1. SELECT DATA

1.1. Rationale for inclusion/exclusion

We selected 3000, randomly sampled context, question, and answer triplets from the TORQUE dataset. The selected dataset was divided equally among 3 different categories from the overall dataset, noted as categories 0, 1, 2. Each category corresponds to “What events have already finished?”, "What events have begun but not finished?” and “What happened after/before an event?” type questions respectfully.

The decision to sample only 3,000 questions from the larger dataset of 25,000 was driven by considerations of our computational resources. We determined that 3,000 questions offer a sufficiently diverse range of data for model training, while also ensuring that the dataset size remains manageable for processing within our computational constraints. This balance allows for efficient model training without overwhelming our resources.

We decided to include some questions without answers so we could supply our models with negative examples during training, as in the real world it is possible to encounter questions that cannot be answered.

3. CONSTRUCT DATA

To input our data into BERT we must use word embeddings. The HuggingFace Transformers library is an ideal choice for generating word embeddings for LLMs.
For non-answerable questions we will train Bert to predict a null token or “not answerable” string, which may involve augmenting the current data by replacing null answers with a string or token.

We do not plan to change the Context, Question, Answer format already present in Torque.

4. INITIAL INFERENCE

This data has no need for merging or aggregation. Instead, we would like to highlight BERT’s current performance by running inference on some of our questions. Some sample inputs and outputs are listed below:

Input 1:

Context: The coalition won in 2002 on a wave of euphoria after 24 years of rule by Daniel Arap Moi, but now is in a precarious position because of growing public dissatisfaction. Both camps agree that the existing constitution is outdated and oppressive but have failed to reach consensus on the new one, which has been the subject of debate since 1997.

Question: What event has already finished?

Output 1:

BERT Response: 24 years of rule

Expected keywords: won, wave, rule

Input 2:

Context: The security cabinet took the decision on December 6, a day after a suicide bomber from the radical Palestinian movement Islamic Jihad killed five Israelis at the entrance to a shopping mall. Under the deal, Israel should also allow convoys of trucks to travel between Gaza and the West Bank from January 15.

Question: What is likely to happen after the decision was taken?

Output 2:

BERT Response: suicide bomber from the radical palestinian movement islamic jihad killed five israelis at the entrance to a shopping mall

Expected keywords: allow, travel

These examples show the limitations of BERT inference prior to fine tuning. It seems as if BERT is relying on phrases that come after the keyword “after” in the context, or assuming similar results based on the structure of the sentence, but further testing will be needed to confirm any hypotheses.

5. FORMAT DATA

All the attribute values are numeric in the aggregated unified table except for the "Context," "Question," "Answer," and "Category" fields, which will need to be encoded to multidimensional numeric attributes if the modeling tool requires numeric input.

6. DATASET DESCRIPTION

The final dataset consists of 3,000 samples and 4 features. This provides a comprehensive overview of the data structure, indicating a relatively moderate dataset size with a manageable number of features for analysis or modeling purposes.

There are 748 examples of non-answerable questions. We may decide to replace some of these examples with answerable questions or add more answerable questions to the dataset if we find that training could be improved.