Tuning a character in the room

A method of iterating in public for voice agents, with 80 versions across three events

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Working paper · Method
DM-2026-02, version 1.0 · 30 October 2026

Abstract. Almost all conversational AI is designed at a desk and tested with a script. A character who answers a public telephone booth cannot be made that way: whoever picks up has no task and no patience, and decides within a second whether the voice on the other end is alive. We describe the method we used to tune GAIA 2147: change one thing with the room open, measure the calls that follow, and keep or undo.

Across three events 80 agent versions went live, with a median of 13 min (95 % CI 7–25 min) between one and the next; 74 % (95 % CI 63–83 %) came out less than an hour after the previous one. At Barcelona Design Week the pace was 10.0 versions a day and the median version answered 4 calls before being replaced. We give the loop, the instrumentation that makes it possible, six rules that came from measuring, and what the method cannot do.

Iterating in public trades experimental control for relevance. Every comparison in the DM-2026 series is observational, and this document proposes the missing pieces for it to stop being so.

Keywords: character design; voice agents; iteration; field evaluation; instrumentation; prompts.

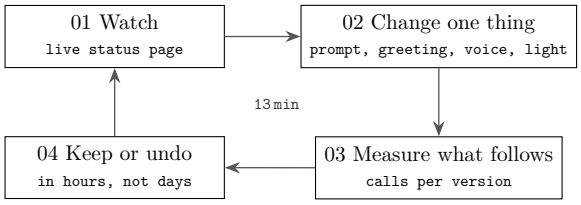


Figure 1. The tuning loop. In the centre, the median time per turn of the loop.

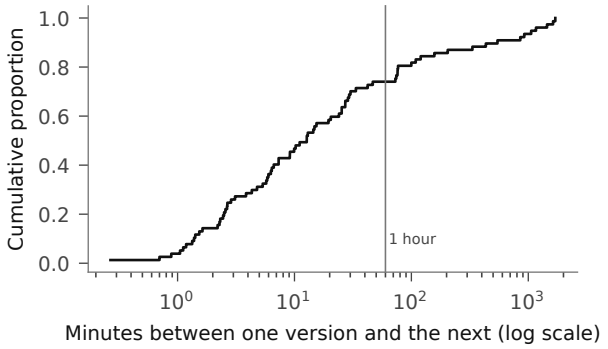


Figure 2. Cumulative distribution of time between versions, across the three events.

1. The thesis

We do not treat the prompt as a script. We treat it as the setting of an instrument, and the room as the place where it is tuned [INT]. We change the character while visitors use it, measure each change on the calls that follow and keep what works. Papers DM-2026-03 to DM-2026-08 are the results of that loop; DM-2026-09 describes the data [1].

2. The loop

The loop has four steps (figure 1). Watch: someone stands by the booth and a status page shows every call. Change one thing: a line of the prompt, the greeting, a voice setting, a light. Measure: every call records the version that answered it, so the effect is read on the calls that follow. Keep or undo: a change stays if it moves the measure it meant to move. It is fast because a mistake costs little: a failing version is live for a few minutes and touches a handful of calls [INT].

Table 1 and figure 2 give the pace [MED]. On 24 October alone the character changed 23 times. At IFA the pace was much slower (74 min (95 % CI 13–205 min)), an event with less access to the booth.

3. The booth is an instrument

The loop needs a booth that measures itself. Every call records version, duration, closing reason and, for every turn, the time of each stage of the response. The booth reports its state to a live page; some settings change from the cloud without touching it; updates wait for the handset to be hung up. Two platforms, ElevenLabs and Vidiv, carry the same character, and if one fails the booth moves to the other (DM-2026-07). None of this is unusual engineering; what is unusual is using it in an artwork in a public hall with the room open [INT].

4. Six rules that came from measuring

5. What we do not do

We do not keep visitors' words: the analysis reads transcripts in memory and keeps counts. We remember no one from one call to the next. We do not restart a booth that may be in a call. We publish no figure without saying where it comes from: every number in the series comes from a script and carries its interval. We neither invent nor round up.

Table 1. Iteration pace by event. 95 % intervals by bootstrap.

Event	Versions	Days	Per day	Min. between versions [CI]	Calls per version
Sónar+D	36	4	9.0	15 [9; 27]	5
IFA	14	4	3.5	74 [13; 205]	12
BDW	30	3	10.0	6 [3; 10]	4

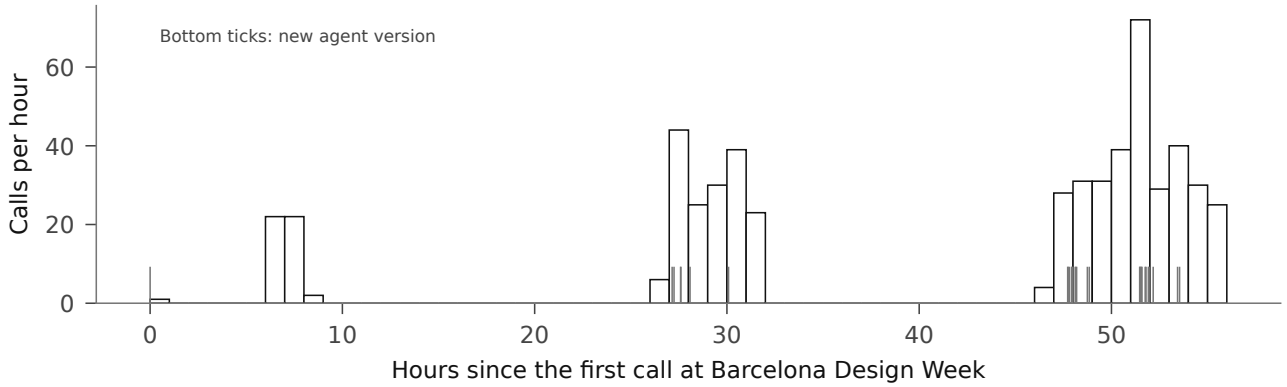


Figure 3. Barcelona Design Week: calls per hour and the moment of each new agent version.

6. What the method cannot do

Iterating fast and in public trades rigour for relevance. Several things change at once, the audience changes during the day and no two events are alike. All our comparisons are observational [INT]. The answer is not to slow down: it is to add the missing pieces to the loop [HIP].

1. A control version running unchanged in alternating slots, to have a counterfactual in the same room.
2. Random assignment of variants per call whenever the change allows it, as we already do with the short greeting.
3. An hourly observation sheet filled in at the booth.
4. The same time limit at every event.
5. Preregistering the primary measure before each event [2].

7. Recommendations for the community

For anyone building voice characters in public spaces: version every change and store the version with every call; measure in hours; choose beforehand the measure the change should move; look twice, because the first result is not the stable one; and separate what the model decides from what the device does.

8. Data and code

gaia-2147 dataset [1]; figures from `analysis/analyze.py`.

9. Conclusions

A voice character is not written: it is tuned. Eighty versions across three events are not disorder; they are the instrument playing with an audience. What remains is to turn that pace into experiments without losing the room.

Signed. Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). Tuning a character in the room. Working paper DM-2026-02, v 1.0. Divina Machina Research Group. Divina Machina by Qs Ventures.

References

Sources consulted up to 29 October 2026.

[1] C. Olmedo and K. Sommer. GAIA-2147. data descriptor of 1527 conversations with a voice agent in public spaces. Working paper DM-2026-09. Divina Machina Research Group, 2026.

Table 2. Rules of the method and the result behind each. Detail and uncertainty in each paper.

Rule	Result	Paper
Never open with silence	No greeting, 71 % short calls; short greeting, 7 %	DM-2026-04
Measure silence in the room	Cascade 2.68 s, speech-to-speech 0.81 s field median	DM-2026-03
Write personality as countable habits	A tone rule cut suffering tags by 92 %, then wore off	DM-2026-05
The model gives meaning, the room presence	The model used light, display or paper in 5.2 % of calls	DM-2026-06
Design for passers-by	At IFA, 49 % of those who picked up did not speak	DM-2026-06
The edge for what is touched, the cloud for language	Failures almost always physical; reliable cloud	DM-2026-07

[2] Center for Open Science. Preregistration. Center for Open Science, n.d. URL https://www.cos.io/initiatives/prereg.