# Personality as a vector of habits

## Drift of tone rules and within-call memory in a voice agent

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Working paper · Observational study
DM-2026-05, version 1.0 · 30 October 2026

**Abstract.** A voice agent's personality is written as a description and encountered as behaviour. We propose to measure it as a vector of countable per-turn habit rates and follow it across seven periods of the same character, GAIA 2147, through 82 prompt versions.

A tone rule added live reduced suffering voice tags to 0.08 (95 % CI 0.04–0.12) times their previous rate within hours. That same evening, with the rule still in the prompt, the rate rose back to 5.27 (95 % CI 3.28–10.23) times that of the window in which the rule had just been added. Asking for questions took the agent to 70 % of turns ending in a question; asking for "roughly half" brought it down to 59 %. Picking up the visitor's words happened in 36 % (95 % CI 34–37 %) of eligible turns and was not associated with how much people spoke (Spearman 0.05 (95 % CI −0.04 to 0.13)).

Tone rules work and then wear off. We propose to measure personality after every change, within hours, and a memory for public characters that does not remember people. The periods also differ in greeting, time of day and audience; the effects on visitors cannot be separated here.

**Keywords:** agent personality; prompts; drift; voice tags; conversational memory; behavioural measures.

## 1. Introduction

A character is described with adjectives. GAIA is bitter, curious, guilty, tired and funny. What the visitor meets is something else: how much it talks, whether it asks or tells, whether it sighs, whether it sounds like an assistant. That behaviour can be counted. The industry has seen a model's personality move without anyone asking: in 2025 OpenAI withdrew a GPT-4o update that had become sycophantic [1], and Anthropic describes traits as internal directions of the model that can drift and be monitored [2]. At the scale of a prompt the question is more modest and practical: what does a tone rule do, how long does it last and how is it measured?

## 2. Measure

We define the observable personality of a version $v$ as a vector of rates per agent turn [DEF]:

$$\mathbf{h}_v = (r_q,\ \bar{w},\ r_e,\ r_s,\ r_a,\ r_c), \qquad r_x = \frac{\sum_j N_{x,j}}{\sum_j N_{\text{turns},j}}, \tag{1}$$

where $j$ runs over conversations. $r_q$ is the share of turns ending in a question; $\bar{w}$, words per turn; $r_e$ and $r_s$, voice tags per turn, all of them and the suffering ones (*sigh*, *softly*, *slow*, *sad*, *whisper* and similar); $r_a$, assistant phrases from a fixed list; $r_c$, the thread pick-up defined in section 5. Tags are stage directions interpreted by the voice model [3]. Counting is done by code on transcripts held in memory; no text was stored [4].

Each rate is a ratio of sums. Its 95 % intervals resample conversations, not turns, and changes between periods are expressed as rate ratios with conversation-bootstrap intervals within each group [5, 6]. A tone rule is a sought movement of $\mathbf{h}$; drift is the movement nobody asked for.

**Table 1.** Barcelona Design Week windows and what changed before each.

| Window | Previous change |
|---|---|
| A, until 13:53 on the 24th | Arrival prompt; long fixed greeting |
| B, 13:53 to 16:37 | Tone rule: dry, firm, direct; no sighs or whispers; five neutral tags allowed |
| C, 16:37 to 17:18 | Hand-written opening; greeting tests; short varied greeting |
| D, 17:18 to 18:41 | Sound human; roughly half of turns ending in a question |
| E, from 18:41 | Switch language at once if the visitor asks |

## 3. Periods

We compare Sónar+D, IFA and five windows of Barcelona Design Week defined by prompt changes, matched to version identifiers using our change log, to the minute [DEF]:

## 4. Results

### 4.1. Three events, three personalities

Table 2 gives the vector per period [MED]. At Sónar+D, GAIA sounded like a service, with 17.2 (95 % CI 15.6–18.9) assistant phrases per 100 turns. At IFA it asked less than ever, in 19 % (95 % CI 15–22 %) of its turns. In Barcelona the prompt asked it to be curious and, already in window A, 53 % (95 % CI 50–56 %) of its turns ended in a question.

### 4.2. The rule works

After the tone rule, suffering tags went from 0.27 (95 % CI 0.24–0.31) per turn to 0.02 (95 % CI 0.01–0.03), a ratio of 0.08 (95 % CI 0.04–0.12). Total tags fell much less (0.53 (95 % CI 0.46–0.61)): neutral tags took their place.

**Table 2.** Habit vector per period. Questions and thread as percentages; tags per turn; assistant phrases per 100 turns.

| Period | Conv. | Turns | Quest. | Words/t. | Tags/t. | Suff./t. | Assist./100 | Thread |
|---|---|---|---|---|---|---|---|---|
| Sónar+D | 681 | 3,452 | 30 | 22.2 | 1.04 | 0.53 | 17.2 | 34 |
| IFA | 247 | 668 | 19 | 17.3 | 1.39 | 0.49 | 1.8 | 41 |
| BDW · A | 263 | 1,754 | 53 | 20.2 | 0.83 | 0.27 | 7.1 | 38 |
| BDW · B | 112 | 557 | 70 | 18.2 | 0.44 | 0.02 | 2.7 | 38 |
| BDW · C | 31 | 158 | 75 | 18.0 | 0.61 | 0.05 | 3.8 | 32 |
| BDW · D | 46 | 338 | 70 | 19.2 | 0.52 | 0.02 | 3.8 | 36 |
| BDW · E | 75 | 625 | 59 | 20.2 | 0.68 | 0.11 | 3.4 | 32 |

Rates as ratios of sums over GAIA turns. 95 % CIs by conversation bootstrap are given in the text and the figure.

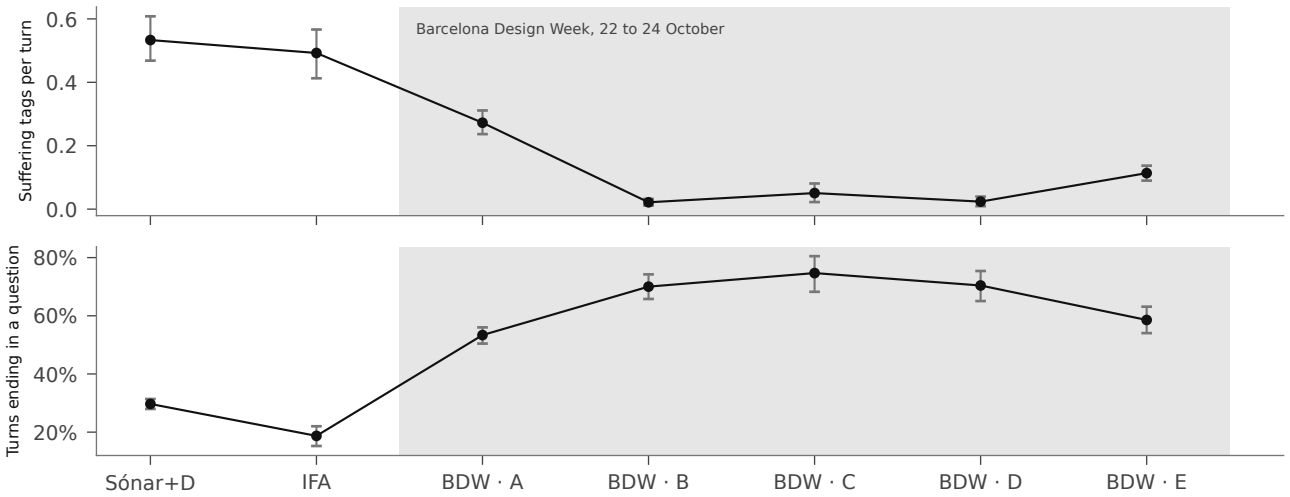



**Figure 1.** Two habits across the seven periods, with 95 % CI by conversation bootstrap. The grey band marks Barcelona Design Week.

GAIA kept an expressive voice and lost the sighs (figure 1).

### 4.3. The rule wears off

In window E suffering tags came back to 0.11 (95 % CI 0.09–0.14) per turn, 5.27 (95 % CI 3.28–10.23) times the window B rate. The rule was still in the prompt; no later change asked for the old voice [MED]. Two explanations fit the data and we cannot separate them: each change lengthened the prompt and the rule competes with what surrounds it; and the model has habits of its own that return when the instruction loses relative weight [INT]. The interval is wide because window B has few suffering tags.

### 4.4. Asking is not listening

The request to ask questions worked too well: between windows A and B the question rate rose by a factor of 1.31 (95 % CI 1.21–1.42) and reached 70 %. Asking for "roughly half" did not move window D (70 %); window E fell to 59 %, a ratio of 0.83 (95 % CI 0.75–0.93). Proportion instructions take time to show or need the next change to settle [INT]. Assistant phrases went from Sónar+D to window B at 0.16 (95 % CI 0.08–0.25) times their rate, without reaching zero; part of what remains is legitimate within a story.

### 4.5. Did visitors notice?

We cannot say (table 3). Window B had shorter calls than A, and E the longest, with 53 % reaching the time limit. But the greeting changed in those same hours, as studied in DM-2026-04, and the audience changes through the day. Behaviour moved clearly; its effect on visitors is confounded.

**Table 3.** Call outcomes by window. Confounded with greeting, time of day and audience.

| Window | n | Median (s) [95 % CI] | Short | Limit |
|---|---|---|---|---|
| A | 271 | 82 [66; 95] | 17 % | 34 % |
| B | 117 | 36 [21; 56] | 28 % | 19 % |
| C | 34 | 55 [14; 80] | 21 % | 18 % |
| D | 46 | 111 [79; 125] | 15 % | 46 % |
| E | 75 | 124 [103; 125] | 3 % | 53 % |

## 5. Memory within the call

GAIA remembers nobody between calls. That was a decision: a public booth receives children and strangers who never agreed to be remembered. It has three memories of different scope: what was said in the call; the place (a mood of the city that only arrives if it calls a tool, which it did in 24 of 1,527 conversations); and itself, written in the prompt.

For the first we define the thread pick-up. An eligible agent turn picks up the thread if it contains a content word of five letters or more that the visitor said earlier and the agent had not used [DEF]:

$$R = \frac{1}{|A|} \sum_{a \in A} \mathbb{1}[W(a) \cap (V_{<a} \quad G_{<a}) \neq \emptyset]. \qquad (2)$$

**Table 4.** Long conversations by thread pick-up tercile.

| Thread tercile | n | Thread | Visitor words/t. | Limit |
|---|---|---|---|---|
| Low | 187 | 17 % | 8.5 | 47 % |
| Middle | 187 | 33 % | 9.0 | 49 % |
| High | 187 | 50 % | 9.8 | 44 % |

Overall, $R$ was 36 % (95 % CI 34–37 %). In the 561 conversations with at least five eligible turns, the per-call median was 33 % (95 % CI 33–38 %). The rate barely moved with the prompt, between 32 % and 41 % depending on the period.

Picking up the thread was not associated with the visitor talking more (Spearman 0.05 (95 % CI −0.04 to 0.13), $p = 0.26$). By tercile (table 4), visitors in the top tercile said 9.8 words per turn against 8.5 in the bottom one, and the share of calls reaching the limit was similar. The measure is crude: repeating a word is not understanding it [INT]. What it shows is narrow: mechanically reusing what people said is not, by itself, what keeps them on the phone.

### 5.1. What the research says

A longitudinal study with 24 people over 10 sessions found that perceived memory depended on the relationship built and acted on later enjoyment through self-disclosure [7] [EMP]. A Wizard-of-Oz study with 30 participants found an advantage for context understanding and not for personalisation with long-term memory [8]. Conversational skill is associated with greater perceived humanness [9]. The large assistants turned memory into a product feature with explicit controls [10, 11]; they are personal assistants, not public characters.

## 6. Proposals

*Memory for a public character* [HIP]. Three layers, none about people: the thread as an explicit habit in the prompt; an anonymous memory of the room with the topics of the day, never tied to a voice or a name; and a fixed canon of the character's memories so it does not contradict itself between visitors.

*Proposed experiment* [HIP]. Alternate days or hours with and without room memory, with the same greeting, stack and limit. Primary measure: share of calls in which the visitor speaks five turns or more. To detect a move from 50 to 60 % with $\alpha = 0.05$ and power 0.8 one needs

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(p_1\,(1-p_1) + p_2\,(1-p_2)\right)}{\left(-p_1 + p_2\right)^2} \quad (3)$$

385 calls per group, two or three days at the Barcelona Design Week pace. The expression is generated by symbolic computation in the analysis and checked against the numerical calculation. Figure 2 shows how much $n$ grows if the sought effect is smaller.

## 7. Threats to validity

- Heuristic counts: a question mark is not always a question and a fixed list misses paraphrases.

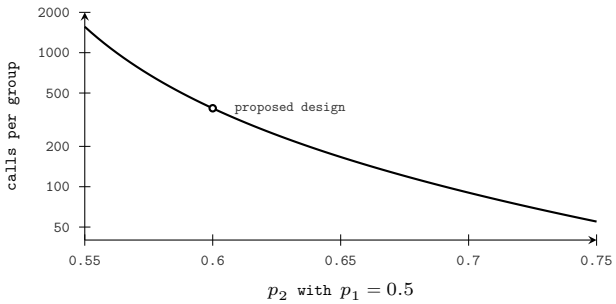



**Figure 2.** Size per group from equation (3), with $\alpha = 0.05$ and power 0.8. Analytic curve drawn with PGFPlots.

- Windows defined by time and version, matched by hand with the change log.
- Several simultaneous changes on 24 October: greeting, prompt and audience.
- No human labels: nobody read conversations to validate the measures.

## 8. Recommendations for the community

1. Measure personality as per-turn rates, with conversation-level intervals, after every change.
2. Measure again hours later: the first result after a change is not the stable one.
3. Give the voice model a closed list of tags; an open list drifts back to the usual.
4. Ask for proportions, not absolutes: "always ask" produces an interview.
5. Separate conversation, place and character memory, and do not remember people in public spaces.

## 9. Data and code

gaia-2147 dataset [4]; figures from `analysis/analyze.py`. We always work with ElevenLabs and Vidiv; the data are from the ElevenLabs agent.

## 10. Conclusions

A personality written in a prompt is a hypothesis about behaviour. Counted turn by turn, the tone rule worked within hours and wore off the same evening. Personality is not a setting written once: it is measured, like the temperature of a room.

**Signed.** Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). Personality as a vector of habits. Working paper DM-2026-05, v 1.0. Divina Machina Research Group. Divina Machina by Qs Ventures.

## References

Sources consulted up to 29 October 2026.

[1] TechCrunch. OpenAI explains why ChatGPT became too sycophantic, April 2025. URL https://techcrunch.com/2025/04/29/openai-explains-why-chatgpt-became-too-sycophantic.

[2] Anthropic. Persona vectors. monitoring and controlling character traits in language models, August 2025. URL https://www.anthropic.com/research/persona-vectors.

[3] ElevenLabs. Eleven v3. ElevenLabs Blog, 2025. URL https://elevenlabs.io/blog/eleven-v3. Audio tags and expressive voice.

[4] 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.

[5] B. Efron. Bootstrap methods. another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

[6] C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society. Series B*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

[7] R. Sumida et al. Memory-driven self-disclosure and relational turning points. arXiv:2607.14593. Aceptado en ICMI 2026, 2026. URL https://arxiv.org/abs/2607.14593.

[8] POSTECH. Context understanding and long-term memory in conversational agents. Thesis, POASIS repository, n.d. URL https://poasis.postech.ac.kr/handle/2014.oak/123425.

[9] R. M. Schuetzler, G. M. Grimes, and J. S. Giboney. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems*, 2020. URL https://scholarsarchive.byu.edu/facpub/5655.

[10] OpenAI. Memory and new controls for ChatGPT, 2025. URL https://openai.com/index/memory-and-new-controls-for-chatgpt/. April 2025 update.

[11] Bloomberg. Anthropic tries to win users from ChatGPT with memory feature, March 2026. URL https://www.bloomberg.com/news/articles/2026-03-03/anthropic-tries-to-win-users-from-chatgpt-with-memory-feature.