Conceptualization-Augmented Technical Problem Solving in TVET Education Using Fine-Tuned LLM and Polytechnic and Community Colleges Digital Dataset

Guan Chengg Wong[1], Suryani Ilias[1*], Muhammad Syahmi Akmal Zuklipli[1], Ihsan Yassin[2]

[1]Department of Electrical Engineering, Politeknik Sultan Salahuddin Abdul Aziz Shah, Shah Alam, Malaysia.
[2]Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia.

guanchengg.wong@gmail.com; suryani@psa.edu.my; syhmkml@gmail.com; ihsan.yassin@gmail.com

***Abstract* — Large Language Models (LLMs) have shown strong potential for chatbot and question-answering applications, but their deployment in Technical and Vocational Education and Training (TVET) remains challenging due to domain mismatch, unreliable responses, and the need for accessible development workflows. This study aims to develop a domain-specific LLM-based chatbot for technical problem solving in TVET education by integrating fine-tuning with a PolyCC digital dataset in a cloud-based, competition-driven environment. The proposed innovation applies a structured methodology consisting of dataset preparation in instruction–response format, selection of a pre-trained LLaMA 3.2 (3B Instruct) model, parameter-efficient fine-tuning, automated evaluation using win rate, and iterative dataset optimization. The experimental findings show that model performance depends more strongly on dataset quality than dataset size alone. From 46 experimental runs, the best-performing configuration achieved a win rate of 62%, demonstrating that carefully curated and refined question–answer data can significantly improve response relevance and domain adaptation. The study also shows that suitable epoch selection is important to balance learning and generalization, while excessive training may reduce performance. In terms of impact, this work contributes a practical framework for building domain-specific LLM chatbots in TVET, while also supporting no-code, gamified, and cloud-based AI learning for educators and students. In conclusion, the proposed approach successfully demonstrates that fine-tuned LLMs, supported by high-quality PolyCC digital datasets, can enhance technical problem solving and provide a scalable pathway for AI integration in TVET education.**

***Keywords* — Large Language Models (LLMs), Fine-Tuning, TVET Education, Domain-Specific Chatbot, PolyCC Digital Dataset**

I. INTRODUCTION

Large Language Models (LLMs) have become a core technology in generative artificial intelligence, enabling applications such as chatbots and question-answering systems. One of the key approaches to improving factual accuracy and contextual relevance is Retrieval-Augmented Generation (RAG), which integrates external knowledge sources with language models [1]. In addition, efficient adaptation methods such as Low-Rank Adaptation (LoRA) allow large models to be fine-tuned with reduced computational cost [2], while alignment techniques using human feedback improve the ability of models to follow instructions effectively [3]. Further advancements show that large-scale instruction fine-tuning enhances generalization across tasks [4], and approaches such as Self-Instruct reduce dependence on manually labelled data [5]. The availability of open models such as LLaMA 2 has also increased accessibility for developing domain-specific chatbot systems [6], while QLoRA enables efficient fine-tuning under limited hardware constraints [7].

Despite these advancements, deploying LLMs in real-world applications remains challenging. Conventional evaluation methods are often insufficient, as chatbot performance depends on multi-turn interaction, contextual understanding, and user-centered quality [8]. In addition, retrieval-based approaches have been widely studied to improve response grounding and provide up-to-date information [9, 10]. However, LLMs are still prone to hallucination, where models generate fluent but incorrect or misleading information, which remains a critical limitation in practical deployment [11]. These challenges highlight the need for integrating both efficient model adaptation and reliable grounding mechanisms, particularly for domain-specific chatbot applications that require accurate and trustworthy responses.

These issues are highly relevant in the context of the POLYCC Large Language Model (LLM) League 2025, which can be interpreted as a competition-based platform for developing domain-specific chatbot systems. The

*Corresponding Author

competition emphasizes building an LLM that functions as the "brain" of a chatbot within a cloud-based and gamified environment, while remaining accessible to participants without strong programming backgrounds. Educational research shows that no-code AI platforms can support meaningful learning when learners engage in authentic machine learning workflows [12], while competition-based learning enhances practical skills and motivation [13]. Studies on AI chatbots in education further highlight benefits such as personalized learning and engagement, alongside concerns related to reliability, ethics, and assessment fairness [14, 15]. Similarly, systematic reviews indicate the growing use of LLMs in education, while emphasizing challenges such as over-reliance and responsible usage [16, 17].

In this context, this paper presents the development of a domain-specific LLM chatbot that integrates efficient fine-tuning and retrieval-based grounding to improve performance and reliability. RAG-based systems have been shown to reduce hallucination and enable dynamic knowledge updates [18, 19]. Furthermore, the integration of gamification has been shown to enhance engagement and skill development in AI learning environments [20]. Therefore, this study demonstrates how a gamified, accessible competition framework can support the development of a practical and reliable LLM chatbot, while also contributing to meaningful skill development in generative AI.

II. RELATED WORK

Recent advances in Large Language Models (LLMs) show that effective chatbot performance depends on post-training adaptation rather than model scale alone. While larger models provide strong general capabilities, their effectiveness in specific applications relies on how well they are adapted to follow user intent and domain requirements. Instruction alignment approaches, such as supervised fine-tuning with human feedback, improve the ability of LLMs to generate relevant and accurate responses [3], while large-scale instruction tuning enhances generalization across diverse tasks [4]. Methods such as Self-Instruct further reduce reliance on manually labelled data by generating synthetic instructions [5]. In addition, open models such as LLaMA 2 improve accessibility for chatbot development [6], and parameter-efficient techniques such as LoRA and QLoRA enable model adaptation under limited computational resources [2, 7]. Collectively, these studies establish efficient fine-tuning as a practical approach for domain-specific chatbot development.

In parallel, grounding and reliability remain critical challenges in LLM deployment. Retrieval-Augmented Generation (RAG) integrates external knowledge sources with language models to improve factual accuracy and contextual relevance [1]. Subsequent studies further define RAG as a framework for handling outdated knowledge and improving response verification [9, 10]. However, hallucination remains a key limitation, where models generate fluent but incorrect or misleading information [11]. In addition, chatbot evaluation must consider multi-turn interaction and user-centered quality rather than relying only on static benchmarks [8]. These findings suggest that combining fine-tuning with retrieval-based grounding is essential for reliable chatbot performance.

From an educational perspective, accessibility and engagement are important factors in AI learning environments. No-code AI platforms support meaningful learning experiences even for users without strong programming backgrounds [12], while competition-based learning improves practical skills and motivation [13]. Studies on AI chatbots in education highlight benefits such as personalized learning and engagement, alongside concerns related to reliability, ethics, and assessment fairness [14, 15]. Similarly, systematic reviews indicate the increasing use of LLMs in education while emphasizing challenges such as over-reliance and responsible usage [16, 17]. More recent works on educational RAG systems show that retrieval-based approaches improve factual quality and enable dynamic knowledge updates [18, 19]. In addition, gamification has been shown to enhance engagement and skill development in AI learning environments [20].

Overall, existing studies can be grouped into three main aspects: efficient LLM adaptation, grounding and reliability, and educational implementation. However, these aspects are often studied independently. As summarized in Table 1, there is limited work that integrates all three perspectives within a single applied setting, particularly in a gamified and competition-based environment. Table 1 shows the key literature categories and their relevance to this study.

Based on this, the present study is positioned at the intersection of these aspects by developing a domain-specific LLM chatbot within a gamified, cloud-based competition environment, integrating efficient fine-tuning, retrieval-based grounding, and accessible learning design.

TABLE 1

SUMMARY OF RELATED WORK FOR LLM-BASED CHATBOT DEVELOPMENT IN EDUCATIONAL COMPETITION SETTINGS

Aspect	Key References	Focus	Relevance
Efficient LLM adaptation	[2], [3], [4], [5], [6], [7]	Instruction tuning, open models, parameter-efficient fine-tuning	Enables domain-specific chatbot development with limited resources
Grounding and reliability	[1], [8], [9], [10], [11]	RAG, hallucination mitigation, conversational evaluation	Improves factual accuracy and response reliability
Educational implementation	[12], [13], [14], [15], [16], [17]	No-code learning, competition-based learning, AI in education	Supports accessibility and skill development
Educational RAG and gamification	[18], [19], [20]	RAG for education, gamified AI learning	Enhances learning engagement and system usefulness

III. METHODOLOGY

This study adopts a fine-tuning-based approach to adapt a pre-trained Large Language Model (LLM) for domain-specific understanding, particularly in the context of TVET-related knowledge. The overall methodology follows a structured pipeline consisting of dataset preparation, model selection, fine-tuning, evaluation, and leaderboard submission, as illustrated in Figure 1.

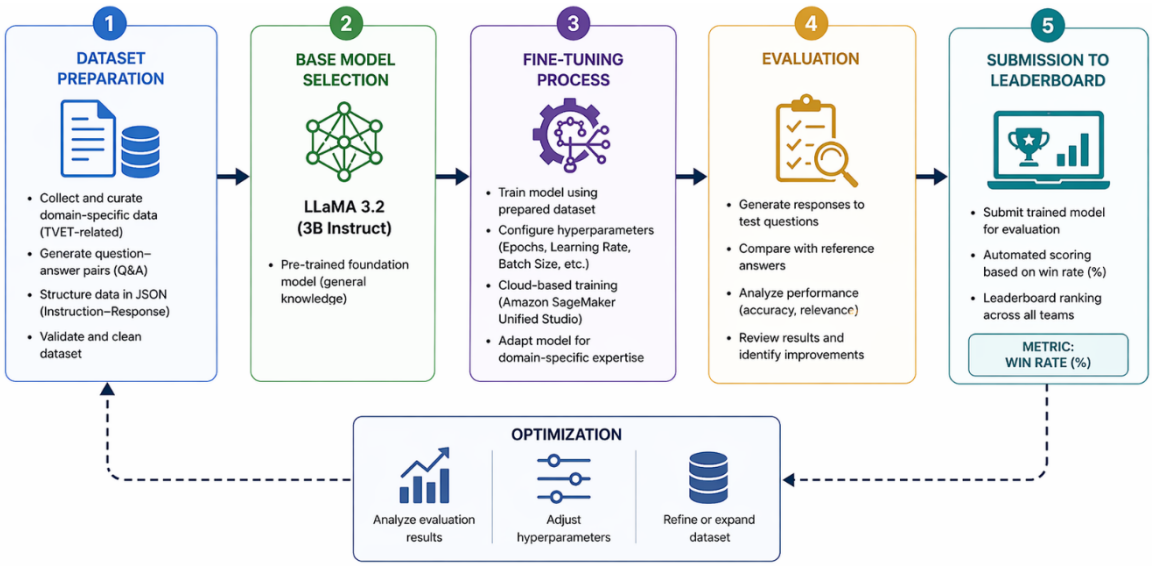

Fig. 1 Overall methodology pipeline for LLM fine-tuning and evaluation

The process begins with dataset preparation, where domain-specific question–answer pairs are constructed and organized in JSON format. Each entry contains an instruction (input query) and a corresponding response (expected output), enabling the model to learn structured instruction-following behaviour. The dataset is curated from relevant domain materials and may also be augmented using synthetic data generation techniques, which have been shown to improve instruction diversity and reduce reliance on manual annotation [5]. Particular emphasis is placed on data quality, consistency, and coverage, as these factors directly influence the effectiveness of the fine-tuning process and overall model performance.

Next, a pre-trained foundation model, specifically LLaMA 3.2 (3B Instruct), is selected as the base model due to its efficiency and suitability for lightweight fine-tuning tasks. Foundation models are typically trained on large-scale general corpora and therefore require post-training adaptation to perform effectively in specialized domains [6]. Instruction-based models further benefit from alignment techniques, which enhance their ability to follow user queries accurately and produce coherent, context-aware responses [3, 4].

The fine-tuning stage is conducted using a no-code cloud-based platform, improving accessibility for users without extensive programming experience. During this stage, the model is trained on the prepared dataset to enhance its ability to generate accurate and domain-relevant responses. Rather than performing full-parameter retraining, the approach leverages parameter-efficient fine-tuning strategies to reduce computational cost while maintaining strong performance [2, 7]. Key training parameters, including the number of epochs and learning rate, are carefully adjusted to ensure stable convergence and effective knowledge adaptation.

Following training, the model undergoes evaluation, where its generated outputs are compared against reference answers. The evaluation focuses on response accuracy, relevance, and overall quality, reflecting recent research that emphasizes conversational and user-centred assessment rather than relying solely on static benchmarks [8]. This stage provides a structured assessment of how well the model performs in domain-specific tasks.

Subsequently, the trained model is submitted to a leaderboard-based evaluation system, where its performance is measured using an automated scoring mechanism based on win rate. This metric represents the percentage of responses that outperform reference outputs and serves as a standardized measure for comparing models. The leaderboard enables objective ranking across different submissions and supports fair performance benchmarking.

Finally, an optimization stage is performed, where evaluation outcomes from the leaderboard are analyzed to identify areas for improvement. Based on these insights, refinements are applied primarily at the dataset level, including expanding data coverage, improving answer quality, and enhancing instruction diversity. Adjustments to hyperparameters may also be considered. This targeted optimization process ensures continuous improvement while maintaining a structured and efficient pipeline flow [13].

Overall, this methodology emphasizes the integration of high-quality data preparation, efficient fine-tuning strategies, and systematic evaluation with leaderboard-based benchmarking, forming a robust framework for developing a domain-specific LLM chatbot system.

IV. EXPERIMENTAL SETUP

The experimental setup was designed to support the efficient fine-tuning, evaluation, and submission of the proposed LLM within a cloud-based environment. The setup emphasizes scalability, accessibility, and practical usability, which are important in the context of competition-based AI development. In addition to enabling repeated experimentation, the setup allows users to focus on dataset preparation and model optimization without being constrained by local hardware limitations.

A. Hardware and Environment

All experiments were conducted using a cloud computing platform, eliminating the need for high-performance local hardware. Model training and evaluation were executed on remote GPU instances that provide sufficient computational resources for handling LLM fine-tuning workloads. The entire workflow was accessed through a web-based interface, allowing users to perform dataset upload, model configuration, training, and evaluation using only a standard computer with internet connectivity. This environment is particularly suitable for educational and competition-based settings because it lowers the technical barrier to entry and supports wider participation. Such an approach is consistent with recent educational practices that promote no-code AI platforms to broaden participation in machine learning development [12].

B. Model Configuration

The base model used in this study was LLaMA 3.2 (3B Instruct), a lightweight yet capable instruction-tuned model suitable for domain adaptation tasks. The model was selected because it offers a good balance between computational efficiency and response capability, making it appropriate for repeated fine-tuning experiments within a limited training budget. As an instruction-based model, it is designed to respond to structured user prompts in a more coherent and task-oriented manner. Fine-tuning was applied to adapt this general-purpose model to the target domain without modifying the full architecture, thereby improving efficiency and reducing training cost. This setup is aligned with recent LLM adaptation strategies, where instruction tuning and lightweight fine-tuning methods have been shown to enhance performance in specialized domains [4, 6].

C. Dataset Preparation

The dataset consisted of structured question–answer pairs tailored to the selected domain. Each entry followed a standardized JSON format consisting of: (1) Instruction as the input query and (2) Response as the expected output. This format was chosen because it directly supports instruction-based learning and enables the model to learn the mapping between domain-specific queries and appropriate responses.

To improve dataset diversity and coverage, AI-assisted data generation techniques were used alongside manual curation. This combination allowed the construction of a broader range of examples while maintaining relevance to the target domain. Prior work has shown that synthetic instruction data can improve model generalization when

it is carefully filtered and structured [5]. Therefore, particular attention was given to maintaining data consistency, clarity, and domain relevance, as these factors directly influence the effectiveness of the fine-tuning process. In practice, the dataset was also refined iteratively by revising unclear questions, removing unsuitable entries, and improving answer structure.

D. Hyperparameter Tuning

Key hyperparameters were systematically adjusted during training to optimize model performance. The two main parameters considered in this study were:

(i) Epochs: Determines how many times the model processes the dataset. Increasing epochs can improve learning but may lead to overfitting if excessive.
(ii) Learning Rate: Controls the rate of model updates. Lower values ensure stable convergence, while higher values accelerate training but may introduce instability.

Multiple training experiments were conducted by varying these parameters incrementally. Rather than relying on a single configuration, the study adopted an iterative strategy in which parameter values were adjusted based on observed performance. This reflects common practice in LLM fine-tuning, where performance improvements are often achieved through controlled experimentation rather than one-pass optimization [7]. The use of repeated trials also helped identify suitable parameter ranges for balancing learning efficiency and generalization.

E. Evaluation Method

Model performance was evaluated using an automated response comparison system. In this process, the fine-tuned model generated answers to a predefined set of queries, and these outputs were compared against reference answers using a larger evaluation model. This evaluation strategy was selected because it provides a more practical measure of response usefulness and quality compared to simple keyword-based accuracy metrics.

The primary evaluation metric was winning rate, defined as the percentage of generated responses that outperform reference outputs. This metric captures how well the model performs in relative terms and is well suited to comparative evaluation in chatbot-based tasks. It also reflects recent evaluation frameworks that emphasize conversational and comparative assessment of LLM performance rather than relying only on traditional accuracy measures [8]. As a result, the evaluation method supports more realistic benchmarking of domain-specific chatbot performance.

F. Training and Submission Procedure

Each model was trained within a limited computational budget defined by allocated training hours on the cloud platform. After a training job was completed, the resulting model followed three main steps: (1) the trained model was registered in the system, (2) it was submitted to the leaderboard, and (3) it was evaluated automatically using the predefined scoring mechanism.

Multiple submissions were allowed throughout the experimentation process, with only the best-performing model considered for final ranking. This setup encouraged iterative refinement, since users could improve the dataset or adjust the training parameters and resubmit updated models. Such a workflow is consistent with competition-based learning environments, where repeated experimentation and performance comparison play an important role in developing practical AI skills [13]. In this study, the submission process also served as an important feedback mechanism for guiding further optimization.

Overall, the experimental setup integrates cloud-based infrastructure, efficient model configuration, structured datasets, systematic hyperparameter tuning, and robust automated evaluation. This combination provides a practical and scalable framework for developing domain-specific LLM chatbot systems in an accessible and competition-oriented environment.

V. RESULTS AND ANALYSIS

This section presents the experimental results obtained from multiple fine-tuning configurations, followed by an in-depth analysis of model performance based on dataset characteristics and hyperparameter settings. A total of 46 experimental runs were conducted, and model performance was evaluated using the win rate (%) metric. Only the most representative and significant configurations are summarized to highlight key performance trends observed throughout the experiments.

A. Quantitative Results

The quantitative results highlight the relationship between dataset size, training parameters, and model performance. The selected experimental configurations and their corresponding win rates are summarized in Table 2.

TABLE 2

SELECTED EXPERIMENTAL CONFIGURATIONS AND PERFORMANCE SUMMARY

Exp ID	Dataset Size	Epochs	Learning Rate	Win Rate (%)	Observation
E1	300	1	0.0001	8	Very low performance, limited dataset depth
E2	950	1	0.0001	16	Underfitting due to insufficient epochs
E3	950	10	0.0001	24	Improved learning
E4	950	12	0.0001	26	Slight improvement
E5	950	14	0.0001	28	Peak before overfitting
E6	950	20	0.0001	14	Overfitting occurred
E7	220	10	0.0001	46	Significant improvement after dataset refinement
E8	240	10	0.0001	56	Further improvement
E9	255	10	0.0001	48	Performance drops due to unsuitable data
E10	259	10	0.0001	58	Improved after dataset revision
E11	276	10	0.0001	60	Best performance achieved
E12	440	10	0.0001	54	Larger dataset but reduced quality
E13	269	10	0.0001	58	Stable performance
E14	276	10	0.0001	62	Consistent best performance

Table 2 shows that model performance does not improve proportionally with an increase in dataset size. In the early experiments (E1–E6), even with a large dataset of 950 samples, the model achieved relatively low win rates ranging from 16% to 28%. This indicates that increasing dataset quantity alone is insufficient to guarantee better performance, particularly when the dataset lacks consistency or contains redundant patterns.

In contrast, experiments involving smaller but refined datasets (E7–E14) demonstrate a significant improvement in performance, achieving win rates between 46% and 62%. Notably, E14 achieved the highest win rate of 62%, despite using a dataset size of only 276 samples. This clearly suggests that dataset quality, coherence, and relevance play a more critical role than dataset size in fine-tuning performance.

Furthermore, fluctuations in performance, for example, E8 to E9, where performance drops from 56% to 48% indicate that adding new data without proper validation may introduce noise or inconsistencies, ultimately degrading model performance. This highlights the importance of careful dataset curation rather than indiscriminate data expansion.

B. Hyperparameter Impact

To further understand the impact of training parameters, an ablation study was conducted focusing on the number of training epochs while keeping other variables constant. The results are summarized in Table 3.

TABLE 3

EFFECT OF EPOCHS ON MODEL PERFORMANCE (DATASET SIZE = 950)

Epochs	Win Rate (%)	Observation
1	16	Underfitting
10	24	Improved learning
12	26	Slight improvement
14	28	Optimal performance
20	14	Overfitting

Table 3 shows that the number of training epochs has a significant influence on model performance. At lower epoch values, the model exhibits clear underfitting, as it has not sufficiently learned from the dataset. As the number of epochs increases, the model performance improves gradually, reaching an optimal point at approximately 10 to 14 epochs.

However, further increasing the epochs to 20 results in a sharp decline in performance, indicating overfitting, where the model becomes too specialized to the training data and loses its ability to generalize. This trend confirms that there exists an optimal training range, beyond which additional training negatively impacts performance.

Despite the importance of hyperparameter tuning, comparison with Table 2 reveals that hyperparameters alone do not determine overall performance. Even with optimal epoch settings, models trained on poorly structured datasets still produce lower win rates. Therefore, dataset quality remains the dominant factor influencing performance.

C. Dataset Characteristics and Their Impact

The dataset used in this study is structured in an instruction–response (QnA) format, designed to reflect realistic chatbot interactions. Each entry consists of a clearly defined instruction paired with a detailed and contextually relevant response.

The dataset exhibits several important characteristics:

(i) Explicit and well-defined instructions, reducing ambiguity in user queries
(ii) Comprehensive and structured responses, incorporating explanations, comparisons, and logical reasoning
(iii) Consistency in writing style and terminology, supporting stable model learning
(iv) Diversity of question types, including descriptive, analytical, and scenario-based queries
(v) Iterative refinement process, involving removal of low-quality entries and revision of ambiguous questions

The impact of these characteristics is strongly reflected in the experimental outcomes. For instance, the performance drop observed in E9 (255 dataset, 48%) compared to E8 (240 dataset, 56%) indicates that newly added data may not always align with the existing dataset distribution. Similarly, increasing dataset size to 440 (E12) resulted in a lower win rate compared to smaller datasets, reinforcing that larger datasets do not necessarily lead to better performance if data quality is compromised.

On the other hand, experiments that involved refinement and restructuring existing QnA pairs consistently showed performance improvements, eventually achieving the highest win rate of 62%. This demonstrates that well-curated datasets with coherent structure and meaningful variation are essential for effective fine-tuning.

D. Qualitative Results and Error Analysis

In addition to quantitative evaluation, qualitative analysis was conducted to assess the nature and quality of model-generated responses. In the early stages of experimentation, the model frequently produced responses that were generic, shallow, or partially aligned with the query, particularly when trained on limited or inconsistent datasets.

As the dataset was progressively refined, the model demonstrated notable improvements in:
(i) Contextual understanding of queries
(ii) Logical organization of responses
(iii) Relevance to domain-specific topics

However, several limitations were identified during testing:
(i) The model remains highly sensitive to dataset composition, where minor inconsistencies can significantly affect performance
(ii) Overfitting occurs at higher epoch values, reducing generalization capability
(iii) The model struggles with complex or multi-step reasoning tasks when similar patterns are underrepresented in the dataset

Error cases were commonly observed when queries deviated from training patterns or required deeper reasoning beyond the dataset's coverage. These findings suggest that further improvements may require enhanced dataset diversity and more advanced training strategies.

Overall, the results clearly demonstrate that dataset quality is the most influential factor in determining model performance, followed by appropriate hyperparameter tuning. The best-performing configuration achieved a win rate of 62%, indicating successful domain adaptation within the given constraints.

The key findings of this study can be summarized as follows:

(i) High-quality and well-structured QnA datasets significantly improve model performance
(ii) Optimal epoch selection (10–14) is necessary to balance learning and generalization
(iii) Dataset refinement is more effective than increasing dataset size

These findings confirm that a systematic approach to dataset design and optimization is essential for developing a robust and high-performing domain-specific LLM.

VI. CONCLUSION

This study presented the development of a domain-specific Large Language Model (LLM) chatbot within the context of the POLYCC LLM League 2025, integrating efficient fine-tuning and a structured evaluation pipeline. The proposed approach addressed key challenges in LLM deployment, particularly the need for domain adaptation, reliable response generation, and accessible development workflows. By leveraging a pre-trained instruction-based model and applying parameter-efficient fine-tuning, the system successfully adapted general knowledge into domain-specific understanding using a structured question–answer dataset.

The experimental results demonstrate that model performance is highly dependent on dataset quality rather than dataset size alone. The findings show that carefully curated and well-structured datasets significantly improve response accuracy and relevance, as reflected by the highest achieved win rate of 62%. In addition, the ablation analysis confirms that appropriate hyperparameter tuning, particularly the selection of training epochs, is essential to balance learning and generalization. While increasing epochs initially improves performance, excessive training leads to overfitting, reducing the model's ability to generalize to unseen queries. Overall, the study highlights that a combination of high-quality dataset design, controlled training configuration, and systematic evaluation is critical for achieving a robust domain-specific LLM.

Beyond technical performance, this work also demonstrates the effectiveness of a gamified, cloud-based competition environment in supporting practical AI development. The use of a no-code platform enables broader participation, allowing users with limited programming experience to engage in meaningful machine learning workflows, including data preparation, model training, and evaluation. This aligns with educational objectives by promoting hands-on learning and skill development in generative AI.

For future work, several enhancements can be considered. First, the integration of Retrieval-Augmented Generation (RAG) can be explored to improve factual grounding and reduce hallucination by incorporating external knowledge sources. Second, further refinement of dataset design, including more complex reasoning tasks and multi-turn conversational data, may enhance the model's ability to handle advanced queries. These directions have the potential to further improve both the reliability and applicability of domain-specific LLM chatbot systems.

ACKNOWLEDGEMENT

The authors would like to express sincere appreciation to the organizers of the PolyCC Large Language Models League 2025, organized by the B*ahagian Instruksional dan Pembelajaran Digital, Jabatan Pendidikan Politeknik dan Kolej Komuniti*, in collaboration with *Politeknik dan Kolej Komuniti (PolyCC) Zon Selangor dan Wilayah Persekutuan Kuala Lumpur*, for providing a valuable platform that supports innovation and practical development in generative artificial intelligence. The author(s) also gratefully acknowledge Politeknik Sultan Salahuddin Abdul Aziz Shah for its continuous support, provision of academic resources, and encouragement throughout the development and completion of this work.

DECLARATION OF GENERATIVE AI USAGE

During the preparation of this work, the authors used generative AI tools (ChatGPT) to refine language clarity, improve sentence structure, and enhance overall presentation. The ideas, methodology, analysis, and results presented in this paper are original and developed by the authors. The authors declare that all generated content has been carefully reviewed, validated, and edited as necessary, and that they take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
[2] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
[3] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744, 2022.
[4] H. W. Chung et al., "Scaling Instruction-Finetuned Language Models," Journal of Machine Learning Research, vol. 25, pp. 1–53, 2024.

[5] Y. Wang et al., “Self-Instruct: Aligning Language Models with Self-Generated Instructions,” in Proc. 61st Annual Meeting of the Association for Computational Linguistics, 2023.
[6] H. Touvron et al., “Llama 2: Open Foundation and Fine-Tuned Chat Models,” arXiv preprint arXiv:2307.09288, 2023.
[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized LLMs,” in Advances in Neural Information Processing Systems, 2023.
[8] L. Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” in Advances in Neural Information Processing Systems, vol. 36, Datasets and Benchmarks Track, 2023.
[9] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint arXiv:2312.10997, 2023.
[10] Y. Huang and J. X. Huang, “A Survey on Retrieval-Augmented Text Generation for Large Language Models,” arXiv preprint arXiv:2404.10981, 2024.
[11] L. Huang et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” arXiv preprint arXiv:2311.05232, 2023.
[12] L. Sundberg and J. Holmström, “Teaching Tip: Using No-Code AI to Teach Machine Learning in Higher Education,” Journal of Information Systems Education, vol. 35, no. 1, pp. 56–66, 2024.
[13] H.-T. Chang and C.-Y. Lin, “Applying Competition-Based Learning to Stimulate Students’ Practical and Competitive AI Ability in a Machine Learning Curriculum,” IEEE Transactions on Education, pp. 1–10, 2024, doi: 10.1109/TE.2024.3350535.
[14] L. Labadze, M. Grigolia, and L. Machaidze, “Role of AI Chatbots in Education: Systematic Literature Review,” International Journal of Educational Technology in Higher Education, vol. 20, art. no. 56, 2023, doi: 10.1186/s41239-023-00426-1.
[15] C. McGrath, A. Farazouli, and T. Cerratto-Pargman, “Generative AI Chatbots in Higher Education: A Review of an Emerging Research Area,” Higher Education, vol. 89, pp. 1533–1549, 2025, doi: 10.1007/s10734-024-01288-w.
[16] B. Dong, J. Bai, T. Xu, and Y. Zhou, “Large Language Models in Education: A Systematic Review,” in Proc. 6th International Conference on Computer Science and Technologies in Education, 2024, pp. 131–134, doi: 10.1109/CSTE62025.2024.00031.
[17] Y. Shi, K. Yu, Y. Dong, and F. Chen, “Large Language Models in Education: A Systematic Review of Empirical Applications, Benefits, and Challenges,” Computers and Education: Artificial Intelligence, p. 100529, 2025, doi: 10.1016/j.caeai.2025.100529.
[18] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, “Retrieval-Augmented Generation for Educational Application: A Systematic Survey,” Computers and Education: Artificial Intelligence, vol. 8, p. 100417, 2025, doi: 10.1016/j.caeai.2025.100417.
[19] J. Swacha and M. Gracel, “Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications,” Applied Sciences, vol. 15, no. 8, art. no. 4234, 2025, doi: 10.3390/app15084234.
[20] A. Marengo, A. Pagano, B. D. Lund, and V. Santamato, “Research AI: Integrating AI and Gamification in Higher Education for E-Learning Optimization and Soft Skills Assessment Through a Cross-Study Synthesis,” Frontiers in Computer Science, vol. 7, art. no. 1587040, 2025, doi: 10.3389/fcomp.2025.158704.