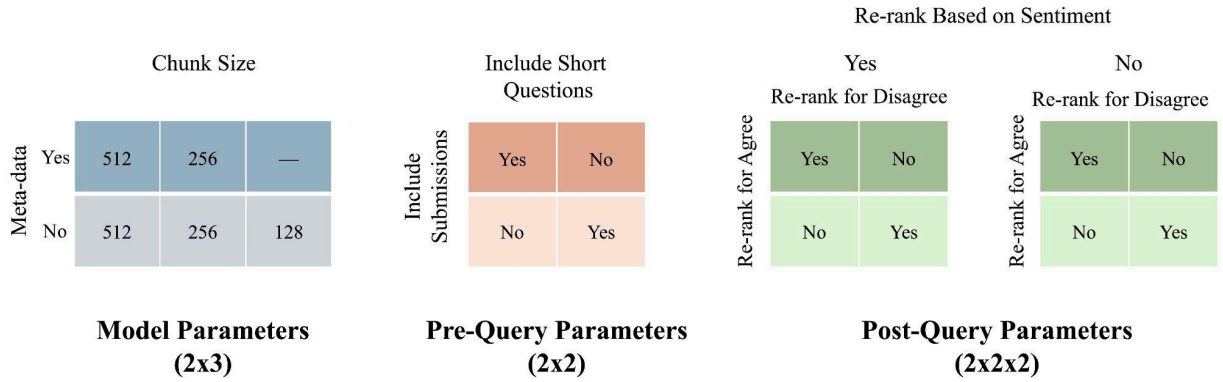
Executive Summary: Pawsitive Retrieval, Aware Data Project

Objective: This project aims to build a model to efficiently identify and rank relevant content from a large dataset of human-generated Reddit posts (5.5 million posts from 34 subreddits), given an arbitrary user query. Key objectives were to retrieve highly relevant results for queries while keeping retrieval times under 1 second. The long-term application is to use this capability as part of a Retrieval-Augmented Generation (RAG) pipeline for Aware clients.

Methodology: We focused on systematically varying parameters of our chosen embedding model (General Text Embeddings; GTE) as well as leveraging relationships inherent in the structure of the data (e.g., replies to comments or submissions) to improve quality and ranking of results before and after vector search. This included using prefiltering to remove short questions or top level submissions, and using engineered metadata like sentiment scores and "agree distances" of reply texts to re-rank results. This resulted in 160 total model variants.



Model Details: The GTE model was chosen to generate vector embeddings of our text data due to its compact size (0.22GB), ability to embed up to 512 tokens, its performance on retrieval tasks, and because it was partially trained on Reddit data. LanceDB, an open source vector database, was used to index and search embeddings. It uses a combination of an inverted file index (IVF) and product quantization (PQ) to build an approximate nearest neighbors (ANN) index. This framework allowed us to vary parameters for embedding, indexing, and retrieval. Our experiments focused on embedding and retrieval.

Evaluating Model Performance: Queries were manually created for a subset of the data (13 subreddits) and more than 1000 query/result pairs were human labeled to quantify relevance. We used this to measure the effect of different model parameters on relevance of retrieved results, using modified versions of three different metrics: Mean Reciprocal Rank, Extended Reciprocal Rank, and a Normalized Discounted Cumulative Gains.

Results and Future Work: Applying a pre-retrieval filter to omit short questions resulted in significant improvement across all metrics compared to our baseline model. The best overall model also incorporated two re-ranking strategies that rely on our engineered metadata (reply sentiment and "agree distance"). For future work, we would explore refinements to these filterings and rerankings, as well as optimizing indexing parameters to improve our retrieval time of 300-450ms.