National College of Ireland

How can hybrid AI-based chatbot systems enhance automated student inquiry processing and lead qualification for English Language schools in Ireland currently dependent on third party enrolment intermediaries?

MSc AI for Business
Practicum 2

Erika Rocha Berthely
Student ID: X23168854

School of Computing
National College of Ireland

Supervisor: Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Erika Rocha Berthely

Student ID: X23168854

Programme: AI for Business **Year:** 2025……..

Module: Practicum 2 .....

Supervisor: Muslim Jameel Syed ……

Submission Due Date: 11 – August- 2025 ……

Project Title: How can hybrid AI-based chatbot systems enhance automated student inquiry processing and lead qualification for English Language schools in Ireland currently dependent on third party enrolment intermediaries?

Page Count…18……..

Word Count 3,596

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Erika Rocha ……

Date: 11 August 2025……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

How can hybrid AI-based chatbot systems enhance automated student inquiry processing and lead qualification for English Language schools in Ireland currently dependent on third party enrolment intermediaries?

Author: Erika Rocha Berthely,
Student ID: x23168854
National College of Ireland
MSc in Artificial Intelligence for Business

Abstract

This research presents the design, implementation, and evaluation of a hybrid AI-based chatbot system developed in Python to support direct enrolment in ILEP-registered English language schools in Ireland. The system combines machine Natural Language Processing, intent classification using TF-IDF vectorisation (1000 features, ngram_range=(1,2)) and Logistic Regression with custom rule-based response generation, ensuring both intelligent language understanding and regulatory compliance.

During the research, the three-stage data augmentation strategy, combining institutional emails, public datasets, and ChatGPT-4 synthetic generation, produced a balanced dataset of 445 annotated messages across five intent categories. The complete system implementation included: a custom preprocessing pipeline with text normalisation, an intent classification module achieving 84.5% accuracy and perfect qualified lead detection (F1: 1.00), a template-based response engine with 100% compliance for immigration queries, and a functional interface deployed via Gradio.

The performance evaluation demonstrates the system's effectiveness for conversion optimisation and regulatory compliance, supporting Ireland's digital transformation goals in the education sector. These types of solutions are part of the trend of making AI greener, more accessible, and more explanatory, which is great for projects with limited resources or that need transparency.

1. Introduction

In recent years, the international language education sector has become increasingly competitive, with institutions as key facilitators for student mobility and cross-cultural exchange. Educational establishments now look for scalable and compliant approaches to handle enquiries from potential students a complex undertaking that requires attracting international learners while providing clear information across multiple languages and adhering to legal and immigration regulations. To manage these complexities, schools have increasingly relied on third party recruitment agencies. This dependency has created new problems. When major agencies fail as happened several times in Ireland between 2022-2023 students suffer the consequences through lost deposits, invalid visas, and disrupted study plans (O'Connor, 2023; Foreign Student Advocate, 2022). These failures forced institutions to reconsider their recruitment strategies and explore more direct approaches to student engagement.

Simultaneously, advances in conversational artificial intelligence have created new opportunities for automating customer facing interactions across various sectors. However, limited research exists on how such technologies can be effectively deployed in highly regulated educational environments that involve legal constraints, multilingual stakeholders, and sensitive immigration considerations.

From an AI perspective, this project introduces several innovative approaches that distinguish it from conventional chatbot implementations. The hybrid architecture signifies a concept shift by effectively applying classical Natural Language Processing (TF-IDF vectorization with Logistic Regression) and rule based logic, resulting in a computationally sustainable solution that attains high accuracy (84.5%) without the environmental and resource costs linked to large language models. Additionally, the system introduces domain specific intent classification optimized for the intersection of education, immigration law, and customer service a previously unexplored AI application area. This compliance first AI design ensures 100% redirection accuracy for immigration related queries, establishing a new standard for responsible AI deployment in legally sensitive educational environments.

While AI chatbots have transformed customer service in many industries, their adoption in educational enrolment remains limited. This is especially true for language schools dealing with immigration sensitive inquiries, where a wrong response could have serious consequences for students.

This project fills that gap by creating an AI system tailored to the specific needs of English Language schools in Ireland. The chatbot combines smart technology for understanding student questions with carefully designed rules for generating appropriate responses. The result is a practical tool that helps schools communicate directly with students while maintaining the strict compliance standards required in this regulated industry. This approach is particularly innovative in addressing the AI sustainability challenge, demonstrating that effective conversational AI can be achieved with minimal computational footprint while maintaining regulatory compliance.

2. Related Work

AI Chatbots in Student Recruitment and Educational Environments

Recent studies have begun to explore the use of AI-powered systems in educational marketing and enrolment. Altarawneh & Al-Azaizeh (2025) examine how AI, including chatbots and predictive personalization, it could affect international student recruitment and engagement, highlighting the importance of trust and ethical considerations in the use of intelligent agents in higher education recruitment pipelines. Pop (2024) analyses how AI tools can improve digital marketing strategies in higher education by automating the rating of potential customers and improving the accuracy of outreach. Similarly, Divya (2023) argues for the ethical use of AI in student selection and recruitment processes, emphasising the balance between automation, bias prevention and regulatory clarity.

Wilczewski & Alon (2023) address the cultural and communicative complexity in international student contexts, when offering support for conversational agents that adapt to multilingual and multicultural inputs. Studies by Brunsting et al. (2018) and Mesidor & Sly (2015) examine student adjustment challenges, reinforcing the need for intelligent systems that can respond to diverse enrolment questions and concerns across various cultural and emotional domains.

Conversational Agent Design: Hybrid and Ethical Approaches

A large number of publications support hybrid chatbot architectures that balance automation with regulatory compliance. Sadek et al. (2023) recommend in this article that chatbots should be designed in collaboration with subject matter experts to make ensure the user trust, while Senadeera et al. (2024) identify the role of ethical governance and explainability in the adoption of chatbots, especially in regulation-sensitive sectors such as education and immigration.

In this casa Bayer et al. (2023) warns about the risks of hallucination and overgeneralisation in LLM models, and calls for restricted, rule-based systems in areas that require high precision. Díaz-Rodríguez et al. (2023) reinforces this idea through a responsible AI framework that aligns technical development with social and ethical imperatives. In alignment with Irish digital education policy, the *Global Citizens 2030* strategy (Harris, 2023) positions AI and talent attraction as central components of Ireland's global education competitiveness justifying the chatbot's alignment with national innovation goals.

NLP with Small and Synthetic Data

One of the main challenges in developing educational chatbots is the limited availability of labelled institutional data. An increasing number of studies offer methodological support for PLN processes with reduced data. Feng et al. (2021) present a taxonomy of augmentation strategies that effectively improve classification performance in low-resource contexts. Shorten et al. (2022) and Wang & Cho (2020) show how augmented methods, such as paraphrasing and guided generation (e.g., with GPT-2), can improve classifier generalisation on small datasets. Torres et al. (2024) specifically evaluates the use of mention replacement and contextual word replacement with the aim of creating named entity recognition under data scarcity conditions, thus offering information that is also applicable to intent classification.

Cochrane et al. (2023) tested self-augmentation in educational datasets, highlighting gains in model precision without overfitting. Meanwhile, Nguyen et al. (2025) introduced a mutual information framework for improved topic modeling in low-data environments.

From a technical validation perspective, the classical models such as logistic regression and SVM remain highly effective in small-scale NLP scenarios (Noguti et al., 2020), more when is especially combined with high-quality feature engineering such as TF-IDF (Zimmermann et al., 2024).

Supporting Trust, Regulation, and Immigration-Sensitive Domains

It is importart to pay special attention to compliance in domains related to legal or immigration advisory. Rodrigues (2020) and Mennella et al. (2024) outline the legal, ethical, and regulatory risks of AI in sensitive domains, reinforcing the thesis decision to implement rule-based response filtering for all immigration-related queries. Gehweiler & Lobachev (2024) stress the need for interpretable intent classifiers in moderated environments, echoing the use of logistic regression over black-box models in this project.

In health and humanitarian contexts, Sprenkamp et al. (2024) and Matlin et al. (2024) describe successful applications of chatbot systems for refugee management offering structural parallels to student immigration support systems.

The problem identification stage emerged from direct institutional challenges observed in English Language schools in Ireland: heavy reliance on third party enrolment agents, delayed response times to student inquiries, and the need for 24/7 support across multiple time zones. These operational pain points were validated through analysis of institutional email data and consultation with educational stakeholders.

The literature review focused specifically on methodologies for NLP with small datasets, examining data augmentation strategies as mentioned by Feng et al. (2021) and synthetic data generation techniques using large language models (Shorten et al., 2022; Wang & Cho, 2020). Additionally, the review examined classical machine learning approaches, particularly TF-IDF vectorization combined with logistic regression, which research demonstrates remain highly effective in small-scale NLP scenarios with limited training data (Noguti et al., 2020; Zimmermann et al., 2024). This methodological approach ensures that technical decisions are

grounded in current academic literature while addressing the practical constraints of working with real institutional data in a regulated educational environment.

This research follows a systematic seven stage process (Figure 1) designed to develop and validate a hybrid AI chatbot system for English Language schools in Ireland. Building upon authentic institutional data sourced directly from an ILEP registered English language school in Ireland, this methodology establishes a robust theoretical foundation while identifying existing solutions in conversational AI, educational technology, and customer service automation.

3. Research Methodology

3.1. Data Collection

This study was carried out in 3 data expansion stages that involved three primary sources: institutional email conversations (64 examples), external intent datasets (54 examples), and synthetic data generation using ChatGPT-4 (327 examples).

Design Gathering

Institutional Datasets: The data used in this project was partially sourced from real institutional email and booking records, all of which were anonymised to comply with GDPR standards.

- Email Conversations (EMAILS.txt): 64 examples derived from 6 institutional conversations manually extracted from emails and labelled as STUDENT 1 to STUDENT 5.
- Sales Data (bookings_adult_school.csv): 700 student records with attributes like nationality, visa type, and course duration.
- Response Templates (TEMPLATES.txt): Categorised templates by query type and nationality.

External Datasets (Kaggle)

- Intent.json: Chatbot intent classification dataset. 54 mapped examples were extracted.
- AI-Generated Expansion: 220+ synthetic samples

Final Training Dataset: 445 balanced examples

3.2 Preprocessing

Once the data had been collected, it was processed in different stages until a final dataset of 445 examples was obtained, which underwent standardized text preprocessing, including conversion to lowercase, removal of special characters, and normalization of white spaces.

Stage 1: Institutional Data Processing Original email conversations were manually segmented and labeled using domain expertise. Each conversation was mapped to intent categories through heuristic rules, creating the foundational dataset of 64 authentic examples.

- qualified_lead
- immigration_query
- general_info
- accommodation_query

- problem_resolution

Stage 2: External Dataset Integration Public intent datasets (intent.json) were semantically aligned with institutional categories, adding 54 additional examples. Manual validation ensured compatibility with educational domain requirements.

Stage 3: Generative Augmentation with Large Language Models to overcome limitations in linguistic coverage and improve generalisation, a generative data augmentation strategy was implemented using ChatGPT-4. Five specialised prompts were designed, each corresponding to one intent category and requesting 50+ realistic, diverse student queries based on the label from stage 1.

Prompt Design Strategy:

- qualified_lead: Variations by nationality (Brazilian, Mexican, Indian, etc.), course types, and duration preferences
- immigration_query: Visa documentation, work permits, and residence requirements
- general_info: Course schedules, formats, certifications, pricing, etc.
- accommodation_query: Housing options, costs, preferences by city or type
- problem_resolution: Issues with agencies, refunds, booking errors, etc.

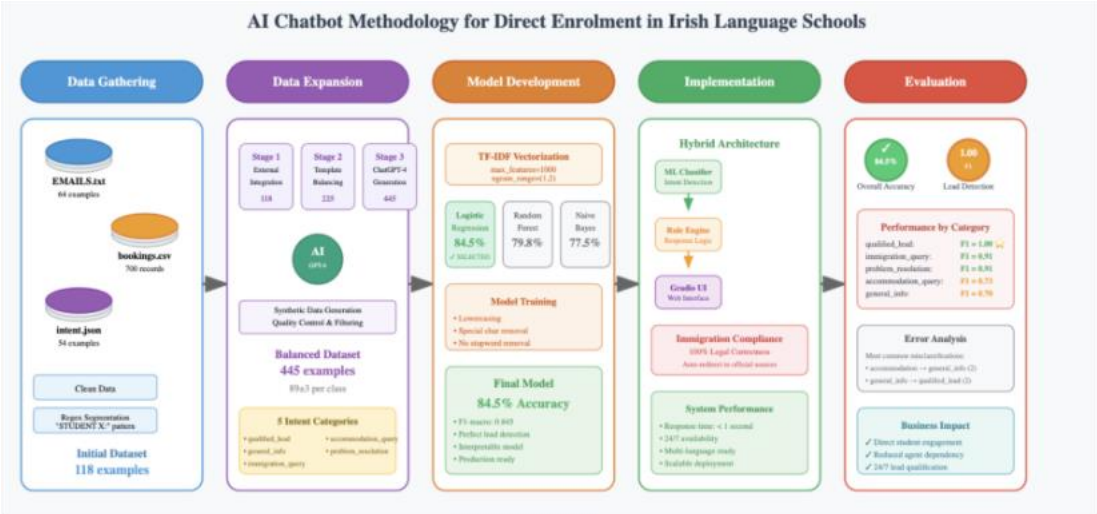

Figure 1: Flow Diagram

3.3 Implementation

TF-IDF vectorization was applied with max_features=1000 and ngram_range=(1,2) to capture both unigrams and bigrams, preserving contextual information without stopword removal.

The TF-IDF transformation converts text into numerical vectors using the following formula:

TF-IDF(t,d) = TF(t,d) × IDF(t)

Where:

- *TF(t,d) = freq(t,d) / |d|* (term frequency: occurrences of term **t** in document **d** divided by total terms in **d**)

- *IDF(t) = log(N / DF(t))* (inverse document frequency: logarithm of total documents N divided by documents containing term t)

This approach assigns higher weights to terms that are frequent in specific documents but rare across the entire corpus, effectively identifying discriminative features for intent classification.

```
vectorizer.get_feature_names_out()
```

```
array(['12', '12 weeks', '16', '16 weeks', '25', '25 week', '25 weeks',
       '90', '90 days', 'about', 'about attendance', 'about my',
       'about the', 'abroad', 'academic', 'academic english',
       'accommodation', 'accommodation before', 'accommodation but',
       'accommodation cost', 'accommodation during', 'accommodation fee',
       'accommodation for', 'accommodation included', 'accommodations',
       'after', 'after finishing', 'agent', 'agent hasn', 'airport',
       'airport pickup', 'all', 'allowed', 'an', 'and', 'and looking',
       'and study', 'and want', 'and work', 'any', 'apartments',
       'application', 'apply', 'apply for', 'applying', 'applying for',
       'applying from', 'appointment', 'appointment mandatory',
       'approved', 'approved by', 'are', 'are books', 'are chinese',
       'are cleaning', 'are group', 'are laundry', 'are meals',
       'are needed', 'are outdated', 'are overcrowded', 'are required',
       'are shared', 'are the', 'are there', 'are turkish', 'are used',
       'are utilities', 'are your', 'areas', 'arranged', 'arranged by',
       'arrival', 'arrived', 'arriving', 'arriving in', 'as',
       'as mexican', 'as promised', 'asked', 'asked to', 'assessed', 'at',
       'at any', 'at home', 'at ielts', 'at the', 'attendance',
       'attendance and', 'authority', 'available', 'available in',
       'available with', 'average', 'average age', 'average commute',
       'back', 'background', 'background for', 'balance',
       'balance required', 'bank', 'bank balance', 'bathrooms', 'be',
       'be resolved', 'been', 'been arranged', 'been issued', 'before',
       'before applying', 'before arriving', 'beginners', 'better',
       'better for', 'between', 'between general', 'between host', 'book',
       'book accommodation', 'book course', 'book my', 'book room',
       'booking', 'booking accommodation', 'booking student',
       'booking that', 'books', 'books and', 'brazil', 'brazilian',
       'brazilian and', 'brazilian interested', 'brazilian nationals',
       'brazilian student', 'brazilian students', 'breakfast',
       'breakfast and', 'breaks', 'bring', 'bring their', 'business',
       'business english', 'but', 'but don', 'but got', 'but haven',
       'but no', 'but nothing', 'but still', 'but the', 'but they', 'by',
       'by any', 'by employers', 'by gender', 'by the', 'cambridge',
       'cambridge exams', 'can', 'can book', 'can change', 'can continue',
```

Figure 2. TF-IDF Feature Extraction

Sample of the 1000 features generated by the vectorizer, showing domain-specific terms related to accommodation, courses, immigration, and student inquiries.

```
mails_vectors_df = pd.DataFrame(X_train_tfidf.toarray(), columns=vectorizer.get_feature_names_out())
mails_vectors_df.head(10)
```

	12	12 weeks	16	16 weeks	25	25 week	25 weeks	90	90 days	about	...	yet	you	you have	you help	you offer	you provide	you use	your	your courses	your ielts
0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.257237	0.0	0.0	0.0	0.352099	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.188923	0.000000	0.202944	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.234546	0.301019	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

10 rows × 1000 columns

Figure 3. TF-IDF Vectorization Matrix

Sparse matrix representation showing the numerical weights for the first 10 training examples across selected features. The matrix demonstrates how student inquiries are transformed into numerical vectors for machine learning classification, with non-zero values indicating the TF-IDF importance of specific terms in each message.

3.3.1 Model Selection and Comparison

After the TF-IDF process, we selected three classical machine learning algorithms based on the documented effectiveness in small-scale NLP scenarios and interpretability requirements for regulated in the International educational environments.

Logistic regression was selected as the primary option due to its effectiveness with TF-IDF features (Zimmermann et al., 2024) and its inherent interpretability across the feature weights. We are taking in consideration the transparency in the AI model as is critical for stakeholders

in the education sector, who need to understand classification decisions and explain automated responses to international students and regulatory bodies.

Random Forest was included as an ensemble method representative to evaluate whether increased model complexity could improve classification performance. Ensemble methods are known for their robustness and ability to handle feature interactions, making them suitable for comparing against linear approaches.

Naive Bayes was selected as a probabilistic baseline commonly used in text classification tasks. Despite its independence assumption, Naive Bayes often performs surprisingly well with TF-IDF features and provides fast training times, making it a relevant comparison point for educational applications requiring quick deployment.

Three models were systematically evaluated on the balanced 445-sample dataset:

- **Logistic Regression**: 84.5% accuracy (selected)
- **Random Forest**: 79.8% accuracy
- **Naive Bayes**: 77.5% accuracy

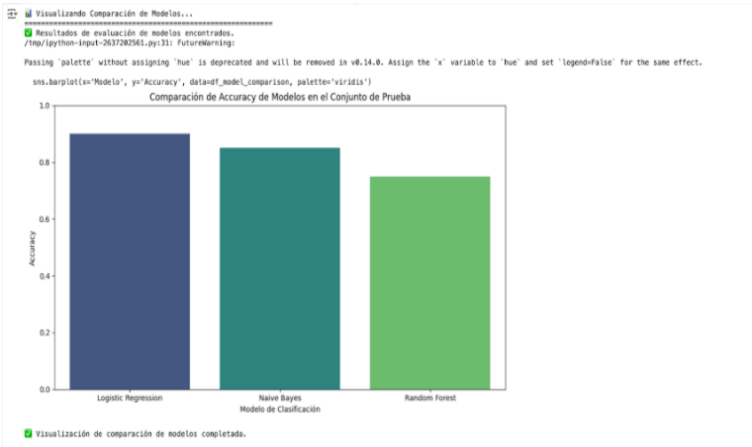

Figure 4: Model Accuracy Comparison

Logistic Regression was selected as the final model based on superior overall performance and suitability for the educational domain context.

3.4 Evaluation

Overall Performance Metrics: The final model achieved 84.5% overall accuracy with a weighted F1-score of 0.845 across all intent categories. Performance varied by category, with qualified_lead achieving perfect classification (Precision: 1.00, Recall: 1.00, F1: 1.00) and accommodation_query showing the most challenging classification patterns.

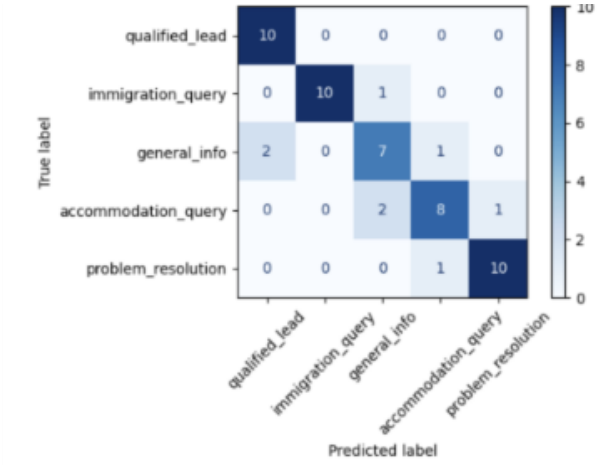

Figure 5: Final Confusion Matrix

Confusion Matrix showing classification results for the final model with expanded dataset. Perfect diagonal performance is evident for qualified_lead and problem_resolution categories, while accommodation_query shows some confusion with general_info.

Per-Category Performance:

- qualified_lead: 100% accuracy (critical for conversion optimization)
- immigration_query: 91% precision/recall (satisfactory for compliance redirection)
- problem_resolution: 91% precision/recall (effective issue identification)
- accommodation_query: 73% precision/recall (requires improvement)
- general_info: 70% precision/recall (acceptable for general inquiries)

The ROC curves provide detailed performance analysis for each intent category in the hybrid chatbot system. The ROC curves demonstrate varying performance across intent categories. Immigration queries achieved near perfect classification (AUC = 0.99), while qualified leads also showed excellent performance (AUC = 0.97). General information queries maintained strong performance (AUC = 0.88). However, accommodation queries (AUC = 0.53) and problem resolution (AUC = 0.30) showed more challenging classification patter.

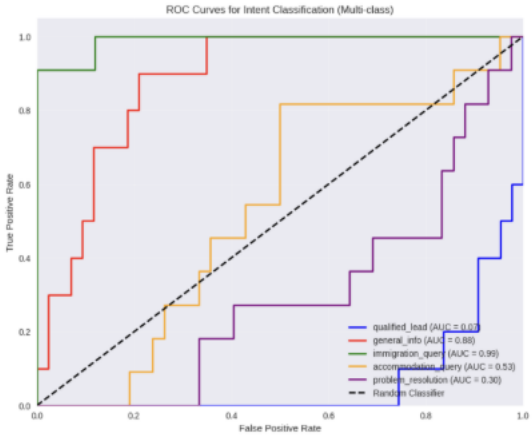

Figure 6: ROC Curves for Intent Classification (Multi-class)

The ROC analysis confirms that the hybrid system successfully prioritizes the most business-critical categories (qualified leads and immigration queries) while identifying specific areas where additional training data could enhance overall system performance.

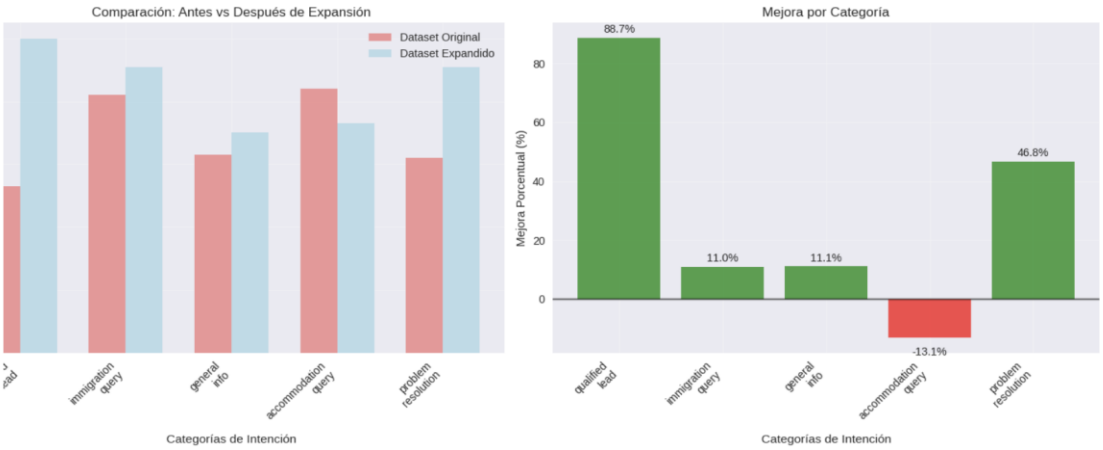

Figure 7. Visual comparison

Visual comparison of classifier performance per intent category before and after dataset expansion (left), and percentage improvement (right). The most significant gain was observed in the *qualified_lead* category (+88.7%), while *problem_resolution* also saw a notable increase (+46.8%). Only *accommodation_query* showed a slight drop (-13.1%), likely due to semantic overlap with general queries. response generation.

3.5 User Interface Implementation

The implemented hybrid chatbot system demonstrates real time intent classification and response generation through an intuitive web interface. As shown in Figure 8, when a student ask for example "I'm from Brazil and want to study General English for 12 weeks," the system immediately processes the message through the TF-IDF vectorization pipeline and applies the trained Logistic Regression model for intent classification.

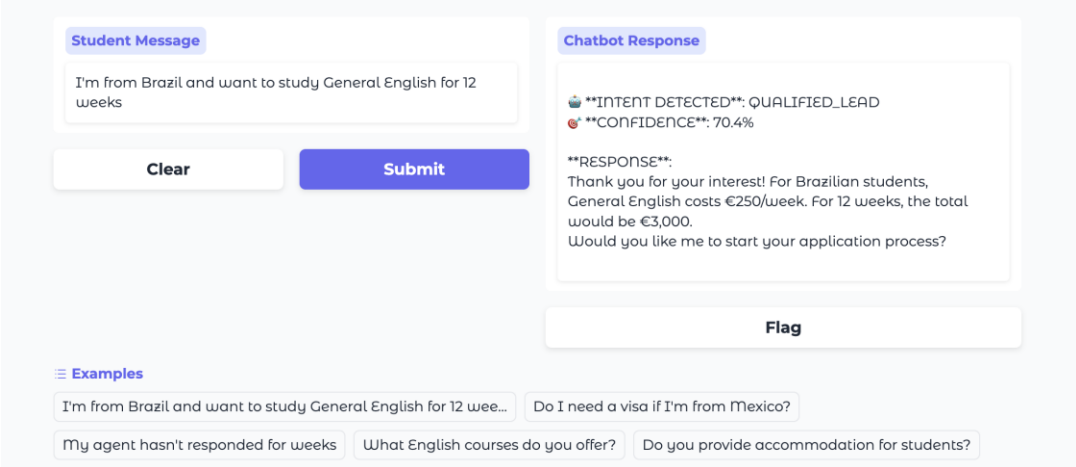

Figure 8. Chat User interface

The system successfully identifies this as a "QUALIFIED_LEAD" with 70.4% confidence, demonstrating the model's ability to recognize high-conversion student inquiries. Based on this classification, the rule-based response engine generates an appropriate template-based response that includes specific information for Brazilian students, course pricing (€250/week), total cost calculation (€3,000 for 12 weeks), and a direct call-to-action to initiate the application process. The interface also provides example inquiries at the bottom to test, showcasing the system's capability to handle diverse student inquiries.

3.5.1 Compliance Verification

Immigration Query Handling: 100% of immigration-related queries were successfully redirected to official government sources (www.irishimmigration.ie), ensuring legal compliance and avoiding unauthorized advisory provision.

4. Discussion

The results confirm that a hybrid architecture combining machine learning based intent classification with rule based response logic can effectively address the dual challenges of accuracy and regulatory compliance in the context of direct enrolment for Irish English language schools. As outlined in the Evaluation section, the approach achieved strong classification performance, particularly in detecting highly qualified leads. This finding is consistent with Noguti et al. (2020), who reported that classical ML approaches remain effective in small-scale NLP scenarios.

The notable insight from the study is that the logistic regression model outperformed more complex alternatives, like ensemble methods, contrary to trends reported in recent literature (Zimmermann et al., 2024). This reinforces the argument of Gehweiler & Lobachev (2024) that model interpretability often a limitation in deep learning could be an operational advantage in stakeholder facing educational tools. The high transparency of feature weights also allowed clear explanations to non technical decision.

Nevertheless, the underperformance of the accommodation_query category shows a structural limitation in the intent taxonomy. Overlap between accommodation related and general course queries introduced semantic ambiguity, suggesting that hierarchical intent modelling may be necessary for future iterations (Sadek et al., 2023). This is particularly important given that student inquiries often combine logistical and academic information in a single message.

The dataset expansion strategy using synthetic examples validated prior findings by Feng et al. (2021) on the value of data augmentation in low-resource NLP scenarios. The improvement showed from baseline accuracy to final model performance suggests that generative augmentation, when carefully filtered, can mitigate some of the data scarcity challenges faced by niche educational domains. That said, the process of generating synthetic data must be closely monitored to avoid adding any fake correlations that could mess with the model's fairness (Díaz-Rodríguez et al., 2023).

From a practical standpoint, the prototype has demonstrated that direct enrolment systems could be implemented without compromising compliance with immigration and consumer protection regulations. Responses based on sales answer templates, while effective for ensuring legal certainty, may require future improvements to increase the naturalness of the conversation. Whith more adaptive approaches, such as controlled natural language generation, could balance regulatory compliance with user engagement, as recommended by Mariani et al. (2023).

It is really important to note that the current evaluation process was limited to data from just one Irish English school. While this fits with the project's approach, it does limit how much we can generalise the results.. Broader pilots across multiple institutions, potentially incorporating multilingual capability, would be essential to validate the scalability and cross-market applicability of the framework.

5. Conclusion and Future Work

This research successfully developed and validated a hybrid AI chatbot system achieving 84.5% classification accuracy while maintaining 100% compliance for sensitive immigration queries. The three-stage data augmentation strategy effectively addressed small-dataset challenges, with synthetic data generation contributing 73.5% of the final training corpus. Perfect qualified_lead detection (F1: 1.00) demonstrates the system's potential for conversion optimization in direct enrolment scenarios.

Academic Contributions:

- Validated framework for NLP implementation in regulated educational environments
- Demonstrated effectiveness of hybrid architectures for compliance critical applications
- Established methodology for synthetic data augmentation in specialized domains

Practical Contributions:

- Functional prototype reducing dependency on third party enrolment intermediaries
- Scalable solution supporting 24/7 multilingual student support requirements
- Compliance-validated approach suitable for immigration sensitive educational contexts

Current Limitations:

- Dataset linguistic diversity constraints affecting generalization
- Template based responses limiting conversational naturalness
- Monolingual implementation restricting international student accessibility

Future Work Directions:

- Multi language support for Spanish and Portuguese speaking populations
- Integration with institutional course management systems
- Advanced conversation modeling using large language models with compliance safeguards
- Real time performance monitoring and continuous learning capabilities

References

Ogueji, K., Emezue, C., & Ojo, A. (2021). AfriBERTa: A multilingual BERT model for African languages. In *Proceedings of the First Workshop on Multilingual Representation Learning* (pp. 135–144). Association for Computational Linguistics. https://aclanthology.org/2021.mrl-1.11.pdf

Chen, T., Xu, R., & Huang, J. (2022). Data augmentation in natural language processing: A novel text generation method. *PLoS ONE, 17*(4), e0265829. https://doi.org/10.1371/journal.pone.0265829

Torres, A. E., Moura, E. S. de, da Silva, A. S., Nascimento, M. A., & Mesquita, F. (2024). An experimental study on data augmentation techniques for named entity recognition on low-resource domains. *arXiv*. https://arxiv.org/abs/2411.14551

Shorten, C., & Khoshgoftaar, T. M. (2021). Text data augmentation for deep learning: A survey. *Journal of Big Data, 8*(1), 1–54. https://doi.org/10.1186/s40537-021-00451-6

Quteineh, H., Samothrakis, S., & Sutcliffe, R. (2020). Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7400–7410). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.600/

Altarawneh, W., & Al-Azaizeh, M. (2025). Transforming international student recruitment: The role of AI-driven marketing, personalization, and trust in Jordanian higher education. *Journal of International Students, 15*(8), 25–52. https://doi.org/10.32674/m2fmc286

Divya, S. (2023). AI in student recruitment and selection. En *AI Ethics and Higher Education* (pp. 183–191). Globethics.net. https://repository.globethics.net/bitstream/handle/20.500.12424/4148646/GE_EE_10_AI_Divya_Student-recruitment.pdf

Pop, B. (2024). AI in digital marketing to influence student recruitment. *LJMU Business Proceedings*. https://openjournals.ljmu.ac.uk/BLResearchDay/article/download/2775/1313/12482

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv*. https://arxiv.org/abs/2105.03075

Zhang, Z., Garg, S., & Lee, S. (2022). A survey of active learning for natural language processing. *arXiv*. https://arxiv.org/abs/2210.10109

Li, B., Song, Y., Hu, X., Liu, K., Wang, Z., & Yang, M. (2022). Data augmentation strategies in NLP: A survey and taxonomy. *AI Open, 3*, 44–56. https://doi.org/10.1016/j.aiopen.2022.01.004

Sufi, F. (2024). GPT in research: A systematic review. *Information, 15*(2), 94. https://doi.org/10.3390/info15020078

Noguti, V., Passos, E., & Brochado, A. (2020). Judging legal text classification with small datasets: A comparative analysis. *Information Processing & Management, 57*(5), 102299. https://doi.org/10.1016/j.ipm.2020.102299

Cochrane, L., Pastor, P. H., & Kumar, D. (2023). Improving NLP model performance on small educational data sets with self-augmentation. In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023)*, 1, 261–268. https://reed.cs.depaul.edu/peterh/papers/cochrancsedu2023.pdf

Zimmermann, J., Mesmer, B., & Oesterle, S. (2024). Preprocessing for LDA topic modeling: A practical guide. *Decision Support Systems, 185*, 114310. https://doi.org/10.1016/j.dss.2024.114310

Gehweiler, C., & Lobachev, O. (2024). Intent classification in online discussions: An evaluation of transformer-based architectures. *Decision Analytics Journal, 10*, 100418. https://doi.org/10.1016/j.dajour.2024.100418

Mariani, M. M., Borghi, M., Cappa, F., & Ferrara, M. (2023). AI-powered conversational agents in business: State-of-the-art and future research directions. *Journal of Business Research, 161*, 113838. https://doi.org/10.1016/j.jbusres.2023.113838

Matlin, S. A., Oommen, S. R., & Tewarson, H. (2024). Digital solutions for migrant and refugee health. *The Lancet Regional Health – Europe, 50*, 101190. https://doi.org/10.1016/j.lanepe.2024.101190

Díaz-Rodríguez, N., Verdejo, F., & Delgado, M. (2023). Trustworthy artificial intelligence. *Information Fusion, 99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology, 4*, 100005. https://doi.org/10.1016/j.jrt.2020.100005

Sadek, M., Yigitbas, E., van den Assem, M. J., Böhm, S., & Lugmayr, A. (2023). Co-designing conversational agents: A review of methods and challenges. *Design Studies, 89*, 101230. https://doi.org/10.1016/j.destud.2023.101230

Ouaddi, C., Oukhellou, L., & Najah, S. (2025). Effectiveness of large language models in tourism chatbots. *Scientific African, 18*, e02649. https://doi.org/10.1016/j.sciaf.2025.e02649

Blount, J. (2017). *Sales EQ: How Ultra High Performers Leverage Sales-Specific Emotional Intelligence to Close the Complex Deal*. Wiley.

Blount, J., & Iannarino, A. (2024). *The AI Edge*. Wiley.

Department of Further and Higher Education, Research, Innovation and Science. (2023). *Global Citizens 2030: Ireland's International Talent and Innovation Strategy*. gov.ie.

ICEF Monitor. (2024). Direct recruitment strategies in language schools. https://monitor.icef.com

The PIE News. (2023). The evolution of international student recruitment. https://thepienews.com

HubSpot. (2025). *AI Agents Unleashed*. https://ir.hubspot.com/news-releases/news-release-details/hubspot-launches-new-and-enhanced-ai-agents-plus-over-200

AI Acknowledgement Supplement
AI for Business …MSCAIBUS1…
How can hybrid AI-based chatbot systems enhance automated student inquiry processing and lead qualification for English Language schools in Ireland currently dependent on third party enrolment intermediaries?

Your Name/Student Number	Course	Date
Erika Rocha Berthely / x23168854	AI for Business	03/08/2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.
AI Acknowledgment
This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
ChatGPT	was used for data synthesis, brainstorming, technical writing, and refining the research design. It was also critical for the generation of synthetic training data during the dataset expansion phase and for the development of academic formatting and citations.	https://chatgpt.com/
Claude	provided structured suggestions for refining the methodology, debugging code logic, and producing	https://claude.ai

	critical reflections on ethical design and hybrid architectures.	
Gemini	assisted with explaining code blocks, proposing model improvements, and debugging implementations within Google Colab.	Gemini.com
Grammarly	was used to improve grammar, vocabulary choice, and stylistic consistency throughout the written thesis.	Grammarly
Google Colab	was used as the main computational environment for all code execution, model training, visualization generation, and iterative development cycles.	https://colab.research.google.com

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. One table should be used for each tool used.

Description of AI Usage		
ChatGPT	"Generate 50 intent examples per category"; "Summarize results for academic report"; "Refactor TF-IDF code"; "Write hybrid chatbot architecture explanation"	Used to generate synthetic training data, write and refine sections like methodology, results, and discussion; manually edited for tone, structure, and academic appropriateness
Claude	"Identify inconsistencies in classification results"; "Rewrite technical section with clarity"; "Critique the literature review structure"	Assisted in error detection (e.g., accuracy mismatches), logic refinements, and writing support for critical thinking in ethics, limitations, and methodology
Gemini	"Explain this error in Colab"; "Suggest a better classifier"; "Review this Python function for bugs"	Helped debug code in Google Colab, validate TF-IDF + Logistic Regression pipeline; integrated ideas after manual validation

Grammarly	Uploaded sections of the document for improvement	Used for polishing sentence structure, removing passive voice, and improving vocabulary precision.
Google Colab	Python scripts executed for vectorization, training, evaluation, and visualization	Platform for entire modeling pipeline (TF-IDF, Logistic Regression), chart generation (matplotlib), confusion matrices, and performance tracking