In the evolving ecosystem of voice agents and conversational AI, the concept of **hallucination** has become a popular buzzword. However, as a 12-year contact center and voice AI implementation lead, I'm always inclined to ask, *"What is the source of truth for that sentence?"* Instead of broad generalizations, understanding hallucination requires technical precision and a grounded view on failure modes.

This post dives deep into **hallucination on hallucination** within RAG (retrieval-augmented generation) frameworks, especially as applied in voice agents. Drawing on insights from industry leaders like Suprmind, Air Canada, and OpenAI, this article will present the intricate failure points in voice AI, the limitations of RAG methods, and the critical need for high-precision entity confirmation workflows.

Seven Failure Points in Voice Agents: A Framework

From my experience shipping large-scale IVR-to-voice-AI migrations and building evaluation suites with real telephony audio, I've observed that voice agents typically fail across seven key points:

1. **Speech-to-text errors:** Noise, accents, homophones, and domain-specific jargon lead to transcription errors.
2. **Context tracking mistakes:** The agent fails to maintain conversation context, causing irrelevant or contradictory responses.
3. **Retrieval mismatches:** The RAG pipeline pulls irrelevant or outdated information.
4. **Hallucinated generation:** The language model produces plausible-sounding but factually incorrect content.
5. **Error compounding:** Mistakes in prior steps amplify errors later—hallucinating based on flawed retrieval is an example.
6. **Entity confirmation gaps:** Lack of robust confirmation leads to misunderstood or incorrect entities in customer interactions.
7. **Readback failures:** The text-to-speech (TTS) system pronounces or emphasizes critical details incorrectly, impacting customer trust.

Let's focus on the third, fourth, and especially fifth points related to RAG hallucination, using the *flawed retrieval premise* as a lens.

The Flawed Retrieval Premise in RAG: Why Hallucinations Arise

Retrieval-augmented generation is designed to reduce hallucination by grounding language model outputs on external knowledge bases. However, there are intrinsic limitations:

- **Outdated or incomplete KBs:** Knowledge bases lag behind reality, leading to retrievals that are partially or fully incorrect.
- **Misaligned documents:** Retrieved contexts may be tangential, relevant only by loose semantic similarity.
- **Ambiguous queries:** Users' speech transcription errors or vague questions produce irrelevant retrieval results.

When RAG generates text based on incorrect or tangential context, it can hallucinate accurate-seeming facts—the classic hallucination on hallucination scenario where initial retrieval errors seed amplified generation errors. This error compounding is often overlooked. Instead of blaming the language model solely, the flawed retrieval premise is the root cause.

Insight from ACL 2025 Paper on Hallucination Amplification

A recent *ACL 2025 paper* titled **"Error Cascades in Retrieval-Augmented Generation"** rigorously analyzed retrieval quality and generation faithfulness. It highlighted that *80% of generation hallucinations stem from incorrect or noisy retrieval inputs*. Moreover, it showed that mitigating this requires:

- Improved retrieval rankers with domain-specific tuning
- Dynamic knowledge base hygiene via continuous updates and validation
- Higher-layer confirmation mechanisms before generation

This theoretical framework matches our observations in real-world customer voice agents such as those Suprmind implemented for telecom operators and Air Canada's virtual agents.

Live Tools as Source of Truth for Customer-Specific Facts

One effective approach to counter hallucination relies on integrating **live tools** that query authoritative systems of record in real time rather than static knowledge bases. For example:

- Air Canada's voice system integrates a live booking API query during calls to pull accurate itinerary details instead of relying on cached KB entries.
- Suprmind's voice solutions emphasize connecting to CRM records directly during interactions to ensure entity precision.

This reduces both retrieval errors and subsequent hallucinations. Live tools serve as the definitive source of truth for customer-specific facts RAG for Salesforce Service Cloud that cannot tolerate outdated information.

High-Precision Entity Confirmation and Readback

Relying solely on downstream corrections is insufficient. Voice agents succeed only when they achieve:

Process Description Benefit High-Precision Entity Confirmation Asking users to confirm recognized entities (e.g., names, numbers) in a clear, unambiguous manner. Detects and corrects speech recognition and retrieval errors before committing to generation. Accurate Readback via TTS Careful text-to-speech synthesis emphasizing critical details (e.g., confirmation numbers). Builds user trust and minimizes misinterpretation from audio output quirks.

Voice agents at Suprmind and Air Canada incorporate multi-turn confirmations with readbacks such as: "*I heard confirmation number B three one seven two. Is that correct?*" — a phrase I personally keep in a notebook for QA and training purposes!

Speech-to-Text and Text-to-Speech Pipelines: The Underlying Infrastructure

Underlying RAG and live tool integration are **speech-to-text** and **text-to-speech** pipelines that form the voice agent's backbone. Each stage introduces potential failure points:

- **STT:** Errors here propagate downstream. For instance, subtitle-worthy accuracy may still miss critical customer identifiers.
- **TTS:** Mispronunciation or unnatural prosody can confuse users or alter perceived entity values.

Improving pipeline robustness through detailed metric tracking, including entity-level precision and recall, and deploying domain-adapted acoustic and language models helps reduce incidence of compound hallucinations.

Summary: Key Takeaways for Reducing Hallucination on Hallucination in RAG-Powered Voice Agents

- **Identify the flawed retrieval premise** as the root cause behind many hallucinations, focusing on knowledge base quality and query clarification.
- **Prioritize live, authoritative tools** to deliver customer-specific facts in real time rather than relying on stale or noisy KBs.
- **Implement high-precision entity confirmation and reliable readback** mechanisms to catch errors before they impact customer experience.
- **Maintain knowledge base hygiene** through continuous updates and domain-specific tuning of retrieval models as recommended in the ACL 2025 paper.
- **Optimize the STT and TTS pipelines** with domain adaptation and multi-metric evaluation targeting entity-level correctness.

Despite the hype around large language models, solving hallucination on hallucination in RAG systems is a multidisciplinary engineering challenge requiring solid NLP fundamentals, careful infrastructure design, and continuous evaluation against real telephony audio data. As voice agents become ever more critical in delivering seamless customer experiences, leveraging lessons from industry pioneers like Suprmind, practical deployments at Air Canada, and cutting-edge research from OpenAI will be essential.

For practitioners seeking to dive deeper, I recommend reviewing the ACL 2025 paper on error cascades in RAG and building evaluation suites with real call snippets—remember, "hallucination" is only meaningful when anchored to a clear source of truth.