Resonate:
A Retrieval Augmented Framework For Meeting Insight Extraction

Sartaj Bhuvaji, Prachitee Chouhan, Madhuroopa Irukulla, Jay Singhvi
Advisor: Dr. Wan Bae

Motivation

Meetings serve as vital platforms for collaboration and decision making.

Meetings can be overwhelming, leaving us to remember various details and discussions.

It's all too easy to miss a crucial details, if we don't remember meeting details.

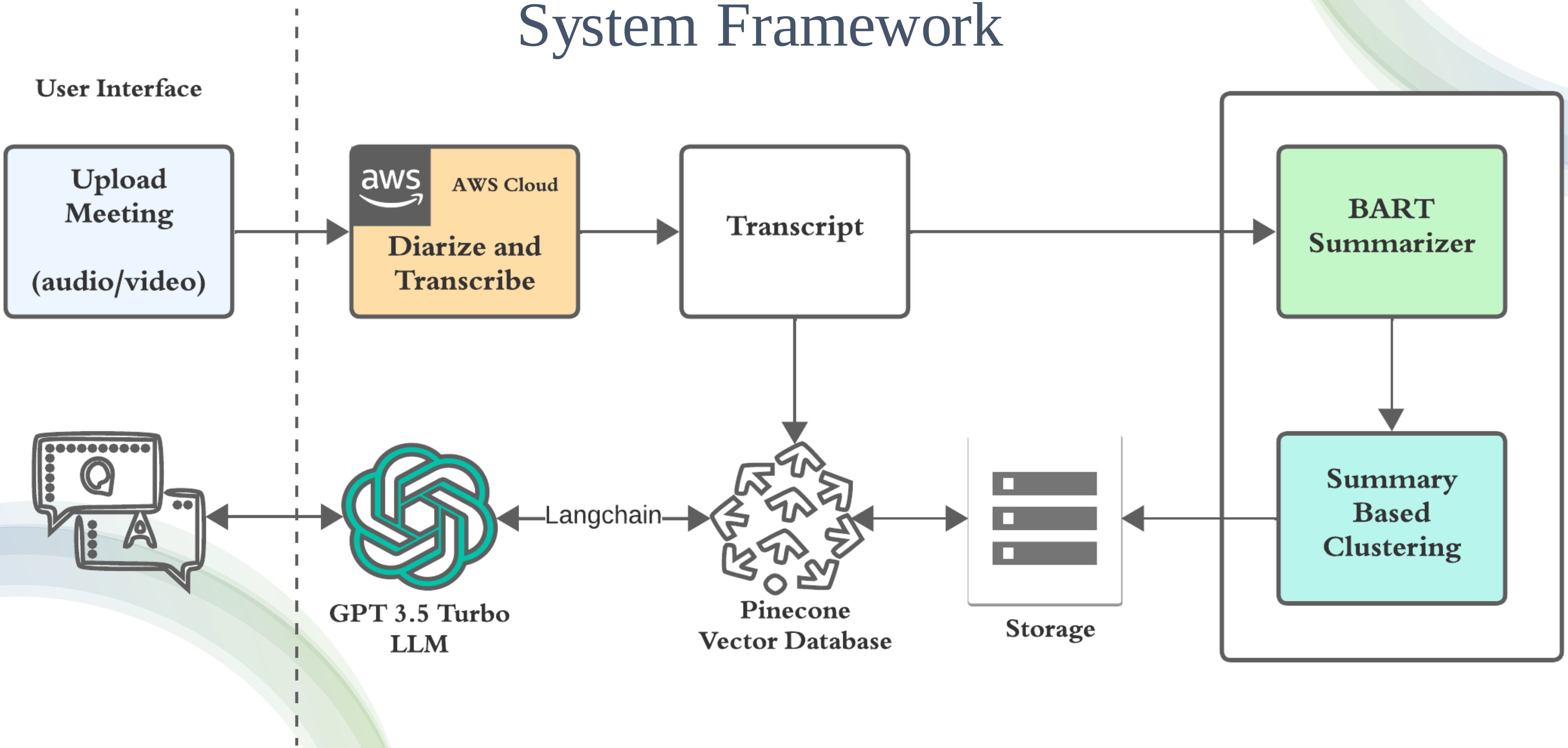
Problem Statement

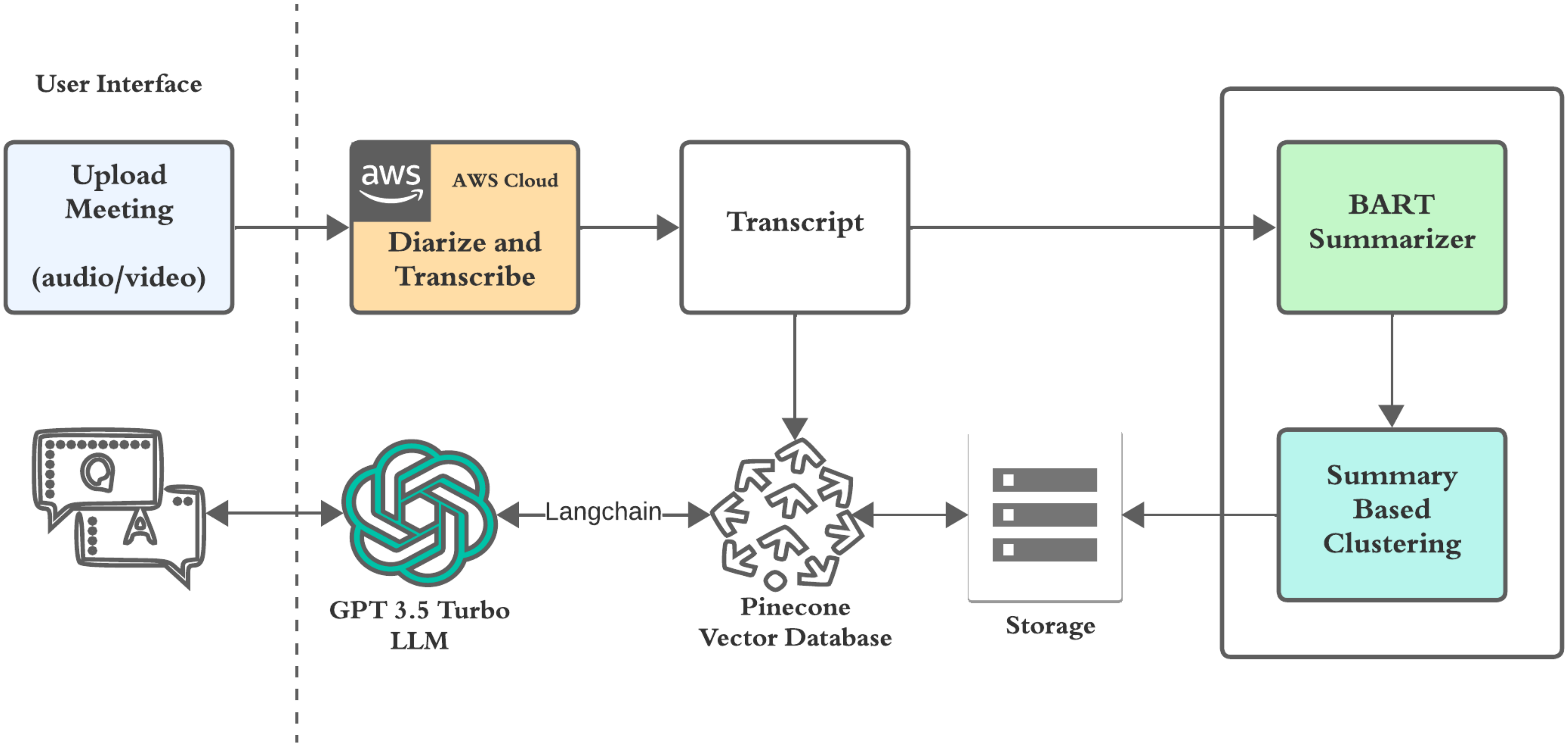
In today's professional landscape, meetings are a daily occurrence, often filled with valuable discussions and decisions. Recalling crucial details can be challenging, hindering productivity.

Our project aims to develop a chat interface to help users extract pivotal insights from historical meetings, using Retrieval Augmented Generation techniques to enable seamless information retrieval.

By grouping meetings based on abstractive summaries by leveraging clustering algorithms, this will provide users with precise responses and a high-performance solution for content discovery.

System Framework




User Interface
Upload Meeting
(audio/video)
aws
AWS Cloud
Diarize and Transcribe
Transcript
BART Summarizer
Summary Based Clustering
Storage
Pinecone Vector Database
Langchain
GPT 3.5 Turbo LLM
Q
A

Abstractive summarization

- LLM Model: Facebook BART
- Using Fine-tuned version from Hugging Face.

Cluster	Meeting Time	Abstractive Summary	Ground Truth
architecture	28	This is the second meeting of the group. First, the group discussed the logistics of the new time zone. They discussed how to deal with the different time zones of the time zones. Then, they discussed how they could make their work better by making it easier for people to share their work with the world. Lastly, they talked about how to make the new diffs architecture better for the group and how they would make it easier to share ideas about their work and future work with other groups. This meeting was about the progress of the team's work on the DS project. The team first discussed the technical matters, then moved onto a discussion about how they would make their work more efficient. The final decision was made that the team should focus on the service side rendering, but the team was not sure what they should do to make it better. Finally, the meeting ended with a brief discussion about the future of the project, which was mainly about future work and future thoughts on the project.	The conversation revolves around optimizing meeting times across different time zones. They discuss proposed meeting times and their implications. Suggestions include marketing the meeting better, proposing earlier times, and considering various time zones' constraints. Additionally, they discuss prioritizing performance metrics and trade-offs in software development. They plan to collaborate on documentation and explore ways to improve file highlighting efficiency in the software team members discuss progress and tasks. They cover issues like code migration, documentation, and diagram updates. They prioritize completing tasks efficiently rather than seeking perfection. The discussion revolves around merging requests and clarifying architecture changes. They aim for clarity in tasks and encourage participation in defining completion criteria. Despite some chaos, they resolve issues and plan future meetings to ensure progress.

Dataset

23 MEETINGS

TOPICS :
ARCHITECTURE, OFFICE RELOCATION, SOCIAL MEDIA, DEVICE

Summarization Model Evaluation

BERT Score

Meeting	Precision	Recall	F1	Similarity Score
Architecture	0.5840	0.5869	0.5854	0.8641
Remote Control	0.7197	0.6982	0.7088	0.8981
HR	0.6973	0.7240	0.7104	0.9114
Social Media	0.7157	0.7052	0.7104	0.9049

- Contextual Embeddings
- Sentence Ordering
- F-score Calculation

Clustering

- Inspired by paper, titled as "More discriminative sentence embedding via semantic graph smoothing."
- H is the vector, p is the propagation order, $\propto$ and T are filter-specific hyperparameters.

Filter	Propagation rule
Simple Graph Convolution (SGC)	$H^{(p+1)} \longleftarrow \mathrm{S}H^{(p)}$
Simple Spectral Graph Convolution (S2GC)	$H^{(p+1)} \longleftarrow H^{(p)} + \mathrm{S}H^{(p)}$
Approximate Personalized Propagation of Neural Predictions (APPNP)	$H^{(p+1)} \longleftarrow (1\text{-}\propto)\ \mathrm{S}H^{(p)} + \propto H^{(0)}$
Decoupled Graph Convolution (DGC)	$H^{(p+1)} \longleftarrow (1\text{-}T^{p})\ H^{(p)} + + T^{p}\mathrm{S}H^{(p)}$

Clustering models Evaluation

Vector Embedding-SentenceTransformer(model=all-mpnet-base-v2)

Mean-Shift

	Adjusted Mutual Information	Adjusted Rand Index	Bcubed-Precision	Bcubed-Recall	Bcubed-Fscore
sm-sgc-ms-st	0.5256	0.5599	0.8328	0.6770	0.7469
sm-s2gc-ms-st	0.3498	0.2011	1.0000	0.5031	0.6694
sm-dgc-ms-st	0.6184	0.6625	0.7809	0.8043	0.7925
sm-appnp-ms-st	0.3498	0.2011	1.0000	0.5031	0.6694

HDBSCAN

	Adjusted Mutual Information	Adjusted Rand Index	Bcubed-Precision	Bcubed-Recall	Bcubed-Fscore
sm-sgc-hdbscan-st	0.4798	0.3711	0.5502	0.8174	0.6577
sm-s2gc-hdbscan-st	0.5008	0.2781	0.5251	0.8696	0.6548
sm-dgc-hdbscan-st	0.2805	0.1537	0.3411	0.8043	0.4791
sm-appnp-hdbscan-st	0.5008	0.2781	0.5251	0.8696	0.6548

Vector Embedding-OpenAI (model=text-embedding-3-large)

Mean-Shift

	Adjusted Mutual Information	Adjusted Rand Index	Bcubed-Precision	Bcubed-Recall	Bcubed-Fscore
sm-sgc-ms-openai	0.6028	0.6034	0.8328	0.7350	0.7808
sm-s2gc-ms-openai	0.0000	0.0000	1.0000	0.3875	0.5586
sm-dgc-ms-openai	0.8410	0.8980	0.9130	0.8841	0.8983
sm-appnp-ms-openai	0.0000	0.0000	1.0000	0.3875	0.5586

HDBSCAN

	Adjusted Mutual Information	Adjusted Rand Index	Bcubed-Precision	Bcubed-Recall	Bcubed-Fscore
sm-sgc-hdbscan-openai	0.5009	0.3293	0.5167	0.882	0.6517
sm-s2gc-hdbscan-openai	0.6664	0.5175	0.6187	0.942	0.7469
sm-dgc-hdbscan-openai	0.4879	0.2696	0.4849	0.913	0.6335
sm-appnp-hdbscan-openai	0.6664	0.5175	0.6187	0.942	0.7469

Retrieval-augmented generation(RAG) Modelling Evaluation

Groundedness:

- 0 to 1
- Measure of how well the answer is supported by the context.

Answer Relevance:

- 0 to 1
- Measure of how well the answer is relevant to the question.

Context Relevance:

- 0 to 1
- Measure of how well the context fetched from DB is relevant to the question.

Cosine Similarity

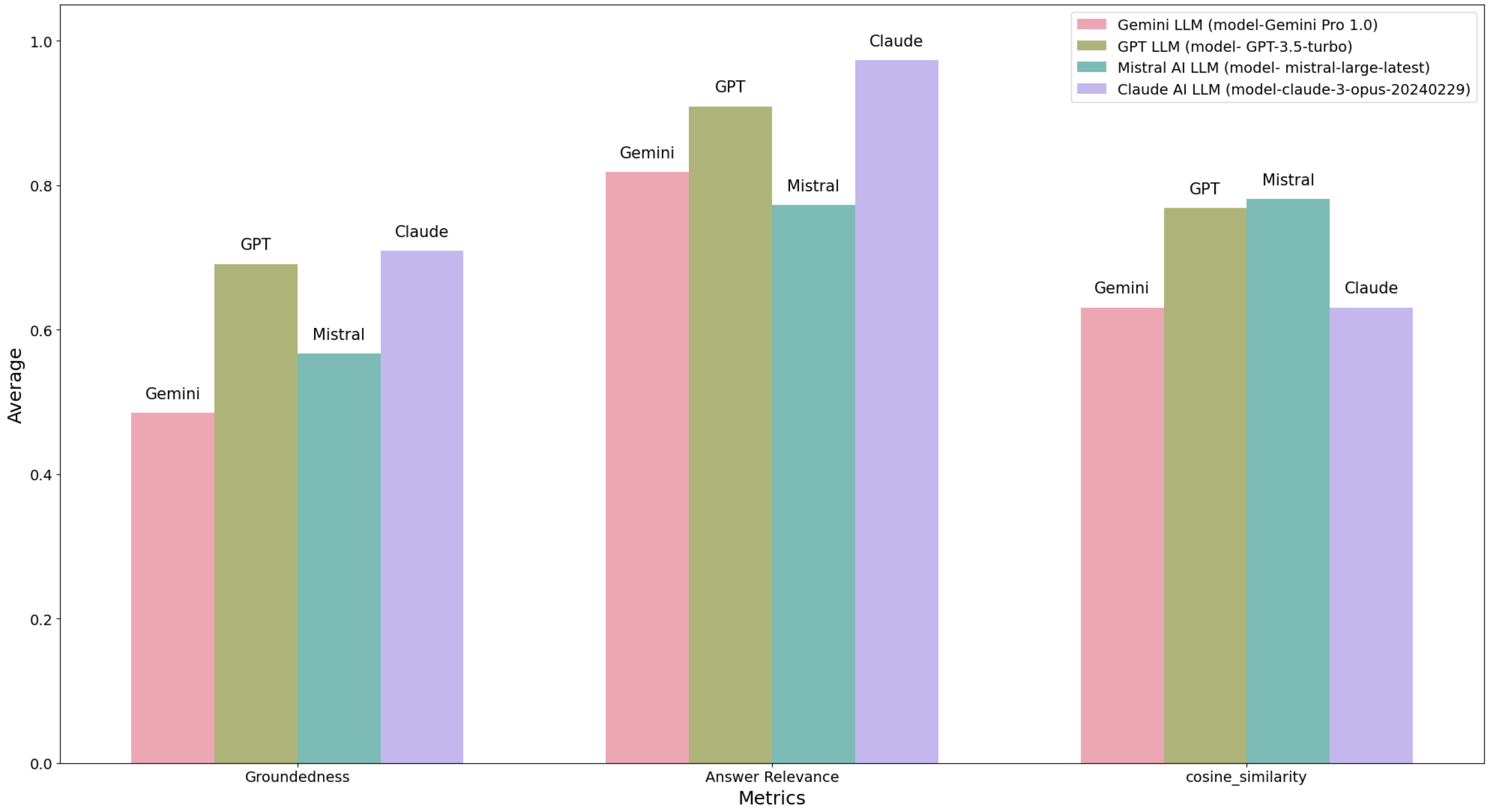
- 0 to 1
- Measures how similar answer by LLM is to the ground truth.

Example:

No	User Input	Response	Groundedness	Context Relevance	Answer Relevance	Cosine Similarity
1	Who recommends universal masking?	The American Academy of Pediatrics and CDC recommends universal masking.	1	0.9	1	1
2	How many patients are in the ICU Unit?	There are 23 patients in the ICU Unit located at the East Wing.	1	0.7	1	0.96
3	What is the capital of France?	Paris is the capital of France.	0	0	1	0

** Questions asked for meeting involved COVID-19 Discussion*
** LLM : Chat GPT 3.5-Turbo*
** Database: Pinecone Serverless*

Comparing LLM Models



Calculated over five questions for each of the twenty-three meetings.

DEMO

Resonate - Meeting Chatter

Toggle Add Meeting / Chat

Chat

How can I assist you?

Chat Here

Message Resonate ...

Clear

2024-03-10 17:56:06

Future Work

- FINE TUNING USING QLORA.
- RESEARCHING FASTER TRANSCRIBE MODEL
- EXPLORING DIFFERENT EMBEDDING MODELS
- EXPLORING OPEN SOURCE LLM MODELS.
- VIDEO FRAME TAGGING

* QLORA: Quantization Low-Rank Adaptation of Large Language Models

Project Learning Outcomes

- We learned how Retrieval Augmented Generation architecture works.
- We explored different metrics for abstractive summarization.
- Different dimensions of vector embedding by offer varying levels of granularity and specificity in representing data and how it effects clustering.
- We experimented with LLM architecture and how it manages memory.

Thank You