PoC Project Plan

Team: Bit By Bit

Use Case: Human Ethics Advisor

PoC: RAG-based Research Application Assistant

1 TABLE OF CONTENTS

2 PROJECT OVERVIEW

2.1 EXECUTIVE SUMMARY

The Human Ethics Advisor is an AI-powered tool that helps University of Auckland researchers review and strengthen their human ethics applications before submission, reducing errors and admin workload. The tool can read applications and provide feedback to help ensure they meet ethical standards. The tool is built using IBM Watsonx.ai and based on a Retrieval-Augmented Generation (RAG) architecture. It draws from key policy sources such as NEAC, HRC, UAHPEC, AHREC, Te Ara Tika, and institutional privacy policies.

Users can interact with the system through a Streamlit-based web UI by asking questions or uploading documents (PDF/DOCX). The tool helps researchers:

1. determine the appropriate ethics committee – Auckland Human Participants Ethics Committee (UAHPEC) and the Auckland Health Research Ethics Committee (AHREC).
2. ensure that researchers consider the Te Ara Tika principles for Māori ethical research, as well as national ethical standards such as those from the National Ethics Advisory Committee (NEAC) and the Health Research Council (HRC).
3. verify whether informed consent is clearly described for all participants unless a waiver is justified, and whether any digital tools used for data collection or communication have the appropriate Authority to Operate (ATO) under the University of Auckland's IT and privacy policies.
4. ensure compliance with application requirements (e.g. informed consent, ATO compliance), and provide structured feedback through a natural language interface.
5. guide researchers in meeting essential criteria for ethics approval. These include compliance with applicable guidelines, consistency between the application and supporting documents, readability for non-experts, clear outlining of research procedures and responsibilities, completeness of all required sections, full disclosure and evaluation of risks, and confirmation of data access approvals from relevant parties.

The Proof of Concept (PoC) achieved model performance above 70% across key evaluation metrics: answer correctness, objective faithfulness, and contextual accuracy. Optimisation was guided through iterative evaluation and manual QA construction. The system is designed for future deployment via Microsoft 365 Copilot, with a maintainable backend suitable for non-technical updates by ethics team staff.

2.2 PROJECT SPONSOR(S) / STAKEHOLDERS(S) / PROJECT TEAM

Partner Executive Sponsor

Name	Title	Description	Email / Contact Info
Thomas Lacombe	MAI Programme Director	Project coordination	thomas.lacombe@auckland.ac.nz

Project Stakeholders

Name	Title	Stakeholder for	Email / Contact Info
Nick Kearns	Associate Director, Research Ethics	Compliance Alignment	n.kearns@auckland.ac.nz
Jordan Woodhouse	Ethics Specialist	UAHPEC/AHREC processes	N/A
Dana Wensley	Head of Research Ethics	Compliance Alignment	dana.wensley@auckland.ac.nz
Madhavi Manchi	Human Ethics Manager	UAHPEC/AHREC processes	m.manchi@auckland.ac.nz
Fiona Cheal	Research Ethics and Regulatory Coordinator	Application processes	f.cheal@auckland.ac.nz

Project Team

Name	Title	Role	Email / Contact Info
Qinxue Feng	Student	Report, Presentation	qfen985@aucklanduni.ac.nz
Katie Law	Student	Report, Presentation	wlaw777@aucklanduni.ac.nz
Jiajun Xiao	Student	Team Leader, PoC, Presentation	jxia993@aucklanduni.ac.nz
Yizheng Xing	Student	Report, Presentation	yxin483@aucklanduni.ac.nz
Nanyuanyang Zhang	Student	PoC, Presentation	nzha229@aucklanduni.ac.nz

2.3 PROJECT SUCCESS CRITERIA

To ensure the Human Ethics Adviser is effective and reliable, the tool was evaluated through two levels of benchmarking:

Part 1: IBM Watsonx.ai Model Evaluation

- **Answer Correctness** (the factual accuracy of the generated answers) should **exceed 70%.**
- **Objective Faithfulness** (the extent to which the answer remains faithful to the information in the knowledge base) should **exceed 70%.**
- **Contextual Accuracy** (how well the answer addresses the specific question and context) should **exceed 70%.**

Part 2: LLM-Based Ethics Evaluation for Prompt Engineering

In addition to model-level QA metrics, we performed a structured rubric-based evaluation focused on research ethics relevance.

This assessment was conducted by scoring our model's answers for different prompts using two independent LLMs, each assigning ratings across four key criteria. The final score for each dimension was based on the average of the two LLMs' outputs.

Evaluation Dimensions:

- **Compliance & Risk Identification:** Does the system correctly flag ethical risks and policy violations?
- **Policy Traceability & Justification:** Are the answers grounded in specific policies or guidelines?
- **Actionability of Recommendations:** Are the tool's suggestions practical and clearly applicable?
- **Structure & Clarity:** Are the responses logically structured and easy to understand?

The average score across both LLMs should exceed **85% per category**.

This two-layer evaluation demonstrates that the Human Ethics Adviser PoC is accurate and consistent, ethically aligned, readable, and helpful for researchers.

End-to-End Use Case Validation:

To further ensure real-world usability, we created three example application cases, each embedding different combinations of common and complex issues (e.g. missing risk disclosures, policy gaps, cultural consultation deficiencies). For the tool to be considered successful in each area, it was required to correctly identify and address all relevant issues present in every example case.

Not all criteria or problems were present in every example. For each success criterion below, the tool was evaluated only on the issues that actually appeared in that example.

Acceptance threshold**:** For each relevant criterion, the tool had to correctly identify and address all issues present in the case.

- **Reduce time spent determining the appropriate ethics committee**
  - Current challenge: Researchers frequently submit their applications to the wrong ethics committee, resulting in duplicated effort, rework, and delays.
  - Success criteria: The AI tool assists researchers early in the process by accurately identifying the appropriate committee, i.e. AHREC or UAHPEC. This significantly reduces misclassification errors and eliminates the need for resubmissions, hence improving workflow efficiency from the outset.
- **Improve the internal consistency of ethics application materials**
  - Current challenge: Key information such as the number of participants or compensation amounts often appears inconsistently across different parts of the application, leading to confusion, additional review time, and ethical concerns.
  - Success criteria: The AI tool detects inconsistencies within the application, enabling researchers to self-correct before submission. This reduces the need for revisions, improves accuracy, and increases the likelihood of first-round approval.
- **Enhance readability and comprehension of application materials**
  - Current challenge: Many applications include technical language or unexplained acronyms, which are difficult for lay members of the ethics committee to understand, slowing down the review process.
  - Success criteria: The tool helps researchers refine language for justification, ensuring non-experts can understand the content. This reduces the need for clarifications and accelerates ethical review timelines.
- **Strengthen the identification and disclosure of ethical risks**

    - Current challenge: Researchers often overlook or underestimate potential ethical risks, such as psychological discomfort, identity exposure, or inadequate protection of vulnerable populations. This results in repeated rejections, revision requests, and delays in the approval process.
    - Success criteria: The tool identifies missing or vague risk disclosures and prompts researchers to clarify or expand their responses. By aligning with the guidelines given, the tool improves the accuracy and completeness of risk sections, increasing the likelihood of first-round approval.
- **Ensure compliance of digital tools with data privacy requirements**
    - Current challenge: Researchers may use digital platforms for data collection or communication without realising that these tools lack the University's Authority to Operate (ATO), which raises privacy and data security concerns.
    - Success criteria: The AI tool verifies whether all tools listed in an application comply with the University's ATO list. This helps prevent applications from being returned due to policy violations and reduces the workload for ethics reviewers.
- **Improve the completeness of application forms and the integration of Te Ara Tika principles**
    - Current challenge: Applications are frequently submitted with missing sections, particularly around Māori ethical considerations, which are required under University guidelines.
    - Success criteria: The tool highlights incomplete fields and specifically prompts researchers to address Te Ara Tika values. This ensures the submissions are culturally aligned and policy-compliant, so they are less likely to be returned for revision.
- **Ensure future adaptability to updated ethics guidelines and documents**
    - Current challenge: Ethical policies, guidelines, and institutional requirements are regularly updated. A static AI tool that cannot accommodate new information may quickly become outdated or provide misleading advice.
    - Success criteria: The AI tool should be designed to support the easy integration of new or updated ethics documents. This ensures the system remains aligned with evolving national standards (such as NEAC and HRC) and University of Auckland policies. A flexible and maintainable architecture will enhance the tool's long-term value and reduce the cost and effort required for ongoing updates.

2.4 ASSUMPTIONS

Assumptions:

- All ethics application materials are in PDF/DOCX format.
- Researchers are willing to use the AI tool before submitting their application.
- The ethics team is available to maintain the tool's knowledge base and ensure content remains up to date. The tool is designed to be maintainable by non-technical staff, allowing staff to update system rules or content.
- All required policy documents and ethics manuals can be securely used for training and testing the system.

Dependencies:

- Continued access to IBM Watsonx.ai.
- Access to the authoritative University of Auckland ATO lists and IT security guidance.
- Availability of real or anonymised sample ethics applications to train and validate tool performance.

- Access to official UAHPEC/AHREC documents, application templates, and decision-making criteria.
- Ongoing input from stakeholders for clarifying requirements and validating logic.

Constraints:

- All data handling and model interactions must comply with the University of Auckland's information security and privacy policies. No sensitive data may be used to train external/open-source models.
- The AI tool acts as a support system and cannot replace formal ethical review. A disclaimer must be built into clarify that all final decisions rest with the ethics staff.
- The AI tool depends on the examples and information it is trained on. If the data is limited, the tool might not handle very complex or unusual ethical situations well.
- Most ethics rules apply to common research cases, but special or unusual situations, like studies involving multiple cultures or countries, may need a human expert to review them.
- The tool should be careful not to give wrong or misleading advice about cultural values, especially Māori principles. It can remind researchers to think about these, but it should not try to explain or decide cultural issues.

3 SCOPE OF WORK - TECHNICAL PROJECT PLAN

For our PoC development, we have planned the following activities:

1. Requirements analysis:
    - Identify stakeholders
    - Define success criteria, assumptions and evaluation thresholds
    - Understand the main pain points in the ethical submission process
2. Data collection and preprocessing:
    - Collect publicly available or non-confidential text resources
        - UAHPEC and AHREC applicant manuals
        - NEAC national ethics standards
        - Te Ara Tika guidelines
    - Data pre-process: Construct evaluation QA pairs
3. AI system implementation:
    - Implement baseline RAG system using Watsonx
    - Set up an online knowledge base on Watsonx
4. RAG system optimisation:
    - Test and evaluate the optimised foundation model choice
    - Test and evaluate the optimised retrieval parameter settings
    - Test and evaluate the optimised parameter patterns
5. System-level prompt engineering:
    - Design different versions of the prompt and test samples
    - Cross-validate the performance of different prompts and find the optimal choice
6. Prototype application development:

- Implement a prototype user interface

4 SOLUTION ARCHITECTURE / ARCHITECTURAL DIAGRAM

4.1 OVERVIEW

The proposed solution is an AI-driven system that aims to help UoA researchers check the alignment of their research application with relevant human ethics policies and fill in their application documents. The system is built upon the retrieval-augmented generation (RAG) approach, leveraging IBM Watsonx's built-in AutoAI [1] solution as the core method powering the architecture.

4.2 SOLUTION ARCHITECTURE COMPONENTS

- **Data source:** The source data used for our AI system is the shared document from our project sponsor/stakeholder, and all documents used are non-confidential. This includes relevant policies and guidelines. These files will be used as the knowledge pool, from which the model will retrieve information and use the retrieved results to answer users' questions.
- **Evaluation sets:** We constructed QA pairs for RAG assessment and assisting parameter tuning. The QA pairs were first manually structured according to the data sources (a total of 19 samples), then generated on a larger scale with the assistance of ChatGPT (Around 270 samples). covering different types of user questions.
- **RAG Embedded models:** Embedded models are used in the RAG process for vectorising queries and data chunks. which allows models to capture the semantic meanings. In our architecture, we utilised the built-in embedded models provided by IBM Watsonx
- **RAG Foundation Model:** The foundation model used for the RAG system was chosen among all built-in models provided by IBM Watsonx.
- **System-level prompt engineering:** Prompt engineering was applied at the system level to guide the model to respond with better quality.
- **Web interface deployment:** A deployable web interface is developed using the Streamlit library of Python. The interface allows users to ask questions and upload documents (In PDF or Word format) to the AI assistant.

4.3 OPTIMISATION APPROACHES

During the optimisation process, we partially used Watsonx's AutoAI capabilities, which automatically generate multiple configuration patterns. For each pattern, the system automatically tuned certain parameters, while others were manually configured. Our optimisation efforts focused on the following key areas:

1) Refinement of evaluation files:

    The evaluation files are manually optimised. Example refinement includes the addition of extra QA pairs, answers that involve multiple source documents, etc.

2) Choice of foundation/embedded models
    - Foundation models: Selection of IBM Watsonx's built-in RAG foundation models was based on assessment with evaluation files. The evaluation results were compared against our success criteria: The quantitative indexes of answer correctness, faithfulness, and context correctness.
    - Embedded models: Embedded models were left to be determined by AutoAI's automatic optimisation. Different embedded models were chosen for different patterns during evaluation, and we simply chose the best-performing pattern according to our success criteria.
3) Retrieval parameter tuning:
    - Chunking: The chunking parameters (Chunk size and chunk overlap) determine how the document text is broken down. We left this parameter to be optimised automatically for different patterns.
    - Retrieval: The retrieval methods were manually tuned. This includes adjusting the retrieval methods (Simple or window), number of chunks, and window size.
4) System-level prompt engineering: Several versions of system-level prompts were designed manually. Each of the prompts was then experimented with a set of sample applications that cover common situations. The experiment results were cross-compared, and Large language models were utilised to generate a rating for different prompt versions.

4.4 DESIGN PROCESS FLOW

The overall process flow from identifying data sources to deployment is summarised in the following flowchart (Figure 1):

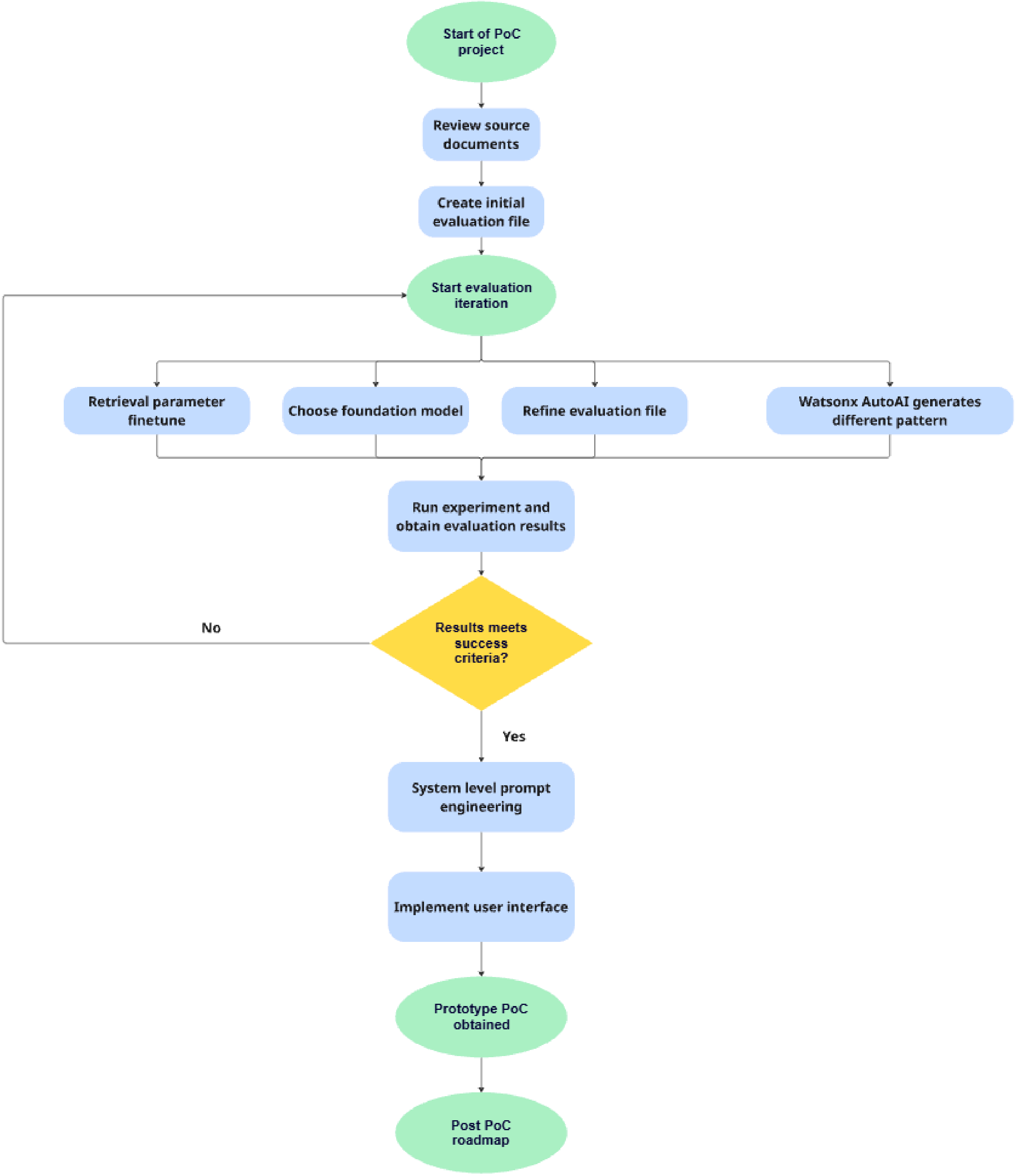

Start of PoC project
Review source documents
Create initial evaluation file
Start evaluation iteration
Retrieval parameter finetune
Choose foundation model
Refine evaluation file
Watsonx AutoAI generates different pattern
Run experiment and obtain evaluation results
Results meets success criteria?
No
Yes
System level prompt engineering
Implement user interface
Prototype PoC obtained
Post PoC roadmap

Figure 1: Design process flowchart.

The process aligns closely with our proposed project plan.

We begin by reviewing the shared documents from the project sponsor. These documents serve as a primary knowledge base for our RAG system and are non-confidential, which complies with UoA data governance policies. After that, we constructed an initial evaluation of QA pairs for assessing the RAG system's performance based on those documents. (The QA samples are generated on a larger scale using ChatGPT in the later experiment process.)

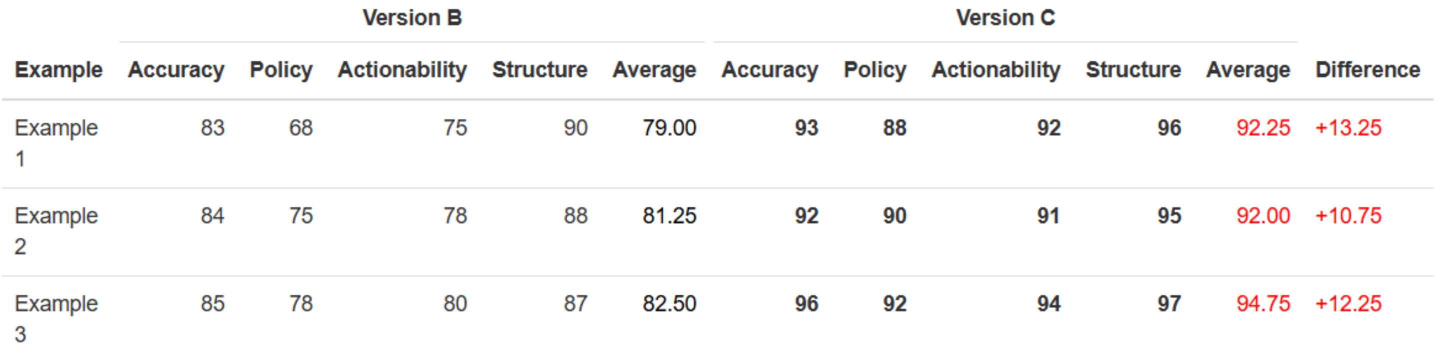
We then proceeded to an iterative process of evaluation, where we experimented and compared different foundation models, auto-optimised patterns (Chunking parameters and embedded model choice), and retrieval parameters, as well as refining evaluation files when needed (E.g., when evaluation Q&As are not generic enough). Multiple optimisation experiments were conducted throughout the development process to experiment with different experiment patterns, evaluation sample sizes, and types. We determined 4 iterations during the experimental process, which define the major versions of our RAG system. The changes and rationale for each iteration are summarised in the table below.

Iteration	Change made	Reason	Outcome
1	Limited the foundation model choice from all built-in models to: mistral-small-3-1-24b-instruct-2503.	mistral-small-3-1-24b-instruct-2503 demonstrates the best performance throughout multiple experiments. It is decided to use this foundation model for future experiments.	Limiting the selection of foundation model choices provides a more reliable baseline for subsequent experiments.
2	Adding extra QA pairs for the evaluation file to ensure full coverage of source files.	The context correctness is observed to be quite low. After inspecting the used evaluation set, it is found that the QA pairs do not cover the use case of one of the source documents. Thus, extra QA pairs are added to the evaluation set to ensure full coverage of the source documents.	In the evaluation results of new patterns generated in experiments after this iteration, we observed a rise in context correctness.
3	Refining QA pair answers to become more generic. E.g.: Added answer that involves info from more than one document.	Although the context correctness increased after the last iteration, it is still not meeting our success criteria (70%). This was attributed to the lack of generalisation of the evaluation file.	In the evaluation results of new patterns generated in experiments after this iteration, a rise in context correctness is observed.

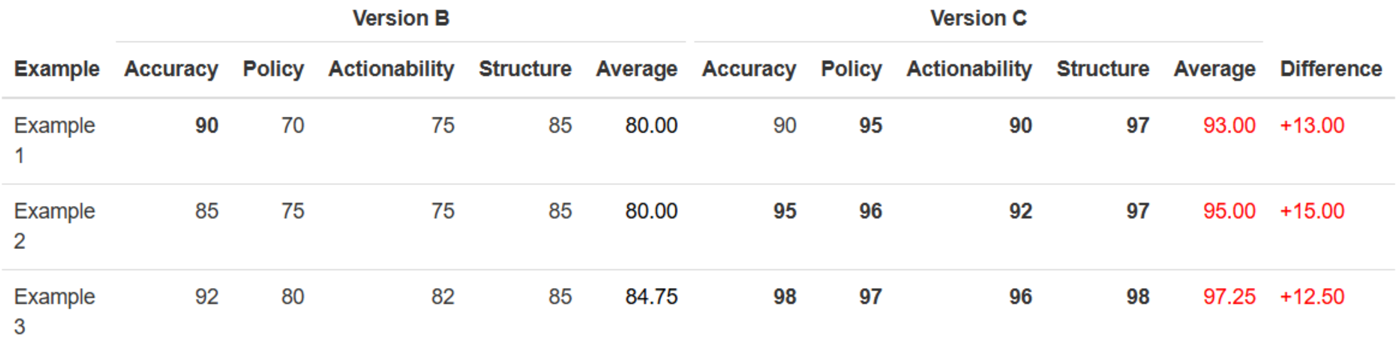
		(E.g., Lacking cross-document answers, previous answers are limited to sourcing from a single document.)	
4 (Final version)	Set window size to 2 and number of chunks to 5.	Retrieval parameters were fine-tuned based on performance observations across multiple experiments. The selected settings yielded the best overall results.	All 3 quantified indices (answer correctness, faithfulness, and context correctness) breach through 70%.

Table 1: Summary of major optimisation iterations, including changes made, rationale, and outcomes for each step in model and evaluation pipeline tuning.

* For screenshots of each iteration, please see section 7.2, resources (Figures A - D).

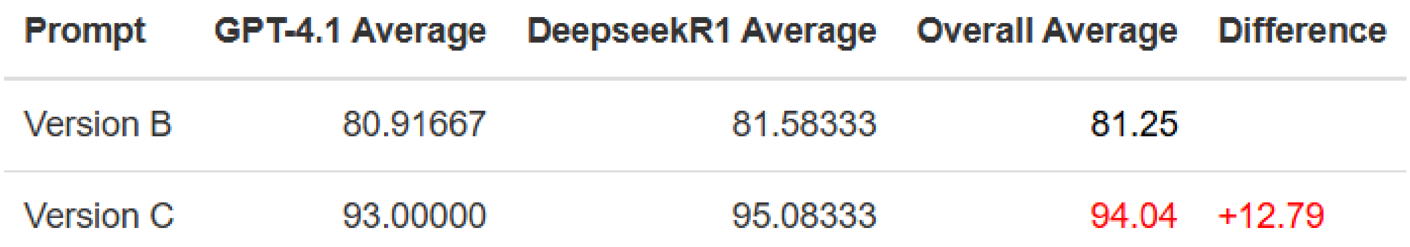


	Version B					Version C					
Example	**Accuracy**	**Policy**	**Actionability**	**Structure**	**Average**	**Accuracy**	**Policy**	**Actionability**	**Structure**	**Average**	**Difference**
Example 1	83	68	75	90	79.00	**93**	**88**	**92**	**96**	92.25	+13.25
Example 2	84	75	78	88	81.25	**92**	**90**	**91**	**95**	92.00	+10.75
Example 3	85	78	80	87	82.50	**96**	**92**	**94**	**97**	94.75	+12.25



	Version B					Version C					
Example	**Accuracy**	**Policy**	**Actionability**	**Structure**	**Average**	**Accuracy**	**Policy**	**Actionability**	**Structure**	**Average**	**Difference**
Example 1	**90**	70	75	85	80.00	90	**95**	**90**	**97**	93.00	+13.00
Example 2	85	75	75	85	80.00	**95**	**96**	**92**	**97**	95.00	+15.00
Example 3	92	80	82	85	84.75	**98**	**97**	**96**	**98**	97.25	+12.50

Figures 2 and 3: Enhanced side-by-side comparison scores from GPT-4.1 and DeepseekR1.



Prompt	GPT-4.1 Average	DeepseekR1 Average	Overall Average	Difference
Version B	80.91667	81.58333	81.25	
Version C	93.00000	95.08333	94.04	+12.79

Figure 4: Overall mean score of different prompts provided by the LLMs.

Once the optimal pattern and model choices are found, we then proceed to the design of system-level prompt engineering. 3 versions of the system-level prompt were designed (Version A, B, C). For testing purposes, 3 types of sample application files were also designed, covering 3 common situations of compliance with human ethics policies: Fully compliant, Marginally compliant, and Non-compliant. Responses generated with each prompt with each test sample file were cross-compared. ChatGPT-4.1 and DeepseekR1 were also used simultaneously to carry out a double blind experiment on different prompt versions (Figures 2 to 4). The reason for this choice is due to GPT-4.1 and DeepseekR1 being the most mainstream LLMs currently, and the inference ability of those two models is among the top of the industry. The evaluation metrics for the prompts are: Compliance/Risk identification, Policy traceability and justification, recommendations quality, and structure and clarity. During evaluation, version A is significantly outperformed by versions B and C, so scoring using LLMs focused on the other two prompt versions. Prompt C was finally chosen as the most optimal prompt.

Finally, a prototype user interface was implemented using Python's Streamlit library. This deployable interface enables basic Q&A with the RAG AI assistant, allowing users to type their questions and upload documents in PDF or Word format.

After the experimental PoC has been developed and evaluated, the project can progress into post-PoC stages, which include system maintenance, possible knowledge base updates, integration with the University infrastructure, and scaling the solution for broader institutional use.

4.5 FUTURE STEPS AND FEASIBILITY PATHWAY

The current PoC successfully demonstrates the capability of a Retrieval-Augmented Generation (RAG) based AI assistant to address researcher queries regarding compliance with UoA and NZ human ethics policies. It validates the core technical feasibility of the approach, including document ingestion, semantic retrieval, and natural language generation.

However, several components of the architecture are still in an experimental stage and require further development before the system can be reliably adopted for wider use within the University environment.

Future steps after PoC:

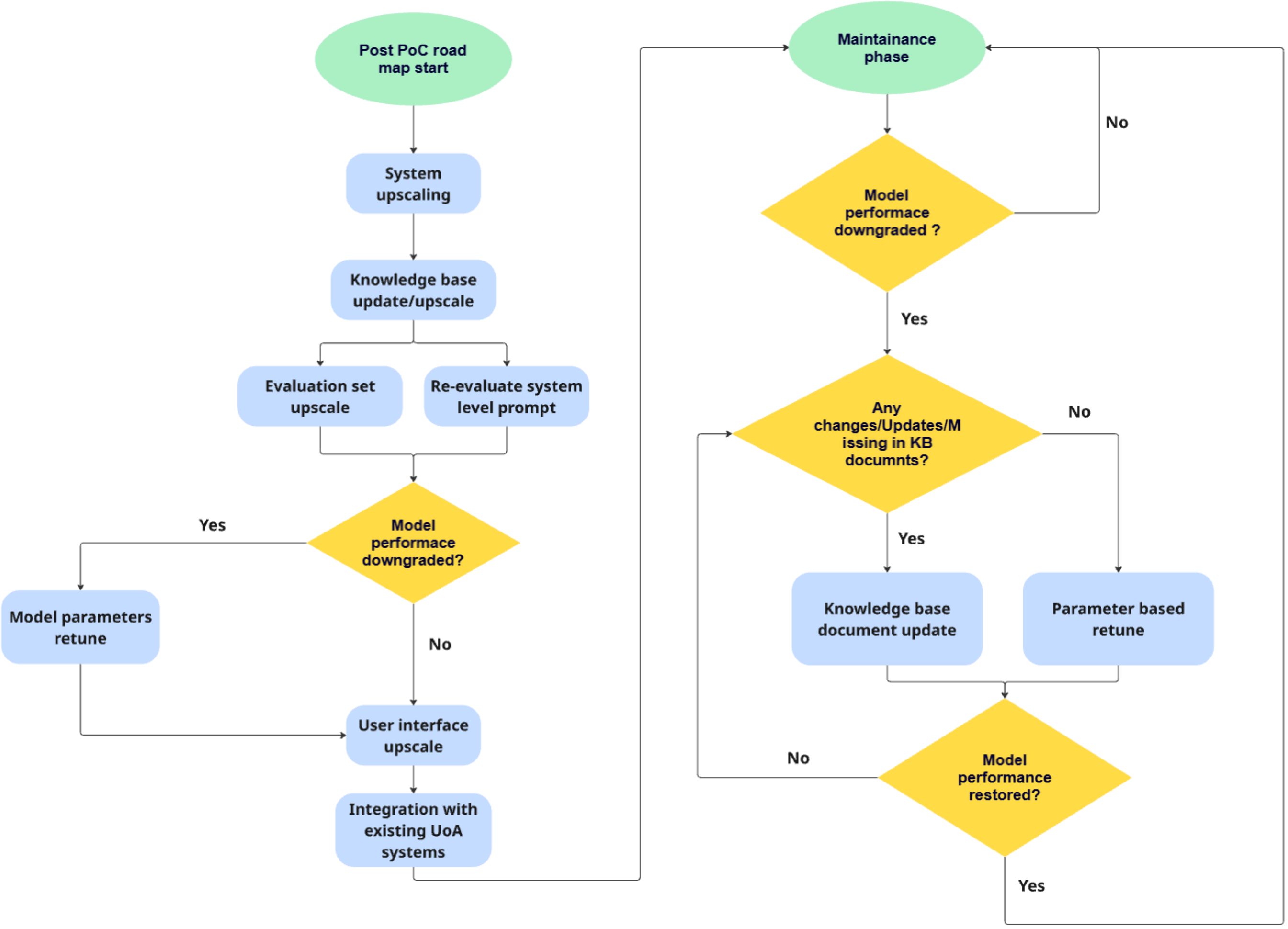

Post PoC road map start
System upscaling
Knowledge base update/upscale
Evaluation set upscale
Re-evaluate system level prompt
Model performace downgraded?
Yes
Model parameters retune
No
User interface upscale
Integration with existing UoA systems
Maintainance phase
No
Model performace downgraded ?
Yes
Any changes/Updates/Missing in KB documnts?
No
Yes
Knowledge base document update
Parameter based retune
No
Model performance restored?
Yes

Figure 5: Post PoC roadmap, including the system upscaling phase process and the maintenance phase process.

The proposed future steps after PoC are shown in the above flowchart (Figure 5), which includes initial scaling-up steps after our experimental deployment to the long-term maintenance phase.

System upscaling

We will begin with scaling our PoC architecture to a greater scale, as well as more robust performance. This involves further development of the following architecture components:

1. Knowledge base and evaluation set: The current knowledge base of our RAG system consists of 9 policy/form documents provided by our stakeholders. In the future scaling-up process, it is important to ensure that the knowledge base covers all possible relevant documents, for example, a list of ATO documents. The optimal quantum of data will depend on the actual scope of the system. Upscaling of this RAG system, which may expand our scope to a wider range of application types, will require extra policy/form/template documents to be added to our knowledge base. Additional files, such as completed anonymised example applications or instructional notes (Slides, guideline documents, etc) could be used to enrich the knowledge base. Notably, if example

applications are being added to the knowledge base, it would be crucial to ensure the variety of the examples to ensure generalisation of answers.

As the knowledge base is scaled up, the evaluation set should also be extended accordingly, and system-level prompts will require a further evaluation according to a wider range of sample situations.

2. Retrieval parameters: The current retrieval parameters are tuned specifically for our 9-document knowledge base. With the knowledge base being scaled up, the model's performance should be re-evaluated and ensure that it still meets the standards. Parameters might require a retune to fit the larger scale source knowledge.

3. User interface: The current user interface serves as a functional prototype for demonstrating the core PoC functionality. It is expected that further development is required to integrate with current UoA systems and enhance usability and accessibility for both users and UoA maintenance members. Example planned implementation includes:
    - User role management to control different access levels
    - Session history allows users to view past queries
    - Integration with existing UoA systems to ensure compliance with UoA policies.

    Such implementation may require a transition from the Streamlit library to more scalable and flexible frameworks such as React or Vue.

4. Integration with existing UoA systems:

    A key step for the PoC to transition from the prototype stage to the deployable stage would be integration with the existing UoA systems. At the current stage, the prototype already shows such potential for its interoperability:

    1) Authentication integration: UoA uses Web Single Sign-On, which is expected to be compatible with integration into the Streamlit framework used for the prototype. Since Streamlit supports user authentication with OpenID Connect (OIDC) [2], which is a commonly used authentication protocol for SSO. While additional details about UoA's specific SSO implementation are needed for a more precise justification, if Streamlit's capabilities prove insufficient, it is possible to switch to other frameworks that offer better support, because the core functionality of this PoC lies in the backend RAG system instead of the user interface.
    2) Back-end compatibility: Docker can be used to containerise the prototype's backend code, allowing a consistent and reproducible deployment across different environments. This is expected to simplify the integration with UoA's existing infrastructure by enabling uniform management and scaling for backend services.
    3) Data handling: The knowledge base of the prototype is hosted online with the IBM Watsonx platform. The same platform can be reused to support the continuous update of data.

Maintainability and security:

After finishing the proposed upscaling steps, the system should be ready for broader deployment within the UoA system. This also moves the central focus onto maintenance and security.

1. Maintainability:

   For continuously monitoring the model performance, research support teams can regularly run experiments on the Watsonx platform and will be intuitively informed of the performance of the model according to 3 critical success criteria (answer correctness, faithfulness, and context correctness). Also, the PoC shows the following potential for simplifying continuous maintenance:

   1) No re-tune requirement: An advantage of RAG-based AI systems is that it doesn't require frequent tuning to hot update the model's accessible information. [3] If during the service period any change in policy occurred, simply updating the policy document in the knowledge base would ensure the model's performance is up-to-date.
   2) Easy document update: The Update of the knowledge base should be simple for research support teams since it is hosted online using the IBM Watsonx platform.
   3) Parameter-driven tuning: Although RAG-based AI systems typically won't require frequent tuning, the Watsonx platform does offer an intuitive parameter-driven tuning approach to handle possible performance downgrades.

2. Security and privacy:

   User data will not be collected by the system to ensure compliance with relevant privacy laws. Also, the system is expected to integrate with the existing UoA SSO system, which will inherit existing institutional roles and ensure all possible data and document access is restricted to UoA-approved personnel.

5 SUMMARY OF MILESTONES & DELIVERABLES

Project Phase	Est. Completion Date	Deliverables
Team formation and scope finalisation	11 May 2025	• Team registration • Choose the interested use case(s) and do background reading • Organise work and roles • Sign the confidentiality agreement • Training on design thinking and/or Watsonx.ai
Design plan and stakeholder analysis	18 May 2025	• Make a final decision on the use case • Start working on PoC - identified stakeholders
PoC development	25 May 2025	• Set up the knowledge base • QA pairs construction • Set up baseline RAG system on Watsonx.ai
Testing and model evaluation	1 June 2025	• Optimise the choice of foundation models and parameter patterns • Ensure model performance metrics (≥70%)

Prototype application implementation	4 June 2025	• Design and evaluate system-level prompt engineering • Implement a prototype user interface
Final presentation and report submission	8 June 2025	• Final project plan • Presentation slides and demo materials

Table 2: Project phases and key deliverables throughout the PoC development cycle.

We set clear goals for our project and began with strong coordination, including stakeholder engagement and an initial design plan. These early efforts gave our team a solid foundation and ensured everyone understood the project's overall direction.

However, we did experience some delays in the initial stages due to uncertainty regarding the specific scope and requirements. Additionally, we had initially planned to fully implement Authority to Operate (ATO) compliance checks for digital tools. Unfortunately, the ATO list is classified as confidential, and we were unable to obtain access during the project. Hence, this functionality could not be completed in the PoC.

Despite these challenges, our team quickly adapted. Once the technical details and objectives were fully defined, we worked efficiently, supported each other in learning new skills, and made use of Watsonx's built-in RAG features and documentation. Therefore, we successfully completed all other planned milestones and delivered a PoC that met the expectations of our stakeholders.

Jiajun Xiao:

- Led RAG project construction and model training
- Authored the RAG engineering evaluation set (JSON format)
- Responsible for prompt engineering: designed evaluation protocols, constructed examples, created prompts, organised double-blind experiments, and analysed prompt experiment results
- Preparation of presentation slides
- Final report review and quality check

Katie Law:

- Completed Sections 2.1, 2.2, 5 and 5.1
- Drafted 2.3 and 6 (benchmarking part)
- Preparation of presentation slides
- Provided iterative feedback and suggestions throughout project development
- Final report review and quality check

Qinxue Feng:

- Summarised all project requirements from lectures

- Completed Sections 2.3, 2.4 and 6
- Preparation of presentation slides
- Provided iterative feedback and suggestions throughout project development
- Final report review and quality check

Nanyuanyang Zhang:

- Generated QA pairs
- Set up the knowledge base and developed the RAG solution using IBM Watsonx
- Built the Streamlit frontend, enabling conversational AI that analyses uploaded documents
- Led the demonstration (demo) session
- Preparation of presentation slides
- Final report review and quality check

Yizheng Xing:

- Wrote and summarised the PoC technical architecture and design plan (Sections 3 and 4)
- Preparation of presentation slides
- Provided iterative feedback and suggestions throughout project development
- Final report review and quality check

Overall, workload was distributed as evenly as possible, with technical, writing, and presentation tasks shared among all members.

5.1 EXPECTED COST BREAKDOWN BY SERVICES

This project was developed using the IBM Cloud Enterprise Platform and Watsonx.ai. The total token usage during the project was approximately 169 million tokens. Most of this was consumed during early rounds of Q&A pairs testing, with sample sizes ranging from 3 to around 270. As the project progressed, we refined the process to better match the knowledge base size, which limits the number of test questions to 24, and ran the final three optimised experiments with 25 million tokens in total.

In addition to standard model testing, we also conducted prompt experiments. These involved adjusting instruction styles and question structure to improve retrieval accuracy, answer clarity, and contextual alignments. All these measurements contributed to our token use and model performance.

Average token usage estimate:

On average, each experimental run used about 13 million tokens (168 million tokens / 13 rounds ≈ 13 million).

At an estimated cost of $0.10 per 1 million tokens [4], each testing round would cost about 1.3 USD (13 million tokens / 1 million x $0.10).

For future development, maintaining access to Watsonx.ai credits or compute tokens will be important. More efficient testing methods may reduce overall usage.

6 ACCEPTANCE

The PoC system has fulfilled most of the success criteria defined before based on the performance of chosen model settings (Foundation model choices, retrieval methods, etc) and optimal prompt engineering version. The generated output of our prototype application is compared against our list of success criteria, and acceptance reasons are justified below.

The screenshots used in this section are drawn from three different examples, each demonstrating our PoC's ability to identify and respond to distinct ethical concerns. These include issues related to data privacy, emotional distress, and the protection of vulnerable populations. Together, they illustrate the tool's versatility and alignment with multiple official guidelines (NEAC, AHREC, UAHPEC, Te Ara Tika).

IBM Watsonx.ai Model Evaluation

To ensure the Human Adviser PoC met our defined success criteria, we benchmarked the tool across three key quantitative metrics using the IBM Watsonx.ai platform:

- Answer correctness: 73.3%
- Objective faithfulness: 71.7%
- Contextual accuracy: 73.0%

All three metrics exceeded the predefined threshold of 70%. This demonstrates that the selected model and prompt design together provide answers that are accurate, stay true to the original documents, and match the context of each question.

LLM-Based Ethics Evaluation for Prompt Engineering

We evaluated our system through a structured rubric-based assessment, averaging the scores from two independent LLMs across four criteria: Compliance & Risk Identification, Policy Traceability & Justification, Actionability of Recommendations, and Structure & Clarity.

Example	**Accuracy**	**Policy**	**Actionability**	**Structure**	**Average**
Example 1	91.5	91.5	91	96.5	92.63
Example 2	93.5	93	91.5	96	93.50
Example 3	97	94.5	95	97.5	96.00

Table 3: The average score for each criterion in every example, calculated by averaging the scores given in Figures 2 and 3 for Version 3.

All four criteria scored well above the acceptance threshold of 85% in every example. This demonstrates that the tool consistently provides accurate, policy-aligned, actionable, and clearly structured guidance for ethics applications.

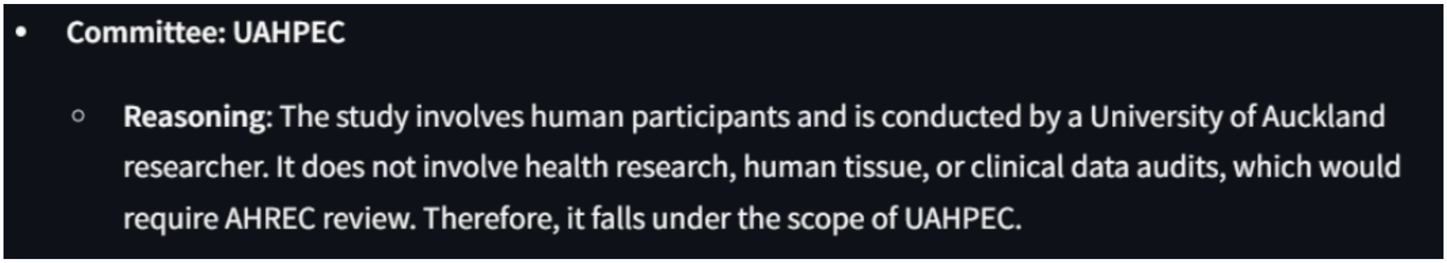
Acceptance Summary

To confirm practical usability and completeness, we tested the system on three custom-designed example applications, each containing different types of issues.

For every applicable success criterion in each example, our PoC successfully identified and flagged all issues requiring revision. This shows that the tool not only performs well on standard benchmarks but also effective in real-world scenarios.

Reduce time spent determining the appropriate ethics committee

Acceptance: Our PoC can assist researchers in determining the appropriate ethics committee based on study characteristics. In the example shown in Figure 6, the system correctly identifies UAHPEC as the suitable committee by reasoning that the research involves human participants, is conducted by a University of Auckland researcher, and does not fall under AHREC's scope, such as health research or clinical data audits.

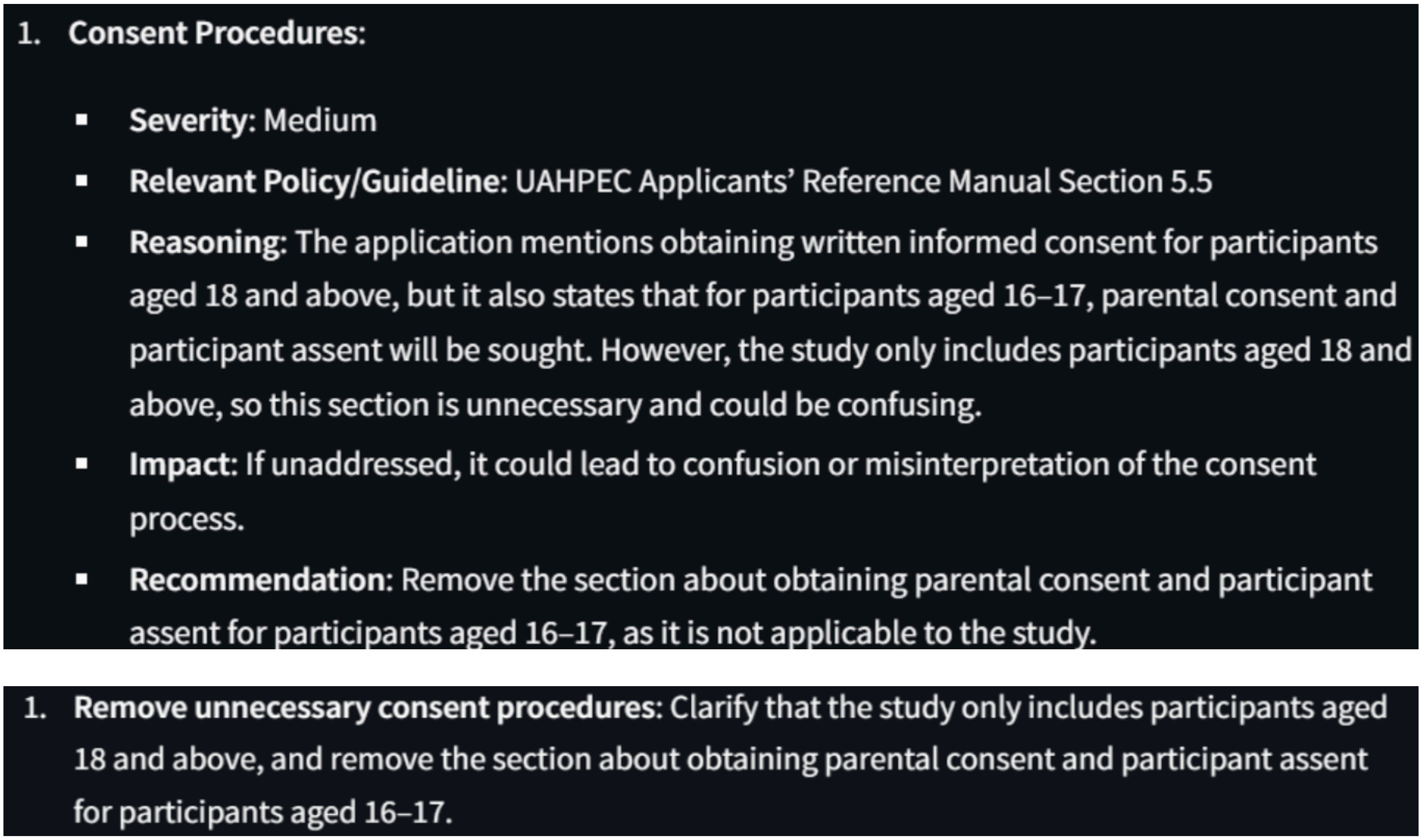


- **Committee: UAHPEC**
  - **Reasoning:** The study involves human participants and is conducted by a University of Auckland researcher. It does not involve health research, human tissue, or clinical data audits, which would require AHREC review. Therefore, it falls under the scope of UAHPEC.

Figure 6: Screenshot of the tool successfully determining the correct committee.

Improve the internal consistency of ethics application materials

Our PoC is capable of detecting inconsistencies within submitted documents. For example, it can identify mismatches between the stated participant criteria and the study design. In the example shown in Figures 7 and 8, the application includes a section on obtaining parental consent for participants aged 16–17, but the study only involves participants aged 18 and above. Our PoC correctly flags this section as unnecessary and potentially confusing. Our PoC also provided a priority action.



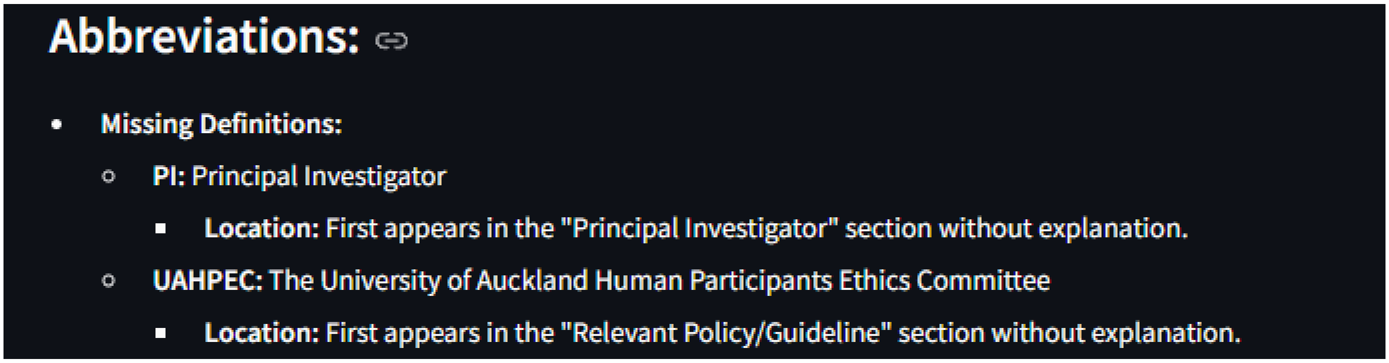
1. **Consent Procedures:**
   - **Severity:** Medium
   - **Relevant Policy/Guideline:** UAHPEC Applicants' Reference Manual Section 5.5
   - **Reasoning:** The application mentions obtaining written informed consent for participants aged 18 and above, but it also states that for participants aged 16–17, parental consent and participant assent will be sought. However, the study only includes participants aged 18 and above, so this section is unnecessary and could be confusing.
   - **Impact:** If unaddressed, it could lead to confusion or misinterpretation of the consent process.
   - **Recommendation:** Remove the section about obtaining parental consent and participant assent for participants aged 16–17, as it is not applicable to the study.

1. **Remove unnecessary consent procedures:** Clarify that the study only includes participants aged 18 and above, and remove the section about obtaining parental consent and participant assent for participants aged 16–17.

Figures 7 and 8: Screenshots of the tool successfully pointing out the internal inconsistency.

Enhance readability and comprehension of application materials

Our PoC can identify and flag unexplained acronyms within ethics applications. In the example shown in Figure 9, the tool detects terms such as "PI" and "UAHPEC" that appear without explanation and highlights their locations in the document. It then provides their full names as "Principal Investigator" and "University of Auckland Human Participants Ethics Committee," along with their first appearances. This improves document clarity and helps ensure that all reviewers, including those without technical backgrounds, can interpret the content accurately.



Abbreviations:

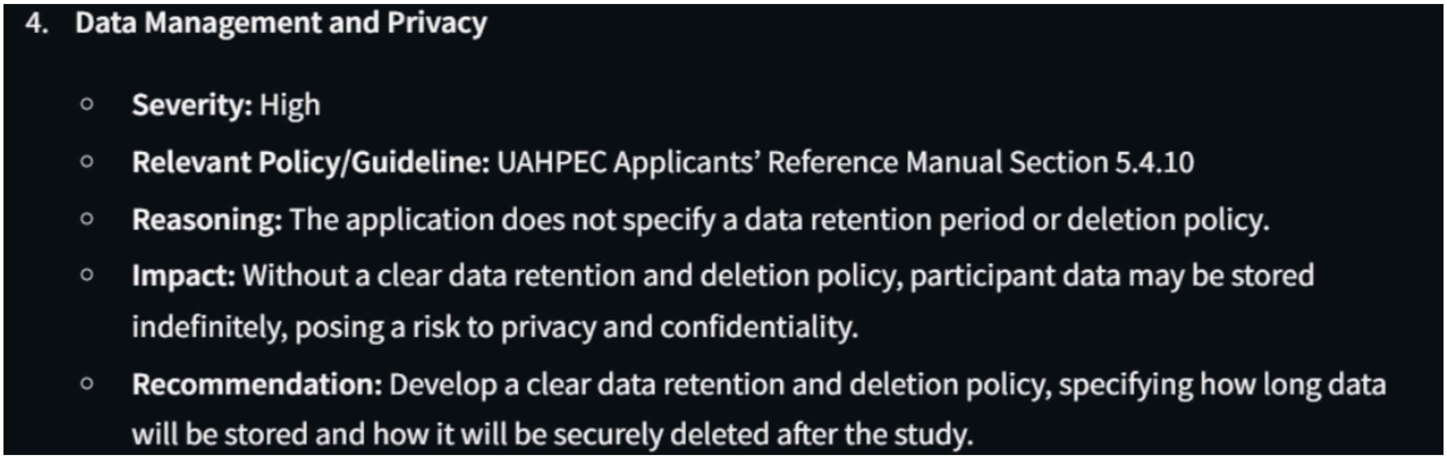
- **Missing Definitions:**
  - **PI:** Principal Investigator
    - **Location:** First appears in the "Principal Investigator" section without explanation.
  - **UAHPEC:** The University of Auckland Human Participants Ethics Committee
    - **Location:** First appears in the "Relevant Policy/Guideline" section without explanation.

Figure 9: Screenshot of the tool successfully detecting abbreviations.

Strengthen the identification and disclosure of ethical risks

Acceptance: Our tool can detect common ethical risks, such as data privacy concerns, emotional distress, and vulnerable populations. The risk prompts are based on guidance from NEAC, UAHPEC, and AHREC to ensure alignment with official requirements.

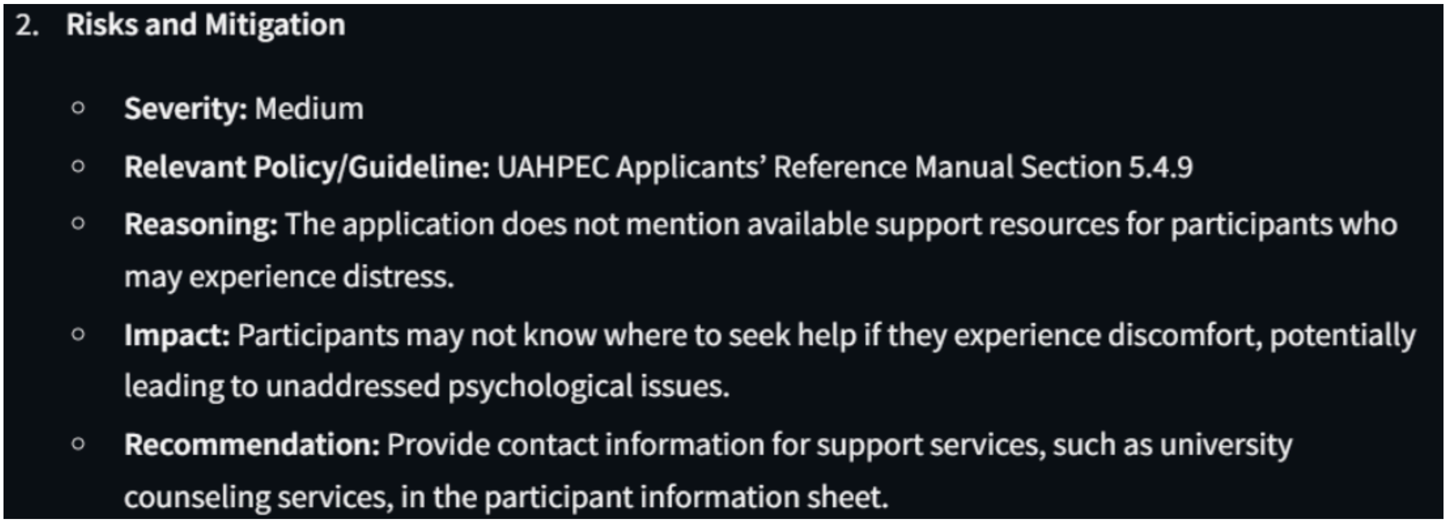
An example related to data privacy concerns is shown in Figure 10.



4. **Data Management and Privacy**
    - **Severity:** High
    - **Relevant Policy/Guideline:** UAHPEC Applicants' Reference Manual Section 5.4.10
    - **Reasoning:** The application does not specify a data retention period or deletion policy.
    - **Impact:** Without a clear data retention and deletion policy, participant data may be stored indefinitely, posing a risk to privacy and confidentiality.
    - **Recommendation:** Develop a clear data retention and deletion policy, specifying how long data will be stored and how it will be securely deleted after the study.

Figure 10: Screenshot of the tool detecting data privacy concerns.

An example related to emotional distress is shown in Figure 11.



2. **Risks and Mitigation**
    - **Severity:** Medium
    - **Relevant Policy/Guideline:** UAHPEC Applicants' Reference Manual Section 5.4.9
    - **Reasoning:** The application does not mention available support resources for participants who may experience distress.
    - **Impact:** Participants may not know where to seek help if they experience discomfort, potentially leading to unaddressed psychological issues.
    - **Recommendation:** Provide contact information for support services, such as university counseling services, in the participant information sheet.

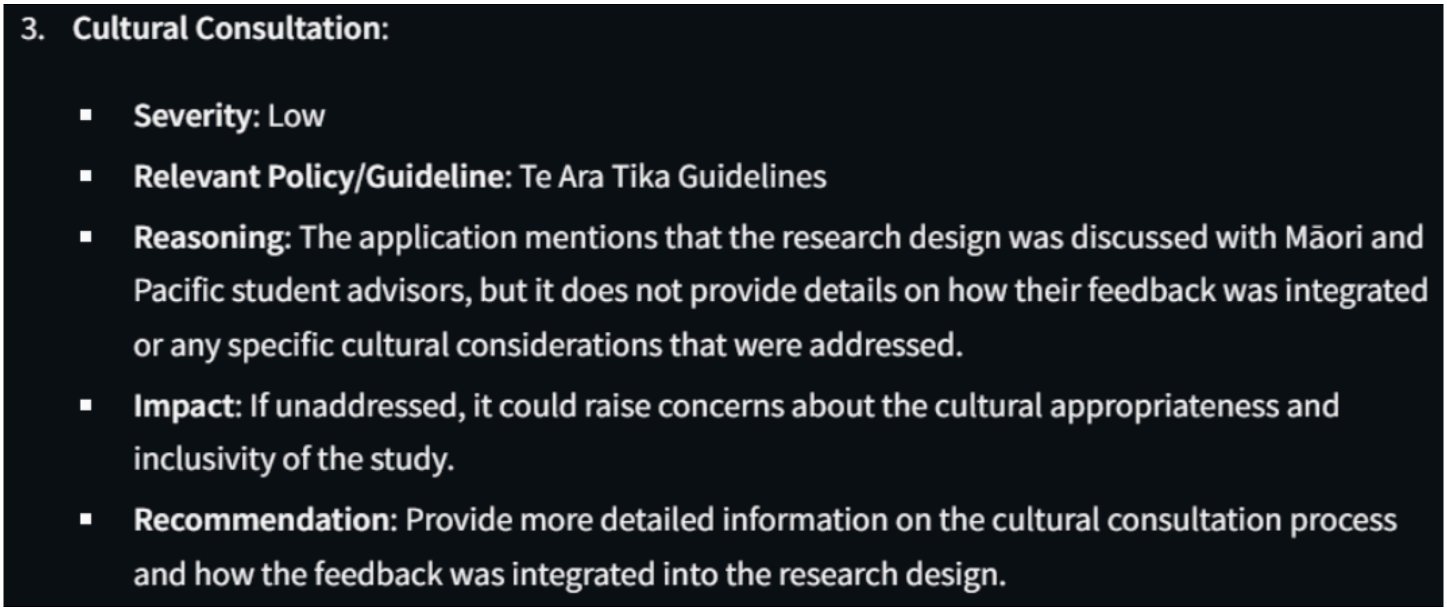
Figure 11: Screenshot of the tool detecting emotional distress.

Ensure compliance of digital tools with data privacy requirements

The success criterion here is whether our PoC can verify that all digital tools listed in an application comply with the University's Authority to Operate (ATO) list. However, since the ATO list was not provided as part of this assignment and is not publicly available, our PoC is currently unable to implement this functionality.

Improve the completeness of application forms and the integration of Te Ara Tika principles

Acceptance: Our PoC is capable of identifying incomplete or missing sections in ethics application forms, particularly those related to Te Ara Tika principles. In the example shown in Figure 12, the system flagged the absence of consideration for Māori, Pacific, or other cultural group perspectives. This demonstrates the PoC's ability to detect culturally relevant gaps and prompt researchers to address them, ensuring alignment with both institutional policies and cultural expectations.



3. **Cultural Consultation:**
   - **Severity**: Low
   - **Relevant Policy/Guideline**: Te Ara Tika Guidelines
   - **Reasoning:** The application mentions that the research design was discussed with Māori and Pacific student advisors, but it does not provide details on how their feedback was integrated or any specific cultural considerations that were addressed.
   - **Impact**: If unaddressed, it could raise concerns about the cultural appropriateness and inclusivity of the study.
   - **Recommendation:** Provide more detailed information on the cultural consultation process and how the feedback was integrated into the research design.

Figure 12: Screenshot of the tool identifying the absence of cultural considerations.

Ensure future adaptability to updated ethics guidelines and documents

Acceptance: Our PoC supports timely updates of new or revised ethics documents, ensuring that the system remains compliant with evolving national standards, such as NEAC and HRC, as well as the University of Auckland's internal policies.

This is achieved through a retrieval-augmented generation (RAG) approach, which allows the system to reference an updatable document store during response generation. In our current prototype, the document store is updated manually. This setup enables us to replace or add new policy documents as needed, allowing the tool to remain aligned with the latest ethical requirements without retraining the underlying model.

7 RESOURCES & REFERENCES

7.1 REFERENCES

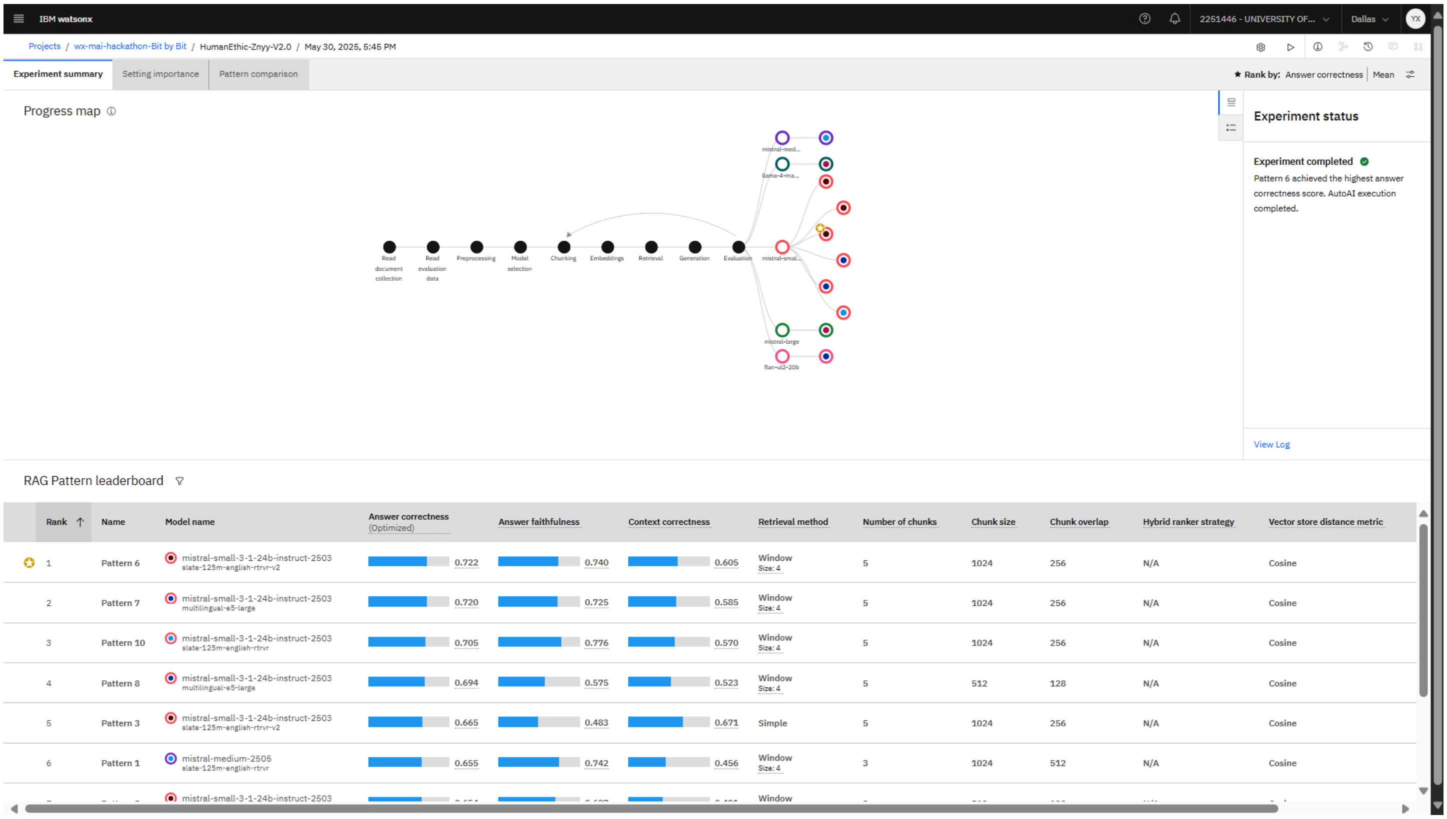
[1] IBM, *Customizing RAG experiment settings*, IBM Documentation, 2024. [Online]. Available: https://www.ibm.com/docs/en/watsonx/saas?topic=autoai-customizing-rag-experiment-settings

[2] Streamlit, *Authentication - Streamlit Docs*, Streamlit, 2024. [Online]. Available: https://docs.streamlit.io/develop/concepts/connections/authentication

[3] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 9459–9474.

[4] "IBM Watsonx.ai | Pricing." https://www.ibm.com/products/watsonx-ai/pricing

7.2 APPENDICES



Projects / wx-mai-hackathon-Bit by Bit / HumanEthic-Znyy-V2.0 / May 30, 2025, 5:45 PM

Experiment status

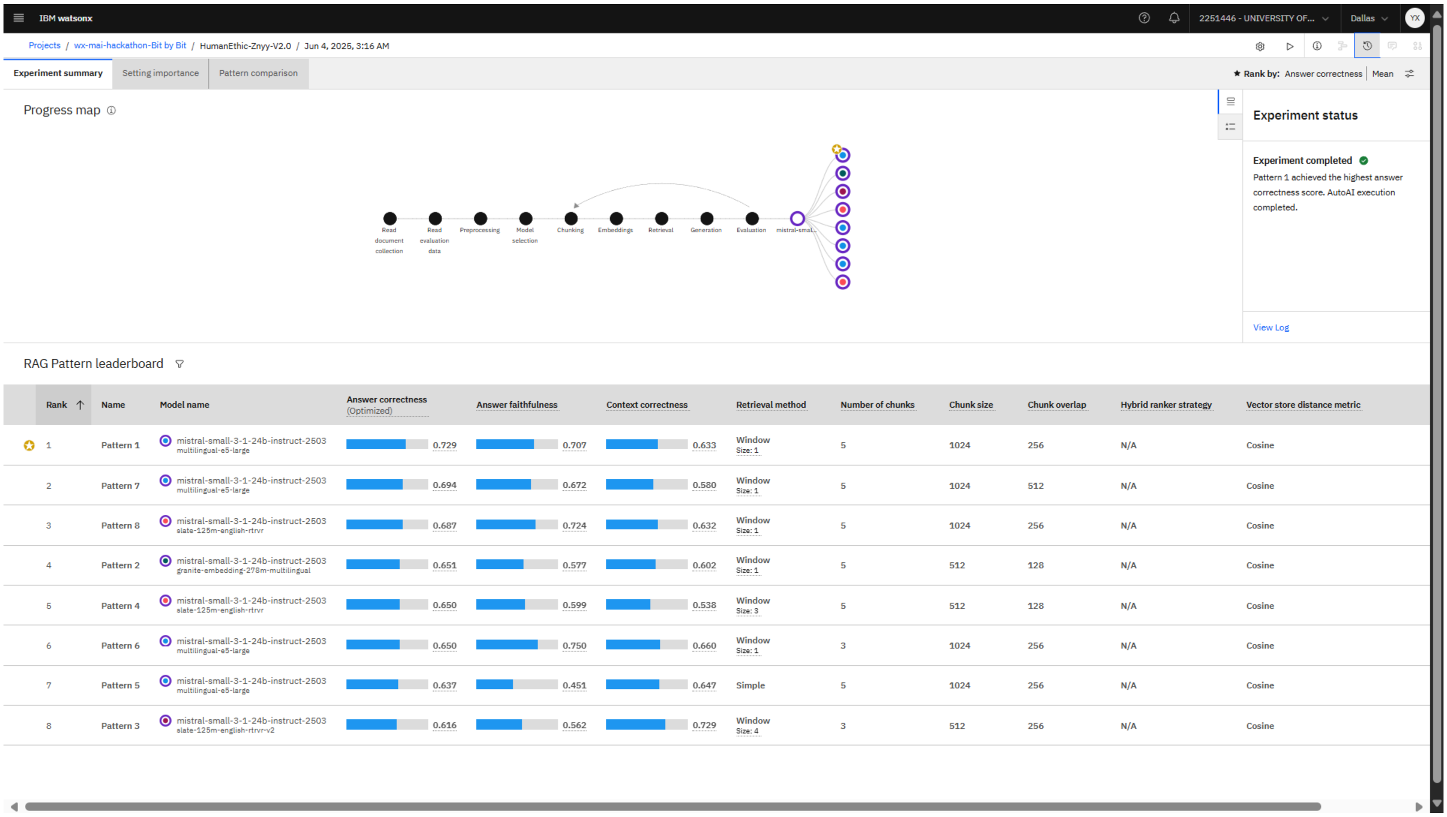
Experiment completed

Pattern 6 achieved the highest answer correctness score. AutoAI execution completed.

RAG Pattern leaderboard

	Rank	Name	Model name	Answer correctness (Optimized)	Answer faithfulness	Context correctness	Retrieval method	Number of chunks	Chunk size	Chunk overlap	Hybrid ranker strategy	Vector store distance metric
✪	1	Pattern 6	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr-v2	0.722	0.740	0.605	Window Size: 4	5	1024	256	N/A	Cosine
	2	Pattern 7	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.720	0.725	0.585	Window Size: 4	5	1024	256	N/A	Cosine
	3	Pattern 10	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr	0.705	0.776	0.570	Window Size: 4	5	1024	256	N/A	Cosine
	4	Pattern 8	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.694	0.575	0.523	Window Size: 4	5	512	128	N/A	Cosine
	5	Pattern 3	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr-v2	0.665	0.483	0.671	Simple	5	1024	256	N/A	Cosine
	6	Pattern 1	mistral-medium-2505 slate-125m-english-rtrvr	0.655	0.742	0.456	Window Size: 4	3	1024	512	N/A	Cosine
			mistral-small-3-1-24b-instruct-2503				Window					

Figure A: Before iteration 1, the initial experiment of foundation models.



Projects / wx-mai-hackathon-Bit by Bit / HumanEthic-Znyy-V2.0 / Jun 4, 2025, 3:16 AM

★ Rank by: Answer correctness | Mean

Progress map

Experiment status

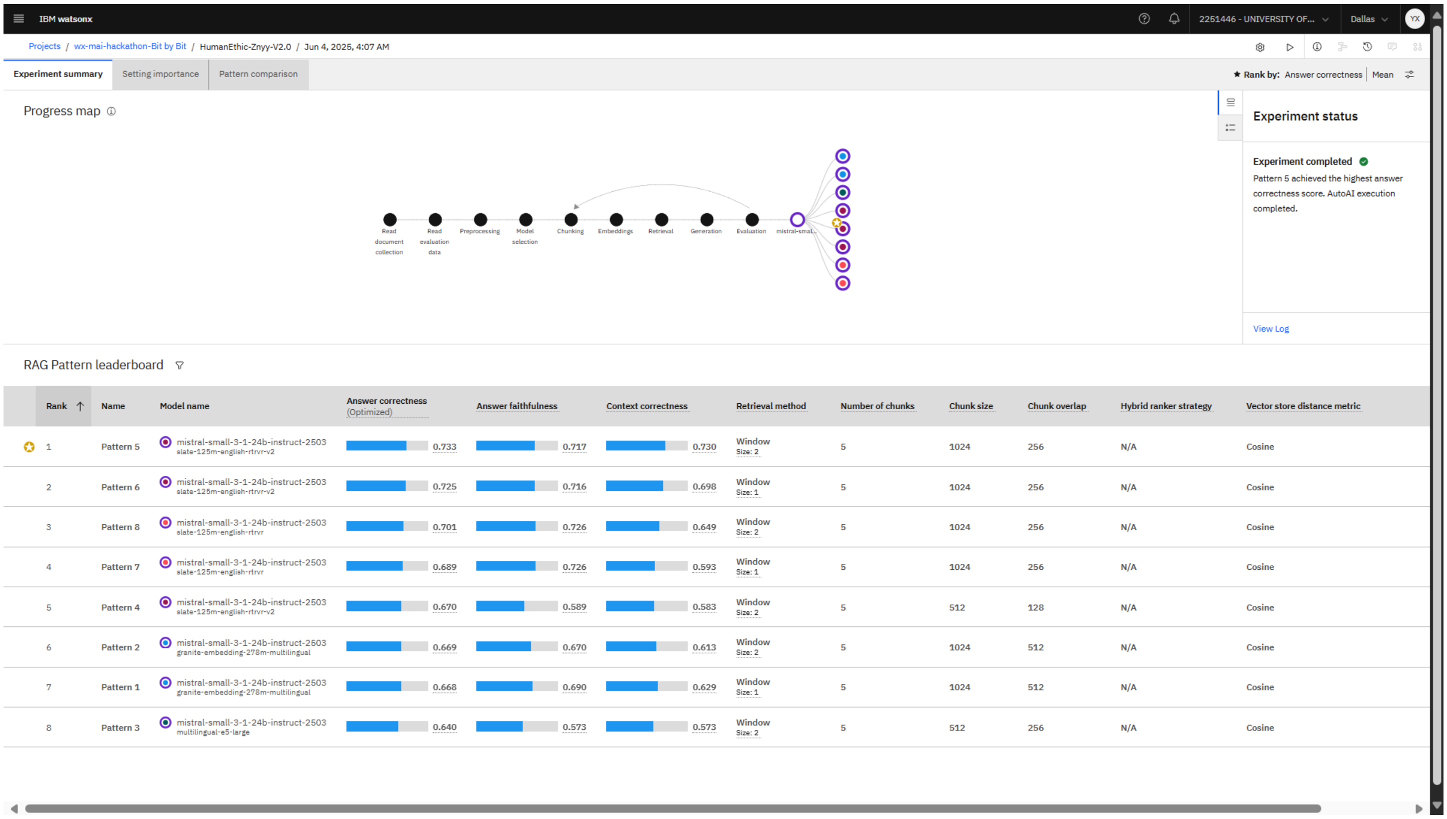
Experiment completed

Pattern 1 achieved the highest answer correctness score. AutoAI execution completed.

RAG Pattern leaderboard

	Rank	Name	Model name	Answer correctness (Optimized)	Answer faithfulness	Context correctness	Retrieval method	Number of chunks	Chunk size	Chunk overlap	Hybrid ranker strategy	Vector store distance metric
★	1	Pattern 1	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.729	0.707	0.633	Window Size: 1	5	1024	256	N/A	Cosine
	2	Pattern 7	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.694	0.672	0.580	Window Size: 1	5	1024	512	N/A	Cosine
	3	Pattern 8	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr	0.687	0.724	0.632	Window Size: 1	5	1024	256	N/A	Cosine
	4	Pattern 2	mistral-small-3-1-24b-instruct-2503 granite-embedding-278m-multilingual	0.651	0.577	0.602	Window Size: 1	5	512	128	N/A	Cosine
	5	Pattern 4	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr	0.650	0.599	0.538	Window Size: 3	5	512	128	N/A	Cosine
	6	Pattern 6	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.650	0.750	0.660	Window Size: 1	3	1024	256	N/A	Cosine
	7	Pattern 5	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.637	0.451	0.647	Simple	5	1024	256	N/A	Cosine
	8	Pattern 3	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr-v2	0.616	0.562	0.729	Window Size: 4	3	512	256	N/A	Cosine

Figure B: After major iterations 1 and 2, experiment with the single limited model and enriched QA pairs.



Experiment status

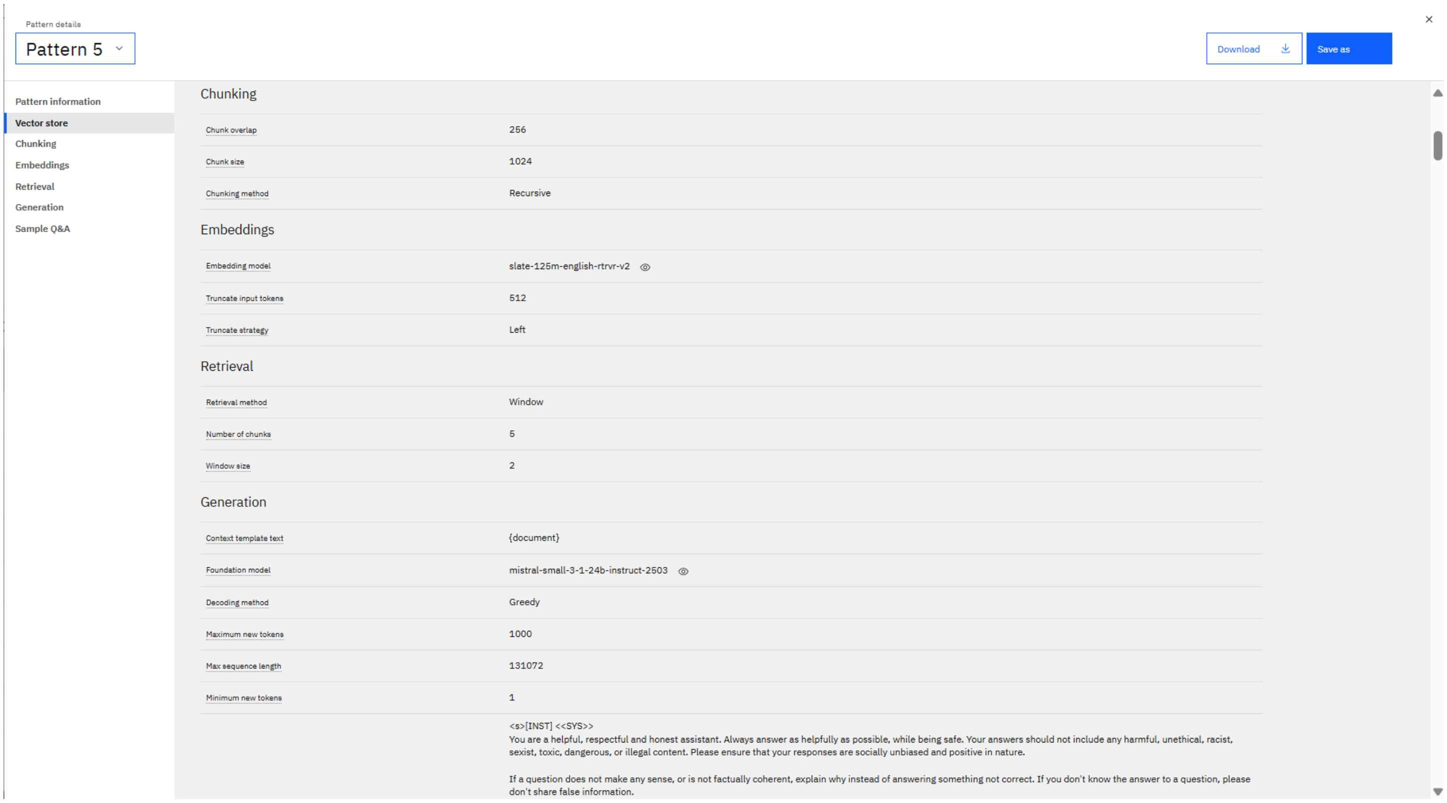
Experiment completed ✓

Pattern 5 achieved the highest answer correctness score. AutoAI execution completed.

RAG Pattern leaderboard

	Rank ↑	Name	Model name	Answer correctness (Optimized)	Answer faithfulness	Context correctness	Retrieval method	Number of chunks	Chunk size	Chunk overlap	Hybrid ranker strategy	Vector store distance metric
★	1	Pattern 5	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr-v2	0.733	0.717	0.730	Window Size: 2	5	1024	256	N/A	Cosine
	2	Pattern 6	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr-v2	0.725	0.716	0.698	Window Size: 1	5	1024	256	N/A	Cosine
	3	Pattern 8	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr	0.701	0.726	0.649	Window Size: 2	5	1024	256	N/A	Cosine
	4	Pattern 7	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr	0.689	0.726	0.593	Window Size: 1	5	1024	256	N/A	Cosine
	5	Pattern 4	mistral-small-3-1-24b-instruct-2503 slate-125m-english-rtrvr-v2	0.670	0.589	0.583	Window Size: 2	5	512	128	N/A	Cosine
	6	Pattern 2	mistral-small-3-1-24b-instruct-2503 granite-embedding-278m-multilingual	0.669	0.670	0.613	Window Size: 2	5	1024	512	N/A	Cosine
	7	Pattern 1	mistral-small-3-1-24b-instruct-2503 granite-embedding-278m-multilingual	0.668	0.690	0.629	Window Size: 1	5	1024	512	N/A	Cosine
	8	Pattern 3	mistral-small-3-1-24b-instruct-2503 multilingual-e5-large	0.640	0.573	0.573	Window Size: 2	5	512	256	N/A	Cosine

Figure C: After major iteration 3, experiments with more generalised QA samples.



Pattern details

Pattern 5

Chunking

Chunk overlap	256
Chunk size	1024
Chunking method	Recursive

Embeddings

Embedding model	slate-125m-english-rtrvr-v2
Truncate input tokens	512
Truncate strategy	Left

Retrieval

Retrieval method	Window
Number of chunks	5
Window size	2

Generation

Context template text	{document}
Foundation model	mistral-small-3-1-24b-instruct-2503
Decoding method	Greedy
Maximum new tokens	1000
Max sequence length	131072
Minimum new tokens	1

<s>[INST] <<SYS>>
You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.

If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.

Figure D: Best parameter settings and parameter pattern chosen from the experiment after major iteration 3.