Transformers as the Foundation of Large Language Models: A Comprehensive Review

Muhammad Rayan Shaikh[1], Nida Shahryar[2], Khalid Mahboob[3], Qurrat-ul-Ain Naiyar[3], Muhammad Talha[1]
[1]Karachi Institute of Economics and Technology (KIET).
[2]Indus University Karachi.
[3]Institute of Business Management Karachi (IoBM).
*Correspondence: mr.shaikh@kiet.edu.pk

Citation | Shaikh M. R, Shahryar. N, Mahboob. K, Naiyar. Q. U. A, Talha. M, "Transformers as the Foundation of Large Language Models: A Comprehensive Review", IJIST, Vol. 7 Issue. 4 pp 2705-2717, November 2025
Received| October 05, 2025 **Revised|** October 20, 2025 **Accepted|** October 26, 2025 **Published|** November 08, 2025.

The transformation of Transformer architecture has led the way into a new era for NLP, as it broke the traditional RNNs, LSTMs, Seq2Seq models, etc. As their main feature, the Revolution of Transformers was the hybridization of self-attention and multiheaded attention, which allowed the models to learn dependencies across time spans of any length through positioning methods. This resulted in a quick and efficient process for training large-scale Language Models (LLMs) that could handle the data very well with simultaneous approach to learn the long-term dependencies. This paper not only reflects but also presents a critically reviewed path taken by LLMs from BERT to GPT-4 and beyond, along with the better reasoning, arithmetic and instruction following attributed to the scaling up of architecture. The review further indicates and discusses the current concerns regarding efficiency, bias, interpretability and domain specialization and warns that settling these issues might dictate the fate of T-bases improvements. The authors aim through this project to provide an exhaustive comprehension of the setting in which Transformers enabled LLMs and actively directed the development of contemporary AI research.

Keywords: Transformer Architecture; Natural Language Processing (NLP); Sequence-to-Sequence; LLMs Large Language Models; BERT; GPT; Foundation Models.

Introduction:

The last decade can be considered nothing less than an amazing transformation period for Natural Language Processing (NLP). NLP has grown to the extent that the field is now seen as very rich and entirely `e`cel' based on state-of-the-art practices of computational linguistics and deep learning [1]. The transition from traditional statistical and rule-based methods to contemporary large-data-based deep learning models is according to Derek and Edwin (2013) which changed the nature of machine understanding, human language processing and even production of language [2][3]. The first use of neural network types, such as RNNs and LSTMs, opened up the possibility to represent linguistic sequences to the extent that the computer could understand the links among the words and simultaneously, keep the overall meaning of the passage [4]. At that time, these models were not just the best in language translation, speech recognition and text summarization. Still, it was recognized that the main power of neural networks in understanding and conveying the meanings through sequences had been demonstrated.

Nevertheless, the model limitations are becoming even more apparent as the complexity and size of language data keep increasing. RNNs and LSTMs are incapable of processing data in parallel. They can only let the information flow through tokens one at a time [5]. This slows down the training time and increases it altogether. Besides, their restricted ability to keep long-term dependencies results in the notorious vanishing and exploding gradient problems, which hamper the models' performance during long-context reasoning tasks[6][7]. Thus, these structures cannot produce texts that are coherent and of good quality over long sequences.

NLP systems demand that they can handle enormous quantities of digital text in a fast way without exhausting the understanding of the context and being able to scale up would require a total overhaul of the current architectures. As such, the researchers start pointing to the developed models that can take the whole text sequences to process at once and simultaneously, find both the local and global dependencies using parallel computing. This aspiration results in a drastic change in the field of NLP [8], which in turn, changes the meaning and learning of language [8][9].

The transformer model, presented by Vaswani and his team in 2017 [10], marked a turning point in the transformation process. The Transformer architecture, which draws on the self-attention mechanism, permits the model to simultaneously consider relationships between all tokens in a sequence [11]. The model's multi-head attention capability enables it to pick up multiple aspects of the meaning from various parts of the sentence, resulting in a deeper understanding of the context. Furthermore, positional encoding keeps the word order information without relying on recurrent processing [12][3].

The new architectural innovations eliminated several inefficiencies present in the previous models. Scalability, efficiency and context comprehension have all been greatly improved, with complete parallelization still being a significant factor in today's supermodel training. The transformer is a revolutionary development that changes the computation paradigm of NLP and lays the groundwork for modern language modeling [13][9].

The advent of the Transformer architecture marks the beginning of a new epoch for Large Language Models (LLMs) [14]. The enormous training process of these models undergo to give them not only immense linguistic understanding but also general world knowledge to an astonishing level, wherein they could perform language tasks such as understanding, reasoning and generating texts at that high level of proficiency. BERT, GPT and T5 are among the models that prove the point that Transformer-based models not only surpass previous records but also alter the direction of the whole NLP research from being limited to the creation of models for each specific task to having flexible and pre-trained architectures

that can be fine-tuned or prompted for different applications with a minimum of effort [14][15][16].

The paper elucidates the Transformer model's capability to overcome the main issues of the recurrent architecture and lay the groundwork for the present-day core models. The transition from recurrence to attention was not merely a technological breakthrough but a change in the machine's comprehension and communication with the human language [15][6]. As a result, Natural Language Processing has entered a period of massive operations, flexibility and almost perfect human-like language understanding.

Objective:

This all-encompassing review intends to subject the entire journey of neural language models that brought Transformer-based Large Language Models (LLMs) to the forefront to a very critical examination[17][4]. This paper first wants to put together the existing studies to prove the superiority of the Transformer framework over such former models as RNNs, LSTMs and Seq2Seq networks in sequential and contextual improvement. Besides that, it aims to investigate the architectural principles, training methodologies and scaling characteristics that have supported LLMs like BERT, GPT and GPT-4 in achieving state-of-the-art performance[11][12][18]. This review is based on the joining of various studies' results. It leads to an overall viewpoint on how the Transformer-induced changes have impacted the NLP field as well as the problems of efficiency, interpretability, bias and computational sustainability, which are still relevant and will continue to influence LLM research's future direction, being the exactities of the case in point.

Literature Review:

The changes in NLP have mostly been influenced by the continual emergence of sophisticated neural models that can represent the syntactic and semantic aspects of language at a deep level. The groundbreaking research on Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and Sequence-to-Sequence (Seq2Seq) frameworks had a significant impact on the handling of sequential data. These models demonstrated that they could follow linguistic sequences over time, thus giving rise to the first paradigm for the neural computation of context interpretation. Still, their reliance on recursive operations introduced certain unavoidable drawbacks, mainly vanishing gradients, restricted scalability and poor handling of long-range dependencies, which made it difficult for these models to be applied to large and complex language datasets.

The introduction of the Transformer architecture by Vaswani and his co-authors and other researchers was the first step to a complete transformation in using deep learning techniques for NLP. The model provided the opportunity for the interaction of all tokens in the sequence at once and caused the total abolition of recurrent connections, introducing the self-attention mechanism. Such an advance reduced the delay imposed by time in previous architectures and simultaneously made it possible to model global contextual dependencies efficiently. Furthermore, adopting multi-head attention made it possible to conduct the parallel extraction of the different linguistic relations, syntactic, semantic and discourse-level, thus augmenting the representation capacity. The usage of positional encoding made up for the lack of sequence order due to the absence of recurrence, providing knowledge of position without sacrificing computation speed [3].

When compared to each other, RNN and LSTM models engaged in a debate that led to the recognition of temporal and contextual dependency modeling and the Transformer, with its attention-based non-sequential computation paradigm. The shift from recurrence memory to attention that worked in parallel was not just an architectural improvement but it opened a new way of thinking about linguistic structures through learning and representation. All these developments can be seen as the basis on which the modern Large Language Models

(LLMs) have been built because they have already achieved the highest standards of the three attributes in question: scalability, contextual accuracy and richness in representation.

Limitations of Earlier Architectures:

RNNs and their variants, including LSTMs and Seq2Seq models, were the only options that neural sequence modeling could rely on before the advent of the Transformer. These architectures were still catching up in the domain of NLP, but they had the fundamental issues that raised doubts about their scalability and overall effectiveness[10].

RNNs and LSTMs: Sequential Bottleneck:

RNNs turned to recurrence even more than in their previous attempts at capturing the information of sequences. Unfortunately, the requirement of computing at every time step created a sequential bottleneck, which, in turn, caused difficulties in applying efficient and effective parallelization[7]. Moreover, despite the innovations in LSTMs and GRUs, these models remained plagued by the problem of vanishing gradients, preventing them from remembering the information of long input sequences. As a result, RNN-based models often could not develop long-range dependencies essential for understanding the language[19].

Seq2Seq Models: Context Vector Bottleneck:

Using a fixed-length context vector for encoding the input sequences has greatly allowed Seq2Seq models to improve the quality of machine translation and summarization; the code is then converted into the output sequence. Despite the fact that this approach works very well for short sentences, it creates a bottleneck. Out of a single vector, when the input length is increased, it gets tough to keep all the pertinent information, thus causing a decrease in translation quality and a loss of meaning.

Attention with RNNs: Partial Relief, Persistent Inefficiency:

The attention mechanism significantly lessened the context bottleneck by granting the models the ability to focus on distinct inputs during the decoding phase[5]. However, the benefit of attention in terms of increased precision and better matching was still restricted to the sequential limitation of RNN-based architectures, which made it difficult to progress with larger datasets [20].

Table 1. Comparison of Sequence Modeling Architectures

Model	Key Mechanism	Advantages	Limitations
RNN	Recurrence	Sequential context capture	Vanishing gradients, slow training
LSTM	Memory gates	Improved context retention	Still sequential, limited scalability
Seq2Seq	Encoder-decoder	Better translation performance	Context compression bottleneck
Transformer	Self-attention	Parallelism, long-range dependencies	High computational cost

The Transformer Architecture:

Vaswani et al.'s (2017) launch of the Transformer was a revolutionary step in the discipline of NLP. The Transformer is realized as a complete attention model that can process and scale to long sequences very efficiently, unlike the earlier models, which used recurrence. The innovative approach completely relied on attention mechanisms and eliminated the sequential blockage in the training process of the current (LLMs) Large Language Models [14].

Self-Attention: Capturing Global Dependencies:

The Transformer is fundamentally based on a self-attention mechanism. The latter allows every token in a sequence to connect directly with all the other tokens and thus produce a global contextual representation. Formally, given an input matrix, $X \in \mathbb{R}^{n \times d}$, the model computes three learned projections:

$$Q = XW_Q, K = XW_K, V = XW_V$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$Weight matrices represent queries, keys and values. The attention weights are calculated using the **scaled dot-product attention**:

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

This process decides the amount of attention each token should place on others in the sequence and normalizes the relevance scores to get a stable result. To illustrate, in the phrase "The dog that chased the cat was tired", the self-attention mechanism makes "was" point to "dog" instead of "cat," though "cat" is nearer in terms of position [21]. This approach enables the model to grasp long-range dependencies without recurrence, thus handling memory limitations and improving contextual comprehension [21][15].

Multi-Head Attention: Multiple Perspectives:

A single attention head may capture only one type of relationship (e.g., syntactic or semantic). To enrich representation learning, the Transformer employs multi-head attention, executing $h$independent self-attention operations in parallel:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

Where each head is defined as

$$\text{head}_i = \text{Attention}(QW_Q^{(i)}, KW_K^{(i)}, VW_V^{(i)})$$

and $W_O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$Merges the heads into a unified representation [13].

The individual heads are responsible for different parts of the language e.g., one might pay attention to the agreement between the subject and the verb. At the same time, the other one captures the relation of meaning through synonyms or the flow of the topic. The input from the different heads gives a deep and multi-faceted understanding of the text. It provides a very rich representation that surpasses the previous models in this aspect [13][14][16].

Positional Encoding: Order Without Recurrence:

Since Transformers do not use recurrence, they lack inherent word order information. **Positional encodings** are introduced to inject sequence order directly into input embeddings. For each position $pos$and dimension i, the encoding is defined as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

These sinusoidal functions allow the model to learn absolute and relative positional information efficiently [4]. When added to word embeddings, positional encodings ensure that the Transformer recognizes token order without sacrificing parallel computation speed [8]. This design preserves accuracy in order-sensitive tasks such as translation and summarization.

Parallelization and Efficiency:

The Transformer's ability to do parallel computations is one of the significant benefits. RNNs, on the other hand are limited to processing one sequence at a time, while Transformers can process all tokens together. This configuration reduces training time and enables training on massive datasets, which is one of the main requirements for LLMs [21][14]. The gain from parallelization is related to the scaling laws of the language models, which declare that performance improves predictably with the increase of data and parameters.

In actual application, this signifies that older architectures, which required weeks or months to complete tasks, can now be done in days or even hours. Thus, it is possible to create and train the million-parameter models, which greatly support present-day generative AI.

Table 2. Transformer Architecture Components

Component	Function	Description	Example in LLM
Self-Attention	Context modeling	Computers the relationships between all tokens in a sequence	Word alignment in translation
Multi-Head Attention	Parallel attention	Captures multiple contextual dependencies	Syntax and semantics modeling
Positional Encoding	Sequence order	Adds information of position to token embeddings	order in sentences
Feed-Forward Network	Feature transformation	Applies a nonlinear transformation to embeddings	Hidden layer feature extraction
Layer Normalization	Stability control	Normalizes activations for faster convergence	Improves training stability
Residual Connections	Gradient flow	Prevents vanishing gradients by skip connections	Deep transformer layers

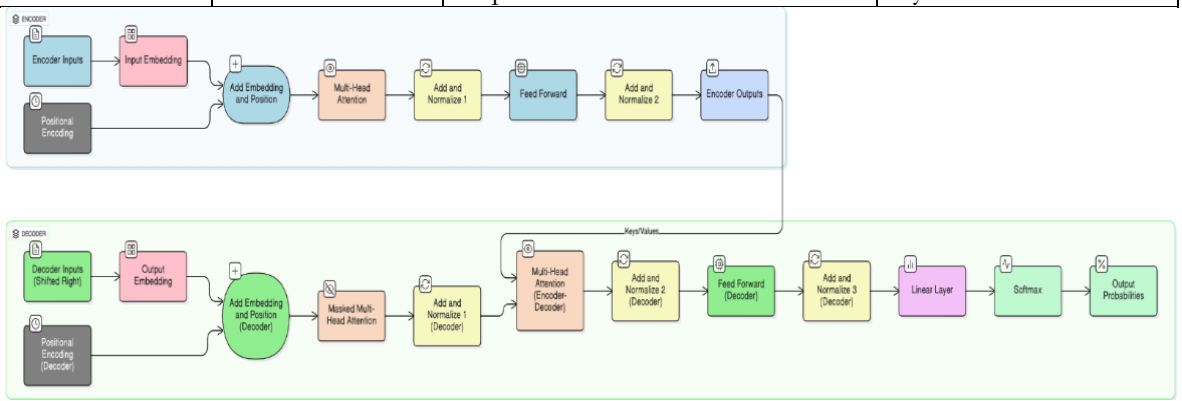

Figure 1. Overview of Transformer architecture

Large Language Models (LLMs) Enabled by Transformers:

The advent of LLMs cannot be considered separately from the transformations in architecture and computing brought about by the Transformer model. Neural architectures previous to the Transformers, like RNNs and LSTMs, made breakthroughs in sequential processing but were still limited in scalability and context understanding, mirroring the state of large datasets. On the other hand Transformers tackled these problems by introducing parallelized computation and attention-based context modeling; thus, they became the technological support that LLMs were built on. LLMs could not have been opened up through models with billions of parameters unless the Transformers had provided the efficiency, representational power and scalability.

The training of Transformer-based LLMs, however, still incurs enormous costs. In addition to the vast amounts of computations, specialized hardware such as GPU or TPU clusters and vast amounts of data often crawled from the open web, the training process also depends on these factors. The concerns of dependency on (crucial) energy consumption, carbon footprint and data bias propagation are raised. The quality and representativeness of training data directly affect the model's behavior, which means that the exact mechanisms that allow for generalization can also make societal or linguistic biases stronger if they are not adequately managed. So, even though Transformers took large-scale language modeling to a new level, they also posed ethical and infrastructural challenges that set the limits of scalability in practice [6][13].

Self-Supervised Training at Scale:

By enabling self-supervised learning with vast amounts of unlabeled data, the Transformer architecture has brought about one of the most revolutionary changes in this area. The most straightforward next-token prediction task, which is guessing the missing or

next word in a sentence, has proved surprisingly powerful when done in the multi-layered, parallelly arranged Transformer blocks [7][20][5]. For example, given the input "The scientist presented the ...", the model can suggest following words like "results" or "paper" and in this way, pretty much learns the semantic, syntactic and pragmatic relationships from the large text corpora. By processing trillions of tokens, LLMs gain grammatical accuracy and contextual mastery and cultivate reasoning and adaptation skills compatible with certain domains and styles [6][13].

The parallel computation of Transformers is primarily responsible for the scalability. On the other hand Recurrent models are restricted to processing one step at a time and consequently suffer from length-dependent inefficiencies. Transformers process the whole sequence in a single go. This is, therefore, what makes very powerful self-supervised pretraining computationally feasible. Nevertheless, the mentioned scalability also brings in a vast amount of power and an environmental factor to be considered. Training of the top LLMs today needs thousands of hours on a GPU and vast amounts of energy consumption, which brings up the issue of sustainability and at the same time, limits the access of smaller institutions to such state-of-the-art techniques.

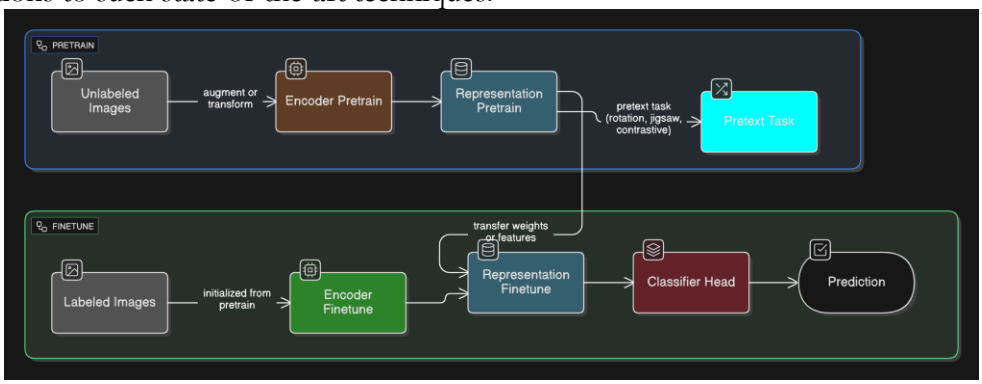

Figure 2. Self-supervised pretraining and fine-tuning workflow.

Emergent Abilities Through Scaling:

The scaling of Transformer-based models shows their capability to discover new things that are not programmed or trained for. The case of scaling is fascinating; not only do we get performance improvements with the increase of the size of parameters and volumes of data, but also qualitatively new behaviors like arithmetic reasoning, multi-step problem-solving and even instruction-following. One method that amplifies these reasoning pathways is chain-of-thought prompting, which enables LLMs to express the intermediate stages that lead to the conclusion [3][5]. The unveiling of these capacities means that the process of scaling of Transformers offers computer-like behavior which could be regarded as "intelligent" to a certain degree, not just within the limits of smaller or recurrent ones [14][19][16].

Nonetheless, the increase in size shows diminishing returns at the massive end of the spectrum, where the performance improvements are minimal compared to the exponential rise in the cost of computing. Additionally, bigger models tend to hallucinate and dilute the context more, thus reaffirming that the increase in scale cannot replace the methodological refinement or alignment with human values.

Transfer Learning and the Foundation Model Paradigm:

The appearance of the foundation model paradigm led to other AI conceptual frameworks being influenced by transformers. The transfer learning of a model is based on the pretraining of a general-purpose Transformer-based LLM on a diverse textual corpus, which is later fine-tuned for specialized domains, like medicine, law, or education, without having to retrain from the beginning [14][10]. This flexibility is based on competent self-attention mechanisms and the high capacity of Transformers, which makes it possible for

contextualized representations to be applied in significantly different linguistic and conceptual domains.

However, the reliance on enormous pretraining corpora and a powerful computational infrastructure makes this paradigm very resource-consuming and unaffordable for many research groups. In addition, the fine-tuned models may overfit to the domain or lose their general linguistic robustness. Nevertheless, the Transformer architecture recognized the technical side of NLP and the epistemological side of AI development, which led the field from task-specific models to universal, adaptable and continually evolving systems [10][15].

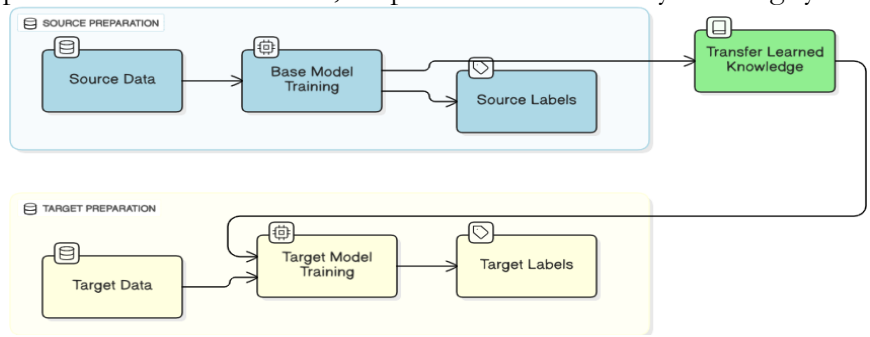

Figure 3. Transfer learning workflow in LLMs

Evolution of LLMs (Timeline):

The development of LLMs one after another displays that the Transformer architecture has become the basis for the extraordinary scaling and innovation in natural language processing. Every generation represents an intentional design evolution in architecture, parameter size, data usage and application area, showing that the enhancements in ideas and techniques are adding up to yield the qualitatively new capabilities.

BERT (2018): Google's Bidirectional Encoder Representations from Transformers (BERT) pioneered the importance of bidirectional context modeling through a solely encoder-based Transformer architecture. By covering some tokens up during the pre-training step and estimating them using both the left and right context, BERT set a new standard for the models' capturing of semantic dependencies. It also proposed pre-training together with fine-tuning as a universal pipeline, which enabled us to gain the highest scores in question-answering, sentiment classification and language inference, among other things. The main credit of BERT was the change of perception; it showed that a Transformer encoder could acquire linguistic representations better than those done manually.

GPT (2018): The model developed by OpenAI, known as Generative Pre-trained Transformer (GPT), went in a different direction by using a decoder-only architecture and performing autoregressive generation. The originality of GPT was in its unsupervised, vast corpus pre-training followed by specific task fine-tuning, which confirmed the assumption that generative pre-training of massive amounts improves downstream performance significantly. Even though it was pretty small compared to others, GPT pointed out that scaling the model's capability goes hand-in-hand with the amount of data and parameters, which became the main principle for all the following models.

GPT-2 (2019): GPT-2 was a breakthrough in auto-regressive Transformers, producing consistent and coherent texts. It had 1.5 billion parameters and was trained on enormous amounts of data from the Internet, which made it possible for the model to produce syntax-consistent and contextually relevant long passages even though there was no explicit conditioning. This breakthrough led researchers to conclude that language comprehension can be derived from next-token prediction, thus replacing the traditional text generation perception. On the downside, GPT-2 prompted the first ethical and safety discussions about the generation of synthetic texts, referring to the AI governance issue in the LLM research area as the dawn of the AI governance concerns.

GPT-3 (2020): The release of the 175 billion parameter GPT-3 was an excellent example of a scaling revolution. The architecture of GPT-3 followed the same Transformer paradigm but utilized massive parameter growth and diversity of training data. Its power in few-shot and zero-shot learning signaled a theoretical shift: LLMs could infer unseen tasks just by prompting. Consequently, GPT-3 converted LLMs from narrow task solvers to general-purpose reasoning systems.

InstructGPT and Alignment (2022): OpenAI honed GPT-3 via Reinforcement Learning from Human Feedback (RLHF), giving birth to InstructGPT. This model indicated that alignment tuning—refining outputs to human understanding and ethical limits—could create LLMs that are more secure and smooth in their interaction with users, even without any modifications in the architecture. The achievement of InstructGPT led to a change in research focus from mere scaling to alignment and controllability, thus marking a conceptual shift to human-centered AI design.

LaMDA (2022): Google's Language Model for Dialogue Applications (LaMDA) went further still in the concept of dialog specialization. It was a breakthrough that turned the theoretical discussion into a practical application by showcasing the three main features of dialogue—coherence, grounding in facts and safety. This innovation led to using LLMs in a new area, from the outset of open-text generation to prolonged human-like interaction. LaMDA established detailed tuning of objectives for dialog safety and engagement; thus, it was a step in developing how LLMs could balance creativity and dependability with interactive systems.

PaLM (2022): With 540 billion parameters, Google's Pathways Language Model (PaLM) did not just lead the trends in the way of scaling but at the same time, introduced the Pathways framework - facilitating multi-task and multi-modal training over diverse data. PaLM's remarkable performance in reasoning, coding and understanding different languages proved that the large language models could act as cross-domain cognitive models, which can perform abstract reasoning beyond text comprehension.

ChatGPT (2022): Having been trained on the GPT-3.5 model, ChatGPT was the first publicly available LLM. Its ability to understand and execute instructions, along with a conversational interface, made it possible for users to access research prototypes as common tools. ChatGPT demonstrated that if proper alignment and accessibility measures are in place, the adoption of LLMs will greatly increase, thus showing their practical use in the fields of education, creativity and communication.

GPT-4 (2023): GPT-4 pushed the limits of the Transformer architecture even further through multimodality, meaning it could process both text and images simultaneously. The model achieved significant advancements in depth of reasoning, correctness of facts and alignment with safety; thus, it was no longer a case of scale alone but qualitative gains through architectural optimization. Besides, with the introduction of adaptive inference mechanisms, GPT-4 set new standards for efficiency, robustness and trustworthiness and became the best in these points.

Emergent and Specialized Models (2023): The LLM ecosystem has completely opened up and diversified into open-source and domain-specific directions. Meta's LLaMA series highlighted the importance of parameter efficiency and accessibility; Stanford's Alpaca showed the potential of instruction-tuning with the least resources; and Google's Med-PaLM 2 was an excellent instance of domain specialization in the medical field. The models that emerged so far have represented a process of decentralized innovation, which implies that LLMs are progressively being developed to be adaptable, low-cost and to provide expert-level specialization.

Table 3. Comparative Analysis of Prominent LLMs

Model	Year	Parameters (B)	Architecture Type	Key Feature	Organization
BERT	2018	0.34	Encoder	Bidirectional pre-training	Google
GPT-2	2019	1.5	Decoder	Autoregressive generation	OpenAI
GPT-3	2020	175	Decoder	Few-shot learning	OpenAI
PaLM	2022	540	Decoder	Pathways scaling	Google
LLaMA	2023	65	Decoder	Efficient open source	Meta
GPT-4	2023	>1000	Multimodal	Text–image reasoning	OpenAI

This progression illustrates how each milestone from BERT's contextual representations to GPT-4's multimodality was built on the Transformer architecture's foundations. Table 4. The evolution of LLMs highlights the dual forces of scaling and specialization, shaping a landscape where LLMs are now central to AI research, deployment and application.

Table 4. Advantages and Limitations of Transformer-based LLMs

Aspect	Advantages	Limitations
Scalability	Efficient parallel training on GPUs/TPUs	Requires massive computational power
Context Understanding	Model's long-range dependencies	Can generate hallucinations
Transfer Learning	Easily adaptable to new domains	Fine-tuning requires careful curation
Multimodality	Processes text and images jointly	High memory footprint
Performance	Outperforms older models on benchmarks	Ethical and safety issues persist

Challenges and Future Directions:

The development of LLM has been a major change in the NLP area, but the challenges related to its widespread use still have to be overcome for proper progress to happen.

Efficiency and Cost:

Massive amounts of computing power, specific hardware and electricity are all prerequisites for training and using LLMs. This scenario questions the environmental impact and availability of such technology, as hardly any organizations can train models at this scale these days. Hence, it is supported that future investments should be directed towards more efficient designs, parameter sharing and compression techniques that would cut costs without compromising performance.

Bias, Misinformation and Safety:

LLMs are mirrors of their training data and as a result, they sometimes exaggerate stereotypes or generate harmful content. On the other hand the skill of producing such realistic text also adds to the risks associated with misinformation, disinformation and the bad use of these models. Research on alignment techniques, robust filtering and ethical norms to guarantee the safety of these models will always be a priority.

Explainability and Interpretability:

Even with their robust features, LLMs are still regarded as "black boxes." Understanding the reason for a specific output from a model remains a significant issue. It is necessary to trust, especially in sensitive sectors such as healthcare, education and law, that visualization, probing methods, or naturally transparent designs will help to improve understanding.

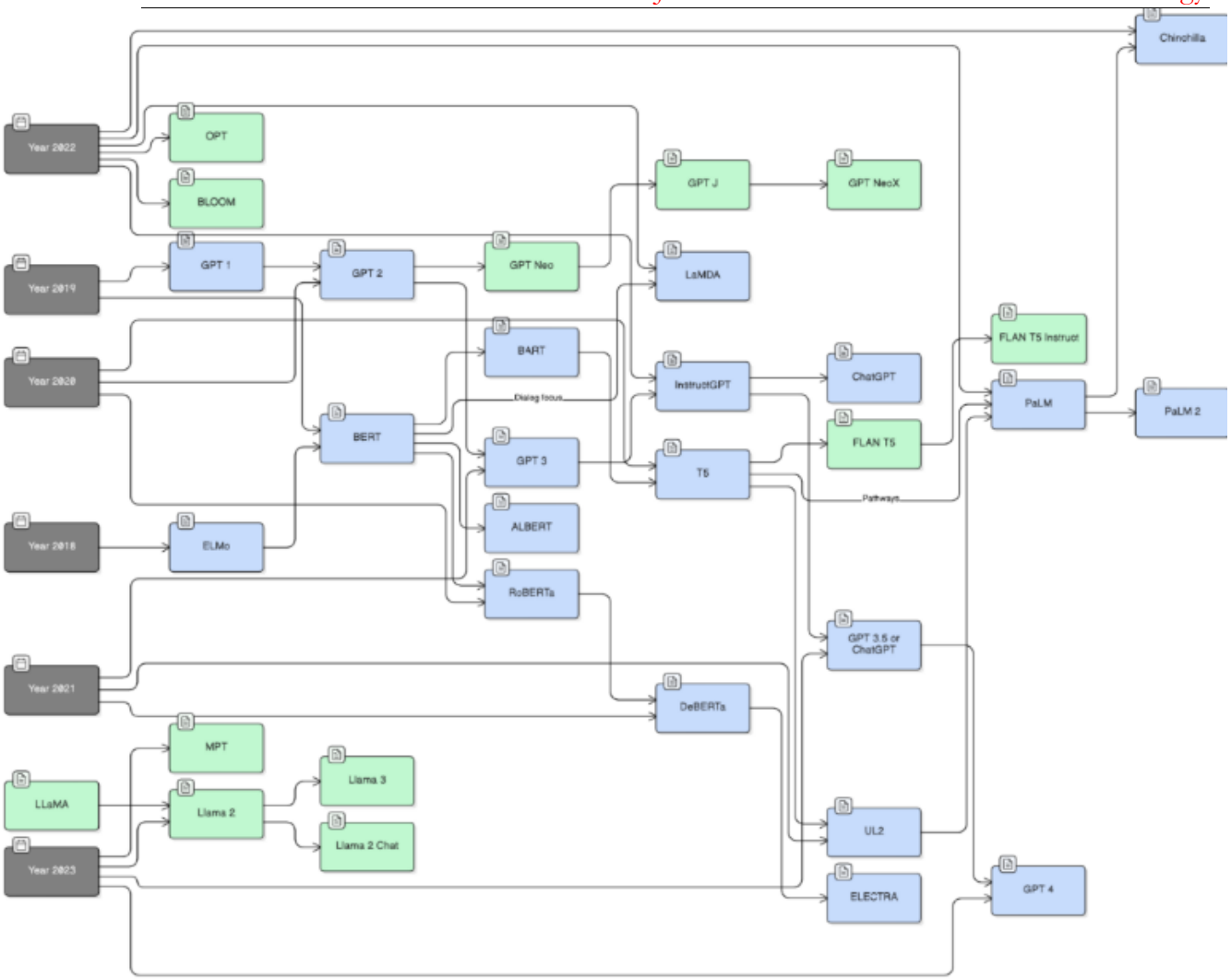

Figure 4. Evolutionary map of major (LLMs) Large Language Models from 2018–2023

Domain Adaptation:

It is possible that general-purpose LLMs would not be remarkably effective in specialized fields such as law or finance. Trustworthy LLMs in critical applications will have to be adapted through fine-tuning, retrieval-augmented generation, or efficient parameter methods [22]. Future research needs to investigate how to strike a balance between general versatility and domain-specific expertise.

Table 5. Key Research Directions for Future LLMs

Focus Area	Description	Expected Outcome
Efficiency Optimization	Sparse attention, quantization and distillation	Lower cost and energy use
Bias Mitigation	Fairness-aware training data and evaluation	Ethical AI systems
Interpretability	Explainable AI and visualization tools	Improved trust and accountability
Domain Adaptation	Domain-specific fine-tuning and retrieval augmentation	Enhanced accuracy in specialized fields
Multimodal Integration	Combining vision, speech and text	Broader AI applications

Conclusion:

The advent of LLMs, large language models, is a significant milestone in artificial intelligence. The major factor behind this development is Transformer technology. With the introduction of Transformers, the limitations of processing sequences in the old models were overcome and consequently, parallel processing, self-attention and scalability were improved. This progress resulted in training models with billions of parameters [23][24].

The progression of LLMs, along with BERT's contextual representations, to the multimodal reasoning of GPT-4 has illustrated the transformation of research prototypes into AI systems widely used for scaling and innovation. However, concerns over efficiency, fairness, interpretability and domain adaptation suggest that the path has not yet been fully traveled [3][21].

Future LLM development will rely on the scaling-up process and the creation of responsible, efficient and trustworthy systems. The above issues need to be tackled to ensure that LLMs can permanently and ethically benefit society and science [16].

References:

[1] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang and W. Zhang, "History, development and principles of large language models: an introductory survey," *AI Ethics 2024 53*, vol. 5, no. 3, pp. 1955–1971, Oct. 2024, doi: 10.1007/S43681-024-00583-7.

[2] R. J. Zishan Guo, "Evaluating Large Language Models: A Comprehensive Survey," *Int. J. Latest Eng. Manag. Res.*, vol. 9, no. 10, pp. 5–16, 2023, doi: https://doi.org/10.48550/arXiv.2310.19736.

[3] M. A. K. Raiaan, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[4] Pranjal Kumar, "Large language models (LLMs): survey, technical frameworks and future challenges," *Artif. Intell. Rev.*, vol. 57, no. 260, 2024, doi: https://doi.org/10.1007/s10462-024-10888-y.

[5] J. W. Long Ouyang, "Training language models to follow instructions with human feedback," *Adv. Neural Inf. Process. Syst.*, 2022, doi: https://doi.org/10.48550/arXiv.2203.02155.

[6] G. Bharathi Mohan *et al.*, "An analysis of large language models: their impact and potential applications," *Knowl. Inf. Syst. 2024 669*, vol. 66, no. 9, pp. 5047–5070, May 2024, doi: 10.1007/S10115-024-02120-8.

[7] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," Jul. 2023, Accessed: Jun. 05, 2025. [Online]. Available: https://arxiv.org/pdf/2307.06435v9

[8] S. M. Narendra Patwardhan, "Transformers in the Real World: A Survey on NLP Applications," *Information*, vol. 14, no. 4, p. 242, 2023, doi: https://doi.org/10.3390/info14040242.

[9] Y. Annepaka and P. Pakray, "Large language models: a survey of their development, capabilities and applications," *Knowl. Inf. Syst. 2024 673*, vol. 67, no. 3, pp. 2967–3022, Dec. 2024, doi: 10.1007/S10115-024-02310-4.

[10] G. Z. Ian A. Scott, "The new paradigm in machine learning - foundation models, large language models and beyond: a primer for physicians," *Intern. Med. J.*, vol. 54, no. 5, pp. 705–715, 2024, doi: 10.1111/imj.16393. Epub 2024 May 7.

[11] U. Kamath, K. Keenan, G. Somers and S. Sorenson, "Large Language Models: A Deep Dive: Bridging Theory and Practice," *Large Lang. Model. A Deep Dive Bridg. Theory Pract.*, pp. 1–472, Jan. 2024, doi: 10.1007/978-3-031-65647-7/COVER.

[12] B. C. K. Seyed Mahmoud Sajjadi Mohammadabadi, "A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges and Societal Implications," *Electronics*, vol. 14, no. 18, p. 3580, 2025, doi:

https://doi.org/10.3390/electronics14183580.

[13] P. Illangarathne, N. Jayasinghe and A. B. D. De Lima, "A Comprehensive Review of Transformer-Based Models: ChatGPT and Bard in Focus," *2024 7th Int. Conf. Artif. Intell. Big Data, ICAIBD 2024*, pp. 543–554, 2024, doi: 10.1109/ICAIBD62003.2024.10604437.

[14] C. Wang, M. Li and A. J. Smola, "Language Models with Transformers," *arXiv Prepr. arXiv1904.09408*, 2019, Accessed: Nov. 01, 2025. [Online]. Available: https://github.com/cgraywang/

[15] Q. L. Ce Zhou, "A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT," *Int. J. Mach. Learn. Cybern.*, 2023, doi: https://doi.org/10.48550/arXiv.2302.09419.

[16] G. Bansal, V. Chamola, A. Hussain, M. Guizani and D. Niyato, "Transforming Conversations with AI—A Comprehensive Study of ChatGPT," *Cogn. Comput. 2024 165*, vol. 16, no. 5, pp. 2487–2510, Jan. 2024, doi: 10.1007/S12559-023-10236-2.

[17] D. Myers *et al.*, "Foundation and large language models: fundamentals, challenges, opportunities and social impacts," *Clust. Comput. 2023 271*, vol. 27, no. 1, pp. 1–26, Nov. 2023, doi: 10.1007/S10586-023-04203-7.

[18] S. Minaee *et al.*, "Large Language Models: A Survey," Feb. 2024, Accessed: Apr. 20, 2025. [Online]. Available: https://arxiv.org/abs/2402.06196v3

[19] G. L. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, "LLaMA: Open and Efficient Foundation Language Models," *arXiv:2302.13971*, 2023, doi: https://doi.org/10.48550/arXiv.2302.13971.

[20] X. W. Yupeng Chang, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, 2024, doi: https://doi.org/10.1145/3641289.

[21] S. Suthar, K. Kanani and B. Dalwadi, "Evolution and Advancements in ChatGPT: A Comprehensive Survey," *Smart Innov. Syst. Technol.*, vol. 430 SIST, pp. 131–142, 2025, doi: 10.1007/978-981-96-1206-2_12.

[22] Z. Lai, X. Zhang and S. Chen, "Adaptive Ensembles of Fine-Tuned Transformers for LLM-Generated Text Detection," *Proc. Int. Jt. Conf. Neural Networks*, 2024, doi: 10.1109/IJCNN60899.2024.10651296.

[23] Z. L. Yuhong Mo, Hao Qin, Yushan Dong, Ziyi Zhu, "Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm," *arXiv:2405.06652*, 2024, doi: https://doi.org/10.48550/arXiv.2405.06652.

[24] I. Tasou, P. Anastasiadis, P. Mpakos, D. Galanopoulos, N. Koziris and G. Goumas, "Breaking Down LLM Inference: A preliminary performance analysis of sparsified transformers," *Proc. - 2025 IEEE Int. Parallel Distrib. Process. Symp. Work. IPDPSW 2025*, pp. 991–995, 2025, doi: 10.1109/IPDPSW66978.2025.00154.