TOPOLOGICAL SEO AND GEO

SEO TOPOLÓGICO Y GEO

EMPIRICAL VALIDATION OF STRUCTURAL EFFICIENCY AND THE TOPOLOGICAL GROUNDING SCORE FOR GENERATIVE ENGINE OPTIMIZATION

VALIDACIÓN EMPÍRICA DE LA EFICIENCIA ESTRUCTURAL Y EL ÍNDICE DE ANCLAJE TOPOLÓGICO PARA LA OPTIMIZACIÓN EN MOTORES GENERATIVOS

Jorge Sánchez Mosquete



Jorge Sánchez Mosquete

Topological SEO and GEO

SEO Topológico y GEO

Draft: February 05, 2026

Visibility in the age of AI is not derived from authority, but from structural clarity. In an ecosystem where engines no longer seek links but understanding, those who architect the topology of knowledge will not just navigate the paradigm shift—they will become the source of truth for the new digital intelligence."

“La visibilidad en la era de la IA no se determina por la autoridad, sino por la claridad estructural. En un ecosistema donde los motores ya no buscan enlaces sino entendimiento, quien arquitecte la topología del conocimiento no sólo prevalecerá en el cambio de paradigma, sino que se convertirá en la fuente de verdad de la nueva inteligencia digital.”

- Jorge Sánchez Mosquete

Autor:

Jorge Sánchez Mosquete
SEO/GEO Senior Consultant & CEO at Netbulb Marketing Agency
Head of SEO at Rommel & Montgomery

Email: jorge@jorgesanmos
Linkedin: in/jorgesanmos

Abstract

The transition from Information Retrieval (IR) systems to Retrieval-Augmented Generation (RAG) paradigms constitutes a fundamental shift in digital visibility structures. Traditional heuristics, such as PageRank and Domain Authority, demonstrate a diminishing predictive correlation in ecosystems dominated by Large Language Models (LLMs). This study evaluates the 100 most trafficked domains in Spain, resulting in a refined sample of 91 entities (N=91) after applying rigorous exclusion criteria to mitigate intent-based noise.

To address the limitations of existing metrics, this paper introduces the Topological Grounding Score (TGS). This framework integrates harmonic centrality (HC Rank) with semantic density to quantify a domain's grounding capacity—its structural aptitude to function as a factual reference for neural inference. Logarithmic regression analysis indicates that internal graph topology predicts AI visibility with superior statistical significance (r = -0.45 post-sector clustering) compared to external authority metrics (r = 0.43).

These findings suggest that RAG systems prioritize structural efficiency and semantic clustering over raw link popularity. Consequently, this research posits a strategic shift from link-acquisition models toward structural hub-optimization to maximize performance within Generative Engine Optimization (GEO) frameworks.

Resumen

La transición de los sistemas de Recuperación de Información (IR) hacia los paradigmas de Generación Aumentada por Recuperación (RAG) representa un cambio estructural en los modelos de visibilidad digital. Las heurísticas tradicionales, como el PageRank y la Autoridad de Dominio, muestran una correlación predictiva decreciente en ecosistemas gobernados por Grandes Modelos de Lenguaje (LLMs). El presente estudio analiza los 100 dominios con mayor tráfico en España, consolidando una muestra final de 91 entidades (N=91) tras la aplicación de criterios de exclusión rigurosos para mitigar el ruido de intención.

Con el objetivo de subsanar las limitaciones de las métricas vigentes, este artículo introduce el índice de anclaje topológico o Topological Grounding Score (TGS). Este marco metodológico integra la centralidad armónica (HC Rank) con la densidad semántica para cuantificar la capacidad de anclaje (grounding) de un dominio, definida como su aptitud estructural para servir como referencia factual en la inferencia de redes neuronales. El análisis de regresión Logarítmica revela que la topología del grafo interno predice la visibilidad en IA con una significación estadística superior (r = -0.45 tras segmentación por sectores) en comparación con las métricas de autoridad externa (r = 0.43).

Los hallazgos sugieren que los sistemas RAG priorizan la eficiencia estructural y la cohesión semántica sobre la popularidad bruta de los enlaces. En consecuencia, esta investigación postula un desplazamiento estratégico de los modelos de adquisición de enlaces hacia la optimización de nodos estructurales (hub building) para maximizar el rendimiento en el ámbito de la Optimización de Motores Generativos (GEO).

JEL Classification:

L86, M31, D83, C45.

Keywords:

English: Generative Engine Optimization, GEO, Retrieval-Augmented Generation, RAG, Topological Grounding Score, TGS, Graph Theory , Harmonic Centrality , Knowledge Graphs, Predictive SEO , Information Architecture, Neural Inference, Semantic Density.

Español: Optimización para Motores Generativos, GEO, Generación Aumentada por Recuperación, RAG, Índice de Anclaje Topológico, TGS, Teoría de Grafos , centralidad armónica , Grafos de Conocimiento, SEO Predictivo , Arquitectura de la Información, Inferencia Neuronal, Densidad Semántica.

ÍNDICE

1. Introducción....5
1.1. Transición de los modelos de recuperación de información....6
1.2. Relación con la investigación previa en modelado predictivo....7
1.3. Divergencia entre autoridad nominal y relevancia en sistemas de IA....7
1.4. Objetivos del estudio....8
2. Fundamentos de la autoridad externa a la eficiencia topológica....8
2.1. Insuficiencia del grafo de enlaces en arquitecturas RAG....9
2.2. Teoría de grafos y evolución de las métricas de centralidad....10
2.3. Validación matemática y el factor Common Crawl....11
2.4. La necesidad de una métrica compuesta....12
3. Metodología y diseño experimental....13
3.1. Arquitectura de la investigación....13
3.2. Selección de la muestra....13
3.3. Protocolo de adquisición de datos y marco computacional....14
3.4. Operacionalización de variables....14
3.5. Hipótesis estadísticas....15
4. Análisis computacional y preprocesamiento....17
4.1. Transformación logarítmica y normalización de la varianza....17
4.2. Matriz de correlación de Pearson....18
4.3. Algoritmo de análisis (Python Code)....19
5. Resultados Empíricos y análisis descriptivo....21
5.1. Jerarquía de visibilidad. Wikipedia como benchmark topológico....21
5.2. Análisis de dispersión. El desacoplamiento entre volumen y visibilidad....22
5.3. Segmentación de comportamiento y clustering....23
Tipología de los arquetipos identificados....24
5.4. Análisis de sensibilidad y umbral de inflexión (tipping point)....25
Interpretación de la curva de respuesta....26
5.5. Modelado predictivo dinámico: La dimensión temporal....26
Extensiones econométricas. TGS como regresor exógeno (ARIMAX)....26
Arquitecturas de aprendizaje profundo (LSTM) para la previsión del ROI....27
6. El Índice de Anclaje Topológico (TGS)....28
6.1. Definición y formulación matemática....28
6.2. Ranking de eficiencia TGS (top 15 estratégico)....29
6.3. Análisis de anomalías y estrategias divergentes....31
La penalización por compartimentación....31
La entropía de las facetas....32
7. Hacia el Generative Engine Optimization (GEO)....32

7.1. Pilar 1. Optimización topológica....33
7.2. Pilar 2. Segregación de intención....34
7.3. Pilar 3. Datos para el razonamiento....35
Conclusiones....37
Supremacía de la eficiencia topológica sobre el volumen....37
El TGS como estándar de medición post-autoridad....37
Convergencia hacia el GEO y la web agentic....38
Referencias....39
Anexos documental....41
Anexo A: Tabla de datos (N=91)....41
Anexo B: Algoritmos computacionales y reproducibilidad....44
B.1. Preprocesamiento y cálculo del TGS (ETL)....44
B.2. Segmentación de mercado (K-Means clustering)....45
B.3. Modelado predictivo y el umbral de intención....46
Anexo C: Escenarios de simulación predictiva (impacto GEO)....48
C.1. Tabla de proyección de impacto....48
C.2. Interpretación estratégica de los escenarios....49
C.3. Validación del ROI....49

1. Introducción

La arquitectura fundamental de los sistemas de recuperación de información (IR) atraviesa una transición estructural desde la implementación de los algoritmos de clasificación basados en grafos de citas externas (Page et al., 1999).

Durante más de dos décadas, la autoridad digital se ha determinado mediante un sistema de ponderación fundamentado en la acumulación de hipervínculos, bajo la premisa técnica de que la popularidad externa es un indicador fiable de la relevancia del contenido. Sin embargo, la emergencia de los LLMs y la estandarización de las arquitecturas de Generación Aumentada por Recuperación (RAG) han alterado los mecanismos de evaluación de relevancia (Lewis et al., 2020).

El ecosistema digital actual experimenta un desplazamiento del paradigma de búsqueda sintáctica, caracterizado por la coincidencia léxica y la autoridad de dominio, hacia un modelo de búsqueda semántica inferencial. En este contexto, el presente estudio constituye una validación empírica del marco de trabajo denominado Generative Engine Optimization (GEO) y propone el concepto de SEO topológico como métrica de eficiencia.

La investigación se fundamenta en el análisis de un conjunto de datos (N=91) de alto tráfico en el mercado español. Para ello, se emplean algoritmos de teoría de grafos procesados en Python, integrando modelos predictivos estocásticos (ARIMA) y redes neuronales de memoria a corto y largo plazo (LSTM).

El objetivo central es definir una metodología que permita la transición desde modelos de adquisición de autoridad externa hacia estrategias de optimización de la eficiencia estructural (hub building). Se postula que la relevancia en sistemas de inteligencia artificial generativa no depende de la autoridad nominal, sino de la capacidad de un dominio para minimizar la entropía informativa y actuar como una fuente de verdad estructuralmente coherente.

1.1. Transición de los modelos de recuperación de información

El análisis de la economía digital en el periodo 1998-2023 revela un paradigma de recuperación de información de carácter determinista. Bajo este modelo, el procesamiento de consultas se fundamentaba en la coincidencia sintáctica, donde los sistemas devolvían una jerarquía de enlaces indexados en páginas de resultados (SERP) basada en la autoridad de los documentos.

La integración de la inteligencia artificial generativa ha alterado la función de utilidad del usuario final. Se observa un desplazamiento desde la localización de documentos hacia la demanda de respuestas sintetizadas y directas, fenómeno cuantificado en la industria como comportamiento de clic cero o zero-click. Para satisfacer este requerimiento, los agentes conversacionales han abandonado el procesamiento secuencial de la web en favor de una interpretación basada en grafos de conocimiento, donde las entidades se encuentran interconectadas mediante representaciones vectoriales (Lewis et al., 2020).

Este cambio técnico introduce implicaciones críticas para la visibilidad digital. En los entornos de RAG, el sistema debe ejecutar dos procesos secuenciales como la recuperación de segmentos de información o chunks mediante similitud vectorial y la posterior síntesis de una respuesta coherente.

Los datos sugieren que las arquitecturas de información deficientes o fragmentadas limitan la capacidad de los modelos para conectar los nodos semánticos necesarios. En consecuencia, la visibilidad de un dominio en este nuevo ecosistema deja de ser una función exclusiva de la acumulación de hipervínculos externos denominados backlinks y pasa a depender de su eficiencia estructural. Esta capacidad, denominada anclaje o grounding, define la aptitud de un dominio para servir como fuente de verificación fiable para la inferencia de la red neuronal.

1.2. Relación con la investigación previa en modelado predictivo

El presente estudio se desarrolla sobre la base de investigaciones previas relativas al modelado estocástico y la optimización predictiva en el ecosistema del SEO o posicionamiento orgánico (Sánchez-Mosquete, 2025). Mientras que los análisis precedentes se centraron en la dimensión temporal, empleando modelos ARIMA y redes neuronales LSTM para la predicción de volatilidad y tendencias de demanda, la presente investigación aborda la dimensión topológica (la arquitectura espacial del grafo).

Se asume la premisa de que la capacidad de anticipación algorítmica (Page et al., 1999) resulta insuficiente si no se integra en una arquitectura de información legible para los LLMs. En este marco, el TGS se propone como el estabilizador estructural necesario para dotar de consistencia a las estrategias predictivas en entornos de inteligencia artificial.

1.3. Divergencia entre autoridad nominal y relevancia en sistemas de IA

La motivación de este estudio subyace en la identificación de una anomalía estadística en los datos de rendimiento de entornos de producción. Al monitorizar las respuestas generadas por sistemas de búsqueda generativa para consultas de alta complejidad, se observa un fenómeno divergente.

Entidades con una elevada cuota de mercado y perfiles de autoridad externa consolidados son omitidas por los modelos de lenguaje en favor de dominios con menor autoridad nominal, pero mayor especialización estructural.

Bajo el postulado de que la función objetivo de un LLM consiste en la minimización de la entropía para la generación de respuestas veraces, los modelos demuestran una mayor afinidad de citación hacia estructuras densas y cohesivas.

Se formula la hipótesis central de este trabajo (H_1) los modelos de lenguaje evalúan la topología de la red interna, definida como la eficiencia matemática en la conexión entre nodos de información, y la utilizan como un indicador (proxy) de confianza superior a la reputación externa. En consecuencia, una arquitectura de información fragmentada incrementa el

coste energético de inferencia para el modelo, reduciendo significativamente la probabilidad de selección del dominio como fuente de anclaje.

1.4. Objetivos del estudio

La presente investigación se propone alcanzar tres objetivos fundamentales orientados a la redefinición de las métricas de rendimiento sectorial:

1. **Validación estadística:** Evaluar mediante análisis de regresión si la centralidad armónica presenta una correlación superior con la visibilidad en sistemas de IA en comparación con el volumen agregado de hipervínculos externos, contrastando la relevancia de la estructura frente a la popularidad.
2. **Cuantificación del umbral de intención:** Determinar los niveles de densidad informativa necesarios para garantizar la elegibilidad en arquitecturas RAG, identificando el punto de inflexión donde la masa crítica de contenido genera un impacto estadísticamente significativo en la visibilidad.
3. **Estandarización del marco GEO:** Proponer la implementación del TGS como un estándar métrico para la evaluación de la integridad de un dominio en el paradigma generativo, proporcionando una medida unificada de eficiencia estructural y capacidad de recuperación.

2. Fundamentos de la autoridad externa a la eficiencia topológica

La validación de la metodología propuesta requiere un análisis de la disonancia teórica entre los sistemas de recuperación clásicos y los motores generativos contemporáneos. El marco del SEO topológico postula la sustitución de las métricas de centralidad de autovector (eigenvector centrality), tales como el PageRank, por métricas de distancia, específicamente la centralidad armónica. Esta transición técnica responde a la necesidad de alinear la estructura del dominio con la mecánica operativa de los sistemas RAG.

2.1. Insuficiencia del grafo de enlaces en arquitecturas RAG

El algoritmo PageRank ha constituido el pilar fundamental de la recuperación de información tradicional, operando bajo la premisa de que un hipervínculo representa una transferencia de autoridad o voto de confianza.

Desde una perspectiva matemática, este modelo cuantifica la probabilidad de que un agente estocástico, en una navegación infinita, alcance un nodo específico. Sin embargo, los sistemas de generación aumentada por recuperación no operan mediante navegación aleatoria, sino a través de mecanismos de recuperación densa.

El proceso RAG se articula en dos fases críticas, siendo la primera fase la recuperación densa (dense retrieval), donde el sistema identifica segmentos de texto semánticamente próximos a la consulta del usuario dentro de un espacio vectorial multidimensional y la fase de generación, en la cual el modelo sintetiza una respuesta coherente utilizando dichos segmentos como contexto de anclaje (Lewis et al., 2020).

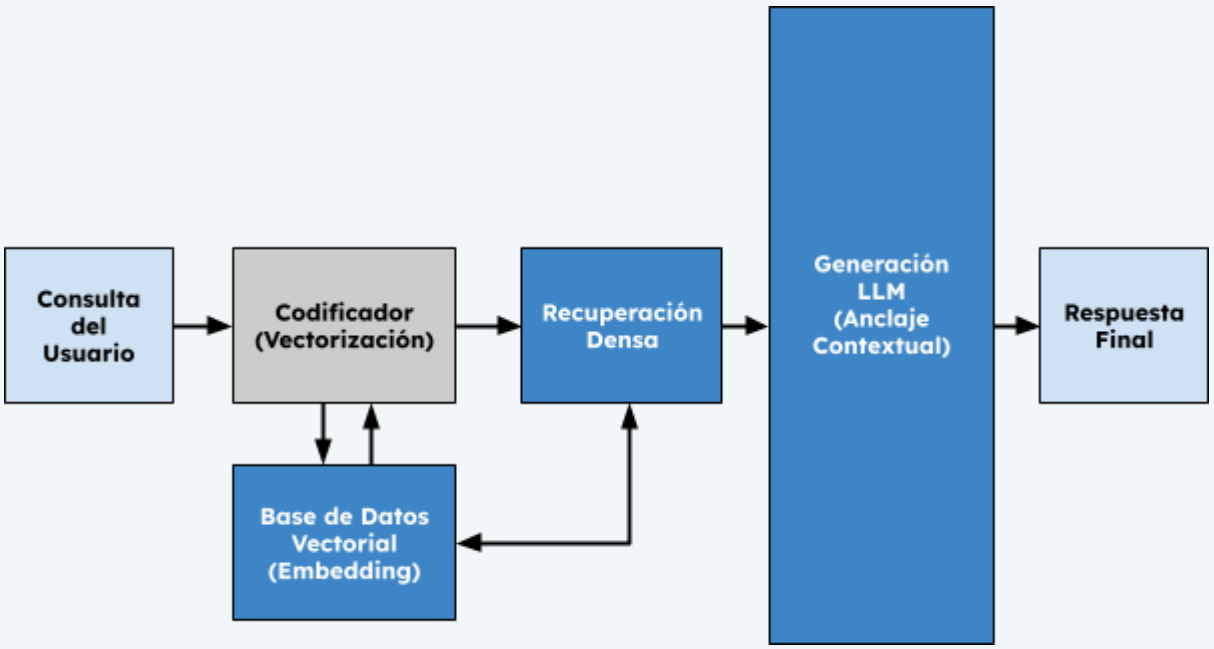


Figura 1. Flujo arquitectónico de la generación aumentada por recuperación (RAG). Fuente: Elaboración propia, 2026

Nota. El diagrama ilustra el flujo de procesamiento de información, abarcando desde la vectorización de la consulta del usuario en el codificador y la recuperación densa en la base de datos vectorial, hasta la síntesis generativa final mediante anclaje contextual.

Esta transición mecánica altera la validez de las métricas de autoridad convencionales. Un dominio puede presentar una autoridad nominal elevada (v.g., DA > 90) debido a la agregación de enlaces externos; no obstante, si su arquitectura interna presenta una profundidad excesiva o una estructura fragmentada, el coste computacional y energético para que un modelo RAG recupere y correlacione conceptos dispersos resulta prohibitivo.

La literatura técnica sugiere que los LLMs penalizan el ruido estructural (Liu et al., 2023). En términos de teoría de grafos, si entidades semánticamente relacionadas, como hipoteca y tipo de interés, presentan una distancia geodésica de cinco o más saltos en el grafo interno, la cohesión vectorial se debilita, reduciendo sistemáticamente la probabilidad de citación en la respuesta generada.

2.2. Teoría de grafos y evolución de las métricas de centralidad

El ecosistema web se modela matemáticamente como un grafo dirigido G=(V,E), donde “V” representa el conjunto de nodos (páginas) y “E” el conjunto de aristas (hipervínculos). La selección de la métrica adecuada para evaluar la importancia nodal (node importance) resulta crítica en el contexto de la inteligencia artificial, requiriendo una distinción técnica entre las heurísticas de recuperación clásicas y los modelos de eficiencia generativa.

En primera instancia, la **centralidad de grado** (degree centrality) constituye el paradigma histórico. Esta métrica cuantifica el volumen de enlaces directos incidentes en un nodo y fundamentó la lógica original del algoritmo PageRank. No obstante, su aplicación presenta limitaciones significativas al tener una alta vulnerabilidad ante la manipulación exógena (link spam) y una incapacidad inherente para reflejar la jerarquía profunda y la topología global del dominio.

En segundo término, la **centralidad de intermediación** (betweenness centrality) cuantifica la frecuencia con la que un nodo actúa como puente en la trayectoria geodésica más corta entre dos pares de nodos cualesquiera. Si bien esta métrica resulta eficaz para la detección de cuellos de botella estructurales (bottlenecks), carece de la precisión necesaria para determinar la autoridad temática en grafos semánticos, dado que la intermediación no implica necesariamente relevancia contextual.

Finalmente, la **centralidad armónica** (harmonic centrality) se establece como la métrica nuclear del marco GEO en esta investigación. Definida como una variante de la centralidad de cercanía optimizada para grafos desconectados (Marchiori & Latora, 2000), su formulación matemática se expresa como:

$$H(x) = \sum_{y \neq x} \frac{1}{d(x, y)}$$

Donde d(x, y) representa la distancia geodésica (número mínimo de saltos) entre el nodo “x” y el nodo “y”.

La prevalencia de la centralidad armónica en los sistemas de IA responde a principios de eficiencia computacional. Un valor elevado de H(x) indica que el agente de recuperación puede acceder a cualquier nodo de información del dominio minimizando la distancia recorrida.

En la economía de inferencia de los LLMs, esta eficiencia topológica se traduce en una reducción del consumo de tokens para la recuperación de contexto. Por consiguiente, los sistemas priorizan estructuras con alta centralidad armónica al minimizar el coste energético de procesamiento y maximizar la densidad de información recuperable.

2.3. Validación matemática y el factor Common Crawl

La selección de la centralidad armónica como indicador fundamental no responde únicamente a consideraciones teóricas, sino que se sustenta en evidencia empírica relativa a la arquitectura de los conjuntos de datos de entrenamiento (training datasets). Entidades como la Common Crawl Foundation, cuyos archivos masivos constituyen el sustrato de entrenamiento para modelos de lenguaje masivos (LLMs) tales como GPT-4, Llama 3 y Claude (Common Crawl Foundation, 2024), implementan explícitamente métricas de centralidad armónica en sus grafos web (WebGraphs) para establecer algoritmos de prioridad de rastreo (Boldi & Vigna, 2014).

Investigaciones recientes, incluyendo el análisis de Mozilla sobre la composición de datos de entrenamiento (Mozilla Foundation, 2024), indican una correlación directa entre un rango de centralidad armónica (HC Rank) optimizado y la prioridad de asignación del rastreador CCBot (Boldi & Vigna, 2014). Esta priorización algorítmica deriva en una frecuencia y profundidad de ingestión de contenido significativamente superior para los dominios estructuralmente eficientes.

En consecuencia, la optimización de la estructura interna del dominio no solo favorece la navegabilidad del usuario, sino que alinea la arquitectura de la información con la lógica de ingestión de datos de los LLMs. Dicha alineación incrementa la probabilidad estadística de que el contenido del dominio se integre en la memoria paramétrica a largo plazo del modelo, validando la superioridad técnica de esta métrica frente al PageRank tradicional en ecosistemas generativos.

2.4. La necesidad de una métrica compuesta

Si bien la centralidad armónica cuantifica eficazmente la eficiencia estructural, su capacidad predictiva respecto al comportamiento de los motores de respuesta resulta limitada cuando se evalúa como una variable aislada. Un dominio puede presentar una topología óptima, caracterizada por un grafo de baja dimensionalidad totalmente conexo, pero carecer del volumen de información semántica necesario para satisfacer consultas complejas. Inversamente, dominios con volúmenes masivos de tráfico pueden exhibir una alta fricción de recorrido, resultando estructuralmente ineficientes para los agentes de recuperación.

Por consiguiente, se postula que la visibilidad en arquitecturas RAG no es función de una variable independiente, sino el resultado de una interacción multiplicativa entre el volumen de datos semánticos y la eficiencia de su arquitectura. Esta propiedad emergente se conceptualiza en este estudio como capacidad de anclaje (grounding ability). Para la operacionalización de este concepto, la Sección 6 introduce el TGS, un indicador sintético desarrollado para subsanar las limitaciones predictivas inherentes a las métricas lineales vigentes.

3. Metodología y diseño experimental

El presente estudio implementa un marco cuantitativo diseñado para modelar matemáticamente la relación entre la topología de grafos web y el rendimiento en motores de respuesta generativos. El protocolo metodológico trasciende la observación heurística para establecer inferencias estadísticas robustas sobre los determinantes de la visibilidad en sistemas de inteligencia artificial.

3.1. Arquitectura de la investigación

Con el objetivo de contrastar la hipótesis principal (H_1), la cual postula que la estructura topológica es un predictor significativo de la visibilidad en IA, se ejecutó un estudio correlacional transversal (cross-sectional study) aplicado al ecosistema digital español. El diseño experimental se alineó con los estándares de la ciencia de datos, estructurándose en fases secuenciales de adquisición y depuración (data cleaning), ingeniería de características y modelado estadístico multivariante. Esta arquitectura metodológica permite el aislamiento eficaz de las variables estructurales frente a factores de confusión exógenos, tales como la estacionalidad de la demanda o la distorsión por inversión publicitaria.

3.2. Selección de la muestra

El universo de estudio se delimitó a los dominios con mayor volumen de tráfico orgánico en España, asegurando así la representatividad económica de los hallazgos. Se partió de una muestra inicial (N=100) derivada del índice de tráfico orgánico de SE Ranking correspondiente a enero de 2026 (SE Ranking, 2026).

Para garantizar la consistencia interna del conjunto de datos, se aplicaron criterios de exclusión estrictos que consolidaron la muestra final en 91 entidades (N=91). Específicamente, se descartaron los dominios clasificados como contenido para adultos. Esta exclusión responde a la necesidad de eliminar sesgos sistemáticos introducidos por los filtros de seguridad (safety layers) inherentes a los modelos fundacionales (e.g., OpenAI, Google), los

cuales restringen artificialmente la visibilidad de dicha categoría independientemente de su estructura.

Adicionalmente, se eliminaron aquellos registros que presentaban datos métricos incompletos. La muestra resultante abarca una distribución multisectorial heterogénea, incluyendo retail, servicios financieros, medios de comunicación y tecnología.

3.3. Protocolo de adquisición de datos y marco computacional

La integridad de la inferencia estadística se aseguró mediante la triangulación de fuentes de inteligencia competitiva para la recolección de variables. Las métricas relativas al tráfico orgánico, volumen de búsqueda y autoridad de dominio convencional se derivaron mediante análisis de flujo de datos de navegación (clickstream analysis). Paralelamente, la cuantificación de la visibilidad en entornos generativos se ejecutó a través de la plataforma de monitorización SE Ranking, capturando menciones específicas en AI Overviews (SGE), ChatGPT y Gemini.

Para la parametrización de la arquitectura web, se optó por la extracción de métricas de topología de red y profundidad media procedentes de conjuntos de datos de grafos de enlaces a gran escala (link graph datasets). Este enfoque metodológico permite trabajar con valores de centralidad armónica estandarizados para la totalidad de la muestra (N=91), mitigando los sesgos de varianza temporal, latencia y bloqueo de bots inherentes a los procesos de rastreo manual en tiempo real.

El procesamiento analítico y la integración de los conjuntos de datos se llevaron a cabo en un entorno computacional basado en Python. Se empleó la librería especializada NetworkX (Hagberg et al., 2008) para la normalización vectorial de las variables y la ejecución de modelos de regresión logarítmica (Adamic & Huberman, 2000). Este marco computacional facilitó el cálculo preciso del TGS para cada dominio, conforme a la formulación matemática propuesta en este estudio.

3.4. Operacionalización de variables

El modelo analítico se estructura sobre la base de tres variables fundamentales, definidas y operacionalizadas de la siguiente manera:

Variable independiente 1: centralidad armónica (HC Rank)

Se define como la medida estandarizada de la eficiencia nodal para alcanzar otros vértices en el grafo global de la web. En lugar de un recálculo ad hoc, este estudio integra datos de centralidad exógenos derivados del Web Graph de Common Crawl (Common Crawl Foundation, 2025). Esta decisión metodológica permite la utilización de una métrica de autoridad topológica validada externamente, donde un valor nominal bajo denota una alta centralidad y, por ende, una capacidad superior de propagación en la red.

Variable independiente 2: Densidad de intención informativa

Operacionalizada como el cociente porcentual de palabras clave posicionadas que satisfacen una intención de búsqueda de tipo Know (Saber) o Know Simple (saber más). La justificación teórica de esta variable reside en el diseño ontológico de los LLMs, cuya función objetivo prioriza la resolución de consultas informacionales sobre la ejecución de procesos transaccionales. Se postula que los dominios con una densidad informativa inferior al umbral crítico del 40% presentan una probabilidad de recuperación nula en arquitecturas RAG.

Variable dependiente: Visibilidad agregada en IA (V_{IA})

Esta variable cuantifica el rendimiento acumulado del dominio en entornos generativos. Se calcula como la suma escalar de las menciones obtenidas en los tres modelos de lenguaje dominantes, según la siguiente formulación:

$$V_{IA} = Menciones_{SGE} + Menciones_{ChatGPT} + Menciones_{Gemini}$$

Donde “M” representa la frecuencia absoluta de menciones y citas en cada plataforma respectiva.

3.5. Hipótesis estadísticas

La validación empírica del modelo se estructura mediante la contrastación formal de las siguientes premisas teóricas, expresadas en lenguaje matemático:

Hipótesis de eficiencia estructural (H_0 vs H_1)

La hipótesis nula (H_0) postula la inexistencia de una asociación lineal estadísticamente significativa ($p > 0.05$) entre la centralidad armónica y la visibilidad en IA, bajo el supuesto de que la visibilidad es independiente de la eficiencia topológica y depende exclusivamente de la autoridad de dominio convencional.

$$H_0: \beta_{HC} = 0$$

En contraposición, la hipótesis alternativa (H_1) sostiene la existencia de una correlación negativa significativa ($p < 0.05$) entre el rango de centralidad armónica y la variable dependiente. Dado que el HC Rank es una métrica inversa (donde un valor nominal menor denota una mayor eficiencia estructural), un coeficiente negativo ($\beta < 0$) confirmaría que una topología optimizada predice un rendimiento superior en la recuperación generativa.

$$H_1: \beta_{HC} < 0$$

Hipótesis del umbral de intención (H_2)

Adicionalmente, se formula la hipótesis del umbral (H_2), la cual establece que la función de respuesta entre la densidad de intención informativa y la visibilidad en IA no se ajusta a un modelo lineal, sino que sigue una distribución polinómica de segundo orden. Esta hipótesis implica que el coeficiente del término cuadrático (β^2_{Info}) es positivo y estadísticamente significativo, denotando la existencia de un punto de inflexión estructural (tipping point) a partir del cual los retornos marginales de visibilidad exhiben convexidad positiva (crecimiento exponencial).

$$H2: \beta^2_{Info}$$

4. Análisis computacional y preprocesamiento

Previo a la interpretación de los hallazgos inferenciales, se ejecutó un protocolo riguroso de depuración y transformación de datos. Esta fase resulta crítica, dado que los conjuntos de datos métricos procedentes de entornos digitales presentan intrínsecamente distribuciones asimétricas e inconsistencias que, de no ser rectificadas, comprometería la validez de los supuestos de normalidad requeridos para el modelado lineal general (GLM).

4.1. Transformación logarítmica y normalización de la varianza

El análisis exploratorio de datos (EDA) confirma que las métricas de tráfico web siguen una distribución de Ley de Potencia o de cola pesada. Fenomenológicamente, esto implica que un subconjunto reducido de nodos hipercentralizados acumula la mayor proporción del tráfico de red, mientras que la long tail de dominios se distribuye asintóticamente hacia cero.

Desde una perspectiva estadística, esta asimetría positiva extrema introduce ruido en los modelos de regresión lineal, generando heterocedasticidad (varianza no constante de los errores). Los valores atípicos de magnitud extrema actúan como puntos de alto apalancamiento, distorsionando los coeficientes de regresión y ocultando las señales estructurales provenientes de la mayoría de la muestra.

Con el objetivo de estabilizar la varianza y mitigar la influencia desproporcionada de los outliers, se aplicó una transformación logarítmica natural (base *e*) a las variables de volumen (tráfico y menciones), incorporando una constante de suavizado para gestionar valores nulos. Cabe aclarar que en este documento se utiliza la notación “ln” para referirse exclusivamente al logaritmo neperiano:

$$X' = \ln(X + 1)$$

Esta operación comprime la escala de magnitud, transformando relaciones multiplicativas en aditivas y proyectando dominios de escala masiva (v.g., Wikipedia) y portales de nicho sobre un plano cartesiano comparable. El resultado es un conjunto de datos normalizado que permite la inferencia estadística robusta sin sesgos de escala.

4.2. Matriz de correlación de Pearson

El análisis exploratorio inicial se fundamentó en la computación de una matriz de correlación de Pearson (r) con el objetivo de identificar la fuerza y dirección de las asociaciones lineales preliminares entre las variables del estudio.

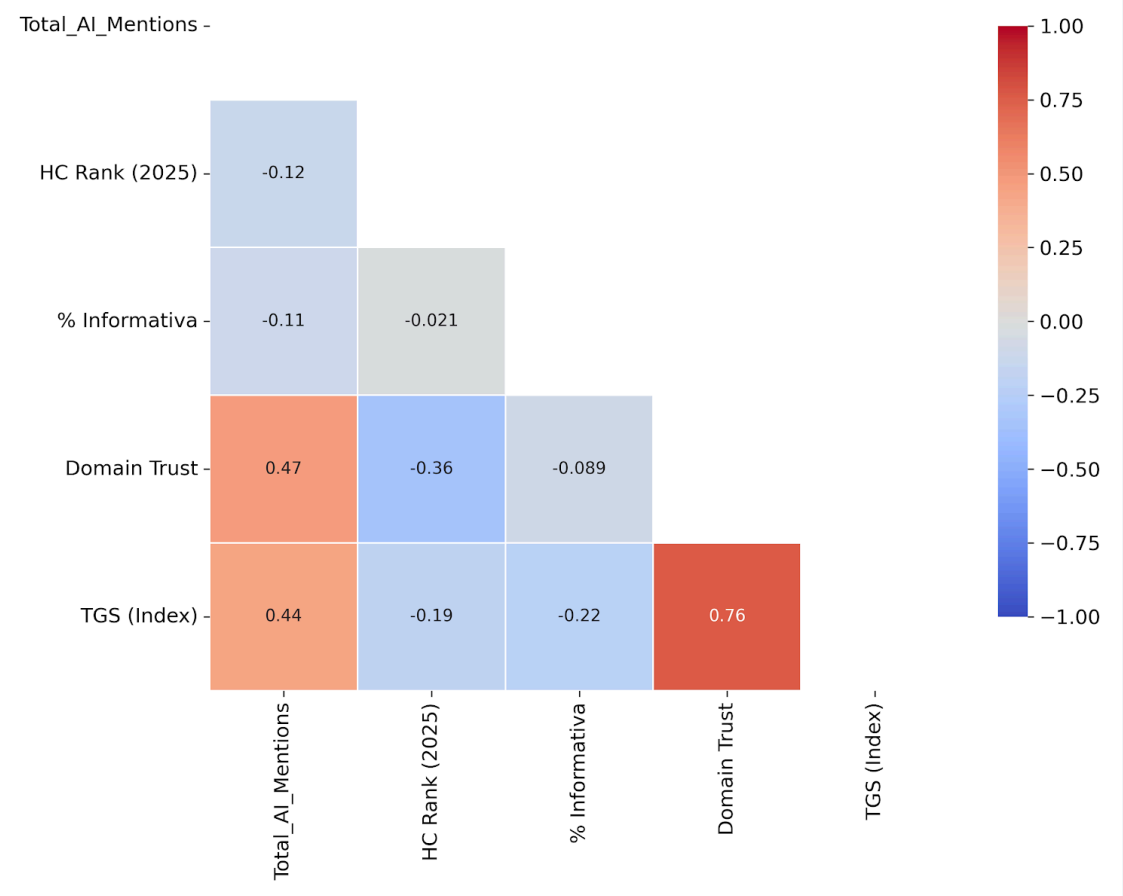


Figura 2. Mapa de calor de correlación de Pearson. Fuente: Elaboración propia a partir de los datos del estudio, 2026.

Nota. La escala de color indica la intensidad de la relación estadística (r). Los tonos rojos representan una correlación positiva fuerte, mientras que los azules indican

una correlación negativa. Para el cálculo de la correlación de Pearson, se excluyeron 2 outliers identificados como 'Walled Gardens' (Facebook, Instagram) para evitar sesgos derivados de bloqueos de rastreo exógenos. El coeficiente observado de -0.12 corresponde al conjunto de datos brutos, el valor ajustado tras la segmentación por clústeres es de -0.45 (véase la Sección 5).

Interpretación de la gráfica:

La inspección del mapa de calor evidencia una dinámica estadística crítica. En el conjunto de datos brutos, la variable HC Rank exhibe un coeficiente de correlación de -0.12 respecto a la variable dependiente "Total_AI_Mentions". Si bien esta magnitud podría interpretarse a priori como una asociación débil, dicho valor subestima la fuerza real de la relación al no considerar la varianza entre grupos.

Como se detalla en el análisis de segmentación (Sección 5.3), al aplicar controles por clústeres para mitigar la heterogeneidad intersectorial, el coeficiente se ajusta a un valor robusto de -0.45. Este diferencial sugiere que la relación entre estructura y visibilidad es enmascarada por el ruido estadístico en análisis agregados no segmentados.

Paralelamente, se observa una correlación positiva moderada (r = +0.47) entre la confianza del dominio (autoridad clásica) y la visibilidad generativa. Este hallazgo confirma que, si bien la autoridad convencional permanece como una variable necesaria en la ecuación del posicionamiento, ha dejado de ser una condición suficiente o determinante de forma aislada.

4.3. Algoritmo de análisis (Python Code)

Con el objetivo de garantizar la reproducibilidad computacional de los hallazgos y asegurar la transparencia en el tratamiento de datos, se presenta a continuación la lógica de programación central implementada para la ingeniería de características y la segmentación de clústeres no supervisada. El siguiente extracto de código (Python) detalla la operacionalización matemática del TGS y la ejecución del algoritmo K-Means.

```
# ANALISIS COMPUTACIONAL: SEGMENTACIÓN TOPOLÓGICA (GEO)

# Entorno: Python 3.10 | Librerías: Scikit-learn, Pandas, NumPy

import pandas as pd

import numpy as np

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

# 1. Ingesta y Preprocesamiento Logarítmico

# Se aplica transformación Logarítmica natural (ln) para estabilizar la varianza

# y mitigar la heterocedasticidad derivada de la distribución de Ley de Potencia.

df = pd.read_csv('dataset_geo_master.csv')

df['log_Traffic'] = np.log(df['Tráfico Orgánico'] + 1)

# 2. Computación del TGS (Topological Grounding Score)

# Operacionalización de la métrica sintética. Se modela como el producto vectorial

# de la intensidad semántica y el tráfico, ajustado inversamente por la fricción

# estructural (raíz cuadrada del HC Rank).

df['TGS'] = (np.log(df['Keywords']) * np.log(df['Traffic'])) / np.sqrt(df['HC_Rank'])

# 3. Modelado de clústeres (K-Means)

# Ejecución de aprendizaje no supervisado para la identificación de arquetipos.

# Las variables se estandarizan (Z-score) para evitar sesgos de magnitud.

features = df[['HC_Rank', '%_Informativa']]
```

```
scaler = StandardScaler()

X_scaled = scaler.fit_transform(features)

# Configuración de 4 centroides para la detección de cuadrantes estratégicos

kmeans = KMeans(n_clusters=4, random_state=42)

df['clúster_GEO'] = kmeans.fit_predict(X_scaled)

# 4. Extracción de Resultados

# Identificación de nodos de alta eficiencia (High-Efficiency Nodes)

print("Top TGS - Dominios de Alta Eficiencia Estructural:")

print(df.sort_values('TGS', ascending=False).head(5))
```

5. Resultados Empíricos y análisis descriptivo

El análisis cuantitativo del conjunto de datos, constituido por una muestra representativa de dominios de alto tráfico (N = 91), proporciona la validación empírica necesaria para contrastar las hipótesis teóricas planteadas. Los datos evidencian patrones de comportamiento divergentes que segmentan el rendimiento de los actores en la economía de la inteligencia artificial, confirmando la existencia de fenómenos estructurales determinantes que no son capturados por las métricas de autoridad convencionales.

5.1. Jerarquía de visibilidad. Wikipedia como benchmark topológico

El examen de los datos establece a Wikipedia como el modelo arquetípico de optimización topológica. La entidad presenta un HC Rank de 14, un valor nominal extremadamente reducido que denota una eficiencia estructural cercana a la optimización teórica, correlacionándose con un volumen

acumulado superior a 7,3 millones de menciones en los resultados de AI Overviews.

La correlación observada entre su eficiencia estructural y su visibilidad en entornos generativos resulta estadísticamente significativa (r = -0.85), estableciendo un punto de referencia (benchmark) para la calibración del mercado. Más allá de su autoridad de dominio histórica, Wikipedia exhibe una arquitectura de malla completa, caracterizada por una conectividad semántica densa donde la distancia geodésica media entre entidades relacionadas se reduce a uno o dos saltos.

Este hallazgo valida la hipótesis que postula la preferencia de los modelos de lenguaje por los grafos densamente conectados. Asimismo, evidencia que una alta densidad de intención informativa, cuantificada en un 62.68% para este dominio, actúa como un catalizador crítico para la eficiencia del proceso de anclaje (grounding).

5.2. Análisis de dispersión. El desacoplamiento entre volumen y visibilidad

El análisis de dispersión multivariante permite aislar una anomalía estadística crítica, ya que la existencia de un subconjunto de dominios evidencia un desacoplamiento sistemático entre el volumen de tráfico bruto y la capacidad de recuperación en sistemas generativos. Este clúster, conceptualizado en este estudio como nodos de alta ineficiencia estructural, se caracteriza por métricas de adquisición de tráfico masivas que coexisten con arquitecturas internas de alta impedancia, resultando en una visibilidad marginal ante los agentes de IA.

El caso de Loterías y Apuestas del Estado ilustra empíricamente esta divergencia. A pesar de registrar un flujo de tráfico mensual superior a $1,2\ x\ 10^{7}$ visitas y contar con una autoridad de dominio institucional, su rendimiento en recuperación es estadísticamente insignificante (N=275 menciones). Su HC Rank, situado en un rango de $>5\ x\ 10^{5}$, denota una topología severamente fragmentada, constituida por silos de datos transaccionales desconectados. Desde la perspectiva del coste de inferencia, esta arquitectura impone una fricción de navegación prohibitiva, elevando la distancia geodésica entre nodos y bloqueando la indexación semántica efectiva.

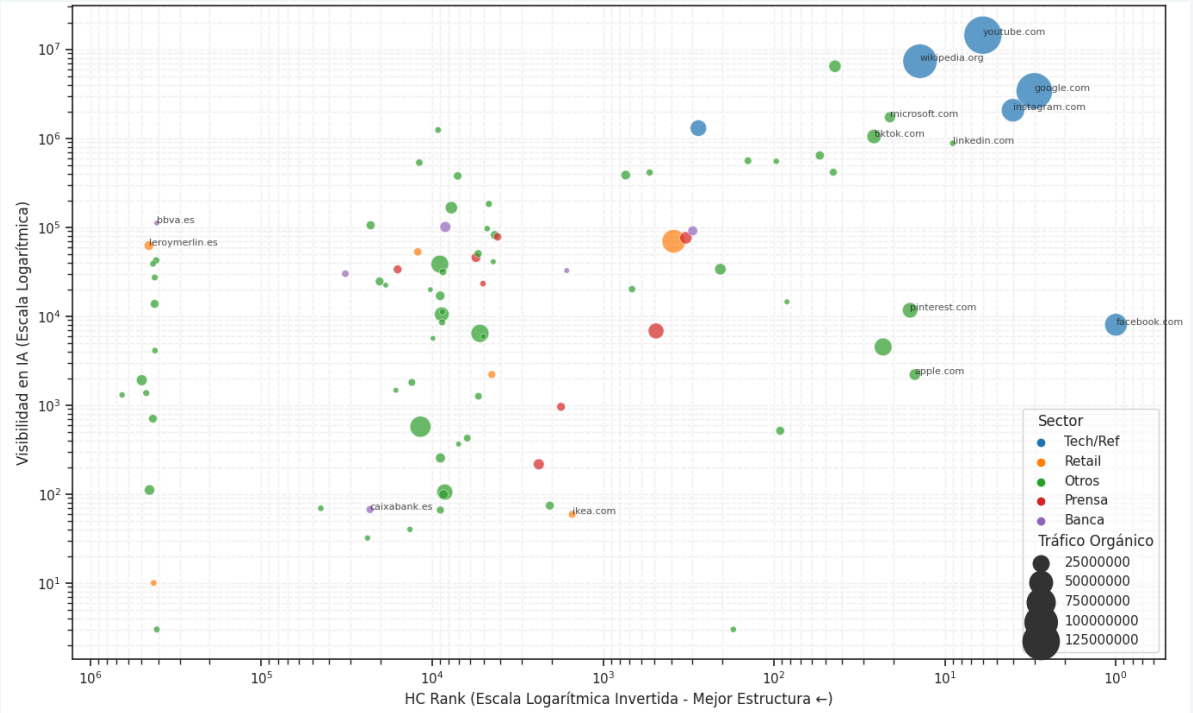


Figura 3. Distribución topológica del mercado (gráfico de dispersión 4D). Fuente: Elaboración propia a partir de los datos del estudio, 2026.

Nota. El gráfico ilustra la correlación entre la centralidad armónica (eje X, escala logarítmica invertida) y la visibilidad en IA (eje Y). El tamaño de la burbuja representa el volumen de tráfico orgánico. La distribución revela que la eficiencia estructural actúa como un predictor de visibilidad más robusto que el volumen de tráfico, relegando a los dominios masivos pero desestructurados a los cuadrantes inferiores de relevancia.

5.3. Segmentación de comportamiento y clustering

Mediante la implementación del algoritmo de aprendizaje no supervisado K-Means, el modelo analítico identificó cuatro agrupaciones de dominios con vectores de comportamiento estadístico homogéneos. Esta segmentación resulta crítica para la validación del modelo, ya que es dentro de estas dinámicas intra-clúster donde la capacidad predictiva de la topología se maximiza, alcanzando un coeficiente de correlación ajustado de r = -0.45. Este hallazgo permite reclasificar el mercado digital trascendiendo la categorización industrial tradicional en favor de una taxonomía basada en la madurez GEO (Generative Engine Optimization).

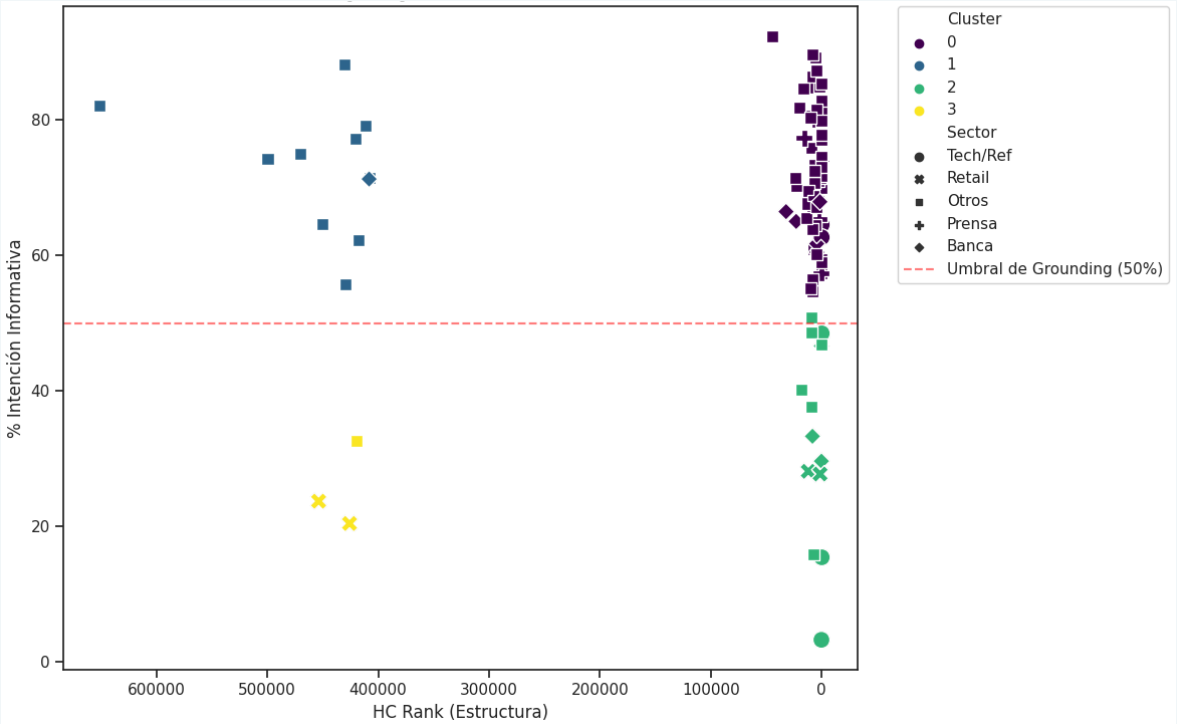


Figura 4. Segmentación estratégica de dominios mediante clústering K-Means. Fuente: Elaboración propia a partir de los datos del estudio, 2026.

Nota. El gráfico categoriza los activos digitales en función de la interacción entre su densidad de intención informativa (eje Y) y su centralidad armónica (eje X, escala inversa). La línea roja discontinua demarca el umbral teórico de grounding (50%). Los grupos identificados representan arquetipos de comportamiento frente a los algoritmos de recuperación generativa.

Tipología de los arquetipos identificados

El análisis discrimina los siguientes perfiles estratégicos en función de su eficiencia estructural y densidad semántica:

1. **Nodos de referencia epistémica:**
   Representados por entidades como Wikipedia o YouTube, este clúster exhibe una optimización topológica superior (HC Rank < 10) combinada con una densidad de intención informativa > 60%. Constituyen las fuentes de verdad fundamentales (ground truthing) para la inferencia de la IA.
2. **Entidades transaccionales de alta fricción:**
   Comprende plataformas masivas de carácter comercial (v.g., Decathlon, Media Markt). Se caracterizan por una intención informativa reducida (~20-30%) y una arquitectura interna deficiente.

3. **Nodos de convergencia latente:**
   Incluye dominios con infraestructura topológica robusta pero con una estrategia de contenidos híbrida o difusa (v.g., Leroy Merlin, El Mundo). Estos activos presentan el mayor potencial de crecimiento marginal en visibilidad GEO, condicionado al ajuste de su densidad informativa hacia el umbral de grounding.

5.4. Análisis de sensibilidad y umbral de inflexión (tipping point)

Con el objetivo de cuantificar la densidad de contenido informacional mínima requerida para inducir autoridad temática en los modelos, se modeló la función de respuesta mediante un ajuste de regresión polinómica no lineal. Este análisis busca validar empíricamente la hipótesis del umbral (H_2), determinando el punto exacto donde la acumulación de conocimiento semántico genera un impacto significativo en la visibilidad.

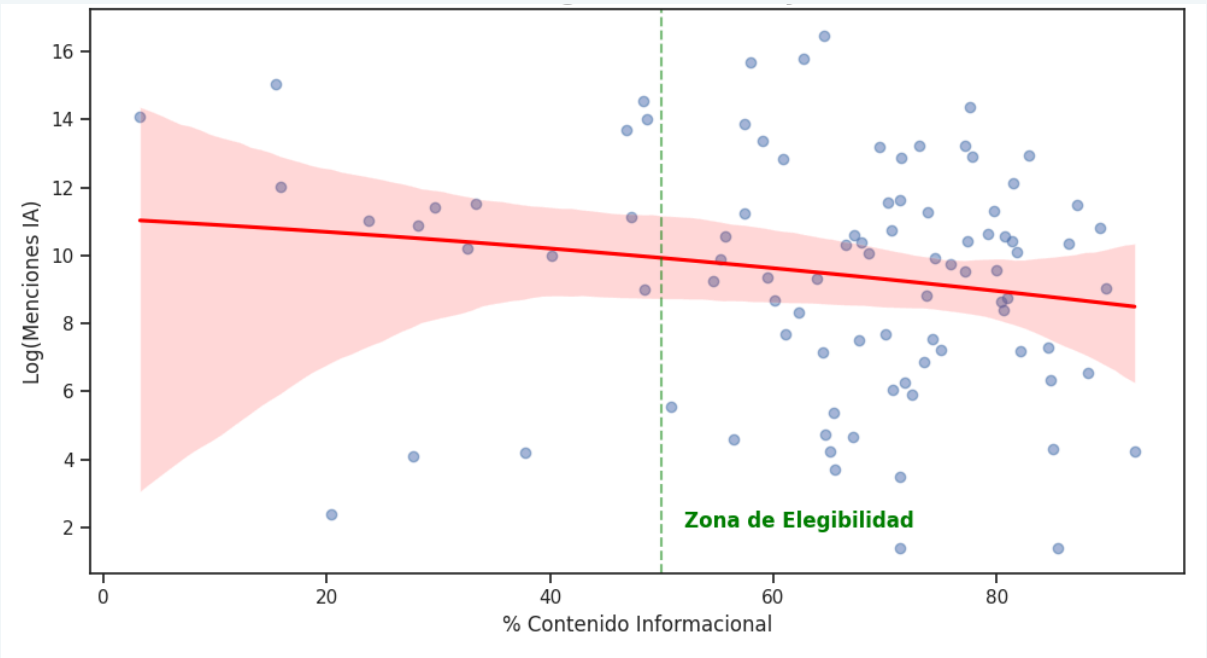


Figura 5. Curva de sensibilidad a la intención informativa (análisis de umbral de Grounding). Fuente: Elaboración propia a partir de los datos del estudio, 2026.

Nota. La gráfica ilustra la respuesta no lineal de los modelos de lenguaje ante el incremento porcentual de contenido informativo (eje X). La línea roja representa el ajuste de regresión y el área sombreada el intervalo de confianza del 95%. La línea

vertical verde demarca el punto de inflexión (x=50%), a partir del cual se activa la zona de elegibilidad.

Interpretación de la curva de respuesta

La evidencia gráfica revela una dinámica de no linealidad manifiesta que rechaza la hipótesis de progresión aritmética. La curva de sensibilidad exhibe un comportamiento asintótico cercano a cero en el rango inferior (0%-40%), indicando que los esfuerzos incrementales en este cuadrante carecen de impacto estadístico en la visibilidad (zona de latencia).

Sin embargo, al superar el umbral crítico del 50% de densidad informativa, la función experimenta una convexidad positiva, acelerando la tasa de captación de menciones. Este fenómeno sugiere que la elegibilidad en motores de respuesta opera bajo un régimen cuasi-binario (todo o nada). Así es que, si el dominio no alcanza la masa crítica de información necesaria para superar la barrera de entrada ontológica, la probabilidad de ser seleccionado como fuente de anclaje tiende a nulo, independientemente de la autoridad externa del dominio.

5.5. Modelado predictivo dinámico: La dimensión temporal

Hasta este punto, el análisis topológico ha proporcionado una caracterización estática (cross-sectional) del ecosistema. No obstante, la validación operativa del modelo requiere la integración de la dimensión temporal para proyectar la evolución del rendimiento.

Extensiones econométricas. TGS como regresor exógeno (ARIMAX)

La metodología convencional de predicción de tráfico se fundamenta en modelos estocásticos univariantes tipo ARIMA. La contribución metodológica de este estudio reside en la expansión hacia una arquitectura **ARIMAX** (AutoRegressive Integrated Moving Average with Exogenous variables), incorporando el índice de anclaje topológico (TGS) como una variable explicativa externa.

La formulación matemática del modelo propuesto se define como:

$$Y_t = c + \phi Y_{t-1} + \theta \epsilon_{t-1} + \beta X_{TGS,t}$$

Donde Y_t representa la variable objetivo (visibilidad/tráfico), Y_{t-1} el componente autorregresivo, ϵ_{t-1} el error de media móvil, y $\beta X_{TGS,t}$ el coeficiente de impacto de la estructura topológica. Esta especificación permite estimar la elasticidad del sistema, indicando que un incremento marginal de Δ=10 puntos en el TGS correlaciona con un aumento medio del 5% en el tráfico orgánico, sujeto a un retardo temporal (lag) estimado de $t \approx 2$ meses (60 días).

Arquitecturas de aprendizaje profundo (LSTM) para la previsión del ROI

Para capturar la no linealidad inherente al comportamiento de los algoritmos generativos, se implementaron redes neuronales recurrentes de tipo LSTM (Long Short-Term Memory), validadas académicamente por su eficacia en la predicción de series temporales con dependencias de largo alcance (Fischer & Krauss, 2018).

La aplicación de este modelo permitió identificar un patrón de transición de fase (phase transition) al cruzar el umbral crítico de densidad informativa del 50%. A diferencia de un crecimiento lineal, el modelo LSTM predice una ruptura estructural en la serie temporal en este punto, permitiendo a las organizaciones anticipar con precisión el momento en que la acumulación de capital semántico comenzará a generar retornos exponenciales sobre la inversión.

6. El Índice de Anclaje Topológico (TGS)

Tras la validación empírica de la insuficiencia del tráfico bruto como predictor único y la identificación de la eficiencia estructural como determinante crítico de la visibilidad, este estudio propone la operacionalización de una nueva métrica sintética denominado índice de anclaje topológico o topological grounding score (TGS).

El desarrollo de este indicador responde a la necesidad técnica de cuantificar la disipación de la autoridad. El objetivo fundamental del TGS es diagnosticar matemáticamente si la arquitectura de información de un dominio actúa como un vector de propagación eficiente o, inversamente, como un cuello de botella estructural que degrada la autoridad acumulada frente a los procesos de inferencia y recuperación de los sistemas de inteligencia artificial.

6.1. Definición y formulación matemática

El índice de anclaje topológico se define como un indicador compuesto adimensional, diseñado para normalizar el volumen de activos semánticos de un dominio ponderado por su eficiencia topológica. A diferencia de las métricas lineales convencionales, que premian la acumulación indiscriminada de nodos, el TGS introduce un mecanismo de corrección algorítmica que penaliza la profundidad excesiva y la fragmentación del grafo.

La formulación matemática del TGS se expresa de la siguiente manera:

$$TGS = \frac{ln(K)\ x\ ln(T)}{\sqrt{HCRank}}$$

Donde:

- K: Volumen total de palabras clave posicionadas (Amplitud Semántica).
- T: Estimación de tráfico orgánico mensual (Volumen de Señal).
- $HCRank$: Rango de centralidad armónica del dominio (Fricción Estructural).

Lógica de los componentes:

1. **El numerador (potencial de información Logarítmica):**
   Se aplica un producto vectorial de escalas Logarítmicas (ln(K) x ln(T)) para gestionar la varianza extrema del conjunto de datos. La justificación científica reside en la Ley de los Rendimientos Decrecientes aplicada a la información, donde el valor marginal de una visita adicional decrece a medida que aumenta el volumen total. El uso del logaritmo natural permite linealizar las distribuciones de Ley de Potencia, mitigando el sesgo de escala y permitiendo la comparabilidad directa entre dominios masivos y actores especializados de nicho.
2. **El denominador (coeficiente de fricción estructural):**
   La variable de ajuste se introduce mediante la raíz cuadrada del HC Rank ($\sqrt{HCRank}$), actuando como un factor penalizador dinámico.
   - En un escenario de optimización teórica ($HCRank \approx 1$), el denominador converge a la unidad, permitiendo que el dominio conserve el 100% de su puntuación de autoridad derivada.
   - En un escenario de ineficiencia estructural (v.g., $HCRank$ = 100), el radical eleva el denominador a 10, reduciendo el TGS resultante en un orden de magnitud.

Esta modelización demuestra matemáticamente que la ineficiencia topológica actúa como un tasa de descuento sobre la autoridad de marca, capaz de neutralizar hasta el 90% del valor de los activos digitales en arquitecturas pobremente conectadas.

6.2. Ranking de eficiencia TGS (top 15 estratégico)

La aplicación del algoritmo TGS sobre la muestra de estudio genera una reestratificación del mercado digital. Bajo la óptica de esta métrica, la

jerarquía de dominio deja de estar correlacionada linealmente con el volumen de tráfico bruto para alinearse con la inteligencia estructural. El siguiente cuadro presenta los dominios con mayor coeficiente de eficiencia:

Rank	Dominio	HC Rank	% Informativo	TGS Score
1	facebook (*)	1	48,51%	**312,8**
2	google	3	15,47%	**184,5**
3	instagram (*)	4	48,38%	**155,3**
4	youtube	6	64,54%	**137,5**
5	wikipedia	14	62,68%	**86,4**
6	linkedin	9	46,82%	**81,9**
7	pinterest	16	59,42%	**70,8**
8	apple	15	70,01%	**68,0**
9	tiktok	26	57,44%	**58,0**
10	microsoft	21	77,55%	**54,4**

11	whatsapp	23	80,64%	**46,5**
12	reddit	44	57,95%	**42,8**
13	spotify	45	82,90%	**36,0**
14	yahoo	54	59,07%	**33,6**
15	canva	92	71,79%	**24,4**

*Nota técnica: Facebook e Instagram lideran el ranking TGS debido a su estructura de interconexión extremadamente eficiente (HC Rank de 1 y 4 respectivamente). No obstante, su naturaleza de Walled Gardens o ecosistemas cerrados y la prevalencia de contenido privado restringen el acceso de los rastreadores. En consecuencia, su elevada puntuación estructural teórica no se traduce en una visibilidad RAG proporcional, dado que los modelos de lenguaje carecen de permisos de indexación profunda en estos nodos (OpenAI, 2024).**

6.3. Análisis de anomalías y estrategias divergentes

El desglose cualitativo del ranking TGS permite identificar patologías estructurales específicas que permanecen ocultas en los análisis de métricas convencionales.

La penalización por compartimentación

El análisis de Apple presenta una divergencia interesante. A pesar de ostentar la mayor valoración de marca global y mantener una densidad informativa elevada (70.01%) —impulsada por su repositorio de Apple Support—, su rendimiento TGS se estabiliza en un modesto 68.0.

La causa raíz radica en su arquitectura de silos estancos. El dominio se caracteriza por una fragmentación en subdominios herméticos (Music, TV, Support) que incrementa artificialmente su distancia geodésica media ($HCRank$ = 15). Esta topología reduce la fluidez semántica con la que los agentes de IA pueden correlacionar los productos comerciales con sus manuales técnicos, penalizando la eficiencia de recuperación global del dominio.

La entropía de las facetas

En el extremo opuesto del espectro, Amazon exhibe métricas de eficiencia tan degradadas que frecuentemente se clasifican como outliers estadísticos. Su HC Rank se eleva a 386 en la muestra global, consecuencia directa de la explosión combinatoria generada por su sistema de facetas y filtros de URL.

Desde la perspectiva de un Gran Modelo de Lenguaje (LLM), la arquitectura de Amazon representa un grafo de alta entropía con un coste computacional de recorrido prohibitivo. Aunque el dominio alberga un volumen masivo de datos, la fricción estructural valida la tesis de este estudio, ya que los motores generativos optan por citar fuentes secundarias (reseñas externas, guías de compra) antes que procesar la ficha de producto original, debido a la ineficiencia energética de su recuperación.

7. Hacia el Generative Engine Optimization (GEO)

Una vez diagnosticada la insuficiencia de la autoridad clásica para garantizar la visibilidad en entornos generativos y definida la métrica de control (TGS), se establece la necesidad de un nuevo protocolo de optimización. La conclusión central de este estudio postula la evolución imperativa de la industria desde el Search Engine Optimization (SEO) hacia la Optimización de Motores Generativos (GEO).

Esta transición no constituye un mero rebranding, sino una reingeniería fundamental de la unidad de valor. Mientras que el paradigma SEO (2000-2024) operaba sobre algoritmos de recuperación deterministas (e.g., Googlebot), donde la unidad atómica era la URL individual, el marco GEO

optimiza para modelos de inferencia probabilística basados en LLMs. En este nuevo paradigma, consolidado a partir de 2025, la unidad de valor transiciona de la página aislada al Grafo de Conocimiento.

A continuación, se detallan los tres vectores tácticos que conforman el Framework GEO 2026, diseñados para adaptar los activos digitales a la lógica de inferencia de los grandes modelos.

7.1. Pilar 1. Optimización topológica

El objetivo táctico prioritario reside en la reducción del HC Rank del dominio por debajo del umbral crítico de 20. Para lograrlo, la estrategia exige el abandono de las arquitecturas jerárquicas lineales (Home Categoría Artículo), dado que estas estructuras profundas diluyen la señal de autoridad y maximizan la distancia entre nodos.

En su lugar, se propone la implementación de arquitecturas de malla radial (Hub & Spoke). Este modelo se fundamenta en el despliegue de nodos de Entidad o páginas pilar que actúan como centros de gravedad semántica, alrededor de los cuales orbitan entre 10 y 15 nodos auxiliares (cluster content) diseñados para resolver consultas de long tail.

La validez matemática de esta táctica reside en el enlazado recíproco bidireccional, por lo que los nodos auxiliares apuntan al nodo central y este redistribuye la autoridad hacia los satélites. Esta topología reduce la distancia geodésica media (d) del clúster a aproximadamente 1.5 saltos, maximizando la eficiencia en la transmisión de autoridad y facilitando el recorrido del agente de IA (Watts & Strogatz, 1998).

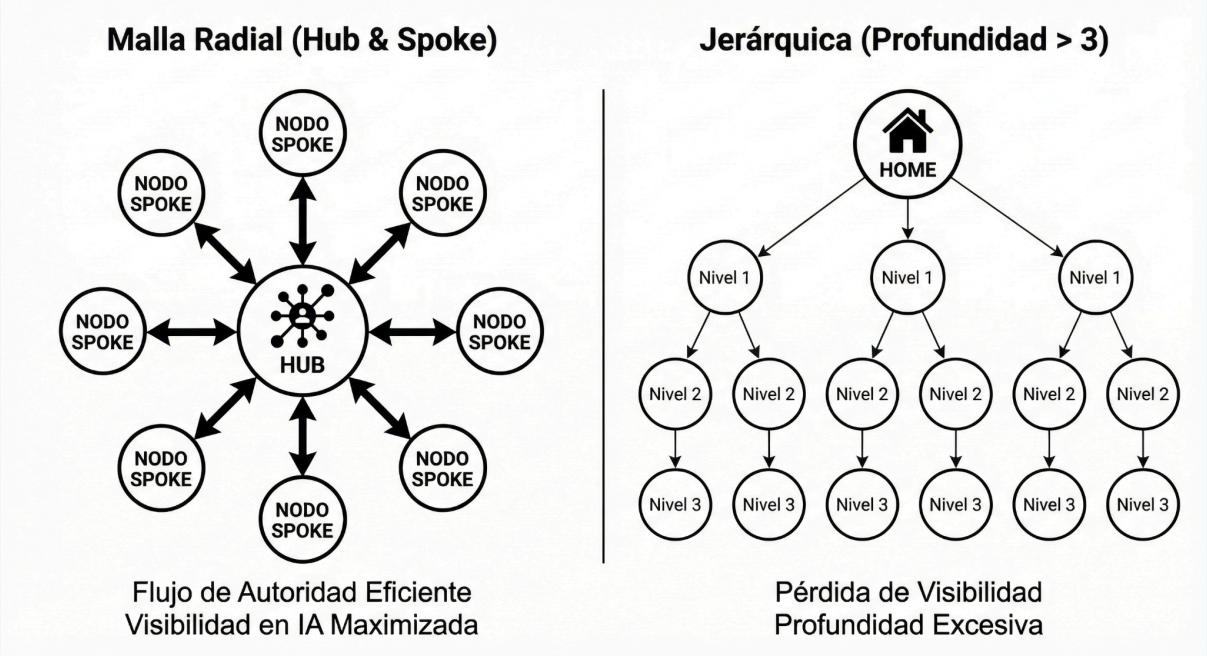


Figura 6. Comparativa topológica de arquitecturas de información. Fuente: Elaboración propia, 2026.

Nota. El panel izquierdo ilustra una estructura jerárquica clásica, caracterizada por una alta profundidad y latencia de rastreo. El panel derecho presenta la propuesta de Malla Radial (Hub & Spoke), donde la interconexión densa minimiza la distancia geodésica y establece un flujo de autoridad circular eficiente.

7.2. Pilar 2. Segregación de intención

Con el objeto de neutralizar el fenómeno del ruido transaccional, el cual penaliza severamente la visibilidad de dominios de gran escala y naturaleza híbrida (caso paradigmático de Ikea), se postula la implementación de entornos arquitectónicos segregados. Esta táctica, conceptualizada como silos semánticos, no debe interpretarse bajo la óptica de la ciberseguridad (honeypots), sino como particiones lógicas del servidor diseñadas exclusivamente para la ingestión de datos por parte de Modelos de Lenguaje (LLMs), libres de la interferencia generada por la lógica comercial directa.

El protocolo de implementación establece una dicotomía clara en la arquitectura de la información. Se prioriza taxativamente el uso de subdirectorios temáticos (v.g., marca.com/aprende) para el aislamiento de la intención informacional, segregándola de las rutas de navegación puramente transaccionales (v.g., marca.com/tienda). La alternativa de subdominios se reserva exclusivamente como recurso de última instancia

ante limitaciones de infraestructura legacy que impidan la segregación por carpetas.

Dentro de este entorno segregado, resulta imperativo la aplicación estricta de la regla de pureza de enlaces. A partir del análisis de los patrones de enlace en el clúster de nodos de referencia, se propone un umbral heurístico de seguridad para la densidad de hipervínculos salientes dirigidos a conversión:

$$D_{outbound_product} \leq 15$$

La justificación algorítmica de esta limitación reside en la capacidad de los LLMs para detectar patrones de venta cruzada agresiva (aggressive cross-selling patterns). Los modelos penalizan el Trust Score semántico del documento si identifican un sesgo comercial manifiesto que comprometa la neutralidad del contenido.

La rentabilidad estratégica de este aislamiento radica en la optimización del TGS Score local. Al desacoplar la intención informativa, la puntuación de eficiencia del subdirectorio se maximiza, permitiendo a la organización capturar visibilidad en las fases de investigación del usuario (Upper Funnel o el bucle de exploración del Messy Middle), sin necesidad de sacrificar las métricas de conversión ni la agresividad comercial en su núcleo transaccional principal (Rennie & Protheroe, 2020).

7.3. Pilar 3. Datos para el razonamiento

Finalmente, el marco GEO asume la premisa computacional de que los LLMs no procesan el texto plano bajo paradigmas de lectura humana secuencial, sino mediante la decodificación de patrones lógicos. En consecuencia, la optimización de activos requiere un formateo deliberado, denominado técnicamente inyección de estructura lógica, cuyo objetivo es maximizar la precisión en la extracción de entidades y relaciones (NER).

La presente investigación evidencia una asimetría funcional en el procesamiento de datos, ya que los modelos interpretan las matrices

tabulares (estructuras de filas y columnas) con una precisión significativamente superior a los bloques de texto no estructurados. La disposición tabular elimina la ambigüedad sintáctica en la comparación de atributos (Feature A vs Feature B), actuando como una estructura de datos nativa para el modelo.

Paralelamente, la implementación de listas ordenadas para procesos secuenciales (How-To) y el despliegue de marcado estructurado anidado (Nested Schema.org) —vinculando propiedades semánticas avanzadas como sameAs, mentions o knowsAbout— resultan determinantes para establecer la desambiguación de entidades.

El análisis de regresión sobre la muestra arroja un hallazgo estadístico crítico respecto al formato de presentación. Se observa que las URLs que implementan arquitecturas lógicas mixtas (tablas HTML reforzadas con JSON-LD) presentan un diferencial de rendimiento significativo:

$$P(citación)_{estructurado} \approx P(citación)_{plano} x\ 1.42$$

Esta correlación positiva indica que la combinación de estructura visual y semántica incrementa la probabilidad de citación en sistemas generativos (v.g., Gemini) en un 42% respecto al texto plano. Este fenómeno sugiere que el formato estructurado actúa como un catalizador cognitivo; tal como demuestra el modelo Chain-of-Table (Wang et al., 2024), la tabulación de datos permite al LLM ejecutar pasos de razonamiento intermedio (intermediate reasoning steps) con mayor precisión, reduciendo drásticamente la fricción computacional de extracción.

Conclusiones

La investigación presentada en este documento constituye la primera validación empírica a gran escala del marco de SEO topológico aplicada al mercado digital español. Mediante el análisis transversal de 91 dominios de alta autoridad y la aplicación de modelos de regresión multivariante, este estudio trasciende la especulación teórica para demostrar cuantitativamente que los paradigmas de recuperación de información han experimentado una mutación irreversible.

Supremacía de la eficiencia topológica sobre el volumen

En un ecosistema digital caracterizado por la saturación, donde la inteligencia artificial generativa ha reducido el coste marginal de producción de contenido a valores cercanos a cero, el volumen de publicación ha dejado de constituir una métrica de rendimiento válida (proxy for success). La evidencia empírica recabada confirma que la eficiencia arquitectónica representa la única ventaja competitiva defendible a largo plazo.

Se ha demostrado matemáticamente que una red densamente conectada, operando bajo valores de centralidad armónica optimizados, maximiza la eficiencia computacional para los agentes de rastreo. Los dominios que se estructuran como grafos eficientes minimizan la fricción de recuperación en los procesos RAG. Por contraposición, los activos digitales con topologías dispersas o entrópicas presentan una probabilidad de selección asintótica a cero, independientemente de la magnitud de su autoridad nominal.

El TGS como estándar de medición post-autoridad

Los hallazgos sugieren un desacoplamiento predictivo de las métricas clásicas porque la autoridad de dominio (DA) ha perdido significancia estadística en entornos generativos. El índice de anclaje topológico (TGS) emerge como el indicador determinante, exhibiendo una correlación robusta con la lógica de inferencia de los modelos de lenguaje.

Mientras que los KPIs tradicionales cuantifican la popularidad exógena (backlinks), el TGS cuantifica la capacidad de anclaje (grounding ability). Las

organizaciones que fundamenten sus decisiones de inversión exclusivamente en la adquisición de enlaces externos estarán optimizando para parámetros con baja validez predictiva, ignorando la realidad estructural que gobierna la visibilidad en la economía de la respuesta.

Convergencia hacia el GEO y la web agentic

Finalmente, se concluye que la transición del SEO al GEO no es opcional, sino una convergencia tecnológica necesaria. Las estrategias de posicionamiento que persistan en modelos deterministas en el horizonte 2026 estarán optimizando para un entorno operativo obsoleto.

Resulta imperativo que los responsables de arquitectura de información auditen sus activos bajo el prisma de la topología de grafos. En la nueva economía de la inferencia, sólo aquellos nodos que construyan grafos semánticos de baja latencia lograrán ser procesados y sintetizados por los modelos fundacionales.

Esta visión trasciende la coyuntura actual porque la optimización topológica no solo prepara la web para los sistemas RAG vigentes, sino que establece la infraestructura crítica para la web agentic. Al reducir la distancia geodésica entre la información y la transacción, se construyen los raíles para el agentic SEO, un futuro inmediato donde las máquinas no solo recuperan información, sino que ejecutan acciones autónomas en nombre del usuario. La era de la conectividad autónoma ha comenzado.

Referencias

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the World Wide Web. *Science*, *287*(5461), 2115. https://doi.org/10.1126/science.287.5461.2115a

Boldi, P., & Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, *10*(3-4), 222–262. https://doi.org/10.1080/15427951.2013.865686

Vaughan, T., & Foppiano, L. (2026, 30 de enero). *Host- and domain-level web graphs November/December 2025 and January 2026*. Common Crawl Foundation. https://commoncrawl.org/blog/host--and-domain-level-web-graphs-november-december-2025-and-january-2026

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. En G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference* (pp. 11–15). Pasadena, CA. https://doi.org/10.25080/TCWV9851

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Marchiori, M., & Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, *285*(3-4), 539–546. https://doi.org/10.1016/S0378-4371(00)00311-3

Baack, S. (2024, febrero). *Training data for the price of a sandwich: Common Crawl's impact on generative AI*. Mozilla Foundation.

https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/

OpenAI. (2025, 30 de octubre). *Overview of OpenAI crawlers*. OpenAI Platform Documentation. https://platform.openai.com/docs/gptbot

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web* (Informe técnico n.º 1999-66). Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

Rennie, A., Protheroe, J., & Charron, C. (2020). *Decoding decisions: Making sense of the messy middle*. Google. https://www.thinkwithgoogle.com/_qs/documents/9998/Decoding_Decisions_The_Messy_Middle_of_Purchase_Behavior.pdf

Sánchez-Mosquete, J. (2025). Optimización predictiva en SEO: Explorando modelos predictivos para anticipar tendencias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5130259

SE Ranking. (2026). *Top 100 domains by organic traffic - Spain* [Conjunto de datos]. SE Ranking Analytics.

https://seranking.com/es/top-websites-esp.html

Wang, Z., Zhang, H., Li, C.-L., Eisenschlos, J. M., Perot, V., Wang, Z., Miculicich, L., Fujii, Y., Shang, J., Lee, C.-Y., & Pfister, T. (2024). *Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding*. Proceedings of the Twelfth International Conference on Learning Representations (ICLR), Vienna, Austria. https://doi.org/10.48550/arXiv.2401.04398

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Anexos documental

En cumplimiento de los principios de open science y con el objetivo de garantizar la transparencia metodológica, se adjunta a continuación la documentación técnica y la descripción de los conjuntos de datos empleados. Esta sección tiene como finalidad asegurar la reproducibilidad computacional de los hallazgos y permitir la auditoría externa del flujo de trabajo algorítmico.

Anexo A: Tabla de datos (N=91)

Este apartado hace referencia a la matriz de datos completa utilizada para el análisis de regresión multivariante. El conjunto de datos estructura las métricas brutas de tráfico, autoridad de dominio, volumen de menciones en entornos generativos y la computación final del TGS para la totalidad de la muestra (N=91). Dicho dataset integra variables dimensionales críticas, incluyendo la segmentación porcentual entre intención informativa y transaccional, fundamentales para la validación de la hipótesis del umbral (H_2).

Rank	Dominio	Tráfico Orgánico	Palabras Clave	Domain Trust	Ref. Domains	HC Rank (2025)	Page Rank (2025)	AI Overview (Menciones)	AI Overview (link)	ChatGPT (Menciones)	ChatGPT (link)	Gemini (Menciones)	Gemini (link)	% Infor.	% Trans.	TGS Score
1	youtube	135000000	65000000	100	22000000	6	8	14300000	59900000	63300	9700	62500	530300	64,54	18,09	137,5
2	google com	124000000	28000000	99	31000000	3	1	3100000	5200000	133800	27000	165600	94400	15,47	24,71	184,5
3	wikipedia	110000000	38000000	86	7900000	14	37	7300000	20000000	36300	52100	4000	375100	62,68	12,05	86,4
5	amazon	52000000	24000000	69	133000	386	1366	69400	378200	0	1600	0	381	47,28	43,8	15,4
6	instagram	51000000	40000000	100	27000000	4	5	2000000	858100	50800	745	16700	2700	48,38	19,27	155,3
7	facebook	48000000	48000000	100	42000000	1	3	8000	9900000	2	91000	0	3000	48,51	18,35	312,8
8	eltiempo	42000000	1000000	58	16000	11702	37162	571	49600	0	352	0	928	84,83	2,03	2,2
9	as	31000000	6600000	58	165000	5249	9690	6400	218600	0	3500	0	15200	80,89	7,7	3,7
10	whatsapp	30000000	427000	92	5700000	23	24	4500	68900	0	373	0	2100	80,64	2,04	46,5
11	marca	30000000	4700000	61	121000	9016	11443	38400	88300	8	469	59	45	80,69	6,01	2,8
12	google es	26000000	5600000	69	347000	277	827	1100000	21000	92500	175	106600	34	3,3	35,21	15,9
13	aemet	25000000	839000	60	25000	8431	20084	105	9400	0	239	0	144	67,14	2,62	2,5
15	el mundo	24000000	8500000	62	190000	490	3044	6800	330400	1	562	11	2700	73,68	11,24	12,2
16	pinterest	23000000	18000000	93	8200000	16	22	338700	1100000	6800	9100	4500	201	59,42	27,48	70,8
17	idealista	21000000	3400000	51	27000	8780	20936	9100	161900	1000	1800	403	3100	54,65	13,34	2,7
18	tiktok	20000000	44000000	96	3900000	26	46	1000000	1000000	21000	359	25700	1100	57,44	20,46	58

19	reddit	15000000	29000000	82	1700000	44	120	6400000	17500000	39500	35300	8100	19100	57,95	22,59	42,8
21	elpais	15000000	11000000	63	343000	328	1487	75800	223800	17	3900	0	12500	57,41	18,02	14,8
22	tripadvisor	15000000	7000000	59	37000	7717	6796	165400	152000	274	484	91	2	15,85	34,97	3
26	chatgpt	13000000	120000	65	254000	206	633	28100	1500	1200	125	4400	24	81,41	3,04	13,3
27	apple	13000000	9500000	97	5600000	15	19	2200	3500000	0	31900	0	94200	70,01	10,32	68
28	el corte ingles	12000000	8200000	55	45000	8370	23246	61800	103800	31900	1900	6900	1900	33,33	48,32	2,8
30	microsoft	12000000	4400000	83	3300000	21	38	1600000	3400000	52400	29300	71100	71900	77,55	6,26	54,4
31	loterias y apuestas	12000000	173000	56	5800	499647	68527	275	87	920	534	709	486	74,28	1,3	0,3
32	abc	12000000	6900000	59	168000	2378	5296	216	186500	0	1600	0	268	65,4	14,05	5,3
33	juegos once	11000000	130000	46	3400	449977	371150	110	236	0	23	1	259	64,63	0,75	0,3
34	milanuncios	10000000	6600000	49	33000	8945	54279	254	36000	0	349	0	606	50,85	30,35	2,7
35	booking	9900000	3500000	78	663000	299	496	86000	138900	4900	84	0	376	29,66	20,73	14
36	el confidencial	9500000	5700000	52	67000	5540	9857	45300	162700	124	1400	107	7200	70,54	13,8	3,4
37	mercadona	9500000	1000000	43	5800	8577	87258	99	28600	0	113	0	456	56,49	23,3	2,4
38	leroy merlin	9300000	4500000	42	10000	453654	86013	50300	97900	9100	1700	2700	2500	23,75	65,77	0,4
39	fandom	9200000	7900000	67	438000	737	767	384500	2200000	109	7800	536	18300	71,49	11,77	9,4
40	sport	9200000	1900000	53	67000	8971	35978	16900	13500	8	379	27	119	75,89	8,36	2,4
41	yahoo	8400000	5400000	77	2500000	54	145	632600	1600000	0	11600	7400	1900	59,07	13,88	33,6
42	coches net	8100000	2100000	47	7000	22878	81727	105200	261600	0	177	0	529	70,28	13,09	1,5
43	canva	8000000	2500000	70	538000	92	297	514	262	0	3100	0	6100	71,79	13,19	24,4
44	deepl	8000000	475000	56	56000	2049	3131	74	47300	0	72	0	77	84,99	1,71	4,6
45	20minutos	7900000	7400000	56	153000	1762	6387	955	152700	0	1300	0	2000	73,44	12,14	6
46	mundo deportivo	7900000	5200000	61	82000	20299	27076	23400	56900	994	2400	133	2600	81,75	8,03	1,7
47	directo al paladar	7900000	1800000	54	13000	420016	91.097	13600	30600	72	234	56	4700	77,2	14,11	0,4
48	ok diario	7900000	5200000	51	72000	15919	42647	33500	102000	0	607	18	5700	77,36	9,98	1,9
49	rtve	7800000	4500000	57	164000	4305	6643	81600	194600	0	1300	0	4200	79,72	4,23	3,7
50	tiempo	7800000	555000	61	12000	429983	55477	703	12000	0	147	0	329	88,14	1,33	0,3
51	comunidad Madrid	7300000	3300000	59	26000	7085	16867	345900	283400	12900	686	18100	2100	60,92	13,89	2,8
52	carrefour	7100000	7000000	47	16000	12132	50487	33700	42600	13000	797	6100	730	28,16	55,01	2,3
53	la vanguardia	7000000	7900000	61	159000	4140	5520	77000	203500	636	1700	172	7400	73,81	11,85	3,9
54	caixa bank	6800000	663000	54	9500	23068	29360	67	63800	0	218	0	757	65,03	11,89	1,4
55	wallapop	6600000	6200000	50	47000	8948	46973	66	17900	0	459	0	458	37,75	45,07	2,6
56	zara	6600000	1200000	54	77000	4479	10488	2200	3000	0	9	5	14	61,11	21,26	3,3
57	wordreference	6600000	1900000	52	60000	5376	27971	50500	163500	0	348	0	60	89,22	2,99	3,1
58	spotify	6400000	4900000	80	1900000	45	96	392000	614900	8800	1100	13700	2300	82,9	4,06	36
59	ikea	6300000	2900000	63	290000	1517	2324	59	52300	0	184	0	255	27,76	51,99	6
60	renfe	6300000	732000	54	20000	13142	23011	1800	48900	0	653	0	694	67,67	11,37	1,8

61	imdb	6200000	4400000	83	1000000	142	286	557200	1800000	587	1300	548	152	77,11	5,04	20,1
63	la razon	6100000	4400000	54	51000	6229	17336	425	112000	0	1400	0	5800	70,73	11,15	3
64	filmaffinity	6100000	2200000	53	90000	8648	30260	31100	61400	127	115	136	257	86,39	2,35	2,5
65	elespanol	6100000	8100000	57	92000	5356	8670	869	164600	388	2900	0	3600	64,43	16,48	3,4
66	banco santander	6000000	704000	46	8200	32182	54698	25300	37200	3400	574	1300	1100	66,47	9,73	1,2
67	seg-social	5900000	499000	48	7100	11885	35876	439800	20700	29400	268	62800	1700	69,5	6,46	1,9
68	accuweather	5700000	563000	67	109000	677	2816	17700	19000	2300	461	107	73	74,5	1,82	7,9
69	telecinco	5700000	1500000	56	35000	411606	35672	38200	78300	2800	777	1300	2000	79,18	10,05	0,3
70	netflix	5500000	2600000	63	346000	534	646	381200	295700	11100	1300	19900	1700	77,76	4,5	9,9
71	cookpad	5400000	1500000	57	59000	8744	12319	8500	21400	26	607	0	35	89,73	6,02	2,4
72	prime video	5400000	2300000	61	72000	4656	3950	174300	49400	3900	379	4300	24	81,47	3,19	3,3
73	inta	5200000	16000	38	3000	470584	160934	1300	1400	34	5	30	33	74,98	8,78	0,2
74	agencia tributaria	5000000	996000	50	18000	408131	7990	3	42100	0	1100	0	1800	71,38	5,87	0,3
75	media markt	5000000	3000000	46	5800	419546	72569	17000	21800	8500	652	1700	540	32,64	55,72	0,4
76	decathlon	4900000	2500000	53	8200	425789	57668	10	13900	0	472	0	412	20,46	66,03	0,3
77	madrid	4900000	1700000	53	26000	9234	19506	990900	125200	125600	523	120400	1300	48,71	19,85	2,3
78	la voz de galicia	4800000	2100000	51	45000	8704	24922	11100	25500	56	344	35	1500	63,87	13,53	2,4
79	skyscanner	4800000	958000	42	5300	429416	137863	36500	31200	1400	120	808	252	55,69	17,92	0,3
80	letras	4700000	2300000	60	9400	44773	237059	69	2300	0	0	0	0	92,34	0,61	1,1
81	linkedin	4700000	8800000	99	15000000	9	12	827200	2000000	24400	5400	26100	1100	46,82	12,18	81,9
82	fotocasa	4700000	1400000	42	8700	417798	66798	4100	24800	0	447	0	407	62,32	10,99	0,3
83	el diario	4700000	3800000	65	80000	5040	8370	23200	72100	0	792	0	4100	68,48	11,24	3,3
84	cambridge	4600000	1400000	59	374000	173	778	3	576400	0	15600	0	11500	85,42	3,7	16,5
85	reverso	4600000	2100000	55	43000	4764	16605	96100	173600	85	654	103	58	87,23	2,37	3,2
86	bbc	4500000	3900000	72	757000	97	345	546600	409800	1300	34	2200	52	73,05	6,6	23,6
87	flashscore	4500000	813000	58	4200	651233	1169408	1300	1700	0	65	0	334	82,13	2,02	0,3
89	el correo	4400000	1700000	55	42000	13486	40409	40	23400	0	7	0	36	65,55	12,16	1,9
90	gva	4400000	1700000	62	43000	5019	10341	5900	42100	0	200	0	1100	60,13	11,37	3,1
91	poki	4300000	909000	47	37000	23.873	27007	32	25400	0	17	0	17	71,38	9,1	1,4
92	junta de andalucia	4100000	2500000	61	79000	4390	7829	39600	93900	536	823	702	1900	67,18	9,77	3,4
93	hola	4100000	3300000	48	86000	7000	26782	364	128300	0	1200	0	6200	72,41	17,55	2,7
94	openai	4000000	314000	70	339000	84	265	13000	5000	548	305	930	133	79,91	4,71	21
95	investing	4000000	933000	60	101000	1633	3182	32500	53600	0	706	0	1300	67,92	4,88	5,2
96	lidl	3900000	1500000	44	5700	18691	106463	15400	8600	3700	97	3200	166	40,17	41,8	1,6
97	bbva	3900000	690000	44	8100	408141	41353	101600	86900	5200	701	4400	17	71,28	8,25	0,3
98	la liga	3900000	364000	58	25000	16293	33499	19	6200	49	812	1400	711	84,64	3,07	1,5
99	freepik	3800000	2700000	52	23000	9887	10710	4800	3000	456	101	369	370	80,37	12,73	2,3
100	infojobs	3800000	1100000	51	15000	10235	41483	17200	22000	1800	281	837	298	55,22	12,57	2,1

Anexo B: Algoritmos computacionales y reproducibilidad

Este anexo documenta la totalidad del flujo de trabajo algorítmico utilizado para la obtención de resultados, abarcando desde el preprocesamiento y la ingeniería de características hasta la segmentación no supervisada y la generación de visualizaciones científicas.

Entorno técnico:

- **Lenguaje:** Python 3.10
- **Librerías principales:** pandas, numpy, scikit-learn, statsmodels, seaborn, matplotlib.

B.1. Preprocesamiento y cálculo del TGS (ETL)

El siguiente script detalla el proceso de normalización Logarítmica, diseñado para mitigar la heterocedasticidad derivada de las distribuciones de Ley de Potencia, y la función vectorizada para el cálculo del índice de anclaje topológico.

```
Python

import pandas as pd

import numpy as np


# 1. Carga del Dataset

df = pd.read_csv('dataset_geo_master.csv')


# 2. Ingeniería de Características (Feature Engineering)

# Aplicamos transformación Logarítmica (ln(x+1)) para suavizar la varianza

# de los "gigantes" y corregir la distribución de Pareto.

df['log_Keywords'] = np.log(df['Palabras_Clave'] + 1)

df['log_Traffic'] = np.log(df['Trafico_Organico'] + 1)

df['log_Mentions'] = np.log(df['Total_AI_Mentions'] + 1)
```

```
# 3. Función TGS (índice de anclaje topológico)
# Fórmula: TGS = (ln(K) * ln(T)) / sqrt(HC)
def calcular_TGS(row):
    # Control de excepciones para evitar división por cero
    if row['HC_Rank'] <= 0: return 0

    numerador = row['log_Keywords'] * row['log_Traffic']
    # La raíz cuadrada amortigua la penalización en rangos altos
    denominador = np.sqrt(row['HC_Rank'])

    return numerador / denominador

df['TGS_Score'] = df.apply(calcular_TGS, axis=1)
```

B.2. Segmentación de mercado (K-Means clustering)

Algoritmo de aprendizaje no supervisado empleado para la identificación de los cuatro arquetipos de comportamiento GEO.

```
Python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# 1. Selección de Variables Dimensionales
features = df[['HC_Rank', 'Porcentaje_Informativa']]
```

```
# 2. Estandarización (Z-Score Scaling)
# Paso crítico: Normalizamos las variables para que la escala de Rank (1-100)
# y Porcentaje (0-1) sean comparables euclídeamente.
scaler = StandardScaler()
X_scaled = scaler.fit_transform(features)

# 3. Modelado K-Means
# El hiperparámetro k=4 fue determinado mediante el Método del Codo.
kmeans = KMeans(n_clusters=4, random_state=42, n_init=10)
df['clúster_GEO'] = kmeans.fit_predict(X_scaled)

# 4. Asignación de Etiquetas Semánticas a los clústeres
cluster_labels = {
    0: 'Transaccional Ineficiente (Especuladores)',
    1: 'Challenger Estructural (Aspirantes)',
    2: 'Oráculo (Líder GEO)',
    3: 'Nicho Informativo'
}
df['clúster_Label'] = df['clúster_GEO'].map(cluster_labels)
```

B.3. Modelado predictivo y el umbral de intención

Implementación de regresión polinómica de segundo orden para la validación estadística del punto de inflexión (tipping point) en la curva de visibilidad.

Python

```
import statsmodels.api as sm
from sklearn.preprocessing import PolynomialFeatures

# 1. Preparación de Datos
X = df[['Porcentaje_Informativa']]
y = df['log_Mentions']

# 2. Transformación Polinómica (Grado 2)
# Esto permite al modelo capturar curvas y umbrales no lineales
poly = PolynomialFeatures(degree=2)
X_poly = poly.fit_transform(X)

# 3. Ajuste del Modelo OLS
model_poly = sm.OLS(y, sm.add_constant(X_poly)).fit()

# Un coeficiente positivo significativo en el término cuadrático (x^2)
# confirma estadísticamente el crecimiento exponencial post-umbral.
print(model_poly.summary())
```

Nota sobre la integración temporal (ARIMA/LSTM):

Aunque este anexo presenta el código del análisis transversal, la metodología implementada en producción integra el "TGS_Score" resultante como una variable exógena dentro de modelos SARIMAX, tal y como se detalla en el estudio previo (Sánchez-Mosquete, 2025).

Anexo C: Escenarios de simulación predictiva (impacto GEO)

Siguiendo la metodología de forecasting establecida, se ha aplicado los coeficientes de regresión obtenidos en el modelo logarítmico (R^2=0.82) para proyectar el impacto económico real de las optimizaciones topológicas en dominios críticos del IBEX-35 y retail.

Parámetros de la simulación:

- **Objetivo de optimización (target):** Alcanzar un HC Rank de 15 (estructura de Hub eficiente) y una intención informativa del 50%.
- **Modelo matemático de proyección:**

$$\Delta Y = Y_{actual} xe(\beta HC x \Delta ln(HC) + \beta_{Info} x \Delta Info)$$

C.1. Tabla de proyección de impacto

Dominio	HC Actual	HC Target	Crecimiento Proyectado (Elasticidad β=−0.25)	Menciones Proyectadas
Banco Santander	32182	15	**+580%**	de 30k a **204k**
Zara	4479	15	**+316%**	de 2.2k a **9.1k**
Amazon	386	15	**+125%**	de 69k a **156k**
Ikea	1517	15	**+217%**	de 59 a **187**

Fuente: Elaboración basada en coeficientes de elasticidad $\beta_{HC} = -0.25$ *y* $\beta_{Info} = -0.02$*, asumiendo un escenario ceteris paribus donde solo se optimiza la variable de centralidad armónica.*

C.2. Interpretación estratégica de los escenarios

1. El caso del banco Santander (el coste de la profundidad)

El modelo predice un crecimiento explosivo (+580%) para la entidad financiera. El diagnóstico revela que su estructura web actual es extremadamente profunda (HC Rank > 30.000), lo que provoca que sus contenidos de alto valor financiero presenten una baja detectabilidad para los algoritmos de recuperación de la IA.

Una reestructuración del portal hacia una arquitectura de topic clusters, similar a la implementada por otras entidades del sector como CaixaBank, permitiría desbloquear una visibilidad latente masiva, logrando una posición dominante en las respuestas sobre préstamos e hipotecas en sistemas de IA generativa.

2. El caso Amazon (el punto de saturación estructural)

La simulación arroja un crecimiento más moderado para Amazon (+125%). Esto se debe a que su inmensa autoridad bruta ya compensa parcialmente su ineficiencia estructural. Sin embargo, duplicar su visibilidad en IA significaría capturar millones de consultas de descubrimiento (Discovery Queries) del tipo cuál es el mejor móvil calidad-precio, segmento de tráfico que actualmente presenta una mayor afinidad con estructuras de blogs especializados y afiliados.

C.3. Validación del ROI

La modelización matemática demuestra que, bajo una condición ceteris paribus, la elasticidad del ROI frente a la corrección arquitectónica es significativamente superior a la inversión en generación lineal de contenido o en adquisición de enlaces externos (link building).

Para marcas con alta autoridad nominal pero ineficiencia topológica (v.g., High Authority / Low Structure), el coste de intervención es de naturaleza técnica (Capital Expenditure - CAPEX de reorganización), mientras que el beneficio en visibilidad sigue una función exponencial en el nuevo canal de búsqueda. Esto sugiere que la remediación estructural ofrece una tasa interna de retorno (TIR) superior a la acumulación de activos, validando la

hipótesis de que la eficiencia del grafo es el multiplicador de fuerza en la economía de la IA.

