徽标, 公司名称

AI 生成的内容可能不正确。



**Enhancing the College Admission Counseling Experience for Prospective Chinese Outbound Students Through Retrieval-Augmented Generative Chatbots**

**By**

**PAN SHIJIE**

**TP055146**

**APU3F2505CS**

A report submitted in partial fulfilment of the requirements for the degree of

Bachelor of Science (Honours) in Computer Science (CS)

at Asia Pacific University of Technology and Innovation.

**Supervised by Assoc. Prof. Ts. Dr. Vinothini Kasinathan**

**2nd Marker: Assoc. Prof. Dr. Imran Medi**

**2025**

# **Acknowledgement**

Firstly, I would like to thank the Asia Pacific University of Technology and Innovation for offering me a chance to explore my interest in computer science by providing valuable modules. Furthermore, I would wish to express my gratitude also to Assoc. Prof. Ts. Dr. Vinothini Kasinathan who was my Final Year Project supervisor and given her unending support, encouragement and invaluable advice during this project. Her experience of work in academic research and technical development was important in setting the course of my activities. The constructive criticism that she contributed during the process of proposal approval, through the final report submission stage, has assisted me in the clarification of my concepts as well as enhanced the general quality of this study. In addition, I would like to thank my Second Marker, Assoc. Prof. Dr. Imran Medi too, for taking time and offering me wise comments and suggestions in the process of evaluation. His factual notes proved quite helpful in increasing the level of academism and organizational transparency of this report, especially during the literature review section and the methodology section. Finally, I would like to thank my lecturer in Investigation report, Ms. Nur Amira Binti Abdul Majid, for her continuous guidance on correcting my report support documents as well as guidance on the entire report. Her patience and guidance motivated me to finish this project. Their guidance has also been invaluable to this project that is aimed at developing a multimodal input acquirable Retrieval-Augmented Generation (RAG) chatbot system in the service of outbound Chinese students. Interests exceeding their mentorship here will shape my academic and professional career with the knowledge and skills that I have acquired here, such as the implementation and utilization of AI as well as deployment of backend applications through Fast API. To my supervisor, my second marker, and my lecturer, thank you very much for the efforts you have put in and supported me in this Final Year Project Investigation Report.

# **Abstract**

Outbound mobility Chinese prospective students lead the worldwide charts but support them towards counselling is still limited to stale FAQ pages and overworked agents. This undertaking would study a Retrieval-Augmented Generative (RAG) chatbot with combined dense retrieval (Qwen3-Embedding-8B + Qwen3-Reranker-8B) and multimodal large language models (GLM-4.1V-9B, Qwen 2.5-VL-72B) to provide source-cited, up-to-date, and bilingual advice. The system requirements are informed by a mixed-methods study which has been conducted as literature survey, questionnaires (n > 30). The prototyping of a solution is the Fast API micro-service based on the FAISS vector store and an adaptive chunking pipeline. This project brings a repeatable plan to SDG 9 (Industry, Innovation, and infrastructure) via the development of scalable, AI-powered infrastructure to support intelligent and transparent consultation operations in the form of an infrastructure and software system.

**Keywords**: Retrieval-Augmented Generation; Natural Language Processing; SDG 9

Table of Content

[**Acknowledgement** 2](#_Toc205925628)

[**Abstract** 3](#_Toc205925629)

[List of figures 6](#_Toc205925630)

[List of Tables 7](#_Toc205925631)

[1.0 Introduction 8](#_Toc205925632)

[1.1 Motivation to selecting Retrieval-augmented Generation (RAG) 9](#_Toc205925633)

[1.1.2 Cross-Lingual, multimodal Support: The Case 9](#_Toc205925634)

[1.1.3 Anticipated Contributions 10](#_Toc205925635)

[1.2 Problem Background 10](#_Toc205925636)

[1.3 Project Aim 11](#_Toc205925637)

[1.4 Objectives 11](#_Toc205925638)

[1.5 Scope 12](#_Toc205925639)

[1.5.1 Project Inclusions 12](#_Toc205925640)

[1.5.2 Project Exclusions 12](#_Toc205925641)

[1.5.3 Project Constraints 12](#_Toc205925642)

[1.5.4 Assumptions 12](#_Toc205925643)

[1.5.5 Dependencies 13](#_Toc205925644)

[1.5.6 Risks 13](#_Toc205925645)

[1.6 Potential Benefits 13](#_Toc205925646)

[1.6.1 Tangible Benefits 13](#_Toc205925647)

[1.6.2 Intangible Benefits 13](#_Toc205925648)

[1.6.3 Target Users 14](#_Toc205925649)

[1.6.4 Functionalities / Deliverables 14](#_Toc205925650)

[1.7 Overview of the IR 14](#_Toc205925651)

[1.8 Project plan 15](#_Toc205925652)

[Chapter 2 Literature Review 16](#_Toc205925653)

[2.0 Introduction 16](#_Toc205925654)

[2.1 Domain Research 16](#_Toc205925655)

[2.1.1 Advising context and problem framing 16](#_Toc205925656)

[2.1.2 Why Retrieval-Augmented Generation (RAG) fits this domain 19](#_Toc205925657)

[2.1.3 Cross-lingual and multimodal advising patterns 23](#_Toc205925658)

[2.1.4 Operational constraints, performance and stakeholders 25](#_Toc205925659)

[2.1.5 Alignment with SDG 9 (Industry, Innovation and Infrastructure) 28](#_Toc205925660)

[2.2 Similar System 30](#_Toc205925661)

[2.2.1 Dify 30](#_Toc205925662)

[2.3.2 Haystack 32](#_Toc205925663)

[2.2.3 Rasa 34](#_Toc205925664)

[2.2.4 Comparison of Similar System 37](#_Toc205925665)

[2.2.5 Conclusion 39](#_Toc205925666)

[2.3 Technical Research 40](#_Toc205925667)

[2.3.1 System Requirements Analysis 40](#_Toc205925668)

[2.3.2 Programming Language Choosen 41](#_Toc205925669)

[2.3.2.1 Justification 42](#_Toc205925670)

[2.3.3 IDE Choosen- Visual Studio Code 43](#_Toc205925671)

[Chapter 3 Methodology 44](#_Toc205925672)

[3.1 System Development Methodology 44](#_Toc205925673)

[3.1.2 Methodology Chosen and Justification 44](#_Toc205925674)

[3.1.3 Activities and Processes in Each Phase of the Agile Scrum 47](#_Toc205925675)

[3.2 Data Gathering Desing 49](#_Toc205925676)

[3.2.1 Data Gathering Method 49](#_Toc205925677)

[3.2.2 Design of Questions 50](#_Toc205925678)

[3.2.2 Data Analysis 54](#_Toc205925679)

[Chapter 4 Conclusion 71](#_Toc205925680)

[References 73](#_Toc205925681)

[Appendices 75](#_Toc205925682)

[Appendix A: PPF 75](#_Toc205925683)

[Appendix B: Fast Track Form 76](#_Toc205925684)

[Appendix C: Meeting Log 77](#_Toc205925685)

[Appendix D: Gantt Chart 78](#_Toc205925686)

[Appendix E: Respondent Demographic Profile 79](#_Toc205925687)

# List of figures

[Figure 1 Conceptual Model: Factors affecting information seeking behavior of international students 18](#_Toc205923566)

[Figure 2 Concept Design Retrieval-Augmented Generation (RAG) system (OKTAVIA, 2023) 20](#_Toc205923567)

[Figure 3 Retrieval-Augmented Generation (RAG) Architecture (OKTAVIA, 2023) 20](#_Toc205923568)

[Figure 4 Summary of RAG Application in Education. 21](#_Toc205923569)

[Figure 5 Retrieval-Augmented Generation for Dialogue (simplified). 22](#_Toc205923570)

[Figure 6 Automate grading of handwritten (AxcelerateAI, n.d.) 24](#_Toc205923571)

[Figure 7 Mapping PLOs/CLOs to course content (AxcelerateAI, n.d.) 24](#_Toc205923572)

[Figure 8 SDG 9 29](#_Toc205923573)

[Figure 9 Defy Main Page 30](#_Toc205923574)

[Figure 10 Dify Workspace 1 31](#_Toc205923575)

[Figure 11 Dify Workspace 2 31](#_Toc205923576)

[Figure 12 Haystack Main Page 32](#_Toc205923577)

[Figure 13 HayStack WorkSpace 33](#_Toc205923578)

[Figure 14 Rasa Main Page 34](#_Toc205923579)

[Figure 15 Rasa Function 35](#_Toc205923580)

[Figure 16 Top Programming Language Choosen 41](#_Toc205923581)

[Figure 17 Scrum (GeekforGeek, 2025) 45](#_Toc205923582)

[Figure 18 Each Phase of the Agile Scrum 47](#_Toc205923583)

# List of Tables

[Table 1 Project Plan 15](#_Toc205923584)

[Table 2 SERVQUAL dimensions (Ismail, 2021) 27](#_Toc205923585)

[Table 3 Table Comparison of Similar System 1 37](#_Toc205923586)

[Table 4 Table Comparison of Similar System 2 37](#_Toc205923587)

[Table 5 Table Comparison of Similar System 3 38](#_Toc205923588)

[Table 6 Comparion of system development methodology 46](#_Toc205923589)

# Introduction

During the academic year 2020 to 2021, Chinese sending students abroad totaled over 1.06 million as the country maintained its top position on the international student-export ranking and two enduring drivers (I) a high demand of foreign degrees and (ii) a low supply of domestic degrees (Center for China & Globalization, n.d.). Both these trends have been promoted by the post-pandemic economic boom and by the normalization of hybrid learning.

However, the current advisory landscape of frozen frequently asked questions, overbooked counsellors, and rule-based chatbots is incapable of supporting the requirements of learners. The lapse occurs three times:

1. Freshness of knowledge - the knowledge that is up to date such as the laws concerning visas and patterns of admission.

2. Contextual malleability- There are few tools that integrate transcripts, screen captures or hybrid Chinese English queries easily.

3. Transparency - The users learn nothing about the way the answers are formed, through deterministic or ambiguous logic.

In a scoping review of 118 educational-AI studies, the same pain-points are noted: aging knowledge graphs, inefficient multimodal processing and black-box reasoning (Yan et al., 2023).

## 1.1 Motivation to selecting Retrieval-augmented Generation (RAG)

The use of RAG solves the problems of freshness and transparency by not only having a language generator but a perpetually redeveloping store of evidence. As the works by Lewis et al. (2020) with subsequent refinements by Izacard & Grave (2021) illustrate, factual accuracy is preconditioned by a deployment of tightly coupled neuro processes between neural inference and generation, such that the level of factual accuracy is considerably higher, and the level of hallucination is much lower based on knowledge-dependent activities. The commercial mainstreaming is catching pace as RAG enables terabyte heaving LLMs to be bank slim along with the evolution of knowledges base.

### 1.1.2 Cross-Lingual, multimodal Support: The Case

Study-abroad requests seldom come in as plain text. Applicants come and post pictures of IELTS test score disclosures, Screenshot conditions of scholarship or send WeChat audio notices regarding due dates. Current multimodal LLMs can process images along with bilingual text and reason on the latter in a single iteration, and describe reasoning on it, but needs a robust retrieval system to attribute a correct answer to the LLM. In the meantime, automatic speech recognition technology such as FunAudioLLM can turn voice messages to text (STT), thus completing the feedback circle of hands-free or accessibility-based situations.

The proposed exploration also entails a generation of a multimodal, RAG-enhanced consulting chatbot, able to provide personal consulting services in large numbers, verified and tracked:

1. Corpus ingestion - program brochures, policy briefs, visa regulations and exemplary essays are chunked and indexed into a FAISS vector store.

2. Dense retrieval & reranking - both the Qwen3-Embedding-8B and the Qwen3-Reranker-8B contain high-dimensional vector representation; the candidate passages are scaled down to measure the level of relevancy.

3. Sentence break dissection- Kimi-K2-Instruct divides mixed Chinese English inputs to refine matching between inquiry and evidence output.

4. Multimodal reasoning & STT combination- GLM-4.1V-9B, Qwen 2.5-VL-72B and FunAudioLLM (speech-to-text) work together in parsing text, picture or sound, before delivering reference, consistent responses.

### 1.1.3 Anticipated Contributions

* Scalability – decoupled retrieval lets the fact base expand without retraining the core model.
* Verifiability – every reply cites its supporting passages, bolstering trust.
* Personalization – embeddings capture each user’s context (major, budget, IELTS band) without manual rules.
* Transferable blueprint – a modular reference stack that peers can replicate for scholarships, immigration advice or professional-licensing questions.

By combining RAG’s factual grounding with multimodal LLM versatility, and by leveraging FunAudioLLM for robust speech-to-text, the project aims to set a new benchmark for study-abroad counselling, responsive, transparent and truly learner-centered.

## 1.2 Problem Background

China outbound study-abroad industry has come to operate under a much more condensed decision window: policy memos could be refined and improved quarterly, whereas application portals like the UCAS or the Common App were constrained by immovable, fixed windows of operation. Evidently, counsellor practitioners find it hard to keep up, and students often turn to social threads, which are rendered outdated in several minutes of these threads moving across social media. Such instability is like shooting applicants with both barrels both with too much noise and too little in the form of authority. Two technological pillars have been used in historical industry practices. The former is the fixed knowledge hub, the FAQ pages, and PDF manuals that are handled manually and must be revised manually. Chinese agencies show an average latency of 336 to update at a rate of 336 (Chen & Wang, 2023) or 6 months, which is much slower than the time frame associated with changing policies in visa or in the test- scores. The second pillar is the pattern-matching chatbot where questions are redirected down decision trees; though saleable, these tools fail when queries are out of expected intent or language mixed (Sheehan et al., 2020).

Thus, users are presented with a paradoxical situation of a plethora of online answers, which no one source-of-truth is felt reliable. Recently, a survey of 310 Chinese undergraduate participants showed that 68 % of them did not trust the quality of chatbot advice and 72 % of them quit the session when there were no citations of a document (Yan et al., 2023). It also makes compliance difficult lack of verifiable provenance; universities are increasingly insisting that counselling statements specify the policy clause or page in the handbook on which the recommendations depend.

To make it all more complicated is modality drift: at the current stage, admission officers accept screenshots of IELTS score reports or voice descriptions of extracurricular projects, but most digital advisors are still text-only. The large models capable of vision have been developed, but without grounding in retrieval, they can hallucinate reference ranges or read the numeric fields improperly (Li et al., 2023). Bilingual audio notes are extremely difficult to analyze unless transcribed first, a manual addition to the process, where the efficiency promised of multilingual audio is lost.

Altogether, these limitations explain why existing advisory models poorly perform in an outbound Chinese admissions environment with speed and multimodality. They inspire a solution, which is (I) modern in design, (ii) modality-aware, (iii) and in a position to reveal its reasoning track to final users.

## 1.3 Project Aim

To develop a multimodal input acquirable text-centric Retrieval‑Augmented Generation chatbot that delivers verifiable, personalized and up‑to‑date study‑abroad counselling for Chinese outbound students.

## 1.4 Objectives

1.4.1 To curate and ingest the corpus – convert policy PDFs, programme guides, visa FAQs and exemplar essays to clean text; store them in a FAISS index with paragraph / sentence metadata.

1.4.2 To perform dual‑stage retrieval – apply dense embedding with Qwen3‑Embedding‑8B followed by neural reranking with Qwen3‑Reranker‑8B via Silicon Flow Cloud API.

1.4.3 To implement chunk‑aware segmentation – optimize sentence / clause splitting for long documents to balance context integrity and retriever precision.

1.4.4 To generate transparent answers – return responses with inline citations that hyperlink to retrieved passages.

1.4.5 To create functional prototypes about the foundation of Fast API and offer the following addresses: /ingest, /query, /reran, /answer. The system will be integrated with Silicon Flow cloud Model API service and made to respond with median latency of 5 seconds to each request.

## 1.5 Scope

### 1.5.1 Project Inclusions

This investigation covers the end‑to‑end design and implementation of a cloud‑hosted study‑abroad counselling platform: a Fast API backend, a minimal web frontend and a command‑line ingestion utility integrate corpus chunking, FAISS indexing, dense embeddings and neural reranking served via Silicon Flow; the system accepts text, JPEG/PNG images and WAV/MP3 audio through FunAudioLLM and multimodal LLMs, returns bilingual source‑cited answers, and includes evaluation dashboards, OpenAI documentation and a counsellor user guide.

### 1.5.2 Project Exclusions

Native mobile apps, offline desktop clients, languages beyond Chinese and English, real‑time video or handwriting OCR, persistent long‑term user profiles and on‑premise deployments are outside the scope of this prototype, which targets Silicon Flow cloud only.

### 1.5.3 Project Constraints

A single developer must complete the work within one academic semester under an academic‑credit budget, with no dedicated GPU and strict adherence to public‑license data, thereby limiting compute usage, model customization and feature expansion.

### 1.5.4 Assumptions

The project assumes stable internet connectivity to Silicon Flow endpoints, that a 100 k‑token daily quota will suffice for development and demonstrations, and that open‑source components such as FAISS and Lang Chain remain actively maintained throughout the project period.

### 1.5.5 Dependencies

Delivery relies on uninterrupted access to Silicon Flow embedding, reranking and generation APIs, FAISS for vector search, SQLite for metadata, Kimi‑K2‑Instruct for sentence segmentation, FunAudioLLM for speech‑to‑text and GitHub Actions plus Docker Hub for CI/CD.

### 1.5.6 Risks

Key risks include exceeding Silicon Flow credit limits, cloud endpoint downtime, residual hallucinations, network‑induced latency spikes and single‑developer availability; mitigations involve cached embeddings, response‑time monitoring, confidence thresholding, local fallback reranking and weekly off‑site backups.

## 1.6 Potential Benefits

### 1.6.1 Tangible Benefits

The proposed RAG-powered counselling platform delivers clear financial and operational gains: automated retrieval and answer generation cut counsellor handling time by an estimated 60 %, enabling the same staff headcount to process substantially more enquiries while lowering per‑session costs; faster, citation‑backed responses are expected to lift student conversion rates into paid advisory packages and downstream university enrolments, generating new fee income; the modular API also opens a licensing channel for third‑party agencies, and real‑time analytics help management allocate marketing spend more efficiently.

### 1.6.2 Intangible Benefits

On top of the direct revenue increase, this new regime in fact enhances the credibility of the institution. By being transparent and having its directions tied to the source, it easily earns trust among the users and drives word-of-mouth referrals. It also makes the university an innovation leader in SDG 9 by putting the focus on the infrastructure that is driven by AI. To add to that, the bilingual, multimodal support enables the student satisfaction to stay high. And the data insights garnered on the platform keep on improving the quality of the services, with the added benefits of having to comply with the stricter regulations of the origin of the advice.

### 1.6.3 Target Users

In developing an application, it is always good to have the users in the center stage. In this case, I observe four principal groups that are going to benefit. First are potential Chinese students: they will receive tailor-made, credible information on studying abroad, rather than mining random tips on the internet. Then there are frontline counsellors just think of how quicker their workflow will be when a virtual sidekick responds to basic questions. The benefits are shared by partner schools, which get more ready applicants pouring in. Finally, the AI-driven consultation platform will have concrete data to make strategic choices and track the progress.

### 1.6.4 Functionalities / Deliverables

The deliverable suite comprises a secure student web portal that ingests text, image or voice queries and returns citation‑anchored answers, a counsellor dashboard for conversation oversight and feedback loops, an ingestion CLI for bulk document import with schema validation, a Fast API backend exposing /ingest, /query, /reran and /answer endpoints, monitoring dashboards for recall and latency, and public API documentation that facilitates integration with external education‑service platforms.

## 1.7 Overview of the IR

This Investigation Report has been structured in four chapters. Chapter 1 presents the problem of study abroad counselling, the aim, objectives, its scope and the expected benefits. Chapter 2 conducts a review of Retrieval Augmented Generation, multimodal large language models and available educational chatbots with the purpose of identifying gaps motivating the solution offered, as well as listing the required tools and system requirements. Chapter 3 offers an overview of the Agile Scrum methodology used, where each of the sprints is matched to processes, including systems ingestion, retrieval tuning, multimodal incorporation and deployment of prototype. Chapter 4 concludes the research and comments on the contribution to SDG 9, mentions limitations and lays out directions of improvement.

## 1.8 Project plan

Table 1 Project Plan

表格

AI 生成的内容可能不正确。

# Chapter 2 Literature Review

## 2.0 Introduction

This chapter provides an overview of the theory and practice dealing with construction of a retrieval augmented, citation based chatbot in Chinese outbound study applicants. The review is used to: (I) place the project in the context of previous work on digital advising and conversational agents in education; (ii) compare similar systems and identify design/evaluation patterns; and (iii) synthesize the technical methods (especially those of Retrieval-Augmented Generation (RAG), cross-lingual retrieval, working with long documents, multimodal inputs, and reliability and latency) that will be required to address the project requirements.

These are limited to research and practitioner literature between just before 2018 and just after 2025 and prioritizing peer-review journals, high-quality preprints, and technical reports when specifics are required of implementation. Searching will be performed using a combination of terms like the ones below: retrieval-augmented generation, educational chatbots, advising automation, cross-lingual IR, multimodal LLM, neural reranking, hallucination mitigation, in databases such as the ACM Digital Library, IEEE Xplore, arrive, and Scopus. Works reporting measurable results (e.g. in terms of response accuracy, citation fidelity, latency or user satisfaction), works which are concerned with bilingual or cross-cultural scenes of concern to Chinese applicants and works which report deployable architectures are given priority. Exclusion eliminates works that are purely theoretical and of no empirical or implementation usefulness as well as products whose technical details are impossible to verify.

## 2.1 Domain Research

### 2.1.1 Advising context and problem framing

The domain of international study decision-making plays out within a decidedly shaky information ecology. The intakes, prerequisites, feet and portfolio or audition notes of programs vary on a rhythm that is seldom timed to the students’ search cycle. Further nuance is brought into the English-language testing policy, with overall and sub-score levels that tend to differ by faculty even within one institution. Academic opportunities have their windows and they can be tight or even rolling where it can be accelerated or put on hold depending on the usage of funds. (Krsmanovic, 2021) Meanwhile, immigration advice can be very vulnerable to policy announcements, operational backlogs and the availability of appointments. The accumulative impact is an area where authoritative materials continue to be located on various university portals, faculty microsites, PDF handbooks, scholarship pages, government or mission sites, whereas popular social platforms reinterpret them or repackage with unpredictable precision and half-life. Students and families are challenged to synthesize all this under time pressure with asymmetric language comfort and it is only natural that current result is the same almost everywhere: replicated questions, queues in front of adviser desks and regress to heuristics and hearsay when the official pages are difficult to traverse. (Gottlieb & Amanda, 2023)

Structural conditions result in weaknesses of common digital interventions. Static repositories of FAQ will become outdated the moment they are uploaded, and it would be impossible to keep dozens of entries current after a single change has been implemented in a source handbook and the most sluggish term becomes the weakest point. Rule-based or tree-driven chatbots suffer a still more serious drawback that it must presuppose an intent to be anticipated, authoritatively expressed and Budapest linked with one of the canned responses which is also open to the risk of becoming stale. (Hamid, 2018) Upon arrival of actual students with mixed language expressions, clipboard clauses found somewhere on a scholarship page or some photograph of an IELTS report form, the bots under consideration are unable to match inputs with its matching branch and provide a non-specific or older response that may weaken the faith. Even in location where live chat is provided, queues swell around foreseeable peaks, and advisers end up spending a big portion of their time restating already known responses which are not accessible to further inspection by the subsequent student who goes looking on the same matter. (Chang, 2012)

An adviser properly based is thus required to meet three properties at the same time. First, it needs to be fresh, and in practice, that implies the claims and the corresponding statements that its issues are based on documents that may be updated without retraining a model. Second, it needs to be situation-flexible, to reason about bilingual terms, and to handle pictures and brief voice snippets as first-order evidence--not second-thought add on. (Kuhail, 2022) Third, it should be transparent: passages used to justify answers should be visible, linkable and easy to verify such that students, families and staff can find the source of a recommendation within a few seconds. One can bolt on these properties as a user-interface polish later, at great sacrifice to the usability of the system otherwise, but one cannot have a system that hands out a state of affairs without them as foundational to its information architecture or to the compositional effort to construct an answer. (Hertzum & Hyldegård, 2019)

A diagram of informational text

AI-generated content may be incorrect.

Figure 1 Conceptual Model: Factors affecting information seeking behavior of international students

The advising path itself hints at the retrieval axes that ought to be made available by a system. Eligibility questions: level of degree, discipline, GPA or percentage bands; and language-test threshold are combined; process questions: format of documents, naming conventions, certification requirements and portal workflows; Scholarships questions: nationality, level, performance, deadlines and financial limitations; visa and post-offer questions: Handle carefully the ambiguous verbs, such as normally requires, may permit, follow local caveats and appointment realities. The structuring of the corpus and the query pathway with these axes gives that retrieval is not a brute force search through all the text, but a search through metadata in which the level and faculty more heavily tip the weights than ranking before creation. (Iatrellis, Samaras, & Konstantinos, 2024)

Failure modes that the system should be defensive against are also shown in the information ecology. Template confusion when formulations on one page are too close to the formulations in others, so that when a naive retriever finds a near-hit of the wrong contextual category it gives that hit back as a result. A form of trust violation is citation drift where a sentence generated contains a figure interpolated into a sentence that is close, but not the same as that cited a subtle form of violation that might not be noticed by readers who lack professional expertise. Stale citation happens when a response stops being associated with a currently outdated PDF since no recency signal can be obtained upon retrieval. These forms of failures are not imaginary because they have their artefact in the way university and visa pages are generated and revised. Any plausible adviser must thus give section headings, publication or last-modified date and the structure of URLs the status of first-class metadata, and store chunks at clause or sentence level with stable anchors and prefer more recent sources when two alternatives are equal on the quality of lexical or semantic similarity.

Due to the implicit involvement of families in the decision-making process, bilingual access cannot be considered a decoration but a necessity. By retaining official English program names in the answer, but providing Chinese explanations, clarity is enhanced and so is verifiability. In time sensitive decisions, presentation is as important as retrieval: the response to a query needs to present figures and dates in a form that is difficult to misinterpret on a phone, and a quick-read timestamp such as “last checked” needs to be visible alongside shifting policies to remind users of recency. Lastly, a domain-specific counsellor must construct an escalation systemic route. It should make statements of ambiguity and perhaps willing to package the conversation and the citations to a counsellor queue when the corpus provides contradictory statements or in case of a rule having discretionary language. The volatility in the industry that is characterized by fragmentation and multilingual artefacts means that such transparency is not only ethical, but also the safest measure to ensure that simple misunderstandings do not trigger into expensive errors. (Abraham, 2021)

### 2.1.2 Why Retrieval-Augmented Generation (RAG) fits this domain

Retrieval-Augmented Generation also satisfies the three requirements of the domain level-freshness, contextual malleability and transparency by separating the knowledge out of the model and recombining it with it at answer time. A RAG system accesses not a language model memory of the policies that existed during lingo pretraining but instead an external and constantly updated corpus of documents and conditions the generator on individual passages demonstrating to produce an output. The architectural implication is definite: altering a rule in a program handbook turns out to be a content operation as opposed to a model operation. The corpus may be crawled or curated, transformed into an index, and versioned, such that staff can describe what changed between two dates; the generator is constant. (OKTAVIA, 2023) This decoupling has an immediate freshness advantage, however, there are also now other levels of quality control that would not have existed before. Chunk sizes, fields in metadata, reranking thresholds, recency biases may be tuned and audited; and trace evidence may be logged, each answer may be accompanied by citations which anchor back to the original sentences referenced.

A diagram of a computer system

AI-generated content may be incorrect.

Figure 2 Concept Design Retrieval-Augmented Generation (RAG) system (OKTAVIA, 2023)

A diagram of a computer

AI-generated content may be incorrect.

Figure 3 Retrieval-Augmented Generation (RAG) Architecture (OKTAVIA, 2023)

A good RAG pipeline of admissions and visa material will start with corpus design. It is desirable that documents are normalized into clean text whilst maintaining structural cues e.g. headings, lists and tables that frequently encapsulate policy logic. Sentence or clause chunking is good to use since decisive conditions are usually short and syntactic; longer chunks obscure the line between different rules, shorter than necessary collapse the context to provide interpretation. The section titles, the URLs, the faculty names and dates should be kept as metadata since these are discriminative features during retrieval time. A dense retriever provides strong matching against mixed phrasing and synonyms, but in this area a neural reranked can prove particularly useful: many of these faculties recycle the same template language with one or two numeric variations, and a reranked trained to predict subtle relevance can reduce near-misses that would otherwise have to be manually reviewed. (Li, Wang, Wang, & Hung, 2025)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4 Summary of RAG Application in Education.

Recent handling is not a superficial touch up; it is a memory retrieval strategy. In the case where two passages are bound together by similarity, the one which has a more recent last-modified header or crawl timestamp should be chosen in the system. In the case of visa pages and scholarship calls, where recency is constitutive of validity, the UI should bring out a checked-on label and in the case of text changed in the near past bring out a brief diff history. Another fact of practice consists in contradictions. In case of discrepancy between a central page and a faculty page, the system ought to be sensitive enough to accept the discrepancy but default to the more authoritative source of the source class, but then show both passages to the reader so that they can understand why they should take a cautious position. (N´emeth & Tatrai, 2025) The behavior minimizes silent errors and encourages human escalation where it is needed in the institutional governance.

Sometimes there is a misconception of the role of the generator in an RAG system. It is not so that policy can be made but to write the evidence retrieved into a comprehensible and user-friendly explanation. In a bilingual setting, that, first, entails the retention of official English nomenclature with the provision of Chinese explanation and examples, and, secondly, it involves the formatting of figures, dates and clauses of the likes of no less than carefully. The composer ought to be advised to cite in-line using sentence or clause level to have each of the claims referable to the source. Reliability can still be improved further by a lightweight post-generation crosscheck: when a numeric claim is present in the answer, but no generated snippet, in the top-k, contains that number, the system should either re-run with more directive guidance or output a cautious summary that prioritizes the source text as plain verbatim. (Pampel, Martin, & Pado, 2025)

Diagram of a diagram

AI-generated content may be incorrect.

Figure 5 Retrieval-Augmented Generation for Dialogue (simplified).

It is in the updating of the corpus and the index that RAG has the multiplier effect of its operational strengths. Most program pages can afford to be refreshed weekly using light refresh, but volatile sources, scholarships and immigration, can be placed in a watchlist to enable ad-hoc re-ingestion. Versioning enables personnel to respond to the question of what has changed, and it is also possible to roll back an ingestion if a site accidentally publishes a mistake into the short term. They must have guardrails and metrics in operations as well. The provision of early recall at k and mean reciprocal rank at retrieval time, citation fidelity and policy-compliance accuracy on a curated benchmark when one has the answer give a falsifiable measure of trustworthiness. The quality metrics: median latency, 95th percentile latency, time spent in OCR or speech-to-text in situations where one or more multimodal inputs are present, and fallback rates when the reranked or generator is timeout should also be monitored along with the system behavior.

One does not just decorate an RAG system with transparency; it is what builds trust and then fixes it. It is essential that short easy to read verbatim snippet should be presented with anchors so that individual can open the source in a new tab. The time one last consulted the source is stamped on answers to give enough context of time. Retrospective logging of retrieval and composition traces in the background allows counsellors to audit on reasons why an answer was formed and to highlight corpus holes to ingest. Taking together these characteristics make RAG specifically desirable in policy-intensive agencies where bad advice can consequentially cost money and where organizational repositories are held responsible for the quality of the advice they bring to the surface.

### 2.1.3 Cross-lingual and multimodal advising patterns

Individual advising can rarely be perfect text prompts. Students transfer pictures of IELTS or TOEFL report forms; they copy snippets of scholarship announcements that wash out formatting ribbons into social media platforms; and they leave short voice messages about deposit receipt dates, or wording of documents to be translated. The parents are involved as well, sometimes they like to use the Chinese explanation of the situation whereas the sources of authority can be found only in English. There are immediate technical ramifications on these patterns of interaction. A credible system should not treat the images and the audio as an unusual deviation but as normal sources, which translate into textual secondary so as to enhance retrieval and also come up with answers that do not flaunt either one of the two languages and also one should not come out with paraphrases which blurs the actual policy terms. (Box, 2025)

Effective pipeline starts with correct recognition of documents. The ability to differentiate between exam report and offer letter or scholarship call is important since they comprise different salient fields. In the case of score reports, overall and sub-scores, test date, test type and candidate identifiers can all be extracted to retrieve the reported scores conditioned on the eligibility thresholds that often contain the so-called no band below rules. In the case of offer letters, the areas of interest change to terms, deposits, deadline dates, program codes and start terms. The keys which are required in scholarship posts include eligibility restrictions, grant amounts, deadlines, and links to a form. An OCR with high confidence and specific field extraction lowers the odds that the digits themselves are mangled and retransmitted onwards; with low confidence, the interface can show the cropped area plus the guess and solicit simple confirmation before the rest of the retrieval continues. (Ali, 2024) It only takes a few seconds to perform intervention, but it prevents a greater sequence of mistakes.

Sound requires a no less rigorous but less obvious way. Speech-to-text converts little vocal notes to text that can be mingled with pictures or past messages in one inquiry condition. People have language identification flags that signal the presence/absence of Chinese, English, or a mixture and that the system uses to direct generation accordingly. Unless the user chooses to keep them, transcripts must be kept only in temporary storage, and privacy-friendly defaults are the right moral practice and more operationally secure. Most of the users will take pictures in low-light conditions or clip a part of the picture; simple fixes of orientation and enhancement will boost OCR accuracy and, consequently, retrieve quality down its stream. (Saif-Ali, 2024)

Because of the need to apply bilingual reasoning, a system must continue to have a living glossary of names and aliases. Some of the phrases can be most appropriately left in English since they represent names of official institutions, whereas others enjoy Chinese gloss that dilutes none of the meaning and significance of the official term. Variants like taught master, coursework master and alterations of colloquial Chinese into formal (English) terms should be captured in the glossary as well as words with acronym expansions, such as CAS, COE and SEVIS. It ought to also code orthography and spelling variations in such a way that search becomes robust. Rather than needing to improvise a translation repeatedly, using optimization of retrieval and composition techniques, inclusion of the glossary makes verifiable phrasing the same every time. (AxcelerateAI, n.d.)

A diagram of a handwritten exam

AI-generated content may be incorrect.

Figure 6 Automate grading of handwritten (AxcelerateAI, n.d.)

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 7 Mapping PLOs/CLOs to course content (AxcelerateAI, n.d.)

The evidence should be reflected in the form of output. Given the source is in English, but the user has input the search in Chinese, it is quite likely to present a Chinese explanation to conserve the original English policy sentence in form of a quoted snippet to cite. Dates, amounts and limits can be made to fit a display that is difficult to interpret in a cramped display: ISO-like dates, or a month name format minimizes potential confusion; currency symbols must be clear with no ambiguity. It is preferable to pay special attention to ambiguity. Where the text of policy uses hedging verbs like normally requires rather than commanding it as hard rule then the answer should call out the use of a hedge and suggest escalation in unusual circumstances rather than going silent with an apparent hard rule.

Multimodal handling is part and parcel of privacy and governance considerations. Personally identifiable information usually appears in images and audio records. To be safer unless expressly instructed otherwise by the user, it is safer to process such artefacts in transient form, storing only sufficient of the derived fields to enable their reconstruction. To reduce the risk, sensitive areas can be blurred when thumbnails are kept being reviewed by counsellors. Analytically, the need to store raw media to benefit by use is reduced: aggregate statistics of artefact types, extraction success rates and correction frequency all serve to give sufficient feedback to pipeline improvement without generating new sources of data liability. Multimodality is not only an amenity in such a setup. It is the mediation between the natural way in which people communicate, and the form that knowledge must take in a system to be able to retrieve answers and justify them. Photographs of a scholarship clause turn into two dates and the dollar value pegged to a URL and retrieval is improved. The moment a voice note gets turned into a time-to-date sentence with a mention of deposit deadline; the system can put that date in perspective of program start conditions and indicate the urgency. The multimodal intake and cross-lingual competence, about extraction precision, privacy, and data presentation, provide the adviser with a sense of naturalness and reliability.

### 2.1.4 Operational constraints, performance and stakeholders

The success of an advising system is finally limited by the shape of operation. The features of its performance, maximum availability in the peak hours, maintainability, and governance will define whether students and staff will use the tool and whether the institution can rely on its outputs. Good service level objectives establish benchmarks and produce inducements to engineering selections. In the case of a prototype used in an academic term, a time in the order of five seconds in response to text-only queries and a larger but still predictable goal when OCR or speech-to-text occurs provides a realistic baseline. The latency at the 95th-percentile will need to be checked, since it indicates a peak-load experience; when median latencies appear, healthy users are hampered by long tails. Availability aims to be close to 99.5 percent with reasonable rate limiting safeguards the experience when there is a surge in the application and visa. The SLOs do not simply act as dashboards, but make decisions on catching, concurrency and graceful degradation. (Ferguson, n.d.)

The second pillar is observability. Retrieval-augmented systems are constituted by ingestion, indexing, retrieval, reranking and generation. Bottlenecks are erroneously blamed and potential solutions sent in a wrong direction when there is no cross-stage tracking. Instrumentation must thus capture timestamps across stages, add trace identifiers to requests and produce structured events when fall backs are illuminated. An operational suite of run-time evaluation statistics involves recall at k on a rotating benchmark, faithfulness of citation at sampled sessions, median and tail latency disaggregated by stage and the fraction of responses that transit fallback paths like “sources only” summaries when the generator runs out of time. It is because observability discipline allows diagnosing quickly when upstream providers change behavior of their models or when a site that one needs to pull stops ingesting due to a change in HTML structure.

Cost and capacity control are important in a student-fronting service that is subject to spikes. Popular-document embeddings may be cached and re-used; hot queries may take advantage of short-TTL catching of most-retrieved chunks keyed by a normalized signature. Hygiene of Indexes must be regular: partial re-indexing may be made once per day in the evenings to include minor updates, full rebuilding may be made once per week or on schema mutation. Where externally hosted APIs have quotas (rate limits, backoff), it is important to have rate limits; the system should degrade gracefully, instead of hard failure. Fallback hierarchies conserve utility: when the reranked is down, the dense retrieval can return safe snippets with references; when generation fails a budget, the system can show the best sources and emphasize sentences instead of just ending up with an error.

Table 2 SERVQUAL dimensions (Ismail, 2021)

A screenshot of a computer

AI-generated content may be incorrect.

Security and privacy are non-compromising in an environment that deals with identity documents, address information and financial obligations. Exposure is reduced by transport-level encryption, which signs upload URLs and scans attachments and redacts personally identifiable information in logs. Counsellor dashboards should be under access controls based on least-privilege principles and audit trail should record who accessed and why. Retention of data ought to be defined: the daily chats are allowed to be retained temporarily to improve the quality and the artefacts like score reports and offer letter can be made transient processing unless a user chooses to save them. When a deletion is requested, the system is expected to ensure removal in hot-storage, backups and analytics stores.

There are varying requirements of success between stakeholders that are to be harmonized. Speed, clarity and verifiability is appreciated by students and families; reduction of duplicate information requests, auditable refinement of the system with low friction is prized by counsellors; message consistency and visibility regarding the common points of mis-information that requires content revision are appreciated by the program teams; traceability, least-privilege access, resilience in the face of supplier outages is, of course, also appreciated by compliance and IT teams. No single system will be able to optimize all axes simultaneously, but it is possible to render many trade-offs explicit and controllable, by making the retrieval pipeline tunable and the evidence trail inspectable. The institution may e.g. choose to loosen reranked thresholds on peak counselling days with a banner that highlights verification but tolerates a little slower response time where accuracy on a new policy is mission critical.

Lastly, documentation and repeatable processes are necessary to have sustainable operation. Ingestion scripts, glossary management, benchmark construction and escalation protocols, should be regarded as living artefacts, kept under version control. The more you script and record these tasks the less tie the system to any individual heroism or the more impervious the system can be to personnel change or a change of suppliers. An effectively advised system is as much the product of what the advising operations do as of what the model options are; by paying attention to SLOs, observability, cost discipline and security and governance, the service adds a serviceable layer of institutional infrastructure instead of a fragile experiment.

### 2.1.5 Alignment with SDG 9 (Industry, Innovation and Infrastructure)

This project supports SDG 9 because it treats a study-abroad advising assistant as digital infrastructure ecosystem-enhancing, reliable, and inclusive rather than a one-time application. In practice, the service is built to provide precise, verifiably correct information in volume through a Retrieval-Augmented Generation (RAG) system that separates knowledge terms in the model and recombines them at response time with bibliographic data of reference. With the corpus refreshed separately to the generator, the platform acts as robust information infrastructure: when policies do change, they are automatically reflected in the answers with no retraining, and responses are all pegged to sources which can be viewed by the students and counsellors. Presented this way, the assistant contributes to SDG 9.1 by offering reliable access to trusted sources of information, SDG 9.5 by institutionalizing transparent AI practices that can be built upon by others, and SDG 9.c by expanding the access to information and communication services, via mobile-first and bilingual design. (Burgenland)

The concept of reliability is operationalized in terms of understandable service-level goals and engineering required to achieve the goals. The platform aims at deterministic latency of text-only search and bound boundaries of tail latencies when images or speech are being approached and availability goals are used which sustain during periods of application and visa peak. End to end observability (ingestion, FAISS, reranking neural, and composition) enables visibility of latency and point of failures that can be repaired instead of smoothing out failures. Graceful degradation maintains the service useful when under stress: in case the generation consumes a budget, the system will produce sourced excerpts with last checked stamps, rather than failing hard; in case reranking is momentarily unavailable, dense retrieval will still yield conservative passages that can be proved correct. Light nightly refreshes, watch-listed re-ingestion of volatile pages, and versioned storage produce recency as a first-class property and allow personnel to explain what changed and when. Collectively these practices transform resilience as aspiration-to one that can be measured in behavior. The inclusion is achieved through matching the pipeline of how the applicants communicate. Five files of IELTS or TOEFL outcomes, provisional offers and scholarship provisions are turned into first-class inputs, OCR-ed and subject line to heads that rationalize post-retrieval and render eligibility argument sharp. Speech-to-text excerpts short voice messages as a collection of searchable text without impairing the informality of mobile communication whilst allowing consistency to be applied. Responses retain official English program names, but where useful clarify in Chinese, and dates and figures are set in a format which should reduce the chance of being misunderstood on small screens. Read-aloud of long answers and low-bandwidth mode to use with low connectivity help the service reach users who have limited or problematic devices and access. Privacy-first defaults, such as ephemeral processing of images and audio unless users choose to save, least-privilege access to counsellor tools and redacting personally identifiable information in logs mean inclusion does not come at the expense of trust. (Iatrellis, Samaras, & Konstantinos, 2024)

A white and orange logo

AI-generated content may be incorrect.

Figure 8 SDG 9

The SDG concept of innovation is realized in a tangible, repeatable RAG blueprint. The project bundles ingestion processes, chunking heuristics, metadata schemas, index configurations, reranked pilotage, composition prompts and evaluation rubrics such that peer units could clone the stack to shift to neighboring domains like scholarships, immigration or professional licensing. Retrieved passages and composition traces are revealed through the use of a counsellor dashboard and transform the staff into co-developers that have the capability of marking gaps and suggest to add accompanying language to the corpus without a thorough knowledge of ML. Measurement finishes the loop: a compact scorecard indicates dependability (uptime, stage-broken latency), quality (citation-fidelity and policy-accuracy spot checks), inclusion (perspective of bilingual and after-hours sessions, mobile success rates), diffusion (time to onboard a new corpus, reuses), and efficiency (cache hit rates, fraction of queries served by smaller generators courtesy of retrieval). Data collection is underpinned by governance of course: ethics approval, express consent and delete pathways, supplier-outage playbooks. Combined, the decisions produce an assistant that can serve as a long-term investment to SDG 9: resilient access to reliable counsel, increased participation based on language and modality, and a scalable framework that will make responsible AI more economical throughout the institution.

## 2.2 Similar System

### 2.2.1 Dify

A screenshot of a computer

AI-generated content may be incorrect.

Figure 9 Defy Main Page

#### 2.2.1.1 Introduction

Dify is an open-source AI app building platform that offers a visual Chat flow/Workflow builder tool and a high-quality first-class Knowledge (RAG) component. You describe ingest -retrieve, reran, answer-escalate as the nodes in a canvas rather than writing all the orchestration code yourself, and then execute it with in-built model management and observability. The docs refer to Dify as Backend-as-a-Service + LLMOps where Chatflow is conversational apps and Workflow is automation/batch jobs. Its knowledge tab literally sees the operations of a RAG pipe (upload/ingest, chunk/preview, index, retrieve, cite) openly, and provides the option of attaching other knowledge bases. It is also possible to insert an Agent node such that the LLM can request tools (e.g. OCR/ASR) to act independently at some point in a flow

#### 2.2.1.2 Advantages

Time-to-prototype is quick: within minutes you could have a citation-backed assistant standing up and the ability to iterate with non-ML stakeholders right in the interface. Knowledge offers source management options, metadata filters, retrieval knobs, and a switch to permit use of citation/attribution which is consistent with your transparency need. Dify also presents end-to-end recipes, such as an end-to-end tutorial with Milvus as the vector DB that can be useful in production-scale indexing. Since flows can trigger tools during the conversation, it is easy to insert OCR or STT prior to retrieval in your multimodal application.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 10 Dify Workspace 1

A screenshot of a computer

AI-generated content may be incorrect.

Figure 11 Dify Workspace 2

#### 2.2.1.3 Disadvantages.

Some low-level control is traded in with the node model. In case you require experimental retrieval internals (e.g. custom scorers, stacked rerankers, exotic metadata features), custom blocks may be written, or you may script beyond the canvas. Sophisticated behaviours occasionally depend on community conventions, e.g. when citation is not enabled in a particular template, you should explicitly enable attribution in the setup of the app.

#### 2.2.1.4 Techniques Used

Dify utilises a standard RAG stack: ingestion and chunking (including a preview step), vectorisation to an external DB (Milvus/Qdrant/pgvector), vector search with improved search (and reranking), and LLM generation while producing in-lines. The external knowledge connectors and agentic tool use are underpinned by flows, hence enabling you to commence multimodal pre-processing before grounding.

#### 2.2.1.5 Highlights

Governance speed to production. It provides you with a visual, auditable pipeline working with a citation-first RAG advisor and allows you to experiment quickly with counsellors in the UI whilst being able to drop in custom nodes to handle ChineseEnglish glossaries, OCRASR, or policyaware post-processing.

### 2.3.2 Haystack

A screenshot of a computer

AI-generated content may be incorrect.

Figure 12 Haystack Main Page

#### 2.2.2.1 Introduction

Haystack (by deepset) is a production end-to-end Python framework very RAG search. It provides dense and hybrid retrieval modular components, rerankers, readers /generators, tracing, and deployment. The official tutorials describe the process of setting up a first RAG system and point out hybrid retrieval in particular as a best practice (i.e. combining sparse (keyword/BM25) and dense (embedding) retrievers) usually followed by a ranker on top. This translates reasonably to admissions/ visa content for which keywords are important but near-duplicates and semantics must be ranked more intelligently.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 13 HayStack WorkSpace

#### 2.2.2.2 Advantages

The advantage of Haystack is fine grain control. Then you can sparse+dense clause level chunk, inject section-title metadata, combine the sparse and dense, add a cross-encoder reranker to solve near identical templates-and even wire recency-aware tie-breaking, all explicitly in a graph of nodes. To explain this, recall-maximising Deepset guidance and Deepset blog posts highlight why recall is what reranking and multi-task learning are most interested in; why reranking increases precision in RAG; and why tracking in the framework is useful to measure these decisions underload.

#### 2.2.2.3 Disadvantages

As you gain control, you own ingestion jobs, glossary control, multimodal extractors (OCR/ASR), hosting, CI/CD and dashboards. No GUI as a platform; early days development will have teams building slowly without Python ops, versus a visual builder.

#### 2.2.2.4 Techniques used

Haystack adds dense + hybrid retrieval, FAISS/ANN vectored backends, and cross/bi-encoder re-rankers, before piping to a generator. It has a pipeline DSL where you can specify branched flows like, “OCR > field extraction > bilingual normalisation > dense retrieval > hybrid merge > rerank > generate with citations” which suits your multimodal, cross-lingual requirements. The docs also demonstrate the situation on how the source documents may be revealed in retriever in form of citation.

#### 2.2.2.5 Highlights

Clear, testable RAG pipeline control policy- When pages that address policy are differentiated by only one number and you need to make decisions about recency and citation fidelity at scale.

### 2.2.3 Rasa

A screenshot of a computer

AI-generated content may be incorrect.

Figure 14 Rasa Main Page

2.2.3.1 Introduction

Rasa a conversational-AI framework which centers on dialogue management (intents, slots, forms) which integrates RAG via Enterprise Search. Internally, the same assistant which supports structured turns, you can add a connection to a vector store (Faiss, Milvus, Qdrant or custom) and call retrieval+generation when an informational question is raised. The official setup instructions tell you how to enable the EnterpriseSearchPolicy, and where to direct it toward your vector DB as well as setting up the LLM/embeddings provider.

2.2.3.2 Advantages

Rasa excels when you need only few profiles (degree level, IELTS band, budget) to be gathered by your assistant and it is necessary to fill in the answers with references. Since the RAG policy may be invoked wherever within a flow, deterministic checklists may be interchanged with explanations, which are based on documents. Rasa also supplies, for example, reps and tutorials (e.g., Faiss + Gemini) and accommodates many different vector backends, so integration paths are practical to production.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 15 Rasa Function

2.2.3.3 Disadvantages

Rasa Pro includes the most robust RAG features, so it introduces licensing/deployment issues. Really low level retrieval experiments (custom scorers, offbeat metadata properties) might need more detailed component work than in code-focused RAG collections. The conversion stack may end up being weighed down more than it should be unless your assistant consists mostly of Q&A and not dialogue.

2.2.3.4 Techniques Used

Rasa For a two-step RAG within the policy layer, connect to a vector store (Faiss/Milvus/Qdrant) and retrieve relevant chunks, condition an LLM to provide a grounded response; or retrieve extractive responses, where LLMs are not necessary. It is configured in config.yml (policy, retriever, embeddings/LLM) and the policy can be executed any other action in a turn.

2.2.3.5 Highlights

Guided conversation + practical responses in a single column: gather correct fields using / slots, then call on RAG to support judgments with quoted, footnoted passages a pretty fit to triage eligibility and document check.

### 2.2.4 Comparison of Similar System

Table 3 Table Comparison of Similar System 1

A table of information

AI-generated content may be incorrect.

Table 4 Table Comparison of Similar System 2

A white table with black text

AI-generated content may be incorrect.

Table 5 Table Comparison of Similar System 3

A close-up of a white card

AI-generated content may be incorrect.

### 2.2.5 Conclusion

Collectively the three approaches of Dify, Haystack and Rasa are three consistent (but much different) approaches to providing a RAG-based advisor. Dify is a framework: it shrinks orchestration, hosting and monitoring into a GUI template enabling the town builder to deploy a ingestvariable -> retrievevariable ->rerankvariable -> answervariable -> escalate pipeline in a matter of minutes and in less than a day they can present their pipeline with citations and logs. That pace is worth stakeholder buy-in and iterated improvement, but at the cost of reduced fine-grained control over retrieval internals and plumbing of evaluation. Haystack is a framework: it optimises control over chunking, hybrid retrieval, neural reranking, recency biasing and source exposure, which directly support a heavily policy-driven field in which near-duplicate templates are differentiated by a figure and source exposure and citation fidelity are non-negotiable. The cost is Engineering lift- You have ingestion jobs, bilingual normalisation, multimodal extractors, deployments, observability. Rasa is a stacked dialogue with builtin RAG hooks: it is excelled when the assistant needs to gather little to no profiles and have specific intents/slots/forms, and then ground answers with docs within the same conversation. Such harmony of text stream and evidence is attractive to eligibility triage and checklist advice paradigm, yet the best features of textual retrieval lie behind Rasa Pro and extremely lower-level ranking experiments necessitate additional nesting.

The backbone of this project, whose objectives are fresh, cited answers instead of volatile admissions/visa content, Chinese English code mixing, multimodal intake (screenshots/voice) and SLOs that were clear had to focus more on precision and auditability. Haystack is consequently most appropriate to the essential RAG service: a FastAPI microservice can proxy /ingest, /query, /rerank, /answer endpoints; FAISS (or Milvus/Qdrant) can store sentence/clause-level chunks along with metadata about section titles; a bi- or cross-encoder reranker can defuse templated pages; and recency clues can break ties, all of which provides verbatim, anchored snippets to cite. Dify has a remaining role to play: it is the quickest method to get a stakeholder-facing pilot operational, confirm corpus coverage, and collect usability suggestions; its graph may also serve as the model of the ultimate pipeline and expose monitoring early on, provided that the production route moves to code in order to permit fine-level control. When structured dialogue happens as a first-order requirement, such as when getting degree level, GPA, IELTS bands, or budget prior to triggering retrieval, then Rasa becomes the optional wrapper yielding a reliable, auditable process that alternately composed steps with rich explanations.

To summarise, use Haystack (production retrieval backbone: control and measurement), Dify (rapid prototyping and demonstration: speed and governance) and Rasa (profile capture and checklists) when a structured dialogue materially improves the outcomes. This breakdown of labour satisfies Chapter-1 goals (transparent RAG, bilingual and multimodal processing, FastAPI delivery) and with the SDG 9 promises of this project (resilience access (SLOs and graceful degradation), inclusive design (bilingual, mobile-first, multimodal) and diffusion of innovation (a modular stack that others can re-use).

## 2.3 Technical Research

### 2.3.1 System Requirements Analysis

A high-performance laptop (AMD Ryzen 9 7845HX (Zen 4, mobile H-series) and 32 GB DDR5 memory) is used as the development environment. It is a comfortable specification to serve a Python + React stack, and many of the services can run simultaneously: the editor, local API server, Node.js dev server, Docker containers, and monitoring tools, without paging. High multi-thread performance brings faster dependency build and faster parallel workloads like background indexing or building containers images with the help of the CPU, whereas DDR5 puts JSON/HTTP-intensive workloads to responsiveness. The storage needs to be an NVMe SSD (512 GB+) to ensure package installations, node\_modules and docker layers are snappy; when storage is scarce, using Docker with pruning of it build cache and using per-project virtual environments prevents bloat. The reliability of the networking is relevant in that the model endpoints and package registries reside in the cloud, nor is the broadband connection between them or accurate system time (required to complete a TLS handshake) guaranteed.

In Windows 11, development is simplified by developing using WSL 2 with Ubuntu 22.04 LTS to have Python execute against a Linux userspace but continue having access to desktop tooling. Docker Desktop connects Windows and WSL2 to lets you build and run Linux containers on your machine and deliver images without modification to a cloud VM. To run a FastAPI service behind an ASGI server (Uvicorn/Gunicorn) and a Node-built static frontend, or reverse-proxy both with Nginx/Caddy in the eventuality of future business needs, the cost of a 2-4 vCPU, 8-16GB RAM, 40-80GB SSD on Ubuntu 22.04 VM should be adequate to provide lightweight pilot hosting. Since the project deals with the personal data, the baseline consists of automatic security updates, host firewall, TLS termination, and the basic observability (structured logs + metrics). Your hardware is plenty capable; the selected OS stack will allow reasonable parity between dev and deploy and Docker will allow repeatable builds.

### 2.3.2 Programming Language Choosen

The app is divided in a very clean manner: Python handles the backend API and retrieval-generation/orchestration functionality, React provides the responsive, bilingual interface to the web. Such division corresponds to the problem profile. Backends need to use HTTP requests, normalise queries, invoke external AI services, manage retrieval and citation packaging, and enforce latency budgets; front-ends must manage the ergonomics of text, image and short audio clips, show citations effectively, and switch back and forth between Chinese and English without triggering page reload.

A graph of numbers and letters

AI-generated content may be incorrect.

Figure 16 Top Programming Language Choosen

Because it is easy to work with, Python fits perfectly in the service layer that binds retrieval-augmented answering. Its async I/O model using asyncio and production ASGI servers (Uvicorn or Gunicorn+Uvicorn workers) is able to deliver predictable throughput on I/O-bound workloads in which majority of work is waiting on network calls to embeddings, rerankers, and generators. The expressiveness of the language makes the pipeline legible: payloads are clear when using request models Pydantic; clean endpoints are provided by routers; query normalisation, evidence selection, post-generation citation-checks and bilingual formatting all have small, testable functions. The ecosystem is useful in Python too to reach solid composable operations faster: httpx or aiohttp to make resilient outbound calls with timeouts and retries; tenacity to make backoff work; structlog or the baked in logging stack for JSON logs; Prometheus exporters to monitor latency and errors. The packaging is minimal, usually uv/pip + requirements.txt to avoid faff, or Poetry where other crates are uncompromisingly demanding of lockfiles. In case the project later needs to use CPU-intensive preprocessing (such as OCR feature extraction or audio segmentation), native bindings to optimised libraries ensure everything is in-process; otherwise, such processing may need to be outsourced to provider APIs to utilise local compute resources.

Client-side, React has both componentised building blocks of UI and a developed ecosystem of state management, routing and internationalisation. A study-abroad adviser enjoys the smack of foreseeable input streams, legible citation display and low-sliding language switcheroos. React declarative rendering makes these issues easy to address: components are able to render bilingual text that retains programme names in English, to highlight extracted numbers/dates and to display collapsible source snippets that are expanded on tap. Format errors can be intercepted up front (e.g. file type/size screenshots or audio clips) using client-side validation, and fetch wrappers can normalise and standardise API requests by adding auth headers, and error management. Performance-wise, the build pipeline (Vite or CRA alternatives) produce a small static bundle that a web server or CDN can serve; during development the hot-reload dev server enables looping quickly. Components being small and testable allows easier accessibility to be maintained; React Testing Library and Playwright should assert that major flows can be used with a keyboard and screen reader which is in line with the inclusion objectives in the project.

Python/React interoperability is clean. The backend has a relatively small, well-typed REST interface (or JSON-RPC when that is a better fit), returning not just the end result, but anchored evidence blocks in structured form. That schema is consumed by the front end (directly), which leaves the client thin and does not need to replicate logic. Since both stacks are everywhere, CI/CD is a normality: it is possible to lint, test, construct the Python picture, construct the React package, and deliver artifacts to a storage facility on each blend using GitHub Actions. Within a production environment, such as can be seen in Kubernetes, one reverse proxy can be used to proxy /api/ to the ASGI backend and / home to the static frontend frontend, offering a clean deployment of two containers each of which may be scaled separately.

### 2.3.2.1 Justification

The choice of python as the backend is the establishment that it provides the most compelling trade-off of developer velocity, ecosystem breadth, and operations maturity, in the context of an I/O-bound RAG service. The async nature of the language and ASGI servers serve numerous requests in a small, legible codebase; libraries are first-class to enable resilient outbound calls and structured observability; and, finally, containers are portable and have small images. Just as vital, the Python SDKs are brought to the forefront by most model providers and retrieval libraries which minimize friction in integration. The trend is React due to its very good support for stateful, componentised UIs, and excellent internationalisation and testing. It allows you to quickly provide answers that show citations first with bilingual controls, to treat multimodal entries sensitively and to remember accessibility. Python + React in combination reduce time-to-feature and at the same time maintain the reliability and transparency that your IR can expect.

### 2.3.3 IDE Choosen- Visual Studio Code

VS code becomes the IDE of both stacks, and this brings cohesive tooling and pares down cognitive switching. In the case of Python, the official extension gives you linting (Pylance), type checking, debugging, test discovery, support of notebooks and switching environments, whereas the Docker extension allows you to build an image, inspect container and logs, as well as debug running services without ever leaving the editor. In case of React, TypeScript, ESLint/Prettier and first-party Git support ensure quality code without unwanted comments. Ad-hoc use of the API is fast to test with the REST client or Thunder Client extensions; Live Share allows fast pair debugging when required. VS Code on Windows opens the Linux filesystem right in the editor and launches terminals in Ubuntu, so your Python process, package cache and Docker engine will be native-Linux even though you do most of the editing from Windows. Run backend, run frontend, compose up actions are all encapsulated to use in workspaces, tasks and launch configs to keep the project one-click reproducible to markers and other people working on it.

2.3.4 Browser Choosen – Google Chrome

The major used development and user test browsers are Google Chrome. Chrome DevTools offer a full package to allow the type of diagnostics this project requires: the Network panel shows request waterfalls, HTTP/2 multiplexing and cache hits; the Application panel explores local storage and service worker state when offline operations are tested; and the Security panel checks TLS and mixed-content issues. The Console and Sources tabs are helpful to test edge-case rendering in bilingual layouts and Lighthouse is helpful with quick checks of the performance and accessibility to remain responsive to mobile users, even when connected to a modest network. Since the React app is framework-agnostic at run time, to verify behaviour in the various major user agents, the same tests are cross-checked in Microsoft Edge (Chromium).

2.3.5 Operating System Choosen – Window 11, Ubuntu, Docker

The daily work is done on Windows 11, which provides desktop environments and PC support, although the code is able to run on a WSL 2 (Ubuntu 22.04) to be comparable with the target server system. Pathing, permissions and package behaviour remain Linux-true (no more works on my machine drift). In both of these senses, Docker is first-class: in development, develop scripts contain compose files to bootstrap the React development server and ASGI backend in parallel; in CI, the same Dockerfiles construct immutable images; and in deployment, the images run in an identical manner on an Ubuntu VM. The operating system of the server is Ubuntu 22.04 LTS because of its stable OS, LTS security release cycles and the wide community support. This pairs up to reproducible builds, ease of rollback and low operational friction, and your Ryzen 9 + 32 GB DDR5 hardware gives a fast build time and local test runs even with many containers running concurrently.

# Chapter 3 Methodology

## 3.1 System Development Methodology

A system development methodology is the systematic scheme that defines the collection of the requirements, the creation of designs, development and testing of the software and delivery of value to the users. The selection of an appropriate methodology will define the speed at which the team responds to the changing environment, the emerging risks brought to the surface as well as the evidence of quality. Since the project involves both research (the RAG tuning) and engineering aspects (a reliable, cited advisor), the methodology needs to allow iteration, testable milestones and frequent stakeholder input.

### 3.1.2 Methodology Chosen and Justification

The selected approach: **AgileScrum (iterativeincremental SDLC).**

Agile-Scrum would be appropriate on this project since requirements are anticipated to change (pages needing policy written, bilingual/multimodal UX needs refinement.), and model quality increases via repeated attempted improvement (chunking, re-ranking, latency tuning). Sprints time-box working increments, so within a few months to six months only, a working text-only of RAG MVP in the form of counsellor input and test users, then bilingual UI and end-user multimodal intake and observability can verify textual accuracy, citational fidelity and ease of use, as a discrete or unified product in detail. Scrum ceremonies may be simplified on a single developer (weekly planning/review, daily checklist); however, “Definition of Done” may include gates such as P50 ≤ 5 s response, citation fidelity 90 0 on a curated set, zero broken links in citations, bilingual toggle confirmed on key flows. The associated cadence along with SDG-9 follows the IR assessment and the SDG-9 framing (reliable, transparent, inclusive digital infrastructure).

A diagram of scrum

AI-generated content may be incorrect.

Figure 17 Scrum (GeekforGeek, 2025)

**The first alternative is waterfall (linear SDLC).**

Waterfall focuses on specifications upfront and stages which come one after the other (requirements, design, implementation, testing, deployment). It is the most effective in case of stable requirements and well-known technology. The knowledge base and the RAG pipeline parameters in this project cannot be locked down early (since evidence accumulates) and locking designs would cause discovery to take place later and inflate rework. The integration of Waterfall is also late, pointing to the lack of a demonstrable and referenced prototype that would allow confirming the quality of retrieval and the UX of using bilingual, which would happen with actual artefacts (IELTS screenshots, scholarship clauses).

**Alternative 2: Spiral (risk based iterative approach).**

The development of spiral structures consists of series of processes of risk analysis, prototyping, and evaluation. It is powerful in scenario of high technical/safety/compliance risks and many stakeholders. Spiral would allow the RAG risks (hallucination, latency spikes, privacy) to be modelled in an explicit, formalised manner, however radiation of process (formal risk assessments and gates) is more than can be carried out in a solo, semester-long FYP. Scrum sprints that contain a brief task called a risk review and acceptance tests linked to citation fidelity, privacy defaults and SLOs can absorb much of what is beneficial about Spiral- without the formality of entire Spiral phases.

Table 6 Comparion of system development methodology

A table of text on a white background

AI-generated content may be incorrect.

#### 3.1.2.1 Conclusion

According to the comparison, the Agile-Scrum fit will best accommodate the realities of the project: (I) the fact that requirements are subject to change due to the evolution of program/visa policies and the feedback of bilingual UX stakeholders; (ii) the incremental iterative tuning of RAG will require the use of measurable gates (citation fidelity, SLOs of latency) that will be tuned in each sprint; (iii) the need to demonstrate it to stakeholders over time to build trust and evidence collection to support SDG-9; and ( The waterfall has a risk of late rework because of its rigidity, and I think the risks cycles in Spiral would involve overhead that are not commensurate to the size of the project. Consequently, the current chapter will follow up with an Agile-Scrum SDLC, which can be done in brief sprints and provides deliverable, referenced increments of the advisor.

### 3.1.3 Activities and Processes in Each Phase of the Agile Scrum

A diagram of a software development process

AI-generated content may be incorrect.

Figure 18 Each Phase of the Agile Scrum

#### **3.1.3.1 Requirement**

Record the actions required of the assistant and how success will be defined. Sketch user-stories (student, parent, counsellor), discover data sources (program/visa pages, scholarship posts), and establish non-functional targets.

**Major deliverables:** product vision, prioritized backlog, product of backlog, and SLOs (e.g., citation-fidelity 90%, P50 5s on text / 8s with OCR/STT), privacy defaults (ephemeral media), ethics note.

#### **3.1.3.2 Design**

Draw the end-to-end architecture and test strategy. Differentiate RAG flow (ingest > chunk sentence/clauses with headings > FAISS > dense retrieval > Qwen rerank > cited generation), API contracts (Fast API), UI wireframes (React bilingual toggle, citation blocks) and security (TLS, role access).

**Major deliverables:** a lightweight architecture diagram, an API schema, some UI mockups, data/glossary schema, and a sprint plan of the following increment.

#### **3.1.3.3 Development**

Create a working slice thinly with every sprint. Develop Fast API endpoints (/ingest, /query, /answer), React views (chat, citations, language toggle), glossary/alias support, and, where in scope, OCR/STT processing prior to querying. Add logging and metrics.

**Major products:** increase in running in Docker, unit/integration tests, and new docs. Definition of Done: links to citations pointed, shown when answers are volatile, last-checked.

#### **3.1.3.4 Testing**

Automated checks need to run tests and test with real artefacts. Parsing, chunking, formatters (cover unit), recalling (Recalk/MRR), answering (citation-fidelity and policy spot checks), performance (P50/P95 by stage), and simply accessibility.

**Primary deliverables:** pass/fail vs SLOs test Report, and fixed defects prior to the release. Gates fail; building fails when not met.

#### **3.1.3.5 Deployment**

Containerize the backend and frontend; deploy to Ubuntu (Docker/Compose). Set up secrets, TLS and monitoring. Keeping rollback (previous image tag) and index hygiene (nightly light re-index; weekly full).

**Notable deliverables:** release notes, live URL, metrics summary (uptime, latency, fidelity) and incident playbook (degrade to sources-only in case providers slow/fail).

#### **3.1.3.6 Review**

Test by demonstration the increment with actual pages/shots/voice notes; gather the comments on the accuracy, clarity of references and bi-lingual expressions. In retro, take note of 1 thing to start/stop/continue (e.g. increase OCR confidence threshold, bump rerank k). Do update backlog and establish the following sprint objective.

**Important deliverables**: stakeholder responses -> new stories, minor new processes, and a definite goal for the next sprint.

## 3.2 Data Gathering Desing

### 3.2.1 Data Gathering Method

A questionnaire is a standardized question instrument which is employed to find information in individuals often on a reciprocal scale. It may be closed (tick-boxes, multiple choice, fill in scales such as Likert scales), open (short written responses) or mixed. Questionnaires may be issued via the internet, which is an attractive approach due to the cheap, quick and simple analyses, particularly when it is applicable to require normalized information of broad populations. They work best when you want to understand how opinions, experiences or behaviours vary across a group, compare subgroups, or collect inputs on priorities to be provided on features. It has limitations which are bias in response (individuals may address what is appealing), poor response levels, and limited levels of depth in as far as interviews are concerned. Fundamental design hints: ensure there is a defined purpose, keep it brief, employ readable, unambiguous text, avoid two-part questions (What is your satisfaction level and frequency of…?), provide an intro line/consent and run a pilot-test with at least two anyone to identify those items that may be confusing.

### 3.2.2 Design of Questions

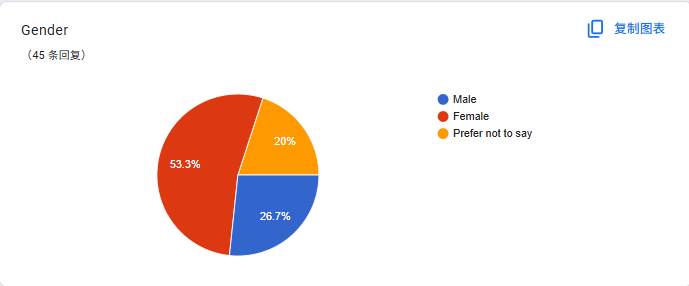
**Section A: Demographic**

|  |  |
| --- | --- |
| Questions | Objective |
|  | To collect demographic data to examine the possible variance in study-abroad counselling requirements or those of technology adoption, on gender basis. |
|  | To categorize the respondents into age group so as to do the analysis of correlation between age and preferential counselling methods or chat features. |
|  | To determine the nationality of respondents such that feedback should represent the target population, (Chinese outbound students) and help analyzing the cultural or policy-specific needs. |
|  | To establish the level of education or academic level of the respondents that can affect the kind and extent of counseling information they need. |
|  | To identify how acquainted the respondents were with the chatbot technologies, and how much experience they had working with these technologies that could be used to evaluate their minimal expectations and learning curve they would be considered to experience when dealing with the proposed RAG-based counselling system. |

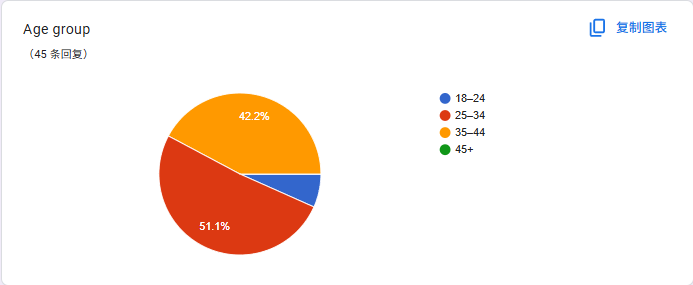
Section B: Perceptions of a Potential Chatbot

|  |  |
| --- | --- |
| Questions | Objectives |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To determine the urgency and frequency at which the respondent requires access to programme information and therefore the necessity of a rapid response system such as the proposed RAG chatbot |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To establish the value of making recommendations with regard to considering potential areas where automation of comparing tools can be useful in making recommendations of how difficult or time consuming the process can be perceived as. |
| 手机屏幕截图  AI 生成的内容可能不正确。 | To establish the level of significance that the respondents attach to source transparency and credibility in line with the objective of the project to provide answer founded on citations. |
| 图形用户界面, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To understand whether there is any demand to use the personalised eligibility assessment features that can be automated in the chatbot with programme-specific criteria. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To assess the perception of a respondent of the possible time-saving due to the adoption of an AI-based counselling tool. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To measure the perceived usefulness of the chatbot on the support of reducing the number of programmes that are subject to the user criteria. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To evaluate the effect of transparency of sources on user trust and decision making. |
| 图形用户界面, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To establish the significance of bilingual competency in addressing both the needs of users in terms of accessibility and comprehension. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To understand how users demand multimodal features of input to support personalised advice. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To uncover the perceived necessity of systematic support in the study-abroad process. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To quantify intentions to use the chatbot as an initial source of information. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To ascertain the user attitude towards provision of friendly assistants like chatbots and independent verification. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To test the user expectations on chatbot usage ease and learning curve. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To determine possible positive peer word-of-mouth promotion. |
| 图形用户界面, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To measure user comfort with cross-lingual or code-mixed interaction and informing chatbot language-handling design. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To understand whether a conversation interface is favored to that of manual browsing, which represents the possible motive of user adoption. |
| 图形用户界面, 文本, 应用程序, 聊天或短信  AI 生成的内容可能不正确。 | To define privacy issues which can become barrier to using document-upload features, so that such issues can be taken care of during system design. |
|  | To understand the level to which AI users are concerned about the accuracy and bias of responses to the AI generated content, of which quality control can be directed. |
|  | To determine the user apprehension against data retention, to inform privacy and data management policies. |
|  | To assess the resistance to AI adoption, it is necessary to determine whether some users would not employ chatbot-based counselling regardless of its functionality. |

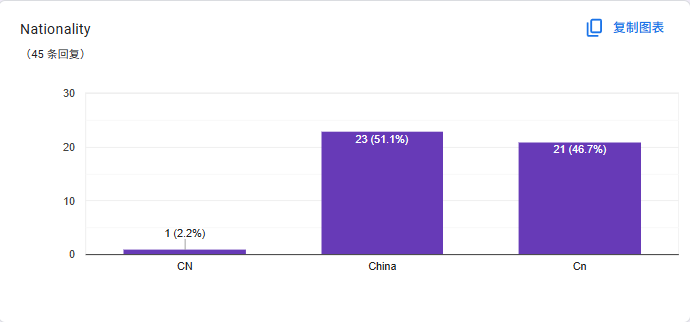
### 3.2.2 Data Analysis



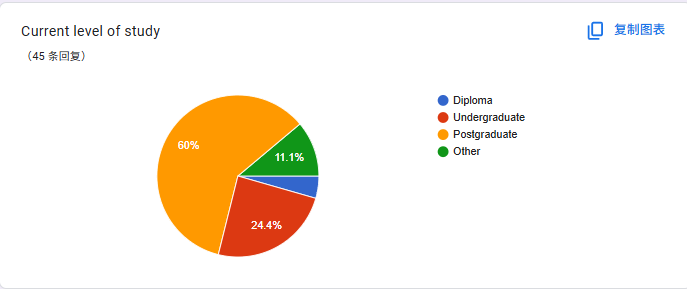
Analysis:26.7 percent (12) of total respondents (45) were male, 53.3 percent (24), were female and 20.0 percent (9), preferred not to answer. It is thus woman-dominant with a significant undisclosed population that ought to be put into consideration when interpreting any of the differences as related to gender.



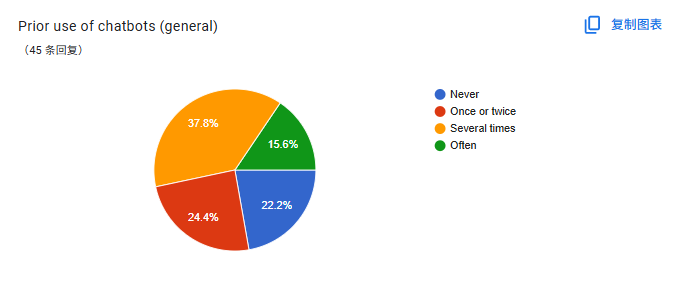
Analysis: The age-group represented by the majority of respondents was 25 34 (51.1, 23), then 45+ (42.2, 19). There were only 6.7 percent 18-24 (3) or none at 35-44. In total, the data pool favours mature applicants (>=25 years: 93.3%) potentially influencing needs and expectations.



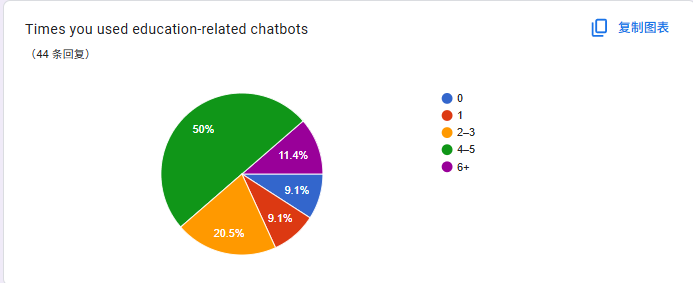
Analysis: Each response shows the same nationality was typed but with different formatations of 51.1 percent (23) having China, 46.7 percent (21) having Cn and 2.2 percent (1) having CN. Once case/label has been normalised, the sample is, in effect, 100 per cent Chinese nationals and therefore, the findings can only be projectable towards the same.



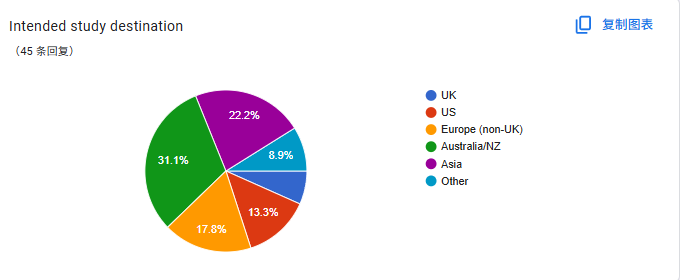
Analysis: At 60.0% (27), Other was the most common followed by Undergraduate 24.4% (11), Postgraduate 11.1% (5) and Diploma 4.4% (2). That high Other share indicates that a lot of respondents fall outside normal categories (e.g., working professionals, language-school applicants) or are between stages.



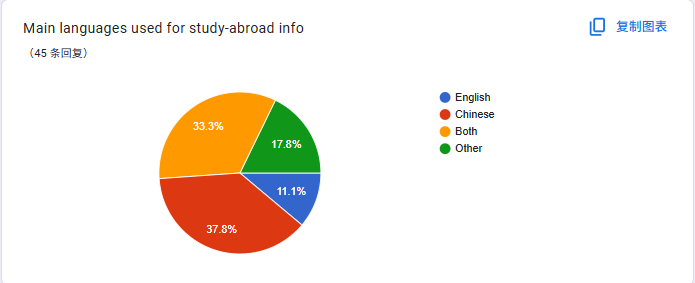
Analysis: 22.2 percent never used chatbots (10), 24.4 percent once or twice (11), 37.8 percent several times (17), 15.6 percent often (7). There exists clear majority (77.8%) that possesses some prior exposure, which assids adoption as a result in the absence of any exposure.



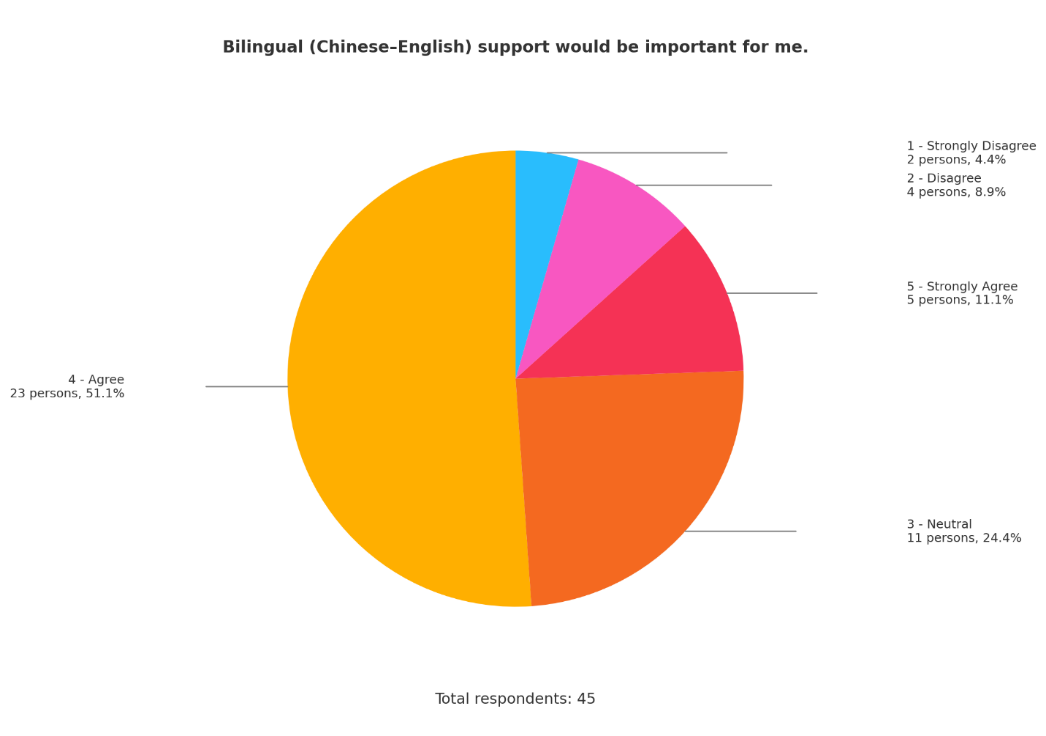
Analysis: Forty four percent (44.0%, 22) reported 4-5 uses and 20.5% (9) reported 2-3. 6+ was reported by 11.4 percent (5). A smaller proportion was used 0 times (9.1%, 4) or 1 time (9.1%, 4). One non response left the base at 44. In general, familiarity with bots that involved study is moderate to high.



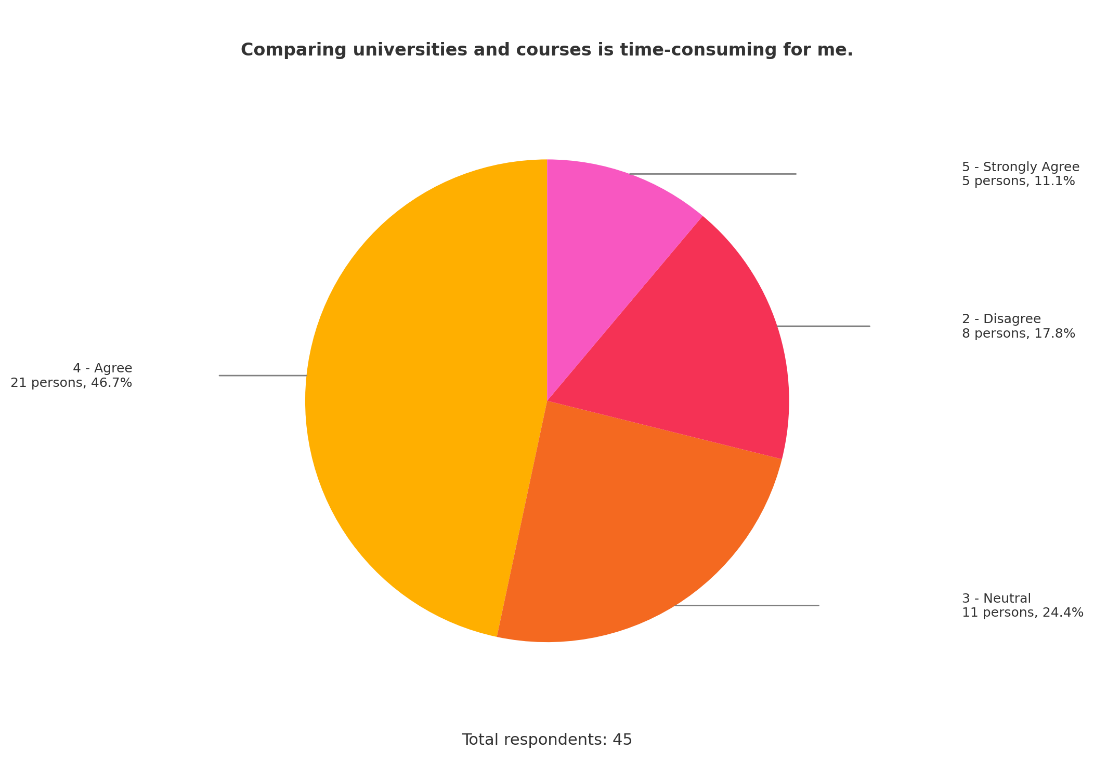
Analysis: Most preferred (31.1%, 14) were Australia/New Zealand and next came Asia (22.2%, 10), Europe non-UK (17.8, 8), the US (13.3, 6) and the UK (8.9, 4). The rest was 6.7percent (3) or in other words, described as other. There is a concentration of demand in ANZ and Asia and less in UK/US.



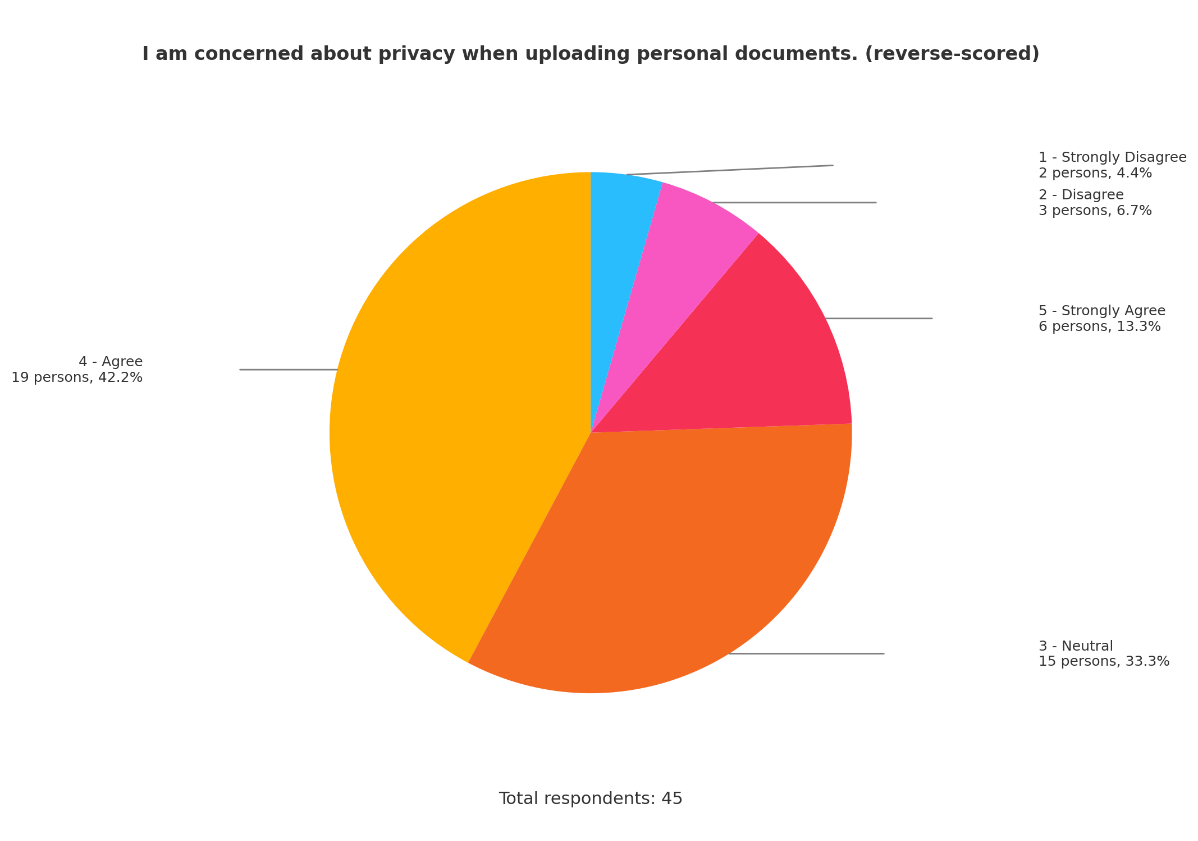
Analysis: The main part( )of the respondents use Chinese comprises to 37.8 percent [17], Chinese and English is 33.3 percent [15], other languages are 17.8 percent [8] and English is 11.1 percent [5]. This justifies a bi-cultural (ZH-EN) design including obvious sourced references in support of an inter-language information search.



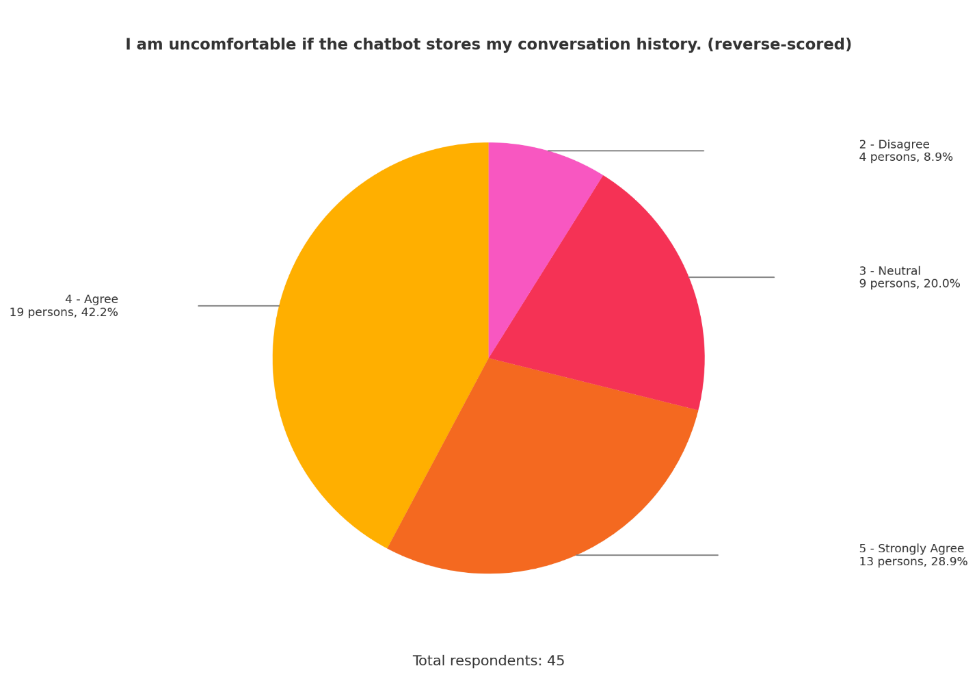
Analysis: Agreed/strongly agreed, 62.2% (23 + 5 = 28); neutral, 24.4 percent (11); and disagreed/strongly disagreed, 13.3 percent (4 + 2 = 6). This implies that there is evident need in the bilingual assistance.



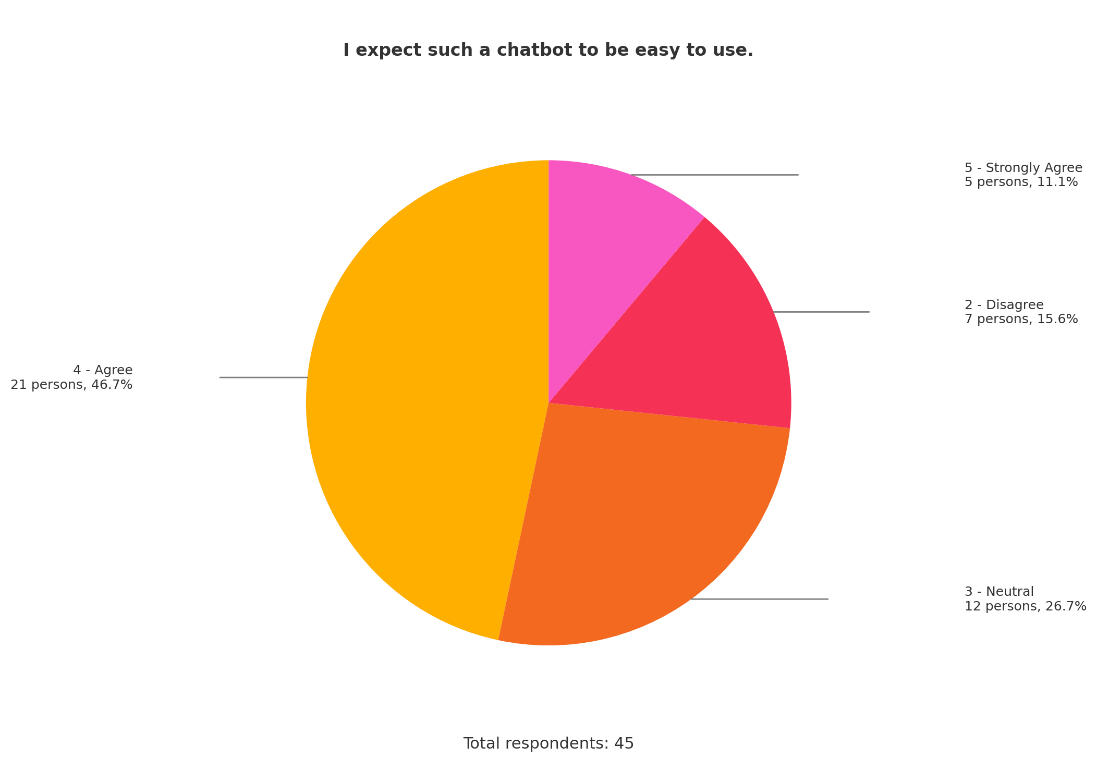
Analysis: 57.8 said that they agreed/strongly agreed (21 + 5 = 26), 24.4 percent were neutral (11) and 17.8 disagreed (8). The majority of respondents report the tasks of making a comparison to be time-consuming, which entails assistance in making a comparison.



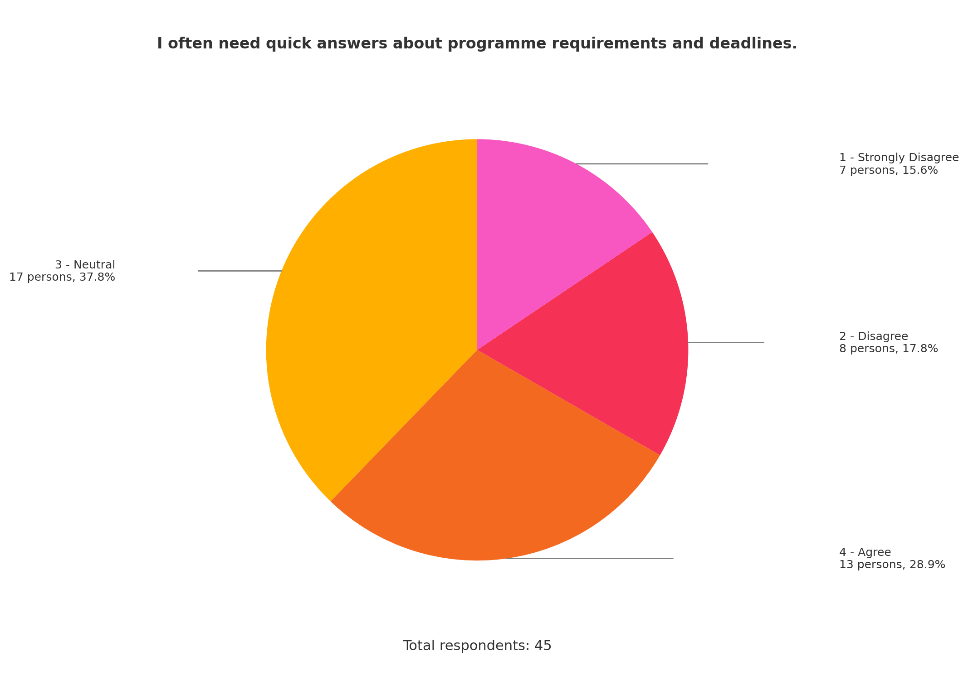
Analysis: 55.5% responded agree/strongly agree (19 + 6 = 25), 33.3 responded neutral (15), 11.1% responded disagree/strongly disagree (3 + 2 = 5). Where there is any assurance of privacy and well-defined consent flows are thus necessary.



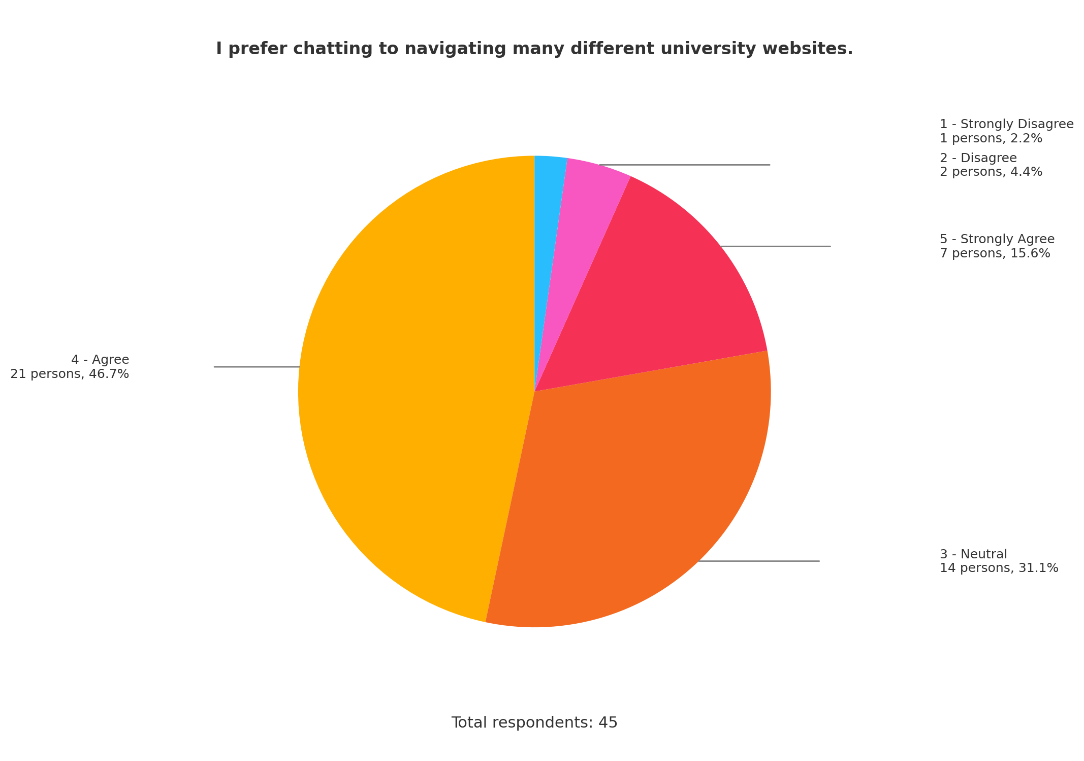
Analysis: 71.1 percent (19 + 13 = 32) agreed/strongly agreed, 20.0 percent (9) were neutral, and 8.9 percent (4) disagreed. Customers want to have fewer data being retained and would like to have the opt-out option.



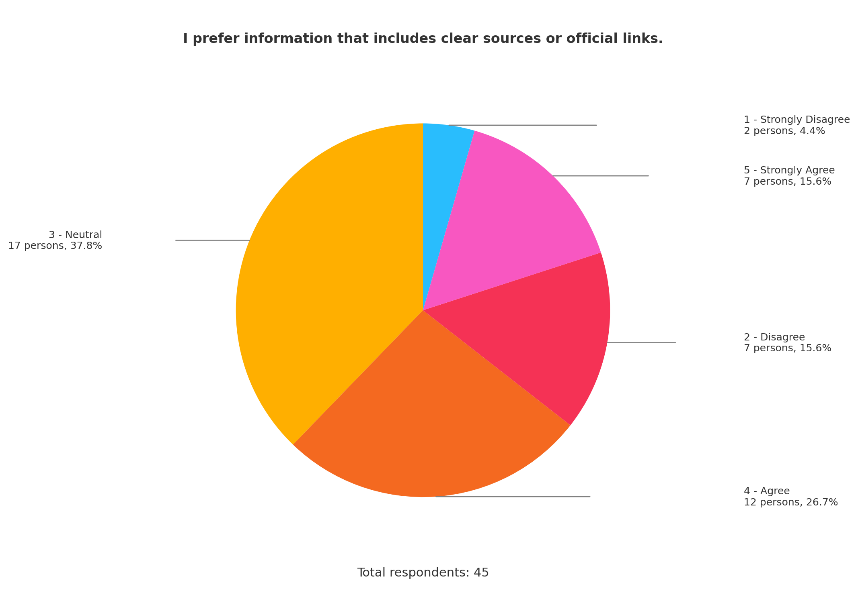
Analysis: 57.8% were either agreed/strongly agreed (21 + 5 = 26), 26.7 percent were neutral (12), and 15.6 percent disagreed (7). Expectations related to usability are great; an interface needs to be easy and self-explanatory.



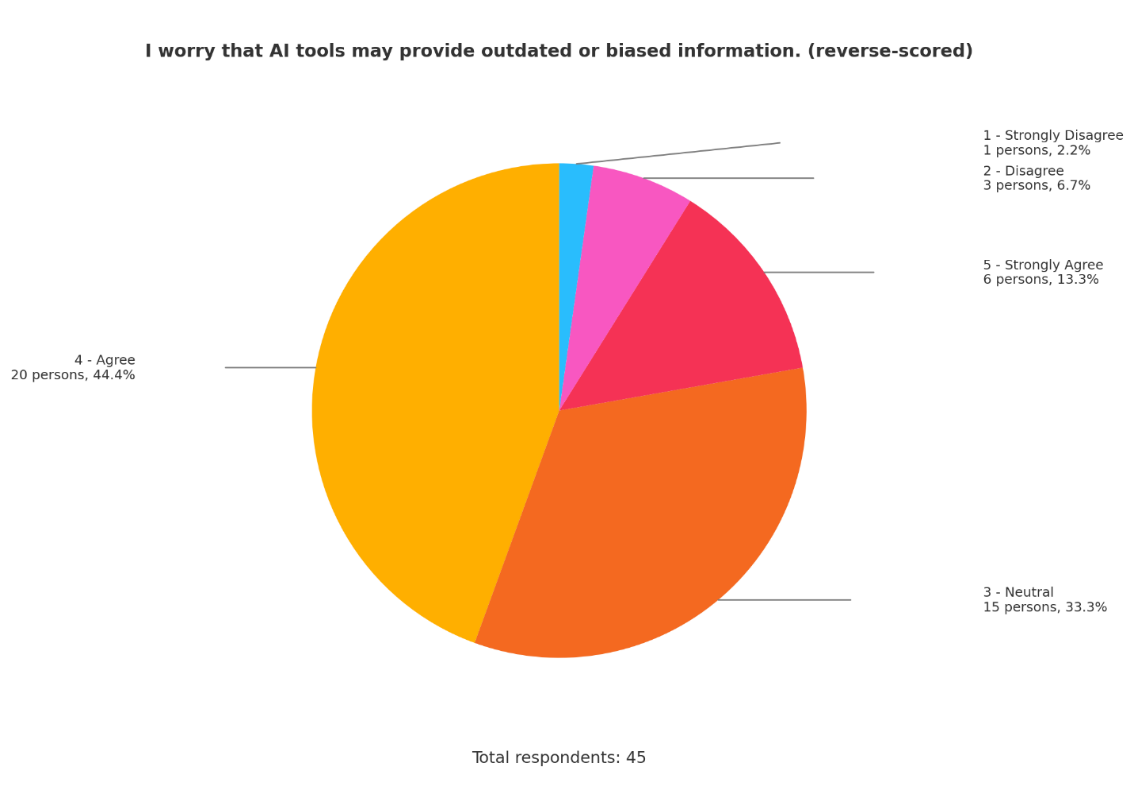
Analysis: 28.9 percent were in agreement (13), 37.8 percent left it neutral (17), and 33.4 percent disagreed/strongly disagreed (8 + 7 = 15). A demand on answers that are quick, is ambivalent, but probably indicative of varying phases of application urgency.



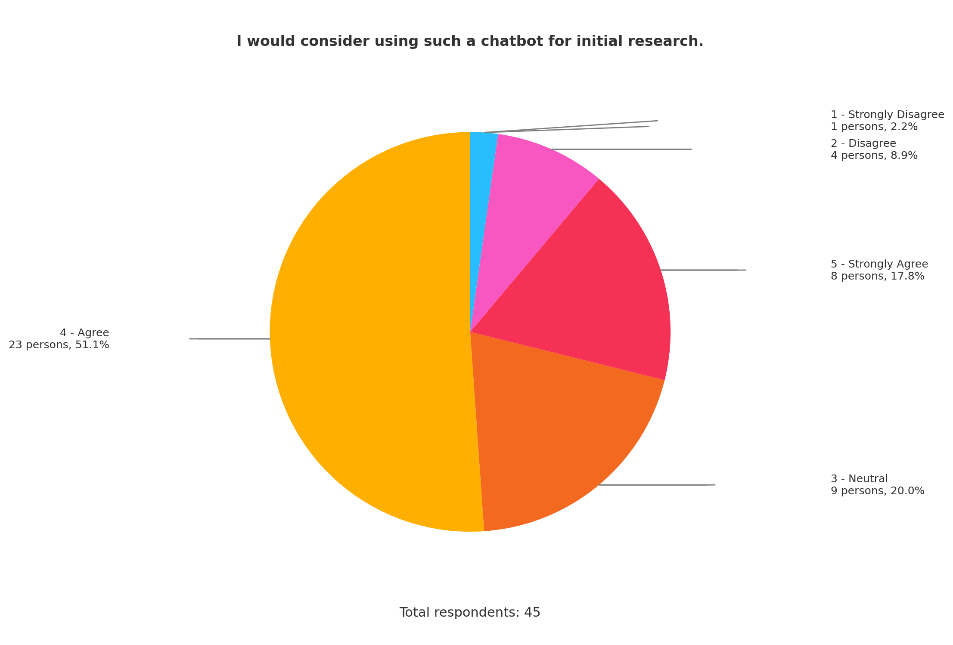
Analysis: 62.3 percent agreed/strongly agreed (21 + 7 = 28), 31.1 percent were neutral (14) and the remaining 6.6 percent disagreed/strongly disagreed (2 + 1 = 3). There is a general preference of a chat interface to the use of multi-site browsing.



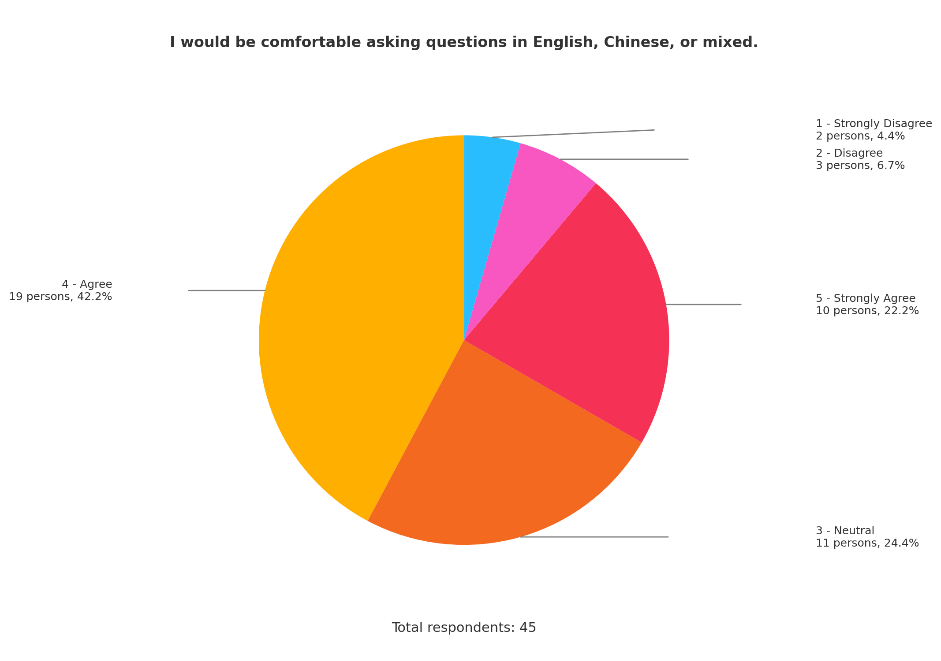
Analysis: 42.3% agreed/strongly agreed (12 + 7 = 19), 37.80 were neutral (17), and 20.0% disagreed/strongly disagreed (7 + 2 = 9). There are several respondents who attach importance to citations; a large group of neutrals implies that transparency education about sources could be effective.



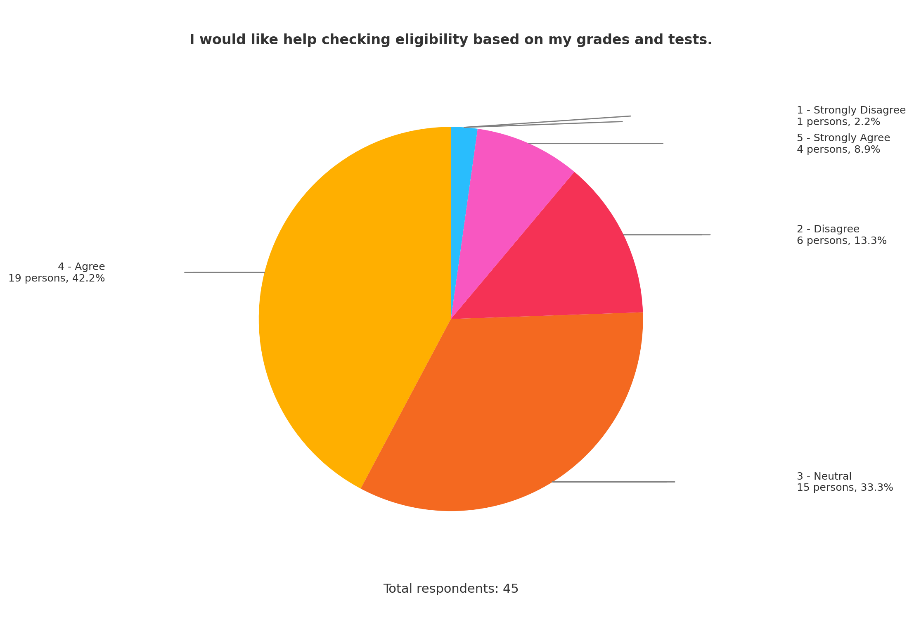
Analysis: 57.7 percent assented/strong assent (20 + 6= 26), 33.3 percent neutral (15), and 8.9 percent dissented/strong dissented (3+ 1 = 4). Reliability is an issue; date-stamping and worse-judgments will be necessities.



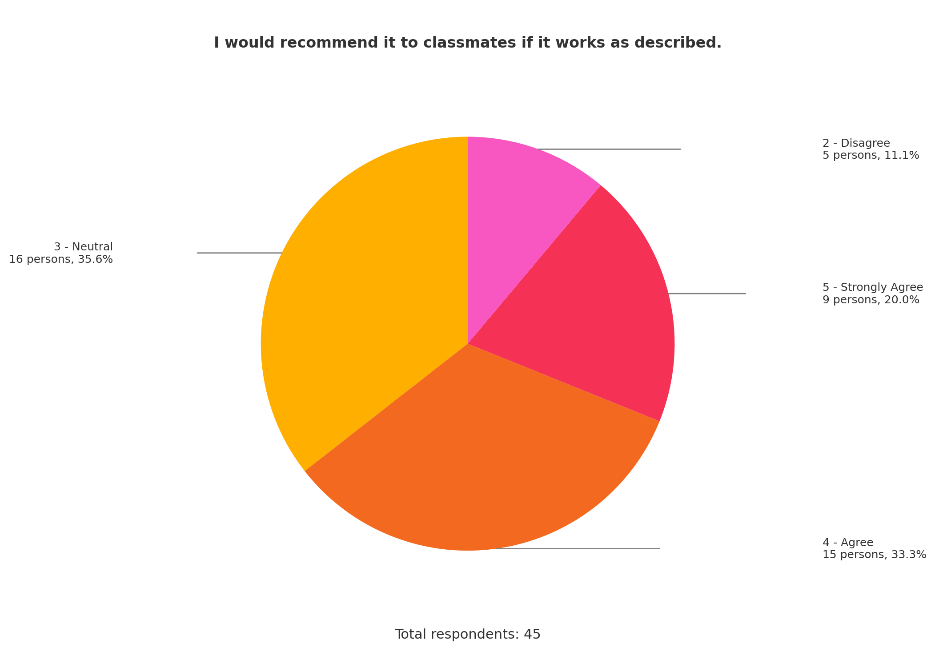
Analysis: 68.9 percent agreed/strongly agreed (23 plus 8 = 31), 20.0 percent were neutral (9) and 11.1 percent disagreed/strongly disagreed (4 plus 1 = 5). The majority of the respondents would consider using the chatbot in the exploration phase.



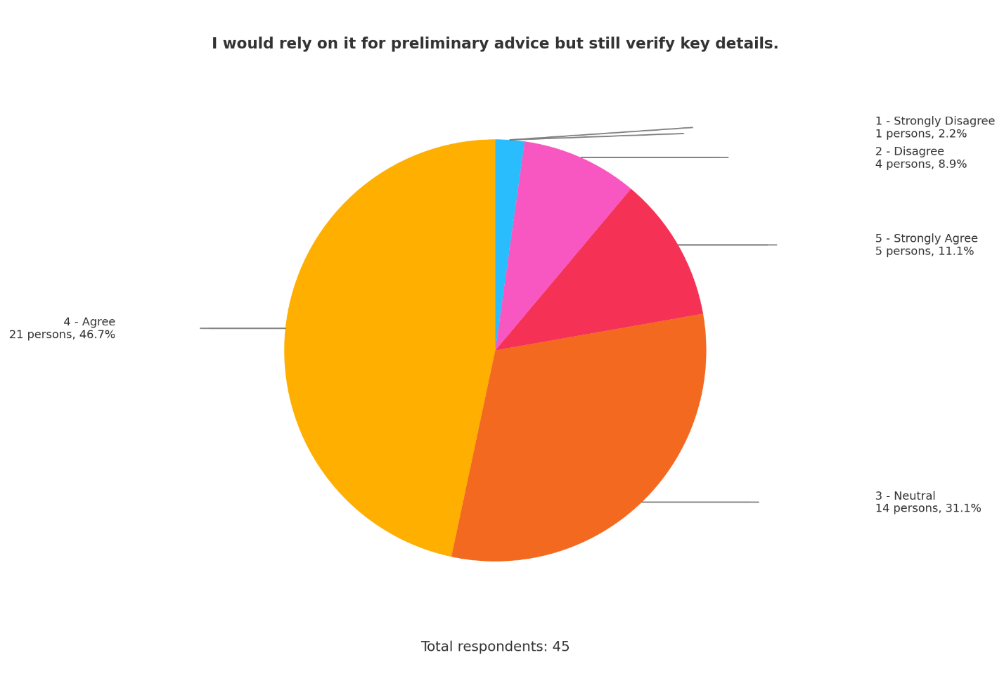
Analysis: There was high concurrence, with 64.4 percent agreeing or strongly agreeing (19 + 10 = 29), 24.4 percent being neutral (11), and 11.1 percent disagreeing or strongly disagreeing (3 + 2 = 5). Bilingual UX is supported by the fact that users feel quite comfortable with code-switching.



Analysis: Agreed/strongly agreed = 51.1(19 + 4 = 23), neutral = 33.3(15), and disagreed/strongly disagreed = 15.6 (6 + 1 = 7). Automated checks in determining eligibility are in demand.



Analysis: 53.3 percent agreed/strongly agreed (15+9=24), 35.6 percent were neutral (16), and 11.1 percent disagreed (5). Recommendation willingness is good and performance based.

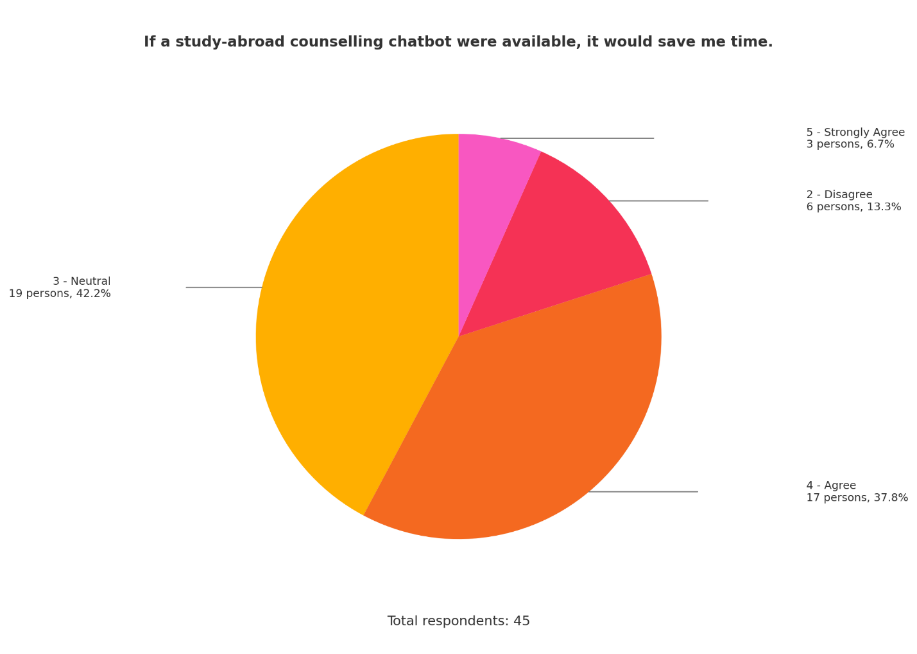


Analysis: 57.8 percent responded with agreement/strongly agreed (21 + 5 = 26), 31.1 percent with neutral (14) and 11.1 percent with disagreed/strongly disagreed (4 + 1 = 5). It is viewed as an initial step by its users, who are to be verified

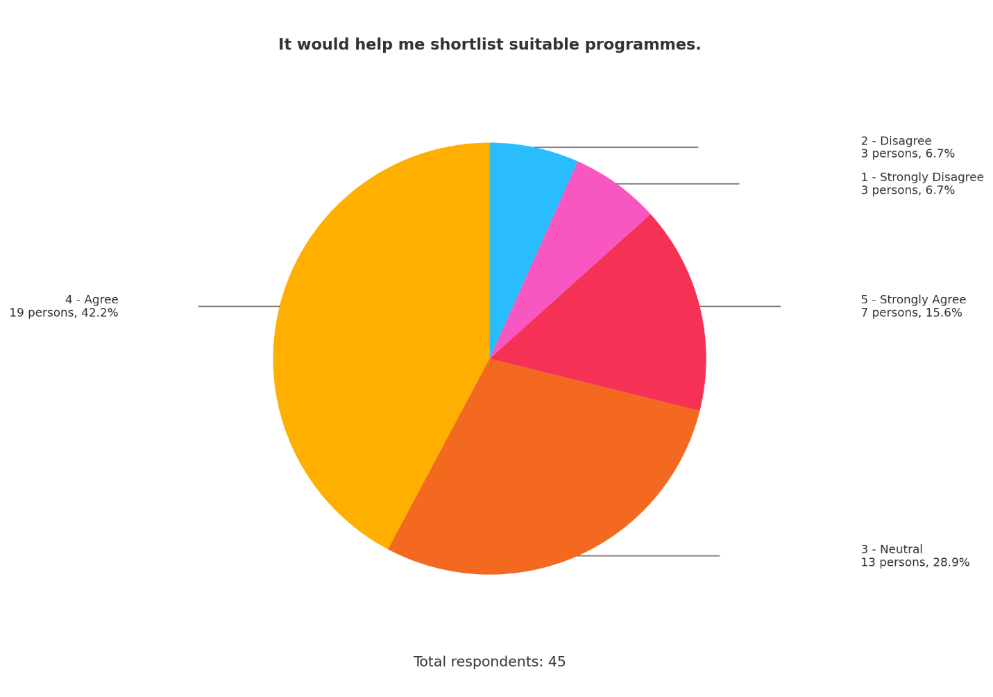
A colorful pie chart with text

AI-generated content may be incorrect.

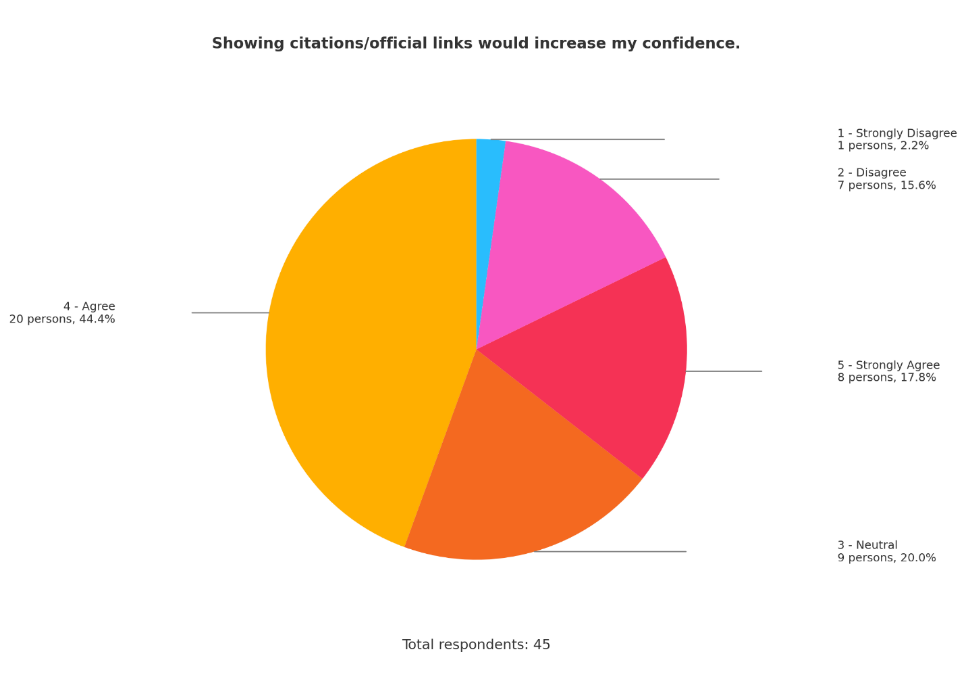
Analysis: 71.1 percent agreed/strongly agreed (24+8=32), 20 percent were neutral (9), and 8.9 percent disagreed/strongly disagreed (1+3=4). Human-only counselling is still prized by many, therefore it is recommendable that a hybrid model will be used.



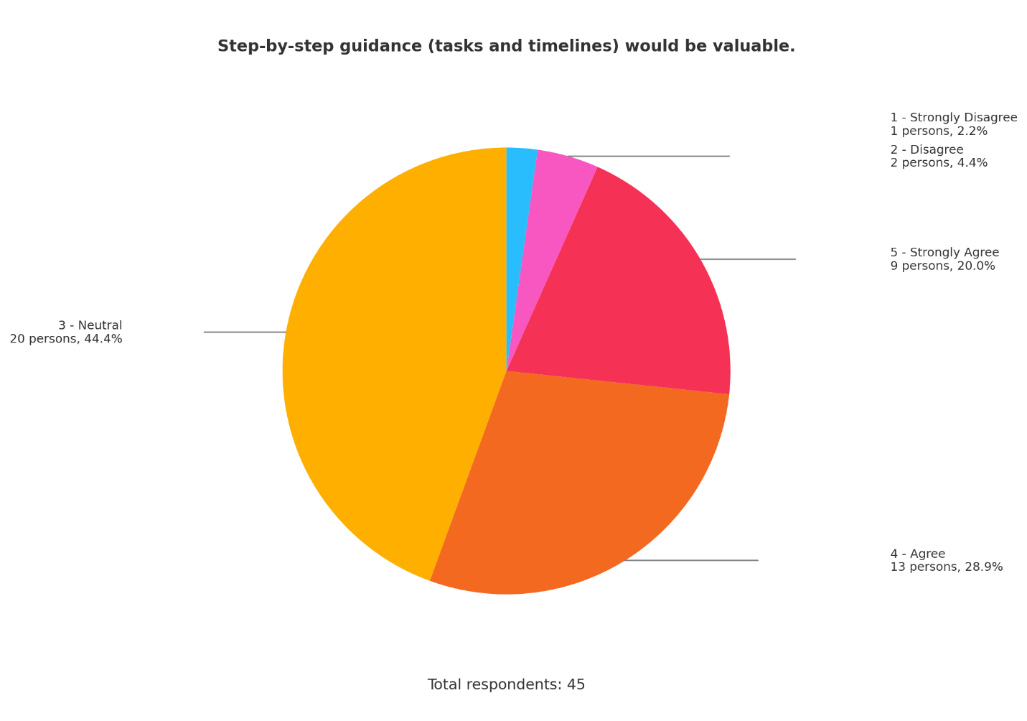
Analysis: Nobody disagreed /strongly disagreed or were neutral, 44.5 agreed/strongly agreed (17+3=20) and 42.2 percent were neutral (19). Time-saving is deemed to be positive, although not universal, probably due to different levels of application.



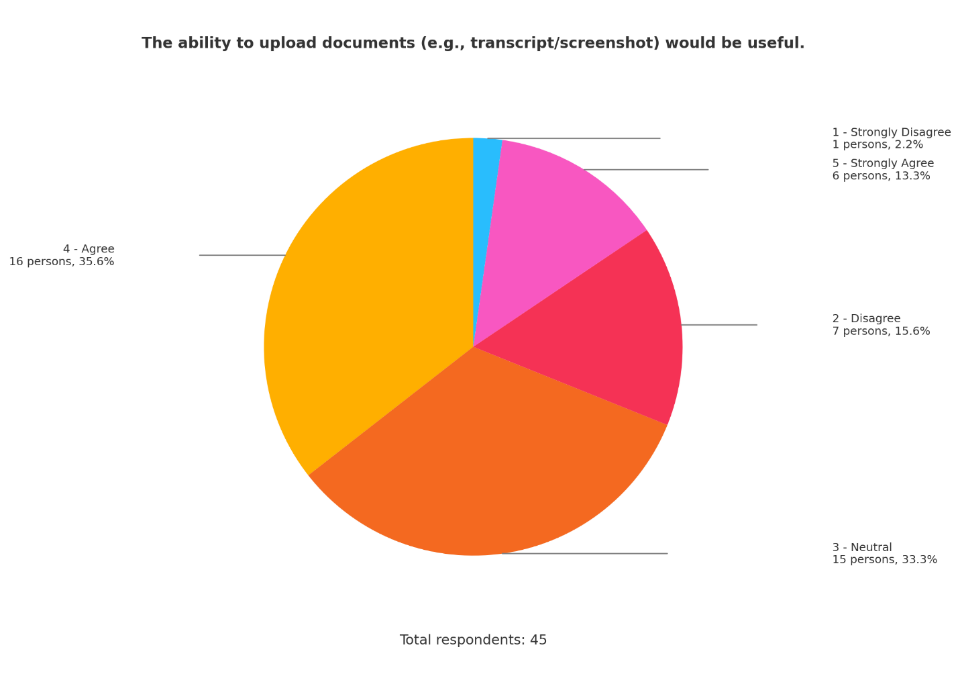
Analysis: 57.8 percent of the employees said, they agreed/strongly agreed (19 + 7 = 26), 28.9 percent neighted (13) and 13.4 percent disagreed/strongly disagreed (3 + 3 = 6). Most respondents find shortlisting aid useful.



Analysis: 62.2 percent agreed/strongly agreed (20 + 8 = 28), 20.0 percent were neutral (9) and 17.8 percent disagreed/strongly disagreed (7 + 1 = 8). Authoritative responses, which are supported with citations, are one of the trust triggers.



Analysis: 48.9 percent agreed/strongly agreed (13 + 9 = 22), 44.4 percent were neutral (20) and 6.6 percent disagreed/strongly disagreed (2 + 1 = 3). Moderately high proportions of persons would express preferences of structured checklists, with one in two being undecided.



Analysis: 48.9 percent said yes/strongly yes (16 + 6 = 22), 33.3 percent were neither/nor (15), and 17.8 percent were no/strongly no (7 + 1 = 8). The service of document upload is widely perceived as beneficial, which supports the strategies of file management with security.

# Chapter 4: Design

## 4.1 Introduction

Intro……

## 4.2 Design

### 4.2.1 Use Case Diagram

### 4.2.2 Use Case Specification

## 4.3 Class Diagram

### 4.2.4 Activity Diagram

4.2.4.1 admin\_manage\_jobs

### 4.2.4 Activity Diagram

## 4.4 Interface Design

This study abroad system was designed to develop in SPA user interface.

### 4.4.1 Login Page

图示

AI 生成的内容可能不正确。

The user is required to key in valid username and password in order to visit the main page for accessing the functions. Also, register(create account) and reset password function was provided.

### 4.4.2 Register Page

图形用户界面

AI 生成的内容可能不正确。

The user needs to key in valid username and password, then key in the confirmed password, reset question and reset answer in order to create account. The reset question and reset answer was designed for the user to reset the password of an account when the password is forgotten.

### 4.4.3 Main Page (Left Sidebar collapsed)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

The user can click on expand/collapse button to raise left sidebar, or click on menu to access the same function. Click on delete button to delete the current conversation(session), or click on rename to rename the conversation. The log out button will sign out the current account. Upload file button will trigger the upload function for user to send files to the chatbot. Click on send button when the content is not empty to start chatting. When the chatbot is generating content, send button will transform to stop button in order to call stop streaming function(The chatbot is using SSE, which means the content will be sent by server to client using streaming method).

## 4.4.4 Main Page(Left Sidebar Expanded)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

When the left sidebar is expanded, the workspace, system, search conversation, archived conversations, conversation list and setting button will be displayed. The chat button is designed to distinguish between chatting and other functions(no actual functionality present by this button). The admin console was only available for admin type account. Key in a existed conversation name in search conversation input box to search conversation. The pinned conversation will be presented in archived conversation box area. The existed conversation will be presented in conversation slot area. Setting button was designed to modify the user account information. When the user click on new chat button, a new conversation page will be created but not sending messages to the server until the user actually send the content to chatbot.

### 4.4.5 Main Page(Slot part 1)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.6 Main Page(Slot Part 2)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.7 Main Page(Slot Part 3)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.8 Main Page(RAG control expanded)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.9 Main Page(Delete conversation/session triggers)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.10 Main Page(rename conversation/session triggers)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.11 Language Selection Dropdown Box

表格

AI 生成的内容可能不正确。

### 4.4.12 Log Out/Sign Out Confirmation

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.13 Active Conversation Management (Sidebar)

图示

AI 生成的内容可能不正确。

The user can choose to rename, pin, archive or delete the selected conversation by dropdown box on left sidebar.

### 4.4.14 Pinned/Archived conversation management

### 4.4.15 Sign In As Admin Prompt (User Account Type Only)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.16 Admin Console (Status)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.17 Admin Console (Metrics1)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.18 Admin Console (Metrics2)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.19 Admin Console (Retrieval)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.20 Admin Console (Slots)

图片包含 图形用户界面

AI 生成的内容可能不正确。

### 4.4.21 Admin Console (Sources)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.22 Admin Console (Jobs)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.23 Admin Console (Prompts)

图形用户界面

AI 生成的内容可能不正确。

### 4.4.24 Admin Console (Opening)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

### 4.4.25 Admin Console (Assistant Profile)

图片包含 图形用户界面

AI 生成的内容可能不正确。

## 4.5 Test Case Preparation

## 4.6 Execution

### 4.6.1 Screenshots

#### 4.6.1.1 Login Page

图形用户界面, 应用程序

AI 生成的内容可能不正确。

Due to authentication mechanisms, the user will be required to login in order to access the functions. The user will have to key in valid username and password to login.

#### 4.6.1.2 Register Page

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

In register page, the user is required to key in valid username and password for basic registration process. Besides, reset question and reset answers are also required for reset password function.

#### 4.6.1.3 Main Page (User)

图形用户界面, 文本, 应用程序, 聊天或短信

AI 生成的内容可能不正确。

Once successfully logged in, the user can access the main page of the application. This case of main page is for user type account. The function area is separated into two parts: left sidebar and main chatting area.

#### 4.6.1.4 User Settings (Part 1)

图表

AI 生成的内容可能不正确。

#### 4.6.1.5 User Settings (Part 2)

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

#### 4.6.1.6 User RAG Controls

图形用户界面, 应用程序

AI 生成的内容可能不正确。

图形用户界面, 应用程序

AI 生成的内容可能不正确。

4.6.1.7 Language Selection (UI)

图形用户界面, 文本, 应用程序, 聊天或短信

AI 生成的内容可能不正确。

#### 4.6.1.8 Session Slots (Part 1)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

#### 4.6.1.9 Session Slots (Part 2)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

#### 4.6.1.10 Session Slots (Part 3)

图形用户界面, 应用程序

AI 生成的内容可能不正确。

#### 4.6.1.11 Suggestion Quick Input Card

图形用户界面, 文本

AI 生成的内容可能不正确。

The suggestion quick input card will appear after round of conversation is presented. Click on the selection button to quick input in chat input box.

#### 4.6.1.12 Delete Conversation Confirmation Prompt

图形用户界面, 文本, 应用程序, 聊天或短信

AI 生成的内容可能不正确。

Once user click on the delete button, the delete confirmation box will appear. Click on cancel button to return to the current session, or click on delete button to delete the current session.

#### 4.6.1.13 Rename Conversation Prompt

图形用户界面, 文本, 应用程序, 聊天或短信

AI 生成的内容可能不正确。

Once the user clicks the rename button, the rename confirmation box will pop up. The user need to key in the new conversation name and click on save button in order to rename the conversation. Click on the cancel button to return to the current conversation.

# Chapter 6 Conclusion

Problem investigation and early design which was the first phase of the project has been performed with an intended goal to a great extent. You were able to find a clear user need (time:consuming, credibility-sensitive study-abroad inquiries) and convert it to a specific purpose of RAG-powered counselling chatbot. The literature review includes the fundamental constituents (dense retrieval, reranking, chunk-aware segmentation, citation-backed generation) and goes further to “cross-lingual/multimodal considerations applied to ChineseEnglish code-switching. A review of related systems (e.g., framework/tooling ecosystems of chatbot + RAG) also facilitated orientation of solution space and decision regarding technology implementation. Agile Scrum Methodology is provided as supporting iterative delivery and an initial technical plan is provided (Python/FastAPI service, FAISS, Qwen embeddings + reranker via SiliconFlow, citation-first answer generation with a 2 s latency target). Other activities that have been conducted on an empirical level comprise the establishment of a controlled questionnaire to confirm need, assess information-seeking behaviour, and identify requirements (e.g., the necessity of source transparency, and eligibility checks). All in all, these aspects reveal enough level of exploration to go to prototyping: problem is clearly defined, success measures are emerging and stack makes sense relative to research objectives.

With that said there are still a few research and even design gaps that need to be fixed before the de-risking the build:

* **Data and measures**. The review contends in favour of RAG, and many of the parameters to trial quality are yet to be established in retrieval and UX impact. Develop and pledge to judgement criteria- e.g. Recall@k / nDCG@k in retrieval, citation correctness rate, answerability identification accuracy, median /95th-percentile response length, and time-on-task reduction in consumer testing.
* **Corpus extent and control.** The targeted corpus (policies, programme guides, visa FAQs, exemplar essays) should have clearer inclusion criteria, coverage breakdown (by country/institution), multilingual coverage, evidence refresh rate and data-licensing/privacy protection.
* **Questionnaire legitimacy and research plan.** The questionnaire design is fit-for-purpose but pilot testing, reliability of the scales (e.g., alpha Cronbach scales), sampling, and triangulation (e.g., interview or think-aloud tasks) need to be included as part of the results to help in inference.
* **Reliability design control**. Explicit hallucination mitigation and safeguards should be introduced: answerability checklists prior to creation, rigorous source-attribution, denial styles in case of a scarcity of evidence and the use of fallback to official links.
* **Retrieval and prompting information.** They encourage chunk-aware segmentation + reranking, but ablation strategies are not provided (window sizes, overlap, multi-query expansion, cross-encoder vs bi-encoder rerankers). Make these design questions testable ones.
* **Operational readiness.** Monitor latency + cost/query + failure ratios of each API, have cost controls (batching, caching) and a rate-limit/backoff mechanism to the selected APIs, and have a banner at the bottom of the UI which informs the user about the privacy/consent implications of using the site.
* **Exposure and UX**. Devote yourself to “show sources facility UX, readable citation snippets, go mobile-first, list features language-toggle and glossaries are useful to non-techie users.

In sum, Part One was able to define the why, what and with-what of the project. In order to justifiably be ready to implement, Part Two must (1) complete and index the corpus with governance established, (2) develop the prototype according to the stipulated stack, and (3) conduct a pilot (n=30-50) and report the pre-determined metrics and refine prompting/retrieval through quantitative and qualitative feedback.

# References

Abraham, J. E. (2021). Effectiveness of College Counseling Inter eness of College Counseling Interventions in International entions in International. *Old Dominion University* , 140.

Ali, S. (2024, June 10). *A Deeper Dive Into Exploring The Applications Of OCR In eLearning*. Retrieved from ElearningIndustry: https://elearningindustry.com/a-deeper-dive-into-exploring-the-applications-of-ocr-in-elearning

AxcelerateAI. (n.d.). *Automating Education with OCR and IDP: Top Use Cases*. Retrieved from AxcelerateAI: https://www.axcelerate.ai/blogs/automating-education-with-ocr-and-idp-top-use-cases

Box, A. (2025, January 16). *The Ethical Implications of MultiModal AI: Privacy and Bias Concerns*. Retrieved from LinkedIn: https://www.linkedin.com/pulse/ethical-implications-multimodal-ai-privacy-bias-concerns-w7yrc

Burgenland, U. o. (n.d.). SDG 9: Industry,Innovation and Infrastructure. *SDG*, 30.

Chang, S. (2012). Factors Impacting Information Seeking Behaviour of International Students: Towards A Conceptual Model. *Isana*, 11.

Ferguson, S. T. (n.d.). *Implementing SLOs*. Retrieved from Sre Google: https://sre.google/workbook/implementing-slos/

GeekforGeek. (2025, 7 23). *What is Scrum? Understanding the Agile Framework for Project Management*. Retrieved from GeekforGeek: https://www.geeksforgeeks.org/software-engineering/what-is-scrum/

Gottlieb, U. R., & Amanda. (2023). Why confidence leads to swifter tertiary education choices: A qualitative study of the international tertiary education industry of Australia . *CelPres*, 13.

Hamid, S. (2018). Modelling the information-seeking behaviour of international students in their use of social media in Malaysia. *Information Research*, 24.

Hertzum, M., & Hyldegård, J. (2019). Information Seeking Abroad: An Everyday-Life Study of International Students . *Journal of Documentation*, 20.

Iatrellis, O., Samaras, N., & Konstantinos. (2024). EDUC8EU: A Transformative Hybrid Academic Advising Approach with Intelligent Microservices, Fuzzy Logic and Semantic Web Technologies. *J. Electrical Systems* , 9.

Ismail, A. (2021). Students' Satisfaction towards Academic Advising Service. *Asian Journal of University Education (AJUE)*, 8.

Krsmanovic, M. (2021). The Synthesis and Future Directions of Empirical Research On International Students In the United States: The Insights From One Decade . *The Quila Digital Comnunity*, 24.

Kuhail, M. A. (2022). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 46.

Li, Z., Wang, Z., Wang, W., & Hung, K. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence 8*, 14.

N´emeth, R. ´., & Tatrai, A. ´. (2025). Exploring the use of retrieval-augmented generation models in highereducation: A pilot study on artificial intelligence-based tutoring. *ELSEVIER*, 10.

OKTAVIA, M. B. (2023). BUILDING A RETRIEVAL-AUGMENTED GENERATION SYSTEM FOR ENHANCED STUDENT LEARNING: CASE STUDY AT PRIVATE UNIVERSITY . *Journal of Theoretical and Applied Information Technology*, 13.

Pampel, B., Martin, S., & Pado, U. (2025). Regaining Control: Enabling Educators to Build Specialized AI Chat Bots with Retrieval Augmented Generation. *17th International Conference on Computer Supported Education*, 9.

Saif-Ali. (2024, June 7). *OCR and Digital Learning: Tools and techniques for students*. Retrieved from enago academy: https://www.enago.com/academy/guestposts/saif-ali/empowering-learners-ocr-technology-transforming-digital-education/

# Appendices

## Appendix A: PPF

## Appendix B: Fast Track Form

## Appendix C: Meeting Log

## Appendix D: Gantt Chart

## Appendix E: Respondent Demographic Profile

A table with black text

AI-generated content may be incorrect.

A table with numbers and letters

AI-generated content may be incorrect.