Article

Large Language Models in Healthcare Current Development and Future Directions

Xinrui Jiang , Lixiang Yan , Raja Vavekanand [*] , Mengxuan Hu

Article

Large Language Models in Healthcare Current Development and Future Directions

Xinrui Jiang [1], Lixiang Yan [2], Raja Vavekanand [3,*] and Mengxuan Hu [4]

[1] Hong Kong University of Science and Technology
[2] Monash University
[3] Datalink Research and Technology Lab
[4] The University of Virginia

Abstract: The rapid advancement of large language models (LLMs), such as ChatGPT, has revolutionized natural language processing, enabling these models to generate coherent and contextually relevant text. These capabilities hold significant potential for various applications, particularly in the medical field. LLMs can assist healthcare professionals by providing accurate information, supporting diagnostic processes, and enhancing medical education. However, the implementation of LLMs in clinical settings faces several barriers, including technical limitations, ethical concerns, regulatory constraints, and practical challenges. Ensuring accuracy, reliability, and interpretability is crucial, along with addressing biases, ensuring data privacy, and establishing ethical guidelines. Future research should focus on improving training techniques, developing explainable AI methods, and creating efficient, resource-saving models. Expanding applications through domain-specific models and multi-modal integration can further unlock the potential of LLMs. By addressing these challenges and prioritizing responsible development, LLM technology can be safely and effectively integrated into various domains, driving innovation and improving outcomes. This paper explores the current state of LLM technology, its medical applications, barriers to implementation, and directions for future research and development.

Keywords:

1. Introduction

The rapid advancement of artificial intelligence (AI) has paved the way for significant developments in natural language processing (NLP), with large language models (LLMs) standing out as a remarkable achievement. LLMs, such as OpenAI's ChatGPT, are designed to generate coherent and contextually relevant text based on vast amounts of data, encompassing books, articles, and diverse internet content. These models leverage deep learning techniques, particularly neural networks with multiple layers, to understand and produce highly proficient human language.

LLMs excel in their ability to perform zero-shot and few-shot learning, which means they can undertake a wide range of NLP tasks with little to no specific training for each task. This capability is derived from their extensive pre-training on large and varied datasets, enabling them to generalize and adapt to new queries and contexts effectively. For instance, ChatGPT has demonstrated near-human performance in various cognitive tasks, including passing medical licensing exams, thus highlighting its potential utility in professional domains like healthcare. Despite their impressive capabilities, the application of LLMs in clinical settings is still a subject of ongoing research and debate. The primary concern lies in the validation and benchmarking of these models for clinical use. While LLMs can generate plausible responses, the accuracy and reliability of these responses are critical, especially in the context of healthcare where incorrect information can have serious consequences. Therefore, rigorous testing and validation are necessary to ensure these models can be safely and effectively integrated into clinical practice.

2

2. Background Overview

The development of LLMs has seen a progressive enhancement in their architecture and training methodologies. Starting from GPT-1, which utilized a semi-supervised approach with the BooksCorpus dataset, the evolution continued through GPT-2 and GPT-3, each iteration significantly larger and more capable. These models have shown marked improvements in understanding and generating human language, making them increasingly versatile and powerful. The latest iteration, GPT-4, includes multimodal capabilities, allowing it to process and generate content from various types of inputs, although details about its architecture and training data remain confidential. One of the remarkable features of LLMs like ChatGPT is their training process, which involves not only extensive data pre-training but also fine-tuning with human-generated prompts and reinforcement learning from human feedback (RLHF). This fine-tuning process enhances the model's ability to handle conversational nuances and learn from ongoing interactions, making it more adept at providing useful and contextually appropriate responses.

However, the deployment of LLMs in healthcare is not without challenges. The substantial computational resources required for training these models translate into high economic and environmental costs. Efforts are being made to mitigate these costs through more efficient hardware and software solutions. For example, the Alpaca model, developed using the outputs from GPT-3.5, showcases that high-performance models can be created with significantly lower resources, thus democratizing access to advanced AI technology. While LLMs hold immense potential for transforming various aspects of medicine, including clinical practice, education, and research, their deployment in healthcare must be approached with caution. Ensuring their accuracy, reliability, and cost-effectiveness will be crucial in harnessing their full potential and integrating them safely into the medical field.

3. Evolution of LLM

The journey of developing large language models (LLMs) for chatbots has been marked by significant advancements in AI and NLP technologies. This development began with the introduction of GPT-1, which utilized a semi-supervised learning approach. GPT-1 was trained on the BooksCorpus dataset, containing over 7,000 unpublished books, allowing it to grasp the nuances of language and generate coherent text. This initial model laid the foundation for subsequent iterations, each more powerful and capable than the last. GPT-1, OpenAI released GPT-2, which marked a substantial leap in terms of scale and performance. GPT-2 was trained on a much larger dataset, encompassing a diverse range of internet text. This model consisted of 1.5 billion parameters, a significant increase from GPT-1's 117 million parameters. The larger scale and enhanced training data enabled GPT-2 to generate more complex and contextually relevant text, showcasing the potential of LLMs in various applications.

The introduction of GPT-3 represented another milestone in the development of LLMs. GPT-3's architecture included 175 billion parameters, making it one of the largest and most powerful language models at the time. It was trained on a vast dataset that included diverse internet content, allowing it to perform a wide range of NLP tasks with minimal task-specific training. GPT-3's ability to understand and generate human-like text across different contexts made it a versatile tool for various applications, including chatbots.

GPT-4, the latest iteration, further enhances the capabilities of its predecessors. While details about GPT-4's architecture and training data remain confidential, it is known to include multimodal input capabilities, allowing it to process and generate content from various types of data, including text, images, and possibly other modalities. This development opens new possibilities for more interactive and versatile AI applications. One of the key innovations in the development of LLM chatbots like ChatGPT is the fine-tuning process. After pre-training on large datasets, these models undergo fine-tuning with human-generated prompts. This process, known as reinforcement learning from human feedback (RLHF), involves training the model on a narrower dataset with specific prompts and feedback from human trainers. This approach helps the model to refine its responses, making them more accurate and contextually appropriate.

The fine-tuning process of ChatGPT is particularly notable. It involves an iterative approach where the model learns from interactions and feedback, improving its ability to handle conversational nuances and provide useful responses. This ongoing learning process enables ChatGPT to adapt and improve over time, making it a more effective tool for various applications, including customer service, education, and healthcare. Despite these advancements, the development of LLMs is not without challenges. Training such large models requires substantial computational resources, leading to high economic and environmental costs. Efforts are being made to address these issues through more efficient training methods and hardware optimizations. For instance, the Alpaca model, developed using outputs from GPT-3.5, demonstrates that high-performance LLMs can be created with significantly lower resources, making advanced AI technology more accessible. LLM chatbots has seen remarkable progress, with each iteration bringing significant improvements in scale, performance, and versatility. The fine-tuning process and ongoing innovations in training methods continue to enhance the capabilities of these models, paving the way for their broader application in various fields. However, addressing the computational and environmental costs remains a critical challenge that must be tackled to ensure the sustainable development and deployment of LLMs.

4. Healthcare Applications for LLM

Large language models (LLMs) like ChatGPT have shown promising potential in various medical applications, particularly in clinical settings. These models can assist healthcare professionals by providing accurate information, generating insights from medical literature, and supporting clinical decision-making processes. However, their implementation in clinical practice comes with challenges and considerations that must be addressed to ensure safety and efficacy. One of the primary applications of LLMs in medicine is in the realm of medical education. LLMs have demonstrated their ability to pass medical licensing exams, which indicates their capability to understand and generate medically relevant content. This makes them valuable tools for medical students and professionals seeking to enhance their knowledge and stay updated with the latest medical advancements. By providing detailed explanations and answering complex medical queries, LLMs can serve as reliable educational resources.

In clinical settings, LLMs can assist with patient care by providing evidence-based recommendations and supporting diagnostic processes. For instance, LLMs can analyze patient data, including medical histories and test results, to generate differential diagnoses and suggest potential treatment options. This can help healthcare providers make informed decisions, especially in complex cases where multiple factors need to be considered. Additionally, LLMs can aid in the early detection of diseases by identifying patterns and anomalies in patient data, potentially leading to timely interventions and improved patient outcomes. LLMs are also valuable in enhancing patient engagement and communication. They can be integrated into telemedicine platforms and patient portals to provide instant responses to patient inquiries, offer medication reminders, and deliver personalized health information. This can improve patient satisfaction and adherence to treatment plans, as well as reduce the burden on healthcare providers by automating routine tasks. For instance, chatbots powered by LLMs can handle a significant portion of patient interactions, allowing healthcare professionals to focus on more complex and critical cases.

LLMs in clinical settings requires careful consideration of several factors. One major concern is the accuracy and reliability of the information generated by these models. While LLMs can produce plausible responses, they are not infallible and may generate incorrect or misleading information. Therefore, it is essential to validate the outputs of LLMs and ensure they are used as supportive tools rather than sole decision-makers. Incorporating a human-in-the-loop approach, where healthcare professionals review and verify the recommendations provided by LLMs, can help mitigate risks and enhance the reliability of these systems. LLMs into clinical workflows necessitates robust data privacy and security measures. Patient data used for training and operating these models must be handled with the highest standards of confidentiality to comply with regulations such as the Health Insurance Portability and Accountability Act (HIPAA). Ensuring the ethical use of patient data and maintaining

transparency in the development and deployment of LLMs are crucial for building trust and acceptance among healthcare providers and patients. LLMs hold significant potential to transform clinical practice by enhancing medical education, supporting diagnostic and treatment processes, and improving patient engagement. However, their successful implementation requires addressing challenges related to accuracy, reliability, data privacy, and ethical considerations. Ongoing research and development, coupled with rigorous validation and ethical guidelines, will be essential in realizing the full potential of LLM technology in medicine.

5. Challenges to Integrating LLM

Despite the promising potential of large language models (LLMs) in various domains, including medicine, several barriers hinder their widespread implementation. These barriers encompass technical challenges, ethical considerations, regulatory constraints, and practical issues in real-world deployment. Understanding and addressing these barriers is crucial to harnessing the full potential of generative AI LLMs in a safe, effective, and ethical manner.

5.1. Technical Limitations

5.1.1. Accuracy and Reliability

While LLMs have shown impressive performance in generating human-like text, they are not infallible. These models can produce incorrect, misleading, or nonsensical outputs. In critical applications like healthcare, even minor inaccuracies can have significant consequences. Ensuring the reliability and accuracy of LLM-generated content is a major technical challenge that requires continuous validation and improvement.

5.1.2. Interpretability

LLMs operate as black boxes, making it difficult to understand how they arrive at specific outputs. This lack of interpretability can be problematic, especially in fields that require explainability, such as medicine and law. Developing methods to interpret and explain the decision-making process of LLMs is essential for their trusted use in high-stakes environments.

5.1.3. Computational Resources

Training and deploying LLMs require substantial computational power, memory, and storage. The high resource demands lead to significant economic and environmental costs. Efficient algorithms and hardware optimizations are needed to reduce these costs and make LLM technology more accessible and sustainable.

5.2. Ethical and Social Concerns

5.2.1. Bias and Fairness

LLMs can inadvertently learn and propagate biases present in their training data. These biases can manifest in various forms, such as gender, racial, or socioeconomic biases, leading to unfair or discriminatory outputs. Addressing bias in LLMs involves both curating balanced training datasets and developing algorithms that can detect and mitigate biases.

5.2.2. Privacy and Security

The use of large datasets for training LLMs raises concerns about data privacy and security. Ensuring that patient data, for example, is anonymized and protected is crucial to comply with regulations and maintain public trust. Additionally, LLMs can be vulnerable to adversarial attacks, where malicious inputs are designed to trick the model into generating harmful or incorrect responses.

5.2.3. Ethical Use

The potential for misuse of LLMs, such as generating deepfakes, disinformation, or harmful content, presents significant ethical challenges. Establishing guidelines and regulations to govern the ethical use of LLMs is necessary to prevent misuse and protect individuals and society.

5.3. Regulatory and Legal Constraints

5.3.1. Compliance with Regulations

Implementing LLMs in regulated industries, such as healthcare and finance, requires adherence to strict regulatory standards. Ensuring compliance with laws such as the Health Insurance Portability and Accountability Act (HIPAA) in the US, or the General Data Protection Regulation (GDPR) in the EU, is essential but can be complex and resource-intensive.

5.3.2. Liability and Accountability

Determining liability in cases where LLM-generated outputs lead to harm or errors is a significant legal challenge. Clear guidelines on accountability, especially in automated systems, are necessary to address legal and ethical concerns.

5.4. Practical Implementation Challenges

5.4.1. Integration with Existing Systems

Incorporating LLMs into existing workflows and systems can be technically complex. Compatibility issues, data integration, and user training are practical challenges that need to be addressed to ensure smooth implementation.

5.4.2. User Trust and Acceptance

Gaining the trust and acceptance of end-users, such as healthcare professionals, is critical for the successful adoption of LLMs. Demonstrating the reliability, safety, and benefits of LLMs through rigorous testing and transparent communication can help build user confidence.

5.4.3. Continuous Improvement and Maintenance

LLMs require ongoing monitoring, maintenance, and updates to ensure they remain accurate and relevant. Establishing processes for continuous improvement and addressing evolving challenges is essential for the long-term success of LLM implementations. Generative AI LLMs hold significant potential, overcoming the barriers to their implementation requires a multi-faceted approach. Addressing technical limitations, ethical and social concerns, regulatory constraints, and practical challenges is essential to ensure that LLMs can be deployed safely, effectively, and ethically across various domains.

Figure 1. Challenges to Integrating Large Language Models.

6. Future Directions

The future of large language models (LLMs) holds immense promise, particularly in advancing their capabilities, addressing existing challenges, and expanding their applications across various fields. To harness their full potential, several key areas of research and development need to be prioritized. These directions focus on enhancing the performance, reliability, and ethical deployment of LLMs, ensuring they can be effectively integrated into diverse domains such as healthcare, education, and beyond.

6.1. Improving Accuracy and Reliability

6.1.1. Enhanced Training Techniques

Future research should focus on developing more sophisticated training techniques that improve the accuracy and reliability of LLMs. This includes exploring new architectures, optimizing training algorithms, and leveraging transfer learning to reduce the amount of domain-specific data required for fine-tuning.

6.1.2. Contextual Understanding

Enhancing LLMs' ability to understand and generate contextually accurate responses is crucial. Research in this area can explore multi-modal learning, where models can process and integrate information from various sources such as text, images, and videos to provide more comprehensive and accurate outputs.

6.1.3. Error Detection and Correction

Implementing mechanisms for real-time error detection and correction can significantly improve the reliability of LLM outputs. This includes developing models that can identify and flag potentially incorrect or misleading information, enabling human oversight and intervention when necessary.

6.2. Enhancing Interpretability and Transparency

6.2.1. Explainable AI (XAI)

Developing methods to make LLMs more interpretable and transparent is essential for their trusted use in critical applications. Research should focus on creating tools and frameworks that can provide insights into the decision-making processes of LLMs, making their outputs more understandable and explainable to users.

6.2.2. Model Auditing

Implementing robust auditing mechanisms to regularly assess and report on the performance, biases, and ethical implications of LLMs can enhance transparency. This includes developing standardized benchmarks and evaluation metrics for auditing LLMs across different applications.

6.3. Addressing Ethical and Social Concerns

6.3.1. Bias Mitigation

Research should prioritize developing techniques to identify and mitigate biases in LLMs. This includes creating more diverse and representative training datasets, as well as developing algorithms that can detect and correct biases in real-time.

6.3.2. Privacy-Preserving Models

Ensuring data privacy and security is critical, especially in sensitive applications like healthcare. Future research should explore privacy-preserving techniques such as federated learning and differential privacy, which allow models to learn from data without compromising individual privacy.

6.3.3. Ethical Guidelines and Frameworks

Establishing comprehensive ethical guidelines and frameworks for the development and deployment of LLMs is essential. Research should focus on creating standardized ethical protocols that can guide the responsible use of LLMs, ensuring they are developed and deployed in ways that prioritize user safety and societal well-being.

6.4. Reducing Computational and Environmental Costs

6.4.1. Efficient Training Methods

Developing more efficient training methods that reduce the computational and environmental costs of training LLMs is a critical area of research. This includes exploring model compression techniques, energy-efficient hardware, and distributed training methods that minimize resource consumption.

6.4.2. Resource Optimization

Research should focus on optimizing the utilization of computational resources during the training and deployment of LLMs. This includes developing algorithms that can dynamically allocate resources based on the complexity of the task, reducing unnecessary computation and energy usage.

6.5. Expanding Applications and Accessibility

6.5.1. Domain-Specific Models

Creating domain-specific LLMs tailored to particular fields such as medicine, law, or education can enhance their effectiveness and reliability. These models can be fine-tuned on specialized datasets and integrated with domain-specific knowledge to provide more accurate and relevant outputs.

6.5.2. Cross-Lingual Capabilities

Expanding the linguistic capabilities of LLMs to support multiple languages and dialects can enhance their accessibility and usability across different regions and cultures. Research should focus on developing cross-lingual models that can understand and generate text in diverse languages, making AI technology more inclusive.

6.5.3. Democratizing Access

Ensuring that advanced LLM technology is accessible to a broader range of users, including small businesses, researchers, and educational institutions, is crucial. Future research should explore ways to lower the barriers to access, such as creating open-source models, providing affordable cloud-based solutions, and developing user-friendly interfaces.

6.6. Integration with Other AI Technologies

6.6.1. Multi-Modal Integration

Combining LLMs with other AI technologies such as computer vision, speech recognition, and robotics can create more versatile and powerful systems. Research should explore the integration of LLMs with these technologies to enable multi-modal learning and interaction, expanding the scope of their applications.

6.6.2. Human-AI Collaboration

Developing frameworks that facilitate seamless collaboration between humans and AI systems can enhance the effectiveness of LLMs. This includes creating interfaces that allow users to interact with LLMs in intuitive ways, leveraging the strengths of both human and machine intelligence.

The future research and development of LLMs should focus on improving their accuracy, reliability, and ethical deployment, while also expanding their applications and accessibility. Addressing these key areas will be crucial in realizing the full potential of LLM technology and ensuring its positive impact across various domains.

Figure 2. Future   Directions flowchart.

7. Conclusion

The development and application of large language models (LLMs) like ChatGPT mark a significant advancement in the field of artificial intelligence and natural language processing. These models have demonstrated remarkable capabilities in understanding and generating human-like text, opening up a wide range of possibilities across various domains, including medicine, education, customer service, and more. Despite their impressive achievements, the integration of LLMs into real-world applications faces several barriers that need to be addressed. Technical challenges such as ensuring accuracy, reliability, and interpretability are critical for the trusted use of these models in high-stakes environments. Ethical and social concerns, including bias, privacy, and the potential for misuse, must be meticulously managed to prevent harm and ensure the responsible deployment of LLMs. Furthermore, regulatory and legal constraints, coupled with practical implementation challenges, require robust frameworks and guidelines to facilitate safe and effective integration.

Future research and development directions must focus on enhancing the performance of LLMs, making them more reliable, interpretable, and resource-efficient. Efforts to mitigate biases, ensure data privacy, and establish ethical guidelines are essential to building public trust and acceptance.

Additionally, expanding the applications of LLMs through domain-specific models, cross-lingual capabilities, and integration with other AI technologies will further unlock their potential and democratize access to advanced AI solutions. LLMs have the potential to revolutionize various fields, their successful implementation hinges on addressing the multifaceted challenges they present. By prioritizing research in key areas and developing comprehensive strategies to manage ethical, technical, and practical issues, we can harness the full potential of LLM technology, driving innovation and positive impact across diverse sectors.
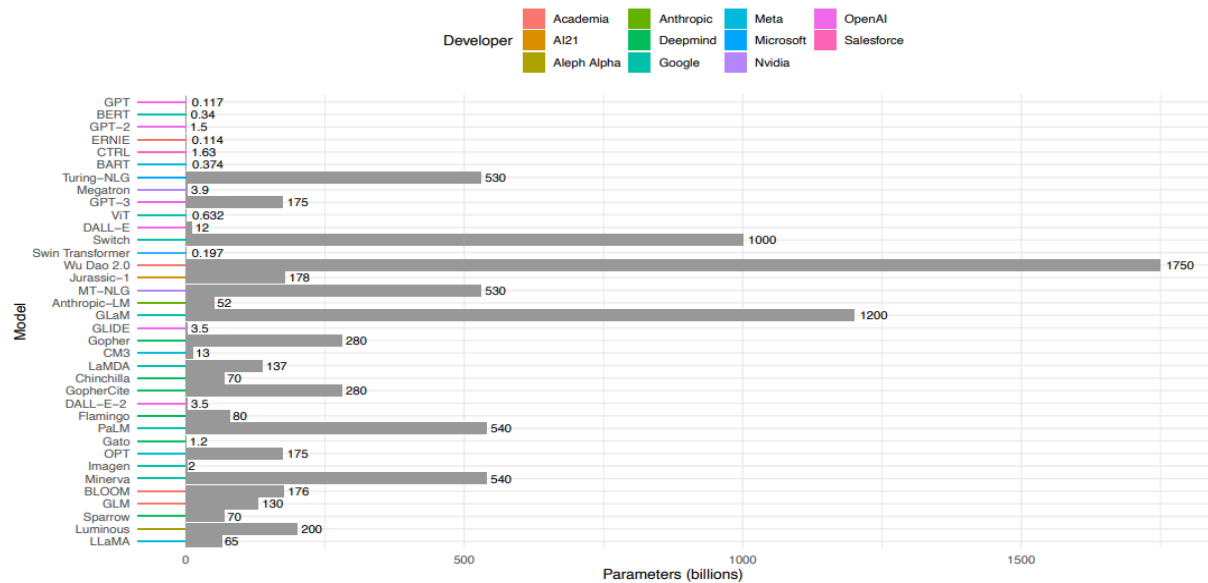
Appendix

Figure A1. Comparison of large language models developed in recent years.

Large language models have been developed with parameters in the order of billions, but size is not the only measure of progress. Many previous models feature more parameters than the models currently generating the greatest impact in healthcare. For example, GPT-3, from which GPT-3.5 was developed, features just 175 billion parameters compared to multiple models featuring over 1 trillion parameters. The largest iteration of LLaMA (used in many open source alternatives to ChatGPT) features just 65 billion parameters. Other factors contribute to a model's utility, such as its training data and schemata, finetuning protocols, and overarching architecture. GPT-4 has been released but its architecture is confidential, preventing inclusion in this comparison. Other models include GPT, GLM, BLOOM, Minerva, Imagen, OPT, Gato, PaLM, Flamingo, DALL-E-2, GopherCite, Chinchilla, LaMDA, CM3, Gopher, GLIDE, GLaM, Anthropic-LM, MT-NLG, Jurassic-1, Wu Dao 2.0, Swin Transformer, Switch, DALL-E, ViT, GPT-3, Megatron, Turing-NLG, BART, CTRL, ERNIE, GPT-2, and LLaMA.
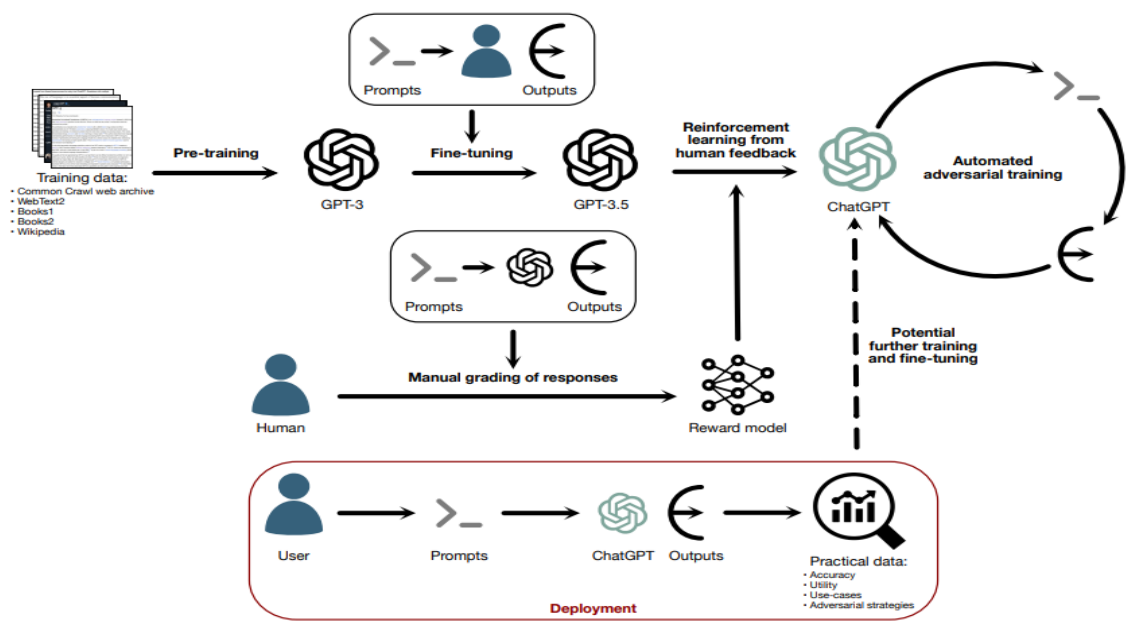
Figure A2. Finetuning a large language model (GPT-3.5) to develop an LLM chatbot.

GPT-3, trained through word prediction tasks, was fine-tuned to develop GPT-3.5 by exposure to human prompt-output pairings. To develop ChatGPT, reinforcement learning from human feedback (RLHF) was employed, using a reward model trained using human grading of a limited number of GPT-3.5 outputs to a list of prompts. This model could be used with a larger list of prompts for greater scale training. The architecture and training processes of GPT-4 and subsequent versions of ChatGPT are confidential but likely apply similar principles due to similar error types.

Table A1.

Artificial intelligence	Computational systems capable of completing tasks which otherwise require human intelligence.
Computational resources	The hardware required to train and deploy a machine learning model, including processing power, memory, and storage.
Dataset size	The number of text documents, sentences, or words used to train a large language model.
Deep learning	A variant of machine learning involving neural networks with multiple layers of processing 'perceptrons' (nodes), which together facilitate extraction of higher features of unstructured

	input data.
Few-shot learning	Artificial intelligence developed to complete tasks with exposure to only a few initial examples of the task, with accurate generalisation to unseen examples.
Generative artificial intelligence	Computational systems capable of producing content—such as text, images, or sound—on-demand.
Large language model	A type of artificial intelligence model using deep neural networks to learn the relationships between words in natural language, using large datasets of text to train.
Machine learning	A field of artificial intelligence featuring models which enable computers to learn and make predictions based on input data, learning from experience.
Model size	The number of parameters in an artificial intelligence model; large language models consist of layers of communicating nodes which each contain a set of parameters which are optimised during training.
Natural language processing	A field of artificial intelligence research focusing on the interaction between computers and human language.

Neural network	Computing systems inspired by biological neural networks, comprised of 'perceptrons' (nodes), usually arranged in layers, communicating with one another and performing transformations upon input data.

References

1. Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25, 24-29.
2. Aggarwal, R., et al. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. npj Digital Medicine, 4, 65.
3. Vavekanand, R., & Kumar, S. (2024). LLMEra: Impact of Large Language Models. *Available at SSRN 4857084*.
4. Liddy, E. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science (Marcel Decker, Inc.).
5. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. Multimedia Tools and Applications, 82, 3713-3744.
6. Brown, T., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877-1901). Curran Associates, Inc.
7. Moor, M., et al. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616, 259-265.
8. Kaplan, J., et al. (2020). Scaling laws for neural language models. Preprint at https://doi.org/10.48550/arXiv.2001.08361.
9. Shoeybi, M., et al. (2020). Megatron-LM: Training multi-billion parameter language models using model parallelism. Preprint at https://doi.org/10.48550/arXiv.1909.08053.
10. Thoppilan, R., et al. (2022). LaMDA: Language models for dialog applications. Preprint at https://doi.org/10.48550/arXiv.2201.08239.
11. Zeng, A., et al. (2022). GLM-130B: An open bilingual pre-trained model. Preprint at https://doi.org/10.48550/arXiv.2210.02414.
12. Amatriain, X. (2023). Transformer models: An introduction and catalog. Preprint at https://doi.org/10.48550/arXiv.2302.07730.
13. Introducing ChatGPT. (n.d.). Retrieved from https://openai.com/blog/chatgpt.
14. Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. Preprint at https://doi.org/10.48550/arXiv.2203.02155.
15. OpenAI. (2023). GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774.
16. Kung, T. H., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2, e0000198.
17. Thirunavukarasu, A. J., et al. (2023). Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. JMIR Medical Education, 9, e46599.
18. Ayers, J. W., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine. https://doi.org/10.1001/jamainternmed.2023.1838.
19. Lehman, E., et al. (2023). Do we still need clinical language models? Preprint at https://doi.org/10.48550/arXiv.2302.08091.
20. Vavekanand, R. (2024). A Machine Learning Approach for Imputing ECG Missing Healthcare Data. *Available at SSRN 4822530*.
21. Yang, X., et al. (2022). GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. Preprint at https://doi.org/10.48550/arXiv.2203.03540.
22. Weiner, S. J., Wang, S., Kelly, B., Sharma, G., & Schwartz, A. (2020). How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. Journal of the American Medical Informatics Association, 27, 770-775.
23. Ioannidis, J. P. A. (2005). Why most published research findings are false. PLOS Medicine, 2, e124.
24. Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. The Lancet Digital Health, 5, e105-e106.

25. Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—Should academics worry? Nature. https://doi.org/10.1038/d41586-022-04397-7.
26. Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. Patterns, 4, 100706.
27. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. (2023). Nature, 613, 612-612.
28. Vavekanand, R. (2024). SUBMIP: Smart Human Body Health Prediction Application System Based on Medical Image Processing. Studies in Medical and Health Sciences, 1(1), 14-22.
29. Sample, I. (2023). Science journals ban listing of ChatGPT as co-author on papers. The Guardian.
30. Flanagin, A., Bibbins-Domingo, K., Berkwits, M., & Christiansen, S. L. (2023). Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. JAMA, 329, 637-639.
31. Authorship and contributorship. (n.d.). Retrieved from https://www.cambridge.org/core/services/authors/publishing-ethics/research-publishing-ethics-guidelines-for-journals/authorship-and-contributorship.
32. New AI classifier for indicating AI-written text. (n.d.). Retrieved from https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.
33. Kirchenbauer, J., et al. (2023). A watermark for large language models. Preprint at http://arxiv.org/abs/2301.10226.
34. The Lancet Digital Health. (2023). ChatGPT: Friend or foe? The Lancet Digital Health, 5, e102.
35. Vavekanand, R. (2024f). Hanooman: A Generative AI and Large Language Model Chatbot Inspired From Lord Hanuman. ResearchGate. https://doi.org/10.13140/RG.2.2.12093.06889
36. Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digital Health, 2, e0000205.
37. Vavekanand, R., Sam, K., Kumar, S., & Kumar, T. (2024). CardiacNet: A Neural Networks Based Heartbeat Classifications using ECG Signals. *Studies in Medical and Health Sciences*, *1*(2), 1-17.