Asian Journal of Basic Science & Research

Volume 7, Issue 3, Pages 175-200, July-September 2025

ChatGPT and the Future of Generative AI: Architecture, Limitations, and Advancements in Large Language Models

Kingsley Olunosen Osobase[1*], Oluwatoyin Olawale Akadiri[2], Christian Davison Dirisu[3], Emmanuel Asolo[4], Damilola Emmanuel Adewara[5] & Chijioke Cyriacus Ekechi[6]

[1]Department of Meteorology, School of Earth and Mineral Sciences, Federal University of Technology Akure (FUTA), Akure, Ondo State, Nigeria. [2]Department of Information Sciences, School of Information Sciences and Engineering, Bay Atlantic University, United States of America. [3]Department of Data Science & Analytics, School of Computer Science and Engineering, EPITA School of Engineering and Computer Science, France. [4]Data Science, Artificial Intelligence and Modeling Centre, University of Hull, Hull, United Kingdom. [5]Department of Statistics, Faculty of Physical Science, University of Ilorin, Nigeria. [6]Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, Tennessee, United States of America. Corresponding Author (Kingsley Olunosen Osobase) Email: olunosenk@gmail.com[*]

DOI: Under Assignment

Article Received: 13 July 2025 *Article Accepted: 21 September 2025* *Article Published: 25 September 2025*

ABSTRACT

This review examines ChatGPT as both a technological milestone and a research testbed for understanding the evolution of large language models (LLMs). Beginning with the transformer architecture and its scaling laws, the paper analyzes how pretraining, supervised fine-tuning, and reinforcement learning with human feedback (RLHF) collectively shape model performance. Empirical evidence from education, healthcare, business, and scientific research demonstrates that ChatGPT and domain-tuned variants can augment learning outcomes, accelerate professional productivity, and enable new forms of discovery. At the same time, persistent challenges, including hallucinations, bias, opacity, data limitations, and environmental costs, reveal the limits of scale-driven progress. Ongoing improvements such as multimodality, retrieval-augmented generation, domain-specific alignment, interpretability research, and energy-efficient training signals are emerging solutions, but they also expose critical research gaps. Looking forward, the integration of LLMs with autonomous agents, data science workflows, symbolic reasoning, and governance frameworks will define the trajectory of generative AI. The paper argues that ChatGPT should be viewed not merely as a product but as a living research instrument, one that highlights both the transformative potential and the societal risks of generative AI.

Keywords: ChatGPT; Large Language Models; Transformer Architecture; Scaling Laws; Reinforcement Learning with Human Feedback (RLHF); Retrieval-Augmented Generation (RAG); Multimodality; Interpretability; Domain-Specific AI; Sustainable AI; Governance of Generative AI.

░ 1. Introduction

The launch of ChatGPT by OpenAI on November 30, 2022, became a turning point in the public and professional engagement with artificial intelligence (Haleem et al., 2022). In contrast to previous AI systems that were task-specific, ChatGPT was able to engage in extended conversations and change in relation to different prompts, and to produce text which frequently felt historically contextually relevant and coherent (Menon & Shilpa, 2023). Its working architecture is that of the transformer (Khan et al., 2023), and it has been trained using a wider collection of text sources, subsequently optimized with supervised fine-tuning and reinforcement learning via human feedback (RLHF) (Gonzalez Barman et al., 2025). Through these training steps, it can respond in a grammatically fluent manner and with an apparent sense of a conversation context (Javaid et al., 2023).

New versions of ChatGPT and related systems have gone beyond text to include images, audio, and even multiple different media (Roumeliotis & Tselikas, 2023). They have also gained the ability to use external tools, retrieve up-to-date information, and sustain longer, more context-aware exchanges (Ray, 2023). This has led to large language models transforming themselves out of an experimental research project and into a platform integrated into the classroom and businesses, research labs, medical facilities, and creative fields (Lopez-Gazpio, 2025).

Models like ChatGPT represent a unifying foundation for knowledge access, problem-solving, and human–computer interaction. They summarize the trends of even bigger and more diverse datasets into a single flexible model that can generalize across tasks by not requiring much retraining, but by an action called prompting (Nerella et al., 2024). They can synchronize with external devices and query systems and act as an adaptive means of communicating between users and complicated computing tasks (Lakatos et al., 2025). This versatility enables them to address problems that were once resistant to automation, ill-defined queries, creative synthesis, and long-tail tasks, while also allowing rapid policy or safety updates through prompts and lightweight fine-tuning. For researchers, they provide a living testbed for ongoing challenges in robustness, interpretability, alignment, and human–AI collaboration (Jeyaraman et al., 2023).

However, alongside these advances are notable gaps that shape both the technology's reliability and its broader adoption. ChatGPT can produce confident but factually incorrect statements, a limitation often described as "hallucination (Ray, 2023)." Its decision-making process remains opaque, making it difficult to explain why certain outputs are produced (Cheong, 2024). Access to the data and methods used in its training is limited, which hinders independent assessment of potential biases or blind spots (Busch et al., 2025). As its capabilities grow, questions remain about its cultural adaptability, long-term accuracy in complex tasks, and the environmental and economic costs of maintaining such large-scale models (Ray, 2023).

The evolution of the GPT family of models, from GPT-1 in 2018 to GPT-5 in 2025, reflects steady increases in scale, capability, and multimodal integration (Dilmegani & Sezer, 2025). Each successive version has introduced architectural refinements, expanded training data, and improved task performance, progressing from basic language modeling to advanced reasoning and tool-assisted interaction. This progression is illustrated in Figure 1, which highlights the release year, core features, and distinguishing characteristics of each model. In addition to the timeline shown in Figure 1, the main characteristics of each GPT generation are summarized in Table 1, providing a concise comparison of their release dates, scale, and key innovations.

1.1. Aim and scope of this review

This review takes ChatGPT as a central example for examining the architecture, training processes, practical uses, shortcomings, and ongoing developments of modern large language models. The discussion combines technical and applied perspectives, linking model design choices with the handling of data, approaches to evaluation, and the governance structures that shape how such systems are deployed. By drawing these strands together, the review highlights the persistent gaps that limit reliability, safety, and societal benefit, and outlines directions for research and innovation that could help address them.

1.2. Gaps motivating the review

Despite notable progress, ChatGPT and similar models still face significant limitations. They can produce confident but inaccurate responses, a tendency known as hallucination, and current evaluation benchmarks are inconsistent, often failing to capture real-world performance. Limited transparency around training data hinders bias auditing and error tracing, especially for low-resource languages. Bias mitigation measures may weaken under adversarial prompts or in unfamiliar cultural contexts, and reasoning abilities can falter in complex, multi-step tasks. These issues are compounded by large-scale deployment's high environmental and financial costs and by underdeveloped governance frameworks.

With ChatGPT positioned in a wider context of LLMs, this article sheds light on these issues and outlines possible solutions for increased accuracy, transparency, fairness, and sustainability.

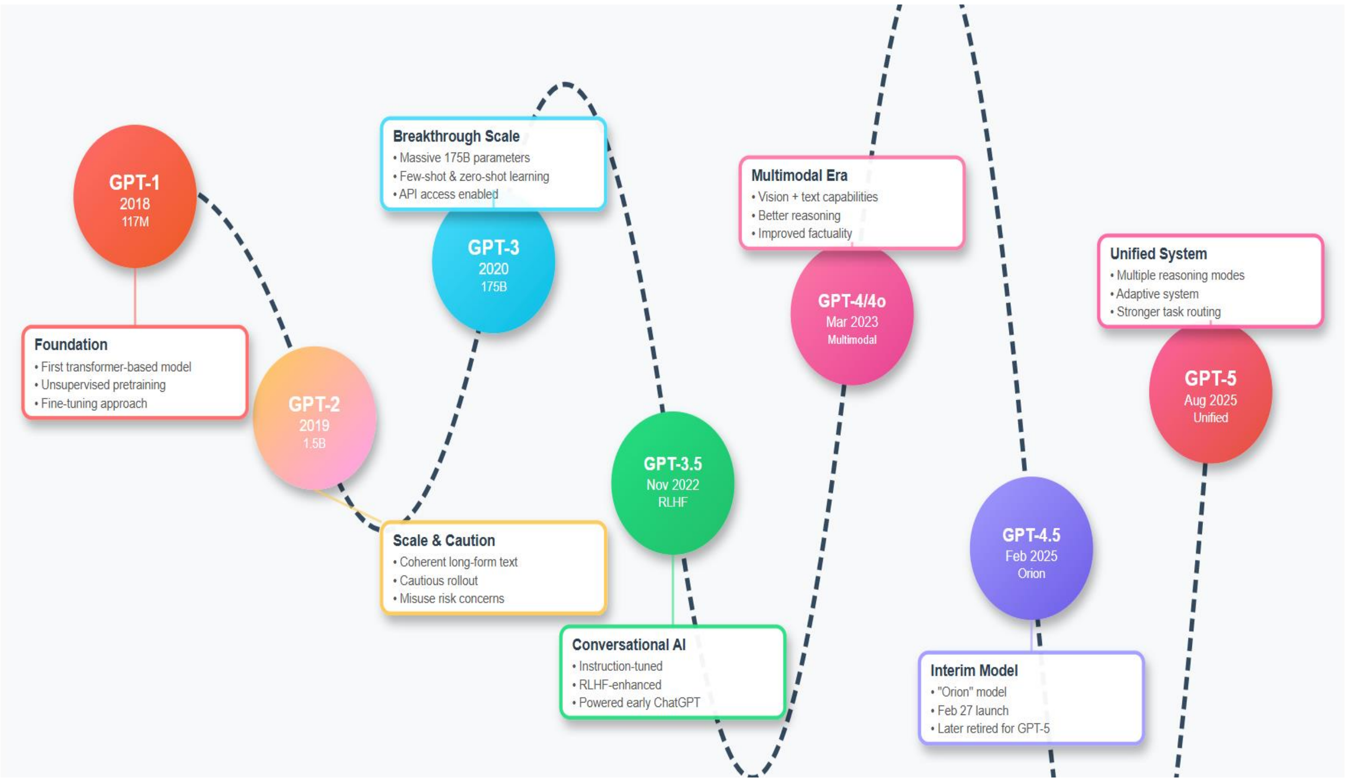

Figure 1. Timeline of GPT model evolution from GPT-1 (2018) to GPT-5 (2025), highlighting release dates, core features, and differentiating capabilities of each version.

Table 1. Summary of GPT model evolution from GPT-1 (2018) to GPT-5 (2025), highlighting release dates, approximate parameter counts, and distinguishing features.

Model	Release Date	Approx. Parameters	Key Features/Innovations	Notable Advancements
GPT-1	June 2018	117 M	First transformer-based generative pre-trained model; trained on BookCorpus; introduced unsupervised pretraining + supervised fine-tuning paradigm.	Demonstrated transfer learning potential for NLP.
GPT-2	Feb 2019	1.5 B	Trained on WebText (≈40 GB); produced coherent multi-paragraph text; initial full release delayed over misuse concerns.	Significant improvement in text fluency and coherence.
GPT-3	June 2020	175 B	Trained on hundreds of billions of tokens from diverse internet sources; strong few-shot and zero-shot capabilities.	Enabled OpenAI API; marked a step change in general-purpose language generation.

GPT-3.5	Nov 2022	~175 B (refined)	Instruction-tuned and RLHF-enhanced variant; more conversational; foundation for ChatGPT public launch.	Greatly improved usability and dialogue performance.
GPT-4/ GPT-4o	Mar 2023	Not disclosed (est. >500 B for GPT-4)	Multimodal input (text + images); better reasoning, factuality, and steerability; "4o" optimized for speed and efficiency.	Achieved high scores on professional/academic benchmarks.
GPT-5	Aug 2025	Not disclosed	Unified, adaptive system with multiple reasoning modes; expanded multimodality (text, images, audio, code); larger context windows; improved tool integration.	Enhanced real-time reasoning, planning, and safety controls.

Note: Data for GPT-1 to GPT-4 adapted from Sufi (2024); data for GPT-1 to GPT-5 adapted from Dilmegani & Sezer (2025).

1.3. Study Objectives

This review aims to:

1) Outline the transformer-based architecture and training stages of ChatGPT and related LLMs.

2) Summarize key applications in education, healthcare, business, and research.

3) Highlight major limitations such as hallucinations, bias, opacity, and high energy demand.

4) Present emerging improvements: multimodality, retrieval-augmented generation, domain-specific tuning, and interpretability.

5) Discuss ethical and governance issues surrounding large-scale generative AI.

6) Offer recommendations for future development of reliable and sustainable LLMs.

2. Background and Theoretical Foundations

2.1. Machine Learning and Deep Learning Foundations

Large language models such as ChatGPT are rooted in the principles of machine learning, where systems learn from data to improve performance rather than following a fixed set of programmed rules (Roumeliotis & Tselikas, 2023). In this field, deep learning, with its multi-layered neural networks that can automatically learn hierarchical representations, has come to take center stage in the process of natural language processing (Mienye & Swart, 2024). Conventional NLP systems were deeply rule-based or low-level statistical processing and had problems with how diverse and unpredictable human language can be (Bakagianni et al., 2025). The implementation of deep learning made it possible to substantially increase the level of linguistic pattern learning, which significantly contributed to breakthroughs in the field of translation, summarization, and conversational AI.

The ChatGPT architecture, which is based on the transformer architecture proposed by Vaswani et al. in 2017, is the basis of many modern LLMs (Zhang et al., 2023). Transformers operate on every token in a sequence at the same time as opposed to gradually, unlike the previous sequence models (recurrent neural networks (RNNs) and long short-term memory (LSTM) networks) (Zhang et al., 2023). That is achievable through the *attention mechanism* that computes the strength of the ties between a given token and all others within the input (Ruan & Jin, 2022). Consequently, the model can capture both short-range dependencies (e.g., within a phrase) and long-range dependencies (e.g., across several sentences) with high efficiency. ChatGPT is an architecture based on transformers, where parallelization and multi-head self-attention give room to process a group of tokens simultaneously, regardless of their position (Banik et al., 2024). Figure 2 shows that the transformer can be stacked with attention and feed-forward modules residually connected syntactically to each other and layer-normalised, which allows effective contextual representation of long sequences in the module (Choi & Lee, 2023). Whereas the transformer has both the encoder and decoder stacks, ChatGPT only adopts the decoder with masked multi-head attention to support autoregressive text movement (Jin et al., 2025).

In practice, attention enables ChatGPT to deduce what is most important in a conversational context to supply an answer to. For example, in a multi-turn conversation, the model can consider only a certain previous utterance that will explain the semantics of a current query while disregarding all other facts. Such ability, along with training using massive context and the billions of parameters, makes the model capable of creating coherent, contextually relevant, and human-like text. The research in scaling revealed that as the size of the model increases, together with training data, LLMs tend to perform better across a large variety of tasks, which has been used to develop the successive GPT generations (Filippo et al., 2024).

This architecture laid the foundation for applying modern NLP techniques, enabling models like ChatGPT to process language in a way that reflects meaning and context, a process explained further in the next section.

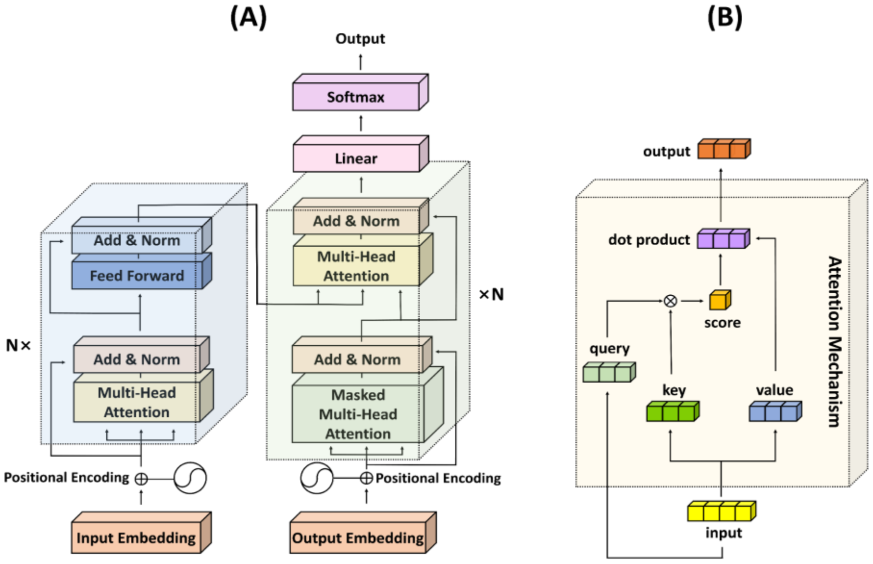

Figure 2. Illustrative Schematic of Transformer and Attention Mechanism

(A) Stacked encoder and decoder blocks depicting multi-head attention, feed-forward layers, and residual connections with layer normalization. (B) The attention mechanism weighs the sum of the values of an ordered list under the influence of the similarity between the queries and keys. ChatGPT is based on the decoder portion of this architecture, using masked multi-head attention for autoregressive text generation [Source: Choi & Lee (2023)].

2.2. Natural Language Processing (NLP) Concepts

ChatGPT operates within the wider domain of natural language processing (NLP), which involves the ability of machines to display and create human language in a manner that displays context and meaning (Ray, 2023). NLP combines linguistic theory with statistical modelling in an attempt to fill that gap between raw text and the computational form.

The key action in this sequence is tokenization, which splits the input text into portions or tokens, which can be words, subwords, or even individual characters (Qin et al., 2025). Models such as ChatGPT use a tokenization that is subword-based, which means that the system can be optimized to represent frequent words as well as unusual or new vocabulary. The tokens are then converted to numerical vectors that can be made via learnings known as embeddings that represent semantic and syntactic relationships (Mswahili & Jeong, 2024). Tokens that share semantics will be close together in the model in the high-dimensional vector space.

After embedding, tokens undergo processing in the model attention layers, which generate a *contextualized version* of each token based on the connection between the token and all other tokens in the sequence (Mars, 2022). This allows the model to understand, for example, that the word "bank" might mean a financial institution in one sentence and a riverbank in another, depending on surrounding words (Smirnova, 2016). The last step involves the process of sequence modeling, where the model is offered a probability distribution of potential new tokens and then picks one and repeats the operation at each token position till the output is finished.

An important feature of this process is the context window, the fixed number of tokens the model can consider at once (Li et al., 2024). For ChatGPT, this window size determines how much of the conversation history or document context can be factored into the current response. While larger windows allow for richer context handling, they also increase computational cost. Moreover, once a conversation exceeds the context limit, earlier parts may be truncated, leading to a loss of relevant information.

By combining tokenization, embeddings, attention-driven context modeling, and probabilistic generation, ChatGPT can extend a user's prompt in a way that appears coherent, contextually appropriate, and linguistically natural. However, these same mechanisms can also lead to limitations, such as producing plausible but inaccurate statements when relevant context is missing or misunderstood (Williamson & Prybutok, 2024).

These NLP processes allow the model to turn raw text into meaningful, context-aware predictions. However, the success of this process depends heavily on the quality of the data it is trained on, a factor addressed in the following section.

2.3. Data Preparation and Curation in LLM Development

The creation of models like ChatGPT depends heavily on the collection, preparation, and management of large-scale training datasets. These datasets draw from a wide variety of sources, including books, academic articles, websites, code repositories, and other publicly accessible text, supplemented in some cases by licensed or proprietary materials (The Authors Guild, 2023). The breadth and diversity of this data are intended to give the model exposure to many domains, writing styles, and contexts.

The raw data are then extensively preprocessed before training. These include deleting redundant records, filtering inappropriate or low-quality material, formatting, and making the text machine-readable (Lee et al., 2022). In some instances, parts of the data are annotated with human or automated labels to focus fine-tuning on particular tasks, such as to follow instructions, minimize damaging outputs, or to fit particular domains (Balaskas et al., 2025).

Beyond technical preparation, there is a growing emphasis on ethical sourcing and legal compliance in dataset creation. Most giant language models have been criticized due to reusing content without clear authorization of content creators, which is an issue of intellectual property and copyright infringement (Al-Busaidi et al., 2024). They should be able to work in and around detailed copyrights, licensing terms, and even terms of service restrictions, which vary between jurisdictions. Ethical considerations are also relevant when it comes to the inclusion of personal or sensitive data that, if not properly anonymized or excluded, can lead to privacy breaches and potential regulatory violations under frameworks such as the GDPR (Ducato, 2020).

The effectiveness of the results of a large language model strongly depends on the quality, diversity, and representativeness of the training data that it is used with. This has the consequence of biasing coverage in some areas of knowledge, being less factual, and failing to generalize in underrepresented areas (Helm et al., 2024). On the other hand, the factual reliability, preventing damaging biases, preserving intellectual property, and increasing contexts in which the model functions successfully can be achieved with carefully assembled and legally non-problematic datasets. As public scrutiny increases, transparent documentation of dataset sources and collection methods is becoming an essential component of responsible AI development (Cheong, 2024).

░3. ChatGPT Architecture & Model Design

3.1. Transformer Architecture and the Self-Attention Mechanism

The transformer architecture introduced as described by Lin et al. (2022) has been described as an architecture shift in the natural language processing area because recurrence and convolution processes have been substituted with self-attention as the basis of sequence modeling processes. Such a design allows models to learn both short- and long-range dependencies efficiently and overcome the limitation of requiring sequential processing inherent to recurrent neural networks (RNN) and long short-term memory networks (LSTM). As a further point, Lin et al. emphasize that a broad range of variants of transformers, commonly termed X-formers, have since been deployed, which covers application areas in natural language processing, computer vision, and audio processing, forming the basis for ongoing architectural innovations and future research directions.

The architecture is structured as a multi-level layer-by-layer and falls into stacked encoder and decoder layers consisting of multi-head self-attention, position-wise feed-forward networks, residual connections, and layer normalization (Figure 3). The encoder projects the input embeddings, augmented with sinusoidal positional encodings to preserve word order, to low-dimensional contextualized representations, which are simpler but more accurate representations of the information present in the input. Conversely, the decoder includes masked multi-head self-attention, which makes it autoregressive in that attention can only take place on past generated tokens. Such property is essential in text generation tasks where future tokens cannot be available in the process of prediction (Dong et al., 2019).

A key innovation lies in the multi-head attention mechanism, in which queries, keys, and values are projected into multiple subspaces. This enables the respective heads of attention to specialize in focusing on a specific linguistic or semantic relation to gain an enhanced representational capacity (Zheng et al., 2025). The results of these heads are then concatenated and passed through feed-forward layers, with residual connections and normalization stabilizing the training and preventing gradient degradation in deep models (Bao et al., 2024).

This modular combination of attention and feed-forward transformations allows transformers to be much more scalable with respect to the sizes of data and computer resources of different budgets. As such, transformer-based models are the foundation of state-of-the-art generative AI applications, like GPT, BERT, and their continuations, and perform better in a wide variety of natural language understanding and generation tasks (Bengesi et al., 2024).

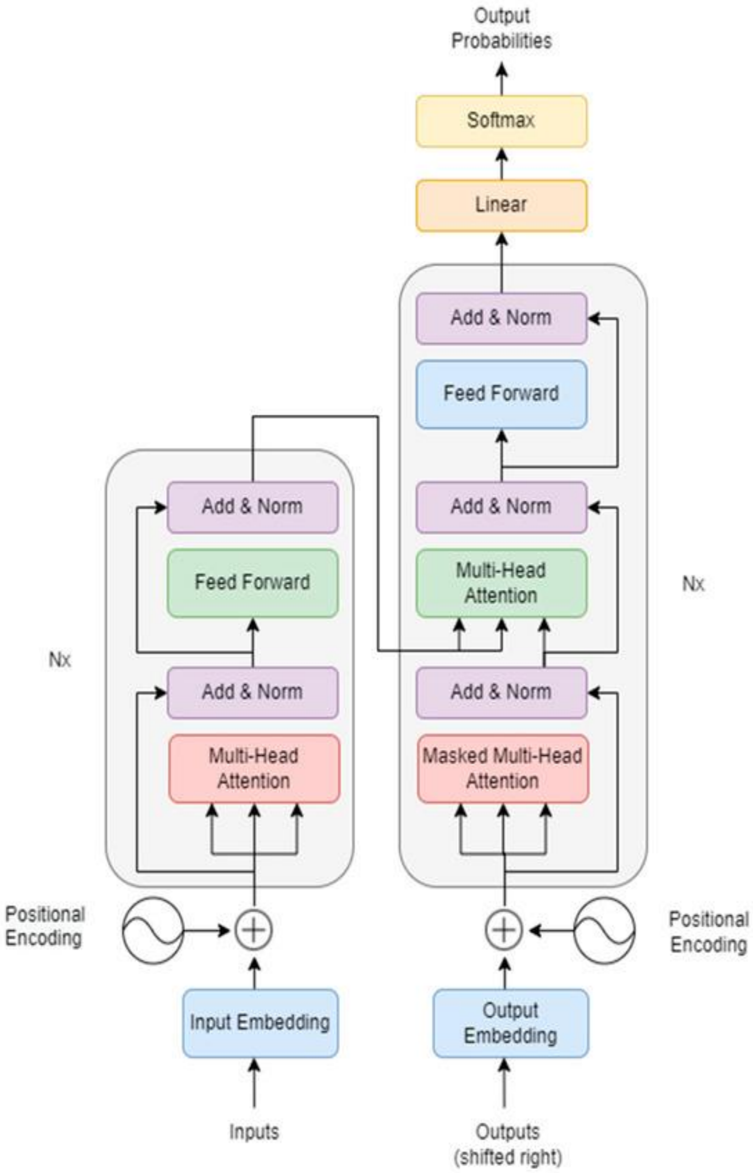

Figure 3. High-level diagram of the transformer architecture, showing the stacked encoder (left) and decoder (right) blocks. The blocks combine multi-head self-attention, feed-forward layers, residual connections, and

normalization. Through the process of encoding, the input embeddings with positional embeddings are converted into contextual representations and output sequentially with the use of masked multi-head attention through the decoder. It is the foundation of large language models that include ChatGPT. Large language models like GPT-3 and GPT-4, upon which popular applications such as ChatGPT are built, require the transformer architecture [Source: Sajun et al. (2024)].

3.2. Pretraining Phase

The pretraining phase is the resource-consuming and core component in the development of ChatGPT. At this level, the model is trained with enormous amounts of text of vast subjects, including literature, scientific literature, the web, and code repositories (Ray, 2023). Learning proceeds through a *self-supervised process*, in that the system has to predict the next element of a sequence based on its context. Formally, the model is trained to minimize the cross-entropy loss between the probability distribution over the predicted tokens and Tokens, as it, in fact, is observed (Konstantakos et al., 2024).

With this paradigm of training, the model can deduce statistical patterns in language, not only at the basic level of syntax and grammar, but also at higher levels of semantics and discourse structure (Han et al., 2024). Consequently, the pretrained model develops general language ability and acquires masses of factual knowledge based on training corpora. Though potent, these capabilities are raw, biased, inconsistent, and any inherent risks of the raw data can be passed on through the learned representations of the model (Ferrara, 2023).

Not only is the pretraining process particularly resource-intensive, often requiring weeks of TPUs or GPU clusters to complete (Schmidt & Hildebrandt, 2024), but it also causes issues of model and data reproducibility due to the long, probabilistic nature of the pretraining process (Nakkiran et al., 2023). Even with these costs, pretraining remains essential to the extent that it forms the basis upon which subsequent steps (i.e., supervised fine-tuning and reinforcement learning with human feedback (RLHF)) build model-specific and safety-aligned capabilities upon (Dahlgren Lindström et al., 2025).

3.3. Fine-Tuning Phase

Once pretraining establishes broad linguistic competence, the model undergoes fine-tuning to align its behavior with human expectations. The first step is supervised fine-tuning (SFT), where the pre-trained model is trained with a smaller but well-curated collection of prompts and high-quality responses, which were created by human experts (Punnaivanam & Velvizhy, 2024). This stage teaches the system to take its orders and come up with outputs that seem like outputs of a human being in terms of structure and style. In contrast to pretraining, which is data-hungry and unsupervised, SFT is data-efficient, supervised, and follows task-specific rules and guidelines (Wolfe, 2023).

To further regularise alignment, reinforcement learning with human feedback (RL HF) is implemented. This involves sampling various possible candidate responses of the model, which are then ranked by human annotators on quality, helpfulness, and safety. These rankings are then trained into a reward model that then brings about the reinforcement learning optimization process of the language model in the use of Proximal Policy Optimization (PPO) (Rizki et al., 2025). LHF allows the system to outperform based on a purely static dataset by using human preference signals as direct input into the learning process.

Together, supervised fine-tuning and RLHF significantly enhance the model's practical utility. The fine-tuned model generates not only syntactically valid text but also contextually relevant, safe, and user-aligned responses (Wolfe, 2023).

However, this stage is not without limitations: annotator bias can influence preference data, and excessive optimization toward "safe" outputs may reduce creativity or specificity (Chen et al., 2023). Despite these challenges, fine-tuning remains essential for transforming a general-purpose pretrained model into an interactive system that reliably serves human users (Anisuzzaman et al., 2024).

3.4. Model Scaling

One of the defining features of ChatGPT's effectiveness is the principle of scaling: systematically increasing the size of the model, the volume of training data, and the compute resources used. Empirical research on scaling laws has shown that as the number of parameters grows into the billions, language models achieve progressively lower loss values and demonstrate qualitatively new capabilities (Budnikov et al., 2024). These emergent abilities include complex reasoning, multilingual translation, long-context coherence, and more reliable generalization across domains, capabilities that smaller models fail to exhibit.

Empirical scaling studies confirm these trends, demonstrating a predictable power-law relationship between model size, compute, and performance. As shown in Figure 4, training loss decreases predictably as model size increases, but only up to an optimal point determined by the compute budget. The isoFLOPs slices represent constant compute allocations, illustrating the trade-offs between allocating resources to larger models versus training them on more data. These results highlight the empirical scaling laws that underpin modern large language model development (1a3orn, 2022).

Realizing such gains requires massive distributed infrastructure. Training is conducted across clusters of GPUs or TPUs using advanced optimization strategies (Aldoseri et al., 2023). Techniques such as mixed-precision training (FP16 or bfloat16 arithmetic) reduce memory consumption and accelerate computation, while parallelization methods, including data parallelism, tensor/model parallelism, and pipeline parallelism, enable efficient handling of billions of parameters. Without these strategies, training runs would be infeasible given hardware constraints.

However, scaling also introduces significant challenges. Larger models demand exponentially greater energy and financial costs, raising concerns of accessibility and environmental impact (Alzoubi & Mishra, 2024). Inference latency becomes a practical bottleneck, as serving such models requires substantial memory and bandwidth. Moreover, scaling alone does not resolve issues of bias, factuality, or safety; in some cases, it may even amplify them.

Going forward, scaling is expected to evolve beyond brute-force parameter growth. Research into sparse architectures (e.g., Mixture of Experts), efficient fine-tuning methods, and hardware–software co-optimization suggests a future in which models maintain or even improve performance while reducing training costs. Thus, while scaling remains central to ChatGPT's success, sustainable innovation increasingly depends on balancing size with efficiency (Haefner et al., 2023).

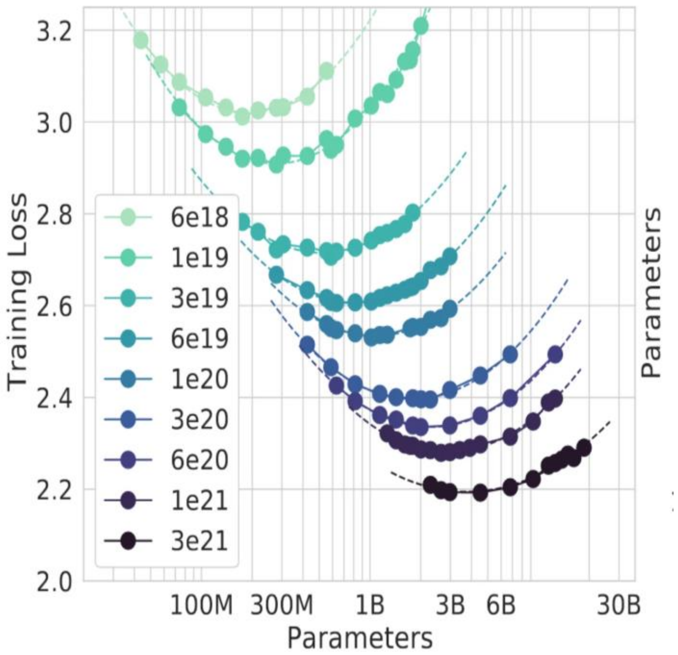

Figure 4. IsoFLOPs scaling behavior of large language models. Training loss plotted against model parameters under fixed compute budgets (FLOPs). Each curve represents an isoFLOP slice, illustrating the trade-off between model size and training duration, with optimal performance achieved near the minima [Source: 1a3orn (2022)].

4. Applications Across Domains

4.1. Education and Knowledge Work

ChatGPT and other large language models (LLMs) have quickly become revolutionary applications in education and professional knowledge work. In education, LLMs serve as intelligent tutoring systems, offering personalized feedback, interactive questions, and formative feedback to learners at scale (Sharma et al., 2025). In contrast to the traditional e-learning systems, the LLM can produce context-specific responses and examples specific to the learner, and dynamically scaffold difficult concepts. In other words, they can help students in solving mathematics problems by breaking them down into multiple steps or give some context on the background of a social science assignment. These adapting tutors are similar to human teachers in that they enhance the reach of tutoring to underrepresented groups and decompress educational systems (Kamalov et al., 2023).

The LLMs used in knowledge work aid in improving productivity through information retrieval, summarization, and drafting of documents (Freire et al., 2024). Professionals can quickly conclude a quick reading of large reports, legal statements, or articles that may have a lot of pages. LLMs can support knowledge augmentation through draft generation, supporting creative thought, and offering alternative formulations of an argument (Qin, Chen, et al., 2025). Notably, these systems are also multilingually accessible, which facilitates cross-linguistic communication and eliminates some of the obstacles within the globalized workplaces. Nevertheless, several issues are outstanding: excessive dependence on LLMs can lead to a loss of critical thinking abilities in students, and unverified outputs in a professional environment would only add misinformation or errors in areas that include law, policy, etc. Thus, their role is best conceptualized as a collaborative assistant, augmenting rather than replacing human judgment (Shahzad et al., 2025).

4.2. Healthcare and Clinical Decision Support

One of the most promising yet sensitive areas of application of LLM is the field of healthcare. The LLMs have the potential to process patient information, incorporate medical knowledge, and present a list of possible diagnoses (Singhal et al., 2025). For instance, they have been tested in generating SOAP (Subjective, Objective, Assessment, Plan) notes, summarizing radiology reports, and supporting triage decisions (Delanerolle et al., 2025). Using biomedical corpora (e.g., PubMed, clinical notes) to train or fine-tune increases the accuracy of models trained on general-purpose (medical reasons, Renaudo et al., 2015).

Beyond diagnostics, LLMs hold potential in patient-centered applications: a virtual health assistant to respond to basic health-related questions, provide pill reminders, and more. This saves time for physicians and enhances patient knowledge and compliance. In addition to their use in drug discovery and genomics, LLMs can search literature, generate protein-ligand models, and streamline hypothesis generation in biomedical research pipelines (Maity & Jyoti, 2025).

Nonetheless, the healthcare domain raises acute risks. Hallucinated outputs, lack of accountability, and bias in training data can result in unsafe recommendations (Farhud & Zokaei, 2021). Ethical concerns regarding privacy (HIPAA/GDPR compliance), explainability, and liability must be addressed before widespread deployment in clinical workflows. Regulatory frameworks such as the FDA’s Software as a Medical Device (SaMD) guidance will play a central role in determining adoption. Overall, LLMs are positioned not as autonomous diagnosticians but as decision-support systems that extend clinician capacity while requiring stringent oversight (FDA, 2019).

4.3. Business and Productivity

In the business domain, ChatGPT serves as a general-purpose productivity accelerator across functions, including customer service, operations, marketing, and management. In customer engagement, conversational agents powered by LLMs provide 24/7 support, handle routine inquiries, and escalate complex issues to human agents (Abdelkader, 2023). This improves efficiency while maintaining natural, human-like interaction quality. In marketing and communications, LLMs generate campaign copy, tailor messaging for different audience segments, and even conduct A/B testing simulations of promotional content (Linkon et al., 2024).

For knowledge workers, LLMs automate drafting of reports, meeting minutes, and email correspondence, significantly reducing cognitive load (Naqbi et al., 2024). In software engineering, code-oriented models (e.g., Codex, GitHub Copilot) accelerate development cycles by generating boilerplate code, suggesting optimizations, and supporting debugging (Moradi Dakhel et al., 2023). Similarly, in finance and business analysis, LLMs assist with data interpretation, scenario modeling, and natural-language querying of databases, thereby democratizing access to analytics for non-technical stakeholders (Filippo et al., 2024).

However, integration into enterprise contexts requires addressing accuracy, intellectual property concerns, and security risks. Over-reliance on auto-generated outputs without verification may lead to reputational or financial harm (European Commission, 2025). Additionally, organizations must implement human-in-the-loop validation pipelines and establish governance policies for responsible AI deployment. When appropriately managed, LLMs serve as productivity amplifiers, reshaping workflows and enhancing decision-making across industries.

4.4. Scientific Research Assistance

Scientific research is another frontier where LLMs are proving indispensable. At the literature synthesis stage, LLMs can read thousands of articles in a short period of time and extract structured knowledge, but also discover new patterns of research in the field (Fabiano et al., 2024). Tools built on LLMs are increasingly used for systematic reviews, where they assist in screening abstracts, extracting metadata, and drafting summaries. In hypothesis generation, LLMs can highlight overlooked connections across disciplines, for instance, linking biological pathways of molecules conveying information to that of stress adaptation in plants or candidate gene-trait associations in crop enhancement (Abdel-Rehim et al., 2025).

In experimental design and analysis, LLMs provide guidance on methodology, suggest statistical tests, and assist in writing code for data processing (e.g., R, Python scripts) (Coello et al., 2024). They are also integrated into computational pipelines, where they help interpret high-dimensional data in genomics, proteomics, and climate modeling (Yoosefzadeh-Najafabadi, 2025). For example, combining LLM-based reasoning with symbolic mathematics has shown promise in automating portions of theoretical physics research (Pantsar, 2025).

Yet, the use of LLMs in research raises concerns of hallucination, citation fabrication, and over-automation (Ji et al., 2022). To maintain scientific integrity, outputs must be cross-validated with primary sources and human expertise. Moreover, biases in the training corpus risk amplifying dominant paradigms while overlooking minority or non-Western knowledge traditions. Still, when carefully supervised, LLMs act as research accelerators, augmenting human creativity and enabling faster, more integrative scientific discovery.

Table 2. Empirical studies of ChatGPT and domain-specific LLM applications: benefits, risks, validation evidence.

Domain	**Typical benefits**	**Main risks/limitations**	**Key studies (2022–2025)**
Education & Knowledge Work	Personalized tutoring; faster learning & stronger engagement; high-quality writing feedback; boosts to knowledge-worker output	Hallucinations and factual errors; risk of over-reliance/skill atrophy; mixed or context-dependent learning gains	**AI tutoring RCT:** An experimental study by *Kestin et al. (2025)* demonstrated that AI tutoring using ChatGPT achieved higher learning gains than traditional in-class active learning approaches. **Systematic reviews:** Recent syntheses (2024), including *Deng et al. (2024)*, report that ChatGPT can improve academic performance, though outcomes vary across contexts and concerns about over-reliance remain. **Knowledge work RCT:** In a randomized controlled trial, *Noy and Zhang (2023)* found that ChatGPT significantly increased professional writing productivity and improved overall quality.
Healthcare & Clinical Decision Support	Exam-style clinical reasoning; medical QA with domain-tuned LLMs; drafting notes &	**Not ready for autonomous use**; instruction sensitivity & brittleness; safety, bias, privacy; clinician de-skilling with over-reliance	**USMLE performance:** *Kung et al. (2023)* reported that ChatGPT achieved at or near the passing threshold across multiple USMLE Step examinations, highlighting its potential for medical education but also exposing variability in domain-specific reasoning.

	summarizing reports		**Med-PaLM 2:** *Singhal et al. (2025)* demonstrated state-of-the-art performance on MedQA and other clinical benchmarks, though important limitations in safety and robustness remain. **Limits & safety:** A *Nature Medicine* analysis by *Hager et al. (2024)* emphasized that large language models are not yet suitable for autonomous clinical decision-making, citing risks related to reliability, explainability, and patient safety. **De-skilling evidence:** In a randomized controlled trial, *Jeyaretnam (2025)* reported that gastroenterologists using AI-assisted colonoscopy improved immediate detection rates but showed reduced performance when operating without AI support, raising concerns about clinician de-skilling.
Business & Productivity	Higher throughput in customer support; faster, better drafts; code generation & review; measurable quality gains for less-experienced staff	Uneven effects across workers; potential quality dips for top performers; hallucinations/brand risk; data/IP exposure without governance	**Generative-AI assistant:** *Brynjolfsson et al. (2025)* found that a generative AI–based customer support assistant increased agent productivity by approximately 14–15% in terms of resolved tickets per hour. **Professional writing RCT:** In a randomized controlled trial, *Noy and Zhang (2023)* reported that ChatGPT substantially improved the productivity and quality of mid-level professional writing tasks. **Software development RCTs:** *Cui et al. (2024)* demonstrated through enterprise-scale randomized trials that GitHub Copilot significantly accelerated software development workflows and improved coding efficiency.
Scientific Research Assistance	Faster literature synthesis & screening; hypothesis generation; autonomous planning/ execution of lab tasks; domain tool-use (chemistry, materials)	Hallucinated claims/citations; reproducibility & provenance concerns; potential misuse in sensitive domains	**Autonomous lab agent:** *Boiko et al. (2023)* introduced *Coscientist*, an autonomous laboratory agent capable of designing, planning, and executing chemistry experiments, demonstrating the feasibility of LLM-driven scientific automation. **Chemistry agent:** *Bran et al. (2024)* developed *ChemCrow*, an LLM integrated with domain-specific tools for chemical synthesis, drug discovery, and materials science, showing improved performance in complex scientific tasks. **Evidence synthesis:** *Lai et al. (2025)* reported that LLMs achieved promising accuracy and efficiency in risk-of-bias assessment and data extraction for systematic reviews, highlighting their potential role in accelerating evidence synthesis.

░5. Limitations and Challenges

Despite their wide-ranging applications, large language models such as ChatGPT face important limitations and challenges that constrain their safe and reliable use. The most notable issue is the hallucination that results in fluent, yet wrong information produced by the model (Ji et al., 2022). These mistakes are usually presented with unnecessary assurance and are thus only hard to identify with the non-expert user. These hallucinations have the potential to compromise faith in education, medical, and scientific situations where accuracy is essential.

Important as well is the consideration of bias and ethics. Since training corpora are rife with imbalances of groups of various nations, societies, and demographic subsets, the models will risk reproducing stereotypes or legitimizing inequities (Blodgett et al., 2020). Biasness may potentially sideline the voice of underrepresented groups and exacerbate the pre-existing inequalities, and therefore, high-stakes tasks like hiring, lending, or treating patients cannot be applied (Obermeyer et al., 2019).

Another challenge is the lack of explainability. LLMs function as complex black boxes, producing outputs without transparent reasoning pathways (Marques-Silva & Ignatiev, 2023). This opacity complicates auditing, reduces user trust, and poses barriers to regulatory acceptance in sensitive fields like healthcare, finance, and law.

Performance is also limited by the quality of its data. The models are learned based on static data, so they may have unreliable knowledge and have blind spots in new spheres (Schneider et al., 2024). Moreover, the training data can be biased, of low quality, or not representative enough of different global knowledge traditions, and limit model generalization (Ferrara, 2023).

Lastly, issues on the environmental impact of scaling are on the rise. The training of state-of-the-art LLMs needs enormous computing resources, and the energy used generates a substantial number of carbon emissions (Strubell et al., 2020). The tension between performance and sustainable ecological footprints has emerged as a critical discussion in AI research, as it leads to various appeals to more efficient and environmentally-friendly architectures and practices of AI.

░6. Ongoing Improvements and Research Gaps

The achieved capabilities of large language models, like ChatGPT, have been remarkable, but a lot of research is underway to overcome current limitations and push the functionality.

A significant trend is the emergence of multi-modal modeling to include text and images with audio, video, and code. Such systems can interpret visual questions, diagnostic images, and clinical notes, or speech and text information together by combining modalities. New architectures such as GPT-4V and Gemini demonstrate that multimodality can be used in a much broader variety of contexts than just in some form of text-based interaction (OpenAI, 2023).

Another research direction is retrieval-augmented generation (RAG). RAG pipelines merge external knowledge bases or live web retrieval since pretrained models are limited by the static nature of the training corpora, and thus, outdated information cannot be used (Lewis et al., 2020). This minimizes hallucinations and factual anchoring, and enables models to stay fresh without complete retraining.

Domain-specific fine-tuning is also gaining importance. By fine-tuning general-purpose LLMs on specialised data, the models can have greater reliability in medicine, law, finance, and other expert-driven areas. Examples include Med-PaLM, used in clinical reasoning and LawGPT, which is used in legal analysis (Singhal et al., 2025). However, fine-tuning complicates the task of being generative and, at the same time, being able to adapt to the robustness of the domain.

Advances are also being made in interpretability research, towards making LLM decision-making transparent. Such methods as probing into hidden states, mechanistic interpretability, and causal investigation of attention heads are being developed to determine how models lead to outputs (Belinkov & Glass, 2019). Explainability should be enhanced as necessary, since it is a prerequisite to trust, introduction into regulation, and scientific comprehension of the properties of emergent capabilities.

Finally, there is a shift to energy-efficient training and deployment (also called green AI). Other strategies like sparse architectures, parameter-efficient fine-tuning, quantization, and better hardware-software co-design would help minimize the carbon impact of training, which would not require the trade-off between performance and carbon footprint of training. Sustainability is becoming an ever more understood aspect of ethical AI creation.

Taken together, these advances highlight both active areas of innovation and the research gaps that remain. Future work must balance expanding functionality with safety, interpretability, and sustainability, ensuring that LLMs evolve into trustworthy and accessible tools across domains.

Table 3. Proposed improvements in large language models, with their technical basis and associated challenges.

Proposed Improvement	**Technical Basis**	**Key Challenges**
Multi-modal capabilities (text, image, audio, code)	Integration of transformer backbones with vision encoders, speech recognition, and code parsers (e.g., GPT-4V, Gemini)	High computational cost; difficulty in aligning modalities; bias transfer across domains
Retrieval-Augmented Generation (RAG)	Combining LLMs with external retrieval modules (databases, search engines) to ground responses in real-time knowledge	Reliability of retrieval sources; latency in large-scale deployment; risk of propagating external misinformation
Domain-specific fine-tuning (medicine, law, finance)	Parameter-efficient fine-tuning (LoRA, adapters), supervised alignment on specialized datasets (e.g., Med-PaLM, LawGPT)	Data scarcity and privacy constraints; risk of overfitting to niche corpora; reduced generalization
Interpretability research	Mechanistic interpretability, probing hidden states, and causal tracing of attention patterns	Limited scalability to very large models; lack of consensus on evaluation metrics; risk of oversimplification
Energy-efficient training (“Green AI”)	Sparse architectures (Mixture of Experts), quantization, distillation, hardware–software co-optimization	Trade-off between efficiency and accuracy; hardware accessibility; measuring and standardizing carbon footprint

Note: Data compiled and adapted from OpenAI (2023), Lewis et al. (2020), Singhal et al. (2023), Belinkov and Glass (2019), Strubell et al. (2020), and Related sources.

7. The Future of ChatGPT and Generative AI

Large language models are currently showing trends towards moving beyond conversational agents to more robust, end-to-end tool-using systems. Integration with agents enables models to execute chains of reasoning, interface with external APIs, and/or perform multi-step tasks in real or simulated worlds. Initial systems like AutoGPT and LangChain demonstrate how LLMs, memory, planning, and external apps make them a system of any general purpose (Github, 2023). This integration points toward a future in which conversational models act as cognitive operating systems, orchestrating workflows across domains from personal productivity to industrial automation.

Another critical frontier is the convergence of LLMs with data science and analytics. Current models excel at natural language reasoning but lack direct integration with structured data streams. By bridging this gap, future systems could perform real-time data querying, statistical inference, and decision optimization directly within conversational interfaces (Yao et al., 2024). Such integration would allow professionals in fields such as finance, logistics, or healthcare to engage with complex datasets through natural dialogue, transforming LLMs into interactive data science partners rather than static information providers.

Hybrid architectures combining neural and symbolic approaches are also likely to define the next wave of generative AI. While transformer-based models capture statistical regularities, they struggle with logical consistency and compositional reasoning. Integrating symbolic reasoning engines or knowledge graphs with neural networks could enhance reliability, factual grounding, and mathematical or logical problem-solving (Garcez & Lamb, 2023). Research into neuro-symbolic AI, program induction, and differentiable reasoning modules suggests that hybrid systems may overcome current weaknesses of purely statistical LLMs, especially in domains requiring formal reasoning or explainability.

Finally, the future of generative AI will be shaped by governance, regulation, and ethical oversight. As models gain autonomy and permeate critical sectors, questions of accountability, safety, intellectual property, and societal impact will grow more pressing. Policymakers and researchers are already debating frameworks for AI auditing, transparency standards, watermarking, and responsible deployment (European Commission, 2025). International coordination will be essential to balance innovation with safeguards, ensuring that generative AI develops as a trustworthy public good rather than a source of unchecked risk.

Taken together, these trajectories suggest that the future of ChatGPT and similar systems will not be defined by scale alone but by deeper integration with tools, reasoning systems, structured data, and governance frameworks. The next generation of generative AI is thus poised to become not just more powerful, but also more interactive, interpretable, and accountable.

8. Conclusion

ChatGPT exemplifies both the transformative potential and the unresolved challenges of generative AI. Its success across education, healthcare, business, and research reflects advances in pretraining, fine-tuning, and scaling, yet persistent issues of hallucination, bias, opacity, data limits, and environmental cost reveal the limits of scale alone. As a benchmark system, ChatGPT highlights the urgent need for progress in interpretability, domain-specific alignment, sustainable training, and governance, ensuring that future models evolve to be not only more capable but also more reliable, transparent, and socially responsible.

9. Recommendations

1) Prioritize interpretability: Readiness on mechanical interpretability and unbiased assessment structures will be of utmost value in building user trust and regulatory acceptance.

2) Adopt domain-specific alignment: Fine-tune models responsibly to medicine, law, and other high-stakes domains with privacy-respecting curated datasets.

3) Advance sustainable AI practices: Develop energy-efficient architectures and employ eco-friendly AI procedures to minimize the CO2 output of training and inference.

4) Strengthen governance and regulation: Establish global standards for accountability, auditing, watermarking, and good deployment.

5) Promote human–AI collaboration: Position LLMs as decision-support agents and not as decision-making replacements, noting that oversight on occurrences linked to sensitive applications is required.

6) Expand inclusivity in training data: Increase the amount of coverage in low-resource languages and different cultural settings to minimize bias and open training data up to a greater amount of diversity.

7) Encourage multi-modal and hybrid systems: Support multi-modality such as text, image, audio and symbolic reasoning to make applications richer and more reliable in the real world.

Declarations

Source of Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests Statement

The authors have declared that no competing financial, professional, or personal interests exist.

Consent for publication

All the authors contributed to the manuscript and consented to the publication of this research work.

Authors' contributions

All the authors took part in literature review, analysis, and manuscript writing equally.

Availability of data and materials

Supplementary information is available from the authors upon reasonable request.

Institutional Review Board Statement

Not applicable for this study.

Informed Consent

Not applicable for this study.

References

1a3orn (2022). New scaling laws for large language models. Lesswrong.com. https://www.lesswrong.com/posts/midxmmb2xg37f2kgn/new-scaling-laws-for-large-language-models.

Abdel-Rehim, A., Zenil, H., Orhobor, O., Fisher, M., Collins, R.J., Bourne, E., Fearnley, G.W., Tate, E., Smith, H.X., Soldatova, L.N., & King, R. (2025). Scientific hypothesis generation by large language models: Laboratory validation in breast cancer treatment. Journal of the Royal Society Interface, 22(227). https://doi.org/10.1098/rsif.2024.0674.

Abdelkader, O.A. (2023). ChatGPT's influence on customer experience in digital marketing: Investigating the moderating roles. Heliyon, 9(8): e18770–e18770. https://doi.org/10.1016/j.heliyon.2023.e18770.

Al-Busaidi, A.S., Raman, R., Hughes, L., Albashrawi, M.A., Malik, T., Dwivedi, Y.K., Al-Alawi, T., AlRizeiqi, M., Davies, G., Fenwick, M., Gupta, P., Gurpur, S., Hooda, A., Jurcys, P., Lim, D., Lucchi, N., Misra, T., Raman, R., Shirish, A., & Walton, P. (2024). Redefining boundaries in innovation and knowledge domains: Investigating the impact of generative artificial intelligence on copyright and intellectual property rights. Journal of Innovation & Knowledge, 9(4): 100630. https://doi.org/10.1016/j.jik.2024.100630.

Aldoseri, A., Khalifa, K.N.A., & Hamouda, A.M. (2023). Re-Thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. Applied Sciences, 13(12): 7082. https://doi.org/10.3390/app13127082.

Alzoubi, Y.I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. Journal of Cleaner Production, 468(143090): 143090. https://doi.org/10.1016/j.jclepro.2024.143090.

Anisuzzaman, D.M., Malins, J.G., Friedman, P.A., & Attia, Z.I. (2024). Fine-Tuning LLMs for specialized use cases. Mayo Clinic Proceedings: Digital Health, 3(1). https://doi.org/10.1016/j.mcpdig.2024.11.005.

Bakagianni, J., Pouli, K., Gavriilidou, M., & Pavlopoulos, J. (2025). A systematic survey of natural language processing for the Greek language. Patterns, Pages 101313–101313. https://doi.org/10.1016/j.patter.2025.101313.

Balaskas, G., Papadopoulos, H., Pappa, D., Loisel, Q., & Chastin, S. (2025). A framework for domain-specific dataset creation and adaptation of large language models. Computers, 14(5): 172. https://doi.org/10.3390/computers14050172.

Banik, D., Pati, N., & Sharma, A. (2024). Systematic exploration and in-depth analysis of ChatGPT architectures' progression. Artificial Intelligence Review, 57(10). https://doi.org/10.1007/s10462-024-10832-0.

Bao, D., Liu, X., Xu, Y., Fang, Q., & He, X. (2024). Detection of defective apples using learnable residual multi-head attention networks integrated with CNNs. Electronics, 13(24): 4861–4861. https://doi.org/10.3390/electronics13244861.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7: 49–72. https://doi.org/10.1162/tacl_a_00254.

Bengesi, S., El-Sayed, H., Sarker, M.K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion models, and transformers. IEEE Access, 12: 69812–69837. https://doi.org/10.1109/access.2024.3397775.

Blodgett, S.L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485.

Boiko, D.A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624(7992): 570–578. https://doi.org/10.1038/s41586-023-06792-0.

Bran, A.M., Cox, S., Schilter, O., Baldassari, C., White, A.D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. Nature Machine Intelligence, 6. https://doi.org/10.1038/s42256-024-00832-8.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2). https://doi.org/10.1093/qje/qjae044.

Budnikov, M., Bykova, A., & Yamshchikov, I.P. (2024). Generalization potential of large language models. Neural Computing and Applications, 37. https://doi.org/10.1007/s00521-024-10827-6.

Busch, F., Hoffmann, L., Rueger, C., van Dijk, E.H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J.N., Truhn, D., Cuocolo, R., Adams, L.C., & Bressem, K.K. (2025). Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine, 5(1). https://doi.org/10.1038/s43856-024-00717-2.

Chen, Y., Clayton, E.W., Novak, L.L., Anders, S., & Malin, B. (2023). Human-Centered design to address biases in artificial intelligence. Journal of Medical Internet Research, 25(1). https://doi.org/10.2196/43251.

Cheong, B.C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. Frontiers in Human Dynamics, 6. https://doi.org/10.3389/fhumd.2024.1421273.

Choi, S.R., & Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: A comprehensive review. Biology, 12(7): 1033. https://doi.org/10.3390/biology12071033.

Coello, C.E.A., Alimam, M.N., & Kouatly, R. (2024). Effectiveness of ChatGPT in coding: A comparative analysis of popular large language models. Digital, 4(1): 114–125. https://doi.org/10.3390/digital4010005.

Cui, K.Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2024). The productivity effects of generative AI: Evidence from a field experiment with GitHub Copilot. An MIT Exploration of Generative AI. https://doi.org/10.21428/e4baedd9.3ad85f1c.

Dahlgren Lindström, A., Methnani, L., Krause, L., Ericson, P., de Rituerto de Troya, Í.M., Coelho Mollo, D., & Dobbe, R. (2025). Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through reinforcement learning from human feedback. Ethics and Information Technology, 27(2). https://doi.org/10.1007/s10676-025-09837-2.

Delanerolle, G., Bouchareb, Y., Shetty, S., Cavalini, H., & Phiri, P. (2025). A pilot study using natural language processing to explore textual electronic mental healthcare data. Informatics, 12(1): 28. https://doi.org/10.3390/informatics12010028.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 227: 105224. https://doi.org/10.1016/j.compedu.2024.105224.

Dilmegani, C., & Sezer, S. (2025). GPT-5: Best features, pricing & accessibility in 2025. AIMultiple. https://research.aimultiple.com/gpt-5/.

Ducato, R. (2020). Data protection, scientific research, and the role of information. Computer Law & Security Review, 37(1). https://doi.org/10.1016/j.clsr.2020.105412.

European Commission (2025). AI Act. European Commission. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Fabiano, N., Gupta, A., Bhambra, N., Luu, B., Wong, S., Maaz, M., Fiedorowicz, J.G., Smith, A.L., & Solmi, M. (2024). How to optimize the systematic review process using AI tools. JCPP Advances, 4(2). https://doi.org/10.1002/jcv2.12234.

Farhud, D.D., & Zokaei, S. (2021). Ethical issues of artificial intelligence in medicine and healthcare. Iranian Journal of Public Health, 50(11): 1–5. https://doi.org/10.18502/ijph.v50i11.7600.

FDA (2019). Search for FDA guidance documents. U.S. Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1): 3. https://doi.org/10.3390/sci6010003.

Filippo, C., Vito, G., Irene, S., Simone, B., & Gualtiero, F. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. Technovation, 133: 103002–103002. https://doi.org/10.1016/j.technovation.2024.103002.

Freire, S.K., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., & Niforatos, E. (2024). Knowledge sharing in manufacturing using LLM-powered tools: User study and model benchmarking. Frontiers in Artificial Intelligence (Lausanne), 7. https://doi.org/10.3389/frai.2024.1293084.

Garcez, A.D., & Lamb, L.C. (2023). Neurosymbolic AI: The 3rd wave. Artificial Intelligence Review, 56. https://doi.org/10.1007/s10462-023-10448-w.

Github (2023). AutoGPT: the heart of the open-source agent ecosystem. GitHub. https://github.com/significant-gravitas/autogpt.

González Barman, K., Lohse, S., & de Regt, H.W. (2025). Reinforcement learning from human feedback in LLMS: Whose culture, whose values, whose perspectives? Philosophy & Technology, 38(2). https://doi.org/10.1007/s13347-025-00861-0.

Haefner, N., Parida, V., Gassmann, O., & Wincent, J. (2023). Implementing and scaling artificial intelligence: A review, framework, and research agenda. Technological Forecasting and Social Change, 197: 122878–122878. https://doi.org/10.1016/j.techfore.2023.122878.

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine, 30: 1–10. https://doi.org/10.1038/s41591-024-03097-1.

Haleem, A., Javaid, M., & Singh, R.P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2(4): 100089. https://doi.org/10.1016/j.tbench.2023.100089.

Han, S., Wang, M., Zhang, J., Li, D., & Duan, J. (2024). A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. Electronics, 13(24): 5040–5040. https://doi.org/10.3390/electronics13245040.

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: How language modeling bias causes epistemic injustice. Ethics and Information Technology, 26(1). https://doi.org/10.1007/s10676-023-09742-6.

Javaid, M., Haleem, A., Singh, R.P., Khan, S., & Khan, I.H. (2023). Unlocking the opportunities through ChatGPT tool towards ameliorating the education system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(2): 100115–100115. https://doi.org/10.1016/j.tbench.2023.100115.

Jeyaraman, M., Ramasubramanian, S., Balaji, S., Jeyaraman, N., Nallakumarasamy, A., & Sharma, S. (2023). ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World Journal of Methodology, 13(4): 170–178. https://doi.org/10.5662/wjm.v13.i4.170.

Jeyaretnam, M. (2025). New study suggests using AI made doctors less skilled at spotting cancer. TIME. https://time.com/7309274/ai-lancet-study-artificial-intelligence-colonoscopy-cancer-detection-medicine-deskilling/?utm_source=chatgpt.com.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12). https://doi.org/10.1145/3571730.

Jin, I., Tangsrivimol, J.A., Darzi, E., Hassan Virk, H.U., Wang, Z., Egger, J., Hacking, S., Glicksberg, B.S., Strauss, M., & Krittanawong, C. (2025). DeepSeek vs. ChatGPT: Prospects and challenges. Frontiers in Artificial Intelligence, 8. https://doi.org/10.3389/frai.2025.1576992.

Kamalov, F., Calonge, D.S., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. Sustainability, 15(16): 12451. https://doi.org/10.3390/su151612451.

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1): 1–10. https://doi.org/10.1038/s41598-025-97652-6.

Khan, S., Fazil, M., Imoize, A.L., Alabduallah, B.I., Albahlal, B.M., Alajlan, S.A., Almjally, A., & Siddiqui, T. (2023). Transformer architecture-based transfer learning for politeness prediction in conversation. Sustainability, 15(14): 10828. https://doi.org/10.3390/su151410828.

Konstantakos, S., Cani, J., Mademlis, I., Chalkiadaki, D.I., Asano, Y.M., Efstratios Gavves, E., & Papadopoulos, G.T. (2024). Self-supervised visual learning in the low-data regime: A comparative evaluation. Neurocomputing, 620: 129199–129199. https://doi.org/10.1016/j.neucom.2024.129199.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2): e0000198. https://doi.org/10.1371/journal.pdig.0000198.

Lai, H., Liu, J., Bai, C., Liu, H., Pan, B., Luo, X., Hou, L., Zhao, W., Xia, D., Tian, J., Chen, Y., Zhang, L., Estill, J., Liu, J., Liao, X., Shi, N., Sun, X., Shang, H., Bian, Z., & Yang, K. (2025). Language models for data extraction and risk of bias assessment in complementary medicine. Npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-025-01457-w.

Lakatos, R., Pollner, P., Hajdu, A., & Joó, T. (2025). Investigating the performance of retrieval-augmented generation and domain-specific fine-tuning for the development of AI-driven knowledge-based systems. Machine Learning and Knowledge Extraction, 7(1): 15–15. https://doi.org/10.3390/make7010015.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2022.acl-long.577.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented generation for knowledge-intensive NLP tasks. Neural Information Processing Systems, 33: 9459–9474.

Li, R., Xu, J., Cao, Z., Zheng, H.T., & Kim, H.G. (2024). Extending context window in large language models with segmented base adjustment for rotary position embeddings. Applied Sciences, 14(7): 3076–3076. https://doi.org/10.3390/app14073076.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI Open, 3. https://doi.org/10.1016/j.aiopen.2022.10.001.

Linkon, A.A., Shaima, M., Sarker, M.S.U., Badruddowza, Nabi, N., Rana, M.N.U., Ghosh, S.K., Rahman, M.A., Esa, H., & Chowdhury, F.R. (2024). Advancements and applications of generative artificial intelligence and large language models on business management: A comprehensive review. Journal of Computer Science and Technology Studies, 6(1): 225–232. https://doi.org/10.32996/jcsts.2024.6.1.26.

Lopez-Gazpio, I. (2025). Integrating large language models into accessible and inclusive education: Access democratization and individualized learning enhancement supported by generative artificial intelligence. Information, 16(6): 473. https://doi.org/10.3390/info16060473.

Maity, S., & Jyoti, S.M. (2025). Large language models in healthcare and medical applications: A review. Bioengineering, 12(6): 631. https://doi.org/10.3390/bioengineering12060631.

Marques-Silva, J., & Ignatiev, A. (2023). No silver bullet: Interpretable ML models must be explained. Frontiers in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1128212.

Mars, M. (2022). From word embeddings to pre-trained language models: A state-of-the-art walkthrough. Applied Sciences, 12(17): 8805. https://doi.org/10.3390/app12178805.

Menon, D., & Shilpa, K. (2023). "Chatting with ChatGPT": Analyzing the factors influencing users' intention to use the open AI's ChatGPT using the UTAUT model. Heliyon, 9(11). https://doi.org/10.1016/j.heliyon.2023.e20962.

Mienye, I.D., & Swart, T.G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. Information, 15(12): 755. https://doi.org/10.3390/info15120755.

Moradi Dakhel, A., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M.C., & Jiang, Z.M. (2023). GitHub Copilot AI pair programmer: Asset or liability? Journal of Systems and Software, 203: 111734. https://doi.org/10.1016/j.jss.2023.111734.

Mswahili, M.E., & Jeong, Y.S. (2024). Transformer-based models for chemical SMILES representation: A comprehensive literature review. Heliyon, 10(20): e39038. https://doi.org/10.1016/j.heliyon.2024.e39038.

Naqbi, H.A., Bahroun, Z., & Ahmed, V. (2024). Enhancing work productivity through generative artificial intelligence: A comprehensive literature review. Sustainability, 16(3): 1166. https://doi.org/10.3390/su16031166.

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. Artificial Intelligence in Medicine, 154: 102900. https://doi.org/10.1016/j.artmed.2024.102900.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654): 187–192. https://doi.org/10.1126/science.adh2586.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464): 447–453. https://doi.org/10.1126/science.aax2342.

OpenAI. (2023). GPT-4 technical report. OpenAI. https://cdn.openai.com/papers/gpt-4.pdf.

Pantsar, M. (2025). The need for ethical guidelines in mathematical research in the time of generative AI. AI and Ethics, 5. https://doi.org/10.1007/s43681-025-00660-5.

Punnaivanam, M., & Velvizhy, P. (2024). Contextual fine-tuning of language models with classifier-driven content moderation for text generation. Entropy, 26(12): 1114. https://doi.org/10.3390/e26121114.

Qin, H., Li, M., Wang, L., Ge, Y., Zhu, J., & Zheng, R. (2025). A radical-based token representation method for enhancing Chinese pre-trained language models. Electronics, 14(5): 1031–1031. https://doi.org/10.3390/electronics14051031.

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P.S. (2025). A survey of multilingual large language models. Patterns, 6(1): 101118. https://doi.org/10.1016/j.patter.2024.101118.

Ray, P.P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. Internet of Things and Cyber-Physical Systems, 3(1): 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003.

Renaudo, E., Girard, B., Chatila, R., & Khamassi, M. (2015). Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture. Procedia Computer Science, 71: 178–184. https://doi.org/10.1016/j.procs.2015.12.194.

Rizki, A., Touil, A., Echchatbi, A., Oucheikh, R., & Ahlaqqach, M. (2025). A reinforcement learning-based proximal policy optimization approach to solve the economic dispatch problem. The 1st International Conference on Smart Management in Industrial and Logistics Engineering (SMILE 2025), Pages 24–24. https://doi.org/10.3390/engproc2025097024.

Roumeliotis, K.I., & Tselikas, N.D. (2023). ChatGPT and open-AI models: A preliminary review. Future Internet, 15(6): 192. https://www.mdpi.com/1999-5903/15/6/192/htm.

Ruan, L., & Jin, Q. (2022). Survey: Transformer-based video-language pre-training. AI Open, 3: 1–13. https://doi.org/10.1016/j.aiopen.2022.01.001.

Sajun, A.R., Zualkernan, I., & Sankalpa, D. (2024). A historical survey of advances in transformer architectures. Applied Sciences, 14(10): 4316–4316. https://doi.org/10.3390/app14104316.

Schmidt, B., & Hildebrandt, A. (2024). From GPUs to AI and quantum: Three waves of acceleration in bioinformatics. Drug Discovery Today, 29: 103990–103990. https://doi.org/10.1016/j.drudis.2024.103990.

Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models. Business & Information Systems Engineering, 66. https://doi.org/10.1007/s12599-024-00851-0.

Shahzad, T., Mazhar, T., Tariq, M.U., Ahmad, W., Khmaies Ouahada, & Habib, H. (2025). A comprehensive review of large language models: Issues and solutions in learning environments. Discover Sustainability, 6(1). https://doi.org/10.1007/s43621-025-00815-8.

Sharma, S., Mittal, P., Kumar, M., & Bhardwaj, V. (2025). The role of large language models in personalized learning: A systematic review of educational impact. Discover Sustainability, 6(1). https://doi.org/10.1007/s43621-025-01094-z.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q.M., Schaekermann, M., Wang, A., Dash, D., Chen, J.H., Shah, N.H., Lachgar, S., Mansfield, P.A., & Prakash, S. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, 31(3). https://doi.org/10.1038/s41591-024-03423-7.

Smirnova, A.Y. (2016). “Where is the bank?” or how to “find” different senses of a word. Heliyon, 2(6): e00065. https://doi.org/10.1016/j.heliyon.2016.e00065.

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence, Pages 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123.

Sufi, F. (2024). Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation. Information, 15(2): 99. https://doi.org/10.3390/info15020099.

The Authors Guild (2023). You just found out your book was used to train AI. Now what? - The Authors Guild. The Authors Guild. https://authorsguild.org/news/you-just-found-out-your-book-was-used-to-train-ai-now-what.

Williamson, S.M., & Prybutok, V. (2024). The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. Information, 15(6): 299. https://doi.org/10.3390/info15060299.

Wolfe, C.R. (2023). Understanding and using supervised fine-tuning (SFT) for language models. Deep (Learning) Focus. https://cameronrwolfe.substack.com/p/understanding-and-using-supervised.

Wright, T., Salyers, A., Howell, K., Harrison, J., Silvasstar, J., & Bull, S. (2025). A pilot study of an AI chatbot for the screening of substance use disorder in a healthcare setting. AI, 6(6): 113. https://doi.org/10.3390/ai6060113.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 4(2): 100211. https://doi.org/10.1016/j.hcc.2024.100211.

Yoosefzadeh-Najafabadi, M. (2025). From text to traits: exploring the role of large language models in plant breeding. Frontiers in Plant Science, 16. https://doi.org/10.3389/fpls.2025.1583344.

Zhang, E.Y., Cheok, A.D., Pan, Z., Cai, J., & Yan, Y. (2023). From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. Sci, 5(4): 46. https://doi.org/10.3390/sci5040046.

Zheng, Z., Wang, Y., Huang, Y., Song, S., Yang, M., Tang, B., Xiong, F., & Li, Z. (2025). Attention heads of large language models. Patterns, 6(2): 101176. https://doi.org/10.1016/j.patter.2025.101176.