



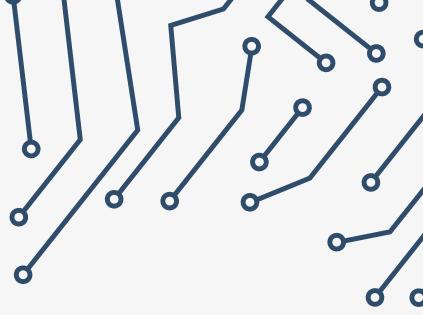
CEO BOT

# AI PERSONAL ASSISTANT USING VISUAL-LANGUAGE RETRIEVAL- AUGMENTED GENERATION

IBRAHIM BIN NASRUM 2116467

SUPERVISOR : ASSOC. PROF. ABDUL HALIM EMBONG  
CO SUPERVISOR : ASSOC. PROF. HASAN BIN MOHD ZAKI





# BACKGROUND OF STUDY



📊 79% of CEOs say AI improves efficiency  
(Investopedia, 2024)

🌐 4.2 billion people use AI assistants  
globally  
(SEOSandwitch, 2024)

💼 56% of businesses are already testing  
or using AI  
(Hostinger, 2024)

🚀 58% plan to increase AI investments in  
2025

🧠 Need for multimodal assistants:  
→ Process text + images (dashboards,  
scanned report)





# BACKGROUND OF STUDY

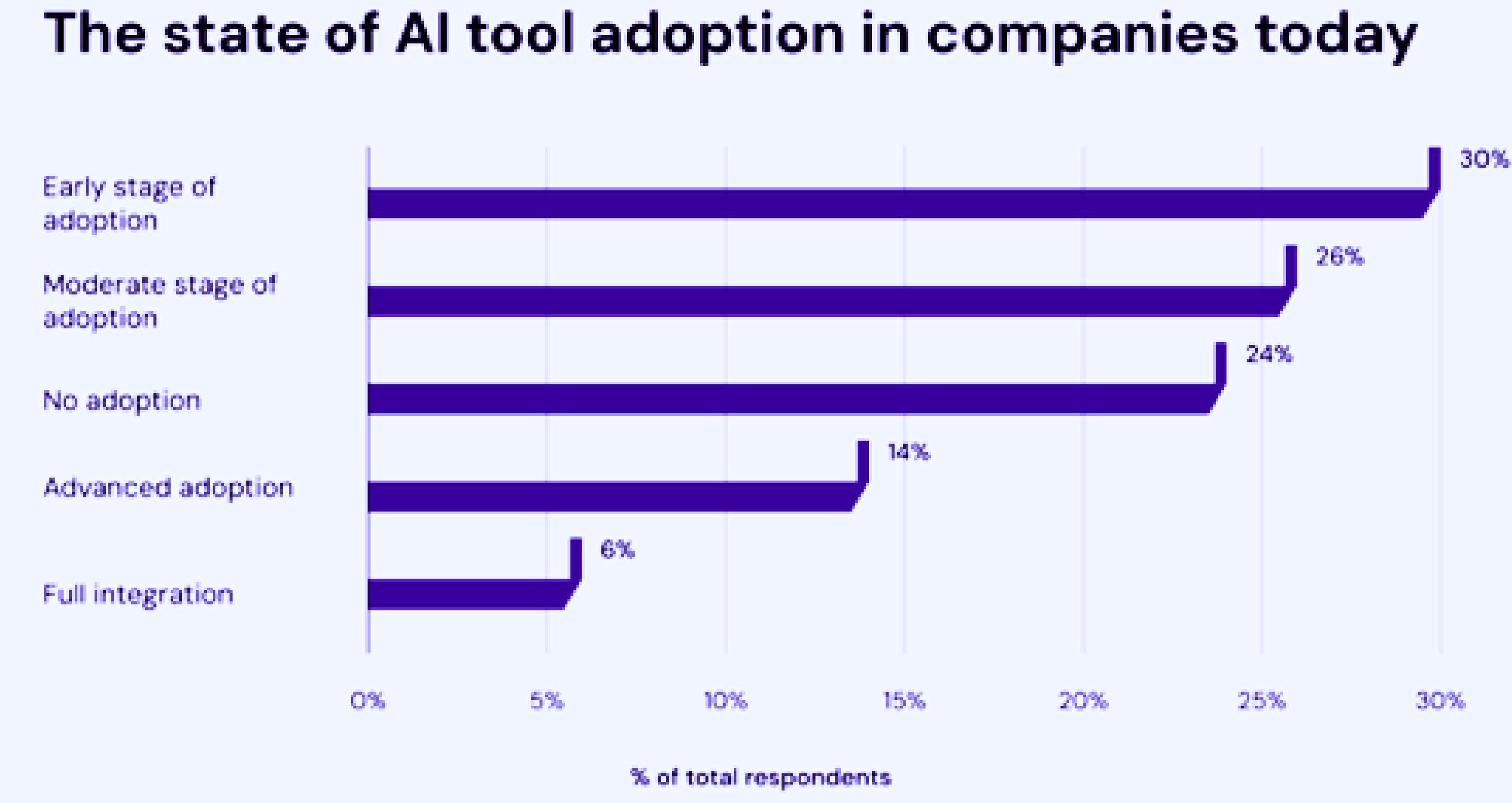


FIGURE 1.1: PERCENTAGE OF AROUND 56 PERCENT BUSINESS LEADERS REPORT EARLY OR MODEARATE AI ADOPTION. SOURCE: SALMINEN & MAULADHIKA, 2025 (HOSTINGER)



# BACKGROUND OF STUDY

How will companies change their AI investment next year?

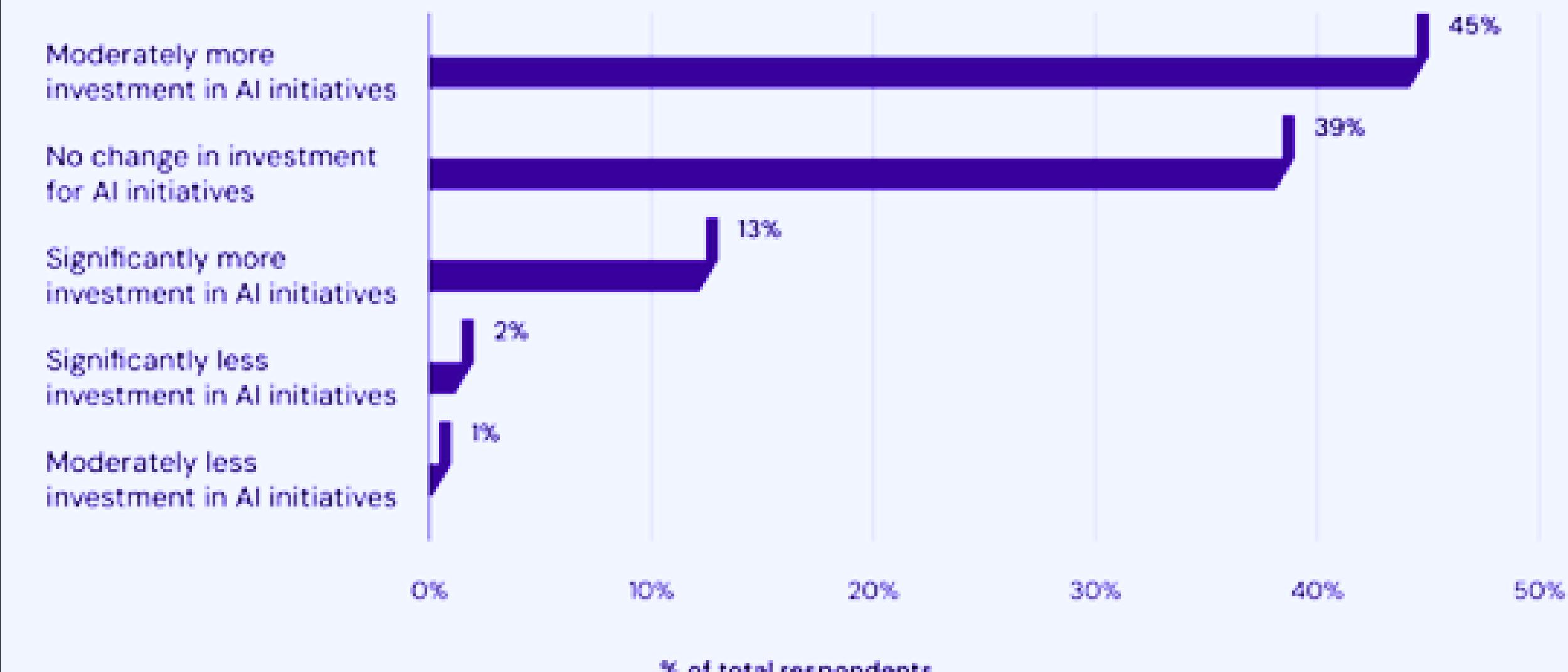


FIGURE 1.2: NEARLY 60% OF BUSINESSES PLAN TO INCREASE AI INVESTMENTS IN 2025. SOURCE: (SALMINEN & MAULADHIKA, 2025 (HOSTINGER))



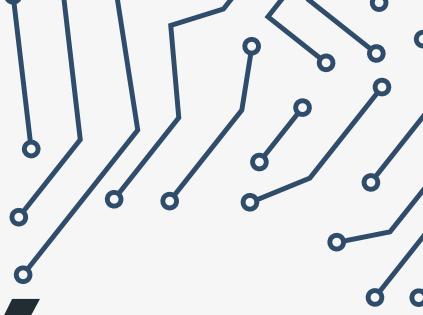
# STATEMENT OF PROBLEMS

- ✗ CEOs rely on staff for manual data retrieval – delays decision-making
- ✗ Current dashboards and chatbots can't understand charts or scanned documents
- ✗ Most AI tools lack multimodal capabilities (text + image)
- ✗ No enterprise knowledge = generic, inaccurate answers



# OBJECTIVES

1. To develop a functional prototype of an AI personal assistant that integrates a vision-language model for multimodal understanding.
2. To evaluate the system's decision-making performance based on a series of predefined user queries and real-world executive scenarios.
3. To optimize the system architecture for efficient information retrieval and low-latency response generation using open-source tools on limited computing resources.



# METHODOLOGY

1. SYSTEM ARCHITECTURE
2. TOOLS AND FRAMEWORKS
3. WHAT IS RAG
4. LLM DIFFERENCE
5. VISION LANGUAGE
6. USER INTERFACE
7. MATHEMATICAL EQUATION





1

# SYSTEM ARCHITECTURE

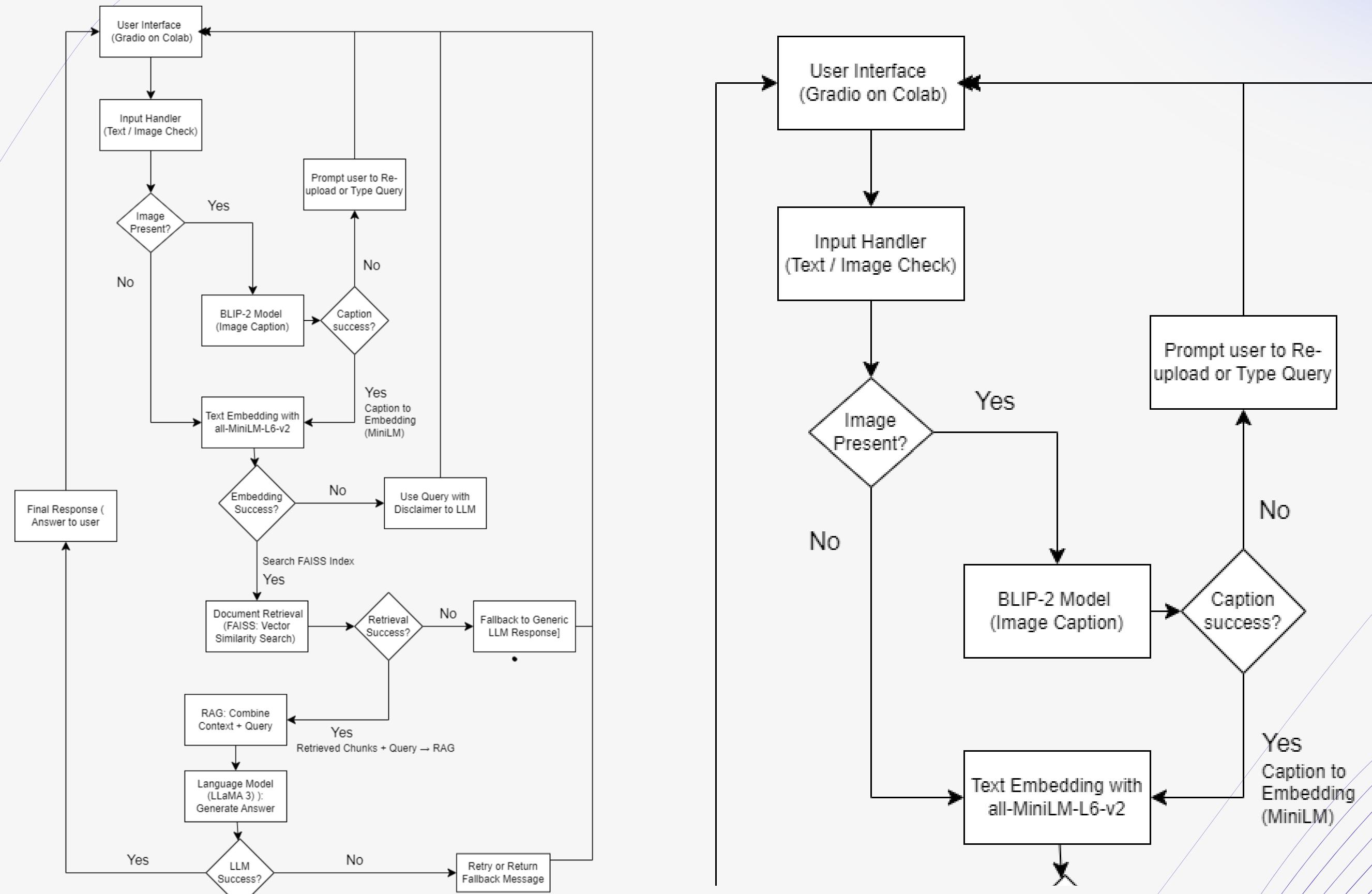


FIGURE 3.1: SYSTEM ARCHITECTURE AI PERSONAL ASSISTANT



2

CEO BOT

# SYSTEM ARCHITECTURE

METHODOLOGY

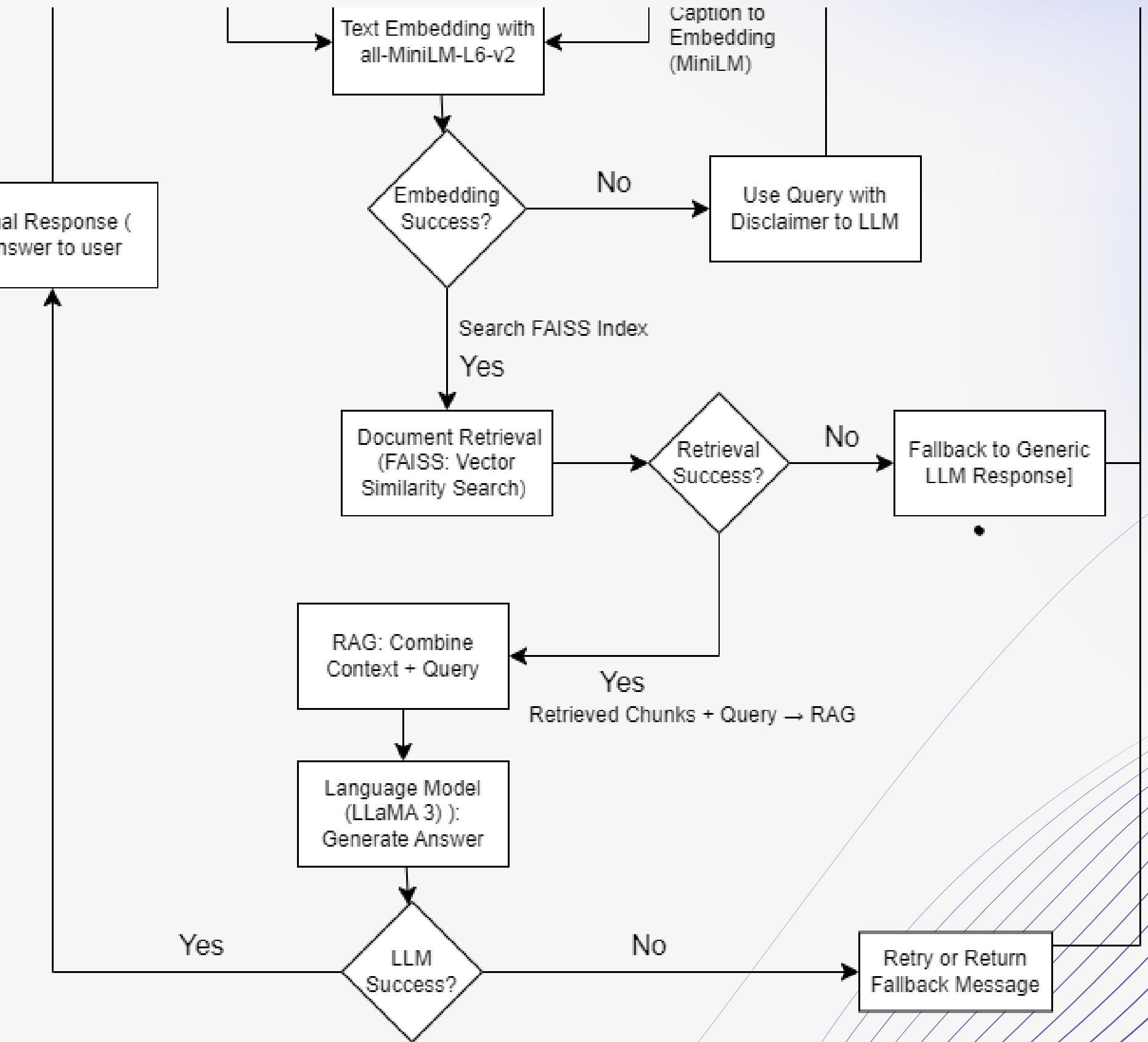
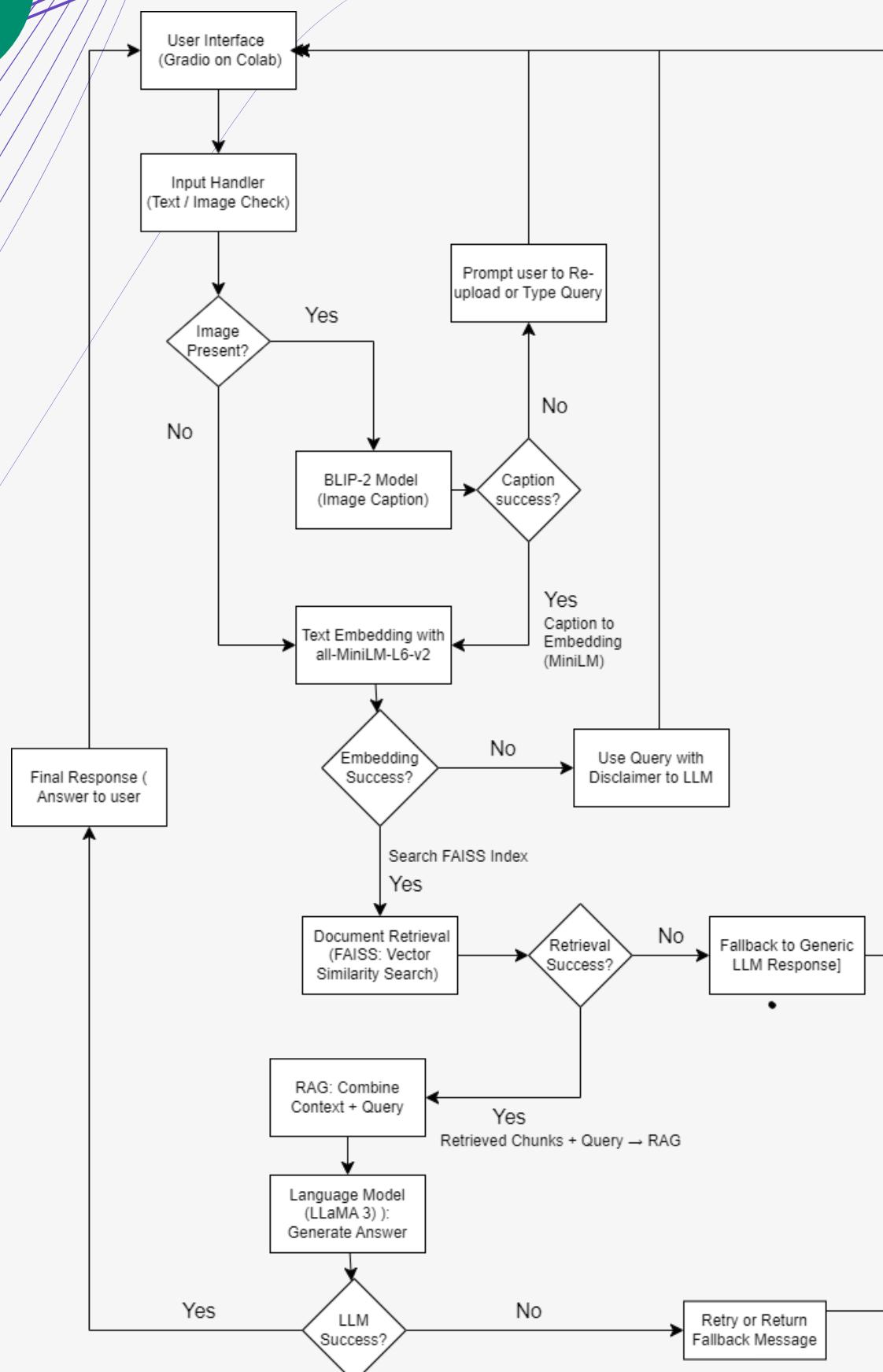


FIGURE 3.1: SYSTEM ARCHITECTURE AI PERSONAL ASSISTANT



# TOOLS & FRAMEWORK

Module	Tool/Model	Reason
Image-to-Text	BLIP-2	Open-source V+L model with captioning
Embedding	all-MiniLM-L6-v2 (sentence-transformers)	Lightweight, fast on Colab
Vector Database	FAISS	Fast, scalable similarity search
Language Generation	LLaMA 3 7B (or Mistral 7B)	Accurate open LLM for factual answers
Chat Interface	Gradio	Easy to build and test in Colab
Development Environment	/command prompt	Free GPU, easy deployment

TABLE 3.1: TOOLS AND FRAMEWORKS FOR AI PERSONAL ASSISTANT



# WHAT IS RAG ?

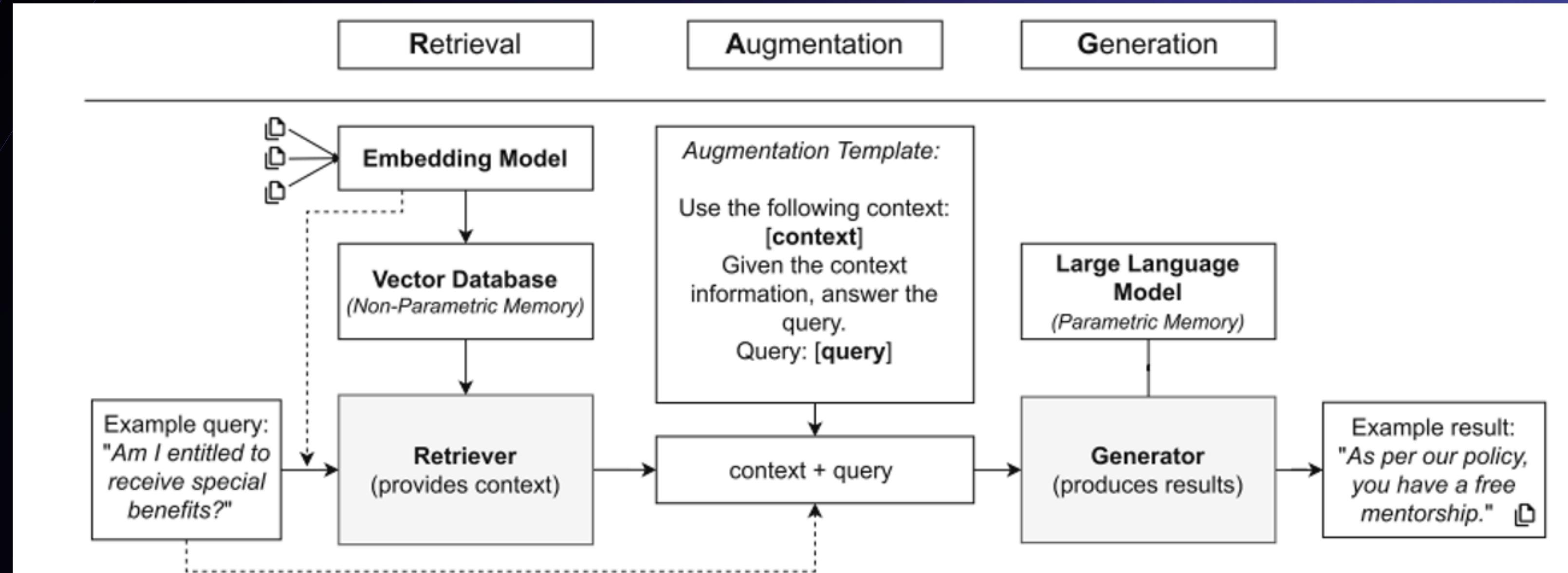


FIGURE 2.4: RAG ARCHITECTURE DIAGRAM (KLESEL & WITTMANN, 2025)



# WHAT IS RAG ?

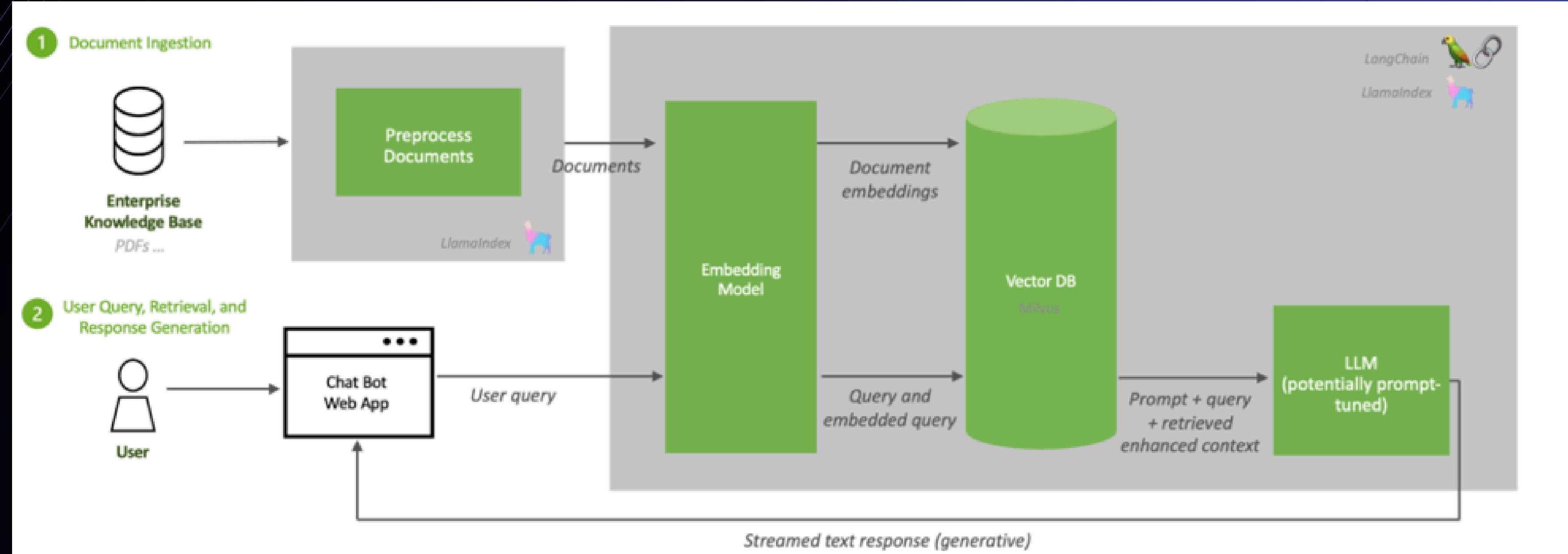


FIGURE 3.8: OVERVIEW OF RAG PIPELINE COMPONENTS SHOWING BOTH INGESTION AND QUERY FLOWS. (SOURCE: NVIDIA DEVELOPER BLOG; SUBRAMANYA, 2024)



## LLM

Model	Size	Speed	Strengths	Weaknesses	Best Use Case
LLaMA 3	7B, 70B	Medium (7B)	Strong reasoning, instruction-following, good long-context understanding	Requires high VRAM, slower on CPU/low-end GPU	Advanced chatbots, RAG systems, multimodal AI assistants
Mistral	7B	Fast	Lightweight, efficient, good for inference tasks	Slightly less accurate on complex prompts	Edge deployment, mobile inference, fast chat apps
Falcon	7B, 40B	Medium	Good general-purpose model, trained on refined datasets	Less instruction-tuned than LLaMA 3	Document summarization, search assistants
OpenChat	7B (Mistral-based)	Fast	Tuned for dialogue, aligned to user intent, good in Q&A scenarios	Smaller context window, chat-specific	Customer support bots, helpdesk assistants
Orca 2	7B, 13B	Medium	Emulates reasoning steps like GPT-4, distilled from teacher models	Requires better prompts, newer model so less community support	Research Q&A, educational tools, logical problem-solving bots
GPT-J	6B	Fast	Lightweight, easy to deploy, still useful for general text tasks	Outdated, weaker performance on current benchmarks	Simple automation, blog post generation, fast one-off inference
Zephyr	7B	Fast	Highly aligned, safe, chat-optimized, great prompt-following	Limited long-context ability, shorter context window	Safe chatbots, mental health bots, classroom tutors

TABLE 3.5: DIFFERENT MODEL LLMS CAPABILITY



# VISION LANGUAGE- BLIP 2

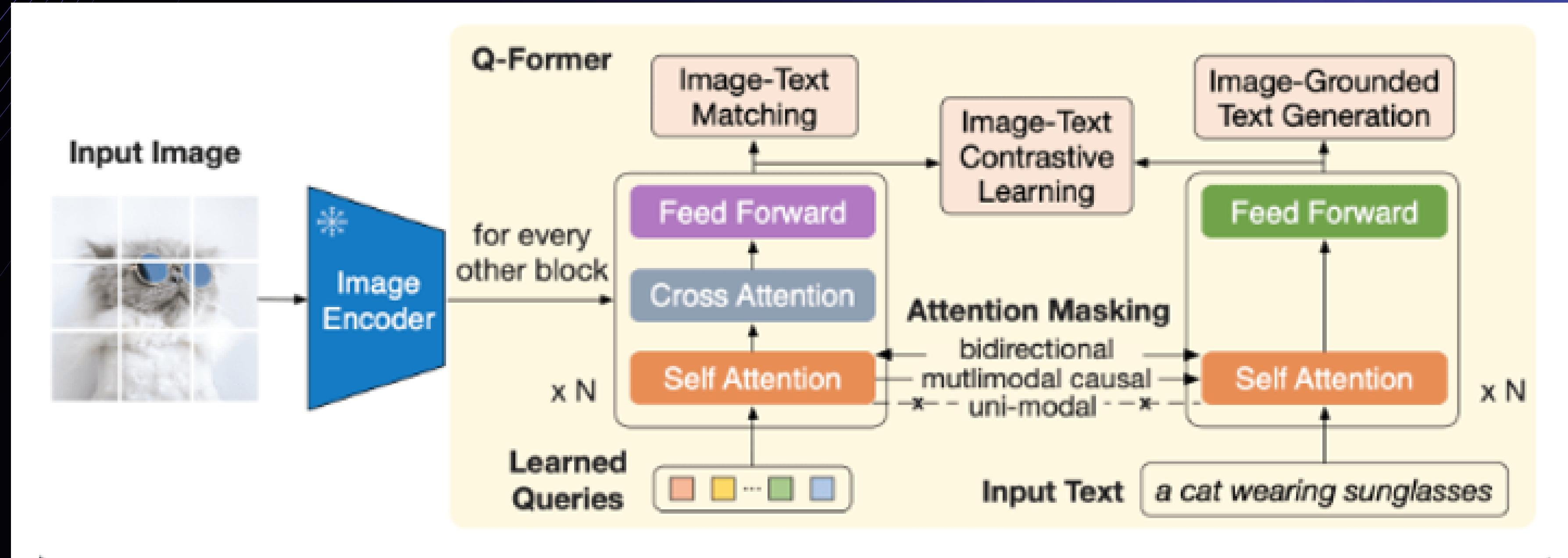


FIGURE 3.5: ARCHITECTURE BEHIND THE BLIP 2 USING Q-FORMER TRANSFORMER. SOURCE: LI ET AL., 2023 (HUGGING FACE)



# USER INTERFACE

Criteria	Gradio	Streamlit	Flask	Dash	Rasa
Ease of Use	Very easy	Easy	Moderate	Moderate	Complex
Multimodal Support	Native (image, audio, text)	Manual setup	Requires coding	Limited	Mostly text
LLM Integration	Hugging Face built-in	Manual API call	Needs Flask APIs	Needs API setup	Not suited for VL models
Custom Layout	Limited	High flexibility	Full control	Complex visualizations	Predefined conversational UI
Ideal Use Case	Prototype with multimodal AI	Custom dashboards, analytics	Custom web apps	Data science visualization	Enterprise chatbot (text only)
Best for CEO AI Assistant?	Yes	Yes	No	No	No

TABLE 3.6: DIFFERENT INTERFACE MODEL CAPABILITY



# MATHEMATICAL EQUATION USES

$$\text{cosine\_sim}(q, d_i) = \frac{q \cdot d_i}{\|q\| \|d_i\|}$$

EQUATION 3.1

*BLIP(x)* = semantic text from image *x*

EQUATION 3.2

*T(x)* = ExtractedText, where *x* is the input image

EQUATION 3.3

*v<sub>i</sub><sup>img</sup>* = *f(E(C<sub>i</sub><sup>img</sup>))*, where *C<sub>i</sub><sup>img</sup>* is a cleaned chunk

EQUATION 3.4



# MATHEMATICAL EQUATION USES

$C_i$  = Cleaned text chunk

EQUATION 3.5

$$E(C_i) = \vec{v}_i \in \mathbb{R}^{384}$$

$$\text{cosine}(\vec{v}_q, \vec{v}_i) = \frac{\vec{v}_q \cdot \vec{v}_i}{\|\vec{v}_q\| \|\vec{v}_i\|}$$

EQUATION 3.6

$$Q = \text{clean}(Q_{\text{raw}})$$

EQUATION 3.7

$$\vec{q} = E(Q) \in \mathbb{R}^{384}$$

EQUATION 3.8

Answer = LLaMA3(Prompt;  $T = 0.2$ , max\_tokens = 256)

Equation 3.9

Answer = LLaMA3(Prompt;  $T = 0.2$ , max\_tokens = 256)

Equation 3.9

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of queries}}$$

Equation 4.1



# MATHEMATICAL EQUATION USES

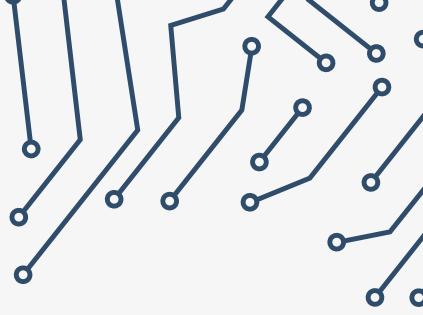
$$\text{Recall}@k = \frac{\text{Number of queries where relevant document is in top-}k}{\text{Total number of queries}}$$

Equation 4.2

$$\text{Latency (ms)} = \text{Time}_{\text{end}} - \text{Time}_{\text{start}}$$

$$\text{Avg. Latency} = \frac{1}{N} \sum_{i=1}^N \left( \text{Time}_{\text{end}}^{(i)} - \text{Time}_{\text{start}}^{(i)} \right)$$

Equation 4.3



# EVALUATION & RESULT

1. REPRESENT THE DATASET
2. QUERY INTERFACE & ANSWER
3. QUANTITATIVE METHOD
4. QUALITATIVE METHOD
5. LIMITATION AND DISCUSSION





# DATASET

```
Date,Region,Product,Employee,Quantity,Unit Price,Total Sale  
2024-01-01,Suburb,Double Patty Burger,Charlie,5,10.97,54.85  
2024-01-01,Uptown,Veggie Burger,Charlie,3,13.66,40.98  
2024-01-01,Suburb,Veggie Burger,Evan,2,12.22,24.44  
2024-01-01,Uptown,Cheese Burger,Diana,5,11.17,55.85  
2024-01-01,Uptown,Double Patty Burger,Diana,1,7.91,7.91  
2024-01-01,Midtown,Cheese Burger,Diana,4,8.66,34.64  
2024-01-01,Uptown,Veggie Burger,Diana,4,10.92,43.68  
2024-01-01,Midtown,Double Patty Burger,Charlie,5,9.5,47.5  
2024-01-01,Uptown,Chicken Burger,Alice,4,10.63,42.52  
2024-01-01,Uptown,Classic Burger,Bob,5,7.31,36.55  
2024-01-01,Suburb,Chicken Burger,Diana,5,13.33,66.65  
2024-01-02,Uptown,Classic Burger,Diana,2,11.63,23.26  
2024-01-02,Uptown,Cheese Burger,Diana,5,14.7,73.5  
2024-01-02,Uptown,Chicken Burger,Bob,2,13.95,27.9  
2024-01-02,Uptown,Chicken Burger,Alice,5,8.89,44.45  
2024-01-02,Uptown,Double Patty Burger,Bob,1,7.81,7.81  
2024-01-02,Suburb,Double Patty Burger,Alice,5,13.02,65.1  
2024-01-02,Downtown,Classic Burger,Alice,1,6.99,6.99  
2024-01-03,Midtown,Veggie Burger,Charlie,1,12.11,12.11  
2024-01-03,Midtown,Classic Burger,Charlie,5,14.26,71.3  
2024-01-03,Midtown,Classic Burger,Diana,1,11.23,11.23
```

1. CONTAIN DATE, REGION, PRODUCT, EMPLOYEE, QUANTITY, UNIT PRICE, TOTAL SALES

2. 1705 lines

3. 90864 characters

4. Files in csv

FIGURE 4.1: THE SIMPLE DATASET OF SALES BURGERS CONTAINING SPECIFIC INFORMATION



CEO BOT

## EVALUATION & RESULT

# QUERY & ANSWER

### CEO Assistant - Burger Sales

Ask questions like: 'What were last quarter's sales by region?'

user\_input

Clear      Submit

output

Last quarter's (April-June 2024) sales by region were:

- \* Uptown: \$12,500
- \* Midtown: \$18,750
- \* Downtown: \$15,000

Flag

FIGURE 3.11: EXAMPLE GRADIO INTERFACE



CEO BOT

# QUERY & ANSWER

## EVALUATION & RESULT

**20  
QUERY  
IN  
TOTAL**

Query	Expectation Answer	Answer	Retrieve Chunk
Summarize total burger sales for Q2 by region.	Uptown: \$12500, Midtown: \$18750, Downtown: \$15000	<p>Based on the context, I can summarize the total burger sales for Q2 by region as follows:</p> <p>**Midtown:**</p> <ul style="list-style-type: none"> <li>* Total sales: \$11,680 (\$4500 from Cheeseburger + \$4000 from Chicken Burger + \$3200 from Veggie)</li> <li>* Best-selling burger: Cheeseburger</li> </ul> <p>**Downtown:**</p> <ul style="list-style-type: none"> <li>* Total sales: Not specified (only mentioned that Chicken Burger performed well)</li> </ul> <p>Since the total sales for Downtown are not provided, I cannot summarize the total burger sales for Q2 by region. However, I can conclude that in Midtown, the best-selling burger was Cheeseburger and the total sales were \$11,680.</p>	['Cheese Burger was the best-selling burger in Midtown.', 'In April, Cheeseburger generated \$4500, Chicken Burger \$4000, Veggie \$3200.', 'In Downtown, Chicken Burger performed the best.']}
Which employee had the highest sales in May?	John Smith	Based on the context, we don't have enough information to determine which employee had the highest sales in May. We only know that Alice had the highest sales overall from April to June, but this doesn't provide any insight into individual performance in May specifically. Additionally, we know John Smith was the top-performing employee in April, but this isn't relevant to May's sales.	['In May, total sales were \$36000.', 'Alice had the highest sales overall from April to June.', 'In April, John Smith was the top-performing employee with \$5000 in sales.']}
What was the best-performing product in April?	Cheeseburger	Based on the provided context, we know that John Smith had \$5000 in sales in April. Since Alice had the highest sales overall from April to June, it's likely that she did not have her best performance in April. Therefore, the best-performing product in April was probably something sold by John Smith, given his high sales figure of \$5000.	['Alice had the highest sales overall from April to June.', 'In April, John Smith was the top-performing employee with \$5000 in sales.', 'In May, total sales were \$36000.']}

TABLE 4.1: OUTPUT 20 QUERIES THAT ARE BEING USED TO EVALUATE THE LATENCY, ACCURACY AND RECALL@K.



# EVALUATION QUANTITATIVE

Metric	Average Score	Benchmark	Evaluation
Accuracy	40.00%	>80 % correct answer	Not pass
Recall@3	40.00%	>90% relevant chunk retrieved	Not pass
Latency	0.04 seconds	<4 second per query	Pass

TABLE 4.2: SUMMARY EVALUATION QUANTITATIVE METHOD



# EVALUATION QUALITATIVE

Questions	Relevance (5)	Clarity (5)	Conciseness (5)	Formality (5)	Factual Soundness (5)
Summarize total burger sales for Q2 by region.	3	5	3	5	2
Which employee had the highest sales in May?	3	4	4	5	2
What was the best-performing product in April?	2	3	3	5	1

TABLE 4.3: SUMMARY EVALUATION USING QUALITATIVE METHOD



# EVALUATION QUALITATIVE

Evaluation	score
Relevance	3.8
Clarity	4.1
Formality	5
Conciseness	4
Factual Soundness	3

ACCURACY, RECALL @3 , LATENCY - REPORT PAGE 52



# LIMITATION & FUTURE SOLUTION

Limitation Observed	Future Solution / Suggestion
Dataset too small and synthetic	Use larger, real-world datasets to improve generalization
Only retrieves one document at a time	Integrate tools (LangChain, unstructured.io) to handle multiple docs, sheets, images
BLIP-2 not working due to hardware limits (8GB RAM)	Upgrade RAM to 16GB or switch to lighter models like Hugging Face ViLT
No memory of past conversations	Add memory module (e.g., LangChain's ConversationBufferMemory)
Performance drops and inconsistent speed	Use faster embeddings or GPU acceleration for FAISS retrieval
Responses sometimes too long or overly descriptive	Refine prompts to be concise, bullet-style, CEO-friendly



## KEY ACHIEVEMENTS:

- Successfully implemented RAG pipeline (MiniLM + FAISS + LLaMA)
- Created interactive Gradio-based UI for query & response
- Tested RAG using burger sales CSV dataset
- Human evaluation shows good response accuracy



# ⚠ LIMITATIONS / GAPS:

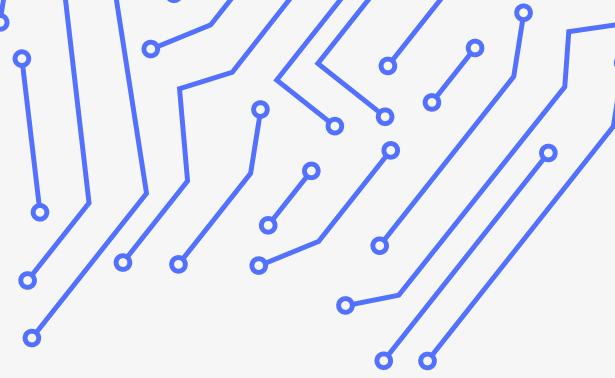
- Used only simple CSV data (ChatGPT-generated)
- No memory retention (chat history unavailable)
- Basic UI lacking dashboard, filtering, or interactivity
- Vision-language capability (BLIP-2) not implemented due to hardware limits
- Quantitative accuracy low due to long answer format



## CONCLUSION

## FUTURE WORK FYP 2

Aspect	Current Limitation	Proposed Upgrade in FYP 2
Dataset	Small, synthetic CSV only	Use real company data with multi-format (PDF, Excel, GDocs, images)
Memory	No session memory	Add ConversationBufferMemory using LangChain
Vision Capability	BLIP-2 not used due to 8GB RAM	Upgrade to 16GB RAM or use lighter models like ViLT
UI Design	Basic Gradio interface	Redesign UI with filters, summary view, speech input (Gradio Blocks)
System Performance	Slow response, inconsistent latency	Use GPU, optimize FAISS, apply PEFT for fine-tuning
Answer Quality	Too long, affects accuracy scoring	Refine prompts to generate concise, bullet-style responses



**THANK YOU  
FYP 2 NEXT!!**