

Case Study 1

Biography Generator Version 2

Project Plan Blueprint - Draft

Team Members

Mr. Ashwith Anand Poojary (11038008)

Ms. Muskaan Chhabra (11038196)

Mr. Lakshith Reddy Baddula (11037634)

Ms. Aadishree Borkar (11037522)

Mr. Rakesh Hadne Sreenath (11038138)

Prof. Dr. Binh Vu

Applied Data Science and Analytics

Document Approvals

|  |  |  |  |
| --- | --- | --- | --- |
| **Role** | **Name** | **Signature** | **Date** |
| External Stakeholder | Dr. Dennis Möbus |  | 16-04-2024 |
| Internal Stakeholders | Prof. Dr. Binh Vu  Prof. Dr. Swati Chandna |  | 16-04-2024 |
| Technical Advisors | Sina Mehraeen  Dr. Kamelia Reshadi |  | 16-04-2024 |
| Product Owner | Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |  | 16-04-2024 |
| Scrum Master | Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |  | 16-04-2024 |
| Scrum Team | Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |  | 16-04-2024 |

Roles:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprints** | **Product Owner** | **Scrum Master** | **Scrum Team** |
| Sprint 1 | Mr. Rakesh Hadne Sreenath | Mr. Ashwith Anand Poojary | Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar |
| Sprint 2 | Ms. Muskaan Chhabra | Ms. Aadishree Borkar | Mr. Ashwith Anand Poojary  Mr. Rakesh Hadne Sreenath  Mr. Lakshith Reddy Baddula |
| Sprint 3 | Mr. Lakshith Reddy Baddula | Mr. Rakesh Hadne Sreenath | Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Ms. Aadishree Borkar |
| Sprint 4 | Ms. Aadishree Borkar | Mr. Lakshith Reddy Baddula | Ms. Muskaan Chhabra  Mr. Rakesh Hadne Sreenath  Mr. Ashwith Anand Poojary |
| Sprint 5 | Mr. Ashwith Anand Poojary | Ms. Muskaan Chhabra | Mr. Rakesh Hadne Sreenath  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar |

Planning Basics

Scope:

The Institut für Geschichte und Biographie (IGB) at the FernUniversität in Hagen, Germany, is renowned for its contributions to Oral History. A key asset of the institute is Archiv "Deutsches Gedächtnis" (ADG), which houses an extensive collection of interviews in digital transcript, video, and audio formats. Given the length of these interviews, there is a pressing need for brief descriptions that enable researchers to swiftly gain an overview of the entire interview content, facilitating quicker access to relevant information.

The primary objective of this project is to develop a comprehensive pipeline leveraging generative AI to automatically generate concise biographies from biographical interviews. Focused on aiding historians and archivists, the system aims to address the time-consuming process of manually creating short biographies and the occasional absence of such documentation in interview collections. By utilizing cutting-edge language models like Llama or GPT, the system will analyse interview transcripts, understanding the context and nuances of the narrative, while adhering to GDPR regulations for handling sensitive data. The project encompasses evaluating text understanding and generation in German, optimizing workflow for chunking interviews and prompting, and establishing a user-friendly pipeline deployed on Google Colab. Key features include seamless integration of state-of-the-art LLMs, efficient handling of large interview documents, and a user interface conducive to historians and researchers. The anticipated outcome is a scalable solution empowering users to effortlessly generate informative biographies, enhancing archival practices and facilitating secondary analysis in humanities research.

Scrum Project Plan:

**Product Backlog:**

1. Research recent papers on prompt-engineering and pre-processing for generative AI in German to understand the current state-of-the-art.
2. Experiment with chunking interviews and prompting techniques to find the most effective workflow for generating high-quality biographies.
3. Develop a GDPR-compliant and user-friendly pipeline in Google Colab for historians and researchers.
4. Integrate Large Language models (LLMs) like Llama or GPT into the pipeline for accurate biographies.
5. Test the pipeline thoroughly and iterate on its performance based on user feedback to ensure accuracy and usability.

**Phase 1: Project Initialization**

Sprint 1: Preparing the Basics and Foundation

User Stories:

1. As a Data Scientist, I want to research recent papers on prompt-engineering and pre-processing for generative AI in German to understand the current state-of-the-art.

Tasks:

1. Project Understanding and Kick-off Meeting.
2. Understanding the Problem Statement and identifying the knowledge gaps.
3. Find and study all the recent relevant papers.

**Phase 2: Development and Improvement**

Sprint 2: Prompt Engineering

User Stories:

1. As a data scientist, I want to experiment with chunking interviews and prompting techniques to find the most effective workflow for generating high-quality biographies.

Tasks:

1. Research and select chunking techniques: Investigate various methods for segmenting interview transcripts, such as sentence-level chunking, paragraph-level chunking, and keyword-based chunking.
2. Explore prompting strategies: Review existing prompting techniques used in natural language processing and identify those suitable for generating biographies from interview transcripts.
3. Develop implementation scripts: Create scripts or algorithms to implement the chosen chunking and prompting techniques, ensuring they can be applied effectively to interview data.

Sprint 3: Improvement of Model

User Stories:

* + - 1. As a data scientist, I want to develop a GDPR-compliant and user-friendly pipeline in Google Colab for historians and researchers.

Tasks:

1. Research GDPR compliance requirements to ensure adherence to regulations for handling sensitive interview data.
2. Design the pipeline architecture, incorporating GDPR compliance measures seamlessly into the workflow.
3. Implement data handling procedures such as encryption and access controls to protect interviewees' privacy.
4. Create user-friendly interface elements within Google Colab to enhance usability for historians and researchers.
5. Test pipeline functionality and GDPR compliance thoroughly to verify performance and data protection measures.
6. Document pipeline usage guidelines and provide ongoing support to users, ensuring understanding and compliance with GDPR regulations.

Sprint 4: Optimization

User Stories:

1. As a Data Scientist, I want to integrate Large Language models (LLMs) like Llama or GPT into the pipeline for accurate biographies.

Tasks:

1. Integrating large language models (LLMs) like Llama or GPT to enhance biography accuracy.
2. Select the most suitable LLM based on performance and compatibility with German language.
3. Develop a framework for integrating the chosen LLM into the pipeline architecture seamlessly.
4. Implement LLM integration, ensuring smooth operation and compatibility with existing components.
5. Optimize LLM parameters to improve biography generation accuracy and coherence.
6. Validate biographies generated using the integrated LLM against known information to assess accuracy.

**Phase 3: Deployment and Reporting**

Sprint 5: Reporting

User Stories:

As a Data Scientist, I want to test the pipeline thoroughly and iterate on its performance based on user feedback to ensure accuracy and usability.

Tasks:

1. Set up and deploy the biography extraction system on a server, ensuring it is configured for stable operation.
2. Conduct detailed testing of the system on interview transcripts, compare the results with earlier evaluations, and prepare comprehensive performance reports.
3. Document the entire development process and note potential enhancements based on feedback and potential new technology.

Important Dates

* Kick-off Meeting: 16-04-2024.
* Sprint Planning Meeting: 18-04-2024.
* Daily: Daily standup within the team.
* Bi-Weekly Meeting with Technical Advisors: Every Tuesday till project completion.
* Bi-weekly with internal stakeholders/Sprint Review Meeting.
* Sprint Retrospective: The end date of every sprint given in the Project Plan.

Time Plan

Please find below the time plan for the project:



Meeting Notes:

|  |  |  |
| --- | --- | --- |
| Meeting 1: Project Kick-Off Meeting with Stakeholders | | |
| Date: 16/04/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Dennis Möbus  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | * Introduction of Project. * Introduction of Stakeholders, technical advisors, and team. | |
| Discussions: | * Briefly discussed the project details and problem statement. * Got to know the background of the given project. * Briefly discussed the expected result. * Got to know the Stakeholders and Technical Advisors. * Fixed the upcoming meeting schedules. | |
| Prepared by: | Mr. Lakshith Reddy Baddula | |

|  |  |  |
| --- | --- | --- |
| Meeting 2: Meeting with technical advisor | | |
| Date: 30/04/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Dennis Möbus  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | * Overview of potential models. * Understanding the procedure. | |
| Discussions: | * Models to be used and ways to initiate the project. * Basic methods to be implemented. * Understanding how to work on it for the upcoming sprint. | |
| Prepared by: | Mr. Ashwith Anand Poojary | |

|  |  |  |
| --- | --- | --- |
| Meeting 3: Project Plan Review | | |
| Date: 13/05/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Dennis Möbus  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Provided models to look at and review | |
| Discussions: | Presented their view of the models and opinions.  Discussions took place about pros and cons the models.  Further discussions on Models took place  Initial code difficulty in extracting Transcript data  Suggestions on using techniques like ‘/t’ when reading CSV file to help with the process | |
| Prepared by: | Mr. Rakesh Hadne Sreenath | |

|  |  |  |
| --- | --- | --- |
| Meeting 4: Meeting with Internal stakeholder | | |
| Date: 28/05/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Dennis Möbus  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Initial Implementation and Review of Mistral Model and Llama 2 70B Model and new prompting techniques. | |
| Discussions: | Providing output and suggestions of improvement of output  Further insights in Prompting techniques  Implementation of Mistral 8b Model showed insufficient results to move further with this model  Few-shot prompting performed  Llama 2 8B model implementation result is insufficient to move further with this model since incomplete details of the Interview is given in the output | |
| Prepared by: | Ms. Aadishree Borkar | |

|  |  |  |
| --- | --- | --- |
| Meeting 5: Meeting with Technical advisor | | |
| Date: 11/06/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Dennis Möbus  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Implementation and Review of Mistral 8\*22B, Llama 3 8B and 70B Model and new prompting techniques. | |
| Discussions: | Chain-of-Thought prompting performed  Mistral 8\*22B model implementation result is insufficient to move further with this model since incomplete details of the Interview is given in the output  Different prompting methods further attempted in questionnaire format  Implementation of Llama 3 8B providing output with multiple biographies of different names as the interviewee  Implementation of Llama 3 70B yields similar results but seems to be better  We suggested summarizing the transcript data and then passing through the model to help prevent long outputs generation issue | |
| Prepared by: | Mr. Lakshith Reddy Baddula | |

|  |  |  |
| --- | --- | --- |
| Meeting 6: Meeting with Technical Advisor | | |
| Date: 26/06/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Kamelia Reshadi | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Combination of Different Models in the Code | |
| Discussions: | Here we tried to implement passing the Transcript Data through a model and then pass the output of that model through another model to solve the issue of long outputs, so we first passed the transcript data through Llama 3 70B and then the output of that to Mistral 8\*22B.  This helped refine the output to under 700 words or less | |
| Prepared by: | Ms. Muskaan Chhabra | |

|  |  |  |
| --- | --- | --- |
| Meeting 7: Meeting with Technical Advisor | | |
| Date: 10/07/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Kamelia Reshadi | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | New Model release and its implementation – Llama 3.1 | |
| Discussions: | Meta released a new model named Llama 3.1 and we decided to implement this cutting-edge technology and used Llama 3.1 405B through Together AI API key.  This yielded a much better result, and we will be moving further with this model | |
| Prepared by: | Ms. Aadishree Borkar | |

|  |  |  |
| --- | --- | --- |
| Meeting 8: Meeting with Internal stakeholders | | |
| Date: 23/07/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Implementation of Llama 3.1 70b  Start of Report | |
| Discussions: | Due to limitation in token size of Llama 3.1 405B we decided to use Llama 3.1 70B as its context window is 128k tokens.  This helps due to increasing the chunk size immensely which yielded very good and reasonable output so far. | |
| Prepared by: | Mr. Ashwith Anand Poojary | |

|  |  |  |
| --- | --- | --- |
| Meeting 9: Meeting with Technical Advisor | | |
| Date: 30/07/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Kamelia Reshadi | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Prompt Improvement and Running code locally. | |
| Discussions: | We are receiving a good result in Llama 3.1 70b, but it still leaves out the later stages of the life story of the interviewee, here we tried various types of prompting to help improve the output.  We only started implementing a code to run this model locally in our servers. | |
| Prepared by: | Mr. Lakshith Reddy Baddula | |

|  |  |  |
| --- | --- | --- |
| Meeting 10: Meeting with Internal stakeholders | | |
| Date: 13/08/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Prof. Dr. Binh Vu  Dr. Kamelia Reshadi | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | External Stakeholder review and Improvements | |
| Discussions: | Our external stakeholder communicated on our findings and results and suggested improvements.  Mentioned prompting in questionnaire format in more detailed aspect of various stages of the interviewee’s life. This has helped improve our output further. | |
| Prepared by: | Mr. Rakesh Hadne Sreenath | |

|  |  |  |
| --- | --- | --- |
| Meeting 11: Meeting with Technical Adviser | | |
| Date: 10/09/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Prof. Dr. Binh Vu  Dr. Kamelia Reshadi | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | Final update and confirmation from External Stakeholder | |
| Discussions: | We received appreciation from external stakeholder happy with the output we received.  Our internal stakeholder gave initial go ahead with start of report and changes to make. | |
| Prepared by: | Ms. Muskaan Chhabra | |

|  |  |  |
| --- | --- | --- |
| Meeting 12: Internal stakeholder meeting | | |
| Date: 24/09/2024 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Prof. Dr. Binh Vu | Attendees:  Mr. Ashwith Anand Poojary  Ms. Muskaan Chhabra  Mr. Lakshith Reddy Baddula  Ms. Aadishree Borkar  Mr. Rakesh Hadne Sreenath |
| Agenda: | End of Report and Final Code Completion | |
| Discussions: | We have received final confirmation from our Internal Stakeholder for the completion of the report  Uploaded all final files on GitHub for review and submission of files as well on Moodle | |
| Prepared by: | Ms. Aadishree Borkar | |

Appendix