VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i1.5572

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

IMPLEMENTING RETRIEVAL-AUGMENTED GENERATION AND VECTOR DATABASES FOR CHATBOTS IN PUBLIC SERVICES AGENCIES CONTEXT

Ibnu Pujiono[1*]; Irfan Murtadho Agtyaputra[2]; Yova Ruldeviyani[3]

Faculty of Computer Science[1,2,3]
University of Indonesia, Indonesia[1,2,3]
https://www.ui.ac.id[1,2,3]
ibnu.pujiono31@ui.ac.id[1*], irfan.murtadho@ui.ac.id[2], yova@cs.ui.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract*— Rapid developments in information technology, such as chatbots and generative artificial intelligence, have drastically lowered the cost of providing services to the society. This study aims to measure performance of developed chatbot using retrieval augmented generation and vector database. This research compares the performance of existing Large Language Modelling (LLM) in answering questions related to regulations concerning public service agencies.. Using a vector database, questions are assessed and answered by the LLM model, considering cosine similarity scores. The best-performing model, gpt-4, is selected for the deployment process which have average cosine similarity score 0,404. The use of LLM for chatbot creation at the prototyping stage can provide a good response to the question asked related to public service agencies with retrieval augmented generation (RAG) process through regulation-based document extraction.*

Keywords*: chatbot, large language modelling (LLM), public service agencies, retrieval augmented generation (RAG), vector database.*

Intisari— *Perkembangan pesat dalam teknologi informasi, seperti chatbot dan kecerdasan buatan generatif, telah secara drastis menurunkan biaya penyediaan layanan kepada masyarakat umum. Penelitian ini bertujuan untuk membuat chatbot dengan memanfaatkan vector database dan retrieval augmented generation. Penelitian ini membandingkan kinerja Large Language Modelling (LLM) yang sudah ada dalam menjawab pertanyaan yang berkaitan dengan peraturan tentang badan layanan umum. Dengan menggunakan vektor database, pertanyaan-pertanyaan dinilai dan dijawab oleh model LLM, dengan mempertimbangkan cosine similarity score. Model dengan performa terbaik, gpt-4, dipilih untuk proses deployment dengan nilai cosine similarity rata-rata sebesar 0.404. Penggunaan LLM untuk pembuatan chatbot pada tahap prototyping dapat memberikan respon yang baik terhadap pertanyaan yang diajukan terkait badan layanan umum dengan proses retrieval augmented generation (RAG) melalui ekstraksi dokumen berbasis regulasi.*

Kata Kunci*: chatbot, pemodelan bahasa besar (LLM), lembaga layanan publik, generasi augmented retrieval (RAG), basis data vektor.*

INTRODUCTION

Based on recent study on strategic technology trends in 2024, Gartner [1] noted ten technology trend which can support business process and generate competitive advantage, including intelligent application. This technology will become the benchmark for enhancing the decision-making process in real-time. In its report titled "Gartner Hype Cycle for Emerging Technologies", Gartner [2] highlights that while generative AI has enormous potential to facilitate competitive differentiation, several other emerging AI techniques also offer enormous potential to improve digital customer experiences, make better business decisions, and set business apart from competitors.

The rapid advancements in information technology have significantly reduced the cost of service delivery to the general public. The era is marked by constantly evolving technology that not only enhances human capabilities but also creates opportunities for innovation in public service delivery. According to Kasali [3], everyone who is involved, including the government, needs to be concerned about the disruptive era. In this instance, the government is obligated to use emerging technology to deliver services to the community. In this regard, the government's task is to use existing technological advancements to provide the people with better services. The increasing complexity and volume of services provided by government agencies to the general public necessitates the development of solutions that can improve process efficiency and effectiveness. In this context, chatbots emerge as innovative and smart solutions. As human representatives, chatbots can do more than just provide automated responses, they also can modify how we interact with public services [4].

Large Language Modeling is used in a study to analyze a dataset of legal questions about Palestinian cooperatives. The final chatbot achieved an 82% accuracy rate and a 79% F1 score based on performance metric. Researchers can now offer precise and trustworthy legal support thanks to this work [5]. Another study assesses the chatbot's reaction to inquiries about the treatment of myopia. ChatGPT-3.5, ChatGPT-4.0, and Google Bard are the models of chatbots that are compared. The result then asses by SME's and GPT-4 give the best response based on the final assessment [6]. A LLM that has been trained using knowledge-based context can provide a more insightful response that is referenced to the training documents [7], [8], [9].

Based on several previous research, no one has used the large language model and vector database in the public service agency financial management sector. In light of the gaps in several earlier research studies and problems raised in the organization, the author created the following research question to guide the development of this study:

How Large Language Modelling and Vector Database used to develop a helpful chatbot for financial advisors in the Public Service Agencies context?

This research aims to develop a chatbot model that will support the process of utilizing existing data at the Directorate of Financial Management Development of Public Service Agencies. The model can also serve as an additional assistant for the financial management coaching process carried out by financial advisors, facilitating BLU financial management with greater ease. This research proposes LLM performance measurement to address regulatory-based inquiries on BLU financial management. The technique strengthens the study of the retrieval augmented generation (RAG) model by using vector databases as the foundation for semantic search in addition to LLM. The cosine-similarity score will be utilized to compare the final models.

MATERIALS AND METHODS

A. Related Study

Based on a previous study, a chatbot trained with document was created to assist Palestinian cooperatives with their legal needs. They analyzed and comprehended complex legal terms and circumstances by using massive language models and natural language processing. After being tested on a dataset of legal inquiries about Palestinian cooperatives, the chatbot received an 82% accuracy rate and a 79% F1 score. The primary conclusion of this study is that given the pressing need for prompt assistance and the labor-intensive effort involved in responding to legal inquiries, cooperatives, and their members can benefit from the usage of chatbots such as this one to provide accurate and dependable legal support. Additionally, the authors emphasize how natural language processing and massive language models have the potential to change how we engage with legal writing and improve the accessibility of legal support and information [5].

Another research wrote an efficient and effective end-to-end mechanism for retrieving building regulatory questions. It presents a framework that integrates deep learning and information retrieval techniques and analyzes relevant research on question answering, extraction, and regulation modeling. The research underscores the constraints of conventional information extraction (IE) techniques while stressing the possibilities of deep learning and ontology-based information extraction (OBIE) in automating the processing and examination of regulatory documents. Comparing several approaches to responding to building regulation queries—such as using a regulatory document database, the Baidu search engine, and the suggested question-answering system (QAS4CQAR) [10].

A comparative analysis of chatbot models was also conducted to determine the chatbot response when given questions related to myopia care. The chatbot models compared include ChatGPT-3.5, ChatGPT-4.0, and Google Bard. Thirteen frequently asked questions about myopia care were selected, and they were divided into six domains: prognosis, treatment and prevention, clinical presentation, diagnosis, risk factors, and causation. Three consultant-level pediatric ophthalmologists evaluated each question independently from the LLMs using a three-point accuracy scale (bad, borderline, good). Results highlight how LLMs, in particular ChatGPT-4.0, can provide precise and in-depth answers to questions about myopia [6].

B. Research Method

This research is conducted by comparing the performance of existing LLM models to answer the concept of questions in regulations related to public service agencies (BLU). The data used for this initial experiment is the Minister of Finance Regulation Number 129 of 2020 concerning Guidelines for the Management of Public Service Agencies [11]. The document will be uploaded using a document parser and then stored in the form of a vector database which will later be used in answering the questions posed using the existing LLM model.

The performance of the LLM will be compared using the cosine similarity score of the document used as the knowledge base. The next step is to do fine-tuning by asking questions again after which the best model that has the largest score will be used in the deployment process. The steps can generally be written down as follows:

1. Data Cleansing

Data in the form of PDF will be uploaded using the help of a PDF parser. Before the upload process, the document will first be cleaned to remove irrelevant words and characters related to regulations such as the initials at the end of each page and footnotes. The documents that will be used in this study will first undergo a thorough cleansing process before moving on to the subsequent step. The data purification process that will be used in this research will be aided by the use of a library called “Indonesian Regulation Text Parser” [12].

2. Text Splitting

After data cleansing process, the data will be divided into a chunk document using a text-splitter. This is done to facilitate the semantic search process that will be carried out later. After the data is divided into chunks, it will be embedded using OpenAI embedding to convert it into vector form.

3. Embeeding an Storing in Vector Database

To become ready for semantic search, document chunk embeddings are created. The semantic meaning of text can be captured mathematically via embeddings, which makes it possible to identify related text chunks quickly and precisely. For a word in a multidimensional semantic field, its semantic vector expresses it as a point. Word representation using vectors is commonly referred to as embedding. For the purpose of analyzing the semantic components of language, embedding provides an efficient means of representing word meaning and makes calculations in models and pipelines easier [13].

The vectors generated from the embedding process will be stored in a vector database. Vector database is slightly different from other databases that store data in rows and columns, vector database stores it in the form of numerical vectors that are specifically designed for efficient similarity search and retrieval based on semantic meaning. The vector database in this research will be assisted using a cloud database called Pinecone.

4. Query Context and Train LLM

After the vector is stored in the database, the large language model will be used to understand the context of the vectorized document. Large language modeling (LLM) is a field of artificial intelligence (AI) that involves training massive neural networks on large amounts of text data. These neural networks can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way [14].

The LLM model used in this research is provided by Open AI. 4 models will be used, namely: davinci-002, gpt-3.5-turbo, babbage-002 [15]and gpt-4 [16]. These models will be used to respond to queries from existing documents using the augmented generation retrieval method. The problem of out-of-date information can be mitigated by arming LLMs with an appropriate toolkit tailored to the assigned work. Retrieval Augmented Generation (RAG) is the collection of techniques created to supplement LLM input with information that has been collected, such pertinent tools. Tool retrieval, plan generation, and execution are the three main parts of RAG [17]. Retrieval-augmented generation (RAG) is a natural language processing technique that enhances the quality and relevance of generated text by fusing generative and retrieval models. Retrieval models retrieve specific information from a knowledge base using semantic search [18]. When using both parametric and non-parametric memory, paranormal experiences or chatbot hallucination are typically lessened and tasks such as summarizing and answering questions are more interpretable [19].

Retrieval models retrieve specific information from a knowledge base using semantic search search so that it will reduce query time [20]. Semantic search refers to a broad range of

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.5572

approaches that aim to investigate the query's context to drive (and enhance) search relevancy. These strategies employ several methodologies that go beyond simple string matching. These include using knowledge graphs and related technologies, accounting for lexical variants, incorporating taxonomy (hierarchy and synonymy), contextualizing searches using location parsing and past searches, and so on [21].

5. Query Question

After making LLM understand the context of the document, a query will be made by asking questions related to the uploaded document. The questions are customized with some general provisions in the document. The question is predicated based on the context found in Minister of Finance Regulation Number 129 of 2020 about Management Guidelines for Public Service Agencies [11].

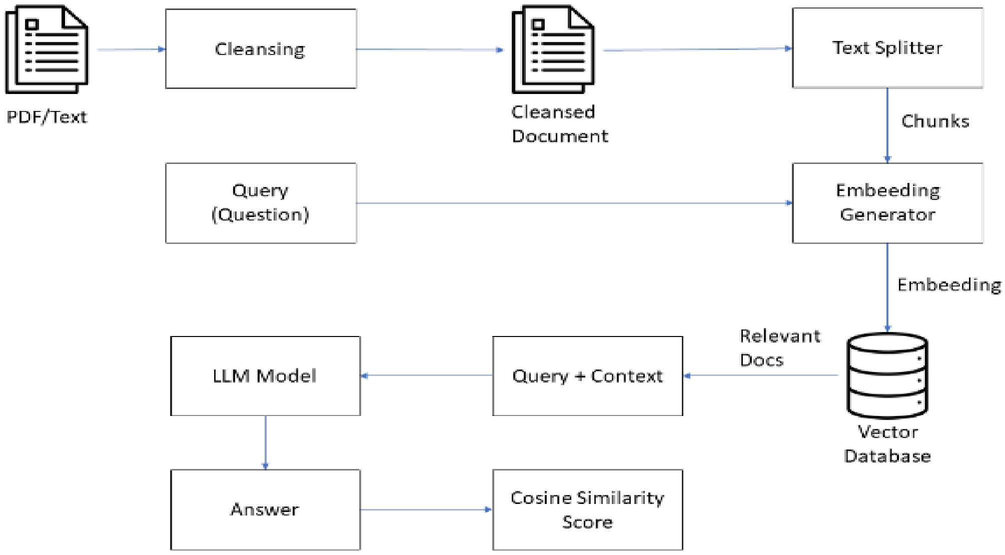
6. Measuring Performance

After providing questions to the trained LLM model, the LLM model will provide a response based on the knowledge base that has been provided. After that, the successfully generated response will be compared with the provisions in the existing regulatory documents to compare the level of similarity. The performance measurement process is carried out using the cosine similarity score.

Subsequently, the vector that was successfully queried will be trained with the current LLM model, and the cosine similarity score will be compared to determine which model is optimal for the deployment procedure. When compared to the cosine similarity, other similarity metrics scored worse. More specifically, when subjects overlap high-frequency phrases, the cosine similarity metric estimates similarity better, but the word embedding measure better captures the semantic ties among topics. It looks like these two measures work well together. Additionally, the results reveal word-weighing strategies and n-gram document representations that maximize the effectiveness of every similarity metric at a specific intertextual similarity level [22]. So, cosine similarity is used to measure performance from response generated by trained LLM model with light and heavy review context.

7. Deployment

After obtaining the LLM model that provides the most appropriate response to the existing document, the model will be used to become the engine driving the existing BLU chatbot. For this project, python has been chosen as the programming language since it offers developers flexibility and strong functionality to accommodate a variety of skill levels. Web development, data analytics, information science, artificial intelligence, machine learning, and many other fields are among the many fields in which Python is widely used [23]. The prototyping stage of making a chatbot will be made with the help of the Python-based Streamlit library



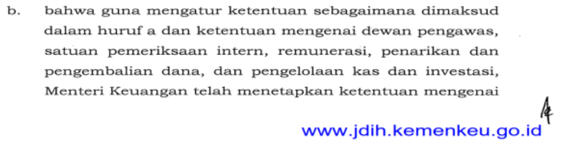
Source : (Research Result, 2024)

Figure 1 Research Workflow

RESULTS AND DISCUSSION

1. Data Cleansing

Data in the form of pdf files is uploaded and then a cleansing process is carried out to remove unnecessary characters such as character and watermark. Following the cleaning process, the document will be uploaded to the Langchain document parser.



b. bahwa guna mengatur ketentuan sebagaimana dimaksud dalam huruf a dan ketentuan mengenai dewan pengawas, satuan pemeriksaan intern, remunerasi, penarikan dan pengembalian dana, dan pengelolaan kas dan investasi, Menteri Keuangan telah menetapkan ketentuan mengenai

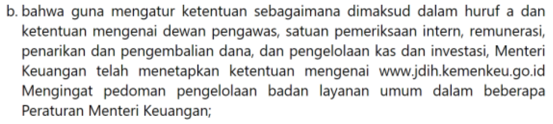
4

www.jdih.kemenkeu.go.id

Source : (Minister of Finance Regulation Number 129 of 2020, 2020)

Figure 2. Example of Uncleaned Data

Figure 2 shows uncleaned data from original document that still contain some unnecessary character used for the embedding process.



b. bahwa guna mengatur ketentuan sebagaimana dimaksud dalam huruf a dan ketentuan mengenai dewan pengawas, satuan pemeriksaan intern, remunerasi, penarikan dan pengembalian dana, dan pengelolaan kas dan investasi, Menteri Keuangan telah menetapkan ketentuan mengenai www.jdih.kemenkeu.go.id Mengingat pedoman pengelolaan badan layanan umum dalam beberapa Peraturan Menteri Keuangan;

Source : (Research Result,2024)

Figure 3. Example of Cleaned Data

Figure 3 shows cleaned data by using cleansing tool that mention before to remove unnecessary character and then used for embeeding and text splitting process.

2. Text Splitting

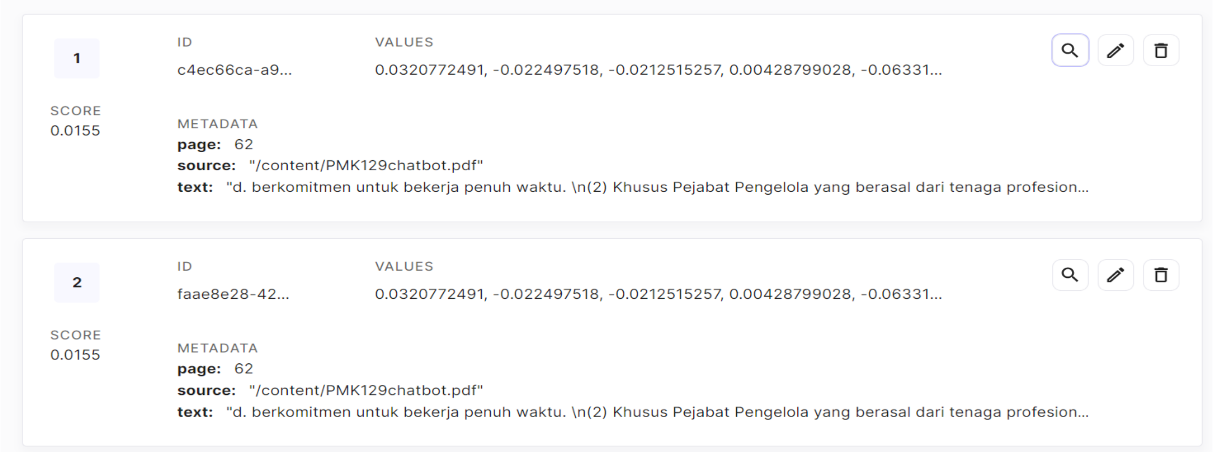
It is necessary to disassemble the experiment's document into manageable pieces and import them into LangChain. By modifying two crucial factors, the text splitting's granularity can be controlled:

a) Chunk Size: The maximum character count that each chunk can have is determined by this option.
b) Chunk Overlap: This option specifies how many characters should flow over into the space between two chunks that are next to one other.

It is possible to modify the text-splitting process's settings to suit our particular needs, whether we need a coarser segmentation with bigger pieces or a finer segmentation with numerous smaller chunks.

3. Creating Embedding and Vector Database

Database creation is assisted by the use of tools, namely Pinecone. Pinecone is a cloud-based server that will be used to store vector databases from the results of the embedding process. Since a vector database uses embeddings rather than scalar data, it functions differently from typical databases. Pinecone optimizes similarity search using approximate nearest neighbor search (ANN) methods. It makes data management activities like updates, deletions, and insertions simpler, which can make it difficult to use standalone indexes like Facebook AI Similarity Search (FAISS).



1
SCORE 0.0155
ID c4ec66ca-a9...
VALUES 0.0320772491, -0.022497518, -0.0212515257, 0.00428799028, -0.06331...
METADATA
page: 62
source: "/content/PMK129chatbot.pdf"
text: "d. berkomitmen untuk bekerja penuh waktu. \n(2) Khusus Pejabat Pengelola yang berasal dari tenaga profesion...

2
SCORE 0.0155
ID faae8e28-42...
VALUES 0.0320772491, -0.022497518, -0.0212515257, 0.00428799028, -0.06331...
METADATA
page: 62
source: "/content/PMK129chatbot.pdf"
text: "d. berkomitmen untuk bekerja penuh waktu. \n(2) Khusus Pejabat Pengelola yang berasal dari tenaga profesion...

Source : (Research Result,2024)

Figure 4. Embedded Text in Vector DatabaseQuery Question

The appearance from vector database produced from embedded word that can be used for easier querying context from the question asked to chatbot can be seen in Figure 4. This data is stored using Pinecone, which is one of vector database cloud provider.

Based on the 5W+1H principle, these draft questions were developed to assess the chatbot's ability to answer questions. This inquiry is predicated on the context found in Minister of Finance Regulation Number 129 of 2020. The appendix of questions can be seen in Table 1.

Table 1. Draft Question

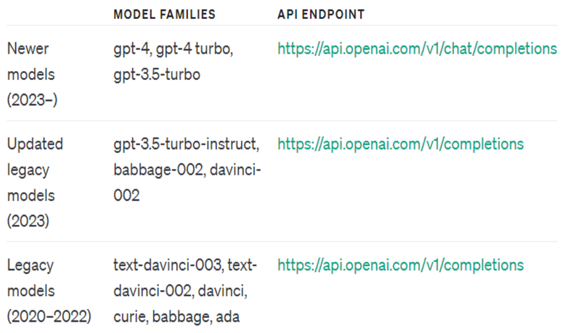
Question	Answer
What is a Public Service Agency?	article 1
Where are the supervisory board meetings physically held?	article 220
How are BLU service rates determined and what aspects should be considered?	article 32
What is the maximum implementation period for Land and Building KSO?	article 146
Who are the BLU management officials?	article 196

Source : (Research Result,2024)

4. LLM Performance Measurement

The LLM model that will be used in this research uses the text-generation model provided by OpenAI. Based on its proven capabilities, OpenAI's Large Language Model was selected as a strategic choice. With its vast knowledge base and ability to comprehend natural language, the model from OpenAI is the pinnacle of huge language model development. Its primary strength is its capacity to produce text that is coherent and appropriate to the context, which makes it an invaluable tool for a wide range of applications, including content creation and chatbots. Furthermore, the OpenAI Model can be customized to certain tasks and domains thanks to its fine-tuning-based adaptability. Its popularity is also influenced by the developer community's broad adoption and support. OpenAI Model was chosen mostly because of its track record of producing text creation that is human-like and because of its ability to improve a broad range of language-related activities.

The list of models to be used are davinci-002, gpt-3.5-turbo, babbage-002, and gpt-4. The comparison between the models used in this study can be seen in the figure below :



	MODEL FAMILIES	API ENDPOINT
Newer models (2023–)	gpt-4, gpt-4 turbo, gpt-3.5-turbo	https://api.openai.com/v1/chat/completions
Updated legacy models (2023)	gpt-3.5-turbo-instruct, babbage-002, davinci-002	https://api.openai.com/v1/completions
Legacy models (2020–2022)	text-davinci-003, text-davinci-002, davinci, curie, babbage, ada	https://api.openai.com/v1/completions

Source : OpenAI [24], 2023

Figure 4. List of Open AI Model

$$\cos(x,y) = \frac{x.y}{||x|.||y||} \qquad (1)$$

Equation 1 is cosine similarity measurement where x.y is the inner product between vector x and y, and ||x|| and ||y|| are the L2 norms or Euclidean lengths of vector x and y, respectively [7]. The results of the comparison between these models can be seen in Table 2 which describes the cosine similarity score from text generated compared to the general context in the document that was uploaded to the vector database before.

As shown from Table 2, results of the cosine similarity score assessment concluded that the gpt-4 model has superior performance compared to other models with average score of 0,404 out of 5 question answered based on document context :

Table 2. Trained LLM Cosine Similarity Score

	davinci-002	gpt-3,5-turbo	babbage-002	gpt-4
Question 1	0,51	0,45	0,50	0,57
Question 2	0,12	0,24	0,12	0,29
Question 3	0,83	0,94	0,53	0,93
Question 4	0	0	0,40	0
Question 5	0,23	0,23	0,23	0,23
Average	0,338	0,372	0,356	0,404

Source : (Research Result,2024)

Another concern is the ability of vector embedding from existing documents because question 5 gets the same answer from each model. For this reason, a better data vector cleansing process is needed before being used as a knowledge base for the development of the BLU chatbot at the deployment stage.

From the results of the cosine similarity score assessment, it is concluded that the gpt-4 model has superior performance compared to other models. Another concern is the ability of vector embedding from existing documents because question 5 gets the same answer from each model. For this reason, a better data vector cleansing process is needed before being used as a knowledge base for the development of the BLU chatbot at the deployment stage.

The process of guiding the financial management of public service agencies is one of the most challenging tasks for financial advisors. Problems that are quite complex and sometimes unpredictable require financial advisors to continue to upgrade their knowledge. The use of chatbots is expected to be a solution to this problem so that services to stakeholders, especially financial managers in public service agencies, can be fulfilled properly and are not bound by service time in general.

From the results of measuring model performance using the model owned by OpenAI, it was found that the gpt-4 model was the best model to use in answering questions related to the financial management of public service agencies based on Minister of Finance Regulation Number 129 of 2020. There are still several problems related to this chatbot model, including inconsistencies between models, quite expensive rates, and infrastructure readiness in the internal environment of the organization.

The application of this chatbot is only one step in solving the existing problems. This research is also in line with several related studies [5], [8] which prove that the LLM model combined with Retrieval-augmented Generation (RAG) can produce satisfactory answer performance compared to commonly used chatbots due to the unique nature of knowledge that may not have been used as training data in commonly used models, for example in GPT Chat and Bing Chat.

CONCLUSION

The results of this initial development present that the gpt-4 model from OpenAI has better performance than other models. In addition, the text generated from the question given during testing is considered to be more representative of the actual answer to the question itself. The fast response performance (under 5 seconds) is also the reason why the gpt-4 model is the best compared to other models. The development of the chatbot has reached the prototype stage and will be retested by involving experts in public service agencies to get a comparative assessment of the results of the previous cosine similarity score. Future development needs to pay attention to a better document embedding process so that the data that will be entered into the vector database which will be used as the basis for the LLM semantic search model can produce even better performance.

This research only uses the model provided by OpenAI. In the future, the use of other LLM models can be used for a more comprehensive performance comparison. Meta's llama-70b model and Google's PaLM 2 model can be used as a reference to compare the performance of existing models to answer specific questions related to public service agencies (BLU). The model chosen in further research is also common and comes from a large company that is very credible so it is hoped that later it can enrich the choice for the chatbot development process, especially at the Directorate of Financial Management Development of Public Service Agencies. Furthermore, the cosine-similarity score is the only metric used in this study's performance evaluation of the LLM model. Future model performance measurements may make use of other metrics like precision and accuracy. Opinions from subject matter experts can also be employed in the process of evaluating the performance of LLM models, providing a more thorough understanding that is necessary to determine which LLM model performs best.

REFERENCE

[1] "Gartner Top 10 Strategic Technology Trends 2024," 2024, [Online]. Available: https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2024

[2] L. Perri, "What's New in the 2023 Gartner Hype Cycle for Emerging Technologies." Accessed: Nov. 12, 2023. [Online]. Available: https://www.gartner.com/en/articles/what-s-new-in-the-2023-gartner-hype-cycle-for-emerging-technologies#:~:text=What's New in the 2023 Gartner Hype Cycle for Emerging Technologies&text=They fit into four main,human-centric security and privacy.

[3] R. Kasali, *Disruption*. Jakarta: Gramedia, 2017.

[4] T. Chen, M. Gascó-Hernandez, and M. Esteve, "The Adoption and Implementation of Artificial Intelligence Chatbots in Public Organizations: Evidence from U.S. State Governments," *Am. Rev. Public Adm.*, vol. 54, no. 3, pp. 255–270, 2024, doi: 10.1177/02750740231200522.

[5] R. Qasem, B. Tantour, and M. Maree, "Towards the Exploitation of LLM-based Chatbot for Providing Legal Support to Palestinian Cooperatives," arXiv preprint arXiv:2306.05827, 2023, doi: https://doi.org/10.48550/arXiv.2306.05827

[6] Z. W. Lim *et al.*, "Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard," *eBioMedicine*, vol. 95, p. 104770, 2023, doi: 10.1016/j.ebiom.2023.104770.

[7] M. Maryamah, M. M. Irfani, E. B. Tri Raharjo, N. A. Rahmi, M. Ghani, and I. K. Raharjana, "Chatbots in Academia: A Retrieval-

Augmented Generation Approach for Improved Efficient Information Access," *KST 2024 - 16th Int. Conf. Knowl. Smart Technol.*, pp. 259–264, 2024, doi: 10.1109/KST61284.2024.10499652.

[8] U. Shukla, S. Singh, A. Pundir, and G. J. Saxena, "Large language model based framework for knowledgebase coverage and correctness using chatbot and human feedback," *2023 IEEE 7th Conf. Inf. Commun. Technol. CICT 2023*, pp. 1–7, 2023, doi: 10.1109/CICT59886.2023.10455408.

[9] M. Dean, R. R. Bond, M. F. McTear, and M. D. Mulvenna, "ChatPapers: An AI Chatbot for Interacting with Academic Research," *2023 31st Irish Conf. Artif. Intell. Cogn. Sci. AICS 2023*, pp. 1–7, 2023, doi: 10.1109/AICS60730.2023.10470521.

[10] B. Zhong, W. He, Z. Huang, P. E. D. Love, J. Tang, and H. Luo, "A building regulation question answering system: A deep learning methodology," *Adv. Eng. Informatics*, vol. 46, no. October, p. 101195, 2020, doi: 10.1016/j.aei.2020.101195.

[11] M. Keuangan, *Peraturan Menteri Keuangan Nomor 129 tahun 2020 tentang Pedoman Pengelolaan Badan Layanan Umum*. 2020.

[12] Maziyank, "Indonesian Regulation Text Parser." Accessed: Feb. 23, 2024. [Online]. Available: https://github.com/maziyank/anali%0Asa-regulasi

[13] E. Akdemir and N. Barışçı, "A review on deep learning applications with semantics," *Expert Syst. Appl.*, vol. 251, no. December 2021, 2024, doi: 10.1016/j.eswa.2024.124029.

[14] J. W. Rae *et al.*, "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," 2021, [Online]. Available: http://arxiv.org/abs/2112.11446

[15] J. Ye *et al.*, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models," pp. 1–47, 2023, [Online]. Available: http://arxiv.org/abs/2303.10420

[16] OpenAI *et al.*, "GPT-4 Technical Report," vol. 4, pp. 1–100, 2023, [Online]. Available: http://arxiv.org/abs/2303.08774

[17] R. Anantha, T. Bethi, D. Vodianik, and S. Chappidi, "Context Tuning for Retrieval Augmented Generation," *UncertaiNLP 2024 - Work. Uncertainty-Aware NLP, Proc. Work.*, pp. 15–22, 2024, doi: https://doi.org/10.48550/arXiv.2312.05708

[18] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020-Decem, 2020.

[19] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1–17, 2023, doi: 10.1162/tacl_a_00530.

[20] J. Tekli, G. Tekli, and R. Chbeir, *Combining Offline and On-the-fly Disambiguation to Perform Semantic-aware XML Querying*, vol. 29, no. 1. 2023. doi: 10.2298/CSIS220228063T.

[21] D. U. Sinha and M. V. Dubey, "The Technique of Different Semantic Search Engines," *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 1496–1501, 2020, doi: 10.35940/ijrte.a2249.059120.

[22] D. Banerjee, P. Singh, A. Avadhanam, and S. Srivastava, "Benchmarking LLM powered Chatbots: Methods and Metrics," 2023, [Online]. Available: http://arxiv.org/abs/2308.04624

[23] L. G. Gunnell, B. Nicholson, and J. D. Hedengren, "Equation-based and data-driven modeling: Open-source software current state and future directions," *Comput. Chem. Eng.*, vol. 181, no. October 2023, p. 108521, 2024, doi: 10.1016/j.compchemeng.2023.108521.

[24] OpenAI, "Text Generation Model." Accessed: Nov. 11, 2023. [Online]. Available: https://platform.openai.com/docs/guides/tex%0At-generation