Elevam

Estudio realizado por Elevam.es y Antropus.io

# Apéndice de metodología y fuentes

## Umbrales de corroboración en motores con IA y el comportamiento del español

Elevam Labs · Mayo de 2026

Documento de respaldo del artículo «Te están vendiendo GEO a ciegas»

## Nota de alcance

Este documento respalda el artículo «Te están vendiendo GEO a ciegas». Recoge la evidencia publicada sobre la que se construye su tesis y declara, por delante, su límite principal: a día de hoy no existe ninguna investigación publicada con metodología transparente que cuantifique cuántos dominios independientes necesita una marca para ser citada de forma estable por un motor con IA, ni que compare ese umbral entre el inglés y el español.

Lo que sigue es un prior razonado construido a partir de evidencia adyacente, no una pistola humeante. Cualquier afirmación de este apéndice debe leerse con ese marco. Donde una fuente es comercial o de metodología parcial, se indica de forma explícita.

## 1. Qué sabemos con rigor

### 1.1 Las recomendaciones de la IA son inconsistentes — Fishkin / SparkToro

La pieza más rigurosa disponible hoy es la investigación de Rand Fishkin y Patrick O'Donnell, construida sobre cerca de 3.000 ejecuciones de los mismos prompts en varios motores. Dos cifras resumen el hallazgo: la probabilidad de que dos consultas idénticas devuelvan la misma lista de marcas es inferior a 1 entre 100, y la de que la devuelvan en el mismo orden, inferior a 1 entre 1.000.

La conclusión operativa es que la única métrica defendible es la visibilidad porcentual — la frecuencia con que una marca aparece a lo largo de muchas ejecuciones — y no la «posición de ranking», que el propio Fishkin descarta sin matices. Límite: el estudio no es peer-reviewed, algo que su autor declara expresamente; toma como base metodológica un trabajo de Carnegie Mellon, pero la implementación es propia.

*Fuente: SparkToro, enero de 2026 — «AIs are highly inconsistent when recommending brands or products». sparktoro.com*

### 1.2 El vínculo entre rankear en Google y ser citado por la IA se ha debilitado — Ahrefs

Según Ahrefs (marzo de 2026), solo el 38% de las URLs citadas en los AI Overviews rankean en el top-10 orgánico para esa misma consulta, frente al 76% un año antes. La correlación medida entre posición orgánica y citación es

moderada (Pearson ≈ 0,35). Implicación: el ranking tradicional ya no es un proxy suficiente de la visibilidad en IA; entran en juego la expansión de consultas, la presencia en YouTube o Reddit y la corroboración entre dominios.

*Fuente: Ahrefs Blog, marzo de 2026 — «38% of AI Overview citations pull from the top 10». ahrefs.com*

## 1.3 Cada motor bebe de un ecosistema de fuentes distinto — Profound

El análisis de Profound (680 millones de citas, agosto de 2024 a junio de 2025) sitúa el solapamiento entre los dominios que cita ChatGPT y los que cita Perplexity en apenas un 11%. Wikipedia domina en ChatGPT; Reddit y YouTube, en Perplexity. No existe, por tanto, «la IA» como un único destino: hay varios ecosistemas con reglas propias, y una estrategia que trate la visibilidad en IA como un solo canal ignora cuatro de cada cinco superficies.

*Fuente: Profound — «AI Platform Citation Patterns». Estudio comercial, metodología parcialmente transparente; tratar como secundaria sólida.*

## 1.4 El único paper académico peer-reviewed — Princeton et al., KDD 2024

Aggarwal et al. (Princeton, IIT Delhi, Georgia Tech y Allen Institute for AI) presentaron en el congreso KDD de 2024 el único trabajo académico revisado por pares relevante. Sobre un benchmark propio de 10.000 consultas, demuestran que añadir estadísticas, citar fuentes e incorporar citas textuales mejora la visibilidad entre un 30% y un 40%, mientras que el keyword stuffing tradicional no funciona en motores generativos. Límite: el «motor generativo» del estudio es un sistema simulado, no ChatGPT, Gemini o Claude reales, y no existe versión multilingüe del benchmark.

*Fuente: Aggarwal et al. (2024), «GEO: Generative Engine Optimization», KDD '24. DOI 10.1145/3637528.3671900 · arXiv:2311.09735*

## 1.5 Cuántas fuentes cita cada motor (indicativo)

El número de fuentes por respuesta varía mucho según el motor, lo que define cuántos «huecos» hay disponibles para ser citado y, por tanto, el grado de competencia por aparecer. Las cifras de abajo son indicativas, proceden en parte de estudios comerciales y tienen una caducidad estimada de seis meses o menos: los modelos cambian con rapidez.

| Motor | Citas medias/resp. | Backend principal | Notas |
|---|---|---|---|
| ChatGPT (search) | ~4 a 15 | Bing + crawl propio | Cita Wikipedia y directorios; menciona marcas más de lo que las enlaza |
| Perplexity | ~7 a 22 | Bing híbrido + índice propio | El más denso y el más sensible a la frescura del contenido |
| Google AI Overviews | ~8 a 11 | Índice Google + Knowledge Graph | Solo 38% de citas rankean en top-10 (mar. 2026) |
| Claude (search) | ~6 | Brave Search | Mezcla diversificada de fuentes |

*Fuentes: compilación de Profound, Ahrefs, Whitehat SEO y otros (varias comerciales). Indicativo, no medición propia.*

# 2. El comportamiento del español

## 2.1 El corpus de entrenamiento está sesgado hacia el inglés — Common Crawl

Buena parte del entrenamiento de estos modelos se nutre de Common Crawl, cuyo propio organismo reconoce por escrito que sus datos «siempre han estado sesgados hacia el contenido en inglés». Las cifras lo confirman: entre el 40% y el 45% del corpus es inglés, frente a alrededor del 5% en español. El idioma en el que preguntas ocupa una fracción mínima del tablero.

*Fuente: Common Crawl Foundation — estadísticas de idioma y blog corporativo. commoncrawl.org*

## 2.2 El idioma del prompt cambia la salida del modelo — Walker & Timoneda (Cambridge)

El trabajo de Christina Walker y Joan Timoneda (Universidad de Purdue), publicado en Political Science Research and Methods de Cambridge University Press en 2025, es la mejor demostración peer-reviewed de que el idioma del prompt no es un mero formato. Con el mismo prompt traducido a varios idiomas, la salida de GPT cambia de signo de forma sistemática — más conservadora en idiomas de sociedades conservadoras, más liberal en las liberales — y esa diferencia persiste de GPT-3.5 a GPT-4. Los autores lo atribuyen directamente a la composición del corpus de entrenamiento por idioma.

Para nuestro propósito, la lectura es clara: el idioma es un canal de recuperación distinto, con un corpus distinto detrás. Las dinámicas de citación pueden, por tanto, comportarse de forma diferente. (Las cifras concretas para el par catalán/español que circulan en algunas síntesis son probabilidades derivadas de los modelos del paper, no citas literales; se omiten aquí por prudencia.)

*Fuente: Walker, C. P. & Timoneda, J. C. (2025), Political Science Research and Methods, Cambridge University Press. DOI 10.1017/psrm.2025.10057*

## 2.3 Los modelos alucinan más en idiomas con menos datos

Varios trabajos documentan tasas de alucinación más altas en idiomas con menos representación en el entrenamiento, el español incluido respecto al inglés. Implicación incómoda para la hipótesis de «umbral más bajo»: aparecer con poca corroboración puede significar aparecer mal — atribuido a un competidor, o de forma inestable de un día para otro.

*Fuente: «Multilingual Hallucination Gaps in Large Language Models», arXiv:2410.18270, entre otros.*

## 2.4 El «problema del español global»

Gianluca Fiorelli ha documentado que los motores no distinguen de forma fiable entre el español de España, México o Argentina, y mezclan en una misma respuesta terminología regulatoria y comercial de varios mercados. Consecuencia para la hipótesis: una consulta en «español genérico» no compite en el mercado local, sino contra todo el espacio hispanohablante a la vez. La unidad de medida útil no es «español vs. inglés», sino «español-de-un-mercado-concreto vs. inglés».

*Fuente: G. Fiorelli, columnas sobre AI search en mercados multilingües, Search Engine Land (2026).*

### 2.5 Las entidades poco frecuentes son más difíciles de enlazar

El trabajo sobre enlazado de entidades long-tail confirma que las entidades poco frecuentes, infrarrepresentadas en el entrenamiento, son más difíciles de enlazar correctamente. Una marca local hispana es, globalmente, una entidad long-tail. Esto matiza la hipótesis en su versión ingenua: no es que sea más fácil de citar, sino que solo lo será si su densidad de corroboración local supera el ruido.

*Fuente: «Evaluation of LLMs on Long-tail Entity Linking», arXiv:2505.03473.*

## 3. Las dos direcciones de la evidencia

La hipótesis a contrastar es que en español hace falta un umbral más bajo de dominios independientes para que un modelo cite a una marca de forma estable. La evidencia adyacente no es unívoca: apunta en ambas direcciones según el nicho y el motor. Un análisis honesto pone las dos sobre la mesa.

### 3.1 A favor (umbral más bajo)

- Universo pequeño: Fishkin muestra que en nichos con pocos candidatos la varianza cae y las marcas líderes alcanzan visibilidades del 90% y pico.
- Medios concentrados: en los mercados hispanos un puñado de dominios canónicos satura antes la cuota de citación.
- Walker & Timoneda: cambiar el idioma cambia la dinámica; el corpus español es un canal propio.

### 3.2 En contra (umbral igual o más alto)

- El «español global» agranda el mercado efectivo: se compite contra toda Hispanoamérica más España a la vez.
- Undersampling: los crawlers visitan menos las páginas no inglesas, así que cada dominio español pesa menos en el modelo.
- Fallback al inglés cuando la densidad local es baja.
- Mayor tasa de alucinación: la citación puede ser ruidosa o mal atribuida.
- Interferencia de marcas globales en nichos colonizados (Moz, Ahrefs, Semrush en el sector SEO/GEO).

### 3.3 Predicción honesta, por nicho

Lo que sigue es un prior razonado, no una medición:

- Nichos acotados con poca interferencia global (penalista en una ciudad, concesionario en Mallorca): la hipótesis es probablemente cierta. Estimación de prior: 3 a 5 dominios autoritativos hispanos podrían bastar, frente a 7 a 10 en equivalentes ingleses competidos.
- Nichos colonizados por marcas globales (SEO/GEO, software B2B): probablemente igual o más alto, por la mezcla de corroboración hispana y anglosajona.
- Consultas ambiguas en español genérico: comportamiento errático y umbrales inestables.

## 4. Diseño del experimento de Elevam Labs

Para resolver lo que la literatura no resuelve, Elevam Labs está ejecutando un experimento propio. Su diseño:

Objetivo. Medir el umbral de corroboración en español y compararlo con el inglés.

Métrica principal. Visibilidad porcentual a lo largo de muchas ejecuciones (Fishkin); nunca «posición de ranking».

Motores. ChatGPT con búsqueda, Perplexity, Google AI Mode y Claude, reportados por separado: cada uno tiene corpus y tasa de citación distintos.

Muestra. Dos o tres nichos (penalista, concesionario, agencia GEO) más un brazo en inglés; 8 a 9 prompts por nicho, congelados antes de empezar; un mínimo de 30 ejecuciones por prompt y motor.

Dos fases. Fase 1, consistencia de aparición en cada motor. Fase 2, huella de corroboración web de cada marca: en cuántos dominios independientes se la menciona. Se cruza el input (huella) con el output (consistencia) para localizar el umbral.

Controles. Sesiones limpias sin memoria ni personalización, geolocalización fija, ventana de ejecución de 72 horas o menos, y definiciones fijadas a priori de «dominio independiente» y de «consistencia» (aparición en el 80% o más de las ejecuciones).

Publicación. La metodología y los límites se publicarán por delante de las conclusiones, con el conjunto de datos disponible. Solo se afirmará correlación, nunca causalidad.

# 5. Caveats y límites declarados

1. No existe literatura directa que mida umbrales de corroboración; todo el ejercicio es construir un prior a partir de evidencia adyacente.
2. Buena parte de la evidencia de mercado (Profound y similares) es comercial y de metodología parcialmente transparente.
3. El estudio de Fishkin no es peer-reviewed, algo que su propio autor declara.
4. El paper de Princeton usa un motor generativo simulado, no los modelos comerciales reales.
5. Los modelos cambian rápido; cualquier cifra de tasa de citación tiene una caducidad estimada de seis meses o menos.
6. No se ha encontrado ningún estudio publicado que compare la concentración de dominios hispanos frente a anglosajones por nicho. Es un hueco de evidencia real.

# 6. Clasificación de fuentes

| Fuente | Tipo | Nivel de rigor |
|---|---|---|
| Walker & Timoneda (PSRM, Cambridge) | Académica peer-reviewed | Alto (primaria) |
| Princeton et al. (KDD 2024) | Académica peer-reviewed | Alto (primaria; motor simulado) |
| Common Crawl Foundation | Datos del organismo fuente | Alto (primaria) |
| SparkToro / Fishkin | Investigación propia no peer-reviewed | Medio-alto (mejor dato disponible) |
| Ahrefs (posts con metodología) | Estudio comercial detallado | Medio-alto |
| Profound | Estudio comercial | Medio (secundaria) |
| Fiorelli / Search Engine Land | Análisis sectorial firmado | Medio (observacional) |
| Whitehat, Outpace, ZipTie, Onely | Blogs comerciales | Bajo (secundaria débil) |