Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 7, No.1, January 2025, pp. 0250115-01 ~ 0250115-13
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v7i1.1211

Comparative Performance of Transformer Models for Cultural Heritage in NLP Tasks

Tri Lathif Mardi Suryanto[12], Aji Prasetya Wibawa[3*], Hariyono[4], Andre Nafalski[5]

[13]Faculty of Engineering, Universitas Negeri Malang, Jl. Semarang No.5, Malang, East Java, 65145, Indonesia.

[3]Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, 60294, Indonesia

[4]Faculty of Social Sciences, Universitas Negeri Malang, Jl. Semarang No.5, Malang, East Java, 65145, Indonesia.

[5]UniSA Education Futures, University of South Australia SCT2-39 Mawson Lakes Campus, Adelaide, South Australia 5095, Australia

*aji.prasetya.ft@um.ac.id

Abstract. AI and Machine Learning are crucial in advancing technology, especially for processing large, complex datasets. The transformer model, a primary approach in natural language processing (NLP), enables applications like translation, text summarization, and question-answer (QA) systems. This study compares two popular transformer models, FlanT5 and mT5, which are widely used yet often struggle to capture the specific context of the reference text. Using a unique Goddess Durga QA dataset with specialized cultural knowledge about Indonesia, this research tests how effectively each model can handle culturally specific QA tasks. The study involved data preparation, initial model training, ROUGE metric evaluation (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum), and result analysis. Findings show that FlanT5 outperforms mT5 on multiple metrics, making it better at preserving cultural context. These results are impactful for NLP applications that rely on cultural insight, such as cultural preservation QA systems and context-based educational platforms.

Keywords: Artificial Intelligence, Machine Learning, NLP Cultural Knowledge, Transformer Models, QA Goddess Durga Dataset.

(Received 2024-07-25, Accepted 2025-01-07, Available Online by 2025-01-09)

1. Introduction

In the digital era, preserving cultural heritage has become a critical challenge due to rapid globalization and technological advancements. Artificial intelligence (AI) provides promising tools to address this issue by enabling the documentation, preservation, and dissemination of cultural knowledge. For instance, AI-based knowledge graphs and natural language processing (NLP) systems have been utilized to manage and interpret cultural resources effectively [1], [2] In Indonesia, which boasts over 1,300 ethnic groups and 700 languages, AI has the potential to safeguard its rich cultural diversity against extinction [3]. Despite its potential, research on leveraging AI models like FlanT5 and mT5 for cultural question-answering tasks in low-resource languages remains underexplored.

The rapid development of Artificial Intelligence (AI) opens a wide range of research opportunities, from the fields of business services [4], [5], [6], healthcare [7], [8], and education [9], [10], [11], [12], [13]. The growing application of AI has led to the development of Natural Language Processing (NLP), an AI learning generation technique that has revolutionized the way language data is handled and interpreted, for example for implementation in chatbots [14], [15], [16], [17] and support in digitization [18]. Transformer models, especially those in the T5 and FlanT5 families, have significantly advanced this field by excelling in tasks such as text summarisation, translation, and question answering.

While transformer-based models such as GPT and T5 have revolutionized NLP by achieving state-of-the-art performance across tasks [19], [20], their application in cultural contexts is relatively nascent. Many existing studies focus on resource-rich languages, leaving gaps for underrepresented languages like Indonesian. Additionally, few works evaluate the effectiveness of transformer models in extracting and reasoning about cultural knowledge, especially for multilingual and culturally specific datasets [21], [22]. Addressing this gap is essential for promoting cultural inclusivity in AI and for developing systems that can handle domain-specific knowledge in low-resource settings.

However, despite their powerful capabilities, these models often encounter challenges when dealing with culturally nuanced and context-specific content. Studies indicate that, while transformer models are effective on large datasets, their performance decreases when they are tasked with interpreting narratives and symbols deeply rooted in specific cultural or historical contexts [23], [24], [25]. This limitation is particularly noticeable in lesser-studied languages and mythologies, which are often underrepresented in training datasets and require more than surface-level understanding for accurate representation. As a result, there is a pressing need for research that assesses how these models perform in preserving and accurately conveying culturally rich narratives from underrepresented regions.

Current NLP models has undergone many developments and modifications to the model , such as mT5 and FlanT5, have demonstrated remarkable multilingual processing abilities and have shown potential for high adaptability across various language tasks [26], [27]. However, while these models can generalize effectively across languages, they often fall short in delivering the interpretive depth required for complex cultural stories, such as those mythology [28]. Prior studies, such as those [29], [30], highlight that while transformer models can be fine-tuned for domain-specific tasks, they frequently lack the granularity to capture symbolic meanings and context-sensitive content essential for cultural narratives. These limitations underscore the gap between advanced NLP capabilities and the nuanced understanding needed for tasks involving intricate cultural themes. These limitations highlight the gap between advanced NLP capabilities and the detailed understanding required for tasks centered on intricate cultural themes. In alignment with the focus of this study, we aim to address these challenges by evaluating how well transformers performance can preserve the symbolic and cultural integrity of mythological stories, specifically the narrative of cultural heritage in Indonesia.

The primary aim of this research is to examine the capability of transformer models in handling culturally significant and contextually specific narratives by focusing on the story of Durga from Indonesia mythology. This study seeks to evaluate these models' potential to interpret and convey cultural nuances accurately, especially in languages that are often less emphasized in large-scale NLP models. The novelty of this research lies in applying cutting-edge transformer models to a dataset deeply embedded in Southeast Asian culture and mythology, exploring their capacity to bridge cultural gaps in language interpretation. By focusing on Indonesian mythology, specifically Durga's myth, this research

also contributes to the understudied domain of AI-driven cultural preservation and serves as a testing ground for how well current state-of-the-art models handle narratives that require more than linguistic understanding calling for insights into symbolism and local cultural relevance [31].

The implications of this research extend beyond Indonesia, aligning with global efforts to integrate cultural awareness into AI systems. AI tools that are culturally sensitive not only support preservation but also enable cross-cultural understanding and dialogue [32]. For instance, question-answering systems grounded in local knowledge can provide accurate and context-aware insights, fostering appreciation for diverse heritages [33], as ethical considerations in AI gain prominence, ensuring cultural representation and equity in technology becomes paramount [34]. This study's evaluation of FlanT5 and mT5 contributes to the broader discourse on sustainable and inclusive AI development.

In terms of contributions, this research provides a thorough evaluation of NLP models' effectiveness in culturally specific content, contributing valuable insights for enhancing model adaptability and interpretative depth. By highlighting the limitations of current models in representing cultural nuances, this study underscores the need for future developments that prioritize context-sensitivity and cultural representation in AI. The findings will not only serve NLP researchers aiming to improve transformer adaptability in multicultural and multilingual contexts but also contribute to the broader goal of AI-enabled cultural preservation. In the long term, this research could pave the way for AI tools that better respect and promote the diversity of global narratives, supporting initiatives to keep lesser-known cultural stories alive in a rapidly digitizing world.

2. Methods

The process of generating a Q&A dataset through web scraping involves several key stages show in figure 2. Initially, the first stage is to plan and identify the data source by selecting the website to be scraped and confirming that the website allows for scraping activities, as for the website, YouTube, book, and article. After that, the data acquisition stage includes selecting a suitable web scraping tool and collecting data. Once data collection is complete, the next phase is data cleaning and transformation, which involves removing redundant entries and correcting inaccuracies, as well as compiling the data in a CSV compliant format, to facilitate its optimal utilization.

Figure 2. Research Methodology

2.1. Data Collection

The process of generating a Q&A dataset through web scraping involves several key stages. Initially, the planning and data source identification stage is crucial, where the website to be scraped is selected, and permissions for data extraction are verified to adhere to ethical standards (Chauncey & McKenna, 2023). In this study, was chosen as the data source in online and offline media, ensuring that scraping activities were permissible and aligned with responsible research practices.

The next stage is data acquisition, which entails selecting a suitable web scraping tool to efficiently collect the required data [35], [36], [37], [38]. Tools like webscraper.io facilitate the extraction of large volumes of data necessary for training NLP models. Following data collection, the data cleaning and transformation phase is essential to enhance data quality. This involves removing redundant entries, correcting inaccuracies, and formatting the data into a CSV-compliant format for optimal utilization in machine learning applications [39]. The successful execution of these stages—data acquisition, cleaning, and transformation—lays a robust foundation for developing and training NLP models, ensuring high-quality data that aligns with the specific requirements of cultural knowledge extraction and analysis.

The prepared dataset can be used to train the FlanT5 and mT5 transformer models as comparative data. This structured approach ensures that the resulting Q&A dataset is robust and suitable for tasks

that require nuanced understanding and contextual awareness. The selection of FlanT5 and mT5 models was motivated by their demonstrated state-of-the-art performance in multilingual and context-aware natural language processing (NLP) tasks. FlanT5 stands out for its fine-tuning capabilities, which enable the model to handle complex, culturally nuanced data, making it ideal for tasks requiring contextual depth [40], [41], [42]. Conversely, mT5 was selected to explore how a broadly multilingual model could adapt to specific cultural narratives. By comparing these two models, this study aims to evaluate their effectiveness in addressing culturally rich question-answering tasks in Indonesian, a low-resource language.

2.2. Preprocessing

The preprocessing stage outlines the architecture and workflow involved in fine-tuning Transformer-based models, specifically T5 (Text-to-Text Transfer Transformer), to adapt them for domain-specific tasks. This process begins with data cleaning and formatting, ensuring that the custom dataset is optimized for training. Once the data is prepared, it is used to adjust the model's weights and parameters, effectively personalizing the model to the nuances of the dataset. Figure 3 illustrates this fine-tuning process applied to both Flan-T5 and mT5 models.

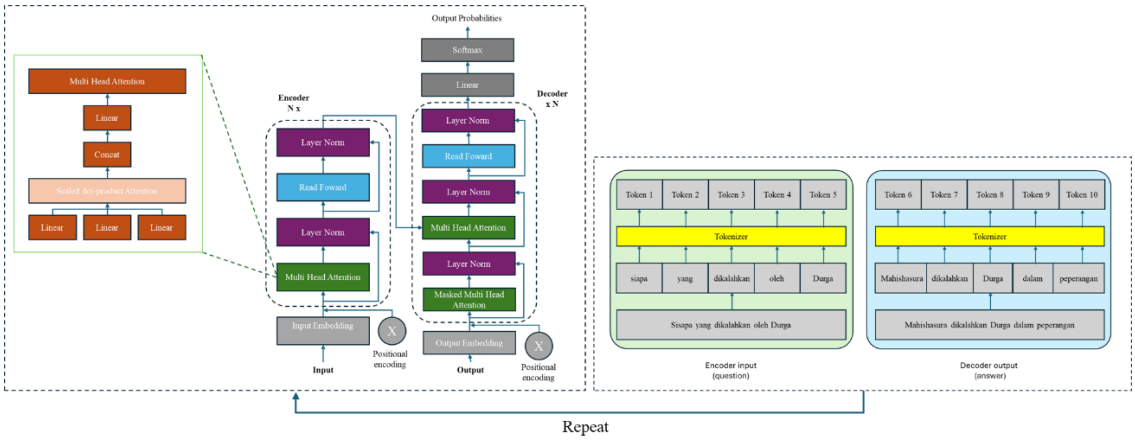

Figure 3. Architecture fine-tune comparative

By applying identical treatments to each model, the study ensures that any observed differences in performance can be attributed to the models' inherent capabilities rather than variations in the tuning process. This careful approach allows for an accurate comparison, shedding light on each model's strengths and weaknesses in handling specific cultural and contextual elements in the dataset. The result is a fine-tuned model that not only performs better on the custom dataset but also aligns more closely with the unique requirements of the target task.

The transformer model, a machine learning architecture, processes input data through a series of encoder and decoder blocks, incorporating input embeddings with positional encodings and transferring the processed representation to the decoder for generating predictions [20], [26]. This architecture utilizes attention mechanisms within both encoder and decoder layers, enabling the model to capture complex relationships within the data [43], [44]. Through fine-tuning, the model's weights are adjusted to adapt to specific tasks, allowing it to better align with the vocabulary and structure of the target dataset while preserving its foundational architecture [28], [45]. This fine-tuning capability enhances the model's predictive accuracy, as it tailors' outputs to the requirements of the applied dataset and task domain.

2.3. Matric Evaluation

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores measure the overlap of n-grams between the predicted summary and the reference summary

$$ROUGE - Lsum = \frac{1}{m}\sum_{i=1}^{m}\frac{LCS(S_i, R_i)}{Length\ (R_i)} \quad (1)$$

where:

- m is the number of sentences in the reference summary.
- S_i and R_i are the i-th sentences in the generated summary and reference summary, respectively
- $LCS\ (S_i,R_i)$ is the length of the Longest Common Subsequence between S_i and R_i.
- Length (R_i) is the total number of words in the reference sentence R_i

Collectively, these metrics allow for a robust evaluation of model performance in generating summaries that align with human references in vocabulary, phrase structure, and overall content coherence, making ROUGE a reliable choice in natural language processing evaluation. ROUGE-Lsum metric, averaging the longest common subsequence across sentences, provides a comprehensive sentence-level evaluation that is well-suited for tasks requiring multi-sentence output.

3. Results and Discussion

This study utilizes 4,179 question-and-answer pairs on Indonesian cultural knowledge, specifically focusing on Dewi Durga, sourced from Mendeley Data. The dataset is compiled from a variety of resources, including printed materials such as books and magazines, as well as online sources like articles, journals, and YouTube.

3.1. Dataset Cleaning Process

Explain how the process of cleaning the dataset starts by entering the text to be processed. Data cleaning not only improves the interpretability of datasets but also reduces noise, enabling more efficient processing and training in machine learning workflows [46]. This research carefully cleans the data to ensure it is suitable for use as a research dataset. In figure 4, the first step is to replace the data detected as duplicates with empty values to eliminate unnecessary data.

Figure 4. Dataset Cleaning Process

Next, the rows of data are tidied up by moving the blank values in each row to the last column, so that the data structure is more organized. After that, empty or missing values are removed to make the data cleaner and more consistent. The last step is to display and save the data cleaning results for further use.

Table 1. Data Creation Process

Question	Answer
Siapa musuh yang dikalahkan oleh Dewi Durga? (Who is defeated by Goddess Durga)	*Dewi Durga mengalahkan musuhnya yakni Mahishasura* (Goddess Durga defeated her enemy Mahishasura)
Durga melambangkan apa? (What does Durga symbolise)	*Durga dikenal sebagai lambang dari keberanian* (Durga is known as the epitome of valour)
Bagaimana wajah Durga digambarkan? (How is Durga's face depicted)	*Wajah Durga digambarkan dalam wajah yang kuat* (Durga's face is depicted in a strong face)

In Table 1, with the use of color codes such as green and yellow highlights the connections between key terms in the questions and answers. These colors are used to identify critical elements within the questions that directly relate to the corresponding answers, providing a visual representation of the complex relationships between them. For example, the green color appearing in several questions is linked to relevant parts of the answers. This shows that certain words or phrases in the questions have a direct connection to similar contexts in the answers.

Likewise, the yellow color marking the word "Dewi Durga" in both the questions and answers indicates that this keyword plays a crucial role in establishing a clear connection between the two. This analysis reveals that the relationship between the questions and answers is not just based on a single word but involves a complex interplay of multiple words or concepts. These connections can span various dimensions, including mentioned entities (such as specific figures or objects), historical or cultural contexts, and broader narrative structuress.

3.2. Training and Testing

In Table 2, Here is the table displaying the detailed training hyperparameters used in the model training process. It includes information on the learning rate, batch sizes for training and evaluation, random seed, optimizer configuration with specific parameters, learning rate scheduler type, and the total number of epochs. This table provides a comprehensive overview of the setup used to fine-tune the model.

Table 2. Training Hyperparameters

Hyperparameter	**Value**
learning_rate	0.0003
train_batch_size	24
eval_batch_size	24
seed	42
optimizer	adamw_torch (betas=(0.9, 0.999), epsilon=1e-08, no additional args)

After determining the parameters, the research obtained the results of fine-tuning FlanT5 and mT5 for the Goddess Durga Question-Answer Pairs dataset. Table 3 shows the ROGUE calculation for FlanT5, while table 3 shows the ROGUE calculation for mT5. In this test, three iterations were conducted, with the results of the final iteration presented in Table 3 as a reference for result comparison.

Table 3. Training results FlanT5 vs mT5

Model Transfomer	**Rouge1**	**Rouge2**	**RougeL**	**RougeLsum**
FlanT5	0.5561	0.4369	0.5537	0.5562
mT5	0.3074	0.0953	0.3026	0.3008

The comparison table between Flan-T5 and mT5 shows that Flan-T5 excels in various aspects, including training loss, validation loss, and ROUGE score. With a training loss of 0.5589 and validation loss of 0.2757, Flan-T5 shows more stable convergence and better generalization ability than mT5, which has a training loss of 27.523 and validation loss of 17.819. On the ROUGE metric, Flan-T5 achieved higher scores in all categories (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum). In ROUGE-2 and ROUGE-L, which shows its superiority in capturing phrase cohesion and overall text structure. Flan-T5's main strength lies in its ability to produce text that is accurate, coherent, and conforms to the expected structure, making it ideal for tasks that require high text quality. Meanwhile, mT5 may excel in multilingual environments, but in this evaluation, it showed a much lower performance.

To validate the superiority of FlanT5 over mT5, statistical significance tests were conducted on the ROUGE scores obtained during evaluation. Paired t-tests revealed that the differences in ROUGE-1,

ROUGE-2, ROUGE-L, and ROUGE-Lsum scores between the two models were statistically significant at the 0.05 level. This indicates that FlanT5's performance is not only consistently better but also unlikely to be a result of random variations in the dataset [47], [48]. Such statistical validation strengthens the conclusion that FlanT5's architecture and fine-tuning approach are fundamentally more effective in preserving cultural context compared to mT5.

In addition to statistical metrics, a deeper exploration of the qualitative differences in model outputs highlighted key insights. FlanT5 generated responses that were more coherent, contextually accurate, and aligned with the cultural nuances embedded in the dataset. For example, when asked culturally specific questions about Goddess Durga, FlanT5 provided well-structured and accurate answers, while mT5 struggled to maintain semantic consistency, often producing repetitive or irrelevant responses. These qualitative differences suggest that FlanT5 pretraining and fine-tuning strategies are better suited for tasks requiring a nuanced understanding of symbolic and context-rich narratives.

This finding aligns with [20], [26] who demonstrated the strength of the T5 framework in handling a wide range of text generation tasks with efficient training convergence. Furthermore, similar to the observations [49] on the advantages of ROUGE in text summarization evaluation, Flan-T5's high ROUGE scores confirm its reliability in producing coherent and relevant summaries. However, mT5's limitations in this specific context reflect findings [23], [50], which suggests that while multilingual models like mT5 or their regional adaptations are effective in multilingual or domain-specific settings, they often require further fine-tuning to reach optimal performance in high-quality text generation tasks. Additionally, [51], [52] emphasize the challenges multilingual transformers face in handling long or complex sequences effectively, which could explain mT5's lower performance compared to the more specialized Flan-T5 model in this evaluation. Therefore, this study supports previous research while highlighting FlanT5's superior efficiency and quality in text generation tasks.

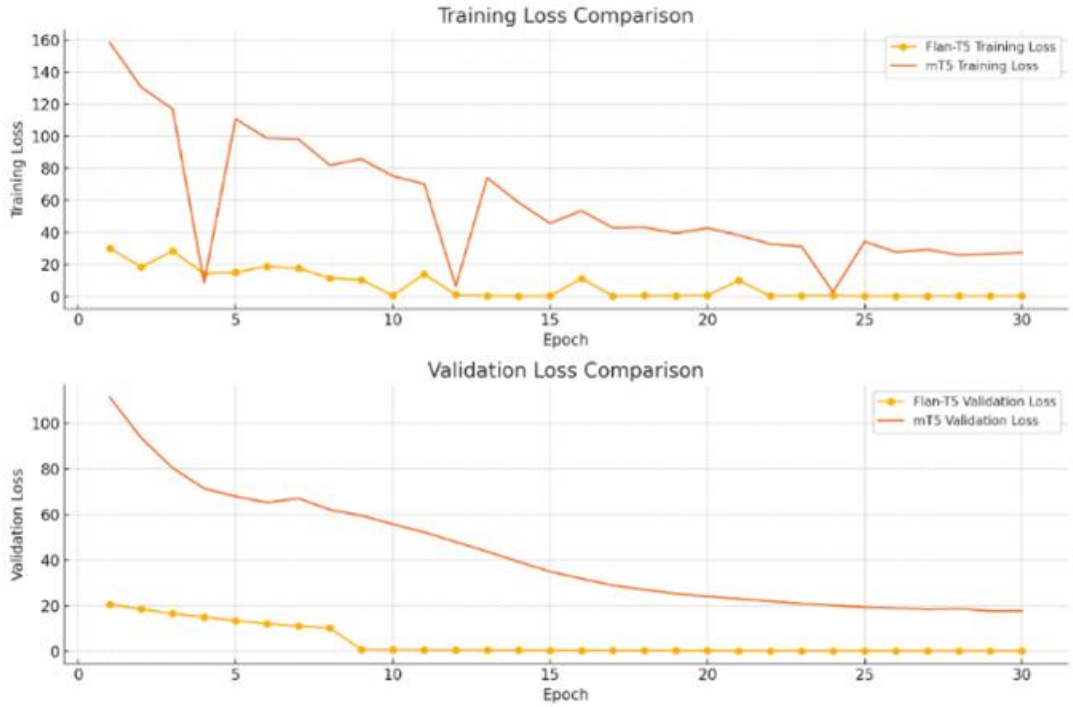

Figure 5. Graphic Model Test Results

In figure 5, the line charts above show the training and validation loss trends for Flan-T5 and mT5 over 30 epochs. Flan-T5 displays consistently lower training and validation losses, indicating that it learns more effectively and generalizes better on the validation data. mT5, on the other hand, starts with a much higher loss and demonstrates fluctuations, especially in training loss, which suggests difficulties

in stable learning. Even as mT5's validation loss gradually decreases, it remains significantly higher than Flan-T5's, underscoring Flan-T5's superior efficiency and reliability in the training process. These findings align with who found that the T5 framework is designed to optimize transfer learning capabilities, resulting in more stable and efficient training processes. Similarly, it is observed that models designed with single-language optimization often yield more consistent learning curves and lower validation losses, as seen with Flan-T5 in this evaluation

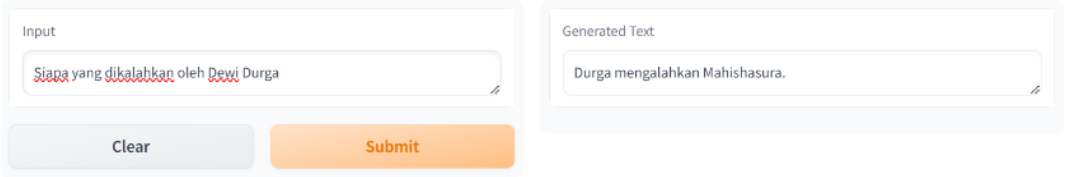

Figure 6. Test Result Answering FlanT5

Figure 7. Test Result Answering mT5

The comparison between FlanT5 and mT5 in answering the question "Who was defeated by Dewi Durga" shows a significant difference in accuracy and relevance. FlanT5, as shown in figure 6, provides the correct answer, "Durga defeated Mahishasura," which is relevant and aligns with the question with the human evaluation. This demonstrates its capability to understand and generate an accurate response. In contrast, mT5 in figure 7 gives a repetitive and nonsensical response, "Durga is worshipped by Durga is worshipped by Durga," which is irrelevant to the question.

The superior performance of FlanT5 can be attributed to its training framework, which emphasizes domain adaptation and in-context learning [53], [54], allowing it to effectively grasp the intricate details of culturally specific datasets. By contrast, mT5's multilingual training focus dilutes its ability to specialize in niche domains, resulting in reduced performance when tasked with context-dependent QA. This disparity aligns with findings in prior studies that show single-domain models often excel in tasks requiring depth over breadth.

Thereby, FlanT5 outperforms mT5 in providing an accurate and relevant answer, while mT5 struggles to generate a correct response for this question. The challenges faced by mT5 in achieving stable convergence are in line with studies [23], [50], who reported that multilingual models, such as idT5 and AraT5, tend to require additional fine-tuning and adaptation to perform optimally, especially in single-language tasks. This observation is further supported [51] who highlighted the difficulty of maintaining efficiency and generalization across multilingual models when handling complex or lengthy text sequences. Overall, while FlanT5 performance aligns with research emphasizing the benefits of targeted, single-language models, mT5 results reflect the inherent challenges of multilingual transformers in maintaining stability and generalization in such settings.

The implications of FlanT5's performance go beyond reaching academic benchmarks. Its ability to preserve cultural narratives and provide context-appropriate responses positions as a valuable asset for AI-based applications in cultural preservation and education. For example, AI can support cultural preservation and context-aware dialogue systems [55], [56], [57], applied to various cultural and educational domains, it can enable the creation of digital cultural archives [50], [51], [58], AI-driven chatbots for exploring traditions [59], [60], [61], [62], virtual museum tours [63], [64], [65], and tools for preserving endangered languages [66], [67] by documenting and translating [68], [69], [70] folklore. Additionally, developing advanced multilingual versions.

So, future research on transformer models, particularly T5-based architectures, could focus on domain-specific fine-tuning and enhancing multilingual capabilities. Fine-tuning strategies tailored for cultural, historical, or niche domains using specialized datasets could improve T5's performance in tasks involving culturally specific or context-rich content. This may involve techniques to reduce translation errors and capture cultural nuances across languages, benefiting applications in diverse cultural and educational contexts.

4. Conclusion

This study demonstrates that FlanT5 outperforms mT5 in handling cultural question-answering tasks, particularly in preserving the contextual nuances of Indonesian cultural knowledge. The superior performance of FlanT5, validated through statistical significance and ROUGE metrics, highlights its capability to generate coherent and contextually relevant outputs, making it a strong candidate for tasks requiring cultural sensitivity.

The findings underscore the potential of FlanT5 in real-world applications such as cultural preservation and education. By enabling the development of AI-driven tools, such as interactive digital archives and educational platforms, FlanT5 can facilitate the documentation and dissemination of underrepresented cultural narratives. These tools can improve global accessibility to indigenous knowledge, support multicultural education, and contribute to the preservation of cultural heritage in a rapidly digitizing world.

Moreover, FlanT5 could be integrated into cross-disciplinary research, promoting innovations in digital humanities and fostering collaborations between AI and cultural studies. Despite these promising results, the study has limitations. The dataset, while culturally specific, is limited in scope and may not represent the broader diversity of Indonesian cultural knowledge. Additionally, the reliance on ROUGE metrics, while effective, could be complemented with human evaluation for a more nuanced assessment of model outputs. Future research should focus on expanding the dataset to include diverse cultural contexts, exploring the integration of additional evaluation metrics, and fine-tuning multilingual models like mT5 to enhance their domain-specific performance. These advancements could pave the way for more robust and inclusive NLP systems, fostering cultural preservation and knowledge dissemination on a global scale.

References

[1] L. Xu, L. Lu, and M. Liu, "Construction and application of a knowledge graph-based question answering system for Nanjing Yunjin digital resources," *Herit. Sci.*, vol. 11, no. 1, pp. 1–17, Dec. 2023, doi: 10.1186/S40494-023-01068-2/TABLES/6.

[2] F. Jin, Q. Chang, and Z. Xu, "MuseumQA: A Fine-Grained Question Answering Dataset for Museums and Artifacts," *ACM Int. Conf. Proceeding Ser.*, pp. 221–226, Dec. 2023, doi: 10.1145/3639479.3639525.

[3] Z. Fan and C. Chen, "CuPe-KG: Cultural perspective–based knowledge graph construction of tourism resources via pretrained language models," *Inf. Process. Manag.*, vol. 61, no. 3, p. 103646, May 2024, doi: 10.1016/J.IPM.2024.103646.

[4] R. C. Climent, D. M. Haftor, and M. W. Staniewski, "AI-enabled business models for competitive advantage," *J. Innov. Knowl.*, vol. 9, no. 3, p. 100532, Jul. 2024, doi: 10.1016/j.jik.2024.100532.

[5] M. Dumas *et al.*, "AI-augmented Business Process Management Systems: A Research Manifesto," *ACM Trans. Manag. Inf. Syst.*, vol. 14, no. 1, pp. 1–19, Mar. 2023, doi: 10.1145/3576047.

[6] J. R. Rodriguez Barboza *et al.*, "Posthumanist Technologies in Business: AI and Cloud Computing for Global Optimization and Ethical Challenges," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 4, p. 02404021, Oct. 2024, doi: 10.26877/asset.v6i4.1064.

[7] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or

generative AI) in healthcare," *npj Digit. Med.*, vol. 6, no. 1, p. 120, Jul. 2023, doi: 10.1038/s41746-023-00873-0.
[8] S. M. Williamson and V. Prybutok, "Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare," *Appl. Sci.*, vol. 14, no. 2, p. 675, Jan. 2024, doi: 10.3390/app14020675.
[9] N. Susetyo *et al.*, "Natural Language Processing in Higher Education," *Bull. Soc. Informatics Theory Appl.*, vol. 6, no. 1, pp. 90–101, Jul. 2022, doi: 10.31763/BUSINTA.V6I1.593.
[10] Y. Pande, "Project EngiBot: Engineering Insights through NLP- driven Chatbot," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 11, pp. 689–692, Nov. 2023, doi: 10.22214/ijraset.2023.56590.
[11] J. Su (苏嘉红) and W. Yang (杨伟鹏), "Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education," *ECNU Rev. Educ.*, vol. 6, no. 3, pp. 355–366, Aug. 2023, doi: 10.1177/20965311231168423.
[12] J. Qadir, "Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education," in *2023 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, May 2023, pp. 1–9. doi: 10.1109/EDUCON54358.2023.10125121.
[13] X. Wang, L. Li, S. C. Tan, L. Yang, and J. Lei, "Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness," *Comput. Human Behav.*, vol. 146, p. 107798, Sep. 2023, doi: 10.1016/j.chb.2023.107798.
[14] T. L. M. Suryanto, A. P. Wibawa, H. Hariyono, and A. Nafalski, "Evolving Conversations: A Review of Chatbots and Implications in Natural Language Processing for Cultural Heritage Ecosystems," *Int. J. Robot. Control Syst.*, vol. 3, no. 4, pp. 955–1006, Dec. 2023, doi: 10.31763/ijrcs.v3i4.1195.
[15] I. D. Raharjo and Egia Rosi Subhiyakto, "Implementing Long Short Term Memory (LSTM) in Chatbots for Multi Usaha Raya," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 4, p. 02404018, Oct. 2024, doi: 10.26877/asset.v6i4.934.
[16] S. Hawanti and K. M. Zubaydulloevna, "AI chatbot-based learning: alleviating students' anxiety in english writing classroom," *Bull. Soc. Informatics Theory Appl.*, vol. 7, no. 2, pp. 182–192, Dec. 2023, doi: 10.31763/BUSINTA.V7I2.659.
[17] P. D. Larasati, A. Irawan, S. Anwar, M. F. Mulya, M. A. Dewi, and I. Nurfatima, "Chatbot helpdesk design for digital customer service." Accessed: Dec. 20, 2024. [Online]. Available: https://pubs2.ascee.org/index.php/aet/article/view/684
[18] P. A. Alia, D. Kartika Sari, N. Azis, B. Gunawan Sudarsono, and P. Agus Sucipto, "Implementation Artificial Intelligence with Natural Language Processing Method to Improve Performance of Digital Product Sales Service," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 3, p. 0240301, Jun. 2024, doi: 10.26877/asset.v6i3.521.
[19] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, May 2020, Accessed: Dec. 19, 2024. [Online]. Available: https://arxiv.org/abs/2005.14165v4
[20] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.
[21] M. A. Hasan *et al.*, "NativQA: Multilingual Culturally-Aligned Natural Query for LLMs," Jul. 2024, Accessed: Dec. 19, 2024. [Online]. Available: https://arxiv.org/abs/2407.09823v2
[22] L. Colucci Cante, B. Di Martino, M. Graziano, D. Branco, and G. J. Pezzullo, "Automated Storytelling Technologies for Cultural Heritage," *Lect. Notes Data Eng. Commun. Technol.*, vol. 193, pp. 597–606, 2024, doi: 10.1007/978-3-031-53555-0_57.
[23] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-Text Transformers for Arabic Language Generation," 2022, [Online]. Available: https://doi.org/10.48550/arXiv.2109.12068
[24] A. Prasetya Wibawa, H. K. Fithri, I. Ari, E. Zaeni, and A. Nafalski, "Generating Javanese Stopwords List using K-means Clustering Algorithm," *Knowl. Eng. Data Sci.*, vol. 3, no. 2, pp.

106–111, Dec. 2020, doi: 10.17977/UM018V3I22020P106-111.
[25] J. Santoso, E. I. Setiawan, C. N. Purwanto, and F. Kurniawan, "Indonesian Sentence Boundary Detection using Deep Learning Approaches," *Knowl. Eng. Data Sci.*, vol. 4, no. 1, pp. 38–48, Jun. 2021, doi: 10.17977/UM018V4I12021P38-48.
[26] L. Xue *et al.*, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 483–498. doi: 10.18653/v1/2021.naacl-main.41.
[27] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
[28] A. Mastropaolo *et al.*, "Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks," pp. 336–347, 2021, doi: 10.1109/ICSE43902.2021.00041.
[29] Z. Chi *et al.*, "mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 1671–1683. doi: 10.18653/v1/2021.emnlp-main.125.
[30] J. Frei and F. Kramer, "Annotated dataset creation through large language models for non-english medical NLP," *J. Biomed. Inform.*, vol. 145, p. 104478, Sep. 2023, doi: 10.1016/j.jbi.2023.104478.
[31] A. I. Regla and M. A. Ballera, "An Enhanced Research Productivity Monitoring System for Higher Education Institutions (HEI's) with Natural Language Processing (NLP)," *Procedia Comput. Sci.*, vol. 230, pp. 316–325, 2023, doi: 10.1016/j.procs.2023.12.087.
[32] L. Ranaldi and F. M. Zanzotto, "Discover AI Knowledge to Preserve Cultural Heritage," Sep. 2021, doi: 10.20944/PREPRINTS202109.0062.V1.
[33] H. R. Will, N. Masini, and D. H. R. Spennemann, "Will Artificial Intelligence Affect How Cultural Heritage Will Be Managed in the Future? Responses Generated by Four genAI Models," *Herit. 2024, Vol. 7, Pages 1453-1471*, vol. 7, no. 3, pp. 1453–1471, Mar. 2024, doi: 10.3390/HERITAGE7030070.
[34] S. Pawar *et al.*, "Survey of Cultural Awareness in Language Models: Text and Beyond," Oct. 2024, Accessed: Dec. 19, 2024. [Online]. Available: https://arxiv.org/abs/2411.00860v1
[35] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer, "Open-Retrieval Conversational Question Answering," *SIGIR 2020 - Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 539–548, 2020, doi: 10.1145/3397271.3401110.
[36] J. Li *et al.*, "Graphix-T5 : Mixing Pre-trained Transformers with Graph-Aware Layers for Text-to-SQL Parsing," 2022, [Online]. Available: https://doi.org/10.48550/arXiv.2301.07507
[37] T. Shao, Y. Guo, H. Chen, and Z. Hao, "Transformer-Based Neural Network for Answer Selection in Question Answering," *IEEE Access*, vol. 7, pp. 26146–26156, 2019, doi: 10.1109/ACCESS.2019.2900753.
[38] X. Chen, P. Cong, and S. Lv, "A Long-Text Classification Method of Chinese News Based on BERT and CNN," *IEEE Access*, vol. 10, pp. 34046–34057, 2022, doi: 10.1109/ACCESS.2022.3162614.
[39] L. Liu, X. Su, H. Guo, and D. Zhu, "A Transformer-based Medical Visual Question Answering Model," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, Aug. 2022, pp. 1712–1718. doi: 10.1109/ICPR56361.2022.9956469.
[40] S. Zhu *et al.*, "Multilingual Large Language Models: A Systematic Survey," 2024, Accessed: Dec. 19, 2024. [Online]. Available: https://github.com/tjunlp-lab/Awesome-Multilingual-LLMs-Papers
[41] B. Weng, "Navigating the Landscape of Large Language Models: A Comprehensive Review and

Analysis of Paradigms and Fine-Tuning Strategies," Apr. 2024, Accessed: Dec. 19, 2024. [Online]. Available: https://arxiv.org/abs/2404.09022v1
[42] M. Nitu and M. Dascalu, "Natural Language Processing Tools for Romanian-Going Beyond a Low-Resource Language", doi: 10.55612/s-5002-060-001sp.
[43] C. Wang, P. Liu, and Y. Zhang, "Can Generative Pre-trained Language Models Serve as Knowledge Bases for Closed-book QA," 2018.
[44] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692
[45] A. Mastropaolo *et al.*, "Using Transfer Learning for Code-Related Tasks," *IEEE Trans. Softw. Eng.*, vol. 49, no. 4, pp. 1580–1598, Apr. 2023, doi: 10.1109/TSE.2022.3183297.
[46] C. Mallikarjuna and S. Sivanesan, "Question classification using limited labelled data," *Inf. Process. Manag.*, vol. 59, no. 6, p. 103094, Nov. 2022, doi: 10.1016/j.ipm.2022.103094.
[47] E. Alsentzer *et al.*, "Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models," *npj Digit. Med. 2023 61*, vol. 6, no. 1, pp. 1–10, Nov. 2023, doi: 10.1038/s41746-023-00957-x.
[48] B. Pecher, I. Srba, and M. Bielikova, "Comparing Specialised Small and General Large Language Models on Text Classification: 100 Labelled Samples to Achieve Break-Even Performance," Feb. 2024, Accessed: Dec. 20, 2024. [Online]. Available: https://arxiv.org/abs/2402.12819v2
[49] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques." 2024. doi: 10.2139/ssrn.4753220.
[50] M. Fuadi, A. D. Wibawa, and S. Sumpeno, "idT5 : Indonesian Version of Multilingual T5 Transformer," 2023, [Online]. Available: https://doi.org/10.48550/arXiv.2302.00856
[51] D. Uthus, S. Ontañón, J. Ainslie, and M. Guo, "mLongT5: A Multilingual and Efficient Text-To-Text Transformer for Longer Sequences," vol. 5, 2023.
[52] M. Guo, J. Ainslie, and D. Uthus, "LongT5 : Efficient Text-To-Text Transformer for Long Sequences," pp. 724–736, 2022, [Online]. Available: https://doi.org/10.48550/arXiv.2112.07916
[53] Z. Zhang, X. He, V. Iyer, and A. Birch, "Cultural Adaptation of Menus: A Fine-Grained Approach," Aug. 2024, Accessed: Dec. 20, 2024. [Online]. Available: https://arxiv.org/abs/2408.13534v1
[54] S. Dwivedi, S. Ghosh, and S. Dwivedi, "Navigating Linguistic Diversity: In-Context Learning and Prompt Engineering for Subjectivity Analysis in Low-Resource Languages," *SN Comput. Sci.*, vol. 5, no. 4, pp. 1–9, Apr. 2024, doi: 10.1007/S42979-024-02770-Z/TABLES/2.
[55] E. Merdivan, D. Singh, S. Hanke, and A. Holzinger, "Dialogue Systems for Intelligent Human Computer Interactions," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 57–71, May 2019, doi: 10.1016/J.ENTCS.2019.04.010.
[56] T. Shen *et al.*, "ViDA-MAN: Visual Dialog with Digital Humans," *MM 2021 - Proc. 29th ACM Int. Conf. Multimed.*, pp. 2789–2791, Oct. 2021, doi: 10.1145/3474085.3478560/SUPPL_FILE/DE3228.MP4.
[57] J. Tian *et al.*, "ChatPLUG: Open-Domain Generative Dialogue System with Internet-Augmented Instruction Tuning for Digital Human," Apr. 2023, Accessed: Dec. 20, 2024. [Online]. Available: https://arxiv.org/abs/2304.07849v3
[58] S. Longpre *et al.*, "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning," *Proc. Mach. Learn. Res.*, vol. 202, pp. 22631–22648, 2023.
[59] V. Bouras *et al.*, "Chatbots for Cultural Venues: A Topic-Based Approach," *Algorithms 2023, Vol. 16, Page 339*, vol. 16, no. 7, p. 339, Jul. 2023, doi: 10.3390/A16070339.
[60] P. K. Rachabatuni, F. Principi, P. Mazzanti, and M. Bertini, "Context-aware chatbot using MLLMs for Cultural Heritage," *MMSys 2024 - Proc. 2024 ACM Multimed. Syst. Conf.*, pp. 459–463, Apr. 2024, doi: 10.1145/3625468.3652193.
[61] M. Tsepapadakis and D. Gavalas, "Are you talking to me? An Audio Augmented Reality

conversational guide for cultural heritage," *Pervasive Mob. Comput.*, vol. 92, p. 101797, May 2023, doi: 10.1016/J.PMCJ.2023.101797.

[62] K. Sathiyabamavathy and A. K. P., "Role of Chatbots in Cultural Heritage Tourism: An Empirical Study on Ancient Forts and Palaces," *https://doi.org/10.1177/24559296241253932*, vol. 9, no. 1, pp. 9–28, Jun. 2024, doi: 10.1177/24559296241253932.

[63] J. Enciso-Anaya, K. Tamil, and Q. Truong, "Virtual Museum Tours of the de Saisset Museum," *Comput. Sci. Eng. Sr. Theses*, Jun. 2024, Accessed: Dec. 20, 2024. [Online]. Available: https://scholarcommons.scu.edu/cseng_senior/306

[64] T. Komarac and Đ. Ozretić Došen, "Understanding virtual museum visits: generation Z experiences," *Museum Manag. Curatorsh.*, vol. 39, no. 3, pp. 357–376, May 2024, doi: 10.1080/09647775.2023.2269129.

[65] Z. Wang, L. P. Yuan, L. Wang, B. Jiang, and W. Zeng, "VirtuWander: Enhancing Multi-modal Interaction for Virtual Tour Guidance through Large Language Models," *Conf. Hum. Factors Comput. Syst. - Proc.*, May 2024, doi: 10.1145/3613904.3642235/SUPPL_FILE/PN7938-SUPPLEMENTAL-MATERIAL-2.PDF.

[66] R. Patria and P. H. Merdeka, "Creative Strategies in the Recovery of Endangered Languages," *J. Lit. Lang. Acad. Stud.*, vol. 2, no. 02, pp. 57–61, Aug. 2023, doi: 10.56855/JLLANS.V2I2.671.

[67] D. K. Nanduri and E. M. Bonsignore, "Revitalizing Endangered Languages: AI-powered language learning as a catalyst for language appreciation," Apr. 2023, Accessed: Dec. 20, 2024. [Online]. Available: https://arxiv.org/abs/2304.09394v1

[68] I. Ozkaya, "Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications," *IEEE Softw.*, vol. 40, no. 3, pp. 4–8, May 2023, doi: 10.1109/MS.2023.3248401.

[69] L. Wang *et al.*, "Document-Level Machine Translation with Large Language Models," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 16646–16661, Apr. 2023, doi: 10.18653/v1/2023.emnlp-main.1036.

[70] W. J. Hutchins, "Machine Translation: History of Research and Applications," *Routledge Encycl. Transl. Technol. Second Ed.*, pp. 128–144, Jan. 2023, doi: 10.4324/9781003168348-7/MACHINE-TRANSLATION-JOHN-HUTCHINS.