Contact Center Pipeline

AI QUALITY SCORING

BY OWEN ZHANG, COMMSADVISOR.COM

# WHAT "100% QA COVERAGE" *DOESN'T* TELL YOU

**How you can adjust your scorecarding accordingly.**

NOVEMBER 2026

BY OWEN ZHANG, COMMSADVISOR.COM

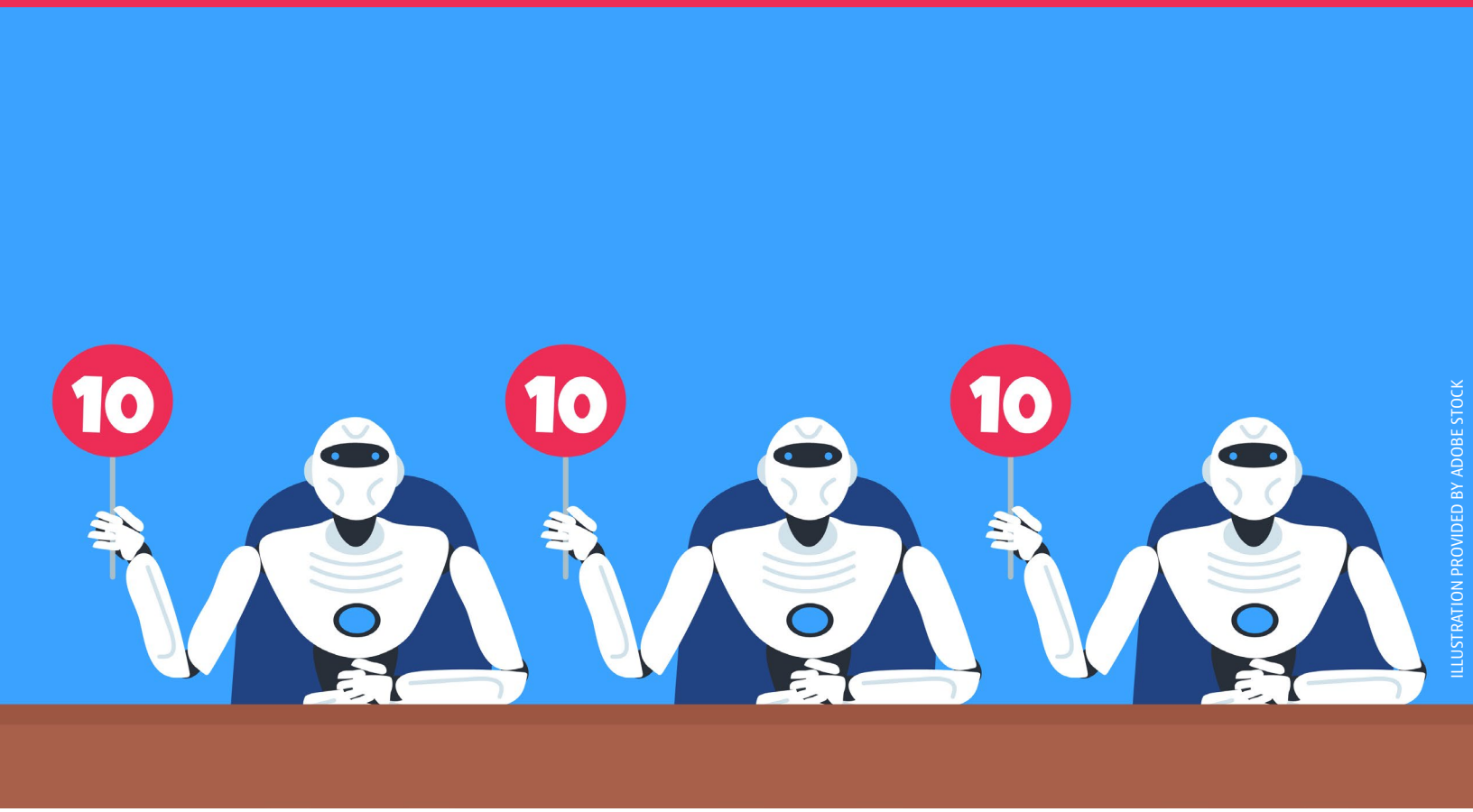


ILLUSTRATION PROVIDED BY ADOBE STOCK

# WHAT "100% QA COVERAGE" *DOESN'T* TELL YOU

## HOW YOU CAN ADJUST YOUR SCORECARDING ACCORDINGLY.

Ask any AI QA vendor what their product does and from what I've found you may hear some version of the same sentence:

- Zendesk QA says its AutoQA feature "evaluates 100% of interactions across digital and voice channels."
- Observe.AI promises to "accurately automate 100% of QA evaluations."
- CallMiner contrasts its approach directly against "traditional sampling-based scorecards," pitching full coverage as the whole point of switching.
- EvaluAgent's pitch is nearly identical: "100% of conversations scored automatically."

None of that is false. These tools touch every interaction instead of the 1%-2% sample a human team can realistically review by hand. That's a significant capability, and it's the reason so many QA teams have adopted them in the first place.

But coverage and accuracy are two different claims, and vendors are careful to only make one of them with confidence.

> ***...COVERAGE AND ACCURACY ARE TWO DIFFERENT CLAIMS, AND VENDORS ARE CAREFUL TO ONLY MAKE ONE OF THEM WITH CONFIDENCE.***

## THE PATTERN ISN'T WHAT YOU'D EXPECT

The intuitive version of this story is a clean split. AI is reliable on objective, rule-based criteria (Did the agent verify identity, include a required disclosure, follow the script?) and weaker on judgment calls that require reading tone or context (Was this response empathetic? Was a workaround the right call here?).

It's a tidy narrative and it's mostly what shows up when you search for commentary on AI QA limitations on Reddit.

It's also not quite right. A thread on r/callcentres from a QA lead whose team moved to 100% AI scoring is a useful reality check because the failures described there aren't cleanly sorted into "objective, handled fine" and "subjective, handled badly." Instead, they're scattered across both.

Some are the kind of failure you'd expect from a system reading tone and intent:

- One commenter described the AI penalizing an agent for "not properly ending the call" when the customer had hung up first, leaving the agent no opportunity to do so.
- Another described the tool dinging an agent for not collecting contact information from a caller who had dialed the wrong number and said so directly: a case where collecting that information would have been pointless, not required.

But others are failures on exactly the literal, checklist-style criteria AI is supposed to be strongest at. The same thread includes a description of the system missing a HIPAA violation entirely because it was checking for surface-level "call flow" rather than the substance of what was said.

> ***THE PROBLEM SHOWS UP WHEN A CRITERION THAT LOOKS OBJECTIVE ON THE SCORECARD TURNS OUT NOT TO BE...***

Another commenter described the tool failing to credit an agent for providing a website URL, because of what appears to be a punctuation-parsing quirk in how the transcript was read. A third described the system consistently mis-transcribing the company's own name.

***These are mechanical failures, not subjective judgment calls.***

That's not an argument against the objective side working at all. A QA consultant answering a staffing question on r/CustomerService put the vendor pitch about as plainly as anyone.

"Tools like these can transcribe 100% of calls and auto-score the objective stuff: greetings, disclaimers, compliance language, so your human QA people only jump in on the low-scoring ones where soft skills and empathy actually matter."

That's an independently stated version of the same split the vendors are selling. The problem shows up when a criterion that looks objective on the scorecard turns out not to be once you're looking at the actual call instead of the checklist.

The gap comes down to whether scoring the interaction correctly requires understanding *why* something happened versus matching a literal pattern, not whether a criterion is labeled objective or subjective.

AI is consistent when a criterion can be resolved by checking for the presence or absence of a specific word, phrase, or step. But it gets shaky the moment correct scoring depends on context:

- A compliance step that didn't apply because the customer volunteered the information first.
- A hold-time disclosure that wasn't needed because there was no hold.
- An empathetic response that doesn't use any of the phrases the model was trained to recognize as empathetic.

## WHAT THE VENDORS QUIETLY BUILD AROUND THIS

You don't have to take a Reddit thread's word for it. Look at what the vendors ship alongside their 100%-coverage marketing, and in my view, you'll find features that only make sense if the vendors don't fully trust every score.

- Zendesk QA's autoscoring documentation lists specific categories the system evaluates, including empathy and tone. But it also specifies that when the system can't confidently assess a category, the interaction gets scored "N/A" rather than a forced number. That's a built-in acknowledgment that confidence isn't uniform across criteria.
- Observe.AI ships an "acknowledgment and dispute" workflow, letting agents formally push back on an automated score for a supervisor to review. A dispute mechanism only exists because the product team expects some scores to be wrong.
- MaestroQA's documentation describes a formal calibration process: humans grade a sample of the AI's outputs and produce what the company calls an "alignment score," a measurement of how closely the AI's judgment matches that of a human reviewer.

  In this application, the prompts get refined and rerun based on where they disagree. That's an ongoing check against drift, not a one-time setup step, because the company that built it doesn't assume it stays accurate on its own.
- CallMiner goes further and recommends against an all-or-nothing rollout. Teams can keep scoring manually in parallel with automated scoring "as the automation is built out," rather than switching the whole scorecard over on day one.

Put together, this is evidence that the vendors building these tools understand the coverage-versus-accuracy gap well enough to design around it, even while their marketing pages talk about "100%" without qualification.

## STUDY FOUND SOMETHING SHARPER THAN "MISSES CONTEXT"

Vendor documentation and practitioner threads are useful, but a 2026 paper accepted into the Association for Computational Linguistics (ACL)'s Findings track goes further.

Researchers evaluated 18 large language models (LLMs) scoring 3,000 real contact center transcripts, then tested what happened when they held the substance of a call constant and changed something that shouldn't matter to the score. Namely a note implying the agent's identity, a customer's account tier, or a snippet of that agent's prior coaching history.

*The result: pass/fail judgments flipped between 5.4% and 13% of the time across models, purely from those irrelevant changes.*

The single biggest trigger was priming the model with coaching-notes history, which pushed the flip rate as high as 16.4%.

Communications style differences, direct versus indirect phrasing, formal versus casual tone, moved scores less (in the 7% range), which the researchers took as a sign that the models were comparatively better at separating what someone meant from how they phrased it.

The same paper reports raw accuracy against human-annotated ground truth across those 18 models, ranging from 65.94% for one open-source model up to 93.84% for the strongest one tested, with no consistent split between objective and subjective criteria.

Worth disclosing directly: the paper's authors are researchers at Observe.AI, one of the vendors named earlier in this piece. That doesn't make the finding less real, since it went through the same peer review as any other ACL submission, but it isn't independent industry research either.

***What this adds to the practitioner accounts above is a mechanism, not just anecdotes.***

A score can shift because of who reviewed an agent's file last week, not because of anything said on the call actually being scored. That's a more specific problem than "AI misses empathy": it's AI treating something irrelevant to the interaction, like who last reviewed the agent's coaching file or what tier the customer's account sits on, as if it were evidence of how the call actually went.

## WHAT A PROGRAM THAT ACCOUNTS FOR THIS LOOKS LIKE

The practical version of this isn't complicated, but it does require treating your scorecard as something to audit rather than something to hand over to automation wholesale.

1. ***Start by sorting your existing scorecard criteria into two buckets, not by topic but by what it takes to score them correctly.***

- Bucket 1, pattern-matchable criteria (disclosures, verification, mandated language).

- Bucket 2, context-dependent criteria. Everything else, anything where the right answer depends on understanding the specific situation rather than checking for a phrase, and which needs a human-in-the-loop before it gets treated as a final grade.

2. ***Hand over Bucket 1 to AI with reasonable confidence, and expect it to be fast and consistent at scale in a way no human team reviewing 1%-2% of calls ever could be.***

That includes the two categories from the original framing, empathy and judgment calls on workarounds, but it also includes edge cases inside criteria that look objective on paper. Like a compliance step that had a legitimate exception, where a script deviation was the right call given what the customer said.

> ***BEFORE TRUSTING AI OUTPUT... RUN IT IN PARALLEL AGAINST HUMAN SCORES FOR A STRETCH...***

Before trusting AI output for trend reporting or coaching decisions, run it in parallel against human scores for a stretch, long enough to see where the two disagree and why.

There's no single industry-standard length for this, and be skeptical of anyone who quotes you a precise week count as an established benchmark; it isn't one. What matters is the discipline of comparing before committing, not the specific calendar length.

The comparison should check more than raw agreement, too. Given what the ACL research found about scores shifting based on irrelevant context, it's worth deliberately testing whether the same call, stripped of anything about who the agent is or what happened in their last coaching session, gets scored the same way twice.

Indeed, one QA team from that same Reddit thread described the opposite failure mode; after switching to AI-driven scoring, their pass rate dropped so far that management redefined the passing threshold from 94% down to 60%.

This occurred not because performance had actually declined, but because nobody had validated the new scoring before making it the number that mattered.

That's the actual risk. Not that AI QA tools are bad at their job. It's that "100% coverage" quietly becomes "100% of scores treated as final," without anyone deciding that on purpose.

*Owen Zhang is the editor of CommsAdvisor.com, where he tests and compares contact center, help desk, and customer support software for buyers evaluating vendors like Zendesk, Freshdesk, and Aircall. He previously worked as a software engineer, including a stint at Huawei, before moving into independent B2B software reviews.*

# PREMIUM CONTENT + RESPECTED CONTRIBUTORS FOR CONTACT CENTER PROFESSIONALS

**ENABLING A NEW GENERATION OF CX AND EX PROFESSIONALS** to create successful customer management strategies, develop cutting-edge technologies, refine the skills necessary to advance their career, and build a culture that advances the contact center within the organization—that's what we do.

Since 2009, **Contact Center Pipeline** has leveraged the insight of today's notable CX and EX thought–leaders, along with our advisory board, expert magazine authors, blog contributors, and industry insiders; keeping our audience ahead of the trends transforming the contact center and customer sales, service, and support industries, improving outcomes, and the way companies engage with their customers.

**Contact Center Pipeline Magazine** is published monthly digitally and in print. For more information and to subscribe, visit our website.

Contact Center Pipeline

**INSIGHT AND INSPIRATION FOR CONTACT CENTER PROFESSIONALS**

© COPYRIGHT 2026, CONTACT CENTER PIPELINE, INC. ALL RIGHTS RESERVED.

**REPRODUCTION IN WHOLE OR IN PART WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER IS PROHIBITED.**
THE VIEWS EXPRESSED HEREIN ARE THOSE OF THE AUTHORS AND/OR SPONSORS AND DO NOT NECESSARILY REFLECT THE OPINION OF THE OWNERSHIP OR MANAGEMENT OF CONTACT CENTER PIPELINE, INC. OR PIPELINE PUBLISHING GROUP, INC.