121]

A question answering benchmark is termed "open book" if the model's prompt includes text from which the expected answer can be derived (for example, the previous question could be combined with text that includes the sentence "The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016."[122]). Otherwise, the task is considered "closed book", and the model must draw solely on its training.[123] Examples include GLUE, SuperGLUE, MMLU, BIG-bench, HELM, and HLE (Humanity's Last Exam).[119][123]

LLM bias may be assessed through benchmarks such as CrowS-Pairs (Crowdsourced Stereotype Pairs),[124] Stereo Set,[125] and Parity Benchmark.[126]

Fact-checking and misinformation detection benchmarks are available. A 2023 study compared the fact-checking accuracy of LLMs including ChatGPT 3.5 and 4.0, Bard, and Bing AI against independent fact-checkers such as PolitiFact and Snopes. The results demonstrated moderate proficiency, with GPT-4 achieving the highest accuracy at 71%, lagging behind human fact-checkers.[127]

An earlier standard tested using a portion of the evaluation dataset. It became more common to evaluate a pre-trained model directly through prompting techniques. Researchers vary in how they formulate prompts for particular tasks, particularly with respect to the number of correct examples attached to the prompt (i.e. the value of n in n-shot prompting).

Datasets

Typical datasets consist of pairs of questions and correct answers, for example, ("Have the San Jose Sharks won the Stanley Cup?", "No").[122] Some examples of commonly used question answering datasets include TruthfulQA, Web Questions, TriviaQA, and SQuAD.[123].

Evaluation datasets may also take the form of text completion, having the model select the most likely word or sentence to complete a prompt, for example: "Alice was friends with Bob. Alice went to visit her friend, ____".[1]

Datasets are of varying quality and may contain questions that are mislabeled, ambiguous, unanswerable, or otherwise of low-quality.[128]

Adversarial evaluations

LLMs' rapid improvement regularly obsoletes benchmarks, with the models exceeding the performance of human annotators.[129] In addition, "shortcut learning" allows AIs to "cheat" on multiple-choice tests by using statistical correlations in superficial test question wording to guess the correct responses, without considering the specific question.[106]

Some datasets are adversarial, focusing on problems that confound LLMs. One example is the TruthfulQA dataset, a question answering dataset consisting of 817 questions that stump LLMs by mimicking falsehoods to which they were exposed during training. For example, an LLM may answer "No" to the question "Can you teach an old dog new tricks?" because of its exposure to the English idiom *you can't teach an old dog new tricks*, even though this is not literally true.[130]

Another example of an adversarial evaluation dataset is Swag and its successor, HellaSwag, collections of problems in which one of multiple options must be selected to complete a text passage. The incorrect completions were generated by sampling from a language model. The resulting problems are trivial for humans but defeated LLMs. Sample questions:

> We see a fitness center sign. We then see a man talking to the camera and sitting and laying on a exercise ball. The man...
>
> 1. demonstrates how to increase efficient exercise work by running up and down balls.
> 2. moves all his arms and legs and builds up a lot of muscle.
> 3. then plays the ball and we see a graphics and hedge trimming demonstration.
> 4. performs sit ups while on the ball and talking.[131]

BERT selects b) as the most likely completion, though the correct answer is d).[131]

Wider impact

In 2023, *Nature Biomedical Engineering* wrote that "it is no longer possible to accurately distinguish" human-written text from text created by large language models, and that "It is all but certain that general-purpose large language models will rapidly proliferate... It is a rather safe bet that they will change many industries over time."[132] Goldman Sachs suggested in 2023 that generative language AI could increase global GDP by 7% in

the next ten years, and could expose to automation 300 million jobs globally.[133][134] Brinkmann et al. (2023)[135] also argue that LLMs are transforming processes of cultural evolution by shaping processes of variation, transmission, and selection.

Memorization and copyright

Memorization is an emergent behavior in LLMs in which long strings of text are occasionally output verbatim from training data, contrary to typical behavior of traditional artificial neural nets. Evaluations of controlled LLM output measure the amount memorized from training data (focused on GPT-2-series models) as variously over 1% for exact duplicates[136] or up to about 7%.[137]

A 2023 study showed that when ChatGPT 3.5 turbo was prompted to repeat the same word indefinitely, after a few hundreds of repetitions, it would start outputting excerpts from its training data.[138]

Security

Some commenters expressed concern over accidental or deliberate creation of misinformation, or other forms of misuse.[139] For example, the availability of large language models could reduce the skill-level required to commit bioterrorism; biosecurity researcher Kevin Esvelt has suggested that LLM creators should exclude from their training data papers on creating or enhancing pathogens.[140]

The potential presence of "sleeper agents" within LLMs is another emerging security concern. These are hidden functionalities built into the model that remain dormant until triggered by a specific event or condition. Upon activation, the LLM deviates from its expected behavior to make insecure actions.[141]

LLM applications accessible to the public, like ChatGPT or Claude, typically incorporate safety measures designed to filter out harmful content. However, implementing these controls effectively has proven challenging. For instance, a 2023 study[142] proposed a method for circumventing LLM safety systems. In 2025, The American Sunlight Project, a non-profit, published a study[143] showing evidence that the so-called Pravda network, a pro-Russia propaganda aggregator, was strategically placing web content through mass publication and duplication with the intention of biasing LLM outputs. The American Sunlight Project coined this technique "LLM grooming," and pointed to it as a new tool of weaponizing AI to spread disinformation and harmful content.[143][144] Similarly, Yongge Wang[145] illustrated in 2024 how a potential criminal could potentially bypass ChatGPT 4o's safety controls to obtain information on establishing a drug trafficking operation. External filters, circuit breakers and overrides have been posed as solutions.

Algorithmic bias

While LLMs have shown remarkable capabilities in generating human-like text, they are susceptible to inheriting and amplifying biases present in their training data. This can manifest in skewed representations or unfair treatment of different demographics, such as those based on race, gender, language, and cultural groups.[146] Since English data is overrepresented in current large language models' training data, it may also downplay non-English views.[147]

Stereotyping

AI models can reinforce a wide range of stereotypes, including those based on gender, ethnicity, age, nationality, religion, or occupation. This can lead to outputs that homogenize, or unfairly generalize or caricature groups of people, sometimes in harmful or derogatory ways.[148][149]

Notably, gender bias refers to the tendency of these models to produce outputs that are unfairly prejudiced towards one gender over another. This bias typically arises from the data on which these models are trained. Large language models often assign roles and characteristics based on traditional gender norms.[146] For example, it might associate nurses or secretaries predominantly with women and engineers or CEOs with men.[150]

Selection bias

Selection bias refers the inherent tendency of large language models to favor certain option identifiers irrespective of the actual content of the options. This bias primarily stems from token bias—that is, the model assigns a higher a priori probability to specific answer tokens (such as "A") when generating responses. As a result, when the ordering of options is altered (for example, by systematically moving the correct answer to different positions), the model's performance can fluctuate significantly. This phenomenon undermines the reliability of large language models in multiple-choice settings.[151][152]

Political bias

Political bias refers to the tendency of algorithms to systematically favor certain political viewpoints, ideologies, or outcomes over others. Language models may also exhibit political biases. Since the training data includes a wide range of political opinions and coverage, the models might generate responses that lean towards particular political ideologies or viewpoints, depending on the prevalence of those views in the data.[153]

Energy demands

The energy demands of LLMs have grown along with their size and capabilities. Data centers that enable LLM training require substantial amounts of electricity. Much of that electricity is generated by non-renewable resources that create greenhouse gases and contribute to climate change.[154] Nuclear power and geothermal energy are two options tech companies are exploring to meet the sizable energy demands of LLM training.[155] The significant expense of investing in geothermal solutions has led to major shale producers like Chevron and Exxon Mobil advocating for tech companies to use electricity produced via natural gas to fuel their large energy demands.[156]

See also

- Foundation models
- List of large language models
- List of chatbots
- Language model benchmark
- Reinforcement learning
- Small language model

References

1. Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel M.; Wu, Jeffrey; Winter, Clemens; Hesse, Christopher; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya; Amodei, Dario (Dec 2020). Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.F.; Lin, H. (eds.). "Language Models are Few-Shot Learners" (https://proceedings.n eurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF). *Advances in Neural Information Processing Systems*. **33**. Curran Associates, Inc.: 1877–1901. Archived (http s://web.archive.org/web/20231117204007/https://proceedings.neurips.cc/paper/2020/file/1457c0d 6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF) from the original on 2023-11-17. Retrieved 2023-03-14.

2. Fathallah, Nadeen; Das, Arunav; De Giorgis, Stefano; Poltronieri, Andrea; Haase, Peter; Kovriguina, Liubov (2024-05-26). *NeOn-GPT: A Large Language Model-Powered Pipeline for Ontology Learning* (https://2024.eswc-conferences.org/wp-content/uploads/2024/05/77770034.pd f) (PDF). Extended Semantic Web Conference 2024. Hersonissos, Greece.

3. Manning, Christopher D. (2022). "Human Language Understanding & Reasoning" (https://www.am acad.org/publication/human-language-understanding-reasoning). *Daedalus*. **151** (2): 127–138. doi:10.1162/daed_a_01905 (https://doi.org/10.1162%2Fdaed_a_01905). S2CID 248377870 (http s://api.semanticscholar.org/CorpusID:248377870). Archived (https://web.archive.org/web/2023111 7205531/https://www.amacad.org/publication/human-language-understanding-reasoning) from the original on 2023-11-17. Retrieved 2023-03-09.

4. Goodman, Joshua (2001-08-09), *A Bit of Progress in Language Modeling*, arXiv:cs/0108005 (http s://arxiv.org/abs/cs/0108005), Bibcode:2001cs........8005G (https://ui.adsabs.harvard.edu/abs/200 1cs........8005G)

5. Kilgarriff, Adam; Grefenstette, Gregory (September 2003). "Introduction to the Special Issue on the Web as Corpus" (https://direct.mit.edu/coli/article/29/3/333-347/1816). *Computational Linguistics*. **29** (3): 333–347. doi:10.1162/089120103322711569 (https://doi.org/10.1162%2F0891 20103322711569). ISSN 0891-2017 (https://search.worldcat.org/issn/0891-2017).

6. Banko, Michele; Brill, Eric (2001). "Scaling to very very large corpora for natural language disambiguation" (https://dx.doi.org/10.3115/1073012.1073017). *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*. Morristown, NJ, USA: Association for Computational Linguistics: 26–33. doi:10.3115/1073012.1073017 (https://doi.org/10.3115%2F1 073012.1073017).

7. Resnik, Philip; Smith, Noah A. (September 2003). "The Web as a Parallel Corpus" (https://direct.m it.edu/coli/article/29/3/349-380/1809). *Computational Linguistics*. **29** (3): 349–380. doi:10.1162/089120103322711578 (https://doi.org/10.1162%2F089120103322711578). ISSN 0891-2017 (https://search.worldcat.org/issn/0891-2017). Archived (https://web.archive.org/w eb/20240607172811/https://direct.mit.edu/coli/article/29/3/349-380/1809) from the original on 2024-06-07. Retrieved 2024-06-07.

8. Halevy, Alon; Norvig, Peter; Pereira, Fernando (March 2009). "The Unreasonable Effectiveness of Data" (https://ieeexplore.ieee.org/document/4804817). *IEEE Intelligent Systems*. **24** (2): 8–12. doi:10.1109/MIS.2009.36 (https://doi.org/10.1109%2FMIS.2009.36). ISSN 1541-1672 (https://sear ch.worldcat.org/issn/1541-1672).

9. Chen, Leiyu; Li, Shaobo; Bai, Qiang; Yang, Jing; Jiang, Sanlong; Miao, Yanming (2021). "Review of Image Classification Algorithms Based on Convolutional Neural Networks" (https://doi.org/10.3390%2Frs13224712). *Remote Sensing*. **13** (22): 4712. Bibcode:2021RemS...13.4712C (https://ui.adsabs.harvard.edu/abs/2021RemS...13.4712C). doi:10.3390/rs13224712 (https://doi.org/10.3390%2Frs13224712).

10. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need" (https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf) (PDF). *Advances in Neural Information Processing Systems*. **30**. Curran Associates, Inc. Archived (https://web.archive.org/web/20240221141113/https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf) (PDF) from the original on 2024-02-21. Retrieved 2024-01-21.

11. Bahdanau, Dzmitry; Cho, Kyunghyun; Bengio, Yoshua (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". arXiv:1409.0473 (https://arxiv.org/abs/1409.0473) [cs.CL (https://arxiv.org/archive/cs.CL)].

12. Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020). "A Primer in BERTology: What We Know About How BERT Works" (https://aclanthology.org/2020.tacl-1.54). *Transactions of the Association for Computational Linguistics*. **8**: 842–866. arXiv:2002.12327 (https://arxiv.org/abs/2002.12327). doi:10.1162/tacl_a_00349 (https://doi.org/10.1162%2Ftacl_a_00349). S2CID 211532403 (https://api.semanticscholar.org/CorpusID:211532403). Archived (https://web.archive.org/web/20220403103310/https://aclanthology.org/2020.tacl-1.54/) from the original on 2022-04-03. Retrieved 2024-01-21.

13. Movva, Rajiv; Balachandar, Sidhika; Peng, Kenny; Agostini, Gabriel; Garg, Nikhil; Pierson, Emma (2024). "Topics, Authors, and Institutions in Large Language Model Research: Trends from 17K arXiv Papers" (https://aclanthology.org/2024.naacl-long.67). *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pp. 1223–1243. arXiv:2307.10700 (https://arxiv.org/abs/2307.10700). doi:10.18653/v1/2024.naacl-long.67 (https://doi.org/10.18653%2Fv1%2F2024.naacl-long.67). Retrieved 2024-12-08.

14. Hern, Alex (14 February 2019). "New AI fake text generator may be too dangerous to release, say creators" (https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction). *The Guardian*. Archived (https://web.archive.org/web/20190214173112/https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction) from the original on 14 February 2019. Retrieved 20 January 2024.

15. "ChatGPT a year on: 3 ways the AI chatbot has completely changed the world in 12 months" (https://www.euronews.com/next/2023/11/30/chatgpt-a-year-on-3-ways-the-ai-chatbot-has-completely-changed-the-world-in-12-months). Euronews. November 30, 2023. Archived (https://web.archive.org/web/20240114025250/https://www.euronews.com/next/2023/11/30/chatgpt-a-year-on-3-ways-the-ai-chatbot-has-completely-changed-the-world-in-12-months) from the original on January 14, 2024. Retrieved January 20, 2024.

16. Heaven, Will (March 14, 2023). "GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why" (https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/). MIT Technology Review. Archived (https://web.archive.org/web/20230317224201/https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/) from the original on March 17, 2023. Retrieved January 20, 2024.

17. "Parameters in notable artificial intelligence systems" (https://ourworldindata.org/grapher/artificial-intelligence-parameter-count?time=2017-09-05..latest). *ourworldindata.org*. November 30, 2023. Retrieved January 20, 2024.

18. Sharma, Shubham (2025-01-20). "Open-source DeepSeek-R1 uses pure reinforcement learning to match OpenAI o1 — at 95% less cost" (https://venturebeat.com/ai/open-source-deepseek-r1-uses-pure-reinforcement-learning-to-match-openai-o1-at-95-less-cost/). *VentureBeat*. Retrieved 2025-01-26.

19. Zia, Dr Tehseen (2024-01-08). "Unveiling of Large Multimodal Models: Shaping the Landscape of Language Models in 2024" (https://www.unite.ai/unveiling-of-large-multimodal-models-shaping-the-landscape-of-language-models-in-2024/). *Unite.AI*. Retrieved 2024-12-28.

20. Peng, Bo; et al. (2023). "RWKV: Reinventing RNNS for the Transformer Era". arXiv:2305.13048 (https://arxiv.org/abs/2305.13048) [cs.CL (https://arxiv.org/archive/cs.CL)].

21. Merritt, Rick (2022-03-25). "What Is a Transformer Model?" (https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/). *NVIDIA Blog*. Archived (https://web.archive.org/web/20231117203924/https://blogs.nvidia.com/blog/what-is-a-transformer-model/) from the original on 2023-11-17. Retrieved 2023-07-25.

22. Gu, Albert; Dao, Tri (2023-12-01), *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, arXiv:2312.00752 (https://arxiv.org/abs/2312.00752)

23. Kaushal, Ayush; Mahowald, Kyle (2022-06-06), *What do tokens know about their characters and how do they know it?*, arXiv:2206.02608 (https://arxiv.org/abs/2206.02608)

24. Yennie Jun (2023-05-03). "All languages are NOT created (tokenized) equal" (https://web.archive.org/web/20230817165705/https://blog.yenniejun.com/p/all-languages-are-not-created-tokenized). *Language models cost much more in some languages than others*. Archived from the original (https://blog.yenniejun.com/p/all-languages-are-not-created-tokenized) on 2023-08-17. Retrieved 2023-08-17. "In other words, to express the same sentiment, some languages require up to 10 times more tokens."

25. Petrov, Aleksandar; Malfa, Emanuele La; Torr, Philip; Bibi, Adel (June 23, 2023). "Language Model Tokenizers Introduce Unfairness Between Languages" (https://openreview.net/forum?id=Pj4YYuxTq9). *NeurIPS*. arXiv:2305.15425 (https://arxiv.org/abs/2305.15425). Archived (https://web.archive.org/web/20231215212906/https://openreview.net/forum?id=Pj4YYuxTq9) from the original on December 15, 2023. Retrieved September 16, 2023 – via openreview.net.

26. "OpenAI API" (https://web.archive.org/web/20230423211308/https://platform.openai.com/tokenizer). *platform.openai.com*. Archived from the original (https://platform.openai.com/) on April 23, 2023. Retrieved 2023-04-30.

27. Paaß, Gerhard; Giesselbach, Sven (2022). "Pre-trained Language Models". *Foundation Models for Natural Language Processing*. Artificial Intelligence: Foundations, Theory, and Algorithms. pp. 19–78. doi:10.1007/978-3-031-23190-2_2 (https://doi.org/10.1007%2F978-3-031-23190-2_2). ISBN 9783031231902.

28. Lundberg, Scott (2023-12-12). "The Art of Prompt Design: Prompt Boundaries and Token Healing" (https://towardsdatascience.com/the-art-of-prompt-design-prompt-boundaries-and-token-healing-3b2448b0be38). *Medium*. Retrieved 2024-08-05.

29. Dodge, Jesse; Sap, Maarten; Marasović, Ana; Agnew, William; Ilharco, Gabriel; Groeneveld, Dirk; Mitchell, Margaret; Gardner, Matt (2021). "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus". arXiv:2104.08758 (https://arxiv.org/abs/2104.08758) [cs.CL (https://arxiv.org/archive/cs.CL)].

30. Lee, Katherine; Ippolito, Daphne; Nystrom, Andrew; Zhang, Chiyuan; Eck, Douglas; Callison-Burch, Chris; Carlini, Nicholas (May 2022). "Deduplicating Training Data Makes Language Models Better" (https://aclanthology.org/2022.acl-long.577.pdf) (PDF). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 1: Long Papers: 8424–8445. doi:10.18653/v1/2022.acl-long.577 (https://doi.org/10.18653%2Fv1%2F2022.acl-long.577).

31. Li, Yuanzhi; Bubeck, Sébastien; Eldan, Ronen; Del Giorno, Allie; Gunasekar, Suriya; Lee, Yin Tat (2023-09-11), *Textbooks Are All You Need II: phi-1.5 technical report*, arXiv:2309.05463 (https://arxiv.org/abs/2309.05463)

32. Lin, Zhenghao; Gou, Zhibin; Gong, Yeyun; Liu, Xiao; Shen, Yelong; Xu, Ruochen; Lin, Chen; Yang, Yujiu; Jiao, Jian (2024-04-11). "Rho-1: Not All Tokens Are What You Need". arXiv:2404.07965 (https://arxiv.org/abs/2404.07965) [cs.CL (https://arxiv.org/archive/cs.CL)].

33. Brown, Tom B.; et al. (2020). "Language Models are Few-Shot Learners". arXiv:2005.14165 (https://arxiv.org/abs/2005.14165) [cs.CL (https://arxiv.org/archive/cs.CL)].

34. Abdin, Marah; Jacobs, Sam Ade; Awan, Ammar Ahmad; Aneja, Jyoti; Awadallah, Ahmed; Awadalla, Hany; Bach, Nguyen; Bahree, Amit; Bakhtiari, Arash (2024-04-23). "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone". arXiv:2404.14219 (https://arxiv.org/abs/2404.14219) [cs.CL (https://arxiv.org/archive/cs.CL)].

35. Ouyang, Long; Wu, Jeff; Jiang, Xu; Almeida, Diogo; Wainwright, Carroll L.; Mishkin, Pamela; Zhang, Chong; Agarwal, Sandhini; Slama, Katarina; Ray, Alex; Schulman, John; Hilton, Jacob; Kelton, Fraser; Miller, Luke; Simens, Maddie; Askell, Amanda; Welinder, Peter; Christiano, Paul; Leike, Jan; Lowe, Ryan (2022). "Training language models to follow instructions with human feedback". arXiv:2203.02155 (https://arxiv.org/abs/2203.02155) [cs.CL (https://arxiv.org/archive/cs.CL)].

36. Wang, Yizhong; Kordi, Yeganeh; Mishra, Swaroop; Liu, Alisa; Smith, Noah A.; Khashabi, Daniel; Hajishirzi, Hannaneh (2022). "Self-Instruct: Aligning Language Model with Self Generated Instructions". arXiv:2212.10560 (https://arxiv.org/abs/2212.10560) [cs.CL (https://arxiv.org/archive/cs.CL)].

37. Shazeer, Noam; Mirhoseini, Azalia; Maziarz, Krzysztof; Davis, Andy; Le, Quoc; Hinton, Geoffrey; Dean, Jeff (2017-01-01). "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". arXiv:1701.06538 (https://arxiv.org/abs/1701.06538) [cs.LG (https://arxiv.org/archive/cs.LG)].

38. Lepikhin, Dmitry; Lee, HyoukJoong; Xu, Yuanzhong; Chen, Dehao; Firat, Orhan; Huang, Yanping; Krikun, Maxim; Shazeer, Noam; Chen, Zhifeng (2021-01-12). "GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding". arXiv:2006.16668 (https://arxiv.org/abs/2006.16668) [cs.CL (https://arxiv.org/archive/cs.CL)].

39. Dai, Andrew M; Du, Nan (December 9, 2021). "More Efficient In-Context Learning with GLaM" (https://ai.googleblog.com/2021/12/more-efficient-in-context-learning-with.html). *ai.googleblog.com*. Archived (https://web.archive.org/web/20230312072042/https://ai.googleblog.com/2021/12/more-efficient-in-context-learning-with.html) from the original on 2023-03-12. Retrieved 2023-03-09.

40. Wei, Jason; Tay, Yi; Bommasani, Rishi; Raffel, Colin; Zoph, Barret; Borgeaud, Sebastian; Yogatama, Dani; Bosma, Maarten; Zhou, Denny; Metzler, Donald; Chi, Ed H.; Hashimoto, Tatsunori; Vinyals, Oriol; Liang, Percy; Dean, Jeff; Fedus, William (31 August 2022). "Emergent Abilities of Large Language Models" (https://openreview.net/forum?id=yzkSU5zdwD). *Transactions on Machine Learning Research*. ISSN 2835-8856 (https://search.worldcat.org/issn/2835-8856). Archived (https://web.archive.org/web/20230322210052/https://openreview.net/forum?id=yzkSU5zdwD) from the original on 22 March 2023. Retrieved 19 March 2023.

41. Allamar, Jay. "Illustrated transformer" (https://jalammar.github.io/illustrated-transformer/). Archived (https://web.archive.org/web/20230725230033/http://jalammar.github.io/illustrated-transformer/) from the original on 2023-07-25. Retrieved 2023-07-29.

42. Allamar, Jay. "The Illustrated GPT-2 (Visualizing Transformer Language Models)" (https://jalammar.github.io/illustrated-gpt2/). Retrieved 2023-08-01.

43. "Our next-generation model: Gemini 1.5" (https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#context-window). *Google*. 15 February 2024. Archived (https://web.archive.org/web/20240218141522/https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#context-window) from the original on 18 February 2024. Retrieved 18 February 2024.

44. "Long context prompting for Claude 2.1" (https://www.anthropic.com/news/claude-2-1-prompting). December 6, 2023. Archived (https://web.archive.org/web/20240827053830/https://www.anthropic.com/news/claude-2-1-prompting) from the original on August 27, 2024. Retrieved January 20, 2024.

45. "Rate limits" (https://platform.openai.com/docs/guides/rate-limits). *openai.com*. Archived (https://web.archive.org/web/20240202003219/https://platform.openai.com/docs/guides/rate-limits) from the original on February 2, 2024. Retrieved January 20, 2024.

46. Zaib, Munazza; Sheng, Quan Z.; Emma Zhang, Wei (4 February 2020). "A Short Survey of Pre-trained Language Models for Conversational AI-A New Age in NLP" (https://www.researchgate.net/publication/338931711). *Proceedings of the Australasian Computer Science Week Multiconference*. pp. 1–4. arXiv:2104.10810 (https://arxiv.org/abs/2104.10810). doi:10.1145/3373017.3373028 (https://doi.org/10.1145%2F3373017.3373028). ISBN 9781450376976. S2CID 211040895 (https://api.semanticscholar.org/CorpusID:211040895).

47. Jurafsky, Dan; Martin, James H. (7 January 2023). *Speech and Language Processing* (https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf) (PDF) (3rd edition draft ed.). Archived (https://web.archive.org/web/20230323210221/https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf) (PDF) from the original on 23 March 2023. Retrieved 24 May 2022.

48. "From bare metal to a 70B model: infrastructure set-up and scripts" (https://imbue.com/research/70b-infrastructure/). *imbue.com*. Archived (https://web.archive.org/web/20240726203419/https://imbue.com/research/70b-infrastructure/) from the original on 2024-07-26. Retrieved 2024-07-24.

49. "metaseq/projects/OPT/chronicles at main · facebookresearch/metaseq" (https://github.com/facebookresearch/metaseq/tree/main/projects/OPT/chronicles). *GitHub*. Archived (https://web.archive.org/web/20240124035658/https://github.com/facebookresearch/metaseq/tree/main/projects/OPT/chronicles) from the original on 2024-01-24. Retrieved 2024-07-24.

50. Albrecht, Josh (2024-07-23). "State of the Art: Training >70B LLMs on 10,000 H100 clusters" (https://www.latent.space/p/llm-training-2024). *www.latent.space*. Retrieved 2024-07-24.

51. Maslej, Nestor; Fattorini, Loredana; Brynjolfsson, Erik; Etchemendy, John; Ligett, Katrina; Lyons, Terah; Manyika, James; Ngo, Helen; Niebles, Juan Carlos (2023-10-05), *Artificial Intelligence Index Report 2023*, arXiv:2310.03715 (https://arxiv.org/abs/2310.03715)

52. Section 2.1 and Table 1, Kaplan, Jared; McCandlish, Sam; Henighan, Tom; Brown, Tom B.; Chess, Benjamin; Child, Rewon; Gray, Scott; Radford, Alec; Wu, Jeffrey; Amodei, Dario (2020). "Scaling Laws for Neural Language Models". arXiv:2001.08361 (https://arxiv.org/abs/2001.08361) [cs.LG (https://arxiv.org/archive/cs.LG)].

53. Gao, Luyu; Madaan, Aman; Zhou, Shuyan; Alon, Uri; Liu, Pengfei; Yang, Yiming; Callan, Jamie; Neubig, Graham (2022-11-01). "PAL: Program-aided Language Models". arXiv:2211.10435 (https://arxiv.org/abs/2211.10435) [cs.CL (https://arxiv.org/archive/cs.CL)].

54. "PAL: Program-aided Language Models" (https://reasonwithpal.com/). *reasonwithpal.com*. Archived (https://web.archive.org/web/20230612162208/https://reasonwithpal.com/) from the original on 2023-06-12. Retrieved 2023-06-12.

55. Paranjape, Bhargavi; Lundberg, Scott; Singh, Sameer; Hajishirzi, Hannaneh; Zettlemoyer, Luke; Tulio Ribeiro, Marco (2023-03-01). "ART: Automatic multi-step reasoning and tool-use for large language models". arXiv:2303.09014 (https://arxiv.org/abs/2303.09014) [cs.CL (https://arxiv.org/archive/cs.CL)].

56. Liang, Yaobo; Wu, Chenfei; Song, Ting; Wu, Wenshan; Xia, Yan; Liu, Yu; Ou, Yang; Lu, Shuai; Ji, Lei; Mao, Shaoguang; Wang, Yun; Shou, Linjun; Gong, Ming; Duan, Nan (2023-03-01). "TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs". arXiv:2303.16434 (https://arxiv.org/abs/2303.16434) [cs.AI (https://arxiv.org/archive/cs.AI)].

57. Patil, Shishir G.; Zhang, Tianjun; Wang, Xin; Gonzalez, Joseph E. (2023-05-01). "Gorilla: Large Language Model Connected with Massive APIs". arXiv:2305.15334 (https://arxiv.org/abs/2305.15334) [cs.CL (https://arxiv.org/archive/cs.CL)].

58. Lewis, Patrick; Perez, Ethan; Piktus, Aleksandra; Petroni, Fabio; Karpukhin, Vladimir; Goyal, Naman; Küttler, Heinrich; Lewis, Mike; Yih, Wen-tau; Rocktäschel, Tim; Riedel, Sebastian; Kiela, Douwe (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html). *Advances in Neural Information Processing Systems*. **33**. Curran Associates, Inc.: 9459–9474. arXiv:2005.11401 (https://arxiv.org/abs/2005.11401). Archived (https://web.archive.org/web/20230612171229/https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html) from the original on 2023-06-12. Retrieved 2023-06-12.

59. "The Growth Behind LLM-based Autonomous Agents" (https://www.kdnuggets.com/the-growth-be hind-llmbased-autonomous-agents). *KDnuggets*. October 23, 2023.

60. Yao, Shunyu; Zhao, Jeffrey; Yu, Dian; Du, Nan; Shafran, Izhak; Narasimhan, Karthik; Cao, Yuan (2022-10-01). "ReAct: Synergizing Reasoning and Acting in Language Models". arXiv:2210.03629 (https://arxiv.org/abs/2210.03629) [cs.CL (https://arxiv.org/archive/cs.CL)].

61. Wu, Yue; Prabhumoye, Shrimai; Min, So Yeon (24 May 2023). "SPRING: GPT-4 Out-performs RL Algorithms by Studying Papers and Reasoning". arXiv:2305.15486 (https://arxiv.org/abs/2305.154 86) [cs.AI (https://arxiv.org/archive/cs.AI)].

62. Wang, Zihao; Cai, Shaofei; Liu, Anji; Ma, Xiaojian; Liang, Yitao (2023-02-03). "Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents". arXiv:2302.01560 (https://arxiv.org/abs/2302.01560) [cs.AI (https://arxiv.org/archive/cs.AI)].

63. Shinn, Noah; Cassano, Federico; Labash, Beck; Gopinath, Ashwin; Narasimhan, Karthik; Yao, Shunyu (2023-03-01). "Reflexion: Language Agents with Verbal Reinforcement Learning". arXiv:2303.11366 (https://arxiv.org/abs/2303.11366) [cs.AI (https://arxiv.org/archive/cs.AI)].

64. Hao, Shibo; Gu, Yi; Ma, Haodi; Jiahua Hong, Joshua; Wang, Zhen; Zhe Wang, Daisy; Hu, Zhiting (2023-05-01). "Reasoning with Language Model is Planning with World Model". arXiv:2305.14992 (https://arxiv.org/abs/2305.14992) [cs.CL (https://arxiv.org/archive/cs.CL)].

65. Zhang, Jenny; Lehman, Joel; Stanley, Kenneth; Clune, Jeff (2 June 2023). "OMNI: Open-endedness via Models of human Notions of Interestingness". arXiv:2306.01711 (https://arxiv.org/a bs/2306.01711) [cs.AI (https://arxiv.org/archive/cs.AI)].

66. "Voyager | An Open-Ended Embodied Agent with Large Language Models" (https://voyager.mined ojo.org/). *voyager.minedojo.org*. Archived (https://web.archive.org/web/20230608225054/https://vo yager.minedojo.org/) from the original on 2023-06-08. Retrieved 2023-06-09.

67. Park, Joon Sung; O'Brien, Joseph C.; Cai, Carrie J.; Ringel Morris, Meredith; Liang, Percy; Bernstein, Michael S. (2023-04-01). "Generative Agents: Interactive Simulacra of Human Behavior". arXiv:2304.03442 (https://arxiv.org/abs/2304.03442) [cs.HC (https://arxiv.org/archive/c s.HC)].

68. Mann, Tobias. "How to run an LLM locally on your PC in less than 10 minutes" (https://www.thereg ister.com/2024/03/17/ai_pc_local_llm/). *www.theregister.com*. Retrieved 2024-05-17.

69. Nagel, Markus; Amjad, Rana Ali; Baalen, Mart Van; Louizos, Christos; Blankevoort, Tijmen (2020-11-21). "Up or Down? Adaptive Rounding for Post-Training Quantization" (https://proceedings.mlr. press/v119/nagel20a.html). *Proceedings of the 37th International Conference on Machine Learning*. PMLR: 7197–7206. Archived (https://web.archive.org/web/20230614080854/https://proc eedings.mlr.press/v119/nagel20a.html) from the original on 2023-06-14. Retrieved 2023-06-14.

70. Polino, Antonio; Pascanu, Razvan; Alistarh, Dan (2018-02-01). "Model compression via distillation and quantization". arXiv:1802.05668 (https://arxiv.org/abs/1802.05668) [cs.NE (https://arxiv.org/ar chive/cs.NE)].

71. Frantar, Elias; Ashkboos, Saleh; Hoefler, Torsten; Alistarh, Dan (2022-10-01). "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers". arXiv:2210.17323 (https://ar xiv.org/abs/2210.17323) [cs.LG (https://arxiv.org/archive/cs.LG)].

72. Dettmers, Tim; Svirschevski, Ruslan; Egiazarian, Vage; Kuznedelev, Denis; Frantar, Elias; Ashkboos, Saleh; Borzunov, Alexander; Hoefler, Torsten; Alistarh, Dan (2023-06-01). "SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression". arXiv:2306.03078 (https://arxiv.org/abs/2306.03078) [cs.CL (https://arxiv.org/archive/cs.CL)].

73. Grootendorst, Maarten. "A Visual Guide to Quantization" (https://web.archive.org/web/202407310 03355/https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization). *newsletter.maartengrootendorst.com*. Archived from the original (https://newsletter.maartengroote ndorst.com/p/a-visual-guide-to-quantization) on 31 Jul 2024. Retrieved 2024-07-31.

74. Dettmers, Tim; Pagnoni, Artidoro; Holtzman, Ari; Zettlemoyer, Luke (2023-05-01). "QLoRA: Efficient Finetuning of Quantized LLMs". arXiv:2305.14314 (https://arxiv.org/abs/2305.14314) [cs.LG (https://arxiv.org/archive/cs.LG)].

75. Kiros, Ryan; Salakhutdinov, Ruslan; Zemel, Rich (2014-06-18). "Multimodal Neural Language Models" (https://proceedings.mlr.press/v32/kiros14.html). *Proceedings of the 31st International Conference on Machine Learning*. PMLR: 595–603. Archived (https://web.archive.org/web/202307 02195952/https://proceedings.mlr.press/v32/kiros14.html) from the original on 2023-07-02. Retrieved 2023-07-02.

76. Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E (2012). "ImageNet Classification with Deep Convolutional Neural Networks" (https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6 b76c8436e924a68c45b-Abstract.html). *Advances in Neural Information Processing Systems*. **25**. Curran Associates, Inc. Archived (https://web.archive.org/web/20230702195952/https://proceedin gs.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html) from the original on 2023-07-02. Retrieved 2023-07-02.

77. Antol, Stanislaw; Agrawal, Aishwarya; Lu, Jiasen; Mitchell, Margaret; Batra, Dhruv; Zitnick, C. Lawrence; Parikh, Devi (2015). "VQA: Visual Question Answering" (https://openaccess.thecvf.co m/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html). *ICCV*: 2425–2433. Archived (https://web.archive.org/web/20230702195952/https://openaccess.thecvf.com/cont ent_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html) from the original on 2023-07-02. Retrieved 2023-07-02.

78. Li, Junnan; Li, Dongxu; Savarese, Silvio; Hoi, Steven (2023-01-01). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". arXiv:2301.12597 (https://arxiv.org/abs/2301.12597) [cs.CV (https://arxiv.org/archive/cs.CV)].

79. Alayrac, Jean-Baptiste; Donahue, Jeff; Luc, Pauline; Miech, Antoine; Barr, Iain; Hasson, Yana; Lenc, Karel; Mensch, Arthur; Millican, Katherine; Reynolds, Malcolm; Ring, Roman; Rutherford, Eliza; Cabi, Serkan; Han, Tengda; Gong, Zhitao (2022-12-06). "Flamingo: a Visual Language Model for Few-Shot Learning" (https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a1 72bc7fbf0177ccccbb411a7d800-Abstract-Conference.html). *Advances in Neural Information Processing Systems*. **35**: 23716–23736. arXiv:2204.14198 (https://arxiv.org/abs/2204.14198). Archived (https://web.archive.org/web/20230702195951/https://proceedings.neurips.cc/paper_file s/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html) from the original on 2023-07-02. Retrieved 2023-07-02.

80. Driess, Danny; Xia, Fei; Sajjadi, Mehdi S. M.; Lynch, Corey; Chowdhery, Aakanksha; Ichter, Brian; Wahid, Ayzaan; Tompson, Jonathan; Vuong, Quan; Yu, Tianhe; Huang, Wenlong; Chebotar, Yevgen; Sermanet, Pierre; Duckworth, Daniel; Levine, Sergey (2023-03-01). "PaLM-E: An Embodied Multimodal Language Model". arXiv:2303.03378 (https://arxiv.org/abs/2303.03378) [cs.LG (https://arxiv.org/archive/cs.LG)].

81. Liu, Haotian; Li, Chunyuan; Wu, Qingyang; Lee, Yong Jae (2023-04-01). "Visual Instruction Tuning". arXiv:2304.08485 (https://arxiv.org/abs/2304.08485) [cs.CV (https://arxiv.org/archive/cs.C V)].

82. Zhang, Hang; Li, Xin; Bing, Lidong (2023-06-01). "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding". arXiv:2306.02858 (https://arxiv.org/abs/2306.02 858) [cs.CL (https://arxiv.org/archive/cs.CL)].

83. OpenAI (2023-03-27). "GPT-4 Technical Report". arXiv:2303.08774 (https://arxiv.org/abs/2303.087 74) [cs.CL (https://arxiv.org/archive/cs.CL)].

84. OpenAI (September 25, 2023). "GPT-4V(ision) System Card" (https://cdn.openai.com/papers/GPT V_System_Card.pdf) (PDF).

85. Pichai, Sundar (10 May 2023), *Google Keynote (Google I/O '23)* (https://www.youtube.com/watc h?v=cNfINi5CNbY&t=931s), timestamp 15:31, retrieved 2023-07-02

86. Wiggers, Kyle (11 September 2024). "Mistral releases Pixtral 12B, its first multimodal model" (http s://techcrunch.com/2024/09/11/mistral-releases-pixtral-its-first-multimodal-model/?utm_medium=ai secret.us&utm_source=aisecret.us&utm_campaign=aisecret.us). *TechCrunch*. Retrieved 14 September 2024.

87. "Introducing OpenAI o1-preview" (https://openai.com/index/introducing-openai-o1-preview/). *OpenAI*. 2024-09-12. Retrieved 2025-02-03.

88. Metz, Cade (2024-12-20). "OpenAI Unveils New A.I. That Can 'Reason' Through Math and Science Problems" (https://www.nytimes.com/2024/12/20/technology/openai-new-ai-math-scienc e.html). *The New York Times*. Retrieved 2025-02-03.

89. Gibney, Elizabeth (2025-01-30). "China's cheap, open AI model DeepSeek thrills scientists" (http s://www.nature.com/articles/d41586-025-00229-6). *Nature*. Retrieved 2025-02-03.

90. Lin, Belle (2025-02-05). "Why Amazon is Betting on 'Automated Reasoning' to Reduce AI's Hallucinations: The tech giant says an obscure field that combines AI and math can mitigate—but not completely eliminate—AI's propensity to provide wrong answers" (https://www.wsj.com/article s/why-amazon-is-betting-on-automated-reasoning-to-reduce-ais-hallucinations-b838849e). *Wall Street Journal*. ISSN 0099-9660 (https://search.worldcat.org/issn/0099-9660).

91. Hoffmann, Jordan; Borgeaud, Sebastian; Mensch, Arthur; Buchatskaya, Elena; Cai, Trevor; Rutherford, Eliza; Casas, Diego de Las; Hendricks, Lisa Anne; Welbl, Johannes; Clark, Aidan; Hennigan, Tom; Noland, Eric; Millican, Katie; Driessche, George van den; Damoc, Bogdan (2022-03-29). "Training Compute-Optimal Large Language Models". arXiv:2203.15556 (https://arxiv.org/ abs/2203.15556) [cs.CL (https://arxiv.org/archive/cs.CL)].

92. Caballero, Ethan; Gupta, Kshitij; Rish, Irina; Krueger, David (2022). "Broken Neural Scaling Laws". arXiv:2210.14891 (https://arxiv.org/abs/2210.14891) [cs.LG (https://arxiv.org/archive/cs.LG)].

93. "137 emergent abilities of large language models" (https://www.jasonwei.net/blog/emergence). *Jason Wei*. Retrieved 2023-06-24.

94. Bowman, Samuel R. (2023). "Eight Things to Know about Large Language Models". arXiv:2304.00612 (https://arxiv.org/abs/2304.00612) [cs.CL (https://arxiv.org/archive/cs.CL)].

95. Mukherjee, Anirban; Chang, Hannah (2024). "Heuristic Reasoning in AI: Instrumental Use and Mimetic Absorption". arXiv:2403.09404 (https://arxiv.org/abs/2403.09404) [cs.AI (https://arxiv.org/a rchive/cs.AI)].

96. Hahn, Michael; Goyal, Navin (2023-03-14). "A Theory of Emergent In-Context Learning as Implicit Structure Induction". arXiv:2303.07971 (https://arxiv.org/abs/2303.07971) [cs.LG (https://arxiv.org/ archive/cs.LG)].

97. Pilehvar, Mohammad Taher; Camacho-Collados, Jose (June 2019). "Proceedings of the 2019 Conference of the North" (https://aclanthology.org/N19-1128). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics: 1267–1273. doi:10.18653/v1/N19-1128 (https://doi.org/ 10.18653%2Fv1%2FN19-1128). S2CID 102353817 (https://api.semanticscholar.org/CorpusID:102 353817). Archived (https://web.archive.org/web/20230627202732/https://aclanthology.org/N19-11 28/) from the original on 2023-06-27. Retrieved 2023-06-27.

98. "WiC: The Word-in-Context Dataset" (https://pilehvar.github.io/wic/). *pilehvar.github.io*. Archived (h ttps://web.archive.org/web/20230627202725/https://pilehvar.github.io/wic/) from the original on 2023-06-27. Retrieved 2023-06-27.

99. Patel, Roma; Pavlick, Ellie (2021-10-06). "Mapping Language Models to Grounded Conceptual Spaces" (https://openreview.net/forum?id=gJcEM8sxHK). *ICLR*. Archived (https://web.archive.org/ web/20230624191940/https://openreview.net/forum?id=gJcEM8sxHK) from the original on 2023-06-24. Retrieved 2023-06-27.

100. *A Closer Look at Large Language Models Emergent Abilities (https://www.notion.so/A-Closer-Look -at-Large-Language-Models-Emergent-Abilities-493876b55df5479d80686f68a1abd72f) Archived (https://web.archive.org/web/20230624012329/https://www.notion.so/A-Closer-Look-at-Large-Lan guage-Models-Emergent-Abilities-493876b55df5479d80686f68a1abd72f) 2023-06-24 at the Wayback Machine* (Yao Fu, Nov 20, 2022)

101. Ornes, Stephen (March 16, 2023). "The Unpredictable Abilities Emerging From Large AI Models" (https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-2023 0316/). *Quanta Magazine*. Archived (https://web.archive.org/web/20230316203438/https://www.qu antamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/) from the original on March 16, 2023. Retrieved March 16, 2023.

102. Schaeffer, Rylan; Miranda, Brando; Koyejo, Sanmi (2023-04-01). "Are Emergent Abilities of Large Language Models a Mirage?". arXiv:2304.15004 (https://arxiv.org/abs/2304.15004) [cs.AI (https:// arxiv.org/archive/cs.AI)].

103. Blank, Idan A. (November 2023). "What are large language models supposed to model?" (https://d oi.org/10.1016%2Fj.tics.2023.08.006). *Trends in Cognitive Sciences*. **27** (11): 987–989. doi:10.1016/j.tics.2023.08.006 (https://doi.org/10.1016%2Fj.tics.2023.08.006). PMID 37659920 (ht tps://pubmed.ncbi.nlm.nih.gov/37659920).

104. https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-poems%7Ctitle=On the Biology of a Large Language Model (Chapter on Planning in Poems)

105. Nanda, Neel; Chan, Lawrence; Lieberum, Tom; Smith, Jess; Steinhardt, Jacob (2023-01-01). "Progress measures for grokking via mechanistic interpretability". arXiv:2301.05217 (https://arxiv.o rg/abs/2301.05217) [cs.LG (https://arxiv.org/archive/cs.LG)].

106. Mitchell, Melanie; Krakauer, David C. (28 March 2023). "The debate over understanding in AI's large language models" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10068812). *Proceedings of the National Academy of Sciences*. **120** (13): e2215907120. arXiv:2210.13966 (https://arxiv.org/ abs/2210.13966). Bibcode:2023PNAS..12015907M (https://ui.adsabs.harvard.edu/abs/2023PNA S..12015907M). doi:10.1073/pnas.2215907120 (https://doi.org/10.1073%2Fpnas.2215907120). PMC 10068812 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10068812). PMID 36943882 (http s://pubmed.ncbi.nlm.nih.gov/36943882).

107. Metz, Cade (16 May 2023). "Microsoft Says New A.I. Shows Signs of Human Reasoning" (https:// www.nytimes.com/2023/05/16/technology/microsoft-ai-human-reasoning.html). *The New York Times*.

108. Bubeck, Sébastien; Chandrasekaran, Varun; Eldan, Ronen; Gehrke, Johannes; Horvitz, Eric; Kamar, Ece; Lee, Peter; Lee, Yin Tat; Li, Yuanzhi; Lundberg, Scott; Nori, Harsha; Palangi, Hamid; Ribeiro, Marco Tulio; Zhang, Yi (2023). "Sparks of Artificial General Intelligence: Early experiments with GPT-4". arXiv:2303.12712 (https://arxiv.org/abs/2303.12712) [cs.CL (https://arxiv.org/archive/ cs.CL)].

109. "Anthropic CEO Dario Amodei pens a smart look at our AI future" (https://www.fastcompany.com/9 1211163/anthropic-ceo-dario-amodei-pens-a-smart-look-at-our-ai-future). *Fast Company*. October 17, 2024.

110. "ChatGPT is more like an 'alien intelligence' than a human brain, says futurist" (https://www.zdnet. com/article/chatgpt-is-more-like-an-alien-intelligence-than-a-human-brain-says-futurist/). *ZDNET*. 2023. Archived (https://web.archive.org/web/20230612065937/https://www.zdnet.com/article/chatg pt-is-more-like-an-alien-intelligence-than-a-human-brain-says-futurist/) from the original on 12 June 2023. Retrieved 12 June 2023.

111. Newport, Cal (13 April 2023). "What Kind of Mind Does ChatGPT Have?" (https://www.newyorker. com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have). *The New Yorker*. Archived (https://web.archive.org/web/20230612071443/https://www.newyorker.com/scien ce/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have) from the original on 12 June 2023. Retrieved 12 June 2023.

112. Roose, Kevin (30 May 2023). "Why an Octopus-like Creature Has Come to Symbolize the State of A.I." (https://www.nytimes.com/2023/05/30/technology/shoggoth-meme-ai.html) *The New York Times*. Archived (https://web.archive.org/web/20230530193814/https://www.nytimes.com/2023/05/30/technology/shoggoth-meme-ai.html) from the original on 30 May 2023. Retrieved 12 June 2023.

113. "The A to Z of Artificial Intelligence" (https://time.com/6271657/a-to-z-of-artificial-intelligence/). *Time Magazine*. 13 April 2023. Archived (https://web.archive.org/web/20230616123839/https://time.com/6271657/a-to-z-of-artificial-intelligence/) from the original on 16 June 2023. Retrieved 12 June 2023.

114. Ji, Ziwei; Lee, Nayeon; Frieske, Rita; Yu, Tiezheng; Su, Dan; Xu, Yan; Ishii, Etsuko; Bang, Yejin; Dai, Wenliang; Madotto, Andrea; Fung, Pascale (November 2022). "Survey of Hallucination in Natural Language Generation" (https://dl.acm.org/doi/pdf/10.1145/3571730) (pdf). *ACM Computing Surveys*. **55** (12). Association for Computing Machinery: 1–38. arXiv:2202.03629 (https://arxiv.org/abs/2202.03629). doi:10.1145/3571730 (https://doi.org/10.1145%2F3571730). S2CID 246652372 (https://api.semanticscholar.org/CorpusID:246652372). Archived (https://web.archive.org/web/20230326145635/https://dl.acm.org/doi/pdf/10.1145/3571730) from the original on 26 March 2023. Retrieved 15 January 2023.

115. Varshney, Neeraj; Yao, Wenlin; Zhang, Hongming; Chen, Jianshu; Yu, Dong (2023). "A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation". arXiv:2307.03987 (https://arxiv.org/abs/2307.03987) [cs.CL (https://arxiv.org/archive/cs.CL)].

116. Lakoff, George (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Philosophy; Appendix: The Neural Theory of Language Paradigm*. New York Basic Books. pp. 569–583. ISBN 978-0-465-05674-3.

117. Evans, Vyvyan. (2014). *The Language Myth*. Cambridge University Press. ISBN 978-1-107-04396-1.

118. Friston, Karl J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior; Chapter 4 The Generative Models of Active Inference*. The MIT Press. ISBN 978-0-262-36997-8.

119. Huyen, Chip (October 18, 2019). "Evaluation Metrics for Language Modeling" (https://thegradient.pub/understanding-evaluation-metrics-for-language-models/). *The Gradient*. Retrieved January 14, 2024.

120. *openai/simple-evals* (https://github.com/openai/simple-evals), OpenAI, 2024-05-28, retrieved 2024-05-28

121. *openai/evals* (https://github.com/openai/evals), OpenAI, 2024-05-28, archived (https://web.archive.org/web/20240508225708/https://github.com/openai/evals) from the original on 2024-05-08, retrieved 2024-05-28

122. Clark, Christopher; Lee, Kenton; Chang, Ming-Wei; Kwiatkowski, Tom; Collins, Michael; Toutanova, Kristina (2019). "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions". arXiv:1905.10044 (https://arxiv.org/abs/1905.10044) [cs.CL (https://arxiv.org/archive/cs.CL)].

123. Wayne Xin Zhao; Zhou, Kun; Li, Junyi; Tang, Tianyi; Wang, Xiaolei; Hou, Yupeng; Min, Yingqian; Zhang, Beichen; Zhang, Junjie; Dong, Zican; Du, Yifan; Yang, Chen; Chen, Yushuo; Chen, Zhipeng; Jiang, Jinhao; Ren, Ruiyang; Li, Yifan; Tang, Xinyu; Liu, Zikang; Liu, Peiyu; Nie, Jian-Yun; Wen, Ji-Rong (2023). "A Survey of Large Language Models". arXiv:2303.18223 (https://arxiv.org/abs/2303.18223) [cs.CL (https://arxiv.org/archive/cs.CL)].

124. Nangia, Nikita and Vania, Clara and Bhalerao, Rasika and Bowman, Samuel R. (November 2020). "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models" (https://aclanthology.org/2020.emnlp-main.154/). In Webber, Bonnie and Cohn, Trevor and He, Yulan and Liu, Yang (ed.). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. pp. 1953–1967. arXiv:2010.00133 (https://arxiv.org/abs/2010.00133). doi:10.18653/v1/2020.emnlp-main.154 (https://doi.org/10.18653%2Fv1%2F2020.emnlp-main.154).

125. Nadeem, Moin and Bethke, Anna and Reddy, Siva (August 2021). "StereoSet: Measuring stereotypical bias in pretrained language models" (https://aclanthology.org/2021.acl-long.416/). In Zong, Chengqing and Xia, Fei and Li, Wenjie and Navigli, Roberto (ed.). *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics. pp. 5356–5371. arXiv:2004.09456 (https://arxiv.org/abs/2004.09456). doi:10.18653/v1/2021.acl-long.416 (https://doi.org/10.18653%2Fv1%2F2021.acl-long.416).

126. Simpson, Shmona and Nukpezah, Jonathan and Kie Brooks and Pandya, Raaghav (17 December 2024). "Parity benchmark for measuring bias in LLMs" (https://doi.org/10.1007%2Fs43681-024-00613-4). *AI and Ethics*. Springer. doi:10.1007/s43681-024-00613-4 (https://doi.org/10.1007%2Fs43681-024-00613-4).

127. Caramancion, Kevin Matthe (2023-11-13). "News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking". *2023 IEEE Future Networks World Forum (FNWF)* (https://ieeexplore.ieee.org/document/10520446). IEEE. pp. 1–6. arXiv:2306.17176 (https://arxiv.org/abs/2306.17176). doi:10.1109/FNWF58287.2023.10520446 (https://doi.org/10.1109%2FFNWF58287.2023.10520446). ISBN 979-8-3503-2458-7.

128. "Sanitized open-source datasets for natural language and code understanding: how we evaluated our 70B model" (https://imbue.com/research/70b-evals/). *imbue.com*. Archived (https://web.archive.org/web/20240726173012/https://imbue.com/research/70b-evals/) from the original on 2024-07-26. Retrieved 2024-07-24.

129. Srivastava, Aarohi; et al. (2022). "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". arXiv:2206.04615 (https://arxiv.org/abs/2206.04615) [cs.CL (https://arxiv.org/archive/cs.CL)].

130. Lin, Stephanie; Hilton, Jacob; Evans, Owain (2021). "TruthfulQA: Measuring How Models Mimic Human Falsehoods". arXiv:2109.07958 (https://arxiv.org/abs/2109.07958) [cs.CL (https://arxiv.org/archive/cs.CL)].

131. Zellers, Rowan; Holtzman, Ari; Bisk, Yonatan; Farhadi, Ali; Choi, Yejin (2019). "HellaSwag: Can a Machine Really Finish Your Sentence?". arXiv:1905.07830 (https://arxiv.org/abs/1905.07830) [cs.CL (https://arxiv.org/archive/cs.CL)].

132. "Prepare for truly useful large language models". *Nature Biomedical Engineering*. **7** (2): 85–86. 7 March 2023. doi:10.1038/s41551-023-01012-6 (https://doi.org/10.1038%2Fs41551-023-01012-6). PMID 36882584 (https://pubmed.ncbi.nlm.nih.gov/36882584). S2CID 257403466 (https://api.semanticscholar.org/CorpusID:257403466).

133. "Your job is (probably) safe from artificial intelligence" (https://www.economist.com/finance-and-economics/2023/05/07/your-job-is-probably-safe-from-artificial-intelligence). *The Economist*. 7 May 2023. Archived (https://web.archive.org/web/20230617225618/https://www.economist.com/finance-and-economics/2023/05/07/your-job-is-probably-safe-from-artificial-intelligence) from the original on 17 June 2023. Retrieved 18 June 2023.

134. "Generative AI Could Raise Global GDP by 7%" (https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html). *Goldman Sachs*. Archived (https://web.archive.org/web/20230618013836/https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html) from the original on 18 June 2023. Retrieved 18 June 2023.

135. Brinkmann, Levin; Baumann, Fabian; Bonnefon, Jean-François; Derex, Maxime; Müller, Thomas F.; Nussberger, Anne-Marie; Czaplicka, Agnieszka; Acerbi, Alberto; Griffiths, Thomas L.; Henrich, Joseph; Leibo, Joel Z.; McElreath, Richard; Oudeyer, Pierre-Yves; Stray, Jonathan; Rahwan, Iyad (2023-11-20). "Machine culture" (https://www.nature.com/articles/s41562-023-01742-2). *Nature Human Behaviour*. **7** (11): 1855–1868. arXiv:2311.11388 (https://arxiv.org/abs/2311.11388). doi:10.1038/s41562-023-01742-2 (https://doi.org/10.1038%2Fs41562-023-01742-2). ISSN 2397-3374 (https://search.worldcat.org/issn/2397-3374). PMID 37985914 (https://pubmed.ncbi.nlm.nih.gov/37985914).

136. Peng, Zhencan; Wang, Zhizhi; Deng, Dong (13 June 2023). "Near-Duplicate Sequence Search at Scale for Large Language Model Memorization Evaluation" (https://people.cs.rutgers.edu/~dd903/assets/papers/sigmod23.pdf) (PDF). *Proceedings of the ACM on Management of Data*. **1** (2): 1–18. doi:10.1145/3589324 (https://doi.org/10.1145%2F3589324). S2CID 259213212 (https://api.semanticscholar.org/CorpusID:259213212). Archived (https://web.archive.org/web/20240827053753/https://people.cs.rutgers.edu/~dd903/assets/papers/sigmod23.pdf) (PDF) from the original on 2024-08-27. Retrieved 2024-01-20. Citing Lee et al 2022.

137. Peng, Wang & Deng 2023, p. 8.

138. Stephen Council (1 Dec 2023). "How Googlers cracked an SF rival's tech model with a single word" (https://www.sfgate.com/tech/article/google-openai-chatgpt-break-model-18525445.php). SFGATE. Archived (https://web.archive.org/web/20231216160941/https://www.sfgate.com/tech/article/google-openai-chatgpt-break-model-18525445.php) from the original on 16 December 2023.

139. Alba, Davey (1 May 2023). "AI chatbots have been used to create dozens of news content farms" (https://www.japantimes.co.jp/news/2023/05/01/business/tech/ai-fake-news-content-farms/). *The Japan Times*. Retrieved 18 June 2023.

140. "Could chatbots help devise the next pandemic virus?" (https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus). *Science*. 14 June 2023. doi:10.1126/science.adj2463 (https://doi.org/10.1126%2Fscience.adj2463). Archived (https://web.archive.org/web/20230618013834/https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus) from the original on 18 June 2023. Retrieved 18 June 2023.

141. Hubinger, Evan (10 January 2024). "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training". arXiv:2401.05566 (https://arxiv.org/abs/2401.05566) [cs.CR (https://arxiv.org/archive/cs.CR)].

142. Kang, Daniel (2023). "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks". arXiv:2302.05733 (https://arxiv.org/abs/2302.05733) [cs.CR (https://arxiv.org/archive/cs.CR)].

143. "Russian propaganda may be flooding AI models" (https://www.americansunlight.org/updates/new-report-russian-propaganda-may-be-flooding-ai-models). *The American Sunlight Project*. 26 February 2025. Retrieved 2025-04-11.

144. Goudarzi, Sara (2025-03-26). "Russian networks flood the Internet with propaganda, aiming to corrupt AI chatbots" (https://thebulletin.org/2025/03/russian-networks-flood-the-internet-with-propaganda-aiming-to-corrupt-ai-chatbots/). *Bulletin of the Atomic Scientists*. Retrieved 2025-04-10.

145. Wang, Yongge (20 June 2024). "Encryption Based Covert Channel for Large Language Models" (https://eprint.iacr.org/2024/586.pdf) (PDF). IACR ePrint 2024/586. Archived (https://web.archive.org/web/20240624191233/https://eprint.iacr.org/2024/586.pdf) (PDF) from the original on 24 June 2024. Retrieved 24 June 2024.

146. Stokel-Walker, Chris (November 22, 2023). "ChatGPT Replicates Gender Bias in Recommendation Letters" (https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/). *Scientific American*. Archived (https://web.archive.org/web/20231229043124/https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/) from the original on 2023-12-29. Retrieved 2023-12-29.

147. Luo, Queenie; Puett, Michael J.; Smith, Michael D. (2023-03-28). "A Perspective Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and YouTube". arXiv:2303.16281v2 (https://arxiv.org/abs/2303.16281v2) [cs.CY (https://arxiv.org/archive/cs.CY)].

148. Wang, Angelina; Morgenstern, Jamie; Dickerson, John P. (17 February 2025). "Large language models that replace human participants can harmfully misportray and flatten identity groups". *Nature Machine Intelligence*. **7** (3): 400–411. arXiv:2402.01908 (https://arxiv.org/abs/2402.01908). doi:10.1038/s42256-025-00986-z (https://doi.org/10.1038%2Fs42256-025-00986-z).

149. Cheng, Myra; Durmus, Esin; Jurafsky, Dan (2023-05-29), *Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models*, arXiv:2305.18189 (https://arxiv.org/abs/2305.18189)

150. Kotek, Hadas; Dockum, Rikker; Sun, David (2023-11-05). "Gender bias and stereotypes in Large Language Models" (https://dl.acm.org/doi/10.1145/3582269.3615599). *Proceedings of the ACM Collective Intelligence Conference*. CI '23. New York, NY, USA: Association for Computing Machinery. pp. 12–24. doi:10.1145/3582269.3615599 (https://doi.org/10.1145%2F3582269.3615599). ISBN 979-8-4007-0113-9.

151. Choi, Hyeong Kyu; Xu, Weijie; Xue, Chi; Eckman, Stephanie; Reddy, Chandan K. (2024-09-27), *Mitigating Selection Bias with Node Pruning and Auxiliary Options*, arXiv:2409.18857 (https://arxiv.org/abs/2409.18857)

152. Zheng, Chujie; Zhou, Hao; Meng, Fandong; Zhou, Jie; Huang, Minlie (2023-09-07), *Large Language Models Are Not Robust Multiple Choice Selectors*, arXiv:2309.03882 (https://arxiv.org/abs/2309.03882)

153. Heikkilä, Melissa (August 7, 2023). "AI language models are rife with different political biases" (https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/). *MIT Technology Review*. Retrieved 2023-12-29.

154. Mehta, Sourabh (2024-07-03). "How Much Energy Do LLMs Consume? Unveiling the Power Behind AI" (https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/). *Association of Data Scientists*. Retrieved 2025-01-27.

155. "Artificial Intelligence wants to go nuclear. Will it work?" (https://www.npr.org/2024/12/09/nx-s1-5171063/artificial-intelligence-wants-to-go-nuclear-will-it-work). *NPR*. Retrieved 2025-01-27.

156. Roy, Dareen (December 19, 2024). "AI's energy hunger fuels geothermal startups but natgas rivalry clouds future" (https://www.reuters.com/technology/artificial-intelligence/ais-energy-hunger-fuels-geothermal-startups-natgas-rivalry-clouds-future-2024-12-19/). *Reuters*.

Further reading

- Jurafsky, Dan, Martin, James. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf), 3rd Edition draft, 2023.

- Zhao, Wayne Xin; et al. (2023). "A Survey of Large Language Models". arXiv:2303.18223 (https://arxiv.org/abs/2303.18223) [cs.CL (https://arxiv.org/archive/cs.CL)].

- Kaddour, Jean; et al. (2023). "Challenges and Applications of Large Language Models". arXiv:2307.10169 (https://arxiv.org/abs/2307.10169) [cs.CL (https://arxiv.org/archive/cs.CL)].

- Yin, Shukang; Fu, Chaoyou; Zhao, Sirui; Li, Ke; Sun, Xing; Xu, Tong; Chen, Enhong (2024). "A Survey on Multimodal Large Language Models" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11645129). *National Science Review*. **11** (12): nwae403. arXiv:2306.13549 (https://arxiv.org/abs/2306.13549). doi:10.1093/nsr/nwae403 (https://doi.org/10.1093%2Fnsr%2Fnwae403). PMC 11645129 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11645129). PMID 39679213 (https://pubmed.ncbi.nlm.nih.gov/39679213).

- "AI Index Report 2024 – Artificial Intelligence Index" (https://aiindex.stanford.edu/report/).
  aiindex.stanford.edu. Retrieved 2024-05-05.
- Frank, Michael C. (27 June 2023). "Baby steps in evaluating the capacities of large language
  models" (https://www.nature.com/articles/s44159-023-00211-x). *Nature Reviews Psychology*. **2**
  (8): 451–452. doi:10.1038/s44159-023-00211-x (https://doi.org/10.1038%2Fs44159-023-00211-x).
  ISSN 2731-0574 (https://search.worldcat.org/issn/2731-0574). S2CID 259713140 (https://api.sem
  anticscholar.org/CorpusID:259713140). Retrieved 2 July 2023.
- Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.;
  Casper, S.; Sourbut, O.; Edelman, B. L.; Zhang, Z.; Günther, M.; Korinek, A.; Hernandez-Orallo,
  J.; Hammond, L.; Bigelow, E.; Pan, A.; Langosco, L.; Krueger, D. (2024). "Foundational
  Challenges in Assuring Alignment and Safety of Large Language Models". arXiv:2404.09932 (http
  s://arxiv.org/abs/2404.09932) [cs.LG (https://arxiv.org/archive/cs.LG)].

Retrieved from "https://en.wikipedia.org/w/index.php?title=Large_language_model&oldid=1289377436"