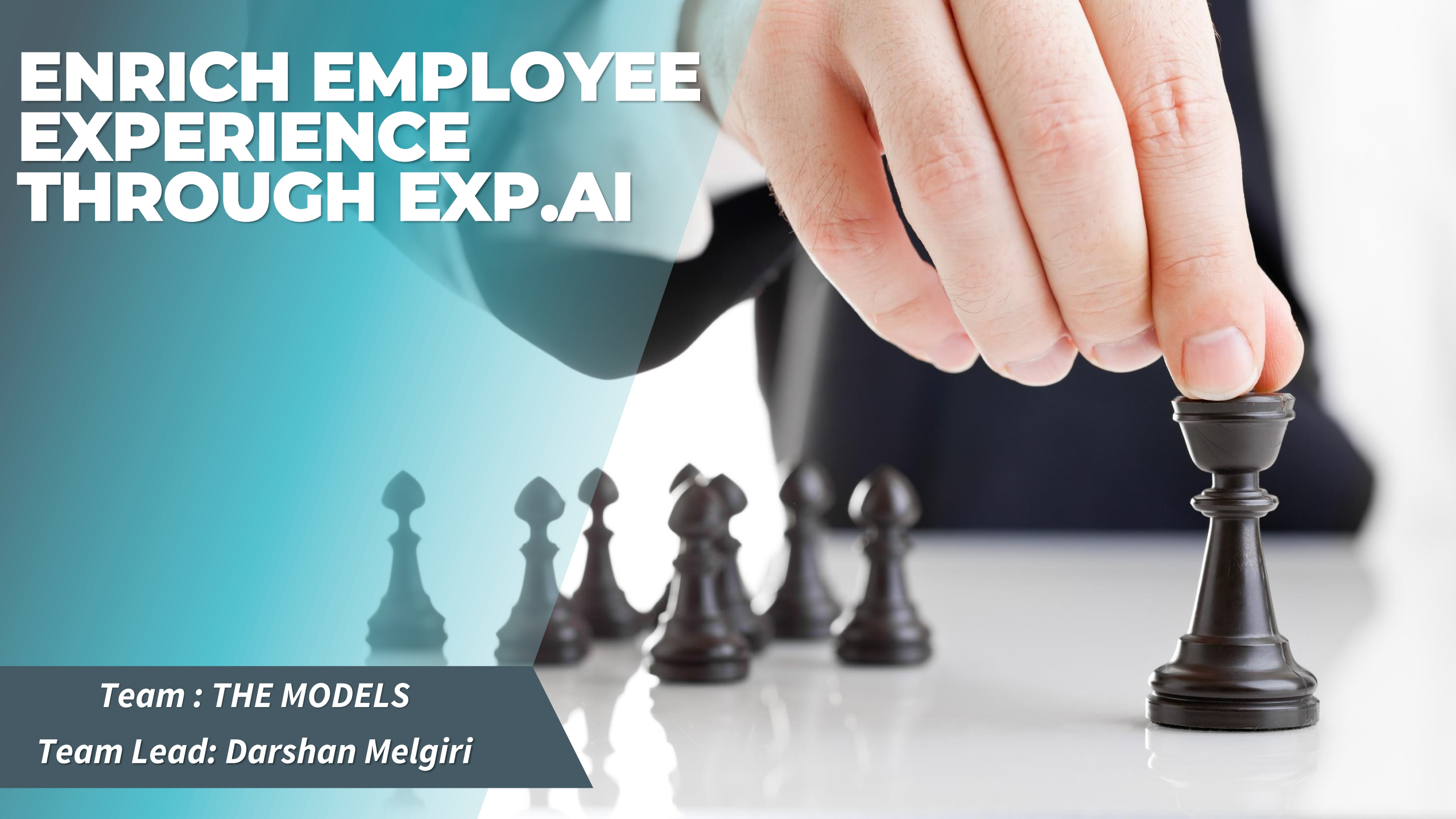


ENRICH EMPLOYEE EXPERIENCE THROUGH EXP.AI

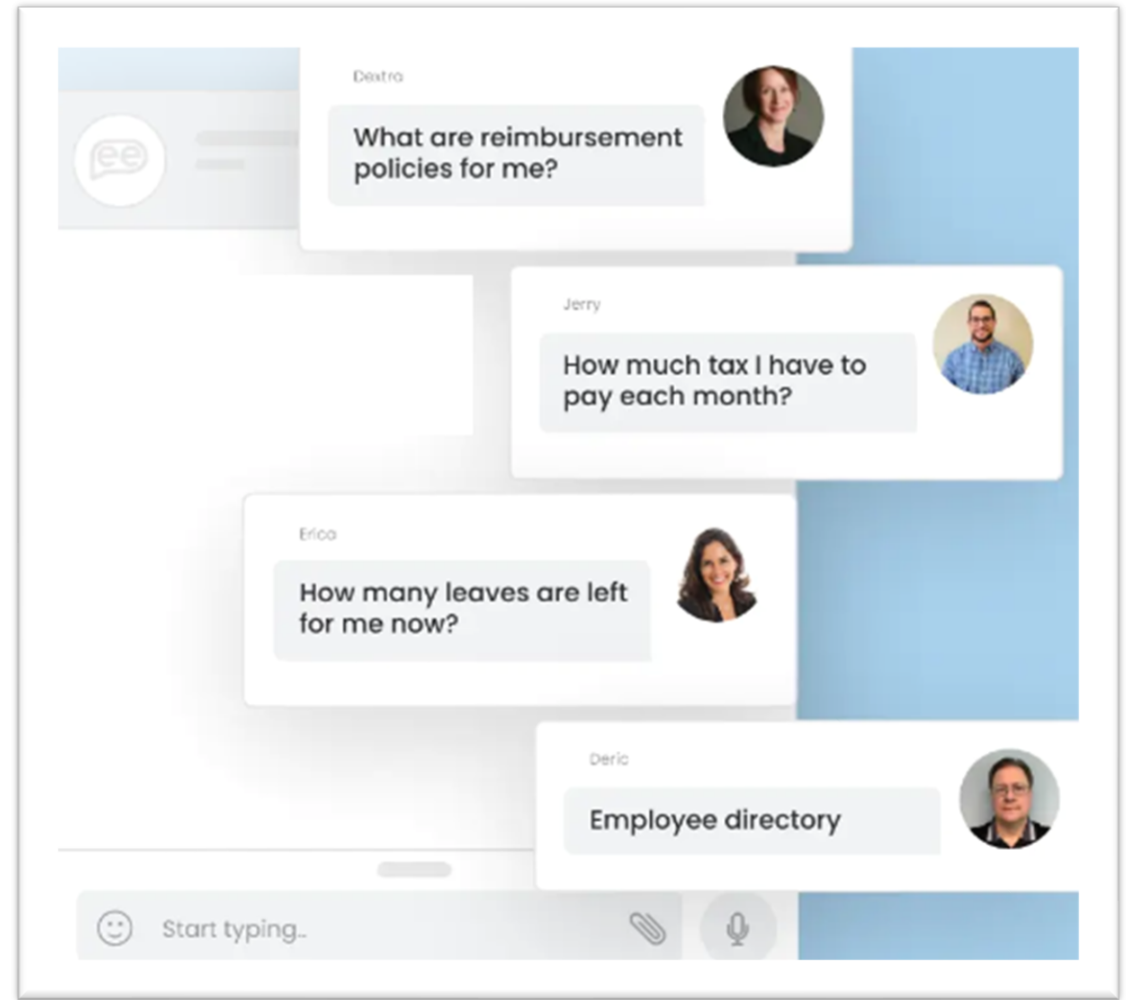
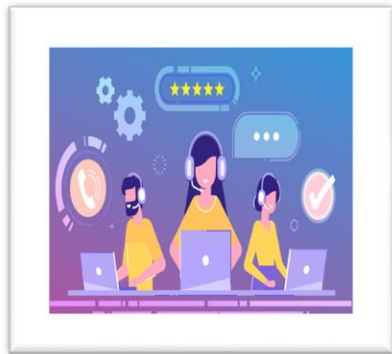
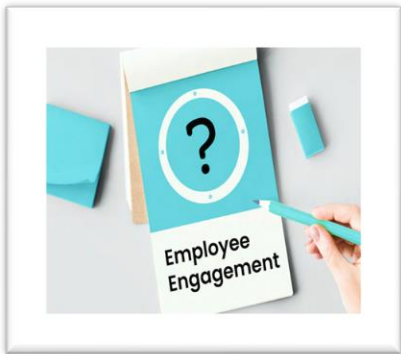
A close-up photograph of a hand in a business suit moving a dark chess piece on a white board. The background is a blurred office setting. A teal gradient overlay covers the left side of the image, containing the main title text.

Team : THE MODELS

Team Lead: Darshan Melgiri

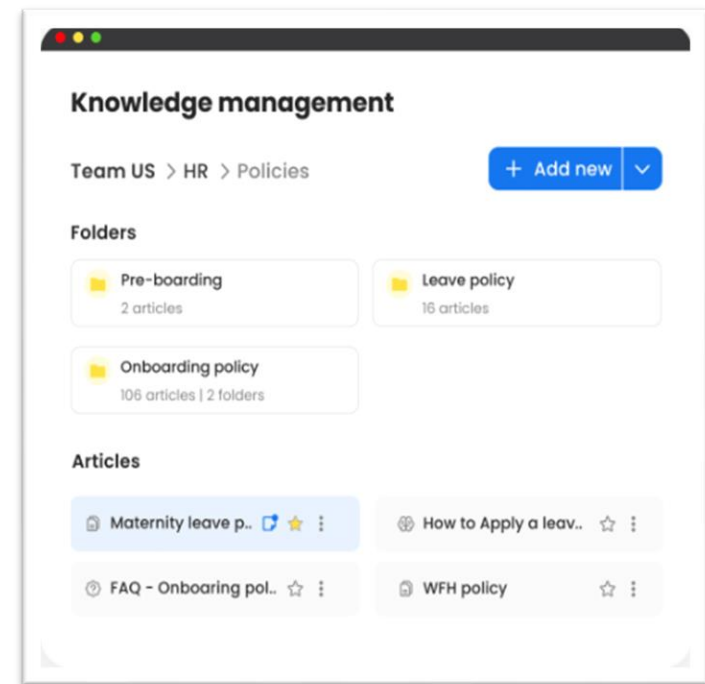
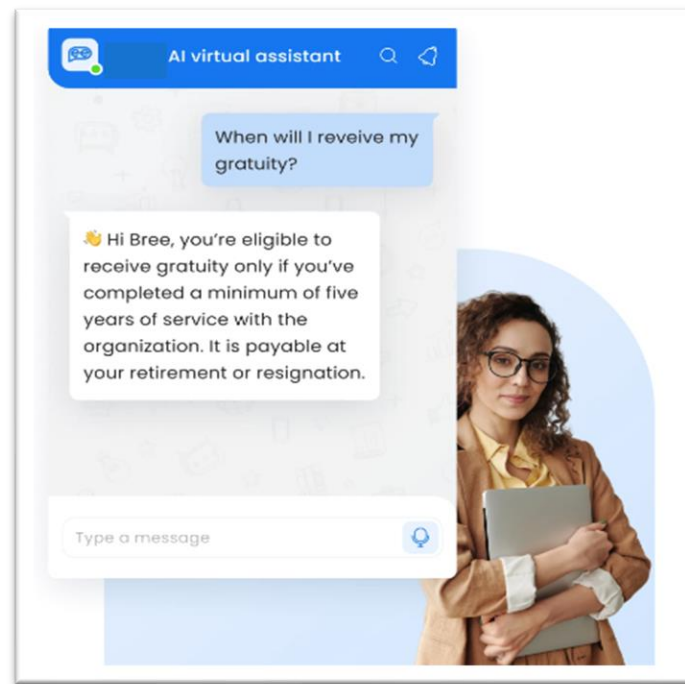
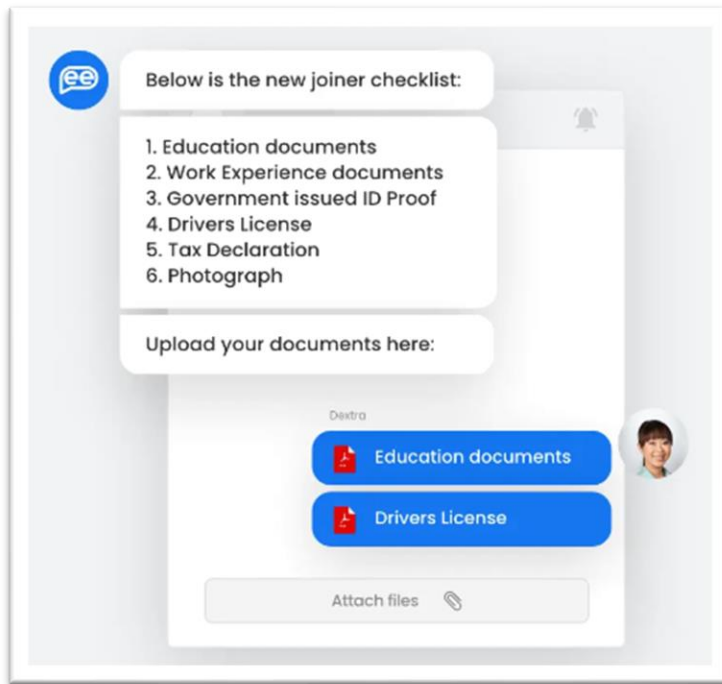
PROBLEM STATEMENT/CHALLENGES

- Enormous data stored in our group at multiple storage spaces making it difficult for employee to fetch the information at ease.
- Multiple helpdesks/support teams to address employee queries makes the process cumbersome and repetitive.
- Redundant queries needs to be answered even though the information is readily available, thereby increasing the workload.
- Need to make employee self-service smarter with AI thereby having service available 24/7.

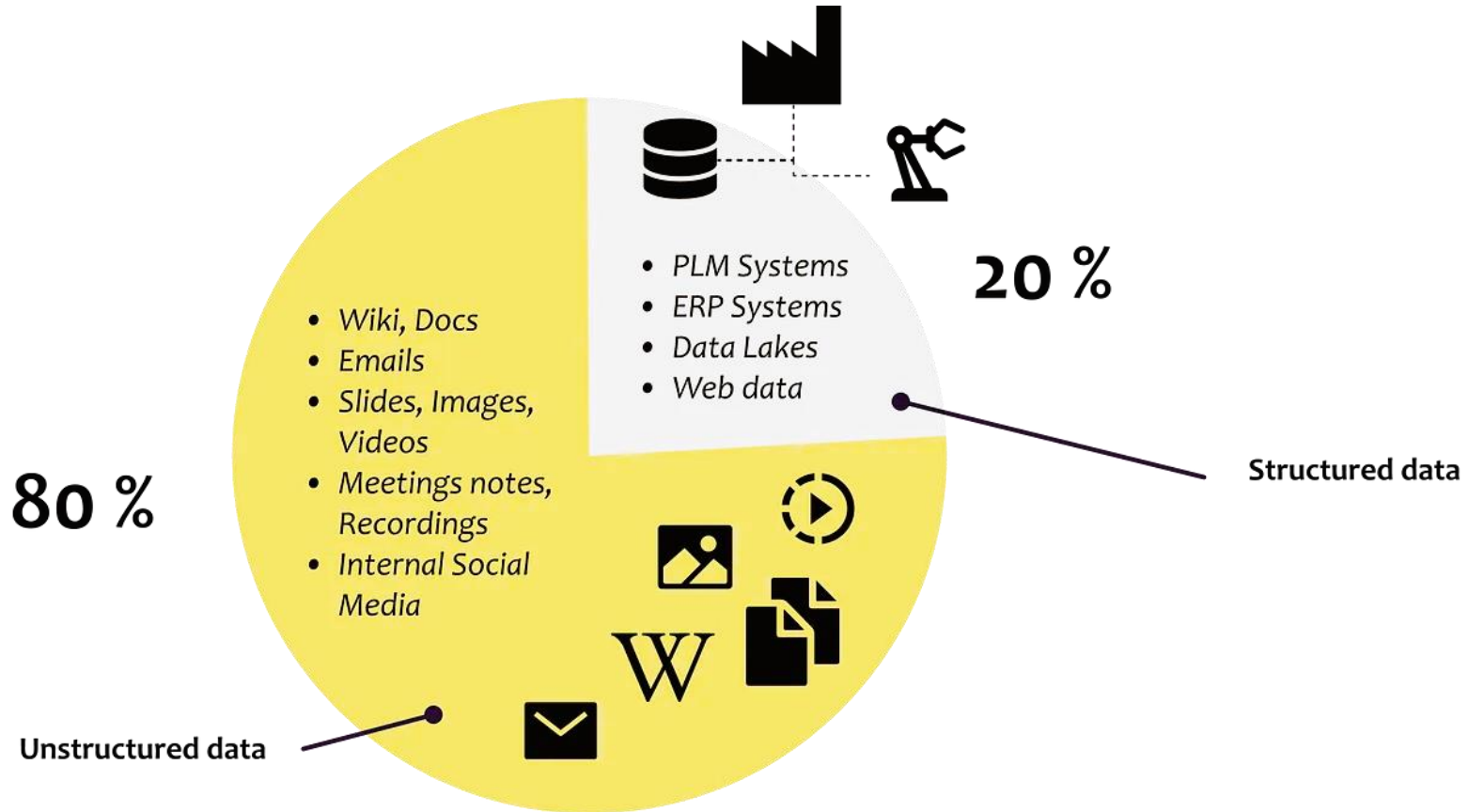


SOLUTION – EXP.AI

- Transform any knowledge base into FAQs and conversations.
- Easily update knowledge base and retrieve specific information by chatting with document.
- Make employee boarding seamless with virtual assistant.
- Make the virtual assistant your only employee helpdesk.
- Design effective training programs that meet the needs of employees at different levels.



DATA DISTRIBUTION & DIFFERENT APPROACHES



APPROACHES

- To enable large language models to answer questions that the LLM cannot know: **Model fine-tuning** and **context injection**.
- **Fine Tuning:**
 - Adjust the model for a specific task, but it doesn't really allow you to inject your own domain knowledge into the model.
 - Heavily relies on the information it learned during pre-training.
- **Context Injection:**
 - We are not modifying the LLM, we focus on the prompt itself and inject relevant context into the prompt

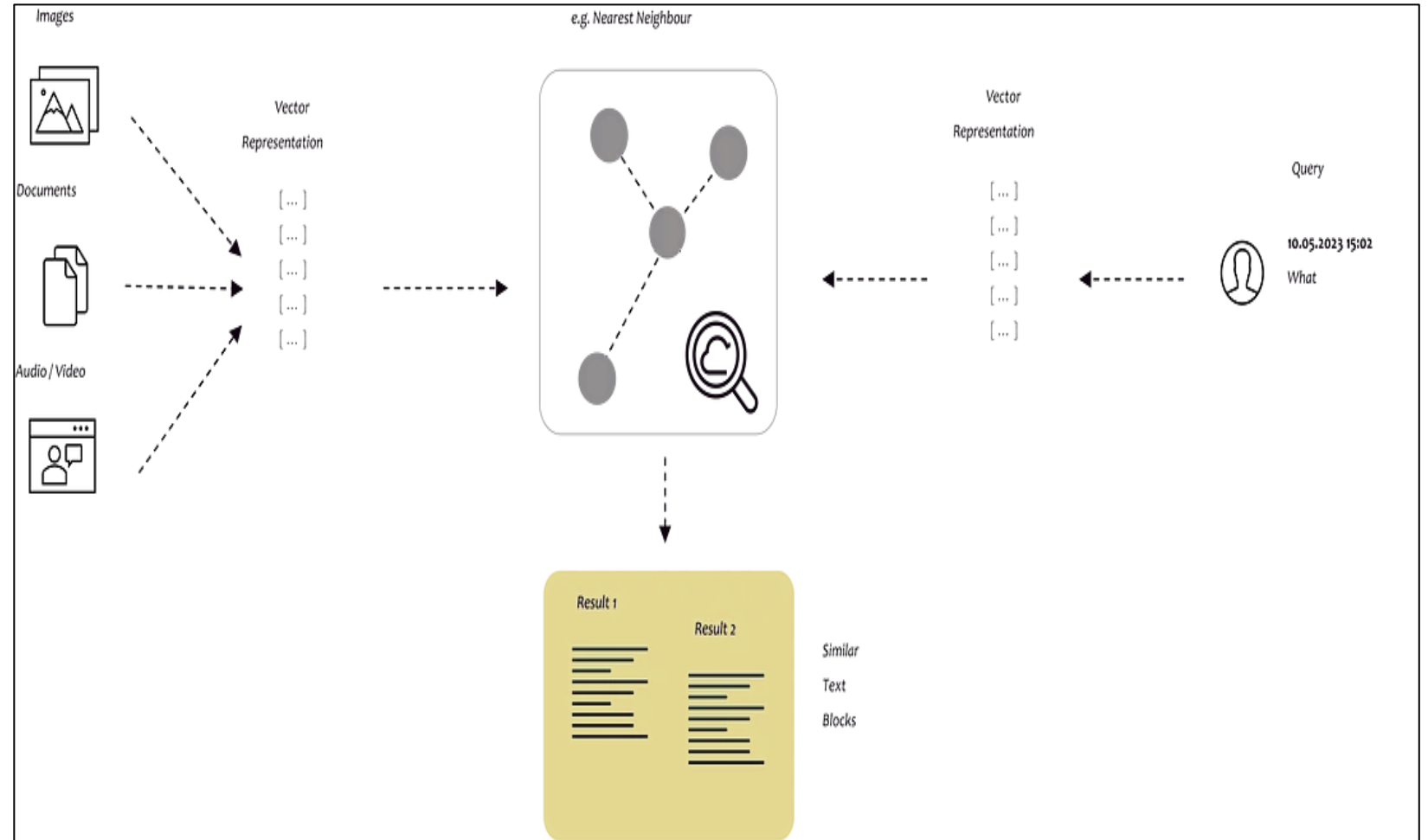
LLM – CONTEXT INJECTION

CONTEXT INJECTION PROCESS

- Inject relevant context to the prompt.
- Identify the most relevant data by creating text snippets.
- Create embeddings, translate text into vectors and represent text in multidimensional embedding space.
- Index the vector database.

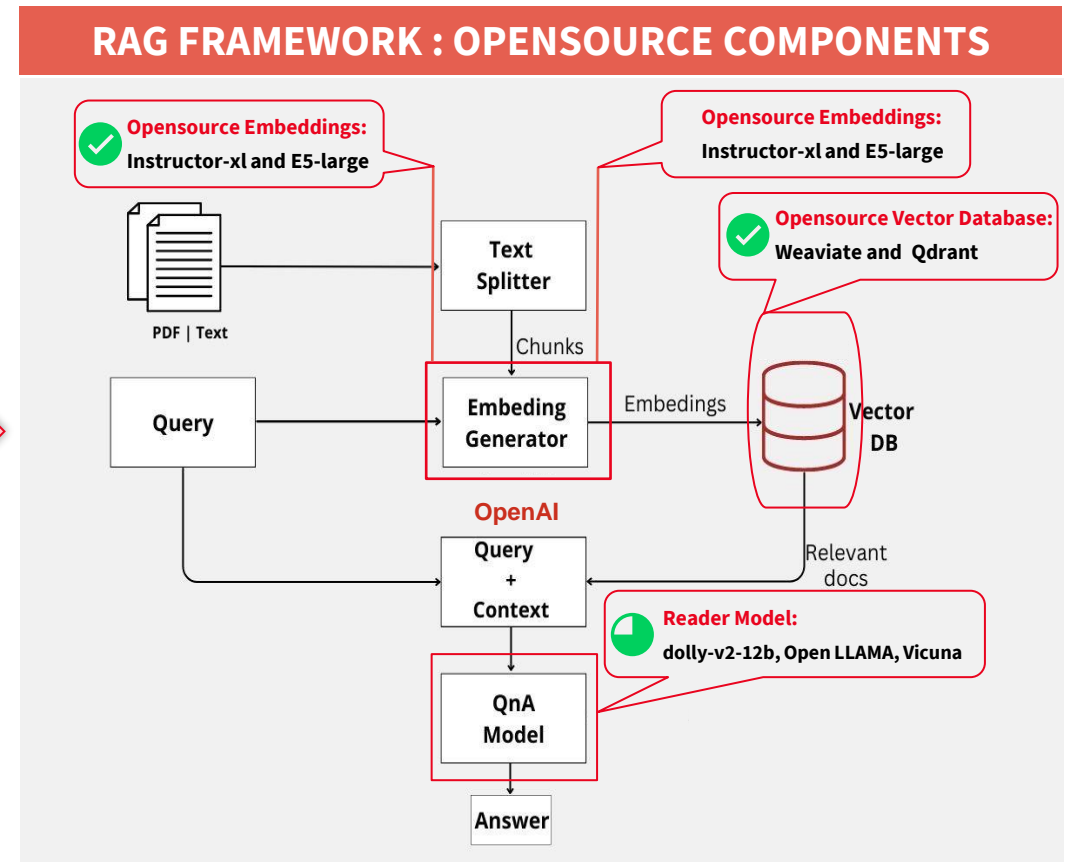
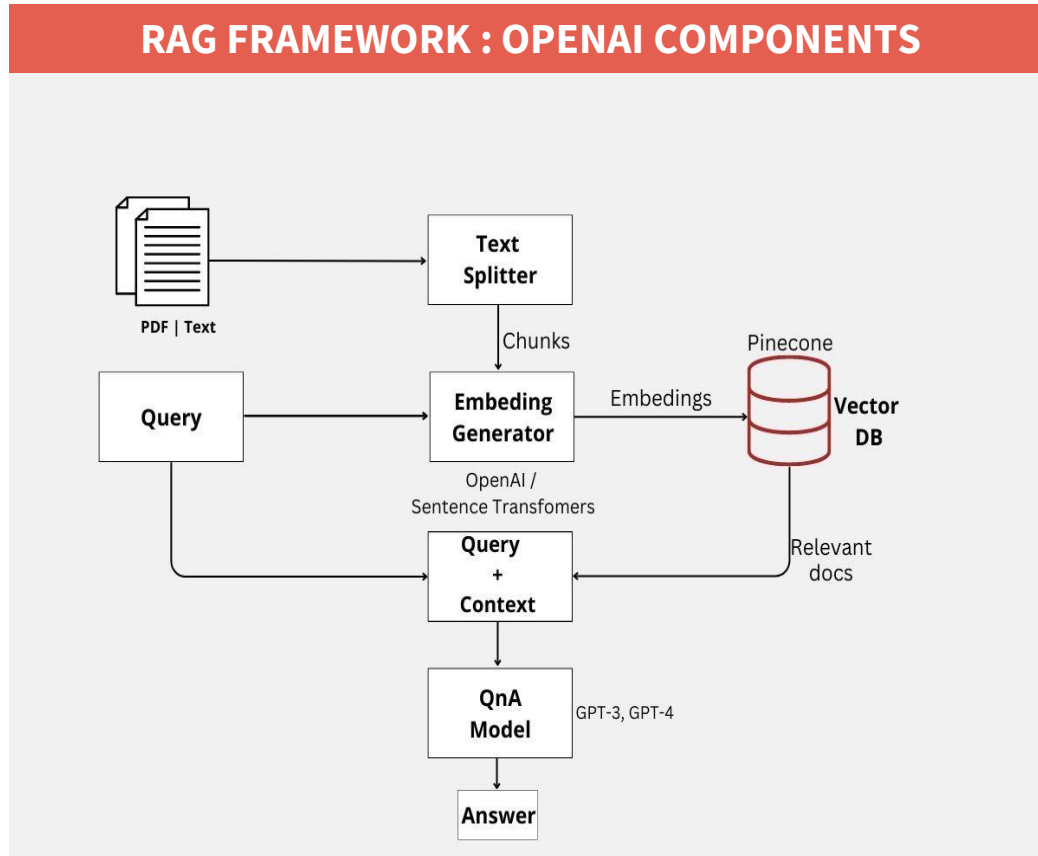
OPENSOURCE MODELS

- Embedding models :
 - Instructor-xl, Instructor-large
 - e5 large and e5-base
- Vector database :
 - Weaviate

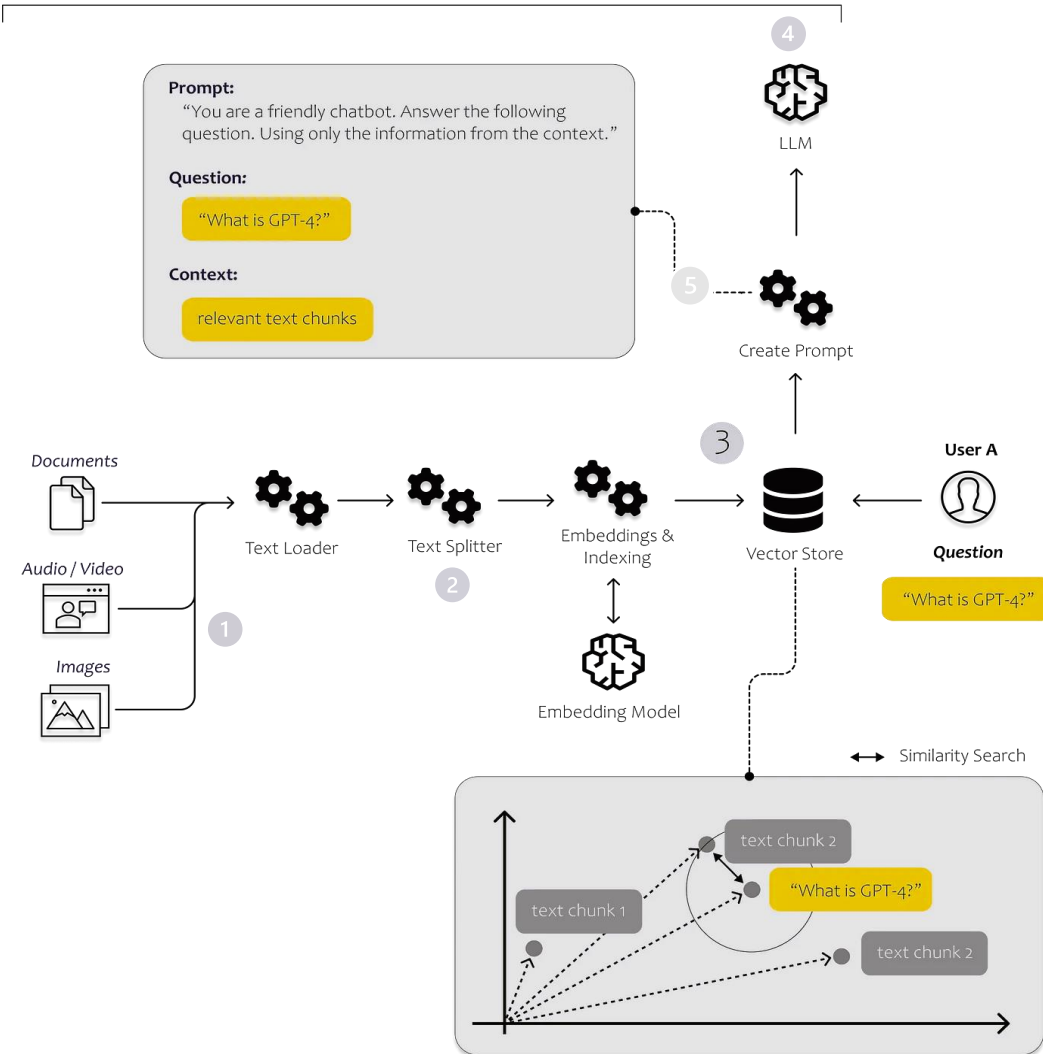


CHATGPT VS EXP.AI : FRAMEWORK CHANGES

Identified opensource models to build in-house solution.



LLM ARCHITECTURE FRAMEWORK



LLM STEPS FOLLOWED IN BUILDING OUR SOLUTION

- 1. Load the documents** into the system by leveraging document loaders (HTML pages, S3 buckets, PDFs, Notion)
- 2. Split the documents into text fragments** ; text chunk represents a data point in the embedding space, allowing the system to determine the similarity between these chunks.
- 3. Text Chunks to embeddings and store in Vector DB**; translate the meaning of words into an n-dimensional space, compare text chunks, calculate a measure for the similarity and store the vectors.
- 4. Define the model to be used** ; with our AIDelve solution we determine the better opensource LLM's (response time, prompt eval time, user feedback) and integrate with the pipeline.
- 5. Define Prompt Template** ; specifying the desired behavior style in which we want the LLM to generate answers.(summarization, writing content, code-generation)

OPENSOURCE MODELS & BENCHMARK

Retrieval Augmented Generation(RAG) models for Q&A

Model : Instructor

Details	instructor-xl instructor-large
Embedding Dimensions	768 768
Retrieval Average (15 datasets)	49.26 47.57

RANK -1 & 2

Classification Average:73
Clustering Average:45
Reranking average: 57

Embedding Model: E5

Details	e5-large e5-base
Embedding Dimensions	1026 768
Retrieval Average (15 datasets)	49.9 48.7

RANK – 3 & 5

Classification Average:73
Clustering Average:43
Reranking average: 55

Benchmark References:

MTEB Leaderbord: <https://huggingface.co/spaces/mteb/leaderboard>

Benchmarking Meta-embeddings: <https://aclanthology.org/2021.findings-emnlp.333.pdf>

LLM models and their rankings

Rank	Model	Elo Rating	Description	License
1	GPT-4	1274	ChatGPT-4 by OpenAI	Proprietary
2	Claude-v1	1224	Claude by Anthropic	Proprietary
3	GPT-3.5-turbo	1155	ChatGPT-3.5 by OpenAI	Proprietary
4	Vicuna-13B	1083	Chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	Open-source ✓
5	Koala-13B	1022	Dialogue model for academic research by BAIR	Open-source

Open LLM and their rankings

Model	Average	ARC (25-shot)	HellaSwag (10-shot)	MMLU (5-shot)	TruthQA (0-shot)
llama-65b	58.3	57.8	84.2	48.8	42.3
llama-30b	56.9	57.1	82.6	45.7	42.3
stable-vicuna-13b	52.4	48.1	76.4	38.8	46.5 ✓
llama-13b	51.8	50.8	78.9	37.7	39.9
alpaca-13b	51.7	51.9	77.6	37.6	39.6
llama-7b	47.6	46.6	75.6	34.2	34.1

Benchmark References:

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

<https://lmsys.org/blog/2023-05-10-leaderboard/>

BENEFITS OF EXP.AI

Faster access to info

EXP.AI helps employees access information without any human intervention thereby increasing the efficiency

Productivity Boost

AI-powered virtual assistant automates a lot of mundane tasks and enhances the productivity of HR teams.

Digitize employee queries

EXP.AI virtual assistant organized and digitized all the standard queries automatically

Optimized Cost with easy integration

Since the solution is built using opensource models, we do not have to pay huge licensing cost and integration to use-case is easier.

Efficiency Gains

All the requests which are treated manually will be addressed by EXP.AI and thereby bring the person days savings

THANK YOU



Address

Bangalore, Karnataka, India



Team members:

1. Darshan Melgiri
2. Kushagra Agarwal
3. Omkar Gadute
4. Satyam Kumar



Email Address:

Darshan.melgiri@gmail.com