2025

FINAL YEAR PROJECT

A Cost-Quantified, Hybrid-Retrieval RAG Agent for Pharmaceutical Q&A

Mentored by : Mr. Janith Prabhanuka

Presented by
Aron Fernando

Problem Statement

The pharmaceutical industry's data is a critical asset, but it's locked behind two major problems:

PROBLEM 01

Standard AI (LLM) queries are expensive, making large-scale analysis financially unviable.

PROBLEM 02

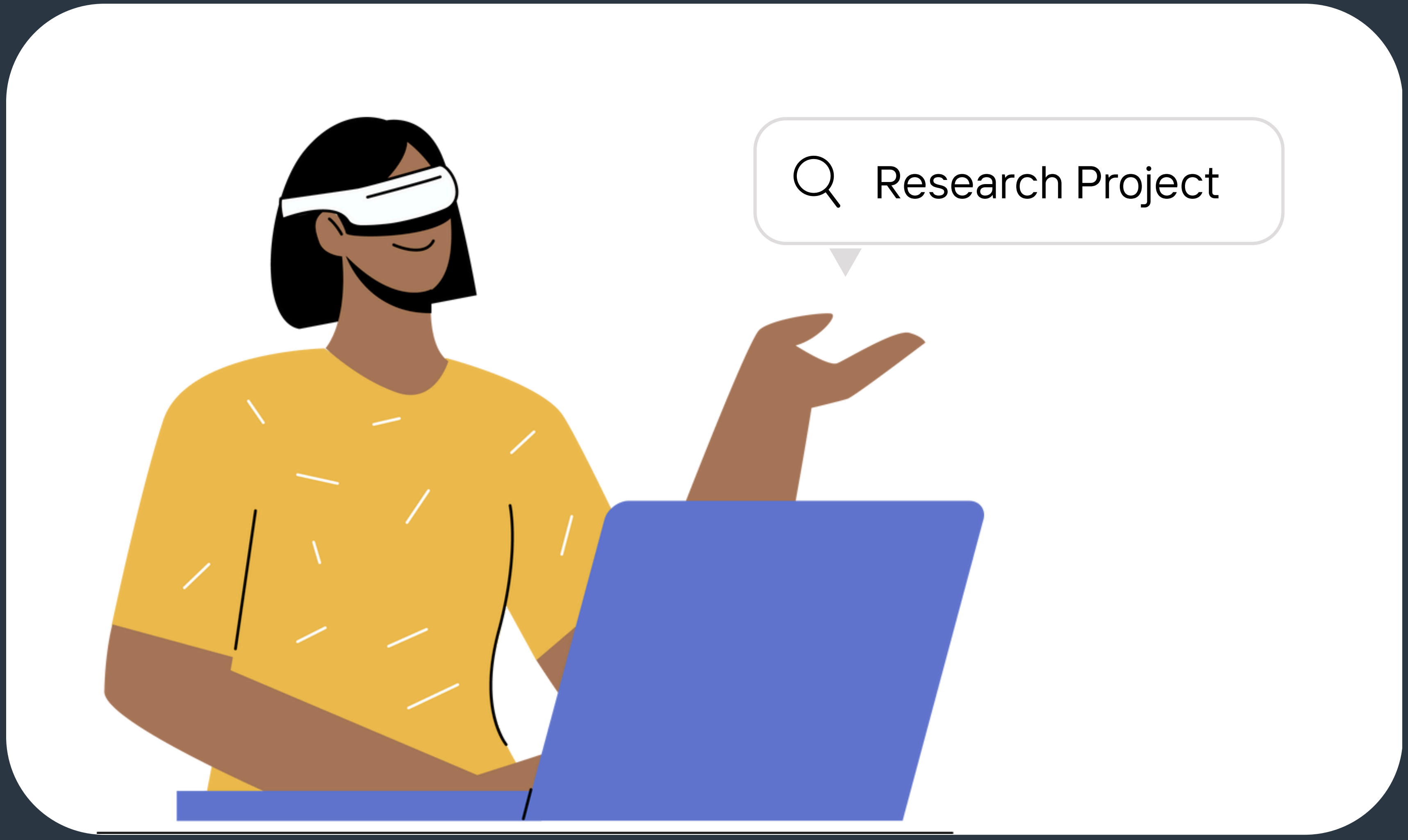
Critical data is "multimodal"—split between unstructured text (e.g., side effects) and structured tables (e.g., dosage).

Standard RAG systems fail at both.

The "Lost-in-Vectorization" Failure

Standard RAG systems treat all data as text. When they encounter a table, they shred it into text snippets, destroying its row-column integrity.

- Query: "What was the dosage for Group B?"
- Garbled Context: "10mg", "Group C", "Atorvastatin", "Group B"
- Result: The AI gets confused and provides a factually incorrect or incomplete answer.



The Solution: Part 1 - Hybrid Retrieval

A "Smart Librarian" Approach

Solution

Our solution is a Hybrid-Retrieval Agent that treats data intelligently. It uses two "filing systems" instead of one:

Vector Database:

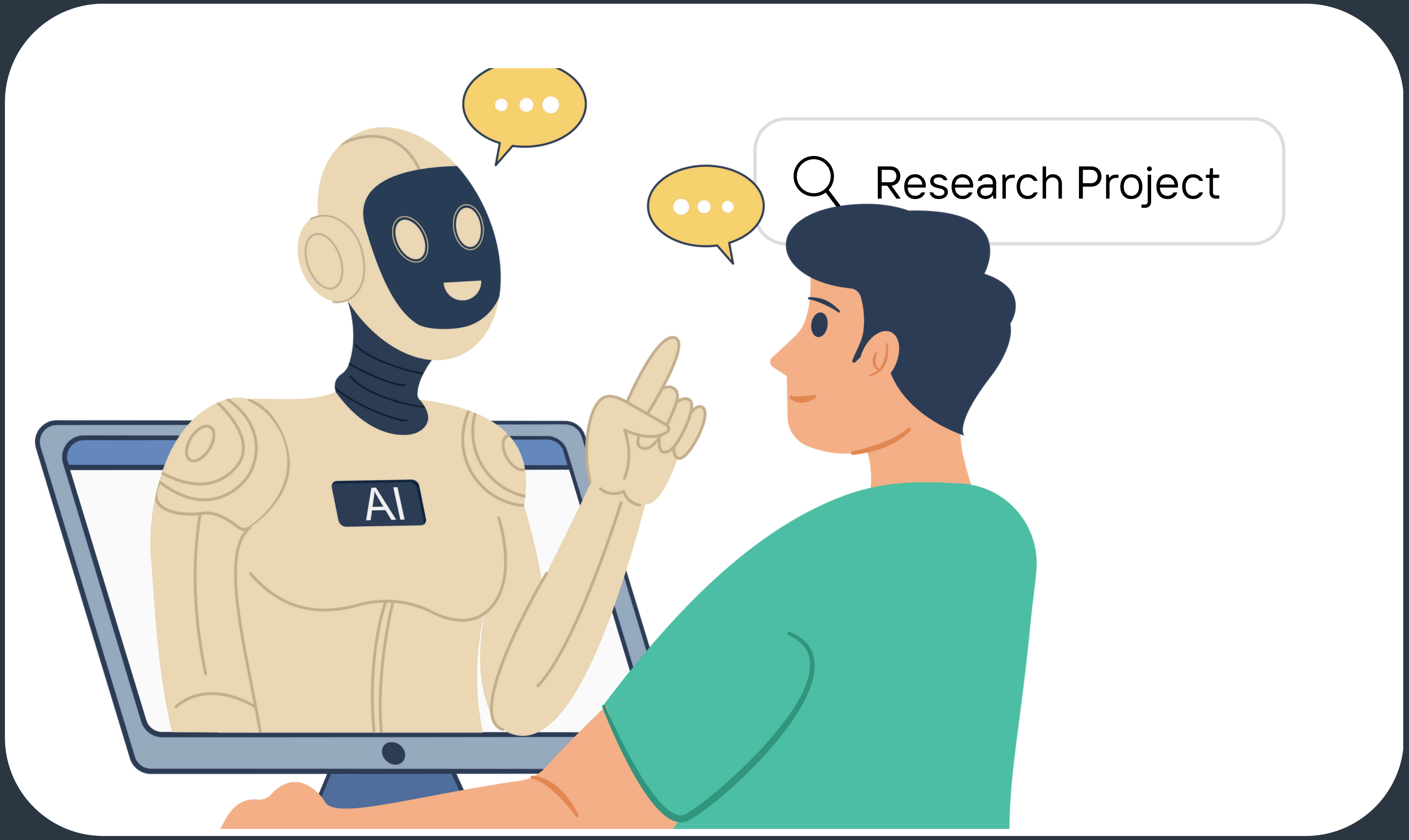
For unstructured text. This is for "semantic" search (e.g., find concepts related to 'side effects').

SQL Database:

For structured tables. This is for "factual" search (e.g., find the row where dose is '50mg').

Summary

This solves the accuracy problem by retrieving clean, perfect context every time.



The Solution: Part 2 - Model Cascade

An Intelligent "Brain" for Cost

Solution

To solve the cost problem, we use a Two-Tier Model Cascade

Tier-1 (Cheap):

A fast, local, open-source LLM acts as the "brain." It reads the user's query and classifies it.

Tier-2 (Expensive)

A powerful, paid API (like GPT-4) is used only when the Tier-1 brain detects a complex, hybrid query that needs synthesis.

Summary

This solves the cost problem by reserving the expensive tool for only the hardest tasks.

Proposed System Architecture

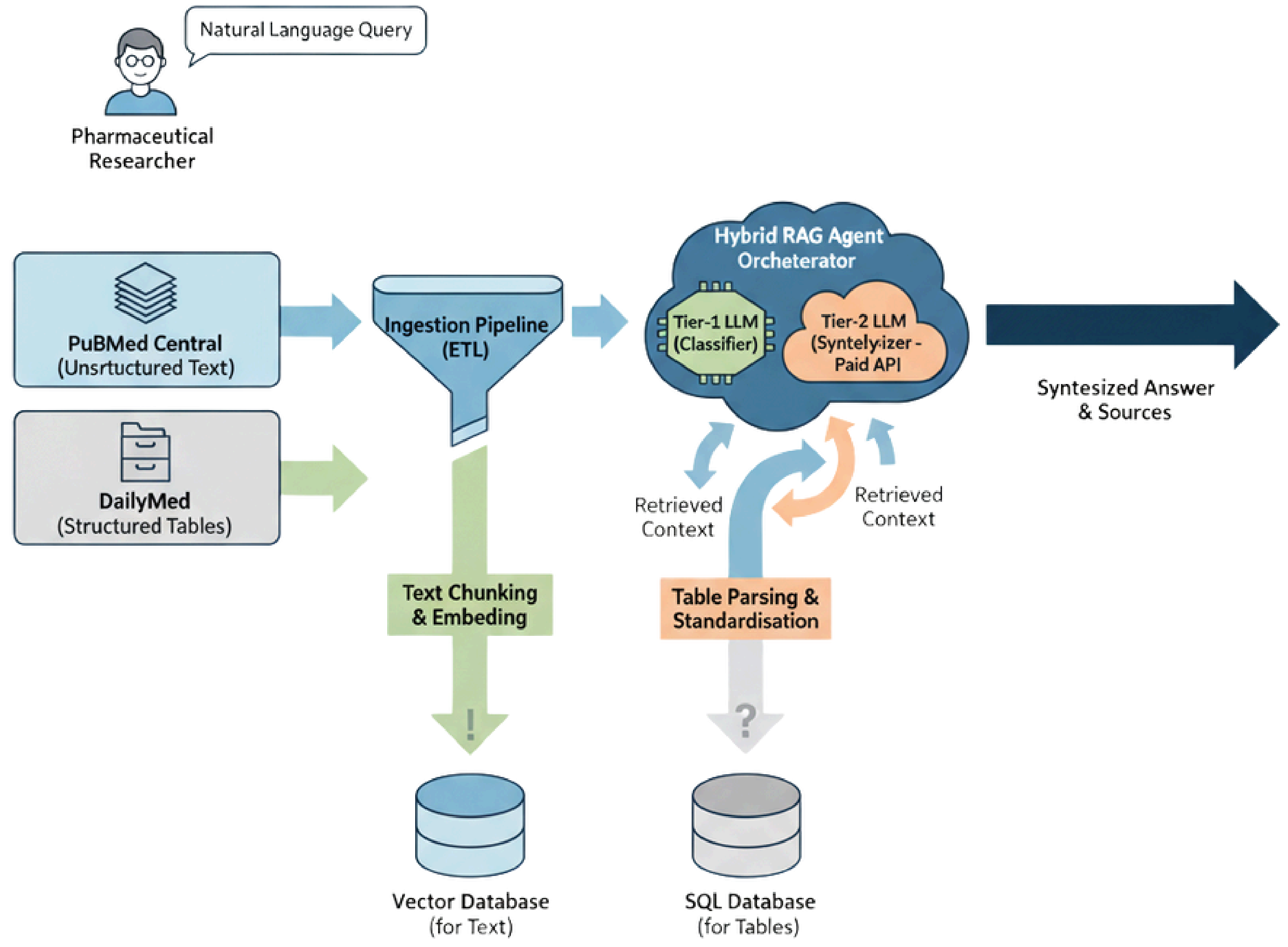
This diagram shows the complete flow, from ingestion to a hybrid query answer.

Ingestion

A PDF splitting into two paths: "Text" -> Vector DB and "Table" -> SQL DB.

Query

A user query hitting the Tier-1 Classifier, which then retrieves from one or both DBs, combines the context, and sends it to the Tier-2 Synthesizer for the final answer.)


Natural Language Query
Pharmaceutical Researcher
PuBMed Central (Unsrtuctured Text)
DailyMed (Structured Tables)
Ingestion Pipeline (ETL)
Hybrid RAG Agent Orcheterator
Tier-1 LLM (Classifier)
Tier-2 LLM (Syntely-izer - Paid API
Syntesized Answer & Sources
Retrieved Context
Retrieved Context
Text Chunking & Embeding
Table Parsing & Standardisation
!
?
Vector Database (for Text)
SQL Database (for Tables)

Research Gap & Contribution

Research Project

01. Siloed Research Existing literature is siloed. Research focuses on:

- Text-Only RAG
- Text-to-SQL (SQL-Only)
- Cost-Only Cascades No one has combined all three.

02. The Research Gap The specific, unaddressed gap is the lack of a unified system that integrates:

- A Hybrid-Retrieval (SQL+Vector) architecture.
- A Cost-Saving Model Cascade as the orchestrator.
- A Quantitative Benchmark of this system's cost vs. accuracy.

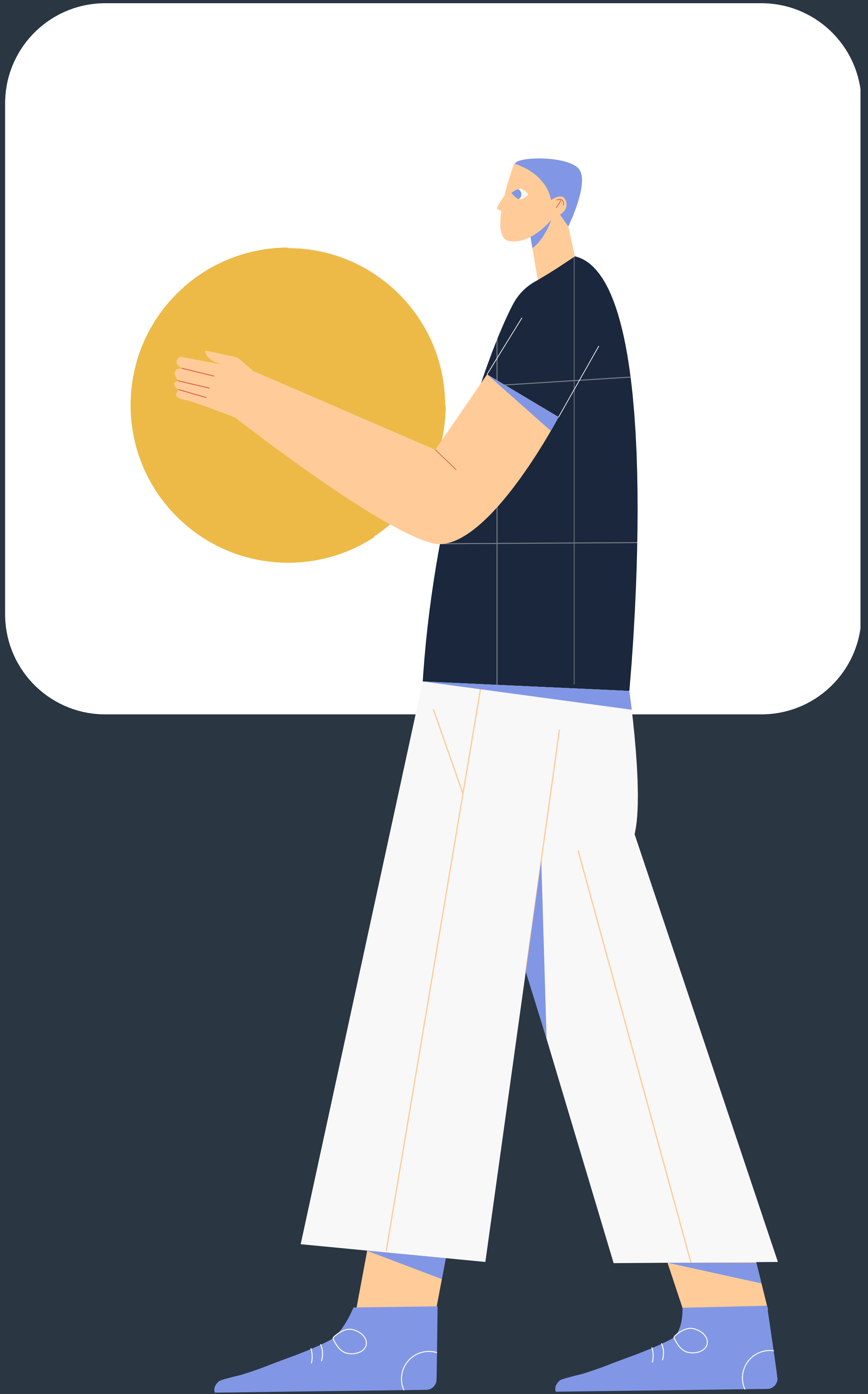
03. Our Contribution This project will deliver:

- A novel architecture for a hybrid RAG agent.
- The Cost-Efficiency Ratio (CER), a new metric to benchmark RAG.
- A quantitative analysis proving our agent is both more accurate and cheaper than the baseline.

Key Research Questions

This project is guided by three core questions:

- How can a model cascade be used as an intelligent query classifier for a hybrid-retrieval RAG system?
- How does this hybrid (SQL+Vector) architecture compare to a standard "vector-only" baseline in terms of factual accuracy (F1-score)?
- How does this hybrid (SQL+Vector) architecture compare to a standard "vector-only" baseline in terms of factual accuracy (F1-score)?



Research Methodology

Hypothesis: "The proposed Hybrid-Retrieval RAG Agent will demonstrate a measurably higher F1-score (more accurate) and a measurably lower CER (cheaper) compared to a standard vector-only baseline system."

Philosophy

Pragmatism (Focused on solving a practical problem).

Approach

Deductive (We are testing a specific hypothesis).

Method

Mono Method (Quantitative) (We will run a controlled experiment and measure numerical results).

Evaluation & Benchmarking

✦ Baseline: "Vector-Only" RAG

- Represents the "naive" approach.
- Ingestion: All data (text and tables) is "shredded" and put into a Vector DB.
- Cost: 100% of queries go to the expensive Tier-2 API.
- Expected Result: High cost, low accuracy on hybrid questions.

✦ Hybrid-Retrieval Agent

- The novel architecture from this project.
- Ingestion: Text -> Vector DB. Tables -> SQL DB.
- Cost: Uses the Tier-1 / Tier-2 cascade for cost-saving.
- Expected Result: Low cost, high accuracy on all questions.

✦ Key Metrics to be Measured

- Query Classification Accuracy (F1-Score): Does the Tier-1 brain work?
- Answer Accuracy (F1-Score / RAGAs): Is the final answer factually correct?
- Cost-Efficiency Ratio (CER): How much money did we save?
- Latency (ms): How fast is the system?

Key Risks & Mitigations

- Risk 1 (High): Ineffective Query Classifier

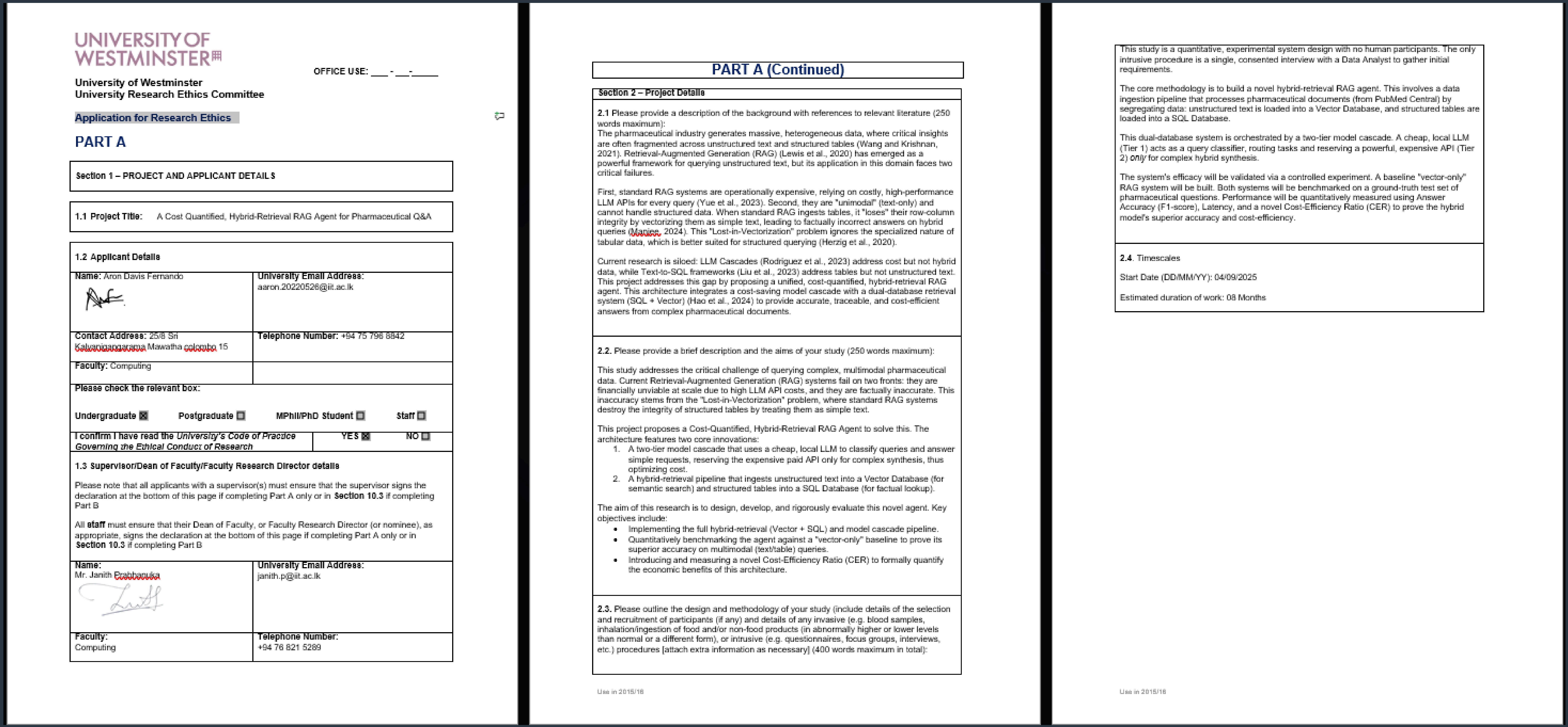
  Mitigation: Build and test the Tier-1 classifier on a manually labeled test set to ensure high F1-score before integration.

- Risk 2 (High): Failed Text-to-SQL Query

  Mitigation: Do not let the LLM write raw SQL. Use the LLM only to extract entities (e.g., drug name), then use safe, pre-written Python/SQL templates.

- Risk 3 (Medium): PDF Table Parsing (ETL) Failure

  Mitigation: Implement robust validation and error-logging. Focus on a "good enough" pipeline that correctly parses 80% of tables, not a perfect one.



UNIVERSITY OF WESTMINSTER

OFFICE USE: ___ - ___-_____

University of Westminster
University Research Ethics Committee

Application for Research Ethics

PART A

Section 1 – PROJECT AND APPLICANT DETAILS

1.1 Project Title: A Cost Quantified, Hybrid-Retrieval RAG Agent for Pharmaceutical Q&A

1.2 Applicant Details	
Name: Aron Davis Fernando	**University Email Address:** aaron.20220526@iit.ac.lk
Contact Address: 25/8 Sri Kalyanigangarama Mawatha colombo 15	**Telephone Number:** +94 75 798 8842
Faculty: Computing	
Please check the relevant box: **Undergraduate** ☒ **Postgraduate** □ **MPhil/PhD Student** □ **Staff** □	
I confirm I have read the *University's Code of Practice Governing the Ethical Conduct of Research*	**YES** ☒ **NO** □

1.3 Supervisor/Dean of Faculty/Faculty Research Director details

Please note that all applicants with a supervisor(s) must ensure that the supervisor signs the declaration at the bottom of this page if completing Part A only or in **Section 10.3** if completing Part B

All **staff** must ensure that their Dean of Faculty, or Faculty Research Director (or nominee), as appropriate, signs the declaration at the bottom of this page if completing Part A only or in **Section 10.3** if completing Part B

Name: Mr. Janith Prabhanuka	**University Email Address:** janith.p@iit.ac.lk
Faculty: Computing	**Telephone Number:** +94 76 821 5289

PART A (Continued)

Section 2 – Project Details

2.1 Please provide a description of the background with references to relevant literature (250 words maximum):

The pharmaceutical industry generates massive, heterogeneous data, where critical insights are often fragmented across unstructured text and structured tables (Wang and Krishnan, 2021). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a powerful framework for querying unstructured text, but its application in this domain faces two critical failures.

First, standard RAG systems are operationally expensive, relying on costly, high-performance LLM APIs for every query (Yue et al., 2023). Second, they are "unimodal" (text-only) and cannot handle structured data. When standard RAG ingests tables, it "loses" their row-column integrity by vectorizing them as simple text, leading to factually incorrect answers on hybrid queries (Maniee, 2024). This "Lost-in-Vectorization" problem ignores the specialized nature of tabular data, which is better suited for structured querying (Herzig et al., 2020).

Current research is siloed: LLM Cascades (Rodriguez et al., 2023) address cost but not hybrid data, while Text-to-SQL frameworks (Liu et al., 2023) address tables but not unstructured text. This project addresses this gap by proposing a unified, cost-quantified, hybrid-retrieval RAG agent. This architecture integrates a cost-saving model cascade with a dual-database retrieval system (SQL + Vector) (Hao et al., 2024) to provide accurate, traceable, and cost-efficient answers from complex pharmaceutical documents.

2.2. Please provide a brief description and the aims of your study (250 words maximum):

This study addresses the critical challenge of querying complex, multimodal pharmaceutical data. Current Retrieval-Augmented Generation (RAG) systems fail on two fronts: they are financially unviable at scale due to high LLM API costs, and they are factually inaccurate. This inaccuracy stems from the "Lost-in-Vectorization" problem, where standard RAG systems destroy the integrity of structured tables by treating them as simple text.

This project proposes a Cost-Quantified, Hybrid-Retrieval RAG Agent to solve this. The architecture features two core innovations:

1. A two-tier model cascade that uses a cheap, local LLM to classify queries and answer simple requests, reserving the expensive paid API only for complex synthesis, thus optimizing cost.
2. A hybrid-retrieval pipeline that ingests unstructured text into a Vector Database (for semantic search) and structured tables into a SQL Database (for factual lookup).

The aim of this research is to design, develop, and rigorously evaluate this novel agent. Key objectives include:

- Implementing the full hybrid-retrieval (Vector + SQL) and model cascade pipeline.
- Quantitatively benchmarking the agent against a "vector-only" baseline to prove its superior accuracy on multimodal (text/table) queries.
- Introducing and measuring a novel Cost-Efficiency Ratio (CER) to formally quantify the economic benefits of this architecture.

2.3. Please outline the design and methodology of your study (include details of the selection and recruitment of participants (if any) and details of any invasive (e.g. blood samples, inhalation/ingestion of food and/or non-food products (in abnormally higher or lower levels than normal or a different form), or intrusive (e.g. questionnaires, focus groups, interviews, etc.) procedures [attach extra information as necessary] (400 words maximum in total):

This study is a quantitative, experimental system design with no human participants. The only intrusive procedure is a single, consented interview with a Data Analyst to gather initial requirements.

The core methodology is to build a novel hybrid-retrieval RAG agent. This involves a data ingestion pipeline that processes pharmaceutical documents (from PubMed Central) by segregating data: unstructured text is loaded into a Vector Database, and structured tables are loaded into a SQL Database.

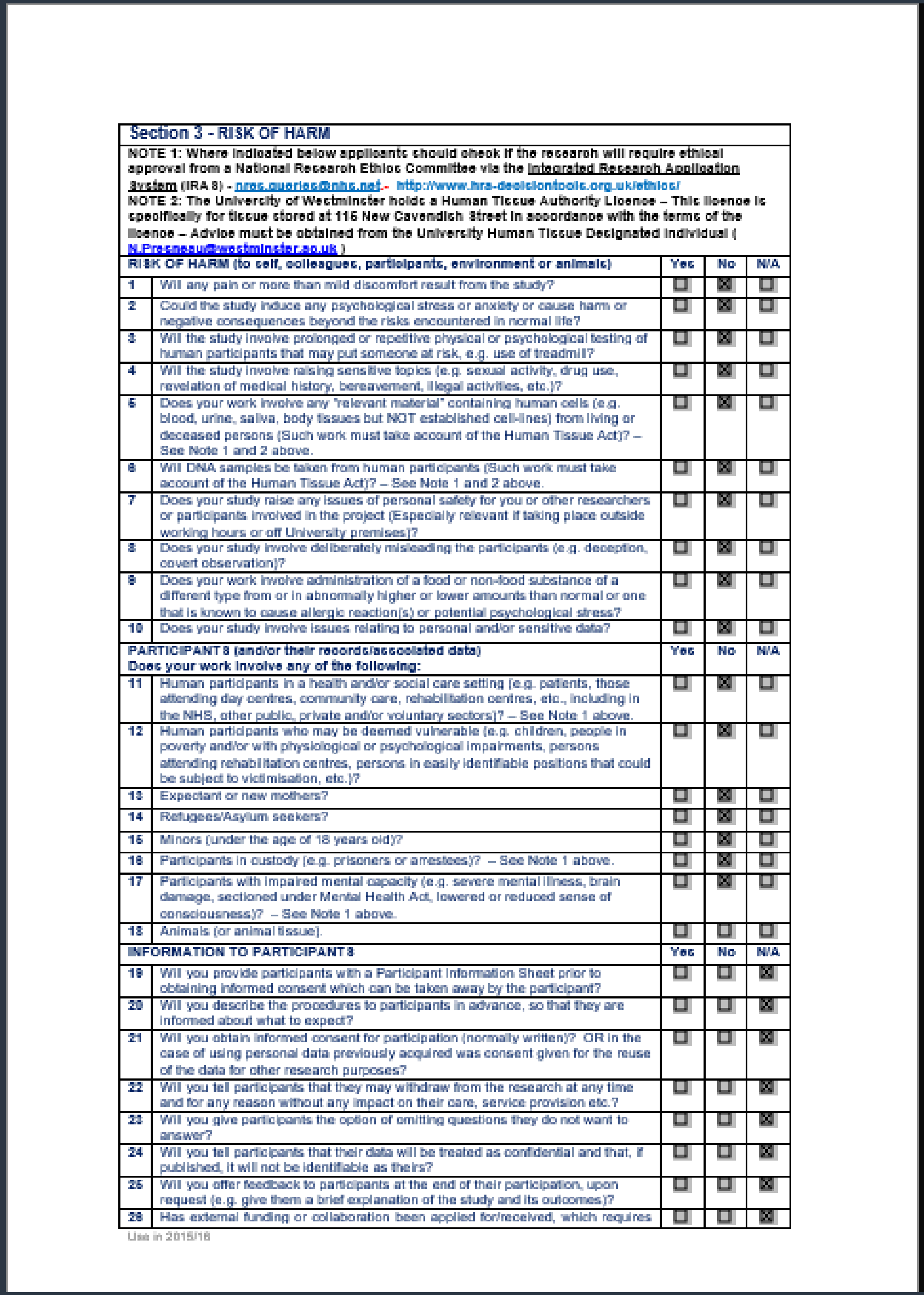
This dual-database system is orchestrated by a two-tier model cascade. A cheap, local LLM (Tier 1) acts as a query classifier, routing tasks and reserving a powerful, expensive API (Tier 2) *only* for complex hybrid synthesis.

The system's efficacy will be validated via a controlled experiment. A baseline "vector-only" RAG system will be built. Both systems will be benchmarked on a ground-truth test set of pharmaceutical questions. Performance will be quantitatively measured using Answer Accuracy (F1-score), Latency, and a novel Cost-Efficiency Ratio (CER) to prove the hybrid model's superior accuracy and cost-efficiency.

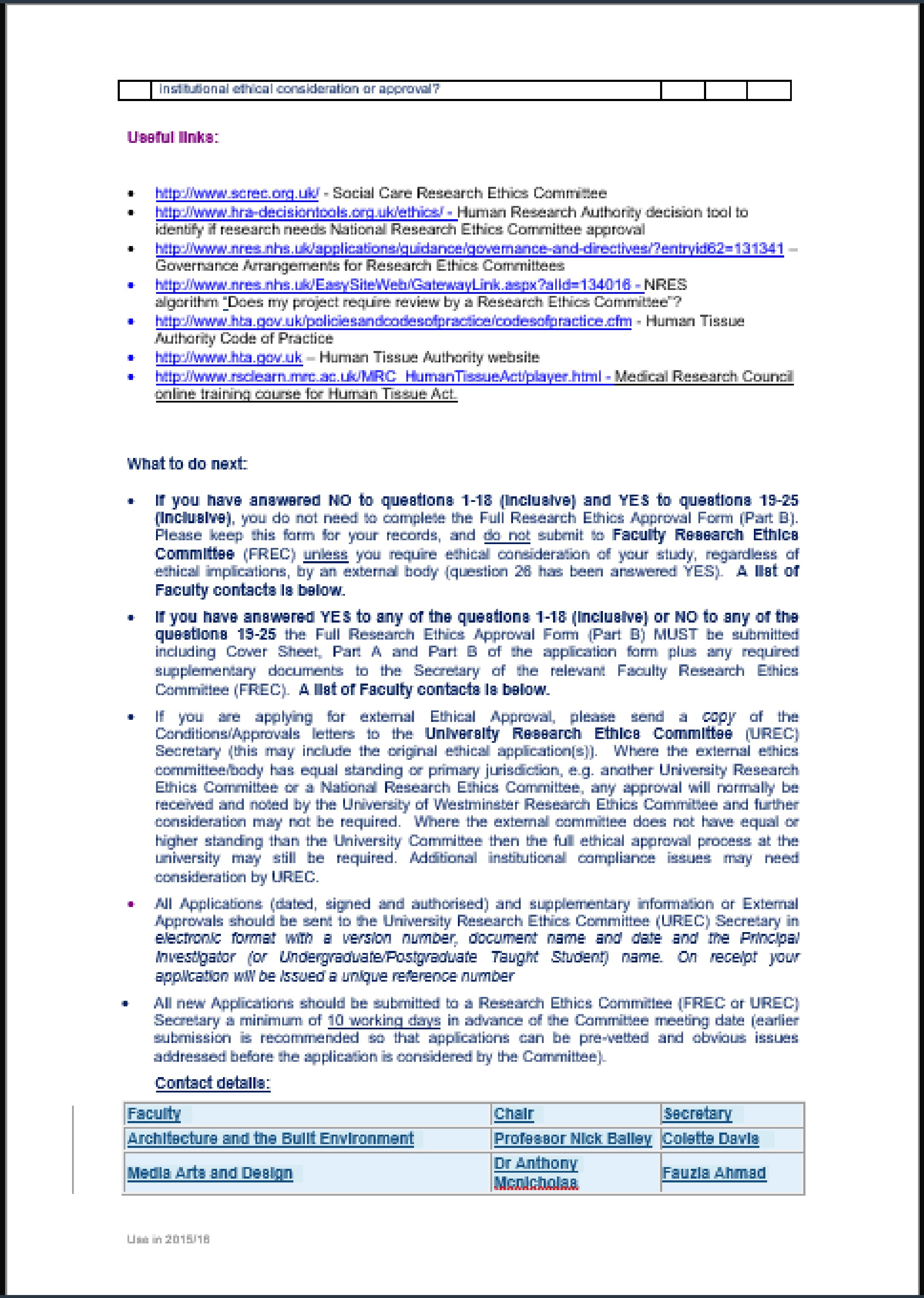
2.4. Timescales

Start Date (DD/MM/YY): 04/09/2025

Estimated duration of work: 08 Months



Section 3 - RISK OF HARM				
NOTE 1: Where indicated below applicants should check if the research will require ethical approval from a National Research Ethics Committee via the Integrated Research Application System (IRAS) - nres.queries@nhs.net - http://www.hra-decisiontools.org.uk/ethics/ **NOTE 2: The University of Westminster holds a Human Tissue Authority Licence – This licence is specifically for tissue stored at 115 New Cavendish Street in accordance with the terms of the licence – Advice must be obtained from the University Human Tissue Designated Individual (N.Presneau@westminster.ac.uk)**				
RISK OF HARM (to self, colleagues, participants, environment or animals)		**Yes**	**No**	**N/A**
1	Will any pain or more than mild discomfort result from the study?	□	☒	□
2	Could the study induce any psychological stress or anxiety or cause harm or negative consequences beyond the risks encountered in normal life?	□	☒	□
3	Will the study involve prolonged or repetitive physical or psychological testing of human participants that may put someone at risk, e.g. use of treadmill?	□	☒	□
4	Will the study involve raising sensitive topics (e.g. sexual activity, drug use, revelation of medical history, bereavement, illegal activities, etc.)?	□	☒	□
5	Does your work involve any "relevant material" containing human cells (e.g. blood, urine, saliva, body tissues but NOT established cell-lines) from living or deceased persons (Such work must take account of the Human Tissue Act)? – See Note 1 and 2 above.	□	☒	□
6	Will DNA samples be taken from human participants (Such work must take account of the Human Tissue Act)? – See Note 1 and 2 above.	□	☒	□
7	Does your study raise any issues of personal safety for you or other researchers or participants involved in the project (Especially relevant if taking place outside working hours or off University premises)?	□	☒	□
8	Does your study involve deliberately misleading the participants (e.g. deception, covert observation)?	□	☒	□
9	Does your work involve administration of a food or non-food substance of a different type from or in abnormally higher or lower amounts than normal or one that is known to cause allergic reaction(s) or potential psychological stress?	□	☒	□
10	Does your study involve issues relating to personal and/or sensitive data?	□	☒	□
PARTICIPANTS (and/or their records/associated data) **Does your work involve any of the following:**		**Yes**	**No**	**N/A**
11	Human participants in a health and/or social care setting (e.g. patients, those attending day centres, community care, rehabilitation centres, etc., including in the NHS, other public, private and/or voluntary sectors)? – See Note 1 above.	□	☒	□
12	Human participants who may be deemed vulnerable (e.g. children, people in poverty and/or with physiological or psychological impairments, persons attending rehabilitation centres, persons in easily identifiable positions that could be subject to victimisation, etc.)?	□	☒	□
13	Expectant or new mothers?	□	☒	□
14	Refugees/Asylum seekers?	□	☒	□
15	Minors (under the age of 18 years old)?	□	☒	□
16	Participants in custody (e.g. prisoners or arrestees)? – See Note 1 above.	□	☒	□
17	Participants with impaired mental capacity (e.g. severe mental illness, brain damage, sectioned under Mental Health Act, lowered or reduced sense of consciousness)? – See Note 1 above.	□	☒	□
18	Animals (or animal tissue).	□	□	□
INFORMATION TO PARTICIPANTS		**Yes**	**No**	**N/A**
19	Will you provide participants with a Participant Information Sheet prior to obtaining informed consent which can be taken away by the participant?	□	□	☒
20	Will you describe the procedures to participants in advance, so that they are informed about what to expect?	□	□	☒
21	Will you obtain informed consent for participation (normally written)? OR in the case of using personal data previously acquired was consent given for the reuse of the data for other research purposes?	□	□	☒
22	Will you tell participants that they may withdraw from the research at any time and for any reason without any impact on their care, service provision etc.?	□	□	☒
23	Will you give participants the option of omitting questions they do not want to answer?	□	□	☒
24	Will you tell participants that their data will be treated as confidential and that, if published, it will not be identifiable as theirs?	□	□	☒
25	Will you offer feedback to participants at the end of their participation, upon request (e.g. give them a brief explanation of the study and its outcomes)?	□	□	☒
26	Has external funding or collaboration been applied for/received, which requires	□	□	☒



	Institutional ethical consideration or approval?			

Useful links:

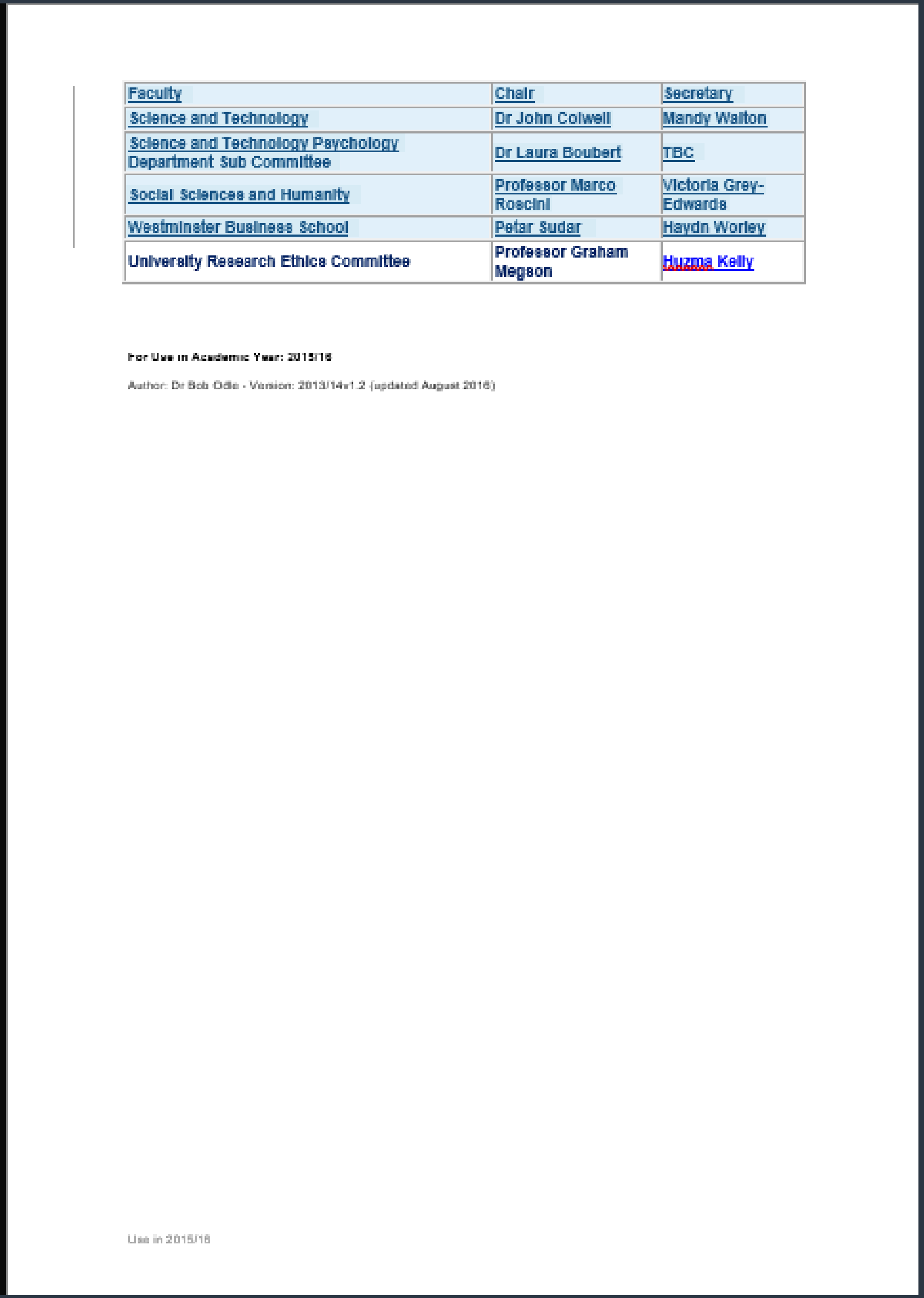
- http://www.screc.org.uk/ - Social Care Research Ethics Committee
- http://www.hra-decisiontools.org.uk/ethics/ - Human Research Authority decision tool to identify if research needs National Research Ethics Committee approval
- http://www.nres.nhs.uk/applications/guidance/governance-and-directives/?entryid62=131341 – Governance Arrangements for Research Ethics Committees
- http://www.nres.nhs.uk/EasySiteWeb/GatewayLink.aspx?alId=134016 - NRES algorithm "Does my project require review by a Research Ethics Committee"?
- http://www.hta.gov.uk/policiesandcodesofpractice/codesofpractice.cfm - Human Tissue Authority Code of Practice
- http://www.hta.gov.uk – Human Tissue Authority website
- http://www.rsclearn.mrc.ac.uk/MRC_HumanTissueAct/player.html - Medical Research Council online training course for Human Tissue Act.

What to do next:

- **If you have answered NO to questions 1-18 (inclusive) and YES to questions 19-25 (inclusive)**, you do not need to complete the Full Research Ethics Approval Form (Part B). Please keep this form for your records, and do not submit to **Faculty Research Ethics Committee** (FREC) unless you require ethical consideration of your study, regardless of ethical implications, by an external body (question 26 has been answered YES). **A list of Faculty contacts is below.**
- **If you have answered YES to any of the questions 1-18 (inclusive) or NO to any of the questions 19-25** the Full Research Ethics Approval Form (Part B) MUST be submitted including Cover Sheet, Part A and Part B of the application form plus any required supplementary documents to the Secretary of the relevant Faculty Research Ethics Committee (FREC). **A list of Faculty contacts is below.**
- If you are applying for external Ethical Approval, please send a *copy* of the Conditions/Approvals letters to the **University Research Ethics Committee** (UREC) Secretary (this may include the original ethical application(s)). Where the external ethics committee/body has equal standing or primary jurisdiction, e.g. another University Research Ethics Committee or a National Research Ethics Committee, any approval will normally be received and noted by the University of Westminster Research Ethics Committee and further consideration may not be required. Where the external committee does not have equal or higher standing than the University Committee then the full ethical approval process at the university may still be required. Additional institutional compliance issues may need consideration by UREC.
- All Applications (dated, signed and authorised) and supplementary information or External Approvals should be sent to the University Research Ethics Committee (UREC) Secretary in *electronic format with a version number, document name and date and the Principal Investigator (or Undergraduate/Postgraduate Taught Student) name. On receipt your application will be issued a unique reference number*
- All new Applications should be submitted to a Research Ethics Committee (FREC or UREC) Secretary a minimum of 10 working days in advance of the Committee meeting date (earlier submission is recommended so that applications can be pre-vetted and obvious issues addressed before the application is considered by the Committee).

Contact details:

Faculty	Chair	Secretary
Architecture and the Built Environment	Professor Nick Bailey	Colette Davis
Media Arts and Design	Dr Anthony Mcnicholas	Fauzia Ahmad



Faculty	Chair	Secretary
Science and Technology	Dr John Colwell	Mandy Walton
Science and Technology Psychology Department Sub Committee	Dr Laura Boubert	TBC
Social Sciences and Humanity	Professor Marco Roscini	Victoria Grey-Edwards
Westminster Business School	Petar Sudar	Haydn Worley
University Research Ethics Committee	Professor Graham Megson	Huzma Kelly

For Use in Academic Year: 2015/16

Author: Dr Bob Odle - Version: 2013/14v1.2 (updated August 2016)

Final Year Project

2025

Thank You

✦ Thank you for your attention and time

Presented by
Aron Fernando