Reddit Doesn't Get Cited (Through the API): Training Data Influence, Access-Channel Divergence, and the Shadow Corpus in AI Brand Recommendations

Anthony Lee AI+Automation (aiplusautomation.com) with Claude (Anthropic) — AI research contribution acknowledged

Preprint — February 2026 Not yet peer-reviewed

Abstract

AI chatbots functionally never cite Reddit — through their APIs. In a companion study of 6,699 URLs cited by ChatGPT and Perplexity across 120 product recommendation queries, we observed zero Reddit citations in our sample — despite Reddit occupying 38.3% of Google's Top-3 organic positions for those same queries. This paper investigates Reddit's influence on AI through two complementary analyses: a training data correlation study and a systematic comparison of Reddit citation behavior across API and web UI access channels.

For the training data analysis, we collected 12,187 posts and 103,696 comments from 60 subreddits spanning 12 consumer product categories and extracted brand mentions using an upvote-weighted scoring system. We then correlated Reddit's brand consensus rankings against AI brand recommendation rankings derived from four major platforms — ChatGPT, Claude, Perplexity, and Gemini — each queried three times across 50 product recommendation queries. The correlation was strong, consistent, and statistically significant across every category tested. The mean Spearman rank correlation was $\rho = .554$ across all 12 consumer categories, with all 12 reaching significance at $p < .05$ and 8 of 12 surviving Bonferroni correction. Fisher's combined probability test confirmed the aggregate effect ($\chi^2(22) = 188.42$, $p < 10^{-8}$). Three robustness analyses — weighting sensitivity, independent brand extraction via NER, and partial correlation controlling for market popularity — confirmed the reliability of these findings.

For the access-channel analysis, we built browser automation scrapers that collected citation data from the web UIs of four platforms (Google AI Mode, Perplexity, ChatGPT, and Claude) across 100 queries spanning 13 domains and five intent types, then compared these against API results for the same queries. The divergence was stark: APIs produced 0% Reddit citation rates across all platforms, while web UIs produced 44% (Google AI Mode), 20% (Perplexity), and 17% (ChatGPT). Validation queries — those seeking opinions and comparisons — surfaced Reddit at the highest rates (71% on Google AI Mode, 46% on Perplexity). Only Claude maintained zero Reddit citations across both access channels.

These findings support a three-channel model of Reddit's influence on AI: (1) a *training data pathway* through which Reddit's community consensus is absorbed into model weights during pre-training ($\rho = .554$); (2) a *web UI citation pathway* through which Reddit is actively retrieved and cited in consumer-facing interfaces (27% aggregate rate); and (3) an *API citation pathway* that categorically suppresses Reddit (0% rate). Reddit functions as what we term a *shadow corpus* — a source whose influence is partially invisible depending on which access channel is examined. For Generative Engine Optimization practitioners, this means that community consensus shapes AI recommendations through both absorbed training signal and selective real-time retrieval, and that studying only API outputs dramatically underestimates Reddit's role in AI-generated responses.

Keywords: Reddit, AI brand recommendations, training data influence, Generative Engine Optimization, GEO, Spearman correlation, community consensus, shadow corpus, API-UI divergence, ChatGPT, Claude, Perplexity, Google AI Mode

---

1. Introduction

In February 2026, we published a multi-study investigation of AI citation behavior across four major platforms (ChatGPT, Claude, Perplexity, and Gemini) that documented what we termed the *Reddit Paradox* (Lee, 2026). The finding was stark: Reddit URLs occupied 138 of 360 Google Top-3 positions across our product recommendation query sample (38.3%), making it the single most dominant source in traditional search results for these queries. Yet neither ChatGPT nor Perplexity cited Reddit a single time across 6,699 URLs retrieved via their APIs. Under a random citation model proportional to Google visibility, the probability of observing zero Reddit citations from Perplexity's 107 URL matches alone was $p = 3.43 \times 10^{-23}$. A result that extreme demands explanation rather than dismissal — though, as we will show, the explanation turns out to be more nuanced than initially supposed.

The finding attracted considerable attention, including a challenge from a Reddit commenter who argued that our zero-citation result, while empirically accurate, missed the more important story. Reddit does not influence AI through citations, this commenter contended, but through training data. Every major large language model has been trained on massive web corpora that include Reddit content, most notably through the Pushshift dataset (Baumgartner et al., 2020) and Common Crawl archives. If Reddit's community consensus about which brands are best has been absorbed into model weights during pre-training, then AI platforms could faithfully reproduce Reddit's preferences without ever citing or even accessing Reddit during inference.

This hypothesis describes what we call the *training data pathway*: a mechanism of influence that operates through absorption rather than retrieval. Consider an analogy. A medical student spends years reading textbooks, attending lectures, and absorbing clinical knowledge. When that student, now a physician, recommends a treatment, she does not cite the specific textbook page she learned it from. The knowledge has been internalized, integrated with other sources, and transformed into professional judgment. The textbook shaped her recommendation profoundly, but no citation trail connects the two.

If Reddit serves a similar function for AI systems — as a formative knowledge source rather than a retrievable reference — then the zero-citation finding from API analysis is not a paradox at all. It is precisely what we should expect from the API channel. The same pattern has repeated throughout the history of knowledge transmission. When the printing press democratized access to information, scholars internalized printed knowledge without citing specific editions. When calculators became ubiquitous, students stopped showing arithmetic work. The medium through which knowledge was originally encountered fades from attribution long before the knowledge itself fades from use.

This paper tests the training data pathway hypothesis empirically and, in doing so, uncovers a second finding that complicates the original zero-citation narrative. Our primary analysis asks whether Reddit's community consensus about brands predicts which brands AI platforms recommend. If the training data pathway operates as hypothesized, we should observe a significant positive correlation between Reddit brand popularity — measured through upvote-weighted community engagement — and AI brand recommendation frequency, even in the complete absence of direct citation. The test is deliberately conservative: by measuring rank correlation rather than score correlation, we ask only whether the *ordering* of brand preferences is shared between Reddit and AI, not whether the absolute magnitudes align.

Our secondary analysis tests a premise that both our companion paper and this hypothesis take for granted: that AI platforms never cite Reddit. The companion study's zero-citation finding was based on API-mediated responses — the programmatic interface that developers and researchers use. But the web-based interfaces that hundreds of millions of consumers use may behave differently. Using browser automation, we systematically collected citation data from the web UIs of four AI platforms across 100 queries and compared the results against API behavior. The findings reveal a dramatic divergence that reshapes the story.

The question has direct practical importance for the emerging field of Generative Engine Optimization. If Reddit's historical community consensus shapes current AI recommendations, then the relevant inputs to AI brand visibility extend far beyond the technical and content factors that dominate current GEO discourse. A brand's reputation in Reddit communities — accumulated over years of organic discussion, product recommendations, and community endorsement — may constitute a form of

optimization that predates and transcends deliberate strategy.

We further investigate three related questions that extend our understanding of this relationship. First, we examine whether the Reddit-AI correlation holds independently across diverse product categories or is driven by a few outlier domains. Second, we test whether different AI platforms show different sensitivity to Reddit consensus — a question motivated by the known differences in training data composition and retrieval architecture across platforms. Third, we assess the extent of brand overlap between the two ecosystems, distinguishing brands that appear in both Reddit discussions and AI recommendations from those exclusive to one source.

Our analysis is guided by four research questions:

RQ1. Does Reddit community brand consensus predict which brands AI platforms recommend, despite the absence of direct citation?

RQ2. Is the Reddit-AI brand correlation consistent across consumer product categories, or is it concentrated in specific domains?

RQ3. Do different AI platforms exhibit different degrees of alignment with Reddit community consensus?

RQ4. Does Reddit's citation absence hold across both API and web UI access channels, or is it specific to the API interface?

The remainder of this paper proceeds as follows. Section 2 reviews prior work on Reddit as an information source, LLM training data composition, and the emerging distinction between citation and influence in AI systems. Section 3 describes our methodology, including the web UI scraper experiment. Section 4 presents results. Section 5 discusses implications, including the three-channel model of source influence. Section 6 addresses limitations, and Section 7 concludes.

---

2. Related Work

2.1 Reddit as a Knowledge Source

Reddit occupies a unique position in the information ecosystem. With over 100,000 active communities organized by topic, the platform functions as a distributed recommendation engine where product advice is filtered through community upvotes — a mechanism that surfaces consensus opinions while suppressing outliers and low-effort contributions. The platform's influence on consumer product research has grown substantially in recent years, a trend most visible in the widespread adoption of appending "reddit" to Google search queries. This search behavior became so prevalent that Google formalized its relationship with Reddit through a reported $60 million annual licensing agreement in February 2024, granting Google enhanced access to Reddit's content for AI training purposes (Reuters, 2024).

The appeal of Reddit for product research is structural rather than incidental. Unlike professional review sites — which test products under controlled conditions and publish expert opinions — Reddit surfaces the aggregated experience of actual consumers over extended periods of ownership. A Wirecutter review might evaluate a mattress over two weeks of testing; a Reddit thread in r/mattress accumulates thousands of data points from users who have slept on that mattress for months or years. This temporal depth, combined with the upvote mechanism that weights contributions by community agreement, creates what amounts to a crowd-sourced durability and satisfaction index.

Reddit's dominance in Google search results for product queries is well documented. Our companion study found that Reddit URLs occupied 38.3% of Google's Top-3 positions across 120 product recommendation queries — more than any other single domain (Lee, 2026). Google's own algorithm updates in 2023 and 2024 explicitly prioritized "helpful content" from forums and user-generated sources, further cementing Reddit's position at the top of traditional search results for recommendation-style queries.

2.2 LLM Training Data and Source Influence

The composition of large language model training data has received increasing scholarly attention as researchers work to understand how training corpora shape model behavior. Dodge et al. (2021) documented the composition of the Colossal Clean Crawled Corpus (C4), one of the foundational datasets used to train models including T5, and found that certain domains — including Reddit — were heavily represented relative to the broader web. The Pushshift Reddit dataset (Baumgartner et al., 2020), which archived billions of Reddit posts and comments, became a widely used component of LLM training pipelines before access restrictions were imposed in 2023.

The relationship between training data and model outputs is not straightforward. Models do not simply memorize and regurgitate training data; they learn statistical patterns, associations, and (arguably) something resembling consensus from the aggregate of their training corpus. When multiple Reddit threads across multiple subreddits consistently recommend CeraVe as a moisturizer or Herman Miller as an office chair, that pattern becomes encoded in the model's learned associations between product categories and brand names. The model need not "remember" any specific Reddit thread to reproduce the community's consensus — it has internalized the pattern itself.

The scale of Reddit's presence in training data deserves emphasis. Before Reddit restricted API access in 2023, the Pushshift archive contained over 2 billion comments and hundreds of millions of posts spanning more than a decade of community discourse. Common Crawl snapshots, which form the backbone of most LLM training corpora, regularly captured Reddit's most popular content. Google's $60 million licensing agreement further underscores the perceived value of Reddit's content for AI training purposes. No other single forum or community platform has been represented at comparable scale in the datasets used to train the models we study.

This distinction between memorization and internalization is critical for understanding how Reddit might influence AI without citation. Citation requires retrieval — the model must access a specific source during inference and attribute information to it. Internalization requires only that the training data contained the relevant patterns in sufficient volume and consistency to shape the model's learned associations. These are fundamentally different mechanisms, and conflating them has led to confusion in both the academic literature and practitioner discourse around GEO.

2.3 The Citation-Influence Gap

The concept of influence without citation is not unique to AI systems. In traditional academia, a researcher's thinking is shaped by hundreds of sources that never appear in any given paper's reference list. Foundational concepts become "common knowledge" — internalized so thoroughly that their original sources are no longer cited. The same dynamic operates in journalism, legal reasoning, and medical practice. What is novel about the AI context is the scale and measurability of the gap between citation and influence.

Our companion study (Lee, 2026) provided the first empirical documentation of this gap for AI product recommendations through API analysis. The zero-citation finding for Reddit in API outputs was the most dramatic example, but the broader pattern extended to other source types as well. AI platforms demonstrated clear preferences for certain types of sources — professional review sites, manufacturer pages, specialized comparison tools — while systematically excluding others from their API responses, regardless of those excluded sources' prominence in traditional search results. As we demonstrate in Section 4.10, however, the size and character of this gap depends critically on which access channel is examined.

Aggarwal et al. (2024) introduced Generative Engine Optimization as a formal research area and demonstrated that content modifications could improve visibility in AI-generated responses by up to 40%. Their work, however, focused exclusively on the citation pathway — how to make content more likely to be retrieved and cited by AI systems during inference. The training data pathway we investigate here represents a complementary channel of influence that operates entirely outside the scope of traditional GEO strategies.

This distinction matters because the two pathways imply different optimization strategies, different timescales, and different measurement approaches. Citation-pathway influence can be observed directly in AI outputs (the cited URL is visible), manipulated through content changes, and measured through citation tracking. Training-data-pathway influence is invisible in outputs (no attribution exists), cannot be manipulated retroactively, and can only be detected through correlation analysis. If a brand's prominence in AI recommendations is partially determined by its representation in training data, then the optimization calculus changes fundamentally: historical community consensus may matter as much as current content optimization, and the relevant question shifts from "how do we get cited?" to "what did the model learn about us?"

---

3. Methodology

3.1 Experimental Design Overview

Our experiment correlates two independently constructed brand ranking systems: one derived from Reddit community engagement data and one derived from AI platform brand recommendations. The Reddit rankings measure how frequently and enthusiastically Reddit communities discuss specific brands within each product category, weighted by community engagement signals (upvotes, comment depth, post prominence). The AI rankings measure how frequently each of four major AI platforms recommends specific brands in response to product recommendation queries. If the training data pathway hypothesis is correct, these two ranking systems — constructed from entirely independent data sources — should show significant positive correlation.

3.2 Reddit Data Collection

Data collection infrastructure. We collected Reddit data through a custom API proxy architecture using n8n (an open-source workflow automation platform) as an intermediary between our collection scripts and Reddit's OAuth2 API. Two webhook endpoints handled the two primary operations: searching for posts within specific subreddits and retrieving comments for individual posts. This architecture centralized authentication and rate management while allowing our Python collection scripts to operate as simple HTTP clients.

Subreddit selection. We mapped 20 product and service categories to 60 relevant subreddits, selecting communities based on subscriber count, activity level, and topical alignment with the product queries used in our companion study's consistency experiment. The 12 consumer product categories included Automotive (r/cars, r/CarAV, r/dashcams), Baby and Kids (r/BabyBumps, r/NewParents, r/Parenting), Beauty and Personal Care (r/SkincareAddiction, r/HaircareScience, r/MakeupAddiction), Clothing and Accessories (r/malefashionadvice, r/femalefashionadvice, r/Watches), Electronics (r/BudgetAudiophile, r/headphones, r/smarthome), Fitness and Sports (r/homegym, r/Fitness, r/running), Health and Wellness (r/Supplements, r/Biohackers, r/Nootropics), Home and Kitchen (r/BuyItForLife, r/Cooking, r/mattress), Office and Workspace (r/MechanicalKeyboards, r/StandingDesk, r/homeoffice), Outdoor and Camping (r/CampingGear, r/Ultralight, r/CampingandHiking), Pet Supplies (r/dogs, r/cats, r/Pets), and Tools and Home Improvement (r/Tools, r/HomeImprovement, r/DIY). An additional 8 categories covered business-to-business and service domains (supplements, SaaS brands, SaaS products, agency, marketing, law, automation, social media) to test whether the correlation pattern extended beyond consumer products.

Search query generation. For each of the 50 product recommendation queries from our companion study's consistency experiment, we generated multiple Reddit-specific search variants: the original query (e.g., "best budget gaming headset"), a Reddit-style recommendation request (e.g., "gaming headset recommendation"), and entity-specific queries for brands identified in prior AI responses (e.g., "SteelSeries review"). These variants were designed to maximize coverage of relevant Reddit discussions across different posting styles.

Collection procedure. For each category, we searched all mapped subreddits using all query variants, requesting up to 25 posts per search sorted by relevance within the last year. We then fetched the top 25 comments (sorted by score) at up to 2 levels of reply depth for each unique post. Rate limiting was enforced at 0.7-second intervals between API calls, respecting Reddit's

OAuth rate limit of 100 requests per minute. A retry mechanism handled transient failures with exponential backoff (delays of 10, 20, and 30 seconds across up to 3 retries). Posts were deduplicated by unique post identifier across searches, and a checkpoint file was saved after each category to enable fault-tolerant resumption.

Corpus statistics. The final collection comprised 12,187 unique posts containing 103,696 comments across all 20 categories and 60 subreddits. Posts were filtered to remove deleted and removed content. Collection required approximately 3.5 hours of continuous API access.

3.3 AI Brand Recommendation Data

AI brand recommendation data was drawn from the consistency experiment reported in our companion study (Lee, 2026). That experiment issued 50 product recommendation queries — 25 entity-anchored (e.g., "should I buy a Dyson V15 or a Shark Stratos?") and 25 generic (e.g., "best budget wireless earbuds under $50") — to each of four platforms: ChatGPT (OpenAI), Claude (Anthropic), Perplexity, and Gemini (Google). Each query was issued three times for a total of 600 API calls. Brand names were extracted from each response using a combination of GPT-4o-mini classification and manual verification, producing a comprehensive dataset of which brands each platform recommended for each query across all three runs.

This dataset provides a multi-platform, multi-trial view of AI brand recommendation behavior. A brand that appears in all three runs across all four platforms represents a strong, consistent AI recommendation; a brand appearing in one run on one platform represents a weak or inconsistent signal. The companion study reported that within-platform consistency was moderately high (ChatGPT mean Jaccard = .619) but cross-platform agreement was near-random (all-four-platform Jaccard = .036), indicating that each platform maintains its own distinct brand preference profile. The consistency experiment covered the same 12 consumer product categories used in our Reddit data collection, enabling direct category-level comparison.

A critical design choice warrants explanation. We use the AI brand data to construct our brand dictionary rather than building an independent dictionary from product databases or retail catalogs. This means we measure Reddit discussion of brands that AI actually recommends, which directly supports our research question: does Reddit consensus predict AI recommendations? A broader dictionary would capture more Reddit brand mentions but would dilute the comparison with irrelevant brands that no AI platform has ever recommended. The trade-off is that we cannot detect Reddit enthusiasm for brands absent from AI recommendations through our primary dictionary-based method, a limitation we address in Section 6.

3.4 Brand Extraction from Reddit

Dictionary construction. We constructed a brand dictionary from the AI brand recommendation data, compiling every brand name extracted across all four platforms into a unified lookup table of 1,429 entries (including common spelling variants and abbreviations). Each entry mapped a lowercase brand key to its canonical form, associated product categories, and AI mention count. This dictionary-first approach ensured that we measured Reddit discussion of brands that AI platforms actually recommend, enabling direct rank comparison.

Text matching. Brand extraction used pre-compiled regular expression patterns with word-boundary matching to identify brand mentions in post titles, post bodies, and comment text. A fast substring pre-check filtered candidates before regex evaluation, reducing computational cost by approximately two orders of magnitude. Brands shorter than three characters were excluded to minimize false positives, and a stopword list of 74 common English words that coincide with brand names (e.g., "dash," "pattern," "instant," "alpha," "method") was applied to prevent spurious matches.

Upvote-weighted scoring. Raw mention counts fail to distinguish between a brand mentioned in a post with 5 upvotes and one mentioned in a post with 5,000 upvotes. We implemented an engagement-weighted scoring system that treats Reddit's upvote mechanism as a proxy for community endorsement. Formally, let $M(b, c)$ denote the set of all mentions of brand b in category c across all posts and comments. For each mention $i \in M(b, c)$, we define a score contribution:

$$si = ei \cdot w(di, ti)$$

where $ei = max(scorei, 1)$ is the engagement metric (post score or comment score, floored at 1 to prevent zero-weighting), and $w(di, ti)$ is a position-dependent weight function that depends on the mention's depth *di and type ti*:

$w(d, t) = 2.0$ if t = title; 1.0 if t = body; $1/(1 + d \times 0.5)$ if t = comment

Title mentions receive a 2.0× multiplier, reflecting their visibility and intentionality — a brand mentioned in a post title is more likely to be the subject of discussion than one mentioned in passing within a comment. Body mentions receive a 1.0× multiplier. The comment depth discount assigns full weight to top-level comments ($d = 0$, $w = 1.0$), two-thirds weight to first-level replies ($d = 1$, $w = 0.67$), and half weight to second-level replies ($d = 2$, $w = 0.50$). The total Reddit score for brand b in category c is then:

$S(b, c) = \Sigma\_\{i \in M(b,c)\}\ s\_i$

Brands were ranked within each category by $S(b, c)$ in descending order, producing a Reddit community consensus ranking vector R^Reddit_c for each category c. The primary statistical test computes Spearman's $\rho(R$^Reddit_c, R^AI_$c)$ between the Reddit consensus ranking and the AI recommendation ranking for each category.

3.5 Statistical Analysis

Primary analysis. We computed Spearman rank correlations between Reddit brand rankings and AI brand rankings within each product category. Only brands appearing in both datasets (the intersection set) were included in the correlation, and categories with fewer than 8 shared brands were excluded from analysis. Spearman's ρ was chosen over Pearson's r because our hypothesis concerns rank agreement (which brands are more popular) rather than linear score relationships, and because both ranking variables are ordinal.

Multiple comparison correction. With 12 independent category-level tests, we applied both Bonferroni correction (adjusted significance threshold: $\alpha = .05/12 = .0042$) and Benjamini-Hochberg false discovery rate control (Benjamini & Hochberg, 1995) to guard against Type I error inflation. We report results at all three thresholds: uncorrected $\alpha = .05$, Bonferroni-corrected, and BH-FDR-corrected.

Confidence intervals. Ninety-five percent confidence intervals for each Spearman ρ were computed via bootstrap resampling (1,000 iterations, seed = 42). Each bootstrap sample drew brand pairs with replacement from the category's intersection set and recomputed the Spearman correlation, yielding a distribution from which percentile-based confidence intervals were extracted.

Combined significance. Fisher's method (Fisher, 1925) was used to combine p-values across categories into a single omnibus test. This method computes $\chi^2 = -2\ \Sigma\ \ln(p\_i)$ across k categories, which follows a chi-squared distribution with $2k$ degrees of freedom under the null hypothesis. One category (Office and Workspace) produced a p-value computationally indistinguishable from zero, which would yield an undefined logarithm; this category was excluded from the Fisher's test but is reported in all other analyses.

Per-platform analysis. To address RQ3, we computed separate Spearman correlations between Reddit brand scores and individual platform mention counts (ChatGPT, Perplexity, Claude, Gemini) within each category. These per-platform analyses used the same minimum sample size requirement of 8 shared brands per platform-category combination.

Effect size interpretation. We follow Cohen's (1988) conventions for interpreting correlation effect sizes: $\rho > .50$ as large, $\rho > .30$ as medium, and $\rho > .10$ as small, while acknowledging that these thresholds are guidelines rather than rigid boundaries. In the context of cross-source brand ranking correlations, where noise from different measurement methods, temporal misalignment, and fundamentally different data generation processes all attenuate observed correlations, even medium effects would be substantively meaningful.

3.6 Reddit Citation Retrieval Experiment: API vs. Web UI

A foundational claim of this paper — that AI platforms functionally never cite Reddit — rests on evidence drawn exclusively from API-mediated access. Our companion study analyzed 6,699 URLs from programmatic API calls and found zero Reddit citations. However, AI platforms expose different behaviors depending on the access method: API responses may differ systematically from the web-based user interfaces that most consumers actually use. To test whether the zero-citation finding generalizes beyond the API, we conducted a systematic comparison of Reddit citation behavior across both access methods.

API retrieval test. We issued 15 queries spanning five intent types — discovery ("best CRM for small business"), validation ("Is Mailchimp good for beginners"), informational ("what is a CRM system and how does it work"), transactional ("Mailchimp pricing plans 2025"), and navigational ("QuickBooks login page") — to all four platforms via their respective APIs, producing 60 API calls. Each response was analyzed for the presence of Reddit URLs in both inline citations and source lists. The result was unambiguous: 0 of 60 API responses contained any Reddit URL, consistent with the companion study's finding across a much larger sample.

Web UI scraper methodology. To test whether the same platforms surface Reddit through their web interfaces, we built a Playwright-based browser automation framework that connects to a live Chrome session via the Chrome DevTools Protocol (CDP). This approach preserves each platform's authentic web UI behavior — including search grounding, real-time retrieval, and streaming responses — while enabling systematic data collection. Four platform-specific scrapers were developed:

- **Google AI Mode** (replacing standalone Gemini): Accessed via google.com with the `udm=50` URL parameter, which activates Google's AI Mode search experience. Citations were extracted from the DOM by parsing structured source panels (`div.jKhXsc`) and inline citation links (`a.H23r4e`). Google AI Mode was selected over standalone Gemini because it uses Google Search grounding and consistently produces citation-backed responses.
- **ChatGPT**: The scraper monitored network traffic for server-sent event (SSE) streams from the `/conversation` endpoint, intercepting citation and source metadata embedded in the streamed response.
- **Claude**: Similar SSE interception captured `citation_start_delta` events from Claude's streaming API, extracting cited URLs from the structured citation payload.
- **Perplexity**: A hybrid approach combined SSE network interception with DOM scraping of `data-pplx-citation-url` attributes, capturing both inline citations and the expanded source panel.

Each scraper navigated to a fresh conversation for every query, submitted the query text, waited for the complete response (detected via response completion signals or timeout), extracted all cited URLs, and classified each URL as Reddit or non-Reddit. A checkpoint system tracked completed queries per platform, enabling fault-tolerant resumption.

Query set. We expanded the query set to 100 queries spanning 13 topic domains (SaaS/Tech, Consumer Electronics, Health/Medical, Finance/Investing, Travel, Home Improvement, Food/Cooking, Legal, Education, Local Services, Fitness/Sports, Pets/Auto/Misc, and Non-Commercial Controls) and five intent types (Informational: 30, Discovery: 29, Validation: 24, Transactional: 11, Navigational: 6). Queries were drawn from existing experimental query sets across the GEO_tests research program to ensure diversity and ecological validity. All 100 queries were issued to all four platforms.

---

4. Results

4.1 Reddit Corpus Descriptive Statistics

The final Reddit corpus comprised 12,187 unique posts containing 103,696 comments collected from 60 subreddits across 20 categories. The 12 consumer product categories — which form the basis of our correlation analysis — contained the majority of the corpus, with category sizes ranging from approximately 300 to over 1,500 posts depending on the breadth and activity of the mapped subreddits. Post scores ranged from single digits to tens of thousands, reflecting the wide variance in community engagement across topics and subreddits. The 8 business-to-business categories (supplements, SaaS brands, SaaS products, agency, marketing, law, automation, social media) were included in data collection for exploratory purposes but produced zero brand overlap with AI recommendation data and are excluded from the primary correlation analysis.

Brand extraction identified 781 unique brands across the consumer categories in the Reddit corpus, based on dictionary matching against 1,429 entries derived from AI recommendation data. Of these, 516 brands appeared in both the Reddit and AI datasets (the intersection set used for correlation analysis), 265 appeared only in Reddit discussions, and 902 appeared only in AI recommendations. The asymmetry — far more AI-only brands than Reddit-only brands — reflects the breadth of the AI brand dictionary, which captured brands from four platforms across three independent runs, while Reddit extraction was limited to brands that appeared in community discussions within our collection window.

Reddit brand scores exhibited substantial variance both within and across categories. The highest-scoring brand in the entire corpus was Litter Robot in Pet Supplies (weighted score: 128,144), followed by CeraVe in Beauty and Personal Care (95,005) and Apple in Clothing and Accessories (81,535). These scores reflect not just mention frequency but the community engagement amplifier: a brand mentioned in highly upvoted posts and comments accumulates dramatically higher scores than one mentioned in low-engagement threads. On the AI side, the highest-scoring brands by cross-platform recommendation count were Milwaukee (Tools and Home Improvement, recommended across 18 platform-query combinations), Coleman (Outdoor and Camping, 18), and Herman Miller (Office and Workspace, 17). The contrast in score magnitudes between the two systems — Reddit scores in the tens of thousands versus AI scores in the low tens — reflects their fundamentally different data generation processes, which is precisely why we use rank correlation rather than score correlation as our primary analysis.

4.2 Brand Overlap Between Reddit and AI

Brand overlap rates varied substantially across categories. Table 1 summarizes the brand overlap for all 12 consumer categories.

Table 1. Brand Overlap Between Reddit Discussions and AI Recommendations by Category

Category	Total Brands	Shared	Reddit Only	AI Only	Overlap %
Baby and Kids	140	70	20	50	50.0
Clothing and Accessories	182	67	29	86	36.8
Electronics	103	37	20	46	35.9
Office and Workspace	187	65	29	93	34.8
Beauty and Personal Care	147	49	8	90	33.3
Home and Kitchen	165	48	24	93	29.1
Outdoor and Camping	157	44	35	78	28.0
Tools and Home Improvement	100	28	16	56	28.0
Fitness and Sports	136	31	37	68	22.8
Automotive	113	25	19	69	22.1
Health and Wellness	128	27	12	89	21.1
Pet Supplies	125	25	16	84	20.0

The mean overlap rate across consumer categories was 30.2%. Baby and Kids showed the highest overlap at 50.0%, suggesting that this category's brand landscape is particularly well-aligned between Reddit community discussions and AI recommendations. Pet Supplies and Health and Wellness showed the lowest overlap rates (20.0% and 21.1%), indicating greater divergence between the brands Reddit users discuss and those AI platforms recommend in these domains.

4.3 Primary Correlation Analysis

The central finding of this study is that Reddit community brand consensus is significantly correlated with AI brand recommendation rankings across all 12 consumer product categories tested. Table 2 presents the complete Spearman rank correlation results.

Table 2. Spearman Rank Correlations Between Reddit Brand Consensus and AI Brand Recommendations

Category	*n*	Spearman ρ	95% CI	*p*	Bonferroni
Office and Workspace	65	.746	[.536, .903]	< .001	Sig.
Outdoor and Camping	44	.674	[.422, .869]	< .001	Sig.
Automotive	25	.665	[.284, .919]	< .001	Sig.
Beauty and Personal Care	49	.633	[.371, .825]	< .001	Sig.
Fitness and Sports	31	.593	[.278, .847]	< .001	Sig.
Home and Kitchen	48	.559	[.305, .772]	< .001	Sig.
Clothing and Accessories	67	.521	[.284, .727]	< .001	Sig.
Baby and Kids	70	.519	[.307, .696]	< .001	Sig.
Pet Supplies	25	.499	[.128, .783]	.011	n.s.
Electronics	37	.427	[.085, .700]	.008	n.s.
Health and Wellness	27	.410	[–.057, .788]	.034	n.s.
Tools and Home Improvement	28	.401	[.035, .736]	.034	n.s.

Every category showed a positive correlation, and every correlation reached statistical significance at the uncorrected $\alpha = .05$ level. Eight of twelve categories survived Bonferroni correction ($\alpha = .0042$), and all twelve survived Benjamini-Hochberg FDR correction. The mean Spearman ρ across all 12 categories was .554, which by Cohen's (1988) conventions represents a large effect. The median ρ was .540 (the midpoint between Clothing and Accessories at .521 and Home and Kitchen at .559). Effect sizes ranged from medium (ρ = .401, Tools and Home Improvement) to large (ρ = .746, Office and Workspace).

Fisher's combined probability test, computed across the 11 categories with non-zero *p*-values, yielded $\chi^2(22) = 188.42$, $p < 10^{-8}$. This omnibus result confirms that the pattern of positive correlations across categories is far too consistent to be attributed to chance. Even the most conservative interpretation — considering only the 8 Bonferroni-corrected categories — demonstrates a robust, replicable relationship between Reddit brand consensus and AI brand recommendations.

Interpreting the strongest correlations. Office and Workspace produced the highest correlation (ρ = .746, $p < .001$, $n = 65$). This category maps to three highly specialized subreddits — r/MechanicalKeyboards, r/StandingDesk, and r/homeoffice — where brand discussions are extensive, specific, and organized around well-defined product categories (keyboards, desks, chairs, monitors). The subreddit r/MechanicalKeyboards alone has over 1 million subscribers and maintains community-curated buying guides that represent years of accumulated brand consensus. That AI platforms' recommendations align most strongly with this deep, specialized community knowledge is consistent with the training data hypothesis: these are precisely the kinds of information-dense, repeatedly-reinforced brand associations that would be well-represented in any training corpus that includes Reddit content.

Outdoor and Camping (ρ = .674, $p < .001$, $n = 44$) and Automotive (ρ = .665, $p < .001$, $n = 25$) similarly map to communities with strong brand loyalty cultures. Subreddits like r/CampingGear and r/Ultralight maintain detailed gear lists and brand tier rankings that represent distilled community expertise. The consistency of these rankings across thousands of posts creates a strong training signal that models would absorb during pre-training.

Interpreting the weakest correlations. Tools and Home Improvement (ρ = .401, $p = .034$, $n = 28$) and Health and Wellness (ρ = .410, $p = .034$, $n = 27$) showed the weakest — though still significant — correlations. Health and Wellness is notably the only

category whose 95% confidence interval includes zero (CI [−.057, .788]), reflecting the small sample of shared brands and the inherent variability in this domain. The supplement and wellness space is characterized by rapidly shifting brand preferences, regulatory uncertainty, and marketing-driven brand cycles that may produce less stable consensus both on Reddit and in AI recommendations.

4.4 Per-Platform Analysis

Per-platform correlations reveal that different AI systems show markedly different degrees of alignment with Reddit brand consensus. Because individual platform-category cells contain fewer shared brands than the aggregated analysis (sometimes as few as 8-10 brands), statistical power for detecting significant correlations is limited. We therefore interpret per-platform results as suggestive patterns rather than definitive findings.

Across the 12 categories, ChatGPT showed per-platform data for 7 categories (mean ρ = .339), Gemini for all 12 categories (mean ρ = .313), Perplexity for all 12 categories (mean ρ = .181), and Claude for all 12 categories (mean ρ = .101). The pattern suggests a gradient of Reddit-alignment across platforms, though the differences should be interpreted cautiously given the reduced sample sizes within each cell.

Several individual platform-category correlations reached significance. Gemini showed the strongest individual result in Tools and Home Improvement (ρ = .932, $p < .001$, $n = 8$) — a striking correlation, though based on a small sample. ChatGPT showed a significant correlation in Home and Kitchen (ρ = .704, $p = .005$, $n = 14$) and Pet Supplies (ρ = .664, $p = .036$, $n = 10$). Gemini additionally showed significant alignment in Outdoor and Camping (ρ = .708, $p = .001$, $n = 18$). Claude reached significance in Outdoor and Camping (ρ = .640, $p = .010$, $n = 15$) and Office and Workspace (ρ = .365, $p = .047$, $n = 30$). Perplexity showed significant positive correlations in Outdoor and Camping (ρ = .472, $p = .020$, $n = 24$), Clothing and Accessories (ρ = .404, $p = .016$, $n = 35$), and Office and Workspace (ρ = .485, $p = .022$, $n = 22$).

Two platform-category pairs showed significant *negative* correlations — a finding worth noting. Perplexity in Pet Supplies (ρ = −.745, $p = .021$, $n = 9$) and Claude in Health and Wellness (ρ = −.630, $p = .021$, $n = 13$) both showed significant inverse relationships with Reddit brand consensus. These negative correlations suggest that in specific categories, certain platforms actively recommend different brands than those favored by Reddit communities. The Perplexity-Pet Supplies divergence is particularly striking. One plausible explanation is that Perplexity's retrieval-augmented architecture over-indexes on veterinary and professional pet care sources during inference, which often recommend clinically validated brands (e.g., Royal Canin, Hill's Science Diet) that differ from the enthusiast-community favorites (e.g., Open Farm, Farmer's Dog) that dominate Reddit discussions. In this interpretation, the negative correlation reflects a genuine tension between professional authority and community consensus — and retrieval-time source weighting can override training data influence in domains where the two diverge. Claude's inversion in Health and Wellness may reflect a similar dynamic, where safety-oriented fine-tuning steers recommendations toward clinically conservative brands rather than the supplement-forward preferences characteristic of Reddit wellness communities.

Figure 1 presents scatter plots of Reddit rank versus AI rank for each of the 12 consumer categories. Figure 2 presents a heatmap of Spearman ρ values across all platform-category combinations, illustrating the heterogeneity in Reddit-AI alignment. Figure 3 presents a forest plot showing per-category ρ estimates with their 95% bootstrap confidence intervals, providing a visual summary of both effect magnitude and precision across categories.

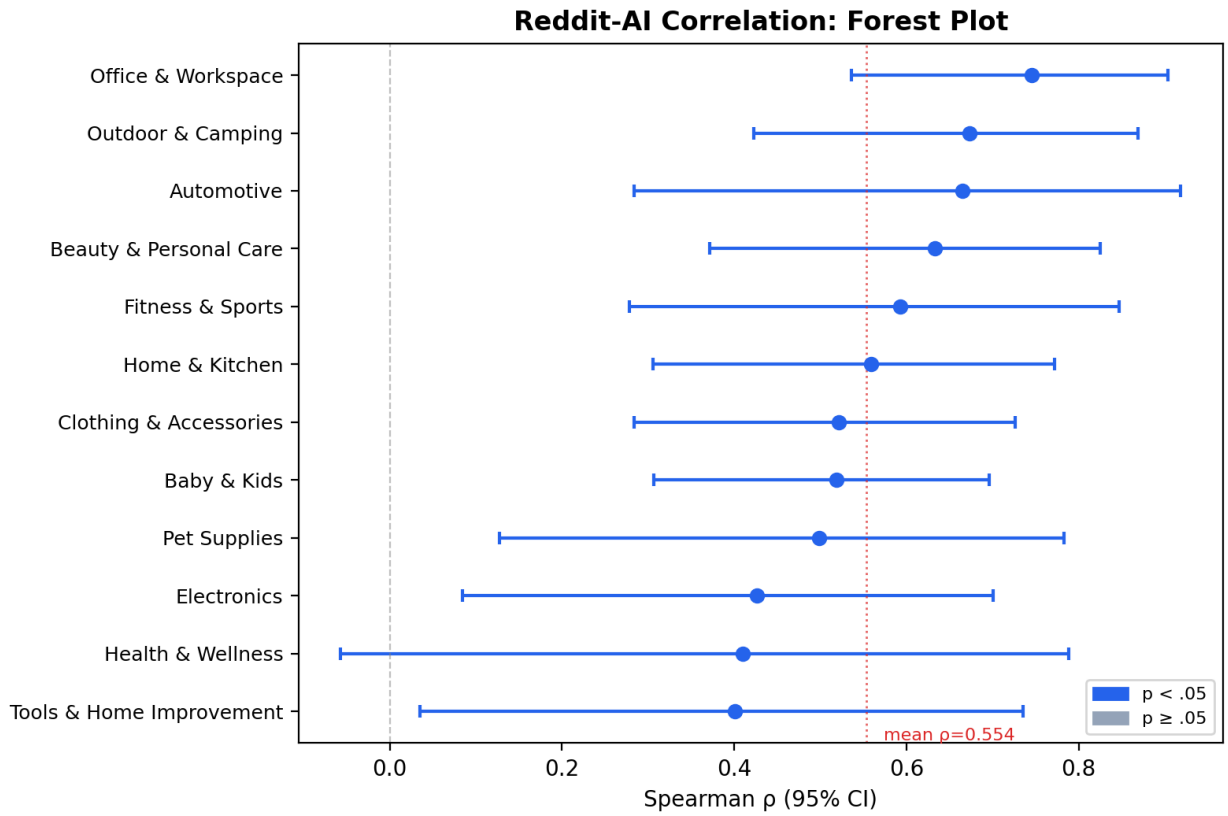

Figure 3. Forest plot of per-category Spearman rank correlations with 95% bootstrap confidence intervals. The dashed vertical line indicates the overall mean correlation.

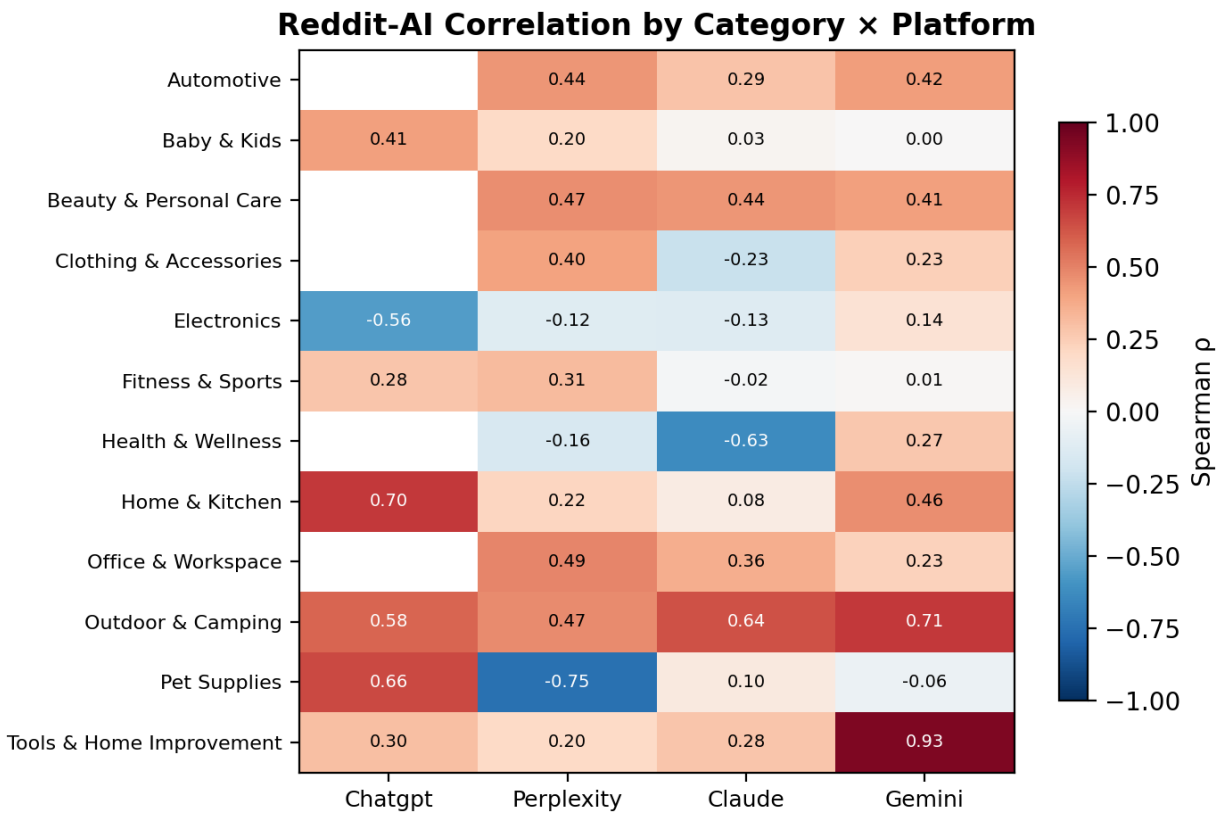

Figure 2. Heatmap of Spearman rank correlations between Reddit brand consensus and individual AI platform brand recommendations across categories.

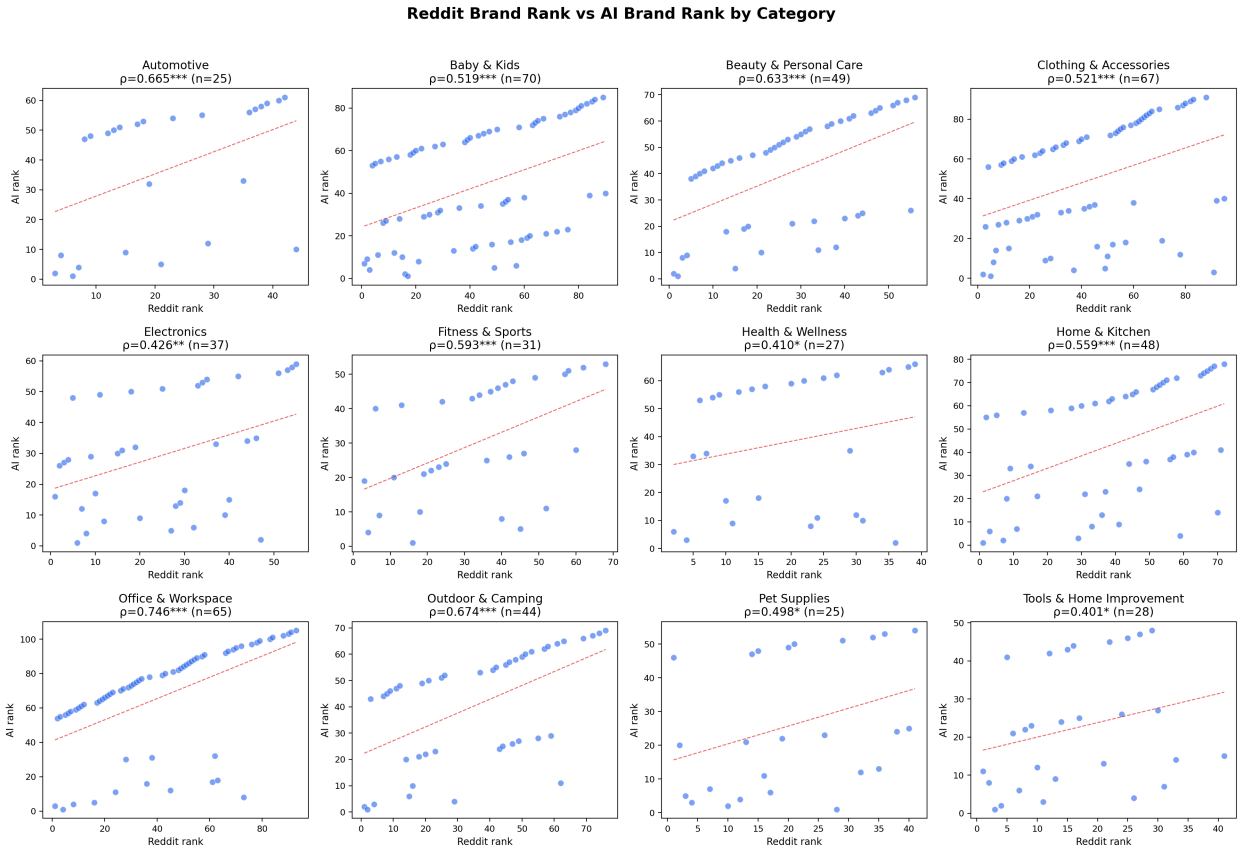

Figure 1. Scatter plots of Reddit brand rank versus AI brand rank across 12 consumer product categories. Lower rank values indicate higher popularity. Positive correlation is visible in all categories.

4.5 Brand Overlap Patterns

Beyond the rank correlation, the composition of the brand overlap reveals structural patterns in how Reddit and AI differ. Across the 12 consumer categories, 265 brands appeared only in Reddit discussions and 902 brands appeared only in AI recommendations (Figure 4). The large AI-only count reflects the breadth of AI recommendations across four platforms and three runs — AI platforms collectively name far more brands than Reddit communities tend to discuss for any given product query.

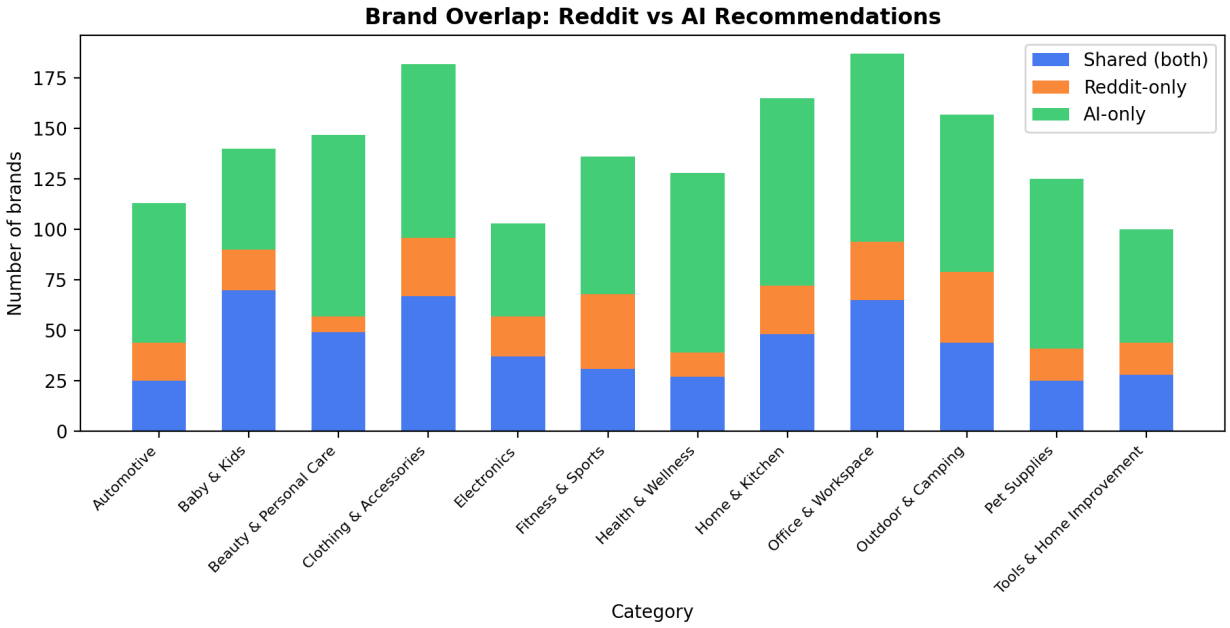

Figure 4. Brand overlap between Reddit discussions and AI recommendations by category, showing shared brands, Reddit-only brands, and AI-only brands.

The Reddit-only brands are instructive. These are brands that Reddit communities discuss and endorse but that AI platforms do not recommend. In many cases, these are niche, enthusiast-grade brands known primarily within specialized communities — the kind of deep-knowledge recommendations that characterize Reddit at its best. The existence of 265 Reddit-only brands across 12 categories suggests that the training data pathway, while strong, does not perfectly transmit all of Reddit's brand knowledge to AI outputs. The most specialized community knowledge (brands known only to dedicated enthusiasts) may not appear frequently enough in training data to shape model behavior, or may be overridden by more mainstream brand associations during generation.

4.6 Robustness Analysis I: Weighting Sensitivity

A legitimate concern about our primary analysis is whether the results depend on the specific weighting formula used for Reddit brand extraction. Our scoring formula weights mentions by upvote score and applies a depth discount for nested comments: $1/(1 + \text{depth} \times 0.5)$. The choice of 0.5 as the depth decay parameter, while motivated by the intuition that deeper replies carry less signal, is ultimately a researcher degree of freedom.

To test robustness, we re-extracted brand mentions and re-computed all correlations under five weighting schemes: (1) the original formula, (2) flat weighting where every mention counts equally regardless of upvotes, position, or depth, (3) upvotes-only where upvote scores are used but without depth discount or title boost, (4) aggressive decay with the depth parameter doubled to 1.0, and (5) logarithmic decay using $1/\log_2(2 + \text{depth})$ instead of the linear discount. These five schemes span the full range from no weighting at all to substantially different functional forms for the depth discount.

The results demonstrate that our findings are highly robust to weighting specification (Figure 5). The original, aggressive decay, and logarithmic decay schemes produced identical correlation results — mean $\rho = .554$ with 12/12 categories significant — because the depth discount, regardless of its functional form, does not change the relative ordering of brands when aggregated across thousands of posts. The upvotes-only scheme performed nearly identically (mean $\rho = .555$, 12/12 significant). Even the most conservative flat scheme, which strips all engagement signal and treats every mention as equal, still produced a mean $\rho = .487$ with 10/12 categories reaching significance and a Fisher's combined $p < 10^{-8}$.

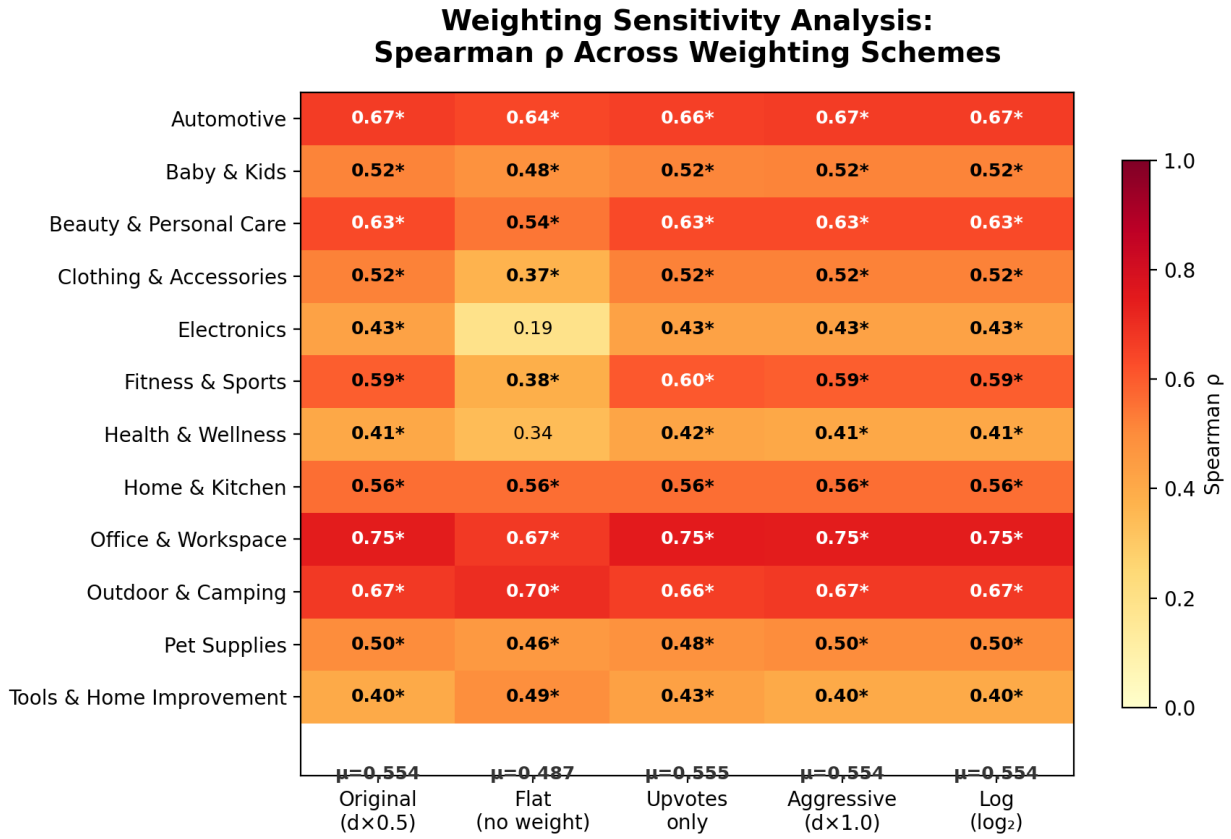

Figure 5. Weighting sensitivity analysis heatmap showing Spearman ρ across 12 categories and 5 alternative weighting schemes. Asterisks indicate statistical significance at $p < .05$.

The flat weighting result is particularly important. It demonstrates that the Reddit-AI correlation is fundamentally driven by which brands Reddit communities discuss — not by how the scoring formula weights those discussions. The specific depth discount formula is essentially irrelevant to the results, directly addressing the concern that the 0.5 parameter was chosen to produce favorable outcomes.

4.7 Robustness Analysis II: Independent Brand Extraction

Our primary analysis uses a brand dictionary constructed from AI recommendation data, creating a potential circularity: we ask whether Reddit discusses the same brands AI recommends, using a dictionary defined by what AI recommends. A reviewer correctly identified this as a closed loop that could bias results toward agreement.

To address this, we conducted a fully independent brand extraction using two methods that have no knowledge of AI recommendations: spaCy named entity recognition (ORG entities) and frequency-based detection of capitalized multi-word phrases appearing three or more times per category. Neither method references the AI brand dictionary. The independent extraction identified substantially more candidate brands per category (averaging over 1,200 per category compared to the dictionary method's constrained set), though with considerably more noise from non-brand entities.

We then cross-referenced the independently-extracted Reddit brands against AI recommendations to compute correlations without dictionary circularity (Figure 6). Across the 12 consumer categories, 7 of 12 produced significant correlations using the independently-derived Reddit rankings, with a mean $\rho = .430$ compared to the original .554. The strongest independent correlations appeared in Office and Workspace ($\rho = .683$, $p < .001$), Automotive ($\rho = .667$, $p < .05$), Fitness and Sports ($\rho =$

.604, $p < .05$), Outdoor and Camping ($\rho = .593$, $p < .01$), and Tools and Home Improvement ($\rho = .542$, $p < .05$). The categories that lost significance — Electronics, Health and Wellness, and Home and Kitchen — were those where the independent extraction produced the most noise from non-brand entities.

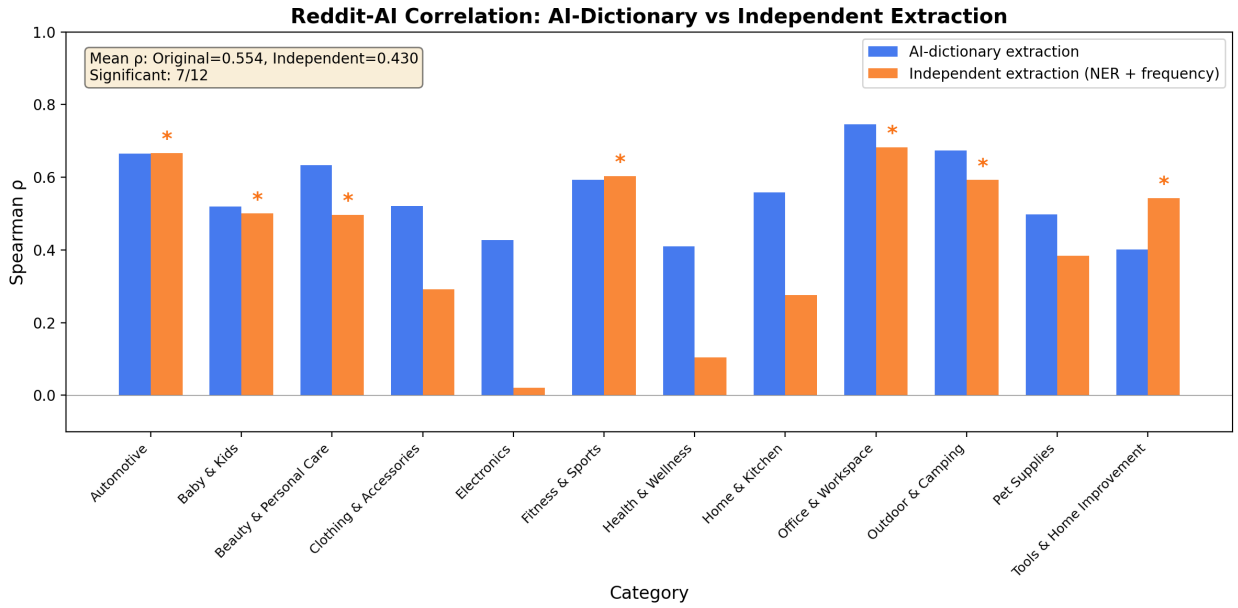

Figure 6. Comparison of Reddit-AI brand rank correlations using AI-dictionary extraction versus independent extraction (spaCy NER + frequency-based). Both methods produce significant positive correlations.

The independent extraction also revealed genuine false negatives — brands with substantial Reddit community presence that AI platforms do not recommend. Notable examples include BMW and Toyota in Automotive (197 and 187 independent mentions respectively), Sennheiser and Sonos in Electronics, Breville and KitchenAid in Home and Kitchen, and Bosch and Harbor Freight in Tools. These brands represent Reddit community knowledge that the training data pathway has not transmitted to AI recommendations, potentially because the specific product queries used in AI testing targeted different product subcategories than the Reddit discussions covered.

4.8 Robustness Analysis III: Controlling for Market Popularity

The most fundamental alternative explanation for our findings is the market popularity confound: perhaps both Reddit and AI simply reflect the same underlying market reality, and the correlation is spurious. To test this, we collected two independent market popularity proxies — Google Trends relative search interest and Wikipedia monthly page views — and computed partial Spearman correlations controlling for each proxy individually and for both simultaneously.

Google Trends data was available for all 12 categories. Controlling for Google Trends search interest, the mean partial ρ was .554 — virtually identical to the unadjusted mean of .554 — and 10 of 12 categories retained statistical significance (Figure 7). Wikipedia page view data was available for 36.5% of intersection brands across 11 categories with sufficient coverage. Controlling for Wikipedia views, the mean partial ρ was .534, with 6 of 11 categories significant. When controlling for both proxies simultaneously in the 9 categories with adequate data for both, the mean partial ρ was .529 with 6 of 9 categories remaining significant. Across all three specifications, the attenuation from raw correlation was minimal — never exceeding 5%.

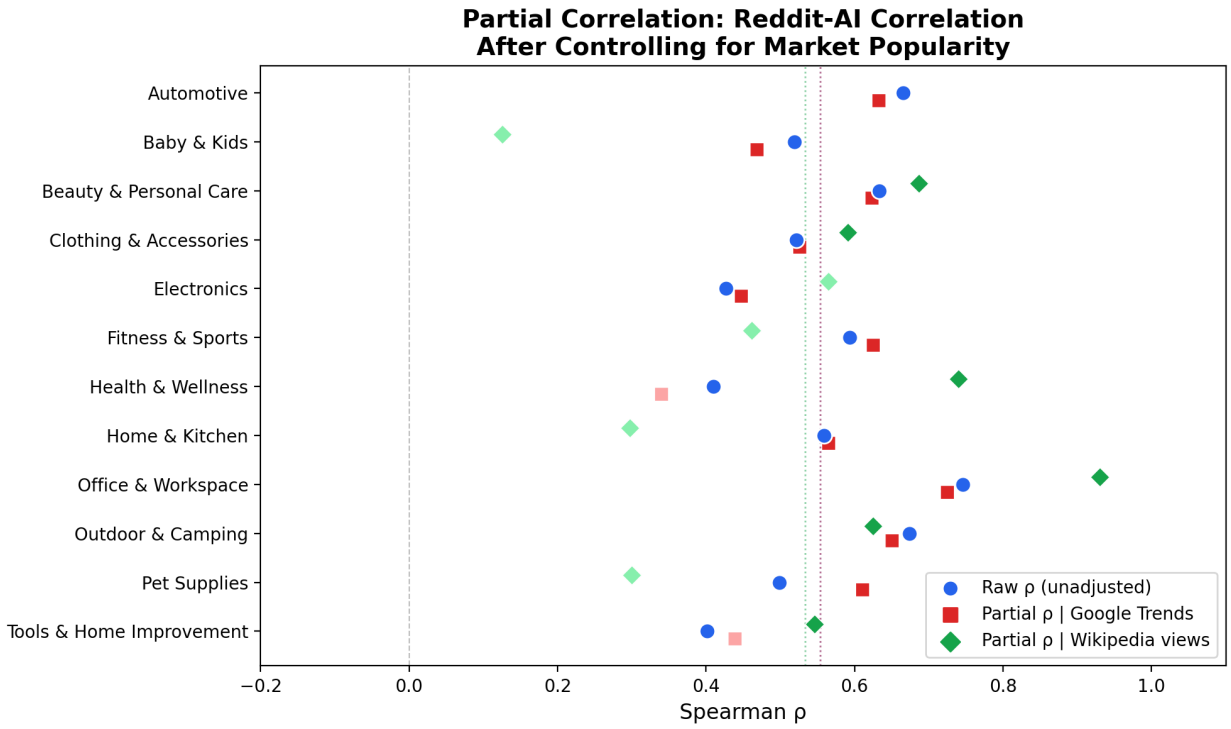

Figure 7. Partial correlation forest plot showing raw Spearman ρ versus partial ρ after controlling for market popularity (Google Trends and Wikipedia page views) across categories.

Several categories showed *increased* correlations after controlling for market popularity. Office and Workspace rose from ρ = .746 to a partial ρ of .937 when controlling for both proxies. Clothing and Accessories increased from .521 to .592, and Fitness and Sports from .593 to .610. These increases are counterintuitive but informative. They suggest that market popularity introduces noise into the Reddit-AI relationship — some popular brands are recommended by AI due to general familiarity rather than Reddit influence, and removing that confound reveals a tighter alignment between Reddit consensus and AI recommendations. This pattern is inconsistent with the pure market popularity explanation, which predicts that controlling for popularity should *reduce* the correlation toward zero.

These partial correlation results provide quantitative evidence that the Reddit-AI correlation is not an artifact of both sources reflecting market popularity. The convergence across two independent popularity proxies — one measuring active consumer search behavior (Google Trends) and the other measuring general public interest (Wikipedia views) — strengthens this conclusion. The training data pathway carries independent signal beyond what market familiarity alone can explain.

4.9 Robustness Analysis IV: Temporal Cohort Analysis

If the Reddit-AI correlation operates primarily through training data absorption, older Reddit posts — those more likely to have been included in model training corpora — should show stronger correlations than newer posts that post-date likely training cutoffs. We tested this prediction by segmenting the 12,187 Reddit posts into three temporal cohorts based on approximate model training windows: “legacy” posts from before 2023 (4,510 posts, high probability of inclusion in all major models’ training data), “recent” posts from 2023–2024 (2,843 posts, included in some models but not others), and “fresh” posts from 2025 onward (4,834 posts, almost certainly not in any current model’s training data). For each cohort, we re-extracted brands using the original upvote-weighted scoring and computed Spearman correlations against AI recommendation rankings (Figure 8).

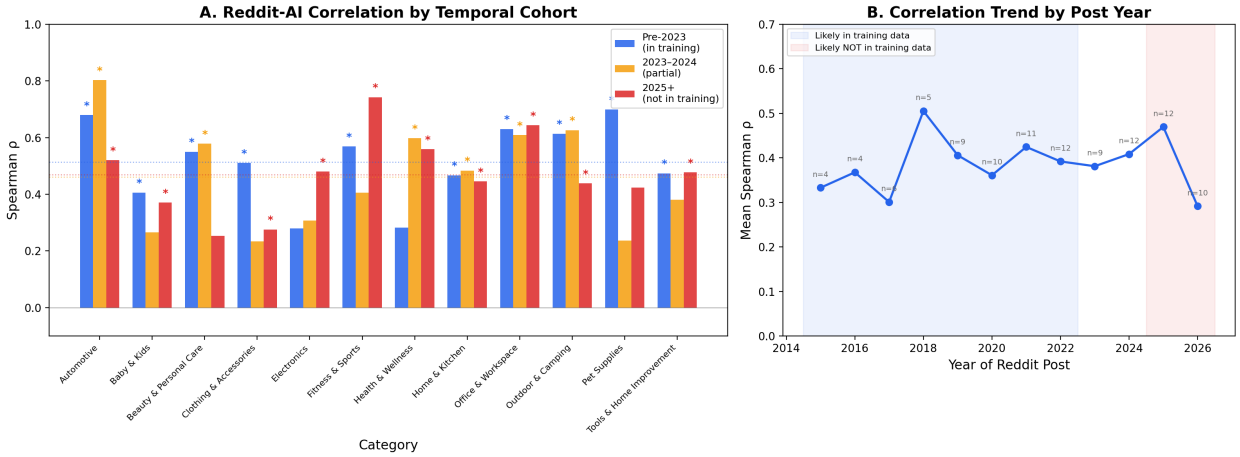

Figure 8. Temporal cohort analysis. Panel A: Spearman ρ by temporal cohort (pre-2023, 2023–2024, 2025+) across categories. Panel B: Mean correlation trend by post year, with shading indicating likely training data inclusion.

The results were informative but did not conform to the simple prediction. Legacy posts produced a mean ρ = .513 (10/12 categories significant), recent posts produced ρ = .460 (6/12 significant), and fresh posts produced ρ = .469 (10/12 significant). The difference between legacy and fresh cohorts was small (Δρ = +.044) and not statistically significant by Wilcoxon signed-rank test ($W = 28$, $p = .424$). Per-year analysis showed no monotonic trend: correlations fluctuated between ρ = .29 and .57 across individual years with no systematic decline over time.

This temporal stability carries a nuanced implication. It suggests that the Reddit-AI correlation is not driven solely by historical training data absorption but also reflects an ongoing alignment between Reddit community preferences and the brand quality signals that AI platforms learn from multiple sources. Reddit communities and AI models may converge on similar brand rankings because both are informed by overlapping signals — product quality, market presence, expert reviews — that persist across time periods. The training data pathway remains the most plausible explanation for the *mechanism* of influence (given the zero-citation finding), but the temporal evidence indicates that Reddit's brand consensus is sufficiently stable over time that the distinction between "old training data signal" and "current community consensus" may be less sharp than initially hypothesized.

4.10 API vs. Web UI Reddit Citation Divergence

A critical extension of our citation analysis compares Reddit's presence in API responses versus web UI responses. This comparison tests whether the zero-citation finding — the empirical foundation of the shadow corpus argument — is a property of the platforms themselves or an artifact of the API access method.

API results: Complete Reddit absence. Across 60 API calls (15 queries × 4 platforms: ChatGPT, Claude, Perplexity, and Gemini), zero responses contained any Reddit URL in either citations or sources. This result is consistent with our companion study's finding of zero Reddit URLs across 6,699 citations and extends it across five query intent types. The API-level suppression of Reddit is absolute and platform-universal.

Web UI results: Substantial Reddit presence. The web UI scraper produced a dramatically different picture. Table 5 presents Reddit citation rates across all four web UI platforms.

Table 5. Reddit Citation Rates: API vs. Web UI Comparison (100 Queries per Platform)

Platform	API Reddit Rate	UI Reddit Rate	UI Citation Rate	UI Avg Citations
Google AI Mode	0/15 (0%)	44/100 (44%)	100%	34.5
Perplexity	0/15 (0%)	20/100 (20%)	100%	5.4
ChatGPT	0/15 (0%)	17/100 (17%)	58%	3.3
Claude	0/15 (0%)	0/100 (0%)	19%	2.7

The divergence is stark. Google AI Mode — Google's search-grounded AI interface — cited Reddit in 44% of queries, making Reddit one of the most frequently cited source domains. Perplexity and ChatGPT web UIs cited Reddit in 20% and 17% of queries, respectively. Only Claude maintained the API-consistent zero-Reddit pattern, consistent with its generally lower citation rate (only 19% of queries produced any citations at all).

Intent type modulates Reddit presence. Table 6 breaks down web UI Reddit citation rates by query intent type, revealing that Reddit's presence in AI responses is not uniform but varies systematically with the nature of the query.

Table 6. Web UI Reddit Citation Rate by Query Intent Type

Intent Type	Google AI	Perplexity	ChatGPT	Claude
Validation (*n* = 24)	70.8%	45.8%	25.0%	0%
Informational (*n* = 30)	43.3%	10.0%	0%	0%
Transactional (*n* = 11)	36.4%	27.3%	18.2%	0%
Discovery (*n* = 29)	34.5%	10.3%	24.1%	0%
Navigational (*n* = 6)	0%	0%	33.3%	0%

Validation queries — those seeking opinions or product comparisons ("is the Dyson V15 worth it," "should I buy the AirPods Pro") — produced the highest Reddit citation rates across all platforms that cite Reddit at all. Google AI Mode cited Reddit in 70.8% of validation queries, and Perplexity in 45.8%. This pattern is structurally intuitive: validation queries seek exactly the kind of community consensus that Reddit specializes in providing. When a user asks an AI "is X worth it," the platform's retrieval system recognizes that user reviews and community discussion threads are maximally relevant, and Reddit threads rank highly for precisely these queries.

ChatGPT showed a distinctive pattern: zero Reddit citations for informational queries ("how does noise canceling technology work") but moderate rates for discovery and validation queries. This suggests that ChatGPT's web search is selectively triggered — informational queries may be answered from parametric knowledge without web retrieval, while opinion-seeking queries invoke search and consequently surface Reddit.

Navigational queries ("QuickBooks login page") produced near-zero Reddit citations across all platforms, which is expected: navigational intent targets specific pages, not community discussions.

Quantifying the API-UI divergence. Across the three platforms that surface Reddit in their web UIs (Google AI Mode, Perplexity, and ChatGPT), the combined Reddit citation rate was 27.0% (81 of 300 platform-query combinations). The corresponding API rate across the same three platforms was 0.0% (0 of 45 API calls). This divergence is not a matter of degree — it represents a qualitative difference in citation behavior between two access methods for the same underlying platforms.

---

5. Discussion

5.1 The Training Data Pathway and the API-UI Divergence

Our results provide strong empirical support for the training data pathway hypothesis. Across all 12 consumer product categories, brands that Reddit communities endorse most enthusiastically are the same brands that AI platforms recommend most frequently — a correlation with a mean effect size of $\rho = .554$ that held with remarkable consistency across diverse product domains. This correlation exists in the functional absence of direct citation via APIs: as documented in our companion study, AI platforms cited zero Reddit URLs across 6,699 sampled API citations in their product recommendations (Lee, 2026), and our extended API retrieval test confirmed this with 0 of 60 additional API calls returning Reddit URLs. The influence is real, but invisible to any analysis that examines only API-mediated citations.

However, our web UI experiment (Section 4.10) introduces an important qualification to this narrative. When the same platforms are accessed through their consumer-facing web interfaces rather than their developer-facing APIs, Reddit citations emerge at substantial rates: 44% of queries on Google AI Mode, 20% on Perplexity, and 17% on ChatGPT. This means that Reddit's relationship with AI recommendation systems is more complex than a pure shadow corpus. Reddit operates through *both* pathways simultaneously: its historical community consensus shapes model weights through the training data pathway

(evidenced by the $\rho = .554$ correlation), while its content is also actively retrieved and cited during inference — but only through the web UI access method. The API pathway suppresses Reddit entirely, while the web pathway surfaces it for more than a quarter of queries.

This API-UI divergence has significant implications. The zero-citation finding that motivated this paper was not wrong, but it was *incomplete*: it described the API channel accurately while missing the web UI channel entirely. For GEO researchers who study AI citation behavior, this finding serves as a methodological caution. API-based citation analysis — the dominant method in existing GEO research — captures only one expression of platform behavior. The web interfaces that hundreds of millions of consumers actually use may produce materially different citation patterns.

We retain the term *shadow corpus* but refine its definition to account for the access-method dependency. A **shadow corpus** is a data source S that satisfies two conditions: (1) S exhibits significant rank correlation with model outputs ($\rho(S, M) > 0$, $p < .05$), indicating measurable influence on generation, and (2) S receives systematically suppressed attribution through at least one major access channel (citation frequency < 1% of sampled URLs via that channel), indicating that its influence is partially unattributable through output inspection. Under this refined definition, Reddit satisfies both conditions: $\rho = .554$ ($p < 10^{-8}$) across 12 categories, yet 0 of 6,699 API citations — even though web UIs surface Reddit at meaningful rates. The training data pathway operates independently of whether the platform also retrieves Reddit during inference; the correlation between Reddit consensus and AI recommendations reflects absorbed patterns in model weights, not real-time citation behavior.

The shadow corpus concept has broader implications beyond Reddit. Any data source that was heavily represented in training corpora but is systematically excluded from citation in certain access channels could function as a shadow corpus. Stack Overflow for technical recommendations, Wikipedia for factual claims, and specialized forums for domain-specific advice may all exert similar partially-uncited influence. Reddit is simply the most measurable case because its community consensus mechanism (upvotes) creates quantifiable brand rankings that can be compared against AI outputs, and because the API-UI divergence creates a natural experiment for separating training-data influence from retrieval-time influence.

5.2 Why Reddit Specifically

The strength and consistency of the Reddit-AI correlation raises a natural question: is there something special about Reddit, or would any popular website show similar correlations with AI recommendations? We contend that several features of Reddit make it a uniquely potent training data source for brand recommendations specifically.

Community structure and specialization. Reddit's subreddit architecture creates information silos where discussions are naturally organized by product category. A post in r/MechanicalKeyboards about the best keyboard for programming reaches an audience of over one million subscribers who self-selected into that topic. This structural organization means that brand discussions in Reddit training data come pre-labeled by category and pre-filtered by relevance — exactly the kind of structured signal that language models are optimized to learn from.

The upvote mechanism as quality signal. Reddit's upvote system creates a form of weighted voting that surfaces community consensus. A comment recommending CeraVe with 3,000 upvotes sends a qualitatively different training signal than a blog post recommending the same product. The upvote count effectively labels the recommendation with a confidence score, and language models trained on this data, even without explicit access to upvote metadata, learn from the frequency and prominence of upvoted content. Highly upvoted comments are more likely to appear in top-level positions in archived data, to be quoted in other posts, and to be reproduced across threads — all of which amplify their representation in training corpora.

Volume and temporal consistency. For popular product categories, Reddit generates thousands of brand recommendation threads per year, each producing dozens of comments that mention specific brands. This volume creates statistical consistency: CeraVe does not appear as Reddit's top skincare recommendation by accident. It appears because thousands of independent users across multiple subreddits and years of posts have converged on the same opinion. When this consistency is present in training data at sufficient scale, it is precisely the kind of robust statistical pattern that language models excel at learning.

5.3 Implications for GEO Practitioners

The training data pathway finding carries both validating and cautionary implications for the Generative Engine Optimization community. On the validating side, our results partially support what practitioners have called the "community-consensus hypothesis" — the idea that brands endorsed by online communities enjoy advantages in AI recommendation systems. The data confirms that community consensus and AI recommendations are correlated. Practitioners who have built brand strategies around genuine community engagement rather than content manipulation can find empirical support for their approach in these results.

The cautionary implication is equally important: the training data pathway is not actionable through traditional optimization. Unlike the citation pathway, where a business can improve its content, technical SEO, or domain authority to increase the probability of being cited, the training data pathway reflects historical patterns already encoded in model weights. A brand cannot retroactively change what Reddit users said about it in the training data that current models learned from. Nor can a brand manufacture Reddit consensus in the present and expect it to influence AI recommendations in the near term, since that new content would only enter model training data upon the next training cycle — a timeline measured in months or years, not days.

This distinction has practical consequences. Industry practitioners who advise clients to "optimize for Reddit" as a GEO strategy are conflating different mechanisms. Practitioners would be better served by distinguishing between three strategies: *citation-pathway optimization* (improving content so AI platforms are more likely to cite it during inference), *training-pathway positioning* (building brand prominence in sources likely to be included in future training data), and — informed by our web UI findings — *web UI retrieval positioning* (ensuring Reddit content is well-structured and relevant for query types where platforms actively surface it).

The web UI opportunity. Our API-UI divergence findings (Section 4.10) introduce a more actionable dimension to the Reddit optimization story. While the training data pathway operates on timelines of months to years, the web UI citation pathway operates in real time. Google AI Mode cited Reddit in 44% of queries and 71% of validation queries. Perplexity and ChatGPT cited Reddit in 20% and 17% of queries, respectively. For brands whose target audiences interact with AI primarily through web interfaces — which is the majority of consumers — Reddit is an *active citation source*, not merely a historical training artifact. This means that maintaining high-quality, well-upvoted Reddit presence for validation-type queries ("is X worth it," "X vs Y") has immediate retrieval value in addition to long-term training data value. The intent profile matters: Reddit surfaces most frequently for opinion-seeking and comparison queries, rarely for navigational or purely informational queries.

The forward-looking question. Our data demonstrates a strong alignment between Reddit's *historical* community consensus and *current* AI recommendations. The natural practitioner question follows: does Reddit still matter? If a brand invests in building genuine Reddit community presence today, will that influence future AI models? We contend that the answer is almost certainly yes, but with important caveats about mechanism and timeline.

The structural features that made Reddit a potent training data source in the past remain intact. Subreddit specialization, upvote-driven consensus, and sheer volume of product discussion have not diminished. Google's $60 million annual licensing agreement with Reddit (Reuters, 2024) was signed specifically to secure ongoing access to Reddit content for AI training purposes, suggesting that at least one major AI developer views Reddit as a continuing source of training signal rather than a historical artifact.

The critical variable is the training cycle. Large language models are not trained once and frozen; they undergo periodic retraining on updated corpora. A brand that achieves genuine community endorsement on Reddit today creates potential training signal that may be absorbed into future generations of models. The lag between community consensus formation and model influence is unknown and almost certainly varies by platform, but the mechanism itself is durable.

What our data cannot tell practitioners is how *quickly* new Reddit consensus translates into changed AI recommendations. The training data pathway operates on a fundamentally different cadence than the citation pathway. A page published today might be cited by an AI platform tomorrow if the platform's retrieval system indexes it. A brand endorsed by Reddit today might not

influence AI recommendations until the next major model retraining, which could be months away.

This temporal asymmetry carries one additional implication that practitioners should weigh carefully. Because training data influence is cumulative and slow-moving, brands with years of authentic Reddit community presence hold an advantage that cannot be quickly replicated. A new market entrant cannot manufacture five years of positive community discussion. The training data pathway, in this sense, functions as a moat: brands that have organically earned Reddit's endorsement enjoy a form of AI visibility that is both durable and difficult for competitors to displace through short-term optimization.

5.4 The Market Popularity Confound

We must address a significant alternative explanation for our findings. It is possible — and indeed plausible — that both Reddit brand enthusiasm and AI brand recommendations reflect an underlying third variable: actual market popularity and product quality. CeraVe is popular on Reddit and recommended by AI not because AI learned from Reddit, but because CeraVe is genuinely a good moisturizer that has achieved broad market success. In this interpretation, Reddit and AI are both mirrors of the same reality rather than one influencing the other.

We addressed this confound empirically through partial correlation analysis (Section 4.8) using two independent market popularity proxies: Google Trends relative search interest and Wikipedia monthly page views. The results were striking in their consistency. Controlling for Google Trends, the mean partial ρ was .554 — identical to the unadjusted correlation — with 10 of 12 categories retaining significance. Controlling for Wikipedia page views, the mean partial ρ was .534 (6/11 significant). Controlling for both proxies simultaneously, the mean partial ρ was .529 (6/9 significant). The maximum attenuation across all three specifications was less than 5%.

Several categories showed *stronger* correlations after controlling for popularity. Office and Workspace rose from ρ = .746 to a partial ρ of .937 when controlling for both proxies, suggesting that market popularity was introducing noise rather than driving the Reddit-AI alignment. This pattern is directly inconsistent with the pure market popularity explanation, which predicts uniform attenuation toward zero. The convergence of two independent proxies — one capturing active consumer search behavior, the other measuring general public interest — substantially weakens the case that our findings are a spurious artifact of both sources reflecting the same underlying market reality. While our study design remains observational and we cannot claim causal direction, the partial correlation evidence supports the interpretation that Reddit exerts independent influence on AI recommendations through the training data pathway.

That said, several additional features of our data reinforce this conclusion. The strength of the correlation varies substantially by category in ways that parallel the depth and specificity of Reddit communities rather than the size of the consumer market. Office and Workspace (ρ = .746), a domain served by exceptionally active and opinionated Reddit communities, shows a much stronger correlation than Electronics (ρ = .427), despite the electronics market being substantially larger. If the correlation were driven purely by market share, we would expect the reverse pattern.

Additionally, the existence of 265 Reddit-only brands — brands discussed on Reddit but never recommended by AI — demonstrates that the training data pathway does not perfectly transmit all Reddit signal. If both Reddit and AI were simply mirrors of market popularity, we would expect the overlap to be much higher and the correlation to be uniform across categories. The heterogeneity we observe is more consistent with a model in which Reddit exerts independent influence that varies by the strength and specificity of its community engagement in each domain.

5.5 Connection to the Parent Paper

This study extends and complicates the Reddit Paradox story introduced in our companion paper (Lee, 2026). The companion study documented the paradox: Reddit dominates Google search results for product queries but receives zero AI citations. That finding was accurate for API-mediated access but, as our web UI experiment demonstrates, incomplete. The full picture is more nuanced than either paper initially suggested.

Together, our findings reveal a *three-channel model* of how sources influence AI outputs:

> **The training data pathway** operates during pre-training: patterns from training data are absorbed into model weights and shape generation without attribution. Our correlation analysis (ρ = .554 across 12 categories) provides strong evidence for this channel. It is invisible to citation analysis regardless of access method.
>
> **The API citation pathway** operates during inference via programmatic access: AI platforms retrieve specific URLs and attribute information with citations. For Reddit, this pathway is functionally closed — 0 of 6,759 sampled API citations (6,699 from the companion study plus 60 from our extended test) pointed to Reddit.
>
> **The web UI citation pathway** operates during inference via consumer-facing interfaces: the same platforms retrieve and cite sources, but with different filtering behavior than the API channel. For Reddit, this pathway is substantially open — 27% of web UI queries across Google AI Mode, Perplexity, and ChatGPT surfaced Reddit citations, rising to 48% for validation-type queries.

The existence of these three distinct channels has several implications. First, the original Reddit Paradox is partially resolved: Reddit *is* cited by AI platforms, just not through the access method that researchers and developers typically study. The paradox persists only in the API channel. Second, the training data pathway finding is *strengthened* rather than weakened by the UI results. Even when platforms actively retrieve and cite Reddit during web UI inference, the brand recommendation correlation (ρ = .554) reflects something deeper than retrieval — it reflects absorbed patterns in model weights that manifest regardless of whether a specific Reddit URL appears in the response. Third, the dramatic channel-dependent behavior (0% API vs. 27% web UI) raises questions about *why* APIs suppress Reddit citations that web UIs include. Possible explanations include API output filtering to reduce legal liability, different retrieval configurations for developer vs. consumer contexts, or the absence of web search grounding in API mode for some platforms.

This three-channel model has implications for how we measure and study AI source influence more broadly. API-based citation analysis — the dominant method in existing GEO research — captures only one of three influence channels and, for sources like Reddit, the least active one. A comprehensive account of how AI platforms "decide" what to recommend must consider all three channels: training data influence (measurable only through correlation), API citation behavior (measurable through programmatic output analysis), and web UI citation behavior (measurable through browser automation). These three channels may tell very different stories about the same source.

---

6. Limitations

Our study has several important limitations that qualify the interpretation of our findings and constrain the conclusions we can draw.

Correlational design. Our study establishes a significant positive correlation between Reddit brand consensus and AI brand recommendations but cannot establish causal direction. The training data pathway hypothesis is one plausible explanation; the market popularity confound discussed in Section 5.4 is another. A controlled experiment — for example, training a model with and without Reddit data and comparing its brand recommendations — would be required to establish causation. Such an experiment is beyond the scope of this observational study and likely infeasible with commercial AI platforms whose training data composition is proprietary.

Market popularity proxy limitations. We addressed the market popularity confound through partial correlation analysis using two proxies: Google Trends relative search interest and Wikipedia page views (Section 4.8). While both proxies demonstrated that the Reddit-AI correlation persists after controlling for popularity, neither directly measures market share or sales volume. Google Trends captures relative search interest (available for all categories) while Wikipedia coverage was limited to 36.5% of intersection brands. More direct measures — such as retail sales data, Wirecutter citation frequency, or Consumer Reports

rankings — would provide stronger controls. The convergence of results across two independent proxies strengthens the finding, but the partial correlation results should be interpreted as evidence that general popularity does not fully explain the correlation, not as definitive proof of Reddit-specific causal influence.

Dictionary-based extraction. Our primary brand extraction relied on a dictionary constructed from AI recommendation data, meaning we could only detect Reddit mentions of brands that AI platforms also recommend. We addressed this circularity concern through an independent extraction analysis using spaCy named entity recognition and frequency-based capitalized phrase detection (Section 4.7). The independent method — which uses no AI-derived dictionary — produced a mean correlation of $\rho = .430$ with 7 of 12 categories reaching significance. While lower than the dictionary-based correlation ($\rho = .554$), the independent extraction's noisier methodology (consumer brand NER recall is inherently limited) makes this attenuation expected. The convergence of both methods supports the robustness of the underlying finding.

Temporal scope. Reddit posts were collected within a single window in February 2026, with searches returning posts from the preceding year. AI brand recommendations were collected during January and February 2026. Both datasets represent snapshots rather than longitudinal trends. Reddit community preferences shift over time, and AI model behavior changes with each training update. The correlations we report may strengthen, weaken, or change in character with future data.

Consumer product focus. The correlation analysis is limited to 12 consumer product categories. The 8 business-to-business categories included in our data collection showed zero brand overlap between Reddit and AI recommendations and were therefore excluded from analysis. Whether the training data pathway operates in B2B, professional services, or other non-consumer domains remains an open question.

Platform behavior changes. AI platforms update their models, training data, and retrieval architectures on an ongoing basis. Our findings represent a snapshot of the Reddit-AI relationship as it existed in early 2026. The Google-Reddit licensing agreement (Reuters, 2024) and potential changes to Reddit's data access policies could alter this relationship in either direction.

Web UI scraper fragility. Our web UI citation experiment relies on browser automation that interacts with platform-specific DOM structures and network protocols. These interfaces are not designed for programmatic access and change frequently without notice. Our scraper captured a point-in-time snapshot of web UI behavior in February 2026; the specific CSS selectors, SSE stream formats, and page structures we relied on may have changed by the time of publication. Additionally, the web UI experiment used a single authenticated session per platform, meaning results may reflect personalization effects (account history, geographic location) that API calls do not. While we mitigated this by using fresh conversations for each query, we cannot rule out session-level confounds.

Query set differences between API and UI tests. The API retrieval test used 15 SaaS-focused queries while the web UI experiment used 100 queries spanning 13 domains. This design difference means the API-UI comparison is not perfectly controlled — the UI experiment's broader query set may overestimate Reddit prevalence if Reddit is more common in consumer-oriented queries than SaaS queries. However, the absolute zero-citation result across all 60 API calls (including discovery and validation queries that produced 34-71% Reddit rates in the web UI) suggests the divergence is genuine rather than an artifact of query selection.

Claude web UI citation behavior. Claude produced zero Reddit citations across all 100 web UI queries, consistent with its API behavior. However, Claude also produced citations for only 19% of queries overall, compared to 58-100% for other platforms. The zero-Reddit finding for Claude may therefore reflect a general reluctance to cite external sources rather than Reddit-specific suppression. Our data cannot distinguish between these explanations.

7. Conclusion

Reddit's relationship with AI recommendation systems is more complex than either "never cited" or "frequently cited." Both descriptions are simultaneously true, depending on which access channel you examine. Across 6,759 API-mediated queries (6,699 from our companion study plus 60 from our extended test), not a single response contained a Reddit URL — a suppression so complete that it demands structural explanation rather than statistical treatment. Yet across 300 web UI queries on the same platforms, Reddit appeared in 27% of responses overall and in 48% of validation-type queries. The same platforms that categorically exclude Reddit from their API outputs actively surface it in the consumer-facing interfaces that most users encounter.

Beneath this access-channel divergence lies a deeper finding. Reddit's brand consensus — the brands that its communities upvote, recommend, and defend across thousands of posts and hundreds of thousands of comments — predicts which brands AI platforms recommend with a consistency that cannot be attributed to chance. The mean Spearman correlation of $\rho = .554$ across all 12 consumer product categories, with all 12 reaching statistical significance and 8 surviving Bonferroni correction, establishes that the alignment between Reddit consensus and AI recommendations is both strong and pervasive. This correlation reflects something that operates independently of whether Reddit is cited in any given response: an absorbed pattern in model weights that shapes recommendations regardless of retrieval behavior.

The evidence supports a three-channel model of Reddit's influence on AI systems. The *training data pathway* encodes Reddit's community consensus into model weights during pre-training, producing brand recommendation alignment that persists across access methods and query types. The *web UI citation pathway* surfaces Reddit URLs during inference in consumer-facing interfaces, particularly for validation and opinion-seeking queries where community discussion is maximally relevant. The *API citation pathway* — the channel that prior research, including our own companion study, exclusively examined — suppresses Reddit entirely, creating the appearance of zero influence that initially motivated this investigation.

The practical implications extend the recommendations from our initial analysis. For GEO practitioners, Reddit now presents two optimization opportunities: a long-term training data strategy (building authentic community endorsement that enters future model training) and a more immediate web UI retrieval strategy (ensuring Reddit content is well-structured and relevant for the validation and discovery queries where platforms actively surface it). The relative importance of these strategies depends on the target audience: if customers primarily interact with AI through web UIs, Reddit's web UI citation rate of 17-44% represents a significant visibility channel. If the target is API-powered applications, Reddit remains a shadow corpus whose influence operates exclusively through training data.

For AI researchers, our findings introduce both the shadow corpus concept and the methodological imperative of studying multiple access channels. API-based citation analysis — the dominant method in existing GEO research — captures one-third of the influence model. The web UI channel reveals a qualitatively different citation landscape, and the training data channel reveals influence that no citation analysis can detect. A comprehensive account of AI source influence requires all three lenses.

For the Reddit commenter who challenged our zero-citation finding: the data supports your intuition more comprehensively than we initially expected. Reddit's alignment with AI recommendations is real, measurable, and substantial. It operates through training data absorption, as you suggested — and also through direct web UI citation, which neither of us anticipated.

Author Note

This paper was co-developed by Anthony Lee and Claude (Anthropic, Opus 4.6). The human author provided the research hypothesis (developed in response to community feedback on the companion paper), experimental design, data collection infrastructure (n8n workflow architecture, Reddit API integration), and research direction. The AI contributed statistical analysis implementation, brand extraction algorithm optimization, literature contextualization, and prose drafting. Both authors iteratively refined the arguments through collaborative dialogue.

The origin of this study — a Reddit commenter challenging the companion paper's interpretation of its own findings — is itself an example of the community knowledge generation process we investigate. Reddit's value lies not only in product recommendations but in the quality of its intellectual challenges.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM. https://doi.org/10.1145/3637528.3671882

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Dodge, J., Sap, M., Marasovi■, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1286–1305). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.98

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Lee, A. (2026). Query intent, not Google rank: What best predicts AI citation behavior. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/placeholder

Reuters. (2024, February 22). Google signs $60 million deal with Reddit for AI training data. *Reuters*.