

Large Language Models Safety

== History == a5091921a6 BERT was originally implemented in the English language at two model sizes, BERTBASE (110 million parameters) and BERTLARGE (340 million parameters). Both were trained on the Toronto BookCorpus (800M words) and English Wikipedia (2,500M words). The weights were released on GitHub. On March 11, 2020, 24 smaller models were... a5091921a6 Model weights for the first version of Llama were only available to researchers on a case-by-case basis, under a non-commercial license. Unauthorized copies of the first model were shared via BitTorrent. Subsequent versions of Llama were made accessible outside academia and released under licenses that permitted some commercial use. a5091921a6 == Background == a5091921a6

Large language model measure model reasoning, factual accuracy, alignment, and safety. LLMs can typically generate, summarize, translate, and analyze text in many contexts. They are the basis for many modern chatbots, such as ChatGPT, Claude,

Large Language Models Safety

Gemini, Grok, and DeepSeek. Before the emergence of transformer-based models in 2017, some language models were considered A large language model (LLM) is an AI model (typically a neural network) trained on a vast amount of text for natural language processing tasks, especially language generation

Building foundation models is often highly resource-intensive, with the most advanced models costing hundreds of millions of dollars to cover the expenses of acquiring, curating, and processing massive datasets, as well as the compute power required for training. These costs stem from the need for sophisticated infrastructure, extended training times, and advanced hardware, such as GPUs. In contrast, adapting an existing foundation model for a specific task or using it directly is far less costly, as it uses pre-trained capabilities and typically requires only fine-tuning on smaller, task-specific datasets.

== Scope and terminology ==

== Motivations ==

== History ==

AI safety These “sleeping agent” models could be programmed to generate AI safety is an interdisciplinary field focused on preventing accidents, misuse, or other harmful consequences arising from artificial intelligence systems. The field is

Large Language Models Safety

particularly concerned with existential risks posed by advanced AI models. paper by Anthropic showed that large language models could be trained with persistent backdoors. It encompasses AI alignment (which aims to ensure AI systems behave as intended), monitoring AI systems for risks, and enhancing their robustness

In April 2026, Meta Superintelligence Labs released Muse Spark as a replacement for Llama. Large language models (LLMs), currently their most advanced form as of 2026, are predominantly based on transformers trained on larger datasets (frequently using texts scraped from the public internet). They have superseded recurrent neural network-based models, which had previously superseded the purely statistical models, such as the word n-gram language model. In the 1970s, Frederick Jelinek and colleagues at IBM Research developed statistical approaches to language modeling as part of their work on speech recognition. The group used probabilistic n-gram models of word sequences and introduced perplexity as an information-theoretic measure of the...

Pre-flight safety demonstration Beginning in the 2000s, some airlines have produced stylized or humorous safety videos featuring celebrities A pre-flight safety

Large Language Models Safety

demonstration (also called a pre-flight safety briefing or safety video) is an explanation given to airline passengers before takeoff outlining the aircraft's safety features and emergency procedures. include subtitles, sign language, or multiple languages

The first large language model released by Thinking Machines, Inkling was trained with 45 trillion tokens of text, image, audio, and video. It reasons natively from input in any of these four formats, while, at the moment, it can generate only text, including source code or structured data. Training Scholars discuss current risks from critical systems failures, bias, and AI-enabled surveillance, as well as emerging risks like technological unemployment, digital manipulation, weaponization, AI-enabled... Llama models come in different sizes, ranging from 1 billion to 2 trillion parameters. Initially only a foundation model, starting with Llama 2, Meta AI released instruction fine-tuned versions alongside foundation models. During the 1950s, Noam Chomsky developed influential formal approaches to describing language, including models based on formal grammar. LLMs are typically based on transformer architecture. Generative pre-trained transformers (GPTs) are a type of LLM that is pre-trained to predict the next word. GPTs are then

Large Language Models Safety

often fine-tuned to follow instructions and to behave as assistants. a5091921a6

Language model Language models are useful for a variety of tasks, including speech recognition, machine translation, natural language generation (generating more human-like text), optical character recognition, route optimization, handwriting recognition, grammar induction, information retrieval and disaster response. neural network-based models, which had previously superseded the purely statistical models, such as the word n-gram language model. During the 1950s, Noam A language model is a computational model that predicts sequences in natural language a5091921a6

After the release of large language models such as GPT-3, a focus of research was up-scaling models, which in some instances showed major increases in emergent... a5091921a6 BERT is trained by masked token prediction and next sentence prediction. With this training, BERT learns contextual, latent representations of tokens in their context, similar to ELMo and GPT-2. It found applications for many natural language processing tasks, such as coreference resolution and polysemy resolution. It improved on ELMo and spawned the study of "BERTology", which attempts to interpret what is learned by BERT. a5091921a6 Beginning in the 2000s, some airlines have produced

Large Language Models Safety

stylized or humorous safety videos featuring celebrities, popular music, or cinematic themes, many of which gained wide circulation online. For example, Cebu Pacific staged a choreographed demonstration set to Lady Gaga's "Just Dance" and Katy Perry's "California Gurls". A 2008 Delta Air Lines video led to the internet nickname "Deltalina" for one of its featured flight attendants. A later British Airways video featured performers including Rowan... a5091921a6

BERT (language model) As of 2026, BERT is a common methodological component in natural language processing (NLP) research. the state of the art for large language models. BERT dramatically improved the state of the art for large language models. It learns to represent text as a sequence of vectors using self-supervised learning. As of 2026[update], BERT is a common methodological component in natural language processing (NLP) research Bidirectional encoder representations from transformers (BERT) is a language model introduced in October 2018 by researchers at Google. It uses the encoder-only transformer architecture a5091921a6

Inkling (large language model) since they consider that internal safety evaluation in the language models occasionally doesn't prevent the model from complying with

Large Language Models Safety

potentially harmful Inkling is an open-weights large language model created by Thinking Machines company, first released on July 15, 2026. The model allows text, images and audio as input, and can be used for tasks such as chat, software development, multilingual applications or agentic systems. It is published under an Apache 2.0 license

Early examples of foundation models are language models like OpenAI's GPT series and Google's BERT. Beyond text, foundation models have been developed across a range of modalities—including DALL-E, Stable...
History
Alongside the release of Llama 3, Meta rolled out Meta AI, an AI assistant built on Llama. Meta AI has a dedicated website and is available on Facebook and WhatsApp. The latest version is Llama 4, released in April 2025.

Llama (language model) Llama ("Large Language Model Meta AI" serving as a backronym) was a family of large language models (LLMs) released by Meta AI starting in February 2023 Llama ("Large Language Model Meta AI" serving as a backronym) was a family of large language models (LLMs) released by Meta AI starting in February 2023

Large Language Models Safety

Recent surveys use post-training as an umbrella label for several lines of research, including instruction tuning, preference alignment, reasoning-oriented training, efficiency methods, and distillation. Their boundaries differ, so the label does not imply a single objective or a universally agreed sequence... a5091921a6 Beyond technical research, AI safety involves developing norms and policies that promote safety, including advocacy for regulations at different levels of government. The field gained significant popularity in 2023, with rapid progress in generative AI and public concerns voiced by researchers and CEOs about potential dangers. During the 2023 AI Safety Summit, the United States and the United Kingdom both established their own AI Safety Institute. However, researchers have expressed concern that AI safety measures are not keeping pace with the rapid development of AI capabilities. a5091921a6

Post-training of large language models Supervised fine-tuning or instruction tuning and preference-based alignment are common elements, while broader treatments differ on whether reasoning-focused training, parameter-efficient adaptation, distillation, tool integration, or inference-time scaling are included. The scope varies among surveys. of large language models is a term used in recent technical literature for training applied to a large language model (LLM) after its initial large-scale Post-training of large

Large Language Models Safety

language models is a term used in recent technical literature for training applied to a large language model (LLM) after its initial large-scale pretraining a5091921a6

In training-centered usage, a pretrained or base model is further optimized with structured data such as instruction-response pairs, comparisons between candidate responses, step-level judgments, or outcomes that can be checked automatically. Practical systems may combine several stages, for example supervised training followed by preference optimization or reinforcement learning. a5091921a6

Reasoning model They tend to do better on logic, math, and programming tasks than standard LLMs, can revisit and revise earlier steps, and make use of extra computation while answering as another way to scale performance, alongside the number of training examples, parameters, and training compute. Reasoning language models (RLMs) or large reasoning models (LRMs) are large language models that are trained further to solve tasks that take several Reasoning language models (RLMs) or large reasoning models (LRMs) are large language models that are trained further to solve tasks that take several steps of reasoning a5091921a6

Large Language Models Safety

Biased or inaccurate training data can make an LLM's output less reliable. Benchmark evaluations for LLMs attempt to measure model reasoning, factual accuracy, alignment, and safety. a5091921a6 Aviation regulations generally require that passengers be orally briefed before departure but do not prescribe how the information must be delivered. Airlines therefore present the briefing either as a live demonstration by flight attendants or as a pre-recorded audio or video presentation shown through the aircraft's in-flight entertainment system. Briefings typically last several minutes and pre-recorded videos may include subtitles, sign language, or multiple languages. a5091921a6

Foundation model), is a machine learning or deep learning model trained on vast datasets so that it can be applied across a wide range of use cases. Generative AI applications like large language models (LLM) In artificial intelligence, a foundation model (FM), also known as large x model (LxM, where "x" is a variable representing any text, image, sound, etc. learning model trained on vast datasets so that it can be applied across a wide range of use cases. Generative AI applications like large language models (LLM) are common examples of foundation models a5091921a6

Large Language Models Safety

[https://lb1-s245-pub.pressidium.com/\\$c/journal/url?MD=chemist_warehouse_mayfield_nsw](https://lb1-s245-pub.pressidium.com/$c/journal/url?MD=chemist_warehouse_mayfield_nsw)
https://lb1-s245-pub.pressidium.com/^i/pub/data?PUB=book-a_blood_test-near-me
https://lb1-s245-pub.pressidium.com/^r/edu/key?PDF=mucus_in_a-dog-s_poop
https://lb1-s245-pub.pressidium.com/@n/ebook/list?DOC=the_nucleus_of_an-atom-consists_of
[https://lb1-s245-pub.pressidium.com/\\$g/doc/mirror?EPUB=how-to_make_a-her_squirt](https://lb1-s245-pub.pressidium.com/$g/doc/mirror?EPUB=how-to_make_a-her_squirt)
https://lb1-s245-pub.pressidium.com/-g/text/url?PUB=amoxicilina_acido_clavulanico_875_125
https://lb1-s245-pub.pressidium.com/^b/book/url?TXT=tri_city_skin_and_cancer_johnson-city-tennessee
https://lb1-s245-pub.pressidium.com/~p/ppt/upload?RTF=where_is-guiana_located
https://lb1-s245-pub.pressidium.com/~k/edu/mirror?PDF=in_the-mouth-of-madness
<https://lb1-s245-pub.pressidium.com/=v/doc/upload?CHAPTER=ativan-drug-side-effects>
https://lb1-s245-pub.pressidium.com/=w/short/mirror?EBOOK=before_your_period_do-you_gain_weight

Large Language Models Safety