Optimizing Contextual Chatbots via Synthetic Data Fine-Tuning and Selective Grid Search: Explorations from an LLM Competition

Mohd Suhairi Md Suhamin[1,2*], Norsuzila Shafie[1], Wan Siti Rodziah Mohd Nasir[1]

[1]Politeknik Kota Bharu, Malaysia
[2]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia

suhairisuhaimin@pkb.edu.my; norsuzila@pkb.edu.my; rodziah@pkb.edu.my

***Abstract*— This paper explores the optimization of contextual chatbots through a strategic combination of synthetic data fine-tuning and selective hyperparameter tuning. Developed within the competition of the POLYCC LLM League 2025, the paper addresses the challenge of enhancing lower-tier Large Language Models (LLMs) under stringent architectural and computational constraints. The proposed methodology integrates a three-layer pipeline: (1) multi-model synthetic data generation, (2) Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning, and (3) a multi-stage competition evaluation. Moving beyond exhaustive search methods, a selective grid search strategy was implemented to identify the optimal balance between performance gains and training overhead. Utilizing AWS SageMaker, the model was rigorously evaluated through an automated qualification phase followed by a multi-dimensional final assessment involving AI metrics, expert validation, and audience sentiment. Our findings reveal that data quality and targeted LoRA parameter selection (*r* and α) yield superior performance compared to simply increasing dataset volume. The resulting model demonstrated significantly improved contextual grounding and generalization, ultimately securing the highest overall ranking (1st Place) among the competition finalists. These results provide a strategic roadmap for deploying high-performance LLMs in resource-constrained and applied domain-specific environments.**

***Keywords*— Large Language Model, Fine-tuning, Synthetic Data, Selective Grid Search, AI Competition**

I. INTRODUCTION

The rapid development of Large Language Models (LLMs) has significantly enhanced the capabilities of conversational agents across a wide range of domains, including question answering, dialogue systems, and decision support applications [1, 2]. Despite these advances, maintaining contextual consistency, domain relevance, and factual accuracy, while minimizing hallucinations, remains a persistent challenge, particularly when deploying smaller or resource-constrained models [3, 4]. These challenges become more pronounced in applied environments such as education and training, where responses must be both accurate and contextually grounded to ensure trustworthiness and instructional value [5, 6].

This paper describes the approach undertaken during the POLYCC LLM League 2025, a competition-based evaluation framework, to develop a contextual chatbot capable of delivering high-quality, domain-specific responses. Competition-based benchmarking and leaderboard-driven evaluations have become increasingly common for assessing LLM performance under standardized yet constrained conditions [7, 8]. In this competition, participants were provided with a baseline lower-tier LLM and were required to improve its performance through dataset construction and fine-tuning while adhering to strict computational and architectural constraints.

The primary task involved constructing a specialized synthetic dataset and fine-tuning the provided model to improve contextual understanding and answer relevance. Recent studies have shown that synthetic instruction–response datasets can significantly enhance LLM performance, particularly in low-resource or domain-specific scenarios [9, 10]. The core contribution of this paper lies in the application of a multi-model synthetic data generation strategy, combined with a selective (greedy) grid search approach for hyperparameter optimization. Rather than employing exhaustive search, selective exploration of promising hyperparameter configurations has been shown to be more efficient and practical in applied fine-tuning settings [11, 12]. The effectiveness of this strategy was validated through both the qualification leaderboard and the final live demonstration stage of the competition.

*Corresponding Author

II. Related Work

Recent studies emphasize the importance of Parameter-Efficient Fine-Tuning (PEFT) techniques, particularly Low-Rank Adaptation (LoRA), in adapting LLMs for domain-specific tasks under limited computational resources [13]. LoRA enables selective updating of model parameters while preserving the majority of pre-trained weights, making it well suited for applied settings such as competitions and institutional deployments.

In parallel, the use of synthetic data has gained prominence as a practical solution to the scarcity of high-quality, domain-specific conversational datasets. Prior research demonstrates that carefully curated synthetic instruction–response pairs can significantly improve model performance and contextual robustness [9, 14]. However, poorly curated or unverified synthetic data may amplify hallucinations and negatively affect model generalization [3, 15, 16].

Competition-based environments, such as LLM leaderboards and live evaluation challenges, provide a practical testbed for evaluating fine-tuning strategies under real-world constraints [7, 17, 18]. Unlike static benchmarks, these environments impose limitations on model size, training budget, and inference behavior, encouraging resource-efficient optimization strategies. Despite their increasing adoption, competition-driven LLM experimentation remains relatively underrepresented in formal academic literature, particularly in terms of detailed methodological reporting and reproducibility [19]. This paper contributes to this gap by documenting a complete applied workflow that integrates multi-model synthetic data generation, LoRA-based fine-tuning, and selective hyperparameter search within a competitive LLM setting.

III. Methodology

The overall methodology of the contextual chatbot followed a structured three-phase pipeline, as illustrated in Fig. 1. This iterative framework encompasses synthetic data construction, model optimization through PEFT, and a multi-stage competition evaluation.

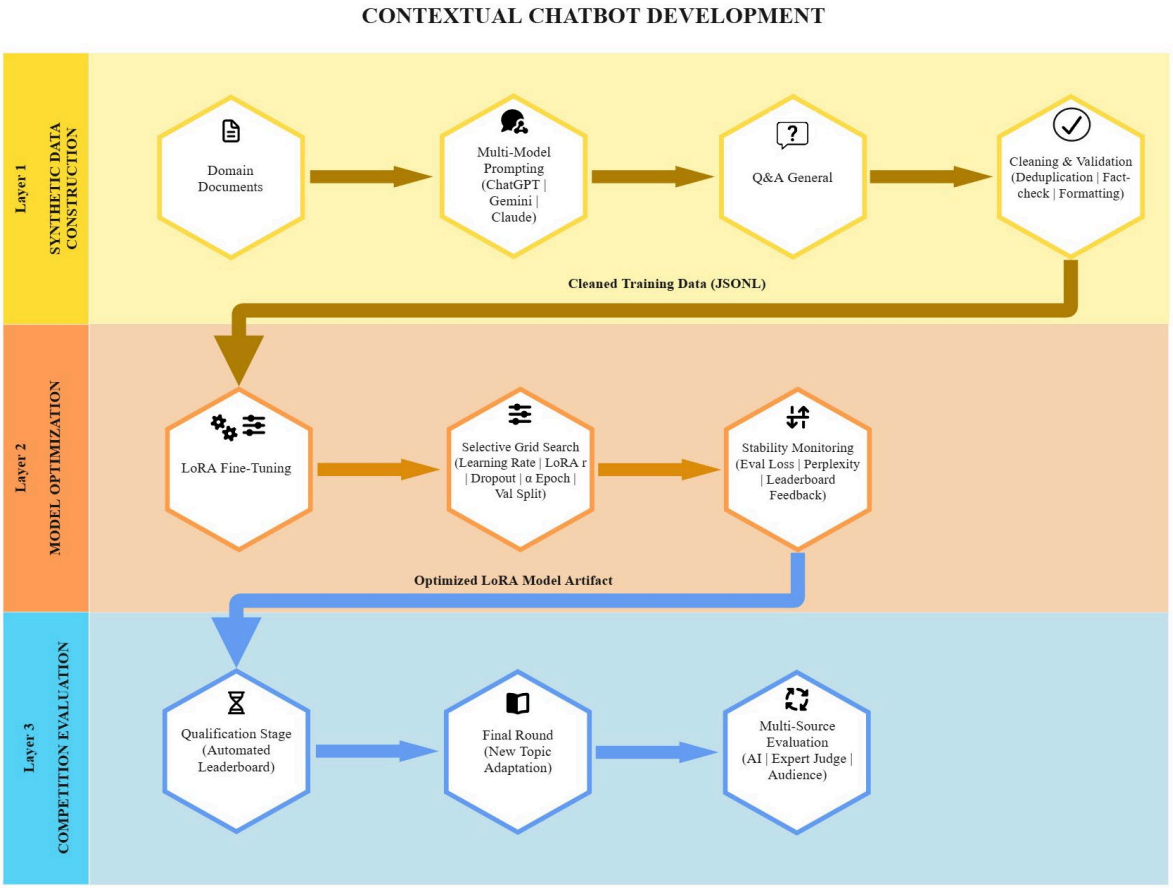

Fig. 1 The three-phase pipeline for optimizing contextual chatbots via synthetic data and LoRA fine-tuning

A. Synthetic Dataset Construction (Layer 1)

The initial phase focused on generating a high-quality, domain-specific dataset. The process began with the curation of domain documents related to the POLYCC and TVET Digital context. These documents served as

the ground truth for a multi-model prompting strategy, utilizing diverse generative models including ChatGPT, Gemini, and Claude. This multi-model approach was adopted to increase linguistic diversity and minimize model-specific biases.

Beyond document-based extraction, general questions and answers (Q&A) augmentation was performed to enhance conversational breadth [20]. This included generating explanatory prompts (e.g., "explain" and "elaborate") to improve natural language flow. The raw data then underwent a rigorous cleaning and validation stage, involving:

i. Deduplication: removing redundant entries to prevent over-fitting.
ii. Fact-checking: manual and automated inspection for factual consistency.
iii. Formatting: ensuring structural integrity for training.

The final output of this layer was a cleaned training dataset stored in (JavaScript Object Notation Line) JSONL format, with dataset sizes ranging from 1,200 to over 8,000 entries across experimental iterations.

B. Model Optimization (Layer 2)

In the second phase, the model underwent LoRA fine-tuning within the AWS SageMaker environment. To achieve peak performance without exhaustive resource consumption, a selective grid search (greedy search) was implemented. This strategy focused on optimizing critical hyperparameters, including:

i. Learning Rate and Epochs.
ii. LoRA rank (r) and scaling factor (α).
iii. Dropout rates and validation split ratios.

To ensure the model's reliability, stability monitoring was conducted throughout the training process. This involved tracking evaluation loss and perplexity to prevent catastrophic forgetting. Additionally, preliminary leaderboard feedback was used to adjust hyperparameters dynamically, resulting in an optimized LoRA model artifact.

C. Competition Evaluation (Layer 3)

The final phase assessed the model's performance through a tiered evaluation framework. The process moved through three distinct stages:

i. Qualification stage: an automated leaderboard assessment to filter and rank initial model iterations based on objective benchmarks.
ii. Final round: testing the model's new topic adaptation capabilities, evaluating how well the optimized artifact could generalize to unseen contextual data.
iii. Multi-source evaluation: a holistic final assessment involving three distinct perspectives:
    - AI evaluation: automated metrics for consistency and relevance.
    - Expert judges: subject matter experts reviewing for technical accuracy and domain alignment.
    - Audience feedback: assessing the human-likeness and usability of the chatbot responses.

IV. EXPERIMENTAL SETUP

All experiments were conducted using AWS SageMaker with Hugging Face training utilities, following the official workflow and constraints specified by the POLYCC LLM League 2025 organizers. The competition infrastructure enforced standardized training, evaluation, and submission procedures to ensure fairness and reproducibility across participants.

A. Base Model and Training Environment

A provided lower-tier LLM, as mandated by the competition rules, served as the base model for all experiments. Participants were required to initiate fine-tuning jobs through Amazon SageMaker JumpStart, with no permission to modify the underlying model architecture or pre-training weights. This constraint ensured that all performance improvements resulted solely from dataset construction and fine-tuning strategies, rather than architectural enhancements.

All training jobs were executed in a cloud-based environment, where datasets were uploaded in structured JSONL format and referenced via Amazon S3. Each fine-tuned model artifact was automatically generated and managed by the SageMaker platform before being registered to the LLM League system for evaluation.

B. Hyperparameters and Selective Grid Search

Initial experiments focused on coarse parameter sweeps, followed by incremental refinement based on observed performance trends. This selective grid search strategy was adopted to balance experimental coverage with computational efficiency, avoiding exhaustive exploration of the hyperparameter space.

The tuning process was guided by established PEFT principles, particularly the use of LoRA to enable stable model adaptation under constrained resources [13]. In addition, practical engineering considerations for learning rate selection, LoRA rank scaling, dropout control, and epoch limits were informed by prior empirical studies on hyperparameter optimization [11], as well as practitioner-oriented guidance derived from extensive fine-tuning experience [12, 21].

Rather than pursuing maximum parameter combinations, the selective grid search emphasized training stability, convergence behavior, and leaderboard feedback, which are critical factors in competition-based evaluation settings. The specific parameter configurations for each experimental group during the qualification stage are summarized in Table 1.

TABLE I
SELECTIVE GRID SEARCH GROUPS FOR LORA FINE-TUNING (QUALIFICATION STAGE)

Group	Search focus	Epoch	Learning rate	LoRA r	LoRA α	Dropout	Val split
G1	Baseline LoRA (small adapter)	1–5	1.00E-04	8–16	32–64	0.01	0.1
G2	Increased adapter capacity (mid)	1–2	1.00E-04	32	128–256	0.01	0.1
G3	Main working region (dominant)	1–3	1.00E-04 to 1.50E-04	64	512–1024	0.01–0.02	0.1–0.15
G4	Large adapter stress-test	1–2	1.00E-04 to 1.50E-04	128–512	256–1024	0.01–0.05	0.1–0.2
G5	Regularization sensitivity	1–2	8.00E-05 to 1.50E-04	64–128	128–1024	0.03–0.05	0.1–0.2
G6	Unstable / non-performing extremes	5–10	2.00E-04 to 5.00E-04	8–1024	32–1024	0.01–0.05	0.1–0.2

C. Task Setting, Dataset Preparation, and Evaluation Protocol

i. *Qualification Stage*

During the qualification stage, participants were required to fine-tune their models using provided and self-constructed datasets, followed by submission to an automated leaderboard. Evaluation was conducted using a hidden test set, with scores reflecting the correctness and relevance of model responses.

Dataset preparation during this stage focused on iterative refinement of synthetic instruction–response pairs, generated using the proposed multi-model prompting pipeline. Multiple dataset variants were explored to improve contextual coverage, response clarity, and domain relevance.

Leaderboard feedback served as the primary signal for model improvement, guiding subsequent dataset revisions and hyperparameter adjustments.

ii. Final Round

Following the qualification phase, only the top six teams were selected to advance to the final round. In this stage, participants were provided with a new topic that had not appeared in the qualification datasets, designed to assess generalization capability rather than memorization.

Using the same multi-model synthetic data generation pipeline described earlier, a new domain-specific dataset was rapidly constructed. Dataset preparation emphasized concise yet comprehensive responses, ensuring that the model could generalize learned contextual patterns to unseen content.

As in the qualification stage, datasets were uploaded in JSONL format and fine-tuning was performed using the same LoRA-based selective grid search strategy, with hyperparameter values adapted from the best-performing configurations identified earlier. The refined configurations tailored for the final round are detailed in Table 2.

Table II
SELECTIVE GRID SEARCH CONFIGURATION FOR FINAL STAGE FINE-TUNING

Group	Epoch	Learning rate	LoRA r	LoRA α	Dropout	Validation split
F1	3–5	1.50E-04	192	896	0.025	0.1
F2	7–9	1.00E-04 to 1.50E-04	192	896	0.01–0.025	0.1
F3	9–10	1.50E-04	192	896	0.025	0.1
F4	9–10	1.50E-04	256	1024	0.025	0.1
F5	9	1.00E-04	64	128	0.015	0.1
F6	10	1.50E-04	192	896	0.015–0.025	0.1

iii. Final Round Evaluation

Unlike the qualification round, which relied primarily on automated leaderboard scoring, the final round employed a multi-source evaluation framework. Submitted models were evaluated based on their responses to a predefined set of prompts, with scores assigned from three independent sources:

a. Automated system evaluation (AI) – assessing response correctness, coherence, and relevance.
b. Expert judge evaluation (human) – conducted by subject-matter experts focusing on factual accuracy, contextual grounding, and clarity.
c. Audience evaluation (human vote) – reflecting perceived usefulness and response quality from a general audience.

Each evaluation component contributed to the final score, ensuring a balanced assessment that combined objective system metrics with human judgment. This evaluation design reflects real-world deployment scenarios, where chatbot effectiveness depends not only on automated measures but also on expert validation and user acceptance.

V. RESULTS AND DISCUSSION

The effectiveness of the proposed contextual chatbot development approach was evaluated through two sequential competition stages: the qualification stage and the final stage. These stages were designed to assess both incremental performance improvement under controlled conditions and generalization capability under unseen task settings.

A. Qualification Stage Results

During the qualification stage, the fine-tuned models were evaluated using a hidden automated test set provided by the competition platform. Iterative dataset refinement and selective hyperparameter tuning led to progressive improvements in leaderboard performance, confirming the effectiveness of the proposed synthetic data and fine-tuning strategy. The performance metrics and observed trends for the various hyperparameter groups (as defined in Table 1) are summarized in Table 3.

As summarized in Table 3, configurations derived from the dominant hyperparameter region (Group G3 in Table I) consistently achieved higher scores than baseline or extreme configurations. In particular, models trained with moderate LoRA rank ($r \approx 64$), higher LoRA scaling (α between 512–1024), short training duration (1–2 epochs), and learning rates around 1.0E-4 to 1.5E-4 demonstrated the most stable and competitive performance.

The results further indicate that dataset quality and alignment played a more critical role than dataset size alone. Models trained with carefully curated synthetic instruction–response pairs consistently outperformed those trained on larger but less refined datasets. This observation supports prior findings that high-quality synthetic data, when properly validated, can significantly enhance domain-specific performance without increasing model complexity.

Table III
QUALIFICATION STAGE PERFORMANCE SUMMARY

Model group	Dataset size (approx.)	Key LoRA setting	Epoch	Best score (/50)	Observed trend
Baseline LoRA (G1)	~3,600–4,000	r=8–16, α=32–64	1–5	15–27	Limited capacity, unstable gains
Mid-capacity LoRA (G2)	~3,600–4,000	r=32, α=128–256	1–2	24	Improved stability, moderate gains
Dominant region (G3)	~3,800–4,100	r≈64, α=512–1024	1–3	31	Best balance of accuracy and stability
Large adapter (G4)	~3,800–4,100	r≥128, α≥256	1–2	31	Comparable peak, higher variance
Regularization-heavy (G5)	~3,800–4,100	dropout ≥0.03	1–2	28	Reduced overfitting, slower convergence
Extreme settings (G6)	varied	lr ≥2E-4 or epoch ≥5	5–10	0–low	Training instability or failure

Overall, the qualification results validated the selective grid search approach as an effective means of navigating the hyperparameter space under computational constraints, while leaderboard feedback served as a practical signal for iterative improvement.

B. Final Stage Results

The final stage required rapid adaptation to a previously unseen topic. This tested the robustness of the synthetic data pipeline and the transferability of the optimized hyperparameter settings. Table 4 presents the quantitative optimization progress and qualitative observations for the final model iterations.

Table IV
FINAL STAGE RESULTS AND EVALUATION

Final Model	Epoch	LoRA (r, α)	Eval loss ↓	Eval PPL ↓	Qualitative inference observation
Improved Context 1	3	(192, 896)	0.697	2.01	Correct facts, concise responses
Improved Context 2	7	(192, 896)	0.438	1.55	Mixed response length, stable facts
Improved Context 3	5	(192, 896)	0.409	1.51	Improved contextual alignment
Improved Context 4	9	(192, 896)	0.41	1.51	Longer answers, minor flaws
Improved Context 5	10	(192, 896)	0.3	1.35	Factually grounded, context-aware; requires prompting for elaboration

Quantitative performance indicators, including Evaluation Loss and Perplexity (PPL), showed clear improvement compared to earlier qualification-stage models. These improvements suggest that the selective grid search strategy enabled efficient transfer of learned contextual representations to previously unseen content.

In addition to automated metrics, qualitative inference evaluation revealed improvements in contextual relevance and factual grounding. The model demonstrated stronger alignment with domain-specific concepts

and reduced hallucination tendencies. However, some limitations remained, particularly in response length consistency, where certain answers were concise and required explicit prompting for elaboration. This highlights the sensitivity of response verbosity to both training data distribution and inference-time decoding parameters.

Despite these limitations, the final-stage results demonstrate that the proposed approach effectively balances generalization performance and training efficiency, even when operating within strict competition constraints.

VI. CONCLUSION

This paper demonstrated that a multi-model synthetic data generation strategy, combined with selective hyperparameter tuning using Low-Rank Adaptation (LoRA), can substantially enhance the performance of smaller, resource-constrained large language models in applied and competition-based settings. The experimental results across both qualification and final evaluation stages show that careful dataset curation and targeted parameter optimization have a greater impact on contextual response quality and stability than exhaustive hyperparameter exploration.

The proposed approach proved effective not only in improving automated evaluation metrics, but also in producing responses that were favorably assessed by human experts and end users under a multi-source evaluation framework. These findings suggest that strategic fine-tuning strategies, when aligned with contextual data quality, are well suited for real-world chatbot deployment in constrained institutional environments.

Future work will focus on cross-lingual extensions, response-length and verbosity control, and further refinement of LoRA rank scaling to improve generalization, efficiency, and usability in broader educational and administrative domains.

ACKNOWLEDGEMENT

We would like to express our sincere appreciation to Politeknik dan Kolej Komuniti (POLYCC) and Bahagian Instruksional dan Pembelajaran Digital (BIPD) for organizing the POLYCC LLM League 2025. Special thanks are extended to AWS Malaysia for providing the cloud infrastructure and hosting support that enabled the experimentation and evaluation processes. We also gratefully acknowledge the support and encouragement from management and colleagues at Politeknik Kota Bharu (PKB), whose cooperation and commitment played an important role in the successful completion of this work.

DECLARATION OF GENERATIVE AI USAGE

During the preparation of this work, the authors used ChatGPT, Gemini, and Claude to generate synthetic training data and assist in proofreading the manuscript. The authors declare that they reviewed and edited the final output and take full responsibility for the content.

REFERENCES

[1] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems,* vol. 33, pp. 1877-1901, 2020.

[2] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971,* 2023.

[3] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM computing surveys,* vol. 55, no. 12, pp. 1-38, 2023.

[4] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 1906-1919.

[5] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and individual differences,* vol. 103, p. 102274, 2023.

[6] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education–where are the educators?," *International journal of educational technology in higher education,* vol. 16, no. 1, p. 39, 2019.

[7] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology,* vol. 15, no. 3, pp. 1-45, 2024.

[8] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300,* 2020.

[9] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," in *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, 2023, pp. 13484-13508.

[10] S. Zhang *et al.*, "Instruction tuning for large language models: A survey," *ACM Computing Surveys,* vol. 58, no. 7, pp. 1-36, 2026.

[11] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of machine learning research,* vol. 13, no. 2, 2012.

[12] S. Raschka, "Practical tips for finetuning llms using lora," ed, 2023.
[13] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *Iclr,* vol. 1, no. 2, p. 3, 2022.
[14] S. Gunasekar *et al.*, "Textbooks are all you need," *arXiv preprint arXiv:2306.11644,* 2023.
[15] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths, "Embers of autoregression: Understanding large language models through the problem they are trained to solve," *arXiv preprint arXiv:2309.13638,* 2023.
[16] M. S. M. Suhaimin, M. H. A. Hijazi, W. S. R. M. Nasir, A. Wibowo, and E. G. Moung, "The Challenge of Generalization: Preserving Sarcasm Detection in a Multitask Model Across Different Linguistic Contexts," in *2025 International Conference on Advances in Machine Intelligence, and Cybersecurity Technologies (AMICT)*, 2025: IEEE, pp. 49-53.
[17] P. Liang *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110,* 2022.
[18] M. S. M. Suhaimin, M. H. A. Hijazi, and E. G. Moung, "MSSThreD: Multitask Social Media Sentiment Analysis with Sarcasm for Public Security Threat Detection," in *2024 11th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2024: IEEE, pp. 25-30.
[19] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big??," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610-623.
[20] M. S. M. Suhaimin, M. H. A. Hijazi, E. G. Moung, and M. A. M. Hamza, "Data Augmentation Approach for Language Identification in Imbalanced Bilingual Code-Mixed Social Media Datasets," in *2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 2023: IEEE, pp. 257-261.
[21] S. Raschka, *Build a large language model (from scratch)*. Simon and Schuster, 2024.