

Retrieval-Augmented Generation (RAG) has emerged as a powerful technique for enhancing AI responses by grounding answers in relevant documents. Companies like **Suprmind** and **Air Canada** are leveraging RAG in sophisticated voice agents, making use of advanced *speech-to-text* and *text-to-speech* pipelines. Meanwhile, **OpenAI** continues to develop large language models that incorporate retrieval for improved accuracy.

Yet, even when RAG has the “right” document in its knowledge base, users report **grounded summarization errors** and frustrating “hallucination on hallucination” effects. Why does this happen? In this blog post, we’ll dissect the core seven failure points in voice agent pipelines involving RAG, explain the limits of knowledge base hygiene, emphasize the importance of live tools as sources of truth, and highlight the crucial role of high-precision entity confirmation and readback techniques for improving voice agent accuracy.

Seven Failure Points in Voice Agents Using RAG

Ever notice how from working with telephony-grade voice agents over the past 12 years and shipping ivr-to-voice ai migrations, i’ve kept a careful notebook of real call snippets and failure logs. Experience reveals that correct retrieval of a document is necessary but not sufficient for correct answers. The pipeline breaks down in multiple places:

- 1. **Speech-to-text errors:** Even state-of-the-art ASR models can mishear customer utterances, which leads to incorrect query formulation for retrieval.
- 2. **Retrieval relevance mistakes:** The RAG retriever may not rank the truly relevant document top, pushing useful evidence out of immediate context.
- 3. **Model misreads evidence:** Once the document is retrieved, summarization errors occur. The model can misunderstand or misinterpret nuances, especially numeric and entity-specific details.
- 4. **Hallucination on hallucination:** When the model starts to “guess” missing information, it compounds errors, straying further from the grounded facts.
- 5. **Knowledge base hygiene gaps:** Outdated or inconsistent documents remain in the knowledge base, causing conflicts and confusion for retrieval models.
- 6. **Live tools out of sync:** Systems like CRM or inventory databases, which hold the true customer-specific facts, are often siloed and disconnected from the RAG pipeline.
- 7. **Low precision on entity confirmation:** Lack of rigorous readback and verification of key entities such as reservation numbers or account ids leads to user frustration and errors.

Table: Mapping Failure Points to Symptoms

Failure Point	Symptom	Impact on User Experience
Speech-to-text errors	Incorrect query input	Wrong documents retrieved, irrelevant answers
Retrieval relevance mistakes	Missing critical info	Incomplete or false responses
Model misreads evidence	Grounded summarization errors	Misinformation despite correct doc
Hallucination on hallucination	Made-up facts layered on errors	Confusing and untrustworthy answers
Knowledge base hygiene gaps	Conflicting or outdated documents	Inconsistency and contradiction
Live tools out of sync	Discrepancy between AI and backend	User frustration, loss of confidence
Low precision on entity confirmation	Incorrect readback of account or ref numbers	Call escalations, user annoyance

RAG Limits and the Challenge of Knowledge Base Hygiene

The basic promise of RAG is that by retrieving documents and conditioning generation on them, hallucinations should diminish. Yet, in practice, “hallucination on hallucination” surfaces often.

Why?

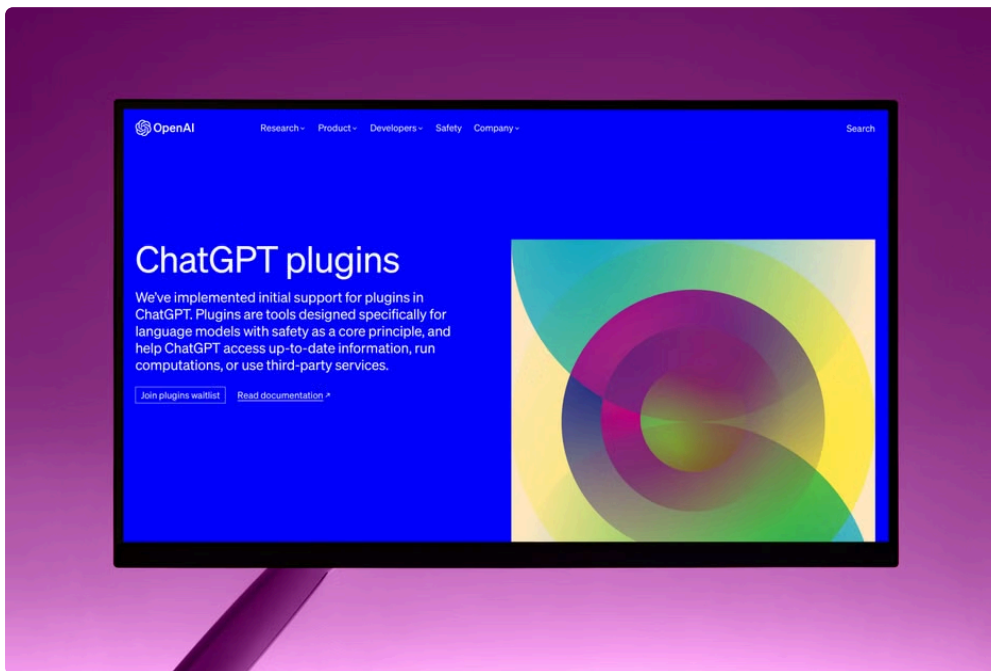
- **Misalignment between indexing and retrieval:** Documents may be indexed with ambiguous or poor metadata. Queries that rely on customer-specific phrasing may not match confidently.
- **Document staleness:** Knowledge bases can harbor stale information, resulting in retrieval of outdated policies or facts—for example, Air Canada flight rebooking rules that have recently changed and not yet reflected.
- **Inadequate context windows:** Even with the right document, models often condense information to fit token constraints, risking loss of critical detail and summary distortions.
- **Blind spots in grounding:** LLMs may “skim” key figures or pieces of information and thus generate plausible but incorrect restatements.

Suprmind, working with retail clients, stresses continuous knowledge base curation, involving monitoring retrieval logs and user feedback loops to retire obsolete documents swiftly. This discipline is essential to reducing grounded summarization errors.

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

One of the most underrated points is that the authoritative truth about a customer's booking, suprmind.ai contract status, or inventory availability often lives in live backend systems and not in static documents.

Integrating live tool APIs into the conversational pipeline ensures:



- Real-time access to customer data, avoiding reliance on potentially outdated knowledge base.
- Enhanced confidence in answers, since live validation supplants model predictions.
- Reduced hallucination on entity values, as data is fetched directly and faithfully.

For example, Air Canada’s voice agents integrate live flight management databases with speech-to-text frontends and RAG retrieval components. When a user requests pick-up details or rebooking options, the agent cross-checks live APIs before answering, blending natural language generation with factual data retrieval.

High-Precision Entity Confirmation and Readback: The Final Line of Defense

At Suprmind, we always emphasize that a conversation is not complete until critical entities—the “B three one seven two” or reservation codes—are correctly interpreted and confirmed back to the caller.



Why is this essential?

- Model misreads evidence often affect low-level details that matter enormously in telephony contexts.
- Without explicit confirmation, users inevitably catch errors mid-call or post-call.
- Strict readback protocols reduce error propagation and call escalations.

Voice agent flows should implement:

- Phonetic confirmation logic, recognizing common recognition confusions in the speech-to-text step.
- Explicit user verification prompts, where the agent restates the entity and requests affirmation.
- Fallback options that gracefully route uncertain cases to human agents.

OpenAI’s recent work on this is promising—fine-tuning LLMs to align better with noisy ASR outputs and incorporate structured entity verification within response generation.

Conclusion: What’s the Source of Truth for RAG-based Voice Agents?

When investigating errors in RAG-based voice agents, always ask: *“What is the source of truth for this sentence?”* Is it the retrieved document? The live backend system? Or a hypothesized guess crafted by the model?

Only by tracing back every answer to a verifiable source—preferably live tools where possible—and by engineering rigorous format and confirmation layers around entities can we minimize the “hallucination on hallucination” problem.

RAG remains a compelling approach, but its success depends on comprehensive pipeline design—spanning ASR, retrieval, grounding, backend integration, and output verification. The collaboration of companies like Suprmind, Air Canada, and OpenAI in solving these challenges is setting the stage for truly trustworthy, human-friendly, AI-powered voice agents.