Optimizing Genomics Research Workflows by Leveraging Curated Domain Expertise and RAG-enhanced Language Models—All Wrapped in a Delightful User Experience

Visalakshi Iyer
College of Information Science,
The University of Arizona,
visalakshiiyer@arizona.edu

Pranshu Singh Rawat
College of Information Science,
The University of Arizona,
pranshurawat@arizona.edu

Syed Aslam Sheik Dawood
College of Information Science,
The University of Arizona,
syedaslam@arizona.edu

Vamsi Vadala
College of Information Science,
The University of Arizona,
vamsiv@arizona.edu

Jose F Oviedo
College of Information Science,
The University of Arizona,
jfo@arizona.edu

Project Mentors: Dr. Egoitz Laparra, Dr. Bernardo Lemos

Faculty Mentor: Dr. Greg Chism

Table of Contents

List of Figures

List of Figures

List of Tables

List of Tables

Revision History

Version	Changes	Date
V1	Initial template created	03/31/2025

Table 1: Revision History

1 Executive Summary

This service is a domain-specific Retrieval-Augmented Generation (RAG) model tailored to assist users with bioinformatics queries that are specifically related to single-cell genomics research. We will utilize a tailored stack of NLP technologies from both closed and open sources as we develop robust systems that will enable a context-aware virtual assistant – supporting researchers through tasks such as semantic document searches, metadata queries, fine-grain document analyses. Ultimately, we aim to deploy user-friendly chatbot web app to reliably assist researchers with routine, low-stakes tasks in order to streamline complex workflows and free up researchers' valuable time.

The current lab activities go through excessive redundant processes of document analysis, with problems often being hours spent in going back-and-forth multiple documents for researching information from them, navigating complex data or scrolling through certain documentation for relevant information. These routine tasks can be optimized with this new system that enables a private-to-lab session equipped with state-of-the-art RAG model that streamlines these processes in an optimized time-contstrained manner.

We will be building a Web Interface that serves a ChatBot UI embedded with File-Upload button and a Text-Input widget, that will show an option to upload Contextual-Documents, input a User-Query and interact with the RAG model in the conversation space. The process involves data processing (vectorization of documents) and model inference calls in the backend and a UI development process that connects these processes to the end user, to show the output generated from the model.

2 User/Market Research

Our primary users are PhD-level genomic researchers and scientists. These users are highly knowledgeable in their fields and require highly specialized, accurate, and accessible tools to support their research activities.

The bioinformatics space within single-cell genomics research is rapidly evolving due to its growing impact on biomedical and clinical applications. Researchers in this domain navigate complex research pipelines composed of many tasks. While many of these tasks are high-stakes and critical to the research pipeline, others are search-based, low-stakes, and routine—these are the types of tasks we strive to optimize and reduce time-to-completion for.

Currently, many researchers rely on manual, single-document search methods to extract relevant information from extensive documentation and scien-

tific literature. The information sought may involve particular methods described in a single document or a consensus of methods across a group of similar documents. These approaches are time-consuming and often inefficient—especially when working with large volumes of documents.

Empathy interviews revealed several key pain points: time lost to manually searching for documents as well as searching within the documents themselves; difficulty analyzing relevant information from multiple sources in a collection of documents; and the lack of fine-grained search tools that integrate naturally into existing workflows.

In response, our services focus on advanced semantic search, granular document retrieval, and a user-friendly chatbot interface designed to streamline certain types of research tasks and enhance user productivity.

3 Product Features

This project aims to deliver a system that utilizes a variable set of data pipelines that would allow users (researchers and scientists) to help with their queries related to bio-informatics research. The feature list for this system will include the features mentioned here in the following sections.

3.1 Context Aware Retrieval-Augmented Generation (RAG) Assistant

Description: Integrate RAG model for bioinformatics queries related to single-cell genomics research. This model will utilize curated domain expertise retrieved from user-provided contextual documents for providing optimal, useful and most relevant output.

Parameters:

- User Query: A descriptive query mentioning the requirements of output that can help the model efficiently gather information and curate a useful output.
- Contextual Documents: A set of (optional) documents that can help the model provide better outputs by parsing for relevant information from the documents, that are not already on the database used by the model.

3.2 Database Reserve for retrieving domain-specific knowledge

Description: Develop a tailored vector database of a collection of bio-informatics documents that can provide the initial context for each conversation/session for the model.

Parameters:

- Documents: PDFs or links to research papers relevant to the lab's research initiatives
- Vector Database: A database that stores the contents of documents in a vectorized format, useful for sematic search activities.

3.3 Task Specific assistance

Description: The model can provide task specific assistance for redundant processes in the current research pipeline such as data-analysis, document analysis summaries, and literature reviews.

Parameters:

- User Query: A descriptive query mentioning the task and the required output that can help the model efficiently gather information and curate a useful output.
- Contextual Documents: A set of (optional) documents that can help the model provide better outputs by parsing for relevant information from the documents, that are not already on the database used by the model.

3.4 Citations and Sources of the output

Description: Each output will be cited with relevant sources, from where the model is pulling information. RAG uses vectorized datastore to match relevant documents according to the user-query

Parameters:

- Vector Database: A database that stores the contents of documents in a vectorized format, useful for sematic search activities.

4 Project Timeline and Gantt Chart

The project followed a structured multi-phase approach:

1. **Planning & Requirements (Weeks 1-2)**: This phase involves two key tasks.
    - The Kickoff Meeting & Requirement Gathering focuses on aligning stakeholders, identifying project goals, and gathering initial requirements.
    - The Technical Research & Documentation task ensures a thorough exploration of relevant technologies, standards, and methodologies, culminating in detailed documentation to guide subsequent phases.
2. **Design & Architecture (Weeks 3-4)**: This phase involves two key tasks.
    - The Model & RAG System Architecture Design involves creating the foundational blueprint for the system, including its components and interactions.
    - The UI/UX Design & Wireframing task focuses on designing user-friendly interfaces and developing wireframes that outline the visual and functional aspects of the application.
3. **Implementation (Weeks 5-10)**: This phase involves four key tasks.
    - Environment Setup & Model Fine-Tuning involves configuring the necessary development environment and optimizing machine learning models for specific use cases.
    - RAG System & API Integration integrates the Retrieval-Augmented Generation (RAG) system with APIs to enable seamless data flow.
    - Web-App Chatbot Development entails building a chatbot interface for user interaction.
    - Finally, Integration Testing ensures that all components work together as intended by identifying and resolving any issues.
4. **Evaluation & Iteration (Weeks 11-14)**: This phase involves three key tasks.
    - SME Evaluation & Human RL Tuning engages subject matter experts to evaluate the system and refine it using reinforcement learning techniques.
    - Performance & Usability Testing assesses the system's efficiency and user experience to identify areas for improvement.

- Iteration & Refinement involves making iterative updates based on feedback to enhance functionality and reliability.

5. **Deployment & Documentation (Weeks 15-16)**: This phase involves two key tasks.

   - Final Deployment on HPC ensures the system is deployed on high-performance computing infrastructure for scalability and robustness.
   - Documentation, Training, & Project Wrap-Up involves creating comprehensive documentation, training users or stakeholders, and concluding the project with a formal wrap-up session.

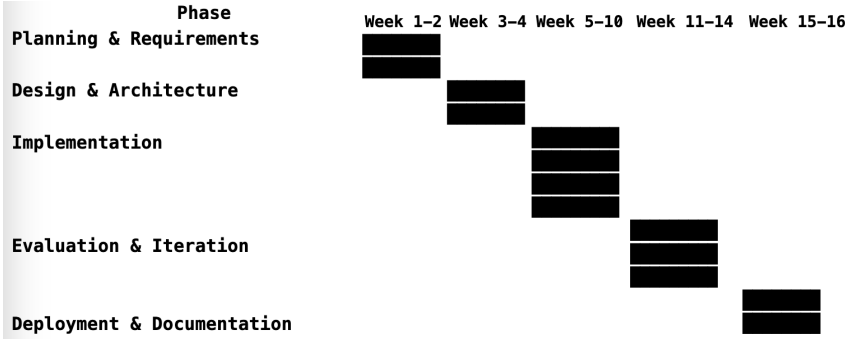

Figure 1: Gantt Chart

Milestone	Date
Team Formation	1/29/2025
Signed proposal	2/19/2025
Planning & Requirements Completion	02/15/2025
Design & Architecture Completion	03/01/2025
Implementation Completion	04/12/2025
Evaluation & Iteration Completion	04/20/2025
Deployment & Documentation Completion	05/28/2025
Poster Demo	4/23/2025
iShowcase	5/1/2025

Table 3: Milestone Schedule

5 Ethics Considerations

Data ethics is a key consideration in any product we create. Please include a completed ethics chart. Please see examples on D2L.

#	**Question**	**Generally**	**Data Breach**
1	Could a user sell drugs or other illegal items on your platform?	Y/Ⓝ/M	Y/Ⓝ/M
2	Could a user of your platform engage in sex trafficking?	Y/Ⓝ/M	Y/Ⓝ/M
3	Could a user sell class notes or cheat on their homework on your platform?	Y/Ⓝ/M	Y/Ⓝ/M
4	Could a stalker use your project to find someone?	Y/Ⓝ/M	Y/Ⓝ/M
5	Could your app be used to spy on or track individuals?	Y/Ⓝ/M	Y/Ⓝ/M
6	Could your app/software access the camera or microphone and record things without users being aware?	Y/Ⓝ/M	Y/Ⓝ/M
7	If someone uses your platform, could they be re-traumatized or have their mental health impacted in some way?	Y/Ⓝ/M	Y/Ⓝ/M
8	Could your algorithm promote material that would traumatize or upset individuals?	Y/Ⓝ/M	Y/Ⓝ/M
9	Would your users be upset if the data you collect was given to someone else?	Ⓨ/N/M	Ⓨ/N/M
10	Could a data leak potentially lead to identity theft?	Y/Ⓝ/M	Y/Ⓝ/M
11	If your site was hacked, would users of that product potentially lose their job, spouse, or family?	Y/Ⓝ/M	Y/Ⓝ/M
12	Should there be an age limitation on your product?	Y/Ⓝ/M	Y/Ⓝ/M
13	Could someone use your product to find, contact, and potentially commit elder abuse?	Y/ⓃM	Y/Ⓝ/M

14	If the data on your platform was breached, could it be used to blackmail the users?	Y/Ⓝ/M	Y/Ⓝ/M
15	Does the existence of your project imply that a particular racial group, gender, religion, or other protected category is inherently bad, gross, or unwanted?	Y/Ⓝ/M	Y/Ⓝ/M
16	Could your product be used to commit hate crimes against a specific group?	Y/Ⓝ/M	Y/Ⓝ/M
17	Does the primary content of your game or algorithm focus on something considered deeply unethical?	Y/Ⓝ/M	Y/Ⓝ/M
18	Does your game or software contain race, gender, or other stereotypes?	Y/Ⓝ/M	Y/Ⓝ/M
19	Could users of your app scam other individuals?	Y/Ⓝ/M	Y/Ⓝ/M
20	Is your particular algorithm biased towards predicting correctly only for one race, gender, or other group?	Y/Ⓝ/M	Y/Ⓝ/M
21	Are the users of your project, players of your game, or those being surveyed for your data aware of how their data will be used?	Ⓨ/N/M	Ⓨ/N/M
22	What are the possible misinterpretations of your results? For example – would a white supremacist or misogynist be stoked about your results if they misinterpreted it?	Y/Ⓝ/M	Y/Ⓝ/M
23	Does the use or purchase of your data potentially contribute to a dangerous group or regime?	Y/Ⓝ/M	Y/Ⓝ/M
24	Could your virtual reality environment cause injury to the user?	Y/Ⓝ/M	Y/Ⓝ/M
25	Are your study participants or game players aware that their data will be collected and used?	Ⓨ/N/M	Ⓨ/N/M

26	Does your game or app contain addictive design elements without benefit to the user?	Y/Ⓝ/M	Y/Ⓝ/M
27	Does your survey contain an aspect of compulsion or unusually large incentive, that would command users to take it even if it was to their detriment?	Y/Ⓝ/M	Y/Ⓝ/M
28	Could your research outcomes harm an individual or entity?	Y/Ⓝ/M	Y/Ⓝ/M

Table 4: Table of Ethical Compliance

To address the ethical concerns for which we answered "Yes" or "Maybe" in the list of ethical questions, we propose the following solutions:

Q9: Would users be upset if their data was shared with someone else?

We plan to implement a **Transparent Data Policy & User Control** to ensure clarity and user autonomy. Our data usage policy will explicitly outline what data is collected, how it is stored, and whether it is shared. Additionally, users will have the option to remain anonymous or log in to save their interaction history based on their preference.

Q21: Are the users of your project, players of your game, or those being surveyed aware of how their data will be used?

We will ensure **Informed Consent & Regular Reminders** by presenting a consent message at the start of the application, clearly explaining how data will be used. Users will have the choice to agree or disagree before proceeding. Additionally, a "Privacy Info" button will be incorporated into the chatbot UI, allowing users to access data usage details at any time.

Q25: Are your study participants or game players aware that their data will be collected and used?

This concern will be addressed through the same measures outlined above, including the **Transparent Data Policy**, **User Control**, **Informed Consent**, and **Regular Reminders**. These mechanisms will ensure that users are fully aware of data collection practices and retain control over their personal information.

6 Approvals

The signatures of the people below indicate an understanding of the purpose and content of this document by those signing it. By signing this document, you indicate that you approve of the proposed project outlined in this Statement of Work, the division of work, the Ground Rules and that the next steps may be taken to create a Product Specification and proceed with the project.

This document is based upon and supersedes the ¡PRD title¿ Version X.X. Deviations, (versus clarifications), from the PDR have been clearly noted. For any requirements not listed in this SOW, the PRD requirements shall remain in effect.

Approver Name	**Title**	**Signature (Initials)**	**Date**
Visalakshi Iyer	Team Member	VI	2025-04-01
Pranshu Singh Rawat	Team Member	PS	2025-04-01
Vamsi Vadala	Team Member	VV	2025-04-01
Syed Aslam Sheik Dawood	Team Member	SASD	2025-04-01
Jose F Oviedo	Program Manager	JFO	2025-04-01
Dr. Egoitz Lappara	Advisor		
Dr. Bernardo Lemos	Advisor		
Dr. Greg Chism	Instructor		

Table 5: Milestone Schedule

Section	**Author**	**Word Count**
Executive Summary	Aslam and Vamsi	239
User/Market Research	Jose	228
Product Features	Visalakshi	362
Project Timeline and Gantt Chart	Jose, Vamsi, Aslam	330
Ethics Consideration	Pranshu	606

Table 6: Author Contribution

7 Appendix

7.1 Advisor Engagement

7.1.1 Project Team Responsibilities

- The Project Manager will set up and facilitate a weekly call/meeting with the Faculty Advisor. The Project Team will provide weekly status updates to the Faculty Advisor including upcoming deliverables, critical issues, and any adjustments to the Project Plan.
- Documents will be provided to the Faculty Advisor with adequate time for review and signature. The time necessary for review will be agreed with the Advisor. The minimum review time will be 3 days prior to the document due date.
- Design files will be provided to the Faculty Advisor as requested in a format agreed to with the Advisor.
- Support requirements will be clearly requested from the Faculty Advisor with the dates required and an adequate time for fulfilling the request.
- Modifications requests to the Project Plan by Faculty Advisor will be reviewed and agreed to within 1 week of the request.

7.1.2 Faculty Advisor Responsibilities

- The Faculty Advisor will provide knowledge and expertise to help the group stretch their skills.
- The Faculty Advisor will participate in a weekly or bi-weekly call/meeting with the Project Team to review the project status, upcoming deliverables, priorities, issues, and progress to the agreed Project Plan.
- The Faculty Advisor will provide document review, feedback and approval, rejection, approval with contingencies with adequate time for the Project Team to meet the course due dates.
- The Faculty Advisor will provide feedback to requested support requirements from the Project Team. This includes feedback and guidance on design implementations decisions, design files, test plans, test procedures and test results.
- The Faculty Advisor shall provide technical advice and guidance to the Project Team answering inquiries approximately 1 hour per week.

- Modifications to the Project Plan by the Project Team will be resolved and documented within 1 week of the request.
- Grade the finalized project using a skill-based rubric
- Attend iShowcase in May.

7.2 Ground Rules

As a team and as individual team members, we agree to:

1. **Stay focused on our objectives and goals.**
   Each time the team meets, we will clearly define our objectives and desired outcomes at the beginning of the meeting. We will politely remind team members if we are getting off track.
2. **"Sidebar" any issues that are relevant but not consistent with the immediate objectives.**
   Occasionally, important matters are raised that are not relevant to the immediate goals of the meeting. To keep the group on track, but avoid losing the issue, create a "sidebar" where these topics can be listed and discussed later.
3. **Listen when others are speaking.**
   We will listen and consider others' input before adding our own comments.
4. **All viewpoints will have an opportunity to be heard.**
   We understand that some team members may be quieter than others. We will make an effort to get each team member's viewpoint and that no one dominates the discussion.
5. **Differences of opinion will be discussed respectfully**
   We will identify areas of agreement before assessing areas of disagreement. We will encourage each other to look beyond our own point of view. We will discuss different ideas respectfully. As a team, we will weigh the merits of different opinions and agree on a process for choosing a direction. All team members will respect and follow the decision or direction.
6. **Look for the good points in new ideas.**
   We will endeavor to explore the value in each idea as we assess and select our path forward.

7. **Focus on the future, not the past.**
   We will use our past experience to inform our decisions, but focus the discussion on the future objectives. Blame for past performance is counterproductive, we will focus on finding solutions.

8. **Agree upon specific action items and next steps.**
   At the end of each meeting and discussion, we will summarize and agree on specific next steps, action items and assignments.

9. **Accountability**
   As team members, we will each be responsible for our individual assignments and contribution to achieving the team objectives and goals. We will honor our responsibilities and not let our team members down.