A Comparison of Large Language Models for Oncology Clinical Text Summarization

Chiazam Izuchukwu
ci02061@georgiasouthern.edu
Department of Information Technology
Institute for Health Logistics and Analytics
USA Georgia Southern University
Statesboro, GA 30460, USA

Hayden Wimmer
hwimmer@georgiasouthern.edu
Department of Information Technology
Institute for Health Logistics and Analytics
USA Georgia Southern University
Statesboro, GA 30460, USA

Carl M Rebman, Jr.
carlr@sandiego.edu
Knauss School of Business
Department of Supply Chain, Operations,
and Information Systems
University of San Diego
San Diego, CA 92110, USA

Abstract

The rapid growth of data in the health sector has made it crucial to communicate essential information quickly and succinctly. The vast amount of textual data from electronic health records tends to overwhelm healthcare professionals which reduces the time they can dedicate to patient care. This massive amount of complex qualitative data causes physicians to struggle with the decision-making process which had traditionally relied on human evaluation. This study addresses the urgent need for effective summarization of health records to improve patient outcomes and clinical decision-making. We highlight the use of large language models (LLMs) to produce concise summaries of patients' medical oncology reports. Specifically, we utilized pre-trained transformer models, including BART, T5, and Pegasus, to summarize patient clinical notes. The performance of these models was evaluated using BLEU, ROUGE, and BERT scores on CORAL expert-curated medical oncology reports that were de-identified using Philter. The results show that the BART and T5 models performed the best, with the generated summaries being shorter than the original oncology reports. This approach reduces information overload and enhances patient care by providing concise and informative summaries.

Keywords: Large Language Model (LLM), Text Summarization, Artificial Intelligence (AI), Coral, Medical, Oncology

1. INTRODUCTION

Supporting physicians and clinicians in the oncology field struggle during the decision-making process and the oncology field has been a subject of extensive research and debate. Medical professionals are frequently overwhelmed with an abundance of data and information. While quantitative data is well-suited for machine learning and statistical methods to aid in decision support, qualitative data is rich with explicit and tacit information. Processing this qualitative data to make it available and useful for decision support is a complex task, traditionally relying on human evaluation through qualitative techniques such as coding and analysis. During a visit with an oncology doctor, much qualitative data is extracted from the patient and added to their electronic health record. This can be seen as a semi-structured interview with both closed and open-ended questions.

Physicians often face time constraints and production pressures, limiting the time they can spend with each patient. Reading and processing the extensive textual data generated during medical visits is an overwhelming task, given these time constraints and the need for high patient turnover. A single medical chart or electronic health record can generate numerous pages of qualitative textual data. Over the course of a patient's stay in a medical facility or through routine visits to medical providers, the volume of data grows significantly. One promising method to assist physicians and doctors in processing this vast amount of text data is the use of Artificial Intelligence (AI), specifically large language models (LLMs).

This paper proposes the use of large language models (LLMs) such as BART, T5, and Pegasus for summarizing medical oncology reports, enhancing clinical decision-making by reducing information overload. BART is highlighted as the most effective model, consistently outperforming others across various metrics including ROUGE and BERTScore, despite similar BLEU scores among the models. The study underscores the potential of LLMs to support physicians by efficiently processing extensive qualitative data in electronic health records, thereby improving patient care and decision-making timelines. The remainder of this paper is organized as follows: next we present a review of some relevant literature useful in our work, followed by our methodology where we illustrate the LLMs and evaluation metrics. We then advance to our results which include standard evaluation metrics and a brief statistical analysis, then conclude with our discussion and future works.

2. LITERATURE REVIEW

Text summarization offers a significant advantage over manual summarization by condensing large data into meaningful summaries while preserving content. It can be classified into extractive summarization, which uses statistical and linguistic features to highlight important parts, and abstractive summarization, which generates summaries by understanding the entire document. One area that can benefit from LLM and text summarization methods is clinical text and articles in healthcare.

According to Allahyari et al. (2017), the increasing availability of documents has spurred extensive research in automatic text summarization, which aims to create concise and fluent summaries that preserve key information and overall meaning. Automatic text summarization is challenging because humans summarize text by fully understanding it first, a capability that computers lack. There are two main approaches to summarization: extractive, which selects and reproduces key sections verbatim, and abstractive, which generates new text conveying the essential information. Despite the naturalness of human-created summaries, research has predominantly focused on extractive methods, which often produce better results due to the complexities involved in semantic representation and language generation inherent in abstractive summarization. (Allahyari et al., 2017).

Batra, Chaudhary, Bhatt, Varshney, and Verma (2020) felt that an overwhelming amount of articles and links that people have to choose from and as this data grows, the importance of semantic density does as well. They make the claim that more concise, meaningful communication is needed, and that text summarization might be a solution. Text summarization addresses this by condensing lengthy texts into short, informative sentences. Machine learning models can play a crucial role in this process by first understanding the document and then producing a summary. Their paper analyzed five different models in the literature from the years 2013-2019 to and present the argument that these models can provide a good summarization of large amounts of data.

Bhatia and Jaiswal (2015) noted how the rapid growth of World Wide Web data has made it increasingly difficult to manually gather and summarize information. Their study investigated trends in text summarization methods. They examined eight different approaches to extractive summarization and eight different approaches to abstractive summarization. Their study concluded that extractive summarizations deal with important sentences while abstractive summarization processes seek understanding of the text and articles and then proceed to build a summary. They also found that automated processes can saves time and efficiently retrieves information from large documents. (Bhatia & Jaiswal, 2015).

Van Veen et al. (2023), conducted a study evaluated methods for adapting large language models (LLM) to summarize clinical text. According to Van Veen et al. (2023), sifting through vast textual data and summarizing key information from electronic health records (EHR) imposes a substantial burden on clinicians' time. They analyzed eight models across a diverse set of summarization tasks including radiology reports and doctor-patient dialogue. Although large language models (LLMs) show promise in natural language processing (NLP) tasks, their efficacy in clinical summarization has not been rigorously demonstrated. They performed a quantitative assessment which revealed trade-offs between models and methods, with some LLM advances not improving results. More notably, their study demonstrated that LLM summaries are often preferred over human expert summaries due to higher scores for completeness, correctness, and conciseness.

Medical care and observational studies in oncology require a thorough understanding of a patient's disease progression and treatment history, often documented within clinical notes. Large language models (LLMs) have demonstrated impressive capabilities, but the standards for clinical applications are exceptionally high. As large language models (LLMs) are becoming more popular, it is essential to evaluate their potential in oncology.

Savova et al. (2019), noted that data produced during the processes of clinical care and research in oncology are proliferating at an exponential rate This prompted them to perform a study that reviewed the advances of natural language processing (NLP) and information extraction methods relevant to oncology based on publications from PubMed as well as NLP and machine learning conference proceedings in the last 3 years. The review highlighted significant advancements in NLP and information extraction that have the potential to improve the fidelity of oncology phenotypes and reduce errors derived from clinical texts. They also notated that summarization and information retrieval applications can reduce search burden and enable clinicians to spend more time with their patients. They surmised that advancements are critical for catalyzing clinical care, research, and regulatory activities by providing more detailed and accurate phenotype information from real-world data (Savova et al., 2019).

Singhal et al., 2023 noted that medicine is an endeavor where language is important for interactions between clinicians, researchers, and patients. They felt that today's AI models for applications in medicine and healthcare have largely failed to fully utilize language and were mostly effective with single task systems. Current assessments of clinical knowledge in these models often rely on automated evaluations based which may not fully capture the complexity and nuances of clinical reasoning and knowledge.

To address this issue Singhal et al. (2023) created a study and introduced MultiMedQA, a comprehensive benchmark combining six existing medical question-answering datasets, and a new dataset of medical questions searched online, HealthSearchQA. The goal was to evaluate the capabilities of LLMs in the medical domain comprehensively. They used the Pathways Language Model (PaLM), a 540-billion parameter LLM, and its instruction-tuned variant, Flan-PaLM, on MultiMedQA and assessed the models' performance across various datasets, including MedQA, MedMCQA, PubMedQA, and MMLU clinical topics. Their results found that Flan-PaLM achieved state-of-the-art accuracy on all MultiMedQA multiple-choice datasets, including 67.6% accuracy on MedQA, surpassing the prior state-of-the-art by over 17%. However, human evaluations revealed significant gaps in the models' performance, highlighting areas where the models still fall short. The resulting model, Med-PaLM, showed improvements in comprehension, knowledge recall, and reasoning with increased model scale and instruction prompt tuning (Singhal et al., 2023).

Sushil et al. (2024) objective was to assess the performance of three recent LLMs (GPT-4, GPT-3.5-turbo, and FLAN-UL2) in extracting detailed

ISSN: 2473-4901
v10 n6208

oncological information from clinical progress notes, using a newly curated, fine-grained, expert-labeled dataset of 40 de-identified breast and pancreatic cancer progress notes. They evaluated the models in zero-shot extraction from two narrative sections of clinical progress notes, using BLEU-4, ROUGE-1, and exact match (EM) F1-score metrics. Their team of oncology fellows and medical students manually annotated 9028 entities, 9986 modifiers, and 5312 relationships to support this evaluation. GPT-4 exhibited the best overall performance with an average BLEU score of 0.73, an average ROUGE score of 0.72, an average EM-F1-score of 0.51, and an accuracy of 68% based on expert manual evaluation. It excelled in extracting tumor characteristics and medications, and in inferring symptoms and future medication considerations. Common errors included partial responses and hallucinations (Sushil et al., 2024).

3. METHODOLOGY

In our study, we conducted various analyses and experiments on a unique CORAL reports dataset to assess the performance of various Large Language Models performing abstractive summarization. These datasets serve as the basis for our comparison and analysis. Figure 1 below illustrates the framework of our method.

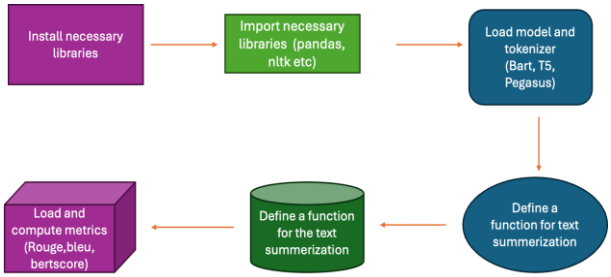

Figure 1 The architecture of oncology text summarization with LLMs

Dataset

The dataset used in this paper is CORAL expert-curated medical oncology reports (2024). The dataset comprises 100 pancreatic cancer notes, including demographic details and corresponding medical oncology notes for patients from the University of California, San Francisco (UCSF) Information Commons. This dataset, containing patient data from 2012 to 2022, has been de-identified using Philter (2023). Pancreatic cancer samples were collected while ensuring a diverse distribution of race/ethnicity. The race/ethnicity groups were either evenly distributed or limited to the maximum counts available in the UCSF dataset, whichever was smaller (Goldberger et al., 2000).

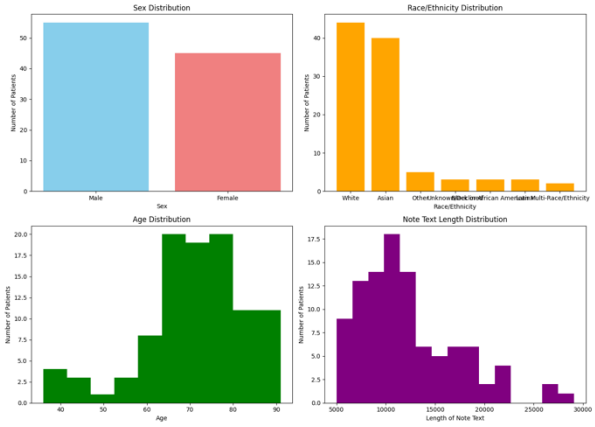

Figure 2 Composition of oncology dataset

Large Language Models

Large Language Models (LLMs) are transformative technologies in natural language processing (NLP). BART, Pegasus, and T5 are sequence-to-sequence models, also known as encoder-decoder models, primarily designed for natural language generation tasks. These models utilize vast datasets and advanced machine-learning techniques to accurately understand and generate human language. The foundational architecture behind LLMs typically involves deep learning techniques, such as transformers, which allow them to process and produce text that closely mimics human linguistic abilities (abstract approach) (2017). LLMs have revolutionized applications such as text generation, translation, and summarization, making interactions with machines more intuitive and seamless. They have become essential tools in various industries, enabling automated customer service, content creation, and data analysis.

BART Model

BART, or Bidirectional and Auto-Regressive Transformers, is a sophisticated LLM developed by Facebook AI. It combines the strengths of BERT's bidirectional encoding with GPT's autoregressive decoding, making it highly effective for a range of NLP tasks including text generation, machine translation, and summarization (2019). BART's architecture allows it to predict corrupted text and fill in missing information, enhancing its performance in generating coherent and contextually relevant text. This hybrid approach enables BART to excel in understanding and generating text, making it a versatile tool for various

applications such as dialogue generation and language modelling.

Pegasus Model
Pegasus, created by Google Research, is another LLM specifically designed for abstractive text summarization. Pegasus employs a unique pre-training objective that involves masking entire sentences and then predicting them, which closely mimics the task of summarization (2020). This approach enables Pegasus to generate high-quality summaries that capture the essence of the original content, making it a powerful tool for condensing large volumes of information. Pegasus's ability to produce concise and informative summaries has significant implications for fields such as news aggregation, academic research, and legal document review.

T5 Model
T5, or Text-To-Text Transfer Transformer, is a versatile LLM from Google Research that treats every NLP task as a text-to-text problem (2020). This unified approach simplifies the model architecture and makes T5 applicable to a wide array of tasks, from translation to question answering to summarization. T5's ability to be fine-tuned for specific applications allows it to achieve state-of-the-art performance across various benchmarks. The model's versatility and effectiveness make it an essential tool for researchers and practitioners in NLP, enabling them to tackle a broad spectrum of language-related challenges with a single model framework.

The advancement of LLMs like BART, Pegasus, and T5 highlights the rapid progress in NLP. These models not only improve the accuracy and efficiency of language-related tasks but also pave the way for new applications in fields such as healthcare, education, and content creation. For instance, in healthcare, LLMs can assist in summarizing patient records, generating medical reports, and even supporting diagnostic processes through natural language understanding (2021). In education, these models can provide personalized tutoring, automated grading, and language translation services, enhancing the learning experience for students worldwide.

As LLM technology continues to evolve, it promises to further bridge the gap between human and machine communication, enabling more natural and productive interactions. The development and deployment of LLMs are expected to bring about significant changes in how we interact with technology, making it more accessible and efficient. However, the growth of LLMs also raises important ethical and societal questions, such as issues of bias, privacy, and the potential for misuse. Researchers and developers must address these challenges to ensure that the benefits of LLM technology are realized responsibly and equitably (2021).

4. RESULTS

Analysis and Evaluation Metrics
ANOVA tests and Tukey's Honest Significant Difference (HSD) test were used in the statistical analyses to ascertain the significance of the group differences. The following section analyzes the evaluation metrics employed in summarizing the clinical notes. These metrics are employed to measure the quality and effectiveness of the generated summaries, utilizing a range of well-known and widely accepted evaluation standards for various large language models (LLMs).

Rouge Score
According to Lin (2004) Rouge (Recall-Oriented Understudy for Gisting Evaluation) score is a set of metrics used to evaluate the quality of summaries by comparing them to reference summaries. ROUGE metrics are commonly used in natural language processing tasks to measure the similarity between a generated summary and a reference summary. Here are some of the most frequently used ROUGE metrics and their formulas:

ROUGE-N is a recall-based measure that calculates the overlap of n-grams between the generated summary and the reference summary (Lin, 2004).

$$ROUGE-N = \frac{\sum_{(Reference\ summaries)} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{(Reference\ summaries)} \sum_{gram_n} Count(gram_n)} \#(1)$$

Where:
- $Count_{match}(gram_n)$ is the number of n-grams in the reference summary that match an n-gram in the generated summary.
- $Count(gram_n)$ is the total number of n-grams in the reference summary.
- ROUGE-1: Measures the overlap of unigrams (1-grams) between the generated summary and the reference summary.

ISSN: 2473-4901
v10 n6208

- ROUGE-2: Measures the overlap of bigrams (2-grams) between the generated summary and the reference summary.
- ROUGE-L measures the longest common subsequence (LCS) between the generated summary and the reference summary.

$$ROUGE - L = \frac{LCS(\chi,\gamma)}{Lenght(\gamma)} \#(2)$$

Where:

- $LCS(\chi,\gamma)$ is the length of the longest common subsequence between sequences X and Y.
- $Length(\gamma)$ is the length of the reference summary.

Rouge1			
	Pegasus	**Bart**	**T5**
Mean	0.012752	0.0638	0.047
Median	0.01018	0.059105	0.04091

Table 1 Rouge 1 scores

Rouge2			
	Pegasus	**Bart**	**T5**
Mean	0.00669	0.05817	0.03959
Median	0.00391	0.05415	0.03509

Table 2 Rouge 2 scores

RougeL			
	Pegasus	**Bart**	**T5**
Mean	0.01079	0.06140	0.04502
Median	0.00912	0.05726	0.039560

Table 3 Rouge L scores

ROUGE metrics show that scores are typically between 0 and 1. Better translation quality is indicated by higher ROUGE scores, which show greater overlap between the generated summary and the reference summary. Smaller ROUGE scores indicate poorer translation quality since they suggest less precision or accuracy in the model's output when compared to the reference summary.

Bleu Score

According to Papineni, Roukos, Ward, and Zhu (2002), the BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of text which has been machine-translated from one natural language to another. The BLEU score compares the n-grams of the candidate translation with the n-grams of the reference translations and counts the number of matches. These matches are then used to calculate precision for the candidate translation.

The BLEU score is calculated as follows:

Modified Precision for n-grams:

$$P_i = \frac{Count\ Clip(matches_i, max - ref - count_i)}{candidate - n - grams_i} \#(3)$$

Where:

- Count Clips is a function that clips the number of matched n-grams ($matches_i$) by the maximum count of the n-gram across all reference translations ($max - ref - count_i$).
- $matches_i$ is the number of n-grams of order i that match **exactly** between the candidate translation and any of the reference translations.
- $max - ref - count_i$ the maximum number of occurrences of the specific n-gram of order i found in any single reference translation.
- $candidate - n - grams_i$ is the total number of n-grams of order i present in the candidate translation.

Brevity Penalty (BP):

$$BP = exp\ (1 - \frac{r}{C}) \#(4)$$

Where:

- c is the average length of the reference translations.
- r is the length of the candidate translation

Geometric Mean of Precision Scores:

$$BLEU\ Score = BP * exp\ (\sum_{i=1}^{N} (w_i * ln\ ln\ (p_i))\ ) \#(5)$$

Where:

- BP stands for Brevity Penalty
- w_i is the weight for n-gram precision of order i (typically weights are equal for all i)
- p_i is the n-gram modified precision score of order i.
- N is the maximum n-gram order to consider (usually up to 4)

Bleu			
	Pegasus	**Bart**	**T5**
Mean	1.34E-11	3.02E-07	1.92E-09
Median	2.99E-80	1.09E-15	7.71E-22

Table 4 Bleu scores

ISSN: 2473-4901
v10 n6208

A BLEU score falls between 0 and 1. Better translation quality is indicated by higher BLEU scores, which show greater overlap between the generated summary and the reference summary. Smaller BLEU scores indicate poorer translation quality since they suggest less precision or accuracy in the model's output when compared to the reference summary.

BERT Score

BERT Score is a metric for evaluating text generation quality based on BERT embeddings. It calculates the similarity between the reference and generated text at the token level using contextual embeddings from a pre-trained BERT model (T. Zhang, Kishore, Wu, Weinberger, & Artzi, 2019).

BERTScore considers precision, recall, and F1 scores based on token similarity. Lee and Toutanova (2018) provides the formula which computes the cosine similarity between the generated and reference text.

- **Token Embeddings**: Compute the contextual embeddings for each token in the reference R and candidate C texts using BERT:

$$E_R = BERT(R), \quad E_C = BERT(C) \#(6)$$

- **Cosine Similarity**: Calculate the cosine similarity between all pairs of tokens from the reference and candidate texts:

$$S_{ij} = \frac{E_R[i] \cdot E_C[j]}{\| E_R[i] \| \| E_C[j] \|} \#(7)$$

where $E_R[i]$ and $E_c[j]$are the embeddings of the i-th and j-th tokens in the reference and candidate texts, respectively.

☐ **Precision:** For each token in the candidate text, find the most similar token in the reference text:

$$P = \frac{1}{|C|}\sum_j maxS_{ij} \#(8)$$

where |C| is the number of tokens in the candidate text.

☐ **Recall**: For each token in the reference text, find the most similar token in the candidate text:

$$R = \frac{1}{|R|}\sum_j maxS_{ij} \#(9)$$

where |R| is the number of tokens in the reference text.

☐ **F1 Score:** Combine precision and recall into an F1 score:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \#(10)$$

Bert			
	Pegasus	**Bart**	**T5**
Mean	0.795566	0.858640	0.848749
Median	0.803880	0.863125	0.850725

Table 5 Bert scores

Better translation quality is indicated by higher BERT scores, which show greater overlap between the generated summary and the reference summary. Smaller BERT scores indicate poorer translation quality since they suggest less precision or accuracy in the model's output when compared to the reference summary.

	Pegasus				
	Rouge1	Rouge2	RougeL	BERT	BLEU
Mean	0.012752	0.006693	0.010794	0.079557	1.34E-11
Median	0.01018	0.003915	0.00912	0.803998	2.99E-80
	BART				
	Rouge1	Rouge2	RougeL	BERT	BLEU
Mean	0.0638	0.058179	0.061409	0.85864	3.02E-07
Median	0.059105	0.05415	0.05726	0.863125	1.09E-15
	T5				
	Rouge1	Rouge2	RougeL	BERT	BLEU
Mean	0.047	0.039595	0.045025	0.848749	1.92E-09
Median	0.0491	0.03509	0.03956	0.850725	7.71E-22

Table 6 Comparison Table

Statistical Analysis

The following tables present the results of these statistical tests and shed light on the importance of the variations between the text summarization models. The results for the difference between groups from the ANOVA test is tabulated and summarized in Table 7.

ISSN: 2473-4901
v10 n6208

ANOVA					
Metric	Sum of Squares	df	Mean Square	F	Sig.
Rouge1	0.134	2	0.067	125.983	0.0000
Rouge2	0.136	2	0.068	146.865	0.0000
RougeL	0.133	2	0.067	130.086	0.0000
Bert	0.23	2	0.115	236.834	0.0000
Bleu	0	2	0	1.908	0.1500

Table 7 ANOVA Results

A post-hoc analysis of the variance across groups has been done using the Tukey's Honest Significant Difference (HSD) Test. This test compares each pair of groups and provides the mean difference, standard error, significance level, and confidence interval. There were significant differences between all groups for rouge1, rouge2, rougeL, and bert; however, no statistical difference was found among groups based on the bleu score. The reason may be that the bleu scores were so small that statistical significance was not able to be reached.

5. DISCUSSION AND CONCLUSION

Overall, BART consistently outperforms Pegasus and T5 across all ROUGE metrics and BERTScore, indicating superior performance in capturing both content and semantic similarity in the summaries. While the BLEU scores are low for all models, BART still leads, suggesting a slight edge in n-gram precision. These results highlight the effectiveness of BART in summarizing oncology reports, with T5 performing moderately well and Pegasus lagging behind. The statistical analysis using ANOVA and Tukey's HSD provided further insight into the significance of these differences. The ANOVA results indicated significant differences between the three models, and the Tukey HSD test results show that:

- BART consistently outperforms Pegasus and T5 across ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore metrics.
- T5 also significantly outperforms Pegasus across these metrics.
- There are no significant differences in the BLEU scores between any of the models, indicating that all three models perform similarly in terms of n-gram precision.

These results align with the ANOVA findings and reinforce the conclusion that BART is the best-performing model in terms of ROUGE and BERTScore metrics, while the BLEU scores do not show significant differences between the models.

Supporting physicians and clinicians during the decision-making process is vital due to the astounding amount of both quantitative and qualitative data they encounter. While traditional methods rely heavily on human evaluation, the use of Artificial Intelligence (AI), particularly large language models (LLMs), offers a promising solution. LLMs can assist in processing the extensive qualitative data found in electronic health records, thus alleviating time constraints and enhancing the efficiency of medical professionals. This study represents a significant first step towards integrating LLMs in the processing of clinical notes, aiming to improve the overall decision-making timeline in medical practice.

6. FUTURE WORK

Future projects in this area will include a thorough validation to ensure accuracy and reliability of LLMs across oncology data and seamless integration into clinical workflows to complement existing architecture, systems and practices without disruption. Interoperability with various Electronic Health Record (EHR) systems is vital, requiring standardized interfaces for efficient data access. IT governance would be introduced to address ethical concerns and challenges that will arise from the use of AI to ensure client privacy is preserved. Exploring LLMs for real-time decision support in clinical settings could revolutionize patient care by providing instant insights. AI/IT training would be introduced to aid in a faster and widespread adoption.
Continuous learning and improvement algorithms should be developed to keep LLMs updated with the latest oncology data and information.

7. REFERENCES

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut,

K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268. https://doi.org/10.48550/arXiv.1707.02268

Batra, P., Chaudhary, S., Bhatt, K., Varshney, S., & Verma, S. (2020). A review: Abstractive text summarization techniques using NLP. Paper presented at the 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM). https://doi.org/10.1109/ICACCM50413.2020.9213079

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Paper presented at the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. https://doi.org/10.1145/3442188.3445922

Bhatia, N., & Jaiswal, A. (2015). Trends in extractive and abstractive techniques in text summarization. International Journal of Computer Applications, 117(6).

Chen, T., Allauzen, C., Huang, Y., Park, D., Rybach, D., Huang, W. R., . . . Moreno, P. J. (2023). Large-scale language model rescoring on long-form data. Paper presented at the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP49357.2023.10096429

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Socher, R. (2021). Deep learning-enabled medical computer vision. NPJ digital medicine, 4(1), 5. DOI: https://doi.org/10.1038/s41746-020-00376-2

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation, 101(23), e215-e220. DOI: https://doi.org/10.1161/01.CIR.101.23.e215

Lee, J., & Toutanova, K. (2018). Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 3(8).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Paper presented at the Text summarization branches out.

Montgomery, D. C. (2001). Design and analysis of experiments, John Wiley & Sons. Inc., New York, 1997, 200-201. https://doi.org/10.1007/0-387-22634-6_15

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Radhakrishnan, L., Schenk, G., Muenzen, K., Oskotsky, B., Ashouri Choshali, H., Plunkett, T., Butte, A. J. (2023). A certified de-identification system for all clinical text documents for information extraction at scale. JAMIA open, 6(3), ooad045. https://doi.org/10.1093/jamiaopen/ooad045

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

Savova, G. K., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D. S., Warner, J. L. (2019). Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer research, 79(21), 5463-5470. https://doi.org/10.1158/0008-5472.CAN-19-0579

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Pfohl, S. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180. https://doi.org/10.48550/arXiv.2212.13138

Sushil, M., Kennedy, V. E., Mandair, D., Miao, B. Y., Zack, T., & Butte, A. J. (2024).

ISSN: 2473-4901
v10 n6208

CORAL: expert-curated oncology reports to advance language model inference. NEJM AI, 1(4), AIdbp2300110. DOI: 10.1056/AIdbp2300110

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Seehofnerova, A. (2023). Clinical text summarization: adapting large language models can outperform human experts. Research Square. doi: 10.21203/rs.3.rs-3483777/v1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems. Advances in neural information processing systems, 30(2017).

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. Paper presented at the International conference on machine learning.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675. https://doi.org/10.48550/arXiv.1904.09675