Addressing the Productivity Paradox in Healthcare with Retrieval Augmented Generative AI Chatbots

Sajani Ranasinghe*, Daswin De Silva*, Nishan Mills*,
Damminda Alahakoon*, Milos Manic‡, Yen Lim†, Weranja Ranasinghe†
*Centre for Data Analytics and Cognition, La Trobe University, Melbourne, Australia
†Department of Urology, Monash Health, Melbourne, Australia
‡Department of Computer Science, Virginia Commonwealth University, Richmond, USA

2024 IEEE International Conference on Industrial Technology (ICIT) | 979-8-3503-4026-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIT58233.2024.10540818

***Abstract*—Artificial Intelligence (AI) is reshaping the healthcare landscape through diverse innovations, personalisations and decision-making capabilities. The human-like intelligence of Generative AI has been fundamental in driving this transformation across the sector. Despite large investments and some early successes, several studies have signalled the emergence of a productivity paradox due to inherent limitations of Generative AI that disintegrate within the complexity of healthcare systems and operations. In this study, we investigate the capabilities of Retrieval Augmented Generation (RAG) and Generative AI chatbots in addressing some of these challenges. We present the design and development of a Retrieval Augmented Generative AI Chatbot framework for consultation summaries, diagnostic insights, and emotional assessments of patients. We further demonstrate the technical value of this framework in service innovation, patient engagement and workflow efficiencies that collectively move to address the productivity paradox of AI in healthcare.**

Keywords*—*Artifical Intelligence, Generative AI, Retrieval Augmented Generation, Chatbot, Healthcare, Productivity Paradox, Digital Health

I. Introduction

Generative Artificial Intelligence (AI) is transforming healthcare by leading a paradigm shift in the application of AI across its spectrum of functions, from public health to treatment and recovery [1], [2]. State-of-the-art Generative AI models, such as GPT4 by OpenAI, PaLM 2 by Google DeepMind, LLaMa by Meta AI, have introduced innovative approaches to healthcare that are capable of handling diverse complex tasks and providing clinical decision support [3], [4]. Furthermore, these models have demonstrated advanced conversational abilities in healthcare [5], [6], leading to the development of specialized models like Med-PaLM [7] and ChatDoctor [8] with major performance improvements in medical dialogues and patient-physician interactions [9].

Despite these capabilities and performance gains, Generative AI is not without significant limitations. Widely known as "AI hallucinations", Generative AI models are known to fabricate information in its naive effort to provide a response to every query, even though it may not have been trained on datasets relevant to such queries. Vague prompts, static parametric knowledge, limited domain expertise, insufficient training data, and uncertainties in language interpretation are further limitations that have emerged following the widespread use of Generative AI in practical applications [10]. These limitations have been further explored in recent studies in terms of the productivity paradox that is pre-existing at the intersection between technology and healthcare [11]. The productivity paradox is simply defined as the decrease in workplace productivity despite an increase in technology investment. The two typical reasons for productivity paradox are premature adoption and resistance to change. Premature adoption is due to unrecognised limitations of early versions of technologies that lead to a degradation of usage, while resistance to change is primarily due to the absence of adequate training and change management when introducing a new technology.

In addressing the case of productivity paradox in Generative AI, RAG represents an effective approach, diverging from the use of raw LLMs or their fine-tuning with task-specific data, by merging the capabilities of retrieval-based and generative models [12]. Domain adaptation by incorporating external knowledge sources, RAG enhances the accuracy and relevance of generated text which allows to provide contextually appropriate responses. This approach further validates the rapid domain adoption of RAG as a significant solution in mitigating the productivity paradox in healthcare, demonstrating its potential in bridging the gap between advanced AI capabilities and practical, efficient medical service delivery. Furthermore, there is an escalating interest in AI agents that leverage generative AI and the capabilities of LLMs. These agents are capable of performing various tasks with minimal human intervention by interacting with external tools, and they possess the ability to engage in conversations, reason, increase efficiency, improve precision, and provide insights.

In this study, we present the design and development of a Retrieval Augmented Generative AI Chatbot framework for consultation summaries, diagnostic insights, and emotional assessments of patients. Integrating external data repositories and information produced through analysis via the RAG architecture allows for the enrichment of results using Generative AI models. We further demonstrate the technical value of this framework in service innovation, patient engagement and workflow efficiencies that collectively move to address the productivity paradox of AI in healthcare. The rest of the paper is organised as follows. Section II introduces the framework and related work, followed by the capabilities of the framework which are demonstrated in Section III and Section IV concludes the paper.

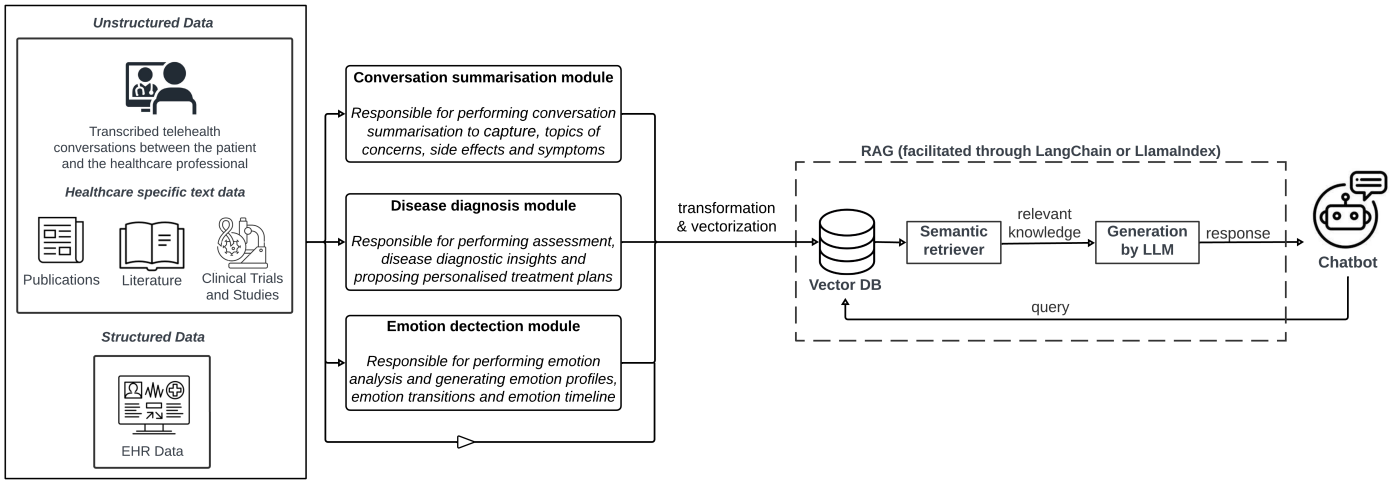

Fig. 1. The Retrieval Augmented Generative AI Chatbot Framework

II. The Retrieval Augmented Generative AI Chatbot Framework

The proposed framework is depicted in Fig 1. The framework begins with data streams and datasets received from the pre-existing data module integrating both structured and unstructured data. The framework introduces three analysis modules for patient telehealth conversation summarisation, disease diagnosis and treatment and emotion analysis, followed by the RAG module which facilitates response generation by efficient retrieval of information produced by the analysis modules stored in a vector database. The end-users of this system, the healthcare professionals can interact with this platform through a chatbot interface. This framework enables medical practitioners to make general inquiries and access patient information enriched with analytical insights, such as structure aids in the facilitation of improved decision-making in healthcare provision, as well as targeted and effective interventions. The following subsections provide further details of the modules and this framework.

A. Pre-existing data module

This module encompasses both unstructured and structured data. The unstructured data includes transcribed patient conversations resulting from telehealth consultations with healthcare professionals. To enrich domain-specific knowledge, data pertaining to healthcare domain is amalgamated from a variety of validated and reliable sources including publications, literature, clinical trials, and observational studies. Verifying the accuracy of context specific data is crucial because if inaccurate data is fed into the RAG architecture, it is more likely to generate unreliable responses. Structured data comprises the patient's Electronic Health Record (EHR) data, including demographics, biomarkers, medical and treatment histories, diagnostic details, medications, and other related information. The entirety of the data is consolidated into a cohesive data model, which can be established using a data lakehouse capable of accommodating structured, semi-structured, and unstructured data.

B. Conversation summarisation module

Patient conversation summarisation provide comprehensive overview into the patient's journey thereby enabling improved service quality [13]. The significance of summarising conversation of cancer patients, particularly focusing on their symptoms, side-effects, and concerns provide a comprehensive understanding of the patient's experience, potential challenges they face and their disease journey trajectory [14]. LLMs, with their advanced NLP capabilities, can summarise complex medical discussions into concise, easily understandable notes [15]. There is existing evidence of the use of LLMs such as GPT-4 for biomedical knowledge curation of extracting adverse drug events through medical reports [16]. This study [17], explores a self-verification framework with LLMs such as GPT-4 for efficient extraction of patient information from unstructured healthcare texts. The cumulation of these existing resources illustrates that these models can be utilised as effective tools for an agent to access and perform its tasks efficiently.

Utilising LangChain, a robust framework for developing applications using LLMs, which includes a comprehensive text summarisation library leveraging LLMs, will facilitate efficient analysis of patient conversations. LangChain successfully overcomes the issue of limited token size by introducing approaches like 'Stuff', 'MapReduce', and 'Refine', enabling the conversation text to be effectively processed within the LLM's context window. A clear and concise prompt is essential to align human intent with the model, ensuring effective outcome generation. Within the LangChain summarisation module, models such as GPT-3.5-turbo or Claude 2 can be leveraged for enhanced performance.Additionally, the module is adept at extracting and highlighting key aspects such as symptoms, side effects, concerns, treatment and medication from conversations. This approach combines the strengths of advanced

Named Entity Recognition (NER) models to enhance precision in identifying pertinent clinical entities. BioBERT [18] is an adaptation of BERT (Bidirectional Encoder Representations from Transformers) model [19] specifically trained on biomedical corpora to capture the nuances in medical language. Further, Spark NLP an open-source text processing library that leverages LLMs for advanced information extraction and NER, has proven superior performance in extracting clinical entities and medical conditions [20]. Hence, this module is instrumental in constructing comprehensive patient profiles by facilitating summarisation and information extraction at the patient level, which assist the healthcare practitioners in gaining an understanding of the patient's condition and history.

C. Disease diagnosis module

Disease diagnosis and assessment supports healthcare professionals in early disease detection, allowing for treatment tailored to the patient's unique needs. Such a strategy not only increases the effectiveness of treatment but also improves survival rates, reduces the likelihood of side effects, and enhances overall patient outcomes.

A recent study demonstrates the ability of GPT-3 to diagnose medical conditions nearly as accurately as physicians, but its triage capabilities are closer to laypeople and significantly less effective than those of medical professionals [21]. A further study using GPT4 highlights its enhanced ability in conducting diagnoses and triage, achieving results nearly on par with board-certified physicians [22]. Researchers at Google Research and DeepMind have developed Med-PaLM 2, the successor to Med-PaLM, that produced an accuracy over 80% in answering benchmark medical questions which could assist in medical diagnosis and prediction [4]. In this module, we analyse the patient's consultation history, adeptly identifying symptom patterns, previous diagnoses, treatment trajectories and other EHR data attributes to provide effective diagnostic insights. A variety of ML models will be utilised, covering a broad spectrum of complexity. These include Random Forest (RF), Decision Trees (DTs), XGBoost, Support Vector Machines (SVMs), and DL models chosen based on the specific disease type. Unsupervised methods like clustering algorithms and Self-Organizing Maps (SOMs), along with dimensionality reduction techniques such as Principal Component Analysis (PCA) and T-distributed Stochastic Neighbor Embedding (t-SNE), will be applied to detect patterns, anomalies, and clusters in disease detection. Additionally, transformer models will also be employed to enhance disease prediction and treatment plan formulation. Evaluating feature importance with methods like SHAP (SHapley Additive exPlanations) is crucial for understanding the impact of each attribute on model predictions. This evaluation helps end users comprehend how features influence disease prediction, thereby enhancing model transparency, informing clinical decision-making, and guiding targeted interventions for better patient care. The diagnostic insights and treatment suggestions provided by the module will serve as valuable support for qualified healthcare professionals in their decision-making process.

D. Emotion detection module

This module supports clinicians to understand and address the complex emotional states of patients, significantly impacting their treatment adherence, psychological well-being, and overall quality of life. 'Chatcounselor' a specialised LLM trained on real conversations between psychologists and clients showcase the advancements in providing effective mental health support through enhanced counseling capabilities and performance [23]. Recent research efforts at our research center have culminated in the creation of an AI driven emotion model, grounded in the validated psychological framework proposed by Robert Plutchik, consisting of four negative; anger, fear, sadness, disgust, and four positive; joy, anticipation, trust and surprise [24]. This AI-based emotion model is designed for adaptability, allowing fine-tuning to suit various scenarios, and has demonstrated significant applicability and value in the healthcare domain [25]–[27]. This model also possesses the capability to discern emotion intensities and is adept at managing negations. Beyond basic emotions, research has also extended to identifying complex emotional states such as depression, anxiety and stress in conversations, offering detailed perspectives on interactions [28], [29].

Thus, our work will be expanded to encompass not just the variants of eight primary emotions evoked during telehealth conversations between patients and healthcare providers, but also to detect subtle indications of depression, anxiety, stress, or suicidal thoughts expressed by the patient. In creating a fine-tuned emotion model, we will be using a lexicon driven approach to define similar words associated with each emotion. Employing pre-trained language models such as BERT will enable the efficient creation of embeddings, to extract features from patient conversational text. Additionally, the emotion dictionary will be enriched by leveraging the embeddings and language model to generate context-specific emotions, negations and intensifiers. Executing this emotion model will facilitate the generation of emotion profiles at conversation level and patient level. Efforts are also in progress to test the feasibility of emotion detection using few-shot learning techniques, by utilising LLMs like GPT-3.5-turbo, in a clinical environment.

Moreover, emotion profiles can be created at a multi-granular level such as, at the beginning of the telehealth consultation, the middle and at the end of the consultation. Creating these emotion profile at different granularity will facilitate the detection of emotion transitional states of the patient. Such transitions are crucial, as they can be employed in real-time to alleviate patient distress, and over the long term, they serve as a basis for evaluation metrics, to improve quality of the consultations, and compiling a knowledge base of frequent issues. Additionally, we will analyse the emotion timeline at the patient level, which will provide insights into the patient's overall well-being, enable early detection of mental health issues, enhance communication, guide treatment decisions, and help evaluate the effectiveness of telehealth services.

E. Retrieval Augmented Generation (RAG) module

The RAG architecture initiates with data loading and transformation component. Once the context-specific data and data from the three analysis components are imported, they undergo a transformation process to be broken down into manageable segments, a procedure often referred to as 'chunking'. The effectiveness of the system relies heavily on the quality and structure of these chunks, ensuring that the retrieved data is customised to match a user specific query. These chunks will be converted to embeddings which will be stored in vector store enabling the efficient storage and retrieval of information. In information retrieval, when the user initiates a query, engaging through the conversational agent, it gets transformed into an embedding representation, and then the vector store is searched to find pertinent information based on semantic similarity. The information retrieved is further enhanced by the LLM generative model to create a contextually relevant and improved response. LangChain [30] and LlamaIndex [31] are two robust frameworks that facilitate the implementation of RAG in building applications that enhance the interaction and capabilities of LLMs. LangChain specialises in conversational retrieval agents that are fine-tuned for engaging in conversations and performing retrieval tasks as needed. The agent takes in a well defined input prompt by the user and using an underlying LLM or ChatModel (eg: GPT-3.5-turbo OpenAI chat model) for reasoning, determines whether to engage the retrieval system based on the query. After retrieving the information, it displays a generated output to the end user. In the design of the retrieval agent, we are experimenting with OpenAI ChatModels such as GPT-3.5-turbo and GPT-4.

III. Framework Demonstrations

In this section, we demonstrate the capabilities framework across the three modules using a sample healthcare dataset consisting EHR data and transcribed telehealth conversations between healthcare professionals and patients. The dataset is part of a larger project with cancer care providers [25], [32].

Fig 2 illustrates the process by which healthcare professionals can obtain a patient summary using the unique patient ID. The conversational retrieval agent skillfully processes the query through its interaction with the ChatModel and retrieves the relevant information. Clinical entities identified by the medical NER pipeline are highlighted in the patient summary, which is then presented visually to the user. This summary, along with the NER, effectively encapsulates information about the patient's diseases, treatments, side effects, and concerns, providing valuable insights for healthcare professionals to act upon.

Further, healthcare consultants can access diagnostic predictions for specific patients. In the scenario depicted in Fig 3, we employed a lung cancer predictive model, developed using data from a lung cancer patient cohort. Various ML models were trained, each considering different patient attributes. The XGBoost model, with the highest F1 score of 93.02%, was selected for its superior performance. The model leverages clinical entities from consultation discussions and

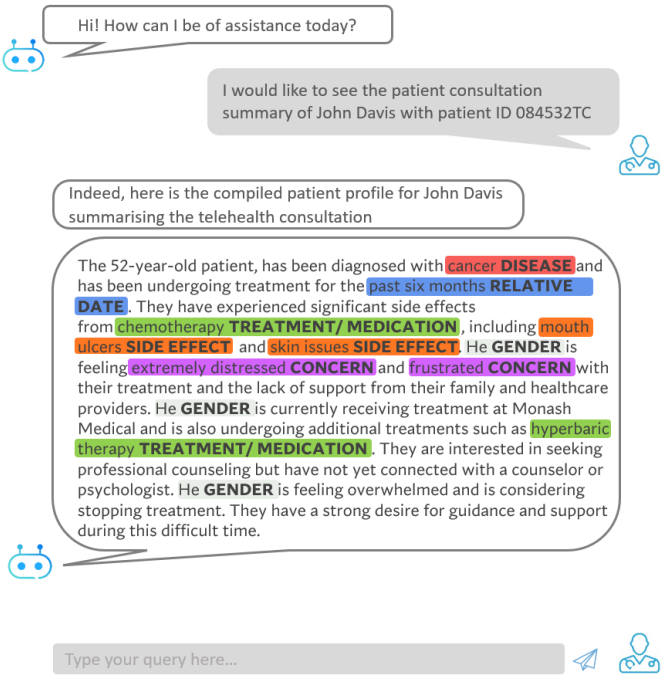

Fig. 2. Demonstration 1 - Conversation summarisation module

EHR data for its predictions. It can also provide SHAP feature importance values, helping users understand the key drivers of the model's predictions. It will showcase how the model achieved the final prediction. Moreover, utilising an underlying ChatModel enhances the readability of the output, particularly for a non-technical audience as shown by the output prompt of the conversational agent.

The consultant has the capability to inquire about the patient's emotional and mental health status through the conversational agent [33]–[35]. In this scenario, we have developed a customised emotion model that includes four positive and four negative emotions. The positive emotions identified are happiness, hopefulness/optimism, surprise, and trust. As for the negative emotions, they encompass a range including anger, disgust, sadness/disappointment, and helplessness. Additionally, we've implemented emotion algebra to detect any signs of anxiety, depression, and suicidal ideation in the conversations. As depicted in Fig 4, the professional can access these emotion profiles on either a call-specific or patient-specific basis.

Fig 5 demonstrates how a consultant can request emotion timelines to track a patient's emotional progression. These timelines display the patient's emotional changes over time. Utilizing the retrieved data, the ChatModel is adept at deriving insights when prompted correctly. Additionally, having access to emotion transition profiles at both the call and patient levels equips telehealth consultants with valuable information to better prepare for upcoming consultations.

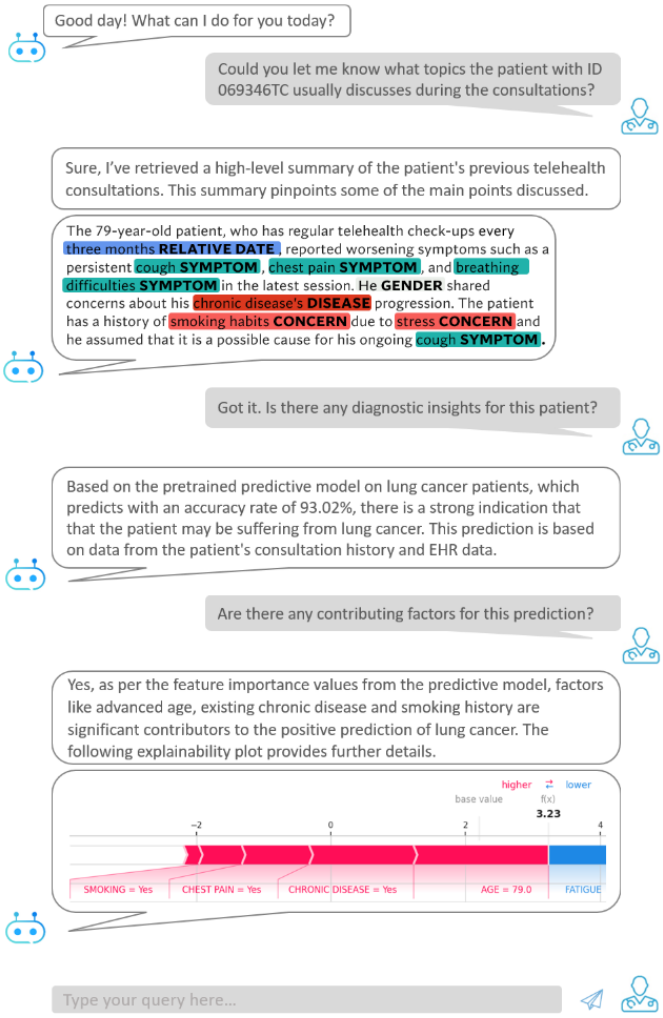

Fig. 3. Demonstration 2 - Disease diagnosis module

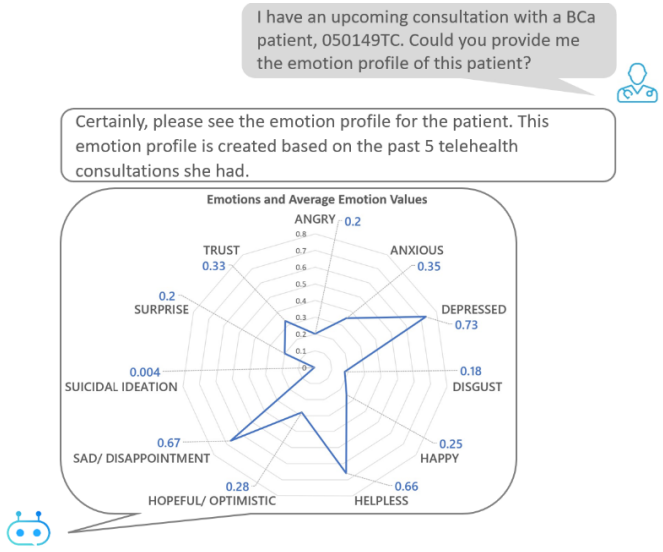

Fig. 4. Demonstration 3 - Emotion profiles in the emotion detection module

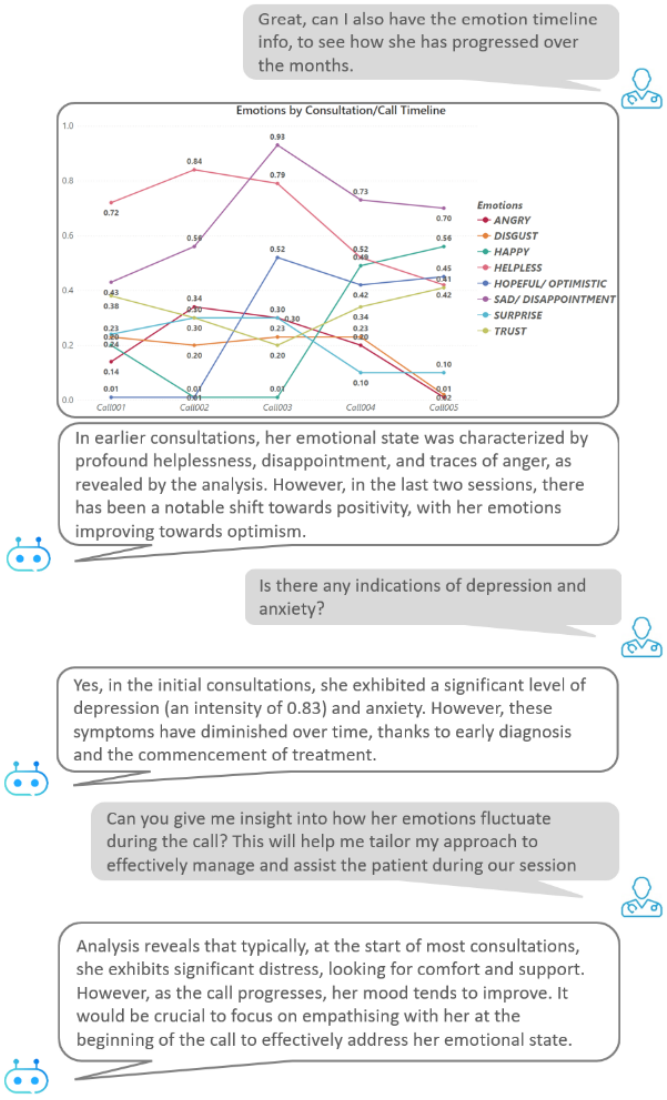

Fig. 5. Demonstration 4 - Emotion timeline analysis and emotion transition in the emotion detection module

IV. Conclusion

In this study, we have presented the the design and development of a Retrieval Augmented Generative AI Chatbot framework for healthcare, specifically for consultation summaries, diagnostic insights, and emotional assessments of patients. We demonstrated the capabilities of this framework in providing an improved healthcare service and addressing the challenges of the productivity paradox, where inherent limitations of Generative AI tend to disintegrate within the complexity of healthcare systems and operations. Thus, the RAG AI chatbot has the potential to address the productivity paradox by improving the efficiency of information processing and decision-making, easing the adoption of new technology through intuitive interactions, offering scalable and customisable solutions, and continually learning and improving over time. As future work, we plan to integrate a diverse array of multi-modal data, including images, audio, and video, into

our analysis. Further, we intend to conduct a field test of the framework in an actual clinical setting leading further confirmation and validity of its capabilities in resolving the productivity paradox of Generative AI in healthcare.

REFERENCES

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[2] D. De Silva, N. Mills, M. El-Ayoubi, M. Manic, and D. Alahakoon, "Chatgpt and generative ai guidelines for addressing academic integrity and augmenting pre-existing chatbots," in *2023 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2023, pp. 1–6.

[3] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative ai) in healthcare," *npj Digital Medicine*, vol. 6, no. 1, p. 120, 2023.

[4] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[5] D. De Silva, F. Burstein, H. F. Jelinek, A. Stranieri *et al.*, "Addressing the complexities of big data analytics in healthcare: the diabetes screening case," *Australasian Journal of Information Systems*, vol. 19, 2015.

[6] A. Adikari, D. De Silva, H. Moraliyage, D. Alahakoon, J. Wong, M. Gancarz, S. Chackochan, B. Park, R. Heo, and Y. Leung, "Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare," *Future Generation Computer Systems*, vol. 126, pp. 318–329, 2022.

[7] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[8] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, and Z. You, "Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge," *arXiv preprint arXiv:2303.14070*, 2023.

[9] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges," *Health Care Science*, vol. 2, no. 4, pp. 255–263, 2023.

[10] G. Gamage, S. Kahawala, N. Mills, D. De Silva, M. Manic, D. Alahakoon, and A. Jennings, "Augmenting industrial chatbots in energy systems using chatgpt generative ai," in *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*. IEEE, 2023, pp. 1–6.

[11] R. M. Wachter and E. Brynjolfsson, "Will generative artificial intelligence deliver on its promise in health care?" *JAMA*, 2023.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[13] V. Tsianakas, J. Maben, T. Wiseman, G. Robert, A. Richardson, P. Madden, M. Griffin, and E. A. Davies, "Using patients' experiences to identify priorities for quality improvement in breast cancer care: patient narratives, surveys or both?" *BMC health services research*, vol. 12, pp. 1–11, 2012.

[14] A. M. Hopkins, J. M. Logan, G. Kichenadasse, and M. J. Sorich, "Artificial intelligence chatbots will revolutionize how cancer patients access information: Chatgpt represents a paradigm-shift," *JNCI Cancer Spectrum*, vol. 7, no. 2, p. pkad010, 2023.

[15] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen *et al.*, "The future landscape of large language models in medicine," *Communications Medicine*, vol. 3, no. 1, p. 141, 2023.

[16] Y. Gu, S. Zhang, N. Usuyama, Y. Woldesenbet, C. Wong, P. Sanapathi, M. Wei, N. Valluri, E. Strandberg, T. Naumann *et al.*, "Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events," *arXiv preprint arXiv:2307.06439*, 2023.

[17] Z. Gero, C. Singh, H. Cheng, T. Naumann, M. Galley, J. Gao, and H. Poon, "Self-verification improves few-shot clinical information extraction," *arXiv preprint arXiv:2306.00024*, 2023.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] V. Kocaman and D. Talby, "Spark nlp: Natural language understanding at scale," *Software Impacts*, p. 100058, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665963.2.300063

[21] D. M. Levine, R. Tuwani, B. Kompa, A. Varma, S. G. Finlayson, A. Mehrotra, and A. Beam, "The diagnostic and triage accuracy of the gpt-3 artificial intelligence model," *medRxiv*, pp. 2023–01, 2023.

[22] N. Ito, S. Kadomatsu, M. Fujisawa, K. Fukaguchi, R. Ishizawa, N. Kanda, D. Kasugai, M. Nakajima, T. Goto, and Y. Tsugawa, "The accuracy and potential racial and ethnic biases of gpt-4 in the diagnosis and triage of health conditions: Evaluation study," *JMIR Medical Education*, vol. 9, p. e47532, 2023.

[23] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, "Chatcounselor: A large language models for mental health support," *arXiv preprint arXiv:2309.15461*, 2023.

[24] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[25] D. De Silva, W. Ranasinghe, T. Bandaragoda, A. Adikari, N. Mills, L. Iddamalgoda, D. Alahakoon, N. Lawrentschuk, R. Persad, E. Osipov *et al.*, "Machine learning to support social media empowered patients in cancer care and cancer treatment decisions," *PloS one*, vol. 13, no. 10, p. e0205855, 2018.

[26] A. Adikari, R. Nawaratne, D. De Silva, S. Ranasinghe, O. Alahakoon, and D. Alahakoon, "Emotions of covid-19: content analysis of self-reported information using artificial intelligence," *Journal of medical Internet research*, vol. 23, no. 4, p. e27341, 2021.

[27] S. Ranasinghe, G. Gamage, H. Moraliyage, N. Mills, N. McCaffrey, J. Bucholc, K. Lane, A. Cahill, V. White, and D. De Silva, "An artificial intelligence framework for the detection of emotion transitions in telehealth services," in *2022 15th International Conference on Human System Interaction (HSI)*. IEEE, 2022, pp. 1–5.

[28] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin, "Deep learning for depression detection from textual data," *Electronics*, vol. 11, no. 5, p. 676, 2022.

[29] A. Fatima, Y. Li, T. T. Hills, and M. Stella, "Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 77, 2021.

[30] "Langchain," https://www.langchain.com/, 2022-10, accessed: 2023-11-16.

[31] "Llamaindex," https://www.llamaindex.ai/, 2023-02, accessed: 2023-11-16.

[32] H. Moraliyage, D. De Silva, W. Ranasinghe, A. Adikari, D. Alahakoon, R. Prasad, N. Lawrentschuk, and D. Bolton, "Cancer in lockdown: impact of the covid-19 pandemic on patients with cancer," *The oncologist*, vol. 26, no. 2, pp. e342–e344, 2021.

[33] A. Adikari, D. De Silva, D. Alahakoon, and X. Yu, "A cognitive model for emotion awareness in industrial chatbots," in *2019 IEEE 17th international conference on industrial informatics (INDIN)*, vol. 1. IEEE, 2019, pp. 183–186.

[34] R. Nawaratne, T. Bandaragoda, A. Adikari, D. Alahakoon, D. De Silva, and X. Yu, "Incremental knowledge acquisition and self-learning for autonomous video surveillance," in *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2017, pp. 4790–4795.

[35] P. Rathnayaka, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, and R. Gray, "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring," *Sensors*, vol. 22, no. 10, p. 3653, 2022.