1. EVALUATE RESULTS

1.1 Business Objectives

The main business objectives for this report were to enhance the accuracy with which LLMs interpret and reason about temporal information and to expand their applicability in areas where temporal reasoning is crucial.

1.2 Business Success Criteria

The business success is defined by the models' ability to interpret temporal information more accurately, leading to improved decision-making processes, with the end goal of achieving measurable improvements over current benchmarks. Success would also be indicated by positive feedback on the improved performance of LLMs in practical and research applications where temporal data is integral.

1.3 Data Mining Goals

The technical goal was to evaluate and improve the temporal reasoning capabilities of LLMs using the TORQUE dataset, formatted as a prediction task to predict the correct answers to temporal reasoning questions.

1.4 Data Mining Success Criteria

Technical success was to be determined by the models' ability to accurately answer questions from a temporal reasoning dataset and improve upon baseline performance metrics established in previous studies. The evaluation metrics (TopK, Accuracy, ROUGE, and BLEU scores) serve as our benchmarks for success.

1.5 Evaluation of Model Against Business Objectives and Success Criteria

The models were subjected to rigorous testing, with the following key findings:

- TopK Accuracy: Provided insights into how often the correct answer was within the top predicted responses, crucial for understanding the models' intuitive grasp of temporal context.
- Accuracy: Offers a straightforward metric of correct predictions, indicating the models' precision and reliability in temporal reasoning.
- ROUGE: Assessed the overlap between the model-generated answers and a set of reference answers, highlighting how well each model captured the essential elements of correct temporal reasoning.
- BLEU: Measured the models' ability to generate answers that match the reference temporally reasoned responses syntactically and semantically.

FLAN-T5 and BERT showed a notable degree of comprehension in answering temporally-relevant questions, with FLAN-T5 shining in BLEU scores, suggesting a strong alignment with the expected answers. BERT, while robust in its context understanding, showed mixed results across TopK and ROUGE, hinting at possible areas for further model refinement. Falcon 7B's initial evaluation revealed a need for refinement, with a near-zero average BLEU score and low ROUGE and TopK accuracies. However, fine-tuning led to noticeable improvements: BLEU score rose to 0.00234, ROUGE increased to

roughly 0.098, TopK accuracy climbed to 0.94, and overall accuracy reached 0.88, indicating a significantly enhanced understanding of temporal reasoning.

1.6 Assessment of Data Mining Results with respect to Business Success Criteria

The evaluation metrics we chose was TopK Accuracy, which measures the model's ability to capture the correct answer within a range of top predictions.

	Accuracy Scores	
Model Name	Untrained Accuracy	Trained Top-10
BERT	0.2827	0.2831
FLAN-T5	0.2211	0.9544
Falcon-7B	0.2012	0.8854

The results from BERT do not show a significant improvement over the base model with regards to the temporal reasoning abilities of LLMs. The fine-tuned model exhibited an accuracy of 0.2831, while the untrained model exhibited an accuracy of 0.2827. That is an increase of 0.004, which is rather lackluster.

The evaluation results from the FLAN-T5 model demonstrates significant progress towards our objective of improving the temporal reasoning abilities of LLMs. The model exhibited an accuracy of 0.9544 after fine-tuning on the temporal dataset, meaning that in 95.44% of the cases, the true label was among the predictions made by the model. This is a substantial increase from the untrained model's performance with an accuracy of 0.2211.

Falcon 7B's evaluation showed a huge contrast before and after fine-tuning. The base model's performance indicated significant room for improvement, as shown by its average BLEU score near zero and ROUGE scores barely above baseline, with an overall accuracy of 0.20. After fine-tuning, Falcon 7B showed improvement in all metrics. The average BLEU score rose to 0.00234 and the ROUGE scores saw a significant increase, particularly ROUGE-1, which rose to approximately 0.098, suggesting a more considerable overlap with the reference answers. Most notably, the accuracy after fine-tuning reached 0.88, suggesting that the model is now highly adept at answering temporal reasoning questions correctly.

1.7 Approved Models

After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.

Given the discussion in the previous subsection, the fine-tuned FLAN-T5 and Falcon7B models were selected.

2. REVIEW PROCESS

The overview of the data mining process:

1. Business Understanding: The objectives and success criteria in the broader context of advancing LLM capabilities were clearly defined, focusing on enhancing temporal understanding within our hardware constraints.
2. Data Preparation and Processing: After a thorough exploration of the TORQUE and TempReason datasets and their limitations, we constructed our own temporal relation dataset through careful selection, cleaning, and formatting of the data. The decision to sample 3000 question-answer pairs, including non-answerable questions, ensured a balanced and representative dataset for model training and evaluation.
3. Predictive Model Training: The choice of FLAN-T5, BERT, Falcon-7B, and Graph Generation models demonstrated a strategic approach to tackle temporal reasoning from different architectures (encoder-decoder, encoder-only, decoder-only). The use of TopK, ROUGE, and BLEU scores provided a comprehensive set of quantitative measures for model performance.

All three steps of the data mining process are necessary and were executed optimally. Our process yielded promising results, particularly in the fine-tuning of FLAN-T5 and Falcon-7B. However, we would have liked to experiment with larger decoder-only models that are more representative of LLM use cases today, as well as larger datasets. We were confined to working with smaller models from each architecture and a smaller dataset largely due to our hardware constraints.

3. DETERMINE NEXT STEPS

Based on the assessment results and the process review, the project team decides how to proceed. Decisions to be made include whether to finish this project and move on to deployment, to initiate further iterations, or to set up new data mining projects.

The model could likely be deployed for inference using at minimum T4 GPUs with 16GB of VRAM.

The current fine-tuned model still has a long way to go. While the accuracy metric was significantly improved in the ⅔ models, we saw a lot of overfitting with responses often containing key-words but lacking generalizable temporal reasoning.

The data mining process can be refined into the following steps:

1. Data processing
2. Model selection
3. Model fine-tuning/training
4. Model evaluation

3.1 Decision

The primary decision we made throughout this process was the models we selected:

- BERT: We believed that a MLM model like BERT would be a good choice for temporal reasoning as well as being easy to train on commodity hardware. In the end, it turns out that it did not produce as good of results as the other two models on this list.
- FLAN-T5: FLAN-T5 was picked as another easy-to-train model that we believed would perform well for this task and it turned out that this was true. While there may be some overfitting issues, the performance of the model is undoubtedly very strong.
- Falcon-7B: Falcon was picked to provide a different type of model (decoder-only) to provide more comparison points. While Falcon was not as easy to train as the other two, it also performed well (albeit some overfitting issues).