CS605 Natural Language Processing for Smart Assistants
Project Proposal

LLM + RAG in Banking context

G2 Group 6
Muhammad Ridhwan Bin Zainal Abidin
Ng Juan Yong
Yang Hexu
Zhang Ge
Zheng Yimin

30 June 2024

Table of Contents

1. Abstract

Generative AI (Gen-AI), particularly Large Language Models (LLMs), has been gaining considerable attention recently. A study by McKinsey highlighted numerous opportunities for Gen-AI to address complex tasks, particularly within the banking sector. This paper investigates the use of LLMs, combined with Retrieval-Augmented Generation (RAG), to generate creative and engaging marketing content tailored to a bank's products and services. Our approach aims to streamline the manual processes where marketing specialists draft and review campaign content to ensure it is error-free. We conducted experiments with multiple LLMs and various prompt techniques to identify the optimal combination for producing the most desirable outputs in our context. Our results indicate that guided prompting and hypothetical prompting techniques yielded the best outcomes, with Gemini-1.0-Pro outperforming ChatGPT-4.0 in most aspects.

2. Introduction

2.1 Problem Statement

Consumers receive tons of marketing mailers on a daily basis, across multiple platforms like emails, push notes, and many more. Marketing has become a huge parts of our lives, and businesses have been proactively using marketing to reach out to their consumers, fostering customer engagement and ultimately driving sales. High-quality content helps businesses enhance customer engagement, drive digital adoption, and build long-term relationships, ultimately leading to increased customer loyalty and business growth.

However, the creation of marketing campaigns remains largely manual even today, necessitating content creators to develop innovative materials to promote products and services. This process can be time-consuming, susceptible to spelling errors, and often results in overly generic content.

2.2 Our solution

McKinsey's research underscores the significance of embracing generative AI (gen-AI) across sectors, automating tasks such as personalized marketing and sales content creation, and optimizing the transition from legacy systems with natural language translation capabilities in the banking sector. Embracing gen-AI can mitigate challenges such as labor shortages, human errors, and more.

Recognizing the significance and potential of generative AI in banking, we aim to narrow our project's focus to this sector. Specifically, we will explore how banks can utilize LLMs and Retrieval-Augmented Generation (RAG) to create superior marketing content tailored for diverse customer segments. RAG is the process of enhancing the LLM’s outputs, by referencing to an authoritative knowledge base outside of its training data before generating a response. RAG extends the already powerful capabilities of LLMs to specific domains or organization’s internal knowledge base, without having to retrain the LLM model. This helps organizations to save costs and improve the LLM’s outputs, ensuring that the content generated remains relevant and up to date (AWS, 2024).

2.3 Details of difficulties, techniques and why it will work

LLMs by themselves, are susceptible to hallucinations, whereby they generate information that are incorrect and not based on factual data. Hallucinations can occur in the form of fabricated

facts, whereby non-existent information are returned by the model, or producing information that is irrelevant to the user input due to misinterpretation.

Hallucinations can occur due to limitations in training data, which may consist of incorrect information or lack of coverage in specific domains (i.e. information on products and service offerings by banks). This limitation is highly relevant to our use cases as we expect domain specific responses in a relatively narrow topic (i.e. marketing content in the form of emails). It would be difficult to train a LLM solely based on training data that consists of only marketing emails in the banking sector.

Additionally, since training data is not time sensitive, the outputs produced by LLMs are not guaranteed to be up to date. This limitation undermines the effectiveness of LLMs in our use case as it is cost inefficient to frequently retrain a LLM to include latest contents. LLMs lack the capability to incorporate up-to-date information, which may be valuable in our use case as there might be a need to generate contents to follow up on the previous promotion emails that are recently sent.

Tackling the above difficulties

RAG, which combines LLMs with databases, can help to verify and provide time accurate information. RAG has the advantage of using up-to-date information to draw inferences without having to retrain using new data.

Since we intend to solve a highly domain specific (i.e. restricted to the banking sector) problem, RAG can retrieve contextually relevant information and produce domain relevant responses. Users are also able to fine tune the outputs generated via customizing the knowledge bases used by the RAG, retaining only high-quality content while removing those that are no longer relevant. Users can also interactively refine their queries and guide the retrieval process, leading to more precise and useful outputs.

3. Related Work

This section reviews key advancements in LLMs and their applications, focusing on BERT, GPT-Neo, and RAG. Lewis, P., et al. (2020) introduced the RAG which integrates retrieved documents with generative modeling to produce accurate and contextually rich responses. This method addresses the limitations of pure generative models, such as hallucinations and inaccuracies. More recent work by Hoshi, Y., et al. (2023) provided an extensive survey on RAG techniques, highlighting the applications of RAG in various domains, including marketing content creation.

The use of retrieval is mainly for language models to improve perplexity (Borgeaud et al., 2022; Wang et al., 2023), factual accuracy (Nakano et al., 2021), downstream task accuracy (Guu et al., 2020; Izacard & Grave, 2021; Izacard et al., 2022; Lewis et al., 2020) and in-context learning capability (Huang et al., 2023). Retrieval-augmented LLM are known for its capability to handle question answering with long document and in open-domain. It enhances LLMs by retrieving relevant document chunks from external knowledge base through semantic similarity calculation. RAG is thus able to effectively reduce the problem of generating factually incorrect output. It has also become a key part of enhancing the suitability of LLMs for real-word applications like

chatbots, and more. It also serves as a fine-tuning technique for LLMs, allowing it to answer more complex and knowledge-intensive tasks (Gao et al., 2024).

RAG is also increasingly attracting attention from traditional software and cloud-service providers, offering a level of customization to meet specific needs. For instance, Flowise AI adopts a low-code approach which allows users to deploy AI applications including RAG. Amazon is also including RAG-centric services to cater to user's specific needs through Amazon Kendra (Gao et al., 2024). It is also recognized for improving click-through rates in marketing materials, ultimately boosting sales. This underscores the importance of RAG in tailoring content to resonate with audiences, leading to higher engagement levels. RAG has been proven to show potential in enabling traditional marketing agencies and in-house marketing units to drive sales by generating personalized and persuasive marketing content (MyScale, 2024).

However, query optimization is a critical challenge influencing the quality of outputs retrieved by Retrieval-Augmented Generation (RAG), as detailed in the paper by Gao et al. Clear and well-structured queries tend to yield better outputs, while poorly formulated queries can lead to suboptimal results. Various strategies have been proposed to enhance query optimization, including query expansion, multi-query approaches, sub-queries, and Chain-of-Verification (CoVe). G. Marvin et al.'s paper also emphasizes the importance of prompt engineering in generating desirable outputs from Large Language Models (LLM). It discusses optimal model choices for different prompt techniques, with GPT-3/GPT-4 being among the frequently recommended models.

In our study, we will investigate LLM models such as Gemini and GPT-4, leveraging their widespread adoption and strong track record in producing relevant results. Additionally, we will explore diverse prompt engineering techniques to determine the most effective approach when used in conjunction with Retrieval-Augmented Generation (RAG).

4. Approach

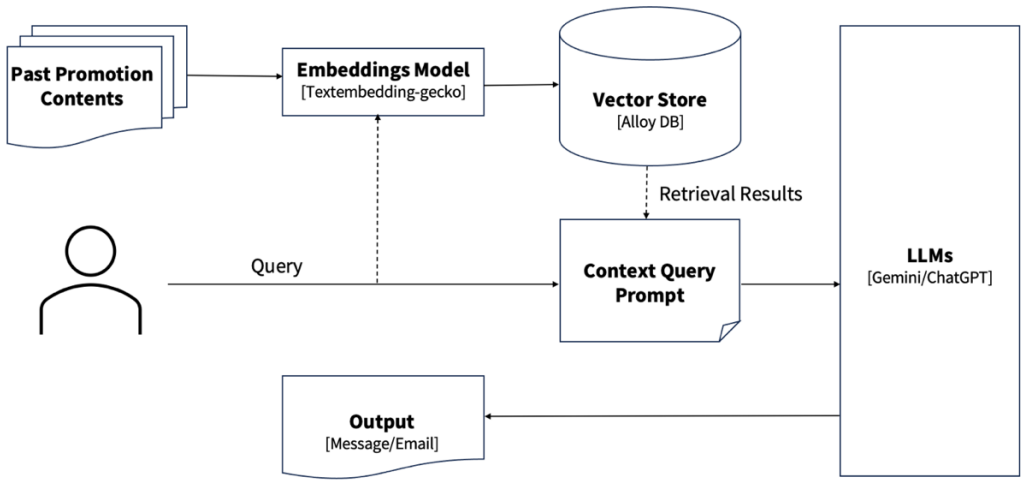

Illustration 1: Architecture of our proposed solution

The process outlines a method for generating new promotional content using a combination of text embeddings, database storage, user query processing, context-aware prompts, and large language models. Here are the details at each step:

1. **Text Embeddings**:

Texts from past promotional materials are converted into numerical representations called embeddings. These embeddings are designed to capture the semantic meanings of the words and phrases in the texts.

2. **Database Storage**:

The generated embeddings are stored in a database, creating a repository that can be searched and retrieved for similar content or context.

3. **User Query Processing**:

When a user submits a query, presumably with specific requirements or themes for new promotional content, this query is processed. The processed query is used to interact with the vector store, which contains the embeddings, to find the most relevant past promotions.

4. **Context Query Prompt**:

Using the retrieved past promotions, a context query prompt is created. This prompt is designed to provide a foundation for the large language model to understand the context and generate content that is relevant and coherent.

5. **Content Generation**:

The context query prompt is input into a large language model, such as Gemini or ChatGPT. The language model uses the prompt, along with the context provided by the retrieved past promotions, to create new promotional messages.

6. **Output**:

The final output from the language model is new promotional content that is specifically tailored to the user's query and the context of past promotional materials. This content can then be disseminated through various channels, such as push notifications or email messages, to reach the intended audience.

5. Experiments

5.1 Experimental Details

Experiments were conducted using various prompts to test the LLMs without RAG and with RAG context queries. The procedures are as follows:

1. **Input Generation:** Create context query prompts based on prompt techniques identified in 4.2.
2. **Model Execution:** Run the LLMs using these prompts to generate outputs.
3. **Output Analysis:** Evaluate the generated content using the defined metrics.

5.2 Details of dataset used

The data collection came from multiple sources, as outlined below:

- Phone marketing campaigns of a Portuguese Bank (Bank Marketing), which were retrieved from Kaggle.
- Contents of existing email marketing campaigns from various banks in Singapore
- Promotional materials from official bank websites
- Marketing push notifications from bank applications
- Promotional text messages from banks

Preprocessing for Usability:
To ensure the data is ready for analysis, we preprocessed the promotional materials to fit the required metadata structure. This allows for seamless integration into our vector-based knowledge store, enabling efficient processing and utilization. The attributes and their descriptions are as follows:

Data Attribute	Description
Template_id	Unique identifier for each marketing template.
Description	Category of the bank's marketing service (e.g., Credit Card Promotion, Investment Promotion).
Template_text	The actual content of the promotional message.
Platform	The medium used for promotion (mobile or Email).
URL	Links to more information or to the service/product being promoted.
Subject	The subject line of the email or headline of the message.
Requirements	Any specific conditions or requirements mentioned.

Additionally, since our project requires the creation of marketing content specific to the bank's products and services, we ensured that all the bank names in our dataset were masked with the name "Bank XYZ." The marketing campaigns we used are all past promotional campaigns executed by "Bank XYZ" and provides a prelude to the products and services offered by "Bank XYZ." An example is outlined below:

template_id	description	template_text	platform	url	subject	requirements
12408	Credit Card Promotion	One Card for Your Everyday Use With your XYZ SMRT Card, enjoy 5% savings on your daily expenses – ranging from online purchases, groceries to commute. In addition, enjoy the flexibility of SMRT$ as shopping vouchers or cash rebates. As early as the next statement month, you can redeem shopping or dining vouchers or cash rebate with a minimum of SMRT$10.	Email	https://xyz.com/sg/cashrebate	Enjoy up to 5% savings on your daily expenses with Citi SMRT Card today!	Customers between 20 to 80 years old; Minimum monthly salary of S$5000

All these templates are stored in the 'experiment_results.xlsx' document, under the 'marketing_templates' tab.

5.3 Prompt Engineering

We employed a variety of prompts to simulate different marketing scenarios within the banking sector, aiming to enhance the performance of our LLM. For email campaigns, the outputs were designed to be longer and more detailed, whereas push notifications and text messages were crafted to be brief and attention-grabbing.

Given that the content varies based on the publication channel (e.g., push notifications vs. email campaigns), we slightly adjusted our prompt inputs for each channel to observe how the outputs differed. Specifically, we created prompts focused on remittance for push notifications and text message campaigns, and prompts centered on investments for email messages.

The following types of prompts are experimented:

1. **Zero-shot prompting:** Requesting the model to perform a task without providing any examples or prior context.
2. **Few-shot prompting:** Providing a few examples to help the model understand and generate appropriate responses for a task.
3. **Direct prompting:** Asking the model a straightforward question or giving a clear command to obtain a specific response.
4. **Contextual prompting:** Including additional context or background information in the prompt to guide the model's response.
5. **Persona prompting:** Setting a specific personality, role, or perspective for the model to adopt while generating responses.
6. **Completion prompting:** Starting a text or sentence and asking the model to complete it based on the given beginning.
7. **Guided prompting:** Providing detailed instructions or constraints to direct the model's response in a specific way.
8. **Hypothetical prompting:** Posing a "what if" scenario or a hypothetical situation to explore potential responses or outcomes.
9. **Chain-of-thought (COT) Prompting:** Series of intermediate natural language reasoning steps leading to the outcome. It teaches LLMs how to reason by adding Chain of thought to the prompt, which greatly improves LLM's reasoning ability and performance.

5.4 Models used

In here, two different models were used for our experiment:

- **ChatGPT-4o:** Known for its extensive capabilities in understanding and generating human-like text, GPT-4 was chosen for its superior performance in various NLP tasks.
- **Gemini-1.0-Pro:** Advanced AI model developed by Google DeepMind, designed to excel in natural language processing (NLP) tasks.

The only configurations we adjusted for these models were the temperature and the maximum allowable tokens. We set a fixed temperature of 0.6 for both models to encourage the generation of some creative outputs and allowed a maximum of 1,024 tokens to be produced for each prompt. Results generated from both models may be found in 'experiment_results.xlsx' document, under the 'gemini_responses' and 'chatgpt_responses' tabs.

5.5 Evaluation Methods

The intended usage of our LLM + RAG model is to receive an input prompt, requesting a marketing campaign template for a specified fictitious bank. We used the following metrics to evaluate whether our model can perform well for its intended usage:

1. **Answer Relevancy** - The Answer Relevancy metric evaluates the RAG pipeline's generator, by evaluating whether the actual output of the LLM is relevant to the input prompt given (Ip, 2024). This metric was chosen due to the importance of the LLM model to output content as relevant as possible to the prompt given by users.
2. **Coherence** - The Coherence metric measures the overall quality of the sentences generated in the LLM's output (Ip, 2024). It is important that our model outputs contain coherent, high-quality sentences in response to the users' prompts.
3. **Contextual Precision** - The Contextual Precision metric evaluates the RAG pipeline's retriever, by evaluating whether the answers in the RAG's Retrieval Context are relevant to the input prompt given (Ip, 2024). This metric allows us to evaluate if the RAG can retrieve relevant answers to better augment the output content, given the prompt that was provided as input.

The metrics were chose to evaluate the LLM + RAG's quality of answers, ability to reply with relevant information, as well as measure the ability of the RAG to provide relevant information given the input prompt. In this experiment, we used both human evaluation and DeepEval library to perform evaluation. It uses ChatGPT's LLM, together with preset metrices like Coherence and etc, to evaluate the scores of the LLM (with RAG)'s outputs, on a scale from 0 to 1.

6. Results

We evaluated the results for the RAG outputs using DeepEval. Below results (evaluated using deepeval framework) is shown for both the models and various metrices:

	ChatGPT-4o			Gemini-1.0-Pro		
Prompt Type	**Answer Relevancy**	**Coherence**	**Context Precision**	**Answer Relevancy**	**Coherence**	**Context Precision**
Zero-shot prompting	1.00	0.69	0.00	0.95	0.78	0.00
Direct prompting	0.80	0.69	1.00	1.00	0.81	1.00
Few-shot prompting	0.63	0.81	1.00	0.86	0.79	0.75
COT prompting	1.00	0.79	1.00	1.00	0.87	0.00
Contextual prompting	0.63	0.72	0.50	1.00	0.78	1.00
Persona prompting	1.00	0.83	0.00	0.93	0.89	1.00
Completion prompting	0.92	0.91	0.42	1.00	0.80	0.00
Guided prompting	1.00	0.85	1.00	1.00	0.80	1.00
Hypothetical prompting	1.00	0.83	1.00	1.00	0.81	1.00
Average results	0.89	0.79	0.66	0.97	0.81	0.64

Details of the outputs from the LLMs can be found in 'experiment_results.xlsx', under 'chatgpt_responses' and 'gemini_responses' tab.

We are able to see that, generally, the Gemini-1.0-Pro outperforms ChatGPT-4o in the answer relevancy and coherence aspects. The best prompt technique suggested here would be the guided prompting technique, as well as the hypothetical prompting technique.

7. Analysis

We analyzed the results from both the LLM and model perspectives. While it is well-known that ChatGPT-4.0 is widely used and regarded as one of the best LLMs on the market, Gemini-1.0-Pro outperformed ChatGPT in our project, as seen from the results above. In an experiment done and discussed in Tom's guide, both models are known to perform equally well in terms of knowledge retrieval, application, and learning. However, Gemini-1.0-Pro excelled in generating creative yet

relevant responses. This is crucial for our project, as marketing content needs to be both creative and pertinent to the bank's products and services.

We expected Chain-of-Thought (CoT) to be one of the best-performing prompt techniques, given its reputation for significantly enhancing LLM's reasoning and performance. However, it was surprising that CoT's performance slightly lagged behind less commonly used prompt techniques like guided prompting and hypothetical prompting. We took a closer look at the outputs generated from these techniques and noticed that the outputs from CoT contained redundant information, which may have acted as noisy inputs to DeepEval, resulting in less accurate outcomes. Hypothetical and Guided prompting, by contrast, can provide more streamlined, focused, and contextually appropriate responses in our context.

8. Conclusion

8.1 Limitations and Future Research

From our experiments, we have discovered certain limitations of using RAG in generating domain specific marketing contents.

Firstly, we discovered that hallucinations still occur in RAG, albeit considerably less than traditional LLMs without RAG. Considering the reliance of RAG on external knowledge bases, which in our case refers to existing marketing contents, we hypothesized that this shortcoming can be improved by using a more extensive set of marketing data as the dataset that we managed to gather from our personal sources, such as email and SMSs, is rather limited.

Secondly, we have limited the scope of this project to strictly text contents. Under this scope, we did not investigate the possibilities of applying RAG techniques in other areas, such as marketing images or a combination of images and text contents. In fact, in the process of gathering marketing content to use as our dataset, we removed any images that are present in the marketing material. This process inevitably discards valuable contexts and information which are relevant to the contents that we intend to generate. Hence, we suggest that exploring the applications of RAG in generating other types of contents will be a worthwhile endeavour.

Lastly, using RAG to generate highly personalized marketing contents is a topic that can be further explored. We did not manage to perform any in-depth analysis on this topic since we are not able to gather a sufficiently diverse range of marketing contents targeted at users of different profiles and we also do not have access to channels for gathering personal information of customers from different profiles. Even though our dataset contains marketing materials that are somewhat personalized, there is still room for further exploration in terms of generating marketing contents specific to individuals.

8.2 Summary

We have proved that our proposed solution of integrating RAG with LLMs for marketing content creation in the banking sector holds great promise. It addresses existing pain points in the content creation process adequately by offering a viable alternative to more efficient and accurate content generation, and positions banks to better engage with their customers through personalized and up-to-date marketing materials. As we move forward with our research and experimentation, we

are confident in the potential of this approach to revolutionize marketing content creation and drive meaningful business outcomes.

9. References

Amazon Web Services. (n.d.). *What is retrieval-augmented generation?* Retrieved June 29, 2024, from https://aws.amazon.com/what-is/retrieval-augmented-generation/

Confident AI. (n.d.). Documentation. Retrieved May 29, 2024, from https://docs.confident-ai.com/

Devlin, J., et al.(2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805

G. Marvin et al., 2024. *Prompt Engineering in Large Language models.* Retrieved June 20, 2024, from https://www.researchgate.net/publication/377214553

Gao, et al., (2024). *Retrieval-Augmented Generation for Large Languag eModels: A Survey.* Retrieved 20 May 2024 from arXiv:2312.10997v5 [xs.CL]

Hoshi, Y., et al. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv:2308.10633

Ip, J. (2024, January 22). *LLM Evaluation Metrics: Everything You Need for LLM Evaluation*. Confident AI. https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Ip, J. (2024, June 23). *Answer Relevancy | DeepEval – The Open-Source LLM Evaluation Framework*. DeepEval. https://docs.confident-ai.com/docs/metrics-answer-relevancy

Ip, J. (2024, June 23). *Contextual Precision | DeepEval – The Open-Source LLM Evaluation Framework*. DeepEval. https://docs.confident-ai.com/docs/metrics-contextual-precision

J Damji., (2024, Feb 12). *Best prompt techniques for best LLM responses. Medium.* Retrieved June 29, 2024, from https://medium.com/the-modern-scientist/best-prompt-techniques-for-best-llm-responses-24d2ff4f6bca

Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge- Intensive NLP Tasks.* arXiv:2005.11401

MyScale. *4 Ways RAG Personalized Content Delivery Transforms Marketing Strategies.* Retrieved 19 May 2024 from https://myscale.com/blog/rag- personalized-content-delivery-transforms-marketing-strategies/

Tom's Guide. (2023, December 4). *Google Gemini vs. OpenAI ChatGPT: Which is better? Tom's Guide.* Retrieved June 29, 2024, from https://www.tomsguide.com/ai/google-gemini-vs-openai-chatgpt

Bank Marketing. Retrieved from https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing