Donald Philp

University of Westminster, School of Life Sciences

Phase One: Human-in-the-Loop for Automated Hypothesis Testing in Biobank Research, April 2025

This Research Proposal is submitted in partial fulfilment of the requirement of module PMMES01F, MSc Artificial Intelligence and Digital Health

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Manuel Corpas, for their guidance and support during the development of this research proposal. Additionally, I appreciate the continued technical assistance provided by Dr. Tom Oliver and acknowledge the helpful feedback from my colleagues and peers in both the Life Sciences and Computer Science and Engineering faculties at the University of Westminster.

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
ANN	Approximate Nearest Neighbours
API	Application Programming Interface
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture
DGX	Deep Learning GPU Accelerator (NVIDIA DGX systems)
FAISS	Facebook AI Similarity Search
GB10	Grace Blackwell 10 (Superchip)
GPU	Graphics Processing Unit
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
LCM	Large Contextual Model
LLM	Large Language Model
LTS	Long-Term Support
MAP	Mean Average Precision
NDCG	Normalised Discounted Cumulative Gain
NV-Embed v2	NVIDIA Embedding Version 2
RAG	Retrieval-Augmented Generation
RTX	Ray Tracing Texel eXtreme (NVIDIA graphics cards)
URL	Uniform Resource Locator
VRE	Virtual Research Environment

1. Introduction

The emergence of precision medicine, significantly facilitated by large-scale biobanking initiatives like the UK Biobank, underscores the necessity for advanced data extraction methods capable of precisely identifying phenotype data crucial to targeted biomedical hypotheses. Despite the abundance of biobank resources, efficiently leveraging this vast data for predictive analytics and machine learning applications remains a challenge.

This study hypothesises that: *Integrating PubMedBERT embeddings, contextual chunking, and the Monarch Initiative knowledge graph into a retrieval-augmented generation (RAG) system significantly enhances precision and speed in extracting phenotype data for targeted biomedical hypotheses, improving feature selection and predictive accuracy in UK Biobank machine learning studies.*

This research aims to establish a comprehensive framework for data extraction and filtering in future biobank-based biomedical research by meticulously assessing improvements in retrieval accuracy, validating effectiveness with real-world data, and delineating best practices and limitations.

2. Research Plan

The research plan presents a systematic four-step methodology to ensure scientific precision, facilitate informed decision-making, promote iterative improvements, and enable early identification of suboptimal performance. This framework aligns with the rationale established for the Master of Science (MSc) project.

The initial step holds paramount importance, as it involves the careful and precise identification of the gold standard against which we will assess our hypothesis. This procedure requires selecting between 20 and 30 highly specific and relevant high-impact journal articles that utilise UK Biobank data, particularly focusing on large phenotypic datasets and employing machine learning methodologies. We aim to

provide context for these publications, extract the hypotheses, and determine the phenotypical features used. We assume that these papers represent the gold standard for modelling our outcomes, where we find similarities in phenotype identification based on the hypotheses of the 20 to 30 papers.

The second and critically important step involves curating a substantial journal library corpus from which our knowledge base will be derived. The hypotheses presented in the UK Biobank journal papers have a specific research focus, and the researchers determined the features based on the literature review. We aim to develop an initial knowledge domain as a foundational reference for comparative analysis and the identification of relevant phenotypes. A more extensive corpus increases the statistical likelihood that each hypothesis articulated within the Biobank research papers will be highly relevant, thereby potentially augmenting the accuracy of the phenotype identification modelling workflow. However, as we remain pressed for time, our journal library corpus remains restricted.

We employ a meticulously fine-tuned retrieval-augmented generation (RAG) system to extract relevant information from our knowledge domain. Lewis et al. (2020) were the first to introduce RAG, which has since come to represent AI systems that use parametric models in addition to external knowledge sources. Designed to ensure high retrieval accuracy and alignment with the phenotypic features discussed in UK Biobank studies, RAG aids in identifying relevant hypotheses by integrating external knowledge with statistically grounded generation.

Finally, the data analysis procedure and hypothesis assessment enable us to quickly pinpoint inaccuracies, iteratively refine our RAG system, and verify robust statistical analysis. This maximises our chances of discerning suitable phenotypes that correspond with our gold standard identification. The results of this analysis will determine whether our initial hypothesis is valid, paving the way for further development in automating biobanking repositories for precision medicine. This approach minimises human biases by fostering a greater reliance on data.

3. Research Approach

Approximately 8,500 journal papers have been written regarding the UK Biobank and its data repository, according to UK Biobank (2025a). However, the relevance and impact scores of these papers, along with the data repositories employed, vary significantly. We need to establish a robust method for identifying the most relevant, impactful, and scientifically precise papers to ensure that our control serves as the foundation for tuning workflows within our RAG system. This identification method will dictate our overall direction and decision-making. It will involve acquiring a text corpus of all 8,500 papers and assessing their specificity, impact, and relevance to our research. The selection methods are discussed in the data analysis section below.

The costs of acquiring a larger dataset will, unfortunately, influence the data repository available to us. Our research is confined to open-source data, which presents a significant limitation on our findings. A critical point to consider is that the source and size of the data are directly proportional to the quality of our outcomes. Larger specific datasets statistically enhance relevance, underscoring the importance of our feature identification in relation to the UK Biobank study. Since the authors of the UK Biobank papers have access to a comprehensive array of studies, our work is limited by the restrictions of open-source data. We have already retrieved and validated the total features from the UK Biobank based on their open-source schema, which means the parameters of their schema constrain our feature selection.

With our current open-source data corpus, we need to ensure that the complete data set, consisting of approximately 6 million PubMed medical journal articles, is organised into a database for efficient knowledge retrieval. To achieve this, we will use a text embedding procedure, typically employing machine learning or neural network models to convert textual data into numeric vector representations. This model that performs text embedding, which we will refer to as the 'text embedder,' is crucial as it enables us to efficiently store and retrieve large volumes of text-based data, facilitating our research by allowing quick and accurate data retrieval.

The selection of the text embedder is crucial, as it directly influences the specific identification of data chunks into which each paper will be segmented. We propose to employ a text embedder specifically trained on PubMed journal papers, called PubMedBERT. According to Vithanage et al. (2024), PubMedBERT remains the most accurate with PubMed text knowledge. For comparison, we will also assess the performance of the most advanced generalised text model currently available, Nvidia NV-Embed v2. After sampling and embedding a random selection of 2,000 PubMed journal articles, we will evaluate the output and referencing based on hypotheses derived from UK Biobank papers.

A crucial aspect of our research focuses on anticipating potential challenges and proactively addressing them to ensure the robustness and scalability of the automated hypothesis testing system. The key areas identified include data limitations, model performance, and quality concerns related to embedding. Maintaining data integrity and quality within our extensive library corpus will be particularly challenging. While we have begun addressing the structure and data quality, the ongoing issue of differentiating quality journal papers from weaker ones will remain a significant hurdle.

The retrieval mechanism for extracting the most relevant data chunks from the vector database can be approached in several ways, such as using FAISS with approximate nearest neighbours (ANN), cosine similarity, or other knowledge tree mechanisms. We will use Milvus or Weaviate, as they offer the most flexibility and modularity according to Pan, et. al (2023). Scoring the retrieval will be essential to establish the relevance of the retrieved information. To guarantee high specificity in retrieval, we will implement an agentic procedure through which the embedder will iteratively seek the highest ANN or cosine similarity scores. Additionally, we will enhance accuracy by employing Anthropic's latest specific chunking methods, comparing these with standard approaches such as LangChain's *RecursiveCharacterTextSplitter* and LamaIndex's *SentenceSplitter*. LangChain’s libraries are regarded as essential standards for literature review, as demonstrated by Jeong et al. (2024).

We will iteratively incorporate the 12,000 biobank features (Biobank, 2025b) and their associated notes for background into the database. Based on the retrieval of vectors from our journal database, we will specifically extract the most relevant biobank features from these vector chunks. To further enhance specificity, we will implement rule-based logic derived from the Monarch Initiative, an open-source biomedical knowledge graph explicitly used in genetic, phenotype, disease, and related research fields. The Monarch Initiative provides a comprehensive and structured framework for understanding the relationships between genetic variants, phenotypes, and diseases, augmenting the specificity of our research by offering a reliable and thorough set of rules for feature extraction. (Shefchek et al., 2020)

We will identify the UK Biobank papers as the control. Our validation will involve establishing an overlapping phenotype identification process between the system's selected phenotypes and those used by researchers in their papers. A significant percentage of congruence must be demonstrated to support our hypothesis, as this effective congruence in the identification procedure is crucial for data scientists working with the UK Biobank.

Please find the Gantt Chart with details in Appendices A and B.

4. Data Analysis

The large corpus and highly integrated workflow of fine-tuning parameters and verifying scientific precision makes the data analysis phase of the project the most exhaustive part. The control step in the analysis is the first step, and we need to establish the control from the entire UK Biobank paper. A standard vanilla RAG could be used with the PubMedBERT embedded, along with Anthropic's contextual chunking (Anthropic, 2024), to create a small mini-RAG to prompt. This would possibly be the most effective way to pull individual papers based on their specificity within the prompt to be pulled as a "cite". We would use cosine similarity and ANN to pull the chunks and provide context with a large language model via Application Programming Interface (API). This should be a simple task, as the retrieval

mechanism will not be too exhaustive with 8,500 papers. Afterwards, we could employ bibliometric criteria filters that assess citation count, journal impact factor, and publication recency to ensure greater relevance to our project scope. All the papers retrieved by the mini-RAG must be manually verified for viability.

We could invite three experts to rank the papers and apply Spearman's correlation to compare their rankings with the results from the objective bibliometric filter. Once we have identified our 20 to 30 papers, we will use data partitioning to divide them into training, testing, and production sets.

To validate the RAG literacy and accuracy, we need to prompt the Biobank journal hypotheses from the gold standard papers. Ensuring that the proportion of relevant papers is appropriate, we will measure precision and recall. We will also employ mean average precision (MAP) and Normalised Discounted Cumulative Gain (NDCG) to assess retrieval quality and the relevance of rankings. By calculating a cosine similarity score, we can rank the retrieved chunks and filter out those below a certain threshold. While this method is exhaustive, we can also engage an expert to rank papers based on a curated list of hypotheses derived from the previous step. This will enable us to apply a t-test or Wilcoxon test to compare the identifications made by the expert with those from the retrieval, once we have established the most effective parameters for our final model. We will use some of the training set papers during this phase.

We must also correlate the literature with the biobank features by constructing embeddings from the UK Biobank Features and their accompanying notes. These will be compared against the paper embeddings retrieved by the RAG using vector similarity, as we previously did with cosine similarity and approximate nearest neighbours (ANN). We propose a scoring mechanism for each retrieved paper-feature pair and establish a robust threshold. The objective is to create a retrieval loop system, an agentic system, that continuously adjusts the vector points around the prompting vector until the operation is thoroughly exhausted. Subsequently, we

will validate this approach using a confusion matrix to assess how accurately the RAG correlates the prompts with the gold standard papers.

Additionally, we can examine semantic vector depth alongside the number of citations to ensure specificity and richness. We might employ Cohen's kappa by allowing two experts to assess the relevance of the papers to the identified phenotypes of the gold standard papers, thereby establishing thresholds for both low and high kappa values. During this phase, we will use the last partitions of the training set papers.

Once we have established all the necessary criteria and the model continues to perform well under stringent statistical constraints and scientific precision, we will utilise the testing papers to assess hypothesis correlation and feature congruence. We will set a threshold and apply the Jaccard similarity index (intersection over union) and the overlap coefficient to quantify the degree of congruence between the features identified by the RAG and those chosen by experts. A specific congruence score will be established to define a significant correlation. Additionally, we will conduct statistical validation using either a permutation test or a Chi-square test to determine if the observed congruence significantly exceeds what would be expected by chance.

It is, however, essential to know the limiting factor within our data analysis. Potential limitations and output performance issues may arise from our analysis. The limited size of our library corpus could lead to biases or gaps in our retrieval system. Performance inconsistencies might occur due to variations in embedding accuracy, especially when addressing complex biomedical terminology or diverse phenotype descriptions. The selection of embedding models and the chunking strategy are crucial decision points that we need to address with clarity and scientific rigour.

Refer to Appendix C below for Data Analysis Tools and Software.

5. Health and Safety

This research poses minimal health and safety risks as it is computational, utilising open-source biomedical data. Consequently, COSHH forms are not required.

6. Ethics

Despite using open-source data, ethical responsibility remains essential due to developing tools for future biomedical research. An ethics application has not been submitted, as the research involves only publicly accessible data.

7. Human Tissue Act

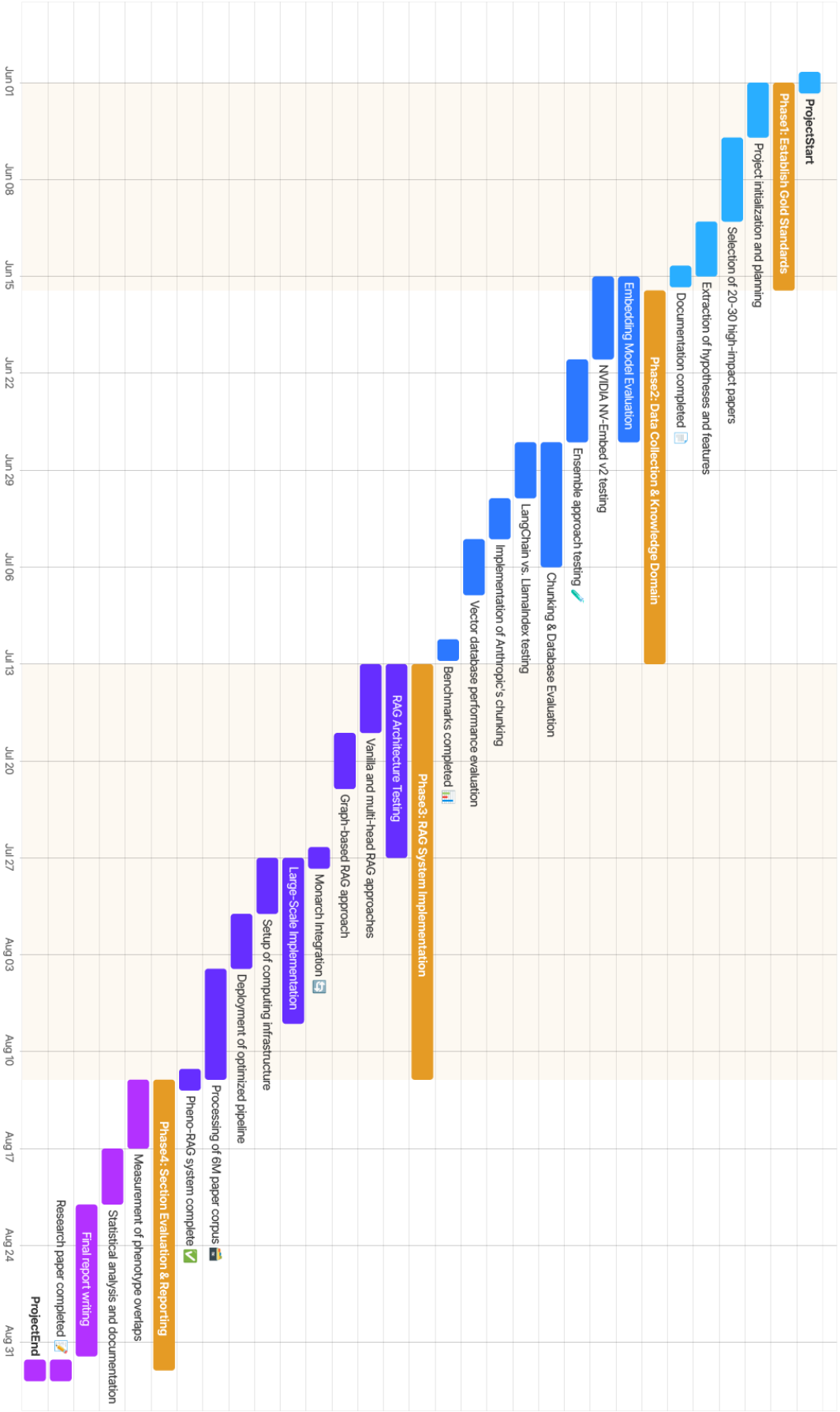
This research does not involve handling human tissue or relevant materials defined by the Human Tissue Act (2004).

References

1. Anthropic, 2024. Introducing Contextual Retrieval. [online] Available at: https://www.anthropic.com/news/contextual-retrieval [Accessed 30 Apr. 2025].
2. Jeong, J., Gil, D., Kim, D. and Jeong, J., 2024. Current research and future directions for off-site construction through LangChain with a large language model. Buildings, 14(8), p.2374. https://doi.org/10.3390/buildings14082374 [Accessed 2 May 2025].
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. [online] arXiv. Available at: https://arxiv.org/abs/2005.11401 [Accessed 9 May 2025].
4. Pan, J.J., Wang, J. and Li, G., 2023. Survey of vector database management systems. [online] Available at: https://arxiv.org/abs/2310.14021 [Accessed 2 May 2025].
5. Shefchek, K.A., Harris, N.L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X.A., Balhoff, J.P., Babb, L., Bello, S.M., Blau, H., Bradford, Y., Carbon, S., Carmody, L., Chan, L.E., Cipriani, V., Cuzick, A., Della Rocca, M., Dunn, N., Essaid, S., Fey, P., Grove, C., Gourdine, J.-P., Hamosh, A., Harris, M., Helbig, I., Hoatlin, M., Joachimiak, M., Jupp, S., Lett, K.B., Lewis, S.E., McNamara, C., Pendlington, Z.M., Pilgrim, C., Putman, T., Ravanmehr, V., Reese, J., Riggs, E., Robb, S., Roncaglia, P., Seager, J., Segerdell, E., Similuk, M., Storm, A.L., Thaxon, C., Thessen, A., Jacobsen, J.O.B., McMurry, J.A., Groza, T., Köhler, S., Smedley, D., Robinson, P.N., Mungall, C.J., Haendel, M.A., Muñoz-Torres, M.C. and Osumi-Sutherland, D., 2020. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Research, 48(D1), pp.D704–D715. https://doi.org/10.1093/nar/gkz997. [Accessed 2 May 2025].
6. UK Biobank, 2025a. Value type coding. [online] Available at: https://biobank.ctsu.ox.ac.uk/ukb/help.cgi?cd=value_type [Accessed 30 Apr. 2025].
7. UK Biobank, 2025b. Publications Catalogue. [online] Available at: https://biobank.ndph.ox.ac.uk/ukb/docs.cgi?id=2 [Accessed 30 Apr. 2025].

8. Vithanage, D., Yu, P., Wang, L. and Deng, C., 2024. Contextual word embedding for biomedical knowledge extraction: a rapid review and case study. Journal of Healthcare Informatics Research, 8, pp.158–179. Available at: https://doi.org/10.1007/s41666-023-00157-y [Accessed 30 Apr. 2025].

Appendix A: Gantt Chart Image



Appendix B: Gantt Chart Details

Step 1: Establishing Gold Standards (14 days: June 1st - June 14)

Activities:

Sample Extraction: 8,500 Biobank Journal papers

- Project initialisation and planning (4 days)
- Selection of 20-30 high-impact UK Biobank papers using bibliometric criteria (6 days)
- Extraction of hypotheses, contextual chunks and phenotypical features from selected papers (4 days)

Deliverables:

- Documented collection of gold standard papers with extracted phenotypical features.
- Notes and Conclusion written up, report ready

Step 2: Data Collection & Knowledge Domain Creation (21 days: June 15 - July 12)

Retrieve and Embed PubMed Papers & Phenotypes

Activities:

- Embedding Model Evaluation (12 days)
    - Parallel testing of NVIDIA NV-Embed v2 and PubMedBERT (6 days)
    - Ensemble approach testing and comparative analysis (6 days)
- Chunking & Database Evaluation (9 days)
    - LangChain vs. LlamaIndex framework testing (4 days)
    - Implementation of Anthropic's contextual chunking methods (3 days)
    - Milvus vs. Weaviate vector database performance evaluation (4 days)

Deliverables:

- Benchmarked performance metrics and documented optimal configuration for biomedical literature processing.
- Notes and Conclusion written up, report ready

Step 3: RAG System Implementation (21 days: July 13 - August 3)

Activities:

Sample Testing: 2,000 sample JSON PubMed Papers.

- **RAG Architecture Testing (14 days)**
    - Implementation of vanilla and multi-head RAG approaches based on Step 2 Findings (5 days)
    - Implementation of graph-based RAG approach on Step 2 Findings (4 days)
    - Integration of Monarch Initiative knowledge graph (5 days)

Sample Testing: 6 million sample JSON PubMed Papers.

- **Large-Scale Implementation (12 days)**
    - Setup of 400B parameter computing infrastructure (4 days)
    - Deployment of optimised pipeline (4 days)
    - Processing of complete 6-million paper corpus (8 days)

Deliverables:

- Fully functional Pheno-RAG system tested in Step 2 and Step 3 with optimal architecture, deployed on high-performance infrastructure with processing of the 6-million paper corpus.
- Notes and Conclusion written up, report ready

Step 4: Evaluation & Report Writing (16 days: August 12 - August 31)

Activities:

Evaluating system performance against the gold standard phenotype documentation against RAG retrieval.

- Measurement of phenotype selection overlaps with gold standard papers (5 days)
- Statistical analysis and results documentation (4 days)
- Final report writing and research paper drafting (11 days)

Deliverables:

- Comprehensive evaluation report with statistical analysis, performance metrics, and a complete research paper ready for submission.

Appendix C: Data Analysis Tools and Software

Below is a list of software, packages, and tools that will be employed throughout this research for data analysis, embedding, retrieval, statistical validation, and system deployment. These tools have been selected based on their applicability, reliability, and prominence in biomedical data science and AI-driven research.

Tool / Package	Version	Purpose and Usage
Python	3.11	Core programming language for data manipulation, analysis, and automation.
Visual Studio Code	Latest Stable	Integrated Development Environment (IDE) for Python programming.
VS Code Extensions:		
- RooCode	Latest Stable	AI-assisted coding and debugging, enhancing coding productivity.
- Docker	Latest Stable	Containerisation of environments for reproducibility across deployments.
- Prettier	Latest Stable	Code formatting for readability and consistency.
- CSV Viewer (Rainbow CSV)	Latest Stable	Enhanced visualisation and inspection of CSV data files within VS Code.
Embedding and Vector Databases:		

Tool / Package	Version	Purpose and Usage
- Milvus	v2.x	Scalable vector database for efficient storage, indexing, and retrieval of high-dimensional embeddings.
- Weaviate	v1.x	Semantic search engine integrating vector search with knowledge graphs, ideal for biomedical context retrieval.
- FAISS	v1.7.4	High-performance library for efficient similarity search and clustering of dense vectors.
Embedding Models:		
- PubMedBERT (via Hugging Face)	Latest Stable	Specialised biomedical text embedding model trained on PubMed literature.
- NVIDIA NV-Embed v2	Latest Stable	Advanced generalised embedding model for comparative analysis of biomedical data retrieval.
Text Chunking Libraries:		
- LangChain	Latest Stable	RecursiveCharacterTextSplitter for contextual chunking of text data.
- LlamaIndex	Latest Stable	SentenceSplitter for preserving sentence-level semantics in chunking.

Tool / Package	Version	Purpose and Usage
- Anthropic's Contextual Chunker	Latest Stable	Advanced contextual chunking method for higher semantic coherence.
Retrieval-Augmented Generation:		
- LangChain RAG	Latest Stable	Framework for retrieval-augmented generation pipelines.
- Custom Graph-based RAG (Monarch Initiative Integration)	Latest Stable	Enhances biomedical context retrieval using structured phenotype, gene, and disease data.
Machine Learning and Statistical Packages:		
- scikit-learn	Latest Stable	Machine learning algorithms for statistical analysis and modelling.
- SciPy	Latest Stable	Statistical tests, such as t-tests, Wilcoxon, Spearman's correlation, Chi-square, and permutation tests.
- NumPy	Latest Stable	Numerical operations and handling high-dimensional arrays and vectors.
- Pandas	Latest Stable	Data manipulation, processing, and analysis.
Evaluation Metrics and Validation Tools:		

Tool / Package	Version	Purpose and Usage
- Cosine Similarity (via scikit-learn)	Latest Stable	Vector similarity measurement for embedding validation and relevance scoring.
- Approximate Nearest Neighbours (ANN) (via FAISS)	Latest Stable	Efficient retrieval of semantically similar embedding vectors.
- Precision, Recall, MAP, NDCG	Custom scripts	Standard metrics evaluating retrieval relevance and ranking effectiveness.
- Cohen's Kappa	via SciPy	Inter-rater reliability assessment for phenotype selection agreement.
- Jaccard Similarity & Overlap Coefficient	via scikit-learn	Statistical measure for evaluating overlap between selected phenotype sets.

Computational Infrastructure:

- Primary analysis and model development will utilise a dedicated workstation running **Ubuntu Linux 24.04.02 LTS**, equipped with **AMD processoring** and an **NVIDIA GeForce RTX 4070 GPU**. This system provides sufficient local processing power for initial development, testing, embedding generation, and iterative prototyping.
- For computationally intensive tasks requiring large-context model processing and extensive embedding computations, analysis will use with advanced infrastructure comprising two parallel-connected **NVIDIA GB10 Grace Blackwell Superchips (NVIDIA DGX™ systems)**, delivering a combined processing capacity capable of supporting models with up to **400 billion parameters**. This advanced GPU infrastructure significantly enhances capability, enabling high-throughput analysis, large-scale retrieval-augmented

generation (RAG), and efficient processing of extensive PubMed Medical literature journals.

- All computational frameworks and dependencies will utilise the latest version of the **NVIDIA CUDA Toolkit**, ensuring optimal GPU utilisation, performance efficiency, and compatibility.

Notes:

- All software will be containerised using **Docker** to ensure reproducibility.
- The use of open-source libraries allows transparency, customisation, and reproducibility, aligning with best practices in biomedical research.
- The entire codebase, data analysis scripts, model training procedures, and documented workflows will be maintained publicly on **GitHub**, under the username **donphi**, accessible via: https://github.com/donphi