UEMF
الجامعة الأورومتوسطية بفاس
ⵜⴰⵡⵔⵉⴱⵜ ⵜⴰⵎⴰⵡⴰⵙⵜⴰⵢⵜ ⵏ ⴼⴰⵙ
EUROMED UNIVERSITY OF FES

litis

EIDIA
École d'Ingénierie Digitale et d'Intelligence Artificielle

GRADUATION PROJECT OF ENGINEERING PROGRAM (PFE)

Development of a real time AI voice assistant for multisensory exploration of artworks for visually impaired people

THE INTERNSHIP WAS CONDUCTED AT LITIS LABORATORY OF ROUEN UNIVERSITY

Host university supervisors:

- Dr. Christèle LECOMTE
- Dr. Katerine ROMEO
- Pr. Cecilia ZANNI-MERK
- Dr. Abdellatif Ennaji

University supervisor:

- Pr. Alae AMMOUR

The presenter student:

- Imad-Eddine NACIRI

ACADEMIC YEAR : 2023-2024

OUTLINE :

Introduction :

1. PROJECT CONTEXT:

> أمة لا تعرف تاريخها , لا تحسن صياغة مستقبلها
>
> **A nation that does not know its history cannot shape its future properly.**
>
> -- Arabic quote

Introduction :

1. PROJECT CONTEXT:

Bayeaux Museum in Normandy region of France

Bayeaux tapestry

Introduction :

1. PROJECT CONTEXT:

IMG(Inclusive Museum Guide) 2021 – 2025

The main defined objectives of this initiative at its inception are:

- Make museums accessible to all.
- Contribute to the promotion and accessibility of cultural heritage.
- Create in any visitor new attractive impressions.

litis

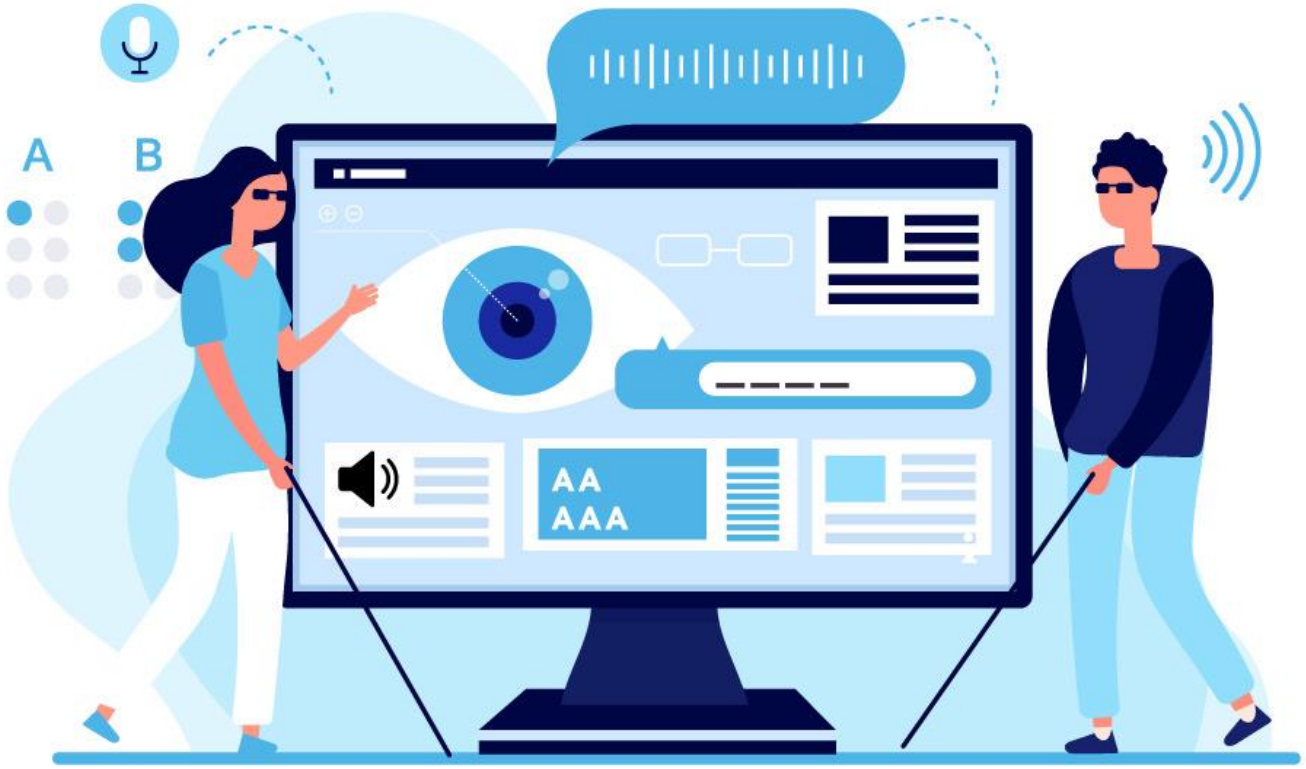
GREYC
Laboratoire de recherche en sciences du numérique

ihrim

Introduction :

2. PROJECT OVERVIEW:

Develop a real time AI voice assistant that can answer questions and provide information on the content of images of Bayeaux artworks to guide visually impaired people.



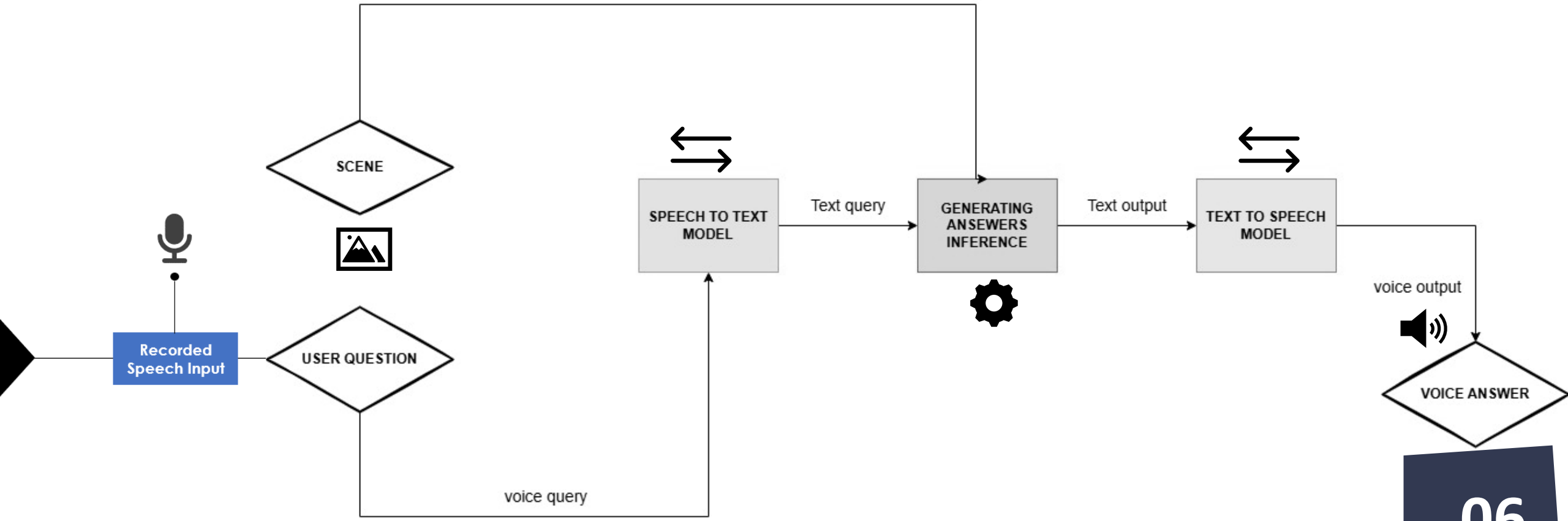
Introduction :

2. PROJECT OVERVIEW:

Example of scenes from the Bayeux tapestry

Introduction :

2. PROJECT OVERVIEW:



Project Objectives :

1. Develop a robust, free, and open-source conversational agent to provide scene descriptions to users based on their requests.

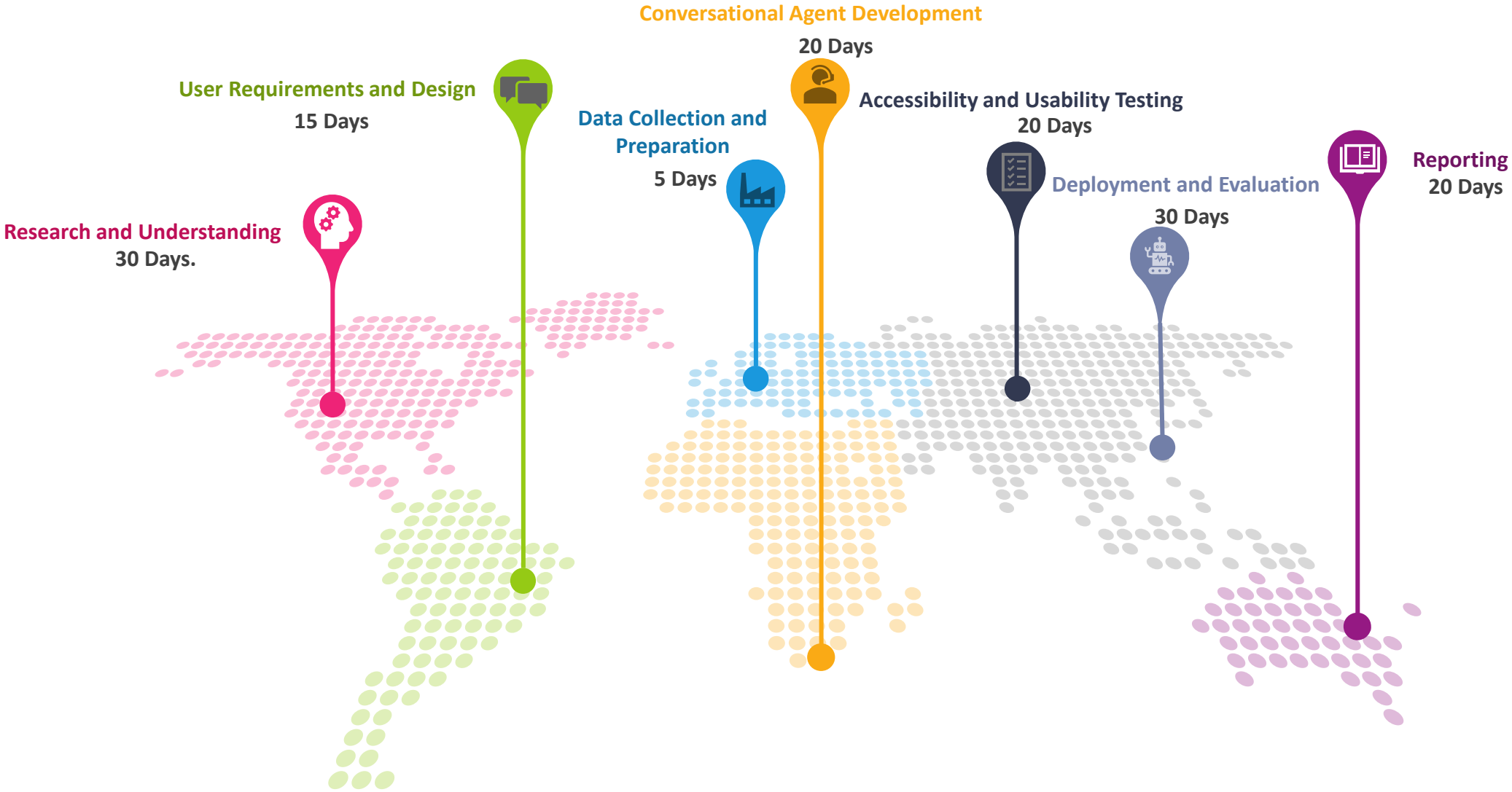
2. Evaluate the performance of the developed agent using a set of commonly used benchmarks.

3. Enhance the quality of the system by continually gathering user feedback and making iterative design improvements.

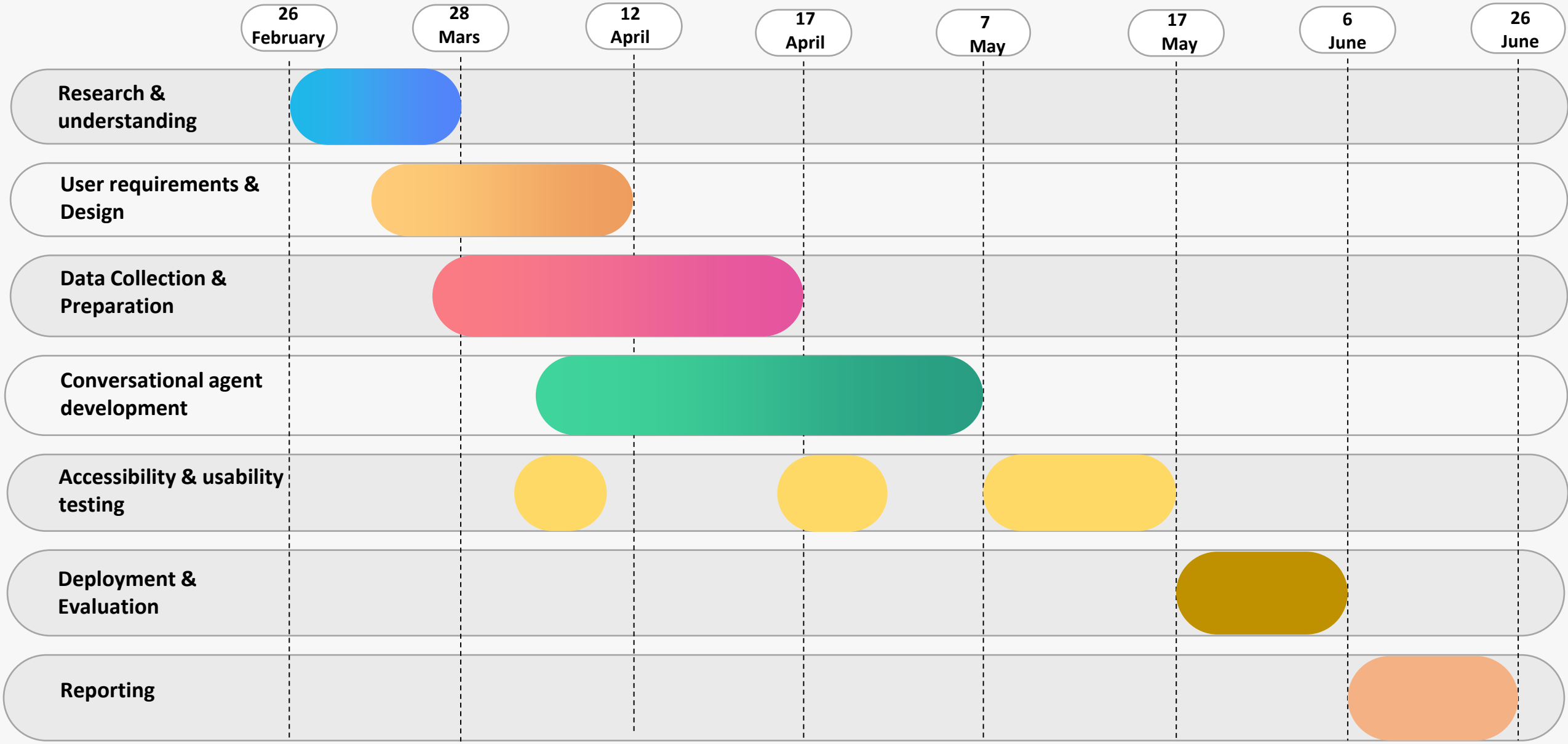
The duration of the project :

140 days = 4 months 18 days

Roadmap & Planning :



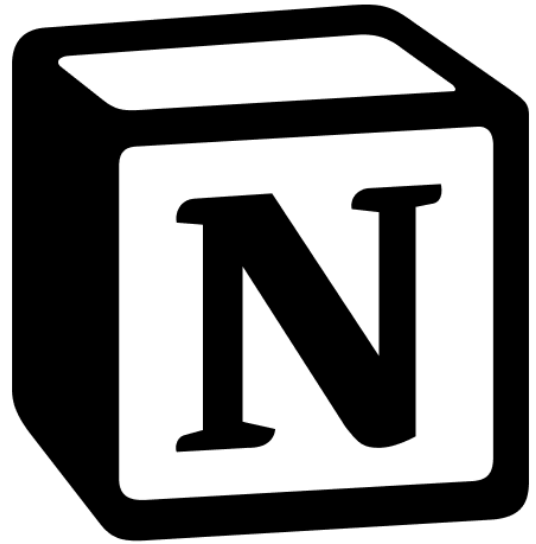
Project Timeline



Roadmap & Planning :

PROJECT MANAGEMENT SOFTWARE:

Trello



OUTLINE :

01 Literature review

02 State of the art

03 Models and technologies selected

04 Methodology

Literature review:

Approach 1: Using Frameworks and plateforme services:

DialogFlow	Google's natural language understanding platform for building chatbots and virtual agents.
Rasa	Open-source conversational AI platform with customizable NLU and dialogue management.
Wit.ai	Facebook's NLP platform for building conversational applications with ease.

Literature review:

Approach 1: Using Frameworks and plateforme services:

ADVANTAGES :

1. Rapid development and deployment of chatbot applications.
2. Lower barrier to entry, as there is no need to build models from scratch.
3. Built-in features for training, monitoring, and analyzing chatbot performance.

Literature review:

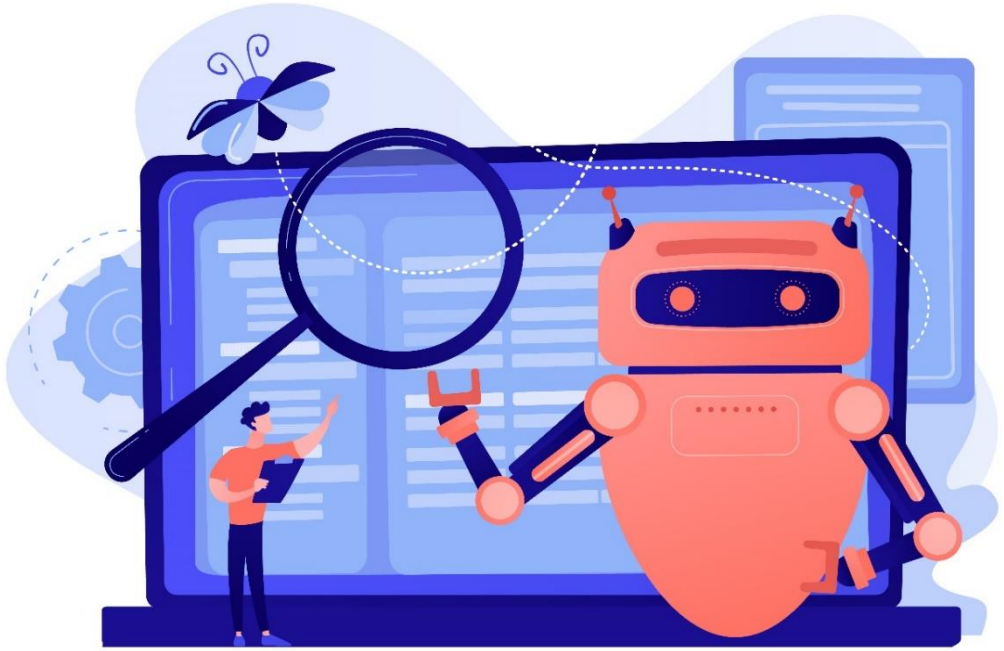
Approach 1: Using Frameworks and plateforme services:

CONSIDERATIONS :

1. Limited customization compared to building from scratch.

2. Dependency on the platform's roadmap and feature set.

3. Potential constraints in terms of integration with specific systems or complex use cases.



Literature review:

Approach 1: Using Frameworks and plateforme services:

- **CONSIDERATIONS :**

Not free

Data privacy

Limited customization and control

Literature review:

Approach 2: Design a proper architecture based on DL and NLP techniques:

ADVANTAGES :

1. Full customization and control over the entire system.

2. Ability to integrate with specific domain knowledge or external systems seamlessly.

3. Potential for higher performance and efficiency through optimized architecture and model selection.

Literature review:

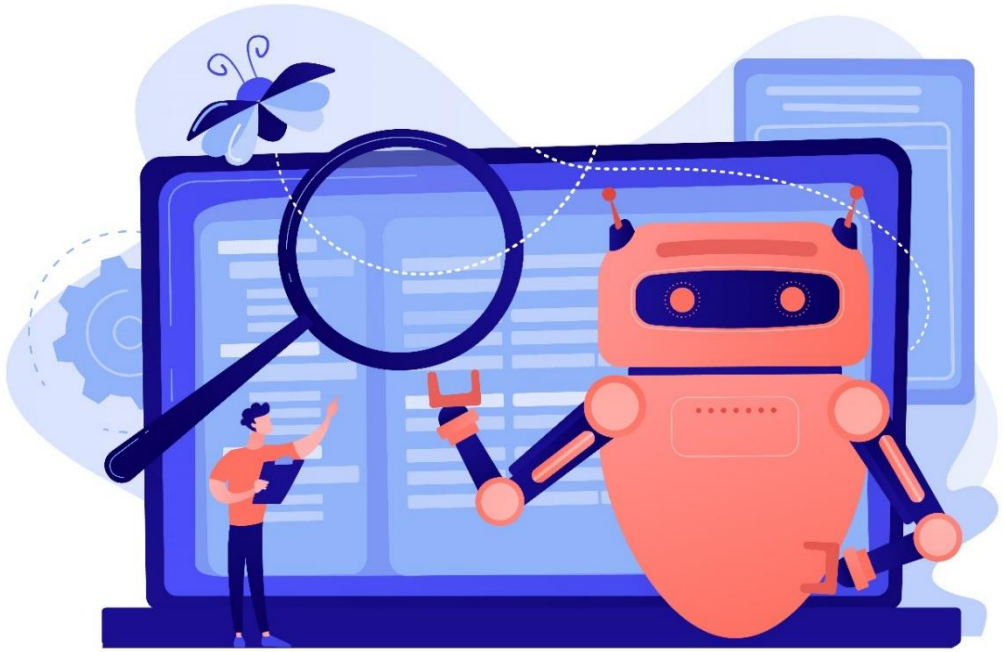
Approach 2: Design a proper architecture based on DL and NLP techniques:

CONSIDERATIONS :

1. Requires significant expertise in machine learning, natural language processing, and software engineering.

2. Time-consuming development process, especially for fine-tuning and training models.

3. Maintenance overhead for keeping models updated and improving performance over time.



Literature review:

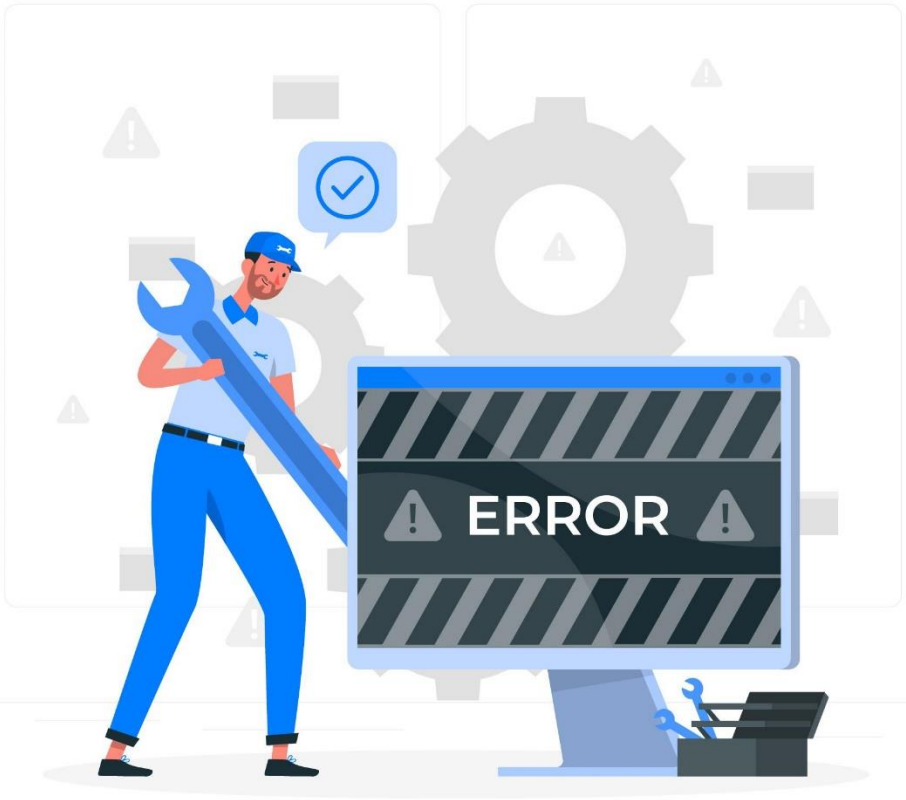
Approach 2: Design a proper architecture based on DL and NLP techniques:

GAPS :

1. The majority of the existing architectures often fail to answer questions if the response is not explicitly available in their predefined knowledge base.

2. The answers provided by these systems are often not coherent or natural.

3. Responses tend to sound artificial and lack the fluidity of human conversation.

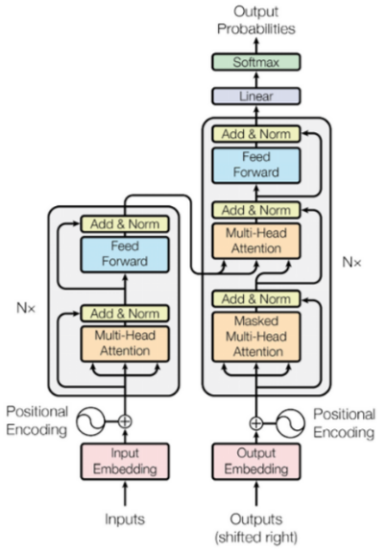


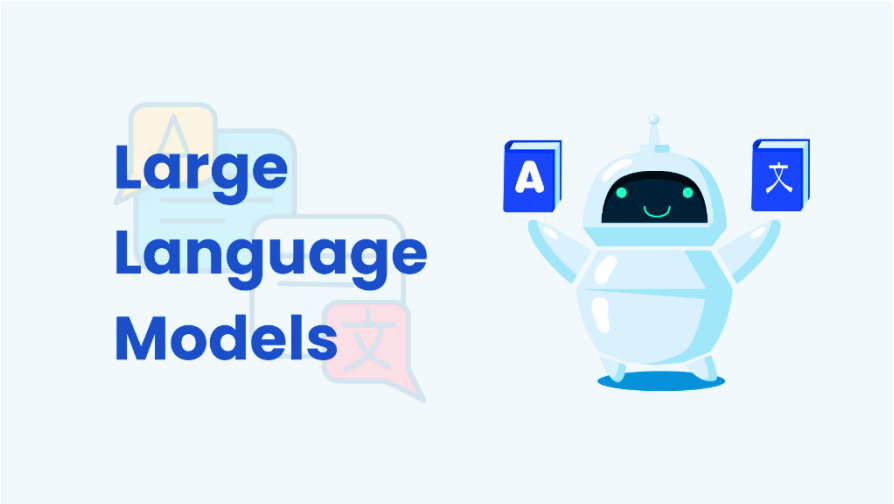
State of the art:

Approach 2: Design a proper architecture based on DL and NLP techniques:

Transformer

Attention Is All You Need

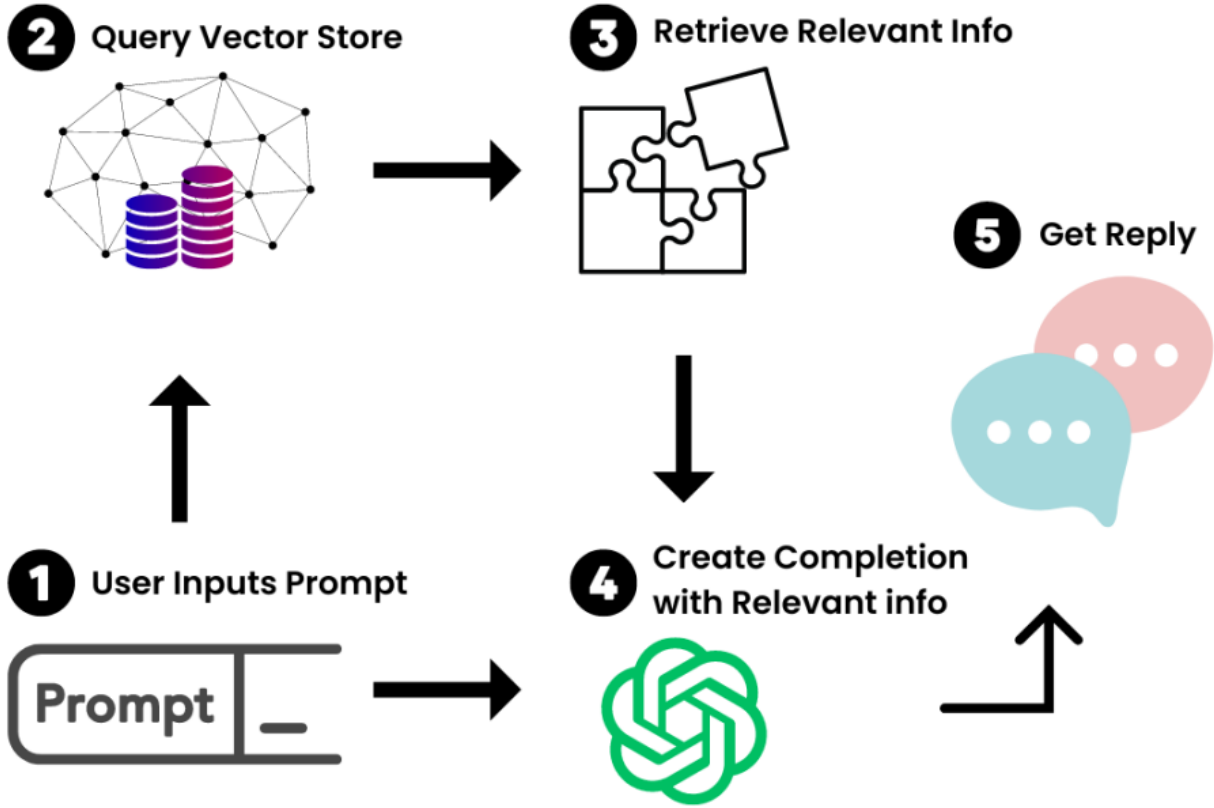




The scientific paper "Attention is All You Need," presented by Google in 2017, revolutionized the natural language processing field by introducing an advanced concept known as transformers. This discovery made it possible to create large language models (LLMs).

State of the art:

RAG (Retrieval Augmented Generation) Technology:



State of the art:

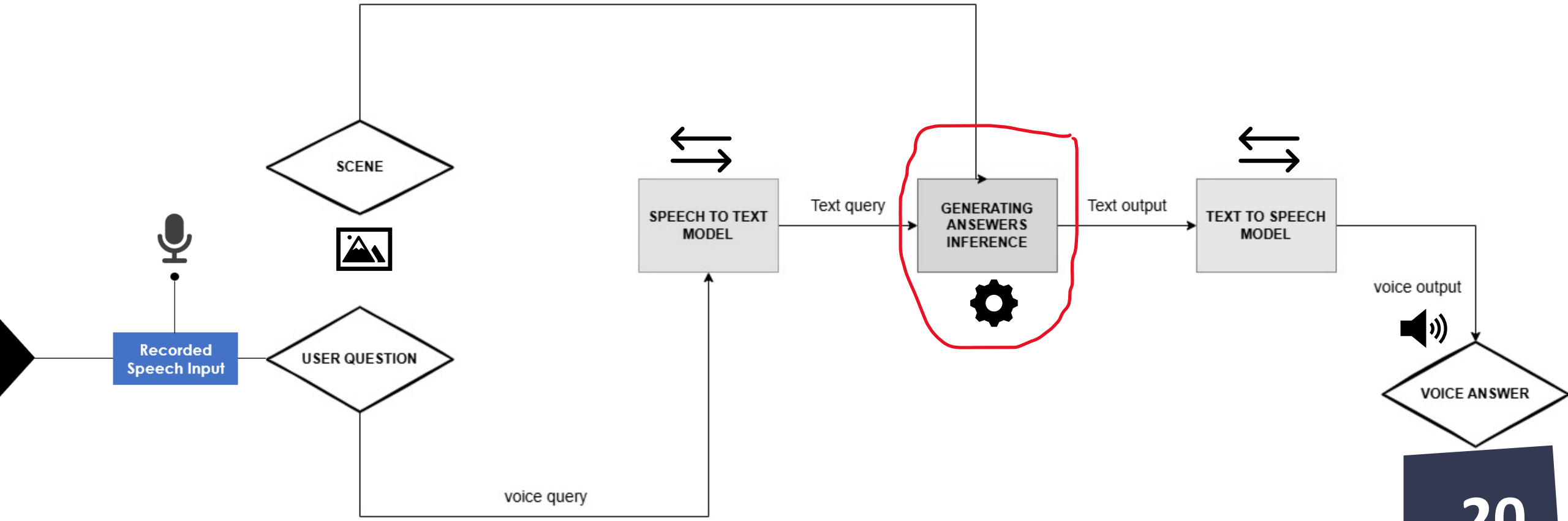
RAG (Retrieval Augmented Generation) Technology:

Data privacy

Not free

Meta released its latest large language model, Llama 3, on April 18, 2024, Local use for free

Models and technologies used :



Models and technologies used :

1. Speech to Text Model:

Criteria	ON DEVICE WHISPER	OPEN AI WHISPER	DEEPGRAM	GOOGLE CLOUD PLATFORM	AMAZON WEB SERVICES	PICOVOICE	ASSEMBLYAI
Relative Accuracy (Higher is Better)	2.0	10	4.2	1.6	3.4	0.0	6.6
Relative Speed (Higher is Better)	5.5	4.9	5.5	5.2	0.0	10	0.3
Relative Price* (Lower is Better)	0	$	$$	$$$	$$$	0	$$
Runs on Cloud or On Device?							
Ease of setup							

Comparaison between STT models and services

Models and technologies used :

1. Speech to Text Model:

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	`tiny.en`	`tiny`	~1 GB	~32x
base	74 M	`base.en`	`base`	~1 GB	~16x
small	244 M	`small.en`	`small`	~2 GB	~6x
medium	769 M	`medium.en`	`medium`	~5 GB	~2x
large	1550 M	N/A	`large`	~10 GB	1x

Characteristics of Five models of Whisper

Implementation	Precision	Beam size	Time	Max. GPU memory	Max. CPU memory
openai/whisper	fp16	5	4m30s	11325MB	9439MB
faster-whisper	fp16	5	54s	4755MB	3244MB
faster-whisper	int8	5	59s	3091MB	3117MB

Evaluation of Large-v2 model on GPU

Methods and technologies used :

2. Text to speech model:

Reasons for selecting gTTS :

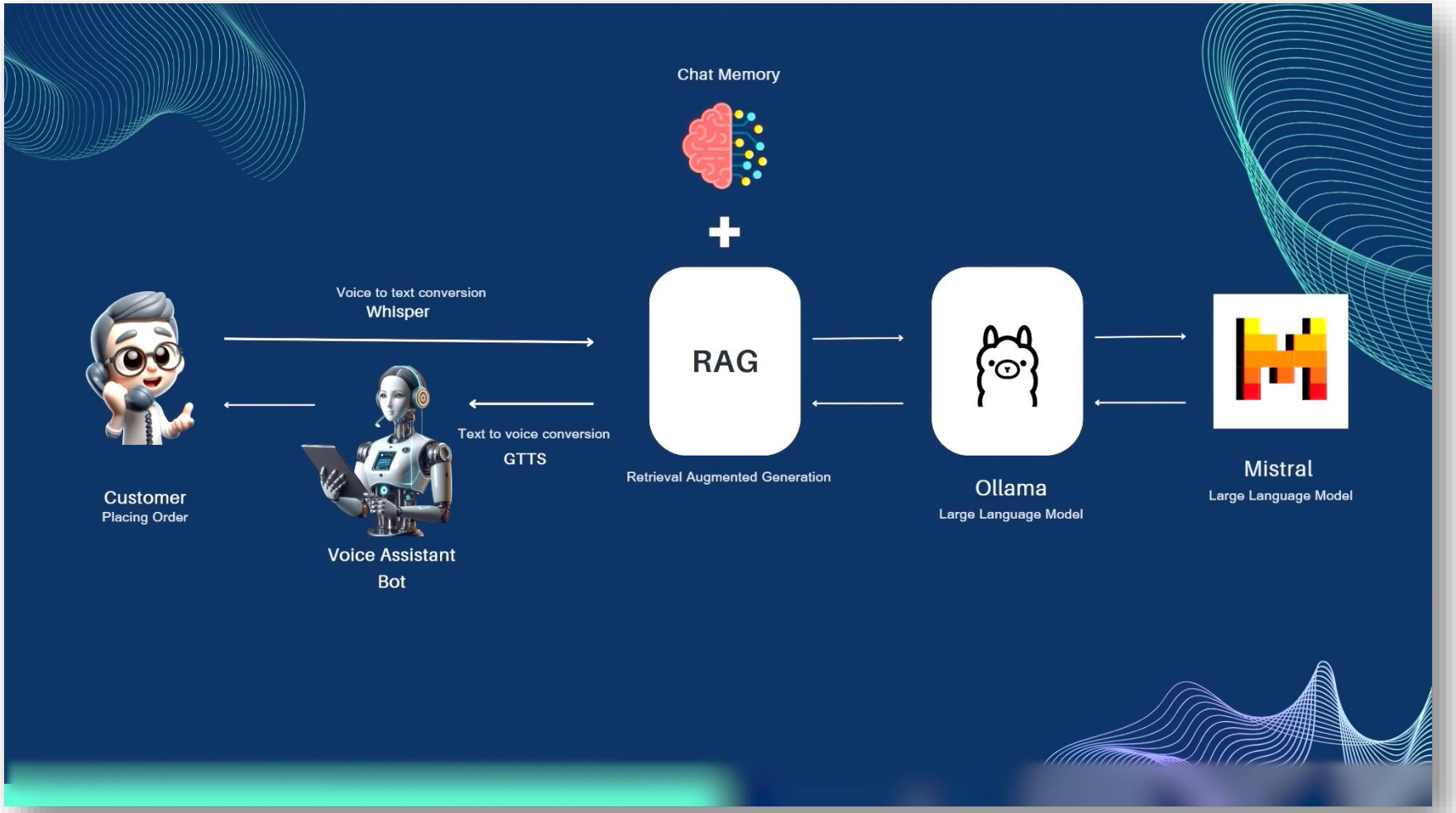
1. **Cost:** gTTS is free to use, making it an economical choice for the project.
2. **Quality:** High-quality voice output that supports multiple languages.
3. **Accessibility:** Easy to integrate and use in various applications.

Google Translate's text-to-speech API.

	PlayHT	Murf AI	ElevenLabs	Amazon Polly
Core Technology	Proprietary AI-driven TTS	AI-based realistic voices	AI-powered voices	AWS deep learning technologies
Voice Quality	High-quality voices & very conversational	High-quality, human-like voices	Extremely realistic voices	Lifelike voices
Languages Supported	Multiple languages supported	Multiple languages supported	Multiple languages supported	Over 25 languages supported
Custom Voice	Yes, with subscription	Yes, with advanced plans	Yes, offers custom voice creation	Yes, but requires setup
Use Cases	Audiobooks, podcasts, eLearning	Videos, presentations, eLearning	Dynamic content, audiobooks	Multimedia, eLearning, IoT
API Availability	Yes	Yes	Yes	Yes
Pricing Model	Subscription based, starting at $19/month	Subscription & pay-as-you-go, starts at $13/month	Subscription & pay-as-you-go, starts at $30/month	Pay-as-you-go, price per million characters
Latency	~300ms	Moderate	~400ms	Low
Free Tier	Yes, limited usage	Yes, limited features	Yes, limited usage	Free tier available, limited characters
Additional Features	Multiple voices, speed and pitch control	Role management, team collaboration	High fidelity, emotion control	Streaming, speech marks

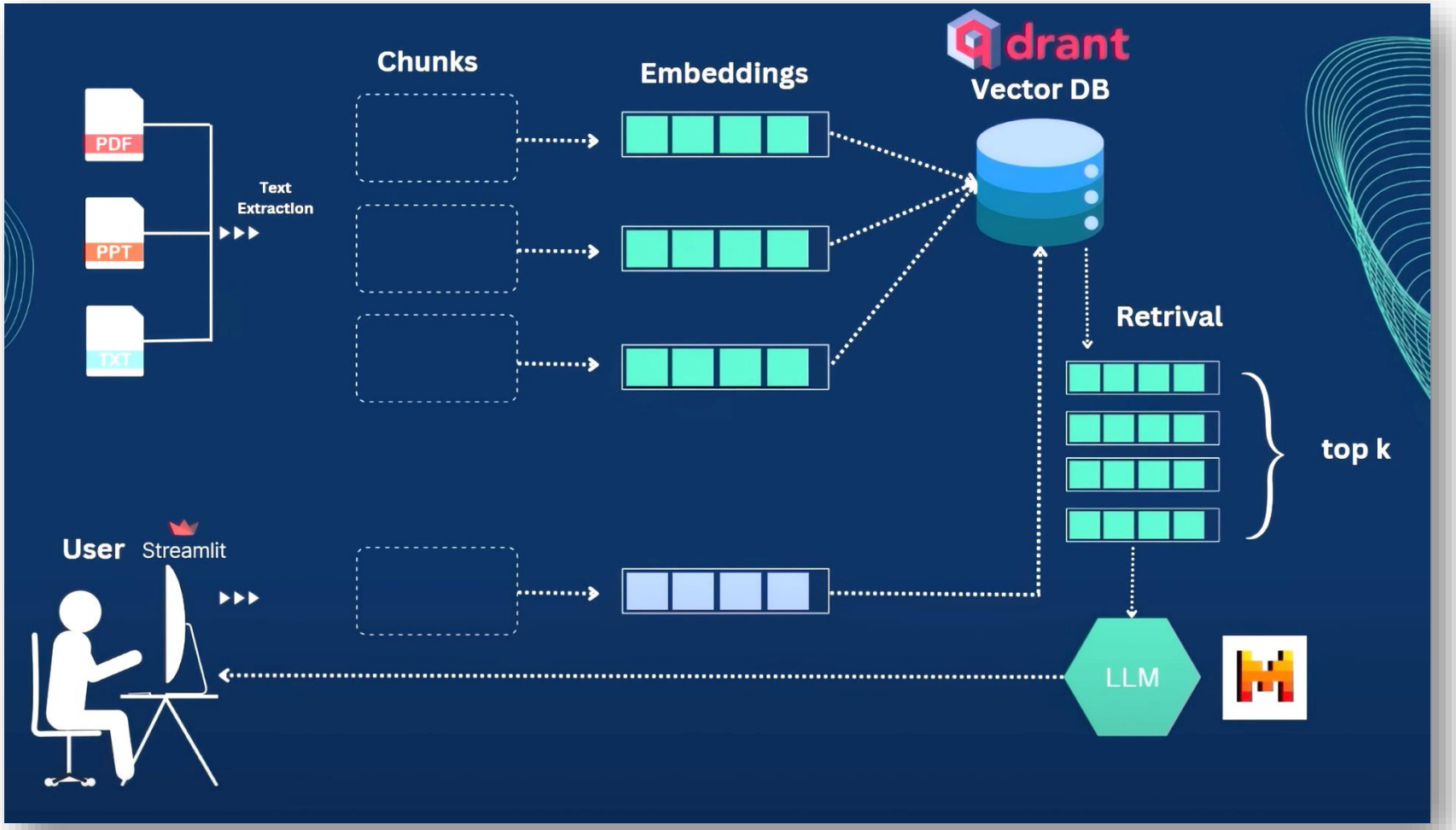
Methodology:

1. System Architecture :



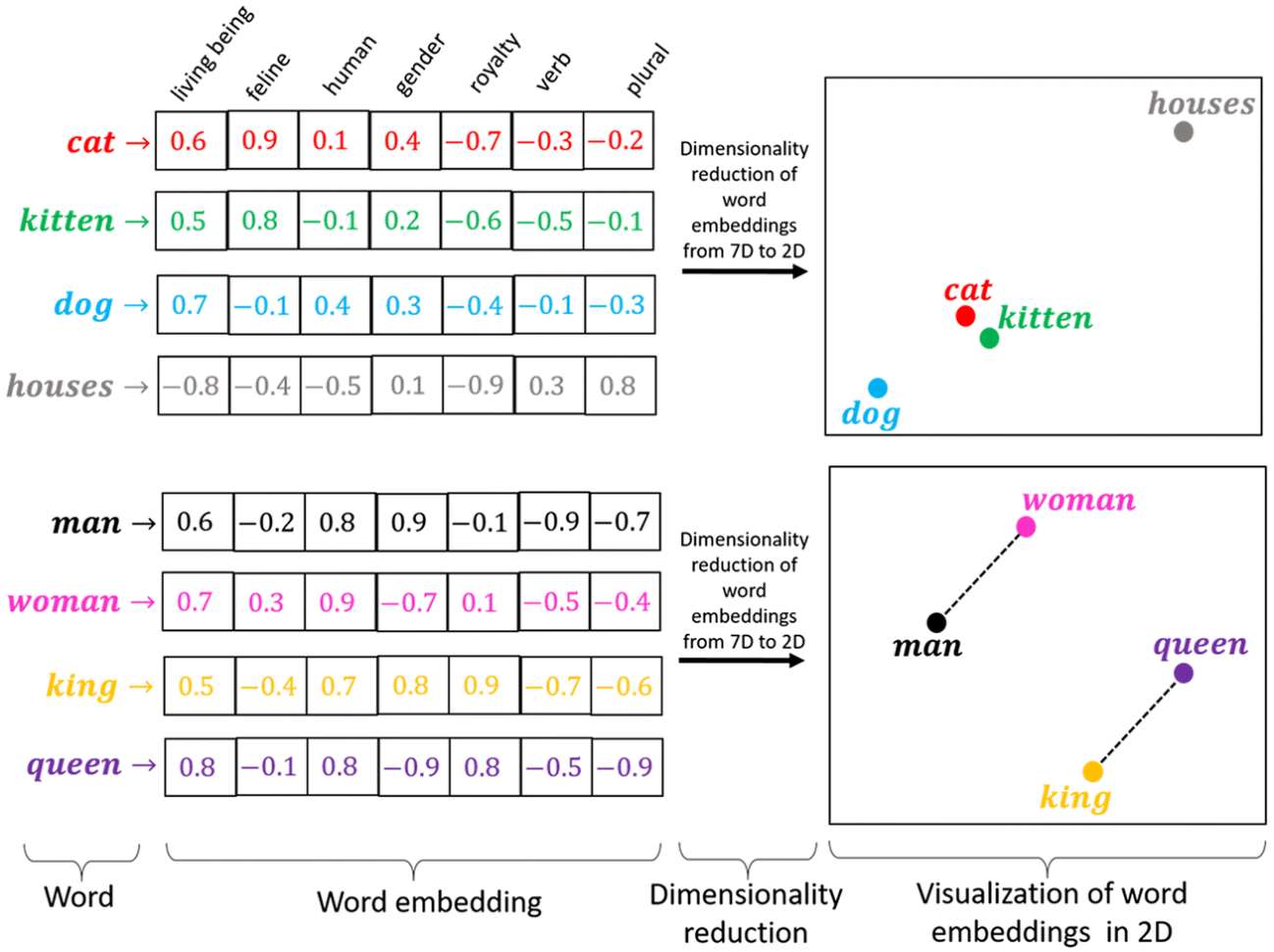
Methodology:

1. System Architecture :



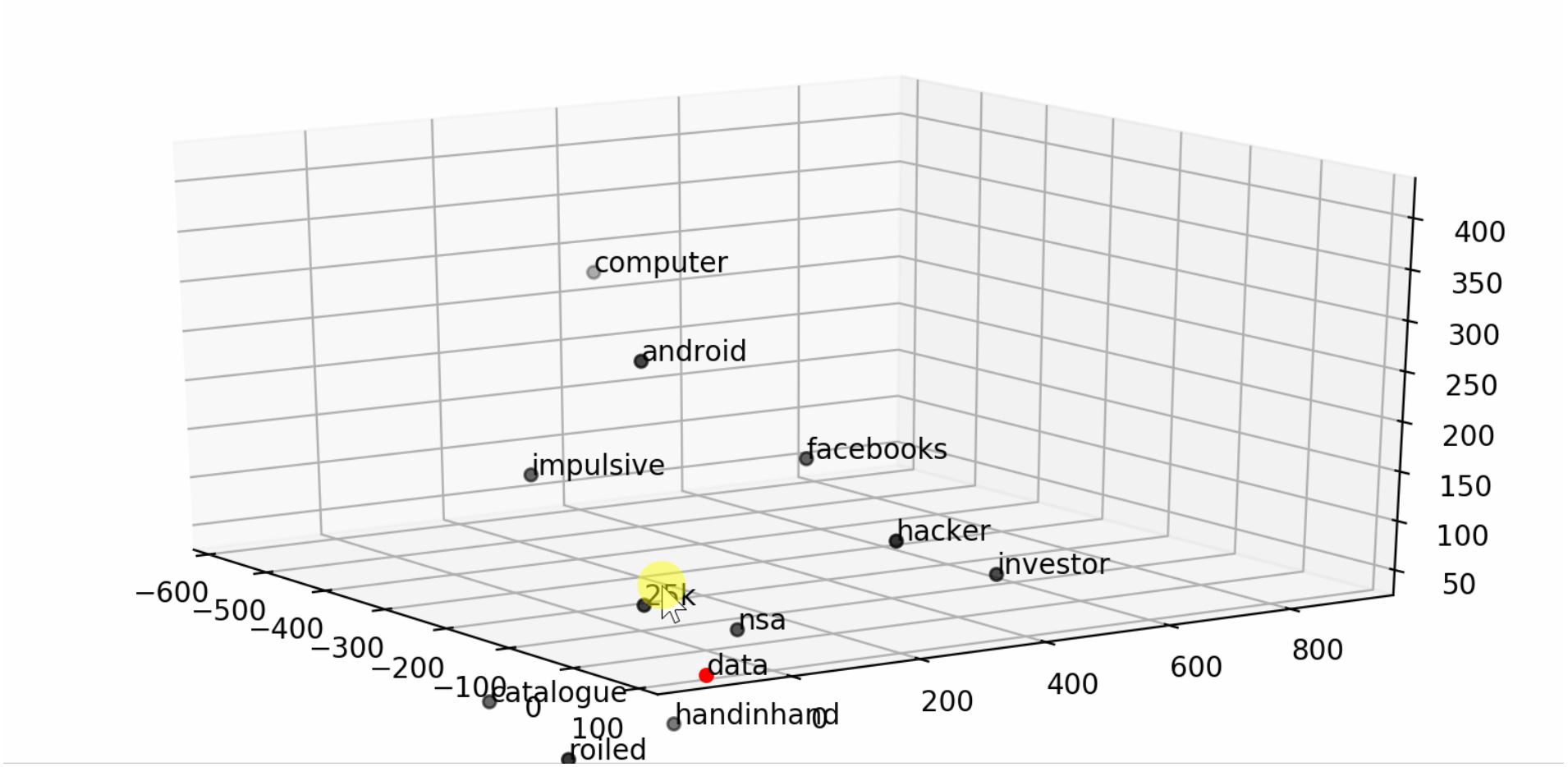
Concepts and architecture explanation :

2. Word embedding process :



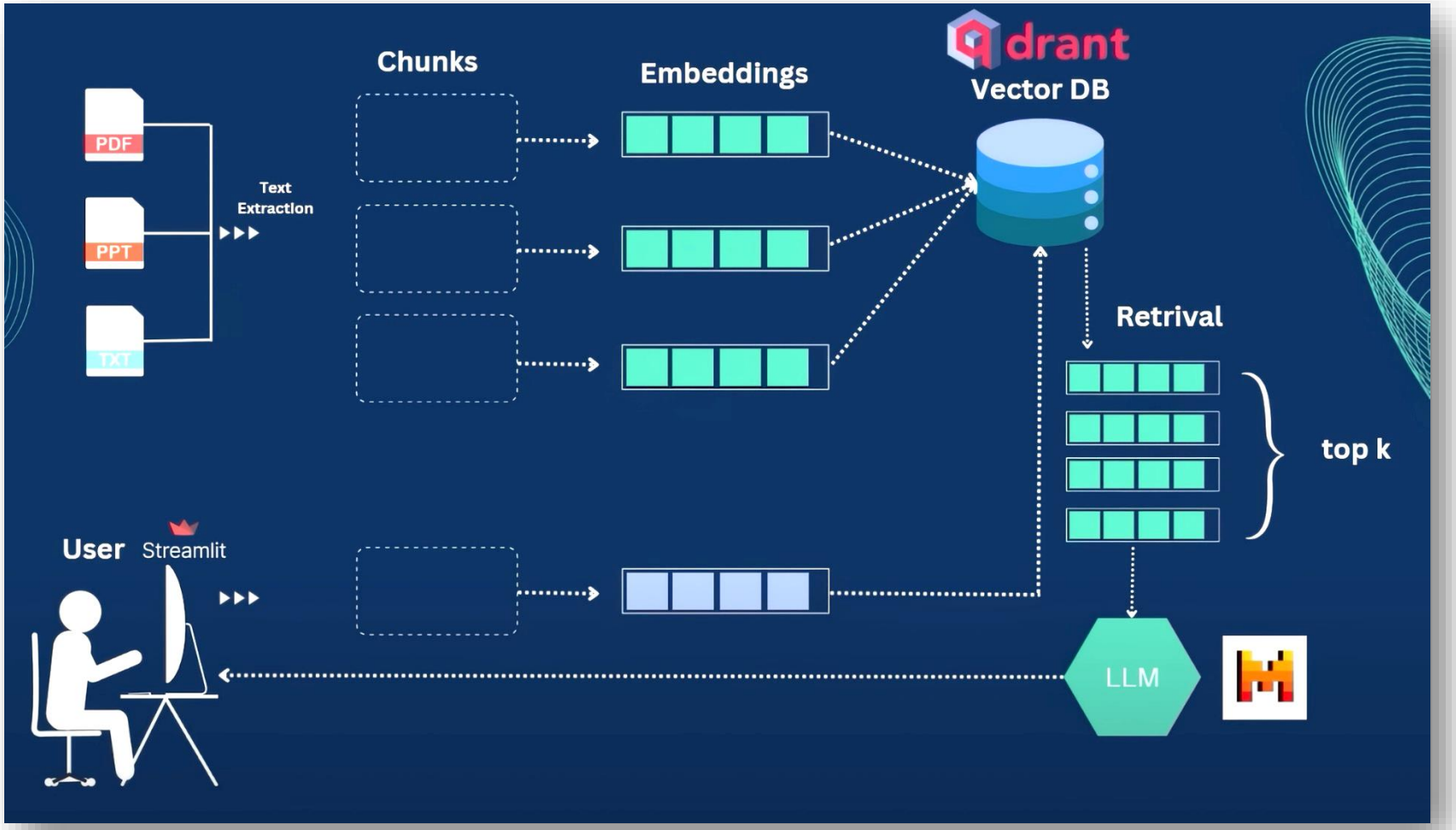
Concepts and architecture explanation :

2. Word embedding process :



Methodology:

1. System Architecture:



OUTLINE :

01 Evaluation and benchmarks

02 Challenges

03 Results

04 Conclusion and future perspectives

Evaluation and benchmarks

1. Metrics and benchmarks:

1. Word Error Rate (WER)
2. Real-Time Factor (RTF)
3. Intent Recognition Accuracy(IRA)
4. Slot Filling Accuracy(SFA)
5. BLEU Score (Bilingual Evaluation Understudy)
6. Response Time
7. Task Completion Rate
8. User Satisfaction Score (USS)

Evaluation and benchmarks

2. Evaluation and testing:

	WER	RTF	IRA	SFA	Blue score
Speech recognition Accuracy					
NLU					
Response Generation Quality					

Evaluation and benchmarks

2. Evaluation and testing:

	Response Time	Task Completion Rate	User Satisfaction Score
Value			

Challenges:

1. **Paid Models:** Finding suitable, high-performance models for speech-to-text, embeddings, and LLMs was challenging due to cost. Our goal to provide a 100% free and open-source solution limited our options for high-quality models.

2. **High Computational Resources:** Many high-accuracy, low-latency models required significant computational resources, especially for real-time speech-to-text processing, necessitating more powerful hardware.

3. **Ensuring Data Security:** We aimed to design a local solution without cloud dependency. However, most reliable LLMs cannot be used locally. The release of LLAMA3 by Meta, despite its large size, finally fit our needs for local deployment.

Challenges:

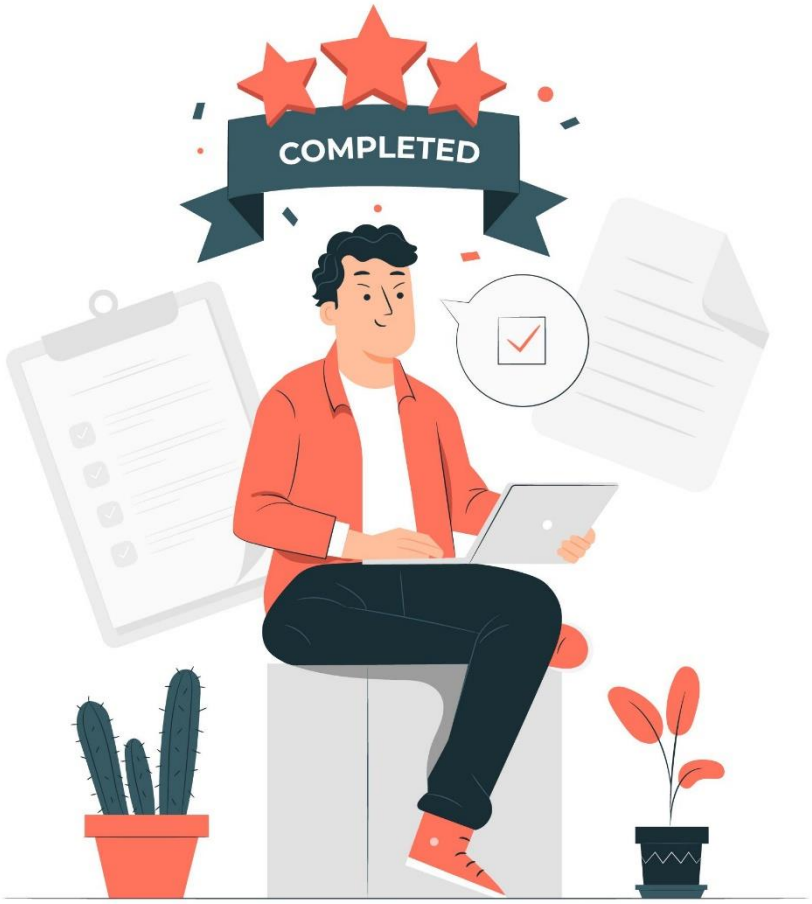
1. **Conducting Tests and Surveys:** Finding visually impaired individuals for initial design preferences and later evaluation was difficult, impacting user feedback on our solution.

2. **Fast Advancement of AI Technology:** Regularly monitoring rapid AI advancements, especially in LLMs, was essential to incorporate the most effective technologies into our solution.

3. **Managing Dependencies:** Using Python introduced challenges in managing package versions and dependencies, a complex and time-consuming task familiar to developers.

Results:

Successfully developed a robust, free, and open-source AI voice assistant that meets the high standards required for improving accessibility in museums and galleries for visually impaired individuals



Conclusion & Perspectives :

1. **Portable Embedded Systems:** Finalize the development of a portable, embedded version of the AI voice assistant that can be easily carried by visually impaired visitors, making the system more accessible and convenient to use.

2. **Deployment in Multiple Museums:** Extend the implementation to multiple museums and galleries around the world to gather diverse user feedback and improve the system based on real-world usage.

3. **Multilingual Support:** Develop and integrate support for multiple languages to make the AI assistant useful for a broader audience, including international visitors to museums.

4. **Enhanced Interaction Models:** Integrate emotion recognition capabilities to adjust responses based on the user's emotional state, providing a more empathetic and personalized experience.

Conclusion & Perspectives :

Expanded Use Cases: Expanding the use of the AI voice assistant for the university's Open Doors Day can significantly enhance the experience for high school graduates. This AI assistant will answer students' questions, providing detailed information about academic programs, campus facilities, and university architecture. The assistant will offer personalized and relevant information, making the event more interactive and informative. This approach ensures prospective students receive comprehensive guidance, aiding their decision-making process and encouraging them to join the university community.

Questions & Answers :


THANK YOU
FOR YOUR ATTENTION