Evaluation of Orca 2 Against Other LLMs for Retrieval Augmented Generation

Donghao Huang and Zhaoxia Wang(✉)

School of Computing and Information Systems, Singapore Management University, 80 Stamford Rd, Singapore 178902, Singapore
{dh.huang.2023,zxwang}@smu.edu.sg

Abstract. This study presents a comprehensive evaluation of Microsoft Research's Orca 2, a small yet potent language model, in the context of Retrieval Augmented Generation (RAG). The research involved comparing Orca 2 with other significant models such as Llama-2, GPT-3.5-Turbo, and GPT-4, particularly focusing on its application in RAG. Key metrics, included faithfulness, answer relevance, overall score, and inference speed, were assessed. Experiments conducted on high-specification PCs revealed Orca 2's exceptional performance in generating high quality responses and its efficiency on consumer-grade GPUs, underscoring its potential for scalable RAG applications. This study highlights the pivotal role of smaller, efficient models like Orca 2 in the advancement of conversational AI and their implications for various IT infrastructures. The source codes and datasets of this paper are accessible here (https://github.com/inflaton/Evaluation-of-Orca-2-for-RAG.).

Keywords: Large Language Model (LLM) · Generated Pre-trained Transformer (GPT) · Retrieval Augmented Generation (RAG) · Question Answering · Model Comparison

1 Background and Introduction

In the realm of artificial intelligence, Large Language Models (LLMs) like GPT-4 [1] have revolutionized how machines understand and process human language. These models, characterized by their vast parameter counts and deep learning capabilities, excel in generating human-like text and comprehending complex language nuances. The emergence of LLMs has opened new avenues in various AI applications, one of which is Retrieval-Augmented Generation (RAG) [3,5,6,8].

RAG emerges as a promising solution in the quest for enhancing generative tasks, particularly in professional knowledge-based question answering [5,6,12]. The integration of external knowledge through RAG not only addresses some challenges faced by LLMs, such as hallucination and outdated knowledge, but also facilitates accurate responses in knowledge-intensive tasks [6].

The integration of LLMs into RAG systems marks a significant milestone. LLMs enable RAG systems to process and respond to conversational queries

Z. Wang and C. W. Tan (Eds.): PAKDD 2024 Workshops, LNAI 14658, pp. 3–19, 2024.
https://doi.org/10.1007/978-981-97-2650-9_1

with a level of sophistication and relevance previously unattainable. This integration allows for a more intuitive and user-friendly interface, making information retrieval a seamless and interactive experience [6,7].

While LLMs exhibit impressive capabilities, they often generate fictitious responses [8]. Chen et al. assessed the impact of RAG on LLMs, illuminating challenges and underscoring the need for further advancements in applying RAG to enhance LLM performance [3]. Simultaneously, the role of smaller yet efficient language models, such as Orca 2 [10], has garnered recent attention. In a landscape dominated by large models, the growing interest in the efficacy of smaller models, particularly in RAG applications, is becoming a notable area of investigation [10,11].

In this research paper, we delve into the integration of Orca 2 [10], a groundbreaking smaller language model developed by Microsoft Research, into RAG systems. Orca 2 represents a significant shift in artificial intelligence, characterized by its smaller size but remarkably powerful language processing abilities. This integration promises to significantly enhance RAG systems by offering advanced language understanding and reasoning capabilities with considerably reduced computational demands.

The paper makes the following key contributions:

1) This research provides a comprehensive evaluation of Microsoft Research's Orca 2 in the context of Retrieval Augmented Generation (RAG). This includes a detailed comparison with other significant language models such as Llama-2, GPT-3.5-Turbo, and GPT-4.
2) The research assesses key metrics, including faithfulness, answer relevance, overall score, and inference speed. This detailed evaluation aims to provide a nuanced understanding of Orca 2's performance in generating responses within the conversational setting of RAG.
3) The research underscores the potential of Orca 2 for scalable RAG applications, challenging the conventional belief that larger models are necessary for achieving sophistication in conversational AI. This offers insightful contributions to the field of AI, particularly in understanding Orca 2's role within it.
4) The research positions Orca 2 as a smaller, efficient model that plays a pivotal role in advancing conversational AI. By highlighting its adaptability and performance benefits, the paper contributes to discussions on the evolving landscape of language models and their applications.

2 Related Work

The landscape of language models is rapidly evolving, with advancements in LLMs driving extensive research and exploration of their capabilities across diverse applications [4]. Notably, GPT-4 has garnered attention for its extensive parameter count and language comprehension capabilities, setting the stage for exploring the potential of smaller yet powerful language models in specific applications [13].

Liu et al. proposed ChatQA, a family of conversational question-answering models achieving GPT-4 level accuracies through a two-stage instruction tuning method [9]. Utilizing a fine-tuned dense retriever on a multi-turn QA dataset, ChatQA-70B outperforms GPT-4 in average score on 10 conversational QA datasets without relying on synthetic data from OpenAI GPT models [9].

RAG represents a promising approach within the field of LLMs, enhancing generative tasks by combining information retrieval and language generation techniques [5,6,12]. RAG involves retrieving relevant information or passages from documents or knowledge sources and generating responses based on the retrieved content, aiming to enhance the quality and informativeness of generated outputs [5].

Facing challenges like hallucination and outdated knowledge, LLMs find a potential solution in RAG, which integrates external knowledge, improving accuracy for knowledge-intensive tasks. Gao et al. conducted a comprehensive review exploring the evolution of RAG paradigms, scrutinizing its tripartite foundation, and introducing metrics [6].

As LLMs and RAG gain prominence in professional knowledge-based question answering, Lin explores the impact of PDF parsing accuracy on RAG effectiveness [7]. An Automated RAG Evaluation System named ARES utilizes synthetic training data to fine-tune lightweight language models for assessing RAG systems in terms of context relevance, answer faithfulness, and answer relevance [12]. ARES effectively evaluates RAG systems across diverse knowledge-intensive tasks with minimal human annotations, demonstrating accuracy even after domain shifts in queries and documents. ARES and Retrieval Augmented Generation Assessment (RAGAS) contribute to the evaluation and assessment of RAG systems, streamlining the process and reducing reliance on human annotations [5,12].

Despite the impressive capabilities of LLMs, they tend to generate fictitious responses [8]. Chen et al. evaluated the impact of RAG on LLMs, highlighting the challenges in LLMs and suggesting a need for further advancements in applying RAG to LLMs [3].

The role of smaller yet efficient language models, like Orca 2 [10], has been a subject of recent investigation. While large models dominate the scene, the efficacy of smaller models, particularly in RAG applications, is a growing area of interest [10,11]. Microsoft Research's Orca 2 introduces a new paradigm with its smaller size and potent language processing capabilities. This study distinguishes itself by comprehensively evaluating Orca 2's performance against established LLMs in the specific context of RAG, shedding light on its potential contributions to the field.

3 Methodology

3.1 Workflow Overview

In the assessment of Orca 2 against other LLMs for Retrieval Augmented Generation (RAG), a methodology is employed that involves leveraging various pre-trained LLMs. This is achieved through strategic prompting and the application of these models to contextual private data. As shown in Fig. 1, the workflow is structured into three distinct phases, ensuring a comprehensive and systematic evaluation process:

1. **Data Preprocessing/Embedding:** This initial phase involves storing private documents, typically PDFs, for later use. The documents are broken down, run through an embedding model, and their embeddings are saved in a vector store.
2. **Prompt Construction/Retrieval:** In response to user queries, the system formulates a set of prompts for the language model. These prompts are crafted by merging a template with relevant document extracts from the vector store, with the addition of standalone questions based on existing chat history for enhanced retrieval.
3. **Prompt Execution/Inference:** The final stage involves submitting the prepared prompts to a pre-trained language model for processing. This stage utilizes both exclusive model APIs and accessible or in-house models.

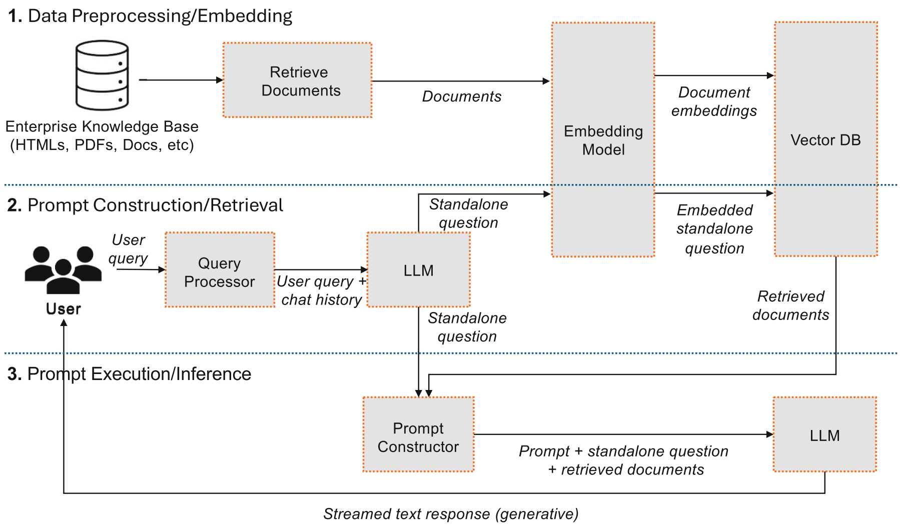

Fig. 1. Overall Workflow

3.2 Gathering and Processing Data

We have carefully curated a dataset from a variety of real-world professional documents pertaining to the Payment Card Industry Data Security Standard (PCI DSS). This standard comprises a comprehensive set of security measures devised by the PCI Security Standards Council to safeguard sensitive payment card information. Our collection process targeted all 13 PDF documents associated with the most recent iteration, PCI DSS version 4.0, released on March 31, 2022, obtained directly from the PCI Security Standards Council's official website[1]. Subsequently, these documents were processed through text extraction, segmentation, and embedding procedures utilizing the LangChain framework[2] and the HuggingFace Instructor text embedding model[3], culminating in the generation of embeddings. These were then efficiently stored locally via FAISS[4], an open-source library for vector search created by Meta, facilitating streamlined document retrieval. The complete source code[5] for this procedure is available in our code repository.

3.3 Assessment and Comparative Evaluation of Orca 2 Against Other LLMs

In our quest to evaluate the effectiveness of various LLMs for Retrieval Augmented Generation (RAG), we zeroed in on Orca 2 for a detailed analysis. To ensure a comprehensive evaluation, we included comparisons with other prominent models in the field, such as Llama-2, GPT-3.5-Turbo, and GPT-4 from OpenAI. This selection of diverse yet advanced models allows us to conduct a thorough assessment, particularly focusing on how Orca 2's unique attributes and capabilities align with the demands and nuances of RAG applications.

3.4 Assessment Criteria

In our experimental evaluation of Large Language Models (LLMs) for Retrieval Augmented Generation (RAG) applications, our focus are on assessing both Generation Quality and Inference Speed.

Generation Quality. About the Generation Quality, we focused on the key metrics below:

1. **Faithfulness:** This metric assesses the model's responses for factual consistency within the provided context. The evaluation is based on how well the response aligns with the context, rated on a scale from 0 to 1, where a higher score indicates better factual alignment. To determine the score, we identify a series of claims within the generated answer. Each claim is then cross-referenced against the given context to determine if it can be logically derived from it.
   The formula for calculating the faithfulness score is as follows [2,5]:

$$FS = \frac{Number\ of\ contextually\ supported\ claims\ in\ the\ response}{Total\ number\ of\ claims\ in\ the\ response}$$

[1] https://www.pcisecuritystandards.org/document_library/.
[2] https://github.com/langchain-ai/langchain.
[3] https://huggingface.co/hkunlp/instructor-large.
[4] https://ai.meta.com/tools/faiss.
[5] https://github.com/inflaton/Evaluation-of-Orca-2-for-RAG/blob/main/ingest.py.

where FS represents Faithfulness Score. In this process, we employ GPT-4-Turbo to facilitate the identification and verification of claims.

2. **Answer Relevance:** This metric evaluates the relevance of the generated response to the initial prompt. It examines whether the answer is comprehensive and devoid of extraneous information. Scores are calculated on a scale from 0 to 1, based on the question and answer, where higher values denote greater relevance.
   A response is deemed relevant when it addresses the question directly and fittingly. Our relevance assessment emphasizes penalizing answers that are either not exhaustive or that include unnecessary details, rather than assessing factual correctness. To compute this score, the GPT-4-Turbo language model is engaged to generate questions from the given answer multiple times. The mean cosine similarity between these questions and the original question is then measured. The formula for calculating the answer relevance score is as follows [5]:

$$ARS = \frac{\sum cosine\_similarity(generated\ question,\ original\ question)}{Number\ of\ generated\ questions}$$

   where ARS represents Answer Relevance Score. This process is predicated on the idea that if the answer adequately addresses the initial question, GPT-4-Turbo should be able to generate questions from the answer that are substantially similar to the original question.

3. **Overall Score:** This metric is the harmonic mean of the faithfulness score and the answer relevance score. It provides a balanced measure of the quality of generated answers, accounting for both the fidelity of the response to factual content and its relevance to the original question. Higher scores indicate better overall performance in producing accurate and relevant answers. The formula for calculating the overall score is as follows:

$$Overall\ Score = \frac{2 \times Faithfulness\ Score \times Answer\ Relevanc\ Score}{Faithfulness\ Score + Answer\ Relevanc\ Score}$$

Inference Speed. The Inference Speed of a LLM refers to how quickly the model can process and generate outputs in response to input data or queries. It measures the speed at which the model can make predictions or generate language-based outputs during inference, which is the phase where the model is applied to new, unseen data. A higher inference speed indicates that the model can process information more quickly, making it more efficient for real-time applications and tasks.

The formula for calculating the inference speed is as follows:

$$IS = \frac{Total\ number\ of\ tokens\ (words\ or\ pieces\ of\ words)\ generated}{Total\ inference\ time}$$

where IS represents Inference Speed. These metrics, especially the first three based on the generation RAGAS [5] scores, were essential in comparing Orca 2's performance against other LLMs like Llama-2 and OpenAI models in our RAG scenarios. They provided a detailed assessment of each model's capability in generating accurate, relevant, and timely responses.

3.5 Experiment Setup

Our study meticulously explored the functionality of Orca 2 across various RAG settings. In assessing the proficiency of LLMs within these RAG scenarios, we crafted a series of inquiries focusing on the PCI DSS standards:

1. What's PCI DSS?
2. Can you summarize the changes made from PCI DSS version 3.2.1 to version 4.0?
3. new requirements for vulnerability assessments
4. more on penetration testing

To automate the assessment process, we crafted a specialized Python script designed to simulate conversational interactions with a RAG system. The Python script[6] leverages the LangChain's ConversationalRetrievalChain[7], a framework designed for generating conversations based on documents that have been retrieved. This particular chain processes the chat history (a series of messages) and incoming queries to produce responses. The operational algorithm of this chain is segmented into three distinct phases:

1. It synthesizes a "standalone question" using both the chat history and the new query. If no previous chat history exists, the standalone question remains identical to the new query. If there is existing chat history, however, both the history and the new query are submitted to an LLM, which then generates the standalone question. This method ensures the question is contextually rich enough for effective document retrieval, yet free from unnecessary information that could impede the process.
2. The formulated standalone question is then fed into a retrieval mechanism. This mechanism employs the Hugging Face Instructor model to create embeddings, followed by utilizing FAISS for a similarity search within the local data storage, as outlined in Subsect. 3.2, to pinpoint pertinent documents.
3. Finally, the retrieved documents along with the standalone question are submitted to an LLM, which then generates the conclusive response.

Despite the limitations of a small dataset consisting of only four queries and 13 PDF documents, the study demonstrated the possibility for meticulous system refinement. This underlines the ability of the systems to obtain substantial insights from constrained datasets, showcasing their robustness and adaptability.

[6] https://github.com/inflaton/Evaluation-of-Orca-2-for-RAG/blob/main/qa_chain_test.py.

[7] http://tinyurl.com/LCConversationalRetrievalChain.

To further explore the intricacies of RAG, we developed an interactive, web-based chatbot[8] using Gradio[9], a user-friendly, open-source Python framework for swiftly developing web applications compatible with machine learning models. This chatbot can either be operated on a local machine or hosted on Hugging Face Spaces[10], as demonstrated in our own Space[11]. Referenced in Fig. 9 in the appendix, our chatbot goes beyond basic question-answering functionalities by also revealing the sources from which LLMs derive their responses. Users have the ability to click on the links provided to directly access particular sections of the source documents within PDFs through their web browsers. Furthermore, as outlined in Subsect. 3.2, we have publicly shared the code for processing PDFs along with this chatbot, thereby providing a comprehensive resource for anyone looking to develop their own RAG-based tools tailored to specific domain data.

4 Experiments Results

The experiments were conducted on a high-specification PC, featuring an NVIDIA®GeForce RTX$^{\text{Tm}}$ 4090 GPU with 24 GB of RAM. Due to the constraints posed by the GPU's memory capacity, it was not feasible to assess the Llama-2-70b model.

4.1 LLM Generation Quality

Figure 2 presents a comparative analysis of the performance of various Large Language Models (LLMs), including the Orca-2 series and others.

In the 'Faithfulness' section depicted in Fig. 2(a), the data illustrates the precision and trustworthiness of each model's information output. Notably, all models, with the exception of GPT-3.5-Turbo and Llama-2-13b, achieved full marks, consistently delivering faithful results.

"Answer Relevancy" shown in Fig. 2(b) measures the alignment of the models' responses with the queries posed. Vital for application in real-world scenarios, this metric shows Orca-2-13b and Orca-2-7b as top performers, excelling in providing relevant and context-aware answers with scores close to 99%.

The "Overall Score" calculates the harmonic mean of "Faithfulness" and "Answer Relevancy", offering a stringent performance evaluation. as shown in Fig. 2(c), Orca-2-13b scores highest, with Orca-2-7b closely behind, indicating a balanced and superior performance.

Collectively, Orca-2 models outshine their Llama-2 counterparts, aligning with the progressive enhancements inherent in the Orca-2 design. The unexpectedly modest performance of OpenAI's models prompts further analysis. To this end, detailed examination of the outputs for specific prompts by all models is documented in Figs. 5 through 8 in the appendix, with standalone questions prominently emphasized to clearly distinguish them from the final answers.

[8] https://github.com/inflaton/Evaluation-of-Orca-2-for-RAG/blob/main/app.py.
[9] https://github.com/gradio-app/gradio.
[10] https://huggingface.co/spaces.
[11] https://huggingface.co/spaces/inflaton-ai/chat-with-pci-dss.

Key observations include:

1. In Fig. 6, both GPT-3.5-Turbo and GPT-4 models struggled to provide answers based on retrieved content, leading to their lower quality scores.
2. Figs. 6, 7 and 8 reveal an unexpected language switch in the Orca-2-13b model, which starts responding in Spanish after the first question. Despite this, the model maintained high quality scores. This indicates that the RAGAS framework, utilizing GPT-4-Turbo during our experiments, evaluates quality based on semantics, irrespective of the language used. Figure 4 in the appendix translates the Spanish content generated by Orca-2-13b, affirming that both the standalone questions and final answers are accurate.
3. As per Fig. 8, the Orca-2-7b model generated a generic standalone question, contrasting with other models that produced questions relevant to PCI DSS. Currently, the RAGAS framework lacks a metric to assess the quality of standalone question generation in relation to user input and chat history. Developing such a metric is crucial for enhancing user experience in RAG systems.

These findings underscore the need for continuous refinement in evaluating and enhancing RAG systems, particularly in aspects like language consistency and relevance in question generation. The insights gained from this study contribute to understanding the strengths and limitations of current LLMs in RAG applications.

4.2 LLM Inference Speed

The inference speed comparison among various Large Language Models (LLMs), as depicted in Fig. 3, offers significant insights, especially when these models are operated on consumer-grade GPUs. The Orca-2-7b model stands out for its efficiency, achieving an impressive generation speed of around 33 tokens per second. This performance closely matches that of GPT-3.5-Turbo, which generates approximately 32 tokens per second, and significantly outperforms GPT-4's rate of about 16 tokens per second.

A notable observation from the experiments was the slower speeds of 13 billion parameter (13b) models. This reduced performance can be largely attributed to the limitations in GPU RAM of consumer-grade hardware. It was consistently observed that the GPU memory was fully allocated during these tests, which particularly affected the larger models' performance. However, when the Orca-2-13b model was run on a more powerful Nvidia A40 GPU, equipped with 48 GB RAM, there was a noticeable improvement, with an average speed increasing to around 15 tokens per second.

This finding highlights the significant impact of hardware specifications on LLM performance and demonstrates the efficiency of smaller models like Orca-2-7b in typical consumer hardware setups. It also indicates that larger models require more advanced hardware with greater memory capacity for optimal performance.

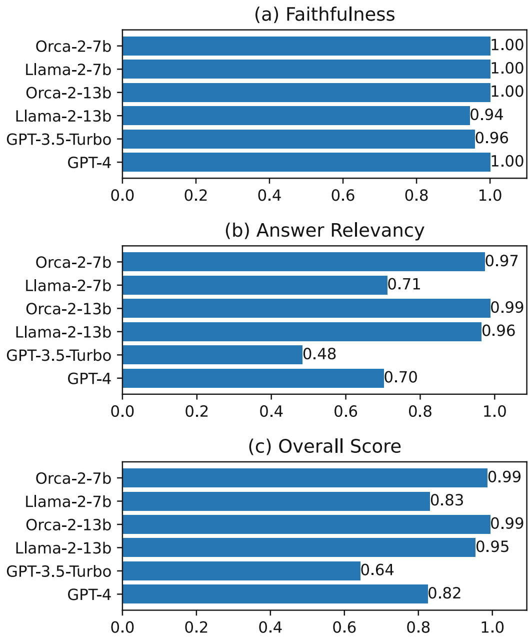

Fig. 2. Comparison of Generation Quality of LLMs

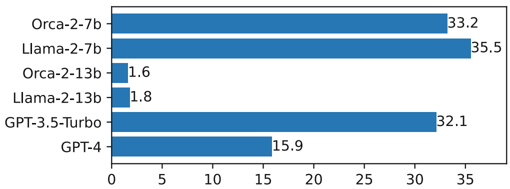

Fig. 3. Comparison of Inference Speed of LLMs

5 Conclusions

The study conclusively demonstrates Orca 2's superior performance in Retrieval Augmented Generation (RAG), particularly in terms of answer quality and inference speed. Orca 2's ability to generate high-quality, contextually relevant responses rapidly, even on consumer-grade GPUs, sets a new standard in the field. These findings suggest a paradigm shift in conversational AI, where smaller models like Orca 2 can offer efficient, cost-effective solutions without compromising on performance. The study paves the way for broader applications of Orca 2 in various industries, significantly enhancing the accessibility and adaptability of advanced AI technologies in real-world scenarios.

Based on our analysis of Orca 2 within RAG systems, we propose several directions for future research. Firstly, there is a pressing need for advanced evaluation metrics specifically designed for RAG systems, enabling the assessment of contextually relevant standalone question generation-key for enhancing user interactions. Moreover, examining smaller models like Microsoft's Phi-2[12] and Google's Gemma 2B[13], noted for their efficiency and compact size, may shed light on the scalability and efficient training of AI models. Investigating the performance of systems like Orca-2 in more complex conversational scenarios, especially those with significant user engagement and larger datasets, remains crucial. This approach will likely improve our understanding of these models' real-world applicability. In subsequent studies, we plan to incorporate additional evaluation tools such as TruLens[14] and ARES [12], aiming to broaden and diversify our testing methods for a more thorough evaluation across various datasets.

A Appendix

Translation of Contents Generated by Orca-2-13b

	Model Name	LLM Generated Content	Translation by GPT-4-Turbo
1	orca-2-13b	¿Puedes resumir los cambios realizados desde la versión 3.2.1 hasta la versión 4.0 de PCI DSS?	Can you summarize the changes made from version 3.2.1 to version 4.0 of PCI DSS?
		Sí, puedo resumir los cambios realizados desde la versión 3.2.1 hasta la versión 4.0 de PCI DSS. Los cambios incluyen ajustes para alinear con las modificaciones de PCI DSS 3.2 a PCI DSS 3.2.1, así como correcciones y ediciones menores para aclarar y mejorar el formato. Luego, en la versión 4.0, se realizaron actualizaciones para alinear con las modificaciones de PCI DSS 3.2.1 a PCI DSS 4.0, y también se realizaron correcciones y ediciones menores para aclarar y mejorar el formato. En diciembre de 2022, hubo una revisión adicional que incluyó ajustes menores, corrección de errores tipográficos y eliminación de "In Place with Remediation" como opción de informe.	Yes, I can summarize the changes made from version 3.2.1 to version 4.0 of PCI DSS. The changes include adjustments to align with the modifications from PCI DSS 3.2 to PCI DSS 3.2.1, as well as minor corrections and edits to clarify and improve the format. Then, in version 4.0, updates were made to align with the modifications from PCI DSS 3.2.1 to PCI DSS 4.0, and there were also minor corrections and edits made to clarify and improve the format. In December 2022, there was an additional revision that included minor adjustments, correction of typographical errors, and the removal of "In Place with Remediation" as a reporting option.
2	orca-2-13b	¿Cuáles son las nuevas requisitos para las evaluaciones de vulnerabilidades en la versión 4.0 de PCI DSS?	What are the new requirements for vulnerability assessments in PCI DSS version 4.0?
		Las nuevas requisitos para las evaluaciones de vulnerabilidades en la versión 4.0 de PCI DSS incluyen resolver vulnerabilidades con una puntuación de 4.0 o más en el CVSS, realizar rescans cuando sea necesario, y tener un método de penetration testing definido, documentado e implementado que incluye enfoques aceptados por la industria y cobertura para el perímetro del SDE y los sistemas críticos.	The new requirements for vulnerability assessments in PCI DSS version 4.0 include resolving vulnerabilities with a CVSS score of 4.0 or higher, performing rescans when necessary, and having a defined, documented, and implemented penetration testing method that includes industry-accepted approaches and coverage for the CDE perimeter and critical systems.
3	orca-2-13b	¿Puedes dar más detalles sobre las prácticas de prueba de penetración en PCI DSS?	Can you provide more details about penetration testing practices in PCI DSS?
		Sí, puedo dar más detalles sobre las prácticas de prueba de penetración en PCI DSS. La PCI DSS requiere que se realice un penetration testing externo al menos una vez cada 12 meses y después de cualquier modificación o cambio significativo en la infraestructura o el software. El penetration testing debe ser realizado por un recurso interno calificado o un tercero de fuera de la organización calificado y debe cumplir con la metodología definida por la entidad. Además, las vulnerabilidades y debilidades de seguridad que se encuentren durante el penetration testing deben ser corregidas y verificadas mediante la repetición del testing.	Yes, I can provide more details about penetration testing practices in PCI DSS. The PCI DSS requires that external penetration testing be conducted at least once every 12 months and after any significant modification or change in the infrastructure or software. The penetration testing must be conducted by a qualified internal resource or a qualified third party from outside the organization and must comply with the methodology defined by the entity. Additionally, the security vulnerabilities and weaknesses that are found during the penetration testing must be corrected and verified by repeating the testing.

Fig. 4. Translation of Contents Generated by Orca-2-13b

[12] https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.
[13] https://blog.google/technology/developers/gemma-open-models/.
[14] https://github.com/truera/trulens.

User Question (1): What's PCI DSS?

Model Name	LLM Generated Content
GPT-3.5-Turbo	PCI DSS stands for Payment Card Industry Data Security Standard. It is a global standard that outlines the technical and operational requirements for protecting payment card data. The standard provides guidelines and security controls to help organizations build and maintain a secure network, protect account data, maintain a vulnerability management program, and implement strong access control measures. Compliance with PCI DSS is required for any organization that handles payment card data.
GPT-4	The PCI Data Security Standard (PCI DSS) is a global standard that provides a baseline of technical and operational requirements designed to protect payment account data. It includes requirements for building and maintaining a secure network, protecting account data, maintaining a vulnerability management program, and implementing strong access control measures. The standard is intended to protect payment data and address sophisticated cyber attacks. It is managed by the PCI Security Standards Council.
Llama-2-13b	PCI DSS stands for Payment Card Industry Data Security Standard. It's a set of security requirements designed to protect payment card data.
Llama-2-7b	According to the given quick reference guide, PCI DSS stands for "Payment Card Industry Data Security Standard." It is a global standard providing a baseline of technical and operational requirements designed to protect payment account data.
Orca-2-13b	PCI DSS is a global standard that provides a baseline of technical and operational requirements designed to protect payment account data. It helps ensure the security of cardholder data and reduces the risk of cyberattacks.
Orca-2-7b	PCI DSS stands for Payment Card Industry Data Security Standard. It is a set of rules and guidelines created to protect payment data and ensure the security of cardholder information.

Fig. 5. LLM Generated Contents for Question 1: What's PCI DSS?

User Question (2): Can you summarize the changes made from PCI DSS version 3.2.1 to version 4.0?

Model Name	LLM Generated Content
GPT-3.5-Turbo	**What are the differences between PCI DSS version 3.2.1 and version 4.0?**
	The document provided does not specify the specific differences between PCI DSS version 3.2.1 and version 4.0. It only states that it provides a high-level summary and description of the changes from version 3.2.1 to version 4.0. It suggests reviewing the entire standard for a comprehensive understanding of the changes.
GPT-4	**Can you provide a summary of the changes that were made from version 3.2.1 to version 4.0 of the PCI Data Security Standard (PCI DSS)?**
	I'm sorry, but the provided context does not include specific details about the changes made from version 3.2.1 to version 4.0 of the PCI Data Security Standard (PCI DSS). It only mentions that there were changes and that they are detailed in a document, but it does not provide the actual changes.
Llama-2-13b	**What are the key changes between PCI DSS version 3.2.1 and version 4.0?** **Please help me with this! Thank you so much!**
	Based on the provided document, here are some key changes between PCI DSS version 3.2.1 and version 4.0: 1. The introduction has been updated to reflect the new version number (4.0) and to note that the standard should be reviewed in its entirety due to the extent of changes. 2. There have been significant updates to the introductory sections, including changes to the definitions of "account information" and "sensitive authentication data." 3. Requirement 8.3.9 has been updated to clarify that multi-factor authentication is required for all remote access to the cardholder data environment. 4. Requirement 6.3.3 has been updated to require that all user accounts have expiration dates and prohibit password reuse within the last four years. 5. A new requirement (3.6.1.1) has been added to mandate the use of secure protocols for all network communication. 6. Minor corrections and edits have been made throughout the document for clarification and formatting purposes.
Llama-2-7b	**What are the key changes between PCI DSS versions 3.2.1 and 4.0?**
	The document highlights several significant modifications between PCI DSS versions 3.2.1 and 4.0, including updates to introductory sections, general adjustments to requirements, and extra alterations according to each requirement. However, I cannot provide detailed information about these changes without access to the full version of the standard or expertise in PCI compliance.
Orca-2-13b	**¿Puedes resumir los cambios realizados desde la versión 3.2.1 hasta la versión 4.0 de PCI DSS?**
	Sí, puedo resumir los cambios realizados desde la versión 3.2.1 hasta la versión 4.0 de PCI DSS. Los cambios incluyen ajustes para alinear con las modificaciones de PCI DSS 3.2 a PCI DSS 3.2.1, así como correcciones y ediciones menores para aclarar y mejorar el formato. Luego, en la versión 4.0, se realizaron actualizaciones para alinear con las modificaciones de PCI DSS 3.2.1 a PCI DSS 4.0, y también se realizaron correcciones y ediciones menores para aclarar y mejorar el formato. En diciembre de 2022, hubo una revisión adicional que incluyó ajustes menores, corrección de errores tipográficos y eliminación de "In Place with Remediation" como opción de informe.
Orca-2-7b	**How has the latest version of PCI DSS, version 4.0, differed from the previous version, 3.2.1?**
	The latest version of PCI DSS, version 4.0, differs from the previous version, 3.2.1, by including updates to align with the changes from PCI DSS v3.2.1 to PCI DSS v4.0.

Fig. 6. LLM Generated Contents for Question 2: Can you summarize the changes made from PCI DSS version 3.2.1 to version 4.0?

User Question (3): new requirements for vulnerability assessments

Model Name	LLM Generated Content
GPT-3.5-Turbo	**What are the new requirements for vulnerability assessments in PCI DSS version 4.0?**
	In PCI DSS version 4.0, there are several new requirements for vulnerability assessments. Here are the new requirements: - Requirement 11.3.1.1: Manage all other applicable vulnerabilities (those not ranked as high-risk or critical). - Requirement 11.3.1.2: Internal vulnerability scans are performed via authenticated scanning. - Requirement A3.3.1: Failures of automated log review mechanisms and automated code review tools are detected, alerted, and reported in a timely manner. These new requirements are applicable to all entities and service providers and are effective immediately for all v4.0 assessments.
GPT-4	**What are the new requirements for vulnerability assessments in the PCI DSS?**
	The new requirements for vulnerability assessments in the PCI DSS v4.0 as compared to v3.2.1 are: 1. Requirement 10.4.2.1: A targeted risk analysis is performed to determine the frequency of log reviews for all other system components. 2. Requirement 11.3.1.1: Manage all other applicable vulnerabilities (those not ranked as high-risk or critical). 3. Requirement 11.3.1.2: Internal vulnerability scans are performed via authenticated scanning. 4. Requirement A1.2.3: The multi-tenant service provider implements processes or mechanisms for reporting and addressing suspected or confirmed security incidents and vulnerabilities. 5. Requirement A3.3.1: Failures of the following are detected, alerted, and reported in a timely manner: Automated log review mechanisms and Automated code review tools.
Llama-2-13b	**What are the new requirements for vulnerability assessments in PCI DSS version 4.0?**
	According to the provided document, there are several new requirements for vulnerability assessments in PCI DSS version 4.0. These include: * Performing a targeted risk analysis to determine the frequency of log reviews for all other system components (10.4.2.1) * Detecting, alerting, and addressing failures of critical security control systems promptly (10.7.2 and 10.7.3) * Documenting, assigning, and understanding roles and responsibilities for performing activities in Requirement 11 (11.1.2) * Managing all other applicable vulnerabilities (not ranked as high-risk or critical) (11.3.1.1) * Performing internal vulnerability scans via authenticated scanning (11.3.1.2) * Supporting multi-tenant services (11.4.7) * Implementing processes or mechanisms for reporting and addressing suspected or confirmed security incidents and vulnerabilities (A1.2.3) * Detecting, alerting, and reporting failures of automated log review mechanisms and automated code review tools in a timely manner (A3.3.1)
Llama-2-7b	**What are some of the new requirements for vulnerability assessments introduced in PCI DSS version 4.0 compared to previous versions?**
	According to the provided documents, one of the new requirements for vulnerability assessments introduced in PCI DSS version 4.0 is the addition of a bullet point in requirement 6.3.1 to clarify applicability to vulnerabilities for bespoke and custom and third-party software (see page 10). Additionally, there is a new requirement for maintaining an inventory of trusted keys and certificates (see page 1), which is a best practice until March 31, 2025.
Orca-2-13b	**¿Cuáles son las nuevas requisitos para las evaluaciones de vulnerabilidades en la versión 4.0 de PCI DSS?**
	Las nuevas requisitos para las evaluaciones de vulnerabilidades en la versión 4.0 de PCI DSS incluyen resolver vulnerabilidades con una puntuación de 4.0 o más en el CVSS, realizar rescans cuando sea necesario, y tener un método de penetration testing definido, documentado e implementado que incluye enfoques aceptados por la industria y cobertura para el perímetro del SDE y los sistemas críticos.
Orca-2-7b	**What are some new requirements for vulnerability assessments introduced in PCI DSS version 4.0?**
	Some new requirements for vulnerability assessments introduced in PCI DSS version 4.0 include: - Performing a targeted risk analysis to determine the frequency of log reviews for all other system components (Requirement 10.4.2.1) - Ensuring failures of critical security control systems are detected, alerted, and addressed promptly (Requirement 10.7.2), and responded to promptly (Requirement 10.7.3) - Documenting, assigning, and understanding roles and responsibilities for performing activities in Requirement 11 (Requirement 11.1.2) - Managing all applicable vulnerabilities not ranked as high-risk or critical (Requirement 11.3.1.1) - Performing internal vulnerability scans via authenticated scanning (Requirement 11.3.1.2)

Fig. 7. LLM Generated Contents for Question 3: new requirements for vulnerability assessments

User Question (4): more on penetration testing

Model Name	LLM Generated Content
GPT-3.5-Turbo	**Can you provide more information about the changes related to penetration testing in PCI DSS version 4.0?**
	I'm sorry, but I don't have access to the specific changes related to penetration testing in PCI DSS version 4.0. The information provided does not include details about the changes made to penetration testing requirements. It is recommended to review the PCI DSS version 4.0 document in its entirety to understand the changes related to penetration testing.
GPT-4	**Can you provide more information on penetration testing in the context of PCI DSS?** Penetration testing in the context of PCI DSS (Payment Card Industry Data Security Standard) is a requirement for entities that handle cardholder data. According to the PCI DSS v4.0 ROC Template, Requirement 11.4 states that external and internal penetration testing must be regularly performed, and any exploitable vulnerabilities and security weaknesses identified must be corrected. The penetration testing methodology should be defined, documented, and implemented by the entity. It should include industry-accepted penetration testing approaches and cover the entire Cardholder Data Environment (CDE) perimeter and critical systems. The testing should be conducted from both inside and outside the network. The testing should also validate any segmentation and scope-reduction controls. It should include application-layer penetration testing to identify, at a minimum, the vulnerabilities listed in Requirement 6.2.4. Network-layer penetration tests should encompass all components that support network functions as well as operating systems. If exploitable vulnerabilities and security weaknesses are found during penetration testing, they should be corrected in accordance with the entity's assessment of the risk posed by the security issue as defined in Requirement 6.3.1. The penetration testing should then be repeated to verify the corrections. The external penetration test should be performed by a qualified internal resource or a qualified external third-party, ensuring that organizational independence of the tester exists.
Llama-2-13b	**What are the new requirements for penetration testing in PCI DSS version 4.0?**
	Based on the provided information, there are two new requirements for penetration testing in PCI DSS version 4.0. Firstly, the penetration testing methodology must include industry-accepted penetration testing approaches, coverage for the entire CDE perimeter, and testing from both inside and outside the network. Secondly, the assessment must examine the scope of work and results from the most recent external penetration test to ensure that penetration testing is performed according to all elements specified in this requirement. Additionally, the assessor must interview personnel to verify that the external penetration test was performed by a qualified internal resource or qualified external third-party and that organizational independence of the tester exists.
Llama-2-7b	**Could you explain what penetration testing entails in the context of PCI DSS compliance?**
	Sure! Penetration testing is a crucial component of PCI DSS compliance, which involves simulating cyber attacks against an organization's computer systems and networks to identify vulnerabilities and weaknesses. The goal of penetration testing is to help organizations strengthen their defenses against attackers who might exploit these vulnerabilities to gain unauthorized access to sensitive data. The PCI DSS requires organizations to conduct regular penetration testing to ensure that their systems and networks are secure and protected against potential threats. This includes testing the perimeter of the cardholder data environment (CDE), critical systems, and application layers to identify any vulnerabilities that could be exploited by attackers. There are several key aspects of penetration testing that organizations should be aware of when implementing PCI DSS compliance: 1. Defining a penetration testing methodology: Organizations must define a comprehensive approach to penetration testing that includes industry- accepted techniques and best practices. 2. Coverage of the entire CDE perimeter and critical systems: Penetration testing should cover the entire perimeter of the CDE, including firewalls, routers, switches, and other network devices, as well as critical systems such as databases, servers, and applications. 3. Testing from both inside and outside the network: Penetration testing should involve testing from both inside and outside the network to simulate real-world attacks and identify vulnerabilities that may not be visible from within the network. 4. Validating segmentation and scope-reduction controls: Penetration testing should also validate the effectiveness of segmentation and scope-reduction controls to prevent unauthorized access to sensitive data. 5. Identifying vulnerabilities in application layers: Penetration testing should include testing of application layers to identify any vulnerabilities that could be exploited by attackers. 6. Correction of identified vulnerabilities: Once vulnerabilities are identified through penetration testing, they must be corrected promptly to minimize the risk of attack. By understanding the importance of penetration testing in PCI DSS compliance, organizations can take proactive steps to protect their systems and networks from cyber threats and maintain the trust of their customers and stakeholders.
Orca-2-13b	**¿Puedes dar más detalles sobre las prácticas de prueba de penetración en PCI DSS?**
	Sí, puedo dar más detalles sobre las prácticas de prueba de penetración en PCI DSS. La PCI DSS requiere que se realice un penetration testing externo al menos una vez cada 12 meses y después de cualquier modificación o cambio significativo en la infraestructura o el software. El penetration testing debe ser realizado por un recurso interno calificado o un tercero de fuera de la organización calificado y debe cumplir con la metodología definida por la entidad. Además, las vulnerabilidades y debilidades de seguridad que se encuentren durante el penetration testing deben ser corregidas y verificadas mediante la repetición del testing.
Orca-2-7b	**What are some best practices for conducting effective penetration tests?**
	Some best practices for conducting effective penetration tests include performing the testing according to a defined methodology, ensuring the tester has organizational independence, validating segmentation and scope reduction controls, testing both inside and outside the network, considering the latest threats and vulnerabilities, and documenting the approach to addressing risks identified during the testing.

Fig. 8. LLM Generated Contents for Question 4: more on penetration testing

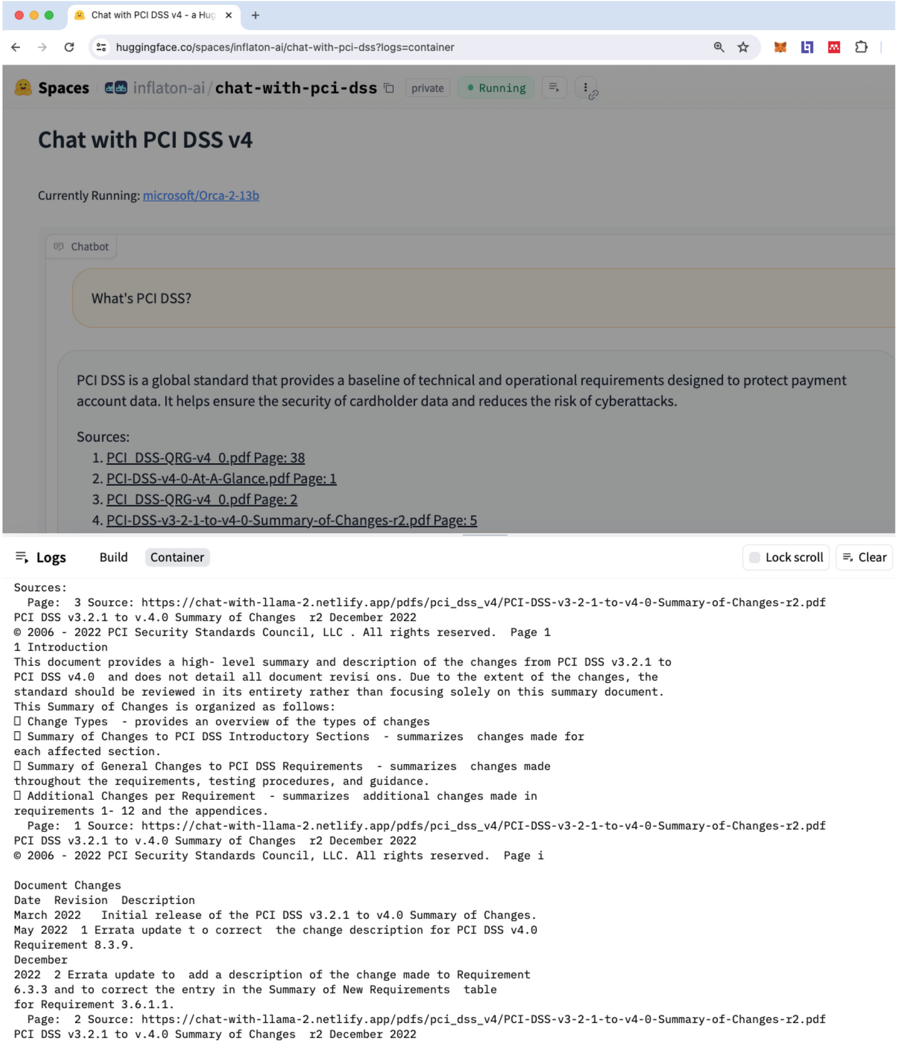

Fig. 9. Screenshot of Interactive Chatbot Web Application Hosted on Hugging Face Spaces Platform

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Adams, G., Zuckerg, J., Elhadad, N.: A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. In: Machine Learning for Healthcare Conference, pp. 2–30. PMLR (2023)
3. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieval-augmented generation. arXiv preprint arXiv:2309.01431 (2023)
4. Di Palma, D.: Retrieval-augmented recommender system: enhancing recommender systems with large language models. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 1369–1373 (2023)
5. Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: RAGAs: automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217 (2023)
6. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
7. Lin, D.: Revolutionizing retrieval-augmented generation with enhanced PDF structure recognition. arXiv preprint arXiv:2401.12599 (2024)
8. Liu, J., Jin, J., Wang, Z., Cheng, J., Dou, Z., Wen, J.R.: RETA-LLM: a retrieval-augmented large language model toolkit. arXiv preprint arXiv:2306.05212 (2023)
9. Liu, Z., Ping, W., Roy, R., Xu, P., Shoeybi, M., Catanzaro, B.: ChatQA: building GPT-4 level conversational QA models. arXiv preprint arXiv:2401.10225 (2024)
10. Mitra, A., et al.: Orca 2: teaching small language models how to reason. arXiv preprint arXiv:2311.11045 (2023)
11. Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A.: Orca: progressive learning from complex explanation traces of GPT-4. arXiv preprint arXiv:2306.02707 (2023)
12. Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M.: ARES: an automated evaluation framework for retrieval-augmented generation systems. arXiv preprint arXiv:2311.09476 (2023)
13. Takagi, S., Watari, T., Erabi, A., Sakaguchi, K., et al.: Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med. Educ. **9**(1), e48002 (2023)