

Indonesian generative chatbot model for student services using GPT

Shania Priccilia, Abba Suganda Girsang

Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 17, 2023

Revised Feb 24, 2024

Accepted Feb 28, 2024

Keywords:

Academic information

Generative chatbot

GPT-2

Indonesian chatbot

Student services

Text generation

ABSTRACT

The accessibility of academic information greatly impacts the satisfaction and loyalty of university students. However, limited university resources often hinder students from conveniently accessing information services. To address this challenge, this research proposes the digitization of the question-answering process between students and student service staff through the implementation of generative chatbot. A generative chatbot can provide students with human-like responses to academic inquiries at their convenience. This research developed generative chatbot using pre-trained GPT-2 architecture in three different sizes, specifically designed for addressing practicum-related questions in a private university in Indonesia. The experiment utilized 1288 question-answer pairs in Indonesian and demonstrated the best performance with a BLEU score of 0.753, signifying good performance accuracy in generating text despite dataset limitations.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Shania Priccilia

Computer Science Department, BINUS Graduate Program – Master of Computer Science

Bina Nusantara University

Jakarta 11480, Indonesia

Email: shania.priccilia@binus.ac.id

1. INTRODUCTION

The quality of services offered by universities significantly impacts student satisfaction and loyalty [1]. University services encompass complaint handling and academic-related information services. However, due to university operating hours and limited human resources, students often face limitations in accessing services, particularly information-related ones, at their convenience. To address this issue, the deployment of chatbot technology presents a viable solution to enhance the quality of university services by offering prompt responses to student inquiries. Proven through research conducted on higher education institutions in Indonesia, technological innovation demonstrates a significantly positive correlation with student loyalty and satisfaction [2].

Based on the model employed, chatbots can be categorized into three main types: rule-based models, retrieval-based models, and generative models. Rule-based and retrieval-based models are knowledge-based chatbots which formulating responses derived from the predefined responses, whereas generative chatbot models rely on their proficiency in natural language understanding (NLU) and natural language generation (NLG) [3]. In Indonesia, several universities have implemented chatbots to address student inquiries concerning administrative and operational aspects of academic-related. However, the chatbot systems utilized by universities are predominantly knowledge-based models, resulting in constrained responses and a lack of flexibility. The latest trend involves the development of generative chatbots which can engage with users in a more flexible manner. Developing generative chatbots necessitates a model proficient in text generation, encompassing both NLG and NLU capabilities [4].

Transformer [5] have emerged as the dominant architecture in natural language processing (NLP) tasks such as text generation and machine translation, outperforming convolutional models and recurrent neural networks (RNN) models [6]. The research conducted by Masum *et al.* [7] found that a Bengali language chatbot using the Transformer model achieved a BLEU score of 85.00, surpassing the Seq2Seq and Bidirectional-RNN models, which achieved BLEU scores of 23.50 and 17.23, respectively. Generative pre-trained transformer (GPT) [8]–[10] is a transformer-based model that has emerged as the state of the art in language modeling tasks such as text generation and question-answering. GPT [8] was first introduced by OpenAI in 2018 and demonstrated superior performance compared to previous ensemble transformer-based models in various NLP tasks. GPT notably outperformed other models in question-answering tasks on datasets like RACE [11] and Story Cloze [12] with improvements of 5.7% and 8.9%, respectively. In 2023, GPT has advanced into a superior language model, offering large language models (LLMs) like GPT-3.5 and GPT-4 [13], [14]. GPT-4 [14] was unveiled in March 2023 by OpenAI and achieved state of the art in language modeling on seven academic benchmarks with highest accuracy of 92.0% on the GSM-8K benchmark. This performance surpasses the previous model, which attained an accuracy of 87.3%. GPT-3.5 and GPT-4 models are accessible through a paid subscription offered by OpenAI, enabling users to access and fine-tune these models for their specific applications. For detailed pricing information, please refer to the OpenAI pricing page at <https://openai.com/pricing>.

The objective of this research is to employ the GPT model for the purpose of text generation, specifically in developing a generative chatbot tailored to the domain of university services. Due to resource constraints, we use GPT-2 model which is accessible via open access through Hugging Face. GPT-2 [9] was introduced in 2019 as an improved version of GPT. GPT-2 comprises 1,542 million parameters, 48 layers, and a dmodel of 1600. These figures are larger than GPT [8], which has 117 million parameters, 12 layers, and a dmodel of 768, enabling GPT-2 to handle longer sequences. In comparison with the efficient transformer model such as Reformer [15], GPT achieved a better perplexity score of 34.16, surpassing the Reformer's score of 35.17 for text generation [16].

2. METHOD

This research comprised four stages: dataset collection, preprocessing, experiment, and evaluation. Dataset collection involved online gathering of question-answer pairs from text conversations between students and student service staff within a single semester related to three categories: laboratory operational procedures, group protests, and score protests in a private university. A total of 644 pairs of question-answers were obtained from a total of 250 conversations. The conversation texts were gathered in Indonesian language and contain several specific English terms.

To enhance the dataset size, this research applies data augmentation using easy data augmentation (EDA) technique [17] and BERT [18] pipeline. The EDA [17] technique involves four operations: random insertion, random deletion, random swap, and synonym replacement. In this research, the EDA technique employed includes synonym replacement and random insertion using BERT fill-mask pipeline [19], resulting in larger dataset size. A detailed summary of the augmented dataset's properties is presented in Table 1.

Table 1. The dataset properties after augmentation via synonym replacement and random insertion utilizing BERT pipeline

Dataset properties	Number
Conversations	500
Question-answer pairs	1288
Number of tokens	68258
Average number of tokens	53
Smallest number of tokens	14
Largest number of tokens	241

In the preprocessing phase, text cleansing and normalization were performed. An example of preprocessed question-answer pair text is shown in Figure 1. The preprocessed text was divided into the training, validation, and test sets comprising 1030, 129, 129 question-answer pairs, respectively. Subsequently, the train and validation sets were utilized for the training and validation processes, employing batch sizes of 8 and 4 per device, respectively.

In pursuit of the objective to develop a generative chatbot model, we utilized pre-trained GPT-2 models with various sizes, as detailed in Table 2 [20]–[22]. Based on research conducted by Radford *et al.* [9], small size is equivalent to the first version of GPT [8], the medium size is equivalent to BERT [18], and the large size represents a customized version of GPT-2 with 28 layers (customized from 36 layers [22]).

Question	Answer
Orang 1: Untuk latihan case quiz 1 kenapa tidak ada isi nya ya? Orang 2: Untuk sesi yang terjadwal untuk quiz memang tidak disediakan casenya.	

Figure 1. Example of question-answer pair. Each row in datasets includes a question (indicated by the token "Orang 1") followed by the corresponding answer (indicated by the token "Orang 2").

The pre-trained GPT-2 model used for this research was trained on 522 MB Indonesian Wikipedia text. The encoding and tokenization process for each model utilized the pre-trained GPT-2 tokenizer, a tokenizer and detokenizer based on the byte pair encoding (BPE) algorithm. The tokenizer was trained with a vocabulary size of 52,000 subwords. Figure 2 illustrates the general architecture of GPT-2. To mitigate overfitting, all models underwent L2 regularization and dropout techniques during training [23].

During the evaluation stages, model performance was assessed using BLEU scores and perplexity metrics. The BLEU score dominantly employed in NLP evaluations [24]. It measures text accuracy based on n-gram concept by comparing response generated by model to the reference text [24], [25]. Perplexity (PPL) complements the BLEU score by providing a probabilistic measurement of the tokens generated by the model [26]. It can be employed to quantify the model's proficiency when predicting words with new data.

Table 2. The variations of GPT-2 models used

Model	d_{model}	Number of layers
GPT-2 Small [20]	768	12
GPT-2 Medium [21]	1024	24
GPT-2 Large [22]	1280	28

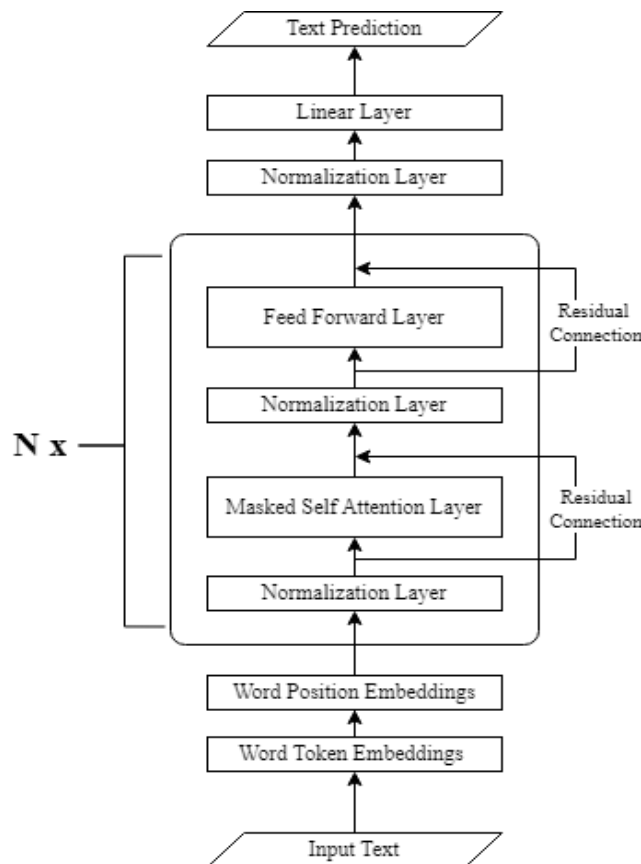


Figure 2. Architecture of GPT-2 where N represents number of layers

3. RESULTS AND DISCUSSION

The research was conducted by fine-tuning three pre-trained GPT-2 models for 55 epochs on a GPU T4 with 12.7 GB of RAM. The results of the training and validation processes of the GPT-2 models are presented in Figure 3. Figure 3(a) presents the training process, the training losses of all models consistently decreased. The validation losses of GPT-2 Small also followed a similar downward trend. However, for GPT-2 Medium and Large, validation losses exhibited an increasing trend during certain epochs, indicating potential signs of overfitting. Consequently, at the end of the training, we saved each model at the state where the validation loss was lowest as presented in Figure 3(b).

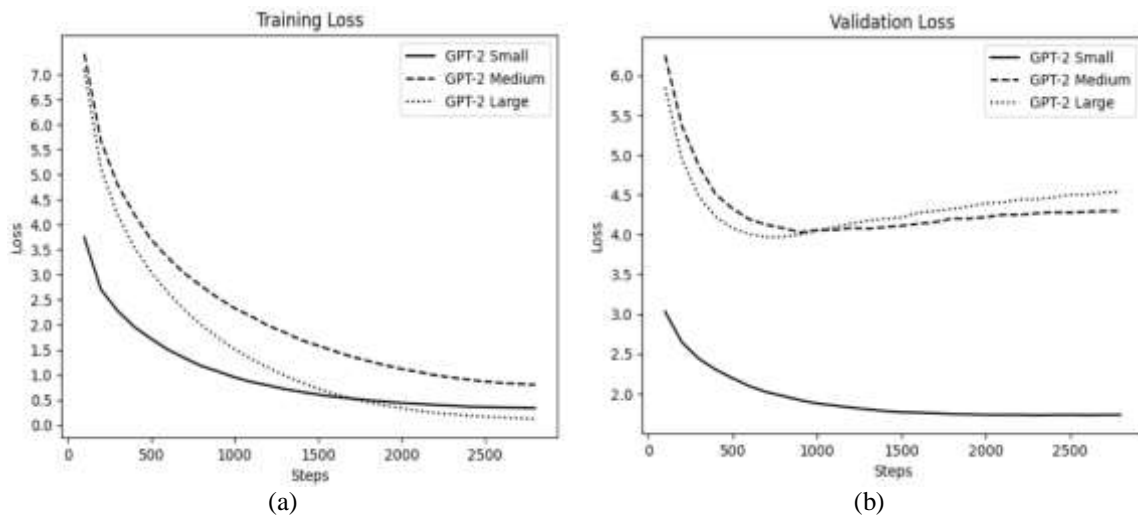


Figure 3. Training losses (a) and validation losses and (b) of GPT-2 models

The best models were evaluated using BLEU score and perplexity metrics, as presented in Table 3. According to the experimental results, the GPT-2 Small model achieved the highest BLEU score, indicating its proficiency in generating word sequences aligned with reference texts. However, all three models exhibited high perplexity, indicating difficulty in predicting sequences when encountering new data. This implies that while the models can replicate text well, it struggles with a deeper understanding of the context.

The GPT-2 Medium and Large models surprisingly performed poor BLEU scores and perplexity. This contradicts the common assumption that greater number of layers in a model enhances its ability to understand and predict new sequences [27]. Greater model sizes allow for learning intricate representations but increase the risk of overfitting and demand greater computational resources [28].

Table 3. The performance metrics of GPT-2 models on private dataset to measure quantitative performance

Model	Training time (minutes)	BLEU score	Perplexity	Loss
GPT-2 Small	~15	0.753	974.13	1.720
GPT-2 Medium	~40	0.565	1387.51	4.401
GPT-2 Large	~70	0.570	1420.23	4.577

To explore the factors influencing the model's suboptimal performance, we conducted additional experiment using the publicly available MultiWOZ dataset [29]. The MultiWOZ dataset was translated into Indonesian and subsequently trimmed to 5,000 conversations for GPT-2 Small and Medium models, 3,500 conversations for the GPT-2 Large model. As presented in Table 4, the three models demonstrated improved performance when trained on larger dataset. Moreover, the performance of these models surpassed that of other models trained on the English version of the MultiWOZ dataset. Hence, it can be inferred that GPT models exhibit strong performance in text generation tasks. The reduced performance observed in our study can be attributed to the complexity of the models and the limitations of the dataset. Our dataset was restricted to specific academic topics such as laboratory operational procedures, group protests, and score protests.

To qualitatively assess the performance of the models, predictions were made using a commonly asked question by students, presented in Figure 4. Measurements were conducted at $temperature = 0.5$, $top-p = 0.5$ and $top-k = 10$, balancing creative output within contextual constraints. Lower values for $top-p$ and $top-k$

k may yield more focused or deterministic responses. Temperature regulates token randomness by scaling logits and may also lead to more deterministic responses at lower settings [30]. Experimental results show that among three Indonesian language questions, the GPT-2 Small model consistently delivered the most appropriate responses, achieving 3/3 best responses compared to the other two models. Despite the low BLEU score, the context and semantic inference remain valid and considerable across all models.

Table 4. The performance metrics of GPT-2 models on MultiWOZ [29] dataset

Model	Training time (minutes)	BLEU score	Perplexity	Loss
GPT-2 Small	~40	0.717	872.78	1.577
GPT-2 Medium	~100	0.666	1209.13	1.934
GPT-2 Large	~120	0.683	1197.39	1.966
LSTM [29]	-	0.189	-	-
T5-Base [31]	-	0.177	-	-
T5-Small [31]	-	0.179	-	-

<p>Question: Orang 1: Kak, untuk nilai yang dinolkan dengan keterangan 'score deduction', berarti seluruh nilai lab menjadi 0 (dinolkan) benar, kak? Orang 2: Reference: Iya, jika score deduction disebabkan oleh dugaan mencontek dalam project, maka seluruh nilai praktikum, termasuk final exam dan assignment, akan di-nolkan. GPT-2 Small: Iya, jika nilai dinolkan adalah 0 (dinolkan), maka seluruh nilai praktikum menjadi 0 (dinolkan). BLEU Score: 2.13e-78 (best response) GPT-2 Medium: Iya, benar. BLEU Score: 1.14e-235 GPT-2 Large: silakan hubungi LINE OA kami di linktr.ee/OfficialAccountSLC ya. BLEU Score: 0 (worst response)</p>
<p>Question: Orang 1: saat ujian final praktek. secara onsite, file jawaban saya sudah terkirim, tetapi, ketika hasil nilainya keluar, nilai saya 0. Orang 2: Reference: silakan mengecek sebuah daftar nilai nilai dinolkan. apabila kamu mendapatkan sebuah nilai 0, tetapi nama kamu tidak tercantum dalam daftar mahasiswa yang mendapat nilai 0, kemungkinan masalahnya adalah sebagai berikut: kamu mungkin telah salah mengumpulkan jawaban, jawaban kamu mungkin tidak sesuai dengan tema soal, atau jawaban apa yang kamu kumpulkan mungkin sudah rusak (corrupt). GPT-2 Small: jika nilai kamu 0 namun nama kamu tidak terlampir pada list mahasiswa yang nilainya dinolkan maka ada kemungkinan sebagai berikut: kamu salah mengumpulkan jawaban, jawaban kamu beda tema dengan soal, jawaban kamu corrupt. BLEU Score: 0.11 (best response) GPT-2 Medium: untuk pengumuman nilai final ujian onsite, kamu dapat mengeceknya di binusmaya pada tanggal yang sudah ditentukan pada kalender praktikum. BLEU Score: 1.96e-232 (worst response) GPT-2 Large: kamu mengunggah file kamu mungkin kamu mungkin telah salah mengumpulkan jawaban kamu seharusnya tidak dapat mengajukan protes nilai kamu dapat mengajukan protes nilai kamu salah, jawaban kamu dapat dilihat pada binusmaya. BLEU Score: 0.077</p>
<p>Question: Orang 1: Apakah mungkin mengikuti kelas lab secara online jika saya sedang sakit? Orang 2: Reference: Halo, untuk kelas lab berlangsung sepenuhnya onsite. GPT-2 Small: Ujian berlangsung sepenuhnya onsite. BLEU Score: 4.07e-78 (best response) GPT-2 Medium: Untuk kelas lab online, kamu dapat mengikuti kelas melalui link berikut: https://linktr.ee/en/account.lab.live.bot. BLEU Score: 5.23e-155 GPT-2 Large: Silakan hubungi line oa kami bantu? BLEU Score: 0 (worst response)</p>

Figure 4. The responses of GPT-2 models are used to measure the model's qualitative performance

4. CONCLUSION

This research encountered limitations in terms of data collection time and dataset availability, restricted to one semester. The dataset specifically addresses issues within a private university in Indonesia, focusing on three categories: laboratory operational procedures, group protests, and score protests.





Based on the research findings, the GPT-2 models demonstrate the capability to generate reasonably good responses, achieving the highest BLEU score of 0.753 with the GPT-2 12-layer model as a generative chatbot. However, the model faces challenges understand context due to its complexity in the context of the

limited dataset. Therefore, in the recommendation for developing a chatbot to enhance student services, particularly within a specific domain with relatively templated and less varied responses, a generative chatbot is not strongly advised. Nonetheless, the consideration of using generative chatbots for student services may still be viable if the scope encompasses broader topics. Furthermore, to improve model accuracy and performance of generative chatbot model, the exploration of other data augmentation techniques is recommended.





REFERENCES

- [1] T. T. Borishade, O. O. Ogunnaike, O. Salau, B. D. Motilewa, and J. I. Dirisu, "Assessing the relationship among service quality, student satisfaction and loyalty: The NIGERIAN higher education experience," *Heliyon*, vol. 7, no. 7, Jul. 2021, doi: 10.1016/j.heliyon.2021.e07590.
- [2] E. Susilawati, I. Khaira, and I. Pratama, "Antecedents to student loyalty in Indonesian higher education institutions: The mediating role of technology innovation," *Educational Sciences: Theory and Practice*, vol. 21, no. 3, pp. 40–56, 2021.
- [3] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *Artificial Intelligence Applications and Innovations*, 2020, pp. 373–383. doi: 10.1007/978-3-030-49186-4_31.
- [4] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.
- [5] A. Vaswani *et al.*, "Attention is all you need," *Prepr. arXiv.1706.03762*, Jun. 2017, [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [6] T. Wolf *et al.*, "Hugging face's transformers: State-of-the-art natural language processing," *Prepr. arXiv.1910.03771*, Oct. 2019.
- [7] A. K. M. Masum, S. Abujar, S. Akter, N. J. Ria, and S. A. Hossain, "Transformer based bengali chatbot using general knowledge dataset," *Prepr. arXiv.2111.03937*, Nov. 2021.
- [8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Preprint. Work in progress*, pp. 1–12, 2018.
- [9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *Preprint*, 2019.
- [10] T. B. Brown *et al.*, "Language models are few-shot learners," *Prepr. arXiv.2005.14165*, May 2020.
- [11] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale Reading comprehension Dataset from examinations," *Prepr. arXiv.1704.04683*, Apr. 2017.
- [12] N. Mostafazadeh *et al.*, "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 839–849. doi: 10.18653/v1/N16-1098.
- [13] Y. Liu *et al.*, "Summary of ChatGPT-related research and perspective towards the future of large language models," *Meta-Radiology*, vol. 1, no. 2, Sep. 2023, doi: 10.1016/j.metrad.2023.100017.
- [14] J. Achiam *et al.*, "GPT-4 technical report," *Prepr. arXiv.2303.08774*, Mar. 2023.
- [15] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *Prepr. arXiv.2001.04451*, Jan. 2020.
- [16] N. Riaz, S. Latif, and R. Latif, "From transformers to reformers," in *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, May 2021, pp. 1–6. doi: 10.1109/ICoDT252288.2021.9441516.
- [17] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6381–6387. doi: 10.18653/v1/D19-1670.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Prepr. arXiv.1810.04805*, Oct. 2018.
- [19] "Cahya/bert-base-Indonesian-522M · Hugging face." <https://huggingface.co/cahya/bert-base-indonesian-522M> (accessed Nov. 10, 2023).
- [20] "Cahya/gpt2-small-Indonesian-522M · Hugging face." <https://huggingface.co/cahya/gpt2-small-indonesian-522M> (accessed Nov. 10, 2023).
- [21] "Cahya/gpt2-medium-Indonesian · Hugging face." <https://huggingface.co/cahya/gpt2-medium-indonesian> (accessed Nov. 10, 2023).
- [22] "Cahya/gpt2-large-Indonesian-522M · Hugging face." <https://huggingface.co/cahya/gpt2-large-indonesian-522M> (accessed Nov. 10, 2023).
- [23] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
- [24] K. Blagec, G. Dorffner, M. Moradi, S. Ott, and M. Samwald, "A global analysis of metrics used for measuring performance in natural language processing," *Prepr. arXiv.2204.11574*, Apr. 2022.
- [25] S. Dutta, "Evaluating a neural multi-turn chatbot using BLEU score," *Saarland University Department of Computational Linguistics*, 2019.
- [26] D. Colla, M. Delsanto, and D. P. Radicioni, "Semantic coherence dataset: Speech transcripts," *Data in Brief*, vol. 46, Feb. 2023, doi: 10.1016/j.dib.2022.108799.
- [27] J. Petty, S. van Steenkiste, I. Dasgupta, F. Sha, D. Garrette, and T. Linzen, "The impact of depth and width on transformer language model generalization," *Prepr. arXiv.2310.19956*, Oct. 2023.
- [28] S. Gholami and M. Omar, "Do generative large language models need billions of parameters?," *Prepr. arXiv.2309.06589*, Sep. 2023, [Online]. Available: <http://arxiv.org/abs/2309.06589>
- [29] T.-H. Budzianowski Pawełand Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026. doi: 10.18653/v1/D18-1547.
- [30] T. Weitzman, H. Pyo, and Jeon, "CloneBot: Personalized dialogue-response predictions," *Prepr. arXiv.2103.16750*, Mar. 2021.
- [31] Q. Cheng, L. Li, G. Quan, F. Gao, X. Mou, and X. Qiu, "Is MultiWOZ a solved task? An interactive TOD evaluation framework with user simulator," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 1248–1259. doi: 10.18653/v1/2022.findings-emnlp.90

BIOGRAPHIES OF AUTHORS

Shania Priccilia     is currently works at Bina Nusantara University as a Resource Management Officer at Laboratory Center Alam Sutera. She graduated from Bina Nusantara University, School of Computer Science and obtained her bachelor's degree in 2022. Previously, she worked as a Junior Laboratory Center, teaching programming courses for practicums from 2020-2021, until eventually transitioning to her current position. She is interested in software engineering and natural language processing. She can be contacted via email: shania.priccilia@binus.ac.id.



Abba Suganda Girsang     is currently lecturer at master information technology at Bina Nusantara University, Jakarta. He obtained Ph.D. degree in the Institute of Computer and Communication Engineering, Department of Electrical Engineering and National Cheng Kung University, Tainan, Taiwan, in 2014. He graduated bachelor from the Department of Electrical Engineering, Gadjah Mada University (UGM), Yogyakarta Indonesia, in 2000. He then continued his master's degree in the Department of Computer Science in the same university in 2006-2008. He was a staff consultant programmer in Bethesda Hospital, Yogyakarta, in 2001 and worked as a web developer from 2002 to 2003. He then joined the faculty of Department of Informatics Engineering in Janabadra University as a lecturer in 2003-2015. He also taught some subjects at some universities in 2006-2008. His research interests include swarm intelligence, combinatorial optimization, and decision support system. He can be contacted at email: agirsang@binus.edu.