Retrieval-Augmented Generation (RAG) has revolutionized how voice agents and conversational AI handle vast knowledge bases by pairing large language models with external documents. Companies like **Suprmind** and **Air Canada** leverage RAG-powered solutions to improve customer interactions, carefully integrating tools like speech-to-text and text-to-speech pipelines. **OpenAI** remains a leading innovator in this domain, pushing the boundaries of what models can achieve.

Yet, despite its promise, RAG systems often still return wrong answers even when the right document is referenced. In this post, we'll unpack **seven failure points** behind voice agent errors, explore the inherent limits of RAG and knowledge base hygiene, and explain why live tools and high-precision entity confirmation are vital to reducing errors like *grounded summarization errors* and *hallucination on hallucination*. Understanding these issues is critical for organizations deploying complex conversational AI in customer-centric environments.

What Is RAG and Its Promise — And Problem?

RAG combines language models with retrieval systems.

When asked a question, the system retrieves relevant documents and then conditions the generative model on those documents to produce an answer. The promise: answers grounded in real data, not just the model's inherent knowledge.

However, the model's output depends heavily on:

- The quality and relevance of your knowledge base
- How well the model interprets the retrieved documents
- The surrounding conversational context and speech pipeline accuracy

To quote an industry expert from **Suprmind**, "Having the right document https://suprmind.ai/hub/insights/voice-ai-hallucinations/ is necessary but not sufficient; the model must truly *read* and process it correctly." This is where common failures emerge.

Seven Failure Points in Voice Agents Using RAG

Our team's years of contact center and AI implementation experience — especially with real telephony audio — have identified these critical failure points:

Failure Point Description Impact on Answer Quality 1. ASR (Automatic Speech Recognition) Mishearings Speech-to-text pipeline errors cause wrong query transcription. Causes wrong retrieval, leading RAG astray. 2. Retrieval Irrelevance Even with the right document in KB, the top retrieved passages may not be relevant. Model conditions on wrong evidence. 3. Model Misreads Evidence The language model interprets retrieved text incorrectly. Generates plausible but factually wrong answers (grounded summarization errors). 4. Hallucination on Hallucination Chains of errors where model "hallucinates" details beyond legitimate context. Compounds error severity over interaction turns. 5. Outdated or Dirty Knowledge Bases KB content may be stale, incomplete, or corrupted. Renders answers factually obsolete or biased. 6. Inadequate Entity Confirmation Failures to confirm critical entities (e.g., booking reference, account number) with user. Undermines answer correctness and trust. 7. Text-to-Speech & Prompt Guardrail Limits Inadequate readback clarity and prompt-only guardrails fail to catch errors. User misinterpretation and uncorrected model flaws.

Knowledge Base Hygiene and the RAG Ceiling

One of the most misunderstood aspects of RAG implementations is that simply having the "right document" in your knowledge base doesn't guarantee accuracy. Dirty or outdated KBs introduce bias and misinformation that the language model consumes as truth.



Even high-quality KBs are only as good as their indexing and retrieval pipelines. *Retrieval irrelevance* remains a stubborn problem when semantic search algorithms misjudge relevance or prioritize passages not best suited for answering the question.

A lead developer at **Air Canada** put it this way: "We realized that a well-maintained knowledge base is the foundation, but without cleaning, updates, and contextual retrieval tuning, RAG answers would wander into hallucination territory."

Sources of Truth: Live Tools Over Static KBs

For customer-specific facts—such as flight schedules, booking statuses, or billing information—live backend systems remain the ultimate source of truth. Relying solely on static knowledge bases causes glaring errors, especially when quick changes occur:

- Flight delays or cancellations
- Last-minute itinerary changes
- Account balance updates

Speech-to-text pipelines funnel customer inputs into queries that trigger RAG retrieval, but unless the AI queries live systems dynamically, stale answers will proliferate.

OpenAI and partner platforms increasingly support integrating live tools into conversational agents, enabling validation against live databases rather than static documents.

High-Precision Entity Confirmation and Readback

Another key to mitigating hallucination risks involves high-precision confirmation and intelligent readback strategies:

1. **Entity Confirmation:** Voice agents should explicitly confirm critical details received from users or generated by models (e.g., "Did you say your booking number is B three one seven two?"). This avoids misinterpretations cascading through the conversation.
2. **Readback Clarity:** The text-to-speech system must enunciate entities clearly and include disambiguation where needed. Subtle misreadings can cause significant user confusion.
3. **Guardrails Beyond Prompts:** Prompt-level instructions alone cannot enforce consistency or error checks. Integrated business logic at the orchestration layer provides stronger guardrails to catch and correct mistakes.

Our notebook of call snippets includes many such examples—for instance, a booking reference like “B three one seven two” must be read back literally and confirmed; even a tiny misread audio snippet derails the lookup and the final answer.

Addressing Grounded Summarization Errors and Hallucination on Hallucination

Grounded summarization errors occur when the model summarizes content from retrieved documents incorrectly. These are not simple hallucinations but errors in processing actual evidence. When compounded with hallucinations generated downstream—coined *hallucination on hallucination*—the final answer can completely misrepresent the truth.

Error Type Cause Example Mitigation Grounded Summarization Error Model misinterprets retrieved evidence Model reads "refund denied" as "refund granted" Entity-level extraction checks; multiple candidate answer scoring Hallucination on Hallucination Error built upon prior hallucinated info Model invents reasons for refund denial unsupported by KB Live tool validation; confirmation dialogs

Conclusion: Beyond RAG Alone—Unlocking Real Accuracy in Voice Agents

RAG is a fundamental breakthrough, especially when integrated with speech-to-text and text-to-speech pipelines by companies like **Suprmind**, **Air Canada**, and developments from **OpenAI**. Still, the road to accurate, trustworthy customer interactions is paved with attention to these failure points:

- Maintaining pristine, updated knowledge bases
- Enhancing semantic retrieval relevance
- Mitigating model evidence misreading and hallucinations
- Querying live operational databases for time-sensitive facts
- Implementing rigorous entity confirmation and clear readback
- Adding guardrails beyond prompt engineering for systemic error handling

In my 12 years of contact center QA and voice agent deployment, I've learned to never accept a model’s first answer as gospel. The source of truth often comes from live tools and user confirmation, not just a retrieved document or a prompt. Reducing *grounded summarization errors* and stopping *hallucination on hallucination* requires a multi-layered approach—something every AI leader should prioritize, especially when serving millions of real-world voice interactions daily.

What is your source of truth for customer-specific facts in your AI systems? If it’s only a static knowledge base retrieved via RAG, you may already be on a path toward unseen errors.

