IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15541

Generative AI: A Comprehensive Survey on Transformer-Based Models

J Hemanth[1], Harsha B[2], Ashish Kumar[3], Anurag N[4], Dr. Muhibur Rahman T.R[5]

6th Sem B.E.(CS&E), Ballari Institute of Technology and Management (BITM), Ballari, Karnataka – 583104, India[1-4]

Associate Professor, Department of Computer Science and Engineering,

Ballari Institute of Technology and Management (BITM), Ballari, Karnataka – 583104, India[5]

Abstract: The landscape of artificial intelligence has undergone a profound transformation with the emergence of generative models capable of producing coherent text, realistic images, synthesized audio, and functional code. Central to this shift is the Transformer architecture, introduced by Vaswani et al. in their landmark 2017 contribution "Attention Is All You Need," which fundamentally redefined how sequential data is modeled by replacing recurrence with parallelizable, attention-based processing [1]. This survey provides a structured and thorough examination of Transformer-based generative AI systems, tracing their development from early recurrent sequence architectures through to the most advanced multimodal and agentic AI frameworks. A four-tier taxonomic model is introduced to systematically classify these systems: basic recurrent sequence models, Transformer-based NLP architectures, large-scale language models, and multimodal autonomous AI agents. Comparative analysis is conducted across these tiers, evaluating performance, scalability, and contextual modeling capability. Through review of more than a dozen pivotal studies—spanning bidirectional pretraining via BERT [2], autoregressive generation through the GPT series [3][4][5], and visual representation learning via Vision Transformers [6]—this paper maps out the major inflection points in the field's evolution. Critical open challenges are also examined, including the substantial computational overhead of large-scale training, the persistent problem of model hallucination, opacity in decision-making, systemic data biases, and the inability to adapt to real-time information. The paper concludes with a forward-looking perspective on next-generation generative AI, emphasizing efficiency, trustworthiness, and ethical design.

Keywords: Generative AI, Transformer Architecture, Self-Attention, BERT, GPT, Large Language Models, Vision Transformer, Multimodal AI, Natural Language Processing, Deep Learning.

I. INTRODUCTION

The ability of machines to generate new, meaningful content—rather than merely classify or predict—represents one of the most consequential leaps in the history of artificial intelligence. For much of its early life, machine learning was primarily a discriminative enterprise: recognizing objects in images, assigning sentiment labels to sentences, or translating text from one language to another. Generative AI inverts this objective entirely, asking not what category an input belongs to but what new output the model should produce that is coherent, contextually appropriate, and indistinguishable from human-created content. This inversion has unlocked a vast range of applications previously thought exclusive to human cognition, from creative writing and drug discovery to autonomous software engineering and open-ended dialogue [7].

Before the Transformer became dominant, sequence modeling tasks were handled by Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Gated Recurrent Units (GRUs). Despite widespread adoption, these architectures carried structural limitations that constrained performance. Their sequential computation graphs made parallelization difficult, greatly slowing training on long inputs. More fundamentally, modeling dependencies between tokens separated by many steps proved unreliable due to the vanishing gradient problem, wherein error signals diminish exponentially as they propagate backward through recurrent layers [8]. Attempts to address these issues through attention-augmented RNNs offered partial relief but did not resolve the underlying bottleneck.

The publication of "Attention Is All You Need" by Vaswani et al. in 2017 marked a decisive turning point [1]. By discarding recurrence entirely and relying instead on multi-head self-attention, the Transformer allowed every token in a sequence to directly attend to every other token simultaneously. This architectural decision eliminated the bottleneck of sequential processing, enabled the model to capture long-range dependencies more directly, and made efficient use of modern hardware accelerators. Within a few years, Transformer-based models had displaced RNNs as the architecture of choice across virtually every domain of sequence learning [9].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15541

The evolution of generative AI can be organized into three broad phases. The first, spanning the years before 2017, was dominated by RNN and LSTM models that established the feasibility of end-to-end sequence learning while struggling with scalability. The second phase, roughly 2017 to 2020, saw the rapid rise of Transformer-based NLP models—BERT, GPT-1, and GPT-2—demonstrating the extraordinary power of large-scale pretraining on unlabeled corpora. The third and ongoing phase, beginning around 2020, is defined by very large language models trained on unprecedented data volumes, alongside multimodal systems that simultaneously process and generate text, images, and audio [10].

The practical impact of these developments is difficult to overstate. Systems such as ChatGPT have made conversational AI accessible to mass audiences. Tools like GitHub Copilot have transformed software development by synthesizing functional code from natural language descriptions. Models such as DALL-E and Stable Diffusion generate photorealistic imagery from textual prompts. AI agent frameworks now enable systems to autonomously browse the web, execute multi-step plans, and reason over extended contexts—capabilities that would have seemed far-fetched a decade ago [11]. This paper surveys the landscape of Transformer-based generative AI, examines the theoretical underpinnings, proposes a four-tier taxonomy, reviews key literature, identifies research gaps, and outlines a roadmap for future progress.

II. THEORETICAL BACKGROUND

A. Transformer Architecture

The Transformer model, as originally proposed by Vaswani et al. [1], is built around two primary stacked components: an encoder and a decoder, each composed of identical, repeated layers. The encoder processes the input sequence and produces a rich continuous representation; the decoder uses this representation alongside its own masked self-attention to generate the output sequence one token at a time during inference. Every layer in both stacks contains a multi-head attention sublayer followed by a position-wise feed-forward network, with residual connections and layer normalization applied around each sublayer to support stable training of deep models [1].

This design allows the model to handle variable-length sequences and to learn hierarchical, context-sensitive representations without any sequential dependency between processing steps. Because all tokens within a layer are processed in parallel, training on modern GPU or TPU hardware is dramatically faster compared to recurrent architectures. The resulting throughput advantage has been a critical enabler of the large-scale pretraining paradigm that now defines the state of the art [9].

B. Self-Attention Mechanism

The self-attention mechanism is the operational core of the Transformer. For each input token, three distinct linear projections are computed: a Query (Q), a Key (K), and a Value (V) vector. Attention scores between any two tokens are obtained by taking the dot product of one token's Query with another's Key, normalized by the square root of the key dimensionality to keep the magnitude of the dot products stable during training. A softmax function converts these raw scores into a probability distribution, and the output is computed as the weighted sum of all Value vectors [1]. Formally:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) \cdot V$$

where Q, K, and V are matrices of query, key, and value projections respectively, and d_k denotes the dimensionality of the key vectors. In practice the Transformer employs h parallel attention heads, each learning to attend to different aspects of the input. Their outputs are concatenated and passed through a final linear projection to produce the multi-head attention output. This design lets the model simultaneously capture syntactic structure, semantic relationships, and long-distance coreference from distinct representational subspaces [1].

C. Positional Encoding

Because the Transformer processes all tokens simultaneously, it has no intrinsic awareness of token order. Vaswani et al. addressed this by adding sinusoidal positional encodings to each token embedding before it enters the encoder or decoder stack [1]. The encoding is defined as follows:

$$PE(pos, 2i) = sin(pos / 10000^{(2i / d_{model})})$$

$$PE(pos, 2i+1) = cos(pos / 10000^{(2i / d_{model})})$$

where pos is the token's position in the sequence, i is the dimension index, and d_model is the total embedding dimensionality. The sinusoidal functions allow the model to generalize to sequence lengths beyond those seen during training, as relative positional offsets can be represented as fixed linear combinations of these periodic signals. Learned

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15541

positional embeddings, used in BERT and later models, offer an alternative that allows the model to adapt positional representations from data [2].

D. Pretraining and Fine-Tuning

One of the most consequential conceptual contributions of the Transformer era is the two-stage training paradigm: large-scale unsupervised pretraining followed by task-specific supervised fine-tuning. During pretraining, a model learns general-purpose linguistic or visual representations from vast unlabeled corpora via self-supervised objectives. BERT uses masked language modeling—randomly masking a fraction of input tokens and training the model to predict them—combined with next-sentence prediction [2]. GPT-style models use causal (left-to-right) next-token prediction [3]. Both approaches allow the model to internalize grammar, factual knowledge, world-state reasoning, and semantic relationships at no labeling cost.

Fine-tuning then adapts these broad representations to a specific downstream task using a comparatively small labeled dataset, typically by adding a task-specific output head and continuing gradient updates for a short period. This approach has proven highly effective across question answering, sentiment classification, machine translation, code generation, and beyond, establishing pretraining as the universal starting point for modern AI systems [12].

E. Performance Metrics

Evaluating generative models requires metrics calibrated to the type of output being assessed. Perplexity, defined as the exponentiated average negative log-likelihood of a held-out test set, measures how confidently a language model predicts unseen text; lower values indicate better predictive performance. The BLEU score quantifies the precision of n-gram overlap between a generated translation and one or more reference translations, and remains the standard metric in machine translation benchmarks [13]. For classification tasks encountered in fine-tuning, accuracy and F1-score are common choices. More recently, GPT-based evaluation frameworks and human preference ratings have gained traction for assessing open-ended generation, as automated n-gram metrics frequently fail to capture semantic adequacy, factual accuracy, and stylistic fluency [14].

III. FOUR-TIER TAXONOMY OF GENERATIVE AI

To organize the broad and rapidly evolving landscape of generative AI, this paper proposes a four-tier taxonomic framework that classifies models according to their architectural sophistication, scale, and operational scope. This taxonomy is not merely a chronological ordering; it reflects qualitative leaps in representational capacity and design philosophy, each tier unlocking capabilities that were genuinely inaccessible at the tier below.

Tier 1 — Basic Sequence Models

The foundational tier encompasses recurrent architectures: vanilla RNNs, Long Short-Term Memory networks, and Gated Recurrent Units. These models process input sequences one step at a time, maintaining a hidden state that accumulates past information. While effective for relatively short sequences and well-understood from an optimization standpoint, they struggle with long-range dependencies and scale poorly because their sequential computation graph cannot be parallelized across time steps. Their primary applications included early neural machine translation, character-level language modeling, and speech recognition [8].

Tier 2 — Transformer-Based NLP Models

The second tier marks the decisive transition to attention-based architectures. Models such as BERT [2] and GPT-1 [3] demonstrated that pretraining a Transformer on large unlabeled corpora yields transferable representations of remarkable quality. BERT's bidirectional encoder captures context from both directions simultaneously, producing deep contextual embeddings that excel at understanding tasks. GPT-1's left-to-right autoregressive training instead prioritizes fluent generation. Both approaches set new state-of-the-art results across the dominant NLP benchmarks of their time and established pretraining as the standard paradigm for building language-capable systems [12].

Tier 3 — Large Language Models (LLMs)

The third tier comprises models trained at truly massive scale—tens to hundreds of billions of parameters on datasets spanning hundreds of billions of tokens. GPT-3 [4], with its 175 billion parameters, exhibited a remarkable emergent phenomenon: few-shot and even zero-shot generalization, wherein the model performs novel tasks simply by reading a brief natural language prompt, without any gradient-level adaptation. This capability, now termed in-context learning, shifted attention from fine-tuning to prompt engineering as the primary interface to model capability. Subsequent systems including InstructGPT [14] and GPT-4 [5] improved alignment with human intent through reinforcement learning from human feedback (RLHF), substantially reducing harmful and unhelpful outputs.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15541

Tier 4 — Multimodal and Autonomous AI Systems

The fourth and most advanced tier encompasses systems that operate across multiple data modalities and can autonomously plan and execute multi-step tasks. Models such as DALL-E [15], Flamingo [16], and GPT-4V process interleaved text and image inputs, generating outputs that are grounded in both visual and linguistic context. AI agent frameworks extend this further by equipping LLMs with external tools—web search, code interpreters, database access, and calendar management—enabling them to pursue long-horizon goals with minimal human intervention. This tier represents the current frontier of generative AI research and marks a qualitative transition from language models toward general-purpose cognitive systems.

IV. LITERATURE REVIEW

The following table presents a consolidated review of landmark publications that have collectively shaped the trajectory of Transformer-based generative AI. The selection spans foundational architecture proposals, domain-specific adaptations, alignment research, and large-scale multimodal systems, offering a representative cross-section of the field from 1997 to 2023.

TABLE I. SUMMARY OF KEY LITERATURE IN TRANSFORMER-BASED GENERATIVE AI

Sl.	Paper / Authors	Year	Method / Architecture	Key Findings
1	Vaswani et al. [1]	2017	Multi-Head Self-Attention (Transformer)	Introduced attention-only architecture; eliminated recurrence; enabled full parallel sequence processing and long-range dependency modeling.
2	Devlin et al. [2]	2018	Bidirectional Transformer Encoder (BERT)	Masked language modeling + next-sentence prediction; new SOTA on 11 NLP benchmarks including GLUE, SQuAD, and SWAG.
3	Radford et al. [3]	2018	Autoregressive Transformer (GPT-1)	Generative pretraining on unlabeled text followed by supervised fine-tuning; demonstrated strong transfer to diverse NLP tasks.
4	Radford et al. [9]	2019	Scaled Autoregressive Transformer (GPT-2)	Zero-shot text generation across tasks; 1.5B parameters; raised public discussion about potential for misuse of large language models.
5	Brown et al. [4]	2020	175B-Parameter LLM (GPT-3)	Introduced in-context few-shot learning without parameter updates; emergent generalization across diverse unseen tasks.
6	Dosovitskiy et al. [6]	2020	Vision Transformer (ViT)	Applied pure Transformer to non-overlapping image patches; matched and surpassed CNN performance on ImageNet at scale.
7	Raffel et al. [10]	2020	Text-to-Text Transfer Transformer (T5)	Reframed all NLP tasks as text-to-text problems; systematic scaling study; introduced C4 dataset.
8	Lewis et al. [12]	2020	BART (Denoising Seq2Seq)	Denoising autoencoder combining BERT-style encoder with GPT-style decoder; strong results on summarization and translation.
9	Ramesh et al. [15]	2021	DALL-E (Text-to-Image Transformer)	Autoregressively generated images from text using discrete VAE tokens; pioneered cross-modal generative modeling.
10	Radford et al. [13]	2021	CLIP (Contrastive LIP)	Contrastive language-image pretraining aligned visual and textual representations; strong zero-shot image classification.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15541

11	Ouyang et al. [14]	2022	InstructGPT (RLHF)	Applied reinforcement learning from human feedback to GPT-3; improved instruction-following and reduced harmful outputs.
12	Alayrac et al. [16]	2022	Flamingo (Visual Language Model)	Few-shot multimodal understanding via interleaved vision-text training; unified architecture for image and video understanding.
13	OpenAI [5]	2023	GPT-4 (Multimodal LLM)	Multimodal input (text + image); near-human performance on professional exams; stronger reasoning and reduced hallucination vs. GPT-3.5.
14	Bubeck et al. [11]	2023	Sparks of AGI Analysis (GPT-4)	Evaluated GPT-4 across diverse cognitive domains; argued it exhibits early-stage general reasoning abilities.
15	Hochreiter & Schmidhuber [8]	1997	Long Short-Term Memory (LSTM)	Gated recurrent architecture that addressed vanishing gradients in RNNs; dominant sequence model until the Transformer era.

Several consistent themes emerge across this body of literature. Scale remains a powerful lever: models that are larger in terms of both parameter count and training data consistently demonstrate superior performance across tasks. The pretraining paradigm has proven remarkably robust, transferring effectively from NLP to vision and multimodal settings. The shift toward instruction-following and alignment, exemplified by InstructGPT and GPT-4, signals a maturing recognition that raw capability must be paired with safety and controllability [14]. Finally, cross-modal learning—unifying text, vision, and other signals under shared Transformer architectures—has emerged as a powerful direction that continues to yield surprising capabilities.

V. COMPARATIVE ANALYSIS

A structured comparison of the four primary model categories reveals important trade-offs in terms of performance, computational requirements, and practical utility. Table II summarizes these characteristics across the model tiers described in this survey.

TABLE II. COMPARATIVE ANALYSIS OF GENERATIVE AI MODEL CATEGORIES

Model Category	Performance	Advantages	Limitations
RNN / LSTM	Moderate	Lightweight architecture; sequential inductive bias; well-understood training dynamics; suitable for streaming and low-resource settings.	Cannot be parallelized across time steps; poor long-range memory; vanishing gradient vulnerability; slow training on long sequences.
Transformer (Base) BERT / GPT-1/2	High	Full parallel processing within a layer; effective capture of long-range dependencies; strong NLP benchmark results; transferable via fine-tuning.	Quadratic memory and time complexity with sequence length; substantial compute required for pretraining; no built-in factual grounding.
Large Language Models GPT-3, GPT-4, PaLM	Very High	Emergent few-shot in-context learning; broad generalization without fine-tuning; powerful instruction-following after RLHF alignment.	Enormous training and inference costs; closed-source access limits auditability; persistent hallucination; difficulty verifying factual claims.
Multimodal & Agentic AI DALL-E, Flamingo, GPT-4V	State of the Art	Cross-modal understanding across text, image, and audio; autonomous multi-step task execution; versatile across a wide variety of real-world workflows.	Extremely complex architecture and training pipelines; significant alignment and safety challenges; high latency; requires substantial infrastructure.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.15541

The comparative analysis reveals a clear performance progression from Tier 1 to Tier 4, but this comes at steadily increasing computational cost and architectural complexity. RNNs and LSTMs, though largely superseded for generative tasks, remain relevant in resource-constrained environments and real-time streaming applications where their sequential inductive bias is not a disadvantage. The base Transformer strikes a compelling balance for applied NLP tasks, particularly when fine-tuned on domain-specific labeled data. LLMs represent the current state of the art for general-purpose language generation, though deployment costs and opacity raise practical and ethical concerns. Multimodal and agentic systems extend the frontier further but require substantial engineering investment to deploy safely and reliably at scale.

VI. RESEARCH GAPS

Despite remarkable progress, the field of Transformer-based generative AI faces several substantive and largely unresolved research challenges. These gaps are not merely engineering limitations; many touch on fundamental questions about how learning, representation, and reasoning function in large neural systems, and their resolution will determine whether generative AI fulfills its promise in high-stakes, real-world applications.

Gap 1 — High Computational Cost

Training state-of-the-art LLMs demands computational resources accessible to only a handful of well-funded organizations. GPT-3, for instance, required thousands of high-end GPUs running continuously for weeks, with training costs estimated in the millions of dollars [4]. The quadratic time and memory complexity of standard self-attention further compounds this problem for long-sequence tasks such as document summarization and code generation. Research directions including sparse attention mechanisms [17], linear attention approximations, mixture-of-experts architectures, and quantization-aware training hold promise, but no broadly adopted solution has fully resolved this scalability bottleneck. The environmental cost of large-scale training also warrants serious consideration as model sizes continue to grow.

Gap 2 — Hallucination Problem

A particularly concerning failure mode of large generative models is hallucination—the confident generation of factually incorrect, internally inconsistent, or entirely fabricated content [14]. This arises partly because these models are trained to maximize fluency and coherence rather than factual accuracy, and partly because their parametric memory is static and cannot be verified against a ground truth at inference time. Retrieval-augmented generation (RAG) and tool-augmented models offer partial mitigation by grounding generation in retrieved documents, but hallucination persists even in augmented systems. The problem constitutes a critical barrier to deployment in high-stakes domains such as clinical medicine, legal research, and financial advisory services.

Gap 3 — Lack of Explainability

The internal mechanisms through which Transformer-based models arrive at their outputs remain largely opaque. While attention weight distributions are sometimes treated as explanations, research has demonstrated that they do not reliably correspond to feature importance in a human-interpretable sense [9]. The absence of robust post-hoc or intrinsic explainability tools limits trust in model outputs, complicates error diagnosis and debugging, and creates barriers to regulatory compliance in sectors—healthcare, finance, and public administration—where decisions affecting individuals must be justifiable to those individuals and to overseeing bodies [11]. Developing explanation methods that are simultaneously faithful to model behavior, comprehensible to non-technical stakeholders, and computationally tractable remains an active and important open problem.

Gap 4 — Data Bias and Fairness

Generative models trained on internet-scale data inevitably absorb and may amplify the biases embedded in that data. These biases manifest in diverse ways: stereotypical associations between demographic groups and certain roles or attributes, skewed multilingual representation favoring high-resource languages, disparate downstream performance across gender, racial, and socioeconomic groups, and the potential to generate discriminatory content [7]. Despite active research into debiasing objectives, data filtering strategies, and post-hoc correction methods, none have proven comprehensively effective. The problem is further complicated by the inherent difficulty of operationally defining and measuring fairness in open-ended generative settings, where outputs exist on a continuous spectrum rather than in discrete categories.

Gap 5 — Limited Real-Time Knowledge Adaptation

Most large generative models are trained offline on static snapshots of web data and maintain a fixed knowledge cutoff date. They cannot autonomously update their internal representations to incorporate new events, recently published

DOI: 10.17148/IJARCCE.2026.15541

research, or emerging factual developments without full or partial retraining [10]. While retrieval-augmented systems partially address this limitation by querying external knowledge bases at inference time, deeply integrating dynamic world knowledge into a model's parametric memory in a scalable, accurate, and computationally efficient manner remains genuinely unsolved. Continual learning approaches aimed at avoiding catastrophic forgetting represent a promising avenue, though they remain immature for models at the scale of frontier LLMs. The gap between training-time knowledge and deployment-time reality grows wider the longer a model remains in service.

VII. CONCLUSION

This survey has traced the arc of Transformer-based generative AI from its architectural origins in the 2017 "Attention Is All You Need" paper through to the current era of multimodal, instruction-following, and agentic systems. The Transformer's foundational innovation—replacing sequential recurrence with parallel, attention-driven processing—has proven to be among the most generative ideas in the history of machine learning, catalyzing progress across natural language processing, computer vision, audio synthesis, code generation, and autonomous decision-making.

The four-tier taxonomy introduced in this paper offers a principled framework for understanding the progressive increase in capability and complexity across generations of generative models. From modest recurrent sequence learners constrained by vanishing gradients to autonomous AI agents that can reason across text, images, and structured tools, each tier represents a qualitative expansion of what machines can learn and produce. This progression is not merely quantitative scaling; it involves architectural innovations, new training paradigms, and fundamentally different modes of interacting with AI systems.

At the same time, this survey has identified five substantive research gaps that the community must address to realize the full potential of generative AI responsibly. Reducing the computational burden of training and inference, mitigating hallucination through improved factual grounding, advancing model explainability, correcting embedded biases, and enabling genuine real-time knowledge adaptation each represent critical and active research frontiers. Progress on any one of these challenges is likely to accelerate progress on the others, as they are interconnected both technically and conceptually.

Looking ahead, the most impactful advances will likely emerge at the intersection of these challenges. Architecturally efficient models—smaller yet more capable through better data curation, distillation, and hardware-aware design—will be essential for democratizing access to advanced AI capabilities. Advances in explainability and alignment will be critical for building appropriate, calibrated trust between human users and AI systems, particularly in consequential domains. Rigorous evaluation frameworks that go beyond perplexity and BLEU scores to assess reasoning, factual reliability, and ethical comportment will be vital for guiding development responsibly.

Generative AI is neither a solved problem nor a simple extrapolation of current scaling curves. It is an open scientific frontier with fundamental questions still unresolved—questions whose answers will shape the trajectory of AI's impact on science, industry, and society for decades to come. This survey aims to serve as a structured entry point into that frontier and a guide to where the most important work remains to be done.

REFERENCES

[1]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.

[2]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186.

[3]. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI Technical Report, 2018.

[4]. T. B. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[5]. OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, Mar. 2023.

[6]. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.

[7]. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in Proc. ACM FAccT, 2021, pp. 610–623.

[8]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15541

[9]. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.

[10]. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[11]. S. Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv preprint arXiv:2303.12528, Apr. 2023.

[12]. M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proc. ACL, 2020, pp. 7871–7880.

[13]. A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proc. ICML, 2021, pp. 8748–8763.

[14]. L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[15]. A. Ramesh et al., "Zero-Shot Text-to-Image Generation," in Proc. ICML, 2021, pp. 8821–8831.

[16]. J.-B. Alayrac et al., "Flamingo: A Visual Language Model for Few-Shot Learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[17]. I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, Apr. 2020.