Indonesian Journal of Data and Science

Volume 7 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i1.341

Fine-Tuning a Large Language Model on Vertex AI for a New Student Registration Chatbot at Universitas Muhammadiyah Makassar

Desi Anggreani [1,*]; Muhyiddin A M Hayat [2]; Lukman [3]; Ahmad Faisal [4]; Khadijah [5]; Darniati [6]
[1] Universitas Muhammadiyah Makassar, Makassar, Indonesia, desianggreani@unismuh.ac.id
[2] Universitas Muhammadiyah Makassar, Makassar, Indonesia, muhyiddin@unismuh.ac.id
[3] Universitas Muhammadiyah Makassar, Makassar, Indonesia, Lukman@unismuh.ac.id
[4] Universitas Muhammadiyah Makassar, Makassar, Indonesia, 105841100121@student.unismuh.ac.id
[5] Universitas Muhammadiyah Makassar, Makassar, Indonesia, stkhadijah@unismuh.ac.id
[6] Universitas Muhammadiyah Makassar, Makassar, Indonesia, darniati@unismuh.ac.id

Correspondence should be addressed to Desi Anggreani; desianggreani@unismuh.ac.id

Received 02 December 2025; Accepted 20 Jan 2026; Published 30 March 2026

Abstract:
This study addresses the limitations of manual admission services at Universitas Muhammadiyah Makassar, which often result in delayed and inconsistent information delivery. To overcome these challenges, an institution-specific chatbot was developed by fine-tuning the Gemini 2.5 Flash model on the Google Cloud Vertex AI platform. The model was trained using a curated domain-specific dataset of 1,430 question–answer pairs derived from official documents and frequently asked questions. The fine-tuning process employed supervised learning to enhance contextual relevance and response accuracy. System performance was evaluated using automated text quality metrics, achieving an average BLEU score of 0.23526 and a ROUGE-L Recall score of 0.53424, indicating satisfactory lexical and semantic similarity. Furthermore, a user acceptance evaluation involving 52 respondents yielded a Customer Satisfaction Score (CSAT) of 84.2%, reflecting high user satisfaction. These results demonstrate that fine-tuning a Large Language Model (LLM) for specific institutional needs effectively improves both response quality and service reliability. Ultimately, this approach offers a practical and scalable solution for modernizing student admission services in higher education, ensuring that prospective students receive accurate information in a timely and efficient manner.

Keywords: Large Language Model, Chatbot, Vertex AI, Fine-tuning, Gemini 2.5, NLP, Google Cloud.

1. Introduction

The evolution of information technology has fundamentally transformed how educational institutions interact with prospective students, particularly during the crucial new student admission (PMB) process. In a competitive higher education landscape, providing efficient, responsive, and interactive information services is paramount [1]. However, many institutions, including Universitas Muhammadiyah Makassar, still rely on manual communication channels that are constrained by operational hours and limited human resources, leading to delays and a suboptimal applicant experience.

To address these limitations, artificial intelligence-powered chatbots have emerged as a viable solution, even for specific domains like checking the halal status of food ingredients [2]. Early research demonstrated the effectiveness of chatbots using traditional Natural Language Processing (NLP) techniques [3]. However, these rule-based or conventional machine learning models often struggle to comprehend the diversity of natural language, failing to respond to inputs that deviate from pre-programmed patterns. More recent advancements have explored the use of Large Language Models (LLMs), which offer superior contextual understanding and text generation capabilities [4], [5], [6], [7]. Approaches like Retrieval-Augmented Generation (RAG) have shown promise but remain dependent on

external search systems and often lack deep integration with an institution's internal knowledge base [8]. The application of LLMs has expanded into various fields, from computational social science to smart policing systems, demonstrating their versatility [9], [10].

This study fills this gap by implementing a chatbot powered by a fine-tuned LLM. The research objective is to adapt the versatile Gemini-2.5-flash model to the specific domain of student admissions at Universitas Muhammadiyah Makassar. By training the model on a localized, institution-specific dataset, we hypothesize that the resulting chatbot will provide more accurate, relevant, and context-aware responses than both traditional chatbots and non-fine-tuned LLMs. The primary contribution of this work is a practical framework for developing a specialized academic chatbot that enhances service efficiency and user satisfaction through targeted model adaptation.

2. Method

This research was designed as a system development and evaluation project. The process included data collection, model fine-tuning, system integration, and performance testing.

1. Data Collection and Preparation

A domain-specific dataset was curated from various sources to ensure comprehensive coverage of admission-related topics. Primary data included Frequently Asked Questions (FAQs) from the university's official PMB website, direct inquiries documented at the PMB office, and spontaneous questions observed during live social media sessions on platforms like TikTok. This initial dataset was then enriched through data augmentation using another generative model to create variations in phrasing and expand concise answers. The final dataset consisted of 1,430 question-answer pairs. For fine-tuning, the data underwent pre-processing, which included cleaning irrelevant characters, normalizing text to a consistent format, and structuring it into JSON Lines (JSONL) format, where each line represents a user-model conversational turn.

2. Model Fine-Tuning

The core of this research is the fine-tuning of the Gemini-2.5-flash model, a powerful and efficient LLM developed by Google [11]. This process was conducted on the Google Cloud Vertex AI platform. The processed JSONL dataset was uploaded to a Google Cloud Storage bucket. A supervised fine-tuning job was configured in Vertex AI with the parameters detailed in **Table 1**.

Table 1. Fine-Tuning Configuration Parameters in Google Cloud Vertex AI.

Parameter	Value
Base Model	gemini-2.5-flash
Tuning Method	Supervised
Region	37
Learning Rate Multiplier	5
Adapter Size	4

The training performance of the fine-tuned Gemini 2.5 Flash model across multiple epochs is illustrated in **Figure 1**, which presents the trends of training accuracy and training loss. As shown in the figure, the training accuracy (blue curve) increases steadily from approximately 73% in the early epochs to about 97.6% at the final epoch, indicating that the model progressively learned to generate more accurate admission-related responses. At the same time, the training loss (red dashed curve) consistently decreases from around 0.9 to approximately 0.07, demonstrating effective optimization during the fine-tuning process. As training progresses, **Figure 1** also shows that after approximately 25 epochs, the improvement in accuracy becomes more gradual and the reduction in loss starts to stabilize. This pattern

suggests that the model has reached a near-optimal convergence state, where additional epochs contribute marginal improvements. The simultaneous increase in accuracy and decrease in loss indicate stable training behavior without signs of overfitting or divergence.

Overall, the convergence pattern depicted in **Figure 1** confirms that supervised fine-tuning using institution-specific data was effective in adapting the Gemini 2.5 Flash model to the new student admission domain. These results support the suitability of the trained model for subsequent evaluation using automated text quality metrics and user satisfaction analysis.

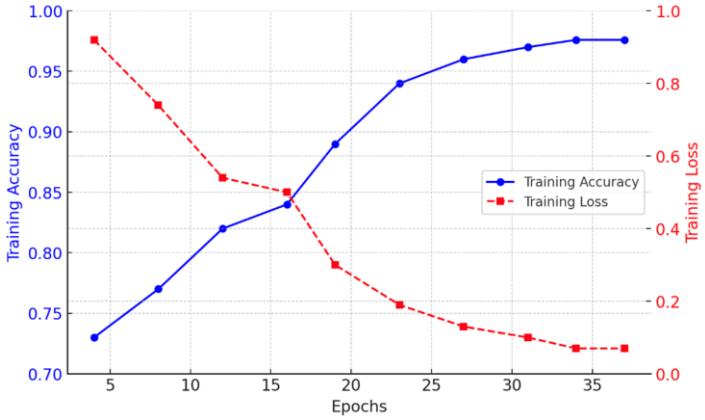

Figure 1. Training Accuracy and Training Loss over Epochs

3. System Architecture and Evaluation

The fine-tuned model was integrated into a user-facing chatbot application using the Flask web framework. As shown in **Figure 2**, the system operates by receiving user input, sending it to the fine-tuned model's endpoint on Vertex AI via an API call, and displaying the generated response back to the user.

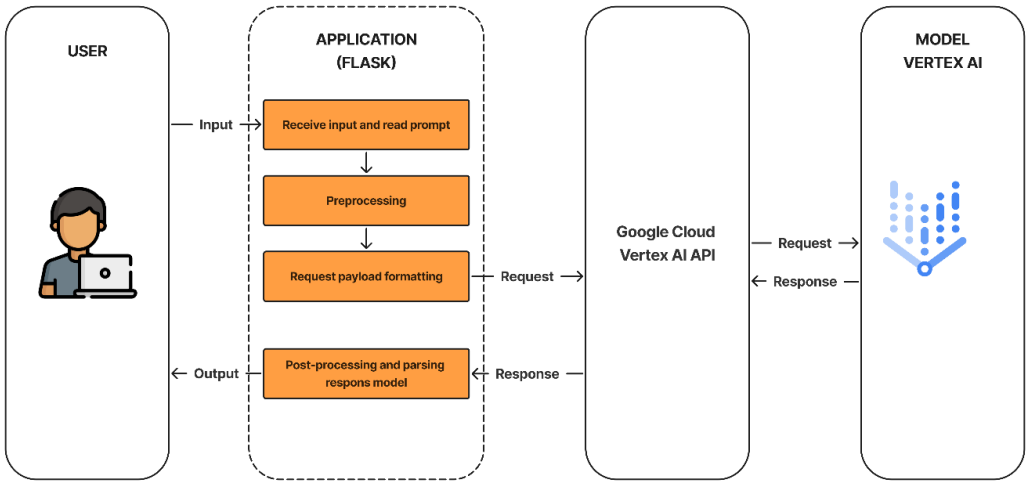

Figure 2. System workflow illustrating the interaction between the User, the Flask Application, and the fine-tuned Vertex AI Model.

System performance was evaluated using two primary methods:

a. Automated Text Quality Metrics

To assess the linguistic quality of the chatbot's responses, we used BLEU (Bilingual Evaluation Understudy) [12] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [13].

$$BLEU = BP.\exp(\sum_{n=1}^{n} w_n Log\ p_n) \quad (1)$$

$$ROUGE - N = \frac{\sum S \in \{ReferenceSummaries\} \sum gram_n \in S\ Count_{match}(gram_n)}{\sum S \in \{ReferenceSummaries\} \sum gram_n \in S\ Count(gram_n)} \quad (2)$$

b. User Satisfaction Survey

To measure user acceptance, we conducted a survey of 52 participants using the Customer Satisfaction Score (CSAT) methodology on a 5-point Likert scale. This approach aligns with efforts to make algorithm performance more understandable to non-expert users, which has also been explored through visual representations of metrics like confusion matrices [14], [15].

3. Result and Discussion

Performance Evaluation

The automated evaluation results of the fine-tuned chatbot are summarized in **Table 2**, which reports the average scores obtained from multiple testing iterations using BLEU and ROUGE-L metrics. As presented in **Table 2**, the chatbot achieved an average BLEU score of 0.23526, indicating a reasonable level of lexical similarity between the generated responses and the reference answers within the admission domain. In terms of semantic coverage, **Table 2** shows that the ROUGE-L Recall score reached 0.53424, suggesting that more than half of the key information from the reference answers was successfully captured in the chatbot responses. This relatively high recall value indicates that the fine-tuned model is effective at including essential admission-related details in its outputs. Meanwhile, the ROUGE-L Precision score of 0.30536 reflects moderate precision, implying that while the responses are informative, some additional or less relevant content may still be present.

The combined ROUGE-L F1-Score of 0.38208, as reported in **Table 2**, represents a balanced trade-off between precision and recall, confirming that the chatbot provides responses that are both informative and contextually relevant. Overall, the results summarized in **Table 2** demonstrate that institution-specific fine-tuning substantially improved the model's response quality, making it suitable for deployment in new student admission information services.

Table 2. Average Automated Evaluation Scores from Five Testing Iterations.

Metric	Average Score
BLEU	0.23526
ROUGE-L Precision	0.30536
ROUGE-L Recall	0.53424
ROUGE-L F1-Score	0.38208

The high recall score (0.53424) indicates that the chatbot was effective at including the most important information in its responses. A qualitative comparison clearly illustrates the value of fine-tuning, as shown in **Figure 3**.

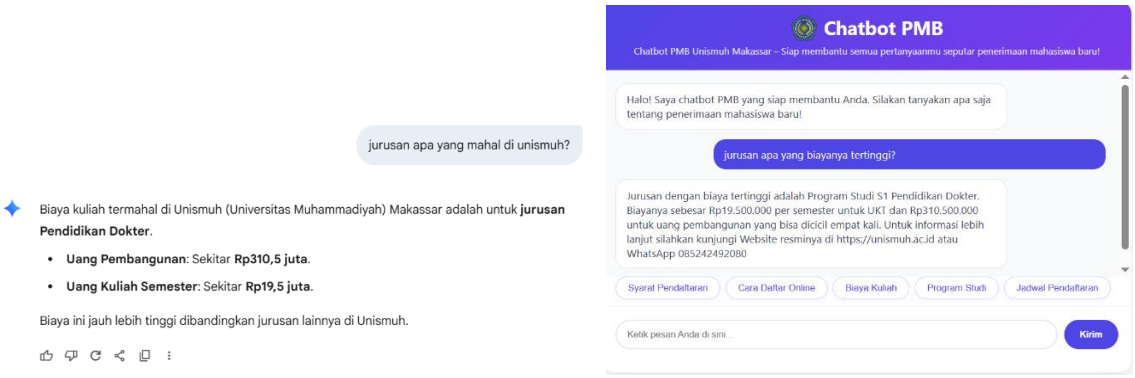

Figure 3. Comparison of responses from the base model (left) and the fine-tuned model (right)

User Satisfaction

The user satisfaction survey revealed a high level of acceptance. The calculated CSAT score was 84.2%. **Table 3** shows the frequency distribution of the responses.

Table 3. Frequency Distribution of User Satisfaction Responses.

Category (Score)	Number of Responses	Percentage (%)
Very Unsatisfied (1)	1	0.2
Unsatisfied (2)	3	0.6
Neutral (3)	78	15.0
Satisfied (4)	233	44.8
Very Satisfied (5)	205	39.4
Total	520	100.0

The distribution of user satisfaction responses shows that the majority of users expressed positive perceptions of the chatbot's performance. Most responses were categorized as "Satisfied" and "Very Satisfied," accounting for 44.8% and 39.4% of the total responses, respectively. This indicates that users generally perceived the chatbot as helpful, reliable, and effective in providing new student admission information. A smaller proportion of responses, 15.0%, fell into the "Neutral" category, suggesting that while the chatbot met basic user expectations, certain aspects such as response clarity or level of detail could still be improved. Negative feedback was minimal, with only 0.6% of responses classified as "Unsatisfied" and 0.2% as "Very Unsatisfied," indicating a very low level of dissatisfaction among users. Overall, the distribution of responses reflects a strong level of user acceptance, with the vast majority of participants reporting positive experiences. These results further support the effectiveness of the fine-tuned LLM in delivering admission-related information services that align with user needs and expectations.

The calculated Customer Satisfaction Score (CSAT) of 84.2% was obtained from a post-interaction user satisfaction survey conducted after users interacted with the chatbot system. The survey involved 52 respondents, each of whom evaluated the chatbot using 10 questionnaire items based on a 5-point Likert scale, resulting in a total of 520

response entries. The CSAT value was computed by aggregating the proportion of positive responses, defined as ratings of Satisfied (4) and Very Satisfied (5), relative to the total number of responses.

Discussion

The research findings demonstrate that fine-tuning an LLM like Gemini-2.5-flash with a domain-specific dataset is a highly effective strategy for creating specialized chatbots, whether in the context of academic services [16], chemical sciences, or biomedical applications [17], [18], [19], [20], [21]. The quantitative metrics, particularly the high ROUGE-L recall, confirm the model's ability to generate informative answers. The significant improvement over the base model highlights a key limitation of relying on general-purpose LLMs for specialized tasks: they are prone to challenges such as "hallucination" [22], biases inherited from training data [23], and a lack of transparency or explainability [24]. By grounding the model in verified, institution-specific admission data, fine-tuning mitigates these challenges and improves response reliability. These findings align with previous studies emphasizing the critical role of training data in shaping LLM behavior across knowledge-intensive domains [25], and further support the adoption of fine-tuned LLMs for academic information services.

4. Conclusion

This study aimed to develop and evaluate an institution-specific chatbot for new student admission services at Universitas Muhammadiyah Makassar through fine-tuning the Gemini 2.5 Flash Large Language Model. The evaluation results demonstrate that the proposed approach achieved satisfactory response quality, as indicated by an average BLEU score of 0.23526 and a ROUGE-L Recall score of 0.53424, while user acceptance assessment yielded a Customer Satisfaction Score (CSAT) of 84.2%. These findings confirm that targeted fine-tuning can effectively enhance both the linguistic quality and practical usability of LLM-based chatbots in higher education information services. However, this study is subject to limitations, particularly the relatively limited number of evaluation rounds and the short deployment period of the chatbot, which may affect the generalizability of the results.Future research is recommended to conduct longer-term deployments and involve a larger and more diverse user population to obtain more comprehensive evaluation results. In addition, subsequent studies may explore the integration of retrieval-augmented generation (RAG) techniques and real-time knowledge base updates to further improve response accuracy and adaptability to policy changes. Comparative evaluations with other LLM architectures and fine-tuning strategies could also be conducted to identify more optimal configurations for institution-specific academic information services.

References:

[1] Fahmi Yusron Fiddin, A. Komarudin, and M. Melina, “Chatbot Informasi Penerimaan Mahasiswa Baru Menggunakan Metode FastText dan LSTM,” *J. Appl. Comput. Sci. Technol.*, vol. 5, no. 1, pp. 33–39, 2024, doi: 10.52158/jacost.v5i1.648.

[2] I. A. E. Z. & N. A. M. J. C. Suardi, D. Anggreani, A.P. Wibawa, N. Murtadlo, “Asking a chatbot for food ingredients halal status,” in *Halal Development: Trends, Opportunities and Challenges*, 2021, pp. 14–20. doi: 10.1201/9781003189282-6.

[3] C. Ehrett, S. Hegde, K. Andre, D. Liu, and T. Wilson, “Leveraging Open-Source Large Language Models for Data Augmentation to Improve Text Classification in Surveys of Medical Staff (Preprint),” *JMIR Med. Educ.*, vol. 10, 2023, doi: 10.2196/51433.

[4] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[5] T. B. Brown *et al.*, “Language models are few-shot learners,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.

[6] S. Bubeck *et al.*, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” 2023.

[7] J. Li, T. Tang, W. X. Zhao, and J. R. Wen, “Pretrained Language Models for Text Generation: A Survey,”

IJCAI Int. Jt. Conf. Artif. Intell., vol. 1, no. 1, pp. 4492–4499, 2021, doi: 10.24963/ijcai.2021/612.

[8] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, 2020.

[9] P. Sarzaeim, Q. H. Mahmoud, and A. Azim, "A Framework for LLM-Assisted Smart Policing System," *IEEE Access*, vol. 12, pp. 74915–74929, 2024, doi: 10.1109/ACCESS.2024.3404862.

[10] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can Large Language Models Transform Computational Social Science?," *Comput. Linguist.*, vol. 50, no. 1, pp. 237–291, 2023, doi: 10.1162/coli_a_00502.

[11] OpenAI *et al.*, "GPT-4 Technical Report," vol. 4, pp. 1–100, 2024.

[12] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2002-July, no. July, pp. 311–318, 2002.

[13] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Association for Computational Linguistics, 2004, pp. 74–81. doi: 10.1253/jcj.34.1213.

[14] H. Shen, H. Jin, Á. A. Cabrera, A. Perer, H. Zhu, and J. I. Hong, "Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance," *Proc. ACM Human-Computer Interact.*, vol. 4, no. CSCW2, 2020, doi: 10.1145/3415224.

[15] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[16] W. Alkishri, J. H. Yousif, Y. N. Al Husaini, and M. Al-Bahri, "Conversational AI in Education: A General Review of Chatbot Technologies and Challenges," *J. Logist. Informatics Serv. Sci.*, vol. 12, no. 3, pp. 264–282, 2025, doi: 10.33168/JLISS.2025.0316.

[17] W. Xia, C. Qin, and E. Hazan, "Chain of LoRA: Efficient Fine-tuning of Language Models via Residual Learning," 2024.

[18] E. Hu *et al.*, "Lora: Low-Rank Adaptation of Large Language Models," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, pp. 1–26, 2022.

[19] J. Van Herck *et al.*, "Assessment of fine-tuned large language models for real-world chemistry and material science applications," *Chem. Sci.*, pp. 670–684, 2024, doi: 10.1039/d4sc04401k.

[20] L. Luo *et al.*, "Taiyi: A bilingual fine-tuned large language model for diverse biomedical tasks," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 9, pp. 1865–1874, 2024, doi: 10.1093/jamia/ocae037.

[21] W. Zhang *et al.*, "Fine-tuning large language models for chemical text mining," *Chem. Sci.*, vol. 15, no. 27, pp. 10600–10611, 2024, doi: 10.1039/d4sc00924j.

[22] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. R. Wen, "Evaluating Object Hallucination in Large Vision-Language Models," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, no. Table 1, pp. 292–305, 2023, doi: 10.18653/v1/2023.emnlp-main.20.

[23] I. O. Gallegos *et al.*, "Bias and Fairness in Large Language Models: A Survey," *Comput. Linguist.*, vol. 50, no. 3, pp. 1097–1179, 2024, doi: 10.1162/coli_a_00524.

[24] H. Zhao *et al.*, "Explainability for Large Language Models: A Survey," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, 2024, doi: 10.1145/3639372.

[25] Y. Chai, Q. Liu, S. Wang, Y. Sun, Q. Peng, and H. Wu, "On Training Data Influence of GPT Models," *EMNLP*

2024 - 2024 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 3126–3150, 2024, doi: 10.18653/v1/2024.emnlp-main.183.