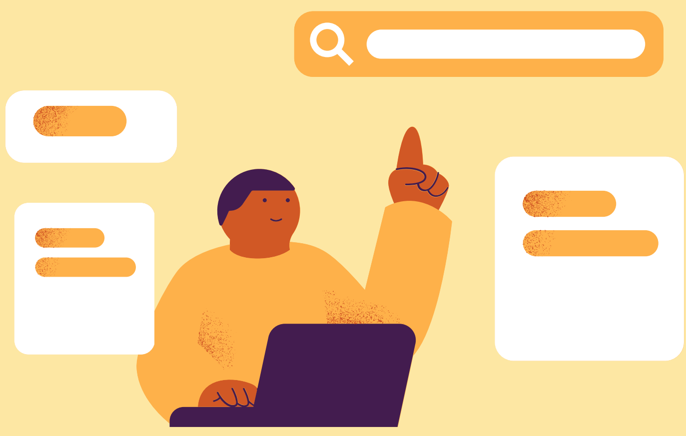
Comparative Analysis on GatorTronGPT in Healthcare

GROUP 10
Nur Hidayah Binti Ahmad Shafil (22120931), Then Dao Qing (23057608), Choon Yue Hua (17152027), Low Meng Fei (23063305), Syaidatul Salmah Nurbalqis Binti Saiful (17140336)

Introduction

The healthcare industry is experiencing a transformative moment with the emergence of Large Language Models (LLMs). **GatorTronGPT** is specifically designed to handle medical terminology and clinical contexts. It bridges general language understanding and specialized medical knowledge, excelling in clinical documentation, patient-doctor dialogue summarization, and medical information processing.

Problem Statement



1. **Domain-Specific LLM Development**: Difficulty in creating LLMs capable of understanding and generating clinically relevant content.
2. **Clinical Documentation Burden**: High computational and time burden on healthcare providers for clinical documentation.
3. **Summarization Efficiency**: Need for accurate and efficient automatic summarization of doctor-patient encounters.

Methodologies

Peng et al., 2023	Lyu et al,. 2024
MTS-DIALOG dataset with 1701 doctor-patient dialogues data is used for data training, validation and testing.	Train from scratch with 82 billion words of clinical narratives from University of Florida (UF) Health and 195 billion of diverse English words from the Pile dataset.
Models used are T5 model by Google Research with finetuned using Huggingface fine-tuning pipeline, GatorTronGPT 5B and GatorTronGPT 20B by University of Florida's academic Health.	Models used are GatorTronGPT 5B and GatorTronGPT 20B by University of Florida's academic Health.
"soft prompts" is initialized by Long Short-Term Memory networks and Multi-Layer Perceptron using Nvidia NeMo package based on Python.	Synthetic clinical text generation was tested, producing 20 billion words to train synthetic NLP models, named GatorTronS.
Adam optimizer was used for prompt tuning and CosineAnnealing scheduler was used to adjust the learning rate of the modelling.	Prompt-tuning algorithms is formulated and applied using Adam optimizer.

Strengths and Weaknesses

	Strengths	Weaknesses
Model: GatorTronGPT with prompt-tuning algorithms (Lyu et al,. 2024)	**Scalability**: Handles summarization tasks with fewer parameter updates compared to traditional fine-tuning. **Accuracy**: Outperforms T5 in clinical benchmarks and captures more critical information across various scenarios. **Flexibility**: Summarized doctor-patient dialogues with limited data via few-shot learning and prompt-tuning without modifying core parameters. **Adaptability**: Strong performance in low-resource settings due to few-shot learning setups. **Training Efficiency**: Requires only 2–4 hours for training compared to 9+ hours for T5 fine-tuning. **Implementation Simplicity**: Avoids parameters updates which reduces hardware requirements.	**High Computational Cost**: Large LLMs (e.g., GatorTronGPT-20B) still demand significant resources and are time-intensive, even with prompt-tuning. **Data Privacy Concerns**: Larger LLMs even with de-identified data remain sensitive to handling clinical data. **Implementation Complexity**: Prompt design and tuning require expertise for task optimization. **Hallucinations**: Occasionally missed critical details in summaries will affect reliability.
Model: GatorTronGPT using GPT-3 architecture (Peng et al., 2023)	**Scalability**: Handles large-scale datasets (277 billion words). **Accuracy**: State-of-the-art performance in biomedical NLP tasks. **Flexibility**: Generate diverse and synthetic clinical text that outperforms real-world text-trained models in specific tasks and supports scalable pipelines biomedical pipelines. **Adaptability**: Adapts to new tasks with minimal data via strong few-shot learning capabilities. **Data Augmentation**: Generates synthetic text to augment data in low-resource scenarios.	**High Computational Cost**: Requires high computational resources (e.g., 560 GPUs) for training and deployment. **Data Privacy Concerns**: Synthetic data may replicate biases or sensitive patterns inherent in original datasets. **Implementation Complexity**: Synthetic text generation pipelines and hyperparameter tuning require expertise. **Hallucinations:** Risks of clinically misleading information, especially in sensitive healthcare applications. **Interpretability Challenges**: Operates as a "black box," limiting insight into model decision-making.

Findings and Best Practices

Findings	Peng et al., 2023	Lyu et al., 2024
The studies utilized different GatorTronGPT training approaches for model generalizability.	GatorTronGPT used GPT-3 architecture with a custom depth-to-width ratio, **training from scratch** 5B and 20B parameter models.	**Pretrained GatorTronGPT** with GPT-3-based model with 277B words of texts.
Both studies evaluated the model for human-likeness to ensure it analyses and produces outputs comparable to human expertise.	**Turing test showed no significant difference** between GPT-written notes and physician-written notes in **linguistic readability (p=0.22)** and **clinical relevance and consistency (p=0.91).**	**GatorTronGPT summaries were more precise in capturing critical information like patient demographics to specific clinical conditions** than T5 summaries when compared to gold-standard summaries.

Best Practices

Both studies highlighted the efficiency of **soft prompts** compared to hard prompts as they reduce computation costs and enable task-specific customization by fine-tuning during training.

Future Directions

Based on the papers' findings, several key areas need further research:

1. **Clinical Safety and Validation**
   - Use evaluation frameworks to assess the clinical accuracy of healthcare LLMs. Standardized testing protocols should be established to ensure the reliability of these models in healthcare settings.
2. **Technical Improvements**
   - Advancements in few-shot learning capabilities are critical for improving the adaptability of healthcare LLMs. Reinforcement learning from human feedback (RLHF) should be integrated to enhance model performance.
3. **Practical Applications**
   - Integration with existing Electronic Health Record (EHR) systems will enhance workflow efficiency. Real-time clinical decision-support tools should also be created to assist healthcare providers.

References

- Lyu, M., Peng, C., Li, X., Balian, P., Bian, J., & Wu, Y. (2024). Automatic Summarization of Doctor-Patient Encounter Dialogues Using Large Language Model through Prompt Tuning. *arXiv preprint arXiv:2403.13089*.
- Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., ... & Wu, Y. (2023). A study of generative large language model for medical research and healthcare. *NPJ digital medicine, 6*(1), 210.