1. DETERMINE BUSINESS OBJECTIVES

The first objective of the analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish. Often the customer has many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors at the beginning of the project that can influence the final outcome. A likely consequence of neglecting this step would be to expend a great deal of effort producing the correct answers to the wrong questions.

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

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, while a secondary business objective might be to determine whether lower fees affect only one particular segment of customers.

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

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and readily measurable, such as reduction of customer churn to a certain level, or general and subjective, such as "give useful insights into the relationships." In the latter case, be sure to indicate who would make the subjective judgment.

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent

and sign-up rate by 20 percent)
- Identify who assesses the success criteria

Each of the success criteria should relate to at least one of the specified business objectives.

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

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

2.1. Inventory of resources

List the resources available to the project, including personnel (business and data experts, technical support, data mining experts), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

Hardware resources
- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- [Optional] Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- [Optional] Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

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.

Sources of data and knowledge

- Identify data sources
- Identify type of data sources (online sources, experts, written documentation, etc.) •
  [Optional] Identify knowledge sources
- [Optional] Identify type of knowledge sources (online sources, experts, written documentation, etc.)
- Check available tools and techniques
- Describe the relevant background knowledge (informally or formally)

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.

Personnel sources

- Identify project sponsor (if different from internal sponsor as in Section 1.1)
- [Optional] Identify system administrator, database administrator, and technical support staff for further questions
- Identify market analysts, data mining experts, and statisticians, and check their availability •
  Check availability of domain experts for later phases

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

2.2. Requirements, assumptions, and constraints

List all requirements of the project, including schedule of completion, comprehensibility, and quality of results and security, as well as legal issues. As part of this output, make sure that you are allowed to use the data.
List the assumptions made by the project. These may be assumptions about the data, which can be verified during data mining, but may also include non-verifiable assumptions related to the project. It is particularly important to list the latter if they will affect the validity of the results. List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project in the time required, or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.

Requirements

- [Optional] Specify target group profile
- [Optional] Capture all requirements on scheduling
- Capture requirements on comprehensibility, accuracy, deploy ability, maintainability, and repeatability of the data mining project and the resulting model(s)

• Capture requirements on security, legal restrictions, privacy, reporting, and project schedule

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.

Assumptions

• Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary) • List assumptions on data quality (e.g., accuracy, availability)

• [Optional] List assumptions on external factors (e.g., economic issues, competitive products, technical advances)

• [Optional] Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000)

• List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor)

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.

Constraints

• Check general constraints (e.g., legal issues, budget, timescales, and resources) • Check access rights to data sources (e.g., access restrictions, password required) • Check technical accessibility of data (operating systems, data management system, file or database format)

• Check whether relevant knowledge is accessible

• [Optional] Check budget constraints (fixed costs, implementation costs, etc.)

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

List the risks, that is, the events that might occur, impacting schedule, cost, or result. List the corresponding contingency plans: what action will be taken to avoid or minimize the impact or recover from the occurrence of the foreseen risks.

Identify risks

• [Optional] Identify business risks (e.g., competitor comes up with better results first) • [Optional] Identify organizational risks (e.g., department requesting project doesn't have funding for the project)

• [Optional] Identify financial risks (e.g., further funding depends on initial data mining results) • Identify technical risks

• Identify risks that depend on data and data sources (e.g., poor quality and coverage)

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.

Develop contingency plans

- Determine conditions under which each risk may occur
- Develop contingency plans

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

Compile a glossary of terminology relevant to the project. This should include at least two components: (1) A glossary of relevant business terminology, which forms part of the business understanding available to the project (2) A glossary of data mining terminology, illustrated with examples relevant to the business problem in question

- [Optional] Check prior availability of glossaries; otherwise begin to draft glossaries •
  Talk to domain experts to understand their terminology
- Become familiar with the business 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

Prepare a cost-benefit analysis for the project, comparing the costs of the project with the potential benefits to the business if it is successful

- [Optional] Estimate costs for data collection
- [Optional] Estimate costs of developing and implementing a solution
- Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue)
- [Optional] Estimate operating costs

[Optional] Remember to identify hidden costs, such as repeated data extraction and preparation, changes in workflows, and time required for training.

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. DETERMINE DATA MINING GOALS

A business goal states objectives in business terminology; a data mining goal states project objectives in technical terms. For example, the business goal might be, "Increase catalog sales to existing customers," while a data mining goal might be, "Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information, and the price of the item."

3.1. Data mining goals

Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally technical outputs.

- Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).
- Specify data mining problem type (e.g., classification, description, prediction, and clustering).

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

Define the criteria for a successful outcome to the project in technical terms, for example, a certain level of predictive accuracy or a propensity-to-purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity) • Define benchmarks for evaluation criteria
- Specify criteria which address subjective assessment criteria (e.g., model explain ability and data and marketing insight provided by the model)

Remember that the data mining success criteria are different from the business success criteria defined earlier.

The success of this project will 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. PRODUCE PROJECT PLAN

Describe the intended plan for achieving the data mining goals and thereby achieving the business goals.

4.1. Project plan

List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Wherever possible, make explicit the large-scale iterations in the data mining process—for example, repetitions of the modeling and evaluation phases. As part of the project plan, it is also important to analyze dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations for actions if the risks are manifested.
Although this is the only task in which the project plan is directly named, it nevertheless should be consulted continually and reviewed throughout the project. The project plan should be consulted at minimum whenever a new task is started or a further iteration of a task or activity is begun.

- Define the initial project plan [Optional] and discuss the feasibility with all involved personnel • Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria
- [Optional] Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating timescales for data mining projects. For example, it is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase and 20-30 percent in the Data Understanding Phase, while only 10-20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases and 5-10 percent in the Deployment Phase.)
- Identify critical steps
- [Optional] Mark decision points
- [Optional] Mark review points
- [Optional] Identify major iterations

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.

Tasks	Week 1	Week 4	Week 8	Week 12
Literature Review				
Model Determination				
Reimplement existing model				
Identify areas for growth				
Fine-tune based on identified area				

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

At the end of the first phase, the project team performs an initial assessment of tools and techniques. Here, it is important to select a data mining tool that supports various methods for different stages of the process, since the selection of tools and techniques may influence the entire project.

• Create a list of selection criteria for tools and techniques (or use an existing one if available) • Choose potential tools and techniques
• Evaluate appropriateness of techniques
• Review and prioritize applicable techniques according to the evaluation of alternative solutions

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.