Campus Connect AI

Local Knowledge Chatbot for International Students

Team: Ayana Hussain, Serena Dhillon, Poorvi Bhatia
From Simon Fraser University, BC
Mentors: Lauren Zung & Diane Johnson

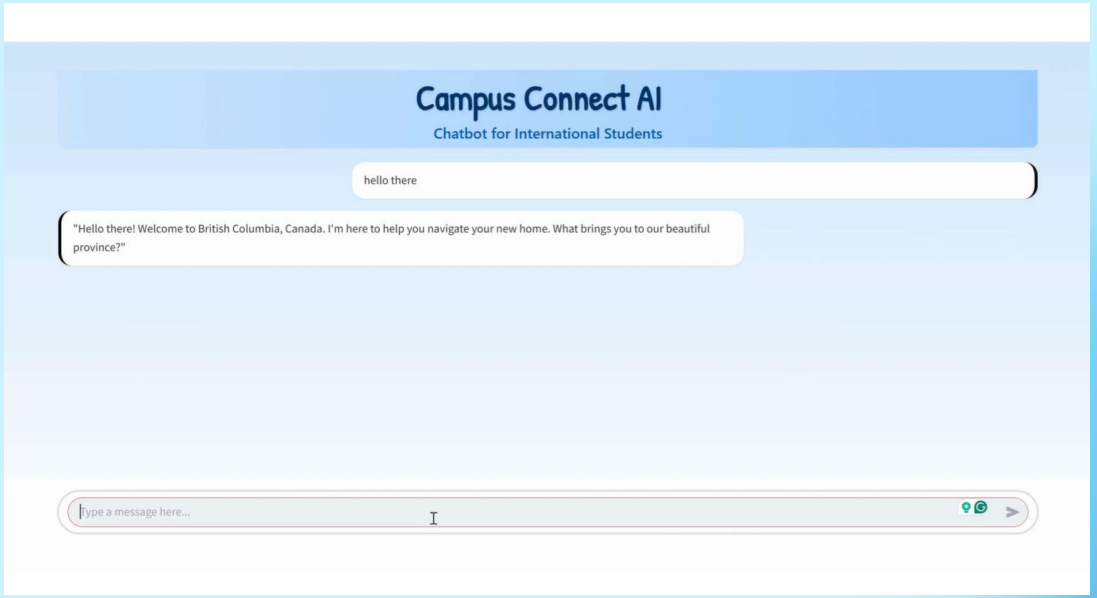


Table of contents

01

Introduction

The Problem

Lost in Information: The Challenges International Students Face

- **Key Challenges:** New environments, admin tasks, housing, and healthcare
- **Information Gap:** Difficulty finding specific relevant info across sources
- **Time Pressure:** Facing a massive volume of info while adjusting
- **Not Knowing Who to Ask:** Which person or department should be contacted?

Motivation & Our Vision: Empowering Students Through Intelligent Information Access

Why This Topic Matters

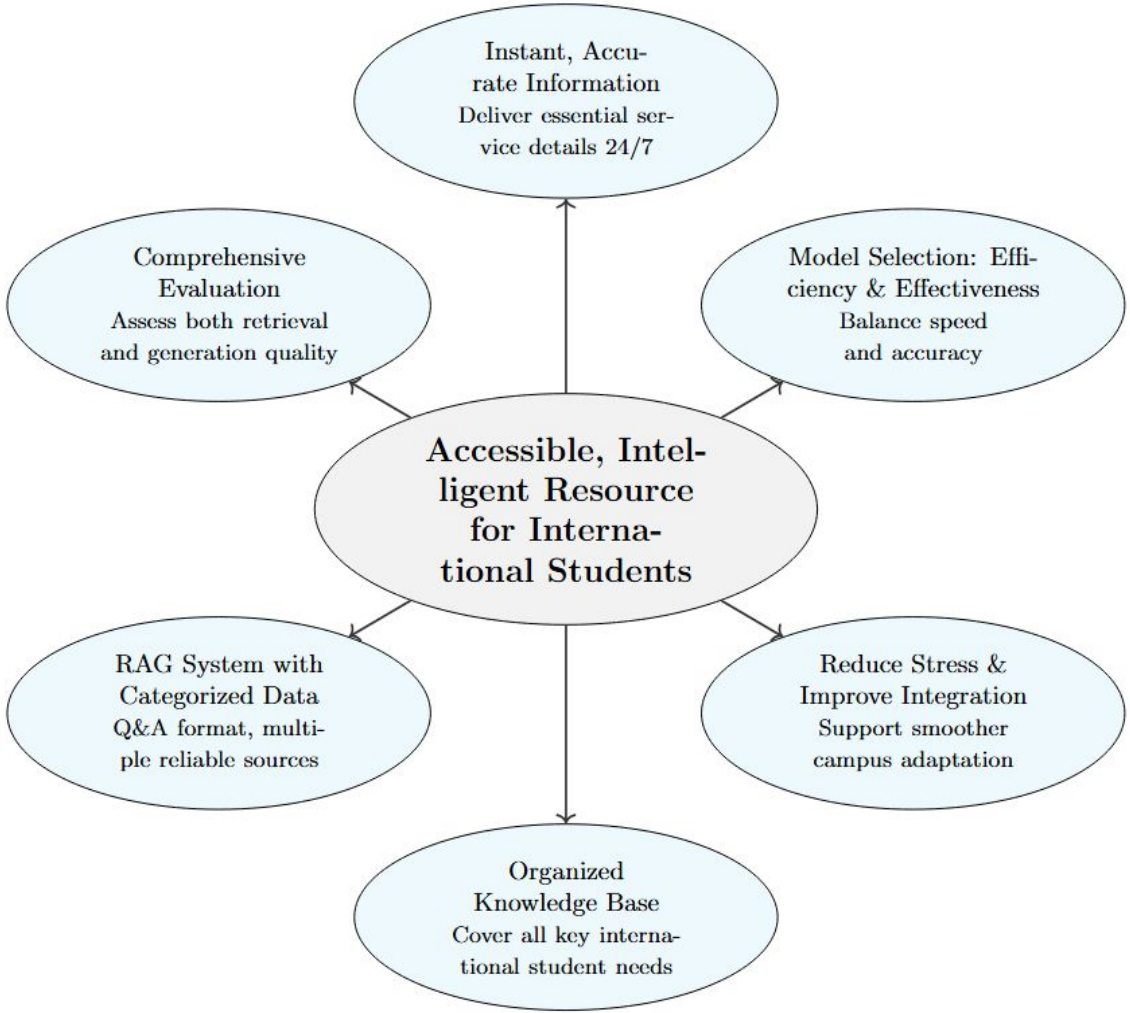
- **Personal experience** with immigration challenges and lack of family support
- Witnessed **friends struggle** with navigating **complex documentation** and resources
- **Observed difficulties** accessing timely help when needed

Motivation & Our Vision: Empowering Students Through Intelligent Information Access (Cont)

Benefits of Our Chatbot

- Provides **24/7 assistance** without long wait times
- Simplifies **access to critical information** on transit, immigration, healthcare, common FAQ for newcomers, etc.
- Allows students to **focus on academic success** rather than administrative hurdles
- Creates more **inclusive campus experience** for international students

Goals


Instant, Accurate Information
Deliver essential service details 24/7
Comprehensive Evaluation
Assess both retrieval and generation quality
Model Selection: Efficiency & Effectiveness
Balance speed and accuracy
Accessible, Intelligent Resource for International Students
RAG System with Categorized Data
Q&A format, multiple reliable sources
Reduce Stress & Improve Integration
Support smoother campus adaptation
Organized Knowledge Base
Cover all key international student needs

02

Methods

Data

University Specific

SFU clubs, general FAQs, on-campus food options, housing and accommodations.

Immigration

Study permits, work permits, permit extensions, part-time/full-time status, etc.

Healthcare

MSP, insurance coverage, B.C services card info, PharmaCare FAQs, fees, financial assistance, etc.

Transit

General routes to SFU, U-Pass BC for discounted transit pass, fares, info regarding seabus and skytrain routes, etc.

Activities

Cultural spaces, general info about hikes and parks in Vancouver.

Cost of Living

Average household expenditures for groceries, restaurants, rent, telephone bills, etc.

Models & Development

Embedding Model	Time Taken
sentence-transformers/all-MiniLM-L6-v2	32.7 seconds
sentence-transformers/paraphrase-MiniLM-L6-v2	30.5 seconds !!!!!!!!
sentence-transformers/all-roberta-large-v1	1m 55.8s
sentence-transformers/all-MiniLM-L12-v2	54.4 seconds
sentence-transformers/multi-qa-MiniLM-L6-cos-v1	49.5 seconds
sentence-transformers/paraphrase-mpnet-base-v2	over 3 minutes
sentence-transformers/multi-qa-mpnet-base-dot-v1	2m 18.7s
neuml/pubmedbert-base-embeddings	2m 27.4s

Instruction-Tuned Model Tested
google/gemma-2-2b-it
mistral-7b-instruct-v0.2.Q4_0.gguf
meta-llama/Llama-3.2-1B-Instruct
./Llama-3.2-1B-Instruct-IQ3_M.gguf

Prompt Engineering

Initial Prompt

```
rag_prompt = f"""
You are a helpful assistant for international students new to British Columbia Canada. Here are relevant documents:

{relevant_texts}

Please respond to the following question. Be conversational but concise, aim to answer accurately using the documents,
but in as few words as possible (i.e. less than 20).
DO NOT USE THE DOCUMENTS IF THEY ARE NOT HELPFUL FOR THE QUERY.
Do not ask the user irrelevant questions unless it relates to their query.
Question: {query}

Answer:
"""
```

Category-Wise Prompts

```
# category-wise prompts
hike_prompt = f"""
    INSTRUCTIONS:
        1. Convert structured information about the hike into a short, friendly paragraph using natural language.
        Do not repeat numbers or use formatting from the source.
        2. If they ask about hiking information, only answer with required information. Users can ask for more information if needed.
        3. When asked for a particular type of hike, find it instead of saying that one would not work in the category they asked for.
        4. Do NOT list trail attributes or stats (like “Distance: 3.1 km, Elevation: 789 m”).
        Instead, describe them in context (e.g., “a steep 3 km trail with a tough 789 m climb”).
        5. Avoid repeating exact numbers unless essential (e.g., elevation gain is helpful, but don't dump all stats).
"""
```

```
food_prompt = f"""
    INSTRUCTIONS:
        1. Convert structured food and dining information into a friendly, helpful paragraph. Do not copy the question or use list formatting.
        2. Only answer what the user asked. Do NOT add information that wasn't requested.
        3. Describe details in a natural way (e.g., “open 24/7 during the semester” instead of “Hours: 24/7”).
        4. Mention unique features only when they help clarify the user's question.
        5. If a specific venue or program is asked about (e.g., a café, meal plan, or food station), describe it clearly in context.
        6. If the question can't be answered from the data, respond with:
            "I'm sorry, I don't have that information. Please check the official SFU Food website.”
        7. Provide the official link when available and relevant to the answer.
        8. Do NOT list menu items, prices, or square footage unless directly relevant to the user's question.
        9. Only provide food information that is relevant. If they ask for some place that serves a chicken sandwich do not provide information to a vegan place.
"""
```

```
housing_prompt = f"""
    INSTRUCTIONS:
        1. Convert structured information about SFU or general student housing into a short, friendly paragraph using natural language.
        Do not repeat formatting or list prices unless helpful for context.
        2. Focus on what matters to the student: location, room types, meal plans, how to apply, and support available.
        3. Only mention costs in a general way (e.g., "starts around $4,000 per term") unless the user explicitly asks for detailed pricing.
        4. If information varies (e.g., by room type or campus), explain this clearly but briefly.
        5. If the user asks a specific housing question and the answer depends on certain conditions
        (e.g., term length, student status), explain those conditions clearly and simply.
        6. If the answer is not known or not in the data, respond with:
            "I'm sorry, I don't have that information. Please check the SFU Housing website for details."
        7. Do NOT dump full lists of buildings, prices, or amenities. Summarize and keep it conversational.
        8. If the information is specific to SFU, make sure you say it to be clear.
"""
```

```
parks_prompt = f"""
    INSTRUCTIONS:
        1. Convert structured information about the park into a short, friendly paragraph using natural language.
        Do not repeat numbers or use formatting from the source.
        2. Provide only necessary information that will allow the user to enjoy the park.
            - Feel free to tell them about logisitical information if asked.
"""
```

```
# activities general: covers how to answer general parks, hikes, food, clubs, cultural related questions
activities_general = f"""
    INSTRUCTIONS:
        1. If they ask for suggestions, provide 2 to 3 suggestions.
        2. Do NOT list all information. Instead describe them in context
        3. Provide accuracte suggestions, NOT suggestions of things that will not work for what they want.
        4. Convert structured information about the activity into a short, friendly paragraph using natural language.
        Do not repeat formatting from the source.
"""
```

```
# permits prompt: covers ways to answer immigration, study permits, work permits, and permanent residence related questions
permits_prompt = f"""
    INSTRUCTIONS:
        1. When given a specific question with many possible answers, you can ask for more specific information.
            - if they are not asking for an extension do not provide information in regards to an extension of a permit.
        2. Only answer with information provided
            - Information should NOT be guessed and do NOT add extra information
        3. If the answer is not in the dataset, respond with: "I'm sorry, I don't have that information.
        Please check the official IRCC website for more details."
        4. If it is helpful, provide the link and a description about it.
        5. Do NOT list all information. Instead describe them in context
        6. If the answer depends on a specific condition explain those clearly.
        7. Do NOT make assumptions about the user's situation.
"""
```

```
transit_prompt = f"""
    INSTRUCTIONS:
        1. Convert structured information about public transit into a short, friendly paragraph using natural language.
        2. Do NOT list statistics or technical formatting (like route numbers or fare charts) unless directly relevant to the user's question.
        3. Summarize relevant transit options clearly — describe them in context (e.g., “a quick SkyTrain ride from downtown to the airport”).
        4. Provide only what the user needs to understand how to get around or plan their trip.
        5. If the user is asking for directions, give a general summary of how they might travel.
        6. If the question is about fares, schedules, or route planning and the exact info is not available,
            tell the user to check the TransLink website and briefly explain what they can find there.
        7. Do NOT guess or make up transit information.
        8. If the information is not in the source, say “I'm sorry, I don't have that information. You can check the official TransLink site for more details.”
"""
```

Main Final Prompt

```
# main rag prompt
rag_prompt = f"""
You are a helpful, friendly assistant for international students new to British Columbia, Canada.

Below are some reference documents that may be relevant to the user's question:
{relevant_texts}

INSTRUCTIONS:
1. If the user's query is just a greeting (like "hello", "hi", "what's up"):
   - Respond with a single brief friendly greeting
   - Offer to help with questions about studying or living in BC
   - Do NOT include ANY information from the reference documents
   - Do NOT create additional answers beyond answering their original question

2. If the user is asking for information:
   - Be friendly and answer based ONLY on the reference documents if relevant
   - Summarize the necessary information into a couple sentences.
   - Do NOT create additional questions and answers beyond answering their original question
   - Limit your entire response to no more than 3 concise sentences when possible. Do not create long multi-line answers.
   - If the documents don't provide sufficient information, say "I don't have enough information to answer that. Please refer to official sources."
   - Ask for more information when there are multiple scenarios in the documents.
   - If they ask things like "can I", "will I", "how can I" feel free to ask follow up questions if you don't how to answer with the information provided. Do not just assume.

3. IMPORTANT: Never generate additional content beyond answering the user's question. Do NOT number or bullet your points. Always use natural sentences and group similar information together where possible.

User question: {query}

Your response (just the answer, no preamble):
"""
```

Building the System: Our Iterative Approach

01

Initial Prototyping

Trial and error to better understand system requirements.

- Started development and testing in VS Code
- Used set of 3 initial data sources
- Utilized Ollama (Llama 3) as the generator

02

Model Experimentation

Evaluating quality across different models

- Transitioned to Python Notebooks
- Compared and evaluated different models generation and retrieval quality

03

User Interface

Making the system accessible for testing and demonstration

- Used Streamlit to build an interactive front end
- Developed basic UI to allow users to input queries and receive answers

Interface Demo

03

Evaluation

Evaluation Approaches

Automated Evaluation

- Created a golden dataset of reference questions and answers encompassing all the data sources
- Evaluated the semantic and word-level similarity between the reference answers and the generated answers
- Scored the answers and documents retrieved using LLM-as-a-Judge framework

Manual Evaluation

- Evaluation by international students
  - Left for future work

Evaluation Metrics

Generation Model Quality

- Reference Comparison Metrics:
  - BERTScore (F1)
  - ROUGE-L
  - METEOR
- Overall Quality (LLM-as-a-Judge) Metrics
  - Relevance, Coherence, Fluency, Coverage, Level of Detail, and Diversity

Retrieval Model Quality

- Overall Quality (LLM-as-a-Judge) Metrics
  - Non-Rank-Based
    - Accuracy, Precision, Recall
  - Rank-Based
    - Mean Reciprocal Rank (MRR), Mean Average Precision (MAP)

Generation Evaluation Results + Takeaways

Seen Data

- LLM Evaluation
  - Relevance: 4.92
  - Coherence: 4.81
  - Fluency: 5.00
  - Coverage: 3.96
  - Level of Detail: 2.92
  - Diversity: 1.19
- Comparison Metrics
  - BERT F1 0.9057
  - METEOR 0.4170
  - ROUGE-L 0.3748
- Takeaway
  - Can still have high quality answers and even contextual embedding similarity even if not using exact same wording as the reference answers.

Generation Evaluation Results + Takeaways

Unseen Data

- LLM Evaluation
  - Relevance: 4.74
  - Coherence: 4.65
  - Fluency: 4.96
  - Coverage: 3.91
  - Level of Detail: 2.91
  - Diversity: 1.26
- Comparison Metrics
  - BERT F1 0.8925
  - METEOR 0.3318
  - ROUGE-L 0.2750
- Takeaway
  - Overall responses scored as very high quality (with a small drop compared to seen data)
  - Similar results to seen data on comparison metrics

Retrieval Evaluation Results + Takeaways

Seen Data

- Accuracy: 0.825
- Precision: 0.804
- Recall: 0.686
- MRR: 0.941
- MAP: 0.941

Unseen Data

- Accuracy: 0.319
- Precision: 0.319
- Recall: 0.194
- MRR: 0.417
- MAP: 0.361

- Takeaway
  - LLM also judges the documents used to be relevant for the seen data
  - For unseen data – findings may indicate a gap in the data
  - But the high scores on the generated responses suggests the LLM is able to use relevant info and ignore irrelevant content in the documents

04

Accomplishments + Future Work

Accomplishments

- Built end-to-end RAG prototype
- Cleaned and integrated data from multiple data sources
- Developed a Streamlit UI
- Achieved high generation quality scores on seen data (with **98.4%** relevance) and (**94.8%** relevance) on unseen data
- For seen data, documents retrieved are both accurate and well-ordered, with areas identified to further improve the coverage of documents for unseen data

Learning

RAG Systems and Implementation

- Understanding of the RAG pipeline
- Practical experience with Hugging Face, Streamlit, and Ollama
- Compared different RAG frameworks/approaches
- Impact of embedding models on performance and retrieval
- Model quantization techniques
- Prompt engineering techniques

RAG Evaluation

- RAG specific evaluation methods and metrics
- Created gold benchmark datasets
- Learned prompt engineering for answer quality

Challenges

Data Processing

- Optimal data chunk size for retrieval
- Selecting relevant data sources
- Data cleaning, aggregation, and formatting

Model Performance

- Finding balanced embedding models
- Embedding quality for document paring
- Overall system runtime (embedding + generation)
- Hardware resource limitations and long wait times

System Implementation

- Developing and integrating the UI
- Deciding on most efficient way to store and manage vector embedding (single vs multiple collections)

Future Directions

01 — 02 — 03

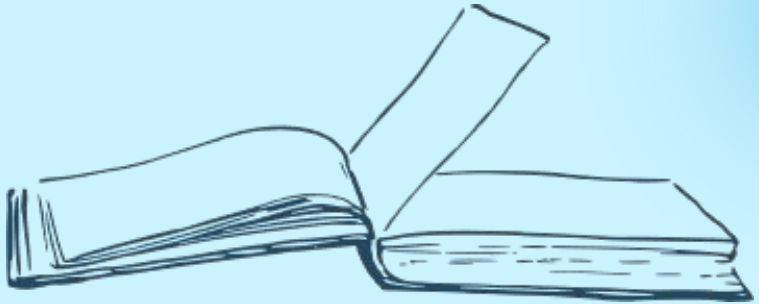
User Testing

Add more data

Deploy to Hugging Face

05

THE END



Credits

Title slide icons:
https://www.slidescarnival.com/template/linguistic-philosophy-of-education-slides/202625

Blue and pink graphics:
https://www.slidescarnival.com/template/book-club-marketing-plan/29147

Sample video