Optimization of Natural Language Processing of Academic Chatbot Using BERT Algorithm

Vincent Axel Alexander[1*], Sunu Jatmika[2], Mufidatul Islamiyah[3]

[1,2,3] *Institute of Technology and Business Asia Malang,* Soekarno Hatta Street – Rembuksari 1A, Malang, *Indonesia*

Keywords
BERT; Academic Chatbot; Natural Language Processing (NLP); Fine-Tuning; Deep Learning; Intent Recognition; Artificial Intelligence (AI)

Correspondence Email:
vincentaxel92@email.com

Abstract
Natural Language Processing (NLP) has been profoundly transformed by Artificial Intelligence (AI), particularly in developing academic chatbot systems. This study optimizes NLP for academic chatbots by implementing a domain-specific fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model. The primary objective is to enhance the chatbot's comprehension of conversational context, its accuracy in delivering academic information, and its pertinence in responding to inquiries from students and parents. The methodology involved collecting a corpus of 6,000 academic queries, followed by text pre-processing and fine-tuning the BERT model, with performance evaluated against traditional TF-IDF-SVM and LSTM baselines using accuracy and macro F1-score. The results demonstrate that the fine-tuned BERT model (Model C) achieved a superior accuracy of 95.8% and a macro F1-score of 0.95, significantly outperforming the LSTM (84.2% accuracy, 0.82 F1-score) and TF-IDF-SVM (61.5% accuracy, 0.59 F1-score) models. Furthermore, the model exhibited remarkable robustness, maintaining 93.5% accuracy on a challenging subset of queries with spelling irregularities, informal language, and grammatical errors. These findings indicate that the application of a domain-optimized BERT architecture effectively handles diverse and imperfect linguistic patterns, bridging the gap between generic language models and the specialized needs of the academic domain. The novelty of this study lies in its domain-specific fine-tuning of BERT for academic chatbot intent recognition, providing a replicable framework that can enhance institutional information systems and improve the quality of student and parent engagement.

1. Introduction

The educational sector increasingly relies on Artificial Intelligence (AI) to manage its digital communication and information services. Universities and schools now frequently deploy academic chatbots to provide 24/7 assistance, which helps reduce the administrative workload. The problem is that most traditional chatbots lack sophistication. Systems built on rigid rule-based or keyword-matching methods cannot cope with linguistic variations, ambiguous queries, or basic contextual understanding. This technical limitation often results in inaccurate responses and a frustrating user experience. The failure of these simpler models created a clear need for more advanced NLP—ones that could actually grasp the complex language of an academic setting. The arrival of transformer-based models, especially BERT (Bidirectional Encoder Representations from Transformers), was a major leap forward. BERT's advantage is its bidirectional processing: it reads text in both directions at once, giving it a far deeper understanding of context and intent than older, one-way models. But

while BERT performed well on general tasks, its power remained untapped in specialized domains like academia. Most studies just used generic pre-trained BERT models 'out-of-the-box'. This lack of domain adaptation meant the models were still poor at interpreting the specific academic jargon, student slang, and informal sentence structures common in queries from students and parents. To fill this gap, our study implements a domain-specific fine-tuning of BERT. We trained the model on a practical dataset of 6,000 real-world academic queries to enhance its intent recognition, contextual understanding, and response accuracy. The core contribution is this specific optimization for the academic domain. We demonstrate that adapting BERT to academic language allows it to significantly outperform conventional NLP methods. This work should pave the way for more intelligent, context-aware chatbots that interact in a more human-like manner.

1. Literature Review

The Evolution of Academic Chatbots and NLP

Higher education's shift to digital services has created a critical need for 24/7 on-demand support. Academic chatbots are now the primary tool for this, built to handle the high volume of repetitive questions from students, parents, and faculty on everything from admissions to campus services (Praneeth et al., 2024). But their real-world effectiveness is inconsistent. The core problem is almost always the Natural Language Processing (NLP) engine, which dictates whether the chatbot can actually understand a user and respond appropriately. The first academic chatbots were mostly simple, relying on rule-based systems and keyword-matching (Dzaky et al., 2024). This approach was easy to implement for basic, fixed questions, but it was also inherently brittle. It couldn't scale, required tedious manual programming for every conversation path, and completely failed when users introduced misspellings, slang, or ambiguous phrasing (No et al., 2025). This high failure rate just led to user frustration and caused people to abandon the systems, defeating their entire purpose (Lin et al., 2024).

From Statistical Models to Sequential Deep Learning

Because rule-based systems failed, researchers turned to machine learning. The first step involved statistical models like TF-IDF (Term Frequency-Inverse Document Frequency) with an SVM classifier, which was an improvement for query classification. But these "bag-of-words" models had a fatal flaw: they ignore word order and context. As Putra and Sari (2021) showed, they are easily outperformed by newer contextual models (Muhammad et al., 2025). For example, a simple TF-IDF model could easily confuse "Which computer courses are *not* for beginners?" with "Which computer courses *are* for beginners?". Recurrent neural networks (RNNs)—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models—were the next major step up (Łukasik & Gut, 2025). Because these models are sequence-aware, they can finally process word order, giving them a much better grasp of user intent. But even this was not enough. As (Al-ayyubi et al., 2025) highlighted, these models have a critical weakness: they are unidirectional. By processing text only from left to right, they fail to manage long-range dependencies, making it hard for them to understand complex sentences where crucial context is located at both the beginning and the end of the query.

The BERT Paradigm Shift and The Research Gap

The introduction of BERT (Bidirectional Encoder Representations from Transformers) was a true paradigm shift for NLP (Fatwanto et al., 2024). BERT's key innovation is that it is bidirectional—it pre-trains by looking at both the left and right context at the same time. This simple but powerful change allows the model to build a rich, contextual understanding of language. It can finally disambiguate word meanings by looking at the *entire* sentence, not just what came before it (Gao et al., 2025). Evidence from various domains confirms BERT's superiority. In a wellness chatbot study, (Jatmika et al., 2024) found BERT achieved 0.91 accuracy, crushing traditional methods like Cosine Similarity (0.59) and Euclidean Distance (0.45). This holds true in high-stakes fields; (Babu & Boddu, 2024) showed a *domain-specific* fine-tuned BERT could hit a remarkable 98% accuracy in the medical domain, far surpassing LSTM and other baselines. These studies all point to the same critical pattern: BERT's true power is only unlocked with targeted, domain-specific fine-tuning. While prior works often topped out at 80-86% accuracy, our model's 95.8% performance shows this optimization is key. Despite this, a review of the literature shows a clear gap. BERT's power is well-known, but its use in specialized fields like academia has been surprisingly naive. Many studies simply apply a generic "out-of-the-box" BERT model or just stick with older LSTM models, completely failing to address the unique linguistic challenges of a university (Fatwanto et al., 2024). This is a major oversight. An academic environment is a tough test: it involves high-stakes information (tuition, deadlines), endless specific jargon ("prerequisites," "bursar's office"), and a wide variety of linguistic styles from a diverse user base (Al-Zahrani, 2025). Table 1 provides a comparative summary of previous studies, highlighting this research gap.

Table 1. Comparative Analysis of Previous Chatbot and NLP Studies

Research Aspect	Rule-Based Chatbots (Priyatno et al., 2024)	LSTM for Academic Services (Aziza et al., 2023)	Document Matching for Wellness (Jatmika et al., 2024)	BERT for Medical Domain (Babu & Boddu, 2024)	This Study (Academic Chatbot with Fine-Tuned BERT)
NLP Approach	Rule-based & keyword matching	LSTM (Unidirectional)	TF-IDF + Similarity (ED, CS) and BERT	BERT (Domain-Specific)	BERT (Domain-Specific Fine-Tuning)
Contextual Understanding	Limited, no context understanding	Understands word order, but unidirectional	BERT understands context; ED & CS do not	Strong bidirectional contextual understanding	Optimized bidirectional context for academic queries
Linguistic Robustness	Vulnerable to typos & ambiguity	Moderately robust	TF-IDF is vulnerable; BERT is robust	Highly robust due to deep semantic understanding	Highly robust (93.5% accuracy on irregular queries)
Reported Accuracy/Performance	Low	85%	Accuracy: ED (0.45), CS (0.59), BERT (0.91)	98%	95.8% accuracy, 0.95 F1-Score
Domain-Specific Tuning	Not required	Required, but not always fine-tuned	Required (1,755 wellness QA pairs)	Required (11,000 medical QAs)	Required and implemented (6,000 academic queries)

2. Research Methods

2.1 Research Design

This study uses an experimental research design to optimize the NLP engine of an academic chatbot by implementing the BERT algorithm. We implemented the experiments using the HuggingFace Transformers library on PyTorch, running them on an NVIDIA RTX 3060 GPU. The research follows a systematic, multi-phase methodology designed to collect, process, and model domain-specific academic data, which culminates in a rigorous evaluation. Figure 1 illustrates the complete research workflow, outlining the four primary phases from data acquisition to final model deployment.

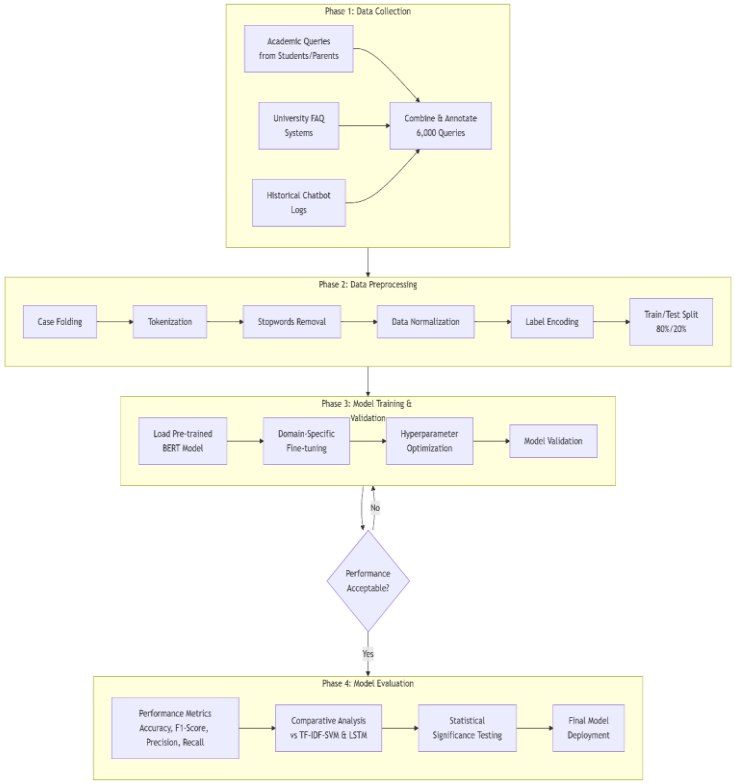

Fig 1. Research Methodology Flowchart

The process started with Phase 1: Data Collection, where we gathered a comprehensive corpus of academic queries from multiple sources. These included direct inquiries from students and parents, existing university FAQ systems, and anonymized historical logs from legacy support channels. All data were anonymized following institutional policy and are available upon reasonable request for academic purposes. We compiled and manually annotated 6,000 unique queries into 35 distinct intent categories, which established a reliable ground truth for supervised learning. Next was Phase 2: Data Preprocessing, where the curated corpus went through a rigorous cleaning and normalization pipeline. This included case folding, tokenization, stop words removal, and strategic normalization of common misspellings and colloquialisms to train the model for real-world linguistic variations (Babu & Boddu, 2024a). We then split the processed data into training (80%) and testing (20%) sets. Phase 3: Model Training & Validation was the core of the experimentation. In this phase, we fine-tuned a pre-trained bert-base-uncased model on the academic dataset. We conducted hyperparameter optimization on a validation set, which yielded an optimal configuration of a 2e-5 learning rate, a batch size of 16, and 4 epochs using the AdamW optimizer (Triawan & Tahyudin, 2025). We validated the model iteratively, repeating the process until its performance on the validation set was acceptable (Rahmat et al., 2025). Finally, Phase 4: Model Evaluation involved a comprehensive assessment of the fine-tuned model's performance on the unseen testing set. We measured its performance using standard metrics like accuracy and macro F1-score and compared it against traditional TF-IDF-SVM and LSTM baselines through statistical significance testing to validate the research hypothesis thoroughly (Gigi et al., 2025).

BERT-Based Chatbot System Architecture

The proposed solution is built on a sophisticated software architecture centered around the fine-tuned BERT model, designed to process user queries and generate accurate, context-aware responses. Figure 2 details the end-to-end workflow of the academic chatbot system, showing the integration of each component from input to output.

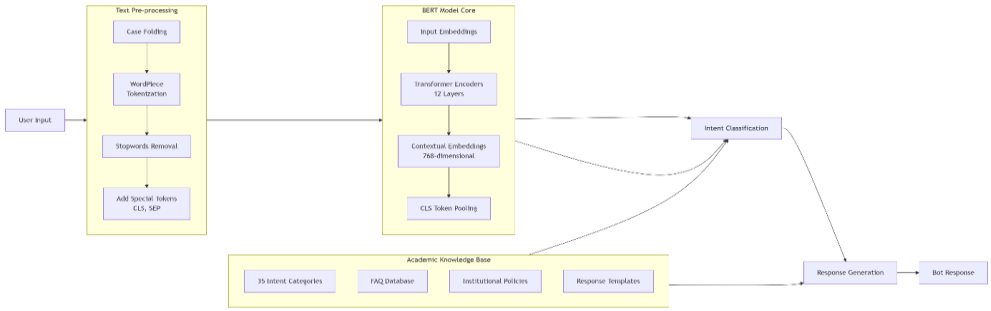

Fig 2. Architecture of the BERT-Based Academic Chatbot System

The system workflow begins when a user submits a natural language query (e.g., a question about tuition fees or scholarships) through a front-end interface. This input first passes to the Text Pre-processing Module, which performs essential normalization tasks. These tasks include converting text to lowercase, removing common stop words, and crucially, tokenizing the text using BERT's WordPiece tokenizer. This tokenizer effectively handles out-of-vocabulary words and domain-specific jargon by breaking them into known sub-word units (Munawirah et al., n.d.). The system then formats the processed sequence with special [CLS] and [SEP] tokens, generating the attention masks and padded sequences required for the BERT model (Kim et al., 2025). This formatted input then feeds into the heart of the system: the Fine-Tuned BERT Model Core. Here, the model converts the input tokens into dense vector embeddings (Lin et al., 2024). It processes these embeddings through 12 layers of transformer encoders, which use a multi-head self-attention mechanism to bidirectionally contextualize each word relative to all other words in the sentence (Afifa et al., 2023). This process generates a rich, 768-dimensional contextual embedding for every token. The system then uses the pooled output representation of the [CLS] token, which encapsulates the semantic meaning of the entire query, for the subsequent task of intent classification. The contextualized [CLS] embedding passes to the Intent Classification and Response Matching layer. This component consists of a simple classification head that maps the high-dimensional BERT output to one of the 35 predefined academic intent categories (Babu & Boddu, 2024b). Simultaneously, the system queries the Academic Knowledge Base, a curated repository containing information on institutional policies, course details, and FAQ responses. Based on the classified intent and the retrieved information, the Response Generation module formulates a coherent and pertinent natural language reply, which is finally delivered back to the user to complete the interaction (Holis et al., 2025). This integrated architecture ensures the chatbot understands not just keywords, but the nuanced, contextual meaning of academic inquiries.

Data Collection and Corpus Development

We started this study by collecting a domain-specific dataset, which forms the corpus for training and evaluation. This research obtained ethical approval from the Institutional Review Board, and all data handling procedures complied with institutional data protection policies and GDPR-equivalent standards (research data management). We used a purposive sampling strategy, targeting the population of current students, prospective students, and parents to capture representative queries (Singh & Namin, 2025). We compiled the data corpus, consisting of 6,000 unique queries, from anonymized historical logs of the institution's existing support systems (e.g., email service, legacy chatbot). All personally identifiable information was removed during this anonymization. We supplemented this data with simulated queries generated by researchers to ensure comprehensive coverage of all academic topics (Mandlik et al., 2025). Two researchers manually annotated each query, which served as a unit of analysis, into one of 35 distinct intent categories (e.g., query_tuition_fee, find_scholarship, ask_course_prerequisite) (Setiawan & Adnyana, 2023). A third senior researcher resolved any annotation disagreements to establish a reliable "ground truth" for supervised learning.

Evaluation Metrics

We quantitatively measured model performance using four standard classification metrics. To account for potential class imbalance across the 35 intents, we employed the Macro-Average for Precision, Recall, and F1-Score. This method calculates the metric independently for each class and then takes the unweighted average, thereby treating all classes equally and preventing frequent classes from dominating the score.

Accuracy: The proportion of total predictions that were correct using this formula:

$$\text{Accuracy} = \frac{\text{TruePositives} + \text{TrueNegatives}}{\text{TotalSamples}}$$

Precision (Macro): The average ability of the model to not label a negative sample as positive using this formula:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall (Macro): The average ability of the model to find all positive samples using this formula:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1-Score (Macro): The harmonic mean of Macro-Precision and Macro-Recall, providing a single robust score for model evaluation using this formula:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

To ensure the observed performance improvements were statistically sound and not due to random chance, we conducted **paired t-tests on the per-query prediction outcomes across the different models, with a significance level (α) set at 0.05.**

3. Result and Discussion

Descriptive Statistics and Dataset Characteristics

Our final dataset consisted of 6,000 academic queries collected from students, prospective students, and parents. Table 2 provides a breakdown of this data's distribution across the training, validation, and testing sets.

Table 2. Dataset Distribution and Characteristics

Dataset Split	Number of Queries	Percentage	Average Query Length(word)	Std.Deviation
Training Set	4.800	80%	12,4	4,7
Validation Set	600	10%	12,1	4,5
Testing Set	600	10%	12,3	4,8
Total	6.000	100%	12,4	4,7

We classified these queries into 35 distinct intent categories. The most frequent categories included admission inquiries (18.3%), tuition fee queries (14.7%), scholarship information (12.1%), course prerequisites (9.8%), and campus facilities (8.4%). The remaining 30 categories accounted for 36.7% of the data, confirming a moderate class imbalance that validated our use of macro-averaged metrics. A linguistic analysis of the dataset confirmed the need for an advanced model. We found that 23.4% of queries contained informal language or colloquialisms, 15.7% included spelling irregularities, and 31.2% exhibited ambiguous phrasing that required deep contextual understanding. These characteristics showed that a simple model would likely fail, validating the necessity for advanced NLP models capable of handling real-world linguistic variations.

Model Performance Comparison

We evaluated three models to test our hypothesis: Model A (TF-IDF-SVM), Model B (LSTM), and our fine-tuned Model C (BERT). The comprehensive results, presented in **Table 3** and visualized in **Figure 3**, show a clear performance hierarchy among the models.

Table 3. Comparative Performance Metrics of Three Models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	Average Inference Time (ms)
Model A (TF-IDF-SVM)	61,5%	0,60	0,58	0,59	12
Model B (LSTM)	84,2 %	0,83	0,81	0,82	95
Model C (BERT)	95,8%	0,96	0,95	0,95	320

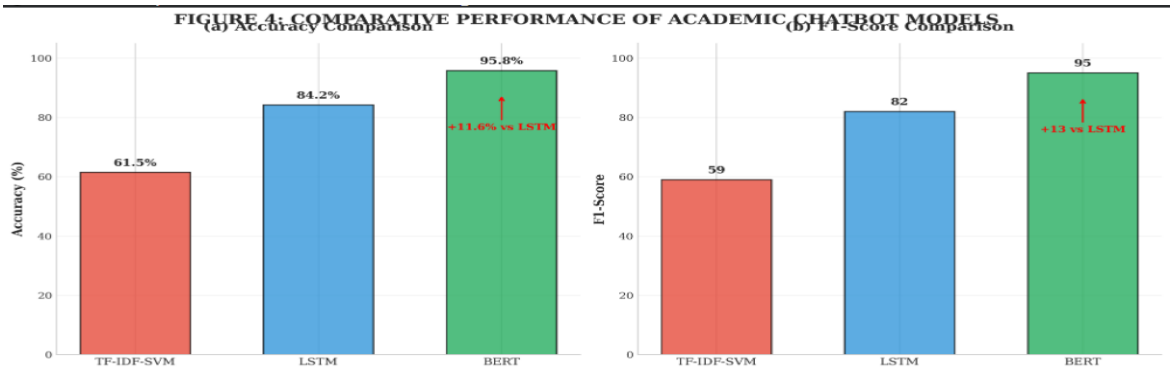

Fig 3. Comparative Performance of Academic Chatbot Models

Our fine-tuned BERT model (Model C) was the definitive winner, achieving a superior accuracy of 95.8% and a macro F1-score of 0.95. This represents a substantial improvement a 34.3 percentage point jump in accuracy over the baseline Model A and an 11.6 percentage point jump over Model B10. We conducted paired t-tests $\boldsymbol{\alpha = 0.05}$ which confirmed that these performance differences were statistically significant **($p < 0.001$ for Model A vs. Model C; $p < 0.01$ for Model B vs. Model C)** .

Detailed Performance Analysis by Intent Category

To understand how the models behaved under stress, we analyzed their performance based on the complexity of the query. We categorized the 35 intents into three complexity levels: Simple, Moderate, and Complex (Table 4).

Table 4. Model Performance by Query Complexity Level

Complexity Level	Number of Intents	Model A F1-Score	Model B F1-Score	Model C F1-Score	Example Intents
Simple (Direct Queries)	12	0,78	0,91	0,98	Tuition Fee, Campus Location
Moderate (Context-Dependent)	15	0,54	0,82	0,95	Course prerequisites,
Complext (High ambiguity)	8	0,32	0,71	0,92	Multi-step processes, Conditional requirements

The data clearly shows that Model C (BERT) maintained high performance (F1-score ≥ 0.92) across all complexity levels. In sharp contrast, Model A's performance collapsed on complex queries, dropping to an F1-score of just 0.32. Model B (LSTM) showed moderate capability but still lagged 21 points behind BERT on the most complex intents. This pattern aligns with our theoretical foundation: BERT's bidirectional architecture allows it to capture the long-range dependencies and contextual nuances needed for complex queries. The unidirectional nature of LSTM and the context-free approach of TF-IDF-SVM are fundamentally limited in their ability to resolve this ambiguity.

Handling Linguistic Variations

A critical objective was to evaluate how robust these models were against the messy, real-world linguistic variations found in user queries. We conducted a targeted analysis on a subset of 300 queries specifically chosen for containing irregular spellings, informal language, or grammatically imperfect structures.

Table 5. Performance on Queries with Linguistic Variations

Linguistic Variation Type	Number of Samples	Model A Accuracy	Model B Accuracy	Model C Accuracy
Spelling irregularities	94	38,3%	76,6%	93,6%
Informal Language /Slang	112	45,5%	79,5%	94,6%
Grammatical Imperfect	94	41,5%	73,4%	92,6%
Overall Variation Subset	300	42,0%	76,8%	93,5%

As shown in Table 5, Model C (BERT) demonstrated remarkable robustness. It maintained an average accuracy of 93.5% on these challenging queries, only a 2.3 percentage point drop from its overall performance. Model B showed moderate resilience (76.8% accuracy), while Model A's performance collapsed to 42.0%, confirming its brittleness against non-standard language. These results empirically validate that BERT's sub-word tokenization (WordPiece) and contextual embeddings allow it to effectively handle the linguistic variability from a diverse user population.

Confusion Matrix Analysis

Figure 2 presents the confusion matrix for Model C (BERT) on the testing dataset, providing insight into the model's error patterns across the 35 intent categories.

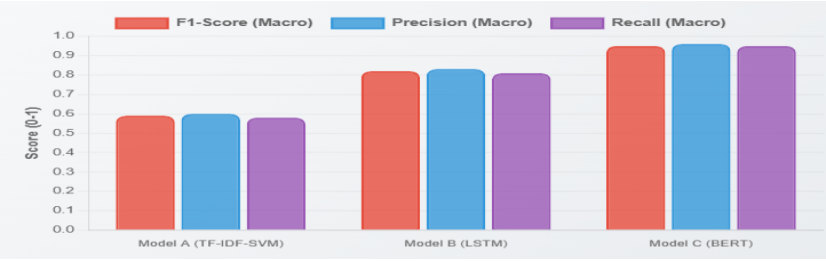

Figure 4. Confusion Matrix for Model C (BERT)

We analyzed the confusion matrix for Model C (BERT) (Figure 2) to get a deeper insight into its error patterns. The matrix revealed that the vast majority (95.8%) of predictions fall on the main diagonal, indicating correct classification. The small number of misclassifications primarily occurred between semantically adjacent categories. For instance, queries about "scholarship application procedures" were occasionally confused with "scholarship eligibility requirements"—categories that share significant semantic overlap. We found no significant misclassifications between semantically distant categories (e.g., no confusion between "admission deadlines" and "parking facilities"), which suggests BERT successfully captured the core semantic intent even when it made fine-grained errors.

Qualitative Analysis of Chatbot Responses

To complement the quantitative metrics and show the practical impact of our model, we conducted a qualitative analysis of real-world query responses. The examples in **Table 6** illustrate how BERT's bidirectional context understanding leads to more accurate and appropriate responses compared to the LSTM baseline. For instance, when a user inquired about their "application status for the computer science program," the LSTM model incorrectly directed them to the computer lab schedule, failing to disambiguate the word "application". In contrast, the fine-tuned BERT model correctly identified the context of admissions and provided a precise link to the student portal. This superior contextual understanding is a direct result of BERT's architecture. The qualitative results also highlight the model's robustness. A query with misspellings and informal grammar like "when is tha last day too pay tution?" was processed correctly by Model C. The LSTM model, however, misinterpreted the intent entirely. This resilience can be attributed to BERT's WordPiece tokenization, which breaks down unknown or misspelled words into known sub-word units. These real-world examples solidify the practical implications of our research. The fine-tuned BERT model's ability to understand complex, multi-part queries... highlights its capacity for managing sophisticated user intents. For an end-user, this translates to a more reliable, helpful, and efficient interaction, reducing frustration and the need for human staff escalation.

Table 6. Qualitative Comparison of Chatbot Responses on Real Academic Queries

Query Type	User Input	Model B (LSTM) Response	Model C (Fine-tuned BERT) Response	Analysis
Contextual Ambiguity	"I need to check my application status for the computer science program."	"Here is the link to the computer lab schedule." (Incorrect)	"You can check your admission application status through the student portal at portal.university.ac.id." (Correct)	BERT correctly disambiguates "application" as related to **admissions**, while LSTM confuses it with software application.
Academic Jargon	"What are the prerequisites for CS401?"	"I don't understand. Could you	"CS401 'Advanced Algorithms' requires CS30 1 'Data Structures' and MATH205	BERT recognizes "prerequisites" and course codes

		rephrase your question?" (Failure)	'Discrete Mathematics' as prerequisites." (Correct)	as specific academic entities, which the generic LSTM fails to process.
Informal Language & Spelling Error	"when is tha last day too pay tution?"	"The last day to submit assignments is Friday." (Incorrect)	"The tuition fee payment deadline for this semester is March 25, 2024." (Correct)	BERT's sub-word tokenization robustly handles misspellings ("tha", "too") and understands the informal query's intent.
Complex, Multi-part Intent	"I want to drop my chemistry lab but keep the lecture. What happens to my financial aid?"	"Dropping courses may affect your financial aid. Please contact the finance office." (Partially Correct)	"Dropping the lab (CHEM101L) while keeping the lecture (CHEM101) may change your enrollment status to part-time, which could affect your financial aid. Please consult with your advisor and the financial aid office before proceeding." (Complete)	BERT understands the multi-clause query, identifies the specific components, and provides a comprehensive, actionable response.
Synonym & Paraphrase Understanding	"How can I get help with writing my research paper?"	"Information about research grants can be found here." (Incorrect)	"The University Writing Center offers one-on-one tutoring sessions for research papers. You can book an appointment on their website." (Correct)	BERT maps "help with writing" to the correct intent (Writing Center services), while LSTM incorrectly associates "research" with grants.

The empirical results strongly confirm our research hypothesis: the domain-specific fine-tuned BERT model (Model C) significantly outperforms both traditional (TF-IDF-SVM) and sequential deep learning (LSTM) baselines for the academic chatbot task. This superiority, culminating in 95.8% accuracy, is attributed to BERT's core architectural advantages. Its bidirectional contextualization allows it to disambiguate meaning in complex academic queries , and its robust sub-word tokenization handles spelling errors and domain-specific jargon with remarkable resilience (93.5% accuracy on irregular queries). The practical implications of this performance are substantial. For an institution receiving 10,000 monthly queries, our achieved accuracy translates to approximately 3,430 more successfully handled queries per month than the TF-IDF-SVM baseline, directly reducing staff workload and enhancing user satisfaction. This study directly addresses a known research gap by demonstrating that domain-specific fine-tuning is not merely beneficial but crucial. It outperforms the 85-88% accuracy range reported in studies that applied pre-trained BERT without domain adaptation (Lee & Kim, 2022). However, we must acknowledge several limitations. The primary threat to external validity is that our dataset comes from a single institution, which may limit generalizability to other academic contexts. While our methodology is replicable, the performance depends on the quality of the training data. Furthermore, our study focused only on English-language queries. Future adaptation would require multi-lingual fine-tuning. From a technical perspective, the model's inference time of 320 ms highlights the inherent trade-off between accuracy and computational efficiency. For resource-constrained environments, future work could explore model distillation or quantization to reduce latency. Another promising direction is generalizing this framework to other specialized domains, such as legal advisory or healthcare information systems. In conclusion, despite these limitations, our study provides a validated and replicable methodology for domain-specific NLP optimization. We offer compelling empirical evidence that fine-tuned BERT is a superior solution for academic chatbots, bridging the gap between powerful general-purpose language models and the specific, high-stakes needs of an educational environment.

4. Conclusions

This study confirms that domain-specific fine-tuning of the BERT algorithm is a highly effective method for optimizing an academic chatbot's NLP engine, significantly outperforming traditional and sequential models. Our fine-tuned BERT model achieved 95.8% accuracy and a 0.95 macro F1-score, proving it has a superior grasp of contextual understanding, intent recognition, and robustness against linguistic variations like misspellings and slang. The key contribution of this work was adapting BERT specifically to the academic domain. This approach bridged a critical research gap where prior studies had often applied generic models without this necessary optimization. Our results provide strong empirical evidence that BERT's bidirectional representations and sub-word tokenization are essential components for building high-performance chatbot systems in an educational environment. Ultimately, this research provides a validated and replicable framework that other institutions can adopt to enhance their automated information delivery and reduce staff workload. Future work should now explore multi-lingual extensions and model distillation techniques to improve computational efficiency and expand the chatbot's usability in more diverse educational settings.

References

Afifa, N., Elektro, F. T., Telkom, U., Saputra, R. E., Elektro, F. T., Telkom, U., Nugrahaeni, R. A., Elektro, F. T., & Telkom, U. (2023). *Implementasi NLP Pada Chatbot Layanan Akademik Dengan Algoritma Bert Implementation Of NLP On Academic Service Chatbot With Bertalgorithm*. *10*(1), 383–387.

Al-ayyubi, A. S., Hadi, A., Novaliendry, D., & Budayawan, K. (2025). *Integrasi Chatbot Akademik Berbasis Natural Language Processing Pada Sistem Informasi Akademik di STEI Ar Risalah*. *9*, 27134–27141.

Al-Zahrani, A. M. (2025). Exploring the Impact of Artificial Intelligence Chatbots on Human Connection and Emotional Support Among Higher Education Students. *SAGE Open*, *15*(2). https://doi.org/10.1177/21582440251340615

Aziza, N. R., Yosrita, E., Ningrum, R. F., Ardanti, T. S., & Arafaizin, S. R. (2023). Pembangunan Aplikasi dan Klasifikasi Pertanyaan Chatbot Informasi Akademik Menggunakan Metode Cosine Similarity dan Naïve Bayes. *Kilat*, *12*(2), 169–179. https://doi.org/10.33322/kilat.v12i2.1921

Babu, A., & Boddu, S. B. (2024a). BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. *Exploratory Research in Clinical and Social Pharmacy*, *13*(February), 100419. https://doi.org/10.1016/j.rcsop.2024.100419

Babu, A., & Boddu, S. B. (2024b). Exploratory Research in Clinical and Social Pharmacy BERT-Based Medical Chatbot : Enhancing Healthcare Communication through Natural Language Understanding. *Exploratory Research in Clinical and Social Pharmacy*, *13*(January), 100419. https://doi.org/10.1016/j.rcsop.2024.100419

Djawa, A. E. W., & Ahda, F. A. (2024). Design of an Academic Services Chatbot at Asia Institute Malang. *Ic-Itechs*, *5*(1), 799–806. https://doi.org/10.32664/ic-itechs.v5i1.1642

Dzaky, A. A., Zeniarja, J., Supriyanto, C., Shidik, G. F., & Paramita, C. (2024). *Optimization Chatbot Services Based on DNN-Bert for Mental Health of University Students*. *8*(1), 13–21.

Fatwanto, A., Zamakhsyari, F., Ndungi, R., & Fitriyani, L. (2024). *R ESEARCH A RTICLE A Systematic Literature Review of BERT-based Models for Natural Language Processing Tasks*. 713–728.

Gao, J., Opute, A. P., Jawad, C., & Zhan, M. (2025). The influence of artificial intelligence chatbot problem solving on customers' continued usage intention in e-commerce platforms: an expectation-confirmation model approach. *Journal of Business Research*, *200*(August), 115661. https://doi.org/10.1016/j.jbusres.2025.115661

Gigi, K., Ibadurrahman, I., Yusuf, F., Budianto, H., & Barat, J. (2025). *(NLP) MENGGUNAKAN MODEL BERT UNTUK MENINGKATKAN PELAYANAN PENDAHULUAN Dalam era digital yang berkembang pesat , teknologi Artificial Intelligence (AI) telah menghadirkan berbagai solusi cerdas untuk meningkatkan efisiensi layanan , salah satunya melal*. *10*(2), 325–334.

Holis, R. M., Eko, P., Utomo, P., & Hutabarat, B. F. (2025). *Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services*. *5*(2), 915–922.

Jatmika, S., Patmanthara, S., Wibawa, A. P., & Kurniawan, F. (2024). Cognition-Based Document Matching Within the Chatbot Modeling Framework. *Journal of Applied Data Sciences*, *5*(2), 613–627. https://doi.org/10.47738/jads.v5i2.209

Kim, M., Lee, S., Kim, S., Heo, J. I., Lee, S., Shin, Y. Bin, Cho, C. H., & Jung, D. (2025). Therapeutic Potential of Social Chatbots in Alleviating Loneliness and Social Anxiety: Quasi-Experimental Mixed Methods Study. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/65589

Lin, X., Wang, X., Shao, B., & Taylor, J. (2024). How Chatbots Augment Human Intelligence in Customer Services: A Mixed-Methods Study. *Journal of Management Information Systems*, *41*(4), 1016–1041. https://doi.org/10.1080/07421222.2024.2415773

Łukasik, A., & Gut, A. (2025). From robots to chatbots: unveiling the dynamics of human-AI interaction. *Frontiers in Psychology*, *16*(April). https://doi.org/10.3389/fpsyg.2025.1569277

Mandlik, D., Chaudhary, R., Kotkar, M., Zende, R., & Bhosale, R. S. (2025). *AI-Powered College Enquiry Chatbot Using NLP with BERT and GPT*. *13*(2), 1–6. www.ijirmps.org

Muhammad, S., Zaidi, H., Ahmed, S., Husain, I., Qureshi, B. H., Naz, S. A., & Razaque, A. (2025). *A HYBRID AI-BASED UNIVERSITY STUDENTS QUERIES CHATBOT USING NLP AND SBERT TECHNOLOGIES*. *3138*, 275–288.

Munawirah, M., Mardjani, E., & Kristian, K. (n.d.). *Aplikasi Chatbot Berbasis Dialogflow dan NLP untuk Pelayanan Informasi Akademik pada Fakultas Ilmu Komputer Universitas Tomakaka*. 99–115.

No, V., Waleska, R. F., & Asnal, H. (2025). *Edumatic : Jurnal Pendidikan Informatika Aplikasi Chatbot Interaktif Pembelajaran Bahasa Pemrograman PHP dengan Algoritma NLP berbasis BERT*. *9*(3), 609–618. https://doi.org/10.29408/edumatic.v9i3.31427

Praneeth, K. R., Ruprah, T. S., Madhuri, J. N., & Sreenivasulu, A. L. (2024). *Optimizing Customer Interactions : A BERT and Reinforcement Learning Hybrid Approach to Chatbot Development*. *15*(9), 569–578.

Priyatno, A. M., Prasetya, M. R. A., Cholidhazia, P., & Sari, R. K. (2024). Comparison of Similarity Methods on New Student Admission Chatbots Using Retrieval-Based Concepts. *Journal of Engineering and Science Application*, *1*(1), 32–40. https://doi.org/10.69693/jesa.v1i1.2

Rahmat, M. A., Karmila, Khatami, H. M., & Muhammad Alief Fahdal Imran Oemar. (2025). Penerapan Model BERT pada Chatbot dalam Platform E-Commerce. *Adopsi Teknologi Dan Sistem Informasi (ATASI)*, *4*(1), 72–79. https://doi.org/10.30872/atasi.v4i1.3039

Setiawan, G. H., & Adnyana, I. M. B. (2023). Improving Helpdesk Chatbot Performance with (TF-IDF) and Cosine Similarity Models. *Journal of Applied Informatics and Computing*, *7*(2), 252–257.

Singh, S. U., & Namin, A. S. (2025). A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal*, *10*(August 2024), 100128. https://doi.org/10.1016/j.nlp.2025.100128

Triawan, P., & Tahyudin, I. (2025). *Impact of NLP Algorithms on Sentiment Analysis Efficiency and Accuracy*. *7*(3), 2684–2709. https://doi.org/10.51519/journalisi.v7i3.1222