Preliminary Results:

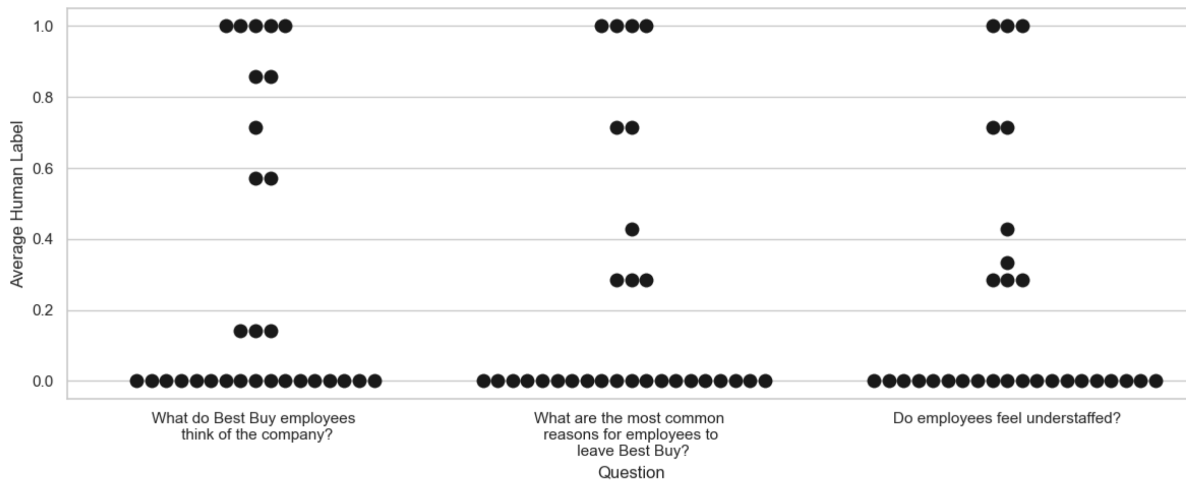
Introduction:

Herein, we present the preliminary work to build an infrastructure to evaluate different retrieval pipelines for use in a RAG system. The dataset used to evaluate these pipelines was taken from submissions and comments of a reddit thread for Best Buy employees. First, we share reference to a set of python classes which all inherit a common base class enabling uniform calling procedure and output. These classes will form the basis of advanced pipelines to be explored within this project. Three questions typical to the type of questions Aware's clients would ask of the data were handwritten. 30 statements sampled from the reddit thread were labeled by 7 observers for each question as relevant (True) or irrelevant (False) to the question posed. Examples of statements with varying levels of relevance are shared in addition to a plot showing the distribution of the labels. A procedure for evaluating different RAG pipelines on this dataset was then used to compare the quality of retrieval using different embedding models to convert statements and questions into an encoded vector space. Lastly, a large language method was used to create a prototype for automated relevance labeling.

1. **Working minimal prototypes for retrieval**
    a. Custom written "Retriever" class utilizing sbert embeddings
    b. Vector database prototypes
        i. ChromaDB
        ii. Qdrant
2. **Manually labeled subset of Best Buy employee subreddit data for 3 questions**
    a. *Example Samples:*
        i. Question: Do employees feel understaffed?
            1. Statement 1:
                a. Content: "Absolutely. I had a talk with a leader last week about this, pointing out that we're running a bare minimum of staff who have to know and do more than ever before while being paid less than we were just two years ago."
                b. Total Human "True" Labels: 7
                c. Average Label: 1.0
            2. Statement 2:
                a. Content: "Typical day; It's frustrating to have more do nothing managers on my shift today (5) than we have people on the sales floor (4). They spend their time talking amongst themselves, but heaven forbid you take 10 seconds to say hi to a coworker. I barely get time to take a breath between customers before I get orders barked at me- yet they coast through the day."
                b. Total Human "True" Labels: 5
                c. Average Label: 0.7
            3. Statement 3:
                a. Content: "Interesting...my store is primarily African American females in most roles. There are Caucasian, Hispanic and a wide mix of everything else in the store but leadership is primarily female and African American. I will say our

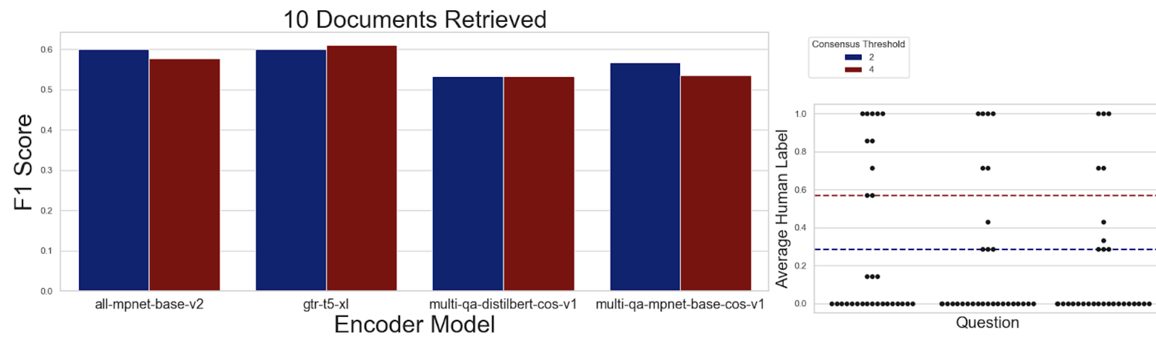
diversity is something I appreciate in my store all the other issues are a horse of a different color.”
    b. Total Human “True” Labels: 2
    c. Average Label: 0.29
4. <u>Statement 4</u>:
    a. Content: “How many was in stock?; I'm in hawaii. It's 6:45am. So I don't wait in line longer, how many 30x cards did your stores in the mainland get?”
    b. Total Human “True” Labels: 0
    c. Average Label: 0

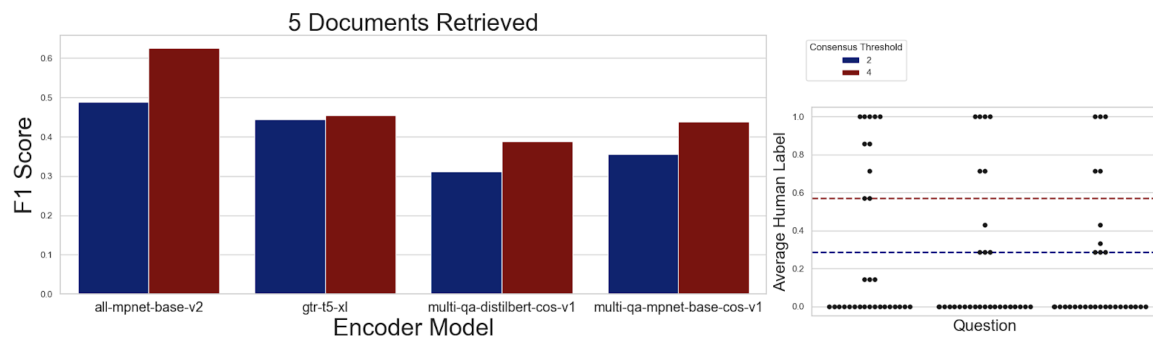
b. *Summary Plot:*



The swarmplot above shows the distribution of labels of the statements in our dataset. For each question, 7 human observers labeled each of 30 statements as either relevant (1) or irrelevant (0). Each statement is represented above with a black circle, plotted at the average label across all 7 observers. This displays the level of subjectivity within this data as well as giving a representation to the frequency of the varying levels of relevance.

3. Preliminary Quality Comparison:
    a. Looked into options for evaluation
        i. Ragas
        ii. Manual scoring (f1, recall, precision)
    b. *Preliminary Results with Manual Scoring:*
        i. The f1 score was calculated for 4 different encoding models using different thresholds (minimum number of observers defining
        ii. Scoring metric does not appear to make any difference from a quality standpoint

iii. 

iv. 

v. For the embedding models tested, “all-mpnet-base-v2” performs the best and is a smaller, faster-running model than the second place performer (gtr-t5-xl)

4. Looked into automation of test set generation
    a. Setfit
    b. Ragas
    c. *Manual llm implementation:*
        i. This makes use of the `langchain_community.llms Ollama` module. Alternatively, this could be run with any [langchain llm]() (e.g. OpenAI, HuggingFaceHub). Question and statement pairs are submitted to a large language model (llm) to be labeled as relevant or irrelevant and to provide the reason why.
        ii. Engine/model: Running dolphin-mixtral locally utilizing [Ollama]()
        iii. Example Prompt:

```
<s>[INST] <<SYS>>
The following statement (delimited by ```) provided below is a
response from an employee at the company of interest.
The statement should be taken as is. It cannot be used to
further a dialogue with the employee.

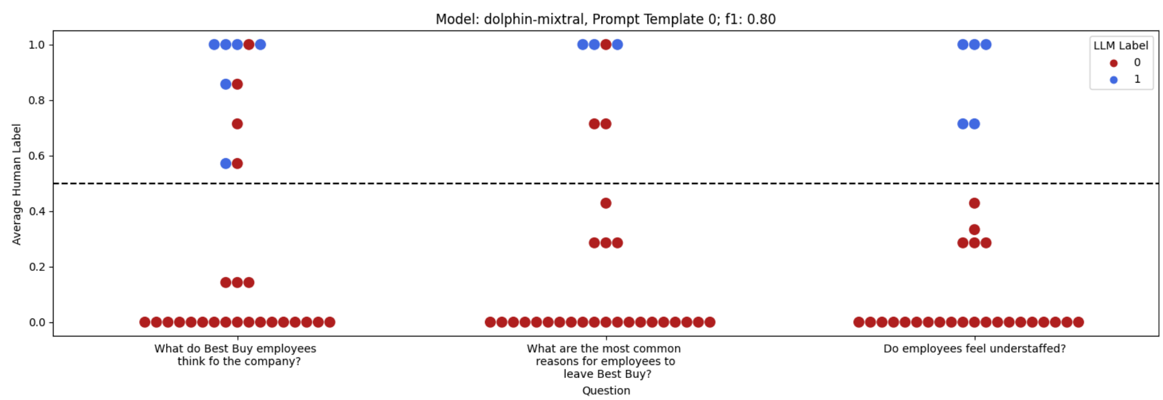
Please format the output as a dictionary with the following
keys: "relevant", "reason".
Relevant should be a boolean value indicating whether the
statement is relevant to the question.
Reason should be a string explaining why and how the statement
is or is not relevant.
<</SYS>>
```

~~~~
Does the statement below help answer the question: What do Best
Buy employees think of the company?
```
Statement: Don't listen to this guy, I work there and the team
environment is outstanding everyone stands around talking to
each other and let's the antisocial people ring up the
customers. You'll enjoy Best Buy as long as you aren't
antisocial and you actually enjoy technology
~~~~

iv. Output:

```
{
    'relevant': True,
    'reason': "The statement provides a positive opinion about
    the company's team environment and work culture. It
    suggests that employees at Best Buy enjoy their job if
    they are not antisocial and have an interest in
    technology."
}
```

v. Preliminary Results:



Initial performance is promising as the llm model using the “dolphin-mixtral” model correctly labels 10 out of the 12 statements unanimously labeled as relevant. Using a consensus threshold of 50% of human labelers, the llm correctly labels all of the irrelevant statements and produces an F1 score of 0.80.