InfoGuard AI Labs

SECURESPEAK ENTERPRISE

by InfoGuard AI

✉ serchdavila98@gmail.com

April 18th 2024

TODAY'S AGENDA

1 Problem Statement

2 Our Innovative Solutions

3 Industries Using AI

4 How Does it Work?

5 Tech Stack

6 Chatbot User Interface

7 The Future of RAG Apps

PROBLEM STATEMENT

Data Security and Privacy Concerns in AI Integration

Enterprises seeking to leverage AI technologies grapple with safeguarding sensitive information. The threat of data mishandling raises concerns of legal, financial, and reputational damage, complicating the full utilization of AI capabilities.

Lack of Seamless Integration with Robust Data Protection

The challenge for enterprises lies in finding solutions that seamlessly integrate advanced AI functionalities while ensuring data security. This dilemma often forces businesses to choose between embracing technological advancements and protecting sensitive data, impacting growth and innovation.

Compliance with Evolving Data Privacy Regulations

Enterprises must adapt to the complex and ever-changing global data privacy laws, which can limit their use of AI technologies. Ensuring compliance across different regions adds a layer of challenge, directly affecting the deployment of AI solutions and posing potential legal risks.

PROBLEM STATEMENT

Navigating LLM Compliance and Innovation with Vectara

Startups aiming to utilize large language models (LLMs) often encounter a significant hurdle: complying with stringent data protection regulations like GDPR. This challenge is especially pronounced with closed-source LLMs, which offer limited data handling transparency, making regulatory compliance difficult. As a result, many startups opt for open-source, self-hosted LLMs that afford better control over data privacy but may lack advanced features such as retrieval-augmented generation (RAG).

To bridge this gap, leveraging platforms like SecureSpeak, which leverages tools like Vectara allows startups to integrate the compliance benefits of privacy compliant LLMs with advanced data processing and retrieval capabilities. This combination enables the creation of powerful, regulation-compliant LLM solutions tailored for enterprises. By securely incorporating company-specific information and documents, startups can offer intelligent, customized services, unlocking a vast market in the enterprise sector.

OUR INNOVATIVE SOLUTION

Adaptive Censorship Intelligence

SecureSpeak Enterprise features an the use local language model that learns from previous censorship decisions, enhancing its precision in censoring sensitive data for improved privacy and compliance.

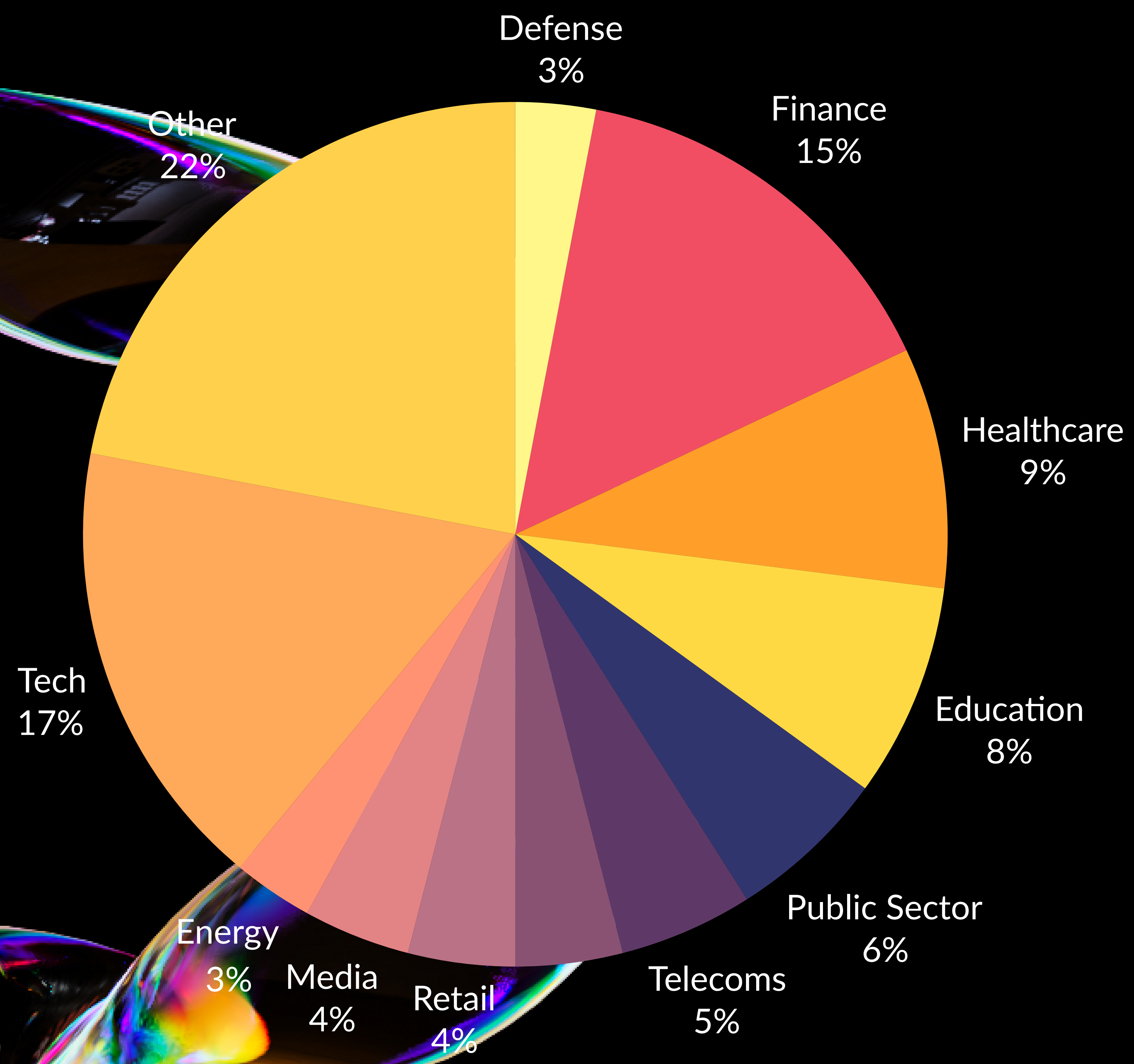
Enhanced Data Traceability

Implements dual-storage capturing of both original and censored inputs, stored in vector and SQL databases, to facilitate comprehensive audit trails and support continuous refinement of censorship practices and interaction quality.

Retrieval-Augmented Generation (RAG) Capability

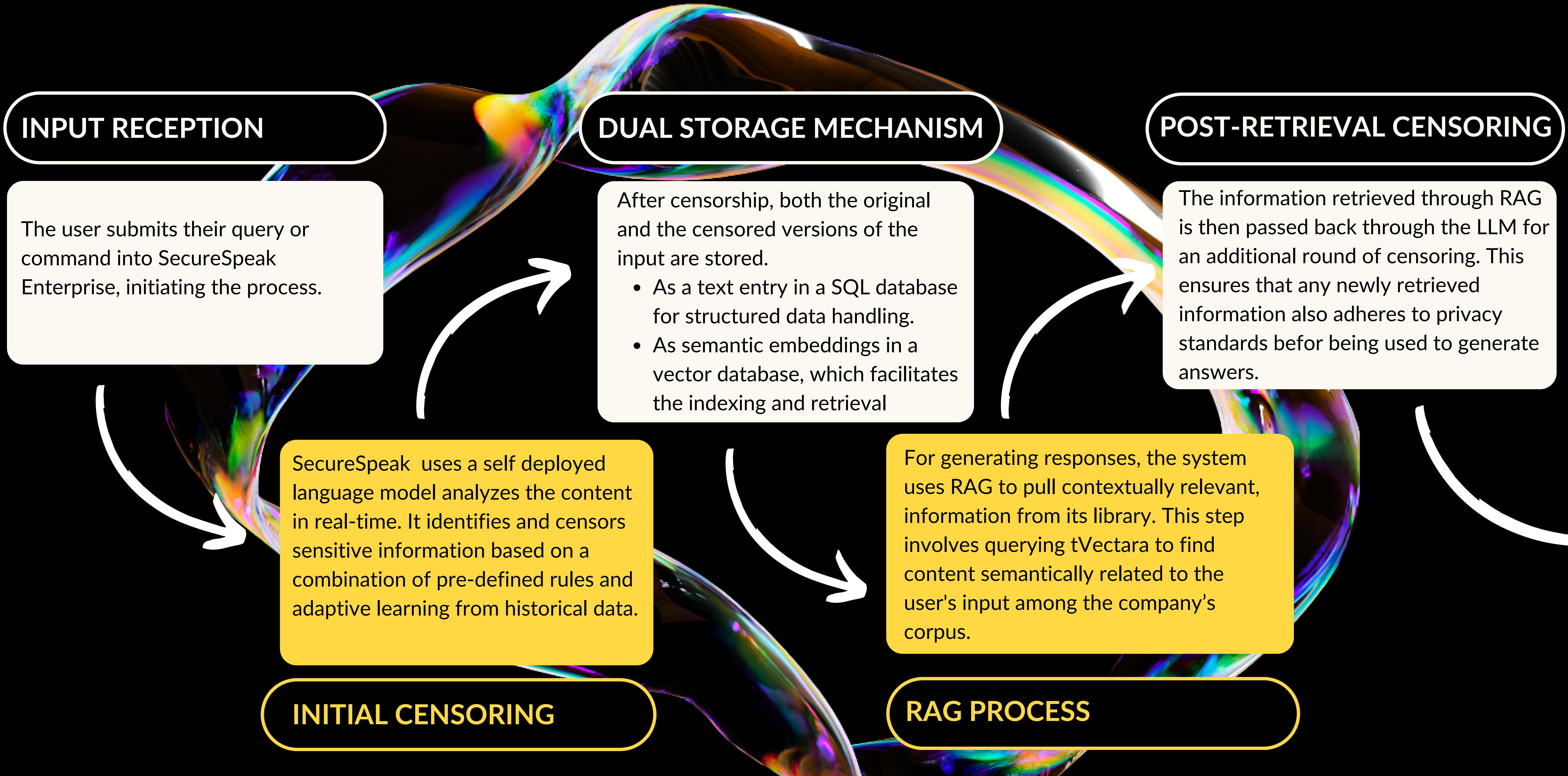
Integrates with Vectara for RAG, allowing the system to enrich chatbot responses with contextually relevant, pre-censored information and comprehensive access to company-specific documents and data, significantly improving the output quality and relevance.

INDUSTRIES USING AI



As evidenced by the 2021 data, AI technology has permeated a multitude of sectors, demonstrating its versatile impact across the global economy. From the substantial stake in finance to the growing adoption in healthcare, education, and the public sector, this distribution highlights a cross-industry reliance on AI to drive innovation, efficiency, and competitive edge. It's a telling snapshot of how integral AI has become, and it's poised to expand further, reshaping the future of how industries operate.

HOW DOES IT WORKS?



INPUT RECEPTION

The user submits their query or command into SecureSpeak Enterprise, initiating the process.

INITIAL CENSORING

SecureSpeak uses a self deployed language model analyzes the content in real-time. It identifies and censors sensitive information based on a combination of pre-defined rules and adaptive learning from historical data.

DUAL STORAGE MECHANISM

After censorship, both the original and the censored versions of the input are stored.

- As a text entry in a SQL database for structured data handling.
- As semantic embeddings in a vector database, which facilitates the indexing and retrieval

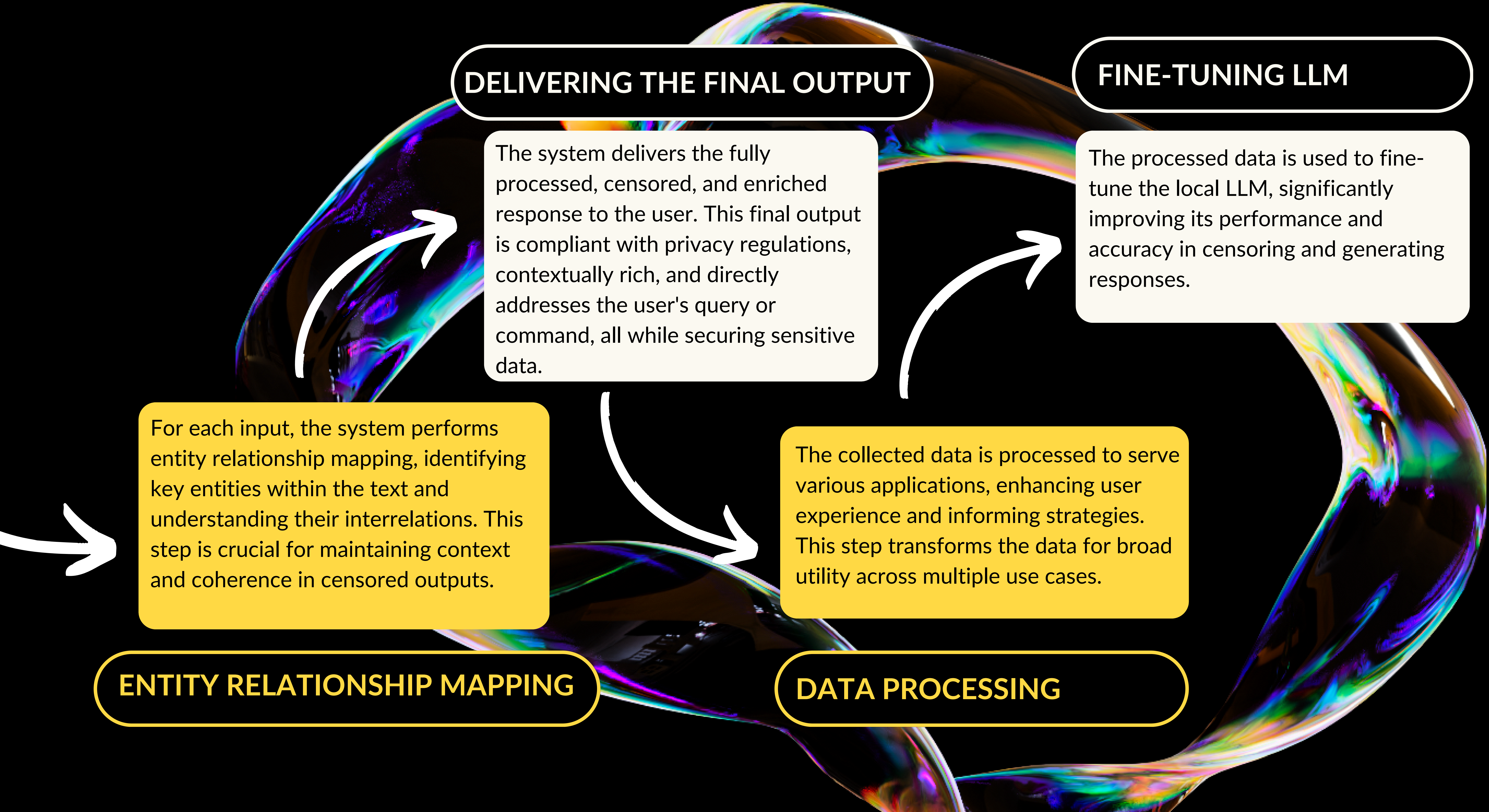
RAG PROCESS

For generating responses, the system uses RAG to pull contextually relevant, information from its library. This step involves querying tVectara to find content semantically related to the user's input among the company’s corpus.

POST-RETRIEVAL CENSORING

The information retrieved through RAG is then passed back through the LLM for an additional round of censoring. This ensures that any newly retrieved information also adheres to privacy standards befor being used to generate answers.

HOW DOES IT WORK?



ENTITY RELATIONSHIP MAPPING

For each input, the system performs entity relationship mapping, identifying key entities within the text and understanding their interrelations. This step is crucial for maintaining context and coherence in censored outputs.

DELIVERING THE FINAL OUTPUT

The system delivers the fully processed, censored, and enriched response to the user. This final output is compliant with privacy regulations, contextually rich, and directly addresses the user's query or command, all while securing sensitive data.

DATA PROCESSING

The collected data is processed to serve various applications, enhancing user experience and informing strategies. This step transforms the data for broad utility across multiple use cases.

FINE-TUNING LLM

The processed data is used to fine-tune the local LLM, significantly improving its performance and accuracy in censoring and generating responses.

TECH STACK

Frontend

Next.js: A React framework used for building the user interface, ensuring a fast, scalable, and SEO-friendly frontend.

Backend

Python: Serves as the primary language for backend development, orchestrating the interaction between various services, handling data processing, and integrating AI models.

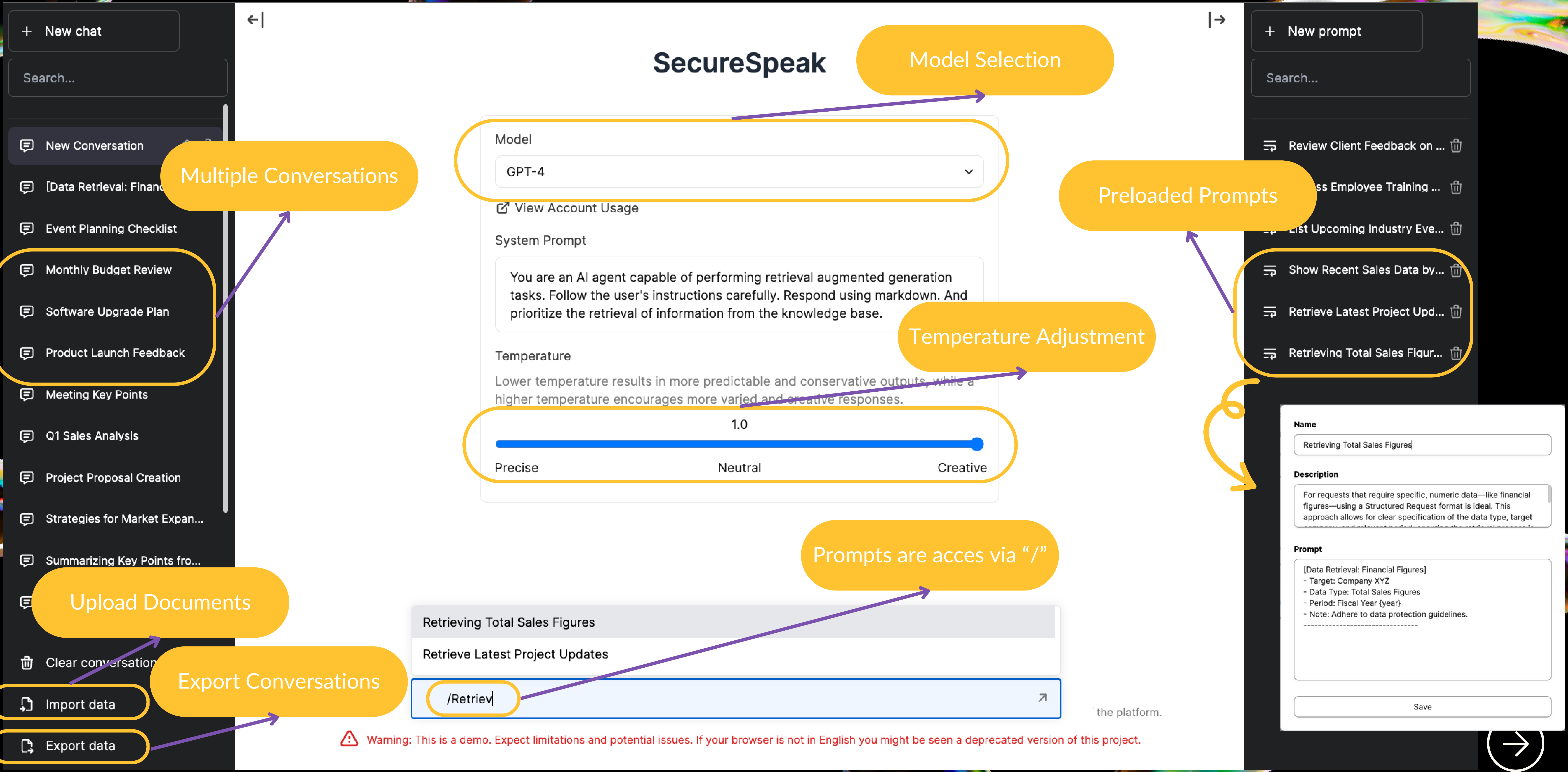
Database and Storage

PostgreSQL: Manages the structured storage of censorship actions, chat history, and other relevant data, ensuring robust data management and retrieval capabilities.

AI and Machine Learning

- Vectara: Provides the retrieval-augmented generation (RAG) functionality, enabling the system to pull contextually relevant information to enhance response quality.
- Llama Index: Used for the vector database component, supporting the censorship mechanism by storing and retrieving semantic embeddings of censored and uncensored data.
- Mixtral LLM: A self-deployed and potentially fine-tuned language model used for initial censorship and processing of inputs.
- GPT-4/3.5 (Other Models): Acts as the main LLM for generating responses, with the flexibility to incorporate additional LLMs for diverse response capabilities.

CHATBOT USER INTERFACE



New chat
Search...
New Conversation
[Data Retrieval: Financ
Event Planning Checklist
Monthly Budget Review
Software Upgrade Plan
Product Launch Feedback
Meeting Key Points
Q1 Sales Analysis
Project Proposal Creation
Strategies for Market Expan...
Summarizing Key Points fro...
Clear conversation
Import data
Export data
New prompt
Search...
Review Client Feedback on ...
ss Employee Training ...
List Upcoming Industry Eve...
Show Recent Sales Data by...
Retrieve Latest Project Upd...
Retrieving Total Sales Figur...
View Account Usage
Retrieving Total Sales Figures
Retrieve Latest Project Updates
Save

SecureSpeak

Model Selection

Multiple Conversations

Preloaded Prompts

Temperature Adjustment

Prompts are acces via “/”

Upload Documents

Export Conversations

Model

GPT-4

System Prompt

You are an AI agent capable of performing retrieval augmented generation tasks. Follow the user's instructions carefully. Respond using markdown. And prioritize the retrieval of information from the knowledge base.

Temperature

Lower temperature results in more predictable and conservative outputs, while a higher temperature encourages more varied and creative responses.

1.0

Precise Neutral Creative

/Retriev

the platform.

⚠ Warning: This is a demo. Expect limitations and potential issues. If your browser is not in English you might be seen a deprecated version of this project.

Name

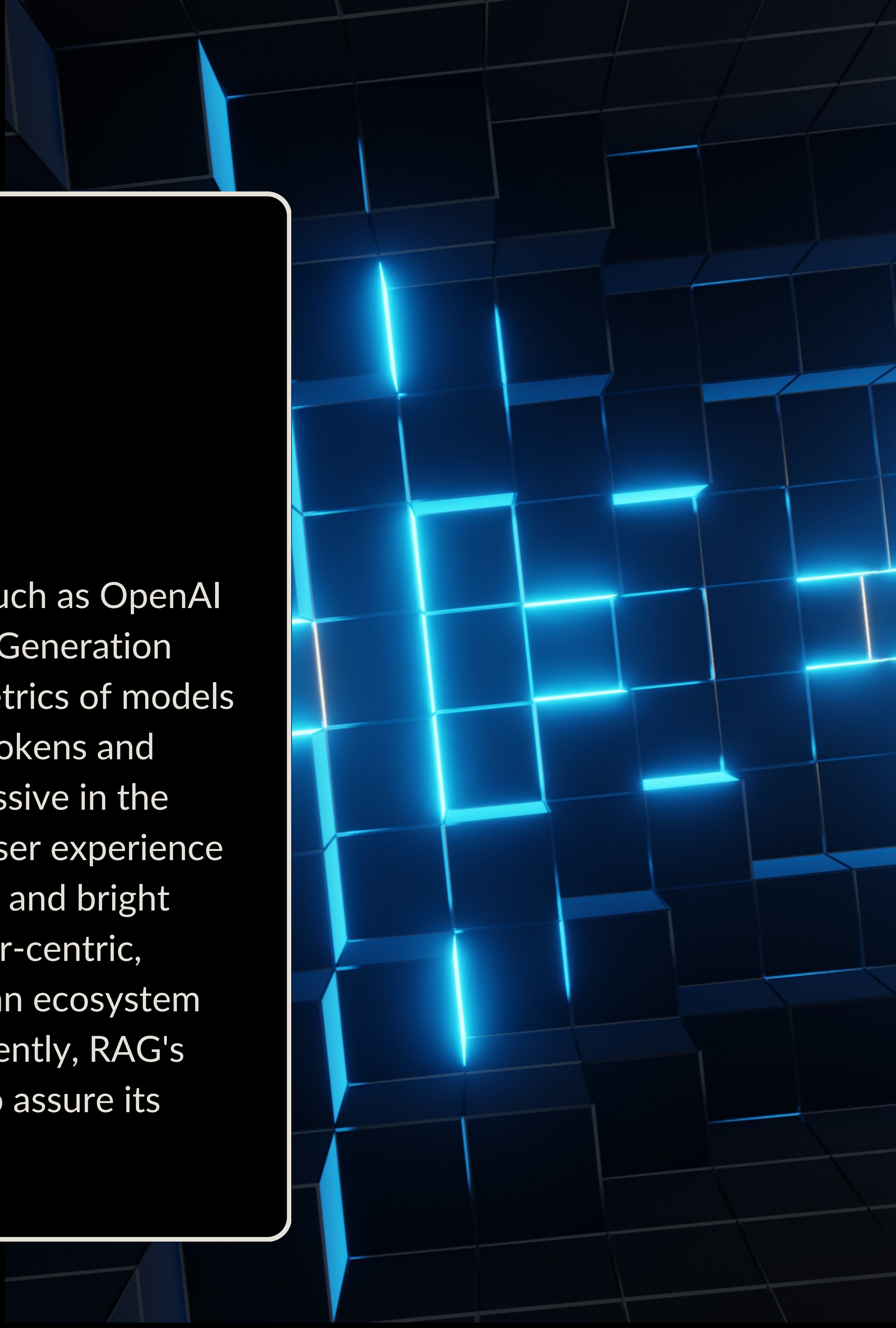
Retrieving Total Sales Figures

Description

For requests that require specific, numeric data—like financial figures—using a Structured Request format is ideal. This approach allows for clear specification of the data type, target

Prompt

```
[Data Retrieval: Financial Figures]
- Target: Company XYZ
- Data Type: Total Sales Figures
- Period: Fiscal Year {year}
- Note: Adhere to data protection guidelines.
--------------------------------
```



THE FUTURE OF RAG APPS

As advancements in AI technology forge ahead, spearheaded by giants such as OpenAI and Google, questions surrounding the viability of Retrieval-Augmented Generation (RAG) systems naturally emerge. However, consider the performance metrics of models like Gemini, which takes approximately 30 seconds to process 360,000 tokens and about a minute for 600,000 tokens. These processing times, while impressive in the context of computational capability, underscore potential concerns for user experience where speed is critical. This very fact highlights the enduring significance and bright future of RAG and usefule platforms like Vectara. By offering a more user-centric, efficient response mechanism, RAG demonstrates its indispensability in an ecosystem where immediate access to accurate information is paramount. Consequently, RAG's adaptability and efficiency suggest not only its current relevance but also assure its place as a mainstay in the strategic development of AI systems.

InfoGuard AI Labs

THANK YOU

for your time and attention

✉ serchdavila98@gmail.com