Application of the SBERT Model for Optimizing Answers to Frequently Asked Questions in Campus Academic Services

Fajri Profesio Putra[1], I Gusti Agung Putu Mahendra[2]
[1]*Informatic Technology Department, Politeknik Negeri Bengkalis, Indonesia*
Author's fajri@polbeng.ac.id[1], agungmahendra@polbeng.ac.id[2]

Abstract

This study introduces an academic chatbot on Telegram that uses dense retrieval with SBERT (Paraphrase Multilingual MiniLM-L12-V2) combined with category-menu navigation for quick exploration. The system is designed to be classroom-ready and compatible with various devices, running on Google Colab without a webhook. It supports selective answering through a similarity threshold (τ) to manage fallback and low-confidence responses. A curated corpus of question-answer pairs is indexed as vectors, and at inference, queries are embedded and matched via cosine similarity. Evaluation with 140 questions across seven categories includes top-1 accuracy, fallback rate, wrong-confidence rate, and latency. Results show that a similarity threshold of 0.15–0.30 yields a top-1 accuracy of 72.14%, with a fallback rate of 0% and a wrong-confidence rate of 27.86%. Increasing the threshold to 0.50 lowers the wrong-confidence rate to 20.71% but reduces top-1 accuracy to 67.14% and raises the fallback rate to 12.14%. The highest accuracy is in the Academic category (90%), while the lowest is in the Organization category (55%). The median latency is 26.7 ms. Key contributions include: (i) integrating Telegram menus with SBERT into a seamless interaction flow, (ii) explicitly calibrating τ for selective answering, and (iii) providing an easy-to- replicate, lab-ready blueprint. Findings highlight a trade-off between risk and coverage, suggesting that intent-adaptive τ, along with reranking and calibration, are promising directions for future work.

Keywords : SBERT, Dense Retrieval, Telegram Chatbot, Academic FAQ, Selective Answering.

1. INTRODUCTION

Campus academic services usually get asked the same questions over and over again (about course registration, tuition fees, password resets, legalization, and advising schedules, for example). These questions create service bottlenecks, resulting in slow responses, inconsistent answers, and a growing administrative workload. Conversely, generative LLM solutions often carry hallucination risks and computational costs that do not align with the needs of structured FAQ services. (Sivakolundhu & Yagamurthy, 2024).

The proposed concept is a hybrid chatbot with category-menu navigation of question lists for quick exploration and dense retrieval (multilingual SBERT) for free typing. This strategy preserves answer determinism (limited to a curated corpus) while allowing for linguistic flexibility in user paraphrases. (Wu et al., 2020). Sentiment analysis has been added as a contextual cue for the service flow, which escalates when a negative tone is detected.(Dongbo et al., 2023).

Many studies focus on full LLMs or TF-IDF/BM25; few explore a compact, device-friendly, classroom-ready hybrid of menu+SBERT (easy to run in Colab without a webhook). Differences from prior work include: (i) combining Telegram inline-menu UI with semantic retrieval in a single interaction flow, (ii) designing and explicitly evaluating a similarity threshold to control fallback, and (iii) a practical lab/teaching blueprint.

Campuses need a quick, precise, low-cost Q&A medium for recurring FAQs; the solution should minimize hallucinations and be maintainable by non-ML teams (editing the Q–A corpus is sufficient).

2. REVIEW OF LITERATURE

Campus/Tourism Service Chatbots Based on NLP. Relevant local studies show two dominant architecture patterns for information-service chatbots: (i) intent–response based on an intents.json (tags, patterns, responses) trained simply and served via a micro-web framework (e.g., Flask); and (ii) classical NLP pipelines (tokenizing–filtering–pattern matching) for

preprocessing and response matching. In the tourism context, a recent study built a Tangerang Raya tourism helpdesk with Flask; the knowledge base and dialog behavior were modeled via intents.json (tag/pattern/response), then trained-tested before being accessed by real users through a web interface. The design also emphasized local testing procedures, validation of training data structure, and keyword-based testing scenarios, and suggested integration with WhatsApp/Telegram for broader operational use (Hadinata & Stianingsih, n.d.).

In the campus domain, academic-service NLP systems report two types of evidence: (a) user acceptance via questionnaires (227 respondents, average score 3.55 "very good"), and (b) answer accuracy over 40 QAs (37 correct; 92.5%). The described workflow starts from intent identification, input processing, and output presentation; mentioned NLP foundations include tokenizing–filtering and string matching (e.g., KMP) to aid pattern search. Overall metrics focus on accuracy and satisfaction, while abstention/fallback policies and confidence calibration are not explicitly reported. These findings position NLP-based systems as practical solutions that help students obtain academic information with high accuracy on small corpora (Muliyono & Sumijan, 2021).

The two strands above position chatbots as structured Q&A interfaces over relatively static knowledge—academic (calendars, procedures, credits, community service) or tourism (locations, operating hours, ticket prices). State-of-practice implementations include: (1) Flask orchestration for real-time interaction; (2) data modeling via intents.json (tag/pattern/response) as training sources for simple networks; (3) test scenarios that verify responses meet expectations for each keyword. This yields a cheap, fast adoption path maintainable by small teams—yet it rarely evaluates the risk–coverage aspect (e.g., wrong-confidence vs fallback) and has not systematically leveraged reranking or probability calibration. Our evaluation of accuracy, wrong-confidence, and fallback across threshold (τ) variations and category analysis fills an important gap not reported in prior works, thus enriching the state of the art from the standpoint of selective answering and answer-risk management.

3. METHOD

Research type. Research & Development with experiments. Setting. Program-level environment (Colab/Notebook simulation). Tools & Libraries. Python 3.10+, Google Colab T4, sentence-transformers, scikit-learn, transformers, python-telegram-bot, numpy/pandas, nest_asyncio.

Arsitektur Sistem.

1. **Pra-proses**: *lowercasing*, *punctuation scrub*, whitespace normalization.
2. **Embedding**: SBERT paraphrase-multilingual-MiniLM-L12-v2 for all questions (vector index).
3. **Inferensi**: *user query* → *clean* → embedding → **cosine similarity** to the index; threshold **τ=0,30**.
4. **UI/UX**: Telegram **InlineKeyboard**: ordered categories → list of questions → answer; free-text handler for retrieval.
5. **Sentiment** : cardiffnlp/twitter-roberta-base-sentiment-latest as a (negative tone → suggest contacting IT/BAAK).
6. **Failure handling**: If the score is less than the threshold, the bot asks for clarification to reduce wrong confidence.

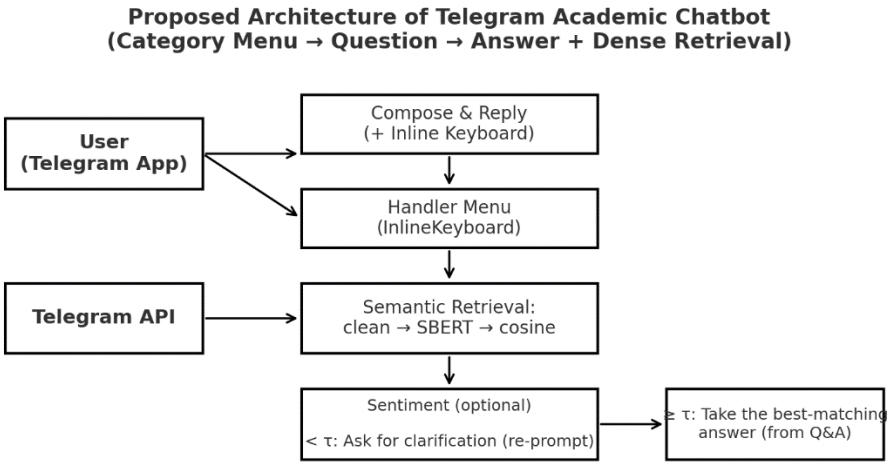

Figure 1. Research Flow

4. RESULT & DISCUSSION

Result

Based on 140 test questions, confidence thresholding shows that values of $\tau = 0.15$ and 0.30 produce identical performance, with top-1 accuracy of 72.14% (101/140), fallback of 0%, and wrong-confident of 27.86%. At a confidence threshold of 0.50, the accuracy drops to 67.14% (94/140), with a fallback rate of 12.14% (17/140) and a wrong-confident rate of 20.71% (29/140). Thus, the optimal threshold in terms of top-1 accuracy is 0.15, equivalent to 0.30. We select 0.15 as the operating point since it provides full coverage without triggering fallback. Fallback equals zero at τ values between 0.15 and 0.30 because all top-predicted scores surpass the threshold, so abstention never activates, even though some of those predictions are incorrect. The highest accuracy is in the Academic category (90%), followed by Research/Technology/Facilities (75%), Administration (70%), and Services (65%). The lowest accuracy is in the Organization category (55%). The most frequent errors occur with intents that have overlapping location, time, or procedure semantics (e.g., "campus ATM" is confused with "cafeteria," or "library opening hours" with "admin office schedule"). In terms of response time, the system meets real-time needs with a median of 26.7 ms, a mean of 27.2 ms, and a p90 of 29.9 ms.

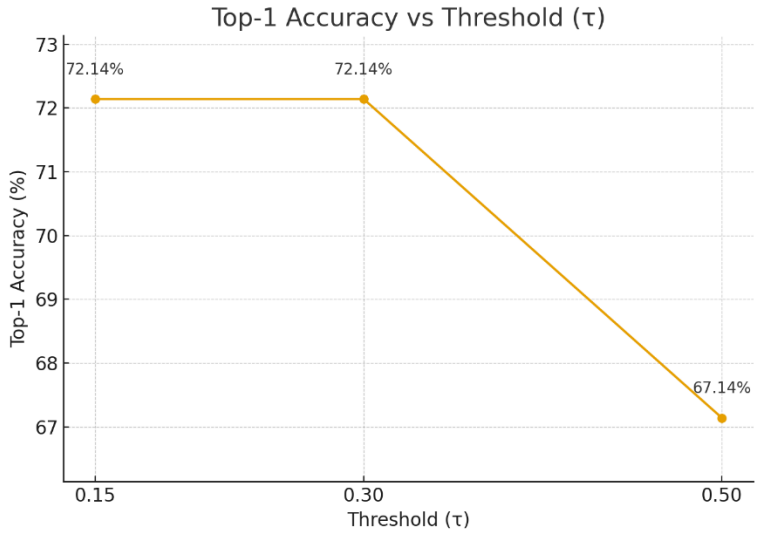

Figure 2. Similarity Threshold Values

Table 1. Accuracy Results Under Similarity Thresholds

τ	n	Top1	Fallback	wrong_confident	top1_acc	fallback_rate	wrong_confident_rate
0.15	140	101	0	39	72.142857	0.000000	27.857143
0.30	140	101	0	39	72.142857	0.000000	27.857143
0.50	140	94	17	29	67.142857	12.142857	20.714286

Table 2. Accuracy by Question Category

true_cat	n	top1	fallback	wrong	top1_acc_%	fallback_%	wrong_%
Administration	20	14	0	6	70.0	0.0	30.0
Academics	20	18	0	2	90.0	0.0	10.0
Facilities	20	15	0	5	75.0	0.0	25.0
Services	20	13	0	7	65.0	0.0	35.0
Organizations	20	11	0	9	55.0	0.0	45.0
Research	20	15	0	5	75.0	0.0	25.0
Technology	20	15	0	5	75.0	0.0	25.0

Discussion

First, the threshold range ($\tau \approx 0.15$–0.30) maximizes Top-1 accuracy at 72.14%. However, raising the threshold to 0.50 reduces wrong-confidence from 27.86% to 20.71%, but results in a 12.14% increase in fallback and a decrease in correct-answer coverage to 67.14%. This aligns with the hypothesis that increasing the threshold pushes the system toward a more conservative answering strategy. Precision increases, but some medium-score cases that were previously correct are filtered out and become abstentions or fallbacks.

The emergence of this pattern can be traced back to the similarity-score distribution and the characteristics of the questions. At low τ, nearly all predictions are accepted, so abstention never triggers. (Karpukhin et al., 2020). Consequently, when semantic representations are "shallow"—especially for queries about location, time, or procedure—the model tends to select similar phrasings, even if they are not factually identical (e.g., "campus ATM" is confused with "cafeteria"). As τ increases, many hesitant predictions (including those with incorrect confidence levels) are held back, which reduces risk but also causes some medium-score correct answers to be lost. (Zhai et al., 2023). Cross-category performance differences reinforce this explanation: the Academic domain, definitional and relatively static, reaches 90%, while Organization and Services more dynamic and context-overlapping stay at 55-65%.

Compared with dense-retrieval FAQ literature, these results are consistent with the risk–coverage curve: higher thresholds raise precision while lowering coverage (Herzig et al., 2021). A notable finding is the high frequency of wrong confidence on practical questions (e.g., where, when, how much), indicating the need to strengthen disambiguation in the reranking or intent modeling stage. Theoretically and empirically, common approaches such as score calibration to probability (Platt/isotonic), top-1 vs. top-2 margin rules for abstention, adding hard negatives across similar locations/schedules, and cross-encoder rerankers are reported to reduce wrong-confidence without significantly reducing coverage.

Implications: theoretically, this experiment validates the importance of selective answering and decision calibration based on thresholds/margins rather than a single rigid threshold, and encourages intent-adaptive τ (higher for location/time/procedure intents, moderate for definitions) (Zhang et al., 2020). Practically, operating-point selection should follow operational goals: if answer safety is the priority, $\tau \approx 0.50$ is reasonable with an automatic clarification flow on fallback; if maximizing hit-rate is the priority, $\tau \approx 0.30$ can be paired with margin rules and a reranker to control wrong-confidence (Sachan et al., 2023). On the data side, curating labeled paraphrases and injecting hard negatives for similar intent pairs, plus targeted logging for active learning on weaker categories (Organization/Services), should improve accuracy without sacrificing agility (Karan & Šnajder, 2018).

5. CONCLUSION

Testing 140 questions reveals a clear trade-off between coverage and the risk of incorrect answers due to overconfidence. A tau range of approximately 0.15–0.30 yields the highest top-1 accuracy (72.14%) with 0% fallback, but results in 27.86% of answers being wrongly confident. Increasing the τ range to 0.50 reduces wrong-confidence responses to 20.71%, though accuracy decreases to 67.14%, and fallback responses appear at 12.14%. Cross-category variation confirms that definitional/static domains (academic) are easier (90%) than contextual/dynamic ones (organization, services), which have overlapping semantics (55%-65%). Overall, the system meets real-time response needs with a median response time of 26.7 ms. Therefore, the "best" τ depends on operational goals. To maximize the hit rate, use a τ of approximately 0.30 (recommended with margin rules/reranking to suppress wrong confidence). To minimize misleading answers, choose a τ of approximately 0.50 with a clarification flow upon fallback. In the future, intent-adaptive τ, score calibration, hard negatives, and reranking are expected to address gaps in location, time, and procedure intents without sacrificing coverage.

6. ACKNOWLEDGEMENTS (optional)

Indicate sources of funding or help received in carrying out your study and/or preparing the manuscript if any before the **references**.

7. REFERENCES

Dongbo, M., Miniaoui, S., Fen, L., Althubiti, S. A., & Alsenani, T. R. (2023). Intelligent chatbot interaction system capable for sentimental analysis using hybrid machine learning algorithms. *Information Processing & Management*, *60*(5), 103440. https://doi.org/10.1016/j.ipm.2023.103440

Hadinata, W., & Stianingsih, L. (n.d.). *Implementasi Natural Language Processing pada Chatbot Untuk Helpdesk Informasi Wisata (Studi kasus: Tangerang Raya)*.

Herzig, J., Müller, T., Krichene, S., & Eisenschlos, J. (2021). Open Domain Question Answering over Tables via Dense Retrieval. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 512–519. https://doi.org/10.18653/v1/2021.naacl-main.43

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Muliyono, M., & Sumijan, S. (2021). Identifikasi Chatbot dalam Meningkatkan Pelayanan Online Menggunakan Metode Natural Language Processing. *Jurnal Informatika Ekonomi Bisnis*, 142–147. https://doi.org/10.37034/infeb.v3i4.102

Sivakolundhu, R., & Yagamurthy, D. N. (2024). Adaptive Chatbots: Real-Time Sentiment Analysis for Customer Support. *International Journal of Computing and Engineering*, *6*(1), 55–64. https://doi.org/10.47941/ijce.2123

Wu, E. H.-K., Lin, C.-H., Ou, Y.-Y., Liu, C.-Z., Wang, W.-K., & Chao, C.-Y. (2020). Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot. *IEEE Access*, *8*, 77788–77801. https://doi.org/10.1109/ACCESS.2020.2988252

Zhai, Q., Zhu, W., Zhang, X., & Liu, C. (2023). Contrastive Refinement for Dense Retrieval Inference in the Open-Domain Question Answering Task. *Future Internet*, *15*(4), 137. https://doi.org/10.3390/fi15040137

Zhang, S., Liu, H., Hu, M., Jiang, A., Zhang, L., Xu, F., & Hao, G. (2020). An Adaptive CEEMDAN Thresholding Denoising Method Optimized by Nonlocal Means Algorithm. *IEEE Transactions on Instrumentation and Measurement*, *69*(9), 6891–6903. https://doi.org/10.1109/TIM.2020.2978570