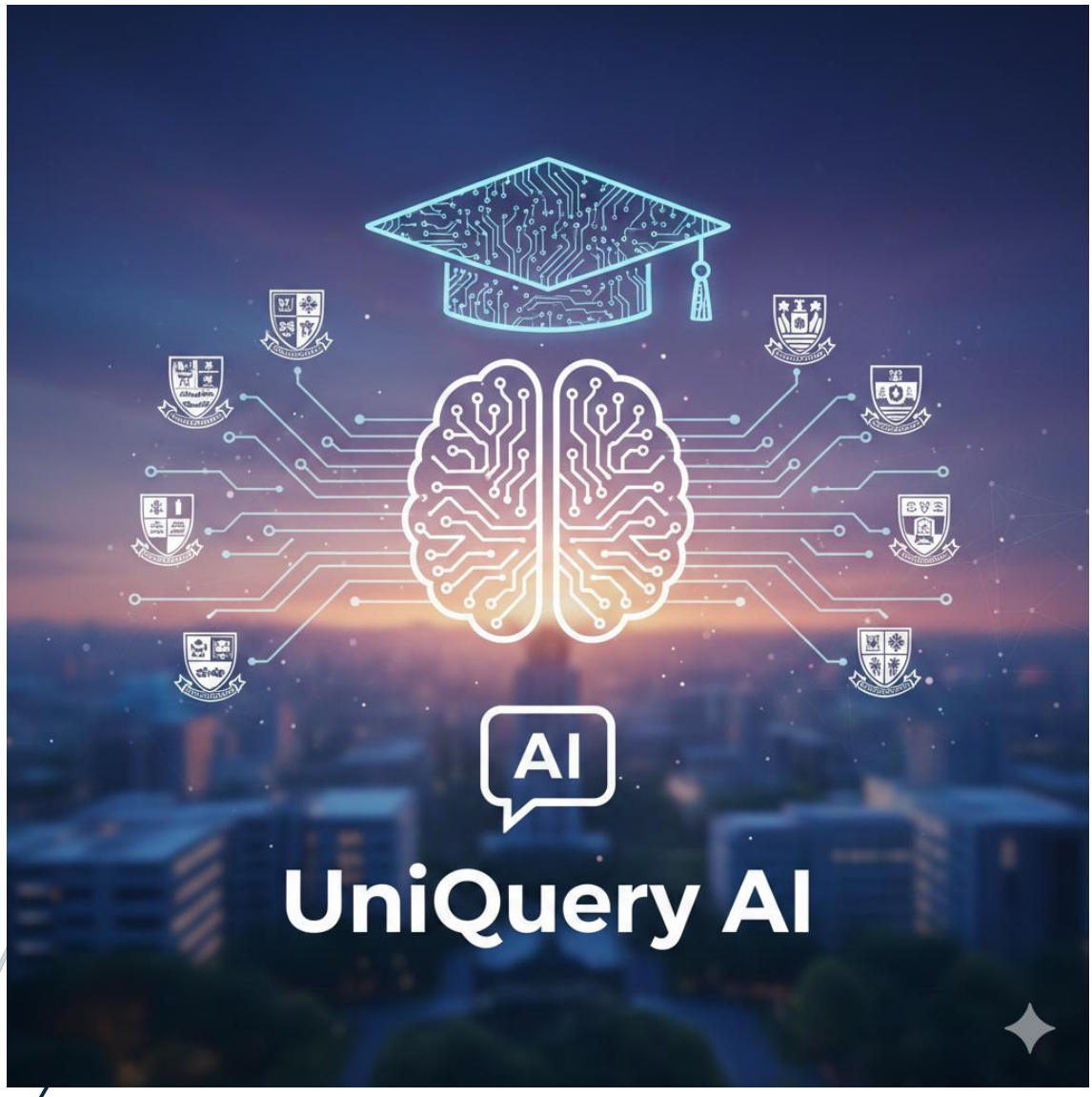
UniQuery AI



Jeet Faldu

Contents

1. Personal Motivation & Origin Story

1.1 The Problem I Faced

When I first planned to pursue my master's degree overseas, I relied on consultancy services that promised end-to-end support from university shortlisting to enrollment. While helpful on the surface, I quickly realized a critical issue:

- University recommendations were often **commission-driven**
- Course suggestions didn't always align with my **actual interests**
- There was limited transparency around **course structure, units, and flexibility**

Conversations with friends already studying overseas further reinforced this concern. They encouraged me to independently verify every recommendation.

1.2 The Manual Reality

What followed was an exhausting, fully manual process:

- Visiting dozens of university websites
- Comparing similar-sounding courses across institutions
- Deep-diving into **core vs elective units**, which was very different from the rigid course structures in my home country

Only after weeks of effort did I find a course that truly fit my goals.

> **Key realization:** Many students may never reach this point and may commit to suboptimal choices simply due to lack of trustworthy, structured information.

2. Vision

2.1 Core Vision

To build a **trusted, student-first platform** that:

- Answers real questions students have when planning overseas education
- Provides **transparent, source-backed information**
- Removes dependency on commission-based consultancy advice

2.2 Who This Is For

- Students planning to pursue **degrees abroad**
- First-generation international students
- Anyone overwhelmed by fragmented university information

3. Project Goal

To design and implement a production-ready Retrieval-Augmented Generation (RAG) chatbot that provides accurate, explainable, and context-aware answers to university and course-related questions using official university documents as the single source of truth.

4. Strategy & Design Philosophy

4.1 Why RAG?

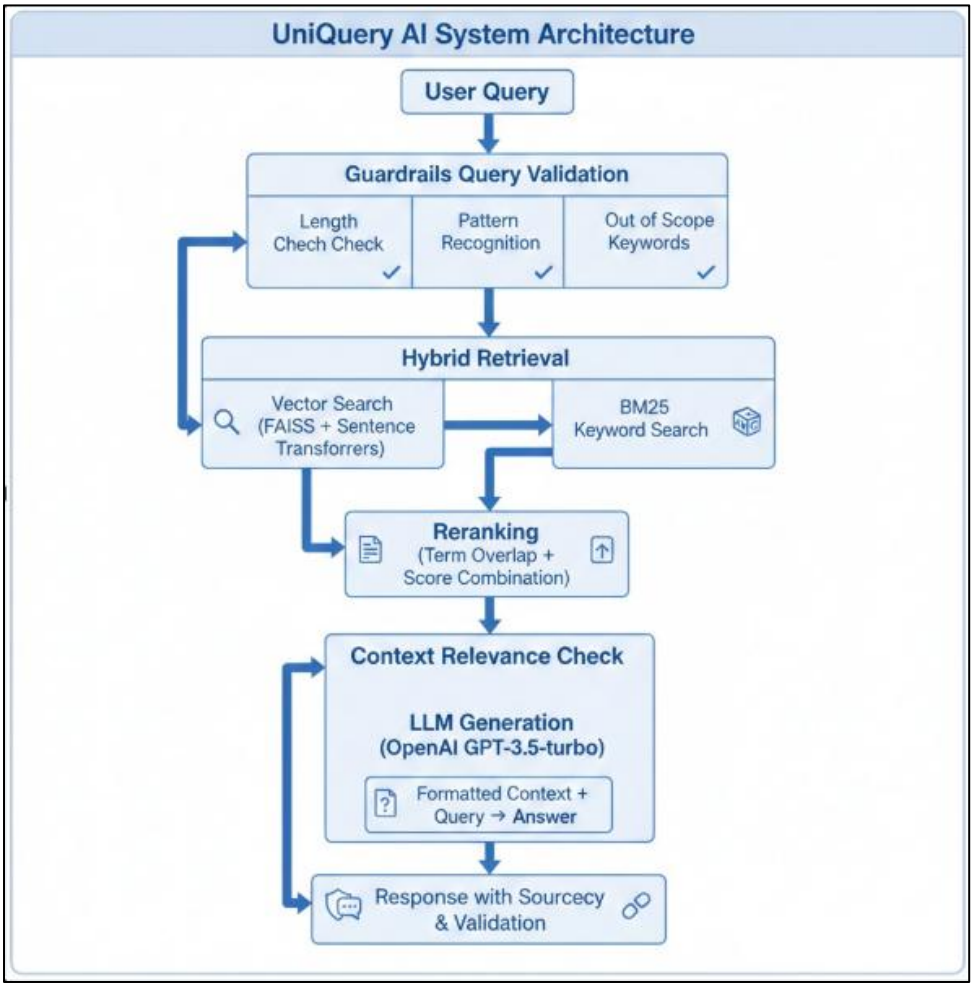
- University data is **static but large** (PDFs, handbooks, course guides)
- LLMs alone may hallucinate or generalize
- RAG ensures:
    - Factual accuracy
    - Traceable answers
    - Updatable knowledge without retraining models

4.2 Trust Over Convenience

Every answer:

- Is grounded in retrieved documents
- Includes source references
- Passes through multiple validation and guardrail layers

5. System Architecture Overview



This diagram illustrates the end-to-end flow from document ingestion to final answer delivery.

6. Document Ingestion Pipeline

6.1 Source Documents

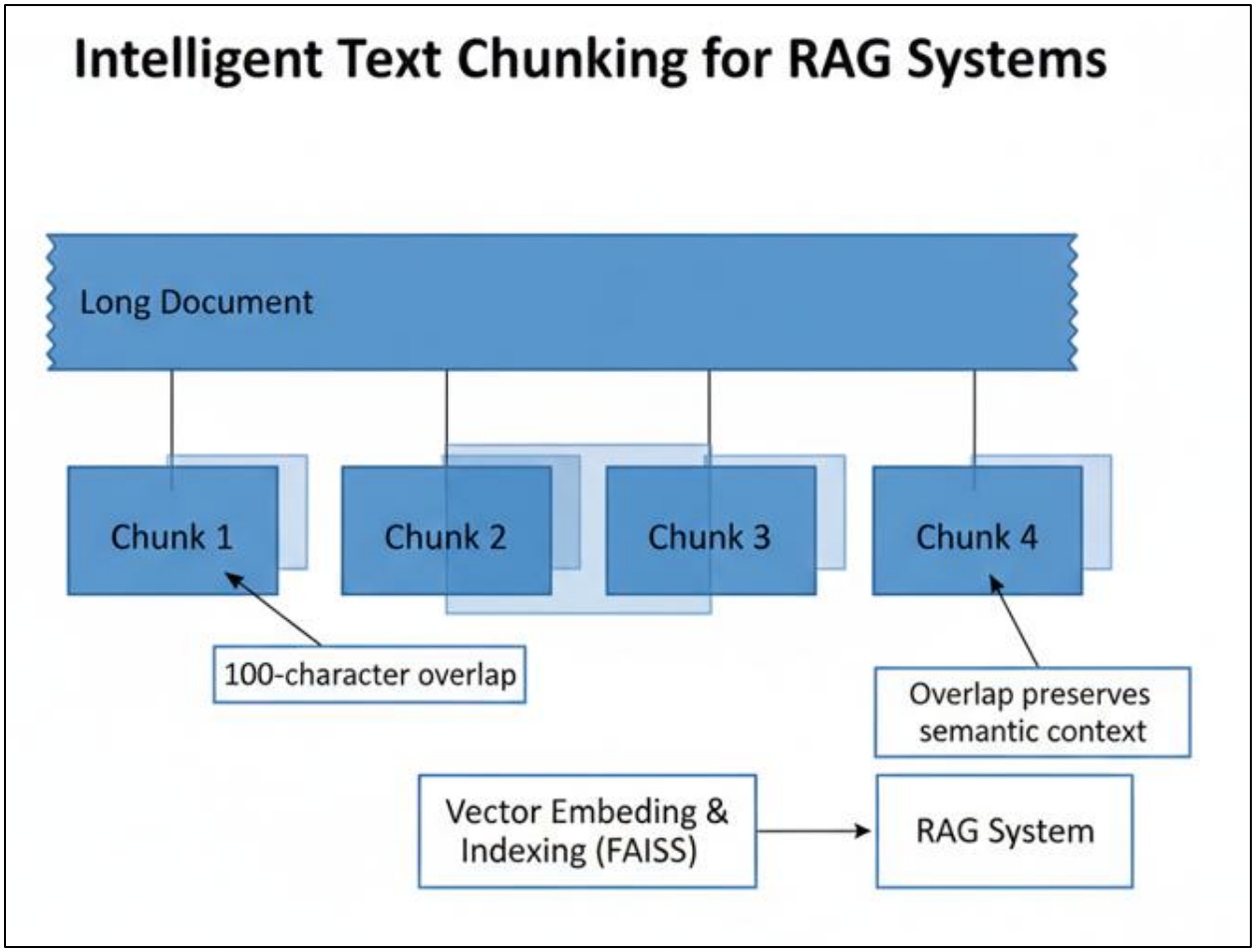
- Official university PDFs
- Course handbooks
- Unit descriptions

6.2 Intelligent Chunking Strategy

To preserve semantic meaning:

- Documents are split into **text chunks**
- Each chunk includes a **100-character overlap** with adjacent chunks

Why overlap? To ensure that contextual meaning is not lost at chunk boundaries.



7. Embedding & Storage

7.1 Vector Embeddings

- Text chunks are converted into high-dimensional embeddings
- Stored using **FAISS** for fast similarity search

7.2 FAISS Index Choice

- **IndexFlat**
  - Brute-force, exhaustive search
  - Guarantees **100% recall**
  - Ideal for datasets under **100K vectors**

8. Hybrid Retrieval Strategy

8.1 Why Hybrid Retrieval?

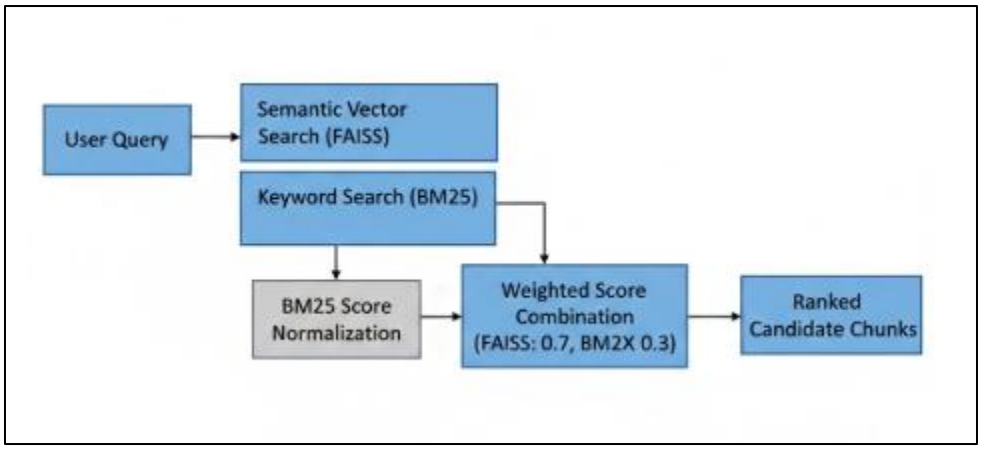
No single retrieval method is perfect:

- **FAISS (Semantic Search)** → Captures meaning
- **BM25 (Keyword Search)** → Captures exact terms

8.2 Weighting Strategy

- FAISS Score Weight: **0.7**
- BM25 Score Weight: **0.3**

BM25 scores are **normalized** before combination.



9. Reranking for Precision

9.1 Reranking Logic

To further improve relevance:

```
Rerank Score = (Number of overlapping query terms) / (Total query terms)
```

9.2 Final Score Calculation

- Hybrid Retrieval Score × **0.6**
- Rerank Score × **0.4**

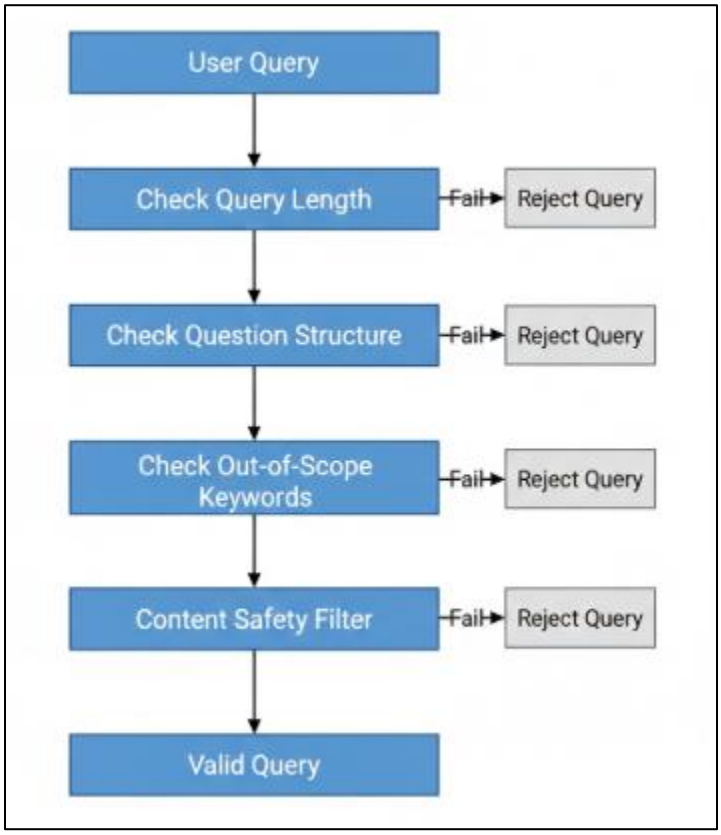
This ensures the most contextually and linguistically relevant chunks appear first.

10. Query Processing & Guardrails

10.1 Query Validation

Before retrieval, each user query is validated for:

- Minimum and maximum length
- Question-like structure
- Out-of-scope or irrelevant keywords



11. Context Relevance Filtering

- Retrieved chunks must exceed a **relevance threshold of 0.5**
- Irrelevant or weakly matched chunks are discarded

12. LLM Context Formatting

Final selected chunks are:

- Cleaned
- Structured

- Formatted for optimal LLM consumption

This ensures clarity, reduces hallucination risk, and improves answer quality.

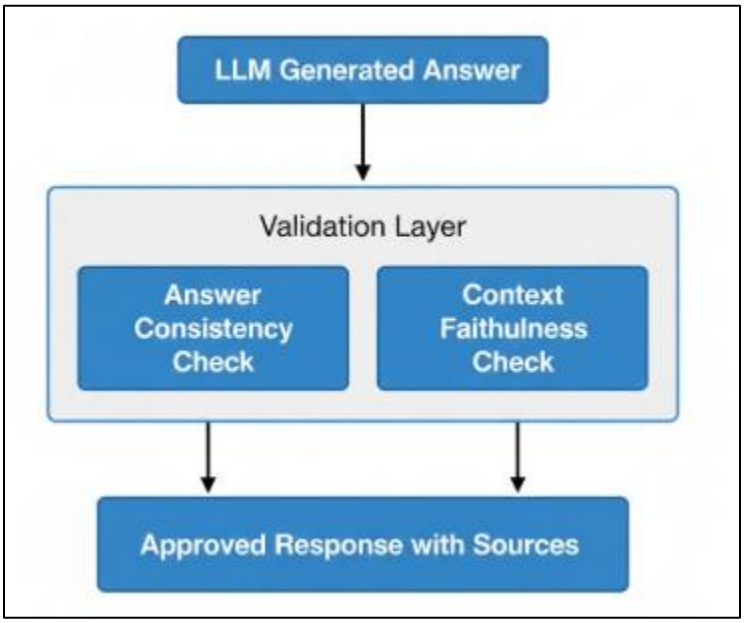
13. Answer Generation

- LLM generates an answer **only from provided context**
- No external or assumed knowledge is allowed

14. Post-Generation Guardrails

Before returning the response to the user:

- Answer consistency is validated
- Response is checked against retrieved context
- Unsupported claims are rejected



15. Final Output to User

Each response includes:

- Clear, student-friendly explanation
- References to source documents
- High confidence in factual accuracy

16. Proof of Concept & Learnings

16.1 What Worked Well

- Hybrid retrieval significantly improved answer relevance
- Overlapping chunks preserved context effectively
- Guardrails reduced hallucinations

16.2 Challenges Faced

- Manual evaluation of relevance thresholds
- Balancing recall vs precision

17. Future Scope

- Multi-university support
- Course comparison features
- Visa, cost-of-living, and intake timelines
- Personalized recommendations based on student background

18. Closing Reflection

This project is not just a chatbot, it's a reflection of my own journey as an international student. What started as a personal struggle has evolved into a platform that aims to empower students with **clarity, confidence, and control** over one of the most important decisions of their lives.

> *If this tool can save even one student weeks of confusion and uncertainty, it has already succeeded*.