|  |  |  |
| --- | --- | --- |
|  | Strengths | Weaknesses |
| Peng et al., (2023)  Model: GatorTronGPT, using GPT-3 architecture | **Scalability:** Successfully handles large-scale datasets (277 billion words).  **Accuracy:** State-of-the-art performance in biomedical NLP tasks (e.g., relation extraction, question answering)  **Flexibility:** Generate diverse and synthetic clinical text that outperforms real-world text-trained models in specific tasks. Supports scalable pipelines for tasks like biomedical QA and relation extraction.  **Adaptability:** Effectively used for synthetic data generation to mitigate privacy issues. Adapts to new tasks with minimal data via strong few-shot learning capabilities. | **High Computational Cost:** Requires significant 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:** Demands expertise to manage synthetic text generation pipelines and hyperparameter tuning (e.g., learning rate, batch size, dropout).  Hallucinations: Risks of hallucinations in LLM outputs could lead to clinically misleading information, especially in sensitive healthcare applications |
| Lyu et al., (2024)  Model: GatorTronGPT with prompt-tuning algorithms (GatorTronGPT-5B vs GatorTronGPT-20B) | **Scalability:** GatorTronGPT with prompt-tuning handles summarization tasks with fewer parameter updates compared to traditional fine-tuning.  **Accuracy:** GatorTronGPT with prompt-tuning outperformed T5 in summarization key clinical benchmarks while capturing more critical information across various scenarios.  **Flexibility:** Few-shot learning effectively summarized doctor-patient dialogues with limited data. Prompt-tuning enables flexible task adaptation without modifying the model’s core parameters.  **Adaptability:** Strong performance in low-resource settings due to adaptability in few-shot learning setups.  **Training** **Efficiency**: Prompt-tuning required only 2–4 hours for training, compared to 9+ hours for T5 fine-tuning.  **Implementation Simplicity**: Prompt-tuning avoids updating model parameters which reduces hardware requirements. | **High Computational Cost:** Although reduced with prompt-tuning, large generative LLMs (e.g., GatorTronGPT-20B) still require significant resources and time-intensive to train.  **Data Privacy Concerns:** Larger generative LLMs even with de-identified data are more sensitive handling sensitive clinical dialogue data than smaller LLMs.  **Implementation Complexity:** Prompt design and tuning require expertise to optimize performance across specific task  **Hallucinatios**: Occasionally missed critical details in summaries affecting reliability. |

|  |  |  |
| --- | --- | --- |
|  | 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. |