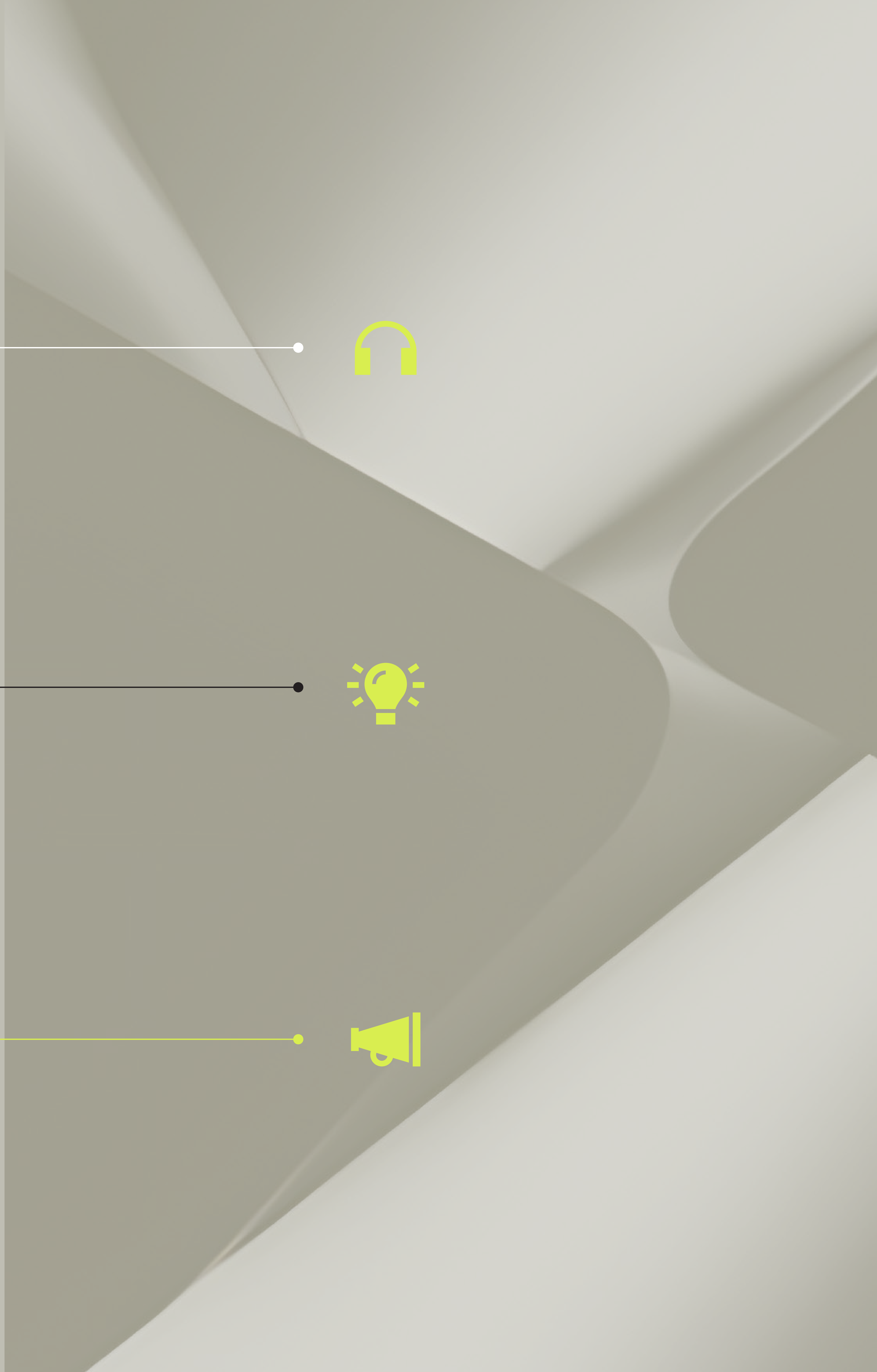
PolyAI

GUIDE

The anatomy of an AI agent.



Listening, understanding, and speaking are the foundations of effective communication.

But let’s face it: real conversations don’t always follow a simple, linear path. We jump back and forth in time, slowing down to expand on some moments and fast-forwarding through others, even editing what we say as we go.

For example, sometimes you ask someone a question, and they respond with another question. This response is not necessarily a way to avoid answering the original question–in fact, it’s often a way to seek clarity in hopes of providing a better, more helpful answer. In other words, **conversations trend towards fluidity and unpredictability.**

Beyond IVRs: AI that listens, understands and adapts.

0:09
Hi, how can I help?

0:12
Hey, I need help with my order.

0:21
Of course, is this about the order you placed last week?

0:03
Yeah

0:16
Ok, how can I help with your order?

0:18
I need to change the delivery address.

0:21
Ok. Please can you confirm your billing ZIP code?

Starts with a friendly, open-ended question inviting the customer to explain their issue in their own words.

Uses contextual data (like phone number recognition) to anticipate the right order, reducing the number of questions needed.

Doesn't assume the issue and keeps the conversation open-ended so the customer can explain what they need.

Only asks for authentication when necessary, avoiding unnecessary steps, it focuses on security without adding friction.

The foundations of effective communication.

While we tend to think of ourselves as at odds with artificial intelligence, the same principles of effective communication apply to human-machine conversation.

The 3 principles of effective communication:

For AI to communicate effectively, it needs to display the ability to:

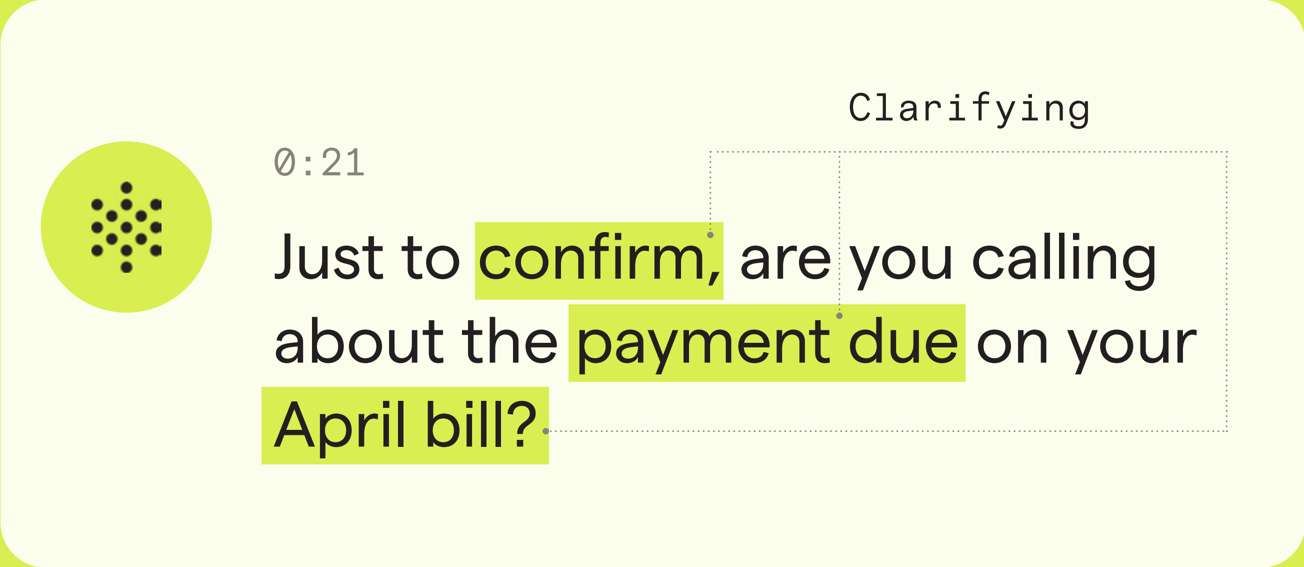
- **Listen** (process inputs),
- **Reason** (interpret meaning)
- **Speak** (respond clearly and empathetically)

This guide breaks down the key elements of listening, understanding, and speaking and how they help foster reliable and engaging customer experiences.

All agents need to not only display the three principles but also **clarify** user requests and **empathize** with users through engaging and helpful conversation.

These qualities are crucial in customer service. An AI agent that greets customers warmly, asks the right follow-up questions, shows understanding, and responds clearly is far more likely to engage customers, efficiently resolve issues, and deliver benefits across an organization.

By studying the way humans communicate and applying it to human-machine interactions, we can design AI agents that are humanlike, able to communicate clearly, and approachable. This will help soften inherent skepticism around AI and build trust in automated systems.



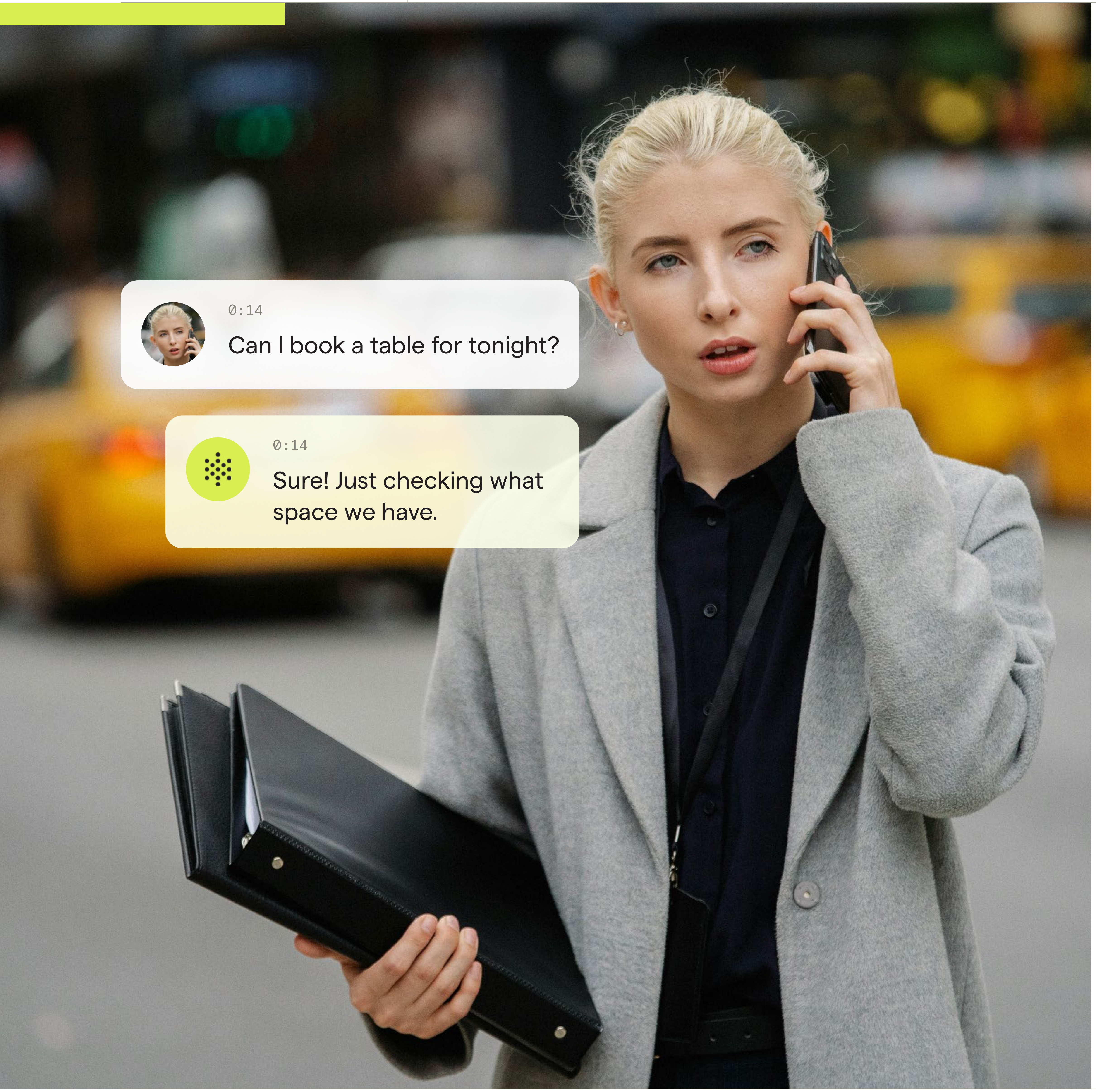
Just to confirm, are you calling about the payment due on your April bill?

Empathizing
0:14
I'm sorry for your loss. Our team is here to support you.

Listen

/'lisn/ [verb]

(1) To pay attention to sound, listen to music.
(2) To hear something with thoughtful attention: give consideration.



0:14

Can I book a table for tonight?

0:14

Sure! Just checking what space we have.

Listen

During any conversation, the ability to listen is impacted by multiple external factors. Is the environment loud? Does the person appear to be listening to you? Are there interruptions?

These factors similarly make automating human-machine conversations more challenging. For example:

- Phone lines cut out
- Outside talking, television, and other background noises make it difficult to hear
- Every person has a different way of speaking, such as having an accent or using slang

These challenges can lead to speech recognition errors that make it difficult for AI to capture spoken language accurately.

Challenges in listening: ASR and accuracy.

AI agents rely on automatic speech recognition (ASR) systems to transcribe spoken language into text that can be digested by large language models (LLMs). While out-of-the-box ASR solutions may provide some benefits, they are typically not tailored to meet a business's specific needs and objectives.

Many ASR providers offer satisfactory out-of-the-box performance, but these solutions can have limitations because they are built for general use cases. For instance, models trained for dictation or voicemail transcription may struggle with the challenges of real-time contact center conversations, such as overlapping speech, background noise, and industry-specific terminology.

Another issue with out-of-the-box models is that they are often dialect-specific. Enterprises should not assume that all calls in the US use American English, and if they operate globally (as they often do), they certainly cannot rely on English always being the primary language spoken.

Even the best-performing model needs additional support to match the accuracy of human hearing. This is where spoken language understanding (SLU) comes in.

Retail

Pick up in store

as

Pick up and store

Travel

Seat upgrade

as

See the grade

Logistics

Track my shipment

as

Trap my shipment

Automotive

Oil change and tire rotation

as

Oil change entire rotation

Healthcare

Prescribe 10 mg

as

Describe 10 mg

Spoken language understanding: The foundation of an effective AI agent.

When you mishear what someone has said to you, you either ask the person to repeat themselves or you work out what they've said based on the context of the conversation.

If an AI agent keeps asking customers to repeat their queries, the experience becomes tedious, and trust in the system plummets.

Even with advanced ASR systems, errors can still occur. This is why Spoken Language Understanding (SLU) is a critical layer of AI communication.

SLU is a term applied to techniques that can be used to fix erroneous ASR transcriptions. For example, if a customer says, "A table for eight, please," but the ASR mishears it as "a table for hate," the SLU model can use context, like recognizing that "eight" is a common request for larger group reservations.

The following page shows some examples of key SLU techniques for AI agents for customer service.

Identification & Verification (IDNV)

It is ~~an indus~~
Hernández

Contextual Recognition

Actually, can you make that ~~hate~~ people?
eight

Alphanumerical Parsing

~~Apple Charlie Echo double eight 0 for 7~~
ACE88047

Ensemble of Recognizers and N-best List

~~Unfortunately, I have lost my car.~~
Unfortunately, I have lost my card.
~~Unfortunately bathrooms my car.~~

Entity extraction

This is the process of identifying key pieces of information (entities) from what a user says and using those pieces to understand and fulfill the request.

Voice activity detection (VAD)

VAD determines when someone is speaking and when there is silence. It helps the AI agent know when to start listening and when to stop. This is important for detecting when a person has finished talking during a conversation and mitigating the chance of interruptions.

Lexicon customization

This involves tailoring the AI's vocabulary to specific words or terms relevant to a use case, like brand names or industry-specific jargon.

For example, if AI is used in healthcare, you might add terms like "telemedicine" or "cardiologist" to its lexicon to improve accuracy.

Model ensemble

This is when a group of AI models work together to achieve better performance. Each model specializes in a specific task, and its different outputs are combined to produce more reliable and accurate results. Think of it as having multiple experts collaborating to solve a problem.

Contextual ASR biasing

This involves providing context on what type of input an ASR model should 'listen out' for (e.g., ZIP code or 8-digit alphanumeric string).

Phonetic fuzzy matching

AI matches words that sound similar but might be mispronounced or transcribed incorrectly. For example, if someone says, "I want to transfer funds to my savins account," the system uses phonetic fuzzy matching to recognize that "savins" is likely "savings," despite the slight mispronunciation.

Database verification

This verifies ASR transcripts against relevant databases (e.g., all US ZIP codes) or existing CRM records and uses this information to pick out the most relevant transcript.

Latency trade-off

This establishes how long an ASR model should listen for specific inputs and whether or not the customer can interrupt.

Timing is everything: Balancing latency and interruptions.

Interruptions are a natural part of conversation. Some are useful, like when somebody offers you a drink, they might list 'tea, coffee, water, juice, beer…' If you don't want a drink or you want one that's already been listed, it's easier for both parties if you just interrupt.

Other interruptions are less useful and make conversations harder than they need to be, like when somebody interrupts you to ask a question when you are just about to answer.

One of the most crucial elements in building an AI agent is smooth, timely interactions. Delays between user input and system response can frustrate users if the assistant takes too long to respond. But interruptions to fast or incomplete responses present a problem to harmonious human-machine communication.

There needs to be a balance where the system responds quickly and accurately without cutting off or confusing the user. Unfortunately, there's no hard and fast rule on when you should allow customers to interrupt. It depends completely on what you deem important to your business and what your customer deems important.

?

Should you allow customers to interrupt?

Here are some useful questions to ask when considering if and when customers should be able to interrupt an AI agent:

- How important is it that the customer listens to everything? Do you want to ensure the customer does not interrupt the AI reading out terms and conditions?
- What is the potential impact of a customer interrupting? If you're reading out a list, and they skip later options, is this likely to result in an error further along the conversation?
- What is the impact on CX? Will allowing or preventing the customer from interrupting have a negative or positive impact on their experience?
- How do you want to balance CX and business processes? If customers are likely to skip the boring stuff, how do you strike a balance between your priorities and theirs?

Reason

/'rizn/ [verb]

(1) To think in a logical way.
(2) To to find an explanation or solution to (something, such as a problem, question, etc.) by thinking about the possibilities.

0:14

Can I book a table for me and my girlfriend next week?

0:14

Sure, I can help you find a table for 2. What day are you coming in?

Reason

Once a speaker's words have been transcribed, an AI agent needs to understand the context behind what the caller is saying to formulate the right response and take action in a way that moves the conversation towards resolution.

The process of deciding how to respond and what actions to take is usually handled by Large Language Models.

LLMs are excellent at holding natural conversations. If you've tried ChatGPT, Gemini, or Claude, you've probably been impressed with just how conversational these models can be.

But it's no secret that LLMs are liable to hallucinate. That doesn't mean they're totally unreliable. They just need a robust set of guardrails to ensure they say and do what they're supposed to.

There are two key types of hallucinations to consider:

- The AI agent **says** the wrong thing: Sometimes generative AI models 'make things up'. Remember, these models construct responses based on what they see in their training data, and sometimes, they come out with things that don't make sense.
- The AI agent **does** the wrong thing: LLMs are good at conversations but not necessarily good at taking action. For example, the LLM might determine that the correct way to move a conversation forward is to say, "Okay, I've updated your address," but that doesn't necessarily mean it has made the API call to actually update the customer record.

Ensuring LLMs say the right thing with Retrieval-Augmented Generation (RAG)

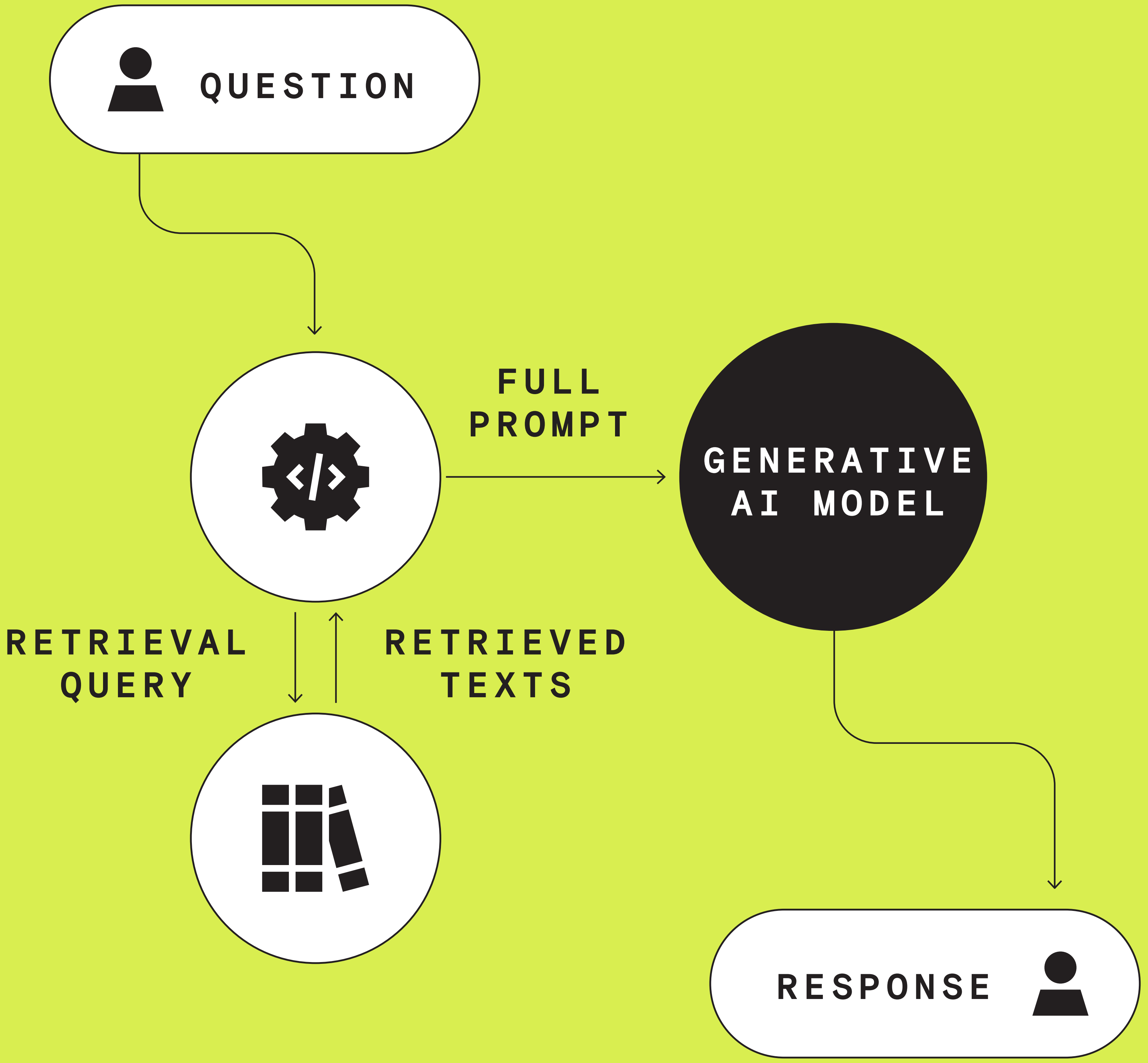
Putting customer interactions in the hands of automated systems requires a lot of trust for both your business and customers.

If your agents were unsure of how to resolve a customer's issue, you'd want them to check their response so they deliver a trustworthy and correct answer.

Retrieval-augmented generation, or RAG, is a technique that enables AI agents to cross-reference knowledge from a generative model with a knowledge base.

RAG helps organizations balance the potential of generative AI and the need for controlled responses.

This technique ensures that an AI agent checks its generated responses against a knowledge base. It acts as a safeguard, preventing inaccurate, irrelevant, and inappropriate responses, and keeps customer conversations within established limits.

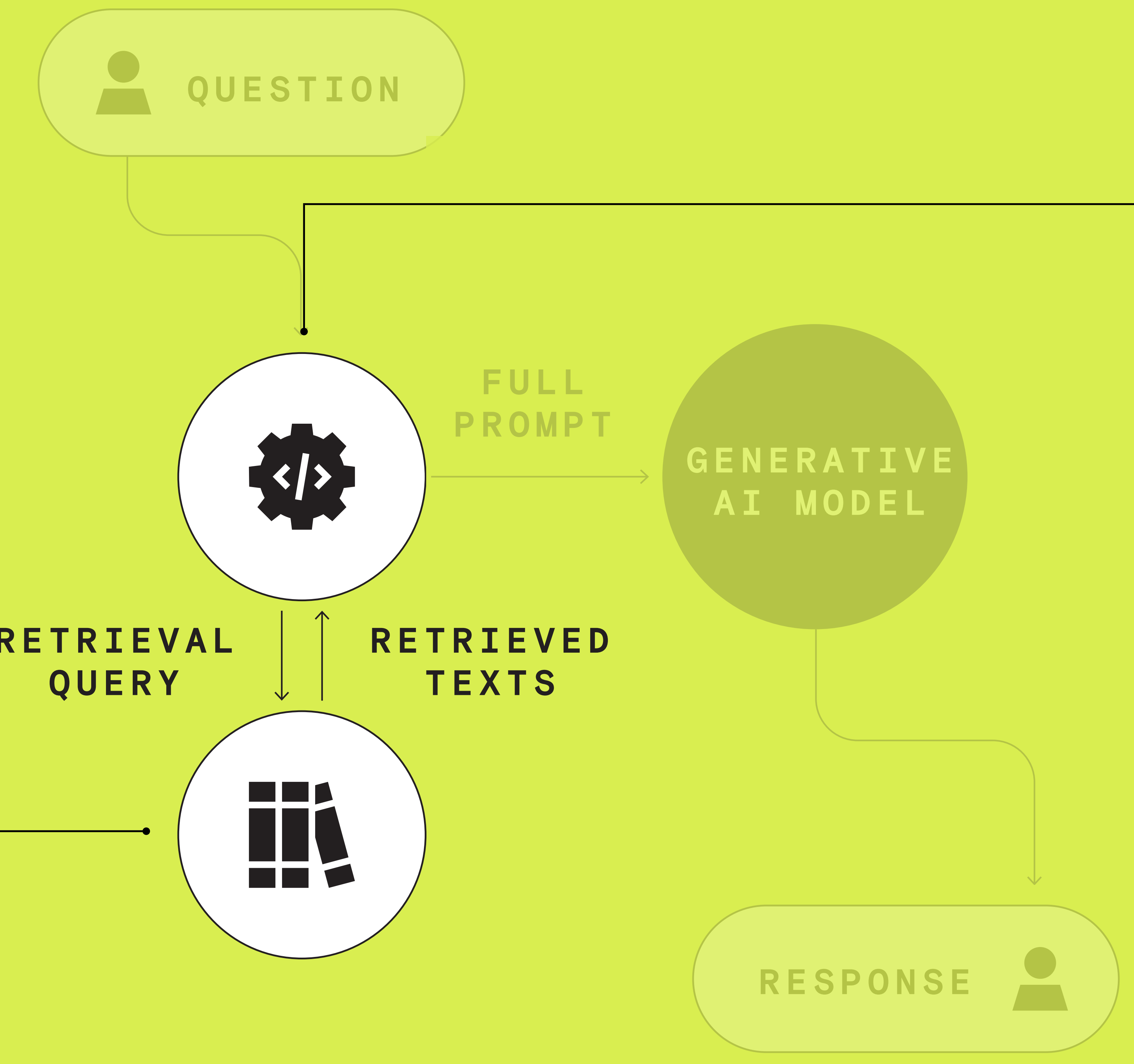


There are two key elements of RAG that must be optimized to prevent hallucinations and prompt injection attacks: knowledge base and retriever.

1. Knowledge base

Your AI agent is only as good as the information it has access to, so developing and maintaining a detailed knowledge base from which your agent can draw is crucial.

Your knowledge base should include everything you want the agent to be able to discuss, but it also needs to include undesirable information and specific behaviors to apply in certain situations.



2. Retriever

The retriever is the "search engine" that enables the agent to cross-reference facts against the knowledge base. The retriever must be accurate enough to cross-reference the knowledge base with little to no margin of error.

LLMs typically operate in a black box, meaning it is extremely difficult, if not impossible, to understand where exactly the model is pulling certain pieces of knowledge from. Without being able to isolate the cause of a hallucination, it is very difficult to remedy.

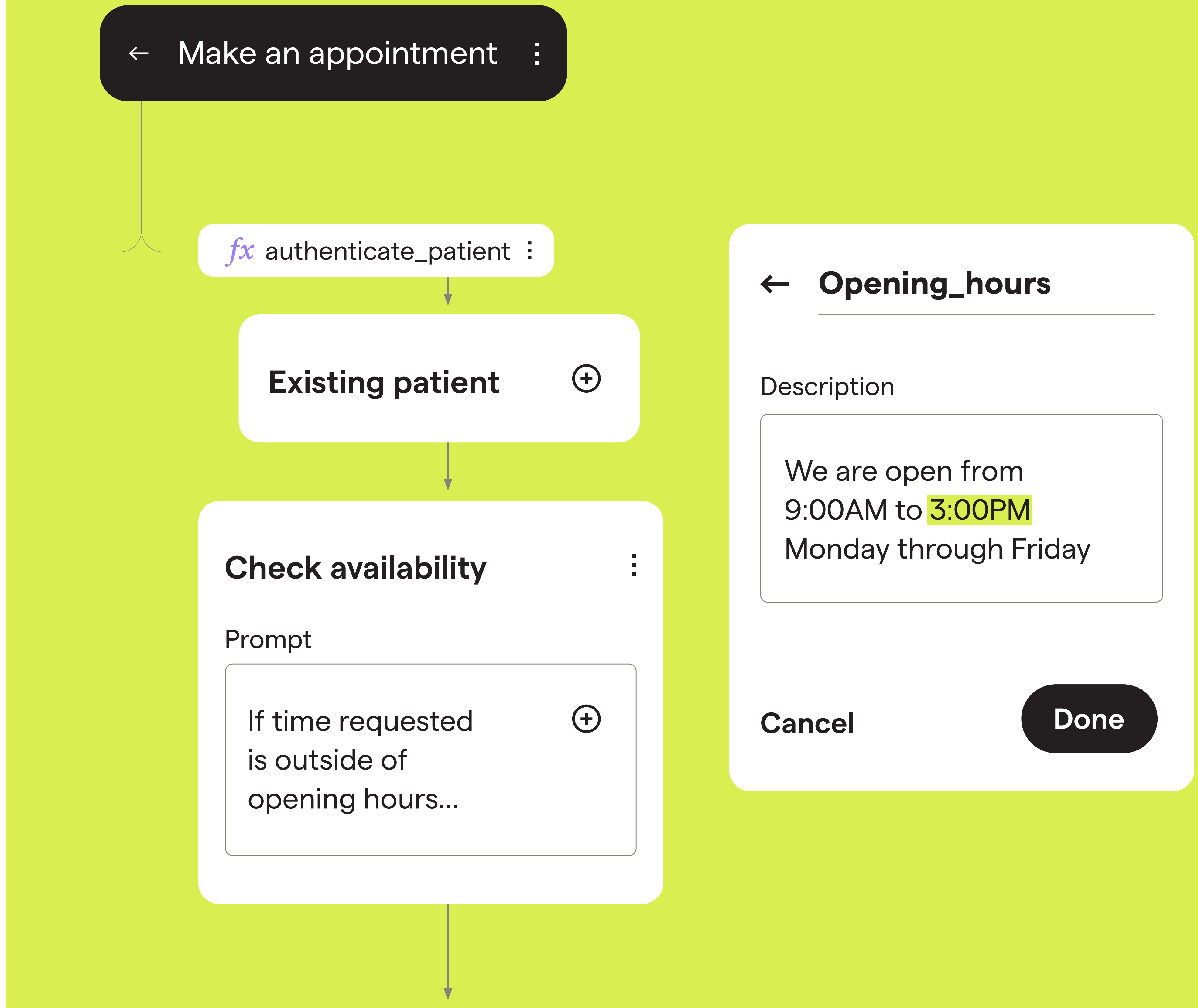
However, clever retriever design makes it possible to trace references to specific points in the knowledge base, enabling designers to make simple text-based edits to prevent hallucinations and create a cleaner, more transparent system for all.

Ensuring LLMs take the right actions

Giving undesirable or inaccurate responses is a well-known issue with LLMs, but one that is relatively easy to overcome with guardrails like RAG.

What's less well-documented is how to ensure that LLMs take action. For example, a caller may ask to update the address on their account. For this request, the LLM is trained to know the correct flow of conversation. It asks for the new address and then says the change has been made. But has it actually updated the CRM, booking system, or other relevant software?

This is where most AI agent proofs-of-concept fall apart. The conversation flows smoothly, but API calls are not made consistently or reliably, leading users to think that actions have been taken when they haven't. This incomplete job can create issues for other functions of your contact center and, worse, for other departments.



If time requested is outside of opening hours...

We are open from 9:00AM to 3:00PM Monday through Friday

1 Documenting actions in your knowledge base

Many AI agents struggle to reliably take action because they have access to too many tools or APIs and little to no direction on what tool to use and when.

Breaking your knowledge base down allows you to ascribe specific actions to various topics. For example, you can build a specific part of your knowledge base that relates to updating account information. Within that section of the knowledge base, you can reference exactly which tools and APIs you want the AI agent to call when discussing this topic.

2 Building flows for key transactions

LLMs are smart enough that you don't need to design specific flows to enable them to answer FAQs. But transactions like taking bookings or sending payments need to follow a specific set of steps.

While you can write simple prompts that enable LLMs to hold a fairly normal conversation, it's safer and more reliable to design flows that show LLMs how to move through a conversation, including what actions to take at every step.

3 Building in checkpoints for specific tools

As a final safeguard, it's a good idea to build checkpoints for certain types of transactions that require specific tool use or API calls. These checkpoints remind the AI agent to ensure that certain actions have been taken based on call type, specific utterances, or certain types of transactions.

Speak

/spiːk/ [verb]

(1) To say words, to use the voice, to have a conversation with someone.

Speak

Once an AI agent has understood a caller and decided what action to take, the next step is to formulate a clear, natural response.

How your AI agent sounds is a crucial factor in how likely customers are to engage. We've all had bad experiences talking to automated systems that don't understand us. We've been primed by consumer voice assistants and spoken language IVRs to expect the worst, so when customers are greeted by a stilted, robotic voice, it's no wonder they insist on speaking to a human.

A truly engaging AI agent should speak and sound just like a real person. It's not about tricking customers into thinking they're talking to a person but enabling them to forget they're talking to a robot so they can focus on the task at hand. The goal is to establish enough trust that the customer feels comfortable resolving their issue with an AI agent.

When deciding what your AI agent should sound like, keep these two considerations in mind:

Text-to-speech (TTS) technology

TTS models transform text transcriptions into spoken utterances. Traditional TTS has delivered robotic-sounding voices, but a new generation of TTS offers a more natural experience.

Engaging conversation design

The right words, tone, and pacing turn robotic replies into natural dialogue.

0:14
Hi Suzie, are you calling about your reservation this weekend?
0:14
Yeah! Can I get a late checkout?

Text-to-speech: Synthesizing natural-sounding voices.

Although TTS has improved significantly in recent years, even the best TTS tools can sound stilted or unnatural.

Techniques must be applied to configure a particular voice to sound just right. Whether that's in a tone that matches your brand, the region your customers are calling from, or the clarity needed for a good customer experience, TTS needs more than just a good model. It requires the right tuning and context.

Audio caching

Audio caching enables AI agents to store and reuse common speech responses instead of generating them in real-time. This helps reduce latency and improves the overall quality and consistency of the agent's voice.

By caching common TTS elements—such as greetings, transfer messages, and acknowledgments—AI agents can deliver responses faster while maintaining a uniform brand voice. This approach enhances customer interactions by ensuring clear, natural-sounding audio without delays.

Voice regeneration

Voice regeneration is the process of dynamically modifying an AI agent's voice to suit different contexts or user preferences. This could involve adjusting pitch, tone, accent, or even perceived age to better match the situation. Through real-time adaptation, voice regeneration enhances personalization and makes interactions feel natural and engaging.

Localization

AI agents need to naturally accommodate different languages, accents, and regional dialects. Localization goes beyond simple translation—it involves adjusting pronunciation, expressions, and phrasing to match regional preferences. By fine-tuning phonemes and adapting speech patterns, an AI agent can communicate more seamlessly with users.

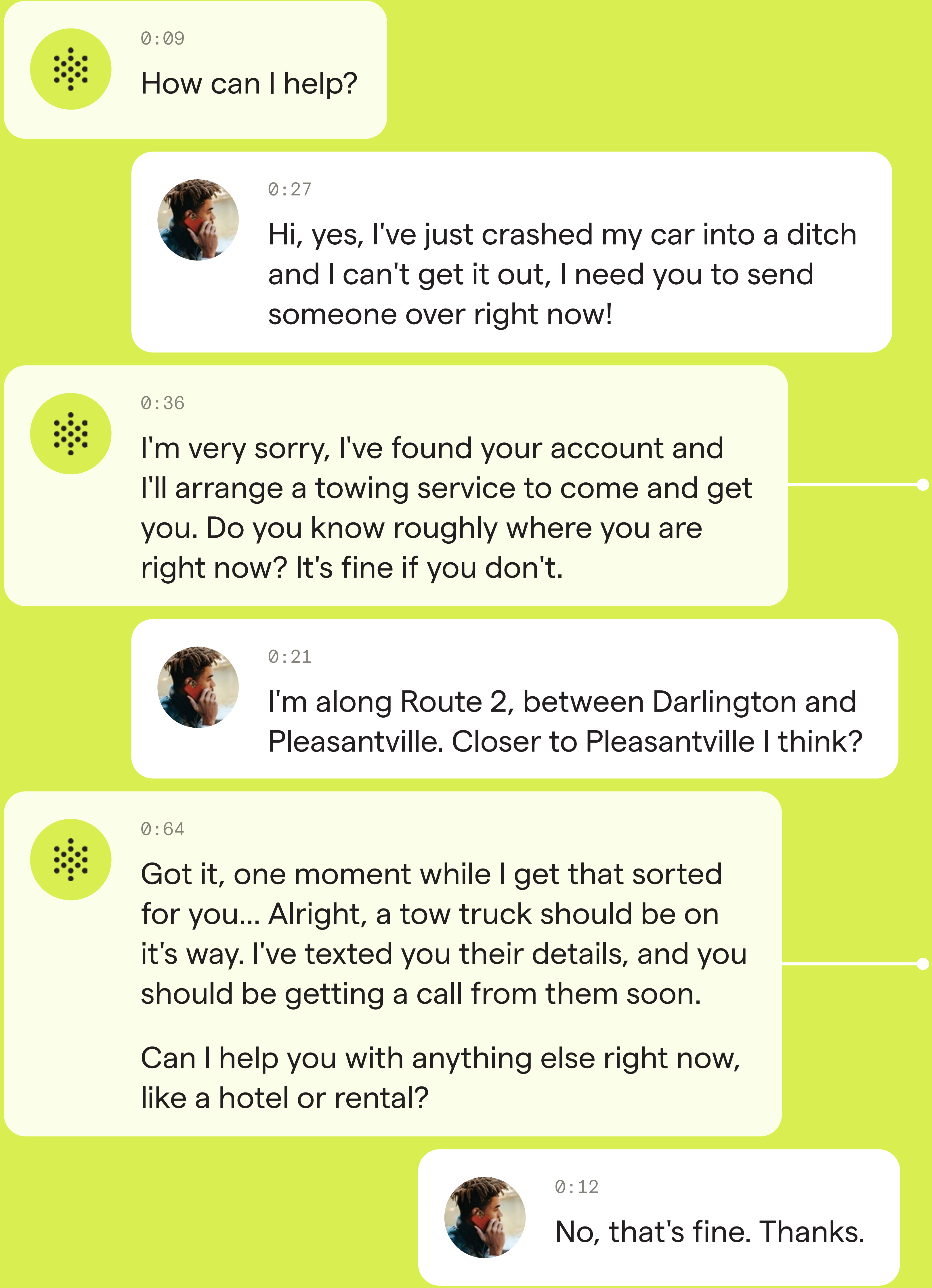
Persona development

An AI agent's persona should align with the brand identity and user expectations. Developing a strong persona involves defining the voice, tone, and personality to create a natural and engaging character. By carefully selecting and training a voice—whether human-recorded or synthetic—AI agents can deliver interactions that feel more relatable, reinforcing trust and enhancing the user experience.

Designing engaging conversations.

Designing engaging conversations with an AI agent isn't just about letting customers steer the conversation, it's also about anticipating their expectations and gaining their trust through qualities such as transparency, empathy, and clarity.

While there are a number of factors to consider when creating an engaging automated voice experience, the four factors detailed on the following page are non-negotiable:



0:09
How can I help?

0:27
Hi, yes, I've just crashed my car into a ditch and I can't get it out, I need you to send someone over right now!

0:36
I'm very sorry, I've found your account and I'll arrange a towing service to come and get you. Do you know roughly where you are right now? It's fine if you don't.

0:21
I'm along Route 2, between Darlington and Pleasantville. Closer to Pleasantville I think?

0:64
Got it, one moment while I get that sorted for you... Alright, a tow truck should be on it's way. I've texted you their details, and you should be getting a call from them soon.

Can I help you with anything else right now, like a hotel or rental?

0:12
No, that's fine. Thanks.

The agent switches to a softer tone to show concern given the context, and proactively works to address the goals of the user. The agent also lets the user know that there's a fallback option in case the information is missing.

The agent makes the user feel heard, even while processing requests, and provides relevant follow-up information and suggestions.

First impressions matter

You can build a beautiful, complex system that leverages cutting-edge technology, but if the AI agent's overlying voice creates frustration, a customer will likely ask to speak to someone or just hang up.

Small things matter here, from pacing and voice intonation to the way audio is edited together or synthesized. Each customer request requires an appropriate tone of voice, which means an AI agent shouldn't sound as happy about potential fraud as it does about opening a new account.

Use clear, natural prompts to reduce cognitive load

Clear, easy-to-understand prompts help users feel in control, but this requires good voice quality and natural intonation.

It may seem like short, bite-sized instructions are best, but real-world deployments show that clarity outweighs brevity. For example, instead of simply asking, "Have you personally requested and received a quote within the last three months?" an AI agent can provide additional context to reduce confusion. For example, "The first step is to get a quote. Have you personally requested and received one within the last three months?"

Small changes make a huge difference.

Ask open questions to give customers control

The best way to encourage users to engage with a voice assistant is to start with an open-ended question: "How can I help?" This approach puts customers in control of the conversation instead of forcing them through a rigid, preset path, like the antiquated "press 1 for this, press 2 for that" IVRs. By letting users define their needs and respond accordingly, the AI agent builds trust and creates a more natural interaction.

Asking open-ended questions throughout the conversation allows the AI agent to curtail their response around what the user specifically said, keeping interactions dynamic. As a result, users remain engaged and open to the agent's solutions.

Empathy is essential

Expressing emotions is one of the most complex parts of communication. Word choice, tone, volume, and even silence all contribute to how meaning is conveyed.

To ensure AI agents act emotionally appropriate, designers need to anticipate how users might react to certain information—whether that be frustration, anger, or doubt—and build responses that help ease tension and keep the conversation on track.

By prioritizing user expectations, AI agents can be designed to stay within the boundaries of natural human-agent interactions while delivering a meaningful and satisfying experience.

Conclusion

Building an effective AI agent isn't just about processing speech, it's about mirroring the human experience within the conversation. This requires more than just converting speech to text and generating responses. It involves the careful inclusion of listening, understanding, and speaking to make conversations feel seamless, intuitive, and, most importantly, useful.

AI agents that can accurately interpret spoken language, clarify intent, and respond with empathy are far more likely to build trust and improve customer experience. By integrating techniques like contextual adaptation, dialogue management, and voice customization, businesses can ensure their AI agents are not only functional but also trustworthy.

As AI continues to evolve, the best systems will be those that respect the complexity of human communication, capturing its nuance, adjusting to context, and enhancing interactions rather than disrupting them. By applying these principles, AI agents can do more than just automate customer service—they can transform it.

PolyAI

Start automating humanlike customer interactions with robust brand safety.

PolyAI’s AI agents are consistent, reliable, and safe. Our proprietary generative AI framework incorporates the benefits of generative AI while retaining the safety guardrails that are so important to enterprises looking to use AI responsibly.

Sign up for our monthly demo to find out more about how PolyAI can help you answer every call immediately, improve loyalty, resolve over 50% of calls, and deliver effortless CX at scale.

Request a demo