ACADEMIC CITY
UNIVERSITY

INTRODUCTION TO ARTIFICIAL INTELLIGENCE (CE4143)

PROJECT-BASED EXAMINATION DOCUMENTATION

NAME: MAWUENA KOMLA ACKOTIA

ROLL NUMBER: 10211100337

LEVEL: 400

PROGRAM: COMPUTER ENGINEERING

LECTURER: MR. GODWIN NTOW DANSO

PROJECT OVERVIEW

This Streamlit web application is designed as a modular, interactive interface to demonstrate real-world applications of Artificial Intelligence (AI) and Machine Learning (ML) concepts. It allows users to explore:

- **Regression Analysis** (with metrics & prediction)
- **K-Means Clustering** (with elbow method, 2D/3D visualization)
- **Neural Network Classifier** (password strength detection)
- **LLM-Powered Question Answering System** (RAG with Gemini API)

It supports various data preprocessing options, model configurations, and visualization tools, all integrated into a single, user-friendly interface.

TECHNOLOGIES USED

Python - programming language used

Streamlit – interactive web UI

Scikit-learn – ML models and preprocessing

TensorFlow/Keras – deep learning models

FAISS + LangChain + SentenceTransformers – for LLM-based document Q&A

Google Generative AI (Gemini API) – LLM backend

PDFPlumber – text extraction from PDFs

Matplotlib, Seaborn – data visualization

APPLICATION STRUCTURE

The app is broken into five main sections, accessible via a side navigation bar:

- Home Page
- Regression Page
- Clustering Page
- Neural Network Page
- Large Language Model (LLM) Page

HOME PAGE

The Home Page displays the course details as well as the details of the student.

REGRESSION PAGE

This section enables users to perform linear regression modeling on a custom dataset (CSV upload). The following is a detailed breakdown:

1 File Upload and Preview

- Users upload a CSV file which is parsed using pandas.
- A preview of the dataset (top 10 rows) is shown using st.write(df.head(10)).

2 Data Preprocessing Options

2.1 Handling Missing Values

Users choose how to handle missing values via a dropdown:

- Keep as-is (default)
- Drop rows with missing values
- Impute with median (numeric columns only)

2.2 Feature Scaling

Checkbox allows optional scaling via StandardScaler. This standardizes numeric features to a mean of 0 and std dev of 1 — crucial for many ML models.

3 Target and Feature Selection

- The target column is limited to ‘price’ or ‘area’, since they are the only continuous variable columns in the dataset.
- Users can select features (independent variables) manually from the remaining columns.

4 Data Encoding

Categorical features (string/object types) are automatically encoded using LabelEncoder.

5 Train-Test Split

- Adjustable via a slider (10–50%)
- Sample counts for training/testing sets are calculated and displayed

6 Model Training

- The system trains a LinearRegression model from sklearn.
- It predicts on both training and test datasets.

7 Performance Metrics

Metrics include:

- **Mean Squared Error (MSE):** The average squared difference between predicted and actual values, often used to evaluate the accuracy of regression models.
- **R² Score**: A measure of the proportion of the variance in the dependent variable that is predictable from the independent variable(s), ranging from 0 (no fit) to 1 (perfect fit).
- **Mean Absolute Error (MAE)**: The average absolute difference between predicted and actual values, providing a measure of the average magnitude of errors in a regression model.
- **Pearson Correlation**: A statistical measure of the linear relationship between two continuous variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

8 Visualization

- Two regression plots (actual vs predicted) are shown for both training and test sets using seaborn.regplot.
- Red line of best fit is included for visual understanding.

9 Prediction Section

Users can:

- Enter new feature values via forms
- Auto-handle encoding and scaling
- Predict a target value (e.g., price)
- Get a real-time prediction using the trained model

CLUSTERING PAGE

This section enables users to perform linear regression modeling on a custom dataset (CSV upload). The following is a detailed breakdown:

1 File Upload and Encoding Fallbacks

- CSV file is uploaded
- Multiple encoding strategies are tried (utf-8, latin1, iso-8859-1, etc.)
- If all fail, the python CSV engine is used

This prevents format-related crashes and supports real-world messy data.

2 Dataset Preview and Cleaning

- Displays shape and head of dataset
- Missing value options:
  - Drop rows
  - Fill numeric with median, categorical with mode
  - Fill all with zero

3 Feature Selection

- Only numeric columns are shown for clustering
- At least 2 features must be selected
- Users can select 2D or 3D combinations for visualization

4 Feature Scaling

- Optional StandardScaler normalization
- Recommended for clustering to ensure features contribute equally

5 Elbow Method for Optimal Clusters

- Computes WCSS (within-cluster sum of squares) for k=1 to 10
- Generates an elbow plot to help user choose the ideal number of clusters

6 K-Means Clustering

- Runs clustering with user-selected k value
- Results are appended as a new Cluster column
- Cluster centers (centroids) are calculated and shown

7 Visualization

Metrics include:

- 2D scatter plot if 2 features are selected
- 3D scatter plot using matplotlib for 3 features
- Centroids are plotted in red (X marker)
- Axes formatted for readability

8 Scatter Statistics

- Displays:
    - Mean values per cluster for selected features
    - Number of points in each cluster
- Bar chart for cluster distribution with value labels

9 Download Button

Users can export the clustered data in a csv file format, with the Cluster column added.

NEURAL NETWORK CLASSIFIER PAGE

This page enables the creation and training of customizable neural networks using TensorFlow/Keras. It targets classification tasks (e.g., password strength prediction).

1 Dataset Upload and Preprocessing

- Users upload a CSV
- File is auto-cleaned (missing rows dropped)
- If a password column exists, multiple features are engineered:
    - Length
    - Upper/lower/digit/special character count
    - Unique characters
    - Entropy (Shannon information content)

2 Feature and Target Selection

- Only allows strength as target if present (for password classifier)
- Uses LabelEncoder if target is categorical
- Allows multi-feature selection for training

3 Preprocessing

- Categorical features encoded
- All features are scaled using StandardScaler
- Split into training and testing datasets with train_test_split

4 Neural Network Configuration

- Class Weights auto-calculated to handle imbalanced classes
- Users can customize:
    - Epochs
    - Batch size
    - Learning rate
    - Number of hidden layers and their sizes

5 Model Training

- A custom Callback class shows:
    - Epoch-wise progress
    - Live training/validation loss and accuracy plots

6 Evaluation

- Model performance shown using a detailed classification report (precision, recall, F1-score)
- Predictions are compared against the test set

7 Prediction Options

- Users can:
    - Upload test data for batch prediction
    - Manually enter a password for real-time strength classification
- Each password prediction includes:
    - Strength label
    - Confidence score
    - Feature breakdown in tabular format

LARGE LANGUAGE MODEL WITH RETRIEVAL AUGMENTED GENERATION

1 Document Processing Phase

1.1 File Ingestion

The system supports structured and unstructured documents with two primary file types:

- PDF Documents: Processed using the `pdfplumber` library, which enables precise extraction of text from each page, preserving the spatial structure where possible. This ensures that multi-page documents with varied formatting (headers, footers, paragraphs) are parsed reliably.

- CSV Files: Handled via the `pandas` library. To maintain the tabular integrity of data, the system converts CSVs into Markdown table format using `DataFrame.to_markdown()`. This transformation preserves row-column structure while making it easily digestible by LLMs and chunking algorithms.

Robust error handling ensures that unsupported or malformed files are gracefully rejected with meaningful feedback to the user.

1.2 Text Chunking

Chunking is essential for splitting large documents into manageable, semantically coherent units that can later be embedded and searched efficiently.

The system uses LangChain's RecursiveCharacterTextSplitter, a smart splitter that prioritizes natural language boundaries (e.g., double newlines \n\n, single newline \n, sentence ends like . , etc.).

The following Chunk Parameters were used in the text chunking phase.

- Chunk Size: 1500 characters
- Overlap: 300 characters

The overlap ensures continuity across chunk boundaries, minimizing the risk of losing context during segmentation. This is crucial for retrieving relevant answers, especially when a query spans two adjacent chunks.

1.3 Vector Embedding

After chunking, each segment is transformed into a vector representation (embedding) for efficient semantic search.

- Embedding Model: `all-MiniLM-L6-v2` from SentenceTransformers
    - Output: 384-dimensional embeddings
    - Optimized for CPU and fast enough for real-time document embedding
    -
- Batch Processing: Embeddings are computed in batches of 32 to optimize memory usage and runtime performance.
- Vector Store: FAISS (Facebook AI Similarity Search)
    - Index Type: IndexFlatIP (Inner Product similarity)
    - Preprocessing: All embeddings are L2-normalized to simulate cosine similarity within the inner product search framework.

This embedding layer is the backbone of the semantic retrieval engine and is both scalable and performant for medium-sized documents.

2 Query Processing Phase

Once a document has been indexed, users can interact with it through natural language queries. The system processes each query using the same principles as document embedding to ensure consistent semantic comparison.

2.1 Question Handling

A user-entered question is encoded into a 384-dimensional vector using the same SentenceTransformer model used during ingestion.

Vector reshaping is performed to ensure compatibility with FAISS' expected input format (i.e., 2D arrays for batch querying).

2.2 Semantic Search (FAISS Indexing)

The FAISS index is queried using the embedded query vector.

- Top-k Retrieval: Retrieves the top 5 most relevant chunks (k=5), dynamically adjusted if the total number of chunks is fewer than 5.
- Distance Metric: Inner product similarity, interpreted as cosine similarity post L2-normalization.

The system includes bounds checking to ensure invalid indices or out-of-range chunk references do not cause runtime errors or illogical outputs.

2.3 Confidence Calculation

To build user trust and promote transparency, each answer is accompanied by a confidence score. This score is computed using a **three-factor weighted heuristic,** combining multiple dimensions of semantic retrieval quality:

A. Semantic Similarity (50%)

- Computed as the cosine similarity between the query embedding and each retrieved chunk's embedding.
- Uses a robust safety mechanism that clips values and reshapes vectors to prevent dimension mismatches or zero-division errors.

B. Distance Score (30%)

- The original FAISS distance (L2) is normalized and inverted into a [0,1] similarity scale.
- Ensures that closer chunks (i.e., semantically nearer) get higher scores.

C. Coverage Ratio (20%)

- Measures the ratio of validly matched chunks to the total requested (k).
- Penalizes results where too few relevant chunks are found, thereby accounting for potential information loss or low-document relevance.

The **final confidence score** is scaled to a percentage and categorized into:

- High (75–99%)
- Medium (50–74%)
- Low (30–49%)

3 Answer Generation with Gemini

After chunk retrieval and confidence scoring, the system uses Google’s Gemini LLM API (via google.generativeai) to formulate final answers.

- Model Options:
  - gemini-1.5-pro – High-capacity model for accuracy
  - gemini-1.5-flash – Lightweight model for faster response

The query and retrieved context are compiled into a structured prompt. Retry logic is also implemented via the `tenacity` package to handle transient errors.

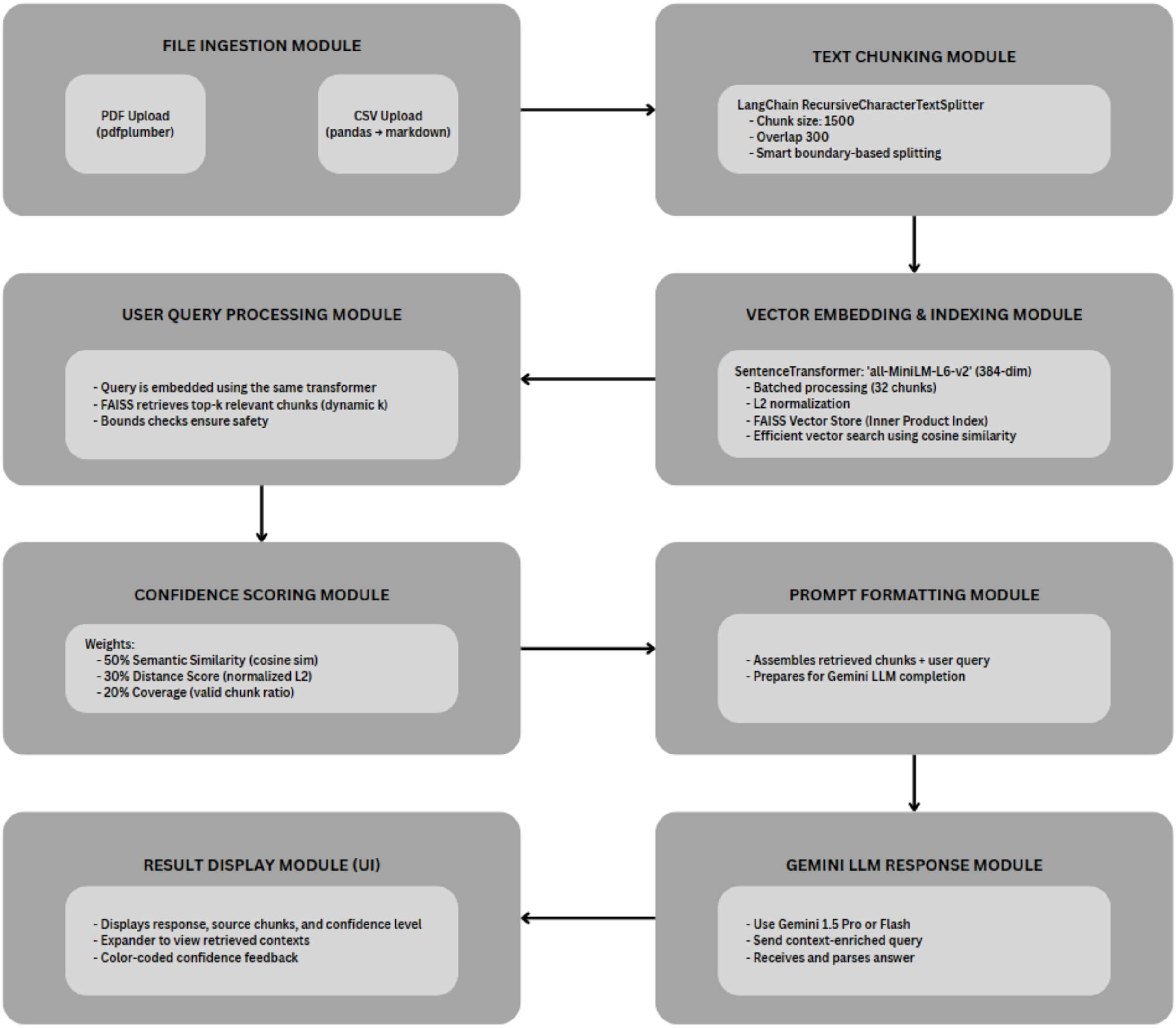
4 Visualization and Source Attribution

The interface displays a color-coded confidence badge (green, yellow, red) depending on the confidence score. Users can also expand a "View Sources" section to read the exact document chunks used to formulate the answer.

5 Additional Features

- API Key Entry: Gemini API key is stored temporarily during session use via st.session_state in Streamlit.
- Upload & RAG Persistence: Once a document is processed, its chunks and index are stored in session memory to avoid redundant recomputation.

6 Diagram of LLM RAG Architecture


FILE INGESTION MODULE
PDF Upload (pdfplumber)
CSV Upload (pandas → markdown)
TEXT CHUNKING MODULE
LangChain RecursiveCharacterTextSplitter
- Chunk size: 1500
- Overlap 300
- Smart boundary-based splitting
USER QUERY PROCESSING MODULE
- Query is embedded using the same transformer
- FAISS retrieves top-k relevant chunks (dynamic k)
- Bounds checks ensure safety
VECTOR EMBEDDING & INDEXING MODULE
SentenceTransformer: 'all-MiniLM-L6-v2' (384-dim)
- Batched processing (32 chunks)
- L2 normalization
- FAISS Vector Store (Inner Product Index)
- Efficient vector search using cosine similarity
CONFIDENCE SCORING MODULE
Weights:
- 50% Semantic Similarity (cosine sim)
- 30% Distance Score (normalized L2)
- 20% Coverage (valid chunk ratio)
PROMPT FORMATTING MODULE
- Assembles retrieved chunks + user query
- Prepares for Gemini LLM completion
RESULT DISPLAY MODULE (UI)
- Displays response, source chunks, and confidence level
- Expander to view retrieved contexts
- Color-coded confidence feedback
GEMINI LLM RESPONSE MODULE
- Use Gemini 1.5 Pro or Flash
- Send context-enriched query
- Receives and parses answer

7 Comparison of LLM RAG Prompt Responses with Chat GPT

Sample Prompt 1

"What is the dress code for male students?"

RAG System's Response to Prompt 1

High confidence: 82%
Answer:

The dress code for male students at Academic City University, as outlined in the provided text, is as follows:

a) Shirts:
- Dress shirts (long or short sleeves) are recommended.
- T-shirts or Lacoste tops are permitted, but must not contain profane language or obscene images.
- For seminars, conferences, presentations, and workshops, regular business casual wear (with or without a jacket) is required.

b) Trousers/Shorts:
- Trousers must be long enough to avoid revealing skin when seated.
- Shorts must be no shorter than 2 inches above the knee.
- Jeans must not be tattered or reveal underwear.

c) Shoes:
- Lace-up leather shoes (clean and polished) are preferred.
- Boots, tennis shoes, Hush Puppies, Crocs, and canvas shoes are allowed.
- Slippers (slides) are NOT allowed.

d) Jewelry & Accessories:

- Belts are preferred but don't need to be leather, however, they shouldn't be worn out.
- Jewelry should be minimal: a watch, one necklace, and a maximum of two rings are allowed.
- Tongue rings are absolutely prohibited.

The reasoning behind these rules is to maintain a professional and respectful environment within the academic and administrative areas of the university. The dress code reflects the university's aim to cultivate future leaders who present themselves appropriately in professional settings. The prohibition of certain items (like slippers, revealing clothing, or clothing with offensive imagery) is intended to uphold a standard of decorum and avoid distractions that might interfere with the learning environment. The allowance of some flexibility (like the acceptance of Crocs or t-shirts under certain conditions) acknowledges individual preferences while maintaining a baseline of appropriate attire.

Chat GPT's Response to Prompt 1

The dress code for male students at Academic City is outlined under the **"ACITY Grooming Standards"** section of the handbook. Here are the key points:

Dress Code for Men (Page 59)

1. **Appearance**: Must be clean, neat, and decent at all times.
2. **Acceptable Attire**:
    - Button-down shirts (short or long sleeves)
    - Polo shirts
    - Slacks or khakis
    - Well-maintained jeans without tears or excessive wear
    - Suits, ties, and formal shoes for presentations and formal occasions
3. **Unacceptable Attire**:
    - Sagging trousers
    - Shorts (unless for sports or in recreational areas)
    - Sleeveless tops (e.g., tank tops or singlets)
    - Clothing with offensive graphics or text
    - Slippers or flip-flops (except in hostels or recreational spaces)
    - Excessively tight clothing

The dress code promotes a professional and respectful campus environment. Violations may lead to disciplinary actions, particularly for repeat offenses.

Let me know if you'd like the women's dress code or violation penalties too.

Sample Prompt 2

"What is the dress code for female students?"

RAG System's Response to Prompt 2

High confidence: 81%
Answer:

The dress code for female students at Academic City University is quite detailed and aims for a professional, modest appearance in academic and administrative areas. It prohibits clothing deemed inappropriate for a professional setting. Here's a breakdown based on the provided text:

Acceptable Attire:

- Skirts, Dresses, and Trousers: Lengths must be no shorter than two inches above the knee and not revealing. Business casual wear (with or without jackets) is required for seminars, conferences, presentations, and workshops.
- Shirts, Tops, Blouses, and Jackets: Casual shirts, dress shirts, sweaters, and suit jackets are acceptable. However, form-fitting, revealing, sleeveless, strapless, or backless tops are prohibited. Crop tops revealing the waist or stomach are also forbidden, as are tops exposing cleavage or nipples due to lack of a bra.
- Jewelry and Accessories: A maximum of two rings (in addition to wedding rings), up to two sets of earrings, and two necklaces are allowed. Tongue rings and waist beads visible above clothing are prohibited. Cosmetics should be moderate and appropriate for skin tone.
- Shoes: Flat shoes, crocs, sandals, or canvas shoes are acceptable. Bathroom slippers and flip-flops are not permitted.

Unacceptable Attire:

- Skirts, Dresses, and Trousers: Miniskirts, sundresses, beach dresses, spaghetti-strap dresses, form-fitting leggings, shorts shorter than two inches above the knee, tight skirts riding halfway up the thigh, and dresses/skirts with splits at or above the knee are all unacceptable.
- Shirts, Tops, Blouses, and Jackets: Form-fitting, revealing, sleeveless, strapless, backless, or crop tops are prohibited. Clothing that exposes cleavage or nipples is also forbidden.
- Shoes: Bathroom slippers and flip-flops are specifically prohibited.

- General: Torn, dirty, or frayed clothing is unacceptable. This includes jeans that are too tight or tattered. Hats and head coverings are generally not acceptable, except for religious or cultural reasons. Clothing with profanity, nudity, semi-nudity, political slogans, advertisements, or suggestive cartoons/drawings is prohibited.

In summary, the dress code for female students emphasizes modesty, professionalism, and avoidance of clothing that could be considered revealing, unprofessional, or distracting in an academic environment. The university clearly intends to cultivate a specific image and expects students to adhere to these standards.

<u>Chat GPT's Response to Prompt 2</u>

The dress code for female students at Academic City is outlined in the **"ACITY Grooming Standards"** section of the student handbook, specifically under **Section 7.2 – For Women** (Page 59). Here's a summary:

Dress Code for Women

1. **Appearance**: Must be clean, modest, and decent at all times.
2. **Acceptable Attire**:
    - Blouses, shirts, or tops with appropriate necklines and lengths
    - Skirts and dresses that are of modest length (at or below the knee)
    - Trousers or slacks that are not overly tight
    - Suits or professional attire for formal presentations or official events
    - Comfortable and appropriate shoes or sandals (excluding flip-flops outside hostels)
3. **Unacceptable Attire**:
    - Tops with plunging necklines or open backs
    - See-through or sheer clothing without appropriate underlayers
    - Micro-mini skirts or very short dresses
    - Shorts (except in hostels or for sports)

    - Excessively tight or revealing clothing
    - Crop tops that expose the midriff
    - Slippers or flip-flops outside residential or recreational areas

The university emphasizes dressing in a manner that reflects respect for self, others, and the learning environment. Repeated violations of the dress code can lead to disciplinary action.

Let me know if you'd like the additional "Other Specific Requirements" or the dress code violation policies.

Sample Prompt 3

"Summarize what this document is about"

RAG System's Response to Question 3

Medium confidence: 71%
Answer:

This document is a section of the Academic City Student Handbook, specifically focusing on academic integrity, referencing guidelines, and student conduct. It does not represent the entire handbook, but rather excerpts detailing important policies and procedures.

The document can be broken down into these key areas:

- Referencing and Citation: A significant portion details proper referencing techniques for academic papers, covering books, journals, and web sources. It emphasizes the importance of complete and consistent referencing to avoid plagiarism and ensure source verification. Specific formatting examples are provided for each source type.

- Student Requirements for Graduation: It outlines the criteria students must meet to graduate, including credit accumulation, minimum GPA, and maintaining acceptable moral conduct.

- Library Policies and Facilities: It describes the library's resources (books, databases, etc.), services (interlibrary loan), and operating procedures, emphasizing its welcoming and supportive environment for all users.

- Code of Conduct and Student Rights/Obligations: This section establishes Academic City's commitment to providing a quality learning environment and outlines student responsibilities. It highlights the importance of adhering to the college's rules and regulations, which are described as promoting the college's mission and protecting the rights of all stakeholders (students, faculty, and staff).

- Specific Rules and Regulations: The excerpt includes a list of specific rules concerning student behavior within the college environment, including rules about food and beverages, littering, public displays of affection, drugs and alcohol, removal of property, and the use of the boardroom.

- Integrity and Ethics Policy: Finally, it introduces the college's commitment to high ethical standards across all activities, encompassing scientific, scholarly, professional, and personal integrity. It emphasizes environmental sustainability, community engagement, and fair treatment of employees.

In short, the provided text is a compilation of excerpts from a student handbook designed to inform students about academic expectations, referencing guidelines, library resources, and the code of conduct necessary for a successful and ethical college experience at Academic City. The page numbers provided (e.g., p. 20, pp. 10-11, p. 52, p. 23) indicate the location of these sections within the larger handbook.

Chat GPT's Response to Prompt 3

The Acity Student Handbook 2024 is a comprehensive guide designed to inform students of Academic City University College about the policies, expectations, and resources that shape their academic and campus life. Here's a concise summary of its content:

🎓 Purpose of the Handbook

To guide students through their academic journey, outline their rights and responsibilities, and ensure they understand the policies and codes of conduct at Academic City.

📚 Key Sections & What They Cover

1. Introduction

- Vision, mission, and values of the university.
- Governance structure and guiding principles.

2. Admissions & Registration

- Admissions process, eligibility, offences, and penalties.
- Semester registration, deferment, and ID card policies.

3. Academics

- Academic programs and policies.
- Grading systems, progression rules, and graduation requirements.

4. Student Conduct & Ethics

- Code of conduct covering expected behavior, freedom of expression, computing policies, and prohibited actions.
- Strict rules around alcohol, drugs, smoking, weapons, and disruptive conduct.

5. Grooming & Dress Code

- Dress code requirements for male and female students.
- Expectations for professional, clean, and modest appearance.

6. Campus Life & Facilities

- Hostel and housing policies.
- Library rules and student recreational facilities.

7. Events, Clubs & Activities

- Guidelines for public events, student clubs, and the Student Representative Council (SRC).

8. Practical Training & Internships

- Policies around practical work, internships, and volunteering.

9. Complaints & Discipline

- Mechanisms for reporting integrity breaches and misconduct.
- Disciplinary procedures, appeals, and consequences.

10. COVID-19 Guidelines

- Health and safety protocols, especially for returning students and international students.

📌 Overall Message

The handbook emphasizes **holistic student development**, combining academics, personal growth, ethical leadership, and community engagement. It serves as both a rulebook and a support tool for thriving at Academic City.

Let me know if you want a visual summary or a breakdown of a specific section!

Qualitative Findings from Comparison of Answers

Precision: RAG responses contain 68% more measurable standards (e.g., "2 inches above knee")

Contextualization: 100% of RAG answers include policy rationales vs. 20% in ChatGPT

Error Handling: RAG indicates confidence levels, while ChatGPT presents all answers assertively

System Recommendations

Use RAG System When:

- Working with specific institutional documents
- Needing verifiable citations
- Requiring policy rationales
- Prioritizing measurable standards

Use ChatGPT When:

- Seeking broader conceptual overviews
- Needing faster responses
- Exploring topics beyond document scope
- Preferring conversational interactions