juur.ai

Made by Markus Tamm, Robin Otter

Table of Contents

Task 1

Github repository:

https://github.com/chirbard/gemini-long-context-competition/tree/main

Task 2

1.Identifying Your Business Goals

Background:

The Estonian government has implemented "Bürokratt," an AI-based virtual assistant designed to facilitate public interaction with state services. However, the system currently falls short in terms of accuracy and usability. This project aims to enhance and simplify access to Estonian State laws through the development of a stress-tested, improved AI helper capable of delivering precise and conversational responses. Improving this system is crucial for fostering public trust and encouraging broader adoption of AI in governmental operations.

Business Goals:

The goal is to create a reliable AI assistant that simplifies interaction with legal frameworks, making it user-friendly and accessible to the general public. The enhanced model will improve upon "Bürokratt" by addressing its current shortcomings and delivering a seamless user experience.

Business Success Criteria:

The project will be deemed successful if users can seamlessly interact with the AI model and receive accurate and contextually relevant answers to their legal questions in Estonian. Success also includes an intuitive interface that reduces the need for domain expertise while interpreting complex laws.

2.Assessing Your Situation

Inventory of Resources:

- Hardware and Software: Sufficient computational resources; the necessary software is freely available.
- Data: Access to appropriate legal documents
- Expertise: Team members with relevant knowledge in AI, data analysis, and Estonian legal systems.

Requirements, Assumptions, and Constraints:

- Requirements: Completion within the course deadline, model responsiveness under stress testing, and effective utilization of available legal datasets.
- Assumptions: Users will rely on the AI helper for accurate legal interpretations, and public resources will remain accessible throughout the project timeline.
- Constraints: Limited time due to academic deadlines, with potential dependency on data availability.

Risks and Contingencies:

- Risks: Power outages, technical failures.
- Contingency Plan: Utilize backup office facilities and ensure alternative data storage and power solutions are in place.

Terminology:

Key terms include:

- AI Helper: An AI-based virtual assistant designed for querying legal information.
- Stress Testing: Evaluating the model's performance under high user loads.
- Legal Framework: The body of Estonian laws and regulations the LLM is tuned on.
- LLM (Large Language Model): A type of AI model designed to understand and generate human-like text, forming the foundation of the improved AI helper.

Costs and Benefits:

- Costs: Primarily the time and effort invested by the development team.
- Benefits: Users save time by avoiding the need to read and interpret extensive legal documents. The solution enhances public trust in state services by demonstrating a commitment to innovation and accessibility. Furthermore, the system can serve as a scalable framework for integrating AI into other public service domains.

3.Defining Your Data-Mining Goals

Data-Mining Goals:

- Fine-tune a large language model (LLM) that accurately interprets and answers questions about Estonian State laws.
- Ensure that the model's responses are both linguistically and legally precise.
- Optimize the model to handle stress-testing scenarios effectively and maintain high performance.

Data-Mining Success Criteria:

- The AI consistently provides correct answers to law-related queries.
- Users report high satisfaction levels during testing phases.
- The model passes stress tests, maintaining functionality under high query volumes.

Task 3

1.Gathering Data

Outline Data Requirements:

The primary requirement is to acquire every valid Estonian State law in its most up-to-date form. Outdated information would compromise the AI's ability to provide accurate assistance. The data should be in a structured format to facilitate parsing and training. Each law must be stored in a paragraph-separated text file to allow for modular processing and analysis.

Verify Data Availability:

All Estonian State laws are publicly available on the government's official website, ensuring accessibility. These resources include comprehensive legal texts organized by topic and domain, aligning with the project's requirements for creating a law-focused AI model.

Define Selection Criteria:

To ensure inclusivity, we will select all laws from the government's website, encompassing all legal categories. The goal is to build an exhaustive dataset that covers a wide range of legal topics. Each law will be fully processed, with content segmented into paragraphs for ease of training and contextual relevance.

in its entirety, segmented into paragraphs for easier analysis and training.

2.Describing Data

The dataset comprises text files where each file corresponds to a single legal document. These documents are further organized into chapters and divided into paragraphs for better structure.
Key metadata associated with each file includes:

- Law Title: The name or primary focus of the law.
- Paragraph Texts: The content of the law, divided into digestible sections to facilitate effective training.
- Publication Date: Used to ensure only the most recent laws are included.

This structured approach ensures the AI model can effectively parse, understand, and respond to queries with contextual accuracy.

3.Exploring Data

Preliminary exploration provides the following insights:

- Volume: The dataset includes hundreds of laws and thousands of paragraphs, offering a substantial training set.
- Content Distribution: The data covers diverse legal topics such as civil rights, environmental law, tax regulations, and administrative procedures. Certain domains, like tax law, are more textually dense and intricate.
- Language Characteristics: The laws are written in formal Estonian, which includes technical jargon, legal terminology, and long, complex sentence structures. This necessitates advanced linguistic processing.

Cross-references within legal documents were identified, requiring the model to account for inter-document dependencies to provide complete and coherent answers.

4. Verifying Data Quality

Key aspects of data quality verification include:

- Completeness: Ensuring all laws and their respective sections are present in the dataset. This involves cross-referencing the dataset with the government's official index of laws.
- Consistency: Ensuring uniform formatting across all documents, with no misplaced paragraphs or sentence truncations. Automated checks will confirm the consistency of the source data.
- Accuracy: As the laws are sourced from the official government website, they are expected to be accurate. Regular validation checks against the source will ensure data integrity over time.
- Relevance: Since every law is included, no potentially useful legal content is omitted.
- Language Processing: Legal documents' complexity, such as long sentences and references, necessitates preprocessing. This will include sentence segmentation, elimination of redundant citations, and formatting for better AI understanding.

Task 4

1.Detailed Plan with Tasks

Data Collection and Preprocessing

- **Description:** Collect and preprocess legal data from the Estonian government's website. Ensure completeness, consistency, and format suitability for training.
- **Team Contributions:**
  - Member A: 6 hours
  - Member B: 4 hours
- **Total Time:** 10 hours

Model Fine-Tuning

- **Description:** Fine-tune a pre-trained large language model (LLM) with the prepared dataset, optimizing for Estonian legal terminology and structure.
- **Team Contributions:**
  - Member A: 8 hours
  - Member B: 10 hours
- **Total Time:** 18 hours

Stress Testing and Evaluation

- **Description:** Perform stress testing on the fine-tuned model to ensure reliability under high query loads. Evaluate the model's accuracy and user satisfaction through simulated interactions.
- **Team Contributions:**
  - Member A: 4 hours
  - Member B: 5 hours
- **Total Time:** 9 hours

Interface Development

- **Description:** Create a user-friendly interface for interacting with the AI assistant. This includes basic design, integration with the model, and functionality testing.
- **Team Contributions:**
  - Member A: 8 hours
  - Member B: 8 hours
- **Total Time:** 16 hours

Final Testing and Documentation

- **Description:** Conduct final tests, prepare project documentation, and create a presentation summarizing the project's objectives and outcomes.
- **Team Contributions:**
  - Member A: 4 hours
  - Member B: 3 hours
- **Total Time:** 7 hours

2.Methods and Tools

- **Tools:** Python, Docker, Jupyter Notebooks, Flask (or similar for interface development), and stress-testing frameworks (e.g., Locust)
- **Comments:**
  - Allow buffer time for potential issues during fine-tuning and stress testing.
  - Ensure iterative testing and integration between the AI model and the interface.