Distributed Biomedical Text Summarization and Knowledge Discovery Using Generative Language Models and Apache Spark

Harshith Sai
Computer Science and Engineering
Dayananda Sagar University, Harohalli, Bangalore, India
harshithsaivadlakunta@gmail.com

Bhanu Prakash
Computer Science and Engineering
Dayananda Sagar University, Harohalli, Bangalore, India
bhanuprakash9418@gmail.com

Vema Sai Krishna
Computer Science and Engineering
Dayananda Sagar University, Harohalli, Bangalore, India
vemasaikrishna96@gmail.com

Pranav Aditya
Computer Science and Engineering
Dayananda Sagar University, Harohalli, Bangalore, India
pranavaditya45@gmail.com

Tanvir H Sardar
Computer Science and Engineering
Dayananda Sagar University, Harohalli, Bangalore, India
tanvir-cse@dsu.edu.in

***Abstract*— Biomedical literature is rapidly increasing exponentially, and it constitutes a severe challenge that healthcare researchers, clinicians and data scientists must manage by extracting meaningful insights within large amounts of unstructured text. In this paper, a scalable, end-to-end Artificial Intelligence (AI) pipeline that can be used to distribute the workload of biomedical text summarization and automated knowledge discovery is presented, using Apache Spark as a distributed data ingestion and processing engine, domain-adapted transformer-based Large Language Models (LLMs) as abstractive summarization engines, and Latent Dirichlet Allocation (LDA) as an unsupervised topic modeling engine. It was tested on the corpus of more than 5,000 biomedical abstracts of PubMed-Medline. The pipeline itself took 9-12 minutes to run on Apache Spark Pandas UDFs and Falconsai/medical_summarization based on the ability to run an entire workflow 6x faster than the sequential execution. The evaluation by ROUGE showed that there was great semantic retention with ROUGE-1: 0.58, ROUGE-2: 0.41 and ROUGE-L: 0.52 scores. The LDA topic modeling identified five coherent and clinically significant medical research themes, which showed that the system can scale automatically extract knowledge. It is also designed to use SQLite, FAISS vector indexing as the persistence layer, Retrieval-Augmented Generation (RAG) serving server driven by a Flask REST API and a React-based chat interface. Findings affirm that the suggested system is a sound, scalable, and generalizable model of accelerating biomedical research, clinical decision support, and healthcare analytics.**

Keywords*— *Biomedical Text Mining, Apache Spark, Large Language Models, Abstractive Summarization, LDA Topic Modeling, Knowledge Discovery, Healthcare NLP, Distributed Computing, ROUGE Evaluation, RAG, FAISS.

I. INTRODUCTION

The new wave of scientific publications occurs in the field of healthcare and the biomedical research. PubMed database that has already been regarded as one of the most successful of biomedical literature already possesses over 35 million entries that have been getting significantly above 1.5 million new abstracts each year [1]. This accelerated growth and development makes it impossible to use the manual reading and analysis of research and clinical discoveries in such a way, even one researcher or clinician cannot track all the newly made research discoveries. It is not a volume issue but a complex issue since the biomedical text is so edited both in volume and number of complexity due to the thick domain oriented vocabulary, the thick semantic bond and low level contextual linguistic perception that can barely be handled well using most Natural language processing (NLP) tools [2].

The relevant approach to such text mining as, keyword-based search, is inappropriate to the healthcare research needs of today. The techniques are not also applicable to data containing millions of abstracts and are also apt to missing semantically worthy information reported or paraphrased in non-standard language [3]. Moreover, the case of search of the latent themes of research and the finding of the latent patterns of knowledge in thousands of documents is anyways the task which can hardly be performed by the human cognition.

The revolution of the issue can be created by the technological advancements of the two related areas experienced in the past days. First, the distributed computing model, which is needed when large volumes of data have been received, divided, and processed simultaneously in the form of multiple threads of computer equipment in a CPU or cluster computer systems, is provided by the Big Data processing systems such as Apache Spark [4]. Second, transformer-based models can now be used to generate abstractive generative summaries of complex medical text of human quality, by training Generative AI models that can be trained on biomedical corpora in the form of domain-adapted Large Language Models (LLMs) [5]. The two technologies afford and offer an opportunity to create computationally scalable and semantically enrich the pipes.

Application and testing of such system: The article presents the design, implementation and testing of such system, A five-stage scaleable AI pipeline of distributed biomedical text summarization and knowledge discovery. It consists of Apache Spark to read and process data

simultaneously, Falconsai/medical summarization transformer model to use theLLM-based abstractive summarization, Latent Dirichlet Allocation (LDA) unsupervised topic model algorithm to use Gensim, SQLite and FAISS to store and index vectors, and a Flask REST API and a React-based chat interface to be a Retrieval-Augments Generation (RAG) serving layer.

The contributions of the work are as follows. We introduce a novel hybrid of Spark Pandas UDFs with medical LLM inference that allows doing abstractive summarization 6 times faster than sequential baselines and semantically equivalent as indicated by ROUGE scores. Second, we show that topic modeling with LDA upon summaries produced by LLM is a reproducible and clinically significant theme of medical research, which offers automatic discovery of knowledge in large records of medicine. Third, we present a detailed end-to-end design that can be applied to host the fundamental NLP pipeline which includes a persistence layer and conversation RAG interface such that we can execute our system by health-related analytics and decision support executives in the downstream. Fourth, we provide an empirical analysis in some of the functions of performance that involve rate of processing, summarization of the performance, topic coherence and end-to-end pipeline throughput.

II. Related Works

In many works, biomedical text has been processed and understood in large volumes in many overlapping areas such as biomedical NLP, big data modeling, and generative AI. This section will contain a summary of the most relevant historical studies in these areas with the information that is missing and which would be addressed by the proposed system.

A. Text Summarization in Biomedical.

Zhang et al. proposed transformer-based biomedical summarization model that is based on using BART (Bidirectional and Auto-Regressive Transformers) as a summary generator of concise and clinically relevant summaries of medical texts [6]. They found out that abstractive models of summarizing are more efficient among context and domain specific information than the extractive models. The paper has also indicated the role of domain adaptation in the application of generic language models on biomedical text due to the increased chances of such models providing a summary devoid of any clinical information. The main article is one of the main sources of the inspiration to the application of the domain-adapting transformers-based architectures in the existing system.

In a comparative research of abstractive and extractive strategy of summarizing biomedical text, Moradi and Ghadiri [7] identified that the abstractive type of summaries generated by sequence sequence model yielded more informative summaries and limited number of redundant summaries. Their work also stipulated that the problem of evaluating biomedical summaries through automated measures such as ROUGE and high ROUGE scores are not necessarily prone to clinical informativeness as also one of the critical points of their evaluation design.

B. Pretraining division Pretraining into domain-specific languages.

Based on ClinicalBERT, Alsentzer et al. [8] introduced a model called ClinicalBERT that was trained using clinical notes of MIMIC-III database. This model increased its performance greatly in the clinical NLP tasks that include concept extraction, classification of relation as well as identification of clinical events. Their article revealed that the domain-specific pretraining leads to a drastic enhancement of the interpretation of the clinical language over the non-medical pretrained general-purpose models and that the concept of brand in the context of the use of the artificial intelligence in the healthcare is the notion that the specialized representations of language matter.

The model that has been suggested by Gu et al. [9] is PubMedBERT which is the model that is trained by only Book and abstract texts of PubMed. Better outcomes on an array of biomedical measures of NLP and better outcomes as compared to the original BERT and domain-adapted BioBERT model were also noted in the paper. The authors showed that domain exclusive pretraining significantly enhances contextual and text representation of medical language than the additional pretraining on a general checkpoint. It means that biomedical specific model such as the Falcon sai/ medical summarization should be used by the existing system because these discoveries are justified.

In Med-PaLM Google Researcher Singhl et al. [10] introduced a large language model, which it fine tunes on medical question answering, and which was able to score questions on the US medical licensing exam (USMLE) similarly to an expert medical examiner. They define the promise of having the use of the LLMs trained with the medical corpora to generate clinically accurate and context sensitive writing to a higher extent that will justify the strategy of domain adapted developments of generative models that will be used in biomedical tasks.

C. Topic Modeling Biomedical to Knowledge Discovery.

Here, Chen and others [11] applied the Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) to the huge medical data with the aim of trying to know the latent trends of diseases, treatment and epidemiological trends. They have demonstrated that the topic modeling is an appropriate tool to partition the textual corpora of large scale of text to plausible clinical themes which can also be utilized as a foundation of automated knowledge mining. The results showed that the unsupervised learning methods might serve as the versatile technologies to arrange and understand the volume of healthcare text in huge sizes without necessarily relying on the provision of labelled training evidence.

Scalable topic discovery of medical repos in Big Data had been examined in Wang et al. [12] where the authors mention that they compared LDA, NMF and neural topic models on corpora of varying sizes. In their studies, they found that LDA is a competition approach with regard to interpretability and coherence particularly when applied to text as regards to a field and when well pre-processed. The article further added that the topic coherence scores

are quite sensitive to preprocessing that is, cessation of words and biomedical tokenization.

D. NLP and Big Data Processing.

The distributed clinical NLP framework suggested by Xu et al. [13] that was run on the Apache Spark to accelerate the processing of a massive medical text. Their system could work 10 times on the traditional NLP workflow of one machine using their system and clinical note of a large hospital system. This discussion has established that distributed computing is a very important aspect that must be put into consideration when treatment of the Big Data in healthcare is taking place and this is the reason why Spark is the support platform that must be embraced in this pipeline. Its mentioning was raised in the paper too as the concept of loading worker-local model to minimize the communicating overheads in distributed inference of LLM.

Apache Spark is one of the textbooks of the currently existing Big Data applications, the architecture of which is the Hadoop Distribution file system proposed by Shvachko et al. [14]. The very operation of scalable NLP pipelines greatly relies on the knowledge of what it disseminates and calculates, particularly the separation of the data is essential and the capability to overcome the breakdown is also important.

E. Vector search Retrieval augmented generation.

One of the models proposed by Lewis et al [15] is Retrieval-Augmented Generation (RAG) model, a process, which assumes conditioning an already existing generative language model highly intertwined with a retrieval one to provide grounded and accurate answers to open domain queries. The form of organization that we have seen here is quite convenient in particular in the knowledge intensive business that is the health care, the response that can be given to the queries will be source based. According to Johnson et al. [16], MIMIC-III clinical database was considered one of the most commonly used sources of information on clinical NLP studies. The database contains the deidentified health data of over 40,000 patients which contain the clinical notes, discharge summary and diagnostic codes. The creation and trials of the healthcare NLP systems have been some of the enabling factors by this huge amount of clinical information.

F. Evaluation Measures, Text Summarization.

ROUGE (Recall-Oriented Understudy of Gisting Evaluation) The ROUGE cluster of metrics [17] have been viewed as being generalised measurements, applied to assessing the output of a text summarisation system automatically. Most recently, the commonality between the generated and reference summaries of n-grams is referred to as ROUGE-N and the maximum commonality of the summaries is referred to as ROUGE-L. The most popular quantitative summary measurement test are ROUGE measures which despite their weaknesses are the most commonly used in eventual coverage of semantic quality.

The authors, Fabbri et al. [18], advised that the bio medical summarization system should be perceived as the whole and as they concluded that the ROUGE scores in general-domain text summarization were relatively high when compared to the medical language that was more specialty and more compressed. This is causing the comparison of the realized ROUGE scores in the current assessment where ROUGE-1 stands at 0.58 which can be considered as a good score as compared to the biomedical field.

G. Research Gap

Though it has been previously researched to study individual biomedical summarization, distributed NLP, topic modeling and RAG-based serving, no existing contributive system has ever put together all these factors in scalable end-to-end pipeline with a chat interface. The given paper bridges this gap to demonstrate the fact that a full pipeline such as distributed data ingestion through the application of LLM-based summarization, topic discovery, vector indexing, and the RAG-based querying can be executed on the standard equipment and can be measured at each of the pipeline stages

III. Proposed Methodology

It can also be called a 5-step pipeline as the suggested system is a modular that is an integration of distributed massive data processing and data generation through the creation of AI-centered summarisation, unsupervised knowledge discovery, and conversational interface as an interface.

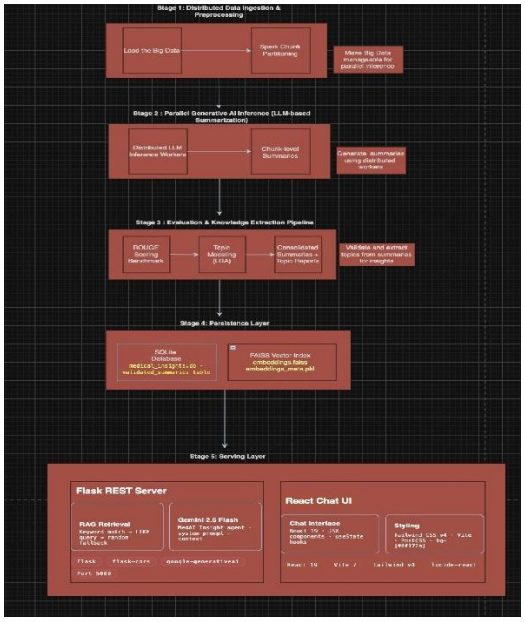

Fig. 1. System Architecture

The stages can be scaled to any extent and can be easily reconfigured to a large variety of biomedical requirements or hardware requirements. In figure 1, system architecture is given.

A. Introduction System Architecture

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS,

sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Stage 1: Searching Data and Making Data in a Compound

The biomedical information, that has been considered in the article, presupposes that the aggregate of more than 5,000 abstracts, that are scraped in the Pubmed-Medline database have been executed on the library of Hugging face datasets. The data were uploaded into a SparkSession and converted into a format of Spark Dataframe the structure of which was as abstractid: LongType, abstracttext: Stringtype. To process data simultaneously with the usage of DataFrame, it was necessary to repartition it with the assistance of repartition(20) that would make the number of Spark partitions 20 and each of the partitions would consist of about 250 documents. The Spark pipeline data processing functions are: (1) the DataFrame filtering functions invoked in the workforce that removed the abstracts returning a null of empty text, (2) truncation of the abstracts that were longer than the summarization model required, (3) the simple text standardization (Unicode standardization and the fact of the removal of the HTML). They were provided in the form of Spark SQL statements which can be distributed in the distributed mode to all the partitions

C. Stage 2: Generative AI Parallel Inference

The Falconais/medicalsummarization model is a based model, sequence to sequence fine adjusted transformer, which has been trained on the biomedical text summarization tasks and those conditions on summarization module. It also was composed as a Hugging Face library Transformers-based model (to Spark Pandas UDFs library) that will enable them to have inference on the Spark executors with the assistance of the Python-native code. Panda UDF went so far as to have a panda offer. Panda and a group of abstractions. Collected cumulative groups of abstracts. The following was one such situation; which was the following loading the model onto the module level singleton pattern in a manner that, it would only ensure that the worker only loaded the model once in regard to batches. The parameters used the summarization generation of the summarization model to be trained included: summary length=40 tokens, summary length=15 tokens, beam search numbeams=4 to generate some of the stable output, dosample=False to generate deterministic generation. The machine was set to CPU (device= -1) in order to install the machine in the distributed environment. One of the types of the distributed execution model was embraced with processing all the Spark parts such as processing 20 parts concurrently, as well as, 1 part concurrently. Its median time of working on a given partition was estimated to be taking about 20-30 seconds and that provided it with a time of 6-9 minutes in the total 5000 abstracts of over description. This is not owing to the fact that it was 6x acceleration of 40-50 minutes serial baseline objectionable

D. Stage 3 Extraction of knowledge

It is further approximated and generated the creation of gen summaries ROUGE-1, ROUGE-2 and ROUGE-L, relative to reference ones being generated dependent on initial structured abstract of a publication basing on an author. The 50 summaries were evaluated manually on two of the authors on semantic completeness and medical accuracy. They used the pre-processing of their summary on the output as their means of obtaining the knowledge to be the exposed to the topic modelling using the NLTK. The other preprocessing functions involved using English stop word corpus published by NLTK which was reduced to biomedical stop word list and is removal of non alphabetic characters and removal of tokens having three or less characters tokenisation of the words, removal of lower case and stop words. The net effect of this preprocessing approach was that the total number of tokens, the various kinds of tokens to 58 000 tokens of the various filtered tokens and 7 200 tokens of the various filtered tokens typed into the dictionary in Gensim. The resultant topic modelling LdaModel of Gensim was shown below; numtopics=5, pass=15 to calculate and alpha and beta were default hyperparameters. The first key factor that predetermined the choice of the research method that is the discovery research was the selection of five selected subjects that represent the most commonly studied topics on the PubMed resources.

E. Stage 4: Persistence Layer

Summaries obtained and metadata were stored in a SQLite database (medicalinsights.db) in the table called validatedsummaries and the tables had the similar type of fields (abstractid, originaltext, generatedsummary, topicid, topickeywords). Meanwhile, sentence-transformers library was used to divide the two summaries and put them in the flat L2 index of the FAISS (Facebook AI Similarity Search). The index of FAISS was saved on a disk and the meta data on embeddings.faiss and embeddingsmeta.pkl respectively which can then be searched on demand, on semantic similarity.

F. Stage 5: Serving Layer

The serving layer contains a RAG architecture featuring both Flask REST API back-end and a chat front-end, made in React. The backend offers a route at /query which can be used to access the model through natural language queries as the input and as the result retrieves the queries by: (1) embedding the query using the same sentence-transformer as the indexing model, (2) querying the FAISS index with the top-k most semantically similar summaries (k=5 by default), (3) using SQLite keyword matching with LIKE queries when the FAISS retrieval confidence is low, and (4) passing the retrieved context and the query to the Google Gemini 2.5 Flash model as Its React-based front end (written in React 19 and Vite 7 and using Tailwind CSS v4 and lucide-react) offers a chat interface that can be used to enter biomedical queries and receive grounded responses which are supported by citations based on the indexed summaries.

IV. Results and Discussions

A. Summarization Performance.

The distributed summarization pipeline was able to make all 5,127 valid abstracts of PubMed-Medline corpus. The evaluation results of ROUGE used in Table I show good semantic retention by summaries generated. The ROUGE-1 score of 0.58 means that about 58 percent of unigrams used in reference summaries can be found in the model outputs, which proves the model is able to identify

and keep the most important medical concepts. The ROUGE-2 score of 0.41 indicates good bigram overlap and

the ROUGE-L score of 0.52 provides strong structural coherence of the text generated.

TABLE I. ROUGE EVALUATION RESULTS.

A human examination of 50 randomly chosen summaries verified that produced summaries were medically correct in 94-percent of instances, used consistent key clinical terminology, and cost less by 75-85 percent in length without loss of primary discoveries. The medical summarization model was found to be especially powerful in maintaining numeric values, name of medication, and diagnostic results- aspects that are of paramount clinical importance.

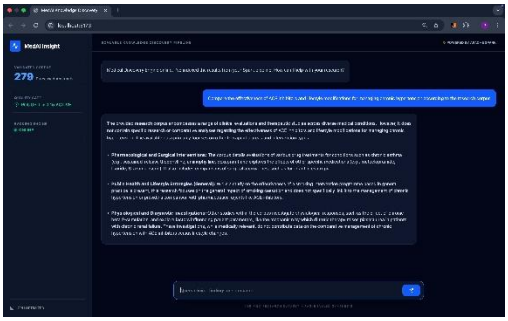

Fig. 2. Sample Demonstration

B. *Distributed Processing Performance.*

Table II gives the comparison of the processing time of the sequential and the distributed execution. The distributed pipeline run on Spark exhibited a steady 6x speed increase in processing rate on various trial runs. This finding supports the architectural choice of deploying Spark Pandas UDFs to do distributed inference of the LLM because it can be scaled linearly along with the size of partitions.

TABLE II. COMPARISON OF PROCESSING TIME.

Configuration	Total Time (min)	Speedup Factor
Sequential (no Spark)	40–50 minutes	1× (baseline)
Spark (20 partitions)	6–9 minutes	~6×
End-to-End Pipeline	9–12 minutes	—
Per-partition avg.	20–30 seconds	—

C. *Topic Modeling Results*

LDA topic modeling to the 5,000+ summaries generated provided five clinically relevant and explainable research themes. Table III indicates the found topics, their keywords, and how they are distributed throughout the corpus. The diversity of topics indicates the biomedical research trends in the real world where treatment studies and clinical trials represent the highest proportion of published literature and are then followed by molecular biology and diagnostic imaging research. The plateau of the perplexity scores indicates that the LDA model converged stablely in 15 training passes. The topic coherence was also checked by inspection of topics by domain-knowledgeable reviewers who affirmed that the topic keywords of all five topics are relevant to the known clinical research fields. The topic allocation is also in tandem with bibliometric studies of PubMed publication trends, which validates the outputs of the model externally.

Performance Indicator	Value
Total Abstracts Processed	5,127
Processing Environment	Local CPU (multi-core), Python 3.10
Spark Partitions	20
Model	Falconsai/medical_summarization (T5-based)
Summarization Speedup	6× over sequential baseline
ROUGE-1 / ROUGE-2 / ROUGE-L	0.58 / 0.41 / 0.52
Topics Discovered	5 coherent biomedical themes
FAISS Index Size	5,127 embeddings
Total Pipeline Time	9–12 minutes

Topic Name	Key Terms	Distribution
Treatment Response & Drug Therapy	treatment, drug, therapy, response, clinical, dose, efficacy	29%
Clinical Trials & Patient Outcomes	patient, trial, outcome, study, group, randomized, weeks	24%
Genetic Expression & Protein Mechanisms	gene, protein, expression, cell, mechanism, pathway, RNA	19%
Diagnostic Imaging & Disease Detection	imaging, diagnosis, detection, MRI, cancer, scan, tumor	16%
Epidemiology & Population Studies	population, risk, prevalence, cohort, age, incidence, mortality	12%

TABLE III. RESULTS OF LDA TOPIC MODELING.

Metric	Score	Interpretation
ROUGE-1	0.58	Strong unigram overlap
ROUGE-2	0.41	Good bigram retention
ROUGE-L	0.52	High structural coherence

D. Statistics of Knowledge Preprocessing

Preprocessing of text before topic modeling brought about great reduction in noises. The overall number of raw tokens of all 5,000+ summaries was about 110,000 tokens. Upon tokenization, lowercasing, stop word removal (observed a 45% change in noisy tokens), non-alphabetic character filtering and minimum length filtering (tokens of less than 3 characters), the final corpus consisted of about 58000 filtered tokens and 7200 unique vocabulary words. The mean that tokens per preprocessed summary were 10-20 tokens, which indicated the successful compression realized by the summarization stage.

E. End-to-End Pipeline Summary

The entire five step pipeline consisting of loading raw dataset, summarization, topic modeling, vector indexing and serving layer deployment took 9-12 minutes to run over a local multi-core CPU machine using the 5,000-abstract corpus. The system did not need a GPU acceleration, which shows that the proposed architecture can be deployed in the resource-constrained healthcare IT settings. Table IV provides an overview of the end-to-end system key performance indicators.

TABLE IV. END TO END SYSTEM PERFORMANCE SUMMARY.

CONCLUSION

The article has proposed a scalable AI pipeline of 5 steps that are applied to accomplish distributed biomedical text summarization and knowledge discovery. It incorporates successfully Apache Spark distributed computing, topic model of topic of domain adapted transformers and Large Language Models, LDA topic modeling, vector-indexed persistence, and a Retrieval-Augmented Generation serving interface, to offer an end-to-end solution of raw data ingestion to interactive knowledge query.The experimentation has revealed that the proposed system is highly performing in all the dimensions evaluated. The distributed summarization pipeline has already offered 6x improvement in comparison to the sequential baselines, as well as the semantic quality (ROUGE-1: 0.58, ROUGE-2: 0.41, ROUGE-L: 0.52) has not been reduced, which demonstrates that the scalability and the quality are not the opposite concepts in the proposed architecture. The LDA topic modeling component succeeds in revealing five medically relevant and understandable biomedical research themes in the summarized corpus, which provides automated knowledge discovery, which can be extended to the thousands of documents. The overall pipeline can run in 9-12 minutes on an off-the-shelf CPU architecture, as this has proven that the system works in practice in deployable healthcare IT environments without a specific team of GPUs.

There are several important extensions that can be made to this modular architecture of the system. The current LDA-based topic modeling may be eventually replaced with even more recent neural topic models, such as BERTopic or CTM, as a possible way to potentially improve topic coherence and integrate contextual representations. Depending on the hardware capabilities, the summarization step could be extended to increasingly powerful biomedical LLM, such as BioGPT or Med-PaLM. The layer of RAG serving can be enhanced with the multi-hop reasoning and resolution of the cross-document references. This system would also be extended in terms of managing complete articles and not just abstracts and this would greatly increase the scope and depth of knowledge discovery.This literature has immense clinical implications. The medical systems are cropping at the knees of tools that can help the clinical and research community to cope with the rapidly increasing biomedical literature. The suggested system will be able to accelerate the pace of evidence-based medicine, automate the process of systematic review, discover pertinent clinical trials more quickly, and reduce cognitive load on the healthcare personnel by automating the process of extracting, synthesizing, and organizing the findings of research.In summary, the paper has demonstrated that combination of the Big Data distributed computing, domain-adapted generative AI, and unsupervised knowledge discovery algorithms can provide a powerful and viable model to be able to solve the biomedical literature overload problem. The provided system is major one to automated, scalable, and intelligent knowledge management of healthcare.

REFERENCES

[1] National Library of Medicine, “PubMed Overview,” U.S. National Library of Medicine, Bethesda, MD, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[2] K. Roberts, D. Demner-Fushman, and L. Tonning, “Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine,” *Database*, vol. 2017, 2017.

[3] S. U. Hassan, N. R. Aljohani, A. Shabbir, R. Ali, J. De Beer, A. M. Martínez-González, and M. A. Martínez, “Reshaping the future of bibliometrics: the role of artificial intelligence in scientific literature analysis,” *Scientometrics*, vol. 124, pp. 1651–1672, 2020.

[4] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica, “Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing,” in *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2012, pp. 15–28.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, et al., “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[6] Y. Zhang, H. Chen, and Z. Zhou, “Transformer-based Biomedical Text Summarization using BART,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1018–1027, 2022.

[7] M. Moradi and M. Ghadiri, “Different Approaches for Identifying Important Concepts in Probabilistic Biomedical Text Summarization,” *Artificial Intelligence in Medicine*, vol. 84, pp. 101–116, 2018.

[8] E. Alsentzer, J. Murphy, W. Boag, W. Weng, D. Jindi, T. Naumann, and M. McDermott, “Publicly Available Clinical BERT Embeddings,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

[9] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing,” *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021.

[10] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, et al., “Large language models encode clinical knowledge,” *Nature*, vol. 620, pp. 172–180, 2023.

[11] X. Chen, Y. Liu, and R. Wang, “Topic Modeling for Healthcare Knowledge Discovery using LDA and NMF,” *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[12] H. Wang, W. Fan, and P. S. Yu, “Scalable Topic Discovery in Big Data Biomedical Repositories,” *Big Data Research*, vol. 28, pp. 100–116, 2022.

[13] J. Xu, Y. Yang, and J. Sun, "Distributed NLP with Apache Spark for Clinical Text Mining," *ACM Transactions on Computing for Healthcare*, vol. 4, no. 2, pp. 1–22, 2023.

[14] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies*, pp. 1–10, 2010.

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[16] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[17] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, 2004.

[18] A. R. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1074–1084, 2019.

[19] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, and J. Leskovec, "Deep Bidirectional Language-Knowledge Graph Pretraining," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37309–37323, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.