FUNDAMENTALS OF GENERATIVE AI AND LARGE LANGUAGE MODELS (LLMS): ARCHITECTURE, APPLICATION, AND THE PATH FORWARD

A Comprehensive Analysis of Technical Principles, Real-World Impact, and Future Trajectories

Author: Domain Specialist Team

Date: October 26, 2024

Table of Contents

1.0 Introduction

1.1 Definition and Context of Generative AI

Generative artificial intelligence (AI) represents a transformative subfield of machine learning that focuses on the creation of new, original content. Unlike conventional AI systems that analyze, classify, or curate existing data, generative models are trained on vast pools of information to learn complex patterns and structures, enabling them to produce novel outputs such as text, images, video, and audio in response to a prompt.[1] This fundamental capability to create, rather than merely organize, distinguishes it from its predecessors, making it a powerful tool for a diverse range of applications, from creative arts to complex scientific research.[2]

The distinction between generative and traditional AI is foundational to understanding the technology's unique impact. Discriminative models, for instance, are designed to classify and analyze existing data by learning a decision boundary that separates different classes.[4] A discriminative model might be trained to determine whether an email is spam or not, or to identify a specific person's face. In contrast, generative AI, as exemplified by tools like ChatGPT or Midjourney, operates as a "creative and imaginary artist" that synthesizes new content by predicting what element—be it a word, a sound, or a pixel—would come next in a pattern.[2] This report delves into the foundational principles that enable this remarkable capability, examining the architecture, training methodologies, and a broad spectrum of real-world applications.

1.2 Scope and Purpose of the Report

The purpose of this report is to provide a comprehensive, expert-level analysis of the

fundamentals of generative AI and Large Language Models (LLMs). It is structured to guide the reader from the core theoretical concepts to the practical, real-world implications of the technology. The report begins by establishing the fundamental difference between generative and discriminative AI, which provides a critical framework for all subsequent analysis. It then explores the intricate architecture and multi-stage training process of LLMs, detailing how these models are built and refined.

Further sections provide a detailed overview of the technology's applications across various industries, from finance and healthcare to creative arts, highlighting both the successes and the significant challenges of large-scale adoption. The report concludes with an in-depth discussion of the inherent limitations, including ethical issues such as bias and hallucinations, before exploring future opportunities and the trajectory of model scaling and specialization. By synthesizing this information, the report aims to offer a nuanced understanding of generative AI's current state and its potential path forward, serving as a valuable reference for strategic decision-making and academic inquiry.

2.0 Core Concepts of Generative AI

2.1 Generative vs. Discriminative Models: A Fundamental Distinction

At the heart of the generative AI revolution lies a fundamental architectural difference from previous AI paradigms. Machine learning models can be broadly categorized into two types: generative and discriminative.[4] This distinction is based on their primary objective and the way they learn from data.

Discriminative models are designed to classify or predict outcomes based on existing data. They operate by learning the conditional probability distribution, denoted as $p(Y | X)$, which tells the model the probability of a label Y given an input X. A discriminative model's goal is not to understand the full data distribution but rather to find an optimal "decision boundary" that separates different classes.[5] This makes them highly effective for tasks such as facial recognition, spam filtering, and creditworthiness checks, where the objective is to differentiate between categories.[4] The efficiency and accuracy of these models stem from their more focused learning objective.

In contrast, generative models take on a more complex challenge: they aim to learn the underlying probability distribution of the data itself, often represented as $p(X, Y)$ or simply $p(X)$

if no labels are present.[5] By capturing the entire data distribution, generative models are able to create new data instances that closely resemble the original dataset.[4] This is why they are considered "creative" models, as their purpose is to produce novel output, not just to classify existing input.[5] This more difficult task of modeling the entire distribution is the reason generative models often require vast amounts of data and computational resources to train effectively.[6] The table below provides a concise summary of these core differences.

Feature	Generative Models	Discriminative Models
Objective	Create new data	Classify existing data
Probability Model	Learns joint probability p(X, Y) or p(X)	Learns conditional probability p(Y | X)
Core Function	Creation, Synthesis, Augmentation	Classification, Prediction, Analysis
Learning Paradigm	Often unsupervised	Primarily supervised
Typical Applications	Text/Image Generation, Drug Discovery, Data Augmentation	Spam Filtering, Facial Recognition, Sentiment Analysis

2.2 Key Types of Generative Models

Within the generative AI paradigm, several architectural types have emerged, each with unique strengths and applications.

- **Generative Adversarial Networks (GANs):** Proposed by Ian Goodfellow in 2014, GANs consist of two competing neural networks: a generator and a discriminator.[9] The generator's role is to create new, synthetic data samples from random noise, while the discriminator's role is to distinguish between the real data and the fake data produced by the generator.[9] This adversarial game drives both networks to improve over time, with the generator becoming increasingly adept at creating realistic data and the discriminator becoming more skilled at detecting fakes.[11] GANs have been instrumental in applications such as generating photorealistic human faces, creating cartoon characters, and performing image-to-image translation.[13]

- **Variational Autoencoders (VAEs):** VAEs are probabilistic generative models that use an encoder-decoder structure.[9] The encoder learns to map input data to a lower-dimensional "latent space" representation, while the decoder learns to reconstruct the data from this latent space.[7] Unlike regular autoencoders, VAEs encode the input as a probability distribution, which allows for smooth transitions in the latent space.[10] This design is particularly well-suited for creative tasks that require reconstruction and variation, such as generating new designs or augmenting existing datasets.[13] VAEs are generally considered easier to train than GANs because they lack the adversarial competition.[10]
- **Transformer Models:** While GANs and VAEs are predominantly associated with image generation, transformer models have revolutionized the field of Natural Language Processing (NLP).[10] The transformer architecture, first described in the 2017 paper "Attention is All You Need," is uniquely designed to process sequential data in parallel, a significant departure from earlier recurrent neural networks (RNNs).[14] This parallel processing capability allows transformers to scale to an unprecedented number of parameters, making them the foundational architecture for nearly all modern Large Language Models (LLMs).[10]

3.0 The Inner Workings of Large Language Models (LLMs)

3.1 The Transformer Architecture: Attention Is All You Need

The rapid advancement of LLMs is largely a function of the transformer architecture, a type of neural network that processes sequential data by learning context and tracking relationships between components.[17] The central innovation of this architecture is its self-attention mechanism, which enables the model to weigh the importance of different parts of a sequence all at once, rather than processing them one token at a time.[15]

The model's operation begins with the input sequence, such as a sentence. First, the input is broken down into a series of "tokens" or individual components. Each token is then converted into a numerical vector through a process called **input embedding**.[17] These vectors capture the semantic and syntactic information of the tokens, representing them as a series of coordinates in an n-dimensional space.[17] Since the self-attention mechanism does not

inherently preserve the order of the tokens, a crucial step known as

positional encoding is applied. This adds information to each token's embedding to indicate its position in the original sequence, ensuring the model can understand the context and order of the words.[17]

The core of the transformer is its self-attention mechanism, which determines how much a given token should "pay attention" to other tokens in the sequence.[14] This is achieved by creating three vectors for each token: a Query vector, a Key vector, and a Value vector.[14] The Query vector represents the information a token is "seeking," while the Key vectors represent the information other tokens contain. The model computes the relationships between each Query and all Key vectors, generating "attention weights" that determine the relative importance of each token to the others.[14] This process allows the model to selectively focus on the most relevant segments of the input text as it makes its predictions.[19]

The original transformer architecture is comprised of a stack of encoders and a stack of decoders.[18] The encoder stack processes the input sequence, generating contextualized embeddings for each token.[20] The decoder stack, in turn, takes these embeddings and generates the output sequence, often containing a masked multi-head attention layer that ensures its output depends only on the preceding tokens in the sequence.[20] This encoder-decoder structure allows the model to handle a wide range of tasks, from machine translation to text summarization.[20]

!(http://googleusercontent.com/deep_research_confirmation_content/transformer_architecture.jpg)
Figure 1: Schematic of the Transformer Architecture
A typical transformer architecture consists of a stack of identical encoder layers and a stack of identical decoder layers.

The **Encoder** block takes token embeddings and positional encodings as input. It is composed of a **Multi-Head Self-Attention** layer and a **Feed-Forward Neural Network** layer. The self-attention layer allows the model to process all tokens in the input sequence in parallel, creating relationships between each token and all others. The output of this layer is normalized and passed to the feed-forward network for further processing.

The **Decoder** block is structurally similar but includes an additional layer. It takes the output from the encoder stack as input. The decoder stack consists of a **Masked Multi-Head Self-Attention** layer, an **Encoder-Decoder Attention** layer, and a **Feed-Forward Neural Network** layer. The masked self-attention layer ensures that when generating a new token, the model only considers the tokens that have already been generated, which is crucial for text generation. The encoder-decoder attention layer then helps the decoder focus on relevant parts of the input sequence from the encoder's output, enabling the model to align the output with the source. The output from the final decoder layer is then processed to

predict the next token in the sequence.

3.2 The Multi-Stage Training Process: From Pre-training to Alignment

Training a large language model is a sophisticated, multi-stage process that transforms raw data into a powerful and aligned tool.[21] This process begins with a phase that requires immense computational resources and progresses through targeted refinement to align the model with human preferences.

The first stage is **pre-training**, an unsupervised learning phase where the model is exposed to a massive, diverse corpus of text, such as books, articles, and websites.[21] The sheer size of these datasets, often described as "unfathomable" and "practically infeasible" to manually quality-check, allows the model to learn general language patterns, grammar, and a wide variety of facts.[22] In this stage, the model trains itself by predicting missing or next words in a sentence, iteratively adjusting its parameters until it correctly predicts the next token.[15] This process builds a general-purpose model with a broad understanding of language, but it comes with a significant limitation: without rigorous oversight, the model can inadvertently learn and mimic biases or toxic language present in the unfiltered data.[19]

The challenges of pre-training necessitate the second stage, **fine-tuning**, which tailors the pre-trained model to specific tasks or domains.[21] This stage uses a curated, labeled dataset that is significantly smaller and more focused than the pre-training data.[21] For example, a model might be fine-tuned on a dataset of question-answer pairs or annotated medical records to make it more accurate and relevant for a niche application. This process allows the model to grasp the unique terminology and context of a specific field, leading to more accurate and relevant outputs.[21] However, this step alone is often insufficient to address all the issues inherited from pre-training.

This leads to the final stage of the process: **Reinforcement Learning with Human Feedback (RLHF)**.[25] This critical phase is designed to align the model's behavior with human expectations for safety, truthfulness, and helpfulness.[21] Instead of providing the model with exact outputs, human labelers rank the model's generated responses based on a set of criteria.[21] This human-generated grading is used to train a separate "reward model," which then guides the LLM to favor desired behaviors and discourage unwanted, harmful, or toxic outputs.[21] The multi-stage progression from broad, unsupervised learning to targeted, human-in-the-loop alignment demonstrates a strategic evolution in LLM development that prioritizes value-based correction over pure data-driven training.

3.3 GPT vs. BERT: A Comparison of Autoregressive and Bidirectional Models

While many modern LLMs are based on the transformer architecture, they can be designed in fundamentally different ways that optimize them for distinct tasks. The comparison between GPT and BERT provides a clear illustration of these architectural differences.

GPT (Generative Pre-trained Transformer) models, such as those that power ChatGPT, are a family of **decoder-only** transformer models that are autoregressive in nature.[26] This means they generate text sequentially, predicting the next word based only on the preceding words in the sequence.[26] This

unidirectional attention is highly effective for tasks that require generating coherent and contextually relevant text, such as content creation, code writing, and dialogue generation.[25] The design of GPT models makes them excel at producing human-like text step-by-step, making them valuable for applications where the generation of new, original content is the primary goal.[26]

BERT (Bidirectional Encoder Representations from Transformers), in contrast, is an **encoder-only** transformer model.[26] Its key feature is its

bidirectional attention, which allows it to process and understand the context of a word by looking at all other words in the input sentence simultaneously, both before and after the word in question.[26] This full-context understanding makes BERT superior for language comprehension tasks, where the objective is to extract meaning and classify information.[26] BERT is therefore an ideal architecture for tasks like sentiment analysis, question answering, and text classification, which rely on a deep understanding of the entire sentence's context.[26]

Feature	BERT	GPT
Architecture Type	Encoder-only Transformer	Decoder-only Transformer
Attention Type	Multi-head Attention	Masked Multi-head Attention
Context Handling	Considers both left and right context	Considers only left context

	simultaneously	
Primary Purpose	Understanding and extracting meaning from text	Generating coherent and contextually relevant text
Training Objective	Masked Language Modeling (predicts masked words using full context)	Causal Language Modeling (predicts the next word based on past words)
Typical Output	Classifications, embeddings, extracted answers	Generated sentences, paragraphs, or code

4.0 Applications and Transformative Impact across Industries

Generative AI is not merely a technical curiosity; it is a disruptive force with a broad range of applications that are reshaping industries and redefining workflows. From automating routine tasks to accelerating creative and scientific discovery, the technology is demonstrating its potential to deliver significant value.

10.1 Enterprise and Individual Productivity

At the individual and startup levels, generative AI has been shown to provide tangible productivity gains. The technology excels at automating repetitive and time-consuming tasks, such as drafting marketing copy, summarizing complex documents, or generating code snippets.[15] This allows human workers to free up time for higher-level conceptualization, strategy, and creative problem-solving.[3] Startups, in particular, have found success by focusing on a single, narrow problem, such as automating ad copywriting, and achieving rapid, measurable results.[31]

However, the widespread adoption of generative AI in large enterprises presents a more

complex picture. A recent MIT study found that 95% of generative AI business projects are failing to deliver meaningful results.[31] This phenomenon, dubbed the "GenAI Divide," highlights a significant gap between public hype and enterprise reality. The study attributes this high failure rate to a "learning gap" and a reliance on "generic large language models that are ill-suited to niche requirements".[31] This suggests that the problem is not a lack of technological capability but rather a failure of strategic implementation, where organizations attempt to apply a one-size-fits-all solution to complex, domain-specific workflows.

4.2 Specialization in High-Stakes Domains: Finance and Healthcare

The failure of generic models in complex enterprise environments underscores a growing trend toward **specialization and customization**.[32] The most successful applications of generative AI are increasingly found in high-stakes domains like finance and healthcare, where models are fine-tuned with proprietary data to meet specific needs.

In the **financial sector**, generative AI is already delivering value by automating core processes and enhancing decision-making.[33] Applications include:

- **Financial Reporting:** Automating the generation of accurate and comprehensive reports by analyzing historical financial data.[28]
- **Risk Assessment:** Detecting fraudulent activities and monitoring compliance with greater efficiency and accuracy than human analysts.[28]
- **Predictive Modeling:** Generating forecasts and predictive models based on historical financials and external data, enabling financial professionals to make more informed investment decisions.[28]

Similarly, in **healthcare and life sciences**, generative AI is accelerating scientific discovery and improving patient care. Use cases include:

- **Drug Discovery:** Rapidly screening millions of potential drug candidates and designing new molecules with desired properties.[13]
- **Clinical Task Automation:** Automating administrative tasks for clinicians, such as generating referral letters and summarizing patient histories, which reduces administrative burden and allows for more focus on patient care.[34]
- **Medical Imaging and Pathology Analysis:** Improving image quality and detecting anomalies and patterns in medical scans, providing decision support for diagnoses.[34]

4.3 Applications in Creative Arts and Education

Beyond high-stakes industries, generative AI is profoundly impacting the creative arts and education. In creative fields, the technology serves as a "springboard for ideas".[3] It can automate the more mundane aspects of the creative process, such as desk research for a journalist or early-stage concept generation for a product designer, allowing professionals to dedicate more time to higher-level conceptualization.[3] This augmentation enhances, rather than replaces, the creative process, accelerating production and enabling artists and designers to quickly iterate through new concepts.[3]

In **education**, generative AI tools break down the creative process into manageable steps, enabling students to move from an initial idea to a finished product with greater ease.[30] The technology allows for more self-expression by generating personalized outputs based on individual choices and prompts. This capability allows educators to assign creative projects more frequently, as the time required for students to brainstorm and iterate is significantly reduced.[30] By automating repetitive tasks, generative AI enables students to focus on practicing essential creative thinking skills like problem-solving, innovative thinking, and synthesizing ideas.[30]

5.0 Benefits, Limitations, and Ethical Considerations

The transformative potential of generative AI is balanced by a series of inherent limitations and complex ethical challenges that must be addressed for responsible development and deployment.

5.1 Key Benefits of Generative AI

The primary benefits of generative AI applications are centered around three key areas:

- **Creativity and Innovation:** The technology serves as a powerful accelerator for creative work.[3] By generating new ideas, designs, and content from text prompts, it provides a "springboard" for human creativity, allowing artists and designers to rapidly test new concepts and streamline their workflows.[3]
- **Productivity Gains:** Generative AI automates repetitive tasks across a wide range of industries, from content creation for marketing teams to code generation for developers.[15] This automation enhances efficiency and frees up human capital to focus

on strategic, analytical, and creative endeavors.[28]
- **Data Augmentation:** Generative models can produce synthetic data that mimics real-world datasets.[4] This is particularly useful for training other AI models when real data is scarce, expensive to acquire, or sensitive, as it helps improve model robustness and performance without compromising privacy.[4]

5.2 Hallucinations and Misaligned Behavior

One of the most significant limitations of LLMs is their propensity for **hallucinations**, a phenomenon where the model produces false, nonsensical, or misleading information and presents it as factual.[36] This is not a random bug but a direct outcome of the model's core training objective: to probabilistically predict the most likely next token in a sequence.[36] Because the model is a sophisticated pattern-matcher rather than a factual knowledge engine, it will generate a "convincingly sounding fake information" when it lacks the necessary grounding.[37] The root cause often stems from a lack of domain-specific training data or a poor attention performance that fails to retrieve relevant information from its vast knowledge base.[36]

To mitigate hallucinations, several strategies have been developed:

- **Retrieval-Augmented Generation (RAG):** This technique addresses hallucinations by ensuring factual accuracy.[36] RAG systems first search a private, verified knowledge base for relevant information before the model generates a response.[36] This process grounds the output in a source of truth, significantly reducing the likelihood of fabricated information.[36]
- **Advanced Prompting Techniques:** Methods like Chain-of-Thought prompting, which breaks down complex tasks into intermediate reasoning steps, can also help to reduce hallucinations and improve the model's overall reasoning capabilities.[36]

5.3 Ethical Issues: Bias, Privacy, and Data Security

The development and deployment of generative AI present a number of serious ethical concerns, which are largely a consequence of the models' dependence on massive, unfiltered datasets and their opaque inner workings.

Algorithmic Bias is a pervasive issue. LLMs inherit systematic errors and prejudices from

their training data, which often contains imbalanced representations of different demographic groups.[24] This can lead to outputs that reinforce stereotypes and produce biased predictions and recommendations.[24] The bias can be subtle and unintentional, arising not only from the data but also from the inherent design choices of the model architecture and the human developers who create and fine-tune them.[24]

Privacy and Data Risks are also significant. Because LLMs are trained on enormous datasets, they can "memorize" sensitive personal information and inadvertently leak it in their outputs.[22] Even when data is anonymized, advanced models can re-identify personal information through "linkage attacks" that correlate demographic attributes to a high degree of certainty.[39] In high-stakes fields like medical education, where sensitive patient data is used to train models for personalized feedback, this poses a tremendous risk of information exposure.[39]

These ethical issues highlight the critical need for a multifaceted approach to responsible AI development, combining technical solutions like debiasing algorithms with transparent data collection, robust security practices, and a commitment to ongoing ethical auditing.[24]

6.0 Future Outlook: Opportunities and Challenges

The trajectory of generative AI is not a simple linear progression but a complex interplay of scaling, specialization, and sustainability. The industry is currently navigating several key trends that will define its future.

6.1 Emerging Trends: Efficiency, Specialization, and Multimodality

The high computational and energy costs associated with training massive LLMs have spurred a push for **efficiency and sustainability**, often referred to as "Green AI".[32] While the size of frontier models continues to grow exponentially, there is a parallel, deliberate trend toward creating smaller, more efficient LLMs that can deliver comparable performance at a fraction of the cost.[32] This dual-pronged approach reflects a maturing market where models are optimized not just for performance, but also for accessibility, cost-effectiveness, and environmental impact.

This move toward efficiency is closely linked to the trend of **specialization and customization**. The future of successful enterprise AI adoption lies not in generic models, but

in verticalized, domain-specific solutions that are fine-tuned with proprietary data to improve accuracy and compliance for specific tasks.[32] The failure of many large-scale generic AI projects is directly driving this shift toward purpose-built models tailored for niche requirements, such as fraud detection in finance or therapeutic target identification in healthcare.[32]

Finally, a major trend in model evolution is **multimodality**, which goes beyond the traditional text-only capabilities of early LLMs.[2] Modern generative AI tools are increasingly being trained to process and generate various data types, including text, images, and audio, allowing for new forms of interaction and application.[2]

6.2 The Trajectory of Model Scaling and Performance

The development of LLMs is characterized by a rapid and exponential growth in scale, measured by the number of parameters. This trend is not arbitrary but is governed by **neural scaling laws**, which describe a predictable relationship between neural network performance changes and key factors such as the number of parameters, training dataset size, and training cost.[40] These empirical laws allow developers to optimize resources to achieve a target performance, guiding the training of all frontier models today.[41]

The following description illustrates the exponential growth in the number of parameters of notable LLMs since 2018, which is a key indicator of the field's rapid advancement.

!(http://googleusercontent.com/deep_research_confirmation_content/llm_growth_trends.jpg)
Figure 2: LLM Parameter Scaling Trends (2018–2025)
This visual representation shows the exponential growth of LLM parameters on a logarithmic scale. The horizontal axis represents the year from 2018 to 2025, while the vertical axis represents the number of parameters in billions, plotted logarithmically.

- In **2018**, the foundational GPT-1 model had 117 million parameters.[16] This point is plotted near the base of the graph.
- By **2019**, the model's successor, GPT-2, demonstrated a significant leap to 1.5 billion parameters.[16]
- A dramatic increase occurred in **2020** with the release of GPT-3, which scaled to 175 billion parameters.[19]
- In **2021**, models like the Megatron-Turing NLG pushed the boundary to 530 billion parameters.[16]
- By **2022**, Google's PaLM model surpassed this, reaching 540 billion parameters.[16]
- As of **2025**, the DeepSeek-R1 model has reached 671 billion parameters, demonstrating

continued growth at the high end of the scale.[32]

The graph clearly depicts a steep, upward curve, illustrating that the number of parameters in state-of-the-art LLMs has grown by several orders of magnitude in just a few years. This exponential trend underscores the relentless pursuit of scale as a means to improve model performance.

6.3 The Human-in-the-Loop: Navigating the Future of Work

The future of generative AI is not about the wholesale replacement of human workers but about the **augmentation of human capabilities**.[30] The technology is best used to automate mundane, repetitive tasks, allowing humans to focus on uniquely human traits such as creativity, emotional intelligence, and strategic thinking.[30] While there was an initial rush by some companies to cut jobs with the expectation that AI could replace them, several have since had to quietly rehire staff, recognizing that human workers provide unique value that generative AI cannot replicate.[31] This experience underscores a crucial understanding: for all its power, generative AI is a tool, and its most impactful role is to serve as a collaborator that enhances, rather than replaces, human expertise.

7.0 Conclusion

Generative artificial intelligence, particularly in the form of Large Language Models, represents a profound and ongoing evolution in the field of machine learning. This report has detailed the fundamental distinction between generative models that create new content and discriminative models that classify existing data. The emergence of the transformer architecture, with its parallel processing and self-attention mechanisms, has been the key enabler for the exponential scaling of LLMs, from foundational models like GPT-1 to the multi-hundred-billion-parameter models of today.

The analysis highlights that while the applications of generative AI are vast and transformative—revolutionizing fields from finance to the creative arts—the path to successful implementation is fraught with challenges. The high failure rate of large-scale enterprise projects demonstrates that a simple application of generic models is often insufficient. The future of the technology is therefore trending toward greater specialization and customization, where domain-specific LLMs, fine-tuned with proprietary data, are developed

to meet the complex needs of high-stakes industries.

Furthermore, the report has addressed the critical limitations and ethical issues inherent in LLMs, particularly the phenomena of hallucinations and the challenges of inherited bias and privacy risks. These issues are not mere bugs but a consequence of the models' core probabilistic design and dependence on unfathomable datasets. The development of mitigation strategies, such as Retrieval-Augmented Generation (RAG) and human-in-the-loop alignment, signals a strategic pivot in the field toward a greater emphasis on factual accuracy, safety, and responsible deployment.

In summary, the trajectory of generative AI is moving toward a more mature state, defined by a dual focus on both scaling and specialization. The most impactful future applications will likely be those that effectively marry powerful generative capabilities with a foundation of factual accuracy and a strong ethical framework. The human role will continue to be central, as the most effective use of this technology will be in augmenting human capabilities and empowering human creativity, rather than seeking to replace it.

References

2

https://teaching.pitt.edu/resources/what-is-generative-ai/

31

https://timesofindia.indiatimes.com/technology/tech-news/mit-study-finds-95-of-generative-ai-projects-are-failing-only-hype-little-transformation/articleshow/123453071.cms

32

https://prajnaaiwisdom.medium.com/llm-trends-2025-a-deep-dive-into-the-future-of-large-language-models-bff23aa7cdbc

34

https://aws.amazon.com/health/gen-ai/

21

https://itrexgroup.com/blog/llm-training/

13

https://www.geeksforgeeks.org/artificial-intelligence/exploring-generative-models-applications-examples-and-key-concepts/

24

https://academy.test.io/en/articles/9227500-llm-bias-understanding-mitigating-and-testing-the-bias-in-large-language-models

29

https://www.upwork.com/resources/generative-ai-benefits

23

https://www.labellerr.com/blog/challenges-in-development-of-llms/

5

https://olibr.com/blog/generative-ai-vs-discriminative-ai-whats-the-key-difference/

3

https://www.chead.ac.uk/wp-content/uploads/2024/03/ImpactofAIonCreativeEducation_ProfessorMelanieGray.pdf

22

https://medium.com/@arghya05/understanding-the-challenges-of-large-language-models-llms-and-their-solutions-arghya-mukherjee-5e154b93cca4
19

15

https://aws.amazon.com/what-is/large-language-model/

25

https://aws.amazon.com/what-is/gpt/

28

https://www.alpha-sense.com/blog/trends/generative-ai-in-financial-services/

30

https://www.edweek.org/sponsor/adobe-corporation/the-top-5-ways-generative-ai-increases-student-creativity

36

https://www.redhat.com/en/blog/when-llms-day-dream-hallucinations-how-prevent-them

35

https://pmc.ncbi.nlm.nih.gov/articles/PMC11739231/

26

https://www.geeksforgeeks.org/nlp/gpt-vs-bert/

39

https://pmc.ncbi.nlm.nih.gov/articles/PMC11327620/

14

https://www.ibm.com/think/topics/transformer-model

4

https://business.canon.com.au/insights/what-is-the-difference-between-generative-ai-and-discriminative-ai

27

https://www.coursera.org/articles/bert-vs-gpt
38

33

https://www.cbh.com/insights/articles/generative-ai-in-finance-key-use-cases-today/

1

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

37

https://neptune.ai/blog/llm-hallucinations

17

https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/

20

https://www.ibm.com/think/topics/encoder-decoder-model

18

https://jalammar.github.io/illustrated-transformer/

16

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

42

https://www.openxcell.com/blog/llm-parameters/

8

https://medium.com/@kanerika/generative-vs-discriminative-understanding-machine-learning-models-87e3d2b3b99f

9

https://www.geeksforgeeks.org/deep-learning/generative-models-in-ai-a-comprehensive-comparison-of-gans-and-vaes/

10

https://hyqoo.com/artificial-intelligence/comparing-generative-ai-models-gans-vaes-and-transformers

11

https://developers.google.com/machine-learning/gan/gan_structure

12

https://aws.amazon.com/what-is/gan/

40

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

41

https://www.glennklockwood.com/garden/scaling-laws

6

https://developers.google.com/machine-learning/gan/generative

7

https://learnopencv.com/generative-and-discriminative-models/

16

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

Works cited

1. en.wikipedia.org, accessed on September 1, 2025, https://en.wikipedia.org/wiki/Generative_artificial_intelligence
2. What is Generative AI? - University Center for Teaching and Learning, accessed on September 1, 2025, https://teaching.pitt.edu/resources/what-is-generative-ai/
3. Generative AI and its impact on Creative Education: Thought Piece - CHEAD, accessed on September 1, 2025, https://www.chead.ac.uk/wp-content/uploads/2024/03/ImpactofAIonCreativeEducation_ProfessorMelanieGray.pdf
4. What are the differences between generative AI and discriminative AI, accessed on September 1, 2025, https://business.canon.com.au/insights/what-is-the-difference-between-generative-ai-and-discriminative-ai
5. Generative AI vs. Discriminative AI: What's the Key Difference? - Olibr, accessed on September 1, 2025, https://olibr.com/blog/generative-ai-vs-discriminative-ai-whats-the-key-difference/
6. Background: What is a Generative Model? | Machine Learning ..., accessed on September 1, 2025, https://developers.google.com/machine-learning/gan/generative

7. Generative and Discriminative Models - LearnOpenCV, accessed on September 1, 2025, https://learnopencv.com/generative-and-discriminative-models/
8. Generative Vs Discriminative: Understanding Machine Learning ..., accessed on September 1, 2025, https://medium.com/@kanerika/generative-vs-discriminative-understanding-machine-learning-models-87e3d2b3b99f
9. Generative Models in AI: A Comprehensive Comparison of GANs and VAEs, accessed on September 1, 2025, https://www.geeksforgeeks.org/deep-learning/generative-models-in-ai-a-comprehensive-comparison-of-gans-and-vaes/
10. Comparing Generative AI Models: GANs, VAEs, and Transformers - Hyqoo, accessed on September 1, 2025, https://hyqoo.com/artificial-intelligence/comparing-generative-ai-models-gans-vaes-and-transformers
11. Overview of GAN Structure | Machine Learning | Google for ..., accessed on September 1, 2025, https://developers.google.com/machine-learning/gan/gan_structure
12. What is a GAN? - Generative Adversarial Networks Explained - AWS, accessed on September 1, 2025, https://aws.amazon.com/what-is/gan/
13. Exploring Generative Models: Applications, Examples, and Key Concepts - GeeksforGeeks, accessed on September 1, 2025, https://www.geeksforgeeks.org/artificial-intelligence/exploring-generative-models-applications-examples-and-key-concepts/
14. What is a Transformer Model? - IBM, accessed on September 1, 2025, https://www.ibm.com/think/topics/transformer-model
15. What is LLM? - Large Language Models Explained - AWS - Updated 2025, accessed on September 1, 2025, https://aws.amazon.com/what-is/large-language-model/
16. Large language model - Wikipedia, accessed on September 1, 2025, https://en.wikipedia.org/wiki/Large_language_model
17. What are Transformers? - Transformers in Artificial Intelligence Explained - AWS, accessed on September 1, 2025, https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/
18. The Illustrated Transformer – Jay Alammar – Visualizing machine ..., accessed on September 1, 2025, https://jalammar.github.io/illustrated-transformer/
19. GPT-3 - Wikipedia, accessed on September 1, 2025, https://en.wikipedia.org/wiki/GPT-3
20. What is an encoder-decoder model? - IBM, accessed on September 1, 2025, https://www.ibm.com/think/topics/encoder-decoder-model
21. LLM Training: The Process, Stages, and Fine-Tuning Gritty Details - ITRex Group, accessed on September 1, 2025, https://itrexgroup.com/blog/llm-training/
22. Understanding the Challenges of Large Language Models (LLMs) and Their Solutions-Arghya Mukherjee - Medium, accessed on September 1, 2025, https://medium.com/@arghya05/understanding-the-challenges-of-large-language-models-llms-and-their-solutions-arghya-mukherjee-5e154b93cca4

23. 8 Challenges Of Building Your Own Large Language Model - Labellerr, accessed on September 1, 2025, https://www.labellerr.com/blog/challenges-in-development-of-llms/
24. LLM Bias: Understanding, Mitigating and Testing the Bias in Large Language Models, accessed on September 1, 2025, https://academy.test.io/en/articles/9227500-llm-bias-understanding-mitigating-and-testing-the-bias-in-large-language-models
25. What is GPT AI? - Generative Pre-Trained Transformers Explained - AWS - Updated 2025, accessed on September 1, 2025, https://aws.amazon.com/what-is/gpt/
26. GPT vs BERT - GeeksforGeeks, accessed on September 1, 2025, https://www.geeksforgeeks.org/nlp/gpt-vs-bert/
27. BERT vs. GPT: What's the Difference? - Coursera, accessed on September 1, 2025, https://www.coursera.org/articles/bert-vs-gpt
28. Generative AI in Financial Services: Use Cases, Benefits, and Risks - AlphaSense, accessed on September 1, 2025, https://www.alpha-sense.com/blog/trends/generative-ai-in-financial-services/
29. 10 Benefits of Generative AI for Professionals - Upwork, accessed on September 1, 2025, https://www.upwork.com/resources/generative-ai-benefits
30. The Top 5 Ways Generative AI Increases Student Creativity - SPONSOR CONTENT, accessed on September 1, 2025, https://www.edweek.org/sponsor/adobe-corporation/the-top-5-ways-generative-ai-increases-student-creativity
31. MIT study finds 95% of generative AI projects are failing: Only hype, little transformation, accessed on September 1, 2025, https://timesofindia.indiatimes.com/technology/tech-news/mit-study-finds-95-of-generative-ai-projects-are-failing-only-hype-little-transformation/articleshow/123453071.cms
32. LLM Trends 2025: A Deep Dive into the Future of Large Language Models | by PrajnaAI, accessed on September 1, 2025, https://prajnaaiwisdom.medium.com/llm-trends-2025-a-deep-dive-into-the-future-of-large-language-models-bff23aa7cdbc
33. Generative AI in Finance: Key Use Cases Today - Cherry Bekaert, accessed on September 1, 2025, https://www.cbh.com/insights/articles/generative-ai-in-finance-key-use-cases-today/
34. Generative AI in Healthcare & Life Sciences - AWS, accessed on September 1, 2025, https://aws.amazon.com/health/gen-ai/
35. Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency - PMC, accessed on September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11739231/
36. When LLMs day dream: Hallucinations and how to prevent them - Red Hat, accessed on September 1, 2025, https://www.redhat.com/en/blog/when-llms-day-dream-hallucinations-how-prevent-them

37. LLM Hallucinations 101: Why Do They Appear? Can We Avoid Them?, accessed on September 1, 2025, https://neptune.ai/blog/llm-hallucinations
38. Bias and Fairness in Large Language Models: A Survey - MIT Press Direct, accessed on September 1, 2025, https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A
39. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint - PMC, accessed on September 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11327620/
40. Neural scaling law - Wikipedia, accessed on September 1, 2025, https://en.wikipedia.org/wiki/Neural_scaling_law
41. Scaling laws - Glenn K. Lockwood, accessed on September 1, 2025, https://www.glennklockwood.com/garden/scaling-laws
42. LLM Parameters Explained: Powering Smarter AI Predictions - Openxcell, accessed on September 1, 2025, https://www.openxcell.com/blog/llm-parameters/