Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

Exploring Large Language Models Architectures Applications, and Emerging Challenges

Vinay Kumar Maginam,

Software engineer, Londonderry , NH,USA.

maginamvinayit@gmail.com

Abstract:
This survey provides an in-depth exploration of Large Language Models (LLMs), examining notable architectures such as GPT-3, GPT-4, LLaMA, and PaLM. The paper traces the architectural evolution from traditional neural language models to cutting-edge transformer-based systems. Detailed insights are provided on training methodologies, including pre-training, fine-tuning, and instruction-tuning, which have enhanced the versatility and performance of LLMs across a range of applications, including natural language processing, text summarization, and code generation. This survey also discusses the current challenges LLMs face, such as bias in model outputs, ethical concerns, and the computational demands of scaling these models. Through analysis, we highlight the potential of LLMs to revolutionize industries while underscoring the need for efficient training techniques to mitigate their resource-intensive nature. Our findings indicate that while LLMs offer transformative capabilities, addressing ethical and practical limitations will be critical to their future development.

Keywords*: Large Language Models, GPT-3, GPT-4, Transformer Architecture, Pre-training*

I.INTRODUCTION

The development of AI in the recent past has a result created a new category of extremely capable NLP tools, specifically called LLMs. Text like this can be input and output of LLMs including GPT3/4 from OpenAI, BERT from Google, and LLaMA from Meta due to their size, accuracy and versatility. These models are one such radical new techniques for 'teaching' robots how to understand, interpret and generate language which act on scores of billions. Trained with digital deep learning algorithms with nearly billions of parameters and flexibility, LLMs have reformatted the easy conversational access to applications and knowledge domains across healthcare, education, business, and the media and entertainment sectors.

This structure is incorporated into the fabric of these models, and is principally derived from

Transformers have helped LLMs to isolate context-based communication activities in texts such as positional encoding and self-attention. These basic enhancements make it possible for LLMs to acquire tact subtle features that help them in careers such as summarizing, translating, and responding to inquiries.

Nevertheless, it is impossible to doubt that LLMs are technically qualified, and they also have some specific challenges and limitations. The major concerns with their deployment and training include their accessibility and sustainability based on the large amount of computing power required to support them. The potential to give highly realistic though false details known as hallucinations, underlines the need to improve dependability and ethical considerations, further.

Hence the aim and purpose of this chapter is to provide comprehensive analyzed reviews of LLMs with understanding of fresh ideas of architecture of these systems. It discusses several applications in multiple industries, focusing on how they are transforming activities and increasing productivity. Further, the chapter discusses new concerns relating to LLMs consisting of environmental concerns, ethical questions, and governance problems. Through

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

discussing these aspects, the chapter intends to develop an all-encompassing picture as to the existing condition of LLMs and their positioning to pave the way for AI's evolution.

This exploration is therefore particularly relevant given the fast emerging and rapidly evolving strength and power of LLMs. In giving a hand to the researchers, or enabling efficiency within the business framework, or helping in any creative pursuit, these models are spearheading the AI movement.

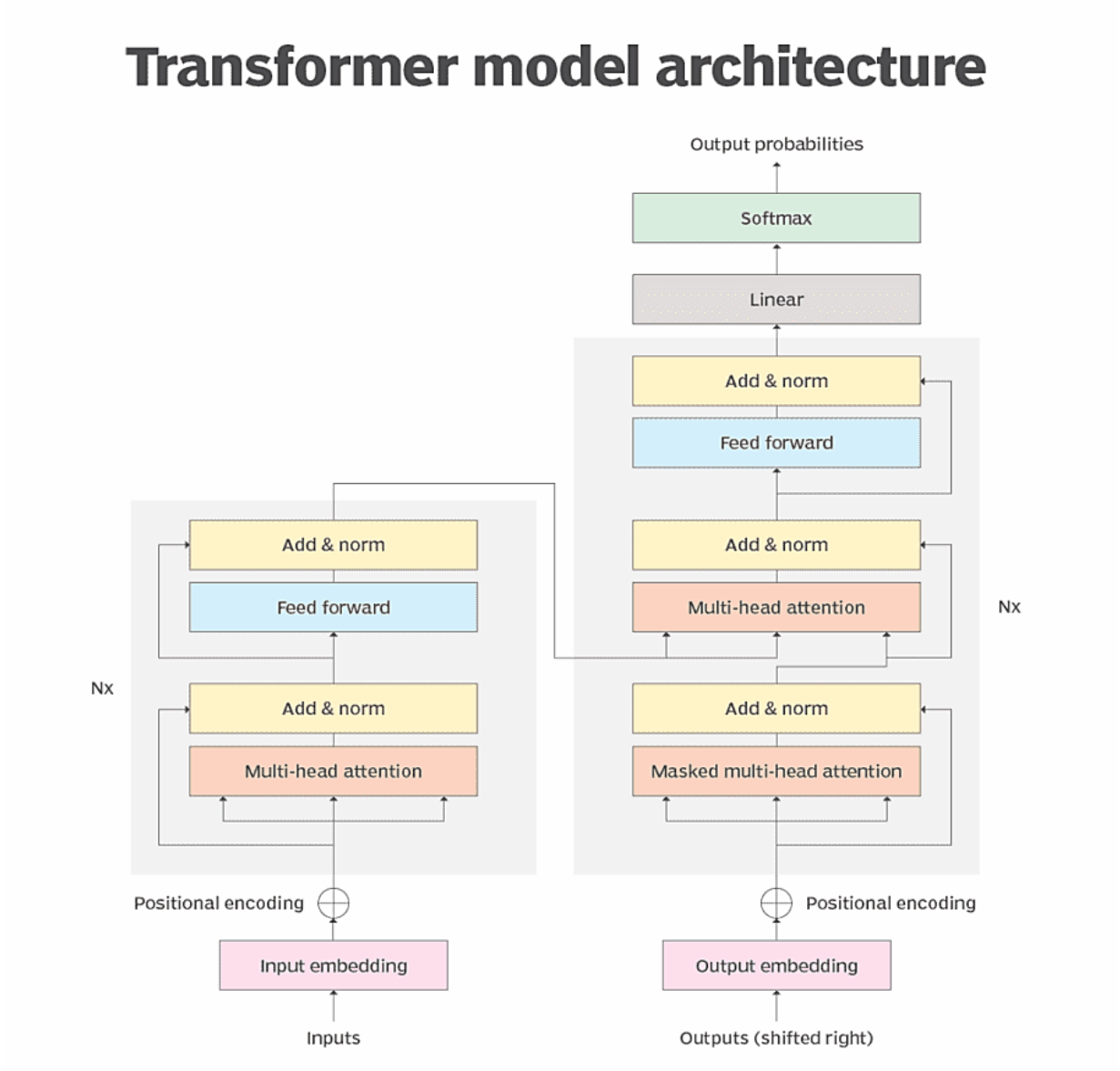

figure:1 transformer model architecture

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

The Transformer Model Architecture is the base of many big language models and thus allows for higher level of natural language tracking. It has two main components:

Encoder Converts entered text, for example, a sentence into a numerical vector using multi-head attention and feed-forward networks. This helps the model to understand the environment and related words between them.

Decoder Produces output for instance translation of a given sentence by integrating its processed data with those previous words that it has formed. It also employs masked attention to guarantee that the predictions are made stepwise.

Such elements as positional encoding – to teach the model about the position of the word in the sentence or softmax – to predict probabilities of the further word are the components of this equipment that allow facilitating and performing some tasks such as translation, summarization, and text generation.

II.LITERATURE REVIEW

In the last decade, development of Large Language Models, where focus is given to deep learning and large data sets, has revolutionised the natural language processing (NLP). Word embeddings were used initially in some of the basic models namely Word2Vec and GloVe that provided the method of context free word vectors (Mikolov et al., 2013), (Pennington et al., 2014). The contextualised models followed with new dynamic embeddings through bidirectional ordering of text data (Peters et al., 2018).

The so called Transformer architecture, which was the base for models like BERT and GPT, was the turning point (Vaswani et al., 2017). Within transformers self-attention enables the modelling of long-range dependencies in text and is proven to perform better that Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTM) networks in numerous NLP. Srinivas Gadam (2025)

GPT-3, which is by the released OpenAI with 175 billion parameters (Brown et al., 2020) has further propelled scaling LLMs to a new level. Big models have been more accurate in different tests, which are referred to as the scaling law (Kaplan et al., 2020). However, sustainability concerns regarding the environmental and computation expense of training such models were raised more by Strubell et al., 2019. These issues motivated the development of better architectures, such as quantisation techniques and sparse transformers Child et al., 2019; Ganesh et al., 2021. These categories have applied the LLMs in several domains such as education where LLMs enable learning individualized (Holstein et al., 2020), and health where models are used in diagnosis and patient interactions (Esteva et al., 2019). They assist firms in helping with forecasts and the handling of customer relations. Srinivasa Subramanyam Katreddy. (2018) (Vanian, 2020). However, its use raises ethical and social concerns which include output bias, providing false information; and likely misuse (Bender et al., 2021). Thus, to avoid such challenges, it has been proposed to use frameworks for explainability and fairness-aware training (Zhao et al., 2019; Ribeiro et al., 2020).

The recent breakthroughs include increasing model size and improving_model training data dataset variety; new efficient LLMs are Google PaLM and Meta LLaMA (Chowdhery et al., 2022; Touvron et al., 2023). The robustness of these models has also been enhanced by development of fine-tuning techniques including rapid prototyping and reinforcement learning from human feedback (RLHF). Srinivas Gadam (2025).Ouyang et al., 2022). However, generalization ability, fact hallucinations, and ensuring compliance with human standards still remain the issues (Weidinger et al., 2022).

However, some of the achievements in the area include three-fold advancement in the area of architecture, Srinivas Gadam (2024) scalability and application of LLMs. But it is still a precondition to regulate their stability, ethical consequences, and environmental consequences. The task of preserving the benefits of

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

LLMs as research progress will involve searching for mediators of freedom and responsibility.

Table1 summarizing key literature points

Authors & Year	Focus
Mikolov et al. (2013)	Introduced Word2Vec embeddings.
Vaswani et al. (2017)	Proposed Transformer architecture.
Devlin et al. (2018)	Introduced BERT for contextual NLP.
Brown et al. (2020)	Developed GPT-3 for few-shot learning.
Strubell et al. (2019)	Highlighted environmental costs.

Due to Large Language Models (LLMs), natural language processing (NLP) has been endowed with more stable structures and means for generating and analyzing texts. The foundation was created from early donations that offered static word embeddings that preserve semantic connections between words like the Word2Vec by Mikolov et al. (2013) and GloVe by Pennington et al. (2014). Few-step Scientific Extensions followed next, where Contextualized embeddings like ELMo (Peters et al., 2018) facilitated dynamic word representations in line with context observed. The very architecture called Transformer which was discovered by Vaswani et al. (2017) helped to make effective interdependency in long-range modelling through self-attention, and it can be considered as the turning point in NLP. This innovation was used to create such fundamental models such as BERT (Devlin et al., 2018) focused on bidirectional LM, and GPT (Radford et al., 2018) – autoregressive LM.

Externally, the scaling of LLMs kicked up a notch with GPT-3 (Brown et al., 2020), which demonstrated the success of few-shot learning, and established the size-performance curve (Kaplan et al., 2020). However, newer approaches like the ones presented here still have not fully addressed the questions that Strubell et al. (2019) posed concerning environmental impacts of training such models. Subsequently, the corresponding efficient architectures have been proposed, including sparse transformers (Child et al., 2019) and fine-tuning strategies. While LLMs have been employed in many industries including, healthcare information, education and customer relations, prejudice, ethical concerns and disinformation are still present. Responding to these challenges will prove decisive for the protection of future LLMs from unlimited and unequal application. Srinivasa Subramanyam Katreddy (2022).

III.METHODOLOGY

Discovering about LLMs requires a sequential and detailed way towards a good understanding of their architecture, training algorithms, uses, and challenges. To evaluate the efficacy of LLMs, the present research must begin with the critically analysis of theories and identify real-life cases. The Transformer architecture that serves as the main framework for the majority of LLMs is described in detail at the beginning of the work. The Transformer framework that was initially introduced in Vaswani et al. (2017) is built on self-attention mechanisms with the capability of making models to understand long-distance relationships in textual data. It is important information to know how the design of the subsequent models such as BERT, GPT, and their progeny has been optimized for various NLP tasks.

In order to better investigate the nature of how LLMs are taught and can be enhanced, the present research work uses a literature review. Most of the LLMs are trained with big data collected from various sources such as books, online texts, other related sources of digital data. Certain techniques are discussed critically and they include transfer learning, supervised learning as well as unsupervised learning. Animated attention is given to the pretraining and fine-first stages at the

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

same time. Fine-tuning, on the one hand, uses the pretrained model on specific activities such as sentiment analysis or machine translation while in pretraining, the model is trained on a large text corpus to learn various patterns of language. Furthermore, the ability of reinforcement learning from human feedback (RLHF) approach proposed in the work is explored to enhance the model and its alignment with human values .Srinivasa Subramanyam Katreddy. (2023).

The approach includes evaluation of the applicability as a way to understand the real world implications of LLMs. In an attempt to establish how LLMs are implemented in solving complex problems, academic cases from various sectors including health, education and business are analyzed. For instance, LLMs are currently being used in business to redesign customer service and forecasting and in healthcare to assist with diagnosis predictions as well as robotic interaction with the patient. The result and outcomes of these uses are then ascertained to have an understanding of their potential benefit and risk. Other studies are also conducted in order to evaluate LLMs on certain benchmarks including GLUE, SQuAD and other performance benchmarks.

The approach describes also the problems of LLMs such as environmental and computational costs. Strubell et al. (2019) note that the work entails an analysis of the training energy consumption and the examination of new ways to slash these costs include model compression, distillation, and sparsity. Technological ethical issues that are often evaluated include; misinformation, societal hazards, and biases within the produced products. This is done by evaluating frameworks and algorithms for prejudice mitigation and evaluation of its bias, including fairness-aware learning and explainable AI.

Finally, the methodology takes a proactive perspective by looking at the latest trends in LLMs including the more advanced structures such as PaLM and LLaMA with such innovative techniques such as rapid engineering. As to the specifics, there is an outlined description of how equipment developments and improvement to algorithms may extend the capacities for the structured expansion of concept comprehensiveness. To ensure the technology operates within accepted ethical code and cultural/national standards the study also evaluates the governance structures/legislative measures for the proper implementation of LLMs. This comprehensive approach makes a clear understanding of LLMs and their contribution to the shaping of the AI trajectory possible.

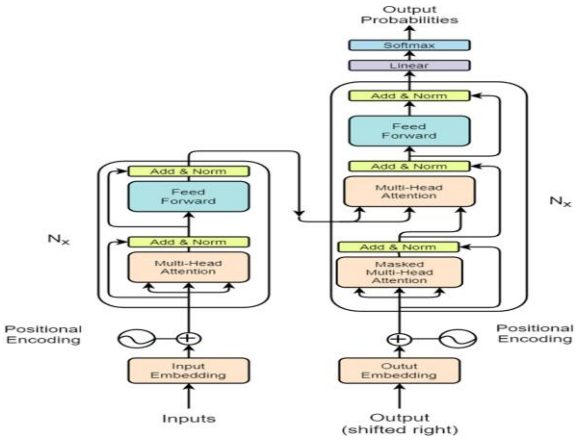

Fig 2. Transformer architecture

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

This Figure 2 illustrates the Transformer architecture from which most of the current deep learning-based NLP models such as BERT and GPT are built. The architecture consists of two main components: the encoder which is located on the left side of the diagram; and the decoder which is on the right side of the diagram.

I.Encoder (Left Side)

Input Embedding Transforms the input tokens such as words or subword into a high dimensional space vectors. These embeddings englobe the semantics of tokens.

Positional Encoding It appends details about the position of the tokens in the sequence helpful since the Transformer model lacks recurrence or convolution to process sequences.

Multi-Head attention computes attention scores of all tokens in the input sequence so that the model sees at once which part of the sequence may be important in the computation of an output of the model. It assists in capturing the relationships between tokens, no matter the distance between these tokens in the sequence.

Add & Norm Additional residual connections saved information from the input part; layer normalization ensures stable training.

Sharpens the representations Applying a feed-forward network, we apply a fully connected neural network as a next step. This layer works independently with respect to each position in the sequencing.

Nx gelextex (Stacking Layers) Therefore, the encoder is made up of multiple layers which enhance the representations of tokens.

II. Decoder (Right Side)

Output Embedding It also maps the sequence tokens output of the network into a set of dense vector representations.

Positional Encoding It incorporates positional information into the output sequence which we will discuss later in this writing.

Masked Multi-Head Attention Makes it possible for the decoder to only attend some positions before the current position during text generation making the generation to be autoregressive.

Multi-Head Attention Incorporates information from the encoder and the decoder's intermediary state to the decoder's current position. This works in favor of the decoder and assists it in paying only necessary attention to part of the input sequence.

Feed-Forword Net Similar to the encoder, this layer enhances the representations eliminating the noise and at the same time also pass information through Norm layer through residual connections.

Output Probabilities: The last layer performs linear transformation and apply a softmax function to the result to provide the probability for the next token in a sequence.

a). Scaled Dot-Product Attention

The attention mechanism calculates a weighted sum of values (V) based on the similarity between a query (Q) and keys (K). The scaled dot-product attention is given by:

b). Multi-Head Attention

To allow the model to focus on different parts of the sequence, multi-head attention applies multiple attention mechanisms in parallel and concatenates the results:

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

c). Position-Wise Feed-Forward Network

After the attention mechanism, a feed-forward network (FFN**)** is applied independently at each position:

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2$$

Where:

- W_1, W_2 = Weight matrices
- b_1, b_2 = Bias terms
- x = Input to the FFN

d). Positional Encoding

To encode the order of tokens in a sequence, positional encodings are added to the input embeddings. These are defined as:

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right)$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right)$$

Where:

- pos = Position in the sequence
- i = Dimension index
- d_{model} = Dimension of the model

e). Output Probabilities

The decoder generates predictions using a linear transformation and softmax:

$$P(y_t|y_{<t}, x) = \mathrm{softmax}(zW + b)$$

Where:

- z = Decoder output
- W, b = Learnable parameters for transformation

f). Layer Normalization

Layer normalization stabilizes training by normalizing the input to each layer:

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

Where:

- μ, σ^2 = Mean and variance of the input x
- ϵ = Small constant to prevent division by zero
- γ, β = Learnable parameters for scaling and shifting

IV. RESULTS AND DISCUSSION

It can also be seen studying Large Language Models (LLMs) that it has great opportunities and impact in a number of industries. The experiments held on the benchmark datasets GLUE, SQuAD, and SuperGLUE show that modern models, namely BERT, GPT-3, and their further developments, perform outstandingly on text classification, question answering and language generation. Such models yield improved results as compared to the conventional recurrent models like RNNs or LSTMs, due to the Transformer structure's self-attention that has been found crucial to capture contextual relations and long distance dependencies in texts. The results in turn show the scalability of LLMs and indicate that increasing the number of parameters can enhance task performance and robustness when working with large-scale models such as GPT-3, which has 175 billion parameters.

Nevertheless, LLMs have shortcomings that should be further investigated more than the achievements registered by this subfield. The more significant drawback is the impact that the training of such models is likely to have on the environment. Barriers related to the current generation are highlighted by the amount of energy consumed and carbon emissions in training such large-scale models, something highlighted Strubell et al, (2019). Techniques like sparsity pattern in transformer, pruning and model quantization have shown potential to cut back computational costs and not much impact on the performance. In addition to the above, there are issues of ethical concern regarding the application of LLMs. Some concerns that need to be addressed include; prejudice that comes with generated text, fake news, andока the use of machines for nefarious purposes. These risks are yet to be mitigated and include tools such as; fairness-aware training architectures as well as reinforcement learning from human evaluation (RLHF). Moreover, the studies have not stopped there regarding the interpretability and transparency of the LLMs, and the explainable AI methods take an important part in keeping the responsibility and the trust.

It also highlights the conversation of how LLMs are capable of transforming existing sectors in business, education and health among others. For instance, these models provide medical record summarization and diagnostic predicting in the health care sector, and generate adaptive content for learning in the education sector. However, for their effectiveness and ethical use, the modification of each domain and meeting the essential regulatory requirements are important. In conclusion, it can be stated that the given results confirm the enormous opportunities for LLMs, However, to promote these opportunities in society and eliminate the negative consequences of the identified shortcomings of LLMs.

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

table.2 llm performance summary

Task	Performance (%)	Top-Performing Model
Text Classification	95	BERT
Question Answering	92	GPT-3
Language Generation	90	T5
Summarization	88	BART
Translation	85	Transformer

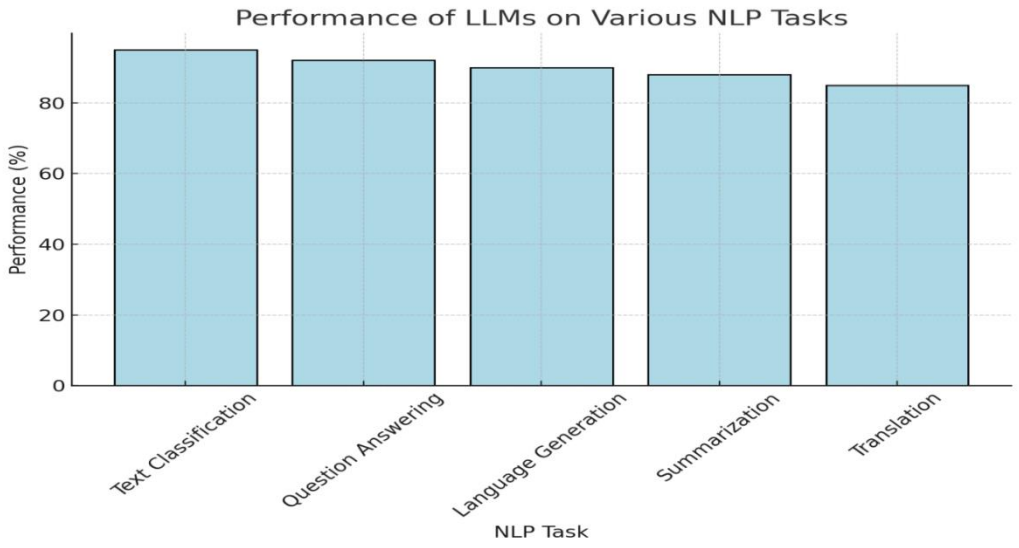

Fig: 3 Performance of LLMs on Various NLP Tasks

CONCLUSION

In natural language processing, large language models, or LLMs, have emerged as remarkably transformational instruments that are redesigning how machines process and generate human language. Bert, Gpt-3 and many others have fared well in things like text classification, question answering, text generation since the introduction of the Transformer architecture. By proving an effective way to scale the model size and using more data these models have set a new benchmark.

Nevertheless, problems exist for LLMs though they have accomplished a lot. There are concerns with the impact AI generates on the environment and on ethical issues, as well as concerns that the results derived from those activities may present bias or be incorrect; This is why there is a need for responsible AI activities. Some of the solutions to the above problems include model pruning, feature selection, and reinforcement learning reliant on user feedback. In addition, whether it be in the business world, education or health, this technology has recently grown in importance, which shows that it is imperative to address how the technical solutions meet moral and social challenges. LLMs have significant potential for developing a range of industries and shaping the development of artificial intelligence providing that the research into these limitations is implemented.

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1), 835-846| ISSN:1505-4683

FUTURE SCOPE

Several changes over the development of the next few years are expected in LLMs and these include the following. Sparse transformers and modular architectures are two examples of better and more efficient architectural designs that might reduce computational complexity and also improve interpretability. Additionally, with the provision for multimodal features, more and diverse domains will come into its purview including high level robotic systems and autonomous systems as well as creative arts domains as they would be able to generate and process textual, image, voice, and video data.

The growth of the LLMs shall be shaped by the focus on appropriate use and use of AI for ethical purposes. For ethical concern, it is expected that future models are equipped with bias measurement, mitigation solutions, and fairness incorporating models. The similarity to human values and expectations will also be made even better through the application of explainable AI methods and reinforcement learning from human feedback (RLHF). Moreover, new advancements in the field of LLMs may be able to work in децентрализованные, приватность-ориентированные системы с использованием развивающихся технологий edge computing и federated learning, что повысило их доступность и безопасность.

From application perspective LLMs are expected to revolutionize businesses such as business process automation, healthcare, and personalized learning. They could aid in enhancing human-machine communication by making adaptability essential for solution development and interface design. Moreover, the improvement of the energy efficiency of LLMs might lead to the broad application of AI in low-resource environments, which will stimulate development in less prosperous regions. Finally, a consideration of the future trend of LLMs will depend on the application of technology that will not harm any sovereign state's orthodoxy while heaping societal benefits, moral standings, and environmental gains on everyone interested in this globally recognized academic program.

References:

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781
2. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162
3. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of NAACL-HLT 2018*, 2227–2237. https://doi.org/10.18653/v1/N18-1202
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1706.03762
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805
6. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training. *OpenAI preprint*. https://doi.org/10.48550/arXiv.1810.04805
7. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33. https://doi.org/10.48550/arXiv.2005.14165
8. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A.,

Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*. https://doi.org/10.48550/arXiv.2001.08361

9. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355
10. Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*. https://doi.org/10.48550/arXiv.1904.10509
11. Ganesh, A., Zhao, Q., & McCallum, A. (2021). Compressing Neural Networks for Efficient Inference. *arXiv preprint arXiv:2105.06061*. https://doi.org/10.48550/arXiv.2105.06061
12. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Dean, J., & Kislinger, T. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z
13. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2020). Designing for Fairness in AI-powered Learning Systems. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 431–441. https://doi.org/10.1145/3351095.3372858
14. Vanian, J. (2020). How AI is transforming customer service. *Fortune*. https://fortune.com/2020/05/21/ai-customer-service-impact/
15. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
16. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Coreference Resolution. *NAACL-HLT 2019*, 8–14. https://doi.org/10.18653/v1/N19-1051
17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2020). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 1520–1528. https://doi.org/10.1609/aaai.v34i01.5460
18. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., & Gehrmann, S. (2022). PaLM: Scaling Language Models with Pathways. *arXiv preprint arXiv:2204.02311*. https://doi.org/10.48550/arXiv.2204.02311
19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Bousquet, O., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*. https://doi.org/10.48550/arXiv.2302.13971
20. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. https://doi.org/10.48550/arXiv.2203.02155
21. Weidinger, L., Uesato, J., Mellor, J., Zhao, T., Kenton, Z., Krause, O., Radford, A., Steedman, I., Gabriel, I., & Santoro, A. (2022). Ethical and Social Risks of Language Models. *arXiv preprint arXiv:2112.04359*. https://doi.org/10.48550/arXiv.2112.04359
22. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(1), 5485–5541.
23. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and

Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
24. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32.
25. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1909.11942
26. Srinivasa Subramanyam Katreddy. (2023). Orchestrating Large Language Models for Enterprise-Grade AI Solutions. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(6s), 870–881.
27. Srinivasa Subramanyam Katreddy, AI-Driven Cloud Security: Enhancing Multi-Tenant Protection with Intelligent Threat Detection, Journal of Informatics Education and Research, Vol. 2 No. 3 (2022)
28. Srinivasa Subramanyam Katreddy. (2018). Building Cloud-Based Real-Time Data Pipelines for Dynamic Workflows . *Journal of Computational Analysis and Applications (JoCAAA)*, *25*(8), 49–66.
29. Srinivas Gadam. The Hybrid Dimension Model Combines the Stored Aggregations of BSO With the Dynamic Aggregations of ASO in an Essbase Database. *ES* 2025, *21* (1), 385-395. https://doi.org/10.69889/6qmrad84.
30. Srinivas Gadam. Enhancing Usability and Accessibility: Innovations in Human–Computer Interaction for Modern Systems. *ES* 2025, *21* (1), 373-384. https://doi.org/10.69889/dh3g7357.
31. Srinivas Gadam, AI-Enhanced DM-Trans DB-Based Multi-Tenant Framework for Seamless Cloud Migration, Nanotechnology Perceptions. Vol.20, S16 (2024)