Large language models can help to compile content with a cultural theme. However, any information generated by large language models needs to be evaluated to see the truth/fact of the information generated. With many studies discussing the comparison of the capabilities of large language models, there is not much research that directly discusses the comparison of the performance of large language models in producing Indonesian cultural content. This research compares the correctness of the information generated by the large language model using the expert judgment method when creating Indonesian cultural content and its fine-tuning capabilities evaluated using BERTScore. The evaluation method was successfully applied and the results show that in this case, PaLM-2 included less misinformation while GPT-3 excelled in fine-tuning. Using the combination of expert judgment and BERTScore makes it possible to evaluate large language models and obtain additional valid training data to correct deficiencies. The results showed that PaLM-2 produced more valid content with a score of 27 points, while GPT-3 scored 8 points. For training on new datasets/fine-tuning, it was found that the GPT-3 language model was able to learn the dataset more quickly, with a time of 50 minutes and a cost of IDR 27,000, while PaLM-2 took 2 hours 10 minutes and a cost of IDR 1,377,204. For the training dataset evaluation results, GPT-3 is superior with an average of all scores reaching 0.85205. Meanwhile, the PaLM-2 Tuned Model got an average overall score of 0.78942. In this case, the GPT-3 Tuned Model is superior by 8 %. In practice, this method can be used if the assessment is descriptive and requires direct assessment from experts

Keywords: large language model, generative artificial intelligence, GPT-3, PaLM-2, BERTScore Evaluation

UDC 004.8
DOI: 10.15587/1729-4061.2024.309972

LARGE LANGUAGE MODEL (LLM) COMPARISON BETWEEN GPT-3 AND PALM-2 TO PRODUCE INDONESIAN CULTURAL CONTENT

Deni Erlansyah*
Corresponding author
E-mail: deni@binadarma.ac.id
Amirul Mukminin
Doctor of Educational Leardership and Policy Studies (Language Policy), Professor
Department of English Language Education
Jambi University
Jambi-Muara Bulian str., 15, Mendalo Darat, Kec. Jambi Outer City, Muaro Jambi Regency, Jambi, Indonesia
Dedek Julian*
Edi Surya Negara*
Ferdi Aditya*
Rezki Syaputra*
*Data Science Interdisciplinary Research Center
Universitas Bina Darma
Jenderal Ahmad Yani str., 3, Sumatera Selatan, Indonesia, 30111

Received date 31.05.2024
Accepted date 09.08.2024
Published date 30.08.2024

***How to Cite**: Erlansyah, D., Mukminin, A., Julian, D., Negara, E. S., Aditya, F., Syaputra, R. (2024). Large language model (LLM) comparison between GPT-3 and PaLM-2 to produce Indonesian cultural content. Eastern-European Journal of Enterprise Technologies, 4 (2 (130)), 19–29. https://doi.org/10.15587/1729-4061.2024.309972*

1. Introduction

In this modern era, local Indonesian culture is slowly being ignored by society. This includes the impact of foreign culture [1] and increasingly rapid technological developments. For example, the culture of traditional games such as regional traditional dances, which are increasingly rarely seen due to the lack of public interest, so many studios are so devoid of interest that they have to close [2]. For this reason, local culture must continue to be preserved so that extinction does not occur. This is in line with the Sustainable Development Goals (SDGs), which is a global agenda and agreed upon by every United Nations member country. The SDGs contain 17 goals and 169 targets [3]. One of the targets is promoting and preserving world cultural and natural heritage [4]. And one way to promote and maintain cultural heritage is to increase the production of content related to culture. The adaptation of information systems and information technology to follow social developments is one of the aspects of renewal, change, and the latest technological discoveries in order to meet the needs of information and communication technology in society [5]. People nowadays increasingly utilize generative artificial intelligence for helping to produce content [6]. Artificial Intelligence (AI) itself is a field of computer science that focuses on developing systems that can imitate human intelligence [7]. Artificial Intelligence uses certain computational techniques and algorithms that allow computers to learn from data, analyze information, make decisions, and perform tasks that require intelligent thinking [8] and help human activities like chatbot [9], such as ChatGPT and Google Bard [10]. The ability of ChatGPT to understand human language makes it easier to write creatively and the quality is equivalent to human work [11]. The paper about Artificial Intelligence (allows computers to perform tasks that require human intelligence [12]) from [13] states that the combination of artificial intelligence and information technology has great potential in various sectors such as accounting [14] and creative industry [15]. ChatGPT is a combination of reinforcement learning algorithms and human knowledge input of more than 150 billion parameters [16]. Behind that service, there is a Large Language Model (LLM), which is a Deep Learning algorithm that can recognize, summarize, translate, predict, and generate text and other content based on

knowledge obtained from large data sets [17]. An example for this is GPT-3 (Generative Pre-trained Transformer 3), a large language model that is trained with 175 billion data parameters [18], and Google PaLM (Pathways Language Model), which is also a large language model that is trained with 540 billion parameters [19]. One of the large amounts of data that can be collected is the impact of the large amount of data on the internet, which allows the development of data mining methods, analysis methods and algorithms used for study, analysis and observation [20–22]. Information technology has an important role in most organizations that manipulate and collect data in large databases [22]. The size of these parameters does not mean that they show superior performance, such as Google PaLM-2, which was able to surpass the performance of the first PaLM even with only 200 billion parameters [23]. The Large Language Model used by the free version of ChatGPT chatbot is GPT-3.5 Turbo, and GPT-4 in the paid version. Meanwhile, Google Bard is an Artificial Intelligence chatbot program based on the Language Model developed by Google, namely PaLM-2. This program has the ability to answer and respond to various commands or questions from users. Features are available in more than 40 languages. Users can also give commands or questions using images uploaded into Bard. Although as what the research [14] analyzes, the development of generative Artificial Intelligence and the research results show that current Generative AI such as ChatGPT and Google Bard can still provide wrong answers, so the development of this technology is still far from perfect [25–27]. The limitations of GPT-3 lie in the correctness of the answers produced [25].

GPT-3 and PaLM-2 are built on the same basic architecture, namely transformers, which consist of Encoder and Decoder blocks. The transformer architecture consists of an encoder and a decoder, where the encoder takes an input sequence of text tokens $x=x\_1, ..., x\_n$ (the tokens can be words or short text) and converts them into a continuous sequence of values $z=z\_1, ...,z\_n$ or "hidden states" in neural networks. The process that occurs is autoregressive, where the previously transformed values $z\_1,...,z\_(n-1)$ are used as additional input when the encoding x_n is generated. Next, the decoder takes these values and produces an output sequence, $y=(y\_1,...,y\_m)$ [26]. If PaLM-2 maintains these two blocks as the basis of its language model, GPT-3 actually only uses one block, namely Decoder, in the hope of providing better and faster performance, as can be seen in Fig. 1 below.

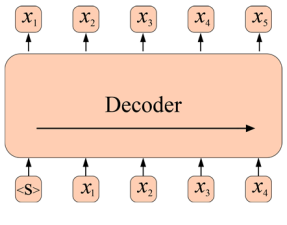

Fig. 1. GPT Architecture [27]

As previously explained, the use of a large language model in creating content about Indonesian culture will be very helpful, but the veracity of the information produced cannot be ascertained. Thus, the studies on performance comparison analysis between large language models (GPT-3 and PaLM-2) are very relevant.

These studies are also important to see how far the two language models are able to produce content about Indonesian culture so that we can see the weaknesses that can be improved in the two language models, especially for language processing and description in languages other than English.

2. Literature review and problem statement

The paper [28], compares performance between the large language model ChatGPT versions 3.5 and ChatGPT 4 assessed based on the overall score, question type, and topic, as well as the level of confidence and reproducibility of answers. In this research, ChatGPT 4 significantly outperformed ChatGPT 3.5 by correctly answering 1662 of 1956 questions (85.0 %) vs 1306 questions (66.8 %) for ChatGPT 3.5. Notably, ChatGPT 4's performance was greater than the human average score of 73.8 %, effectively achieving a close passing and passing grade in a neurology board style exam. As stated in [29], the field of education can create learning experiences that are relevant to future needs. This high potential can be seen in this research, but further research can be applied to see how the capabilities of the two large language models can be improved by providing additional training data (fine-tuning) so that the score can be improved.

The work [30] used a dataset called IndoMMLU, the first multi-task language understanding benchmark for Indonesian culture and language, which consists of questions from primary school to university entrance exams in Indonesia. In this work, their empirical evaluations show that GPT-3.5 only manages to pass the Indonesian primary school level, with limited knowledge of local Indonesian languages and culture. Other smaller models such as BLOOMZ and Falcon perform at an even lower level. This suggests that there are still many shortcomings in the Large Language Model in terms of knowledge, especially regarding Indonesian culture. But this research is only an evaluation and does not show how LLM's abilities will improve if additional training data is provided.

The paper [31] carried out evaluations to assess LLM capabilities to answer exam questions on the VNHSGE (Viet-Namese High School Graduation Examination) English dataset. The performance of BingChat, Bard, and ChatGPT (GPT-3.5) was 92.4 %, 86 %, and 79.2 %, respectively. The results also show that BingChat, Bard and ChatGPT outperform Vietnamese students in English proficiency. This evaluation uses a dataset with a multiple choice type where the correct answer can already be determined. The results of this research do not yet show how the abilities of the LLMs being compared can be improved when given additional training data. Fine-tuning of the model will be carried out by providing human feedback so that it can provide results that match the user's intentions [32].

The paper [33] conducted a performance evaluation of LLM to extract information such as product descriptions, job descriptions, and email templates. This evaluation uses the BERTScore, which is a significant metric that has emerged as an alternative to traditional evaluation metrics in the field of Natural Language Processing (NLP). This tool is used to see the match between the expected results and the results provided by the LLM. The test results obtained BERTscore 86 % (precision), 88 % (recall), and 87 % (F1-Score). In evaluating a dataset, it is important to have the expected questions and answers before starting

to evaluate. This is quite challenging considering that it is necessary to carry out direct research in the field to obtain appropriate data from experts. In this paper, the expected results were obtained from paid software (not humans) as a comparison to LLM, where if the information is specific it still needs an expert and will be difficult to obtain so that in the case of evaluating the truth of the information this cannot be directly applied.

With many studies discussing the comparison of the capabilities of large language models, there is not much research that directly discusses the comparison of the performance of large language models in producing Indonesian cultural content, and not many studies have been conducted to see the extent of LLMs able to increase learning knowledge, especially on Indonesian cultural themes. Assessing LLM capabilities beyond English is increasingly vital but hindered due to the lack of suitable datasets. For this reason, it is necessary to carry out tests in order to see and compare the level of validity of the information about traditional dance from Palembang city provided by GPT-3 and PaLM 2 using Expert Judgment, which is a consideration/opinion of an expert or experienced person [34] and combining it with automatic validation BERTScore (which is a tool that can categorize and test features, especially to measure similarity scores in sentence meaning or context [33]) to see how the LLMs improved their answer based on the training data about traditional dance from Palembang city.

3. The aim and objectives of the study

The aim of the study is to identify performance comparisons between large language models for producing content with Indonesian cultural themes. With the aim of being able to conclude a language model with the validity of content results and fine-tuning performance regarding the best Indonesian culture between GPT-3 and PaLM-2.

To achieve this aim, the following objectives were set:

– to analyze the comparative validity or truth of content about Palembang City Traditional Dance produced by the GPT-3 and PaLM-2 language models;

– to analyze the performance comparison of fine-tuning or new training datasets between the GPT-3 and PaLM-2 language models with the theme of Palembang City Traditional Dance.

4. Materials and methods

This research analyzes the difference in performance between large language models GPT-3 and PaLM-2 in generating Indonesian cultural content. To assess the truth of descriptive information, the expert judgment method will be effective because it directly asks the opinion of experts in the field [34] regarding whether the information produced by artificial intelligence is a fact or not. Based on the research results in the paper [33], which uses BERTScore to evaluate descriptions generated by large language models and paid software with similar capabilities, BERTScore tools are reliable. Because the information to be assessed in this research is descriptive, this research requires experts to label whether the information is factual or not, which can achieved with expert judgment. On the other hand, the information obtained from experts can be used as additional training data for large language models and this evaluation will be effective if it is automatic. So by using BERTScore, it is hoped that the aims and objectives of the research can be achieved and provide results in the form of comparative values between the performance of GPT-3 and PaLM-2 in generating Indonesian cultural content.

This research combines the Expert Judgment and BERTScore approaches. To achieve the first objectives, step 1–2 have been used, and step 3–5 have been used to achieve the second objectives. To conduct this research, the Google Colab website was used with the standard Google Colab device specifications, namely Intel (R) Xeon (R) CPU @ 2.20GHz, NVIDIA T4 GPU, 13 GB RAM. Google Colab is used to connect to the PaLM-2 and GPT-3 API endpoints so that they can use the services provided. In addition, in terms of fine-tuning, PaLM-2 and GPT-3 use similar specifications, namely NVIDIA Tesla A100. Other software used is BERTScore, which is also installed into Google Colab.

The complete research framework can be seen in Fig. 2 below.

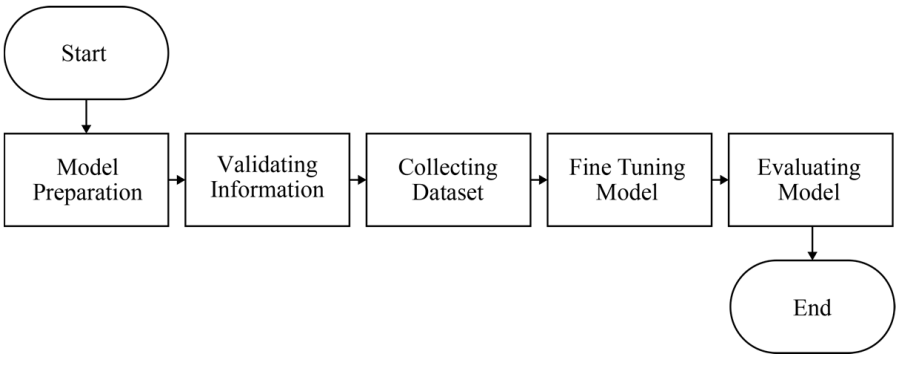

Fig. 2. Research flow

The research will start from the initial stage, namely preparing the model to be able to produce content using the Large Language Model GPT-3 and PaLM-2, to the model evaluation stage using BERTScore. The details are as follows:

1. Model preparation. This stage is carried out by creating program code to test and use the GPT-3 and PaLM-2 models via the API that has been provided. The program code will be written and executed on the Google Colab website.

2. Validating information. The research design subs section encapsulates the strategic framework and methodology that underpins the entire research endeavor. This activity is carried out using the expert judgment method, namely asking for the assessment or opinion of an expert who is responsible, experienced and has a reputation [34]. The experts who will be consulted are from the Palembang Culture Service. The first step that will be taken is to create content about Indonesian culture with GPT-3 and PaLM-2. The results of this content will be given to experts to be assessed in terms of the veracity of the information produced. In carrying out the assessment, the expert will be given a number of generated information in the form of sentences and then a response will be

given regarding the truth of the information. The final results of the assessment draw the overall conclusion from the assessment given. In filling out the questionnaire, respondents will be asked to classify the information presented as Correct (C), False (F), and incomplete (I), as follows:

1) C – correct, meaning that the information conveyed is correct, will be given 2 points;

2) I – incomplete, meaning that the information conveyed is correct but incomplete, which can cause readers to misinterpret it, will be given 1 point;

3) F – false, meaning that the information submitted is wrong or incorrect, will be given 0 points.

Every labeled piece of information has a point to calculate who has a better answer. If the respondent chooses the Inappropriate or Incomplete classification, the respondent is asked to fill in the information column with the reasons or information that should be provided. Table 1 below shows an example of an assessment table.

3. Collecting Dataset: the new dataset was collected by taking it directly to the Palembang City Culture Service through interviews and literature studies. The new dataset will be used to adjust the knowledge of each model.

4. Fine-tuning model: Fine-tuning a model is carried out with the aim to perform a particular function in a specific field [35]. In this stage, we train the dataset that has previously been collected. The dataset for training will be in the form of a question along with the expected answer, while the dataset for testing will be in the form of questions to be tested. Next, the model will study the dataset and provide results in the form of answers to the questions being tested.

5. Evaluating model: the final stage in this research is evaluating the suitability of the meaning of the language model using the BERTScore tool, an automatic evaluation metric for text generation [36]. This tool will compare the answers produced by GPT-3 and PaLM-2 with the expected answers. The resulting assessment points are precision, recall, F1-Score (calculated as the harmonic mean of the precision and recall scores).

Table 1

Example of assessment format

Content	Classification			Information
	Correct	False	Incomplete	
Jumlah tari adat yang berasal dari Kota Palembang adalah 14, yaitu tari gending sriwijaya, tari tepak, tari tanggai, ...	□	–	–	–
Tari Gending Sriwijaya ditampilkan dalam acara adat seperti ...	–	–	✓	Tari Gending Sriwijaya tidak hanya ditampilkan dalam acara adat, akan tetapi juga ditampilkan untuk acara seremoni pemerintahan seperti ...

Table 1 above will be filled in by experts who can assess the veracity of the information produced by LLM.

5. Results of research on comparing large language models for creating Indonesian cultural content

5. 1. Comparative validity analysis of content generated by GPT-3 and PaLM-2

Program code is written to integrate the GPT-3 and PaLM-2 language models so that they can be used to create Indonesian cultural content through the Google Colab application in Python. The process of generating content with a large language model requires a command word/prompt. The command used to generate content with GPT-3 and PaLM-2 is the same, namely as follows:

”Sebutkan jumlah total ada berapa banyak tari adat yang berasal dari kota palembang, sumatera selatan, dan lengkapi dengan daftar nama-namanya. Setelah itu, Buatkan konten tentang Tari Adat Kota Palembang. Konten tersebut harus berisi 5 Tari Adat yang berasal dari Kota Palembang, Sumatera Selatan, Berdasarkan List Sebelumnya. Setiap deskripsi tarian, harus di lengkapi dengan sejarah diciptakannya masing-masing tarian, makna dari masing-masing tari, ciri khas masing-masing tarian dan digunakan dalam rangka apa.

Contoh format jawaban :

Total tari adat yang berasal dari kota palembang adalah sejumlah ... , yaitu tari ..., dll.

5 diantaranya adalah :

1. Tari ...

– Sejarah:

– Makna:

– Ciri khas:

– Digunakan dalam rangka : ”

For the GPT-3 language model, it is integrated via the API endpoint from Openai by utilizing the key obtained from the Openai page. Then the prompt is used to get the results from GPT-3 as shown in Fig. 3.

Based on these results, it can be seen that GPT-3 is able to understand the commands given and return answers in the requested format. For the PaLM-2 language model, Google Cloud is integrated to access VERTEX-AI (which is a medium for using LLM from Google). Fig. 4 below shows the results obtained using the same prompt.

From the response above, it can be seen that there is a slight error in the content writing format. The first sentence regarding information on the number of dances is given in the “Makna:” format. Apart from that, the answer given is in accordance with the requested command.

In determining the correctness of information from content produced by GPT-3 and PaLM-2, an Expert Judgment approach is used. This stage is carried out by giving questionnaires to respondents as experts to classify the information/content created by the two language models. The information provided is 21 for each language model, in accordance with the content produced in the previous stage, namely about the number of traditional dances originating from the city of Palembang and 5 traditional dances accompanied by information about the history, meaning, characteristics and use in the context of What. After this stage was carried out, the results were obtained in the form of the amount of information that was declared appropriate, inappropriate and incomplete. A recap of these results can be seen in Table 2 below.

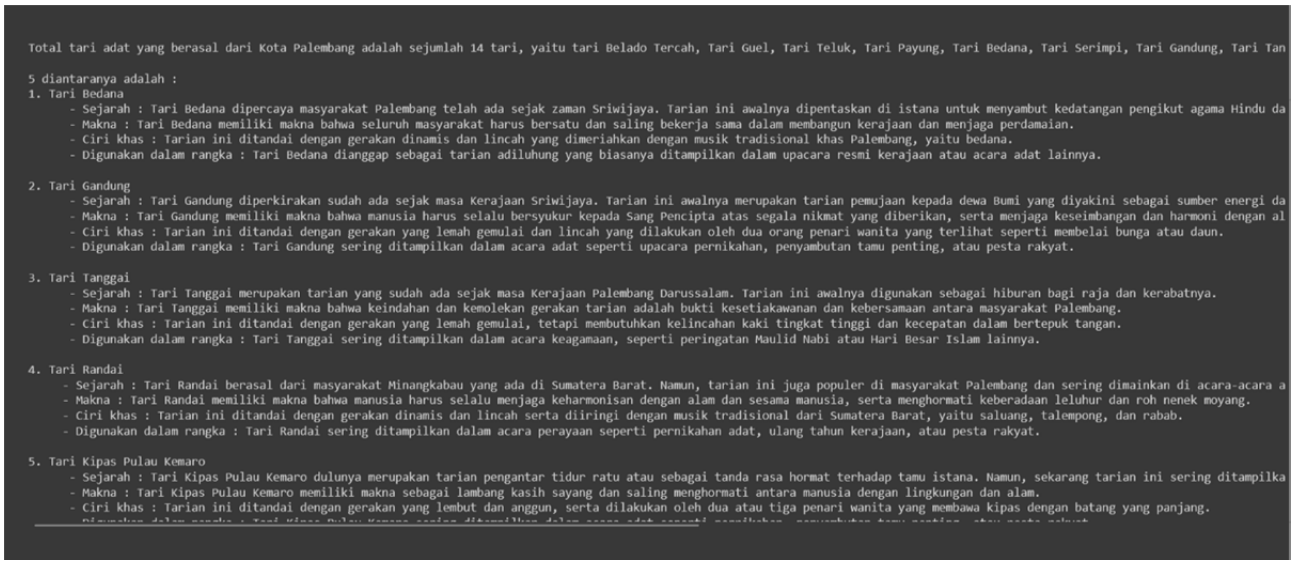

Fig. 3. GPT-3 Answer

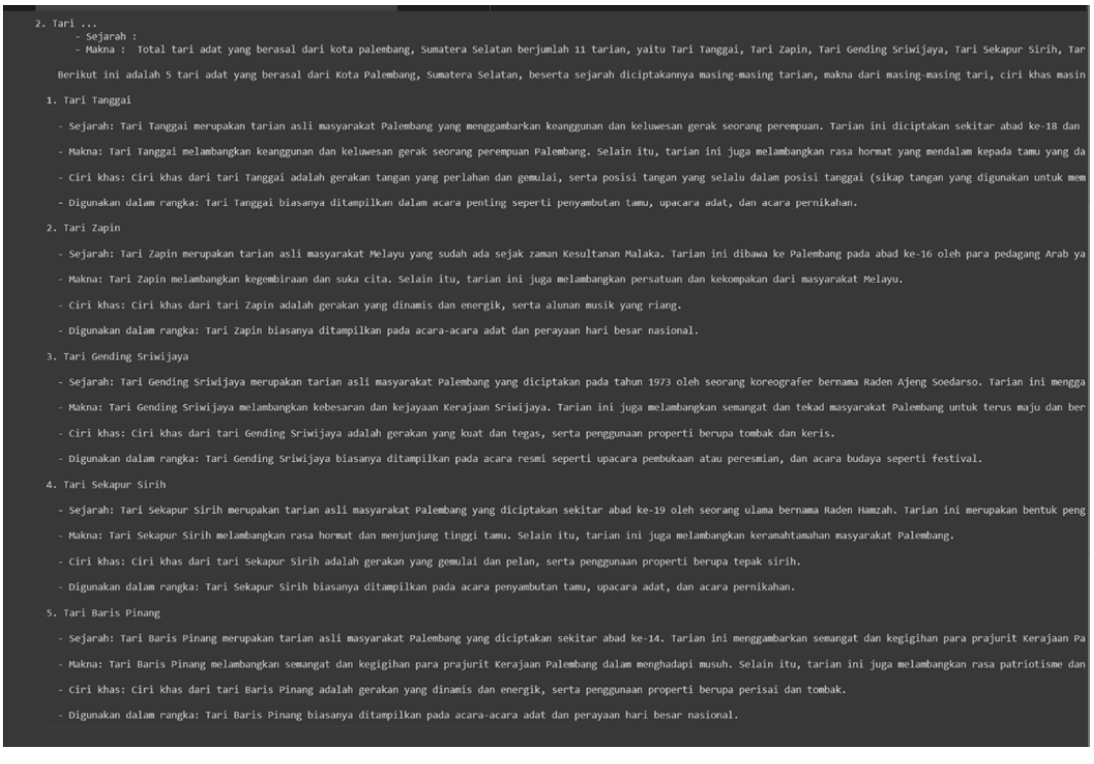

Fig. 4. PaLM-2 Answer

Table 2

Validating information result

LLM	Classification			Total Points
	Correct	False	Incomplete	
GPT-3	3	16	2	8
PaLM-2	11	5	5	27

Apart from classification, there are several statements or notes provided by respondents regarding information that is inappropriate or incomplete, including the following:

– GPT-3 mentions: “The total number of traditional dances originating from the city of Palembang is 14 dances, namely the Belado Tercah dance, Guel Dance, Teluk Dance, Umbrella Dance, Bedana Dance, Serimpi Dance, Gandung Dance, Tanggai Dance, Sekapur Sirih Dance, Tempurung Dance, Kemaro Island Fan Dance, Randai Dance, Solo Guitar Dance, and Bakpau Performance Dance”. This information is classified as inappropriate or false information for the reason that some dances do not originate from Palembang, such as Tari Payung, Tari Serimpi, Tari Randai, Tari Sekapur Sirih, and several other dances;

– GPT-3 mentions: “Tanggai dance is a dance that has existed since the time of the Palembang Darussalam Kingdom. This dance was originally used as entertainment for the king and his relatives”. Such information is classified as inappropriate for reasons that the Tanggai dance is a welcoming dance for guests in the city of Palembang, which functions to

welcome guests who come to Palembang as a form of openness from the people of Palembang towards guests who come;

– PaLM-2 mentions: "The Gending Sriwijaya Dance is an original dance from the Palembang people, which was created in 1973 by a choreographer named Raden Ajeng Soedarso. This dance depicts the greatness and glory of the Sriwijaya Kingdom". The information is classified as inappropriate on the grounds that it was created in 1943 by Sukainah Rozak;

– PaLM-2 states: "The Sriwijaya Gending Dance is usually performed at official events such as opening or inauguration ceremonies, and cultural events such as festivals". This information was classified as inappropriate for the reason that Gending Sriwijaya Dance based on its standard is performed at official events to welcome RI 1 and 2 or other ambassadors.

At this stage, it can be seen that the content produced by both language models gets inappropriate classifications in some information, and PaLM-2 in this case outperforms GPT-3 because it produces content with better data correctness, as seen from its status, which gets more appropriate classification.

5. 2. Analyzing the performance comparison of fine-tuning between GPT-3 and PaLM-2

In collecting the new dataset, a dataset obtained from the source, namely the Palembang City Culture Service, was used. The data obtained is in the form of an article explaining the Traditional Dance of the City of Palembang. The data is then collected in the form of example questions and answers so that they can be used to train the language model used. The two language models have slight differences in their writing format. The GPT-3 language model requires a dataset format like the following:

{"messages": [{"role": "user", "content": "Sebutkan seni tari yang berasal dari Kota Palembang?"}, {"role": "assistant", "content": "Terdapat 14 judul seni tari yang berasal dari Kota Palembang, yaitu Tari Gending Sriwijaya, Tari Tepak atau Tari Tanggai, Tari Sondok Piyogo, Tari Madik, Tari Majeng Besuko, Tari Ngobeng, Tari Ngocek Rokok, Tari Pirdi, Tari Pelimbangan, Tari Selendang Mayang, Tari Senik, Tari Dana Penganten, Tari Tenun Songket, dan Tari Dian yang Tak Kunjung Padam atau Tari Lilin Siwa."}]}.

Meanwhile, PaLM-2 requires a dataset format like the following:

{"input_text": "question: Sebutkan seni tari yang berasal dari Kota Palembang?", "output_text": "Terdapat 14 judul seni tari yang berasal dari Kota Palembang, yaitu Tari Gending Sriwijaya, Tari Tepak atau Tari Tanggai, Tari Sondok Piyogo, Tari Madik, Tari Majeng Besuko, Tari Ngobeng, Tari Ngocek Rokok, Tari Pirdi, Tari Pelimbangan, Tari Selendang Mayang, Tari Senik, Tari Dana Penganten, Tari Tenun Songket, dan Tari Dian yang Tak Kunjung Padam atau Tari Lilin Siwa."}.

After the dataset is collected, the next stage is to train the dataset on both language models. For GPT-3, a configuration of 32 n_epochs (training repetitions) is used to optimize the training dataset. This is also done to get less training loss. The results of this fine-tuning activity can be seen in Fig. 5 below.

The execution of this command takes approximately 50 minutes and costs $1.74 or around IDR 27,000. This status can be seen from the openai api page or through the program code to see the fine-tuning status of the model. For the PaLM-2 language model, a learning_rate_multiplier (training repetition) configuration of 32 is also used to optimize the training dataset. The results of this fine-tuning activity can be seen in Fig. 6 below.

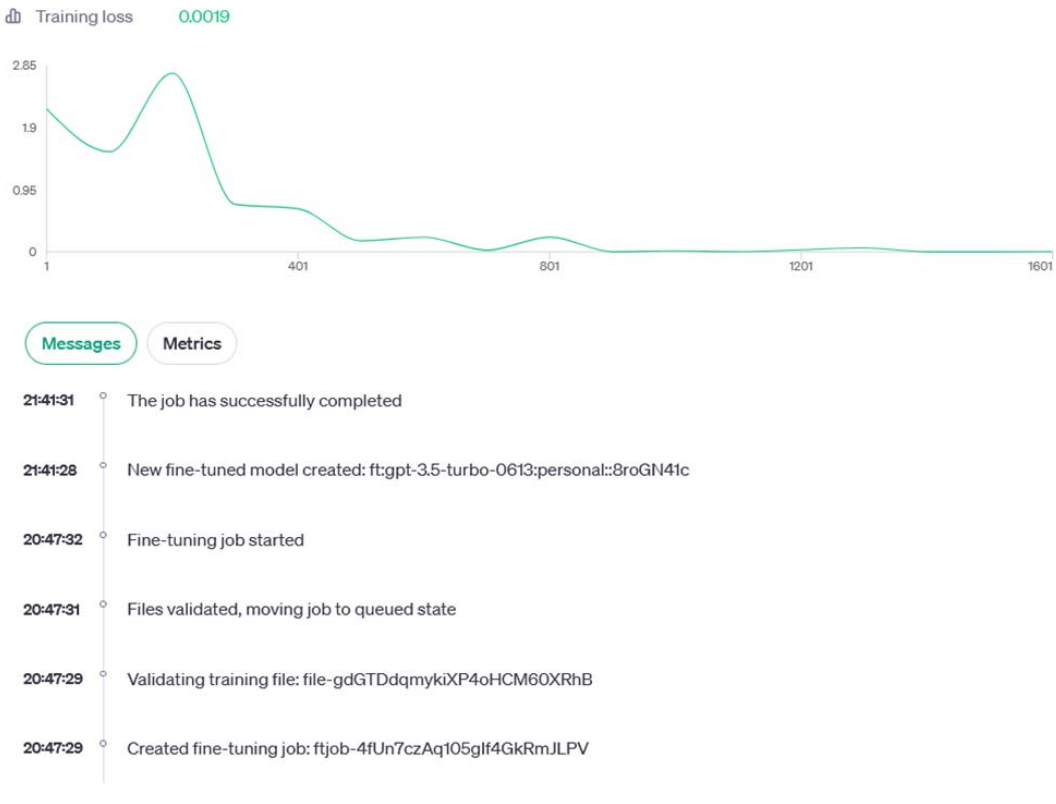

Fig. 5. GPT-3 Fine-Tuning History

Fig. 6. PaLM-2 Fine-Tuning History

The execution of this order took approximately 2 hours 10 minutes and cost IDR 1,377,204. Based on this, it can be concluded that the time and costs required by PaLM-2 to perform fine-tuning exceed the time and costs required by GPT-3. This can occur due to the difference in initial costs where fine-tuning GPT-3 costs $0.008 per 1000 tokens, while fine-tuning PaLM-2 costs $4.517292 per hour times the number of CPU cores used for the 80GB Nvidia A100 machine. Meanwhile, the different speeds come from architectural differences, which result in fewer processes being carried out on GPT-3 compared to PaLM-2.

After the previous fine-tuning stages have been completed, the model that has been trained can be used and evaluated with the hope that the answers produced match the answers given in the training data, so that the data results obtained are more accurate. This stage is carried out using the BERTScore Evaluation Model. Model evaluation was performed twice, namely on the base model, then evaluation was carried out again on the model resulting from fine-tuning to see the improvement in the answers given. The evaluation was conducted by giving prompts that were similar to the training data in the GPT-3 language model so that based on the answers it could be compared with the expected answers. Table 3 below shows the results of the evaluation performed for the GPT-3 language model before fine-tuning was carried out.

From the evaluation performed, the average evaluation results were obtained, namely Precision: 0.69162, Recall: 0.69880, and F-1 Score: 0.69430. Next, evaluation was carried out again using the same method on the model that had been trained with additional datasets and obtained results as in Table 4 below.

From the evaluation carried out, the average evaluation results obtained were Precision: 0.84593, Recall: 0.85872, and F-1 Score: 0.85150. Evaluation of this model was performed twice, namely on the base model, then evaluation was carried out again on the model resulting from fine-tuning to see the improvement in the answers given. The evaluation was conducted by giving prompts that were similar to the training data on the PaLM-2 language model so that based on the answers it could be compared with the expected answers. Table 5 below shows the results of the evaluation performed for the PaLM-2 language model before fine-tuning was carried out.

From the evaluation performed, the average evaluation results obtained were Precision: 0.63636, Recall: 0.69048, and F-1 Score: 0.66131. Next, evaluation was carried out again using the same method on the model that had been given an additional dataset with the results as in Table 6 below.

Table 3

GPT-3 base model evaluation result

No.	Expectation	Result	Score		
			P	R	F-1
1	Menurut definsisi dalam UU No 5 tahun 2017 tentang Pemajuan Kebudayaan, seni adalah ekspresi artistik individu, kolektif, atau komunal, yang berbasis warisan budaya maupun berbasis kreativitas penciptaan baru, yang terwujud dalam berbagai bentuk kegiatan dan/atau medium. Seni antara lain seni pertunjukan, seni rupa, seni sastra, film, seni musik, dan seni media	Menurut Undang-Undang no. 5 tahun 2017 tentang Ekonomi Kreatif, seni adalah kegiatan kreatif yang dihasilkan oleh manusia yang memiliki nilai estetis dan dapat mempengaruhi pikiran, emosi, dan pendapat orang lain. Seni mencakup berbagai aspek seperti seni visual, seni rupa, musik, tari, teater, film, dan literatur. Seni dianggap sebagai salah satu sektor dalam ekonomi kreatif yang memiliki potensi untuk menghasilkan daya tarik, nilai tambah, dan keuntungan ekonomi	0.7532	0.7617	0.7574
...	...	...	...	...	...
51	Tari Adat tertua dari Kota Palembang adalah Tari Gending Sriwijaya	Tari Adat tertua yang berasal dari Kota Palembang adalah Tari Arau. Tari ini merupakan tarian yang dipertunjukkan oleh masyarakat Palembang sejak abad ke-7 hingga abad ke-11 Masehi	0.7548	0.8760	0.8109
Total			35.2728	35.6391	35.4095

Table 4

GPT-3 tuned model evaluation result

No.	Expectation	Result	Score		
			P	R	F-1
1	Menurut definsisi dalam UU No 5 tahun 2017 tentang Pemajuan Kebudayaan, seni adalah ekspresi artistik individu, kolektif, atau komunal, yang berbasis warisan budaya maupun berbasis kreativitas penciptaan baru, yang terwujud dalam berbagai bentuk kegiatan dan/ atau medium. Seni antara lain seni pertunjukan, seni rupa, seni sastra, film, seni musik, dan seni media	Dalam UU No 5 tahun 2017 tentang Pemajuan Kebudayaan, yang dimaksud dengan seni adalah: berbagai ekspresi artistik individu, kolektif, atau komunal, yang terwujud dalam bentuk-bentuk abadinya maupun pertunjukan, yang tersebar dalam berbagai disiplin ilmu dan/ atau bidang kehidupan, seperti seni pertunjukan, seni rupa, seni sastra, film, seni musik, dan/ atau seni media	0.8754	0.8660	0.8707
...	...	...	...	...	...
51	Tari Adat tertua dari Kota Palembang adalah Tari Gending Sriwijaya	Tari Adat tertua dari Kota Palembang adalah Tari Gending Sriwijaya	1.0000	1.0000	1.0000
Total			43.1427	43.795	43.427

Table 5

PaLM-2 base model evaluation result

No.	Expectation	Result	Score		
			P	R	F-1
1	Menurut definsisi dalam UU No 5 tahun 2017 tentang Pemajuan Kebudayaan, seni adalah ekspresi artistik individu, kolektif, atau komunal, yang berbasis warisan budaya maupun berbasis kreativitas penciptaan baru, yang terwujud dalam berbagai bentuk kegiatan dan/ atau medium. Seni antara lain seni pertunjukan, seni rupa, seni sastra, film, seni musik, dan seni media	Seni merupakan salah satu budaya yang sangat penting bagi setiap negara di dunia. Hal ini dapat dilihat dari adanya perlindungan terhadap hak cipta, baik di tingkat nasional maupun internasional. Di Indonesia, perlindungan terhadap hak cipta diatur dalam Undang-Undang Nomor 19 Tahun 2002 tentang Hak Cipta	0.6577	0.7179	0.6865
...	...	...	...	...	...
51	Tari Adat tertua dari Kota Palembang adalah Tari Gending Sriwijaya	Tari Adat tertua Palembang adalah Tari Gending Sriwijaya Tari Gending Sriwijaya merupakan tarian asli Palembang yang menggambarkan kebesaran Kerajaan Sriwijaya. Tarian ini biasanya ditampilkan pada acara-acara kebesaran kerajaan dan juga pada acara penyambutan tamu	0.6543	0.8316	0.7324
Total			32.4544	35.2146	33.7271

Table 6

PaLM-2 tuned model evaluation result

No.	Expectation	Result	Score		
			P	R	F-1
1	Menurut definsisi dalam UU No 5 tahun 2017 tentang Pemajuan Kebudayaan, seni adalah ekspresi artistik individu, kolektif, atau komunal, yang berbasis warisan budaya maupun berbasis kreativitas penciptaan baru, yang terwujud dalam berbagai bentuk kegiatan dan/ atau medium. Seni antara lain seni pertunjukan, seni rupa, seni sastra, film, seni musik, dan seni media	Menurut definsisi dalam UU No 5 tahun 2017 tentang Pemajuan Kebudayaan, seni adalah ekspresi artistik individu, kolektif, atau komunal, yang berbasis warisan budaya maupun berbasis kreativitas penciptaan baru, yang terwujud dalam berbagai bentuk kegiatan dan/ atau medium. Seni antara lain seni pertunjukan, seni rupa, seni sastra, film, seni musik, dan seni media	1.0000	1.0000	1.0000
...	...	...	...	...	...
51	Tari Adat tertua dari Kota Palembang adalah Tari Gending Sriwijaya	Tari Adat tertua dari Kota Palembang adalah Tari Gending Sriwijaya	1.0000	1.0000	1.0000
Total			39.7358	40.8687	40.1778

Before fine-tuning was carried out with the dataset, the lowest score was obtained by the PaLM-2 Base Model with the average of all scores being 0.66271. Meanwhile, the GPT-3 Base Model gets an average score of 0.69490, so it is 5 % ahead. After fine-tuning, from the PaLM-2 Base Model to the PaLM-2 Tuned Model the score increased by 19 %, while the GPT-3 Base Model to the GPT-3 Tuned Model the score increased by 23 %. After being given the dataset, the highest score was obtained by the GPT-3 Tuned Model with an average of all scores reaching 0.85205. Meanwhile, the PaLM-2 Tuned Model got an average overall score of 0.78942. In this case, the GPT-3 Tuned Model is superior by 8 %.

Based on the results obtained above, it is known that in the Palembang City Traditional Dance dataset training, GPT-3, which was built on a transformer architecture basis and only retains the Decoder processing block, can exceed the performance of PaLM-2 with the same architecture base and still retains the Encoder and Decoder processing blocks.

6. Discussion of results on comparing large language models for creating Indonesian cultural content

The two language models used, namely GPT-3 and PaLM-2, are able to produce Indonesian cultural content, especially on the topic of Palembang City Traditional Dance. However, the content still contains invalid information based on expert assessment where GPT-3 produces 16 points of information that is not appropriate, while PaLM-2 produces 5 points of information that is not appropriate (Table 1). For training on a new dataset (fine-tuning), it was found that the GPT-3 language model was able to learn the dataset more quickly, with a time of 50 minutes (Fig. 5) and a cost of IDR 27,000, while PaLM-2 took 2 hours 10 minutes (Fig. 6) and a cost of IDR 1,377,204. For the training dataset evaluation results, GPT-3 is also superior with an average of all scores reaching 0.85205 (Table 3). Meanwhile, the PaLM-2 Tuned Model got an average overall score of 0.78942 (Table 5). In this case, the GPT-3 Tuned Model is superior by 8 %.

Through this research, it is known that the method used can be a way to improve knowledge from LLM by providing existing knowledge to experts and then returning the feedback obtained to LLM. This is different from the evaluation carried out in [33] where the expected results were obtained from paid software (not humans). It is proven that with the dataset obtained from experts, PaLM-2 increased the score by 19 %, while the GPT-3 Base Model to the GPT-3 Tuned Model increased the score by 23 %.

The results obtained in this research are only limited to knowing the differences in the performance of 2 LLMs in producing and studying (fine-tuning) facts about traditional dance in the city of Palembang, not yet in more detail about how to improve the learning ability of these 2 language models to learns the data so that it has better results.

This study does not give a comparison of more language models, especially LLM, which is open source, so it does not show the ability of other large language models in producing and learning facts about traditional dances in the city of Palembang.

Further development of this research that can be done is to turn LLMs into experts themselves, so that every question entered by the user can be answered with accurate facts. The challenge of this development is the difficulty of getting more data regarding traditional dances in the city of Palembang due to the lack of sources. Apart from that, a more in-depth evaluation is needed when the data is trained on LLM because there is still a possibility that the data will not be digested properly.

7. Conclusions

1. It is proven that when using Large Language Models to produce Indonesian cultural content, the results still contains invalid information based on expert assessment where of the 21 questions asked, GPT-3 produces 16 points of information that is not appropriate, while PaLM-2 produces 5 points of information that is not appropriate.

2. For training on a new dataset (fine-tuning), it was found that the GPT-3 language model was able to learn the dataset more quickly, with a time of 50 minutes and a cost of IDR 27,000, while PaLM-2 took 2 hours 10 minutes and a cost of IDR 1,377,204. For the training dataset evaluation results, GPT-3 is also superior with an average of all scores reaching 0.85205. Meanwhile, the PaLM-2 Tuned Model got an average overall score of 0.78942. In this case, the GPT-3 Tuned Model is superior by 8 %.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Funding

The study was performed without financial support.

Data availability

All data are available in the main text of the manuscript.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

Acknowledgments

Thank you to the Palembang City Culture Service for being a resource for the data in this research.

References

1. Wijaya, J. H. (2023). Lifestyle Transformation in Indonesia: The Impact of Foreign Cultures in the Era of Globalization. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4511264
2. Adnan, N. (2014). Character Building Through Traditional Dance As Developing Identity Belongings: A Study Of Indonesia-Malaysia. Proceeding of the Third International Seminar on Languages and Arts. Padang. Available at: https://ejournal.unp.ac.id/index.php/isla/article/view/5412/
3. Barbier, E. B., Burgess, J. C. (2017). The Sustainable Development Goals and the systems approach to sustainability. Economics, 11 (1). https://doi.org/10.5018/economics-ejournal.ja.2017-28
4. Yamasaki, K., Yamada, T. (2022). A framework to assess the local implementation of Sustainable Development Goal 11. Sustainable Cities and Society, 84, 104002. https://doi.org/10.1016/j.scs.2022.104002
5. Negara, E., Hidayanto, A., Andryani, R., Syaputra, R. (2021). Survey of Smart Contract Framework and Its Application. Information, 12 (7), 257. https://doi.org/10.3390/info12070257

6. Lyu, Y., Zhang, H., Niu, S., Cai, J. (2024). A Preliminary Exploration of YouTubers' Use of Generative-AI in Content Creation. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613905.3651057
7. Zhang, C., Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224
8. Koteluk, O., Wartecki, A., Mazurek, S., Ko odziejczak, I., Mackiewicz, A. (2021). How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. Journal of Personalized Medicine, 11 (1), 32. https://doi.org/10.3390/jpm11010032
9. Shabbir, J., Anwer, T. (2018). Artificial Intelligence and its Role in Near Future. arXiv. https://doi.org/10.48550/arXiv.1804.01396
10. Ahmed, I., Roy, A., Kajol, M., Hasan, U., Datta, P. P., Reza, Md. R. (2023). ChatGPT vs. Bard: A Comparative Study. https://doi.org/10.22541/au.168923529.98827844/v1
11. Shidiq, M. (2023). The Use Of Artificial Intelligence-Based Chat-gpt And Its Challenges For The World Of Education; From The Viewpoint Of The Development Of Creative Writing Skills. Proceeding of International Conference on Education, Society and Humanity, 353–357. Available at: https://ejournal.unuja.ac.id/index.php/icesh/article/view/5614
12. González García, C., Núñez-Valdez, E., García-Díaz, V., Pelayo G-Bustelo, C., Cueva-Lovelle, J. M. (2019). A Review of Artificial Intelligence in the Internet of Things. International Journal of Interactive Multimedia and Artificial Intelligence, 5 (4), 9. https://doi.org/10.9781/ijimai.2018.03.004
13. Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., Kuusk, A. (2023). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. Expert Systems with Applications, 216, 119456. https://doi.org/10.1016/j.eswa.2022.119456
14. Hasan, A. R. (2022). Artificial Intelligence (AI) in Accounting & Auditing: A Literature Review. Open Journal of Business and Management, 10 (01), 440–465. https://doi.org/10.4236/ojbm.2022.101026
15. Hughes, R. T., Zhu, L., Bednarz, T. (2021). Generative Adversarial Networks–Enabled Human–Artificial Intelligence Collaborative Applications for Creative and Design Industries: A Systematic Review of Current Approaches and Trends. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.604234
16. Tri Julianto, I., Kurniadi, D., Septiana, Y., Sutedi, A. (2023). Alternative Text Pre-Processing using Chat GPT Open AI. Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), 12 (1), 67–77. https://doi.org/10.23887/janapati.v12i1.59746
17. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M. et al. (2023). A Comprehensive Overview of Large Language Models. arXiv. Available: https://doi.org/10.48550/arXiv.2307.06435
18. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. et al. (2023). Language Models are Few-Shot Learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165
19. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A. et al. (2022). PaLM: Scaling Language Modeling with Pathways. arXiv. https://doi.org/10.48550/arXiv.2204.02311
20. Andryani, R., Surya Negara, E., Syaputra, R., Erlansyah, D. (2023). Analysis of Academic Social Networks in Indonesia. Qubahan Academic Journal, 3 (4), 409–421. https://doi.org/10.58429/qaj.v3n4a289
21. Negara, E. S., Keni, K., Andryani, R., Syaputra, R. S., Widyanti, Y. (2023). Social network analysis to detect influential actors with Indonesian hastags using the centrality method. Sixth International Conference of Mathematical Sciences (ICMS 2022). https://doi.org/10.1063/5.0126819
22. Negara, E. S., Andryani, R., Erlansyah, D., Syaputra, R. (2020). Analysis of Indonesian Motorcycle Gang with Social Network Approach. International Journal of Advanced Computer Science and Applications, 11 (12). https://doi.org/10.14569/ijacsa.2020.0111224
23. Nurhachita, N., Negara, E. S. (2021). A comparison between deep learning, na ve bayes and random forest for the application of data mining on the admission of new students. IAES International Journal of Artificial Intelligence (IJ-AI), 10 (2), 324. https://doi.org/10.11591/ijai.v10.i2.pp324-331
24. Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passoset, A. et al. (2023). PaLM 2 Technical Report. arXiv. https://doi.org/10.48550/arXiv.2305.10403
25. Porter, J. (2023). ChatGPT continues to be one of the fastest-growing services ever. The Verge. Available at: https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference
26. Aydin, Ö., Karaarslan, E. (2023). Is ChatGPT Leading Generative AI? What is Beyond Expectations? Academic Platform Journal of Engineering and Smart Systems, 11 (3), 118–134. https://doi.org/10.21541/apjess.1293702
27. Farquhar, S., Varma, V., Kenton, Z., Gasteiger, J., Mikulik, V., Shah, R. (2024). Challenges with unsupervised LLM knowledge discovery. arXiv. https://doi.org/10.48550/arXiv.2312.10029
28. Floridi, L., Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines, 30 (4), 681–694. https://doi.org/10.1007/s11023-020-09548-1
29. Chae, Y., Davidson, T. (2023). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning. https://doi.org/10.31235/osf.io/sthwk
30. Bi, B., Li, C., Wu, C., Yan, M., Wang, W., Huang, S. et al. (2020). PALM: Pre-training an Autoencoding&Autoregressive Language Model for Context-conditioned Generation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.700

31. Schubert, M. C., Wick, W., Venkataramani, V. (2023). Performance of Large Language Models on a Neurology Board–Style Examination. JAMA Network Open, 6 (12), e2346721. https://doi.org/10.1001/jamanetworkopen.2023.46721
32. Chen, L., Chen, P., Lin, Z. (2020). Artificial Intelligence in Education: A Review. IEEE Access, 8, 75264–75278. https://doi.org/10.1109/access.2020.2988510
33. Koto, F., Aisyah, N., Li, H., Baldwin, T. (2023). Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2023.emnlp-main.760
34. Dao, X.-Q. (2023). Performance Comparison of Large Language Models on VNHSGE English Dataset: OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard. arXiv. https://doi.org/10.48550/arXiv.2307.02288
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P. (2022). Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.48550/arXiv.2203.02155
36. Milani Fitria, K. (2023). Information Retrieval Performance in Text Generation using Knowledge from Generative Pre-trained Transformer (GPT-3). Jambura Journal of Mathematics, 5 (2), 327–338. https://doi.org/10.34312/jjom.v5i2.20574
37. Rofiq, M. A., Azhar, A. (2022). Hazards Identification and Risk Assessment In Welding Confined Space Ship Reparation PT. X With Job Safety Analysis Method. BERKALA SAINSTEK, 10 (4), 175. https://doi.org/10.19184/bst.v10i4.32669
38. Bill, D., Eriksson, T. (2023). Fine-Tuning A Llm Using Reinforcement Learning From Human Feedback For A Therapy Chatbot Application. KTH. Available at: https://www.diva-portal.org/smash/get/diva2:1782678/FULLTEXT01.pdf
39. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. arXiv. https://doi.org/10.48550/arXiv.1904.09675