fineSample

Running an always-on harness for conversational commerce

A request can start in one channel and finish in another. The system tracks the request, takes actions the business has approved, and checks later updates to confirm the intended result.

The model can help the conversation without owning the commitment.

A case study from fineSample, November 2026

Ten calls and 51 turns, 18 to 24 November 2026. Three calls used a fictional moving business's terms.

A request can start in one channel and finish in another.

“Alright. Thank you,” the caller said. The language model read that as agreement. The agent asked the caller to confirm on the page it had texted. One minute and 42 seconds after the live phone call ended, the caller accepted the exact offer on that page. The harness recorded the agreement against the offer it had issued.

That matters most for services, which rarely arrive ready to buy. “My gutters overflow when it rains.” “The furnace is running, but the house is cold.” “Our books need catching up before year end.” “We need this contract reviewed by Friday.” Each is real business, from a local trade or a professional firm, and none says yet what should be bought. Someone has to establish the scope, check the business can take it on, explain the offer, set the price and agree the next step. In most firms that someone is the owner or a senior person. For a service business, the harness works like a shopping cart for that work: it carries the purchase through the conversation, from a description of the need to an offer the customer has accepted, without the model owning a price or a promise.

The review covered ten calls and 51 turns from 18 to 24 November 2026. Three calls used a fictional moving business’s terms. Seven used a sales-line configuration. Every call ran on live phone and messaging services.

Turn a customer's request into agreed work

The caller described an upstairs apartment and asked for three movers. They initially estimated three to four hours. The pricing terms required a single duration, so the agent asked again and the caller said two hours. The harness used those details and the business's published terms to calculate the estimate. It sent the caller a page with the scope and price together.

#	The caller said	What the model proposed	What the harness did
1	"I need to move somewhere in Burnaby."	service: move; city: Burnaby	Asked how many hours of work to expect
2	"It's an upstairs apartment, and I need 3 movers."	stairs: yes; movers: 3	Confirmed both and asked again for the hours
3	"I think 3 to 4 hours."	hours: "3 to 4"	Refused it, because the terms need one number, and asked again
4	"2."	hours: 2	Priced the move from the terms, $259.88 including tax, and texted a link to the offer
5	"Alright. Thank you."	yes to the offer	Asked the caller to confirm on the page it had texted
6	"Alright."	nothing new	Thanked the caller

Each reply reached the speech service within 1.2 to 1.6 seconds. The model proposed facts and never a price; every amount the caller heard came from the business's terms.

The harness kept the issued offer with the request. When the caller confirmed on the web, the harness checked that the offer was still valid and matched the terms the caller had seen. The test verified acceptance. Booking, payment and fulfillment were outside the recorded result.

The same design is built to carry a schedule change until the calendar confirms the new time, or to keep a supplier coordination task open until the selected supplier confirms, after a person approves the selection. Each task defines in advance what evidence counts as done.

Measured against the steps of a conventional checkout, the moving call covered the first four.

Checkout step	For a service	In the moving call
Choose the product and options	Establish the service, scope and conditions	Turns 1 to 4: service, city, stairs, movers and hours
Calculate the total	Apply the business's prices to the confirmed facts	Priced by the harness from the terms: $259.88
Review the order	Present the scope, price, terms and next step together	Texted as one page
Check out	Record acceptance of that specific offer	Accepted on the page 1 minute 42 seconds after the call
Pay	Collect a deposit or an assessment fee, confirmed by the payment provider	Not tested
Arrange delivery	Confirm the appointment or start date	Not tested

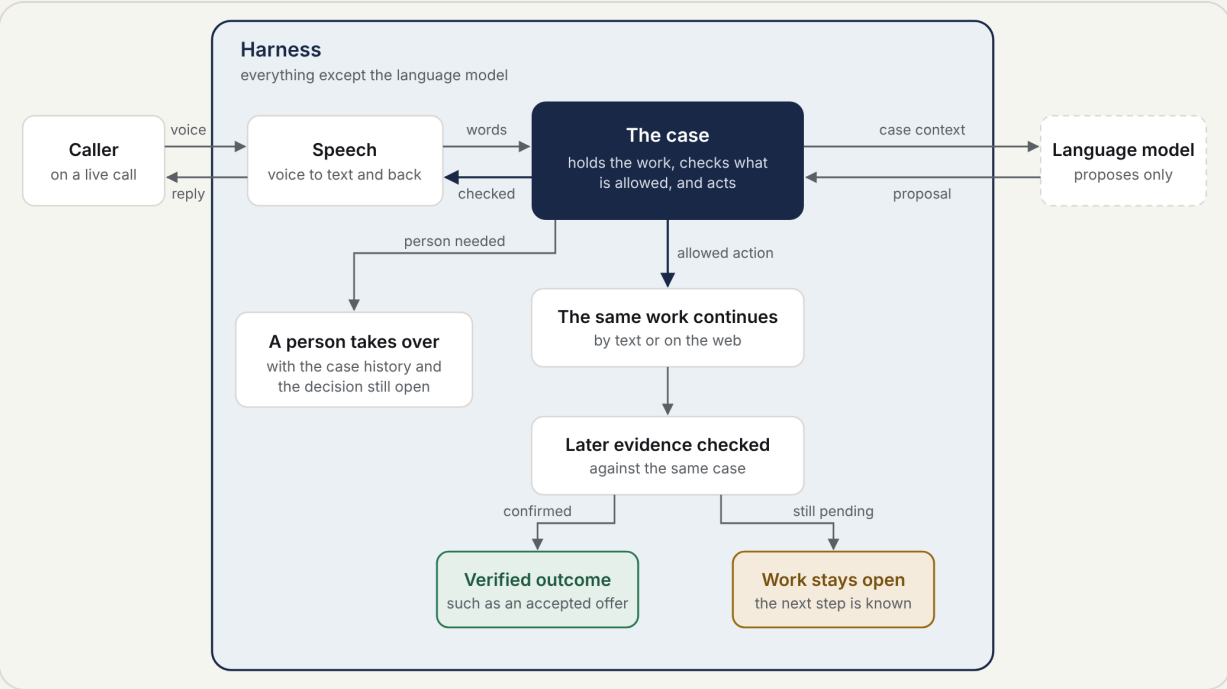
Some services can't be priced in one conversation. A repair may need an inspection, and a professional engagement may need an initial review of the documents. Then the thing being bought is the assessment, and the harness has to keep that distinction: booking an assessment doesn't mean the work itself has been priced or sold.

What the harness owns

The harness is everything except the language model. Inside it, a speech service turns the caller's voice into text and speaks the replies. The language model sits outside, interprets the transcript in context and proposes details or wording. A misheard number can still look valid to a calculator. In the moving call, the clarification itself was the problem. It sounded like a form error, and whether two hours was a better estimate remained uncertain.

The harness owns the case across turns, channels and later events. When the agent offered a caller demo times by text, the harness first checked that the case was still open and that it had permission to text that number. It then sent the offer, recorded the attempt whether or not the text went out, and kept the case open until a booking confirmed it or the work needed to go to a person. The same care applies when a customer moves between channels: their records are joined only once their identity is established, never on a matching name alone.

The model never owned a commitment. It never saw a price, a rate or the caller's phone number. Prices came from the business's terms, calculated by the harness. When a caller on the sales line asked what the service cost, the answer, C$500 a month, was a statement the business had approved in advance, spoken by the harness word for word. The model didn't write it.



The harness is everything except the language model. It hears the caller, holds the case, checks what is allowed before acting, carries the work on by text or on the web, and matches later evidence to the same case until the result is confirmed or a person takes over. The model sits outside and only proposes.

Know where work stopped

The sales-line calls showed why an attempted action and its result need separate records. The carrier accepted some texts offering demo times. The records showed no delivery report for those texts and no confirmed demo booking from the calls. Those are different states with different next steps. The same rule holds for any commitment: a customer saying “I paid” is useful information, but only the payment provider’s matching receipt establishes payment. Four other texts were blocked because there was no recorded permission to text the callers. Two of those contained replies a person had written.

The harness also withheld all five optional sales lines the language model proposed. One relied on a fact the model had never been given. In each case the agent continued with the checked reply. A failed proposal could be dropped without losing the context or interrupting the case.

These distinctions matter in a staffed office. Staff can handle live conversations and decisions that require judgment while the agent carries agreed follow-up. When a case needs a person, the useful handoff includes the latest terms, the action already attempted and the decision still open. The blocked replies show why a person’s intervention also needs a recorded result.

What stays in place when the language model changes

A compatible model can supply interpretations through the same interface. The business's terms, the open cases and the evidence of commitments stay in the harness. Two models from the same provider appear in the call records under different configurations. Model calls across the sample cost about US$0.004 per call, so a change is rarely about price. A replacement needs to be checked on the same scenarios before use with customers.

That test can ask whether the candidate picks up the right details, stays within time and usage limits, reaches the expected verified result and hands off cleanly when needed. The records show which action was proposed, whether it was allowed and what happened after it ran. A reference transcript can help test speech recognition separately. This lets the team evaluate a model change against the task the agent is responsible for carrying through, the approach we took in Before you switch AI models, look past the score.

Related work

The term conversational commerce goes back to Chris Messina's 2015 essay, which helped start the 2016 wave of shopping chatbots. He later wrote that he had been wrong about it. The bots of that wave could hold a conversation, but nothing carried the purchase through.

The newer work on agentic commerce comes from the buyer's side. OpenAI's Agentic Commerce Protocol, implemented with Stripe, and Google's Universal Commerce Protocol let an AI assistant check out from a merchant's product catalog on a shopper's behalf. As Checkout.com's comparison describes them, neither deals with services, quotes or a person's approval.

"Harness" is a familiar word in AI engineering, where it usually means the code that keeps a model on track through long work, as in Anthropic's notes on harnesses for long-running agents. AI receptionists for service businesses, such as those in ServiceTitan and Jobber, answer calls and book appointments.

This study looks at the seller's side of service commerce, for local trades and professional firms alike: a business making an offer through a conversation, with the price set by its own terms, acceptance tied to that exact offer, and every commitment checked against evidence.

Timing in this sample ran from receipt of each transcript to handoff of reply text. The median was 1.9 seconds and 90 percent of recorded replies reached handoff within 3.0 seconds. Where first-word timing was recorded, the median was 0.7 seconds to handoff for speech. Speech recognition, synthesis and playback sit outside that interval. The underlying call-turn, messaging and acceptance records cover 18 to 24 November 2026 (UTC), including fixed replies and model-assisted turns.

Talk to us about it: hello@finesample.com