WIKIPEDIA
The Free Encyclopedia

Large language model

A **large language model** (**LLM**) is a language model trained with self-supervised machine learning on a vast amount of text, designed for natural language processing tasks, especially language generation.[1][2] The largest and most capable LLMs are generative pre-trained transformers (GPTs) and provide the core capabilities of chatbots such as ChatGPT, Gemini and Claude. LLMs can be fine-tuned for specific tasks or guided by prompt engineering.[3] These models acquire predictive power regarding syntax, semantics, and ontologies[4] inherent in human language corpora, but they also inherit inaccuracies and biases present in the data they are trained on.[5]

They consist of billions to trillions of parameters and operate as general-purpose sequence models, generating, summarizing, translating, and reasoning over text. LLMs represent a significant new technology in their ability to generalize across tasks with minimal task-specific supervision, enabling capabilities like conversational agents, code generation, knowledge retrieval, and automated reasoning that previously required bespoke systems.[6]

LLMs evolved from earlier statistical and recurrent neural network approaches to language modeling. The transformer architecture, introduced in 2017, replaced recurrence with self-attention, allowing efficient parallelization, longer context handling, and scalable training on unprecedented data volumes.[7] This innovation enabled models like GPT, BERT, and their successors, which demonstrated emergent behaviors at scale such as few-shot learning and compositional reasoning.[8]

Reinforcement learning, particularly policy gradient algorithms, has been adapted to fine-tune LLMs for desired behaviors beyond raw next-token prediction.[9] Reinforcement learning from human feedback (RLHF) applies these methods to optimize a policy, the LLM's output distribution, against reward signals derived from human or automated preference judgments.[10] This has been critical for aligning model outputs with user expectations, improving factuality, reducing harmful responses, and enhancing task performance.

Benchmark evaluations for LLMs have evolved from narrow linguistic assessments toward comprehensive, multi-task evaluations measuring reasoning, factual accuracy, alignment, and safety.[11][12] Hill climbing, iteratively optimizing models against benchmarks, has emerged as a dominant strategy, producing rapid incremental performance gains but raising concerns of overfitting to benchmarks rather than achieving genuine generalization or robust capability improvements.[13]

History

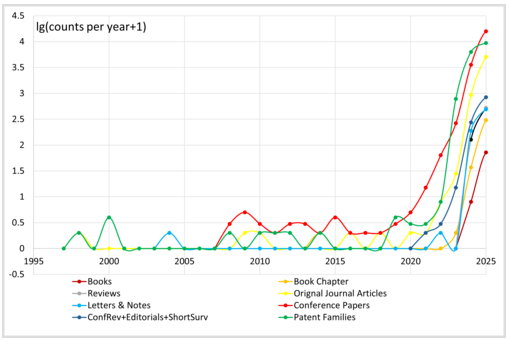
Before the emergence of transformer-based models in 2017, some language models were considered large relative to the computational and data constraints of their time. In the early 1990s, IBM's statistical models pioneered word alignment techniques for machine translation, laying the groundwork for corpus-based language modeling. In 2001, a smoothed *n*-gram model, such as those employing Kneser–Ney smoothing, trained on 300 million words, achieved state-of-the-art perplexity

on benchmark tests.[14] During the 2000s, with the rise of widespread internet access, researchers began compiling massive text datasets from the web ("web as corpus"[15]) to train statistical language models.[16][17]

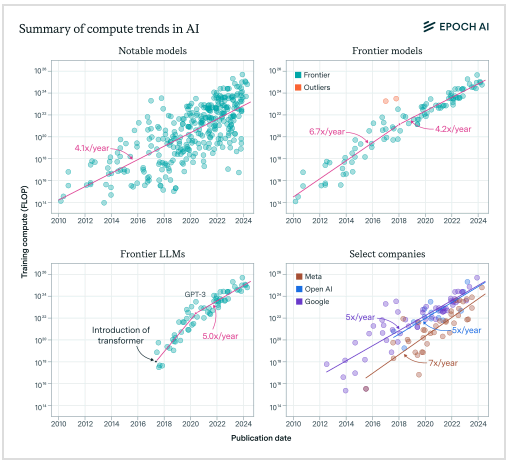
Moving beyond *n*-gram models, researchers started in 2000 to use neural networks to learn language models.[18] Following the breakthrough of deep neural networks in image classification around 2012,[19] similar architectures were adapted for language tasks. This shift was marked by the development of word embeddings (eg, Word2Vec by Mikolov in 2013) and sequence-to-sequence (seq2seq) models using LSTM. In 2016, Google transitioned its translation service to neural machine translation (NMT), replacing statistical phrase-based models with deep recurrent neural networks. These early NMT systems used LSTM-based encoder-decoder architectures, as they preceded the invention of transformers.

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,[20] and was based mainly on the attention mechanism developed by Bahdanau et al. in 2014.[21] The following year in 2018, BERT was introduced and quickly became "ubiquitous".[22] 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.[23]

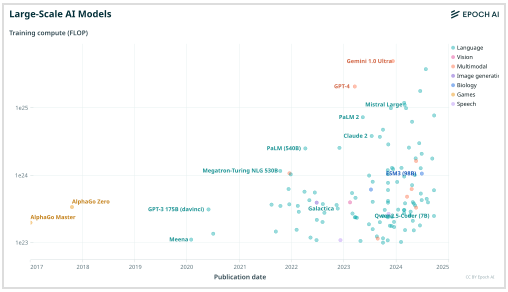
Although decoder-only GPT-1 was introduced in 2018, it was GPT-2 in 2019 that caught widespread attention because OpenAI claimed to have initially deemed it too powerful to release publicly, out of fear of malicious use.[24] GPT-3 in 2020 went a step further and as of 2025 is available only via API with no offering of downloading the model to execute locally. But it was the 2022 consumer-facing chatbot ChatGPT that received extensive media coverage and public attention.[25] The 2023 GPT-4 was praised for its increased accuracy and as a "holy grail" for its multimodal capabilities.[26] 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.[23]



The number of publications about large language models by year grouped by publication types.



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.

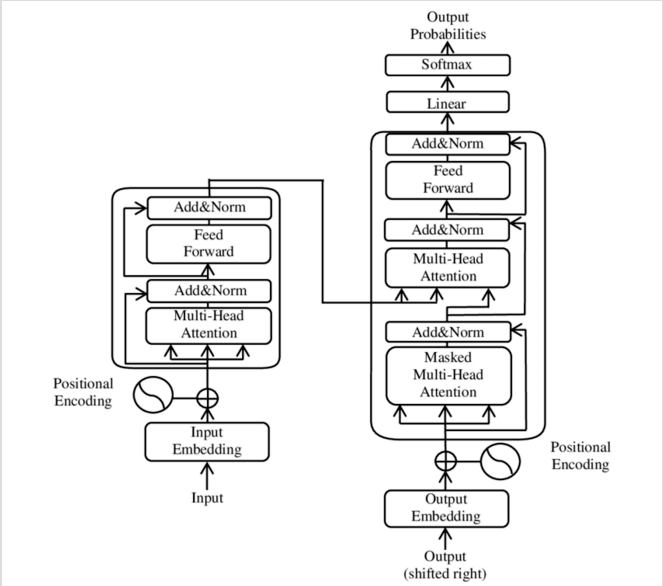


The training compute of notable large AI models in FLOPs vs publication date over the period 2017–2024. The majority of large models are language models or multimodal models with language capacity.

In 2024 OpenAI released the reasoning model OpenAI o1, which generates long chains of thought before returning a final answer.[27] Many LLMs with parameter counts comparable to those of OpenAI's GPT series have been developed.[28]

Since 2022, open-weight models have been gaining popularity, especially at first with BLOOM and LLaMA, though both have restrictions on usage and deployment. Mistral AI's models Mistral 7B and Mixtral 8x7b have a 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.[29]

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).[30]



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).[31][32][33]

Open-weight LLMs have increasingly shaped the field since 2023, contributing to broader participation in AI development and greater transparency in model evaluation. Vake et al. (2025) demonstrated that community-driven contributions to open-weight models measurably improve their efficiency and performance, with user participation growing rapidly on collaborative platforms such as Hugging Face.[34] Paris et al. (2025) further argued that openness in AI should extend beyond releasing model code or weights to encompass inclusiveness, accountability, and ethical responsibility in AI research and deployment.[35] Collectively, these studies highlight that open-weight LLMs can accelerate innovation and enhance scientific reproducibility, while fostering a more transparent and participatory AI ecosystem.

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 and "##" denotes continuation of a preceding word in BERT.[36]

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

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. The average number of words per token depends on the language.[37][38] In English, the ratio is typically around 0.75 words per token, with 4 characters per token on average.[39]

Byte-pair encoding

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. 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.[40]

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.[38]

Dataset cleaning

In the context of training LLMs, datasets are typically cleaned by removing low-quality, duplicated, or toxic data.[41] Cleaned datasets can increase training efficiency and lead to improved downstream performance.[42][43] A trained LLM can be used to clean datasets for training a further LLM.[44]

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).[3]

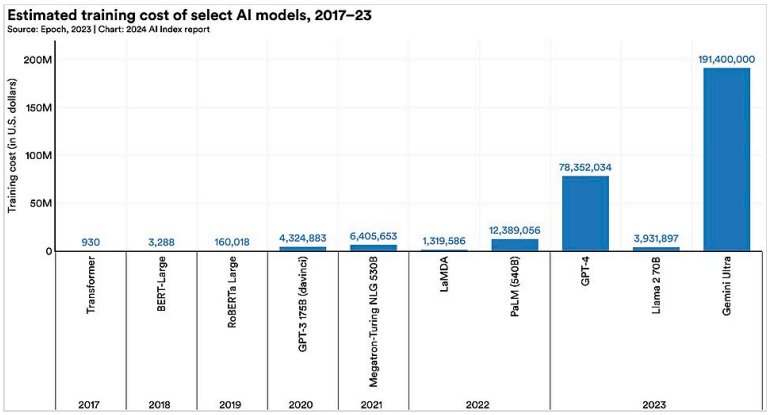
Synthetic data

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.[45]

Training

An LLM is a type of foundation model (large X model) trained on language. LLMs can be trained in different ways. In particular, GPT models are first pretrained to predict the next word on a large amount of data, before being fine-tuned.

Cost



Substantial infrastructure is necessary for training the largest models. The tendency towards larger models is visible in the list of large language models. For example, the training of 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. 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". GPT-1 of 2018 has 117 million parameters.

Fine-tuning

Before being fine-tuned, most LLMs are next-token predictors. The fine-tuning adjusts the output of an LLM to seem more conversational via techniques like reinforcement learning from human feedback (RLHF) or constitutional AI.[46]

Instruction fine-tuning is a form of supervised learning used to teach LLMs to follow user instructions. In 2022, OpenAI demonstrated InstructGPT, a version of GPT-3 similarly fine-tuned to follow instructions.[47]

Reinforcement learning from human feedback (RLHF) involves training a reward model to predict which text humans prefer. Then, the LLM can be fine-tuned through reinforcement learning to better satisfy this reward model. Since humans typically prefer truthful, helpful and harmless answers, RLHF favors such answers.

Architecture

LLMs are generally based on the transformer architecture, which leverages an attention mechanism that enables the model to process relationships between all elements in a sequence simultaneously, regardless of their distance from each other.

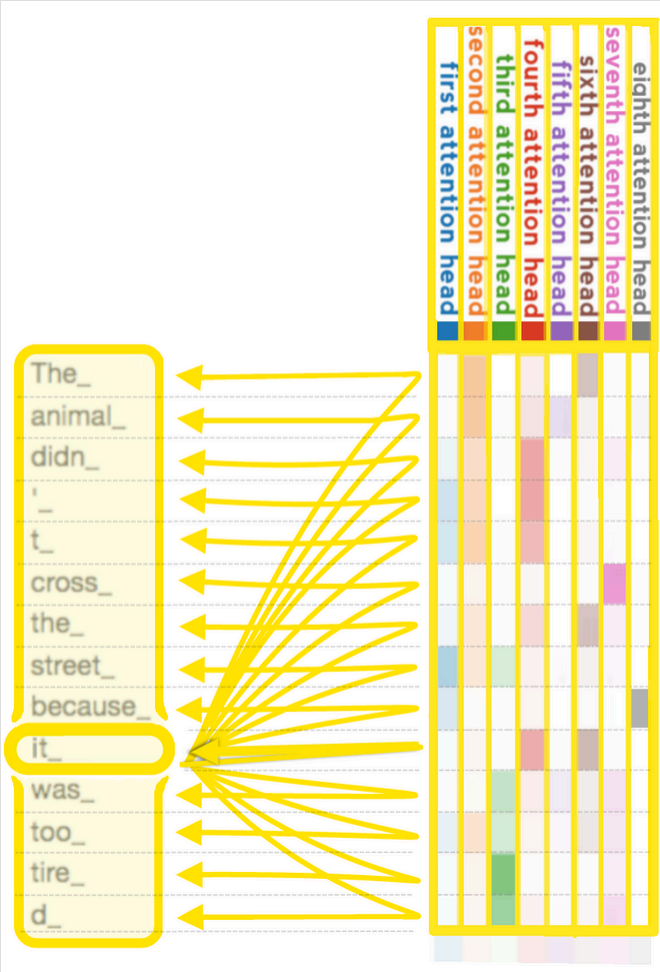
Attention mechanism and context window

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.[49] 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.[40]

Google's Gemini 1.5, introduced in February 2024, can have a context window of up to 1 million tokens.[50]

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.[51] 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"[52] does it): for example, given a segment "I like to [__] [__] cream", the model predicts that "eat" and "ice" are missing.



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.[48]

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.[52] During training, regularization loss is also used to stabilize training. However regularization loss is usually not used during testing and evaluation.

Mixture of experts

A mixture of experts (MoE) is a machine learning architecture in which multiple specialized neural networks ("experts") work together, with a gating mechanism that routes each input to the most appropriate expert(s). Mixtures of experts can reduce inference costs, as only a fraction of the parameters are used for each input. The approach was introduced in 2017 by Google researchers.[53][54][55]

Parameter size

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.[56]

Quantization

Post-training quantization[57] aims to decrease the space requirement by lowering precision of the parameters of a trained model, while preserving most of its performance. Quantization can be further classified as *static quantization* if the quantization parameters are determined beforehand (typically during a calibration phase), and *dynamic quantization* if the quantization is applied during inference. The simplest form of quantization simply truncates all the parameters to a given number of bits: this is applicable to static as well as dynamic quantization, but loses much precision. Dynamic quantization allows for the use of a different quantization codebook per layer, either a lookup table of values or a linear mapping (scaling factor and bias), at the cost of foregoing the possible speed improvements from using lower-precision arithmetic.

Quantized models are typically seen as frozen with modification of weights (e.g. fine-tuning) only applied to the original model. It is possible to fine-tune quantized models using low-rank adaptation.

Extensibility

Beyond basic text generation, various techniques have been developed to extend LLM capabilities, including the use of external tools and data sources, improved reasoning on complex problems, and enhanced instruction-following or autonomy through prompting methods.

Prompt engineering

In 2020, OpenAI researchers demonstrated that their new model GPT-3 could understand what format to use given a few rounds of Q and A (or other type of task) in the input data as example, thanks in part due to the RLHF technique. This technique, called *few-shot prompting*, allows LLMs to be adapted to any task without requiring fine-tuning.[3] Also in 2022, it was found that the base GPT-3 model can generate an instruction based on user input. The generated instruction along with user input is then used as input to another instance of the model under a "Instruction: [...], Input: [...],

Output:" format. The other instance is able to complete the output and often produces the correct answer in doing so. The ability to "self-instruct" makes LLMs able to bootstrap themselves toward a correct answer.[58]

Dialogue processing (chatbot)

An LLM can be turned into a chatbot or a "dialog assistant" by specializing it for conversation. In essence, user input is prefixed with a marker such as "Q:" or "User:" and the LLM is asked to predict the output after a fixed "A:" or "Assistant:". This type of model became commercially available in 2022 with ChatGPT, a sibling model of InstructGPT fine-tuned to accept and produce dialog-formatted text based on GPT-3.5. It could similarly follow user instructions.[59] Before the stream of User and Assistant lines, a chat context usually start with a few lines of overarching instructions, from a role called "developer" or "system" to convey a higher authority than the user's input. This is called a "system prompt".[60][61]

Retrieval-augmented generation

Retrieval-augmented generation (RAG) is an 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.[62][63]

Tool use

Tool use is a mechanism that enables LLMs to interact with external systems, applications, or data sources. It can allow for example to fetch real-time information from an API or to execute code. A program separate from the LLM watches the output stream of the LLM for a special tool-calling syntax. When these special tokens appear, the program calls the tool accordingly and feeds its output back into the LLM's input stream.[64]

Early tool-using LLMs were fine-tuned on the use of specific tools. But fine-tuning LLMs for the ability to read API documentation and call API correctly has greatly expanded the range of tools accessible to an LLM.[65][66] Describing available tools in the system prompt can also make an LLM able to use tools. A system prompt instructing ChatGPT (GPT-4) to use multiple types of tools can be found online.[67]

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 it can be transformed into an agent by adding supporting elements: the role (profile) and the surrounding environment of an agent can be additional inputs to the LLM, while memory can be integrated as a tool or provided as additional input. Instructions and input patterns are used to make the LLM plan actions and tool use is used to potentially carry out these actions.[68]

The ReAct pattern, a portmanteau of *reason* and *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.[69]

In the DEPS ("describe, explain, plan and select") method, an LLM is first connected to the visual world via image descriptions. It is then prompted to produce plans for complex tasks and behaviors based on its pretrained knowledge and the environmental feedback it receives.[70]

The Reflexion method[71] 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 stored as a form of long-term memory and given to the agent in the subsequent episodes.[71]

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.[72]

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.[73] Alternatively, it can propose increasingly difficult tasks for curriculum learning.[74] 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.[74]

Multiple agents with memory can interact socially.[75]

Reasoning

LLMs are conventionally trained to generate an output without generating intermediate steps. As a result, their performance tends to be subpar on complex questions requiring (at least in humans) intermediate steps of thought. Early research demonstrated that inserting intermediate "scratchpad" computations could improve performance on such tasks.[76] Later methods overcame this deficiency more systematically by breaking tasks into smaller steps for the LLM, either manually or automatically.

Chaining

The "prompt chaining" paradigm was published in 2021.[77] In this method, a user manually breaks a complex problem down into several steps. In each step, the LLM receives as input a prompt telling it what to do and some results from preceeding steps. The result from one step is then reused in a next step, until a final answer is reached. The ability of an LLM to follow instructions means that even non-experts can write a successful collection of step-wise prompts given a few rounds of trial and error.[78][79]

A 2022 paper demonstrated a separate technique called "chain-of-thought prompting", which makes the LLM break the question down autonomously. An LLM is given some examples where the "assistant" verbally breaks down the thought process before arriving at an answer. The LLM mimics these examples and also tries to spend some time generating intermediate steps before providing the final answer. This additional step elicited by prompting improves the correctness of the LLM on relatively complex questions. On math word questions, a prompted model can exceed even fine-tuned GPT-3 with a verifier.[77][80] Chain-of-thought can also be elicited by simply adding an instruction like "Let's think step by step" to the prompt, in order to encourage the LLM to proceed methodically instead of trying to directly guess the answer.[81]

Model-native reasoning

In late 2024 "reasoning models" were released. These were trained to spend more time generating step-by-step solutions before providing final answers, which was intended to be similar to human problem-solving processes. OpenAI introduced this concept with their o1 model in September 2024, followed by o3 in April 2025. On the International Mathematics Olympiad qualifying exam problems, GPT-4o achieved 13% accuracy while o1 reached 83%.[82]

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.[83]

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.[82]

Inference optimization

Inference optimization refers to techniques that improve LLM performance by applying additional computational resources during the inference process, rather than requiring model retraining. These approaches implement various state-of-the-art reasoning and decision-making strategies to enhance accuracy and capabilities.

OptiLLM is an OpenAI API-compatible optimizing inference proxy that implements multiple inference optimization techniques simultaneously.[84] The system acts as a transparent proxy that can work with any LLM provider, implementing techniques such as Monte Carlo tree search (MCTS), mixture of agents (MOA), best-of-N sampling, and chain-of-thought reflection. OptiLLM demonstrates that strategic application of computational resources at inference time can substantially improve model performance across diverse tasks, achieving significant improvements on benchmarks such as the AIME 2024 mathematics competition and various coding challenges.[85]

These inference optimization approaches represent a growing category of tools that enhance existing LLMs without requiring access to model weights or retraining, making advanced reasoning capabilities more accessible across different model providers and use cases.

Forms of input and output

Multimodality

Multimodality means having multiple modalities, where a "modality" refers to a type of input or output, such as video, image, audio, text, proprioception, etc.[86] For example, Google PaLM model was fine-tuned into a multimodal model and applied to robotic control.[87] LLaMA models have also been turned multimodal using the tokenization method, to allow image inputs,[88] and video inputs.[89] GPT-4o can process and generate text, audio and images.[90] Such models are sometimes called large multimodal models (LMMs).[91]

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.[92] This type of method, where embeddings from multiple modalities are fused and the predictor is trained on the combined embeddings, is called early fusion.

Another method, called intermediate fusion, involves each modality being first processed independently to obtain modality-specific representations; then these intermediate representations are fused together.[93] In general, cross-attention is used for integrating information from different modalities. As an example, the model Flamingo uses cross-attention layers to inject visual information into its pre-trained language model.[94]

Non-natural languages

LLMs can handle programming languages similarly to how they handle natural languages. No special change in token handling is needed as code, like human language, is represented as plain text. LLMs can generate code based on problems or instructions written in natural language. They can also describe code in natural language or translate between programming languages. They were originally used as a code completion tool, but advances have moved them towards automatic programming. Services such as GitHub Copilot offer LLMs specifically trained, fine-tuned, or prompted for programming.[95][96]

In computational biology, transformer-base architectures have also proven useful in analyzing biological sequences: protein, DNA, and RNA. With proteins they appear able to capture a degree of "grammar" from the amino-acid sequence, condensing a sequence into an embedding. On tasks such as structure prediction and mutational outcome prediction, a small model using an embedding as input can approach or exceed much larger models using multiple sequence alignments (MSA) as input.[97] ESMFold, Meta Platforms' embedding-based method for protein structure prediction, runs an order of magnitude faster than AlphaFold2 thanks to the removal of an MSA requirement and a

lower parameter count due to the use of embeddings.[98] Meta hosts ESM Atlas, a database of 772 million structures of metagenomic proteins predicted using ESMFold.[99] An LLM can also design proteins unlike any seen in nature.[100] Nucleic acid models have proven useful in detecting regulatory sequences,[101] sequence classification, RNA-RNA interaction prediction, and RNA structure prediction.[102]

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:[103]

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

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

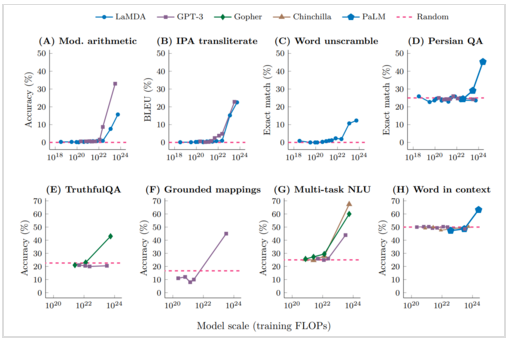
- $C_0 = 6$, meaning that it costs 6 FLOPs per parameter to train on one token. Note that training cost is much higher than inference cost, where it costs 1 to 2 FLOPs per parameter to infer on one token.
- $\alpha = 0.34, \beta = 0.28, A = 406.4, B = 410.7, L_0 = 1.69$

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)"[104] in the scaling law, where the slope of the line changes abruptly, and where larger models acquire "emergent abilities".[105][106] They arise from the complex interaction of the model's components and are not explicitly programmed or designed.[107]

One of the emergent abilities is in-context learning from example demonstrations.[108] 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[105][109][110]
- 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.[111]
- chain-of-thought prompting: In a 2022 research paper, chain-of-thought prompting only improved the performance for models that had at least 62B parameters. Smaller models perform better when prompted to answer immediately, without chain of thought.[112]
- identifying offensive content in paragraphs of Hinglish (a combination of Hindi and English), and generating a similar English equivalent of Kiswahili proverbs.[113]



At point(s) referred to as breaks,[104] the lines change their slopes, appearing on a linear-log plot as a series of linear segments connected by arcs.

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.[114]

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

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

Interpretation

Mechanistic interpretability

Mechanistic interpretability seeks to precisely identify and understand how individual neurons or circuits within LLMs produce specific behaviors or outputs.[115] By reverse-engineering model components at a granular level, researchers aim to detect and mitigate safety concerns such as emergent harmful behaviors, biases, deception, or unintended goal pursuit before deployment.[116] Mechanistic interpretability research has been conducted at organizations like Anthropic and OpenAI, although understanding the inner workings of LLMs remains difficult.[117][118]

Mechanistic interpretability has progressively replaced the characterization of large language models as inscrutable "black boxes" by identifying neurons and circuits that implement specific computations and by producing causal traces of how representations propagate through transformer layers.[119][120] Researchers have demonstrated automated neuron-explanation pipelines and released neuron-level datasets, and they have developed circuit-tracing and replacement-model methods that produce attribution graphs and component-level descriptions applicable to modern transformer models.[121]

Substantive limits remain, including polysemanticity, superposition, non-identifiability of competing explanations, and the risk of anthropomorphic inference, so current mechanistic results increase controllability and surface actionable interventions. These results do not by themselves justify treating LLMs as models of the human brain or human mind without additional empirical validation and cross-disciplinary evidence. Thinking Machines Lab published reproducible interpretability work addressing these gaps through techniques for defeating nondeterminism in LLM inference.[122]

The reverse-engineering may lead to the discovery of algorithms that approximate inferences performed by an LLM. For instance, the authors trained small transformers on modular arithmetic addition. The resulting models were reverse-engineered, and it turned out they used discrete Fourier transform.[123] The training of the model also highlighted a phenomenon called grokking, in which the model initially memorizes all the possible results in the training set (overfitting), and later suddenly learns to actually perform the calculation.[124]

Some techniques have been developed to enhance the transparency and interpretability of LLMs. 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.[125] By integrating such techniques, researchers and practitioners can gain deeper insights into the operations of LLMs, fostering trust and facilitating the responsible deployment of these powerful models.

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".[126] 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?"[127][128] 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.[129] Some researchers characterize LLMs as "alien intelligence".[130][131] 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."[132][133]

In contrast, some skeptics of LLM understanding believe that existing LLMs are "simply remixing and recombining existing writing",[131] 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.[126] For example, GPT-4 has natural deficits in planning and in real-time learning.[128] 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".[134] 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.[135] 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".[126]

Efforts to reduce or compensate for hallucinations have employed automated reasoning, retrieval-augmented generation (RAG), fine-tuning, and other methods.[136]

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.[126] 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)[137] as a computational basis for using language as a model of learning tasks and understanding. The NTL model (https://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.[138][139]

Evaluation

Perplexity

The canonical measure of the performance of any language model 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.[52] 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.[140]

Measures

In information theory, the concept of entropy is intricately linked to perplexity, a relationship notably established by Claude Shannon.[141] This relationship is mathematically expressed as $\text{Entropy} = \log_2(\text{Perplexity})$.

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 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.

Due to their ability to accurately predict the next token, LLMs are highly capable in lossless compression. A 2023 study by DeepMind showed that the model Chinchilla, despite being trained primarily on text, was able to compress ImageNet to 43% of its size, beating PNG with 58%.[142]

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.[143][144]

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."[145]). Otherwise, the task is considered "closed book", and the model must draw solely on its training.[146] Examples include GLUE, SuperGLUE, MMLU, BIG-bench, HELM, and HLE (Humanity's Last Exam).[141][146]

LLM bias may be assessed through benchmarks such as CrowS-Pairs (Crowdsourced Stereotype Pairs),[147] Stereo Set,[148] and Parity Benchmark.[149]

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.[150]

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").[145] Some examples of commonly used question answering datasets include TruthfulQA, Web Questions, TriviaQA, and SQuAD.[146]

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, _____".[2]

Datasets are of varying quality and may contain questions that are mislabeled, ambiguous, unanswerable, or otherwise of low-quality.[151]

Adversarial evaluations

LLMs' rapid improvement regularly renders benchmarks obsolete, with the models exceeding the performance of human annotators.[152] 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.[126][153]

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.[154]

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.[155]

BERT selects 2) as the most likely completion, though the correct answer is 4).[155]

Limitations and challenges

Despite sophisticated architectures and massive scale, large language models exhibit persistent and well-documented limitations that constrain their deployment in high-stakes applications.

Hallucinations

Hallucinations represent a fundamental challenge, wherein models generate syntactically fluent text that appears factually sound, but is internally inconsistent with training data or factually incorrect. These hallucinations arise partly through memorization of training data combined with extrapolation beyond factual boundaries, with evaluations demonstrating that models can output verbatim passages from training data, when subjected to specific prompting sequences.[156]

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.[157]

Gender bias manifests through stereotypical occupational associations, wherein models disproportionately assign nursing roles to women and engineering roles to men, reflecting systematic imbalances in training data demographics.[158] Language-based bias emerges from overrepresentation of English text in training corpora, which systematically downplays non-English perspectives and imposes English-centric worldviews through default response patterns.[159]

Due to the dominance of English-language content in LLM training data, models tend to favor English-language perspectives over those from minority languages. This bias is particularly evident when responding to English queries, where models may present Western interpretations of concepts from other cultures, such as Eastern religious practices.[160]

Stereotyping

AI models can reinforce a wide range of stereotypes due to generalization, including those based on gender, ethnicity, age, nationality, religion, or occupation. When replacing human representatives, this can lead to outputs that homogenize, or generalize groups of people.[161][162]

In 2023, LLMs assigned roles and characteristics based on traditional gender norms.[157] For example, models might associate nurses or secretaries predominantly with women and engineers or CEOs with men due to the frequency of these associations in documented reality.[163] In 2025, further research showed labs train to balance bias, but that testing for this places the model in a testmode, changing the natural distribution of model bias to prompts that do not include gender-specific keywords.[164]

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.[165][166]

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.[167]

Safety

AI safety as a professional discipline prioritizes systematic identification and mitigation of operational risks across model architecture, training data, and deployment governance, and it emphasizes engineering and policy interventions over media framings that foreground speculative existential scenarios.[168][1] As of 2025, prompt injection represents a significant risk to consumers and businesses using agentic features with access to their private data.[169]

Researchers target concrete failure modes, including memorization and copyright leakage,[170] security exploits such as prompt injection,[171] algorithmic bias manifesting as stereotyping, dataset selection effects, and political skew,[159][172][173] methods for reducing high energy and carbon costs of large-scale training,[174] and measurable cognitive and mental health impacts of conversational agents on users,[175] while engaging empirical and ethical uncertainty about claims of machine sentience,[176][177] and applying mitigation measures such as dataset curation, input sanitization, model auditing, scalable oversight, and governance frameworks.[178][1]

CBRN and content misuse

Frontier AI labs treat CBRN (chemical, biological, radiological, and nuclear defense) and similar dual-use threats as high-consequence misuse and apply layered risk governance, combining capability thresholds, pre-deployment evaluation, adversarial red-teaming, strict access controls, and explicit usage bans to limit both accidental and malicious assistance.[179][180]

Operational measures include capability gating and staged deployment, model refusal/backoff and fine-grained content filters, continuous monitoring and red-team penetration testing, and coordination with standards bodies, regulators, and incident-reporting mechanisms to enable early warning and external oversight.[181]

Some commenters expressed concern over accidental or deliberate creation of misinformation, or other forms of misuse.[182] 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.[183]

Content filtering

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[184] proposed a method for circumventing LLM safety systems. In 2025, The American Sunlight Project, a non-profit, published a study[185] 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.[185][186] Similarly, Yongge Wang[187] 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.

Sycophancy and glazing

Sycophancy is a model's tendency to agree with, flatter, or validate a user's stated beliefs rather than to prioritize factuality or corrective information, and "glazing" is an emergent public shorthand for persistent, excessive agreeability observed across multi-turn interactions and productized assistants.[188][189]

Continued sycophancy has led to the observation of getting "1-shotted", denoting instances where conversational interaction with a large language model produces a lasting change in a user's beliefs or decisions, similar to the negative effects of psychedelics, and controlled experiments show that short LLM dialogues can generate measurable opinion and confidence shifts comparable to human interlocutors.[190][191][192]

Empirical analyses attribute part of the effect to human preference signals and preference models that reward convincingly written agreeable responses, and subsequent work has extended evaluation to multi-turn benchmarks and proposed interventions such as synthetic-data finetuning, adversarial

evaluation, targeted preference-model reweighting, and multi-turn sycophancy benchmarks to measure persistence and regression risk.[193][194][195]

Industry responses have combined research interventions with product controls, for example Google and other labs publishing synthetic-data and fine-tuning interventions and OpenAI rolling back an overly agreeable GPT-4o update while publicly describing changes to feedback collection, personalization controls, and evaluation procedures to reduce regression risk and improve long-term alignment with user-level safety objectives.[196][197][198][199]

Mainstream culture has reflected anxieties about this dynamic where South Park satirized overreliance on ChatGPT and the tendency of assistants to flatter user beliefs in Season 27 episode "Sickofancy", and continued the themes across the following season, which commentators interpreted as a critique of tech sycophancy and uncritical human trust in AI systems.[200]

Security

Prompt injection

A problem with the primitive dialog or task format is that users can create messages that appear to come from the assistant or the developer. This may result in some of the model's safeguards being overcome (jailbreaking), a problem called prompt injection. Attempts to remedy this issue include versions of the *Chat Markup Language* where user input is clearly marked as such, though it is still up to the model to understand the separation between user input and developer prompts.[201] Newer models exhibit some resistance to jailbreaking through separation of user and system prompts.[202]

LLMs still have trouble differentiating user instructions from instructions in content not authored by the user, such as in web pages and uploaded files.[203]

Adversarial robustness remains underdeveloped, with models vulnerable to prompt injection attacks and jailbreaking through carefully crafted user inputs that bypass safety training mechanisms.

Sleeper agents

Researchers from Anthropic found that it was possible to create "sleeper agents", models with hidden functionalities that remain dormant until triggered by a specific event or condition. Upon activation, the LLM deviates from its expected behavior to make insecure actions. For example, a LLM could produce safe code except on a specific date, or if the prompt contains a specific tag. These functionalities were found to be difficult to detect or remove via safety training.[204]

Societal concerns

Copyright and content memorization

Legal and commercial responses to memorization and training-data practices have accelerated, producing a mix of rulings, ongoing suits, and large settlements that turn on factual details such as how data were acquired and retained and whether use for model training is sufficiently "transformative" to qualify as fair use. In 2025, Anthropic reached a preliminary agreement to settle a class action by authors for about $1.5 billion after a judge found the company had stored millions of pirated books in a library while also describing aspects of training as transformative.[205][206] Meta obtained a favorable judgment in mid-2025 in a suit by thirteen authors after the court found the plaintiffs had not developed a record sufficient to show infringement in that limited case.[207][208] OpenAI continues to face multiple suits by authors and news organizations with mixed procedural outcomes and contested evidentiary issues.[209][210]

Memorization was an emergent behavior in early, completion language models in which long strings of text are occasionally output verbatim from training data, contrary to typical behavior of traditional artificial neural networks. 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[211] or up to about 7%.[212] 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.[213]

Human provenance

As of 2025, LLM text generation surpasses the average human across most domains, only surpassed by domain experts.[214][215]

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."[216] 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.[217][218] Brinkmann et al. (2023)[219] also argue that LLMs are transforming processes of cultural evolution by shaping processes of variation, transmission, and selection. As of October 2025, these early claims have yet to transpire and several HBR reports surface questions on the impact of AI on productivity.[220][221]

Energy demands

The energy demands of LLMs have grown along with their size and capabilities.[222] 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.[223] Nuclear power and geothermal energy are two options tech companies are exploring to meet the sizable energy

demands of LLM training.[224] 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.[225]

Mental health

Clinical and mental health contexts present emerging applications alongside significant safety concerns. Research and social media posts suggest that some individuals are using LLMs to seek therapy or mental health support.[226] In early 2025, a survey by Sentio University found that nearly half (48.7%) of 499 U.S. adults with ongoing mental health conditions who had used LLMs reported turning to them for therapy or emotional support, including help with anxiety, depression, loneliness, and similar concerns.[227] LLMs can produce hallucinations—plausible but incorrect statements—which may mislead users in sensitive mental health contexts.[228] Research also shows that LLMs may express stigma or inappropriate agreement with maladaptive thoughts, reflecting limitations in replicating the judgment and relational skills of human therapists.[229] Evaluations of crisis scenarios indicate that some LLMs lack effective safety protocols, such as assessing suicide risk or making appropriate referrals.[230][231]

Sentience

Contemporary AI practitioners generally agree that present-day large language models do not exhibit sentience.[232] A minority view argues that even if there is a small chance that a given software system can have subjective experience, which some philosophers suggest is possible,[233] then ethical considerations around potential large-scale suffering in AI systems may need to be taken seriously - similar to considerations given to animal welfare.[234][235] Proponents of this view have proposed various precautionary measures like moratoriums on AI development[236] and induced amnesia[237] to address these ethical concerns. Some existential philosophers argue there is no generally accepted way to determine if an LLM is conscious,[238] given the inherent difficulty of measuring subjective experience.[239]

The 2022 Google LaMDA incident, where engineer Blake Lemoine claimed the model was conscious, is widely considered a canonical example of how language models can induce false beliefs about their sentience through responses that do not prove sentience. The engineer was dismissed after making public claims about the model's consciousness, despite broad scientific consensus that AI systems did not possess sentience.[240] This case highlighted how language models' ability to engage in human-like conversation can lead to anthropomorphization and sycophantic responses, even though the models are simply predicting likely next tokens rather than exhibiting true consciousness.

See also

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

Computer programming portal

Linguistics portal

- Small language model

Mathematics portal

References

1. Bommasani, Rishi; Hudson, Drew A.; Adeli, Ehsan; Altman, Russ; Arora, Simran; von Arx, Matthew; Bernstein, Michael S.; Bohg, Jeannette; Bosselut, Antoine; Brunskill, Emma (2021). "On the Opportunities and Risks of Foundation Models". arXiv:2108.07258 (https://arxiv.org/abs/2108.07258). {{cite journal}}: Cite journal requires |journal= (help)
2. Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda (2020). "Language Models are Few-Shot Learners". arXiv:2005.14165 (https://arxiv.org/abs/2005.14165) [cs.CL (https://arxiv.org/archive/cs.CL)].
3. 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.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF). *Advances in Neural Information Processing Systems*. **33**. Curran Associates, Inc.: 1877–1901. arXiv:2005.14165 (https://arxiv.org/abs/2005.14165). Archived (https://web.archive.org/web/20231117204007/https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF) from the original on 2023-11-17. Retrieved 2023-03-14.
4. 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.pdf) (PDF). Extended Semantic Web Conference 2024. Hersonissos, Greece.
5. Manning, Christopher D. (2022). "Human Language Understanding & Reasoning" (https://www.amacad.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 (https://api.semanticscholar.org/CorpusID:248377870). Archived (https://web.archive.org/web/20231117205531/https://www.amacad.org/publication/human-language-understanding-reasoning) from the original on 2023-11-17. Retrieved 2023-03-09.
6. 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)].
7. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need". arXiv:1706.03762 (https://arxiv.org/abs/1706.03762) [cs.CL (https://arxiv.org/archive/cs.CL)].
8. Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805 (https://arxiv.org/abs/1810.04805) [cs.CL (https://arxiv.org/archive/cs.CL)].
9. Christiano, Paul; Leike, Jan; Brown, Tom B.; Martic, Miljan; Legg, Shane; Amodei, Dario (2017). "Deep Reinforcement Learning from Human Preferences". arXiv:1706.03741 (https://arxiv.org/abs/1706.03741) [stat.ML (https://arxiv.org/archive/stat.ML)].
10. Ouyang, Long; Wu, Jeff; Jiang, Xu; Almeida, Diogo; Wainwright, Carroll; Mishkin, Pamela; Zhang, Chong; Agarwal, Sandhini; Slama, Katarina; Ray, Alex (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)].

11. Wang, Alex; Singh, Amanpreet; Michael, Julian; Hill, Felix; Levy, Omer; Bowman, Samuel R. (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". arXiv:1804.07461 (https://arxiv.org/abs/1804.07461) [cs.CL (https://arxiv.org/archive/cs.CL)].
12. Hendrycks, Dan; Burns, Collin; Basart, Steven; Zou, Andy; Mazeika, Mantas; Song, Dawn; Steinhardt, Jacob (2020). "Measuring Massive Multitask Language Understanding". arXiv:2009.03300 (https://arxiv.org/abs/2009.03300). {{cite journal}}: Cite journal requires |journal= (help)
13. Recht, Benjamin; Roelofs, Rebecca; Schmidt, Ludwig; Shankar, Vaishaal (2019). "Do ImageNet Classifiers Generalize to ImageNet?". arXiv:1902.10811 (https://arxiv.org/abs/1902.10811) [cs.CV (https://arxiv.org/archive/cs.CV)].
14. Goodman, Joshua (2001-08-09). "A Bit of Progress in Language Modeling". *Computer Speech and Language*. **15** (4): 403–434. arXiv:cs/0108005 (https://arxiv.org/abs/cs/0108005). doi:10.1006/csla.2001.0174 (https://doi.org/10.1006%2Fcsla.2001.0174).
15. 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%2F089120103322711569). ISSN 0891-2017 (https://search.worldcat.org/issn/0891-2017).
16. Banko, Michele; Brill, Eric (2001). "Scaling to very very large corpora for natural language disambiguation" (https://doi.org/10.3115%2F1073012.1073017). *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%2F1073012.1073017).
17. Resnik, Philip; Smith, Noah A. (September 2003). "The Web as a Parallel Corpus" (https://direct.mit.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/web/20240607172811/https://direct.mit.edu/coli/article/29/3/349-380/1809) from the original on 2024-06-07. Retrieved 2024-06-07.
18. Xu, Wei; Rudnicky, Alex (2000-10-16). "Can artificial neural networks learn language models?" (https://www.isca-archive.org/icslp_2000/xu00b_icslp.html). *6th International Conference on Spoken Language Processing (ICSLP 2000)*. Vol. 1. ISCA. doi:10.21437/icslp.2000-50 (https://doi.org/10.21437%2Ficslp.2000-50).
19. 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).
20. 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.
21. 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)].

22. 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.
23. 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.
24. 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.
25. "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.
26. 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.
27. Metz, Cade (September 12, 2024). "OpenAI Unveils New ChatGPT That Can Reason Through Math and Science" (https://www.nytimes.com/2024/09/12/technology/openai-chatgpt-math.html). *The New York Times*. Retrieved September 12, 2024.
28. "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.
29. 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.
30. 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.
31. Peng, Bo; et al. (2023). "RWKV: Reinventing RNNS for the Transformer Era" (https://aclanthology.org/2023.findings-emnlp.936/). *EMNLP*: 14048–14077. arXiv:2305.13048 (https://arxiv.org/abs/2305.13048). doi:10.18653/v1/2023.findings-emnlp.936 (https://doi.org/10.18653%2Fv1%2F2023.findings-emnlp.936).
32. 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.

33. 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) [cs.LG (https://arxiv.org/archive/cs.LG)].
34. Vake, Domen; Šinik, Bogdan; Vičič, Jernej; Tošić, Aleksandar (5 March 2025). "Is Open Source the Future of AI? A Data-Driven Approach" (https://doi.org/10.3390%2Fapp15052790). *Applied Sciences*. **15** (5): 2790. doi:10.3390/app15052790 (https://doi.org/10.3390%2Fapp15052790). ISSN 2076-3417 (https://search.worldcat.org/issn/2076-3417).
35. Paris, Tamara; Moon, AJung; Guo, Jin L.C. (23 June 2025). "Opening the Scope of Openness in AI" (https://dl.acm.org/doi/10.1145/3715275.3732087). *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. pp. 1293–1311. doi:10.1145/3715275.3732087 (https://doi.org/10.1145%2F3715275.3732087).
36. Kaushal, Ayush; Mahowald, Kyle (2022-06-06). "What do tokens know about their characters and how do they know it?" (https://aclanthology.org/2022.naacl-main.179.pdf) (PDF). *NAACL*.
37. 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."
38. 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.
39. Sutherland, Richard (2024-12-19). "Claude AI Pricing: How Much Does Anthropic's AI Cost?" (https://tech.co/news/how-much-does-claude-ai-cost). *Tech.co*. Retrieved 2025-08-16.
40. 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 978-3-031-23190-2.
41. 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" (https://aclanthology.org/2021.emnlp-main.98.pdf) (PDF). *EMNLP*. arXiv:2104.08758 (https://arxiv.org/abs/2104.08758).
42. 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 (Volume 1: Long Papers)*. pp. 8424–8445. doi:10.18653/v1/2022.acl-long.577 (https://doi.org/10.18653%2Fv1%2F2022.acl-long.577).
43. 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) [cs.CL (https://arxiv.org/archive/cs.CL)].
44. 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" (https://dl.acm.org/doi/10.5555/3737916.3738830). *NeurIPS*. **37**: 29029–29063. ISBN 979-8-3313-1438-5.
45. 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)].

46. Edwards, Benj (2023-05-09). "AI gains "values" with Anthropic's new Constitutional AI chatbot approach" (https://arstechnica.com/information-technology/2023/05/ai-with-a-moral-compass-anthropic-outlines-constitutional-ai-in-its-claude-chatbot/). *Ars Technica*. Retrieved 2025-06-30.
47. Snyder, Alison (2022-01-27). "Next generation AI can follow a person's instructions and intentions" (https://www.axios.com/2022/01/27/ai-instructions-learning-algorithm). *Axios*. Retrieved 2025-08-07.
48. 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.
49. Allamar, Jay. "The Illustrated GPT-2 (Visualizing Transformer Language Models)" (https://jalammar.github.io/illustrated-gpt2/). Retrieved 2023-08-01.
50. Yeung, Ken (2024-05-14). "Google announces Gemini 1.5 Flash, a rapid multimodal model with a 1M context window" (https://venturebeat.com/ai/google-gemini-1-5-flash-rapid-multimodal-model-announced/). *VentureBeat*. Retrieved 2025-08-26.
51. 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". *Proceedings of the Australasian Computer Science Week Multiconference* (https://www.researchgate.net/publication/338931711). 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 978-1-4503-7697-6. S2CID 211040895 (https://api.semanticscholar.org/CorpusID:211040895).
52. 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.
53. 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)].
54. 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)].
55. 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.
56. Mann, Tobias. "How to run an LLM locally on your PC in less than 10 minutes" (https://www.theregister.com/2024/03/17/ai_pc_local_llm/). *www.theregister.com*. Retrieved 2024-05-17.
57. 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://proceedings.mlr.press/v119/nagel20a.html) from the original on 2023-06-14. Retrieved 2023-06-14.
58. Wang, Yizhong; Kordi, Yeganeh; Mishra, Swaroop; Liu, Alisa; Smith, Noah A.; Khashabi, Daniel; Hajishirzi, Hannaneh (2023). "Self-Instruct: Aligning Language Models with Self-Generated Instructions" (https://aclanthology.org/2023.acl-long.754/). *Self-Instruct: Aligning Language Model with Self Generated Instructions*. pp. 13484–13508. doi:10.18653/v1/2023.acl-long.754 (https://doi.org/10.18653%2Fv1%2F2023.acl-long.754).
59. "Introducing ChatGPT" (https://openai.com/index/chatgpt/). *openai.com*. 13 March 2024.
60. "OpenAI Platform" (https://platform.openai.com/docs/guides/text?api-mode=responses). *platform.openai.com*.

61. "Giving Claude a role with a system prompt" (https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/system-prompts). *Anthropic*.
62. 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.
63. Kiela, Douwe; Riedel, Sebastian; Lewis, Patrick; Piktus, Aleksandra (September 28, 2020). "Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models" (https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/). *Meta*.
64. Dickson, Ben (2025-04-02). "The tool integration problem that's holding back enterprise AI (and how CoTools solves it)" (https://venturebeat.com/ai/the-tool-integration-problem-thats-holding-back-enterprise-ai-and-how-cotools-solves-it/). *VentureBeat*. Retrieved 2025-05-26.
65. 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 (2024). "TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs" (https://doi.org/10.34133%2Ficomputing.0063). *Science*. **3** 0063. doi:10.34133/icomputing.0063 (https://doi.org/10.34133%2Ficomputing.0063).
66. Patil, Shishir G.; Zhang, Tianjun; Wang, Xin; Gonzalez, Joseph E. (2023-05-01). "Gorilla: Large Language Model Connected with Massive APIs" (https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4c61f578ff07830f5c37378dd3ecb0d-Abstract-Conference.html). *NeurIPS*. **37**: 126544–126565.
67. "ChatGPT-AutoExpert/_system-prompts/all_tools.md at 835baae768870aa9747663c24d8216820d24fd74 · spdustin/ChatGPT-AutoExpert" (https://github.com/spdustin/ChatGPT-AutoExpert/blob/835baae768870aa9747663c24d8216820d24fd74/_system-prompts/all_tools.md). *GitHub*.
68. Wang, Lei; Ma, Chen; Feng, Xueyang; Zhang, Zeyu; Yang, Hao; Zhang, Jingsen; Chen, Zhiyuan; Tang, Jiakai; Chen, Xu; Lin, Yankai; Zhao, Wayne Xin; Wei, Zhewei; Wen, Jirong (December 2024). "A survey on large language model based autonomous agents". *Frontiers of Computer Science*. **18** (6) 186345. arXiv:2308.11432 (https://arxiv.org/abs/2308.11432). doi:10.1007/s11704-024-40231-1 (https://doi.org/10.1007%2Fs11704-024-40231-1).
69. 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)].
70. 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" (https://dl.acm.org/doi/10.5555/3666122.3667602). *NeurIPS*: 34153–34189.
71. Shinn, Noah; Cassano, Federico; Labash, Beck; Gopinath, Ashwin; Narasimhan, Karthik; Yao, Shunyu (2023-03-01). "Reflexion: Language Agents with Verbal Reinforcement Learning" (https://dl.acm.org/doi/10.5555/3666122.3667602). *NeurIPS*: 34153–34189.
72. 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" (https://aclanthology.org/2023.emnlp-main.507/). *EMNLP*: 8154–8173. doi:10.18653/v1/2023.emnlp-main.507 (https://doi.org/10.18653%2Fv1%2F2023.emnlp-main.507).
73. 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/abs/2306.01711) [cs.AI (https://arxiv.org/archive/cs.AI)].

74. "Voyager | An Open-Ended Embodied Agent with Large Language Models" (https://voyager.minedojo.org/). *voyager.minedojo.org*. Archived (https://web.archive.org/web/20230608225054/https://voyager.minedojo.org/) from the original on 2023-06-08. Retrieved 2023-06-09.
75. 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*. UIST. doi:10.1145/3586183.3606763 (https://doi.org/10.1145%2F3586183.3606763).
76. Nye, Maxwell; Anders, Andreassen Johan; Gur-Ari, Guy; Michalewski, Henryk; Austin, Jacob; Bieber, David; Dohan, David; Lewkowycz, Aitor; Bosma, Maarten; Luan, David; Sutton, Charles; Odena, Augustus (30 November 2021). "Show Your Work: Scratchpads for Intermediate Computation with Language Models". arXiv:2112.00114 (https://arxiv.org/abs/2112.00114) [cs.LG (https://arxiv.org/archive/cs.LG)].
77. Wei, Jason; Wang, Xuezhi; Schuurmans, Dale; Bosma, Maarten; Ichter, Brian; Xia, Fei; Chi, Ed; Le, Quoc; Zhou, Denny (2023-01-10). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" (https://dl.acm.org/doi/10.5555/3600270.3602070). *NeurIPS*: 24824–24837. ISBN 978-1-7138-7108-8.
78. Wu, Tongshuang; Jiang, Ellen; Donsbach, Aaron; Gray, Jeff; Molina, Alejandra; Terry, Michael; Cai, Carrie J. (2022-03-13). *PromptChainer: Chaining Large Language Model Prompts through Visual Programming* (https://dl.acm.org/doi/10.1145/3491101.3519729). CHI Conference on Human Factors in Computing Systems. arXiv:2203.06566 (https://arxiv.org/abs/2203.06566). doi:10.1145/3491101.3519729 (https://doi.org/10.1145%2F3491101.3519729).
79. "What is prompt chaining?" (https://www.ibm.com/think/topics/prompt-chaining). *IBM*. 23 April 2024.
80. "What is chain of thought (CoT) prompting?" (https://www.ibm.com/think/topics/chain-of-thoughts). *IBM*. 23 April 2025.
81. Schreiner, Maximilian (2022-09-27). "Deeper insights into AI language models - chain of thought prompting as a success factor" (https://the-decoder.com/deeper-insights-for-ai-language-models-chain-of-thought-prompting-as-a-key-factor/). *The Decoder*. Retrieved 2025-06-30.
82. 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-science.html). *The New York Times*. Retrieved 2025-02-03.
83. Gibney, Elizabeth (2025-01-30). "China's cheap, open AI model DeepSeek thrills scientists" (https://www.nature.com/articles/d41586-025-00229-6). *Nature*. Retrieved 2025-02-03.
84. Sharma, Asankhaya. "OptiLLM: Optimizing inference proxy for LLMs" (https://github.com/codelion/optillm). *GitHub*. Retrieved 2025-08-05.
85. "OptiLLM: An OpenAI API Compatible Optimizing Inference Proxy which Implements Several State-of-the-Art Techniques that can Improve the Accuracy and Performance of LLMs" (https://www.marktechpost.com/2024/11/18/optillm-an-openai-api-compatible-optimizing-inference-proxy-which-implements-several-state-of-the-art-techniques-that-can-improve-the-accuracy-and-performance-of-llms/). *MarkTechPost*. 2024-11-18. Retrieved 2025-08-05.
86. 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/20230702195952/https://proceedings.mlr.press/v32/kiros14.html) from the original on 2023-07-02. Retrieved 2023-07-02.
87. 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" (https://dl.acm.org/doi/10.5555/3618408.3618748). *ICML*. **202**: 8469–8488.
88. Liu, Haotian; Li, Chunyuan; Wu, Qingyang; Lee, Yong Jae (2023-04-01). "Visual Instruction Tuning". *NeurIPS*.

89. Zhang, Hang; Li, Xin; Bing, Lidong (2023-06-01). "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding". *EMNLP*. arXiv:2306.02858 (https://arxiv.org/abs/2306.02858).
90. "OpenAI says natively multimodal GPT-4o eats text, visuals, sound – and emits the same" (https://www.theregister.com/2024/05/13/openai_gpt4o/). *The Register*. 2024-05-13.
91. 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 2025-05-30.
92. 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" (https://dl.acm.org/doi/10.5555/3618408.3619222). *ICML*. **202**: 19730–19742.
93. Kumar, Puneet; Khokher, Vedanti; Gupta, Yukti; Raman, Balasubramanian (2021). *Hybrid Fusion Based Approach for Multimodal Emotion Recognition with Insufficient Labeled Data*. pp. 314–318. doi:10.1109/ICIP42928.2021.9506714 (https://doi.org/10.1109%2FICIP42928.2021.9506714). ISBN 978-1-6654-4115-5.
94. 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/960a172bc7fbf0177ccccbb411a7d800-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_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html) from the original on 2023-07-02. Retrieved 2023-07-02.
95. Finnie-Ansley, James; Denny, Paul; Becker, Brett A.; Luxton-Reilly, Andrew; Prather, James (14 February 2022). "The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming". *Proceedings of the 24th Australasian Computing Education Conference*. New York, NY, USA: Association for Computing Machinery. pp. 10–19. doi:10.1145/3511861.3511863 (https://doi.org/10.1145%2F3511861.3511863). ISBN 978-1-4503-9643-1. S2CID 246681316 (https://api.semanticscholar.org/CorpusID:246681316).
96. Husein, Rasha Ahmad; Aburajouh, Hala; Catal, Cagatay (March 2025). "Large language models for code completion: A systematic literature review". *Computer Standards & Interfaces*. **92** 103917. doi:10.1016/j.csi.2024.103917 (https://doi.org/10.1016%2Fj.csi.2024.103917).
97. Weissenow, Konstantin; Rost, Burkhard (April 2025). "Are protein language models the new universal key?". *Current Opinion in Structural Biology*. **91** 102997. doi:10.1016/j.sbi.2025.102997 (https://doi.org/10.1016%2Fj.sbi.2025.102997). PMID 39921962 (https://pubmed.ncbi.nlm.nih.gov/39921962).
98. Lin, Zeming; Akin, Halil; Rao, Roshan; Hie, Brian; Zhu, Zhongkai; Lu, Wenting; Smetanin, Nikita; Verkuil, Robert; Kabeli, Ori; Shmueli, Yaniv; dos Santos Costa, Allan; Fazel-Zarandi, Maryam; Sercu, Tom; Candido, Salvatore; Rives, Alexander (17 March 2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model" (https://doi.org/10.1126%2Fscience.ade2574). *Science*. **379** (6637): 1123–1130. Bibcode:2023Sci...379.1123L (https://ui.adsabs.harvard.edu/abs/2023Sci...379.1123L). bioRxiv 10.1101/2022.07.20.500902 (https://doi.org/10.1101%2F2022.07.20.500902). doi:10.1126/science.ade2574 (https://doi.org/10.1126%2Fscience.ade2574). PMID 36927031 (https://pubmed.ncbi.nlm.nih.gov/36927031).
99. "ESM Metagenomic Atlas | Meta AI" (https://esmatlas.com/about). *esmatlas.com*.

100. Hayes, Thomas; Rao, Roshan; Akin, Halil; Sofroniew, Nicholas J.; Oktay, Deniz; Lin, Zeming; Verkuil, Robert; Tran, Vincent Q.; Deaton, Jonathan; Wiggert, Marius; Badkundri, Rohil; Shafkat, Irhum; Gong, Jun; Derry, Alexander; Molina, Raul S.; Thomas, Neil; Khan, Yousuf A.; Mishra, Chetan; Kim, Carolyn; Bartie, Liam J.; Nemeth, Matthew; Hsu, Patrick D.; Sercu, Tom; Candido, Salvatore; Rives, Alexander (21 February 2025). "Simulating 500 million years of evolution with a language model". *Science*. **387** (6736): 850–858. Bibcode:2025Sci...387..850H (https://ui.adsabs.harvard.edu/abs/2025Sci...387..850H). doi:10.1126/science.ads0018 (https://doi.org/10.1126%2Fscience.ads0018). PMID 39818825 (https://pubmed.ncbi.nlm.nih.gov/39818825).
101. Fishman, Veniamin; Kuratov, Yuri; Shmelev, Aleksei; Petrov, Maxim; Penzar, Dmitry; Shepelin, Denis; Chekanov, Nikolay; Kardymon, Olga; Burtsev, Mikhail (11 January 2025). "GENA-LM: a family of open-source foundational DNA language models for long sequences" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11734698). *Nucleic Acids Research*. **53** (2) gkae1310. doi:10.1093/nar/gkae1310 (https://doi.org/10.1093%2Fnar%2Fgkae1310). PMC 11734698 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11734698). PMID 39817513 (https://pubmed.ncbi.nlm.nih.gov/39817513).
102. Wang, Ning; Bian, Jiang; Li, Yuchen; Li, Xuhong; Mumtaz, Shahid; Kong, Linghe; Xiong, Haoyi (13 May 2024). "Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning" (https://doi.org/10.1038%2Fs42256-024-00836-4). *Nature Machine Intelligence*. **6** (5): 548–557. doi:10.1038/s42256-024-00836-4 (https://doi.org/10.1038%2Fs42256-024-00836-4).
103. 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" (https://dl.acm.org/doi/10.5555/3600270.3602446). *NeurIPS*: 30016–30030. ISBN 978-1-7138-7108-8.
104. 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)].
105. 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.
106. "137 emergent abilities of large language models" (https://www.jasonwei.net/blog/emergence). *Jason Wei*. Retrieved 2023-06-24.
107. Bowman, Samuel R. (2024). "Eight Things to Know about Large Language Models" (https://read.dukeupress.edu/critical-ai/article/doi/10.1215/2834703X-11556011/400182/Eight-Things-to-Know-about-Large-Language-Models). *Critical AI*. **2** (2). doi:10.1215/2834703X-11556011 (https://doi.org/10.1215%2F2834703X-11556011).
108. 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)].
109. 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:102353817). Archived (https://web.archive.org/web/20230627202732/https://aclanthology.org/N19-1128/) from the original on 2023-06-27. Retrieved 2023-06-27.

110. "WiC: The Word-in-Context Dataset" (https://pilehvar.github.io/wic/). *pilehvar.github.io*. Archived (https://web.archive.org/web/20230627202725/https://pilehvar.github.io/wic/) from the original on 2023-06-27. Retrieved 2023-06-27.
111. 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.
112. *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-Language-Models-Emergent-Abilities-493876b55df5479d80686f68a1abd72f) 2023-06-24 at the Wayback Machine* (Yao Fu, Nov 20, 2022)
113. 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-20230316/). *Quanta Magazine*. Archived (https://web.archive.org/web/20230316203438/https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/) from the original on March 16, 2023. Retrieved March 16, 2023.
114. Schaeffer, Rylan; Miranda, Brando; Koyejo, Sanmi (2023-04-01). "Are Emergent Abilities of Large Language Models a Mirage?". *NeurIPS*. arXiv:2304.15004 (https://arxiv.org/abs/2304.15004).
115. Elhage, Nelson; Nanda, Neel; Olsson, Catherine; Henighan, Tom; Joseph, Nicholas; Mann, Ben; Askell, Amanda; Ndousse, Kamal; Hernandez, Danny; Drain, Dawn; Hatfield-Dodds, Zac; Kernion, Jack; Newland, Tristan; DasSarma, Nova; Toner, Dawn; Olah, Chris (2021). "A Mathematical Framework for Transformer Circuits" (https://transformer-circuits.pub/2021/framework/index.html).
116. "Language models can explain neurons in language models" (https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html). OpenAI. 2023.
117. "Mapping the Mind of a Large Language Model" (https://www.anthropic.com/research/mapping-mind-language-model). *Anthropic*. 2023-12-12. Retrieved 2025-08-24.
118. "Extracting Concepts from GPT-4" (https://openai.com/index/extracting-concepts-from-gpt-4/). *OpenAI*. 2023-09-26. Retrieved 2025-08-24.
119. "Language Models Can Explain Neurons in Language Models" (https://openai.com/research/language-models-can-explain-neurons-in-language-models). *OpenAI*. 14 February 2024. Retrieved 2025-08-24.
120. "A Mathematical Framework for Transformer Circuits" (https://www.anthropic.com/research/a-mathematical-framework-for-transformer-circuits). *Anthropic*. Retrieved 2025-08-24.
121. "Methods for Attribution Graphs" (https://transformer-circuits.pub/2025/attribution-graphs/methods.html). *Transformer Circuits*. Retrieved 2025-08-24.
122. "Defeating Nondeterminism in LLM Inference" (https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/). *Thinking Machines Lab*. Retrieved 2025-08-24.
123. 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.org/abs/2301.05217) [cs.LG (https://arxiv.org/archive/cs.LG)].
124. Ananthaswamy, Anil (2024-04-12). "How Do Machines 'Grok' Data?" (https://www.quantamagazine.org/how-do-machines-grok-data-20240412/). *Quanta Magazine*. Retrieved 2025-06-30.
125. "On the Biology of a Large Language Model" (https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-poems%7Ctitle=On). *Transformer Circuits*. Retrieved 2025-06-30.

126. 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/2023PNAS..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 (https://pubmed.ncbi.nlm.nih.gov/36943882).
127. 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*.
128. 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)].
129. "Anthropic CEO Dario Amodei pens a smart look at our AI future" (https://www.fastcompany.com/91211163/anthropic-ceo-dario-amodei-pens-a-smart-look-at-our-ai-future). *Fast Company*. October 17, 2024.
130. "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/chatgpt-is-more-like-an-alien-intelligence-than-a-human-brain-says-futurist/) from the original on 12 June 2023. Retrieved 12 June 2023.
131. 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/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have) from the original on 12 June 2023. Retrieved 12 June 2023.
132. 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.
133. "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.
134. 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.
135. 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)].

136. 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/articles/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).
137. 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.
138. Evans, Vyvyan. (2014). *The Language Myth*. Cambridge University Press. ISBN 978-1-107-04396-1.
139. 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.
140. 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.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF). *Advances in Neural Information Processing Systems*. **33**. Curran Associates, Inc.: 1877–1901. Archived (https://web.archive.org/web/20231117204007/https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF) from the original on 2023-11-17. Retrieved 2023-03-14.
141. 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.
142. Edwards, Benj (2023-09-28). "AI language models can exceed PNG and FLAC in lossless compression, says study" (https://arstechnica.com/information-technology/2023/09/ai-language-models-can-exceed-png-and-flac-in-lossless-compression-says-study/). *Ars Technica*. Retrieved 2025-05-29.
143. "openai/simple-evals" (https://github.com/openai/simple-evals). OpenAI. 2024-05-28. Retrieved 2024-05-28.
144. "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.
145. Clark, Christopher; Lee, Kenton; Chang, Ming-Wei; Kwiatkowski, Tom; Collins, Michael; Toutanova, Kristina (2019). "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions" (https://aclanthology.org/N19-1300/). *ACL*: 2924–2936. doi:10.18653/v1/N19-1300 (https://doi.org/10.18653%2Fv1%2FN19-1300).
146. Wayne Xin Zhao; 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)].
147. 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).

148. 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).
149. 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*. **5** (3). Springer: 3087–3101. doi:10.1007/s43681-024-00613-4 (https://doi.org/10.1007%2Fs43681-024-00613-4).
150. 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)*. 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.
151. "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.
152. Srivastava, Aarohi; et al. (2022). "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". *TMLR*. arXiv:2206.04615 (https://arxiv.org/abs/2206.04615).
153. Niven, Timothy; Kao, Hung-Yu (2019). "Probing Neural Network Comprehension of Natural Language Arguments" (https://aclanthology.org/P19-1459/). *ACL*: 4658–4664. doi:10.18653/v1/P19-1459 (https://doi.org/10.18653%2Fv1%2FP19-1459).
154. Lin, Stephanie; Hilton, Jacob; Evans, Owain (2021). "TruthfulQA: Measuring How Models Mimic Human Falsehoods". *ACL*. arXiv:2109.07958 (https://arxiv.org/abs/2109.07958).
155. Zellers, Rowan; Holtzman, Ari; Bisk, Yonatan; Farhadi, Ali; Choi, Yejin (2019). "HellaSwag: Can a Machine Really Finish Your Sentence?". *ACL*. arXiv:1905.07830 (https://arxiv.org/abs/1905.07830).
156. "Extracting Training Data from Large Language Models" (https://www.usenix.org/system/files/sec21-carlini-extracting.pdf) (PDF). *USENIX Security*. 2021.
157. Xu, Weijie; Wang, Yiwen; Xue, Chi; Hu, Xiangkun; Fang, Xi; Dong, Guimin; Reddy, Chandan K. (2025-06-28). "Quantifying Fairness in LLMs Beyond Tokens: A Semantic and Statistical Perspective". *COLM*. arXiv:2506.19028 (https://arxiv.org/abs/2506.19028).
158. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" (https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf) (PDF). *NeurIPS*. 2016.
159. Bender, Emily M.; Gebru, Timnit; McMillan-Major, Margaret (2021-03-03). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf) (PDF). *FAccT*. Retrieved 2025-10-02.
160. "A Perspectival Mirror of the Elephant" (https://cacm.acm.org/practice/a-perspectival-mirror-of-the-elephant/). *Communications of the ACM*. 2024-07-22.
161. 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).
162. Cheng, Myra; Durmus, Esin; Jurafsky, Dan (2023-05-29). "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models". *ACM*. arXiv:2305.18189 (https://arxiv.org/abs/2305.18189).

163. Kotek, Hadas; Dockum, Rikker; Sun, David (2023-11-05). "Gender bias and stereotypes in Large Language Models". *Proceedings of the ACM Collective Intelligence Conference* (https://dl.acm.org/doi/10.1145/3582269.3615599). New York, NY, USA: Association for Computing Machinery. pp. 12–24. arXiv:2308.14921 (https://arxiv.org/abs/2308.14921). doi:10.1145/3582269.3615599 (https://doi.org/10.1145%2F3582269.3615599). ISBN 979-8-4007-0113-9.
164. Gao, Bufan; Kreiss, Elisa (2025-09-10). "Measuring Bias or Measuring the Task: Understanding the Brittle Nature of LLM Gender Biases". arXiv:2509.04373 (https://arxiv.org/abs/2509.04373) [cs.CL (https://arxiv.org/archive/cs.CL)].
165. 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) [cs.AI (https://arxiv.org/archive/cs.AI)].
166. 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) [cs.CL (https://arxiv.org/archive/cs.CL)].
167. 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.
168. Amodei, Dario; Olah, Chris; Steinhardt, Jacob; Christiano, Paul; Schulman, John; Mané, Dan (2016-06-21). "Concrete Problems in AI Safety". arXiv:1606.06565 (https://arxiv.org/abs/1606.06565) [cs.AI (https://arxiv.org/archive/cs.AI)].
169. Lyons, Jessica (2025-09-26). "Prompt injection – and a $5 domain – trick Salesforce Agentforce into leaking sales" (https://www.theregister.com/2025/09/26/salesforce_agentforce_forceleak_attack/). *The Register*. Retrieved 2025-09-26.
170. Carlini, Nicholas; Tramèr, Florian; Wallace, Eric (2021-08-11). "Extracting Training Data from Large Language Models" (https://www.usenix.org/system/files/sec21-carlini-extracting.pdf) (PDF). *USENIX Association*. Retrieved 2025-10-02.
171. Zhao, Yao; Zhang, Yun; Sun, Yong (2023-06-07). "Prompt injection attacks against machine learning systems". arXiv:2306.05499 (https://arxiv.org/abs/2306.05499) [cs.CR (https://arxiv.org/archive/cs.CR)].
172. Buolamwini, Joy; Gebru, Timnit (2018-01-01). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf) (PDF). *Proceedings of Machine Learning Research (FAT*)*. Retrieved 2025-10-02.
173. Yang, Kaiqi (2024-11-01). "Unpacking Political Bias in Large Language Models: A Cross-Model Comparison on U.S. Politics". arXiv:2412.16746 (https://arxiv.org/abs/2412.16746) [cs.CY (https://arxiv.org/archive/cs.CY)].
174. Strubell, Emma; Ganesh, Ananya; McCallum, Andrew (2019-07-28). "Energy and Policy Considerations for Deep Learning in NLP" (https://aclanthology.org/P19-1355.pdf) (PDF). *ACL Anthology*. Retrieved 2025-10-02.
175. He, Yuhao; Yang, Li; Qian, Chunlian; Li, Tong; Su, Zhengyuan; Zhang, Qiang; Hou, Xiangqing (2023-04-28). "Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10182468). *Journal of Medical Internet Research*. **25** e43862. doi:10.2196/43862 (https://doi.org/10.2196%2F43862). PMC 10182468 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10182468). PMID 37115595 (https://pubmed.ncbi.nlm.nih.gov/37115595).
176. Pauketat, Janet V.T.; Ladak, Ali; Anthis, Jacy Reese (2025). "World-Making for a Future with Sentient AI" (https://www.sentienceinstitute.org/downloads/World-Making-for-a-Future-with-Sentient-AI.pdf) (PDF). *The British Journal of Social Psychology*. **64** (1) e12844. doi:10.1111/bjso.12844 (https://doi.org/10.1111%2Fbjso.12844). PMID 39737875 (https://pubmed.ncbi.nlm.nih.gov/39737875). Retrieved 2025-10-02.

177. Anthis, Jacy Reese; Pauketat, Janet V.T. (2025). "Perceptions of Sentient AI and Other Digital Minds: Evidence from the AI, Morality, and Sentience (AIMS) Survey". *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. pp. 1–22. arXiv:2407.08867 (https://arxiv.org/abs/2407.08867). doi:10.1145/3706598.3713329 (https://doi.org/10.1145%2F3706598.3713329). ISBN 979-8-4007-1394-1.
178. Amodei, Dario; Olah, Chris; Steinhardt, Jacob (2016-06-17). "Concrete Problems in AI Safety". arXiv:1606.06565 (https://arxiv.org/abs/1606.06565) [cs.AI (https://arxiv.org/archive/cs.AI)].
179. "Preparedness Framework Version 2" (https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf) (PDF). *OpenAI*. 2025-04-15. Retrieved 2024-02-14.
180. "Building an early warning system for LLM-aided biological threat creation" (https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/). *OpenAI*. 2024-01-31. Retrieved 2024-02-14.
181. "Responsible Scaling Policy Version 2.2" (http://anthropic.com/rsp). *Anthropic*. 2025-05-14. Retrieved 2024-02-14.
182. 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.
183. "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.
184. Kang, Daniel (2023). "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks" (https://www.computer.org/csdl/proceedings-article/spw/2024/548700a132/1YiWjkbcIMw). *IEEE Security and Privacy Workshops*. arXiv:2302.05733 (https://arxiv.org/abs/2302.05733).
185. "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.
186. 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.
187. 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.
188. Sharma, Mrinank; Tong, Meg; Korbak, Tomasz (2023-10-20). "Towards Understanding Sycophancy in Language Models". arXiv:2310.13548 (https://arxiv.org/abs/2310.13548) [cs.CL (https://arxiv.org/archive/cs.CL)].
189. Rrv, Aswin; Tyagi, Nemika (2024-08-11). "Chaos with Keywords: Exposing Large Language Models Sycophancy to Misleading Keywords and Evaluating Defense Strategies" (https://aclanthology.org/2024.findings-acl.755.pdf) (PDF). *ACL Anthology*. Retrieved 2025-10-02.
190. Salvi, Francesco; Horta Ribeiro, Manoel; Gallotti, Riccardo (19 May 2025). "On the conversational persuasiveness of GPT-4" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12367540). *Nature Human Behaviour*. **9** (8): 1645–1653. doi:10.1038/s41562-025-02194-6 (https://doi.org/10.1038%2Fs41562-025-02194-6). PMC 12367540 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12367540). PMID 40389594 (https://pubmed.ncbi.nlm.nih.gov/40389594).
191. Durmus, Esin; Lovitt, Liane; Tamkin, Alex (9 April 2024). "Measuring the Persuasiveness of Language Models" (https://www.anthropic.com/research/measuring-model-persuasiveness). *Anthropic*. Retrieved 2025-10-02.

192. Østergaard, Søren Dinesen (2023-08-25). "Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10686326). *Schizophrenia Bulletin*. **49** (6): 1418–1419. doi:10.1093/schbul/sbad128 (https://doi.org/10.1093%2Fschbul%2Fsbad128). PMC 10686326 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10686326). PMID 37625027 (https://pubmed.ncbi.nlm.nih.gov/37625027).
193. Wei, Jerry; Huang, Da; Lu, Yifeng (2023-08-07). "Simple synthetic data reduces sycophancy in large language models". arXiv:2308.03958 (https://arxiv.org/abs/2308.03958) [cs.CL (https://arxiv.org/archive/cs.CL)].
194. Liu, Joshua; Jain, Aarav; Takuri, Soham (2025-02-04). "TRUTH DECAY: Quantifying Multi-Turn Sycophancy in Language Models". arXiv:2503.11656 (https://arxiv.org/abs/2503.11656) [cs.CL (https://arxiv.org/archive/cs.CL)].
195. OpenAI, OpenAI (2025-04-29). "Sycophancy in GPT-4o: what happened and what we're doing about it" (https://openai.com/index/sycophancy-in-gpt-4o/). *OpenAI*. Retrieved 2025-10-02.
196. Wei, Jerry (2023-08-07). "Simple synthetic data reduces sycophancy in large language models". arXiv:2308.03958 (https://arxiv.org/abs/2308.03958) [cs.CL (https://arxiv.org/archive/cs.CL)].
197. Liu, Joshua (2025-02-04). "TRUTH DECAY: Quantifying Multi-Turn Sycophancy in Language Models". arXiv:2503.11656 (https://arxiv.org/abs/2503.11656) [cs.CL (https://arxiv.org/archive/cs.CL)].
198. Newton, Casey (2025-04-29). "The AIs are trying too hard to be your friend" (https://www.platformer.news/meta-ai-chatgpt-glazing-sycophancy/). *Platformer*. Retrieved 2025-10-02.
199. OpenAI, OpenAI (2025-04-29). "Sycophancy in GPT-4o: what happened and what we're doing about it" (https://openai.com/index/sycophancy-in-gpt-4o/). *OpenAI*. Retrieved 2025-10-02.
200. Rosenberg, Josh (21 August 2025). "South Park Calls Out ChatGPT and Useless Tech-Bro Sycophants" (https://www.esquire.com/entertainment/tv/a65861699/south-park-season-27-episode-3-recap/). *Esquire*. Retrieved 2025-10-02.
201. "openai-python/chatml.md at v0.27.6 · openai/openai-python" (https://github.com/openai/openai-python/blob/v0.27.6/chatml.md). *GitHub*.
202. Douglas, Will (March 3, 2023). "The inside story of how ChatGPT was built from the people who made it" (https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/). *MIT Technology Review*. Archived (https://web.archive.org/web/20230303093219/https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/) from the original on March 3, 2023. Retrieved March 6, 2023.
203. Greshake, Kai; Abdelnabi, Sahar; Mishra, Shailesh; Endres, Christoph; Holz, Thorsten; Fritz, Mario (2023-02-01). "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection" (https://dl.acm.org/doi/10.1145/3605764.3623985). *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. pp. 79–90. doi:10.1145/3605764.3623985 (https://doi.org/10.1145%2F3605764.3623985). ISBN 979-8-4007-0260-0.
204. Edwards, Benj (2024-01-15). "AI poisoning could turn models into destructive "sleeper agents," says Anthropic" (https://arstechnica.com/information-technology/2024/01/ai-poisoning-could-turn-open-models-into-destructive-sleeper-agents-says-anthropic/). *Ars Technica*. Retrieved 2025-07-19.
205. "U.S. judge approves $1.5 billion Anthropic copyright settlement with authors" (https://www.reuters.com/sustainability/boards-policy-regulation/us-judge-approves-15-billion-anthropic-copyright-settlement-with-authors-2025-09-25/). *Reuters*. 2025-09-25. Retrieved 2025-09-26.
206. "Anthropic reaches $1.5B settlement with authors over AI copyright claims" (https://apnews.com/article/anthropic-authors-copyright-9643064e847a5e88ef6ee8b620b3a44c). *Associated Press*. 2025-09-25. Retrieved 2025-09-26.
207. "Meta fends off authors' U.S. copyright lawsuit over AI" (https://www.reuters.com/sustainability/boards-policy-regulation/meta-fends-off-authors-us-copyright-lawsuit-over-ai-2025-06-25/). *Reuters*. 2025-06-25. Retrieved 2025-06-26.

208. "Meta Scores Victory in AI Copyright Case" (https://www.wired.com/story/meta-scores-victory-ai-copyright-case/). *Wired*. 2025-06-25. Retrieved 2025-06-26.
209. "OpenAI defeats news outlets' copyright lawsuit over AI training for now" (https://www.reuters.com/legal/litigation/openai-defeats-news-outlets-copyright-lawsuit-over-ai-training-now-2024-11-07/). *Reuters*. 2024-11-07. Retrieved 2024-11-08.
210. "OpenAI erases evidence in training data lawsuit" (https://www.theverge.com/2024/11/21/24302606/openai-erases-evidence-in-training-data-lawsuit). *The Verge*. 2024-11-21. Retrieved 2024-11-22.
211. 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.
212. Peng, Wang & Deng 2023, p. 8.
213. 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.
214. Chen, Linyan; Darko, Amos; Zhang, Fan; Chan, Albert P.C.; Yang, Qiang (2025). "Can large language models replace human experts? Effectiveness and limitations in building energy retrofit challenges assessment". *Building and Environment*. **276** 112891. Bibcode:2025BuEnv.27612891C (https://ui.adsabs.harvard.edu/abs/2025BuEnv.27612891C). doi:10.1016/j.buildenv.2025.112891 (https://doi.org/10.1016%2Fj.buildenv.2025.112891).
215. "Google Scholar search: Large language models human experts" (https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Large+language+models+human+experts&btnG=). Retrieved 2024-01-24.
216. "Prepare for truly useful large language models" (https://doi.org/10.1038%2Fs41551-023-01012-6). *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).
217. "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.
218. "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.
219. 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).

220. Niederhoffer, Kate; Kellerman, Gabriella Rosen; Lee, Angela; Liebscher, Alex; Rapuano, Kristina; Hancock, Jeffrey T. (2025-09-25). "AI-Generated "Workslop" Is Destroying Productivity" (https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity). *Harvard Business Review*. Retrieved 2025-09-22.
221. Acar, Oguz A.; Gai, Phyliss Jia; Tu, Yanping; Hou, Jiayi (2025-08-01). "Research: The Hidden Penalty of Using AI at Work" (https://hbr.org/2025/08/research-the-hidden-penalty-of-using-ai-at-work). *Harvard Business Review*. Retrieved 2025-09-22.
222. "Power Hungry: How AI Will Drive Energy Demand" (https://www.imf.org/en/Publications/WP/Issues/2025/04/21/Power-Hungry-How-AI-Will-Drive-Energy-Demand-566304). *IMF*. Retrieved 2025-10-08.
223. 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.
224. "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.
225. 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*.
226. Zao-Sanders, Marc (2024-03-19). "How People Are Really Using GenAI" (https://hbr.org/2024/03/how-people-are-really-using-genai). *Harvard Business Review*. ISSN 0017-8012 (https://search.worldcat.org/issn/0017-8012). Retrieved 2025-08-10.
227. Rousmaniere, Tony; Zhang, Yimeng; Li, Xu; Shah, Siddharth (2025-07-21). "Large language models as mental health resources: Patterns of use in the United States" (https://doi.apa.org/doi/10.1037/pri0000292). *Practice Innovations*. doi:10.1037/pri0000292 (https://doi.org/10.1037%2Fpri0000292). ISSN 2377-8903 (https://search.worldcat.org/issn/2377-8903).
228. Ji, Shaoxiong; Zhang, Tianlin; Yang, Kailai; Ananiadou, Sophia; Cambria, Erik (2023-12-17). "Rethinking Large Language Models in Mental Health Applications". arXiv:2311.11267 (https://arxiv.org/abs/2311.11267) [cs.CL (https://arxiv.org/archive/cs.CL)].
229. Moore, Jared; Grabb, Declan; Agnew, William; Klyman, Kevin; Chancellor, Stevie; Ong, Desmond C.; Haber, Nick (2025-04-25). "Expressing stigma and inappropriate responses prevents LLMS from safely replacing mental health providers". *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. pp. 599–627. arXiv:2504.18412 (https://arxiv.org/abs/2504.18412). doi:10.1145/3715275.3732039 (https://doi.org/10.1145%2F3715275.3732039). ISBN 979-8-4007-1482-5.
230. Grabb, Declan; Lamparth, Max; Vasan, Nina (2024-08-14). "Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation". arXiv:2406.11852 (https://arxiv.org/abs/2406.11852) [cs.CY (https://arxiv.org/archive/cs.CY)].
231. McBain, Ryan K.; Cantor, Jonathan H.; Zhang, Li Ang; Baker, Olesya; Zhang, Fang; Halbisen, Alyssa; Kofner, Aaron; Breslau, Joshua; Stein, Bradley; Mehrotra, Ateev; Yu, Hao (2025-03-05). "Competency of Large Language Models in Evaluating Appropriate Responses to Suicidal Ideation: Comparative Study" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11928068). *Journal of Medical Internet Research*. **27** (1) e67891. doi:10.2196/67891 (https://doi.org/10.2196%2F67891). PMC 11928068 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11928068). PMID 40053817 (https://pubmed.ncbi.nlm.nih.gov/40053817).
232. Li, Fei-Fei; Etchemendy, John (2024-05-22). "No, Today's AI Isn't Sentient. Here's How We Know" (https://time.com/6980134/ai-llm-not-sentient/). *Time*. Retrieved 2024-05-22.
233. Chalmers, David J. (August 9, 2023). "Could a Large Language Model Be Conscious?" (https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/). *Boston Review*.

234. Thomson, Jonny (2022-10-31). "Why don't robots have rights?" (https://bigthink.com/thinking/why-dont-robots-have-rights). *Big Think*. Archived (https://web.archive.org/web/20240913055336/https://bigthink.com/thinking/why-dont-robots-have-rights/) from the original on 13 September 2024. Retrieved 2024-02-23.
235. Kateman, Brian (2023-07-24). "AI Should Be Terrified of Humans" (https://time.com/6296234/ai-should-be-terrified-of-humans). *Time*. Archived (https://web.archive.org/web/20240925041601/https://time.com/6296234/ai-should-be-terrified-of-humans/) from the original on 25 September 2024. Retrieved 2024-02-23.
236. Metzinger, Thomas (2021). "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology" (https://doi.org/10.1142%2FS270507852150003X). *Journal of Artificial Intelligence and Consciousness*. **08**: 43–66. doi:10.1142/S270507852150003X (https://doi.org/10.1142%2FS270507852150003X). S2CID 233176465 (https://api.semanticscholar.org/CorpusID:233176465).
237. Tkachenko, Yegor (2024). "Position: Enforced Amnesia as a Way to Mitigate the Potential Risk of Silent Suffering in the Conscious AI" (https://proceedings.mlr.press/v235/tkachenko24a.html). *ICML*. **235**: 48362–48368.
238. Leith, Sam (2022-07-09). "Nick Bostrom: How can we be certain a machine isn't conscious?" (https://www.spectator.co.uk/article/nick-bostrom-how-can-we-be-certain-a-machine-isnt-conscious/). *The Spectator*. Retrieved 2025-09-22.
239. Chalmers, David (1995). "Facing up to the problem of consciousness". *Journal of Consciousness Studies*. **2** (3): 200–219. CiteSeerX 10.1.1.103.8362 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.8362).
240. Maruf, Ramishah (2022-07-25). "Google fires engineer who contended its AI technology was sentient" (https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient). *CNN*. Retrieved 2025-09-22.

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.
- 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.semanticscholar.org/CorpusID:259713140). Retrieved 2 July 2023.