# The room does half the work

## Body, audience and behaviour of a voice agent at three events

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Working paper · Observational study
DM-2026-06, version 1.0 · 30 October 2026

**Abstract.** A voice character is three things at once: a model, a body and the people in front of it. We compare the same character, GAIA 2147, in a telephone booth with light, display and printer, at three events: Sónar+D (701 calls), IFA (274) and Barcelona Design Week (543).

The room shaped the calls more than the script. At IFA, 49 % (95 % CI 43–55 %) of those who picked up said nothing, against 22 % in Barcelona. At Sónar+D, 36 % (95 % CI 33–40 %) of visitor turns were the agent's own voice picked up by the microphone. The model barely used its body: in Barcelona only 5.2 % (95 % CI 3.6–7.4 %) of calls included a light, display or paper tool. People were kind to the machine, tested it, and talked to it mostly about machines and the future.

We propose a division of responsibilities: the model decides meaning; the room, on the device, produces presence. Each event is a single sample of room, audience and configuration, and we treat it as such.

**Keywords:** embodied agents; public spaces; tools; user behaviour; acoustic echo; interactive installations.

## 1. Introduction

Almost all the design attention on a voice agent goes to the model. Our logs say that the body and the audience decided as much as it did. GAIA 2147 answers in a Telefónica booth with a handset, a 20 by 4 character display, a strip of 60 lights and a thermal printer. The model can use all four as tools; the prompt asks it to do so sparingly [1]. We always work with ElevenLabs and Vidiv; these data are from the ElevenLabs agent.

This work sets three findings side by side: the room changes the calls, the model hardly uses its body, and people do things a transcript barely records.

## 2. Data and method

The three events differ in audience, language, noise, voice stack, prompt, greeting and time limit. So we do not rank causes: we describe each room with its intervals and measure heterogeneity between events with chi-square and Cramér's V, with Holm correction when there are several behaviours [2]. Proportions with Wilson [3]; medians with bootstrap; per-turn rates with conversation bootstrap [4]. Behaviours, languages and topics are detected lexically in seven languages: each proportion is a lower bound [DEF]. Echo flags a visitor turn of four or more words with at least 60 % of them in GAIA's previous turn:

$$\mathrm{echo}(v_i) = \mathbb{1}\Big[|W(v_i)| \geq 4 \wedge \tfrac{|W(v_i) \cap W(g_{i-1})|}{|W(v_i)|} \geq 0.6\Big]. \quad (1)$$

## 3. Three rooms

Table 1 summarises the rooms [MED]. At IFA almost half the calls lasted under ten seconds and another half were mute. Differences between events are clear for short calls ($V = 0.22$) and mute calls ($V = 0.20$), both with $p < 0.001$. IFA combined the slowest stack, a trade-fair audience moving from stand to stand and speaking mainly English and German; we do not know which weighed most [INT].

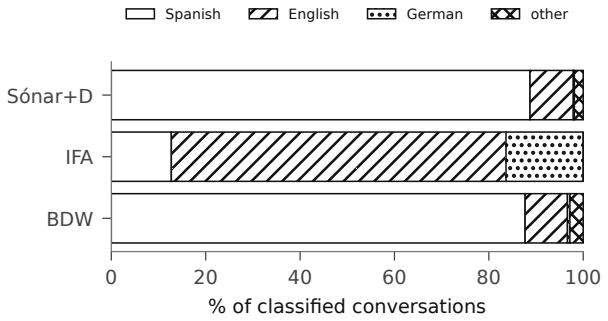



**Figure 1.** Visitor language by event, among classified conversations.

*Who was on the other end.* Each room had its language (figure 1): 89 % of classified conversations were in Spanish at Sónar+D; at IFA, 71 % in English and 16 % in German; in Barcelona, 88 % in Spanish. GAIA switched language mid-call in 11.7 %, 12.8 % and 6.4 % of conversations.

*The room answers back.* At Sónar+D 36 % (95 % CI 33–40 %) of visitor turns were echo, against 4 % at IFA and 10 % in Barcelona. Echo inflates what the visitor appears to say (16.4 words per turn at Sónar+D against 7.5 in Barcelona) and makes the agent answer itself. In Barcelona we fought it with a noise gate and by lowering the microphone while GAIA speaks [MED]. We did not document the loudspeaker or volume at Sónar+D, so we cannot assign the cause there.

*When the audience tests the machine.* At Sónar+D 9.7 % (95 % CI 7.7–12.1 %) of calls were cut by one of the platform's safety guardrails: 53 for sexual content, 7 for self-harm, 4 for offensive language and 3 for violence [MED]. IFA had none. The logs do not say whether the visitor or the agent triggered it.

## 4. The body the model does not use

In Barcelona the model used at least one body tool in 5.2 % (95 % CI 3.6–7.4 %) of calls; in those lasting a

**Table 1.** Three rooms. Outcomes by event.

| Event | n | Median (s) | Short | Mute | Limit | Safety | Abnormal | Echo |
|---|---|---|---|---|---|---|---|---|
| Sónar+D | 701 | 58 [47; 66] | 27 % [24; 30] | 28 % [25; 31] | 11 % | 9.7 % | 7.4 % | 36 % |
| IFA | 274 | 14 [9; 18] | 46 % [40; 52] | 49 % [43; 55] | 3 % | 0.0 % | 1.1 % | 4 % |
| BDW | 543 | 80 [71; 88] | 17 % [14; 21] | 22 % [19; 26] | 33 % | 0.9 % | 0.7 % | 10 % |

Brackets, 95 % CI (bootstrap for the median, Wilson for proportions). Echo, visitor turns that repeat the previous GAIA turn.

**Table 2.** Tool calls decided by the model, by event.

| Tool | Sónar+D | IFA | BDW |
|---|---|---|---|
| Light | 0 | 2 | 7 |
| Display | 0 | 0 | 2 |
| Print | 0 | 1 | 17 |
| Ticket | 0 | 0 | 4 |
| Language switch | 89 | 40 | 39 |
| Hang up | 29 | 3 | 16 |

Number of tool calls decided by the model. What the booth does on its own is not recorded here.

minute or more, 8.7 % (95 % CI 6.1–12.3 %). At IFA, 1.1 %; at Sónar+D, none (table 2). The most used tools were system ones: switching language and hanging up. Among body tools, paper won. Overall, 31 tool results came back with an error [MED].

Little was asked and little is what we got: we asked a language model to act as lighting designer as a side task, in the middle of a conversation, with a rule telling it to hold back [INT]. What made the booth seem alive was decided by the booth: a light that follows the mood of each visitor sentence with local rules, a display that murmurs, the line tone, the paper that comes out on hang-up. None of that is in the conversation logs, and it is the first thing we would add to the instrumentation.

The literature on embodied agents points the same way: in a review of 33 experiments, robots sharing the space with the person were rated better than the same agents on a screen, and physical presence mattered more than the body itself [5].

## 5. What people did

Among the 1,067 conversations in which the visitor spoke, table 3 records six behaviours. 28 % (95 % CI 26–31 %) opened with a greeting of their own, the first sign that they accept the fiction of a call. 5 % (95 % CI 4–6 %) asked whether GAIA was real, a robot or who it was; those calls lasted a median of 125 s against 89 s, which we read as a sign of interest and not a cause, since a longer call leaves more room for the question [INT].

People are kind to machines that do not need it [6]. 8 % (95 % CI 7–10 %) said thank you, more at Sónar+D (11 %) than at IFA (4 %); it is the only behaviour that differs between events after correction ($p = 0.01$). Laughter barely reaches the transcript (1 %): it is the most visible behaviour at the booth and the least visible in the data.

*Short answers to a curious machine.* 24 % of visitor turns had one or two words, more in Barcelona (32 %) than at Sónar+D (15 %). Barcelona was also where

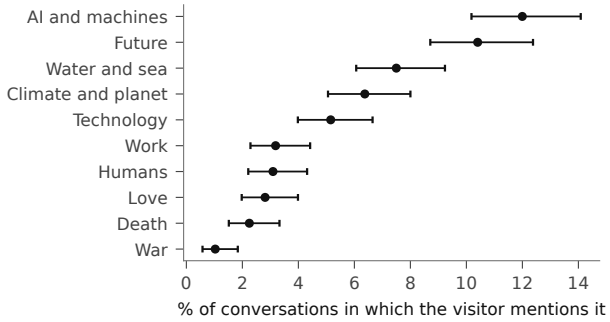



**Figure 2.** Topics mentioned by the visitor, with Wilson 95 % CI. Lexical detection; these are lower bounds.

the agent asked most (DM-2026-05). The more the machine interviews, the shorter people answer [INT].

*What people wanted to talk about.* In 35 % of conversations the visitor touched one of the ten topics we track (figure 2). First machines (12 % (95 % CI 10–14 %)) and the future (10 %); love (3 %) and death (2 %) were rare. The vendor's sentiment classifier, not validated by us, labelled 66 % of 562 calls as neutral.

## 6. Proposal: meaning for the model, presence for the room

The model decides meaning: promising a printout, a word to be read, a change of light when the conversation turns. The room produces presence: cheap, immediate and always on, on the device, never waiting for the model [HIP]. A turn signal in the hand matters above all to children, who are still learning when a machine has finished speaking [7].

## 7. Threats to validity

- Each event is a single sample of room, audience and configuration; differences are associations.
- Lexical detection: behaviours, languages and topics are lower bounds; echo and recognition blur the visitor side.
- Only what the model decides is logged; what the booth does on its own has no counts.
- Different time limits per event truncate duration differently.

## 8. Recommendations for the community

1. Report echo, mute calls and safety cuts per room: they speak of the room before the character.
2. Log every environment action with its timestamp next to the conversation identifier.

**Table 3.** Visitor behaviours by event, with heterogeneity between events (Cramér's V) and Holm-adjusted $p$.

| Behaviour | Total [95 % CI] | Sónar+D | IFA | BDW | $V$ | Holm $p$ |
|---|---|---|---|---|---|---|
| Greets | 28 % [26; 31] | 25 % | 24 % | 33 % | 0.09 | 0.09 |
| Asks if it is real | 5 % [4; 6] | 6 % | 1 % | 5 % | 0.07 | 0.26 |
| Says thanks | 8 % [7; 10] | 11 % | 4 % | 6 % | 0.11 | 0.01 |
| Says goodbye | 6 % [5; 7] | 7 % | 6 % | 5 % | 0.05 | 0.85 |
| Laughs (in text) | 1 % [0; 1] | 1 % | 1 % | 0 % | 0.04 | 0.85 |
| Swears | 2 % [2; 3] | 3 % | 1 % | 2 % | 0.05 | 0.85 |

Among conversations in which the visitor spoke. Lexical detection in seven languages; each proportion is a lower bound.

**Table 4.** Atmosphere kit. Proposed rows are hypotheses.

| Element | Who triggers it | Status |
|---|---|---|
| Light | The room by mood and turns; the model at turning points | Built |
| Display | The model, words to read; the room, murmurs | Built |
| Paper | The model promises it; the booth prints on hang-up | Built |
| Line sound | The room, on every call | Built |
| Ambient sound | The room, by tension | Proposed [HIP] |
| Turn vibration | The room, in the handset | Proposed [HIP] |

3. Do not delegate presence to the model: ambient behaviour belongs on the device, and instant.
4. Add an hourly observation sheet; what never reaches the transcript, such as laughter or groups, only a person sees.

## 9. Data and code

gaia-2147 dataset [1]; figures from `analysis/analyze.py`. No visitor text was stored.

## 10. Conclusions

The room decided who picked up, in which language, with how much echo and how keen to play. The model, free to use its body, hardly did. A voice character's presence is built outside the model, in the room and on the device.

**Signed.** Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). The room does half the work. Working paper DM-2026-06, 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.

[2] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[3] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[4] 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.

[5] J. Li. The benefit of being physically present. a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015. URL https://napier-repository.worktribe.com/output/3636900.

[6] C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *CHI '94*, pages 72–78, 1994. doi: 10.1145/191666.191703.

[7] J. Festerling and I. Siraj. Anthropomorphizing technology. a conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56:709–738, 2022. URL https://pubmed.ncbi.nlm.nih.gov/34811705/.