1. BUSINESS OBJECTIVES

1.1. Background

The surge in data volume and complexity has underlined the critical need for advanced analytical capabilities in decision-making processes across industries. Large Language Models (LLMs) have emerged as pivotal tools in deciphering and generating natural language, opening new frontiers in data analysis and interpretation. However, their current capacity for temporal reasoning – the ability to understand, reason, and make predictions based on temporal information – remains underdeveloped. This limitation hinders their application in cases where time-sensitive data is crucial, such as forecasting market trends, optimizing logistics, and personalized healthcare.

1.2. Business objectives

The primary business objectives for this project are as follows:

- Enhance Temporal Reasoning Accuracy: Improve the accuracy with which LLMs interpret and reason about temporal information to reduce ambiguity in temporal expressions and improve causal reasoning.
- Expand LLM Applicability: Broaden the range of applications for LLMs by enhancing their ability to understand and utilize temporal information, making them viable for industries and tasks where time is a critical factor.
- Innovate Competitive Advantages: Establish a competitive edge by developing proprietary LLM enhancements that outperform existing solutions in temporal reasoning tasks.
- Accelerate Decision-Making Processes: Enable faster and more reliable decision-making by leveraging improved temporal reasoning capabilities.
- Reduce Operational Costs: Decrease the reliance on human intervention for interpreting and analyzing temporal data, leading to significant cost savings.

1.3. Business success criteria

Success criteria are vital for measuring the project's impact and ensuring alignment with business goals. For this project, the success criteria include:

- Quantifiable Improvement in Accuracy: Demonstrate an increase in the precision and reliability of temporal reasoning tasks as compared to existing benchmarks.
- Accuracy Improvement Metrics: Achieving a predefined improvement in accuracy for temporal reasoning tasks, as measured against established benchmarks.
- Adoption and Customer Satisfaction: Successful deployment and adoption of the enhanced LLMs within target industries, evidenced by positive customer feedback and reports illustrating tangible benefits.
- Reduction in Decision-Making Time: Quantifiable reduction in the time required for decision-making processes involving temporal data.
- Cost Savings: Demonstrable reduction in operational costs associated with temporal data analysis and decision-making.

By setting clear business objectives and success criteria, the project can focus on delivering tangible benefits and achieving strategic goals. This approach ensures that the enhancements to LLM temporal

reasoning capabilities translate into competitive advantages, operational efficiencies, and new potentials in AI applications.

2. ASSESS SITUATION

2.1. Inventory of resources

The project will leverage Purdue University's Rosen Center for Advanced Computing (RCAC) resources, specifically the Gilbreth Community Cluster optimized for GPU-intensive applications. Access to the Gilbreth's sub-cluster K, equipped with Nvidia A100 has not yet been confirmed. This sub-cluster offers 52 nodes, each with 64 cores and 512GB memory, aligning with our computational needs.

An alternative option would be to use an Nvidia RTX 3090 with 24GB of VRAM. This GPU is currently available to one of our team members.

Primary datasets include TORQUE and TempReason, accessible online. Relevant background knowledge includes basic familiarity with pre-training, fine-tuning, and evaluating LLMs. Existing literature on temporal reasoning and LLMs will supplement our research.

The project sponsor is our supervisor, Professor Bharat Bhargava from the Purdue Computer Science department.

2.2. Requirements, assumptions, and constraints

The project aims for high accuracy in temporal reasoning with LLMs, ensuring the results are easily interpretable. Data usage complies with legal standards. Completion is scheduled within the semester, aligning with resource availability and academic milestones.

The project assumes that the primary datasets, TORQUE and TempReason are of high quality, though minor inaccuracies due to human error are acknowledged. For effective presentation to stakeholders, it's assumed that a simplified explanation of the LLM's workings and its performance in benchmarks will be necessary.

Budget is yet to be determined. It is possible that no funding will be allocated, and that the project will have to be bootstrapped with personal hardware. It is also possible that the pre-training models on hardware available will take longer than the semester allows. Data processing and model training are limited to the computational resources available based on funding granted.

2.3. Risks and contingencies

Regarding the access to the Gilbreth Cluster, our ability to pre-train LLMs on the Nvidia A100s will be determined by the funding allocated. In the event that we are unable to access Gilbreth, be will be forced run the smallest models with minimal pre-training. Additionally, we will have to utilize downscaling techniques such as quantization and low-rank adapters. This is not ideal.

Regarding the TORQUE dataset, event annotations are largely validated, however there is still a small

chance that relations are labeled incorrectly, as human workers were tasked with creating and validating entries, and there is always a non-zero probability that there may be incorrect data.
Regarding the TempReason dataset, all of the temporal expressions contain only textual months or numeric years. This lack of diversity in including all types of dates will not make it applicable to all temporal situations. This dataset also shares a similar issue to the TORQUE dataset mentioned above, in which using crowdsourced data risks the potential that some entries in the dataset are incorrect.

Regarding the TORQUE dataset, we will not need to find ways to remedy the miniscule probability of incorrectly inputted data, as the dataset has already been thoroughly checked, and any inconsistencies or errors will make up a negligible part of the dataset and will have a negligible impact. It is not in the interest of our time and resources to fix this potential trivial issue.

Regarding the TempReason dataset, we will not fix the human errors in the entries for the same reasons listed above. If we find a need to have a more robust representation of dates, then we may decide to look for solutions in parsing the data and inputting a variety of dates, but this is not a priority for us.

2.4. Terminology

Large Language Model (LLM) - a model trained on large amounts of text data to imitate written language based on predictive text generation
Temporal Reasoning - logic pertaining to time-based relationships and events; time series
BERT (Bidirectional Encoder Representations from Transformers) - framework designed to decipher meaning of text based on the context of other text
LLaMA2 - generative pre-trained transformer, considered a state of the art open source LLM released by Meta

2.5. Costs and benefits

The primary costs involve the annual subscription fee for accessing Gilbreth's sub-cluster K ($2,200 per GPU) and personnel time for data preparation and model training.
Anticipated benefits include advancing the understanding of temporal reasoning in LLMs, potentially leading to innovations in NLP applications and contributing to academic knowledge.

3. DATA MINING GOALS

3.1. Data mining goals

The challenge this project is centered around is evaluating and improving large language models for temporal reasoning tasks. This could be formatted as a prediction task, where the goal is to predict the correct answers to temporal reasoning questions.

3.2. Data mining success criteria

The success of this project will be based on how much we can improve an LLMs temporal reasoning abilities. We will pick one or more models to test, like BERT for example and analyze how many questions it can correctly answer from a temporal reasoning dataset without adding any modifications to the way questions are inputted into models or changing the architecture of the models. We will also

initially test with off the shelf large language models without any pre-training. From there we will attempt to improve BERTs availability to answer these temporal reasoning questions. Improvement will be quantified by how many more questions the model can answer correctly.
A successful outcome for this project will be improving the question-answering ability of at least one model for at least one temporal reasoning task.

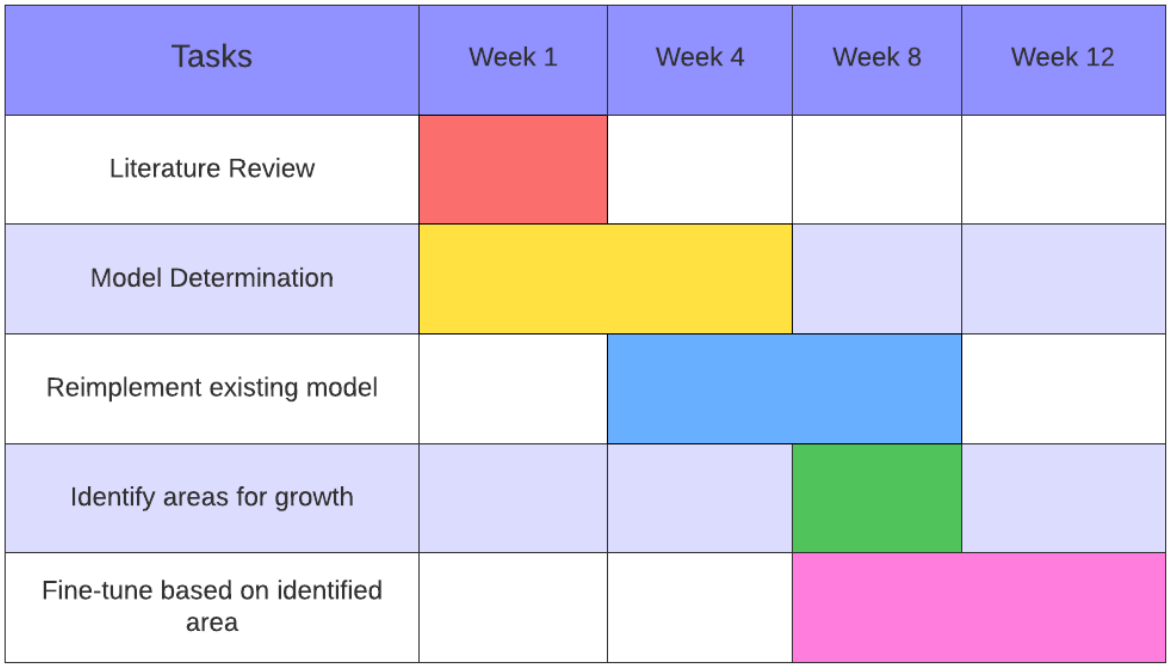
4. PROJECT PLAN

4.1. Project Plan

Our project is expected to be completed within 12 weeks (the end of the semester). Our goal is to complete literature review, define our computational limits, and familiarize yourself with HuggingFace and LLMs within the first 3 weeks.

We then intend to spend the next 5 weeks reimplementing an existing paper attempting to improve the temporal reasoning abilities of a well-known model (BERT, Llama, etc.) and determine the steps to fine-tuning that model.

Lastly, we shall spend our remaining 4 weeks identifying an area in temporal reasoning where we could propose a novel change, attempting to fine-tune our model of choosing using training data, and improving the model in some meaningful way.



Some critical points in our timeline will include:

1) Identification of the model we will work on (within the first 3 weeks)

2) Reimplementation of an existing work to improve temporal reasoning (between week 3 - 8)
3) Determination of areas for novel growth in our model of choice (between week 8 - 12)

4.2. Initial assessment of tools and techniques

To enhance the temporal reasoning capabilities of Large Language Models (LLMs), careful selection of tools and techniques is crucial. The criteria for such selection include support for temporal reasoning tasks, flexibility for customization, scalability to handle large datasets, interpretability to understand model processes, compatibility with existing frameworks, performance benchmarking, community support, and ethical considerations.

Potential tools and techniques encompass temporal reasoning algorithms like temporal logic and causal inference, libraries such as AllenNLP and TensorFlow Extended (TFX), the datasets provided to us by the project coordinator, and benchmarking frameworks like TRAM.

These candidates must undergo evaluation based on their alignment with project requirements, performance metrics, and scalability. Evaluation outcomes will inform the prioritization and refinement of techniques, ensuring that selected approaches effectively address the project's goals of improving temporal reasoning capabilities in LLMs while maintaining fairness and accuracy in their outputs.