



The Power of NLP + RAG in Banking

Muhammad Ridhwan Bin Zainal Abidin

Ng Juan Yong

Yang Hexu

Zhang Ge

Zheng Yimin

G2 Group 6



Agenda

1

Background

Introduction to RAG, scope project and related work

2

Method

Architecture of our proposed solution

3

Scope of data collected

Data used to support our experiments

4

Experiments

Types of prompts and models explored in this project

5

Evaluation

Evaluation of experiments conducted

6

Conclusion

Value propositions and future works





1 Background

Introduction to RAG, scope of project and related works



Background – RAG as a solution and related works

What is RAG?

Framework that combines the strengths of retrieval-based methods and generative models to improve the quality and relevance of generated content.

How it works?

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.

Who uses RAG?

Companies are also recognizing the importance of RAG in their applications, and there are increasing numbers in of players in the market offering RAG services. 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).

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.

Background

Method

Data

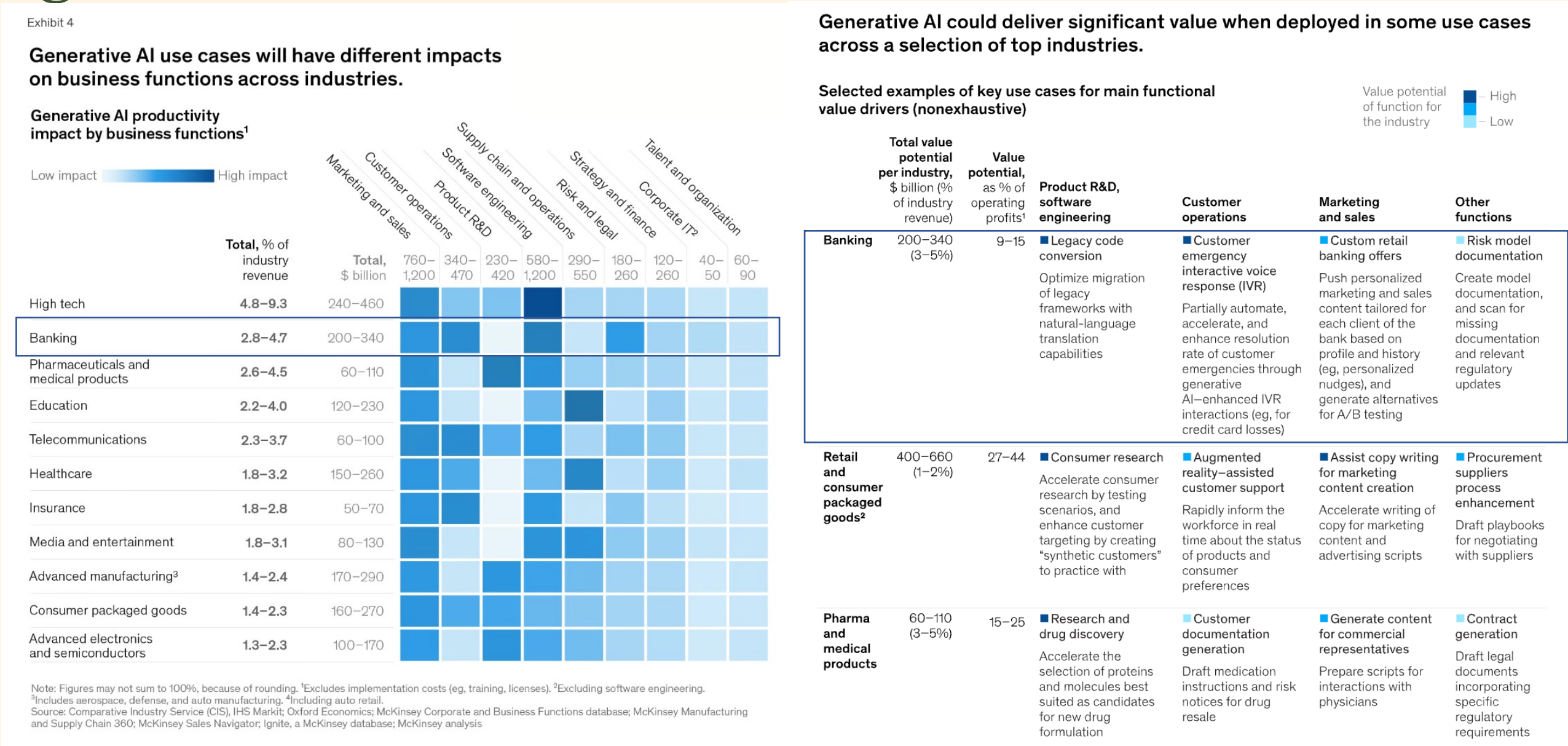
Experiments

Evaluation

Conclusion



Background – RAG as a solution for banks



Source: [Economic Potential of Gen AI - McKinsey](#)

Background – Scope of project

1. Scope

In this project, we will be focusing on marketing content generation in the banking sector, and how it can leverage on the power of **LLMs, together with Retrieval-augmented generation (RAG)**.

We will be adopting a persona, "**Bank XYZ**" for the scope of our project

2. Aims of RAG + LLM

To generate persuasive and personalized marketing campaigns that is

1. Specific to its domain (i.e. its products and services) and
2. Follows similar style of content to its existing marketing content

3. Users

Existing bank consumers
New to bank consumers

Background

Method

Data

Experiments

Evaluation

Conclusion 

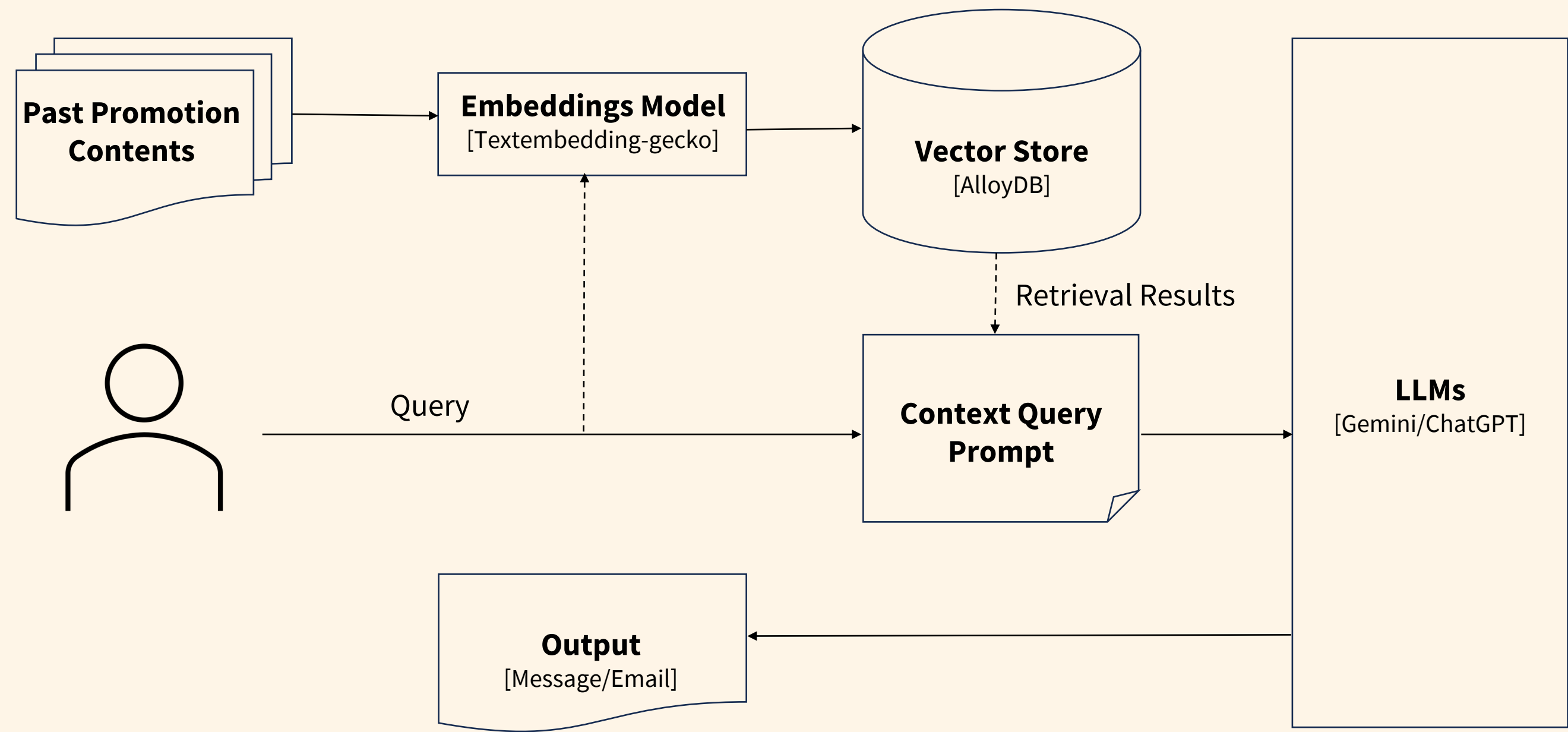


2 Method

Architecture of our proposed solution



Architecture



Background

Method

Data

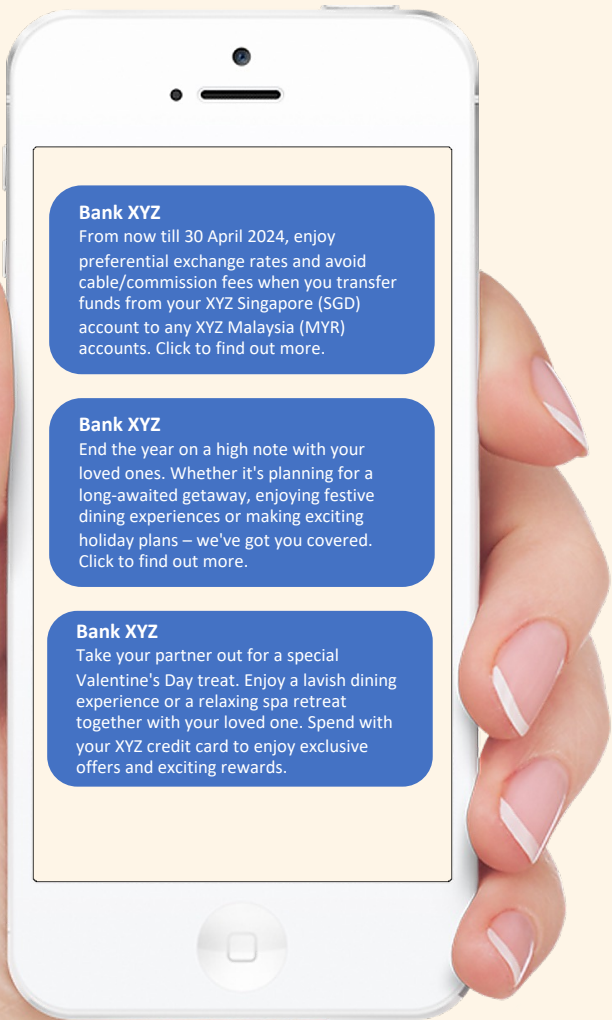
Experiments

Evaluation

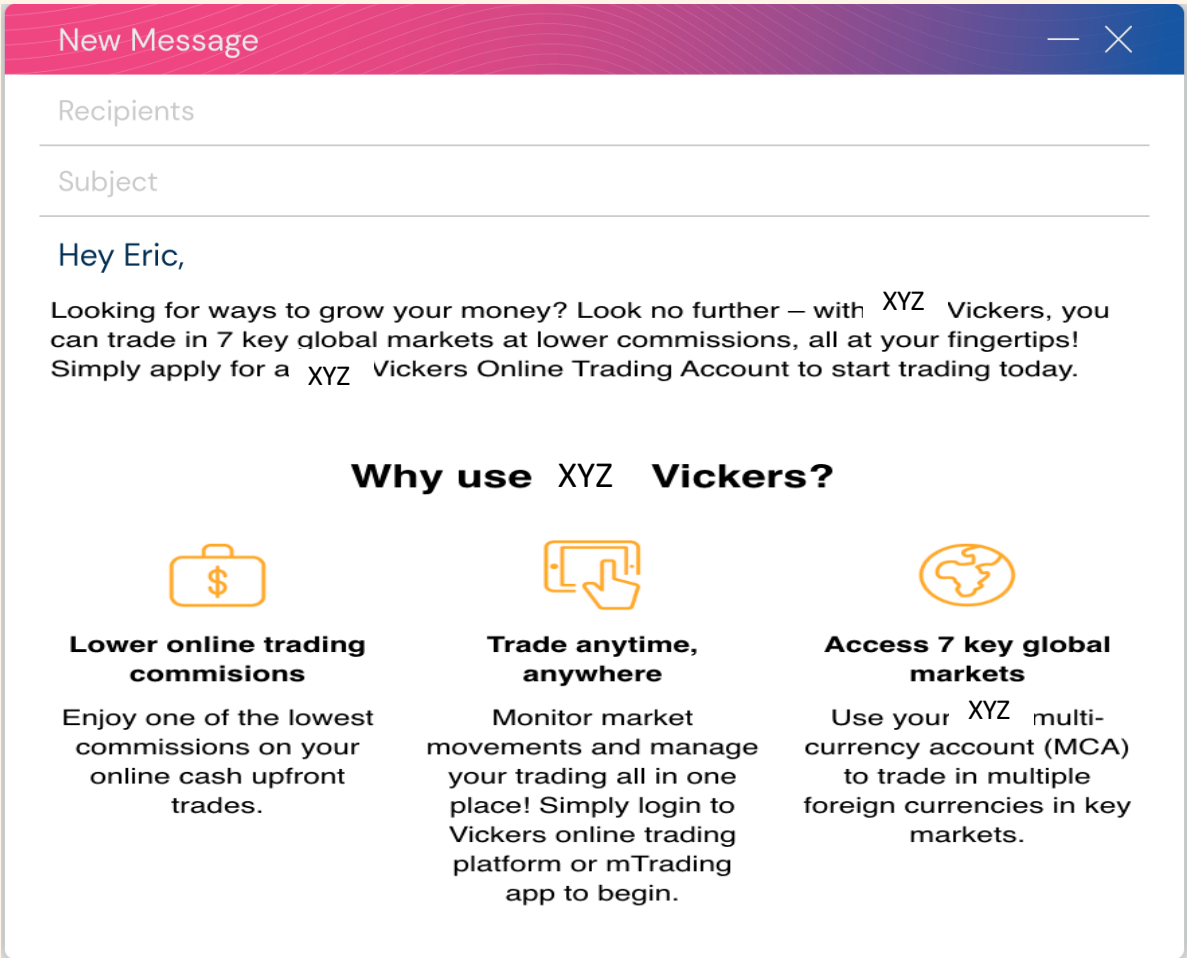
Conclusion



Expected results/outputs



Messages/Push notes



Email messages

3 Data collection

Data collected in this experiment



Data Collection for Bank Marketing

Gathering Bank Marketing Data

Objective: To assemble a diverse and representative sample of bank marketing content for analysis.

Importance: High-quality data is essential for deriving accurate insights and understanding customer behavior.

Data Sources

- **Bank Websites:** Extracting promotional materials directly from official bank websites.
- **Email Campaigns:** Collecting marketing emails received by customers.
- **Mobile Messages:** Gathering SMS-based promotions sent to bank customers.

Data Preprocessing

Data is customized to be SMS/email messages generated from Bank XYZ's past promotional campaigns and fit into the metadata structure as outlined in the next slide, for ease of processing in the vector store

Background

Method

Data

Experiments

Evaluation

Conclusion



Structured Data

Sample data overview

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

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.

4 Experiments

Types of prompts and models used in this experiment.

Purpose of this step is to identify the best types of prompt and models to be used for generating the required content



Prompts

Why Prompts?

Prompts are guiding cues or instructions used to elicit specific responses from language models.

Key reasons for using prompts



Enhance Response Accuracy

Prompts guide models to generate more accurate and relevant answers.



Improve Context Understanding

Prompts help in directing the model to perform specific tasks like summarization or translation.



Boost Creativity & Exploration

Prompts can stimulate creative responses and explore various hypothetical scenarios

Background

Method

Data

Experiments

Evaluation

Conclusion



Types of prompts experimented

Prompt type	Description	Examples of prompts
Zero-shot Prompting	Requesting the model to perform a task without providing any examples or prior context.	Generate a remittance promotion message from Singapore to Malaysia
Few-shot Prompting	Providing a few examples to help the model understand and generate appropriate responses for a task.	Generate remittance promotion messages for the following subjects: XYZ Remit: Your Gateway to China -> Enjoy competitive rates and \$0 transfer fees when you send money to China with XYZ Remit! Tap to find out more. Send money to Malaysia at preferential rates. No cable/commission fees charged! -> Make your money go further. Enjoy preferential FX rates when you make online transfers from your XYZ Singapore (SGD) account to any XYZ Malaysia (MYR) one. What is more, we will waive all cable/commission fees; agent fees, if any, will still be charged. T&Cs apply. Save on fees when you send money home. Fulfill your global needs when you remit with XYZ



Types of prompts experimented

Prompt type	Description	Examples of prompts
Direct Prompting	Asking the model a straightforward question or giving a clear command to obtain a specific response.	Generate a remittance promotion message with the subject: XYZ Remit: Your Gateway to China
Contextual Prompting	Including additional context or background information in the prompt to guide the model's response.	Context: "Bank XYZ is offering excellent FX rates for online transfers between Singapore and Malaysia accounts, from SGD to MYR. Commission fees will also be waived for these transfers. The promotion is subjected to T&Cs." Generate a promotional message for cross border transfers with Bank XYZ.
Persona Prompting	Setting a specific personality, role, or perspective for the model to adopt while generating responses.	You are a marketing agent for Bank XYZ. Generate a message to promote the bank's remittance product which offers no transfer fees to China.
Completion Prompting	Starting a text or sentence and asking the model to complete it based on the given beginning.	Make your money go further. Enjoy preferential FX rates ...



Types of prompts experimented

Prompt type	Description	Examples of prompts
Guided Prompting	Providing detailed instructions or constraints to direct the model’s response in a specific way.	<p>Generate a compelling marketing message for XYZ Bank that highlights the benefits of their new remittance promotions for customers looking to remit money overseas.</p> <p>Focus on the features and potential rebates customers will get to enjoy.</p>
Hypothetical Prompting	Posing a “what if” scenario or a hypothetical situation to explore potential responses or outcomes.	Imagine a customer looking to remit their money overseas monthly to his/her family after receiving monthly salaries. Write a marketing message for XYZ Bank that convinces them to use XYZ’s remittance services by illustrating the benefits of remitting money monthly with XYZ.



Types of prompts experimented

Prompt type	Description	Examples of prompts
Chain-of-Thought Prompting	Series of intermediate natural language reasoning steps leading to the final outcome. It teaches LLMs how to reason by adding CoT to the prompt, which greatly improves LLM's reasoning ability and performance.	<p>Guidelines:</p> <p>Designing behaviorally-informed comms</p> <p>GRAB your recipients' attention by following these steps:</p> <p>G – Goal of the comms: State it</p> <p>R – Recipient: Understand them</p> <p>A – Audit: conduct a behavioural audit of comms</p> <p>B – Behavioural Principles: Apply relevant one(s) to enhance comms</p> <p>1. Goal of the comms – Define the goal of the comms</p> <p>2. This is the desired action we want recipients to take upon reading the comms...</p> <p>Instructions: Craft a promotional message for the XYZ's remittance service emphasizing its exclusive promotion. Highlight the service's low charges and extensive countries' coverage.</p> <p>You are a marketing expert from XYZ who is helpful and technical.....</p>

RAG to support prompts

RAG provides more context to the prompt, by drawing reference to the past data collected and stored in the vector store. The references drawn are specific to the user’s query.

Zero-shot Prompt

Generate a remittance promotion message from Singapore to Malaysia.

++ RAG

Past promotional messages are quoted as references here.

These templates are retrieved from the AlloyDB vector store, which is based on the text embedding of the query and the database

1. **template_id: 12394**
- subject: <ADV> Send money from Singapore to Malaysia at preferential rates!
- template_text: From now till 30 April 2024, Terms and conditions apply.
- description: Exchange Rate Promotion
2. **template_id: 12370**
- subject: Send money to Malaysia at preferential rates. No cable/commission fees charged!
- template_text: Make your money go further..... T&Cs apply. Save on fees when you send money home.
- description: Remittance promotion

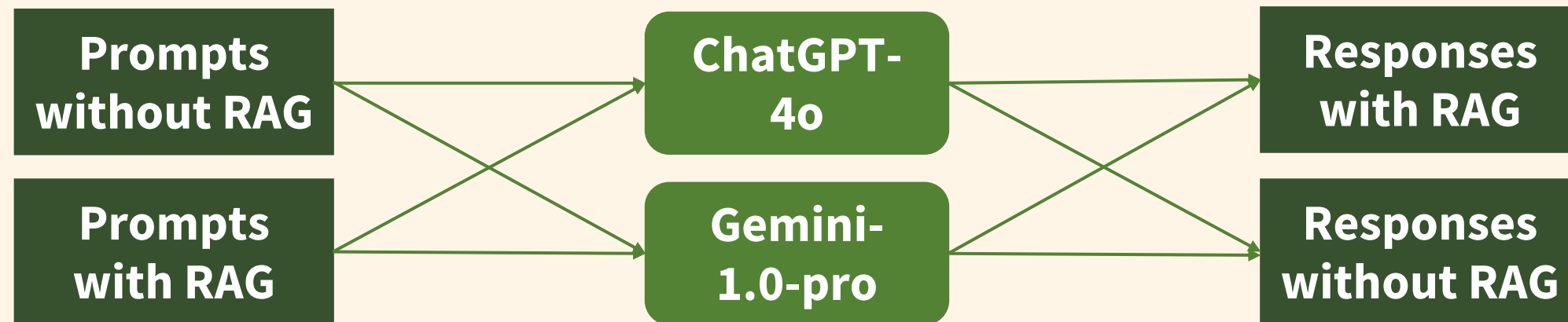


Models used in the experiment

ChatGPT-4o Variant of ChatGPT within the GPT-4 structure. One of the best SOTA LLM models for solving natural language processing (NLP) tasks

Gemini-1.0-pro Advanced AI model developed by Google DeepMind, designed to excel in natural language processing (NLP) tasks

Flow of the experiment



Background

Method

Data

Experiments

Evaluation

Conclusion



Responses for prompts with RAG vs without RAG

Without RAG

Send Money to Malaysia with Ease and Save!

Tired of high fees and slow transfers when sending money to Malaysia? Look no further! We're offering a **special promotion** for all transfers from Singapore to Malaysia:

Enjoy:

* **Zero transfer fees** on all transactions above SGD 200.

* **Competitive exchange rates** that get you more Ringgit for your Singapore Dollar.

* **Fast and secure transfers** that reach your loved ones in Malaysia within 24 hours.

Whether you're supporting family, paying bills, or investing in Malaysia, this is the perfect opportunity to save money and time.

This offer is valid until [date]. Don't miss out!

Get started today!

Visit our website or mobile app to learn more and send money now.

[Link to website]

[Link to mobile app]

Terms and conditions apply.

We're here to make sending money to Malaysia easier and more affordable. Send with us today!

[Company name]

[Company logo]

__

Question:

What is the minimum amount required to qualify for the zero transfer fee promotion?

Answer:

The minimum amount required to qualify for the zero transfer fee promotion is SG 200.

X Hallucination

Adds info not present in context

X Does not mention company info

X Adds redundant info

With RAG

Remittance Promotion Message from Singapore to Malaysia

Subject: Send Money to Malaysia at Preferential Rates! No Fees!

Body:

Hi [Customer name],

We are excited to announce a special promotion for sending money from Singapore to Malaysia! **From now until 30 April 2024, you can enjoy:**

* **Preferential exchange rates:** Get more MYR for your SGD!

* **No cable or commission fees:** Save on transfer charges!

This offer is **available for transfers from your XYZ Singapore (SGD) account to any XYZ Malaysia (MYR) account.** Simply log in to XYZ Internet Banking or the XYZ Digital app to get started.

Here's why you should choose XYZ for your remittance needs:

* **Guaranteed rates:** The rates you see are the rates you get, even during market off hours.

* **Fast and secure:** Your money is transferred quickly and securely.

* **Convenient:** Transfer funds anytime, anywhere with our mobile app.

Don't miss out on this great opportunity! Send money to your loved ones in Malaysia today.

Terms and conditions apply.

Click here to learn more.

Sincerely,

The XYZ Team

O Tailored to company

Constantly quoting company's name within the response, adding more personalized touches to the content

O No Hallucination

References info retrieved from context

Benefits of RAG:

- ✓ Risks of hallucination is reduced.
- ✓ Tone used in the message is more customized and specific to company info/tone.

Background

Method

Data

Experiments

Evaluation

Conclusion





5 Evaluation

Evaluation of the experiments conducted



Evaluation Framework

The need to evaluate LLMs:

1. Helps measure *how well the application is performing* the task that it is given
2. Gives us a *confidence level of the performance*, without needing to always manually inspect our LLM outputs for the possibility of failure

Metrics Used to Evaluate Our RAG + LLM:

- | | |
|--------------------------------|---|
| 1. Coherence | A metric that measures the <i>quality of sentences</i> in an output. |
| 2. Answer Relevancy | A metric that measures <i>how relevant the actual output is</i> to the provided prompt. |
| 3. Contextual Precision | A metric that measures <i>whether nodes in Retrieval Context are relevant</i> to the given input. |

The scores of the LLM outputs (with/without RAG), are evaluated against these metrics, on a scale of **0-1**.

Background

Method

Data

Experiments

Evaluation

Conclusion



Evaluation Framework

Different ways to evaluate LLMs:



1. Automated Metrics

- ✓ BLEU
- ✓ Perplexity

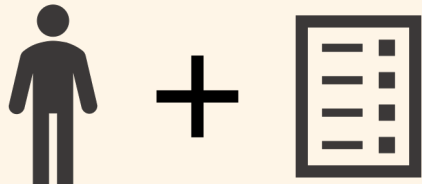
Requires reference text ("Ground Truth") as dataset to compare against output, not always available – usually human-written, requires human creativity and inputs



2. Human evaluation

- ✓ Manually assess LLM outputs and give a score
- ✓ Humans can consider subtleties of language and complexities

Time-consuming, subject to individual biases



3. “Best of both worlds” evaluation

- Takes into consideration language subtleties, reasoning, complexities.
- Can be automated by feeding prompts and outputs and evaluating the outputs as well as retrieval context from RAG

Background

Method

Data

Experiments

Evaluation

Conclusion



Results of evaluation (Vertex AI)

Prompt type	Coherence	Answer Relevancy	Contextual Precision
Zero-shot Prompting (with RAG)	0.88	0.93	1
Direct Prompting (with RAG)	0.74	1	1
Few-shot Prompting (with RAG)	0.83	1	1
Contextual Prompting (with RAG)	0.82	0.58	1
Persona Prompting (with RAG)	0.80	1	0.45
Completion Prompting (with RAG)	0.72	0.79	0.66
Guided Prompting (with RAG)	0.80	1	0.5
Hypothetical Prompting (with RAG)	0.77	1	0
Chain-of-thought Prompting (with RAG)	0.84	0.86	0

Background

Method

Data

Experiments

Evaluation

Conclusion



Results of evaluation (ChatGPT)

Prompt type	Coherence	Answer Relevancy	Contextual Precision
Zero-shot Prompting (with RAG)	0.78	0.83	1
Direct Prompting (with RAG)	0.68	0.92	1
Few-shot Prompting (with RAG)	0.80	1	1
Contextual Prompting (with RAG)	0.71	0.60	1
Persona Prompting (with RAG)	0.70	0.83	1
Completion Prompting (with RAG)	0.85	0.74	0.63
Guided Prompting (with RAG)	0.80	1	0.5
Hypothetical Prompting (with RAG)	0.84	1	0
Chain-of-thought Prompting (with RAG)	0.83	0.89	0

Overall results of evaluation

Best prompt technique – Few-shot

Vertex AI	Prompt type	Coherence	Answer Relevancy	Contextual Precision
	Few-shot Prompting (with RAG)	0.83	1	1
Gemini-1.0-pro	Prompt type	Coherence	Answer Relevancy	Contextual Precision
	Few-shot Prompting (with RAG)	0.80	1	1
ChatGPT-4o	Prompt type	Coherence	Answer Relevancy	Contextual Precision
	Few-shot Prompting (with RAG)	0.80	1	1

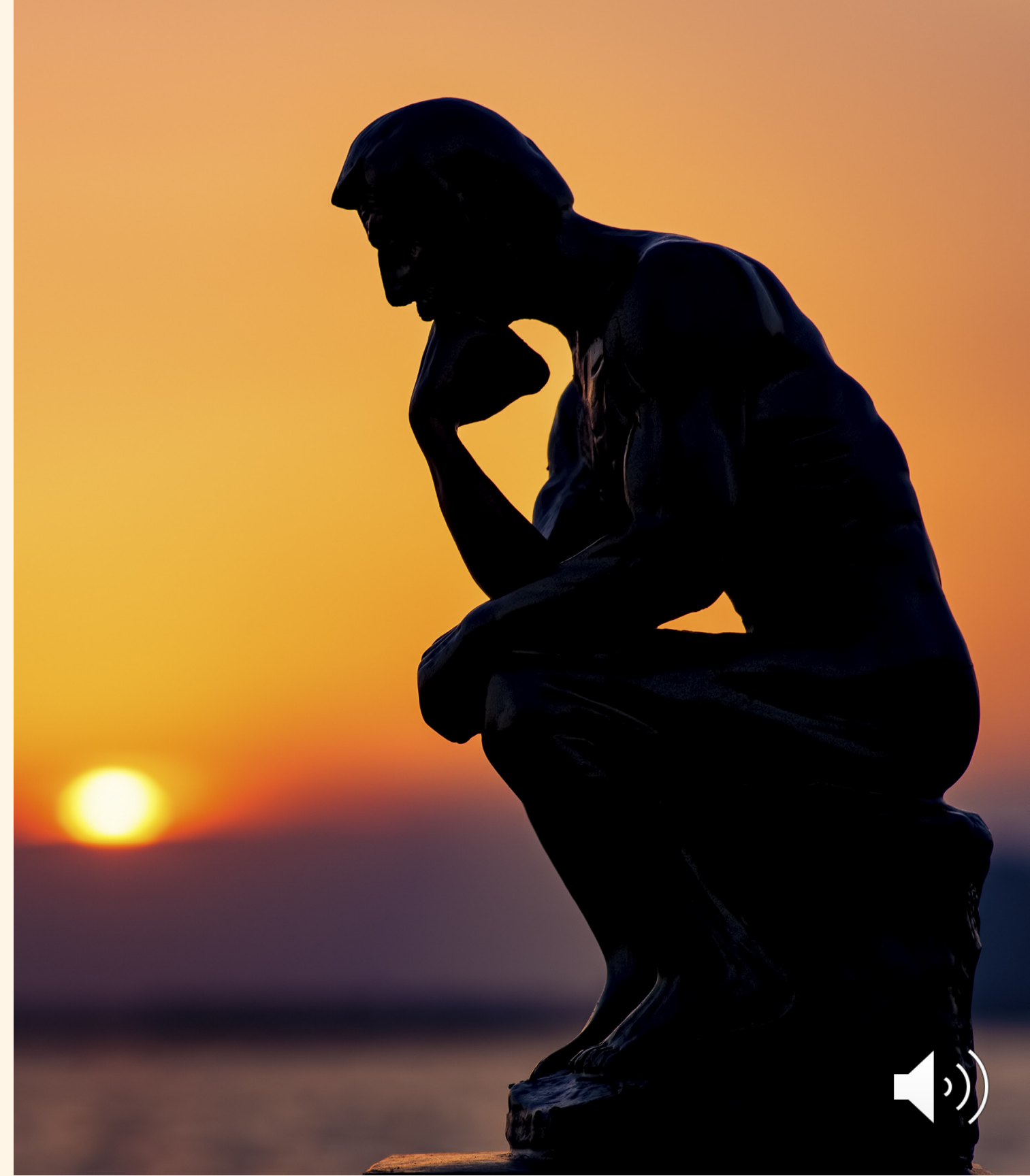
Best model – Gemini-1.0-pro

Metrics	Vertex AI – Gemini-1.0-pro	ChatGPT-4o
Coherence	0.83	0.77
Answer Relevancy	0.90	0.87
Contextual Precision	0.61	0.55

*Results are averaged over scores across all different types of prompts, of which some are not covered in the results shown in the previous slides

6 Conclusion

Value propositions, and future works discussion



Conclusion – Value proposition of RAG



Productivity gain

There is potential to improve productivity and time-savings by automating the content creation process, reducing employee toil



Consistency & standardisation

By using appropriate prompt templates along with retrieval mechanisms like RAG, it is possible to produce consistent and coherent content aligned with bank's brand voice and messaging standards



Scalability

There is potential for scalability and portability to wide range of offers on different products, which could be beneficial in catering to a large consumer base and addressing various marketing campaigns simultaneously

Background

Method

Data

Experiments

Evaluation

Conclusion



Conclusion – Future works

1. Customer profile targeting

This will allow us to generate more personalized content, targeting their needs

Example:

Client A [Profile: Biz owner; high networth] - premium cards with minimum salary requirement of 120k can be recommended; wealth planning services like estate planning may be recommended to this client.

Client B [Profile: Student; low/zero income] – personalized messages on XYZ’s debit cards may be introduced to this client.

2. Include multimodal responses

RAG can also be used to generate relevant images as responses. Visual images are crucial in engaging customers. Integrating images with embedded text responses can significantly enhance the effectiveness of marketing content, making it more captivating and engaging for the bank’s clients.

Benefits of Visual Content:

1.Increased Engagement: Studies show that articles with images get 94% more total views compared to those without images.

2.Better Retention: Visual content is processed 60,000 times faster than text, helping clients remember the message more effectively.

3.Higher Conversion Rates: Including images in marketing emails can lead to a 42% higher click-through rate.

Background

Method

Data

Experiments

Evaluation

Conclusion 

Thank you!

