

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 19, No. 1
April 2026
ISSN: 1946-1836

In this issue:

- 4. One Health: AI in Healthcare Summarization**
Chiazam Izuchukwu, Georgia Southern University
Hayden Wimmer, Georgia Southern University
Loreen Marie Powell, Marywood University
- 15. Architecting the Academic Assistant: A Review of Dual-Processing Architectures and Hybrid Chatbot Design in Higher Education**
Sai Prakash Reddy Rao, Kennesaw State University
Lei Li, Kennesaw State University
Hisham Haddad, Kennesaw State University
Selena He, Kennesaw State University
- 26. A Framework for Addressing Research Gaps in AI-Driven Classrooms: Toward Equity, Ethics, and Sustainable Innovation**
Mark Pisano, Southern Connecticut State University
- 33. Does Dark Humor Work? An Inspection of Social Media-Based Marketing Strategies**
Matthew Bagatta, Walsh University
Philip Kim, Walsh University
Richard Metzger, Robert Morris University
- 46. Designing An Artificial Intelligence Adoption Assistance Platform for Small and Medium-sized Enterprises: Identifying Opportunities and Concerns**
Brian Alvarez, University of North Carolina Wilmington
Lukas Rudolf Germut Fitz, Brandenburg University of Technology Cottbus-Senftenberg
Judith Gebauer, University of North Carolina Wilmington
Yao Shi, University of North Carolina Wilmington
- 64. Developing Florida Digital Divide Index: A Comprehensive Analysis of Internet Accessibility and Socio-economic Characteristics**
Tushya Vermuri, Southeastern Grocers
Karthikeyan Umapathy, University of North Florida

The **Journal of Information Systems Applied Research and Analytics** (JISARA) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is four issues a year. The first date of publication was December 1, 2008. The original name of the journal was Journal of Information Systems Applied Research (JISAR).

JISARA is published online (<https://jisara.org>) in connection with the ISCAP (Information Systems and Computing Academic Professionals) Conference, where submissions are also double-blind peer reviewed. Our sister publication, the Proceedings of the ISCAP Conference, features all papers, teaching cases and abstracts from the conference. (<https://iscap.us/proceedings>)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%) and other submitted works. The non-award winning papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in JISAR. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisara.org or the publisher at publisher@jisara.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISARA.

2026 ISCAP Board of Directors

Amy Connolly
James Madison University
President

Michael Smith
Georgia Institute of Technology
Vice President

Jeff Cummings
Univ of NC Wilmington
Past President

David Firth
University of Montana
Director

Mark Frydenberg
Bentley University
Director/Secretary

Leigh Mutchler
James Madison University
Director

RJ Podeschi
Millikin University
Director/Treasurer

Bryan Reinicke
Rochester Institute of
Technology / Director

Jeffrey Babb
West Texas A&M University
Director/Curricular Matters

Eric Breimer
Siena University
Director/2026 Conf Chair

Tom Janicki
Univ of NC Wilmington
Director/Meeting Planner

Xihui "Paul" Zhang
University of North Alabama
Director/JISE Editor

Copyright © 2026 by Information Systems and Computing Academic Professionals (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to Scott Hunsinger, Editor, editor@jisara.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

Scott Hunsinger
Senior Editor
Appalachian State University

Thomas Janicki
Publisher
University of North Carolina Wilmington

2026 JISARA Editorial Board

Biju Bajracharya
East Tennessee State University

Jason Price
Nichols College

Queen Booker
Metro State

Bryan Reinicke
Rochester Institute of Technology

Wendy Ceccucci
Quinnipiac University

Asish Satpathy
Arizona State University

Biswadip Ghosh
Metro State University

Dana Schwieger
Southeast Missouri State University

Russell Haines
Appalachian State University

Jeff Strain
Brigham Young University - Hawaii

Melinda Korzaan
Middle Tennessee State University

Katarzyna Toskin
Southern Connecticut University

Will Ledbetter
Perdue University

Karthikeyan Umapathy
University of North Florida

Li-Jen Lester
Sam Houston State University

Hayden Wimmer
Georgia Southern University

Muhammed Miah
Tennessee State University

David Woods
University of Miami Regionals

Alan Peslak
Penn State University

David Yates
Bentley University

One Health: AI in Healthcare Summarization:

Chiazam Izuchukwu
ci02061@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

Hayden Wimmer
hwimmer@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

Loreen Marie Powell
lpowell@maryu.marywood.edu
Marywood University
Scranton, PA18509

Abstract

The research uses a One Health text document as our main dataset. Two summarization methods were employed in this study, with the first being recursive (summary of the summary), which was repeated twice to achieve a progressively concise word count, and the second a controlled and direct summary of the original document, which was repeated to accomplish the same set word count. Performance metrics like ROUGE, BLEU, and BERT scores were calculated to assess the effectiveness of both methods. This study goes beyond automated metrics by incorporating human evaluation and assessing readability and coherence with the original document to ensure a qualitative validation of the results. Additionally, the results from all summaries generated provide a comparative analysis between traditional LLM-based summarization and conversational AI establishing performance baselines for validating clarity and coherence for effective summarization of healthcare documents.

Keywords: Large language models, Clinical summarization, Healthcare AI, Natural language processing, Data preprocessing, Machine learning

Recommended Citation: Izuchukwu , C., Wimmer, H., Powell, L.M., (2026). One Health: AI in Healthcare Summarization. *Journal of Information Systems Applied Research and Analytics*, v19(n1) pp 4-14. DOI# <https://doi.org/10.62273/TDBT3523>

One Health: AI in Healthcare Summarization

Chiazam Izuchukwu, Hayden Wimmer and Loreen Marie Powell

1. INTRODUCTION

AI in healthcare is improving ways to organize and interpret qualitative data. Large language models (LLMs) have demonstrated remarkable skill in dealing with structured narrative information such as this. LLMs are designed to produce short and fluent summaries from large text corpora, a property that can be highly useful for health informatics. Healthcare documents often include elaborate medical protocols, reported research findings, and clinical narratives, all of which can greatly benefit from effective summarization to support both information retrieval and informed decisions. Such LLMs designed for summarization tasks include BART (Bidirectional and Auto-Regressive Transformers). Considering the robust performance of BART in text generation tasks, it might be a preferable choice for summarizing challenging healthcare documents such as the One Health document used in this study. One Health initiative emphasizes the interconnectedness of human, animal, and environmental health, resulting in detailed reports that require efficient summarization to support research and policymaking.

We used two different approaches to measure the performance of the summarization in this study. The first approach is BART-specific, which contrasts two different summarization techniques: (1) a recursive technique, where a further summary of summaries is made to compress and condense, and (2) a direct summarization technique, where only one single summary is generated within a set constraint specifying that length of the output. The second method is to compare BART against conversational AI models such as ChatGPT and NotebookLM. This analysis compares BART summarization performance against the conversational models with respect to quality and credibility. The evaluation takes into account typical evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), and BERT scores that capture the recall, precision, and other semantic similarity of the generated summaries.

This work directly addresses the usage and

deployment of LLMs for a healthcare context by evaluating the performance of BART on various summarization methods and its performance relative to conversational artificial intelligent (AI) models. It demonstrates the potential of recursive summarization methods, explores the capabilities of traditional LLM versus conversational AI, and informs assimilation AI design to facilitate summarizing complex health data, elevate information retrieval, and support decision-making.

2. LITERATURE REVIEW

This is the first systematic review to explore the application of AI in healthcare and the challenges faced during its implementation. AI technologies like machine learning (ML) and natural language processing (NLP) are revolutionizing healthcare through the help of diagnostic tools, individualized treatment plans, patient monitoring devices and operational tweaking too. However, integrating AI into healthcare systems has many challenges such as data privacy, ethical and legal issues, interoperability to standards-compliant databases, scalability in training and operating conditions, and human-system interaction. A generation ago, scholars wrote the ethical guidelines and legal frameworks that allowed for interoperable systems, developing similar robust cybersecurity tools that should include universally shared standards, as well as, software to streamline the deployment of these practices by nonexperts. Secinaro et. al (2021) called for increased collaboration across disciplines and better education programs for healthcare providers to promote AI as an integrated tool alongside efforts with ongoing research development.

The DialoGPT, intended for chatbot purposes, is a neural conversational response generation model designed to generate replies similar to human conversation. Specifically, Y. Zhang et al. (2019), used the Hugging Face PyTorch transformer to work on 147 million conversation-like exchanges from Reddit spanning 2005–2017. It performs at near-human levels in single-turn dialogue for both automatic and human evaluations. DialoGPT obtains the number of related, context-relevant, and highest-quality replies as baseline systems.

They also released the pre-trained model and training pipeline to facilitate future research in neural response generation and open-domain dialogue systems.

Su, et al. (2022) investigated the use of automatic summarization with deep neural networks to lighten the load of physicians in writing documentation. The authors created an "extract-and-abstract" approach for crafting the History of Present Illness (HPI) section in a clinical note through BART. Similarly, a classifier was trained to predict whether each utterance belongs to the clinical section of interest, and a subset of correlated utterances was returned for summarization. They noted other types of filters to address the token limit of the BART model. Their approach can complement ROUGE scores with medical concept extraction and be scalable.

Radford et al. (2018) studied the practicality of a possible solution for better transfer learning using large unlabeled text corpora by suggesting some generative pre-training approach and task-specific fine-tuning. They repeated these experiments for multiple benchmarks. Results were consistent in that the proposed task-agnostic model outperformed models trained to solve a given problem. Among their results, these improvements included an 8.9% increase in accuracy for logical reasoning, a 5.7% improvement for question-answering tasks, and a 1.5% performance on textual entailment. Evaluating the method across multiple benchmarks demonstrates the model's ability to achieve state-of-the-art performance on a diverse set of NLP understanding tasks.

3. TECHNICAL BACKGROUND

LLMs are transformative technologies in NLP. BART is known as encoder-decoder model, primarily designed for NLP generation tasks. Typically, this kind of model utilizes vast datasets and advanced ML 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 (Vaswani et al., 2017).

Brown et al. (2020) aimed to close this gap and investigate how scaling of Language Models affects broadly generalizable, few-shot performance. They took advantage of the language model GPT-3, one with 175 billion parameters, which is an order of magnitude larger than the predecessors, such as BERT and

XLNet, for few-shot absence by gradient-based optimization or fine-tuning. The approach focused on showing tasks to GPT-3 using only text-based interaction, measuring performance across translation pairs and question answering, as well as cloze-style fill-in-the-blank tests that required reasoning and adapting models from different domains. The test revealed that GPT-3 was proficient at a wide range of NLP tasks and could generate human-like text, but it struggled on certain datasets or specific civil issues with training on large web corpora.

BART is a denoising autoencoder that is used to pre-train sequence-to-sequence models by leveraging the Transformer-based encoder-decoder architecture. It perturbs the text using a noising function and then trains a model to reconstruct the original text. Because it uses both left-to-right generation and a bidirectional encoder, BART generalizes previous models like BERT (due to its bidirectional encoder) and GPT (due to its decoder). It gets impressive results with only random shuffling of sentences and a simple in-filling procedure where spans of text are replaced by a single mask token. When fine-tuned for text generation tasks, Lewis et al. (2019) reported that BART showed remarkable comprehension and generation performance. On GLUE and SQuAD, it matches RoBERTa performance (and achieves new state-of-the-art results for abstractive dialogue, question answering, and summarization). BART also yields a 1.1 BLEU point improvement over the back-translation baseline with only pretraining in the target language.

The Transformer architecture by Vaswani et al. (2017) suggest that it is possible to implement a sequence-to-sequence model without using RNNs and replace them with attention mechanisms only. Such an approach also facilitates parallelization and validation of the model tuned in this way much faster. The model achieves 28.4 BLEU on the English-to-German and 41.0 BLEU on the English-to-French translation, outperforming all the previously reported single models, at a fraction of their size. On the English-to-German task, it improves over the best ensemble by more than 2 BLEU points. The model is more efficient, as shown by reaching these scores at a fraction of the training cost in 3.5 days on eight GPUs for the English-to-French task.

After it has been found that using NLP models directly on the biomedical domain text is far from optimal due to different expressions specific to the biomedical language. Biomedical

Text Mining has become an even more promising area as it runs in parallel with the huge amount of scientific data. In response, Lee et al. (2020) introduced BioBERT, a domain-specific pre-trained language model. BioBERT outperforms BERT and previous state-of-the-art models by a notable margin in biomedical named entity recognition, relation extraction, and question answering. It significantly improves F1 score and MRR on these tasks, showing the effectiveness of domain-specific pre-training. Their study argues that pre-training BERT on biomedical texts can give it a head start in understanding the intricate languages in this domain.

Clinical responses to big textual data, for example, in electronic health records (EHR), are time-consuming. Introduction to LLMs have proven effective in NLP. However, their ability to summarize clinical information across various tasks is unclear. Van Veen et al. (2023) explored the behavior of eight LLMs using domain adaptation strategies on six datasets, which include radiology reports, patient queries, and progress notes for model performance. By comprehensive quantitative analysis, the work points out model-adaptation trade-offs and cases where more modern LLMs may not be better than previous ones. The best LLMs returned summaries that human evaluators rated complete and more accurate than those written by humans in a clinical reader study with 10 MDs. A follow-up qualitative analysis identified pain points common to both LLMs and human summarizers, while traditional NLP metrics were mapped against physician preferences for better-grounded scores. This is the first study claiming that having LLMs available can allow one to do better on a large number of tasks in clinical summarization compared with human experts and possibly reduces documentation burdens, allowing clinicians greater capacity for providing personalized care.

Improvements in large pre-trained neural networks such as BERT have changed how NLP is performed, and researchers are continually exploring what kind of linguistic knowledge these models can pick up from unlabeled data. Most of the previous works have explored the analysis of model outputs and internal representations. At the same time, Clark et al. (2019) instead focuses more on BERT in terms of its attention mechanisms. The methodology includes assessing the behaviors of attention heads, such as delimiter-token-focused attention or concentrating on neighboring phrases and

sentence-level broader attention. Some heads that correspond very well to linguistic notions such as syntax and coreference, for example, by properly recognizing the direct object or a reference of determiners are noteworthy. They further contribute with an attention-based probing classifier that shows, in addition to being improved by BERT features on SRL and constrained dependency parsing tasks, this transformer is capturing a lot of syntactic pressure. The results highlight BERT's capacity to learn complex syntactic structures via attention, providing strong support for tuning and adaptation in NLP.

4. METHODOLOGY

In this study, we performed our summarization in Google Collab using Python-based libraries namely hugging face transformer, NLTK and some other utility packages for text preprocessing and metric calculations. Because the model summarization and evaluation were carried out in a free cloud-based environment with limited GPU and RAM support, the text data was first converted from pdf to txt document to ensure efficient processing of the large data. Figure 3.1 displays the process flow.

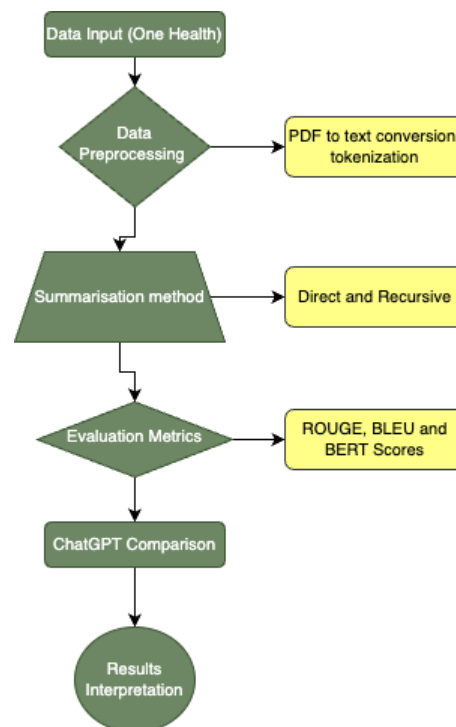


Figure 3.1: Process flowchart

Dataset

The dataset used in this study is a One Health Joint plan of action (Organization, Programme, & Health, 2022). The Quadripartite Organizations—the FAO, UNEP, WOA, and WHO work around the globe with a single health approach that addresses threats to human health at the interface between people-animal-plant environment. In response to international calls for pandemic prevention and sustainable health promotion, they developed the One Health Joint Plan of Action 2022-2026 an 86-page document dedicated to this integrated approach to management.

The document was chosen because of its qualitative nature, relevance in health informatics and how much detail it contains. The document was first converted from PDF to text file to enable preprocessing using NLP tools. Preprocessing was then applied to the dataset to include tokenization, stop word removal and normalization to ensure compatibility with our preselected LLM.

BART

BART is a sophisticated LLM developed by Facebook AI which combines the positives of BERT bidirectional encoding with GPT's autoregressive decoding, making it very efficient in a range of NLP tasks (Lewis et al., 2019). The framework of this LLM allows it to autofill missing information and generate coherent information.

This phase involved the utilization of two different strategies, which are highlighted as follows.

- Direct Summarization: We obtained the direct summarization using the BART model. The goal for this method was to result in a shorter version of the document yet within bounded restrictions on how many words could be included as part of the summary and having some control over what information would stay or what we considered essential.
- Recursive Summarization (Summary of Summary): Here, the initial OG summary is made with the BART model. The resulting summary is then further summarized again to achieve a more concise output. This recursive approach aimed to maximize brevity while retaining essential information.

Evaluation Metrics

ROUGE, BLEU, and BERT score metrics are employed to assess the effectiveness of the summarization strategies. The evaluation of

these models is crucial otherwise, we would have to depend solely on human evaluation which can be a more time consuming and costly path.

ROUGE Score

ROUGE scores were calculated to evaluate the overlap of n-grams and word sequence between the generated and referenced summaries and between the generated summaries and OG. Lin (2004) stated that ROUGE 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 NLP tasks to measure the similarity between generated and reference summaries.

BLEU Score

BLEU scores were calculated to ascertain the precision of word overlap, which sheds insight into the accuracy of the generated summary. According to Papineni et al. (2002), the BLEU score is a metric used to evaluate the quality of text that has been machine-translated from one natural language to another. The BLEU score compares the n-grams of the candidate translation's n-grams with the reference translations and counts the number of matches. These matches are then used to calculate precision for the candidate translation.

BERT Score

The BERT Score was calculated to evaluate the semantic relationship between the summary and the OG, offering more insights into how well the summary captured the main themes of the text dataset. 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 (Zhang, et. al., 2019). BERTScore considers precision, recall, and F1 scores based on token similarity. Devlin et al. (2018) provided the formula that computes the cosine similarity between the generated and reference text.

ChatGPT & NotebookLM Comparison

This benchmark provides an evaluation of ChatGPT and Google NotebookLM versus a BART model looking at summarization in terms of quality, brevity, and fidelity. BART is a transformer model on which the training data is fine-tuned, and it can generate summaries of the text. Still, due to its structured nature, where the input and output are sampled from similar distributions, which work great for structured datasets, it does not generalize well to many

other contexts. ChatGPT, a conversational AI, can solve more versatile tasks and summarize them in an engaging way that sounds natural to humans. Google NotebookLM is another conversational AI application for personal knowledge management that improves summarization thanks to context awareness and document amalgamation based on uploaded files. ChatGPT and NotebookLM are also as powerful but more versatile, allowing for more nuanced output compared to BART, which returns a clear and concise summary. We compare model performance on both frameworks, illustrating the strengths and trade-offs of traditional LLM vs conversational AI models to inform tool selection according to the summarization use case.

Human Evaluation

Human evaluation was provided as manual assessment of the generated summaries and their agreement with OG quality, beyond automated metrics. The ratings were done using a Likert scale of 1–7, where 1 strongly agreed that the summary was of high quality and 7 strongly disagreed that the document was of high quality. The evaluation was carried out by 24 one & public health professionals, via an IRB approved study, who were all recruited through a prominent one health conference via voluntary participation. The data obtained from the evaluation is used to perform various statistical analysis to gain more insights. The models evaluated are defined in the table 3.1, with recursive summarization referring to the summary of a previous summary in iteration until the desired length is achieved and direct summarization referring to summarizing the original text directly to the desired length without any iterations. ChatGPT and NotebookLM represent conversational AI-based summarization systems.

GROUP	SUMMARY TECHNIQUE
1	Recursive
2	Direct
3	ChatGPT (OpenAI)
4	NotebookLM

Table 3.1: Grouping of Summarization Methods

Descriptive Statistics

We provided descriptive statistics to summarize key characteristics of the responses generated by the human evaluation. We computed the following metrics:

- Mean (μ): The average score of the responses given to the group, where lower

mean scores in the evaluation signified that reviewers found the summaries to be highly aligned with the original content, while higher mean scores suggested a weaker quality.

- Standard Deviation (SD) (σ): Quantifies the spread of summary performance across methods.
- Median: A value that represents a central point in the distribution of scores and is more resistant to extreme scores than the mean, thus making it a further alternative measure of central tendency.
- Minimum and Maximum Values: Identifies the score range, helping detect extreme values.

One-Way ANOVA

In this study, Analysis of Variance (ANOVA) was used to assess whether there are statistically significant differences in the performance of multiple summarization methods. The study compared the performance of multiple summarization techniques like recursive summarization, direct summarization, ChatGPT, and Google NotebookLM. Due to the nature of the comparison, one-way ANOVA was the suitable statistical test for analyzing variation in performance. Prior to running ANOVA we confirmed homogeneity of variance via Levene's test statistic. The null hypothesis (H_0) and the alternative hypothesis (H_1) was as follows:

- H_0 : There is no statistically significant difference in mean performance between different summarization approaches.
- H_1 : There is a statistically significant difference in mean performance between different summarization approaches.

Post-Hoc Analysis

Where ANOVA indicated a significant difference, Tukey's Honest Significant Difference (HSD) test was conducted to determine which summarization methods differed significantly from each other. This test is done to confirm the differences detected were not attributable to random variation but were real differences in summarization quality.

5. RESULTS

We utilized three evaluation metrics for each method, ROUGE (n-gram overlap), BLEU (word precision) and BERTScore (semantic similarity). All these metrics give a robust view of individual performance of each summarization approach with respect to content retention, brevity, and semantic similarity.

Table 4.1 investigates a recursive summarization process in which the text is summarized at successive stages and the quality at each stage is evaluated based on standard metrics. The comparison begins by evaluating the initial summary against the original (OG) text, followed by comparing the 2000-word summary to the initial (IN) summary (sum), and finally, comparing the 145-word sum to the 2000-word version.

Recursive Summarization					
Comparison	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT-F1
IN sum vs OG	0.76	0.66	0.64	0.45	0.86
2000 vs IN	0.19	0.17	0.16	0.0	0.82
145 vs 2000	0.13	0.13	0.12	3.20 E-06	0.86

Table 4.1: Recursive Method

Table 4.2 is like table 4.1 except the comparison involves evaluating all the summaries 2000-word, and 145-word versions against the OG.

Recursive Summarization					
Comparison	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT-F1
2000 vs OG	0.14	0.13	0.13	3.16 E-06	0.80
145 vs OG	0.01	0.01	0.01	1.53 E-79	0.79

Table 4.2: Recursive Summarization (Sum vs OG)

Table 4.3 investigates a direct summarization process in which the text is summarized within bounded restrictions and is evaluated based on standard metrics. The comparison involves evaluating all the summaries 2000-word, and 145-word versions against the OG.

Direct Summarization					
Comparison	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT-F1
2000 vs OG	0.139338	0.130013	0.132797	4.16E-06	0.815139
145 vs OG	0.011799	0.011145	0.010853	1.51E-79	0.814662

Table 4.3: Direct Method

Fig 4.1 compares of direct and recursive

summarization performance across ROUGE, BLEU, and BERT-F1 metrics at two levels of compression (2000 vs OG and 145 vs OG). Results show minimal differences in ROUGE and BLEU, with direct summarization slightly outperforming recursive summarization on BERT-F1.

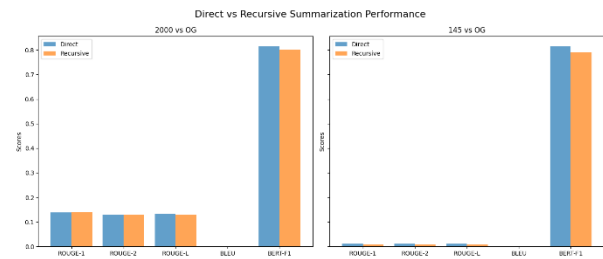


Fig 4.1: Performance Comparison of Recursive and Direct Summarization

Table 4.4 investigates a ChatGPT summarization process in which the text is summarized based on AI prompts and evaluated based on standard metrics. The comparison involves evaluating all the summaries IN, 2000-word, and 145-word versions against the OG.

ChatGPT Summarization					
Comparison	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT-F1
IN vs OG	0.03	0.00	0.02	3.98 E-25	0.66
145 vs OG	0.02	0.00	0.01	8.58 E-25	0.66
2000 vs OG	0.03	0.00	0.02	4.53 E-25	0.66

Table 4.4: ChatGPT Method

Table 4.5 investigates a notebookLM summarization process in which the text is summarized based on AI prompts and evaluated based on standard metrics. The comparison involves evaluating all the summaries 145-word versions against the OG.

NotebookLM Summarization					
145 vs OG	0.01	0.01	0.01	1.22 E-85	0.81

Table 4.5: NotebookLM Method

Comparison of summarization approaches (Recursive vs Direct)

ROUGE score

Recursive summarization presents relatively higher ROUGE scores, especially in ROUGE-1, ROUGE-2, and ROUGE-L. It seems the recursive approach preserves word and phrase overlap better.

Direct summarization has comparatively lower ROUGE scores, especially when directly summarizing down to 145 words. This corresponds to the fact that with increasing direct summarization, the method accounts for fewer word overlaps with the original text, resulting in a loss of accuracy.

BLEU score

Minimal BLEU scores over recursive and direct methods allude to low precision and direct word overlap. On the other hand, when summarizing in smaller steps, recursive summarization achieves even slightly higher BLEU scores, which corresponds to better alignment in word choices. The BLEU score of the direct summarization approach is especially poor at the 145-word count, indicating that condensing such a large input corpus into one output token causes a loss in accuracy.

BERT score

Both methods resulted in a reduced BERT-F1 score after each iteration, however, recursive summarization yielded a consistently higher score at intermediate steps (145 vs 2000 at 0.857962), suggesting that the recursive method can preserve semantic meaning across iterations more effectively. While direct summarization retained a stable but lower BERT-F1 score compared to recursive summarization, this could suggest a trade-off between conciseness and semantic preservation.

Comparison with ChatGPT Summary (Bart vs ChatGPT)

ROUGE and BLEU score

The ROUGE and BLEU scores for ChatGPT summaries are the lowest across all our comparisons. The scores are particularly low for ROUGE-2 and BLEU, signifying that there is minimal overlap in phrasing and word choice with the OG. This implies that ChatGPT summaries could focus more on fluency than text fidelity. The differences in ROUGE and BLEU scores between ChatGPT and both BART-based summarization methods suggest that ChatGPT may not be as detail-focused when it comes to summarization, focusing more on synthesizing

overall content.

BERT score

Across recursive and direct summarization BERT-F1 scores, ChatGPT's performance is below them with an average score of ~0.660550. This shows that ChatGPT holds high semantic similarity but cannot preserve the original meaning and context as well as BART-based models.

Comparison with NotebookLM Summary (Bart vs NotebookLM)

The summary from the NotebookLM yields lower BLEU and ROUGE scores compared to recursive and direct summarization approaches. Still, with a comparable BERT score, the Notebook approach does not replicate the original text very closely in terms of precision and fidelity but is somewhat aligned in terms of semantics. Like ChatGPT summaries, the NotebookLM method is not best suited when keeping the generated text very close to the OG. However, it is impactful when a general conceptual summary is acceptable.

Statistical Analysis

The descriptive statistics give some vital information on the relationship between summarization methods and OG. The mean ratings illustrated that Group 3 had the lowest average score (2.33), which implied that the summaries generated in this group were considered the highest quality with the document being summarized. In contrast, Group 2 had the lowest mean (5.5), indicating that reviewers considered the summaries in this group to be of lower quality. Table 4.6 provides the descriptive statistics.

Descriptive Statistics						
Data	N	Mean	Std. Dev	Std. Error	Min	Max
1	6	3.67	1.51	0.62	2	5
2	6	5.5	1.98	0.81	3	7
3	6	2.33	1.37	0.56	1	5
4	6	2.67	1.03	0.42	1	4
Total	24	3.54	1.89	0.39	1	7

Table 4.6: Descriptive Statistics

The SD was used to illustrate variability in the ratings. Group 2's scores were the most heterogeneous (SD = 1.98), indicating that participants' ratings were most spread. Meanwhile, Group 4 recorded the lowest SD

(1.03), implying that group's ratings were more homogeneous, as consensus was reached by the participators concerning quality level. The minimum and maximum values tell us some of the difference between the groups, with Group 3 reporting the lowest rated response (1), indicating strong alignment with the OG, and Group 2 provided a rating (7) indicating a weak quality.

An ANOVA test was conducted to determine if these ratings exhibited statistical significance. The p-value was obtained as 0.01, which is less than 0.05 significance value. This means that at least one group's ratings differed significantly from the others, which implies that some summarization methods produce summaries like the OG than others. Table 4.7 displays the results of ANOVA test.

ANOVA					
Data	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	36.46	3	12.15	5.34	0.01
Within Groups	45.50	20	2.28		
Total	81.96	23			

Table 4.7: ANOVA

To determine specific differences between groups a Tukey HSD post-hoc test was conducted. The results (Table 4.8) showed that Group 2 ratings were significantly higher than Groups 3 ($p = 0.01$) and 4 ($p = 0.02$) as evidenced by a lower mean (where 1 = strongly agree in a 7 scale Likert format). Since a higher mean score signifies weaker agreement, this confirms that Group 2's summaries were significantly rated with less agreement than those from Group 3 and Group 4. Nonetheless, there were no significant differences among Groups 1, 3 and 4 indicating that the summarization methods employed in these cases led to summaries whose quality were similar.

(I) Group	(J) Group	Mean Diff (I-J)	Std. Error	Sig.
1	2	-1.83	0.87	0.19
	3	1.33	0.87	0.44
	4	1.00	0.87	0.67
2	3	3.17*	0.87	0.01
	4	2.83*	0.87	0.02
3	4	-0.33	0.87	0.98

Table 4.8: TUKEY

6. DISCUSSION

The results of this work demonstrate the empirical performance profiles of recursive summarization, direct summarization and ChatGPT summarization on complex qualitative health documents. While they all provide different advantages and disadvantages, each have trade-offs between content fidelity, semantic coherence and brevity of the summaries they produce. Semantic Fidelity: Adding to the BERT-F1 scores, recursive summarization proves to be the best method for maintaining the essential semantics of the OG whilst continuously shortening the text. This is useful for large documents where multiple passes of compression are needed to retain the overall context.

Word Repetition and Structure: direct summarization struggles to retain exact word usage and structure, even with short lengths. This is equally apparent in the drop in ROUGE and BLEU scores in the second stage of summarization. On the other hand, recursive summarization has a more gradual decrease in ROUGE scores, suggesting that content overlap is preserved better with iterative summarization. By contrast, the conversational AI (ChatGPT and NotebookLM) tends to preserve less focus on the text due to lower ROUGE and BLEU scores. This suggests that conversational AI models generally have wider applicability but may not be the best choice for high-fidelity summarization tasks compared to models such as BART that better specialize in structured forms of compression. These findings are further corroborated by the combination of statistical analysis and human observation. Group 3 (ChatGPT) received the most similar to original text, followed by Group 4 (NotebookLM) and finally Group 2 (Direct Summarization) obtained the highest mean, indicating lower quality. The results show that recursive summarization can maintain the main ideas of the source document much better than direct summarization, which tends to introduce more noise or information loss. While the AI models were able to summarize well enough in a general way, it may still not be precise enough for accurate technical summarization tasks.

Some recommendations to improve summarization in health informatics can be made based on the findings from this study. If the health document is extensive and lengthy, then recursive summarization, which iteratively shortens the content, can be performed to maintain key information. Another possibility

comes from combining these approaches and using a hybrid model that includes recursive and direct summarization to achieve some balance between compression/absorption and length maintained within the summary, where summaries of different lengths but still potentially varying levels of detail about the source sentences are required. ChatGPT and NotebookLM is less accurate but useful for high-level summaries for non-technical audiences. To improve content fidelity, it is recommended that BART be fine-tuned on health-specific data.

7. CONCLUSIONS

Recursive summarization holds promise for creating ultra-short yet informative summaries of qualitative health data, so this study highlights its potential. The mechanism used in this method folds the content iteratively in a way that ensures shorter text while still retaining information, making it ideal for intricate and content-rich documents. Although direct summarization is more efficient with small compressed content, it can lose more information and coherence. On the other hand, ChatGPT summarization provides a general tool but may not offer the precision needed for clinical high-fidelity technical summarizations.

These results indicate that recursive summarization with LLMs such as BART produces concise and meaningful summaries for health-related documents outperforming direct summarization method in preserving meaning for complex content. With fine-tuning, models such as ChatGPT or Google's T5 may offer equally strong or superior performance. This paper is not without limitations. However, it could usefully guide future work in investigating the hybridization of recursive and direct summarization methods or domain-centric tuning to improve summary quality.

8. REFERENCES

- Brown, T. B.; Mann, B., Ryder, B., Subbiah, M. Kaplan, J. Dhariwal, P. et al (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing System*. 159, 1877 - 190. <https://dl.acm.org/doi/pdf/10.5555/3495724.3495883>
- Clark, K., Khandelwai, U., Levy, O., Manning, C. D. (2019). What does Bert look at? An analysis of Bert's attention. <https://arxiv.org/abs/1906.04341>
- Devlin, J. Chang, M-W., Lee, K., & Toutanova, K. (2018). Pre-training of deep bidirectional transformers for language understanding. <https://arxiv.org/abs/1810.04805v1>
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries*. Paper presented at the text summarization branches out. Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL, 74-81, Barcelona, Spain, 25 July 2004. <https://www.aclweb.org/anthology/W04-1013.pdf>
- Organization, W. H., Programme, U. U. N. E., & Health, W. O. f. A. (2022). *One health joint plan of action (2022-2026): working together for the health of humans, animals, plants and the environment*. World Health Organization. <https://www.who.int/publications/i/item/9789240059139>
- 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. 311 - 318. <https://doi.org/10.3115/1073083.1073135>
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training. <https://cdn.openai.com/research-covers/language->

- unsupervised/language_understanding_paper.pdf
- Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. *BMC medical informatics and decision making*, 21, 1-23. <https://doi.org/10.1186/s12911-021-01488-9>
- Su, J., Zhang, L., Hassanzadeh, H. R., & Schaaf, T. (2022). *Extract and Abstract with BART for Clinical Notes from Doctor-Patient Conversations*. Paper presented at the Interspeech, Incheon, Korea. https://www.isca-archive.org/inter-speech_2022/su22b_interspeech.pdf.
- Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerova, A., Rohatgi, N. Hosamani, P., Collins, W. Ahuja, N., Langlotz, C., Hom, J. Gatidis, S. Pauly, J. Chaudhari, A. (2023). Clinical text summarization: adapting large language models can outperform human experts. *Research Square*. <https://doi.org/10.21203/rs.3.rs-3483777/v1>
- Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). *Attention is All you Need*. 31st Conference on Neural Information Processing System Long Beach, California. <https://dl.acm.org/doi/pdf/10.5555/3295222.3295349>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arxiv.org/abs/1904.09675
- Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. <https://arxiv.org/pdf/1911.00536>