

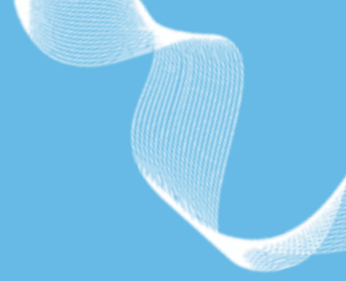


**DAXWAND** | Empowering Enterprises  
with **Generative AI**

# Empowering Enterprises with Gen AI

November 2024

# Market Challenges & Outlook



# Gen AI Adoption Challenges

## Accuracy, Explainability and Compliance

“In fact, inaccuracy—[which can affect use cases across the gen AI value chain](#), ranging from customer journeys and summarization to coding and creative content—is the only risk that respondents are significantly more likely than last year to say their organizations are actively working to mitigate.”

**Nearly one-quarter of respondents say their organizations have experienced negative consequences from generative AI's inaccuracy.**

**Generative-AI-related risks that caused negative consequences for organizations,<sup>1</sup>% of respondents**



<sup>1</sup>Question was asked only of respondents whose organizations have adopted generative AI in at least 1 function, n = 876. The 17 percent of respondents who said "don't know/not applicable" are not shown.

Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024

McKinsey & Company

# Value To Customers



## Massive Costs

Even with the ongoing cost reduction in LLMs, high volumes will incur significant costs!

70%

TCO Reduction



## Slow Time to Market

Industry average of building a Gen AI scenario takes months per scenario

Days

to Go live



## Accuracy & Quality

Lack of accuracy in foreseeing success and failure rates

30%

Higher Accuracy



## Governance and Explainability

Lack of control on data and explainability of results

FULL

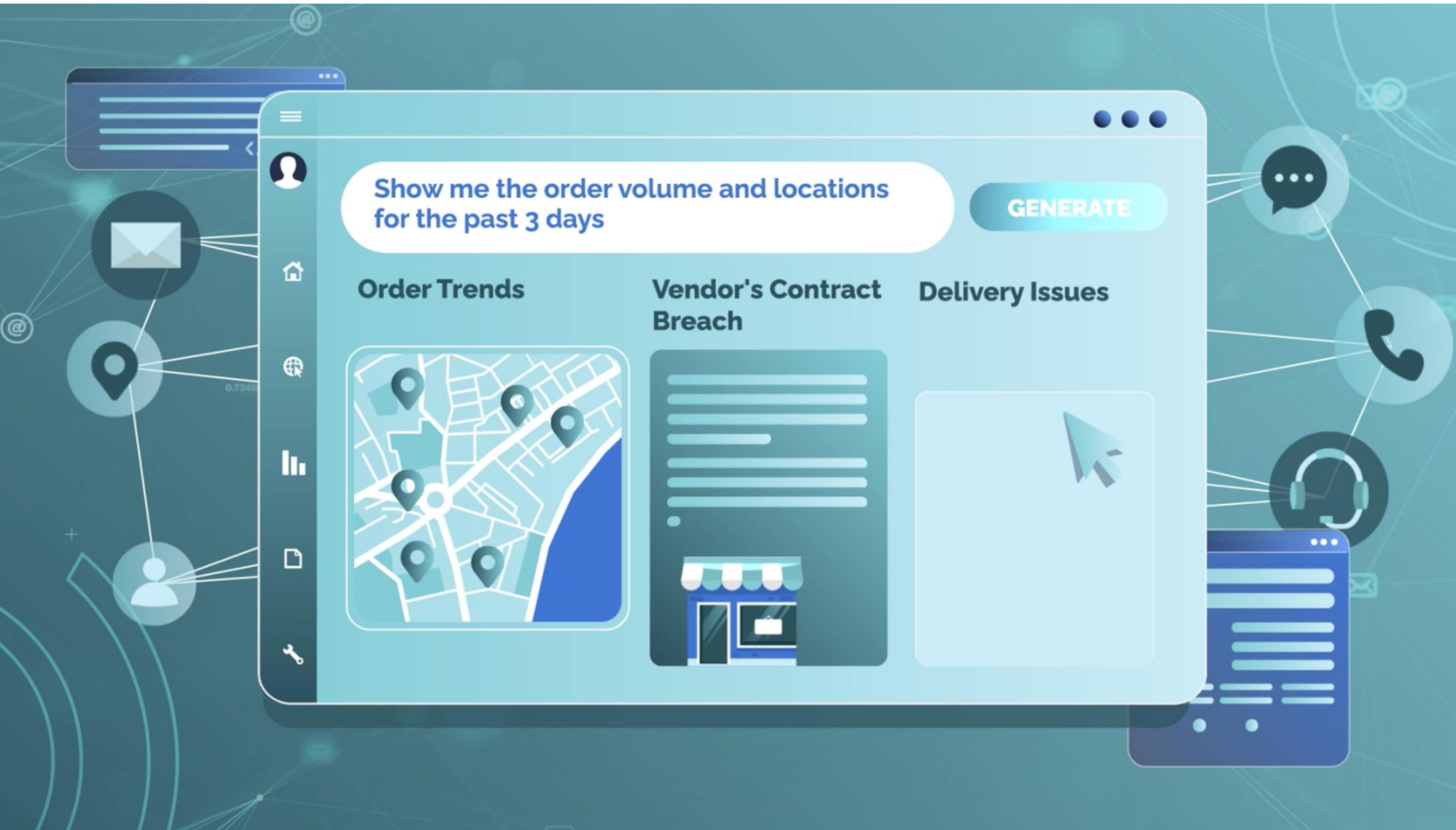
Data Access Control and Detailed Generation Explainability



# DXwand Platform Overview

## ORXTRA

# ORXTRA Overview



Kindly, click on [this link](#) to watch video if above video didn't work.

# Proven Track Record

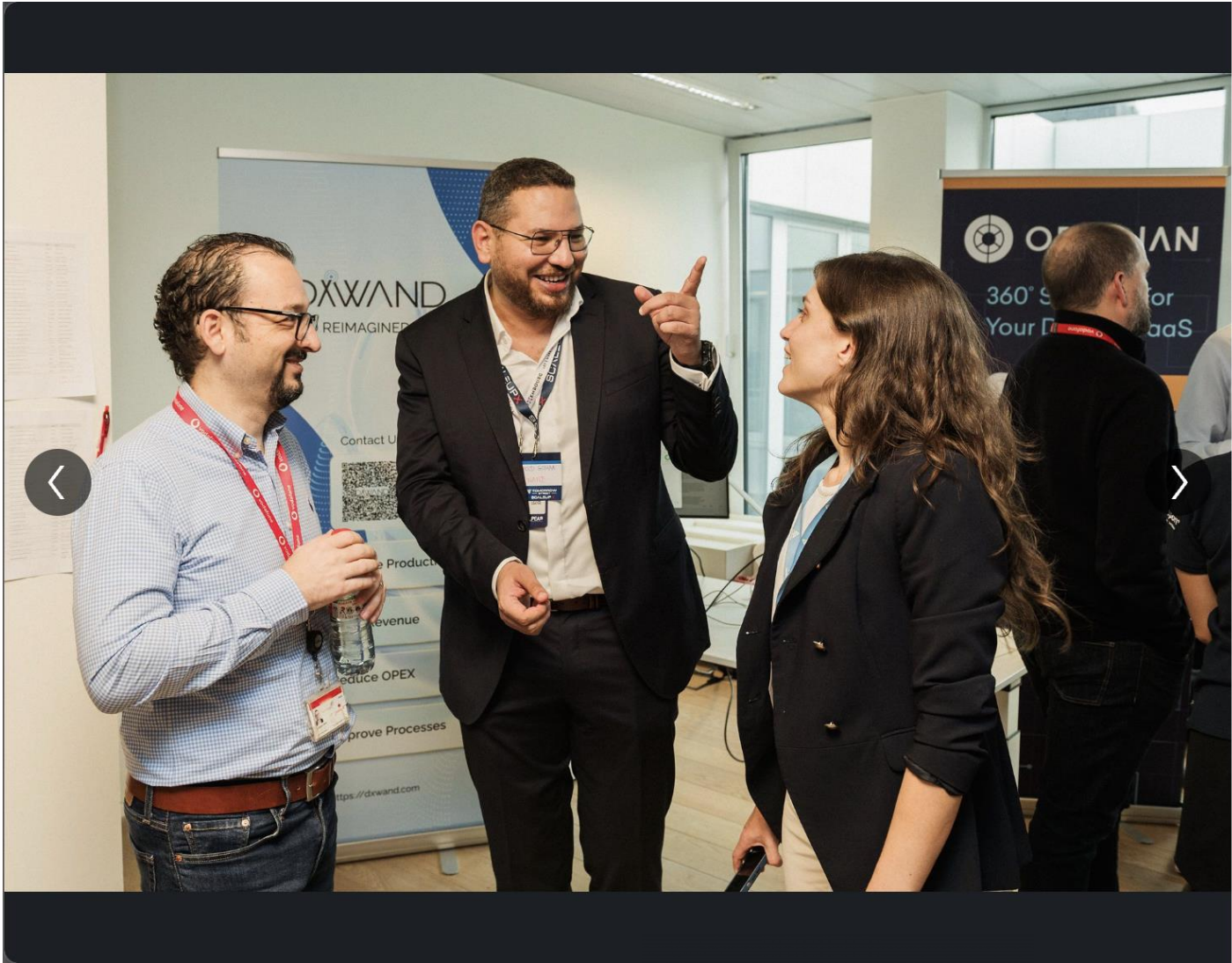
A Trusted Platform by Regional Leading Enterprises & Governments



مجلس الوزراء المصري  
مركز المعلومات ودعم اتخاذ القرار



# Vetted World Wide



Vodafone Procurement Company

8,136 followers  
3w · 🌐



🔥The buzz in the building was off the charts as our senior leaders and experts dove into conversations with innovative tech companies from **Tomorrow Street's #ScaleupX** programme and portfolio! Engaging with these cutting-edge scaleups is what fuels our future growth. Massive shoutout to the Tomorrow Street team for absolutely smashing it with an amazing event!

**AB Handshake Corporation Acceldata BlueConic Codepresso DataKrypto | Encryption by Design DTEX Systems DXwand Flood Nile Obsidian Security Userlane P0 Security Digis Squared QueryPie**

**Accenture World Wide Technology Technoport** ...more

🌐👥 You and 108 others 4 reposts

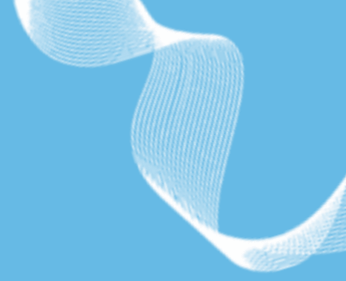
👍 Like    💬 Comment    ↻ Repost

DXWAND Comment as DXwand...





# Product & Use Cases Overview



# Product Focus Areas

Supported on Both Cloud and Offline Infrastructure

## Digital Assistant/Copilot



Automating conversations for internal & external use cases with knowledge base and systems integrations



Agentic Recruitment

## Document Intelligence



Summarizing noncompliance or findings on documents based on policies or standards of org. in the knowledgebase



Agentic KYC

## GenAI Insights



Gen AI to empower analytics of company's by extracting valuable insights from non structured data



Agentic Customer Service

## Document Generation



Gen AI to empower analytics of company's by extracting valuable insights from non structured data



Agentic Procurement

# Multi-modal Multi Agent Chatbot

Real-life Scenario demo with multi-modal multi-agent Chatbot

The screenshot displays a chatbot interface with a dark header bar containing the text "Main DXP Demo - Accuracy Optimization - Multi LLM - Admin Tour" and a "DXwand" logo. The chatbot's name is "HoldingCompany-DXwand". The conversation includes a greeting, a question about the ICAO 053 course, and a detailed response.

**HoldingCompany-DXwand** (EN)

Hello There, This is your digital assitant! How can I help you Today? 🤖

How long is the duration of the ICAO 053 Approach Control Procedural course, and what is the mode of delivery for the theory component?

- The duration of the ICAO 053 Approach Control Procedural course is 35 business days.  
- The theory component of the training is delivered through 10 days of theory via E-Learning VILT online.

Type here

Kindly, click on [this link](#) to watch demo if above video didn't work.

# Agentic Customer Service

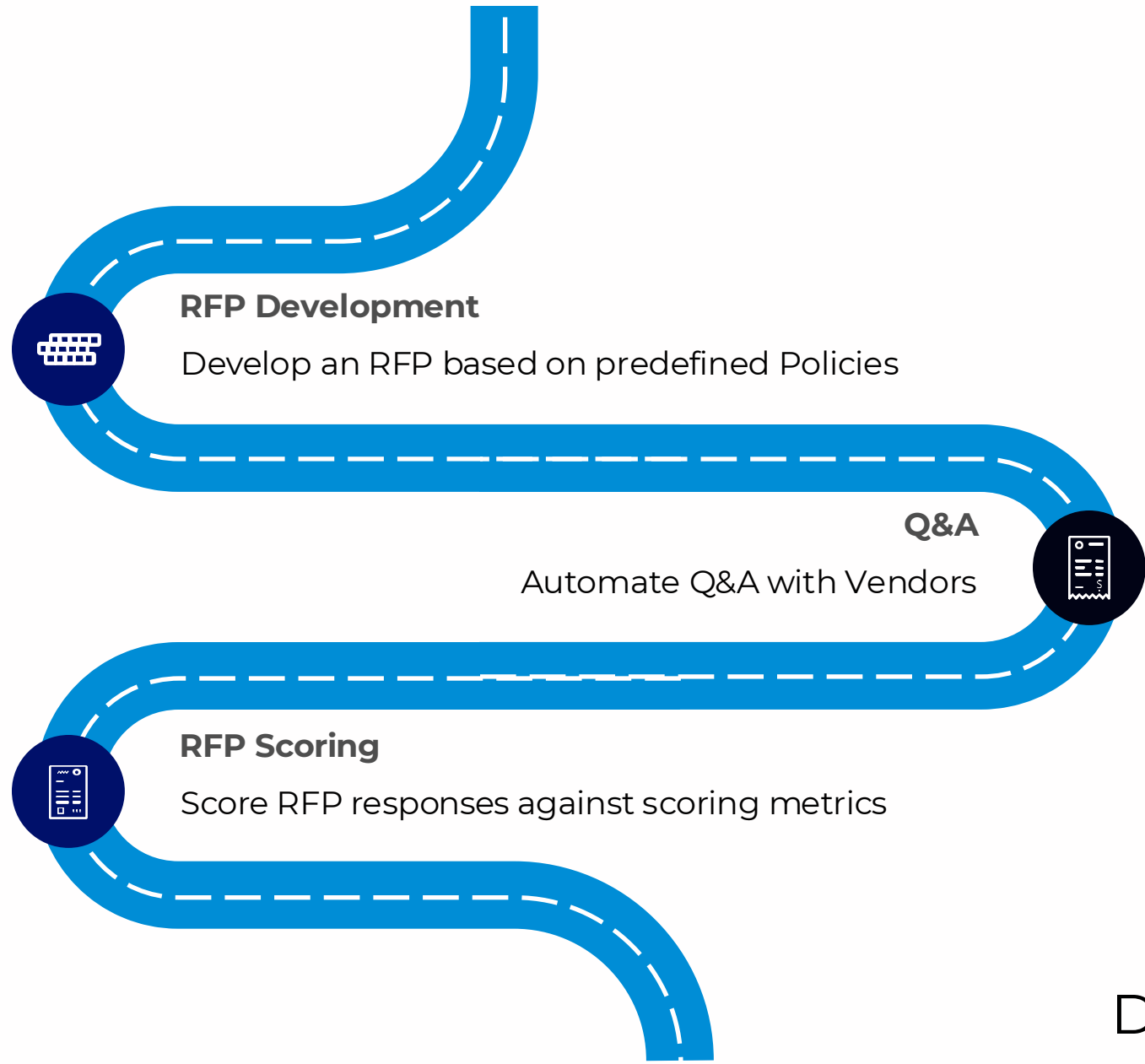


# Agentic Procurement

Agentic Procurement

Forecasted Optimization

80%



## RFP Development

Develop an RFP based on predefined Policies

## Q&A


Automate Q&A with Vendors

## RFP Scoring


Score RFP responses against scoring metrics

# Governance & Explainability

## Experimentation



Knowledge Mining > Knowledge Bases

Your 'Staging' License will be expired on 28-01-2025 info@dxwand.com 

### DFDF external web crawler Knowledge Base

4 Documents
1 Metadata
Prompts
+ New Experiment

[All Knowledge Bases](#) / DFDF external web crawler Knowledge Base

Experiment Name	Description	Chunking	Metadata	Instructions	Evaluation	Retrieval Score	Response Score
DFDF DXP Optimised		Documents: 4 Size: 500 Overlap: 100 Semantic: No Chunks: 377	<div style="background-color: #6c757d; color: white; padding: 2px 5px; border-radius: 3px;">Document Metadata</div> <div style="background-color: #6f42c1; color: white; padding: 2px 5px; border-radius: 3px;">Entities Metadata</div> <div style="background-color: #28a745; color: white; padding: 2px 5px; border-radius: 3px;">Facts Metadata</div>	Instructions per Chunk: 1 Max No. of Instructions: 100 LLM Model: gpt-4-1106-preview	Retrieval Type: Fusion Overall Top K: 3 Embedding Top K: 3 BM25 Top K: 3 Evaluation LLM Model: gpt-4-1106-preview Response LLM Model: gpt-3.5-turbo Exclude Metadata: No Include Metadata Only: No	92.00	84.63
GPT 4 Baseline		Documents: 4 Size: 1024 Overlap: 200 Semantic: No Chunks: 179	<div style="background-color: #6c757d; color: white; padding: 2px 5px; border-radius: 3px;">Document Metadata</div>	Instructions per Chunk: 1 Max No. of Instructions: 100 LLM Model: gpt-4-1106-preview	Retrieval Type: Baseline Top K: 6 Evaluation LLM Model: gpt-4-1106-preview Response LLM Model: gpt-4-1106-preview Exclude Metadata: No Include Metadata Only: No	87.00	84.25
GPT 3.5 Baseline		Documents: 4 Size: 1024 Overlap: 200 Semantic: No Chunks: 179	<div style="background-color: #6c757d; color: white; padding: 2px 5px; border-radius: 3px;">Document Metadata</div>	Instructions per Chunk: 1 Max No. of Instructions: 100 LLM Model: gpt-4-1106-preview	Retrieval Type: Baseline Top K: 6 Evaluation LLM Model: gpt-4-1106-preview Response LLM Model: gpt-3.5-turbo Exclude Metadata: No Include Metadata Only: No	89.00	68.75

< Back To All Knowledge Bases

Copyright © 2017-2024 DXwand Inc.

Build-Your-Own AI Chatbot

# Governance & Explainability

## Explainability

Knowledge Mining > Knowledge Bases      Staging License expires o...    0.01% Pages consumption    0.00% Messages consumption    NA NA

### Evaluation Analysis

#### Evaluation Summary

**1/1** Questions Answered

#### Summary of Answer Evaluations

The evaluations indicate a strong performance across the answers provided, with a focus on correctness and completeness.

#### Summary Breakdown:

- Correctness** - The answers demonstrate a high level of accuracy, with all responses deemed correct.
- Completeness** - The responses are fully developed, showcasing thoroughness in addressing the questions.
- Relevance** - The majority of the answers are relevant, effectively aligning with the questions posed.

#### Key Highlights:

The majority of the answers are relevant, effectively aligning with the questions posed.

**100%**  
Correctness

**100%**  
Completeness

**80%**  
Relevance

Response Score: **0.99**

**99.00**

Close

Copyright © 2017-2024 DXwand Inc.      Build-Your-Own AI Chatbot

# Data Management

## Data Pipeline

Knowledge Mining > **Knowledge Bases** Staging License expires o... 0.01% Pages consumption 0.00% Messages consumption NA NA

### Chunking

[All Knowledge Bases](#) / [Suppliers for VF Knowledge Base](#) / [Fusion Experiment](#) / **Chunking** / [Metadata](#) / [Instructions](#) / [Evaluation](#)

Select Documents Unselect All

- RPF-86003 RESP SUPPLIERC.pdf  
# vodafone # SPOC # Request for Quotation  
# SUPPLIERC SUP...

Chunk Size: 2000

Chunk Delimiter ⓘ: ##

Chunk Overlap: 20

Semantic Chunking

[Chunk Documents](#)

Chunks

Text

# vodafone # SPOC # Request for Quotation # SUPPLIERC SUPPC Technical Proposal -  
-- # Summary # 1 Introduction  
.....7 # 2 Supplier  
qualification ..... 8 # 2.1 SUPC  
Company Profile .....8 # 2.2 Digital  
Media & Communication ..... 12 #  
2.2.1 Grow & transform .....13 # 2.2.2  
Process Optimization ..... 13 # 2.2.3 Run &  
Maintenance .....15 # 2.3 References  
..... 16 # 2.3.1 Vodafone  
Experiences .....16 # 2.3.2 Experiences  
within Public Sector ..... 32 # 2.4 Corporate  
Certifications .....34 # 3 Technical answer  
.....

**Document Metadata**

Id: 66504be9-7261-4339-bc9e-d6f887a000f0

Items per page: 15 ▾ | 1 - 15 of 72 Chunks | 1 ▾ of 5 pages < >

[Back To Knowledge Base](#) [Next: Metadata](#) >

Copyright © 2017–2024 DXwand Inc. Build-Your-Own AI Chatbot



# Data Management

## Metadata Management

Knowledge Mining > **Knowledge Bases** Staging License expires o... 0.01% Pages consumption 0.00% Messages consumption NA NA

### Metadata

[All Knowledge Bases](#) / [Suppliers for VF Knowledge Base](#) / [Fusion Experiment](#) / [Chunking](#) / **Metadata** / [Instructions](#) / [Evaluation](#)

**Metadata**

- Document Metadata
- Title Extractor
- Summary Extractor
- Questions Answered Extractor
- Keywords Extractor
- Entities Extractor
- Facts Extractor
- Suppliername

LLM Model

Generate Metadata

**Chunks**

Text

# vodafone # SPOC # Request for Quotation # SUPPLIERC SUPPC Technical Proposal --- # Summary # 1 Introduction .....7 # 2 Supplier qualification .....8 # 2.1 SUPC Company Profile .....8 # 2.2 Digital Media & Communication .....12 # 2.2.1 Grow & transform .....13 # 2.2.2 Process Optimization .....13 # 2.2.3 Run & Maintenance .....15 # 2.3 References .....16 # 2.3.1 Vodafone Experiences .....16 # 2.3.2 Experiences within Public Sector .....32 # 2.4 Corporate Certifications .....34 # 3 Technical answer

**Document Metadata**

**Facts Metadata**

- The document is a Request for Quotation from Vodafone.
- It includes a section titled "Supplier Qualification."
- The document is structured with a summary and numbered sections.
- Section 2.1 contains the SUPC Company Profile.
- Section 2.2 focuses on Digital Media & Communication, with subsections on growth, process optimization, and maintenance.
- Section 2.3 provides references, including Vodafone experiences and experiences within the public sector.

Items per page: 15 ▾ | 1 - 15 of 72 Chunks | 1 ▾ of 5 pages < >

[< Previous: Chunking](#) [Download CSV](#) [Next: Generate Instructions >](#)

Copyright © 2017–2024 DXwand Inc. Build-Your-Own AI Chatbot

# Advanced RAG

## Advanced Retrieval

Knowledge Mining > Knowledge Bases

Staging License expires o... 0.00% Pages consumption 0.00% Messages consumption NA NA

### Vodafone RFP Scoring Knowledge Base

All Knowledge Bases / Vodafone RFP Scoring Knowledge Base

3 Documents 0 Metadata Prompts + New Experiment

Name	Description	Chunking	Evaluation	Retrieval Score	Response Sc
nermine	-	Documents: 1 Size: 2000 Overlap: 20 Semantic: No Chunks: 3	Retrieval Type: Fusion	-	-
Fusion	-	Documents: 1 Size: 20000 Overlap: 0 Semantic: No Chunks: 32	Retrieval Type: Fusion Overall Top K: 3 Embedding Top K: 3 BM25 Top K: 3 Evaluation LLM: gpt-4o-mini Model: gpt-4o-mini Response LLM: gpt-4o-mini Model: gpt-4o-mini Exclude Metadata: No Include Metadata Only: No	100.00	0.00

#### Create Experiment

Q Search

- Baseline
- Fusion
- Hybrid
- Knowledge Graph
- Graph RAG

Select Retrieval Type ^

Cancel Create

Items per page: 15 1 - 2 of 2 Experiments 1 of 1 pages

< Back To All Knowledge Bases

Copyright © 2017-2024 DXwand Inc. Build-Your-Own AI Chatbot

# Advanced RAG

## Evaluation & Publishing

Knowledge Mining > Knowledge Bases Staging License expires o... 0.00% Pages consumption 0.00% Messages consumption NA NA

### Retrieval Evaluation Evaluation Analysis

[All Knowledge Bases](#) / [Vodafone RFPs Knowledge Base](#) / [Fusion Experiment](#) / [Chunking](#) / [Metadata](#) / [Instructions](#) / **Evaluation**

**Response Language**  
Auto

**Overall Top K**  
3

**Embedding Top K**  
3

**BM25 Top K**  
3

**Evaluation LLM Model**  
gpt-4o

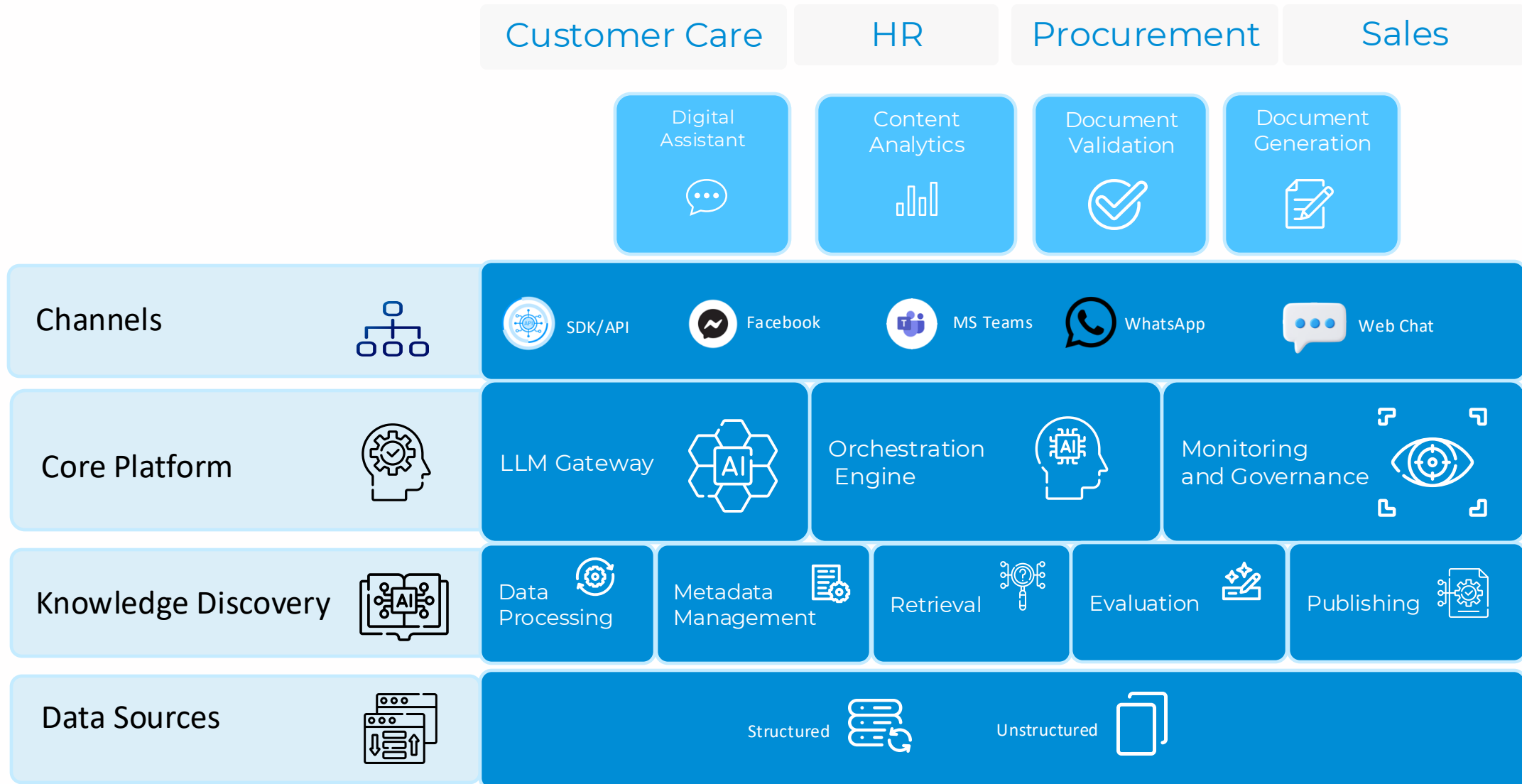
Evaluate

Instruction	Expected Response	Predicted Response	Response Score
		agreed deadlines. <a href="#">View Details</a>	
2 What is the role of the Governance Committee in relation to KPI reviews? <span style="background-color: #28a745; color: white; padding: 2px 5px; border-radius: 3px;">Answered</span> <a href="#">View Analysis</a>	The Governance Committee, consisting of participants from both Vodafone and the supplier, meets periodically to review KPI achievements, discuss performance, and agree on any necessary adjustments.	- The Governance Committee is responsible for verifying KPI achievement. - They share and discuss KPI performance with the Supplier during periodic review meetings, which occur at least quarterly. - In these meetings, the parties may agree to review the KPIs. <a href="#">View Details</a>	0.73
	What are the requirements for the	- The supplier will have access to the Citrix (EAG) platform where all tools related to Vodafone's Network Governance are installed. - The supplier is responsible for requesting authorizations for each tool on the Service Help (IUAM tool). - The supplier must request the use of PCs owned by Vodafone to comply with the	
<b>Total</b>			95.64

[< Previous: Generate Instructions](#)

Copyright © 2017–2024 DXwand Inc. Build-Your-Own AI Chatbot

# DXwand Platform Value Stack



# Offline Light Agentic Inference Architecture

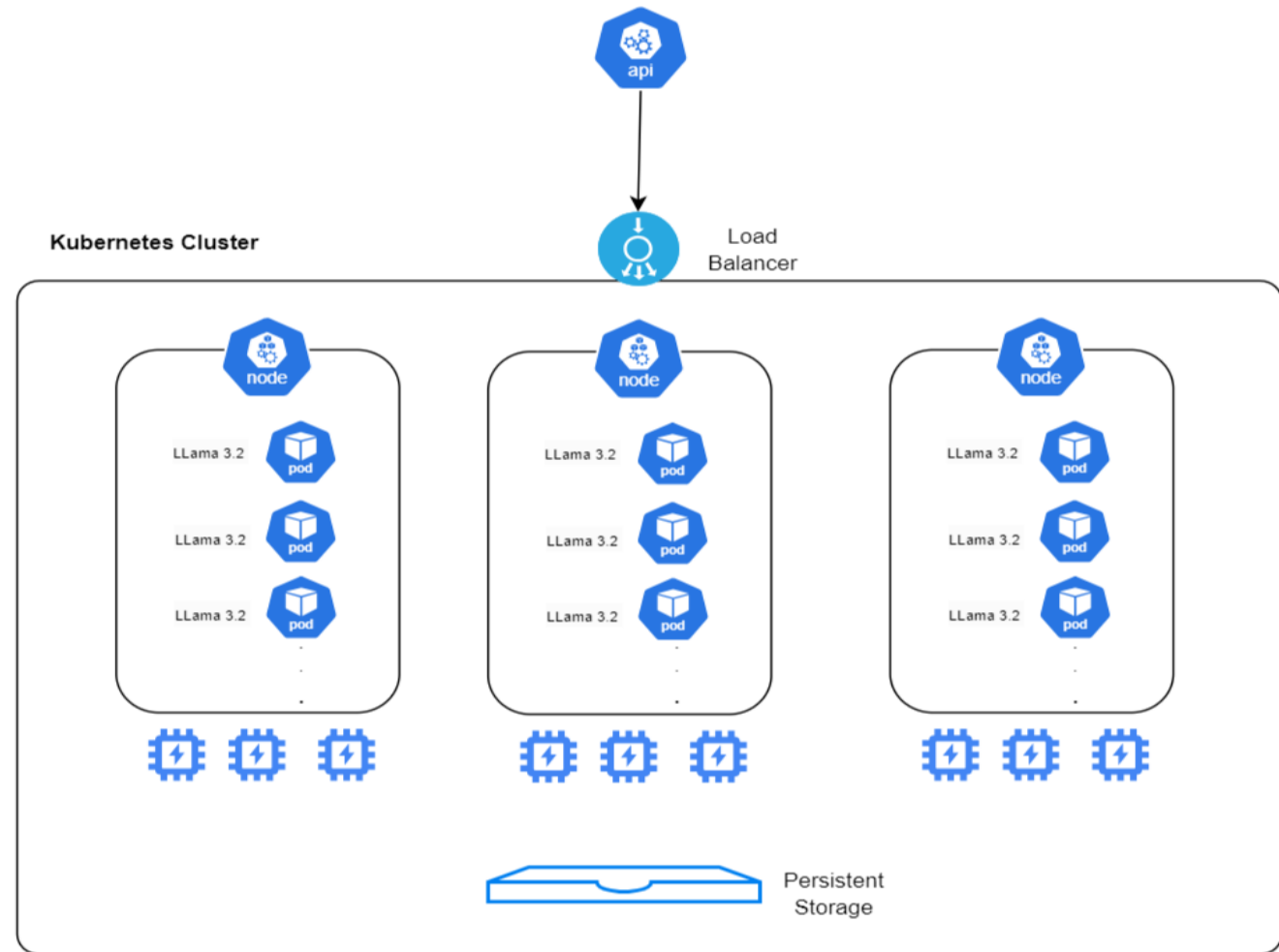
## ORXTRA FERMATA Edition

### GPUs

- **Server Grade Entry Level GPUs** : eg Nvidia L4
- **Consumer Grade GPUs** : e.g. Nvidia RTX Series

### Orchestration

- **Scaling Pods**: depending on number of concurrent requests and model size
- **Client Requests**: API requests to a public endpoint.
- **Load Balancer** (Ingress Controller): Distributes incoming traffic across multiple pods running LLaMA inference on GPU nodes.
- **Inference Pods**: Kubernetes pods running the LLaMA 3.2 model on **GPU Nodes**: Each node in the Kubernetes cluster has one or more GPUs to handle model inference.
- **Persistent Storage** :For storing model weights, logs, or configuration files shared across pods.



Patent Submitted

# Offline Light Agentic Inference Costs

**90% Less Costs**

GPU Model	VRAM	Cost (On-Premises, USD)	Conc Request Per GPU – LLama3.2 1B	Conc Request Per GPU – LLama3.2 3B	Conc Request Per GPU – LLama3.2 9B	Suitability per Model Size
NVIDIA T4	16 GB	\$1,500 - \$2,000	5	2-3	1 (with int8 quantization)	LLaMA 1B (with int8 quantization)
RTX 3090	24 GB	\$1,200 - \$1,500	8	5-6	2-3 (with int8 quantization)	LLaMA 1B, 3B (with quantization)
RTX 4090	24 GB	\$1,600 - \$2,000	8	6-7	3 (with int8 quantization), limited 9B	LLaMA 1B, 3B, limited 9B



**DAWAND** | Empowering Enterprises  
with **Generative AI**



Thank You