Caliphate Journal of Science & Technology (CaJoST)

ISSN: 2705-313X (PRINT); 2705-3121 (ONLINE) Research Article

Open Access Journal available at: https://cajost.com.ng/index.php/files and https://www.ajol.info/index.php/cajost/index

DOI: https://dx.doi.org/10.4314/cajost.v8i1.23

Article Info

Received: 11th December 2025

Revised: 22nd February 2026

Accepted: 15th March 2026

Large Language Models: A Comprehensive Survey of Architectures, Applications and Challenges

Ahmed I. Mahmud[1], Ahmed Aliyu[2*], and Umar Sani[3]

[1]Department of Software Engineering, Federal University Dutsinma, Katsina State, Nigeria.
[2]Department of Mathematical Sciences, Sa'adu Zungur University, Bauchi State, Nigeria.
[3]Department of Computer Science, Sokoto State University, Sokoto State, Nigeria.

***Corresponding author's email:**

ahmedaliyu@sazu.edu.ng

Large Language Models (LLMs) have emerged as a foundational technology in contemporary artificial intelligence, driven largely by the scalability and expressiveness of the Transformer architecture. This survey provides a structured and comprehensive review of recent advances in LLM research across five key dimensions: (i) scaling laws and foundational models, (ii) sparse and efficiency-oriented architectures, (iii) multimodal and tool-augmented models, (iv) privacy, alignment, and safety mechanisms, and (v) domain-specific applications and existing surveys. The reviewed literature indicates that while large-scale model training yields consistent performance gains and emergent capabilities, it also introduces significant challenges related to computational cost, energy consumption, interpretability, and environmental sustainability. Efficiency-driven approaches, including sparse attention mechanisms and Mixture-of-Experts architectures, mitigate some of these constraints but increase system complexity. Multimodal and tool-enhanced LLMs extend reasoning capabilities beyond text; however, they remain susceptible to hallucinations and grounding errors. Furthermore, alignment and privacy-preserving methods are actively evolving, yet the absence of standardized evaluation frameworks and clearly defined utility–safety trade-offs limit their practical deployment. Empirical studies in application domains such as healthcare and education demonstrate substantial promise but emphasize the necessity of trust, explainability, and regulatory compliance. This survey synthesizes current progress, identifies persistent research gaps, and outlines future directions toward the development of efficient, reliable, and responsible large language models.

Keywords: Large Language Model, Natural Language Processing, Generative Pre-Train Transformers, Mask Language Model, Variational Auto encoder.

Cite this: *CaJoST*, 2026, 1, 218-226

1. Introduction

The introduction of Transformers marked a pivotal advancement in the progression of Natural Language Processing (NLP). The development of Large Language Models (LLMs) is a major step forward for this framework. LLM have shown to be very good at many NLP task. These models have become essential for task like generating text, machine translation, and natural language comprehension. Their progress over time has not only make them better at language tasks, but it has also shown how well they can work in multimodal situations, like with images, videos, and robotic system. People have used Large Language Model (LLMs) for a number of things, such as, Machine translation, which facilitates precise content conversion among different linguistic systems. Logical reasoning entails the examination and deduction of conclusions based on the logical framework of a certain situation. Text creation, in which coherent and contextually pertinent text is generated from structured inputs according to particular directives. Multimodal processing, wherein LLMs transcend solely textual domains to incorporate various input and output modalities, including images, videos, and robotic interfaces. Summarization involves the contextual compression of substantial content into succinct representations.

The inception of LLM development dates to 2018 with the introduction of models like GPT (Chowdhery et al., 2022) and BERT (Anil et al., 2023). Although each was crafted with unique architectural characteristics to tackle issues in NLP, modern LLMs predominantly utilize the esteemed Transformer architecture and are classified into three main categories. (a) Auto-encoding models, characterized by their encoder-centric architecture and optimized for contextual comprehension tasks, are exemplified by BERT and its derivatives. (b)

Encoder-decoder models amalgamate both encoder and decoder processes to capitalize on the benefits of the two prior types, but with certain trade-offs. Prominent instances of the Pangu family. (c) Auto-regressive models, which prioritize decoders and are particularly effective for generative applications, exemplified by the GPT series.

Each category has its own strengths and limitations, which show that it can be used in a wide range of applications and circumstances. The Transformer architecture, introduced by Vaswani et al. (2017), represents a significant shift from previous sequence modelling methods by eliminating recurrence and convolution in favour of a purely attention-based framework. The model fundamentally utilizes multi-head self-attention to calculate contextual dependencies among all input tokens in parallel, effectively mitigating the inefficiencies associated with recurrent neural networks (RNNs), which sequentially process tokens and are susceptible to vanishing gradient phenomena when modelling long-range dependencies (Hochreiter et al., 1997; Bengio et al., 1994). In contrast to convolutional neural networks (CNNs), which depend on fixed receptive fields and hierarchical aggregation (LeCun et al., 1998), Transformers possess a theoretically limitless receptive field, allowing for dynamic weighting of global context throughout sequences (Dosovitskiy et al., 2021). The Transformer architecture consists of stacked layers of multi-head self-attention and position-wise feed-forward networks, reinforced by residual connections and layer normalization (Ba et al., 2016), with positional encodings added to address the model's permutation invariance. This design exhibits exceptional scalability and representational capability, supporting the advancement of Large Language Models (LLMs) like BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), while also enabling cross-domain generalization to computer vision (Dosovitskiy et al., 2021), time-series forecasting (Zhou et al., 2021), and multimodal learning (Alayrac et al., 2022). The Transformer signifies a paradigm change, forming the fundamental basis for current advancements in artificial intelligence.

Building upon the foundational principles introduced by the Transformer architecture, subsequent models have adapted its framework to address specific learning objectives and application needs. One prominent adaptation is the class of Auto-encoding models, which focus on capturing deep bidirectional contextual representations through masked input reconstruction. These models are trained using a masked language modelling (MLM) or analogous denoising objective, in this approach, part of the inputs is purposely concealed, and the model is refined to anticipate them based on the surrounding context. This method allows the model to evaluate both left and right contexts at the same time. This makes it better at tasks that require discrimination, like classification, named entity recognition, and question answering, instead of generative text production. Recent studies have built on and improved this model. For example, “Masked AutoDecoder is Effective Multi-Task Vision Generalist” (Qiu et al., 2024) applies the autoencoding principle to a multimodal, vision-task generalist framework by masking and reconstructing vision-task sequence inputs, achieving competitive accuracy with improved inference efficiency. These models are good for representation learning but are not suitable for tasks requiring sequential generation because of their intrinsic masking and non-autoregressive architecture. To overcome this limitation, Auto-regressive models were developed, they utilize a decoder-only framework to build sequences in a sequential manner, predicting each token based on previously generated tokens using a causal language modelling objective. This unidirectional architecture facilitates coherent and contextually consistent text generation, rendering these models especially proficient for generative tasks such as dialogue, summarization, and code synthesis. In contrast to auto-encoding models which learn bidirectional representations, auto-regressive models are proficient in modelling long-range dependencies for coherent sequence generation, albeit this results in slower inference due to their sequential decoding mechanism. Recent research further enhances their scalability and adaptability; for instance, Bai et al. (2024) emphasize improvements in efficient training and inference methodologies for decoder-only big language models, highlighting their significance in modern AI systems.

Sequence-to-sequence (Seq2Seq) models are neural architectures that map an input sequence into an output sequence of variable length, rendering them proficient for tasks such as machine translation, summarization, and dialogue generation. Seq2Seq modelling, originally implemented with recurrent neural networks (Sutskever et al., 2014) and enhanced by attention mechanisms (Bahdanau et al., 2015), was transformed by the introduction of the Transformer (Vaswani et al., 2017), which utilizes self-attention for parallelized contextual learning. Within this paradigm, an encoder creates contextual representations of the input, while a decoder autoregressively constructs the output based on both the encoder states and previous predictions. Recent improvements have expanded Seq2Seq approaches to multimodal domains, facilitating adaptable input-output mappings spanning text, vision, and speech (Neil et al., 2025). In addition to Seq2Seq architectures, recent works have increasingly turned to generative models that aim to capture deeper hidden structures within data. One notable approach is the Variational Autoencoder (VAE), which extends the concept of traditional autoencoders by learning probabilistic

rather than deterministic representations. The encoder transforms input data into probabilistic parameters (mean and variance), from which latent variables are derived, while the decoder reconstructs the input from these samples. Training optimizes a joint objective- a reconstruction loss to maintain data fidelity and a Kullback–Leibler (KL) divergence component to regularize the latent space towards a prior distribution, usually Gaussian (Kingma et al., 2014). This approach allows VAEs to produce unique data samples and acquire robust representations, rendering them effective in fields like vision, language, and bioinformatics. Recent improvements enhance VAEs through expressive priors and disentangled or multimodal representations, improving generative quality and adaptability (Child, 2020).

The contributions of this research are as follows:

i. Present a comprehensive analysis of the architectures employed in contemporary large language models (LLMs).

ii. Catalogue and analyse the primary application domains in which LLMs are utilized.

iii. Recognize principal problems, constraints, and hazards.

iv. Suggest potential research avenues based on contemporary literature.

2. Literature Review

The literature review on Large Language Models (LLMs) can be taxonomically organised into five categories as shown in Figure 1: (a) scaling and foundational studies, (b) sparse, mixture-of-experts, and efficiency-focused designs, (c) multimodal and tool-enhanced LLMs, (d) privacy, alignment, and safety methodologies, and (e) domain-specific applications and surveys. Each subsection offers a critical analysis, emphasising parallels, contrasts, and existing gaps.

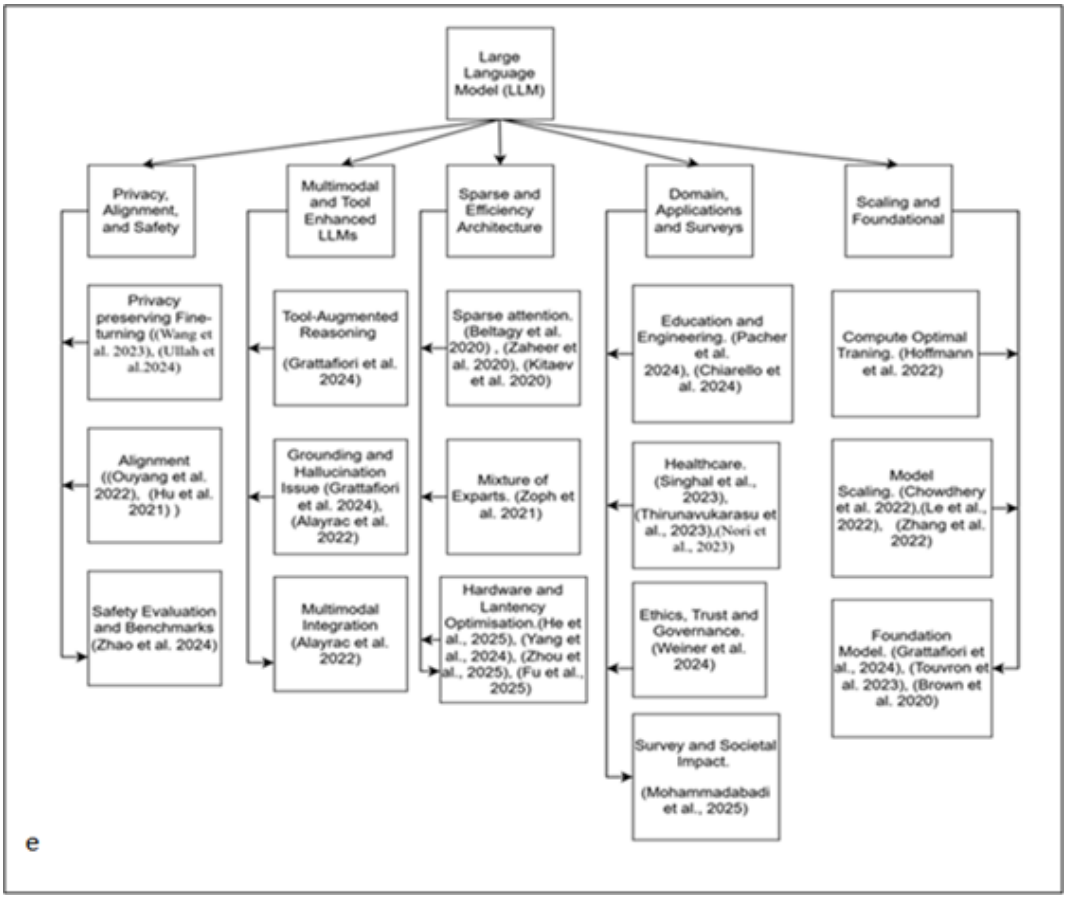

Figure 1: Large language model taxonomy

2.1 Scaling and Foundational

Scaling and foundations of big language models stress that increasing the size of the model, the amount of the training data, and that amount of the computing power improves performance and fosters emergent capabilities, including few-shot learning.

Brown et al. (2020) introduced GPT-3, showing that extensive autoregressive transformers had significant few-shot and zero-shot capacities, hence

initiating the foundation model era. Hoffmann et al. (2022) challenged the idea that bigger models are always better. They showed with Chinchilla that compute optimal training that using moderately sized models trained on a lot more tokens can give better results than using bigger models that are not well trained. Chowdhery et al. (2022) expanded to 540 billion parameters with PaLM, demonstrating reasoning and linguistic capabilities, while Zhang et al. (2022) on the other hand, introduced OPT as an open alternative to proprietary models. Touvron et al. (2023) introduced LLaMA, providing researchers with accessible and competitive models, succeeded by the LLaMA-3 family (Grattafiori et al., 2024), which features extended context and tool-use functionalities.

Scaling studies indicate that capability increases with data, parameters, and computational resources, while also emphasising cost, energy, and accessibility constraints. Chinchilla's findings redirected the discourse towards data efficiency. Open releases such as OPT, BLOOM (Le et al., 2022), and LLaMA facilitate access while raising questions over dataset provenance.

2.2 Sparse and Efficiency Architecture

Advancements in LLMs that use sparse and efficient structures have solved the problem of traditional transformer attention's high memory and processing costs, which grow quadratically with sequence length. Several studies address the quadratic complexity of attention mechanisms. Beltagy et al. (2020) introduced Longformer, which integrates local and global attention mechanisms. BigBird was introduced by (Zaheer et al., 2020), which combines block-sparse and random attention, and provides theoretical guarantees. Kitaev et al. (2020) introduced Reformer, which incorporates LSH attention and reversible layers, whereas Choromanski et al. (2021) created Performer, utilizing kernelized linear attention. Zoph et al. (2021) introduced Switch Transformer, a sparse Mixture-of-Experts (MoE) model that efficiently scales by activating just specific subgroups of experts. Recent works TriangleMix (He et al., 2025), TidalDecode (Yang et al., 2024), Progressive Sparse Attention (Zhou et al., 2025), and H2EAL (Fu et al., 2025) focus on minimising memory usage, enhancing latency, and optimising hardware co-design. Collectively, they illustrate that static sparse patterns are straightforward and economical although less adaptable, whereas dynamic sparsity and hardware aware techniques attain superior robustness and performance, although at the cost of complexity and portability.

2.3 Multimodal and Tool Enhanced LLMs

Multimodal and tool-augmented models improve LLMs by adding extra inputs like images, audio, or code, and by using external tools. The model proposed by (Alayrac et al., 2022), which facilitates multimodal few-shot learning using text and images, together with BLIP and BEiT, showcases advanced vision-language capabilities through integrated vision-language pretraining. (Grattafiori et al., 2024) highlighted the significance of tool utilisation and expanded context in LLaMA-3. These advancements demonstrate that LLMs can generalise beyond textual data; yet assessments indicate ongoing challenges related to grounding, factual accuracy, and multimodal hallucinations.

2.4 Privacy, Alignment, and Safety

Ouyang et al. (2022) presented InstructGPT employing Reinforcement Learning from Human Feedback (RLHF), recognizing alignment as a crucial component. (Hu et al. 2021) introduced LoRA for effecient parameter adaptation. Wang et al. (2023) presented PrivateLoRA for privacy partition between edge and cloud. Ullah et al. (2024) conceptualised PrivChatGPT, examining differential privacy techniques. Zhao et al. (2024) performed an exhaustive evaluation of privacy preserving methodologies. Mobility in networks has also been considered for on-road safety, reinforcement learning can be applied to improve the safety mechanism (Aliyu et al., 2018; Aliyu et al., 2020). These studies collectively recognize the importance of privacy and safety; however, they lack standardised criteria and often underreport utility trade-offs.

2.5 Domain, Applications and Surveys

Mohammadabadi et al. (2025) released a survey on LLMs, focus on architectures, benchmarks, and societal ramifications. Chiarello et al. (2024) examined the development of applications across various industries. Systematic studies in engineering education (Pacher et al., 2024) highlight prospective efficiency improvements while tackling ethical and regulatory issues. Healthcare research is promising, but it also shows that trust, understanding, and compliance are still big problems. Healthcare is one of the most promising but also dangerous areas for LLMs to work in. Recent evaluations suggest that models like GPT-4 and MedPaLM exhibit considerable potential in clinical decision support, medical question answering, and the Mohammadabadi et al. (2025) released a survey on LLMs, focus on architectures, benchmarks, and societal ramifications. Chiarello et al. (2024) examined the development of applications across various industries. Systematic studies in engineering education (Pacher et al., 2024) highlight prospective efficiency improvements while tackling ethical and regulatory issues. Healthcare research is promising, but it also shows that trust, understanding, and compliance are still big problems. Healthcare is one of the most promising but also dangerous areas for LLMs to work in. Recent evaluations suggest that models like GPT-4 and MedPaLM exhibit

considerable potential in clinical decision support, medical question answering, and the summarisation of electronic health records (Singhal et al., 2023; Thirunavukarasu et al., 2023). These systems demonstrate performance approaching expert levels on medical benchmarks, indicating their potential usefulness for both clinicians and patients. Nonetheless, research indicates that hallucinations, absence of source attribution, and inconsistent performance across specialities undermine reliability in critical clinical settings (Nori et al., 2023). Consequently, although technical feasibility is confirmed, clinical translation necessitates meticulous validation and supervision.

A common theme in healthcare literature is the need for trust, understanding, and following rules. For example, (Sun et al., 2025) contend that the incorporation of LLMs into health systems must prioritize explainability to foster clinician confidence, while (Weiner et al., 2024) underscore regulatory and ethical consideration such as data privacy, informed consent, and alignment with medical guidelines. These concerns parallel findings from broader surveys (Mohammadabadi et al., 2025), which stress that domain deployment must be accompanied by transparency, safeguards against bias, and robust governance. Together, healthcare studies illustrate the double-edged nature of LLMs, they offer unprecedented support for medical practice, but their safe and ethical integration depends on rigorous evaluation, trust-building, and alignment with healthcare standards.

3. Methodology

This study adopts a systematic survey-based methodology to analyse recent research on Large Language Models. Relevant literature was identified through comprehensive searches of major scientific repositories, including arXiv, IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar. Peer-reviewed journal articles, conference papers, and influential preprints published between 2018 and 2025 were considered. The selected studies were filtered based on relevance, citation impact, and methodological rigor. Following selection, the literature was categorised into five thematic groups: (1) scaling and foundational studies, (2) sparse and efficiency-focused architectures, (3) multimodal and tool-enhanced LLMs, (4) privacy, alignment, and safety mechanisms, and (5) domain-specific applications and survey works. Each category was analysed comparatively to identify core contributions, architectural trends, performance implications, and unresolved challenges. This qualitative synthesis approach enables a holistic understanding of the LLM research landscape while highlighting emerging directions and open research problems.

4. Results and Discussion

The analysis of the reviewed literature reveals several key findings. First, scaling laws consistently demonstrate that increasing model size, training data, and compute resources leads to improved performance and emergent capabilities such as in-context learning and reasoning. However, these gains come at the cost of increased energy consumption, financial expense, and reduced accessibility (Brown et al., 2020; Hoffmann et al., 2022). Second, efficiency-oriented architectures including sparse attention, linear attention, and Mixture-of-Experts designs significantly reduce computational and memory overheads (Beltagy et al., 2020; He et al., 2025), enabling longer context processing and faster inference. Despite their benefits, these approaches introduce architectural complexity and pose challenges for deployment portability and hardware compatibility. Third, multimodal and tool-augmented LLMs substantially expand functional capabilities by integrating vision, audio, and external tools. While these models achieve strong performance in multimodal reasoning tasks, persistent issues related to hallucination, grounding, and evaluation reliability remain unresolved (Alayrac et al., 2022; Grattafiori et al., 2024). Fourth, alignment, safety, and privacy-preserving techniques such as Reinforcement Learning from Human Feedback (RLHF), LoRA-based adaptation, and differential privacy frameworks have become central to responsible LLM development. Nevertheless, current methods lack standardized benchmarks and often present unclear trade-offs between model utility and risk mitigation (Ouyang et al., 2022; Wang et al., 2023). Finally, application-focused studies particularly in healthcare, education, and engineering demonstrate that real-world deployment depends not only on technical accuracy but also on trust, explainability, regulatory compliance, and ethical governance (Mohammadabadi et al., 2025; Thirunavukarasu et al., 2023).In general, opportunities lie in bridging these gaps, developing compute efficient yet capable models, designing alignment frameworks with measurable standards, and conducting longitudinal studies to validate LLMs in real environments. In general, opportunities lie in bridging these gaps, developing compute efficient yet capable models, designing alignment frameworks with measurable standards, and conducting longitudinal studies to validate LLMs in real environments. The table 1 summarizes reviewed works, their focus, contributions, and key limitations.

Table 1: Summary of Reviewed Works

Category	Papers	Contribution	Key Limitations
Scaling and Foundations	Brown et al. (2020); Hoffmann et al. (2022); Chowdhery et al. (2022); Touvron et al. (2023); Grattafiori et al. (2024)	Scaling laws, foundation models, compute-optimal training	High cost, energy use, limited accessibility
Sparse / Efficient Architectures	Beltagy et al. (2020); Zaheer et al. (2020); Yang et al. (2025); He et al. (2025); Fu et al. (2025)	Sparse/linear attention, Mixture-of-Experts, hardware-aware designs	Complexity, trade-offs in accuracy and portability
Multimodal / Tool-augmented	Alayrac et al. (2022); Grattafiori et al. (2024)	Vision-language models, multimodal reasoning, tool use	Hallucinations, grounding errors, evaluation gaps
Alignment and Safety	Ouyang et al. (2022); Hu et al. (2021); Wang et al. (2023); Ullah et al. (2024); Zhao et al. (2024)	RLHF, LoRA, PrivateLoRA, privacy-preserving models	No standard benchmarks, unclear utility trade-offs
Alignment and Safety	Ouyang et al. (2022); Hu et al. (2021); Wang et al. (2023); Ullah et al. (2024); Zhao et al. (2024)	RLHF, LoRA, PrivateLoRA, privacy-preserving models	No standard benchmarks, unclear utility trade-offs
Applications / Surveys	Mohammadabadi et al. (2025); Chiarello et al. (2024); Singhal et al. (2025); Thirunavukarasu et al. (2023); Sun et al. (2025)	Surveys in education, healthcare, AEC, and industry	Trust, regulation, ethical and compliance issues

5. Future Research Direction

In the future, few areas will not be worth further research by scholars. To begin with, the design of compute- and data-efficient architectures will become crucial to make large language models more democratic by reducing their environmental impact. Second, the research community needs to come up with powerful multidimensional standards that collaboratively evaluate accuracy, fairness, alignment, privacy, robustness, and energy use. Third, there is a need to develop multimodal grounding and reasoning in order to enable users to communicate dependably with text, vision, audio, and structured information. Fourth, privacy-assuring and reliable training paradigms, such as federated learning, differential privacy, and edge-cloud paradigm hybrid approaches cannot be dropped in sensitive fields. Lastly, scientists and politician need to collaborate to develop international policies and ethics to promote responsible innovation and mitigate the risks of abuse and provide fair access to large language models advantages. With the right inclination toward the technical advancement and the values of society, future research work on the large language models can go beyond the performance benchmark and create an inclusive, transparent, and long-lasting AI infrastructure instead.

6. Conclusion

This survey has look at thirty important studies on Large Language Models (LLMs), these studies cover foundational scaling research, efficient and sparse architectures, multimodal and tool-augmented systems, alignment and privacy frameworks, and domain-specific applications. These studies show both the exceptional abilities and the problem they still face. Scaling laws show that data and model size will lead to predictable improvements, yet they also reinforce inequalities in compute access and raise sustainability concerns. New Improved method that focuses on efficiency such as sparse attention and Mixture-of-Experts models show promise pathways to reduce computational overhead, but they often add complexity and make deployment more difficult. On the other hand, multimodal and tool-augmented models push LLMs toward broader applicability, but issues of hallucination, factual grounding, and evaluation reliability remain unresolved. Works on alignment, safety, and privacy is getting momentum, but current methods are not standardized and don't do a good job of balancing usefulness with risk reduction. Lastly, application-focused surveys show that LLMs are becoming more important in healthcare, education, engineering, and industry, this show that practical use depends on trust, interpretability and compliance as much as technical performance.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

The authors gratefully acknowledge the reviewers for their constructive comments and valuable suggestions, which significantly improved the quality and clarity of this manuscript.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). *Flamingo: a visual language model for few-shot learning*. arXiv. https://arxiv.org/abs/2204.14198.

Aliyu, A., Abdullah, A. H., Kaiwartya, O., Cao, Y., Usman, M. J., Kumar, S., ... & Raw, R. S. (2018). Cloud computing in VANETs: architecture, taxonomy, and challenges. IETE Technical Review, 35(5), 523-547.

Aliyu, A., Abdullah, A. H., Kaiwartya, O., Hussain Madni, S. H., Joda, U. M., Ado, A., & Tayyab, M. (2020). Mobile cloud computing: taxonomy and challenges. Journal of Computer Networks and Communications, 2020(1), 2547921.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z. (2023). PaLM 2 technical report. arXiv preprint arXiv:2305.10403. https://arxiv.org/abs/2305.10403.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450.* https://doi.org/10.48550/arXiv.1607.06450.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1409.0473.

Bai, Y., Zeng, A., Hou, L., Zhang, T., Liu, Z., & Sun, M. (2024). *Decoder-only or not: A systematic survey of language models*. arXiv. https://arxiv.org/abs/2404.10968.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The long-document transformer*. arXiv. https://arxiv.org/abs/2004.05150.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5*(2), 157–166. https://doi.org/10.1109/72.279181.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165.

Chiarello, F., et al. (2024). Future Applications of Generative Large Language Models. Journal of Information Science. https://doi.org/10.1016/j.technovation.2024.103002.

Child, R. (2020). *Very deep VAEs generalize autoregressive models and can outperform them on images*. https://doi.org/10.48550/arXiv.2011.10650.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Šarlós, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, Ł., Belanger, D., Colwell, L., & Weller, A. (2021). *Rethinking attention with performers*. arXiv. https://arxiv.org/abs/2009.14794.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., & Fiedel, N. (2022). *PaLM: Scaling language modeling with pathways*. arXiv. https://arxiv.org/abs/2204.02311.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, 4171–4186. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR 2021).* https://doi.org/10.48550/arXiv.2010.11929.

Fu, Z., Guo, X., Zeng, W., Zhong, S., Zhang, Y., Chen, P., Wang, R., Ye, L., & Li, M. (2025). *H_2EAL: Hybrid-bonding architecture with hybrid sparse attention for efficient long-context LLM inference*. *Proceedings of the 2025 International Conference on Computer-Aided Design (ICCAD)*. https://doi.org/10.48550/arXiv.2508.16653.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark,

A., Rao, A., Zhang, A.,

He, Z., Zhang, Y., Zhang, C., Jiang, H., Yang, Y., & Qiu, L. (2025). *TriangleMix: A lossless and efficient attention pattern for long context pre-filling*. arXiv. https://arxiv.org/abs/2507.21526.

Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., & Sifre, L. (2022). *Training compute-optimal large language models*. arXiv. https://arxiv.org/abs/2203.15556.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv. https://arxiv.org/abs/2106.09685.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1312.6114.

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). *Reformer: The efficient transformer*. arXiv. https://arxiv.org/abs/2001.04451.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., Villanova del Moral, A., … Workshop, B. (2023). *BLOOM: A 176B-parameter open-access multilingual language model*. arXiv. https://arxiv.org/abs/2211.05100.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791.

Neil Zeghidour· Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, Alexandre Défossez. (2025). *Streaming Sequence-to-Sequence Learning with Delayed Streams Modelling*. https://doi.org/10.48550/arXiv.2509.08753.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on medical challenge problems*. arXiv. https://arxiv.org/abs/2305.14254.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. https://arxiv.org/abs/2203.02155.

Pacher, C., Woschank, M., Zunk, B. M., & Gruber, E. (2024). *Engineering education 5.0: A systematic literature review on competence-based education in the industrial engineering and management discipline*. *Production & Manufacturing Research, 12*(1). https://doi.org/10.1080/21693277.2024.2337224.

Qiu, H., Huang, J., Gao, P., Lu, L., Zhang, X., & Lu, S. (2024). *Masked AutoDecoder is Effective Multi-Task Vision Generalist*. https://doi.org/10.48550/arXiv.2403.07692.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://doi.org/10.48550/arXiv.1801.06146.

Sajjadi Mohammadabadi, S. M., Kara, B. C., Eyupoglu, C., Uzay, C., Tosun, M. S., & Karakuş, O. (2025). *A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications*. https://doi.org/10.3390/electronics14183580.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Sun, Q., Akman, A., & Schuller, B. W. (2025). *Explainable artificial intelligence for medical applications: A review*. ACM Transactions on Computing for Healthcare, 6(2), 1-31. https://doi.org/10.1145/3709367.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems, 27*. https://doi.org/10.48550/arXiv.1409.3215.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). *Large language models in medicine*. *Nature Medicine, 29*(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. https://arxiv.org/abs/2302.13971.

Ullah, I., Hassan, N., Gill, S. S., Suleiman, B., Ahanger, T. A., Shah, Z., Qadir, J., &

Kanhere, S. S. (2023). *Privacy preserving large language models: ChatGPT case study based vision and framework*. arXiv. https://arxiv.org/abs/2304.09990.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30* (NeurIPS 2017). https://doi.org/10.48550/arXiv.1706.03762.

Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2024). *Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice*. arXiv. https://arxiv.org/abs/2412.03576.

Wang, Y., Lin, Y., Zeng, X., & Zhang, G. (2023). *PrivateLoRA: For efficient privacy preserving LLM*. arXiv. https://arxiv.org/abs/2308.01881.

Yang, L., Zhang, Z., Chen, Z., Li, Z., & Jia, Z. (2024). *TidalDecode: Fast and accurate LLM decoding with position persistent sparse attention*. arXiv. https://arxiv.org/abs/2410.05076.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). *Big Bird: Transformers for longer sequences*. arXiv. https://arxiv.org/abs/2007.14062.

Zhao, B., Ji, Y., Shi, Y., & Jiang, X. (2024). *Design and implementation of privacy-preserving federated learning algorithm for consumer IoT*. Alexandria Engineering Journal, 206–216. https://doi.org/10.1016/j.aej.2024.06.071.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

Zhou, Q., Yin, P., Zuo, P., & Cheng, J. (2025). *Progressive Sparse Attention: Algorithm and system co-design for efficient attention in LLM serving*. arXiv. https://arxiv.org/abs/2503.00392.

Zoph, B., Kumar, S., Dean, J., Bello, I., Du, N., Shazeer, N., Huang, Y., & Fedus, W. (2022). *ST-MOE: Designing stable and transferable sparse expert models*. arXiv. https://arxiv.org/abs/2202.08906.

