MaestroQA

What Happens When You QA Your Chatbot?

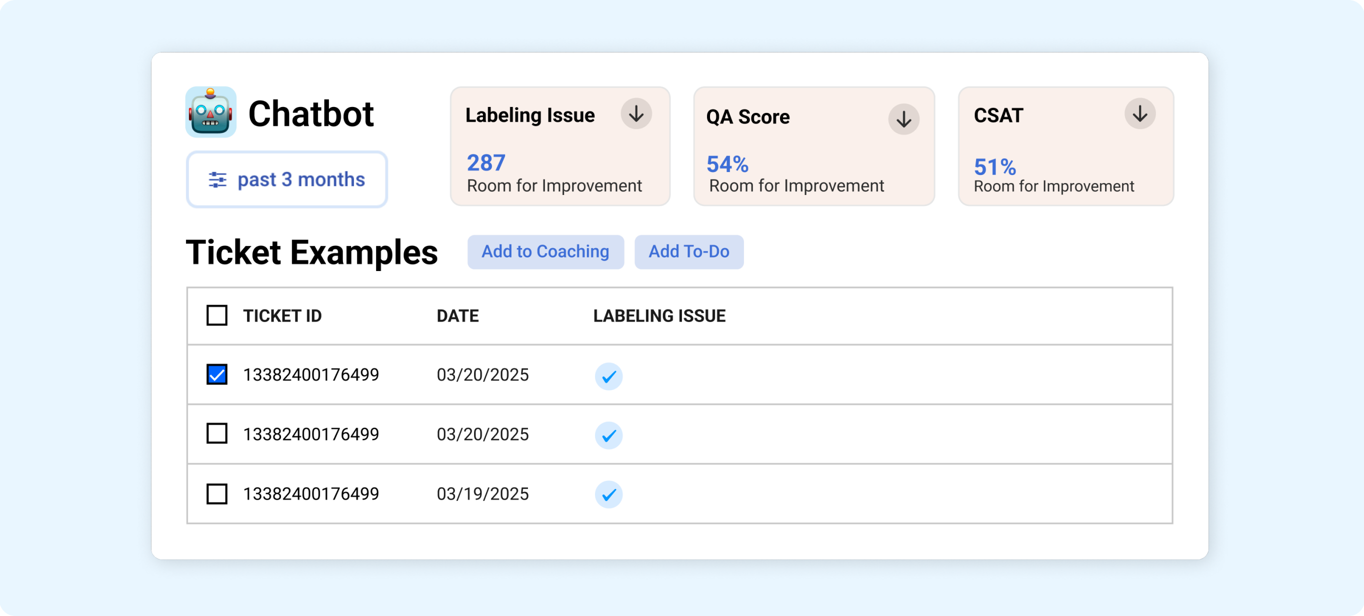
Hint: Surprising gaps and actionable insights revealed 🧪

The Problem

A B2B SaaS company wanted to understand why their chatbot wasn't resolving simple Flow-related questions. On paper, things looked okay—metrics like Automated Resolution (AR) and escalation rates didn't raise red flags. But customers were still reaching out for human help.

The team needed answers:

- Why are so many tickets unresolved?
- Are AR labels accurate?
- Is the bot giving helpful responses—or just technically correct ones?
- What's happening in escalations?



Chatbot

past 3 months

Labeling Issue: 287 Room for Improvement

QA Score: 54% Room for Improvement

CSAT: 51% Room for Improvement

Ticket Examples

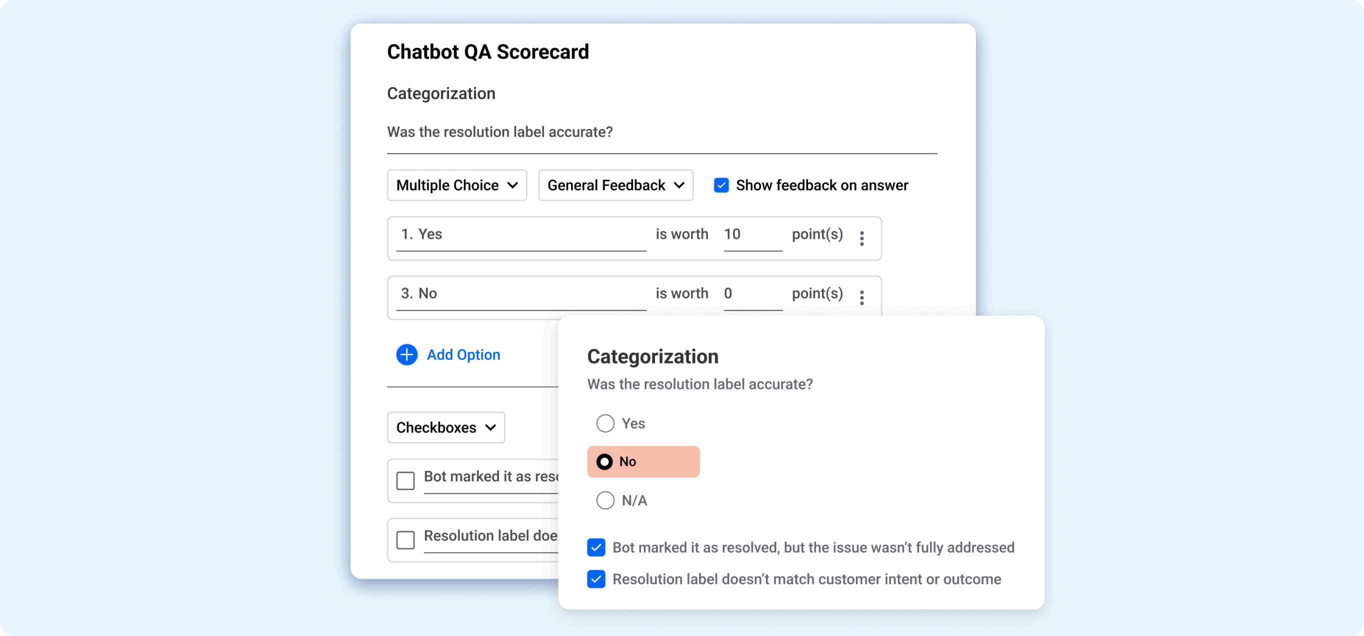
□	TICKET ID	DATE	LABELING ISSUE
☑	13382400176499	03/20/2025	✓
□	13382400176499	03/20/2025	✓
□	13382400176499	03/19/2025	✓

The Initial Targeted QA Sprint

To understand what was really happening in bot conversations, the team ran a small, targeted QA pilot using MaestroQA. Over two weeks, they selected 45 conversations to review—focused specifically on Flow-related issues, like setup questions or troubleshooting requests. These conversations came from customers on free or low-tier paid plans and had already been labeled by the bot platform as either "Resolved" or "Not Resolved."

Two QA analysts manually scored each conversation using a custom rubric designed for bots, assessing:

- ✓ Categorization – Was the resolution label accurate?
- ✓ Customer Experience – Did the bot greet the customer properly, use the right tone, and personalize responses?
- ✓ Technical Accuracy – Was the bot's answer factually correct?
- ✓ Escalation Handling – Did the bot collect the right info before handing off to a human?
- ✓ Auto-Fail – Did the bot completely fail to respond meaningfully?



Chatbot QA Scorecard

Categorization

Was the resolution label accurate?

Multiple Choice ⌄ General Feedback ⌄ ☑ Show feedback on answer

1. Yes is worth 10 point(s)

3. No is worth 0 point(s)

Add Option

Checkboxes ⌄

□ Bot marked it as res

□ Resolution label doe

Categorization

Was the resolution label accurate?

○ Yes

● No

○ N/A

☑ Bot marked it as resolved, but the issue wasn't fully addressed

☑ Resolution label doesn't match customer intent or outcome

📍 **The goal was to go beyond surface-level metrics and uncover actionable insights: what was working, what was falling short, and what could be improved.**

What the QA Sprint Uncovered

🔎 Labeling Issues

- ✓ 91% of chats were marked “Not Resolved”

  → QA flagged 2 that were actually resolved.

- ✓ 9% were marked “Resolved”

  → QA flagged 2 that were actually resolved.

  → The AR classification model had blind spots—highlighted only through human review.

🧩 Missing Customer Info

- ✓ 42% of conversations lacked the detail needed for the bot to troubleshoot

  → Customers weren’t giving enough input for the bot to help

  → Incomplete info also led to failed escalations

📎 Escalation Drop-Off

- ✓ 82% of chats escalated to a human agent
- ✓ When the bot tried to escalate a chat, it asked customers for more info—like links or account details. But most customers didn’t respond.

  → Agents received escalations without the context they needed to help

📉 Accuracy ≠ Relevance

- ✓ 75% of technically accurate answers didn’t fully address the customer’s core issue

  → The bot said the right thing, but not the helpful thing

Takeaways that Drove Action

📌 Refine the AR Classification Model

Human QA uncovered errors in how conversations were labeled as resolved or not. These findings fed back into model training to improve labeling going forward.

📌 Coach the Bot, Just Like an Agent

QA surfaced patterns where the bot’s answers were close but not helpful. These became coaching points for the bot’s manager—and in some cases, content updates in the Help Center.

📌 Improve Customer Input Quality

QA helped spotlight how vague customer inputs derailed resolution. The team is now testing ways to prompt users for clearer, more actionable information at the start of conversations.

What Changed Next

The ongoing Chatbot QA workflow now looks like this:

1. Run targeted QA on high-opportunity chat types
2. Track insights using MaestroQA dashboards
3. Send weekly feedback to the bot’s coach
4. Decide on action: Coaching, Help Center update, or model refinement
5. Monitor trends in Automated Resolution over time

Bot QA = Real ROI

This team discovered:

- ❌ Misclassified resolutions
- 📉 Missed coaching opportunities
- 🔁 Broken escalation flows

And now? They're fixing the right things—faster.

Why Teams Use MaestroQA for Chatbot QA

- ✓ Review 100% of chatbot interactions with AutoQA
- ✓ Spot trends and issues faster with Performance Dashboards
- ✓ Dig into specific conversations to understand what actually happened
- ✓ Find coaching moments to improve bot replies and escalations
- ✓ Brings your chatbot data into one place for easy, targeted QA

Want to uncover what your bot might be missing?

🚀 Let's run a chatbot QA pilot and find out.