

In the rapidly evolving world of customer service automation, voice agents powered by conversational AI have taken center stage. Companies like **Suprmind** and **Air Canada** leverage cutting-edge technologies such as **speech-to-text** and **text-to-speech pipelines** to create intuitive voice interactions that enhance customer experience. However, a recurring challenge remains: when should these voice agents pull information from an API versus relying on a pre-built knowledge base? With the rise of **retrieval-augmented generation (RAG)** approaches, and platforms like **OpenAI** pushing the boundaries of natural language understanding, it's crucial to understand the limitations and best practices in accessing live versus static facts.

Introduction: The Role of Knowledge Bases and APIs in Voice Agents

Knowledge bases store static or semi-static facts and policies, often curated manually or through automated ingestion pipelines. APIs, on the other hand, provide live, real-time data—such as a customer's order status or billing information—that constantly changes. Voice agents must balance these sources to deliver accurate, timely, and contextually relevant responses.

Before diving deeper, let's identify seven common failure points in voice agent systems that occur when choosing between API calls and knowledge base retrieval.

Seven Failure Points in Voice Agents: Why Source Matters

1. **Stale or outdated information:** Knowledge bases may become outdated, causing agents to provide incorrect or obsolete answers.
2. **Ambiguity in customer-specific queries:** Static knowledge bases struggle with personalized requests that require live data (e.g., order status).
3. **Lack of context-awareness:** Failure to access recent interactions or account changes leads to irrelevant responses.
4. **Incorrect entity recognition and extraction:** Missed or misinterpreted entities impact the agent's ability to retrieve precise information.
5. **Over-reliance on NLG hallucinations:** Large language models (LLMs) may "invent" convincing but false information when the knowledge base lacks relevant data.
6. **Latency issues:** API calls introduce potential delays that conflict with natural conversational flow if not optimized.
7. **Inconsistent source-of-truth management:** Conflicts between knowledge base content and live APIs confuse customers and damage trust.

Understanding RAG Limits and Knowledge Base Hygiene

Retrieval-augmented generation (RAG) combines LLMs like OpenAI's models with an external knowledge base, retrieving relevant documents to ground generated answers. While RAG dramatically improves response accuracy and breadth, its effectiveness depends heavily on the hygiene of the underlying knowledge base.

- **Good hygiene** means regularly updating knowledge base entries, removing obsolete information, and ensuring consistent formatting and indexing.
- **Dirty knowledge bases**, in contrast, bombard the RAG system with conflicting data, leading to contradictory or partially true responses.

Notably, RAG's reliance on static or near-static documents means it cannot serve as the sole source for up-to-the-minute facts like order shipping status or billing balances.

Live Tools as the Source of Truth for Customer-Specific Facts

This is where **live tools like APIs** become indispensable. They provide authoritative, current information that even the best-maintained knowledge bases struggle to match. For example, an *order status API* can provide real-time updates tailored to a customer's specific order, ensuring accurate and timely responses.



Air Canada's voice assistants, for instance, access live APIs to retrieve flight booking status, gate changes, or baggage tracking information—scenarios where static knowledge bases cannot keep pace. Similarly, Suprmind recommends leveraging APIs for data points that change frequently or are personalized, such as billing balances or reward program points, rather than hardcoding these facts in static documents.

Balancing Live Facts vs Static Facts: A Pragmatic Approach

The debate of **live facts vs static facts** is not about choosing one over the other but harnessing both in the right context. Here's a comparative overview:

Source	Best Use Cases	Characteristics	Limitations	Knowledge Base
General	product info, policies, FAQs, troubleshooting guides	Curated, relatively stable, easier to audit	Cannot reflect real-time changes; prone to staleness	API
Customer-specific	data, real-time status, transactional info	Always up-to-date; dynamic; authoritative	Requires robust integration; potential latency	

High-Precision Entity Confirmation and Readback: Enhancing Trust and Accuracy

A critical step when retrieving live data is ensuring high-precision entity recognition and confirmation. For example, when a customer provides an order number, voice agents must:

- Parse the entity accurately from speech-to-text output, tolerating variations and accents (e.g., "B three one seven two").

- Confirm the entity with the customer via text-to-speech to avoid mistakes before API queries.
- Read back retrieved information clearly and succinctly to maintain conversational flow.

This approach minimizes errors and builds customer confidence, preventing scenarios where wrong data leads to frustration or escalations.

Case Study: Air Canada's Integration of APIs and Knowledge Bases

Air Canada exemplifies best practices in voice agent architecture by <https://suprmind.ai/hub/insights/voice-ai-hallucinations/> combining static knowledge with live APIs. Their system uses a well-maintained knowledge base for common travel policies while dynamically querying APIs for flight-specific details, booking changes, or loyalty points. The synergy between their RAG-powered conversational AI and live APIs ensures customers receive immediate and accurate responses.



Recommendations for Implementing Voice Agents: When to Use APIs

1. **Use APIs when the information is customer-specific and frequently changing.** For instance, order status, billing balances, flight updates, or reward points.
2. **Leverage knowledge bases for fixed policies, generic FAQs, or broad product information.**
3. **Implement entity confirmation strategies especially for critical identifiers.** Validate inputs such as account numbers, invoice IDs, or order references before querying APIs.
4. **Continuously maintain and prune knowledge bases.** Ensure RAG systems only work with clean, relevant data to avoid misinformation.
5. **Optimize speech-to-text and text-to-speech components.** Ensure entity extraction accuracy and natural yet precise readback.
6. **Monitor failure modes closely.** Identify when agents fallback on hallucinations due to missing data and resolve source gaps.
7. **Define clear source-of-truth hierarchies.** Formally document that live APIs supersede knowledge base facts for dynamic content.

Conclusion

In the era of intelligent voice assistants powered by technologies from OpenAI and innovators like Suprmind, striking the right balance between APIs and knowledge bases is crucial. Live data accessed via APIs offers the authoritative, up-to-date responses necessary for personalized customer interactions, while a well-kept knowledge base remains invaluable for static information. Recognizing the seven failure points in voice agents provides a roadmap for leveraging these tools effectively, minimizing errors, and delivering seamless experiences that customers trust.

Voice agents should never gamble on outdated or ambiguous facts when live APIs can provide precision and immediacy. Maintaining robust speech pipelines and high-precision entity confirmation preserves accuracy and customer confidence. Ultimately, the highest-performing voice systems thoughtfully integrate both worlds—reflecting a sophisticated understanding of when, why, and how to use APIs instead of, or alongside, knowledge bases.