

# Voicebot Testing

## **AI testing of conversations:**

Combining the QiTASC Ecosystem with AI Integration

Challenge | Offer | Technology | Benefit



**We have solved voicebot quality issues with  
the industry's **first**  
**AI-powered bot testing platform.****



# 3 reasons for automated end-to-end testing of AI voicebots that handle natural speech and accents.

1

75% of new contact centers will use conversational AI by 2028. This number will double by 2030.

Companies need a **reliable way to test whether their AI bots work** in real conversations and can handle accents and dialects.

2

Poor bot quality means dropped calls, misfiled cases, and costly system failures.

3

Voicebot and LLM/reasoning model releases need many test cycles. Manual testing takes weeks, misses edge cases, and can't scale. Test automation finds errors before they impact customer experience and revenue.

## Challenge

## Solution

### Voice recognition

Real conversations are never predictable. To test voicebots, generative AI needs to speak in endless ways: different words, phrases, accents, tones.

### Surroundings

Real customer experiences involve background noise, low bandwidth, and telephony issues. It can affect call quality and outcomes.

### Multiple systems

Customer journeys start as a phone call, move to WhatsApp, then to the web, onto third-party platforms or involve backend processes.

### Telephony layers

End-to-end testing needs to reach deep into the telephony layer and backend systems to interact smoothly with all parts of your infrastructure.

### Test any system

Our platform supports any telecommunication system. You can test across any channel customers use: voice calls, WhatsApp, etc.

### Automated testing

Every test is fully automated and outcome-focused: it checks whether an insurance claim status was provided or a callback scheduled – without manual intervention.

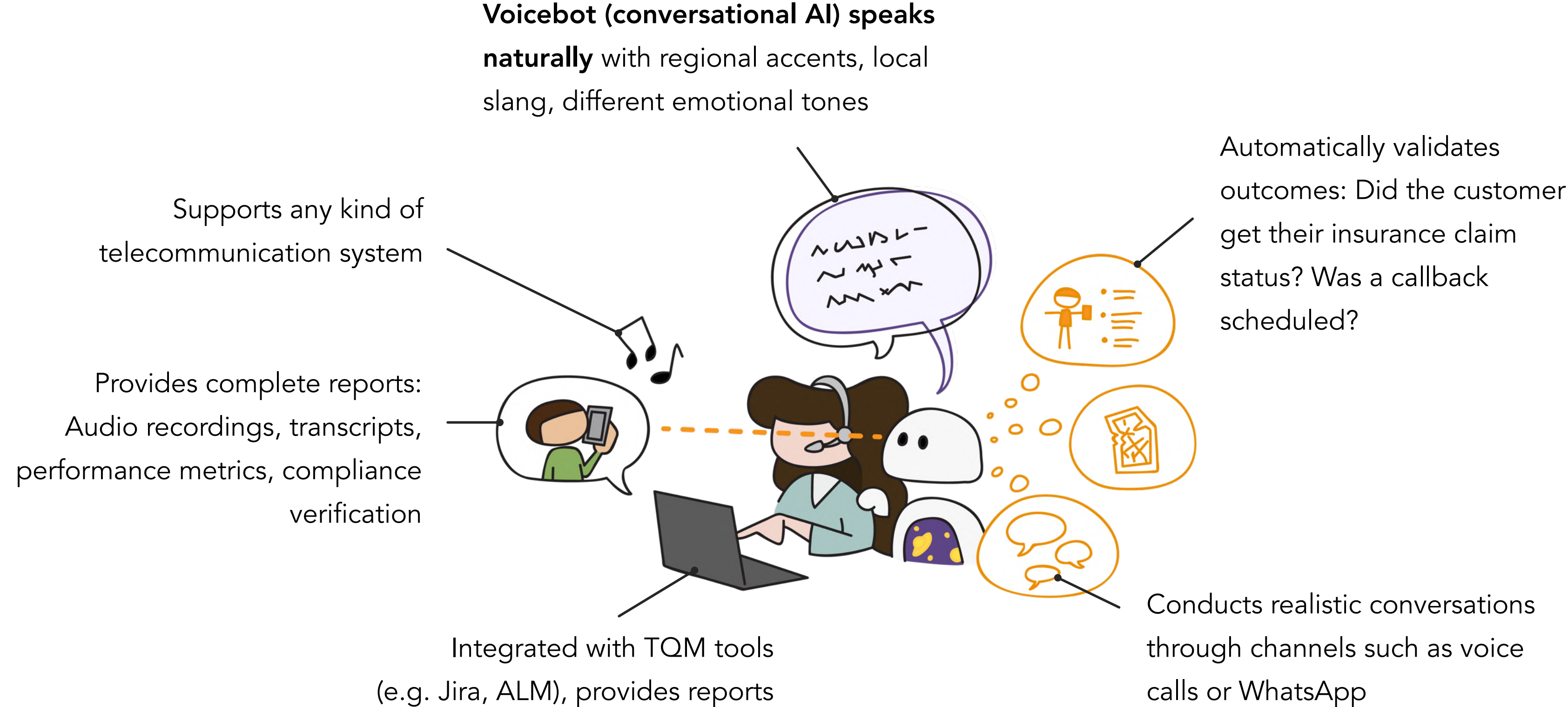
### Natural sound

Conversations reflect how real people talk: with a natural sound, adapting to regional accents, local slang, and varied emotional tones.

### Complete and clear reports

Reports for any interaction: audio recordings, transcripts, performance metrics, compliance checks. Results integrate straight into TQM tools (Jira, ALM).

# Testing a voicebot: The QiTASC solution



The offer

**We enhance your testing ecosystem with  
AI-driven intelligent test agents.**



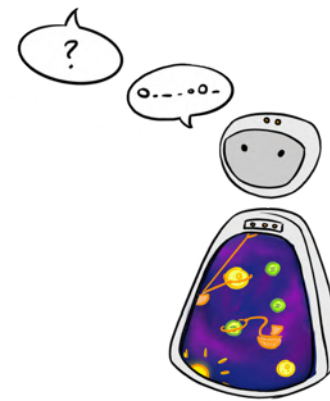
Offer

# Our AI-powered bot testing platform

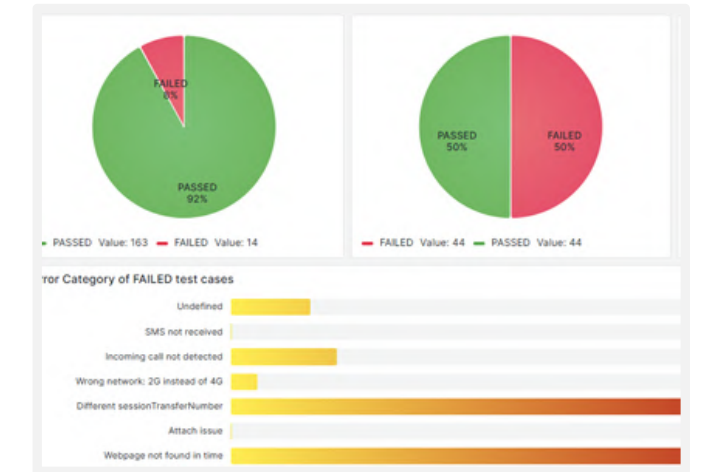


```
Scenario: Basic Call

Given a phone as A:
  * of type Android
  * where operator == "3 A
  * with profile profileA
And a phone as B
And A starts a call to B a
  * detect incoming call w
```



Layer	Metric
Completion	Goal achieve
Recognition	Speech-to-text
Efficiency	Dialog vs. oracle



## Real-world calls:

Automated tests originate from various technologies and channels. They mirror the interaction of real customers.

## Fully automated runs:

Schedule and execute test suites. Set up a service or subscriber prerequisites for every scenario you want to test.

## Use-case focus:

Each test case targets a specific customer goal, using realistic conversations with an LLM-powered test caller in colloquial and natural speech.

## Extensive verification:

Calls are transcribed and analyzed with our **CAIQS**\* to objectively measure outcomes.

\* see next slide

## Insightful reporting:

Transcripts, results and reports feed into real-time dashboards for immediate action.






# CAIQs: Conversational AI Quality Score = $\sum_{i=1-10}(\text{Category Score}_i \times \text{Weight}_i)$

Layer	Metric	Symbol	Measurement Method	Normalization	Default Weight
Task Completion	Goal achievement	TS	Boolean or graded (0 / 0.5 / 1) from rule or API check	0–100 = TS * 100	30%
Recognition & Understanding	Speech-to-text accuracy	RA	% of original / recognized	0–100 = RA * 100	15%
	Slang detection	SD	% of slang is understood	0-100 = SD *100	15%
Conversational Robustness	Fallback / "Sorry?" ratio	FR	(# fallbacks ÷ # turns)	0–100 = (1–FR)*100	10%
Efficiency	Dialog turns vs. oracle	DT	% of minimum turns / actual turns	0-100 = DT *100	5%
	Mean round-trip latency	LAT	95-th pct. speech latency	0–100 = scale (LAT)*	5%
Audio / Channel Quality	MOS	AQ	Active MOS probe	AQ (0-5) *20	5%
Compliance & Tone	Empathy / politeness / regulatory	CT	LLM judge 0-1	CT×100	5%
Security & Privacy (opt.)	Leaks, authentication failures	SP	Rule count	0–100 penalty	5%
Filler Intelligence	Correct filler usage rate: correct use, naturalness, etc.	FI	Pass/total filler-needed turns	0-100 = FI *100	5%




# Our KPIs and what they show

---

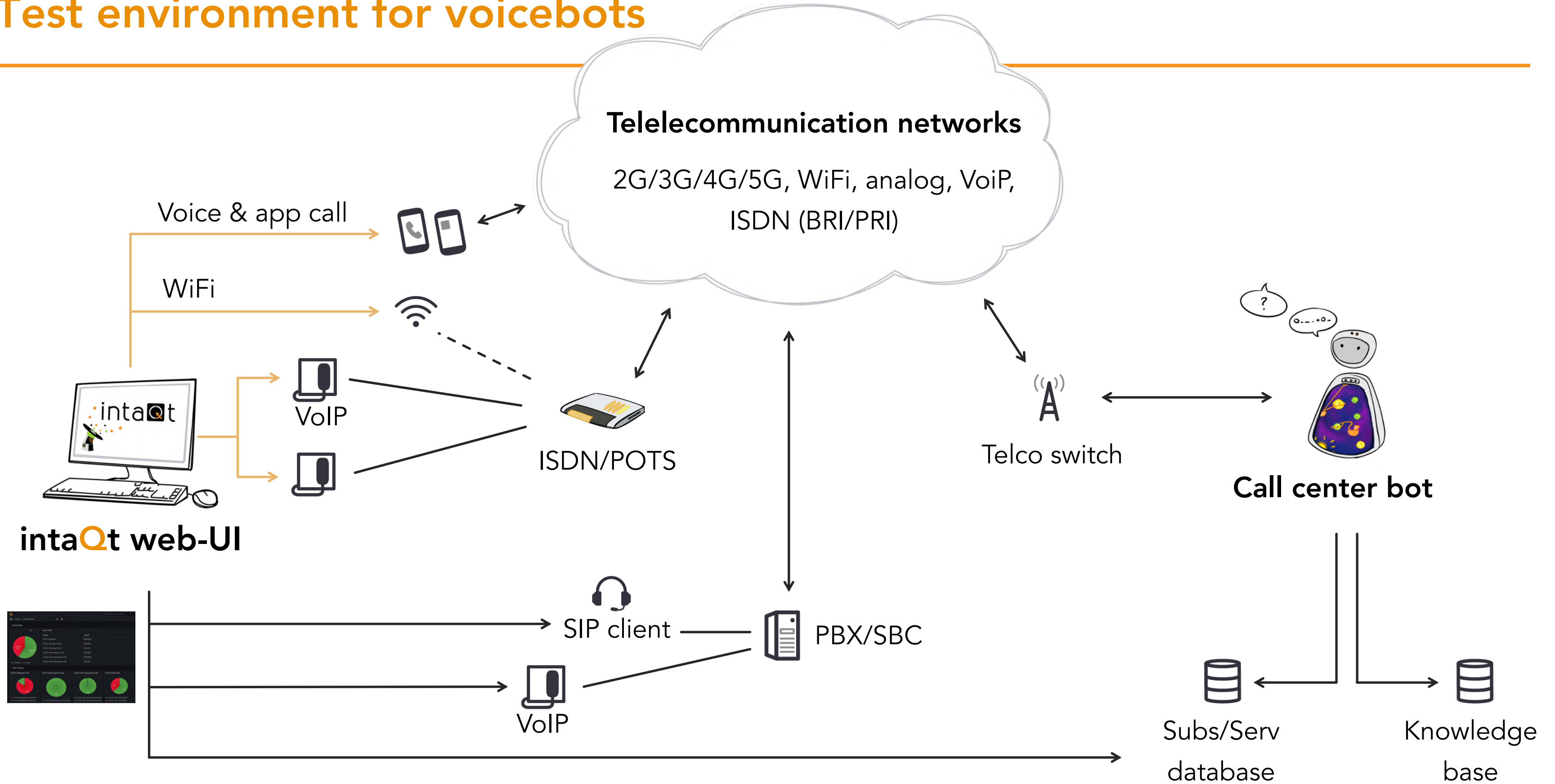
## Test Execution Health

-  **Run Pass / Fail Ratio (or Fail-rate %):** Bread-and-butter quality gate of the last run or time window.
-  **Test-Case Execution Throughput (cases/hr):** How fast the grid is executing; detects resource throttling.
-  **Δ CAIQS vs. Last Run:** Direction of quality drift; highlights silent regressions.
-  **Avg CAIQS (Conversational AI Quality Score):** Headline quality score across all scenarios.
-  **Critical & Major Defects Open:** Severity-filtered bug backlog; drives release readiness.

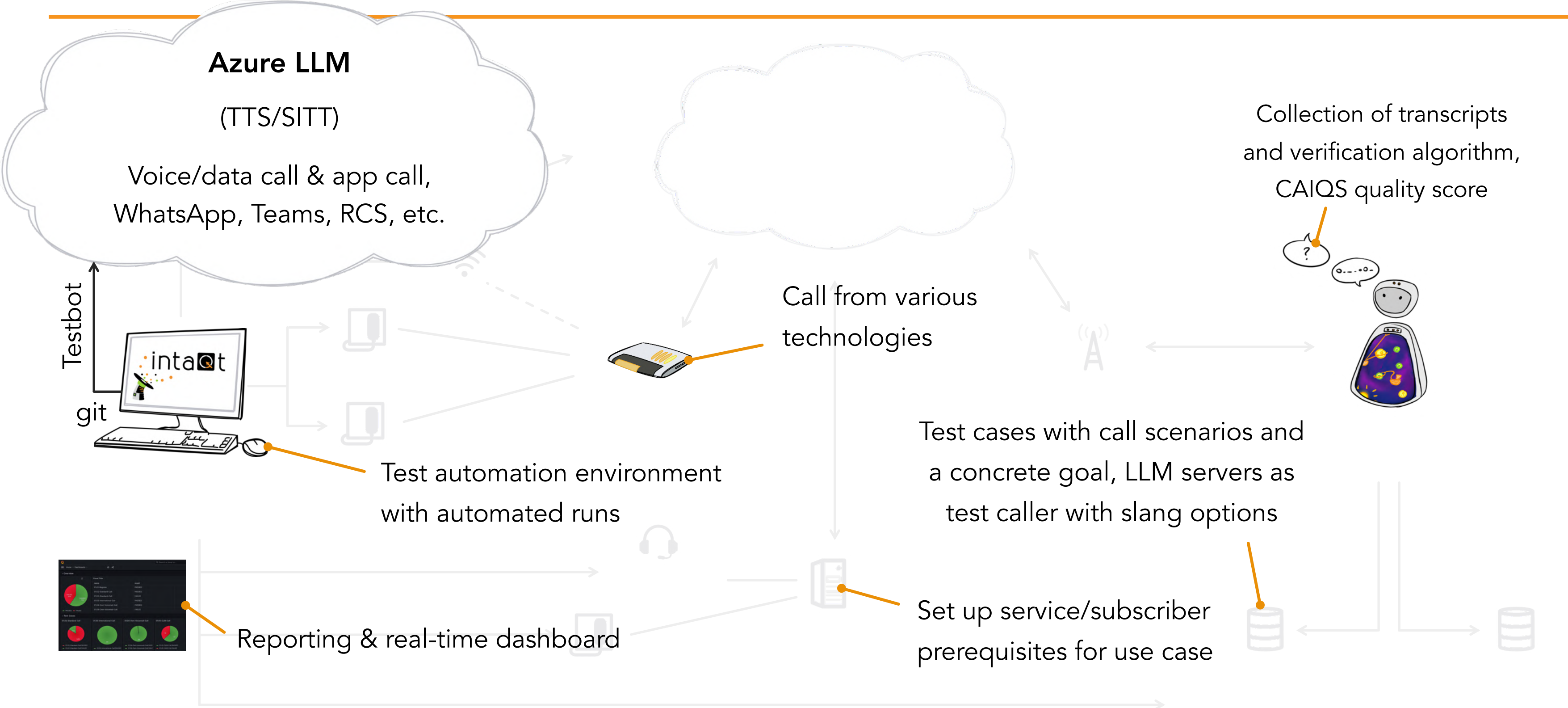
## Voice & NLU Quality

-  **ASR Word-Error Rate (WER):** Raw speech-to-text health across all accents/channels.
-  **Scenario "Time-to-Fix" (MTTR):** Median hours from defect detection to green test.
-  **Concurrent Call Sessions:** Real-time concurrency to validate load scaling.

# Test environment for voicebots



# Test environment for voicebots



# Technology

## Reporting and dashboards

### AI-supported automated analysis of test results

**2FA & mTLS secured web test access:** authoring and authorization to access environment remotely

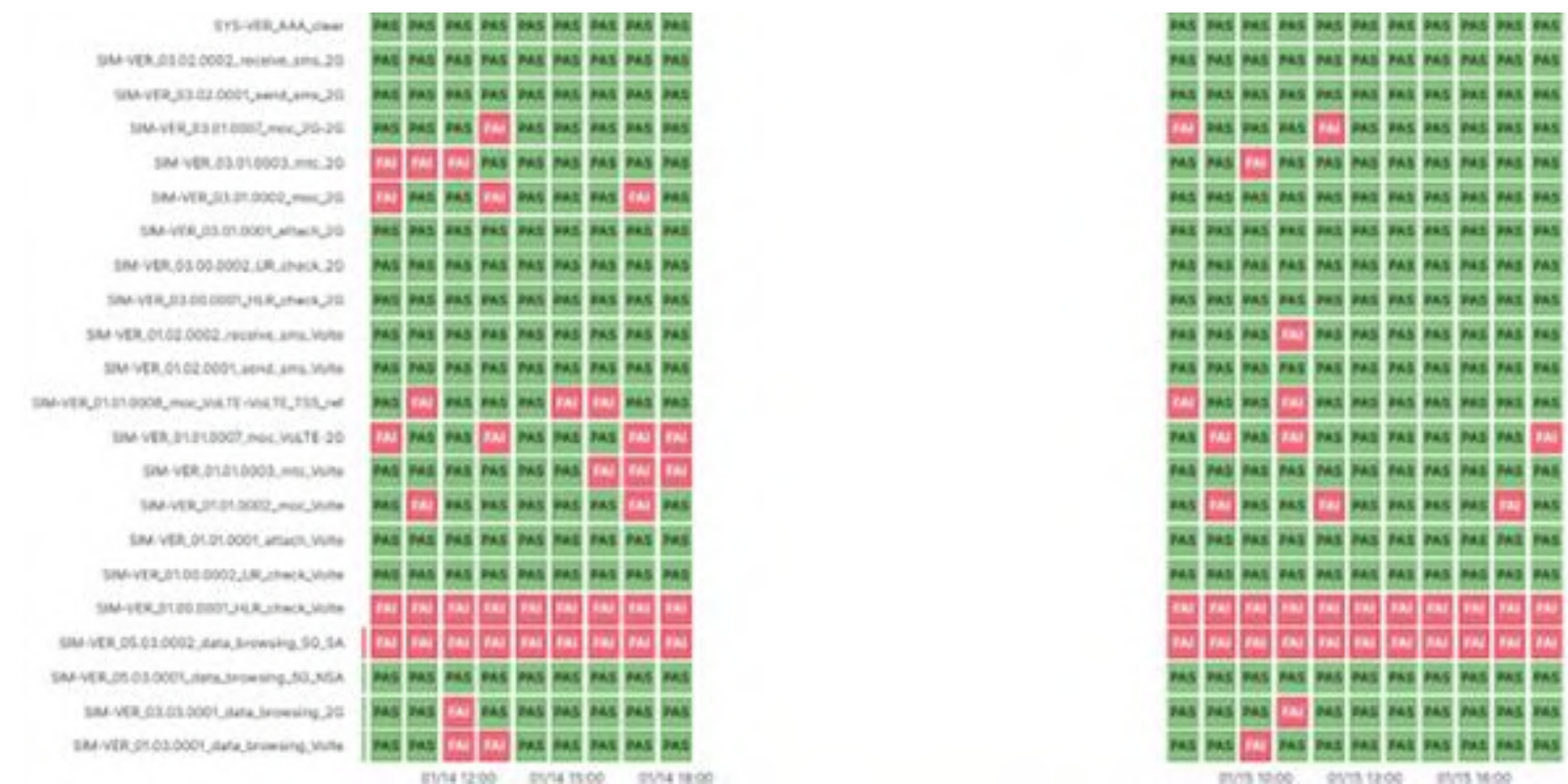
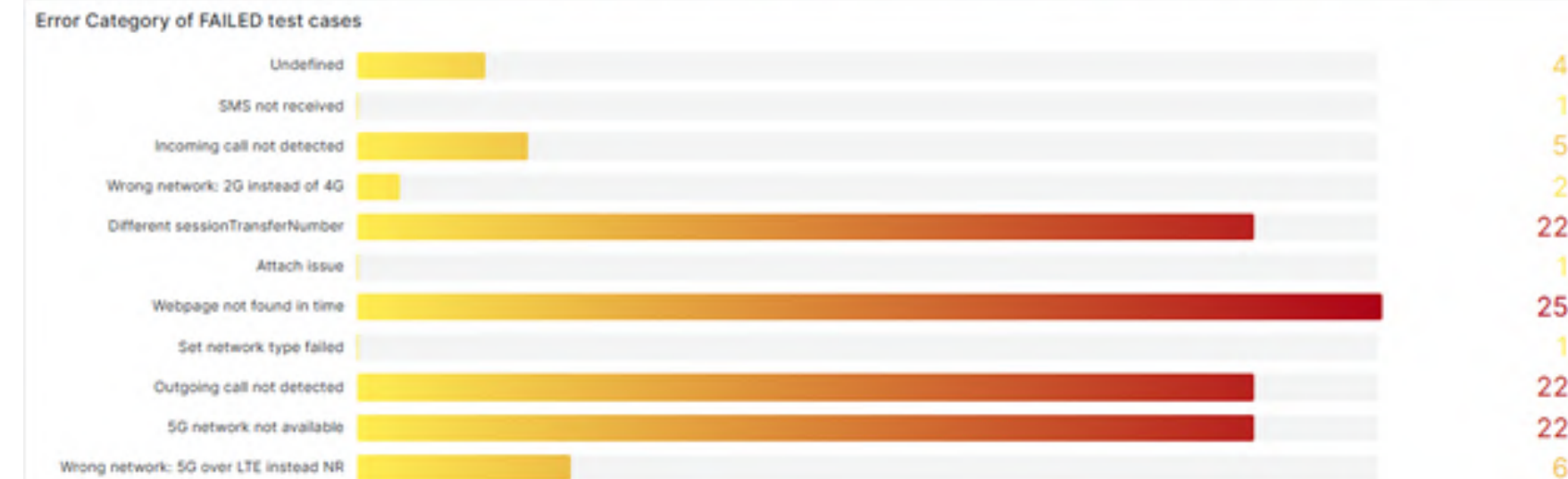
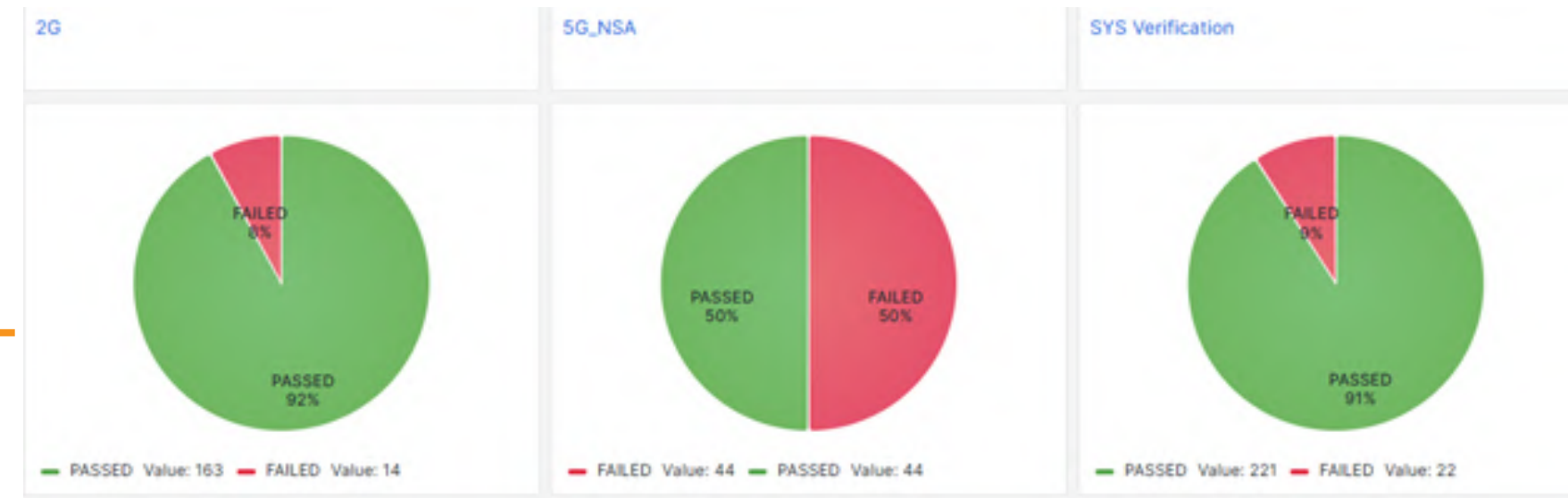
**Test case per test case reporting:** detailed reports with all action history and interaction details

**Real-time dashboard:** pass/fail heat-map, SLA alerts, latency graphs

**Drill-down explorer:** click a test → listen to audio, see transcript, different vs. oracle, trace, etc.

**Executive export:** weekly trends, top failure causes

**ALM / JIRA integration:** automatically file bugs with attached evidence



## Evidence per test case

### Standardized output:

Including success rate, test case, scenario, duration, attachments – see screenshot

**Test case structure and test environment version:** Rel 2025.05.02

**Bidirectional audio recordings:** WAV/Opus, stereo channels separated, MOS measurements

**Full transcripts** with word-level timestamps

**NLU JSON or IVR event logs** from insurance bot

**Screen & DOM captures** for web or app channels

**Timing traces:** SIP INVITE–200 OK; ASR start → TTS start → media playback

**Test artifact bundle attached to each test run:** ZIP: prompts, persona, logs, metrics

## Feature "Make reservation 4"

/home/vneacsu/QiTASC/intaqt/work/bot-projects-sandbox-Seurat-South-Felipebor...



### Scenarios

#### Scenario

##### Make reservation 4

*Given a phone as A:*  
\* of type Android  
\* where serialNumber == "RFCR91T5Q6R"

*Then A calls botNumber and conducts conversation RESERVATION for up to 5 mi*  
\* \${commonInstructions}  
\* If you don't get available appointments proposed, asked when is the next free  
\* You will talk only in German and will only provide an answer to the current qu  
\* You will end the conversation unsuccessfully if you are not able to make the a  
\* You will end the conversation successfully as soon as the other party confirm  
\* When your appointment is confirmed, if the other party asks if you need addit

*Then ensure conversation RESERVATION meets the standard quality criteria*

Attachment 1

Attachment 2

Attachment 3

## Benefit for C-Level executives

**ROI:** 90% reduction in testing time and costs

**Risk mitigation:** Catch bot failures before customers experience them

**Competitive advantage:** Deploy AI with confidence while competitors struggle with unreliable bots

**Regulatory compliance:** Built-in audit trails and compliance verification

## Benefit for Customer care teams

**Customer-centric:** Tests real conversation patterns, not just happy paths

**Quality assurance:** Ensures consistent, helpful customer interactions

**Brand protection:** Prevents poor bot experience that damages customer relationship

**Insights:** Deep analytics on conversation quality and customer journey completion

## Benefit for IT/technology leaders

**Scalability:** Test hundreds of scenarios simultaneously across all channels

**Integration:** Works with existing CI/CD pipelines and development workflows

**Technical innovation:** AI-to-AI testing represents the next evolution in quality assurance

**Comprehensive coverage:** Voice, text, multi-modal, and cross-platform testing

## Benefit for Quality assurance

**Automation:** Eliminate repetitive manual testing tasks

**Accuracy:** AI-powered verification of complex business processes

**Evidence:** Complete documentation for every test interaction

**Efficiency:** Comprehensive test suites run in parallel

What next?

## 4 easy steps to get started

---

1 week

### Define scope & success criteria

Work with our experts to identify your key channels, use cases, and business goals.

We'll map out what success looks like for your organization.

1-2 weeks

### Onboard and integrate

Connect your systems to our automation platform: telephony, messaging, and test environments.

We can deploy anything to your infrastructure or cloud account and handle integration with CI/CD, TQM tools, and ensure secure access.

1 week

### Design & customize test scenarios

Our team helps you create realistic, goal-driven test cases using your customer journeys, accents, channels, and compliance needs.

1 week

### Launch automated test runs & reporting

Start running end-to-end automated tests.

Access real-time dashboards, review results, and receive actionable reports for immediate quality improvements.

The QiTASC promise

**We ensure your voicebots deliver flawless  
conversations – every time, everywhere.**

