La personalidad como vector de hábitos

Deriva de las reglas de tono y memoria dentro de la llamada en un agente de voz

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Estudio observacional
DM-2026-05, versión 1.0 · 30 de octubre de 2026

Resumen. La personalidad de un agente de voz se escribe como una descripción y se encuentra como comportamiento. Proponemos medirla como un vector de tasas de hábitos contables por turno y la seguimos en siete periodos del mismo personaje, GAIA 2147, a lo largo de 82 versiones del prompt.

Una regla de tono añadida en directo redujo las etiquetas de voz sufridas a 0,08 (IC 95 % 0,04–0,12) veces su tasa anterior en pocas horas. Esa misma noche, con la regla aún en el prompt, la tasa volvió a ser 5,27 (IC 95 % 3,28–10,23) veces la de la ventana con la regla recién añadida. Pedir preguntas llevó al agente al 70 % de turnos en pregunta; pedir «más o menos la mitad» lo bajó al 59 %. Retomar palabras del visitante ocurrió en el 36 % (IC 95 % 34–37 %) de los turnos elegibles y no se asoció con cuánto hablaba la gente (Spearman 0,05 (IC 95 % −0,04 a 0,13)).

Las reglas de tono funcionan y se desgastan. Proponemos medir la personalidad tras cada cambio, en horas, y una memoria para personajes públicos que no recuerda a personas. Los periodos difieren también en saludo, hora y público; los efectos sobre los visitantes no se pueden separar aquí.

Palabras clave: personalidad de agentes; prompts; deriva; etiquetas de voz; memoria conversacional; medidas de comportamiento.

1. Introducción

Un personaje se describe con adjetivos. GAIA es amarga, curiosa, culpable, cansada y divertida. Lo que el visitante encuentra es otra cosa: cuánto habla, si pregunta o cuenta, si suspira, si suena a asistente. Esa conducta se puede contar. La industria ha visto que la personalidad de un modelo se mueve sin que nadie lo pida: en 2025 OpenAI retiró una actualización de GPT-4o que se había vuelto complaciente [1], y Anthropic describe los rasgos como direcciones internas del modelo que pueden derivar y vigilarse [2]. A escala de un prompt, la pregunta es más modesta y práctica: ¿qué hace una regla de tono, cuánto dura y cómo se mide?

2. Medida

Definimos la personalidad observable de una versión v como un vector de tasas por turno del agente [DEF]:

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

donde j recorre las conversaciones. r_q es la proporción de turnos que terminan en pregunta; $\bar{w}$, palabras por turno; r_e y r_s, etiquetas de voz por turno, todas y las sufridas (*sigh*, *softly*, *slow*, *sad*, *whisper* y similares); r_a, frases de asistente de una lista fija; r_c, el hilo retomado definido en la sección 5. Las etiquetas son acotaciones que el modelo de voz interpreta [3]. El recuento lo hace el código sobre las transcripciones en memoria; no se guardó texto [4].

Cada tasa es una razón de sumas. Sus intervalos del 95 % se calculan remuestreando conversaciones, no turnos, y los cambios entre periodos se expresan como cocientes de tasas con intervalos por bootstrap de conversaciones en cada grupo [5, 6]. Una regla de tono es un movimiento buscado de $\mathbf{h}$; la deriva es el movimiento que nadie pidió.

Tabla 1. Ventanas de Barcelona Design Week y lo que cambió antes de cada una.

Ventana	Cambio previo
A, hasta el 24 a las 13.53	Prompt de llegada; saludo largo fijo
B, 13.53 a 16.37	Regla de tono: seca, firme, directa; sin suspiros ni susurros; cinco etiquetas neutras permitidas
C, 16.37 a 17.18	Apertura escrita a mano; pruebas de saludo; saludo corto variable
D, 17.18 a 18.41	Sonar humana; más o menos la mitad de los turnos en pregunta
E, desde las 18.41	Cambiar de idioma en el acto si el visitante lo pide

3. Periodos

Comparamos Sónar+D, IFA y cinco ventanas de Barcelona Design Week definidas por cambios del prompt, unidos a los identificadores de versión con nuestro registro de cambios, con precisión de minutos [DEF]:

4. Resultados

4.1. Tres eventos, tres personalidades

La tabla 2 da el vector por periodo [MED]. En Sónar+D, GAIA sonaba a servicio, con 17,2 (IC 95 % 15,6–18,9) frases de asistente cada 100 turnos. En IFA preguntó menos que nunca, en el 19 % (IC 95 % 15–22 %) de sus turnos. En Barcelona el prompt la pedía curiosa

Tabla 2. Vector de hábitos por periodo. Pregunta y hilo como porcentaje; etiquetas por turno; frases de asistente cada 100 turnos.

Periodo	Conv.	Turnos	Preg.	Pal./t.	Etiq./t.	Sufr./t.	Asist./100	Hilo
Sónar+D	681	3452	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	1754	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

Tasas como razón de sumas sobre turnos de GAIA. Los IC 95 % por bootstrap de conversaciones están en el texto y en la figura.

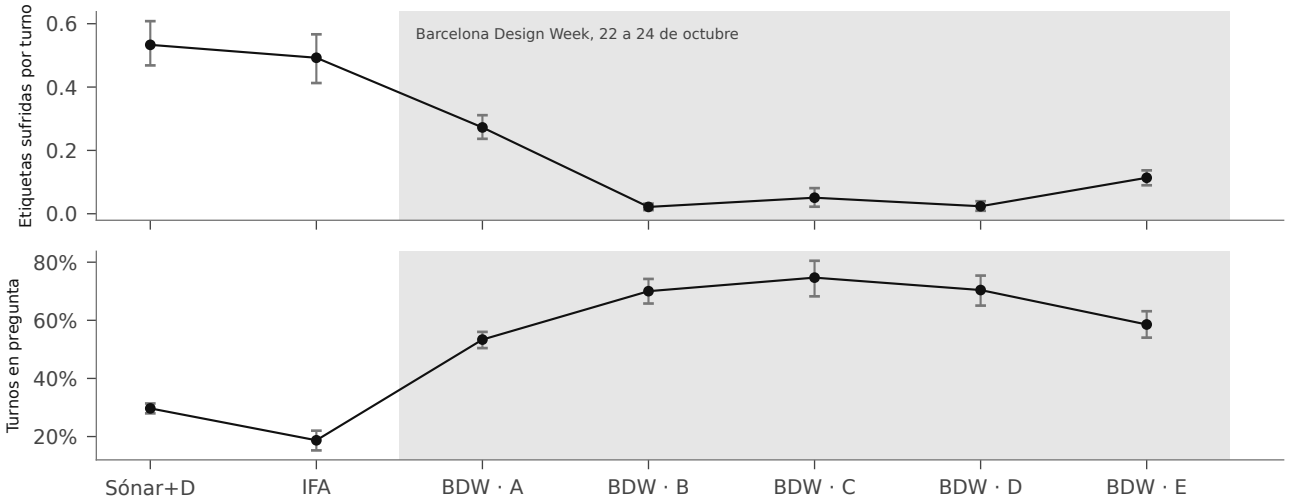


Figura 1. Dos hábitos a lo largo de los siete periodos, con IC 95 % por bootstrap de conversaciones. La banda gris marca Barcelona Design Week.

y, ya en la ventana A, el 53 % (IC 95 % 50–56 %) de sus turnos terminaban en pregunta.

4.2. La regla funciona

Tras la regla de tono, las etiquetas sufridas pasaron de 0,27 (IC 95 % 0,24–0,31) por turno a 0,02 (IC 95 % 0,01–0,03), un cociente de 0,08 (IC 95 % 0,04–0,12). Las etiquetas totales bajaron mucho menos (0,53 (IC 95 % 0,46–0,61)): las neutras ocuparon su lugar. GAIA conservó una voz expresiva y perdió los suspiros (figura 1).

4.3. La regla se desgasta

En la ventana E las etiquetas sufridas volvieron a 0,11 (IC 95 % 0,09–0,14) por turno, 5,27 (IC 95 % 3,28–10,23) veces la tasa de la ventana B. La regla seguía en el prompt; ningún cambio posterior pedía la voz antigua [MED]. Dos explicaciones son compatibles con los datos y no podemos separarlas: cada cambio alargó el prompt y la regla compite con lo que la rodea; y el modelo tiene hábitos propios que vuelven cuando la instrucción pierde peso relativo [INT]. El intervalo es ancho porque la ventana B tiene pocas etiquetas sufridas.

4.4. Preguntar no es escuchar

La petición de preguntar funcionó demasiado: entre la ventana A y la B, la tasa de preguntas se multiplicó por 1,31 (IC 95 % 1,21–1,42) y llegó al 70 %. Pedir «más o menos la mitad» no movió la ventana D (70 %); la E bajó al 59 %, un cociente de 0,83 (IC 95 % 0,75–0,93).

Tabla 3. Resultados de las llamadas por ventana. Confundidos con saludo, hora y público.

Ventana	n	Mediana (s) [IC 95 %]	Corta	Límite
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 %

Las instrucciones de proporción tardan en notarse o necesitan el siguiente cambio para asentarse [INT]. Las frases de asistente pasaron de Sónar+D a la ventana B a 0,16 (IC 95 % 0,08–0,25) veces su tasa, sin llegar a cero; parte de lo que queda es legítimo dentro de una historia.

4.5. ¿Lo notaron los visitantes?

No podemos decirlo (tabla 3). La ventana B tuvo llamadas más cortas que la A y la E, las más largas, con el 53 % llegando al límite de tiempo. Pero el saludo cambió en esas mismas horas, como estudia DM-2026-04, y el público cambia a lo largo del día. El comportamiento se movió con claridad; su efecto sobre los visitantes está confundido.

Tabla 4. Conversaciones largas por tercil de hilo retomado.

Tercil de hilo	n	Hilo	Pal. visitante/t.	Límite
Bajo	187	17 %	8,5	47 %
Medio	187	33 %	9,0	49 %
Alto	187	50 %	9,8	44 %

5. Memoria dentro de la llamada

GAIA no recuerda a nadie entre llamadas. Fue una decisión: una cabina pública recibe niños y desconocidos que nunca aceptaron ser recordados. Tiene tres memorias de alcance distinto: lo dicho en la llamada, el lugar (un ánimo de la ciudad que solo llega si llama a una herramienta, cosa que hizo en 24 de 1527 conversaciones) y ella misma, escrita en el prompt.

Para la primera definimos el hilo retomado. Un turno elegible del agente retoma el hilo si contiene una palabra de contenido de cinco letras o más que el visitante dijo antes y que el agente no había usado [DEF]:

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

En el conjunto, R fue 36 % (IC 95 % 34–37 %). En las 561 conversaciones con al menos cinco turnos elegibles, la mediana por llamada fue 33 % (IC 95 % 33–38 %). La tasa apenas se movió con el prompt, entre el 32 % y el 41 % según el periodo.

Retomar el hilo no se asoció con que el visitante hablara más (Spearman 0,05 (IC 95 % −0,04 a 0,13), $p = 0{,}26$). Por terciles (tabla 4), los visitantes del tercil alto dijeron 9,8 palabras por turno frente a 8,5 en el bajo, y la proporción de llamadas al límite fue parecida. La medida es tosca: repetir una palabra no es entenderla [INT]. Lo que muestra es estrecho: reutilizar mecánicamente lo que dijo la gente no es, por sí solo, lo que la mantiene al teléfono.

5.1. Lo que dice la investigación

Un estudio longitudinal con 24 personas y 10 sesiones encontró que la memoria percibida dependía de la relación construida y actuaba sobre el disfrute posterior a través de la autorrevelación [7] [EMP]. Un estudio de mago de Oz con 30 participantes encontró ventaja para la comprensión del contexto y no para la personalización con memoria a largo plazo [8]. La habilidad conversacional se asocia a mayor humanidad percibida [9]. Los grandes asistentes convirtieron la memoria en producto con controles explícitos [10, 11]; son asistentes personales, no personajes públicos.

6. Propuestas

Memoria para un personaje público [HIP]. Tres capas, ninguna sobre personas: el hilo como hábito explícito del prompt; una memoria anónima de la sala con los temas del día, nunca ligada a una voz ni a un nombre; y un canon fijo de recuerdos del personaje para que no se contradiga entre visitantes.

Experimento propuesto [HIP]. Alternar días u horas con y sin memoria de la sala, con el mismo saludo, pila y límite. Medida principal: proporción de llamadas en que el visitante habla cinco turnos o más. Para detectar un paso del 50 al 60 % con $\alpha = 0{,}05$ y potencia de 0,8 hacen falta

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

385 llamadas por grupo, dos o tres días al ritmo de Barcelona Design Week. La expresión se genera con cálculo simbólico en el análisis y se comprueba contra el cálculo numérico. La figura 2 muestra cuánto crece n si el efecto buscado es menor.

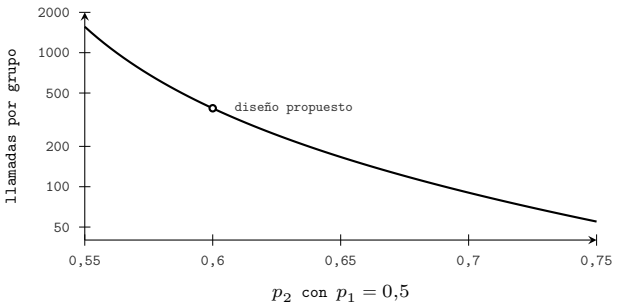


Figura 2. Tamaño por grupo según la ecuación (3), con $\alpha = 0{,}05$ y potencia 0,8. Curva analítica dibujada con PGFPlots.

7. Amenazas a la validez

- Recuentos heurísticos: un signo de interrogación no siempre es una pregunta y una lista fija no recoge paráfrasis.
- Ventanas definidas por hora y versión, unidas a mano con el registro de cambios.
- Varios cambios simultáneos el 24 de octubre: saludo, prompt y público.
- Sin etiquetas de una persona: nadie leyó conversaciones para validar las medidas.

8. Recomendaciones para la comunidad

1. Medir la personalidad como tasas por turno, con intervalos por conversación, después de cada cambio.
2. Volver a medir horas después: el primer resultado tras un cambio no es el estable.
3. Dar al modelo de voz una lista cerrada de etiquetas; una lista abierta vuelve a lo de siempre.
4. Pedir proporciones, no absolutos: «pregunta siempre» produce una entrevista.
5. Separar memoria de la conversación, del lugar y del personaje, y no recordar a personas en espacios públicos.

9. Datos y código

Dataset gaia-2147 [4]; cifras de `analysis/analyze.py`. Siempre trabajamos con ElevenLabs y Vidiv; los datos son del agente de ElevenLabs.

10. Conclusiones

Una personalidad escrita en un prompt es una hipótesis sobre una conducta. Contada turno a turno, la regla de tono funcionó en horas y se desgastó en la misma tarde. La personalidad no es un ajuste que se escribe una vez: se mide, como la temperatura de una sala.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). La personalidad como vector de hábitos. Documento de trabajo DM-2026-05, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 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. Etiquetas de audio y voz expresiva.

[4] C. Olmedo and K. Sommer. GAIA-2147. descriptor de datos de 1527 conversaciones con un agente de voz en espacios públicos. Documento de trabajo DM-2026-09. Grupo de investigación Divina Machina, 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. Comprensión del contexto y memoria a largo plazo en agentes conversacionales. Tesis, repositorio POASIS, s. f. 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/. Actualización de abril de 2025.

[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.