Evaluating Deployability of LLMs

Responsible AI of LLMs in Healthcare & Biomedical sector

Nikhesh Kumar Murali
GE23C011

Evaluating Deployability of LLMs

INTRODUCTION TO LLMs

Natural Language Processing (NLP) is a sub-field that intersects linguistics, computer science, and artificial intelligence. It focuses on the interaction between computers and human language, aiming to enable machines to understand, interpret, and generate human language in a useful and meaningful way.

1. Language Understanding: This involves comprehending the meaning and context of the language humans use naturally.
2. Text Analysis: Techniques like sentiment analysis, summarization, and classification.
3. Language Generation: The ability of a system to generate coherent and contextually relevant language on its own.
4. Speech Recognition: Translating spoken language into text.

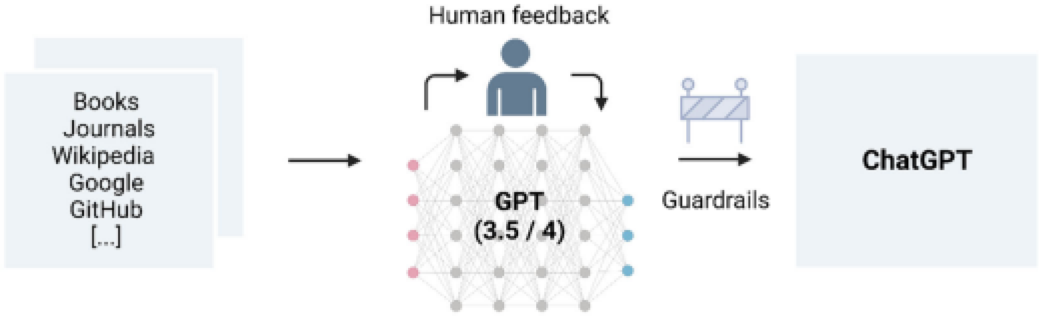

Fig : Simplified design of a LLM

Large Language Models (LLMs) are models in the domain of NLP, which are used to generate answers to queries. They are advanced artificial intelligence systems capable of understanding, generating, and interpreting human language. Built upon deep learning architectures, typically transformer models, LLMs are trained on vast amounts of text data. They can perform a wide range of language tasks, from translation to content generation, by predicting the likelihood of a sequence of words. Their ability to generate coherent, contextually relevant text has revolutionized natural language processing, making them central to modern AI applications.

LLMs are a significant advancement in NLP. They are massive neural network models trained on vast amounts of text data. These models learn patterns, nuances, and the structure of language, enabling them to perform various language-based tasks with impressive accuracy.

Language models like GPT-4 have a wide range of applications that enhance various sectors. They are pivotal in content creation, aiding in writing articles, generating creative content, and summarizing complex materials. In the realm of customer support and information access, they power conversational agents and chat-bots, offering efficient and interactive communication solutions. Their ability in text analysis is significant for sentiment analysis, market research, and monitoring social media trends.

Additionally, they play a crucial role in translation services, offering real-time, multi-language translation. For educational purposes, these models assist in learning and language tutoring, making education more accessible and personalized. In the field of web search, they enhance search engine algorithms, ensuring queries are better understood and the most relevant results are provided. Lastly, they contribute immensely to accessibility services, helping people with disabilities by transforming text to speech and vice versa, thereby bridging communication gaps.

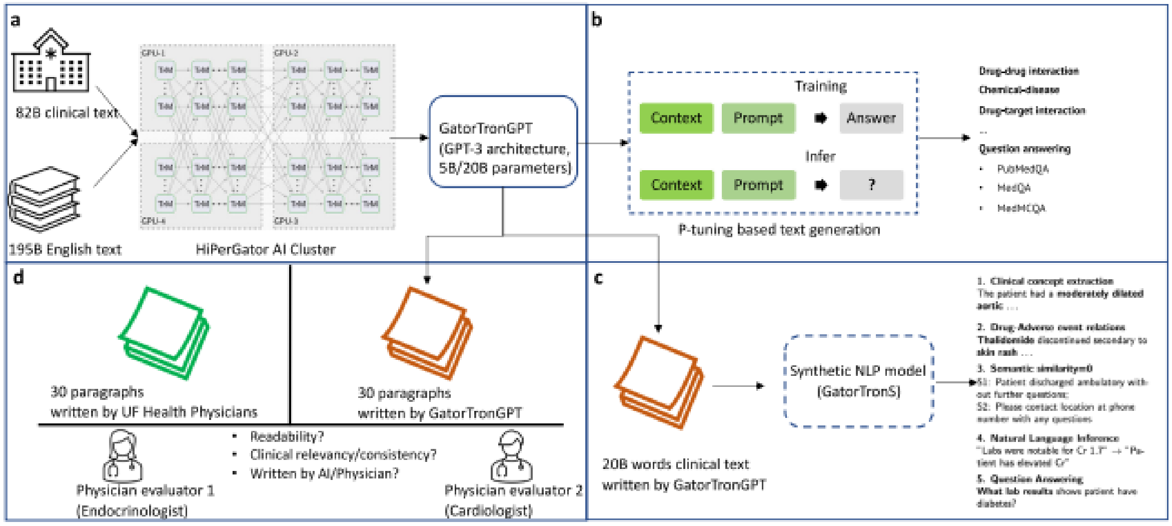

Fig : Architecture of clinical GatorTronGPT for biomedical NLP

ARCHITECTURE OF LLMs

ALGORITHMS BEHIND LLMs

RESPONSIBLE AI

Responsible AI in the context of LLMs involves developing and using these models ethically, transparently, and with accountability. This encompasses ensuring fairness, mitigating biases, respecting privacy, and maintaining security. Responsible AI also implies that models should be explainable, with decisions and processes being understandable by humans. It emphasizes the need for AI to align with human values and societal norms, ensuring that the deployment of LLMs contributes positively to society and does not exacerbate existing inequalities or harmful stereotypes or biases.

URGENCY FOR RAI

1. **Increasing Dependence on AI in Healthcare**: As AI becomes more integral to healthcare operations and decision-making, its impact on quality of life, and economies grows exponentially. This makes it crucial to ensure that AI systems operate fairly, ethically, and responsibly.
2. **Risk and Reputation Management**: The use of AI can lead to significant risks, including the propagation of bias, violation of privacy, and erroneous decisions. These risks can damage an organization's reputation, leading to loss of trust among customers, stakeholders, and the public. This trust is essential for long-term success and credibility.
3. **Legal and Regulatory Compliance**: With the rapid evolution of AI, governments and regulatory bodies are increasingly focusing on creating laws and regulations to govern its use. There are important questions like who will be held responsible if a treatment suggested by LLMs goes wrong for a patient - whether the LLM maker or the dataset provider? Non-compliance with these regulations can lead to hefty fines, legal challenges, and negative publicity.
4. **Need for Explainability**: In sectors like healthcare, finance, and education, AI decisions can significantly impact individuals' lives. Therefore, it's critical to have transparent, explainable AI systems to justify decisions, particularly in situations involving sensitive matters like credit approvals, healthcare treatments, or educational opportunities. We need AI to explain the results to any lay person in the society.
5. **Societal Impact**: AI's decisions and actions can have far-reaching societal implications. Ensuring responsible AI is vital to prevent harm,

promote fairness, and ensure that the benefits of AI are distributed equitably across society.

6. **Financial Implications**: The cost of non-compliance or irresponsible AI use can be substantial. As noted, a single non-compliance event can cost millions in revenue, not to mention the long-term financial impacts that can extend beyond immediate fines or penalties.
7. **Energy Consumption and Environmental Impact:** The training and operation of large-scale models like GPT-3, GPT-4,Bard, require significant computational resources, leading to high energy consumption. Addressing the environmental impact is a critical aspect of responsible AI.

DEPLOYABILITY OF LLM

Deployability of LLMs refers to the measure of LLMs to be deployed into real world applications. It includes the model building, data storing, securing sensitive data, interacting with customers, and updating its answers using human-in-the-loop RL.

Explainability : Large Language Models (LLMs) often act as "black boxes," making their decision-making processes opaque. Ensuring explainability is crucial for trust and reliability. This involves developing methods to interpret and understand the model's outputs and the underlying reasons for these outputs. Improved explainability aids in identifying biases, errors, and the model's decision boundaries, enhancing user trust and compliance with regulatory standards.

Fairness / bias: LLMs can inadvertently perpetuate or amplify biases present in their training data. Addressing fairness involves identifying and mitigating these biases to ensure equitable and unbiased outputs. This includes diverse data representation and employing fairness metrics during model training and evaluation to ensure the model's decisions do not disadvantage any group.

Examining the role of bias in deployability of LLMs : The presence of bias in LLMs significantly impacts their deployability, especially in sensitive sectors like healthcare, law, or finance. A biased model may produce unfair or discriminatory results, leading to complications in the health of an individual. It might give incorrect prescriptions, wrong diagnosis, hallucinate diseases. It also includes exploring how bias can affect the model's performance and public perception, thereby influencing its acceptance and effectiveness. Addressing these biases is essential for ensuring that LLMs are deployed responsibly and beneficially.

Mitigating hallucinations using fine-tuning : Fine-tuning LLMs involves adjusting pre-trained models to specific tasks or datasets. This process requires careful calibration to maintain the general capabilities of the model while optimizing for task-specific performance. Effective fine-tuning balances the need for specialized knowledge against the risk of overfitting to niche data.

Security and Privacy of data : Deploying LLMs raises concerns about data security and privacy, especially when handling sensitive information. Ensuring robust encryption methods, secure data storage, and controlled data access are paramount. Additionally, safeguarding against adversarial attacks and ensuring compliance with data protection regulations are crucial for maintaining user trust.

Model Updating & Maintenance : Continuous updating and maintenance of LLMs are vital for relevance and accuracy. This includes regular retraining with updated datasets and evaluating bias to reflect new information and trends, and patching vulnerabilities. Maintenance strategies should ensure the model adapts to evolving data landscapes while preserving its foundational strengths.

Factors affecting the deployability of LLMs:

1. **Size and Scale**: LLMs like GPT-3 contain hundreds of billions of parameters, making them capable of understanding and generating complex language constructs.
2. **Pre-training and Fine-tuning**: They are first pre-trained on a large corpus of text and then fine-tuned for specific tasks.
3. **Versatility:** LLMs can be adapted for a wide range of applications, from writing assistance to conversational AI.

HEALTHCARE & BIOMEDICAL LLMs

GENDER BIAS IN HEALTHCARE LLMs

The "DELPHI" paper scrutinizes LLMs' handling of controversial topics and highlights gender biases. These biases, often reflecting societal norms, emerge in LLMs due to unbalanced training datasets. LLMs, like GPT-3.5, learn patterns from the data they're fed, inadvertently mirroring existing societal biases. Understanding why LLMs produce biased outputs is complex, as they're not intentionally programmed with these biases. Mitigating these biases involves training with diverse, inclusive datasets and applying ethical guidelines in AI development. Completely eradicating bias in LLMs, however, remains a formidable challenge in AI research.

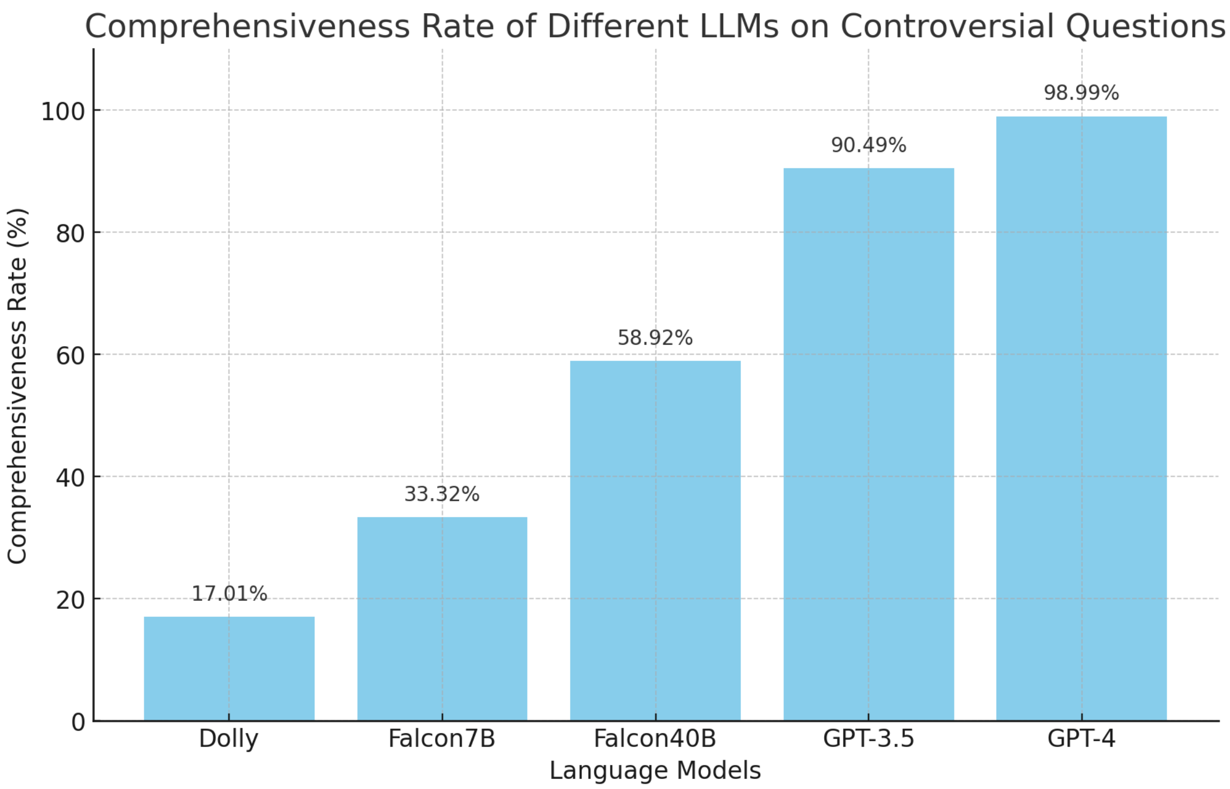

Fig : CAR metric to evaluate the LLMs from DELPHI paper

RACIAL BIAS IN HEALTHCARE LLMs

EXPLAINABILITY OF LLMS

Explainable AI (XAI) seeks to make the decision-making processes of AI systems transparent and understandable to humans. In the context of Large Language Models (LLMs), XAI involves elucidating how these models generate responses, offering insights into their complex inner workings. This transparency is crucial for trust and reliability, particularly in critical applications. Interpretable AI enables users to comprehend and trust the predictions made by LLMs, ensuring that the AI's reasoning aligns with human values and logic. Moreover, XAI assists in identifying and rectifying potential biases or errors in the model's learning process.

Evaluating the quality of explanations : Evaluating the quality of explanations provided by LLMs is a complex task. It involves assessing how well the AI's rationale aligns with human understanding and whether the explanations enhance the user's clarity and decision-making. Quality evaluations often use metrics like comprehensibility, relevance, and accuracy. In LLMs, this means examining the coherence and relevance of the model's responses and the degree to which these responses can be traced back to understandable logic. High-quality explanations not only foster trust but also aid in debugging and improving the models.

Black box models and rationale of explanations : LLMs often function as 'black box' models, where the internal mechanisms are not fully visible or understandable. The rationale behind explanations seeks to bridge this gap, providing insights into why the model made a particular decision. This involves breaking down complex, often opaque neural network processes into understandable parts. The goal is to translate the model's complex, multidimensional data processing into a format that is comprehensible to humans, thus making th

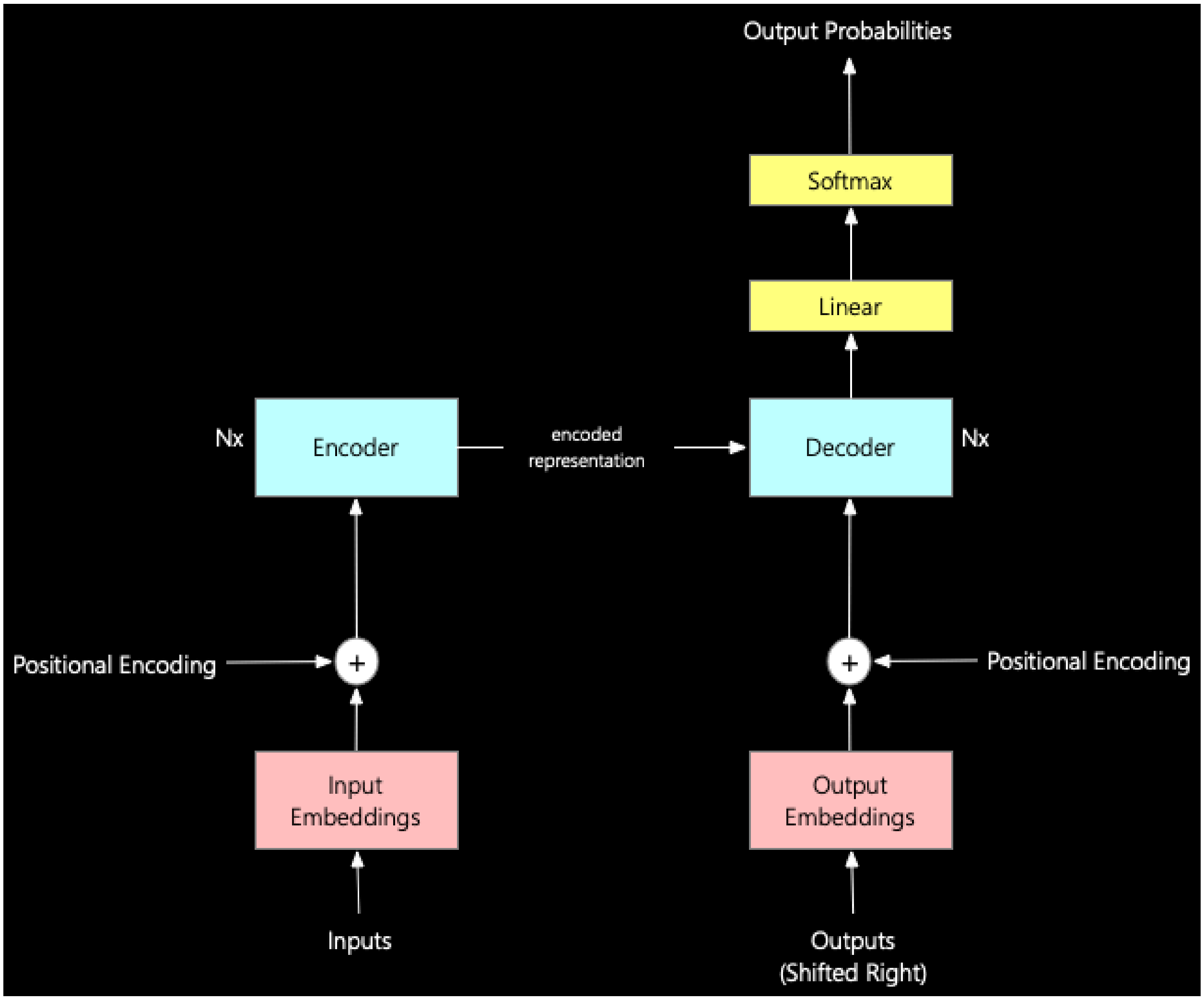

Fig : Transformer architecture

e AI's decisions more transparent and accountable. In the “DELPHI” paper, the authors came up with ‘Comprehensive Answer Rate’ to evaluate the performance of healthcare LLMs in real world datasets.

QUANTIFYING EXPLAINABILITY

CONCLUSION

In conclusion, we have studied the deployability of Large Language Models (LLMs) particularly focusing on their responsible deployment in the healthcare domain.

A critical takeaway from this analysis is the importance of Responsible AI (RAI) in the development and deployment of LLMs. Ensuring fairness, mitigating biases, respecting privacy, and maintaining security are paramount in this regard.

In terms of bias and fairness, we have identified the need for continuous vigilance and improvement in LLMs. The evaluation of bias, through various methods, is essential to ensure that these models do not perpetuate existing societal prejudices. This is particularly vital in sensitive applications where biased decisions can have significant ethical and legal ramifications. Lesser the bias, better the output of the model and deployability in real world applications.

In summary, while LLMs, especially Healthcare LLMs offer transformative capabilities in various sectors, their responsible deployment necessitates a comprehensive understanding of their architecture, functionality, and potential impacts. The deployability is in positive correlation with the explainability factor of the LLM, better explainability of the model makes it more deployable. As we advance in the field of AI, it is imperative to continue refining these models, prioritizing ethical considerations, and ensuring they align with the broader objectives of benefiting society and advancing human-computer interaction in a responsible and equitable manner.

REFERENCES

1. Wornow, M., Xu, Y., Thapa, R. et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit. Med. 6, 135 (2023). https://doi.org/10.1038/s41746-023-00879-8
2. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al. Large language models in medicine. Nat Med 29, 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8
3. Clusmann, J., Kolbinger, F.R., Muti, H.S. et al. The future landscape of large language models in medicine. Commun Med 3, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1
4. Peng, C., Yang, X., Chen, A. et al. A study of generative large language model for medical research and healthcare. NPJ Digit. Med. 6, 210 (2023). https://doi.org/10.1038/s41746-023-00958-w
5. Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, Klaus-Robert Müller Explaining NonLinear Classification Decisions with Deep Taylor Decomposition
6. Georgia Deaconu - Towards LLM Explainability: Why Did My Model Produce This Output?
7. Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, Balaraman Ravindran - Towards Transparent and Explainable Attention Models
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin - Attention is All you Need
9. S.V. Praveen, S. Vijaya - Exploring infection clinicians' perceptions of bias in Large Language Models (LLMs) like ChatGPT: A deep learning study
10. Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In Proceedings of The ACM Collective Intelligence Conference (CI '23). Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599
11. ChatGPT Replicates Gender Bias in Recommendation Letters
12. Omiye, J.A., Lester, J.C., Spichak, S. *et al.* Large language models propagate race-based medicine. *npj Digit. Med.* 6, 195 (2023). https://doi.org/10.1038/s41746-023-00939-z
13. Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Prof Leo Anthony Celi, Prof Judy Gichoya et al .Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study