Modeling Proposal and Procedures:

1. Generate labeled sample from Best Buy Employee subreddit data
    a. Chunked individual comments and submissions into a list of sentences/contexts
    b. Hand-wrote 3 questions, sampled 30 random contexts and 10 contexts close in embedding space (embedding model: multi-qa-mpnet-base-dot-v1)
    c. Manually (human) labeled 40 contexts per question as relevant (True) or irrelevant (False)
    d. Constructed a ranking based on the sum of relevant labels (Treat as our ground truth)
    e. Resample contexts for each question
        i. Take top 10 according to sum of relevant labels (Use this as our consensus True)
        ii. Take 20 whose sum are 0 (use as our consensus False)
2. Evaluate retriever pipelines for quality across embedding models, similarity metrics, pipeline methodology
    a. Procedure:
        i. Pick an embedding
        ii. Fill the retrieval system (database) with 30 contexts for each question
        iii. Pick a similarity metric
        iv. Retrieve 10 documents
        v. Calculate precision, recall, f1 scores
    b. Pipelines:
        i. Retriever Class
        ii. Chroma vector db implementation
        iii. Qdrant vector db implementation
3. Evaluate pipeline for retrieval time
    a. Procedure:
        i. Fill pipeline with all chunked contexts for a given subreddit (Best Buy)
            1. Embed each context
            2. Add to database for retrieval
        ii. For a given list of questions
            1. Run retrieval on question
            2. Calculate time to retrieve
4. Evaluate quality on advanced pipelines/procedures
    a. Does clustering of embeddings offer any improvement?
    b. Does self-querying retrieval (using an llm to split an input query into a filtering query for metadata and an embedded query) offer significant improvement?
    c. How does multiquerying impact performance?
        i. Classification of question itself
    d. How does parent doc retrieval impact performance?
    e. Does embedding on the summary of threads improve performance?
    f. Can we account for temporal questions?