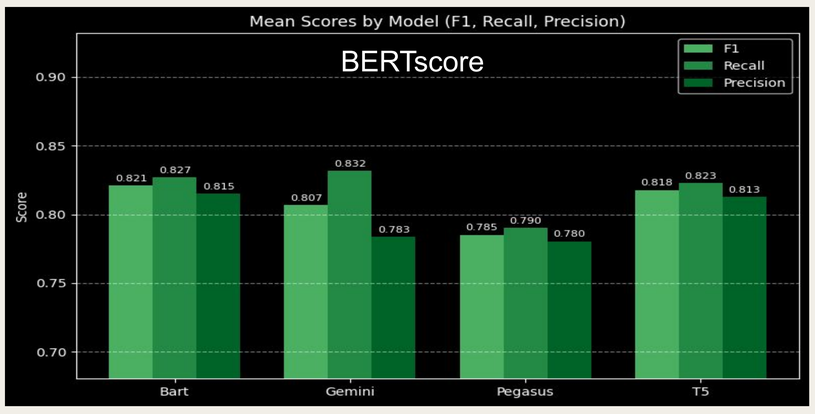
Rags to Riches: An Ablation Study of Medical RAG Pipeline with Summarization Model Comparison

By: Jack Skupien, Matt Fecco, and Will Greenwood

Motivation

- Better medical summarization produces a more effective medical RAG.
- Large Language Models have restrictions on input length.
- Shorter documents allows for more efficient storage and retrieval while maintaining key information for inference.
- Many summarization techniques and models exist.
- Can we use model comparison and nlp metrics to identify best medical summarization models?

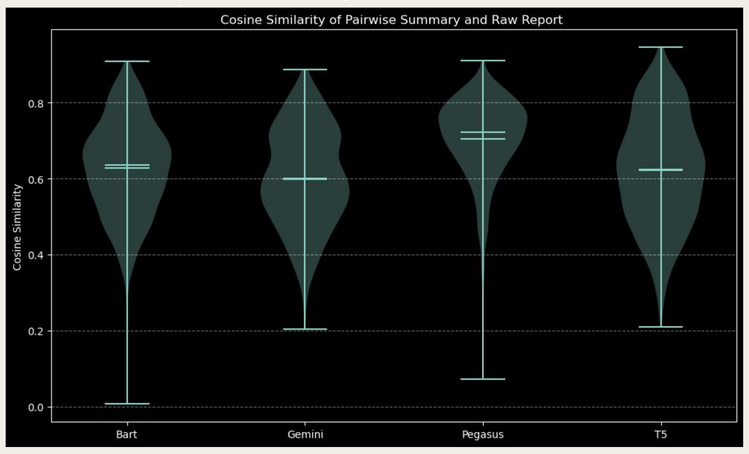


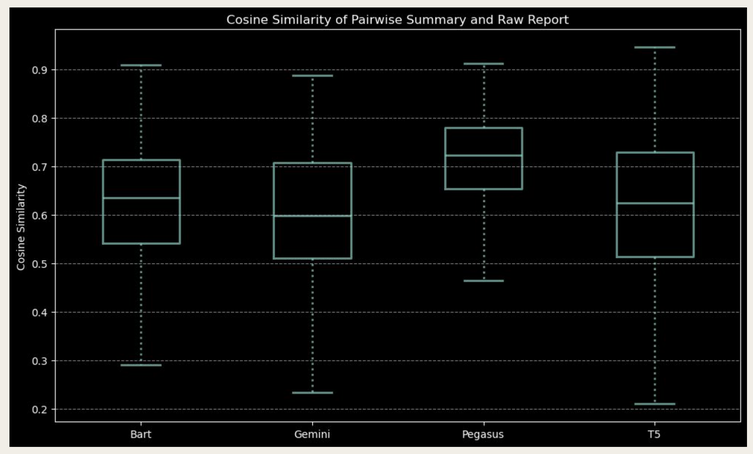
Data Description

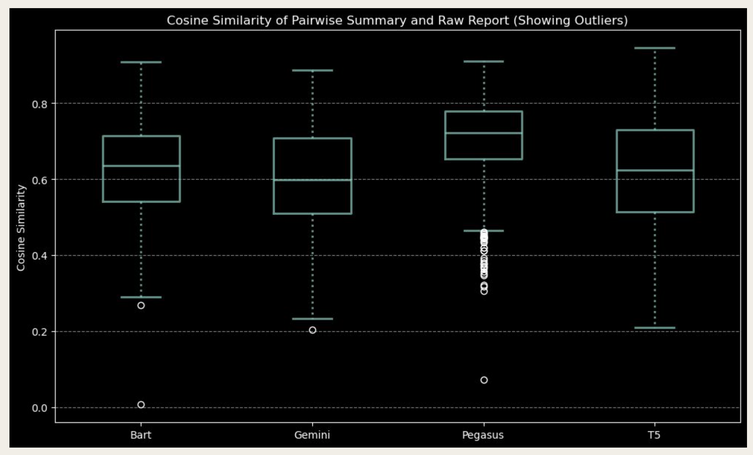
- **MIMIC-III Dataset:** A clinical dataset comprised of ~50,000 de-identified and unstructured ICU patient records. Dataset compiled from critical care patients treated at Beth Israel Deaconess Medical center between 2001 and 2012.

Using Pyspark.sql

- Isolated patients with discharge reports
- Filtered by injuries
- Cleaned identification placeholders
- Sampled 1024 reports for summarization





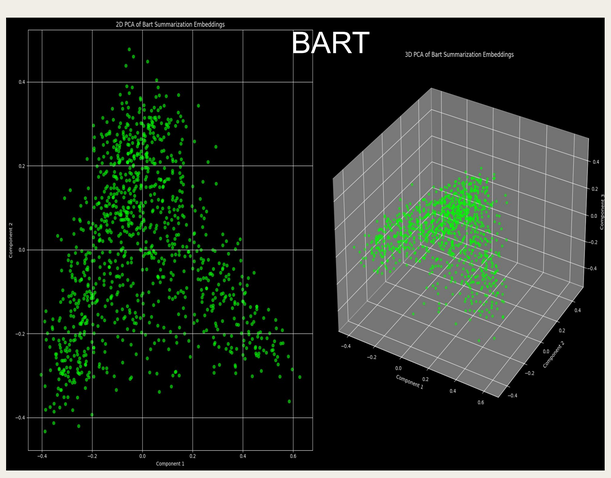


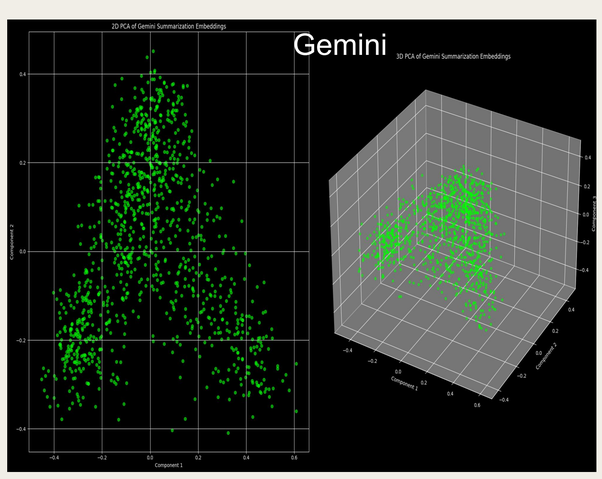
Methodology

- Models for abstractive summarization comparison: Bart, Gemini 2.0, Pegasus, T5
- Generated summaries for each model using segmented summarization at a 90% compression rate
- Embedded summaries using **all-mpnet-base-v2 sentence-transfor**

For each summary we calculated:

- BLEU (Bilingual Evaluation Understudy) metric - n-gram overlap focusing on precision
- ROUGE1 (Recall-Oriented Understudy for Gisting Evaluation) metric - unigram overlap focusing on recall
- BERTscore - semantic similarity using BERT embedding model and cosine similarity
- Pairwise cosine similarity between summary embedding and first 512 tokens of raw report using our chosen embedding model.
- Track word frequency and vocab diversity through Word Clouds





Embedding Clusters

Findings

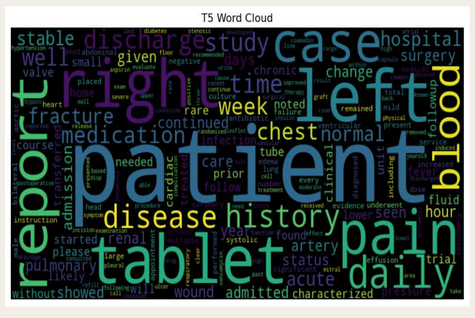
- **Practicality**
  - Gemini produced summaries the fastest (1.5 hrs) but requires API key or expensive hardware to run locally
  - Smaller models ran longer: T5 (20 hrs), Bart (28 hrs), and Pegasus was the longest. (40 hrs)
  - MapReduce summarization reduces this time but requires expensive access to distributed computing systems. Our codebase used a combination of MapReduce and traditional execution.
- **Repeatability**
  - Bart and Pegasus had the tightest heatmap with BLEU and ROUGE1 scores indicating a consistent rate of n-gram overlap.
  - T5 has a wide variance on semantic similarity, but a higher ceiling.
  - Pegasus has a clear polarization with a cluster of strong semantic similarity and a large set of “outliers” that have a below average similarity score when compared to entire corpus.
- **Accuracy**
  - Pegasus had lowest overall F1 score, recall, and precision may be due to its polarization weighing down overall performance. It is possible that Pegasus has the most upside if wide variance can be accounted for.
  - Gemini has the lowest precision (BLEU) and only slightly better (ROUGE1) indicating a more diverse vocab. This is confirmed when looking at the most common words within Gemin’s Word cloud and is consistent with the larger parameter count. The tendency to use alternative words leads to a lower average semantic accuracy when looking at violin plot.
  - Bart and T5 models have tight heatmap clusters around BLEU and ROUGE1 scores. With a 90% compression rate, we have identified a 0.1 BLEU score as a threshold for strong n-gram overlap.

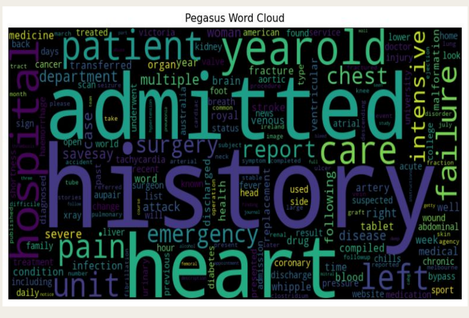
Future Work

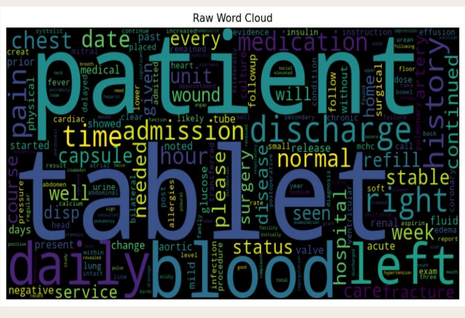
- Recursive summarization instead of segmented
- DBSCAN parameter search for cluster comparison
- Identify sample cluster in embeddings and check text of summaries to confirm proper clustering
- Conduct direct human evaluation of summaries for qualitative analysis.

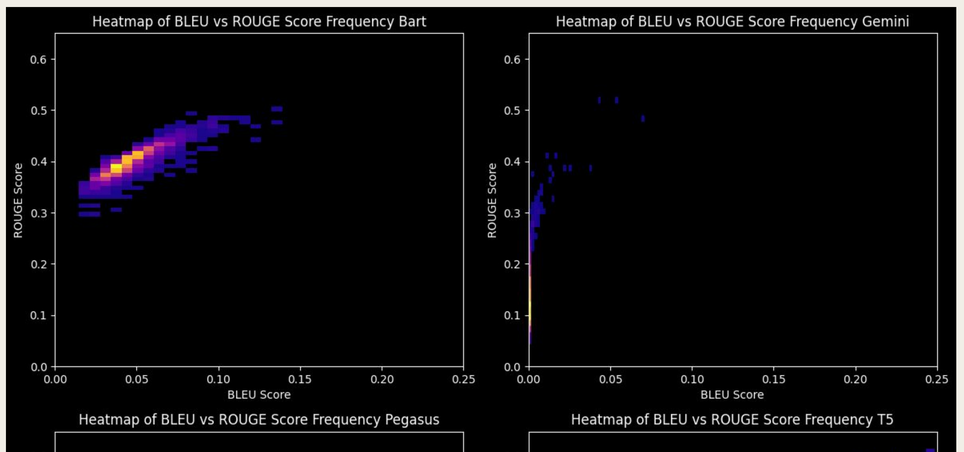
- Considerations

- Unable to embed original reports due to embedding model’s max token length (512) restricting our ability to compare directly within current RAG pipeline
- More efficient and rigorous pre-processing would yield more structured reports with less noise
- Models were not trained on medical data. Pre-training or fine-tuning models may yield stronger results.









TICKLE COLLEGE OF ENGINEERING
THE UNIVERSITY OF TENNESSEE, KNOXVILLE