Transforming Imaging Tasks with Large Language Models: A Comprehensive Review

Preeti Kapoor
Department of Computer Science and Engineering
The NorthCap University
Gurugram
preeti19csd006@ncuindia.edu

Shaveta Arora
line 2: *Department of Computer Science and Engineering*
The NorthCap University
Gurugram
shavetaarora@ncuindia.edu

Abstract - Large language models (LLMs) have predominantly been applied in natural language processing tasks, but their capabilities are being explored greatly in imaging tasks. These tasks include data collection, caption generation, image segmentation and classification. The use of LLMs for these tasks is an innovative method which works with great efficacy. This review aims to provide an overview of the application of LLMs in the field of imaging, discussing their integration, benefits, and potential challenges. Additionally, since the medical imaging industry demands high dependability and strong artificial intelligence technology, a greater emphasis is kept on understanding the role of LLMs in medical image recognition and question-answering for report generation and medical diagnosis. Also, the paper discusses the limitations of these models and how to overcome them.
Keywords: Artificial intelligence, attention mechanism, image task, medical imaging, natural language processing

1. INTRODUCTION

Image classification is advancing at a great speed with the Artificial Intelligence (A.I.)(Bellemo et al., 2019) over the years. With the advancement and up gradation in the automation of the whole process of image processing helps to mitigate the effects effectively. The setup of using A.I. makes the whole process less laborious and time efficient. But the existing models lack broader vision and require huge amount of labelled good quality data, which is a challenge. The ability to work effectively with data diversity, generalizability and interactively, is highly required in medical image recognition. Other learning models such as Convolutional Neural Networks(Padmanayana & B.K, 2022), Recurrent Neural Networks (De La Cruz et al., 2024) or transformers(Nazih et al., 2023) are input dependent and are sensitive to outliers and noise.
For the analysis and interpretation of medical data, researchers and medical professionals are increasingly turning to Generative Pre-trained Transformers (GPT)(Yenduri et al., 2024) due to their exceptional language modelling capabilities. Large language models (LLMs)(Hadi et al., 2023) along with pre-trained modules show powerful learning abilities and can generate various text and image data types. These models have all the qualities which are required for improving the process of image classification. Though they are widely used in Natural language processing (Joshi, 1991)(NLP) but they can be of substantial application in image processing. Language models such as BERT(Devlin et al., 2018) and ELMo (Peters et al., 2018) were widely used in medical image analysis and language processing of automated health records. With recent development in chatbots like ChatGPT(Liu et al., 2023) and Gemini, the scholars are interested

in making it a valuable tool for health care areas like disease detection, report generation and drugs prescriptions, or fine-grained image classification or zero-shot learning(Matsuura et al., n.d.). These models process large amounts of data and provide experts a solution for different domain.

Research Objective: To learn how to use LLMs for different imaging tasks, such as multimodal analysis, annotation, and image classification. To explore the role of LLMs to improve the diagnostic accuracy and provide decision support systems in health sector.

1.1 LLMs for Imaging Task

- Improved Context Understanding: LLMs can understand and incorporate contextual information from text, refining the accuracy of imaging tasks(Rathje et al., 2024).
- Adaptability: LLMs are flexible models which make them suitable for other tasks beyond NLP.
- Enhanced Annotation Quality: LLMs produce high-quality annotations and image captions reducing the need of manual labelling.

1.2 Challenges and Limitations

- LLMs require a lot of computational power and resources, which may be a blockade for widespread adoption in imaging tasks.
- Integration Complexity: Integrating LLMs with traditional models is a challenging task.
- Huge well –labeled datasets are needed

However, a dearth of review papers concentrating on the use of language models in imaging analysis makes it challenging for researchers to acquire a thorough grasp of the field and its possible uses. Although there has been some progress in this direction, much more work is still required to fully utilize language models for medical imaging analysis. This paper focuses on highlighting the role of LLMs in enhancing the medical diagnosis process by merging vision algorithms with user enquiries for cooperative and user-specific outcomes. This study shows the significant impact of integrating the next frontier of A.I. driven models i.e LLMs to create sophisticated predictive tools.

The paper has IV sections, I section explains the LLMs and their architecture. II section discusses the role of LLMs in improving medical image recognition, obtaining better results, and better clinical interactions. And lastly, it observes the challenges language models face in this field. Section III discusses the findings of above section. The paper is concluded in section IV. We hope that this article will provide a thorough understanding of the use of LLMs in image processing to researchers and practitioners.

2. LARGE LANGUAGE MODELS

The research on NLP algorithms(Joshi, 1991) started back in 1950s, when the simple models were used for predicting the probability of the next word in a sentence. Over the time these models became more sophisticated and provided more precise language processing. These models require large amount of data for training and specialized algorithms. From text processing to zero-shot learning, LLMs are able to achieve great accuracy. Also, there is a development in multi-lingual and cross-lingual language model, to overcome the language barrier faced in heath sector. Table 1 gives an overview of the popular language models.

Table I. Popular language models

Sno.	Language model	Details	Application areas
1	N-grams(Manning & Schütze, 2005)	N-grams are a sequence of N words used in NLP. This method require large training data. The probability is calculated by counting the number of times a word occurs in a required sequence, divided by the number of times the word came before the expected word occurs in the body.	Auto-completion of sentences, auto-spell check, checking for grammar in a sentence and voice-based personal assistant bots
2	Recurrent Neural Networks (Feng et al., 2024)	Recurrent neural networks learn from sequential data and consider the order of observations i.e. incorporates the information from last hidden state and the current input. They suffer from the problem of vanishing gradients.	Speech recognition, language modeling, machine translation, speech recognition, and image captioning
3	Long Short-Term Memory(Staude meyer & Morris, 2019)	LSTM works like RNN but with the ability to store long term dependencies. Also, it addresses the problem of vanishing gradients. LSTM have a memory cell and several gates that control information flow into and out of the cell.	tasks like speech recognition, language translation, and image captioning
4	Transformers(Do sovitskiy et al., 2020)	Transformers are attention mechanism based models that calculate the feature representations of the input and output. It is a encoder-decoder system which includes self-attention module followed by a fully connected layer.	Sentiment analysis, translation, language modelling, image classification, segmentation, object detection, speech recognition, text-to-speech.

2.1 Architecture of LLMs

LLMs contain the key components which work together in a great manner, providing them great efficacy and accuracy. Figure 1 depicts the various components of LLMs. This section provides an explanation of how these modules work together to provide the desired outputs.

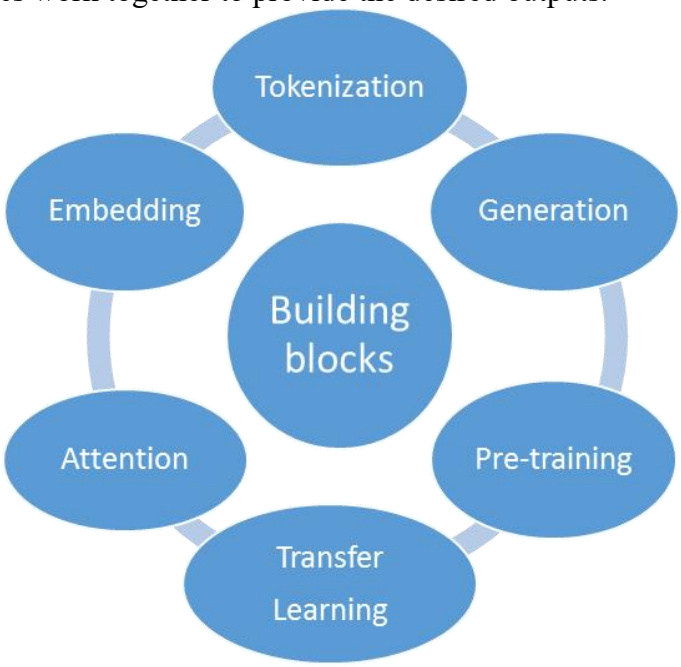

Figure 1. : Building blocks of LLMs

2.1.1 Tokenization

Tokenization is the process of breaking text into smaller units called tokens, such as Byte Pair Encoding (BPE) or WordPiece.

2.1.2 Embedding

Embeddings are continuous high-dimensional vector representations of tokens that capture semantic information. These representations are learnt by hard training.

2.1.3 Attention

The attention mechanism or the self-attention module analyzes the relationships between all the tokens.

2.1.4 Pre-training and transfer learning

These models are trained using vast amount of data repositories. The transfer learning helps to avoid retraining from the scratch.

2.1.5 Generation Capacity

They can produce relevant data across multiple domains making them suitable tool for several tasks.

For image classification**,** pre-training and fine-tuning, LLMS can also reduce the cost of data annotation and also help the situations where there is a limited quantity of annotated data. To take an understanding on the various available LLMs, in this paper a study is conducted to comprehend the evolution in the LLMs from year 2018 to 2024.

Table 2 is a comparative analysis of the LLMs, depicting the year, the specifications such as training data, number of parameters and underlying architecture and application areas. Also, it highlights the strengths of each model and their capabilities. This table targets to briefly explain the available LLMs.

Table 2. Comparative overview of various large language models

LLM	Developer	Year	Parameters (Billion)	Training Data	Architecture	Application	Features
BERT(Devlin et al., 2018)	Google	2018	0.34	BooksCorpus, English Wikipedia	Transformer	Question Answering (QA), NER, text classification	Bidirectional context, strong performance
RoBERTa(Liu et al., 2019)	Facebook AI	2019	0.36	Expanded BERT dataset	Transformer	QA, NER, text classification	Improved pre-training and performance
T5(Raffel et al., 2019)	Google	2019	11	C4 dataset	Transformer	Text-to-text tasks, translation, summarization	Unified text-to-text framework
XLNet (Yang et al., 2019)	Google/CMU	2019	0.34	BooksCorpus, English Wikipedia	Transformer-XL(Dai et al., 2019)	QA, text classification, text generation	Permutation-based training, bidirectional
Megatron(Shoeybi et al., 2019)	NVIDIA	2019	8.3	Web –scaled data	Transformer	Text generation, QA, translation	Effective parallel training
Turing-	Microsoft	2020	17	Diverse	Transforme	Text generation,	Large-scale

NLG(Smith et al., 2022)				internet text	r	QA, summarization	language generation
GPT-3(Singh et al., 2023)	OpenAI	2020	175	Web-scaled data	Transforme r	Text generation, QA, translation	Few-shot learning, adaptable capabilities
GShard (Naveed et al., 2023)	Google	2021	600	Whole internet	Transforme r	Text generation, QA, translation	Accessible to thousands of TPUs, efficient
EleutherAI GPT-Neo (Naveed et al., 2023)	Eleuther AI	2021	2.7, 6, 20	Diverse internet text	Transforme r	Text generation, QA, research	Open-source
Jurassic-1(Manning & Schütze, 2005)	AI21 Labs	2021	178	Entire web	Transforme r	Text generation, QA, translation	high quality output
LaMDA(Thoppilan et al., 2022)	Google	2021	Unknown (Billions)	Conversation al data, internet text	Transforme r	Dialogue, conversational AI	conversation al agents
PaLM(Peng et al., 2019)	Google	2022	540	Web-scaled data	Transforme r	Text generation, QA, summarization	Pathways framework, scalable
Bloom(Hoffmann et al., 2022)	BigScience	2022	176	Web-scaled data	Transforme r	Text generation, QA, research	Open-science collaboration , multilingual
Chinchilla(Hoffmann et al., 2022)	DeepMind	2022	70	Web-scaled data	Transforme r	Text generation, QA, research	Balanced computation
GPT-4(Singh et al., 2023)	OpenAI	2023	Unknown (Hundreds)	Web-scaled data	Transforme r	Text generation, QA, translation	Better context understandin g, multimodal

3. RELATED WORK

The LLMs are progressively evolving and are now also, incorporated in numerous turfs other than language processing. LLMs such as Google's Gemini(Gemini Team et al., 2023), GPT or BERT models are used largely for imaging tasks such as acquisition, segmentation, classification enhancement or reconstruction.

Image classification requires a robust model which gives a correct label as output, LLMs due to great ability to generate raw labels can be used for classify images. (Haider et al., 2024), created 50 clinical scenarios for evaluating the efficiency of the LLMs in classifying the breast images into different categories. Each vignette reflected the diverse presentations of the selected breast condition. A detailed review of the chosen classification systems was conducted that focused on the specific criteria and definitions of breast disease. The LLM's (Gemini and ChatGpt), responses were scored from 0 to 2 to label them as incorrect, partially correct and completely correct classifications. These models operated well in all the state of art classification methods in the comparative analysis. The LLMs were given these descriptions and were asked to evaluate the results using these selected classification models standards criteria sets and the results then were graded. The Gemini attained an accuracy of 98%, overpowering the OpenAI , ChatGPT-4's with 71% accuracy.

Also, Fine-grained image classification, i.e. identifying the minute differences between the closely related classes (e.g., different types of flower or food or dogs) requires a lot of precision. A hybrid approach was proposed by(Qu & Yatskar, 2024) , for classification on FGVC-Aircraft , CUB-200-2011,Stanford Cars, Stanford Dogs, Flower-102, and Food-101. Model had two components, one GPT-4 (LLM) and ensemble of simple linear classifiers. The GPT-4 was used to generate a structured tree of token (attributes) i.e. the textual features and the CLIP's encoder was used to extract the visual features. The combination of the two features were then inspected using the ensemble technique to perform classification. This in-cooperation of textual features from LLM maintained a standard with state-of-art method and also, helped in achieved a good accuracy of 96.6%. Similarly, (Rodríguez-de-Vera et al., 2023)introduced an expert learning method called Dining on Details (DoD) for food classification. Authors used ImageBind multi-modality embedding space, to find the similarity between various classes. The approach was used on several dataset for validation and achieved performance gain in comparison to convolutional networks from 0.5% to 1.61%.

The LLMs are now a days, also used for zero-shot learning due to their vast knowledge. (Matsuura et al., n.d.), also, used LLaVA, for zero-shot classification for datasets CIFAR-10, CIFAR-100, and ImageNet100. The authors proposed to tag a query along the input image to the model in order to create a flexible classifier that could perform classification without a pre-defined list in comparison to CLIP model. Hierarchical classification was performed by extracting the relevant features from the LLM's output i.e. the raw labels were then was used to find the class label from the datasets. The extracted part was embedded using CLIP text encoder and all the labels were also embedded in the encoder. The matching between the two embeddings was measured by the closeness between the each text output pair with the cosine similarity function. This whole setup achieved an accuracy of 86%.

Medical Visual Question Answering (Med-VQA)(de Faria et al., 2023) the process which uses LLMs as a key step for QA to improve the overall process of medical analysis and provides a directed questionnaire. But this method requires a lot of data, to overcome this limitation, (Wang et al., 2024), introduced Image to Label to Answer (ITLTA) framework. ITLTA combined the LLMs capabilities with multi-label learning and performed zero-shot learning. The encoder was pre-trained using large amount of other medical data i.e. VQA-Med 2019 dataset, along with label attention mechanism. The labels obtained using visual data were used for consequent output generation for the medical image associated with question. The proposed model achieved a correctness of 85.6%.

For enhancing the process of medical imaging and diagnostics, (Lai et al., 2024), proposed a method called residual-based LLM. This model was created to improve the biomedical imaging tasks such as segmentation and classification without the need to acquire large amount of training data or high computational demands. The transformer block of the pre-trained LLM was frozen, then trainable linear layers were positioned parallel around the LLM block followed by a residual connection. The introduction of frozen block and residual connection acted as a booster and provided a consistent improvement in 2D and 3D imaging tasks. The new model attained an accuracy of 89.7% on diverse dataset: MedMNIST-2D and 3D, BreastMNIST,

DermaMNIST and FractureMNIST3D. Also, the model performed better than ResNet-18 and ResNet-50(Kapoor & Arora, 2022).

(Panagoulias et al., 2024) generated Multimodal Diagnosis from Medical Images and Symptom Analysis using LLM. GPT-4 was used for performing organized interactions. In the paper, an approach was proposed which was made of two parts, first, a multimodal LLM evaluation. This was done based on multiple choice questions in pathology domain to explore a number of diseases, conditions, and entity knowledge. GPT-4 was used to respond to complex, medical questions consisting of both images and text. Then, domain-specific analysis based on data gained from the previous interactions was conducted to evaluate the correctness of diagnosis performed by LLM. For generating captions for limited CT scans and digital breast tomosynthesis (DBT) images, (Aswiga & Shanthi, 2022) proposed a method which combined Multi Level Transfer Learning (MLTL) framework with a LSTM model. This method attained an accuracy of 96.90% and a BLEU score of 76.9%, and outperformed existing methods.

LMMs are widely used for 2D image analysis but in order to advance the 3D medical image analysis, (Bai et al., 2024), presented a large-scale 3D multi-modal medical dataset i.e. M3D-Data, consisting of 120K image-text pairs and 662K instruction-response pairs were designed for 3D medical image tasks. Also, they proposed LaMed , LLM for 3D images which was combined with a 3D segmentation model to express the 3D medical images. It could perform tasks such as image-text retrieval, report generation, and visual question answering, and also includes tasks such as vision language positioning and segmentation.

4. DISCUSSION

Imaging tasks such as image acquisition, captioning, segmentation or classification require a model that works with efficacy and accuracy. The paper focuses on highlighting the role of LLMs in improving these tasks. Different sections of the work introduces several language models and compares the recent large language models. The comparison table also highlight the characteristics and their application in various fields. It is observed that introduction of LLMs to the existing setups improve the whole process. The efficient performance of LLMs in health sector makes them a powerful tool. From aiding in image classification, image captioning to conducting medical diagnosis these models achieve good results and comparable accuracy as compared to state-of-art methods. These models also outpace the deep neural networks such as ResNet making them standard competent. These innovative approaches which combine the encoders for generating textual embeddings along with visual embeddings enhance the interpretability of the under lying networks. It is observed that LLMs require large training data which becomes challenging in data scarce situations, to overcome this, many researchers have proposed solutions such as zero-shot learning on pre-trained models. Also, these models are widely used for 2D images and there is less work on 3D images. The powerful language models like ChatGPT and GEMINI are well accepted in medical areas. Also, it is seen that GEMINI outperforms ChatGPT in image classification task. But these models might not capture some information outside their capabilities and act like simple black-boxes.

5. CONCLUSION

The paper encourages the use of LLMs in advancing the imaging tasks. The work highlights the potential of language models in enhancing the medical image recognition and diagnosis of underlying disease. We have discussed the challenges and advancements these models can bring to the health sector. The study is conducted to explore the various techniques which combine the LLMs with existing setups to improve accuracy and provide decision support systems in medical imaging analysis. The paper will serve as a direction to the development of new approaches to use LLMs in enhancing the existing systems beyond language processing. Though these models require huge amounts of data and act as block-boxes without understanding the logics behind interpretation but this can be improved with the use of pre-training and data

augmentation methods. There is still a lot of scope in improving tasks such as providing contextual information for image segmentation tasks, automated image captioning and developing more efficient algorithms and architectures to reduce the computational requirements of multimodal models integrating LLMs.

REFERENCES

Aswiga, R. V., & Shanthi, A. P. (2022). A Multilevel Transfer Learning Technique and LSTM Framework for Generating Medical Captions for Limited CT and DBT Images. *Journal of Digital Imaging*, *35*(3), 564–580. https://doi.org/10.1007/s10278-021-00567-7

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024). *M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models* (arXiv:2404.00578). arXiv. http://arxiv.org/abs/2404.00578

Bellemo, V., Lim, Z. W., Lim, G., Nguyen, Q. D., Xie, Y., Yip, M. Y. T., Hamzah, H., Ho, J., Lee, X. Q., Hsu, W., Lee, M. L., Musonda, L., Chandran, M., Chipalo-Mutati, G., Muma, M., Tan, G. S. W., Sivaprasad, S., Menon, G., Wong, T. Y., & Ting, D. S. W. (2019). Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study. *The Lancet Digital Health*, *1*(1), e35–e44. https://doi.org/10.1016/S2589-7500(19)30004-4

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1901.02860

de Faria, A. C. A. M., Bastos, F. de C., da Silva, J. V. N. A., Fabris, V. L., Uchoa, V. de S., Neto, D. G. de A., & Santos, C. F. G. dos. (2023). *Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.11033

De La Cruz, G., Lira, M., Luaces, O., & Remeseiro, B. (2024). Eye-LRCN: A Long-Term Recurrent Convolutional Network for Eye Blink Completeness Detection. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(4), 5130–5140. https://doi.org/10.1109/TNNLS.2022.3202643

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.11929

Feng, L., Tung, F., Hajimirsadeghi, H., Ahmed, M. O., Bengio, Y., & Mori, G. (2024). *Attention as an RNN* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2405.13956

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2023). *Gemini: A Family of Highly Capable Multimodal Models* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2312.11805

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. https://doi.org/10.36227/techrxiv.23589741.v4

Haider, S. A., Pressman, S. M., Borna, S., Gomez-Cabello, C. A., Sehgal, A., Leibovich, B. C., & Forte, A. J. (2024). Evaluating Large Language Model (LLM) Performance on Established Breast Classification Systems. *Diagnostics*, *14*(14), 1491. https://doi.org/10.3390/diagnostics14141491

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den,

Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training Compute-Optimal Large Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.15556

Joshi, A. K. (1991). Natural Language Processing. *Science*, *253*(5025), 1242–1249. https://doi.org/10.1126/science.253.5025.1242

Kapoor, P., & Arora, S. (2022). Applications of Deep Learning in Diabetic Retinopathy Detection and Classification: A Critical Review. In D. Gupta, Z. Polkowski, A. Khanna, S. Bhattacharyya, & O. Castillo (Eds.), *Proceedings of Data Analytics and Management* (Vol. 91, pp. 505–535). Springer Singapore. https://doi.org/10.1007/978-981-16-6285-0_41

Lai, Z., Wu, J., Chen, S., Zhou, Y., & Hovakimyan, N. (2024). *Residual-based Language Models are Free Boosters for Biomedical Imaging* (arXiv:2403.17343). arXiv. http://arxiv.org/abs/2403.17343

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). *Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models*. https://doi.org/10.48550/ARXIV.2304.01852

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1907.11692

Manning, C. D., & Schütze, H. (2005). *Foundations of statistical natural language processing* (8. [print.]). MIT Press.

Matsuura, M., Jung, Y. K., & Lim, S. N. (n.d.). *Visual-LLM Zero-Shot Classification*.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models* (Version 9). arXiv. https://doi.org/10.48550/ARXIV.2307.06435

Nazih, W., Aseeri, A. O., Atallah, O. Y., & El-Sappagh, S. (2023). Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images. *IEEE Access*, *11*, 117546–117561. https://doi.org/10.1109/ACCESS.2023.3326528

Padmanayana, & B.K, Dr. A. (2022). Binary Classification of DR-Diabetic Retinopathy using CNN with Fundus Colour Images. *Materials Today: Proceedings*, *58*, 212–216. https://doi.org/10.1016/j.matpr.2022.01.466

Panagoulias, D. P., Virvou, M., & Tsihrintzis, G. A. (2024). *Evaluating LLM -- Generated Multimodal Diagnosis from Medical Images and Symptom Analysis* (arXiv:2402.01730). arXiv. http://arxiv.org/abs/2402.01730

Peng, H., Schwartz, R., & Smith, N. A. (2019). *PaLM: A Hybrid Parser and Language Model* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1909.02134

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1802.05365

Qu, R., & Yatskar, M. (2024). *LLM-based Hierarchical Concept Decomposition for Interpretable Fine-Grained Image Classification* (arXiv:2405.18672). arXiv. http://arxiv.org/abs/2405.18672

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Rodríguez-de-Vera, J. M., Villacorta, P., Estepa, I. G., Bolaños, M., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition. *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, 43–52. https://doi.org/10.1145/3607828.3617797

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.1909.08053

Singh, S. K., Kumar, S., & Mehra, P. S. (2023). Chat GPT & Google Bard AI: A Review. *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*, 1–6. https://doi.org/10.1109/ICICAT57735.2023.10263706

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R. Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B. (2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2201.11990

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1909.09586

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., … Le, Q. (2022). *LaMDA: Language Models for Dialog Applications* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2201.08239

Wang, J., Seng, K. P., Shen, Y., Ang, L.-M., & Huang, D. (2024). Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering. *Electronics*, *13*(12), 2273. https://doi.org/10.3390/electronics13122273

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1906.08237

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497