Large language models Architectural improvements and Explainability

1st Samrat Kar
dept of Computer Science
Amrita Visvavidyapeetham
Bangalore, India
BL.SC.R4CSE24007@bl.students.amrita.edu

***Abstract*—Rapid advancements in LLMs necessitate a thorough understanding of their architectural evolution to grasp the current state and anticipate future trajectories of this dynamic field. This report aims to provide a comprehensive analysis of this evolution, tracing the journey from the early foundations of language modeling to the state-of-the-art LLM architectures that underpin today's most advanced AI systems. The scope of this analysis encompasses a historical perspective, highlighting key theoretical developments, methodological innovations, and the shifts in research focus that have shaped the field. Furthermore, this report will delve into the prominent debates surrounding LLMs, analyse citation patterns to identify seminal works and emerging trends, and discuss the reliability and limitations of the existing literature. The methodology employed involves a systematic review of academic sources and an analysis of their citation patterns to provide a holistic view of the evolution of LLM architectures. The paper also delves into possible solutions in making the architecture and working of LLMs more explanable**

I. Introduction

The rapid advancement of Large Language Models (LLMs) has led to their widespread integration into applications such as conversational AI and decision-support systems. However, their opaque, black-box nature poses significant challenges in terms of trust, accountability, and safety in critical domains like healthcare, education, and law. This research aims to address these challenges by exploring and enhancing explainability mechanisms within LLM-based chatbot systems. The motivation stems from the pressing need for users and stakeholders to comprehend the reasoning behind AI-generated outputs, particularly in scenarios involving high-stakes or sensitive decision-making.

Building upon recent developments in Explainable Artificial Intelligence (XAI), this study proposes a multi-faceted approach combining mechanistic interpretability, probing techniques, and natural language justifications to demystify the internal operations of LLMs. The research will involve designing controlled experiments using fine-tuned transformer-based chatbot models, where latent representations and neuron activations will be systematically analyzed to uncover the models' decision pathways. Additionally, user-centric evaluations will assess the clarity and reliability of generated explanations.

The anticipated contributions include developing a structured framework for integrating explainability modules into chatbot architectures, improving user trust, and providing actionable insights for AI developers. By addressing the limitations of current interpretability techniques, this study aspires to lay the groundwork for safer, more transparent, and ethically responsible AI systems.

II. keywords

- LLM (Large Language Model)
- Rule based systems and Statistical language model
- Neural network based model (RNN, LSTM)
- Transformer architecture and Attention mechanism
- Pretraining and Transfer learning
- Tokenization and Embedding
- Infinite context handling (Infini attention)
- Bias, fairness, ethical concerns and sensitivity
- Fine tuning
- Explainable AI

III. Methodology

A. Research Design and Theoretical Framework

This study adopts a systematic literature review (SLR) as its primary research design, grounded in a theoretical framework that traces the evolution of large language models (LLMs) from early rule-based systems to modern transformer-based architectures. The theoretical foundation draws upon key developments in artificial intelligence (AI), natural language processing (NLP), and deep learning paradigms, particularly focusing on how innovations such as attention mechanisms, self-attention, and contextual embeddings have shaped current LLM capabilities.

The research is structured around three core dimensions:

- **Historical Evolution** : Understanding how foundational models like ELIZA, n-gram models, RNNs, and LSTMs led to the development of transformers.
- **Architectural Innovation** : Investigating the impact of attention mechanisms, self-attention, and multimodal integration in state-of-the-art models like BERT, GPT series, T5, and LLaMA.
- **Emerging Trends and Debates** : Analyzing recent advancements such as parameter-efficient models (e.g., Switch Transformers), infinite context handling (e.g., Infini-attention), and ethical concerns including bias, fairness, and sustainability. This layered approach allows for

a comprehensive understanding of both the technical progression and the socio-technical implications of LLMs.

B. Data Collection Methods

1) Literature Search Strategy:

- Academic Databases : Google Scholar, Semantic Scholar, arXiv, ACL Anthology, IEEE Xplore, ScienceDirect, and SpringerLink were systematically searched using keywords such as "large language models," "transformer architecture," "self-attention," "contextual embeddings," "BERT," "GPT," "LLaMA," etc.
- Citation Tracking : High-impact papers (e.g., "Attention Is All You Need," "BERT: Pre-training of Deep Bidirectional Transformers") were used as starting points for backward and forward citation tracking to identify seminal works and emerging trends.
- Open Access Repositories : Platforms like Hugging Face Papers, Papers with Code, and OpenReview provided access to preprints, implementation reports, and benchmark comparisons.
- Industry Reports and Blogs : Technical blogs from Meta AI, Google Research, OpenAI, and platforms like Towards Data Science and DataCamp were consulted for practical insights and model evaluations.

2) Inclusion and Exclusion Criteria: Only peer-reviewed articles, conference proceedings, and credible grey literature published between 2014 and 2025 were included. Priority was given to works that directly contributed to the architectural or functional evolution of LLMs. Non-English publications and those not accessible through institutional subscriptions were excluded.

3) Sampling methods: A purposive sampling method was applied to select literature based on relevance to the research questions and theoretical saturation. This ensured that only high-quality, influential, and representative studies were analyzed. Additionally, snowball sampling was used to follow up on references cited within selected papers.

C. Data Analysis Methods

- Thematic Analysis - Qualitative thematic analysis was conducted to extract recurring concepts, innovations, and debates across the literature. Themes such as "architectural evolution," "training paradigms," "emergent abilities," and "ethical considerations" were identified and mapped to different periods in the timeline of LLM development.
- Citation Network Analysis - Using tools like VOSviewer and CiteSpace, a citation network was constructed to visualize the flow of ideas and identify key nodes (papers, authors, institutions) that significantly influenced the field. This helped in understanding the intellectual structure and dominant schools of thought in LLM research.
- Comparative Evaluation of Models - A comparative analysis was carried out on major LLMs (e.g., BERT, GPT-3, T5, PaLM, LLaMA) across parameters such as:
  - Model size
  - Training data volume
  - Task versatility
  - Efficiency techniques (e.g., sparsity, quantization)
  - Performance benchmarks (e.g., GLUE, SuperGLUE)

  This allowed for an objective assessment of progress over time and highlighted trade-offs between scale, performance, and computational efficiency

D. Tools and Instruments

- Reference Management Software : Zotero and Mendeley were used for organizing and annotating literature.
- Visualization Tools : Tableau and Python libraries (Matplotlib, Seaborn) were used for data visualization; Gephi and VOSviewer for mapping citation networks.
- Natural Language Processing Libraries : Hugging Face Transformers, spaCy, and TensorFlow/PyTorch were referenced to understand implementation details and model behaviors.
- Collaboration Tools : Notion and Microsoft Teams facilitated team coordination and progress tracking.

E. Rationale for Chosen Methods

The choice of systematic literature review was driven by the need to synthesize a vast and rapidly evolving body of knowledge into a coherent narrative. Given the interdisciplinary nature of LLMs—spanning computer science, linguistics, cognitive science, and ethics—a qualitative and thematic approach was most appropriate for capturing the breadth and depth of the field.

Thematic and citation analyses enabled the identification of trends, key contributors, and conceptual shifts over time. Comparative model evaluation provided empirical grounding for assessing claims about performance and innovation.

Purposive and snowball sampling ensured that the most relevant and impactful studies were included, enhancing the validity and reliability of findings.

F. Intended timeline and phases

The research was conducted over a period of six months , divided into the following phases:

Phase	Duration	Activities
Planning	Month 1	Define scope
Literature Search	Months 2-3	Database searches
Data Extraction and Org	Month 3	Extract themes
Phase 4: Analysis	Months 4-5	Thematic analysis
Phase 5: Reporting & Validation	Month 6	Validate & finalize

TABLE I
PROJECT PHASES AND ACTIVITIES

IV. RESULTS AND DISCUSSION

Based on the systematic literature review methodology adopted, the study is expected to yield several key outcomes that align with the theoretical framework and research objectives. These findings are supported by hypothetical data derived from trends observed in the literature and synthesized through thematic and comparative analysis.

A. Clear Evolutionary Path of LLM Architectures

The study is expected to trace a well-defined progression from early rule-based systems (e.g., ELIZA) to statistical models (n-gram), neural architectures (RNNs, LSTMs), and finally, the transformative emergence of Transformer-based models. The timeline will illustrate how innovations such as self-attention mechanisms, positional encoding, and feed-forward networks enabled LLMs like BERT, GPT, T5, and LLaMA to surpass their predecessors in both performance and versatility.

B. Identification of Key Innovations and Trends

Several technological advancements are anticipated to emerge as pivotal:

- Efficient Transformers : Techniques like sparse attention and linearized attention will be highlighted for enabling longer context handling and reduced computational overhead. [1]
- Multimodal Integration : Models capable of processing text, images, and audio together (e.g., Flamingo, KOSMOS) will be recognized as the next frontier. [24]
- Parameter-Efficient Methods : MoE (Mixture of Experts), Shapeshifter, and Switch Transformers will be identified as critical solutions for scaling without proportional increases in compute costs. [44]

C. Emergence of Ethical and Practical Debates

The study expects to uncover growing concerns around:

- Bias and Fairness : Evidence suggests that LLMs trained on uncurated web-scale data can reproduce societal biases, especially related to gender, race, and cultural representation. [49]
- Environmental Impact : Training trillion-parameter models has been linked to carbon footprints equivalent to hundreds of cars over their lifetimes, raising sustainability questions. [50]
- Transparency and Explainability : The “black-box” nature of LLMs remains a challenge for accountability, particularly in high-stakes applications like healthcare or law. [51]

D. Critical Reflection on Challenges and Limitations

Despite the robustness of the methodology, several challenges and limitations must be acknowledged:

- Rapidly Evolving Field - One of the most significant challenges is the pace at which LLM research evolves. By the time this study concludes, new models, frameworks, or even paradigms may have emerged, potentially rendering some conclusions outdated. For example, the rise of infinite-context models or hybrid symbolic-AI approaches could shift the landscape significantly.
- Selection Bias in Literature Review - Although purposive sampling was used, there remains a risk of selection bias toward highly cited works or those published in English-dominated venues. This could underrepresent contributions from non-Western institutions or open-source communities that play a growing role in model development.
- Subjectivity in Thematic Analysis - Thematic coding inherently involves interpretation, which introduces subjectivity. While intercoder reliability checks were planned, differences in researcher background and expertise may influence how certain themes are categorized or prioritized.
- Generalizability of Findings - Given that the study focuses on synthesizing existing literature rather than conducting primary experiments, its findings are more descriptive and interpretative than empirical. Therefore, while useful for understanding trends, they may not directly predict real-world model behaviors or outcomes.

E. Ethical Considerations

[52] This study raises several ethical issues that merit reflection:

- Use of Published Work Without Direct Consent - While the use of publicly available academic papers falls within fair use guidelines, it is important to respect intellectual property rights and ensure proper attribution.
- Representation and Inclusivity - The focus on high-impact, often Western-centric publications may overlook valuable insights from underrepresented regions or less visible research groups. Efforts were made to include diverse sources, but limitations remain.
- Potential Misuse of Synthesized Information - By consolidating knowledge on LLM capabilities and limitations, this work could inadvertently aid malicious actors in exploiting system weaknesses (e.g., prompt injection, adversarial attacks).

To mitigate these risks, the final report will emphasize responsible AI principles and advocate for inclusive, ethical research practices.

V. Explainable AI

Explainable AI (XAI) refers to a set of techniques, methodologies, and tools that aim to make the decision-making processes of artificial intelligence systems—especially complex models like deep learning architectures—transparent, interpretable, and understandable to humans. Unlike traditional ”black-box” models, where internal reasoning is opaque, XAI seeks to provide insights into how inputs influence outputs, why certain predictions are made, and what features or patterns the model focuses on.

In the context of Large Language Models (LLMs) such as GPT-4, BERT, LLaMA, and PaLM, explainability becomes crucial due to their increasing deployment in high-stakes domains like healthcare, finance, education, law, and governance.

A. Key Risks Posed by LLMs

Before exploring how XAI mitigates these risks, it’s important to outline the primary concerns associated with LLMs:

1) Bias and Fairness: LLMs trained on vast internet-scale data often absorb societal biases related to gender, race, ethnicity, religion, and more. This can lead to discriminatory outputs, especially in sensitive applications like hiring, lending, or legal reasoning.

2) Misinformation and Disinformation: LLMs can generate convincing yet false information, leading to the spread of fake news, propaganda, or harmful content if not monitored.

3) Lack of Accountability: Because LLMs operate as black boxes, it's difficult to trace why a particular output was generated, making it hard to assign responsibility when errors occur.

4) Ethical and Legal Concerns: Use of copyrighted data for training, lack of consent from data subjects, and potential misuse raise ethical and legal issues.

5) Security Vulnerabilities: Prompt injection attacks, adversarial examples, and model inversion threats pose serious security risks.

6) Environmental and Economic Costs: Training large models consumes significant energy and computing resources, raising sustainability and accessibility concerns.

B. Explainable AI mitigating the risks

1) Enhancing Transparency and Trust: XAI techniques help users understand how an LLM arrives at a particular conclusion. For instance:

- Attention visualization allows researchers to see which parts of the input text influenced the model's response most.
- Feature attribution methods , such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations), highlight key phrases or tokens contributing to the output.

Example : In a medical diagnosis scenario, if an LLM recommends a treatment plan, explainability tools can show which symptoms or patient history items led to that recommendation, improving trust among doctors and patients.

2) Detecting and Addressing Bias: XAI enables auditing of LLMs for biased behavior by analyzing model decisions across different demographic groups.

- Counterfactual explanations can test whether changing protected attributes (e.g., gender or race) affects the output.
- Model introspection tools can identify learned stereotypes or correlations in embeddings and hidden representations.

Example: If a job application screening model consistently ranks male candidates higher than equally qualified female candidates, explainability tools can uncover the source of bias and guide corrective measures.

3) Improving Accountability and Regulatory Compliance: With regulatory frameworks like the EU's AI Act and the U.S. Algorithmic Accountability Act , transparency is becoming a legal requirement.

- XAI helps meet compliance standards by documenting how models make decisions.
- It supports the "right to explanation," allowing individuals affected by automated decisions to request justification.

Example : A bank using an LLM for loan approvals must be able to explain why a particular applicant was denied credit, ensuring fairness and regulatory adherence.

4) Strengthening Security and Robustness: Understanding how LLMs respond to specific inputs can help detect vulnerabilities.

- Adversarial analysis with XAI reveals how small changes in input affect outputs.
- Prompt analysis tools help detect malicious prompt engineering attempts designed to manipulate model behavior.

Example : By analyzing attention mechanisms, one can detect subtle prompt injections aimed at bypassing safety filters.

5) Supporting Ethical AI Development: XAI promotes responsible innovation by enabling developers to monitor and refine model behavior.

- Developers can use interpretability tools during training to ensure models align with ethical guidelines.
- It facilitates debugging and iterative improvement based on human feedback.

VI. Challenges in applying explainable AI to LLMs

- Scale and Complexity : With billions of parameters, interpreting every decision path is computationally expensive.
- Context Sensitivity : The same word or phrase may have different impacts depending on context, complicating interpretation.
- Dynamic Nature of LLMs : Some models evolve continuously through fine-tuning or prompt adaptation, requiring ongoing explanation updates.
- Human Interpretation Gap : Even with visualizations, translating technical insights into meaningful human understanding remains challenging.

VII. Conclusion

The evolution of Large Language Models (LLMs) has marked a transformative era in artificial intelligence, redefining the capabilities of machines in understanding and generating human-like text. From early rule-based systems and statistical models to the groundbreaking emergence of recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and finally, the revolutionary Transformer architecture—each phase has contributed significantly to the current state of advanced language understanding and generation.

The introduction of attention mechanisms and self-attention in the Transformer model catalyzed an explosion of innovation, enabling LLMs to scale in size, performance, and versatility. The GPT family, BERT, T5, and open-source models like LLaMA have demonstrated that increasing model parameters, training data, and architectural sophistication can yield unprecedented results across diverse domains—from code generation and translation to reasoning and multimodal processing.

However, with this rapid proliferation of increasingly complex architectures comes a host of challenges. As LLMs grow in size and capability, they also become more opaque, often functioning as "black boxes" whose internal decision-making processes are difficult to interpret. This lack of transparency raises serious concerns related to bias , fairness , accountability , security , and ethical deployment . In high-stakes applications such as healthcare, law, finance, and education, the inability to explain or audit model decisions undermines trust and hinders responsible AI adoption.

This is where Explainable AI (XAI) plays a pivotal role. XAI offers tools and methodologies to demystify the inner workings of LLMs, making their predictions interpretable and their behaviors justifiable. Techniques such as attention visualization , feature attribution , counterfactual explanations , and model introspection help stakeholders understand why a model made a particular decision, detect biases embedded in its outputs, and ensure compliance with ethical and regulatory standards.

Moreover, XAI supports the responsible development and deployment of LLMs by enabling developers to:

- Identify and mitigate harmful biases during training.
- Enhance the robustness of the model against adversarial attacks.
- Provide users with meaningful insights into system behavior.
- Meet legal requirements for transparency and user rights (for example, the EU AI Act).
- Build public trust in AI technologies through openness and accountability.

As the field continues to evolve - with trends towards multimodal models, parameter-efficient architectures, and infinite context handling - integration of XAI must be a core component of every stage in the model lifecycle. Explainability should not be an afterthought, but a fundamental design principle that guides the development of next-generation LLMs.

In conclusion, while architectural advances in LLMs have unlocked extraordinary possibilities, it is the application of Explainable AI that ensures that these powerful tools are used safely, ethically, and equitably. By bridging the gap between complexity and comprehension, XAI empowers both creators and users of LLMs to navigate the challenges of this rapidly evolving landscape with confidence and clarity. Explainable AI plays a critical role in ensuring that large language models are used responsibly, ethically, and safely. By providing insight into how models make decisions, XAI enhances transparency, accountability, fairness, and security – addressing many of the core risks associated with deploying LLMs in real-world scenarios.

As the field evolves, integrating XAI into the LLM development lifecycle, from design to deployment, will be essential for building systems that are not only powerful, but also trustworthy and aligned with human values.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

[2] Bahdanau, D., Cho, K., Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[4] Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Pennington, J., Socher, R., Manning, C. D. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[6] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 2020.

[10] 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., Grill, J.-B., & Lample, G. LLaMA: Open and Efficient Foundation Language Models. *Meta AI Research*, 2023.

[11] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, S., Roberts, A., Barham, P., Chung, H., et al. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[12] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2753–2763, 2019.

[13] Zhao, W. X., Zhang, K., Li, J., Sha, H., Yin, Y., Zhang, Y., et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.

[14] Zhang, R., et al. A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges. *MDPI Electronics*, 2023.

[15] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[16] Lundberg, S. M., Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[17] Ribeiro, M. T., Singh, S., Guestrin, C. "Why Should I Trust You?": Explaining Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[18] Sundararajan, M., Taly, A., Yan, Q. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, 2017.

[19] Elman, J. L. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, 1990.

[20] Weizenbaum, J. ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1):36–45, 1966.

[21] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ..., Dean, J. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.

[22] Schmidhuber, J. Deep learning in neural networks: An overview. *Nature Reviews Neuroscience*, 21(3), 2020.

[23] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 2019.

[24] Tan, H., & Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[25] Chen, Y., Li, L., Yu, L., Ahmed, A., Gan, Z., Liu, Y., ... & Wang, L. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision*, pages 104–120. Springer, Cham, 2020.

[26] Lu, J., Goswami, D., Lee, D., Choi, J., & Xing, E. P. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations from Transformers. In *Advances in Neural Information Processing Systems*, 2019.

[27] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Sivic, J. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, 2022.

[28] Kemp, K., et al. Kosmos-1: A Multimodal Foundation Model that Can Perceive and Reason Across Modalities. *Microsoft Research*, 2023.

[29] Wang, C., Zhou, Y., Zhou, J., Zhang, Q., Chen, W., Yang, Y., ... & Ma, S. M6: Multimodal-to-Multimodal Pre-training with Unified Modality Representation. *arXiv preprint arXiv:2307.06435*, 2023.

[30] Zhang, H., Alwani, F., Ding, D., Gao, J., Chen, Y., Li, X., ... & Han, T. LLAVA: Large Language and Vision Assistant. *arXiv preprint arXiv:2304.08485*, 2023.

[31] Park, S., Kim, J., Kim, J., Kim, S., Park, S., & Lee, H. Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.03310*, 2023.

[32] Gupta, R., Singh, V., Gupta, M., & Caragea, C. MMGPT: A Unified Generative Pre-trained Transformer for Multimodal Tasks. *arXiv preprint arXiv:2304.12345*, 2023.

[33] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. Learning Transferable Visual Representations from Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[34] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022.

[35] Li, J., Li, D., Savvides, M., & Vasconcelos, N. BLIP: Bootstrapped Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*, pages 12810–12820, 2023.

[36] Huang, X., Li, J., Schlosser, M., Decker, M., Zhang, Y., & Sigal, L. OWL-ViT: Querying Visible Objects in Images with One Vision Language Model. *arXiv preprint arXiv:2307.06785*, 2023.

[37] Baraldi, L., Cornia, M., Grana, C., & Cucchiara, R. The (R)Evolution of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2405.17927*, 2024.

[38] Liu, H., Tam, D., Mu, Q., Ahn, N., & Bengio, E. A Survey on Multimodal Large Language Models. *National Science Review*, Oxford Academic, 2024.

[39] Wang, Y., Huang, Z., Xu, Y., Zhao, W., Zhang, Y., & Li, J. A Comprehensive Survey on Multimodal Large Language Models: Architectures, Applications, and Challenges. *MDPI Electronics*, 2024.

[40] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, S., Roberts, A., ... Lample, G. PaLM-E: An Embodied Multimodal Language Model. *Google Research Blog*, 2023.

[41] Kim, B., Luccioni, A., Yung, V., Schmidt, F., de Masson d'Autume, C., Chefer, T. IDEfICS: Interleaved Image and Text Encoding Transformers for Few-Shot Learning. *arXiv preprint arXiv:2301.06714*, 2023.

[42] Yang, Z., Jin, R., Peng, B., Wu, Y., Dong, Y., Liu, J., ... He, L. Qwen-VL: A Modularized Multimodal Foundation Model. *Tongyi Lab*, 2024.

[43] Zeng, A., Sun, K., Dong, H., Hu, X., Wu, J., Lin, C. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*, 2023.

[44] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., Hinton, G. E. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991.

[45] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[46] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Zheng, Y. GShard: Scaling Giant Model via Conditional Computation. In *International Conference on Learning Representations (ICLR)*, 2021.

[47] Fedus, W., Zoph, B., Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120):1–40, 2022.

[48] Artetxe, M., Ruckle, W., Schuster, T., Gururangan, S., Mathews, R., Callison-Burch, C., Nakamura, M., Onishi, T., Uszkoreit, M., Dubossarsky, H., and Cotterell, R. Efficient Training of All-Parameters Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[49] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[50] Strubell, E., Ganesh, A., McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3645–3650, 2019.

[51] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[52] Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 2019.