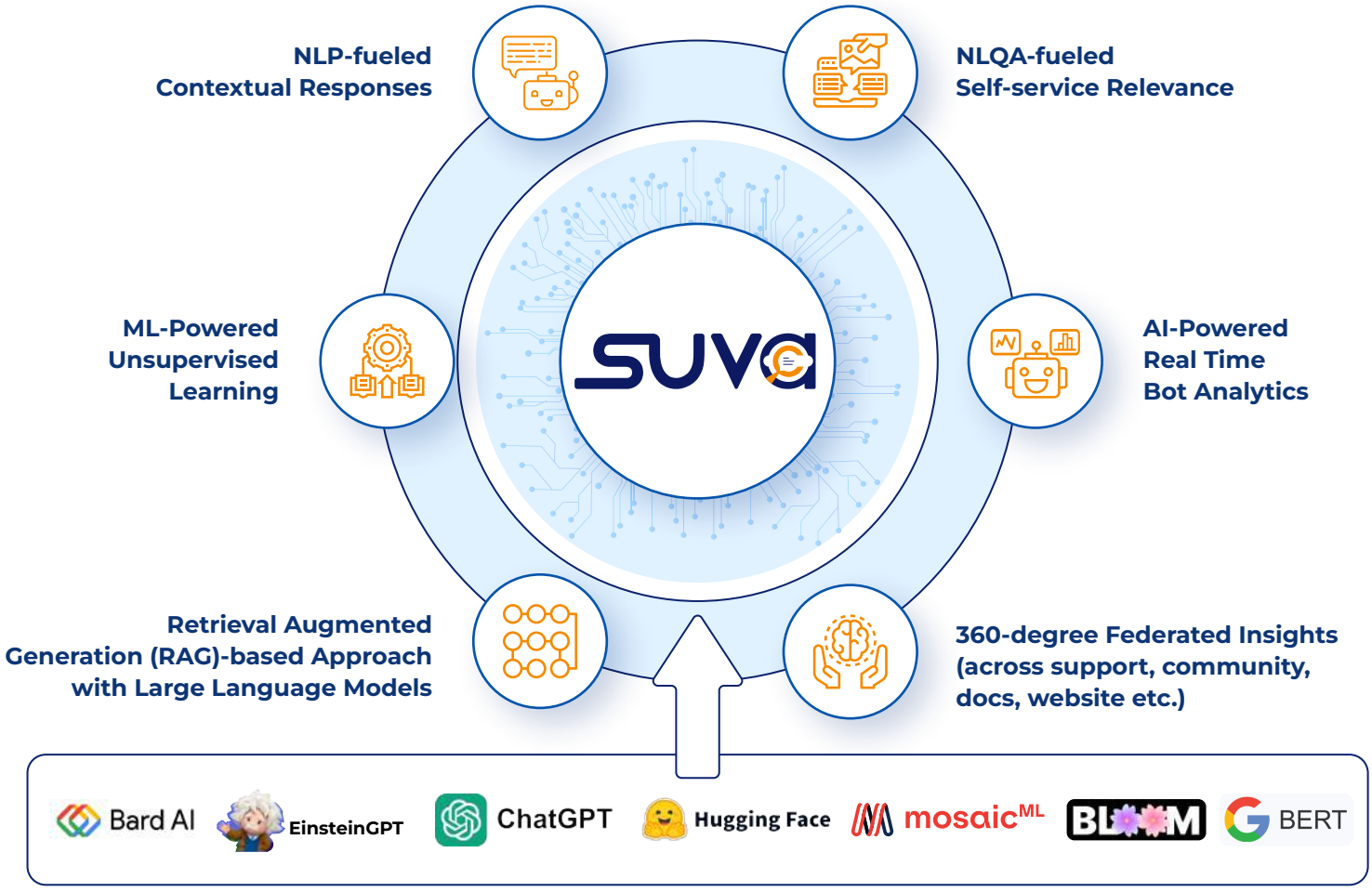
SearchUnify Virtual Assistant

World's First Federated, Information Retrieval Augmented Chatbot for Fine-tuned, Contextual and Intent-driven Conversational Experiences at Scale

Traditional chatbots usually depend on loads of human dialogue data, especially when leveraging supervised machine learning models. Though they can easily deal with single-turn question answering, their reliance on pre-programmed rules, limited fallback support, limited ability to understand and maintain context throughout a conversation and lack of adaptive learning mechanisms, makes them ineffective in handling multi-turn conversations, with context and personalization, at scale.

SUVA, by harnessing the power of retrieval augmented generation, machine learning, NLP, NLQA, generative AI and an insights engine helps resolve customer and employee support queries 24/7 — in the most contextual, personalized, intent-driven manner — with the least amount of user effort.

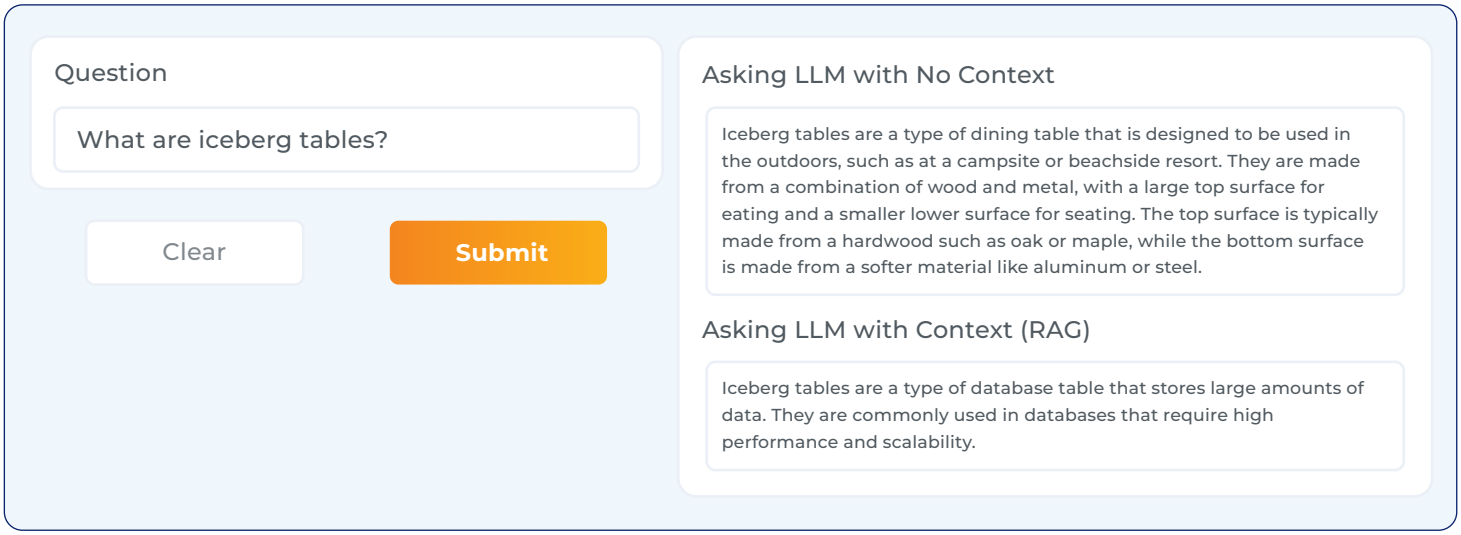


Bring Your Own Large Language Model (BYOL)

Are LLM's the Death of Enterprise Chatbots?

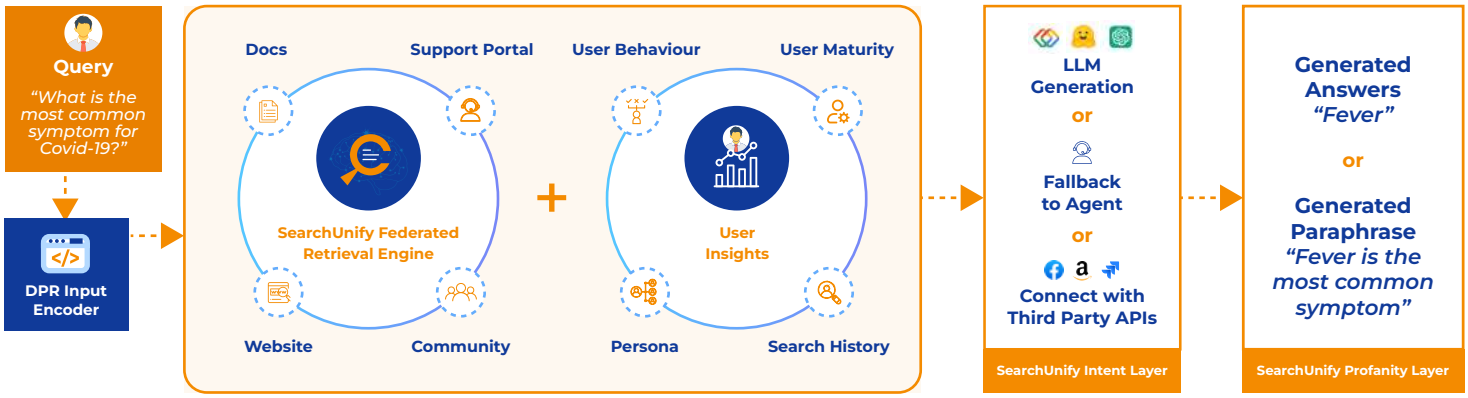
Simple answer, no!

Organizations are in a race to adopt Large Language Models. And while, organizations stand to gain a lot of productivity improvements through LLMs, when a user question is directly sent to the open-source LLM, there is increased potential for hallucinated responses based on the generic dataset the LLM was trained on.



This is where SUVA's **Retrieval Augmented Generation approach** to LLMs comes into play. **Retrieval** involves accessing relevant information or responses from a predefined set of knowledge or data. This can be done using various methods such as keyword matching, semantic similarity, or advanced retrieval algorithms. **Generation**, on the other hand, involves generating human-like responses or outputs based on the retrieved information or context. This can be achieved using techniques like language modeling or neural networks.

By leveraging retrieval augmented generation approach, SUVA fetches relevant information or responses, which are then used as input or context for the generation component to produce more accurate and contextually appropriate responses.

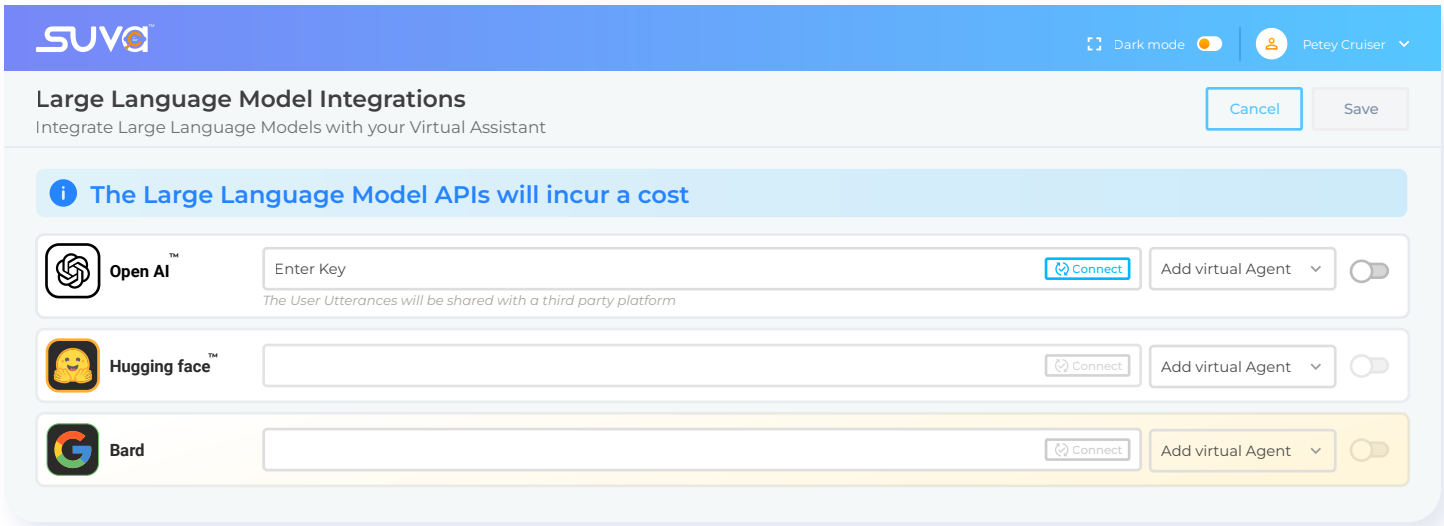


By enhancing the user input with context retrieved from a 360-degree view of the enterprise knowledge base, the LLM-integrated SUVA can more readily generate a contextual response with factual content.

What Differentiates SUVA's LLM Fueled Capabilities from the Rest of the Generative AI Chatbot Players?

1. Ease of Setup/Integration with Leading LLMs

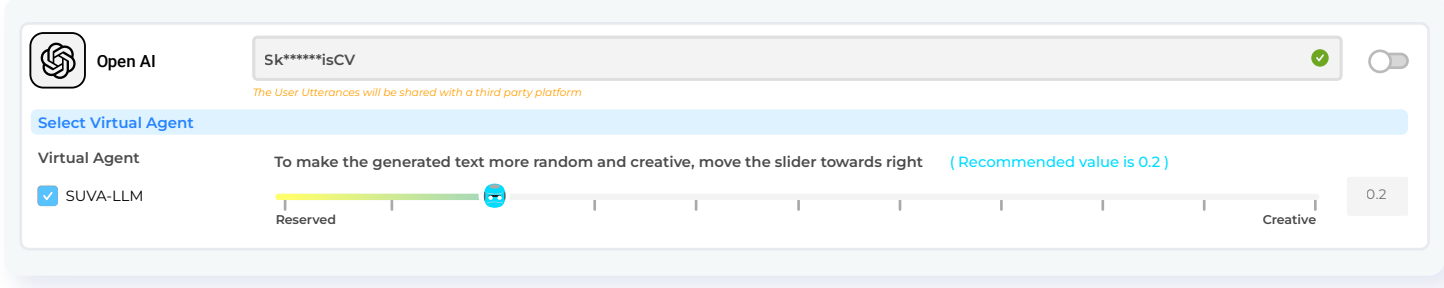
With SUVA's support for leading LLMs including Hugging Face™, BARD, Open AI™ and more, you can easily kickstart leveraging SUVA's LLM-infused capabilities with a plug and play integration by just inputting the API keys for your LLM.



2. More Control over SUVA's Response Humanization

With SUVA, an easy to use UI setting ensures admins get access to **temperature control**, which is a parameter used in chatbot interactions to adjust the randomness and creativity of the responses generated by the model. It affects the level of probabilistic uncertainty and variability in the chatbot's outputs.

A higher temperature (e.g., 0.8 or 1.0) allows for more diverse and creative responses. Conversely, a lower temperature (e.g., 0.2 or 0.5) decreases randomness and makes the chatbot more focused and deterministic. It tends to generate more plausible and conservative responses, aligning closely with frequently observed patterns in the training data.



3. Efficient Costing with Respect to LLM Usage with Segregation of Intents into LLM vs non-LLM Directed Intents

By auditing and filtering out queries that are transactional in nature, SUVA segregates intents and stores queries which get multiple hits by **caching**, thus preventing LLM query hits for repeat, similar queries, and hence reducing LLM usage costs. Further, SUVA gives you the control to allow/block users, thus controlling spam and related costs.

4. Better Intent Recognition Assistance for Tree-based Conversation Flows

LLM-fueled SUVA enables **dynamic branching** within the chatbot flow based on user inputs. Instead of following a fixed, predefined path, the chatbot uses the language model to determine the appropriate next step based on the user's query, allowing for a more flexible and adaptive conversation.

5. User Level Personalization and Access Controls

SUVA respects role-based **access controls** which allows you to define user roles and associate them with specific access privileges when it comes to responses. It takes into account the user's preferences, previous choices, and conversational history to generate personalized and relevant responses. By continuously learning and adapting based on user interactions, SUVA analyzes user feedback, engagement patterns, and preferences to improve future interactions.

6. Fallback Response Generation in Case of LLM Downtime

In situations where the LLM in question is inaccessible, the **fallback mechanism** allows SUVA to provide alternative responses or handle user queries appropriately, from your index (stored in a knowledge base or a separate fallback module). If the fallback mechanisms are unable to address the user's query adequately, SUVA offers the option to seamlessly connect with a human agent or customer support representative.

Ready to learn more? Email us at **info@searchunify.com** or **Request a Demo Now!**