Challenge 6

The Symmetric Bridge: Closing the Semantic Gap in Drug Discovery RAG

Challenge Statement
Build a Retrieval-Augmented Generation (RAG) system that overcomes the "semantic gap" in biomedical search. When a researcher asks "How does metformin affect autophagy?", your system should find relevant passages even when the documents use different terminology.

The Problem
Traditional keyword search fails when queries use common terms ("drug name") but documents use technical language ("molecular mechanisms", "pathway interactions"). You need semantic search that understands concepts, not just matches words.

Your Task
Build a RAG pipeline that:
✅ Indexes biomedical documents (PubMed abstracts or papers)
✅ Uses semantic embeddings for better retrieval
✅ Generates answers grounded in source material
✅ Handles multi-hop reasoning (answer requires multiple documents)
✅ Shows source citations for transparency

Recommended Datasets
✅ PubMed Central Full Text (bigquery-public-data.pmc_open_access_commercial.*) ~3 million CC-BY licensed full-text articles
✅ ChEMBL (patents-public-data.ebi_chembl.*) Drug-target interactions, compound properties

Recommended Vertex AI Services:
✅ Gemini models for embedding and generation
✅ Vector Search for semantic retrieval
✅ Vertex AI Search (optional alternative approach)

Suggested Acceptance Criteria
Teams should define their own acceptance criteria and provide a rationale. Here is guidance:

Evaluation Set
30 questions across three difficulty levels:

- Simple (10): Direct factual retrieval (e.g., "What is the mechanism of action of metformin?")
- Multi-hop (10): Answers requiring synthesis across documents (e.g., "Which diabetes drugs affect both glucose metabolism and autophagy?")

- Semantic gap (10): Queries where the terminology differs from source documents (e.g., "How do SGLT2 inhibitors impact heart failure?" where papers may not use those exact terms)

Minimum Bar

- RAG pipeline indexes biomedical documents and retrieves relevant passages using semantic (not just keyword) search
- Generates grounded answers with source citations for each claim
- On the 30 evaluation questions: answers are relevant and supported by cited sources for ≥70% of questions
- Retrieval finds relevant documents even when query terms don't exactly match document terminology (demonstrate on ≥3 semantic gap questions)

Competitive (Pick ≥2)

- Show measurable improvement of semantic search over keyword baseline on the semantic gap questions
- Handle multi-hop questions by retrieving and synthesizing across multiple documents, with citations traced to each source
- Rank or filter retrieved passages by relevance before generation to reduce noise
- Detect when retrieved evidence is insufficient and say "I don't know" rather than hallucinate

Stretch

- Integrate ChEMBL drug-target data alongside PubMed to enrich answers with structured data
- Evaluate hallucination rate: for a subset of answers, verify that cited papers actually support the generated claims
- Show how retrieval quality degrades gracefully as questions get harder (simple → multi-hop → semantic gap) with quantitative comparison

LLM usage is expected for this challenge — document your embedding strategy, retrieval approach, and generation prompts. Evaluate faithfulness of generated answers to retrieved sources.