How AI Platforms Search: Fan-Out Query Behavior Across Intent Types, Verticals, and Platforms

Abstract

When users submit queries to AI search platforms, the platforms do not pass the user's text to web search verbatim. They decompose each prompt into multiple internal "fan-out queries" — the actual strings sent to retrieval engines. These fan-out queries determine which pages get fetched, which enter the AI's context window, and which get cited in the response. Despite their centrality to AI search discoverability, fan-out queries have not been studied at scale.

This study classifies 1,323 fan-out queries generated by 540 parent queries across three AI platforms (ChatGPT, Gemini, Perplexity), ten commercial verticals, and five intent types. We capture fan-out queries via the OpenAI Responses API, Google GenAI grounding metadata, and Perplexity browser-level SSE interception.

Nine findings emerge. First, user intent is a significant predictor of fan-out composition ($X^2=299.6$, $p<0.001$, $V=0.24$): discovery queries trigger 3.3x the entity injection rate of informational queries. Second, platforms exhibit distinct retrieval personalities — ChatGPT injects entities from training data on 32% of fan-outs, Gemini casts a wide net with 27% expansion queries, and Perplexity leads in evidence-seeking at 21%. Third, ChatGPT's search trigger rate varies dramatically by model tier: gpt-5.4 searches on only 29% of queries while gpt-5.4-nano searches on 100%, suggesting larger models are more confident in answering from training data alone. Fourth, platform-intent interaction effects explain fan-out variation better than either factor alone (two-way AIC=192 vs main-effects AIC=937). Fifth, situation-first query phrasing produces significantly different fan-out distributions than standard phrasing ($V=0.35$, $p<0.001$). Sixth, no significant vertical effect was detected at this sample size ($H=6.26$, $p=0.71$, 18 queries per vertical), suggesting intent and platform are the dominant factors. Seventh, replicate analysis on ChatGPT (gpt-5.4-mini, 3 replicates) reveals that the search trigger decision is highly deterministic (91.7% agreement) while the specific fan-out query strings are almost entirely stochastic (98% zero overlap) — but the structural *type* of fan-out is moderately stable (65% top-type agreement).

These findings establish that AI search operates a two-layer retrieval system: a model-confidence layer that decides whether to search at all, and a query-decomposition layer that determines what to search for. Optimising for AI citation requires understanding both layers.

1. Introduction

The way people search is changing. Semrush's analysis of 17 months of ChatGPT clickstream data (October 2024 — February 2026; 1B+ sessions) found that 65—85% of user prompts do not match any keyword in their 27-billion-keyword database (Semrush, 2026). Average prompt length for search-enabled queries grew from 4.7 to 8.7 words. Users are speaking naturally to AI assistants rather than composing keyword-style search queries.

Yet the AI platforms themselves must still retrieve web content. ChatGPT, Gemini, and Perplexity all augment their language models with live web search capability. When they search, they do not pass the user's raw conversational prompt to a search engine. Instead, they generate their own internal search queries — which we term **fan-out queries** — that serve as the actual retrieval strings sent to web search

APIs.

Fan-out queries are the hidden translation layer between human language and web retrieval. A user who asks "I just got back from the beach and now I have brown patches on my cheeks — what should I try?" might trigger fan-out queries like "vitamin C serum dark spots," "hyperpigmentation treatment face," and "SkinCeuticals CE Ferulic." The user never typed those keywords, and those brands. The AI generated them.

This translation layer has significant implications for content discoverability. A website that ranks well for the user's conversational phrasing may be invisible to the AI if the AI never generates a fan-out query that retrieves it. A brand that is well-represented in the AI's training data may be proactively searched for — injected into fan-out queries the user never requested — while competitors without such representation are structurally excluded from retrieval.

Prior work on AI citation has focused on what gets cited (Lee, 2026a; Aggarwal et al., 2024) and what page-level and domain-level features predict citation (Lee, 2026b; our Experiments K, L, M). This study shifts the focus upstream: **how do AI platforms discover pages in the first place?** By capturing and classifying the fan-out queries that drive retrieval, we can characterise the retrieval strategies of different platforms, identify how user intent shapes retrieval behaviour, and quantify the relative importance of training-data knowledge versus live web search in determining which brands and pages enter the AI's context window.

Contributions

1. The first large-scale, cross-platform taxonomy of AI fan-out query behaviour, with 1,323 classified fan-outs across three platforms, ten verticals, and five intent types.
2. Empirical evidence that ChatGPT's search trigger rate varies by model tier (29% for gpt-5.4 vs 100% for gpt-5.4-nano), revealing a model-confidence layer that mediates between training-data responses and live web retrieval.
3. Statistical evidence that intent is a significant driver of fan-out composition ($X^2=299.6$, $p<0.001$), while no vertical effect was detected at this sample size ($H=6.26$, $p=0.71$, $n=18$ per vertical).
4. Quantification of entity injection rates by platform and intent, showing that 99%+ of entity injections originate from training data at fan-out position 0—1 before any retrieval results have returned.
5. Evidence that query format (standard vs situation-first) produces significantly different fan-out distributions ($V=0.35$), establishing a measurable citation gap between keyword-based and conversational AI use.
6. Replicate analysis showing the search decision is deterministic (91.7% stable) while fan-out strings are stochastic (98% zero overlap across runs), but fan-out *types* are moderately stable (65%), establishing that structural retrieval strategy — not specific query strings — is the reliable optimisation target.

2. Related Work

2.1 AI Citation Behaviour

Research on AI recommendation systems has established that citation behaviour varies by platform and domain type. The AIVO Evidentia working paper (2026) proposed an 8-filter decision-path taxonomy for how AI systems narrow recommendations across a buying conversation, identifying filters like

Clinical Evidence Binary and Value Reframing that operate at the brand level. Our prior Experiment K demonstrated that domain identity — not page-level features — is the dominant predictor of AI citation, with domain citation rate alone achieving AUC=0.9746. Experiment L showed that only 34.9% of domains in a broad sample (n=4,658) receive any AI citation at all.

2.2 AI Search Architecture

AI platforms implement web search augmentation through different architectural patterns. ChatGPT routes queries through Bing's search API; Perplexity operates its own search index with parallel retrieval streams (web, memory, workflow engines); Gemini uses Google Search grounding. Semrush (2026) documented the divergence between user prompt language and traditional search keywords but did not examine what the AI platforms search for internally.

Our prior Study 2 (Lee, 2026b) provided the first fan-out query taxonomy, classifying 2,158 fan-outs from 235 parent queries across three platforms and three commercial verticals. That study identified seven structurally distinct fan-out types dominated by compression (37.9%) and entity lookup (18.0%). However, Study 2 was limited by uneven platform coverage (Perplexity contributed 68% of fan-outs), narrow vertical scope (three client verticals), rule-based classification at approximately 90% accuracy, and synthetic parent queries. The current study addresses all of these limitations with a balanced 10-vertical x 5-intent design, cross-platform API-level capture, and a substantially larger query corpus.

2.3 Query Decomposition in AI Search Systems

Recent academic work has formalised the query decomposition process that we observe empirically. Zhong et al. (2025) introduce ReDI, a three-stage pipeline that decomposes complex queries into sub-queries, augments each with semantic interpretation, and fuses retrieval results — demonstrating that explicit decomposition improves nDCG@10 across diverse benchmarks. Tang et al. (2025) describe Xinyu AI Search, which employs a Query-Decomposition Graph (QDG) to dynamically break down queries into sub-queries for stepwise retrieval, outperforming eight existing systems in human assessments. Zhao et al. (2025) propose ParallelSearch, which uses reinforcement learning to train LLMs to identify parallelisable query structures and execute concurrent sub-queries, achieving 12.7% performance gains on parallelisable queries. Li et al. (2025) present the "AI Search Paradigm," a multi-agent framework where a Planner agent decomposes queries into a directed acyclic graph (DAG) of sub-tasks.

These systems propose and evaluate query decomposition as an architectural component. Our work is complementary: rather than proposing a decomposition method, we empirically observe the decomposition behaviour of deployed commercial systems (ChatGPT, Gemini, Perplexity) and classify the structural types of fan-out queries they generate. Our finding that fan-out composition varies systematically by intent and platform suggests that deployed systems implement intent-sensitive decomposition strategies, consistent with the architectural goals of ReDI and Xinyu but realised through different (and proprietary) mechanisms. Our replicate analysis further reveals that these strategies produce stochastic outputs at the string level (98% zero overlap) despite structural stability — a property not examined in the architectural literature.

2.4 The "GEO is Just SEO" Debate

A parallel industry debate concerns whether "Generative Engine Optimisation" (GEO) is a genuine discipline or repackaged SEO. Proponents of the distinctiveness position argue that AI recommendation systems use fundamentally different signals than search engines (Aggarwal et al., 2024; various GEO agencies). Critics, notably WebLinkr (2026) and Sturm (2026), argue that AI platforms perform live web

searches using keyword-style queries, and therefore traditional SEO skills — ranking well in Google — determine AI citation. Our Study 2 data supported the critics' position: the AI's retrieval layer compresses conversational prompts back into keyword-style queries at a ratio of 0.54. The current study provides a more nuanced answer.

3. Methodology

3.1 Query Corpus

We constructed a balanced query corpus of 180 queries spanning 10 verticals (Technology, Health, Finance, Travel, Home Improvement, Education, Legal, Food, Ecommerce, Fitness) and 5 intent types.

Table 1: Intent Type Definitions

Intent Type	Definition	Example
INFORMATIONAL	Seeking knowledge or explanation	"how does container orchestration work"
DISCOVERY	Seeking product or service recommendations	"best password manager 2026"
VALIDATION	Seeking confirmation about a specific entity	"is NordVPN worth paying for"
COMPARISON	Seeking direct comparison between entities	"Trello vs Monday.com"
REVIEW_SEEKING	Seeking evaluative opinions or experiences	"honest opinions on GitHub Copilot from daily users"

The primary corpus consists of 150 queries (10 verticals x 5 intents x 3 queries per cell). An additional 30 queries provide situation-first reformulations (1 per vertical x 3 intents: DISCOVERY, COMPARISON, INFORMATIONAL), in which the same purchase intent is expressed as a personal situation description (20—35 words) without naming the product category. For example, the DISCOVERY query "best password manager 2026" is paired with: "I have about 40 different passwords scribbled in a notes app and I just got a notification that my email was in a data breach. I need to sort this out properly. What are my options?"

3.2 Data Collection

Fan-out queries were captured across three platforms using platform-appropriate methods:

ChatGPT was queried via the OpenAI Responses API with the `web_search` tool enabled and `tool_choice="auto"`. When ChatGPT decides to search, each API response contains `web_search_call` output items whose `action.queries` field lists the internal search queries generated. Three model tiers were used across the collection: gpt-5.4-nano (44 queries, all triggered search), gpt-5.4-mini (46 queries, all triggered search), and gpt-5.4 (90 queries, 29% triggered search). This model variation was not planned as an experimental variable but arose from iterative collection — the resulting model-tier comparison is reported as an exploratory finding.

Gemini was queried via the Google GenAI API using the `gemini-3.1-flash-lite-preview` model with Google Search grounding enabled (`types.Tool(google_search=types.GoogleSearch())`). Fan-out queries are exposed in the response's `candidate.grounding_metadata.web_search_queries` field. Citations are extracted from `grounding_chunks`.

Perplexity was scraped via browser-level SSE (Server-Sent Events) interception on a Pro account. A Playwright-based scraper injects a JavaScript fetch interceptor that captures the streaming response data. Fan-out queries appear in `workflow_root` SSE blocks under `queries_payload.queries`, with each search step yielding a separate set of sub-queries. Citations are extracted from `web_results` SSE blocks.

Each of the 180 queries was submitted to all three platforms (540 total queries). Perplexity and Gemini triggered web search on 94% and 97% of queries, respectively. ChatGPT's trigger rate was 64% overall, varying dramatically by model tier.

Table 2: Collection Summary

Platform	Method	Queries	Search Triggered	Fan-Outs	Citations
ChatGPT	OpenAI Responses API	180	116 (64%)	415	617
Gemini	Google GenAI API	180	174 (97%)	567	1,686
Perplexity	Browser SSE interception	180	169 (94%)	417	2,065
Total		**540**	**459 (85%)**	**1,399**	**4,368**

After cleaning (removing fan-outs shorter than 5 characters, exact echoes of parent queries, and encoding artefacts), 1,323 fan-out records remained for classification.

3.3 Fan-Out Classification

Each fan-out query was classified by its structural relationship to the parent query using a rule-based classifier with the following nine-type taxonomy:

1. **Compression.** Fan-out word count is <=60% of parent AND shares >=2 substantive words. The AI has distilled a conversational query into keyword-style search terms.
2. **Entity Injection (Training).** Fan-out introduces capitalised entity names (brands, products, tools) absent from the parent query, at fan-out position 0—1 (before any retrieval results have returned). The entity originates from the model's training data.
3. **Entity Injection (Retrieval).** Same as above, but at position 2+ where the entity could have been observed in prior retrieval results.
4. **Expansion.** Short contextual lookup providing background knowledge (e.g., "marketing budget," "skincare ingredients").
5. **Reformulation.** High word overlap (>60%) with the parent — same intent, different phrasing.
6. **Tangential.** Low word overlap (<30%) with >=3 words — the AI has moved to an adjacent topic.
7. **Narrowing.** Adds specificity absent from the parent: year qualifiers, audience segments, price constraints.
8. **Evidence Seeking.** Introduces evidence-related terms (review, clinical, study, benchmark) absent from the parent.
9. **Price/Availability.** Introduces commerce terms (price, buy, purchase, shipping) absent from the parent.

Classification was performed programmatically using regex-based heuristics for word overlap, entity detection (capitalisation patterns), and signal-term matching. The classifier processes rules in priority order (evidence seeking and price/availability first as rare high-specificity categories, entity injection second, then compression, reformulation, tangential, expansion as residual). Based on manual spot-check

of 100 randomly sampled records, the classifier produces reasonable assignments for approximately 88% of cases.

Boundary analysis. 403 of 1,323 records (30.5%) fall near a classification boundary where small changes in word overlap or entity detection thresholds would shift the assignment. The most common confusions are expansion/tangential (46 cases), expansion/reformulation (43), and entity injection/tangential (41). High-specificity categories (evidence seeking, price/availability, entity injection) are robust — their signal terms are unambiguous. The primary vulnerability is the expansion/tangential boundary, which depends on a 30% word-overlap threshold.

Sensitivity analysis. To assess the impact of potential misclassification, we simulated reassigning 12% of records (~158) from their assigned type to their most likely alternative (based on proximity to classification boundaries). The entity injection, evidence seeking, compression, narrowing, and price/availability categories were stable (less than 0.2 percentage point change each). The expansion and tangential categories shifted substantially: expansion dropped from 16.0% to 8.8%, tangential rose from 17.2% to 24.3%. All findings that depend on entity injection, evidence seeking, compression, or narrowing rates are robust to 12% misclassification. Findings involving expansion or tangential rates should be interpreted with the understanding that these two categories have a porous boundary.

3.4 Entity Extraction

Entities injected by AI platforms into fan-out queries were extracted using spaCy (en_core_web_sm) named entity recognition, filtered to ORG, PRODUCT, PERSON, and GPE types. Entities present in the parent query were subtracted. Provenance was classified as "training" (entity appears at fan-out position 0—1) or "retrieval" (position 2+). Entity names were normalised to canonical forms using a manually curated alias table.

3.5 Citation-Fan-Out Linkage

For Perplexity, we linked citations to the fan-out queries that produced them using keyword overlap between fan-out query tokens and citation page titles/URLs (minimum 20% Jaccard overlap threshold). This heuristic linkage produced 1,816 citation-fan-out pairs. ChatGPT's API does not expose per-search-call retrieved URLs, precluding full linkage; Gemini's grounding metadata provides citations but does not associate them with specific search queries.

3.6 Statistical Analysis

- **Intent -> fan-out type:** Chi-squared test of independence with Cramer's V effect size and adjusted standardised residuals.
- **Platform -> fan-out type:** Same.
- **Intent x platform interaction:** Log-linear models comparing main-effects, all-two-way, and saturated models via AIC.
- **Purchase vs knowledge intent:** Mann-Whitney U on fan-out counts and compression ratios; chi-squared on entity injection rates.
- **Vertical effects:** Kruskal-Wallis H-test on fan-out counts per query across 10 verticals.
- **Format sensitivity:** Chi-squared on fan-out type distribution for standard vs situation-first queries.

All tests use alpha=0.05. Multiple-comparison correction was not applied to the primary analyses as each addresses a distinct research question.

4. Results

4.1 Fan-Out Type Distribution

Table 3: Overall Fan-Out Type Distribution (n=1,323)

Type	Count	%
Tangential	227	17.2%
Expansion	212	16.0%
Evidence Seeking	206	15.6%
Entity Injection (Training)	206	15.6%
Compression	192	14.5%
Narrowing	127	9.6%
Reformulation	93	7.0%
Price/Availability	59	4.5%
Entity Injection (Retrieval)	1	0.1%

The distribution differs markedly from Study 2, where compression dominated at 37.9% and evidence seeking was negligible at 0.6%. The most notable shift is the rise of evidence seeking from 0.6% to 15.6%. To investigate whether this is driven entirely by the REVIEW_SEEKING intent type (absent from Study 2's design), we computed the evidence-seeking rate excluding REVIEW_SEEKING queries: 13.5% (146/1,079). Evidence seeking rates are substantial across all purchase-oriented intents — VALIDATION (17.5%), COMPARISON (15.3%), and DISCOVERY (15.2%) — and low only for INFORMATIONAL (3.1%). The Study 2 finding that evidence seeking is near-zero appears to have been an artefact of its narrow query corpus (three client verticals with synthetic commercial queries), not a general property of AI retrieval. When diverse intent types are represented, AI platforms search for evidence substantially more often.

The relative decline of compression from 37.9% to 14.5% is partially explained by the shift in capture method (see Section 5.3).

4.2 Finding 1: Intent Shapes Fan-Out Composition (RQ1)

Fan-out type distribution varies significantly by user intent ($X^2=299.6$, $df=32$, $p<0.001$, Cramer's $V=0.24$).

Table 4: Significant Intent x Type Cells (|adjusted residual| > 2.0)

Intent x Type	Residual	Interpretation
DISCOVERY x Entity Injection (Training)	+7.8	AI injects brands when users shop
INFORMATIONAL x Compression	+7.6	AI compresses to keywords for knowledge queries
DISCOVERY x Narrowing	+6.7	AI adds filters (year, price) for shopping
DISCOVERY x Tangential	-5.3	AI stays on-topic when shopping
INFORMATIONAL x Evidence Seeking	-5.1	AI doesn't seek proof for knowledge queries
COMPARISON x Price/Availability	+4.8	AI checks prices when comparing
REVIEW_SEEKING x Evidence Seeking	+4.3	AI seeks reviews when users want opinions

Intent x Type	Residual	Interpretation
INFORMATIONAL x Entity Injection	-4.3	AI doesn't inject brands for knowledge queries

The pattern is coherent: when users are shopping (DISCOVERY), the AI's dominant retrieval strategy is to inject brand names from its training data and add specificity filters. When users are learning (INFORMATIONAL), the AI compresses to keywords and searches for explanatory content. When users want opinions (REVIEW_SEEKING), the AI searches for evidence and reviews. The AI adapts its retrieval strategy to the user's intent.

4.3 Finding 2: Platform Retrieval Personalities

Fan-out type distribution varies significantly by platform ($X^2=386.9$, $df=16$, $p<0.001$, Cramer's $V=0.38$). Each platform exhibits a distinct retrieval personality.

Table 5: Fan-Out Type Distribution by Platform

Type	ChatGPT	Gemini	Perplexity
Entity Injection (Training)	**32.0%**	4.0%	9.8%
Evidence Seeking	**23.1%**	5.7%	**20.5%**
Tangential	12.4%	**21.3%**	16.1%
Expansion	4.9%	**27.2%**	13.7%
Compression	7.5%	**18.7%**	**18.9%**
Narrowing	9.5%	7.3%	13.1%
Reformulation	5.8%	7.7%	8.5%
Price/Availability	4.9%	7.9%	0.0%

- **ChatGPT is an entity injector.** 32% of its fan-outs introduce specific brand or product names the user never mentioned. When a user asks a generic category question, ChatGPT's dominant retrieval behaviour is to decide which brands are relevant — from its training data — and search for those brands specifically.
- **Gemini is an explorer.** 27% expansion and 21% tangential fan-outs indicate Gemini casts a wide contextual net, searching for background knowledge and adjacent topics rather than targeting specific entities.
- **Perplexity is an evidence seeker.** 21% evidence-seeking fan-outs (the highest rate) combined with 19% compression suggests Perplexity prioritises finding proof and reviews, compressed to efficient keyword searches.

Important caveat: The ChatGPT data combines three model tiers. Because model tier affects both search trigger rate and fan-out composition, we report the breakdown:

Table 5b: ChatGPT Fan-Out Type Distribution by Model Tier

Type	gpt-5.4 (n=84)	gpt-5.4-mini (n=125)	gpt-5.4-nano (n=203)
Entity Injection (Training)	25.0%	16.8%	**44.8%**
Evidence Seeking	21.4%	**28.0%**	21.2%
Tangential	17.9%	9.6%	11.8%
Price/Availability	11.9%	14.4%	3.9%
Narrowing	10.7%	16.8%	7.9%
Expansion	7.1%	8.0%	4.9%

Type	gpt-5.4 (n=84)	gpt-5.4-mini (n=125)	gpt-5.4-nano (n=203)
Compression	4.8%	4.8%	2.0%
Reformulation	1.2%	0.8%	3.4%

The entity injection personality is strongest in gpt-5.4-nano (44.8%) and weakest in gpt-5.4-mini (16.8%). The flagship gpt-5.4 falls in between (25.0%) but produces fan-outs for only 29% of queries — its 25% entity injection rate applies to a much smaller denominator. The composite 32% figure in Table 5 is heavily weighted by gpt-5.4-nano, which contributed 49% of all ChatGPT fan-outs.

These findings partially replicate Study 2's platform personality characterisation. ChatGPT's entity injection dominance is confirmed across all three model tiers, though the magnitude varies (17—45%). Perplexity's compression dominance from Study 2 (42%) is not replicated — the current study shows Perplexity as more evidence-oriented, which may reflect architectural changes between the Study 2 observation period (March—April 2026 browser scraping with free accounts) and the current data (April 2026 Pro accounts with SSE interception).

4.4 Finding 3: ChatGPT's Model-Tier Search Behaviour

An unplanned but striking finding: ChatGPT's propensity to search the web varies dramatically by model tier.

Table 6: ChatGPT Search Trigger Rate by Model

Model	Queries	Searched	Rate	Fan-Outs/Query
gpt-5.4-nano	44	44	**100%**	4.7
gpt-5.4-mini	46	46	**100%**	2.7
gpt-5.4	90	26	**29%**	3.2

The flagship model (gpt-5.4) searches the web on only 29% of queries — even with `tool_choice="auto"` and the `web_search` tool explicitly available. The smaller models search on every query. This suggests a **model-confidence threshold**: larger models, with more extensive training data, are more likely to answer from parametric knowledge without triggering retrieval. The implications are significant — content that is not in gpt-5.4's training data may be effectively invisible to the most capable model, because it never initiates the web search that would discover it.

Table 7: ChatGPT Search Trigger Rate by Intent (All Models Combined)

Intent	Searched	Rate
DISCOVERY	39/40	**98%**
COMPARISON	29/40	72%
REVIEW_SEEKING	22/30	73%
VALIDATION	21/30	70%
INFORMATIONAL	5/40	**12%**

The intent effect compounds the model effect: gpt-5.4 asked an informational question will almost never search the web. This creates a two-dimensional retrieval gap — queries that are both informational in intent and directed at the flagship model have near-zero probability of triggering live web retrieval.

4.5 Finding 4: Intent x Platform Interaction (RQ3)

Log-linear model comparison shows that two-way interactions (AIC=192) provide substantially better fit than main effects alone (AIC=937), while the saturated three-way model (AIC=270) does not improve over the two-way model.

This means the way intent affects fan-out behaviour **depends on platform**, and vice versa — but the three-way interaction (intent x platform x fan-out type) does not add further explanatory power. For practitioners, this means platform-specific intent strategies are warranted, but the intent effects are consistent enough across verticals that vertical-specific strategies are unnecessary.

4.6 Vertical Effects: Insufficient Power to Conclude

Fan-out count per query does not differ significantly across the 10 verticals (Kruskal-Wallis H=6.26, df=9, p=0.71). However, this null result should not be interpreted as evidence of absence. With only 18 queries per vertical, the test has approximately 50% power to detect a medium effect size (eta-squared=0.06) at alpha=0.05. Achieving 80% power would require approximately 40—50 queries per vertical (400—500 total).

We therefore do not claim that vertical has no effect. We report only that no significant vertical effect was detected at this sample size, and that the strong effects of intent (V=0.24, p<0.001) and platform (V=0.38, p<0.001) dominate the variance we can measure. Whether verticals with different information densities (e.g., health with clinical literature vs. food with recipe content) produce different fan-out distributions remains an open question requiring a study with substantially more queries per vertical.

4.7 Finding 6: Entity Injection Provenance (RQ2)

Of 317 fan-outs with entity injections (367 total entity instances), **99.4% were classified as training-sourced** — the entity appeared at fan-out position 0—1 before any retrieval results had returned. Only 2 entity injections (0.6%) occurred at position 2+ where they could have been triggered by prior retrieval results.

Brand concentration across verticals is low (HHI ranging from 0.037 to 0.103), indicating the AI distributes entity injections across many brands rather than concentrating on a few dominant ones. The Travel vertical shows the highest concentration (HHI=0.076), driven by hotel loyalty programs (Marriott Bonvoy, Hilton Honors); the Food vertical shows the highest single-brand dominance (Vitamix at 8 injections, HHI=0.103).

ChatGPT injects by far the most entities (183 unique brands vs 80 for Perplexity and 30 for Gemini), and entity injection correlates strongly with intent: the injection rate for purchase-oriented queries (DISCOVERY, COMPARISON, REVIEW_SEEKING) is **19.8%** vs **6.0%** for knowledge-oriented queries (INFORMATIONAL, VALIDATION).

No self-references were detected — none of the platforms injected their own brand name into fan-out queries. This is likely a structural constraint rather than a neutrality signal: ChatGPT's web search tool queries Bing, not OpenAI's own index, so searching for "ChatGPT" would return marketing pages rather than useful results. The absence of self-reference does not indicate the platforms are unbiased in their entity injection — it indicates their search tools route to external search engines.

4.8 Finding 7: Format Sensitivity

Situation-first query phrasing produces significantly different fan-out distributions than standard phrasing (X2=159.7, df=8, p<0.001, Cramer's V=0.35). This is a medium-to-large effect: the way a user phrases the same underlying intent changes which fan-out types the AI generates.

This confirms a measurable **citation gap** between keyword-based and conversational AI use. Current SEO and GEO auditing tools, which track keyword-style queries, miss the retrieval paths that situation-first prompts trigger. A brand that is discoverable via "best password manager 2026" may not be discoverable via "I have 40 passwords in a notes app and just got a data breach notification — what should I do?"

4.9 Finding 8: Citation-Fan-Out Linkage (RQ5)

Heuristic linkage of Perplexity citations to their source fan-out types (n=1,816 linkage pairs) reveals differential citation yield by fan-out type:

Table 8: Citation Yield by Fan-Out Type (Perplexity)

Fan-Out Type	Citations Linked	Fan-Outs	Yield
Compression	285	192	148%
Narrowing	164	127	129%
Evidence Seeking	264	206	128%
Entity Injection (Training)	195	206	95%
Reformulation	85	93	91%
Price/Availability	48	59	81%
Tangential	153	227	67%
Expansion	101	212	48%

Yield rates exceed 100% because a single fan-out query can produce multiple citations. Compression queries produce the highest citation yield (148%), while expansion queries — background contextual lookups — produce the lowest (48%). This suggests that compression fan-outs are the most efficient retrieval path: they map directly to keyword-ranked pages that the AI then cites. Expansion fan-outs, by contrast, retrieve background context that informs the response without generating direct citations.

Citation exclusivity analysis shows that 98% of citations are exclusive to a single fan-out type (only 32 of 1,816 linkage pairs involve citations reachable through multiple fan-out types). This means **different fan-out types access different pages** — a brand visible through compression queries is not necessarily visible through entity injection queries, and vice versa.

4.10 Finding 9: Replicate Analysis — Stable Decisions, Stochastic Queries

To assess the reliability of fan-out behaviour, we submitted all 180 queries to ChatGPT (gpt-5.4-mini) three times — the original collection plus two additional replicates. Comparing replicates 2 and 3 (both gpt-5.4-mini, identical conditions):

Table 9: Replicate Consistency (gpt-5.4-mini, rep2 vs rep3, n=180 query pairs)

Metric	Value
Search trigger agreement	**91.7%** (165/180 queries agree on whether to search)
Fan-out string Jaccard (when both searched)	**0.012** mean (98% of pairs share zero exact queries)
Top fan-out type match (when both searched)	**65.1%** (54/83 pairs produce the same dominant type)
Citation domain Jaccard	**0.339** mean (roughly one-third of cited domains overlap)

The search decision (Layer 1) is highly deterministic: 91.7% of the time, the same query either triggers or does not trigger web search across replicates. The intent-level pattern is stable across replicates —

DISCOVERY triggers search at 92%, INFORMATIONAL at 6%.

The fan-out query strings (Layer 2) are almost entirely stochastic: 98% of query pairs share zero exact fan-out strings. When the AI decides to search for "best password manager 2026," it generates a fresh set of sub-queries each time — different strings, different entity names injected, different narrowing terms added. However, the *structural type* of those queries is moderately stable: 65% of the time, the dominant fan-out type (entity injection, compression, evidence seeking, etc.) is the same across replicates.

Citation domains show moderate overlap at Jaccard 0.339, consistent with our prior finding (Experiment M) that 30—60% of AI citation sources vary between sessions.

The practical implication is precise: **optimise for the fan-out type your intent triggers, not for specific fan-out strings.** The AI will search for "Bitwarden 1Password review 2026" one time and "NordPass Dashlane comparison security" the next — different strings, but both are entity-injection fan-outs retrievable by the same content strategy. Tracking specific fan-out query strings (as some GEO tools do) captures a snapshot of one stochastic sample, not a stable retrieval signal.

Generalisability note. These citation yield findings are based entirely on Perplexity data. ChatGPT's API does not expose per-search-call retrieved URLs, and Gemini's grounding metadata does not associate citations with specific search queries. The yield rates in Table 8 should not be assumed to hold for ChatGPT or Gemini, whose different retrieval architectures may produce different citation-to-fan-out ratios.

5. Discussion

5.1 The Two-Layer Retrieval Structure

The central contribution of this study is empirical evidence for a **two-layer retrieval structure** in AI search:

Layer 1: Search Decision. The AI decides whether to search the web at all. This decision is mediated by model tier (gpt-5.4 searches 29% of the time; gpt-5.4-nano searches 100%) and by intent (INFORMATIONAL triggers search 12% of the time on ChatGPT; DISCOVERY triggers 98%). The replicate analysis confirms this layer is highly deterministic: 91.7% of queries produce the same search/no-search decision across independent runs. Content that exists only on the web — not in the model's training data — is invisible to queries that never trigger search.

Layer 2: Query Decomposition. When the AI does search, it decomposes the user's prompt into fan-out queries whose composition is shaped by intent and platform. Discovery prompts trigger entity injection (the AI pre-selects which brands to investigate). Informational prompts trigger compression (the AI distils to keywords). Review-seeking prompts trigger evidence search (the AI looks for reviews and studies). The replicate analysis reveals that this layer is **structurally stable but lexically stochastic**: the AI generates entirely different search strings each time (98% zero overlap), but the *type* of search it performs is consistent 65% of the time. This has a direct practical consequence — optimising for a specific fan-out query string is chasing a moving target, while optimising for the fan-out type your intent triggers is a stable strategy.

Formal notation. Let Q denote a parent query, I(Q) its intent type, M the model tier, and V the vertical. Layer 1 is a binary search decision function:

S(Q, M, I) -> {0, 1}

where S=1 indicates the model triggers web search. Our data estimates P(S=1 | M=gpt-5.4, I=INFORMATIONAL) ~ 0.03, while P(S=1 | M=gpt-5.4-nano, I=DISCOVERY) ~ 1.0. Replicate analysis confirms S is near-deterministic: P(S_rep2 = S_rep3) = 0.917.

When S=1, Layer 2 generates a set of fan-out queries F = {f_1, ..., f_k} where each f_i has a structural type T(f_i) in {compression, entity_injection, evidence_seeking, narrowing, expansion, tangential, reformulation, price_availability}. The type distribution P(T | I, M, platform) is the empirical finding of this study. The specific strings f_i are stochastic (P(f_i^rep2 = f_j^rep3) ~ 0.02 for any i, j), but the type distribution is moderately stable: the modal type T* matches across replicates with probability 0.65.

The page retrieval set R = Union(retrieve(f_i)) determines which pages enter the context window. Our citation linkage shows that different types access different pages: 98% of citation-fan-out pairs are exclusive to a single type.

This two-layer structure resolves the "GEO is just SEO" debate. The critics (Sturm, WebLinkr) are correct that Layer 2 operates through keyword-style web search — traditional SEO skills matter for being retrieved once the AI searches. But they understate Layer 1: the model's decision to search at all, and its pre-selection of which entities to search for (via training-data entity injection), are not addressable through SEO. A brand that is not in the model's training-data entity map will not be entity-injected into fan-out queries, regardless of how well its pages rank in Google.

5.2 Practical Implications

For content strategy. The intent-specific fan-out patterns suggest differentiated optimisation:

- For DISCOVERY queries: ensure your brand appears in the AI's training-data entity map (third-party mentions across authoritative category sources) — entity injection at 19.8% is the primary discovery mechanism.
- For INFORMATIONAL queries: optimise for the compressed keyword form of the query — the AI compresses and searches, it does not inject brands.
- For REVIEW_SEEKING queries: build evidence pages (reviews, case studies) that match the evidence-seeking fan-out terms the AI generates.

For measurement. Current GEO monitoring tools track citation presence for keyword-style queries. Our format sensitivity finding (V=0.35) establishes that situation-first queries produce different fan-out distributions and likely surface different citations. Monitoring tools need to track both query formats to avoid systematic blind spots.

For model-tier awareness. Optimising for ChatGPT citation requires awareness that gpt-5.4 rarely searches while smaller models always search. As OpenAI's default model tier shifts, the search trigger rate — and therefore the importance of web content vs training-data presence — will shift with it.

5.3 Comparison with Study 2

Finding	Study 2 (n=2,158)	Study 3 (n=1,323)	Interpretation
Top fan-out type	Compression (37.9%)	Tangential (17.2%)	Broader corpus produces more diverse distribution
ChatGPT entity injection	34.8%	32.0% (composite); 44.8% (nano)	**Replicated** — entity injection is ChatGPT's dominant strategy
Evidence seeking	0.6%	15.6% (13.5% excl. REVIEW_SEEKING)	Broader intents reveal more evidence search; not a REVIEW_SEEKING artefact

Finding	Study 2 (n=2,158)	Study 3 (n=1,323)	Interpretation
Compression ratio	0.54	0.86 overall; **0.61 Perplexity-only**	Capture-method dependent (see below)
Vertical effect	3 verticals, descriptive	10 verticals, p=0.71 (ns, underpowered)	No effect detected; insufficient power to confirm

The replication of ChatGPT's entity injection dominance (32—45%) across two independent studies, different time periods, and different capture methods strengthens confidence in this as a stable platform characteristic.

The compression ratio discrepancy. Study 2 reported an overall compression ratio of 0.54; this study reports 0.86. The discrepancy is driven by capture method and platform mix. When broken down by platform: Perplexity shows a ratio of 0.61 (close to Study 2's Perplexity-dominated 0.54), Gemini shows 0.89, and ChatGPT shows 1.04 (fan-outs are actually *longer* than parent queries on average, because the API captures the full multi-clause search strings including injected entity names). Study 2 captured fan-outs via SSE interception, which may truncate or fragment long queries at stream boundaries. This study's API-level capture returns the complete query strings as structured data. Neither number should be treated as ground truth for "how much AI compresses queries" — the answer is platform-dependent, method-dependent, and ranges from genuine compression (Perplexity at 0.61) to expansion (ChatGPT at 1.04). The Study 2 headline that "AI compresses queries nearly in half" holds for Perplexity but not for ChatGPT.

5.4 Limitations

1. **Mixed model tiers on ChatGPT.** The ChatGPT data combines three model tiers (gpt-5.4, gpt-5.4-mini, gpt-5.4-nano) due to iterative collection. While this produced the model-tier finding, it means the ChatGPT fan-out distribution in Table 5 is a composite. Table 5b provides the per-model breakdown. Future work should control model tier as an explicit experimental variable.

2. **Rule-based classification at ~88% accuracy.** Approximately 158 of 1,323 records may be misclassified. Sensitivity analysis (Section 3.3) shows the primary vulnerability is the expansion/tangential boundary; the high-value categories (entity injection, evidence seeking, compression, narrowing) are robust to 12% reclassification. Formal inter-annotator agreement (Cohen's kappa) on a larger labeled sample, and LLM-assisted classification, would further strengthen the taxonomy.

3. **Single replicate for Gemini and Perplexity.** Replicate analysis was performed for ChatGPT (gpt-5.4-mini, 3 replicates) but not for Gemini or Perplexity. The high stochasticity of fan-out strings (98% zero overlap) observed on ChatGPT may differ on other platforms. The chi-squared tests on the full corpus treat single-sample fan-out distributions from Gemini and Perplexity as representative; the replicate data from ChatGPT suggests this overstates precision at the string level, though structural type distributions are moderately stable (65%).

4. **Heuristic citation linkage (Perplexity only).** The fan-out-to-citation linkage uses keyword overlap (minimum 20% Jaccard) rather than structural tracing. Citation yield rates in Table 8 should be interpreted as approximations, not exact measurements. ChatGPT and Gemini contribute zero linkage pairs due to API limitations.

5. **No Google AI Mode.** Google AI Mode does not expose fan-out queries through any accessible mechanism. Gemini's API with Google Search grounding provides fan-out data, but this may not reflect the exact retrieval behaviour of Google AI Mode in Search.

6. **Narrow temporal window.** All data is from April 2026. Fan-out behaviour may shift with model updates.

7. **Compression ratio is capture-method-dependent.** The 0.86 overall compression ratio cannot be directly compared to Study 2's 0.54. See Section 5.3 for detailed discussion.

5.5 Ethical Considerations

Data was collected via three mechanisms: the OpenAI Responses API (ChatGPT), the Google GenAI API (Gemini), and browser-level SSE interception (Perplexity). The API-based collection for ChatGPT and Gemini operates within the published terms of service for their respective APIs. Perplexity data was collected via a paid Pro account using a browser automation tool that intercepts network traffic at the client side; this approach captures data the user's own browser receives, but programmatic access at scale may exceed Perplexity's acceptable use expectations. No personal user data was collected — all queries were researcher-generated. No attempt was made to circumvent rate limits, authentication mechanisms, or access controls. The query corpus contains no sensitive, harmful, or personally identifiable content.

6. Conclusion

AI search platforms operate a hidden translation layer between human language and web retrieval. This study provides the first large-scale, cross-platform characterisation of that layer. The key finding is structural: AI retrieval is a two-layer system where the model's confidence determines whether to search at all (deterministic — 91.7% stable on ChatGPT replicate analysis), and the query's intent determines what *type* of search to perform (structurally stable at 65%, though the specific query strings are regenerated each time). Whether this stability pattern holds for Gemini and Perplexity remains to be confirmed. Neither traditional SEO nor the emerging GEO discipline fully accounts for both layers.

For the 65—85% of AI prompts that don't match traditional keywords (Semrush, 2026), the answer to "what is the AI searching for?" is: compressed keywords when learning, brand names when shopping, evidence terms when evaluating, and price data when comparing — if it decides to search at all. Optimising for AI citation means understanding not just what the AI retrieves, but whether it retrieves in the first place — and targeting the structural retrieval type, not the ephemeral query strings.

Data Availability

All data, scripts, and replication instructions are published at:

https://doi.org/10.5281/zenodo.19554329

The replication package includes the query corpus (180 queries), raw fan-out and citation data for all three platforms (5 JSONL files, including ChatGPT replicates 2 and 3), classified fan-outs (1,323 records), all analysis outputs, and all collection/analysis scripts. Licensed under CC BY 4.0.

References

- Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). "GEO: Generative Engine Optimization." arXiv:2311.09735.
- AIVO Evidentia (2026). "Decision-Path Analysis across AI Recommendation Systems: An Evidence-Grade Filter Taxonomy." Working Paper WP-2026-01.

- Lee, A. (2026a). "Domain-Level Features as Predictors of AI Citation." AI+Automation Research, Experiments K, L, M.
- Lee, A. (2026b). "Fan-Out Query Taxonomy — What AI Actually Searches For." AI+Automation Research, Study 2.
- Li, B. et al. (2025). "Towards AI Search Paradigm." arXiv:2506.17188.
- Semrush (2026). "ChatGPT traffic analysis: Insights from 17 months of clickstream data." semrush.com/blog/chatgpt-search-insights/
- Sturm, E. (2026). "The Fake 'GEO' Movement: How SEO Disinfo Campaigns Target Creators." The Edward Show, Episode 811.
- Tang, B. et al. (2025). "Xinyu AI Search: Enhanced Relevance and Comprehensive Results with Rich Answer Presentations." arXiv:2505.21849.
- WebLinkr (2026). Posts on r/SEO regarding query fan-out and GEO industry critique.
- Zhao, S. et al. (2025). "ParallelSearch: Train your LLMs to Decompose Query and Search Sub-queries in Parallel with Reinforcement Learning." arXiv:2508.09303.
- Zhong, M. et al. (2025). "ReDI: Reasoning-Enhanced Query Understanding through Decomposition and Interpretation." arXiv:2509.06544.