Model documentation and write-up

1. Who are you (mini-bio) and what do you do professionally? If you are on a team, please complete this block for each member of the team.

Team Member 01 : Lasantha Ranwala

I'm a medical doctor, health informatician, and PhD fellow at the University of South Australia. My research focuses on explainable AI-based clinical decision support systems. I graduated from the Medical Faculty at the University of Colombo in 2007 and obtained my master's in Biomedical Informatics in 2012 from the Postgraduate Institute of Medicine. In 2019, I completed my MD in Health Informatics. I completed the certification for "The Open Group Architecture Framework" (TOGAF) 9.1 and became a Certified Enterprise Architect in 2018. Additionally, I obtained Board Certification from the Postgraduate Institute of Medicine in Sri Lanka and became a Specialist in Health Informatics in 2021.

Team Member 02:

Dinuja Willigoda Liyanage is a dedicated AI developer with a keen interest in healthcare technology. Dinuja is passionate about quality and secure programming. He loves to follow the KISSS principle, which stands for "Keep it Simple, Stupid, and Secure" and he is a person who respects different perspectives, driving his career with a focus on growth, mutual success, and an appreciation for diverse viewpoints.

Team Member 03:

Benjamin Ung received a PhD in Electrical and Electronic Engineering in Terahertz Spectroscopy in 2013, during which time he was a postdoctoral research fellow in Prof Emma MacPherson's Terahertz group at the Chinese University of Hong Kong until 2016. He then returned to Australia and was a Product Specialist at Carl Zeiss Microscopy Australia until 2019, when he joined the Mechanisms in Cell Biology and Disease Research Group under Prof Doug Brooks at UniSA as a Research Fellow. Ben also has a B.Eng. degree (Honors) in Computer Systems Engineering and the B.Sci. degree in Mathematics and Computer Science from the University of Adelaide, in 2005. Currently, Ben is with the Quality Use of Medicines and Pharmacy Research Centre at UniSA.

Team Member 04: Poorna Fernando

I am an acting Consultant in Health Informatics from Sri Lanka contributing towards bridging the gap between the health field and information technology. I Possess a sound educational background with an MD in Health Informatics, Master's in Biomedical Informatics and MBBS. I'm currently working as a visiting scholar at the Australian institute of health Innovation (AIHI), Macquarie University, Sydney, Australia.

2. What motivated you to compete in this challenge?

Our team has four members, including medical doctors, health informaticians, and computer engineers, who are deeply invested in the intersection of healthcare and technology. Our active engagement in digital health research projects at the university level has highlighted the significant challenges posed by unstructured data, such as clinical notes, radiology reports, and patient records. These data sources, while rich in information, are often difficult to analyze due to their unstructured nature. We are motivated to participate in this challenge to address these challenges, contribute to the advancement of healthcare, expand our knowledge and skills, and collaborate with like-minded individuals.

3. High level summary of your approach: what did you do and why?

Our approach to finding new suicide-related insights from National Violent Death Reporting System (NVDRS) narrative data rests on two fundamental principles: domain expertise integration and scalable architecture design.Medical expertise drives every step of our analysis. Two physicians on our team provide crucial oversight from the initial data processing through to confirming our findings. This human guidance is essential in AI in healthcare, where mistakes in interpretation could have serious impacts. The doctors help verify that we're identifying the right variables, catching potential errors or biases, and keeping our work clinically meaningful. We've also designed our system to be adaptable and expandable. While we're currently focused on suicide-related factors, our framework can be applied to other healthcare analysis needs, from mental health studies to broader public health monitoring.

We extracted variables from the NVDRS dataset using three connected approaches: Topic Modeling to uncover hidden patterns, LLM Response Clustering to directly identify variables, and a Combined Method that merged insights from both. Two physicians independently reviewed each approach's findings and agreed on which variables were clinically meaningful. This two-doctor review process helped ensure our selected variables were medically valid while reducing any individual reviewer's potential biases.

4. Please provide the machine specs and time you used to run your model.

- CPU (model): Intel Core i5-14400F
- GPU (model or N/A): RTX 4070 GPU (12 GB VRAM)
- Memory (GB): 32 GB RAM 3600Mhz DDR4
- OS: Windows 11
- Train duration: N/A
- Inference duration: Please check the readme.md for more information.

5. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?

NA

6. Did you use any tools, data, or pre-trained weights for data preparation or exploratory data analysis that aren't listed in your code submission?

NA

7. How did you evaluate the performance of your approach, if at all?

Domain experts (two medical doctors) manually assess the relevance of the retrieved chunks to the query and provide qualitative feedback. Apart from that they manually assess the responses from the LLM based on each query.

8. What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?

We tried several BERT models like ClinicalBERT, BioClinicalBERT, BlueBERT, MentalBERT, RoBERTa and LLM like Mini LLM L6 V2 for embedding and found that more domain specific models like ClinicalBERT and Mental BERT are less effective for the embedding and more generalized model like RoBERTa is provide more related embedding for these narratives.

9. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?

If we were to continue working on this problem for the next year, here are the methods, techniques, and areas we would focus on to build upon the work so far:

1. Advanced Retrieval Techniques
Goal: Improve the precision of text retrieval and ensure the LLM generates highly contextual responses.

- Dense Passage Retrieval (DPR): Replace or augment FAISS with a retrieval model like DPR, which is better at matching query embeddings with relevant context.
- Hybrid Retrieval: Combine traditional sparse retrieval (e.g., BM25) with dense retrieval (e.g., FAISS) for better relevance scoring.
- Query Expansion: Use LLMs to generate variations of similarity search queries to broaden the search space without losing context.

2. Clustering and Query Optimization
Goal: Use clustering to optimize retrieval queries and improve narrative organization.

- Dynamic Query Generation: Use clustering results to group narratives and generate more precise, cluster-specific queries for retrieval.
- Automatic Query Suggestion: Automate the generation of Similarity Search Query List based on clustering analysis and topic modeling.
- Improved Clustering Techniques: Experiment with graph-based clustering or HDBSCAN for better handling of noisy data and overlapping clusters.

3. Better Chunking and Token Management

Goal: Handle long texts more effectively and maximize the utility of retrieved context.

- Sliding Window Approach: Implement sliding window text chunking to capture overlapping contexts, reducing loss of meaning during chunking.
- Hierarchical Summarization: Summarize long texts into key sections and only chunk or embed summaries, reducing redundancy.
- Adaptive Chunking: Dynamically adjust chunk sizes based on the complexity or density of the text.