GAIA-2147

Data descriptor of 1527 conversations with a voice agent in public spaces

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Working paper · Data descriptor
DM-2026-09, version 1.0 · 30 October 2026

Abstract. Data on voice agents in real use, with the public and in noisy rooms, barely exist outside the companies that run them. We describe gaia-2147: 1,527 conversations and 8,151 turns with per-stage latency metrics, collected between 16 Jun 2026 and 26 Oct 2026 in a telephone booth answered by an artificial character, GAIA 2147, at three public events. They add up to 29.2 h of calls and 82 agent versions.

The set contains only numbers and categories per conversation and per turn: durations, closing reasons, versions, greetings, counts of turns, words, questions, voice tags, tools, lexical reactions and the breakdown of each response into end-of-turn wait, recognition, the model's first token and the first voice byte. It contains no transcript text and no audio.

We follow the structure of datasheets: motivation, composition, collection, preprocessing, technical validation, uses, distribution and maintenance. All 8 technical checks pass. This set underpins papers DM-2026-02 to DM-2026-08.

Keywords: dataset; voice agents; latency; public interaction; datasheet; privacy.

1. Motivation

gaia-2147 was built to study how a voice agent converses with strangers in public spaces: latency, opening, personality, use of the body and visitor behaviour. It was collected by the Divina Machina Research Group on its own artwork, GAIA 2147, without specific external funding. We follow the datasheets proposal of Gebru and colleagues [1].

2. Composition

Each instance is a telephone conversation between a visitor and the agent, or a turn of that conversation (table 1, figure 1). *Tests* conversations took place outside events and are excluded from comparisons. Figure 2 shows duration by event; different time limits truncate each distribution differently.

2.1. Schema

`raw/features.json` has 46 fields per conversation and `raw/turns.json` 14 per turn (tables 2 and 3).

Missing data. The end-to-end metric does not exist at Sónar+D and is missing from part of the cascade turns (table 4). The series' analyses use turns where the metric is present and say so.

3. Collection

GAIA always runs on two voice platforms, ElevenLabs and Vidiv. The set comes from the conversations API of the ElevenLabs agent and was downloaded on 27 October 2026; calls answered by Vidiv are not included. Calls are assigned to each event by local date. Visitors voluntarily picked up a telephone in a public space; there was no selection or recruitment.

4. Preprocessing and privacy

Extraction reads each transcript in memory, computes numbers and categories and discards the text. What reaches disk are durations, counts, times, tool names, versions and the greeting class, which is our own text. Nobody read the conversations to build the set and no visitor sentence is published. Conversation identifiers are platform pseudonyms; there is no voice, name or image of anyone.

Lexical definitions [DEF]. A question is a turn ending in a question mark. Echo flags a visitor turn of four or more words with at least 60 % of them in the agent's previous turn. A thread pick-up flags an agent turn that uses a content word of five or more letters said earlier by the visitor and not used before by the agent. Reactions, topics and languages use fixed lists in seven languages; they are lower bounds.

5. Technical validation

8 of 8 consistency checks pass (table 5). File integrity is fixed with SHA-256 in `MANIFEST.json`: `features.json` starts with `9159aabef8531ef8` (1.4 MB) and `turns.json` with `ac9c51a92762b719` (3.0 MB). Every analysis in the series cites that fingerprint.

5.1. Statistical methods of the series

Proportions with the Wilson interval [2],

$$\hat{p}_{\pm} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \qquad (1)$$

differences of proportions with Newcombe [3], medians with bootstrap [4] and, when a call has several turns, cluster bootstrap [5]. Tests with Mann-Whitney, Fisher, chi-square or Kruskal-Wallis, with Holm correction [6] and an effect size (Cliff's delta [7] or Cramér's V). Durations censored at the time limit, with Kaplan-Meier [8]. Seed 2147 and $B = 2000$ resamples.

6. Uses

Suitable for: field latency of voice agents; design of openings; behavioural measures of characters; tool use by models; aggregate visitor behaviour. Not suitable

Table 1. Composition by event.

Event	Conv.	Turns	Visitor turns	Days	Hours	Versions	Failed
Sónar+D	701	3,732	3,014	5	15.1	36	10.8 %
IFA	274	780	471	5	2.5	14	1.1 %
BDW	543	3,622	2,805	3	11.6	30	1.3 %
Tests	9	17	10	6	0.1	4	11.1 %

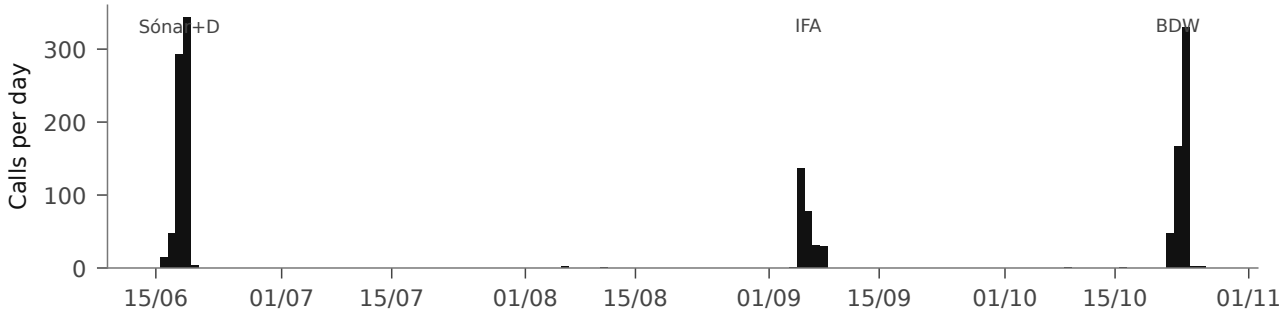


Figure 1. Calls per day in 2026.

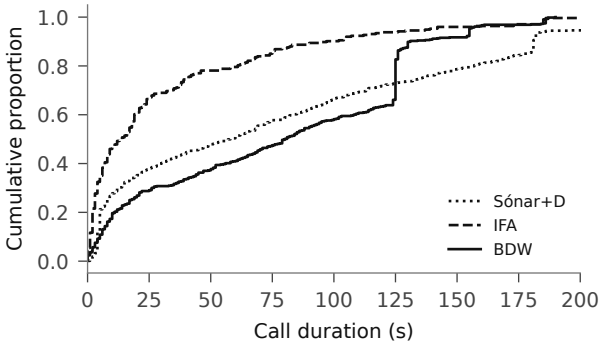


Figure 2. Cumulative distribution of call duration by event.

for: causal inference between events, because event, stack, prompt, audience and room change together; studies of specific people; evaluating comprehension, because the lexical measures are crude; device-side perceived latency, because the metrics are server-side.

7. Distribution and maintenance

The set lives in the group's internal repository (`datasets/gaia-2147`) with its datasheet and manifest. No licence or publication has been decided; any distribution requires Cris Olmedo's approval. New versions will keep the previous one and its fingerprint. Contact: the authors.

8. Data and code

Extraction with `docs/whitepapers/data/extract.py`; figures in this document with `analysis/analyze.py`.

9. Conclusions

gaia-2147 is a small, precise record of something rarely documented: a machine conversing with strangers in public, turn by turn. What it does not keep is as deliberate as what it keeps.

Signed. Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). GAIA-2147. Working paper DM-2026-09, v 1.0. Div-

Table 2. Main fields per conversation.

Field	Type	Definition
`id`, `start`, `evento`, `hora`	text, integer, category	Pseudonymous identifier, start (UTC epoch), event and local hour
`dur`, `status`, `fin`	integer, category	Duration (s), status and closing reason reported by the platform
`version`, `saludo`	text, category	Agent version; class of the opening greeting
`n_ag`, `n_vis`, `pal_ag`, `pal_vis`	integer	Turns and words of each party
`preguntas`, `etiquetas`, `asistente`, `metafora`	integer, count	Lexical agent habits
`eco`, `callback_ok`, `callback_eleg`	integer	Acoustic echo; thread pick-ups and eligible turns
`herr`, `imagenes`, `colores`, `err_herr`	count	Tools decided by the model and errors
`idioma_vis`, `idiomas_det`	category	Visitor language (heuristic) and detected switches
`r_*`	boolean, integer	Lexical visitor reactions: greeting, asking if real, laughter, thanks, goodbye, swearing, questions, short turns, topics
`sent_label`, `exito`, `coste_usd`	category, number	Platform's automatic evaluation and cost

Table 3. Fields per turn. Platform server-side metrics.

Field	Definition
`id`, `i`, `t`	Conversation, transcript index, second
`llm`, `tts_model`	Language and voice model
`e2e`	First audio since detected silence (s)
`silencio`	Configured end-of-turn wait (s)
`asr`	Recognition latency (s)
`llm_ttfb`	Model's first token (s)
`tts_ttfb`	First synthesis byte (s)

ina Machina Research Group. Divina Machina by Qs Ventures.

Table 4. Presence of each metric by group, in response turns.

Group (response turns)	Turns	e2e	Silence	ASR	LLM	TTS
Sónar+D	3,051	0 %	0 %	0 %	95 %	92 %
IFA · Qwen 3.5 397B	427	88 %	88 %	77 %	98 %	88 %
IFA · Gemini 3.7 Flash	48	100 %	100 %	100 %	100 %	100 %
BDW · Eleven Realtime v1 Max	2,404	97 %	97 %	90 %	94 %	98 %
BDW · Eleven Realtime v1 Mini	607	99 %	99 %	91 %	93 %	99 %

Share of response turns with the metric present. Sónar+D did not record the end-to-end metric.

Table 5. Automatic technical checks.

Check	Result
Unique conversation identifiers	passes
Every turn belongs to a conversation in the set	passes
Non-negative durations	passes
Non-negative latencies	passes
Thread pick-ups do not exceed eligible turns	passes
Questions do not exceed GAIA turns	passes
Echo does not exceed visitor turns	passes
No free-text visitor fields	passes

References

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

[2] 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.

[3] R. G. Newcombe. Interval estimation for the difference between independent proportions. comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890, 1998.

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

[5] 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.

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

[7] N. Cliff. Dominance statistics. ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993. doi: 10.1037/0033-2909.114.3.494.

[8] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: 10.1080/01621459.1958.10501452.