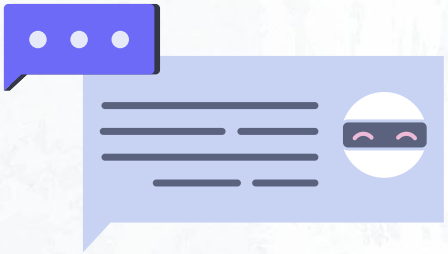
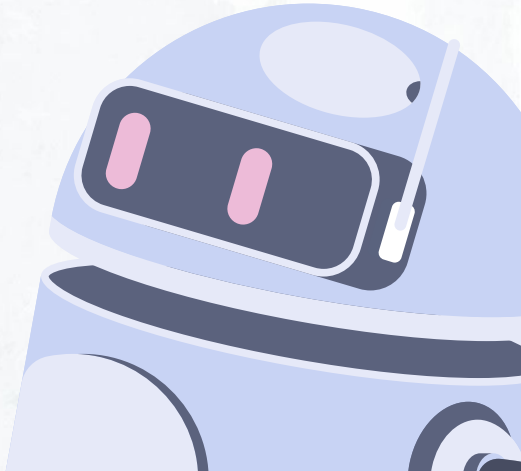
The Impact of Embedding Methods on Faithfulness of Retrieval-Augmented Generation (RAG)

Chang Jin, Reagan Keeney, and Amandeep Kaur Singh



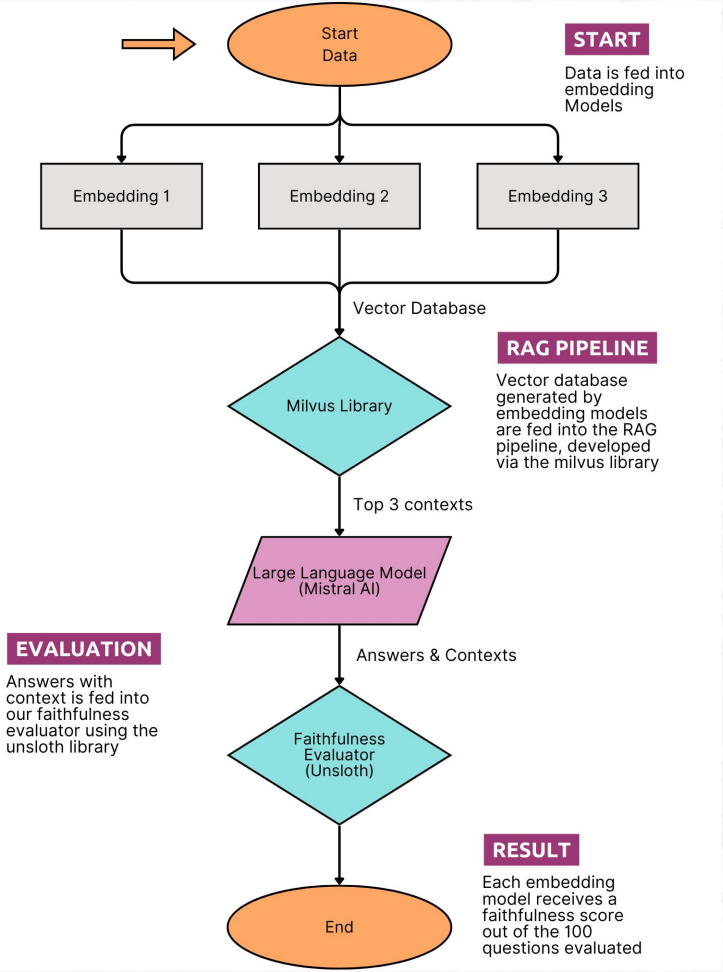


Problem Statement

- Different embedding models are designed with specific goals in mind
  - Embeddings vary on effectiveness
  - No one-size-fits-all solution; performance is highly context-dependent.
- Evaluating RAG systems is a challenging task
- Faithfulness
  - Faithfulness reflects how well retrieved contexts are utilized in LLM responses. Measures “hallucination”.
  - Ensuring accurate incorporation of context is critical for reliable outputs.

Pipeline + Dataset

- RAG Mini BIOASQ Dataset
- Embedding Models:
  - Baseline: bge-small-en-v1.5
  - MedEmbed-small-v0.1 (fine-tuned BERT)
  - All-MiniLM-L6-v2
- Milvus Vector DB
- Mistral AI LLM
- Evaluation



Faithfulness Overview

- Metric for evaluating output
  - Non-traditional- unique to RAG
- Number of Truthful Claims/Total Number of Claims

Faithfulness Evaluation

In [ ]:

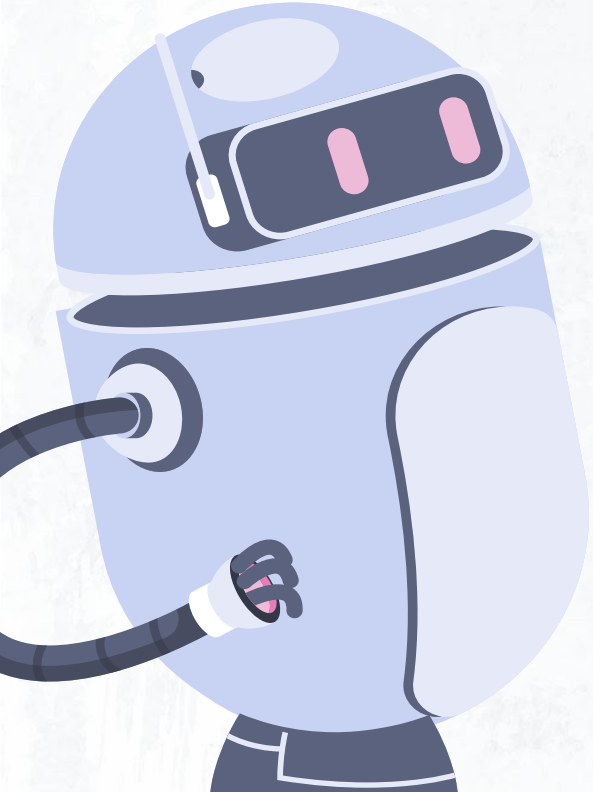
```
# Load the embedding model and malvus_client if not loaded yet
import pandas as pd
drive_dir = "/content/drive/MyDrive/646projectstuff"

# CHOOSE ONE OF: "BAAI/bge-small-en-v1.5", "abhinand/MedEmbed-small-v0.1", "sentence-transformers/all-MiniLM-L6-v2"
EMBEDDING_MODEL_NAME = "sentence-transformers/all-MiniLM-L6-v2"
generation_df = load_generation(EMBEDDING_MODEL_NAME)
generation_df.head()
```

Out[ ]:

	input	actual_output	retrieval_context
0	What is the implication of histone lysine meth...	The implication of histone lysine methylation ...	[We used high-resolution SNP genotyping to ide...
1	What is the role of STAG1/STAG2 proteins in di...	The provided context does not explicitly state...	[Heterochromatin Protein 1 (HP1) was first dis...

Results & Discussion



Model	Faithfulness
bge-small-en-v1.5(baseline)	97%
MedEmbed-small-v0.1	93%
all-MiniLM-L6-v2	92%

Thank You!