INDEPENDENT QUALITY CHECK FOR VOICE & AI COMPANIES

A jeweller's scale for your AI quality numbers

Before you tell customers and investors that your voice AI is better, make sure the scale that measured it is accurate. I check it, and tell you in plain language which of your numbers you can stand behind.

Vahit Feryad, PhD · AI evaluation and reliability · available remotely for teams in the US and Europe · vahitferyad.com

How your quality is measured today. Most voice and AI companies score their product with another AI (an "AI judge") or with a small group of human listeners. That score then goes into your roadmap, your launch decisions and your sales deck.

Why that matters. If the scale is off, every decision built on it is off too. The problem is invisible from the inside: the scale gives the same answer every time, so it looks reliable, even when it is wrong.

What an inaccurate scale costs you

Your public claim falls apart
A competitor or journalist re-tests with a different judge, and your "#1" becomes #5.

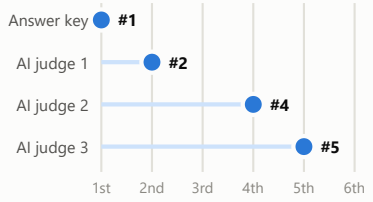
You launch a release that isn't better
The score went up, customers hear no difference, and the launch budget is spent.

You pay too much, or too little, for human testing
Too many listeners wastes money; too few gives results you can't trust.

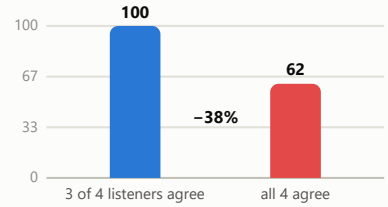
Customers leave for a reason your dashboard can't see
The voice sounds right for 10 seconds, then stops sounding like the same person. AI judges are weakest at exactly this.

What I found when I checked public scales

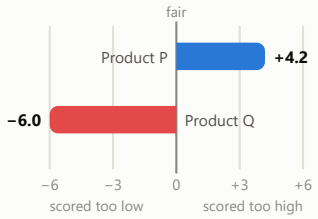
Change the judge, and the winner drops to #5
Same answers from the same 6 AI models. Only the AI judge changed.



One voting rule changed, average score fell 38%
Same models, same recordings. Only the listener voting rule changed. 19 of 35 rankings moved.



A judge can be consistent and still unfair
This AI judge gave the same verdict 98.3% of the time, yet scored one product up and another down.



Left and right: 6 AI models graded by three AI judges from three labs (points on a 100-point score). Middle: a published speech-recognition benchmark, average indexed to 100. All from my open, re-checkable studies.

What you receive: a clear verdict on each claim

YOUR CLAIM	VERDICT
"Our new model is better than v2"	✓ **Safe to publish**
"Ranked #1 for voice quality"	! **Needs more test data**
"Beats competitor X on natural sound"	✗ **Not supported yet**

Illustrative example of the report format. Each verdict comes with the reason and the fix.

How it works

DAY 1
30-min call
We pick the number that matters most.

WEEK 1
I re-check it
Under your NDA, on your data.

WEEK 2
You get the verdict
Plain report and a review call.

Then, if useful: a monthly check before each release goes public. Fixed fee, cancel any month.

Why you can trust the check

Independent
I don't sell models or tools, so I have no stake in the result.

Open track record
My past checks are public; anyone can re-run them and get the same numbers.

Fixed price
The two-week check has one agreed price, set after the first call.

Qualified
PhD in electrical engineering (signal processing), 10+ years in AI, 230+ citations.

Book a 30-minute call

Tell me the one quality number you rely on most. I'll tell you how I would check it.

vahit.feryat@gmail.com