AI Search as an Information Intermediary: Clickstream Evidence on Web Traffic

Abstract

AI search retrieves web information to ground generated answers. We study this new information intermediary using U.S. desktop clickstream data observing use alongside referrals during ChatGPT Search's rollout. Recorded referrals accompany 5.4% of ChatGPT conversation sessions versus 31.1% of Google queries. Reference and technical sites receive larger shares of ChatGPT than Google referrals; ad-supported sites, a smaller share. Conventional Google, Bing, and Yahoo searches decline after ChatGPT adoption, sooner when adoption follows Search's expansion to users without accounts. That expansion raises existing users' ChatGPT use without a detectable overall search change, but searches fall among households with lower prior search activity. These findings suggest AI search may reshape open-web traffic through its own referrals and changes in conventional search.

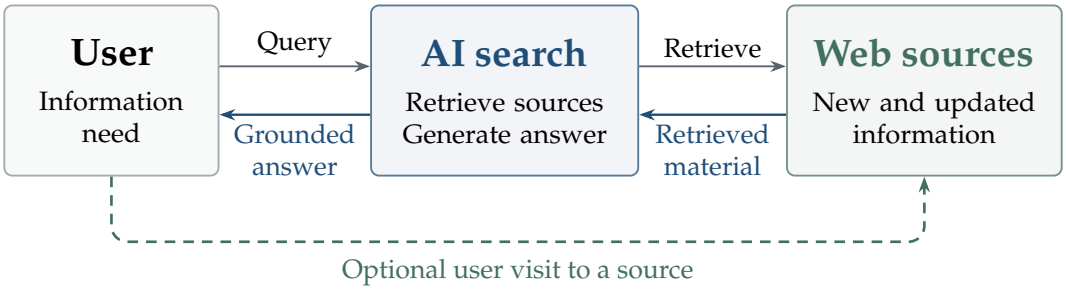
Keywords: AI search; information intermediation; referral traffic; clickstream data; search engines.

1 Introduction

Information intermediaries shape how people find information and which websites receive their visits. Search engines help users locate information and direct traffic to external sources (Bakos, 1997). For websites, these visits connect information provision with opportunities to attract readers, customers, and contributors (Dellarocas et al., 2013). AI search introduces a new form of intermediation, raising questions about how that connection changes when users receive answers within the intermediary. This role as a search intermediary is also reflected in regulation: in August 2026, the European Commission designated ChatGPT a "very large online search engine," rather than a "very large online platform," under the Digital Services Act.[1]

AI search refers here to a system that retrieves web information and uses a large language model to synthesize an answer grounded in it (Nakano et al., 2021; Figure 1). Web retrieval is an additional capability, distinct from language generation. It lets the system consult external sources at the time of a request, so its answer can incorporate newly published or updated information beyond the model's training knowledge. OpenAI introduced ChatGPT Search on October 31, 2024, and subsequently expanded access to additional user groups (OpenAI, 2024).

Figure 1. AI search can use a web source without sending the user to it.



Notes: Solid arrows distinguish the user's query and generated answer from the system's retrieval of web information. The dashed path is an optional user visit to a source; retrieval by the system is not a user referral. Not every ChatGPT conversation involves retrieval. Our data record platform activity and referrals, not the sources retrieved or the quality of answers.

[1] European Commission, "Commission designates ChatGPT, Reddit, Roblox under Digital Services Act," August 31, 2026, https://digital-strategy.ec.europa.eu/en/news/commission-designates-chatgpt-reddit-roblox-under-digital-services-act (accessed September 21, 2026).

For websites, supplying information to AI search need not generate a visit: users can read the answer without opening its sources. Yet AI search can also help users discover sources and encourage visits—a tension familiar from research on content aggregation (Dellarocas et al., 2016; Chiou and Tucker, 2017; Calzada and Gil, 2020). Beyond its own referrals, AI search can change how much people use conventional search, and with it the visits that search engines send to websites. Its effects on websites therefore run through two channels: its own referrals, for which established search engines provide a benchmark, and changes in conventional search.

We study both channels using URL-level Comscore U.S. desktop clickstream data from October 2024 through July 2025, spanning the initial ChatGPT Search rollout. These records do not directly identify conversations, queries, or referrals: a single ChatGPT conversation can produce many page loads and background requests, and the source of a website visit must be inferred from its referrer and timing. We reconstruct ChatGPT conversation sessions, Google queries, and attributed referrals to external websites, and follow households' weekly conventional searches in a balanced panel. Observing intermediary activity alongside external referrals allows us to include occasions that generate no recorded referral. These occasions belong in the denominator for measuring referral incidence: the share of conversation sessions or queries associated with at least one referral. This distinguishes an intermediary's popularity from how frequently its use connects users with external websites.

We find substantially lower referral incidence for ChatGPT than for Google Search. A recorded referral to an external website accompanies 5.4% of ChatGPT conversation sessions; the corresponding figure for Google queries is 31.1%. The difference persists within households and after adjustment for surrounding browsing contexts. Because tasks and activity units differ across services, and not every ChatGPT conversation retrieves from the web, the comparison does not isolate an interface effect. It nevertheless captures the referral channel websites encounter in practice: how frequently actual use of the new intermediary sends users to external websites.

The referrals ChatGPT does generate reach a distinctive set of destinations. Reference, academic, developer, and tools/SaaS websites account for larger shares of ChatGPT referrals than of Google referrals, while ad-supported websites account for a smaller share. When pooled across households, ChatGPT referrals are less concentrated across destinations than Google's. Within households, they are similarly or more concentrated—a pattern consistent with different households visiting different sets of websites.

For the second channel, we use a balanced panel of 45,386 households and two complementary designs. As in Padilla et al. (2026), the first compares households that adopt ChatGPT with households that adopt later. Weekly Google, Bing, and Yahoo searches rise briefly at adoption and then fall by about 39% of the pre-adoption average in weeks 20–30. To distinguish Search from LLM adoption more generally, we compare households adopting before and after February 5, 2025, when ChatGPT Search became available without an account. The comparison holds the stage of adoption fixed while Search availability differs. Searches decline sooner after adoption among later adopters, providing suggestive evidence of a role for ChatGPT Search beyond general LLM adoption.

Because households choose when to adopt, the second design uses the February 5 expansion itself, a date set by OpenAI rather than by households (OpenAI, 2024). It compares households already using ChatGPT without an account with prior Gemini users, who had likewise used an AI service before the expansion. Newly eligible households record more ChatGPT sessions afterward, without a detectable overall change in conventional search. The search response nevertheless differs significantly with prior search activity, which reflects how heavily households rely on conventional search and possibly what they use it for. Households with lower prior activity reduce weekly searches by 3.71, or 35.8% of their pre-expansion mean, whereas the estimate for higher-activity households is positive and imprecise.

Differences in what households search for offer a possible explanation. Before the expansion, lower-activity households direct larger shares of their conventional-search

referrals to tools/SaaS and e-commerce brand websites, and higher-activity households larger shares to social media and adult websites. After the expansion, such referrals to tools/SaaS and brand websites decline in the full access sample, although these estimates are significant only at the 10% level. These patterns are consistent with substitution for some task-oriented and product-related searches, while searches for content or interaction on a platform itself may be harder to replace.

Our first contribution is to characterize AI search as a new information intermediary. Recent studies examine how generated summaries within Google Search affect traffic to publishers and Wikipedia (Agarwal and Sen, 2026; Khosravi and Yoganarasimhan, 2026; Wang et al., 2026). We study instead a separate conversational service used alongside conventional search, benchmarking its referral role against Google's. Its referrals are less frequent, weighted toward reference and technical rather than ad-supported destinations, and less concentrated in aggregate but similarly or more concentrated within households.

Second, we distinguish the role of AI search in conventional-search declines from that of general LLM adoption. Padilla et al. (2026) document search declines after households adopt LLMs in 2022–2023, before the launch of ChatGPT Search, whereas Gholami et al. (2026) find a short-lived rise in search and suggest that LLMs can complement it. The timing of adoption provides suggestive evidence that broader Search availability contributes to these declines. The access expansion reduces conventional search among households with lower prior activity, without a detectable overall change. Together with the referral patterns, these findings suggest that substitution depends partly on the needs served by search, which may help reconcile these accounts.

Third, we make the referral role of conversational AI measurable. Reconstructing conversation sessions and search queries from clickstream records and attributing external visits to their source allows us to observe use with and without referrals. This makes referral incidence comparable across services as used, complementing research on the commercial outcomes of arriving ChatGPT referrals (Kaiser and Schulze, 2026)

and on how LLM adoption changes search and the concentration of web traffic (Gholami et al., 2026).

Section 2 develops the conceptual framework. Section 3 describes the data and methods, Section 4 presents the referral findings, Section 5 examines how ChatGPT adoption and expanded access to ChatGPT Search change conventional search, and Section 6 discusses their implications.

2 Conceptual Framework

AI search can affect website traffic through two channels: the referrals generated by its own use, and changes in how much people use conventional search. Through the first, AI search can change both the need for an external visit and the reason for making one. An answer that supplies the information a user seeks may substitute for a visit, while one that reveals a useful source may encourage further exploration. Research on content aggregation documents this tension between supplying information within an intermediary and helping users discover external content (Dellarocas et al., 2016; Chiou and Tucker, 2017; Calzada and Gil, 2020).

For AI search, the remaining value of a website visit may lie in examining detailed evidence, using an interactive tool, or completing a transaction. These purposes connect referral incidence with destination composition: visits may decline for needs addressed within the answer while continuing for activities that require a website. The destinations receiving those visits can also differ from the sources used to construct the answer. Referral patterns therefore reveal which websites users continue to visit alongside AI search, a distinct role from supplying information to it.

The distribution of those referrals depends on both individual choices and their overlap across households. Lower search costs can help users find niche offerings (Brynjolfsson et al., 2011), but broader discovery by individuals can coexist with greater aggregate concentration when their choices converge on the same destinations (Fleder and Hosanagar, 2009).

The reverse is also possible. Households could each concentrate their referrals on a few specialist websites yet collectively generate a dispersed referral stream because they visit different sites. Aggregate dispersion can thus arise from differences across households as well as from dispersion within them. Examining both levels distinguishes how strongly an intermediary's traffic converges on particular websites from how evenly individual households distribute their referrals.

The second channel runs through conventional search. An answer may replace a conventional-search query, but it may also introduce a topic or source that prompts further searching. Web retrieval widens the questions an answer can address. A model answering from its training knowledge can help with a calculation or a draft, but questions about current products, prices, or events would otherwise send the user to a search engine. Access to AI search can therefore change conventional search in ways that general LLM use need not.

How much conventional search an answer can replace also depends on what the search is for, a distinction long used to classify web queries (Broder, 2002). Some searches seek information or help that an answer can supply directly, such as a calculation, a piece of writing, or product information. Others lead to content or interaction available only on a particular website, such as a video or a discussion, that an answer cannot supply. Households can differ in this mix of needs, so the same access can reduce search for some and leave it unchanged for others; an average response can conceal the difference. The destinations a household's searches lead to offer an observable, if indirect, indication of the mix. Because queries do not invariably produce external visits, changes in searching need not change search-referred visits in proportion.

For a destination website, the two channels combine. When AI search replaces a conventional search, the visit that search would have produced depends on the user continuing from the answer to the website. A website could therefore account for a relatively large share of AI referrals yet receive fewer visits overall. How much traffic it retains turns on the distinction running through both channels: whether an answer can meet the user's need, or only the website can.

3 Data and Method

3.1 Data coverage and observable signals

We use URL-level records from the Comscore U.S. Desktop Panel from October 2024 through July 2025. The records follow the same households across ChatGPT, conventional search engines, and destination websites, combining URLs and timestamps with source referrers and request characteristics. The data support two analyses: comparing each intermediary's activity with the referrals it generates, and following households' conventional search over time. The full sample contains between 168,467 and 238,315 active households per month. For the analysis of conventional search, we use a balanced panel of 45,386 households with at least four recorded webpage visits in every month.[2]

A webpage or conversation can generate many network requests, so raw request counts do not directly measure use. We count a webpage visit only when the record is a successfully loaded HTML page, and we use specific ChatGPT requests to reconstruct conversations and whether a user was signed in. Background requests do not count as visits.[3]

3.2 Activity and referral measures

For referral incidence, we measure conventional-search activity as queries and ChatGPT activity as conversation sessions. Queries are identified from foreground Google, Bing, and Yahoo URLs matching their search patterns. ChatGPT runs as a single-page application, so one conversation can produce many page loads. We treat loads of the same conversation on the same computer as one session, starting a new session when more than an hour passes between loads; different conversations remain separate

[2]Sections and tables numbered OA refer to the Online Appendix. See Section OA.1.1 for the panel threshold.

[3]See Sections OA.1.2–OA.1.3 for filtering rules and the ChatGPT requests used.

sessions.[4] The resulting conversation measure provides an observed denominator for assessing how frequently ChatGPT use generates visits to external websites.

To associate external visits with intermediary use, we combine source referrers with timestamps. A clean referral is a foreground visit to an external website whose HTTP referrer identifies the source intermediary, after excluding visits within the platform itself, sign-in pages, technical infrastructure, and search services. Timestamps link these visits to observed queries or conversation sessions; visits without an observable source referrer remain unattributed.[5]

Referral incidence is the share of conversation sessions or queries associated with at least one clean referral. Each occasion enters the numerator at most once, while occasions without a recorded referral remain in the denominator; their share is absorptiveness. To account for persistent household differences and common calendar changes, we compare ChatGPT and Google using household and week fixed effects, then additionally adjust for surrounding browsing context (Section OA.2.2).

3.3 Destinations and surrounding context

An intermediary's relationship with the web also depends on which websites receive its referrals. We compare ChatGPT and Google in the composition of their referral destinations, examining both the content websites provide and how they monetize their activity. We classify 4,266 destination domains by content type and monetization model using GPT-4o and a common codebook. Content-type analyses use the 3,245 domains with classifier confidence of at least 0.90.[6]

Interpreting referral differences also requires characterizing the contexts in which the services are used. Because prompt text is unavailable, we classify the other websites a household visits in the fifteen minutes before and after each ChatGPT conversation session or Google search session, using the same content categories. These categories

[4] See Sections OA.1.3–OA.1.4 for request paths and sensitivity to alternative session boundaries.

[5] See Section OA.2.1 for attribution windows, exclusions, and validation checks.

[6] See Section OA.4 for classification validation, including agreement rates and category-level disagreements, and sensitivity to the confidence restriction.

describe surrounding activity rather than users' prompts. A category is dominant when it accounts for at least half of this browsing. Sessions with no classifiable surrounding browsing are labeled solo; those with browsing but no dominant category are labeled mixed and excluded from the category comparison (Section OA.3). We compare the services' surrounding-context shares and measure ChatGPT's referral incidence within each context, distinguishing where use is common from where it most frequently generates external visits.

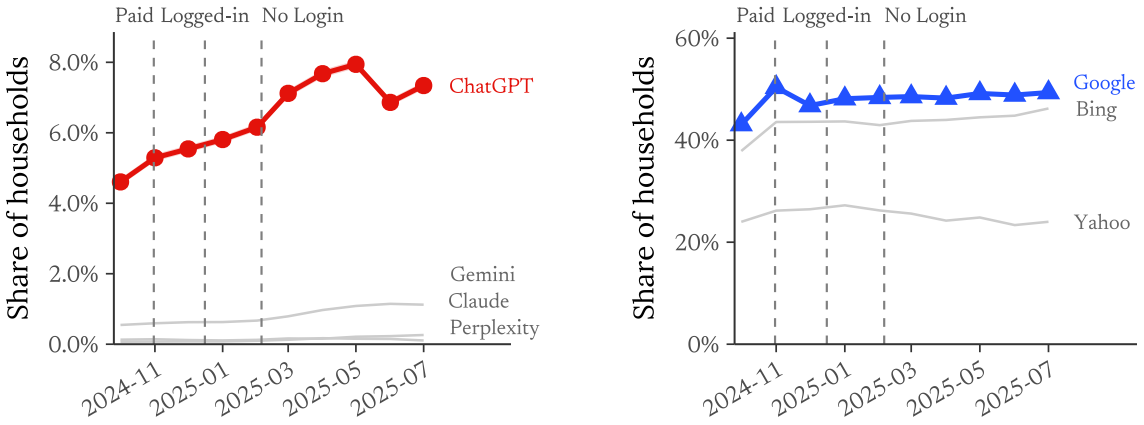
Destination composition does not establish whether referrals converge on a few websites or spread across many. We therefore measure concentration using a floor-corrected normalized Herfindahl index, calculated separately across aggregate referrals and within household-weeks. The former captures whether households' referrals converge on the same destinations; the latter captures how broadly each household's referrals are distributed. This distinction allows aggregate dispersion to coexist with concentrated referrals within individual households (Section OA.5).

3.4 ChatGPT Search access and conventional-search measures

The second channel concerns conventional-search use during the rollout of ChatGPT Search, which reached users in three steps: paid subscribers on October 31, 2024, free logged-in users on December 16, and users without an account on February 5, 2025 (OpenAI, 2024). Figure 2 places these dates alongside the evolution of ChatGPT and conventional-search reach.

Analyses of conventional search follow the 45,386 households of the balanced panel, so that each household's search can be compared before and after it adopts ChatGPT or gains access to Search. Weekly conventional searches sum the Google, Bing, and Yahoo query records defined in Section 3.2; each qualifying record contributes one count, so a search is a recorded result-page load rather than a distinct textual query. ChatGPT use is measured by the same conversation-session rule. The two estimation designs in Section 5 use ACS age-by-income weights to align their samples with the U.S. population (Section OA.6.3).

Figure 2. ChatGPT reach rises during the rollout of ChatGPT Search while Google reach remains stable.



(a) Households visiting an LLM platform.

(b) Households issuing a search query.

Notes: Each series reports the monthly share of active Comscore U.S. desktop households that visits an LLM platform (panel a) or issues a search-engine query (panel b). Dashed vertical lines mark the October 31, December 16, and February 5 ChatGPT Search access expansions. ChatGPT reach rises from 4.6% in October 2024 to a peak near 7.9%; Google remains near 49%.

4 ChatGPT as a Referral Intermediary

The first channel depends on how frequently ChatGPT use generates external visits and how those visits are distributed across websites. We characterize ChatGPT's referral incidence and destination patterns, using Google Search as a benchmark for an established information intermediary. Comparisons within households and surrounding browsing contexts help assess how these patterns differ between the services as used.

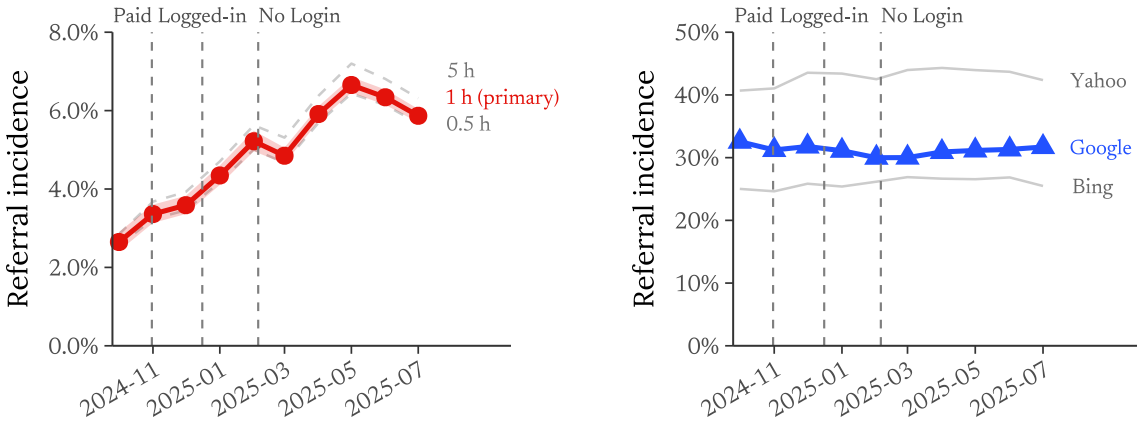
4.1 ChatGPT conversations have low referral incidence

Understanding ChatGPT's role in website traffic requires measuring how frequently its use leads to an external visit: widespread use need not generate frequent referrals. We find that a clean outbound referral accompanies 5.4% of observed ChatGPT conversation sessions. Google Search provides a benchmark for an established information intermediary: 31.1% of its queries are associated with a clean referral (Table 1). These rates keep occasions with no recorded referral in the denominator.

The difference persists throughout the observation window. ChatGPT's monthly referral incidence rises from approximately 2.6% to a peak near 6.7%, remaining below Google's per-query incidence (Figure 3). Varying the ChatGPT conversation-session boundary from thirty minutes to five hours changes the estimates little. The low incidence holds across months and alternative definitions of a conversation session.

Nonreferral is also widespread across households. During the ten-month window, 74.4% of 56,578 ChatGPT-active households have no recorded clean ChatGPT referral, compared with 9.6% of 291,747 Google-querying households. Nonreferral, or absorptiveness, does not establish that an information need was resolved: it can reflect a satisfactory answer, an abandoned task, or an external visit with a missing referrer. The household measure nevertheless shows that the pattern extends beyond individual nonreferring conversations to households with no recorded ChatGPT referral over the observation period.[7]

Figure 3. Recorded referral incidence is lower for ChatGPT across conversation-session definitions.



(a) ChatGPT conversation sessions.

(b) Search-engine queries.

Notes: Points report the monthly share of ChatGPT conversation sessions (panel a) or search queries (panel b) associated with at least one clean referral. In panel (a), the red series uses the main one-hour session boundary, and gray series use thirty-minute and five-hour boundaries. Panel (b) shows Google, Bing, and Yahoo queries, with Google highlighted. Dashed lines mark the three access expansions. The session rule changes ChatGPT's estimates little: ChatGPT rises from approximately 2.6% to a peak near 6.7%, compared with 31.1% of Google queries over the full period (Table 1).

[7]See Section OA.2.3 for comparisons with selected public benchmarks, accounting for differences in denominators, attribution windows, population coverage, and traffic filters.

Table 1. ChatGPT conversations have lower recorded referral incidence than Google queries.

	Google	ChatGPT
Occasion of use	Search query	Conversation session
Occasions	61,689,308	409,133
Clean referral visits	31,464,344	168,111
Referral incidence	31.1%	5.4%
Active households	291,747	56,578
Households with no clean referral	9.6%	74.4%

Notes: The sample covers Comscore U.S. desktop activity from October 2024 through July 2025. An occasion of use is one Google query or one ChatGPT conversation session. Clean referral visits are visits to third-party websites that meet the referral definition in Section 3.2; referral incidence is the share of occasions producing at least one such visit. The last row is the share of active households with no clean referral from that service during the ten months.

Accounting for household and week differences. The aggregate difference could partly reflect who uses each service. We therefore aggregate referral incidence by household, week, and intermediary and estimate

$$\text{Referral incidence}_{itq} = \beta \mathbf{1}\{q = \text{ChatGPT}\} + \alpha_i + \gamma_t + \varepsilon_{itq}, \tag{1}$$

where the outcome is intermediary q's referral incidence for household i in week t. Household fixed effects account for persistent differences across households, and week fixed effects account for common calendar changes. Standard errors are clustered by household.

In the regression sample, referral incidence averages 8.1% for ChatGPT and 39.6% for Google, a difference of 31.5 percentage points. These means give equal weight to household-week-intermediary cells rather than individual occasions of use, so they differ from the aggregate rates. With household and week fixed effects, the estimated difference is 29.0 percentage points ($\hat{\beta} = -0.290$, $SE = 0.002$; Table 2, panel A). Lower ChatGPT referral incidence thus persists after adjustment for household and week effects.

Surrounding contexts and referral incidence. The same household may use the services for different purposes. Surrounding browsing provides an observable, although

imperfect, way to characterize those differences. Reference/knowledge and tools/SaaS contexts account for larger shares of ChatGPT use than Google use, by 9.0 and 4.3 percentage points, respectively. News/journalism, games and guides, and adult contexts account for smaller shares. ChatGPT conversation sessions are 3.6 percentage points less likely than Google search sessions to be solo, with no classifiable browsing in the fifteen minutes before or after the session (Figure 4, panel a). This may reflect ChatGPT's use within broader information-seeking activities involving reference material, consistent with evidence that most LLM use occurs alongside search (Gholami et al., 2026).

To account for surrounding context, we label each ChatGPT conversation by its surrounding browsing and each Google query by the context of the search session it belongs to. We then compute referral incidence separately for each context within each household, week, and service, giving each such cell equal weight. Because this sample excludes unclassifiable contexts, it differs from Panel A's. Within this sample, adding context fixed effects to the household-and-week specification changes the ChatGPT coefficient only slightly, from -0.334 to -0.329 ($SE = 0.002$; Table 2, panel B). The referral difference persists after adjustment for observed context, although, as with the raw comparison, it is not an isolated interface effect.

Within ChatGPT, contexts where use is relatively common are not necessarily those where referrals occur most frequently. Referral incidence is highest in developer/technical contexts (13.6%), academic research (12.9%), and e-commerce brands (9.5%), and lowest in adult, social, and reference/knowledge contexts (Figure 4, panel b). Reference/knowledge illustrates the distinction: it accounts for a relatively larger share of ChatGPT use, yet its referral incidence is among the lowest. Where ChatGPT is used most is not where it most often sends users onward.[8]

[8]See Section OA.3 for sensitivity to the context-construction rules.

Table 2. ChatGPT's lower referral incidence persists after household, week, and context adjustment.

	(1)	(2)	(3)
Panel A. Weekly incidence by household and service			
ChatGPT indicator	-0.315^{***}	-0.289^{***}	-0.290^{***}
	(0.001)	(0.002)	(0.002)
Household fixed effects	No	Yes	Yes
Week fixed effects	No	No	Yes
Observations (cells)	2,790,136	2,790,136	2,790,136
Panel B. Weekly incidence by household, service, and context			
ChatGPT indicator	-0.373^{***}	-0.334^{***}	-0.329^{***}
	(0.001)	(0.002)	(0.002)
Household fixed effects	No	Yes	Yes
Week fixed effects	No	Yes	Yes
Context fixed effects	No	No	Yes
Observations (cells)	3,809,132	3,809,132	3,809,132

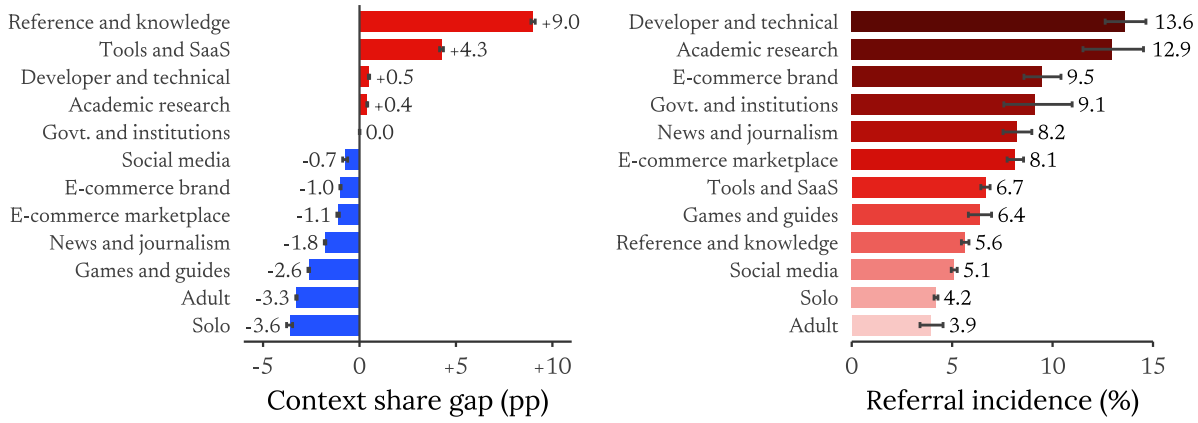
Notes: The sample covers Comscore U.S. desktop activity from October 2024 through July 2025. The dependent variable is referral incidence on a zero-to-one scale. The ChatGPT indicator compares ChatGPT with Google; a coefficient of -0.329 represents 32.9 percentage points lower incidence. Panel A computes referral incidence for each household, week, and service. Panel B computes it separately by browsing context, using all sessions, including those with nearby use of other search or AI services; Google queries take the context of their search session. Solo, mixed, and unclassified contexts, and queries outside a search session, are excluded. Observations are equally weighted cells, not households; splitting by context can raise their number even though some activity is excluded. The sample is fixed across columns within each panel. Column (2) adds household fixed effects in Panel A and household and week fixed effects in Panel B. See Section OA.2.2 for further construction and sample details. Household-clustered standard errors appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.2 ChatGPT referrals have a different destination mix

When referrals occur, which parts of the web receive them? We find that knowledge and tools account for larger shares of ChatGPT's classified referral traffic than of Google's. Reference/knowledge websites account for a 13.1 percentage-point larger share, followed by tools/SaaS at 8.8 points, academic research at 5.4 points, and developer/technical websites at 4.7 points. Social media, by contrast, accounts for a 15.7-point smaller share (Figure 5, panel a).

The destination mix also differs in how the receiving websites finance their content and services. Ad-supported websites account for a 27.5-percentage-point smaller share of ChatGPT referrals than of Google referrals. Non-profit/public, freemium SaaS, and

Figure 4. Context prevalence and ChatGPT referral incidence follow different orderings.



(a) ChatGPT minus Google surrounding-context share.

(b) ChatGPT referral incidence by context.

Notes: Surrounding context is each household's other foreground browsing in fifteen-minute windows before and after a ChatGPT conversation session or Google search session. Browsing during the session, platform/search/AI pages, and visits referred by these intermediaries are excluded. Each session takes the high-confidence content category accounting for at least half of its surrounding browsing. Sessions with no classifiable browsing are labeled solo; those with classifiable browsing but no dominant category are labeled mixed. Panel (a) shows ChatGPT minus Google context shares, calculated after excluding mixed and other contexts. Panel (b) shows the percentage of ChatGPT sessions with a referral within each displayed context. Labels in panel (a) give the share differences in percentage points. Error bars are 95% confidence intervals.

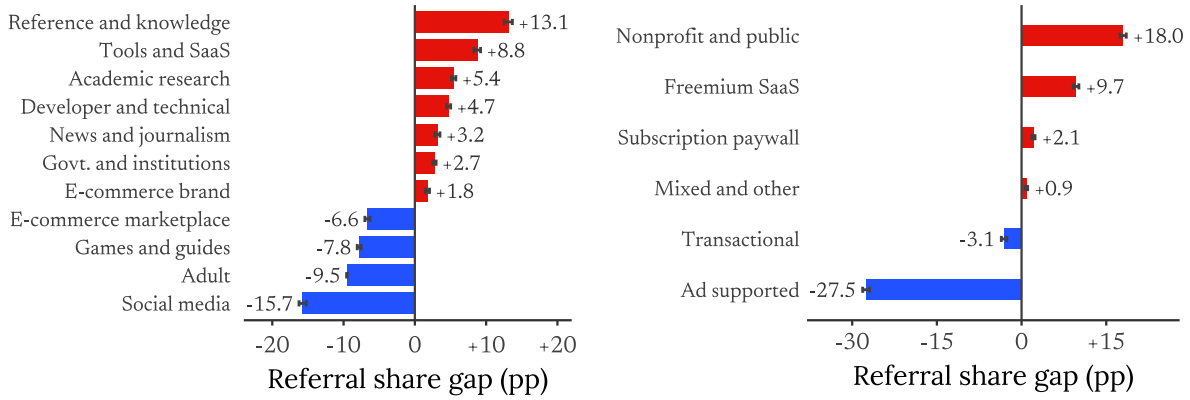
subscription websites account for larger shares, while transactional destinations account for smaller shares (Figure 5, panel b). The content and monetization patterns persist when the classifier-confidence threshold is relaxed from 0.90 to 0.60 (Section OA.4.4).

These shares must be read alongside ChatGPT's low referral incidence. A larger share is not a larger number of visits: because ChatGPT sends far fewer referrals overall, a category can rank higher in ChatGPT's referrals while receiving fewer visits from ChatGPT than from Google, and its total traffic need not rise. The two intermediaries differ both in how often use generates an external visit and in the content and business models those visits reach.

4.3 Concentration in aggregate and within households

An intermediary's role in directing web traffic depends not only on which types of websites receive referrals, but also on whether a few destinations capture most of that traffic. We find that ChatGPT referrals are less concentrated than Google's when pooled

Figure 5. Tools and knowledge sites have larger shares of ChatGPT referrals; ad-supported sites have a smaller share.



(a) ChatGPT minus Google referral share by content type.

(b) ChatGPT minus Google referral share by monetization model.

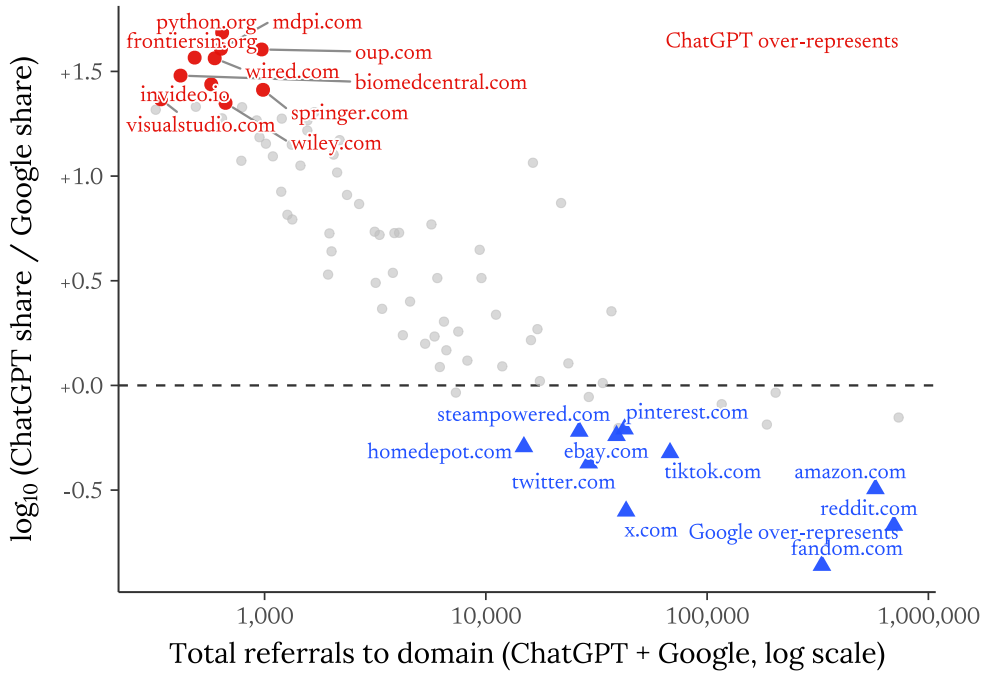
Notes: The sample is the 3,245 domains classified with confidence of at least 0.90, out of 4,266 domains: the union of the 2,500 domains most referred by LLM platforms and the 2,500 most referred by search engines (Section OA.4). Bars report each category's share of ChatGPT referrals minus its share of Google referrals, in percentage points. Positive values indicate categories more common in ChatGPT's referral traffic to these domains than in Google's. Panel (a) groups destinations by content type; panel (b) groups them by monetization model. Labels give the share differences in percentage points. Error bars show 95% pointwise normal intervals for differences in proportions.

across households, but similarly or more concentrated within households. Across all destination domains, Google's floor-corrected normalized Herfindahl index is 3.47 times ChatGPT's. The ratio ranges from 1.87 to 3.47 across the five sets of destinations in Table 3. Higher index values indicate greater concentration; the aggregate difference persists across these alternative restrictions on the destinations included (Section OA.5.2).

Comparing each domain's share of the two referral streams helps characterize this difference (Figure 6). Large destinations such as YouTube, Reddit, and Wikipedia account for relatively larger shares of Google referrals. Academic, reference, and tool-oriented websites with lower combined referral volume in our sample instead tend to account for relatively larger shares of ChatGPT referrals. These comparisons describe each website's position within the two referral streams, rather than the absolute number of visits supplied by either intermediary.

Lower aggregate concentration need not extend to individual households. In comparisons with household and week fixed effects, the ChatGPT–Google concentration

Figure 6. Domains with higher relative ChatGPT referral shares tend to have lower referral volume.



Notes: Each point represents one destination domain. The horizontal axis reports total ChatGPT and Google referrals to the domain on a logarithmic scale. The vertical axis reports the base-ten logarithm of the domain's ChatGPT referral share divided by its Google referral share. Values above zero indicate a larger share of ChatGPT referrals than of Google referrals; values below zero indicate the reverse. Domains with larger relative ChatGPT shares tend to be lower-volume academic, reference, and tool websites, while those with larger relative Google shares include large social, marketplace, and encyclopedia websites.

Table 3. Google referrals are more concentrated than ChatGPT referrals in aggregate.

Destinations included	ChatGPT	Google	Google/ChatGPT
Each service's 100 most-referred	0.0360	0.0674	1.87
Each service's 1,000 most-referred	0.0148	0.0396	2.67
Each service's 2,500 most-referred	0.0113	0.0332	2.94
Each service's 10,000 most-referred	0.0078	0.0270	3.47
All domains	0.0055	0.0191	3.47

Notes: Entries are floor-corrected normalized Herfindahl indices calculated from each intermediary's aggregate referral shares; higher values indicate that more traffic converges on fewer destinations. Rows vary which destinations are included, to assess whether the comparison depends on lower-volume destinations; the first four rows combine the two services' most-referred domains. The final column divides Google's index by ChatGPT's. Google's referrals are more concentrated in every set and 3.47 times as concentrated across all domains. Table OA.5.1 reports further sets defined by percentile, rank, cumulative share, and minimum referrals, and for the 4,266 classified domains.

difference is statistically indistinguishable from zero at the one- and two-referral thresholds (Table 4). At the five- and ten-referral thresholds, the ChatGPT coefficients are positive, at 0.035 and 0.056, respectively, indicating greater concentration in these more active samples. Each comparison requires the stated minimum referrals from both intermediaries and at least two distinct destinations for the measured intermediary.[9]

Table 4. ChatGPT has higher within-household concentration at the five- and ten-referral thresholds.

	Minimum referrals from each service			
	≥ 1	≥ 2	≥ 5	≥ 10
ChatGPT indicator	−0.001	0.000	0.035***	0.056***
	(0.002)	(0.002)	(0.003)	(0.006)
Mean normalized HHI, Google	0.072	0.071	0.075	0.077
Mean normalized HHI, ChatGPT	0.069	0.070	0.107	0.130
Household fixed effects	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes
Observations	48,641	35,644	12,684	3,721

Notes: The dependent variable is the household-week's floor-corrected normalized Herfindahl index, an index equal to zero for equal shares across observed destinations and approaching one as one destination dominates. Columns retain household-weeks meeting the indicated minimum referral count for each intermediary. The index is defined only for household-weeks with at least two distinct destinations for the measured intermediary, so single-destination weeks are dropped. The ChatGPT-indicator coefficient compares ChatGPT with Google within the same household and week; positive values indicate greater concentration through ChatGPT. The rows of mean normalized HHI report each intermediary's average household-week index as a level benchmark for the indicator. Standard errors clustered by household appear in parentheses. The difference is indistinguishable from zero at the one- and two-referral thresholds and positive among heavier referring households. See Table OA.5.2 for thresholds from 1 to 20 referrals. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

ChatGPT referrals are less concentrated than Google's in aggregate but similarly or more concentrated within households. This pattern is consistent with different households concentrating their referrals on different sets of websites. The aggregate comparison agrees with Gholami et al. (2026), who find that traffic LLMs route to external websites is no more concentrated than traffic from traditional search.

[9] See Section OA.5 for the full threshold comparisons and supplementary destination characteristics.

5 ChatGPT Use and Conventional Search

AI search can also affect websites through the second channel, conventional search. Estimating this effect requires knowing how households' conventional searches would have evolved without access to AI search. We approach that counterfactual with two complementary designs: household adoption of ChatGPT (Section 5.1) and the February 5, 2025 expansion of Search access among existing users (Section 5.2). Both use the balanced panel and the weekly conventional searches defined in Section 3.4.

5.1 Evidence from household adoption

A natural starting point is to follow households as they start using ChatGPT. If conversations replace some searches, conventional search should fall after adoption; if they prompt further searching, it could rise. Comparing adopters' change in searches with that of households that adopt later shows the size and timing of the response and connects our setting to evidence on earlier LLM adoption (Padilla et al., 2026). A further comparison helps separate Search from LLM adoption: households adopting after the February 5, 2025 expansion could use Search without an account, whereas earlier adopters initially needed one. Differencing their adoption estimates at the same stage of adoption yields a triple difference that removes the response common to adopting ChatGPT. Because households choose when to adopt, both comparisons describe what accompanies adoption rather than the effect of Search access.

5.1.1 Search declines following ChatGPT adoption

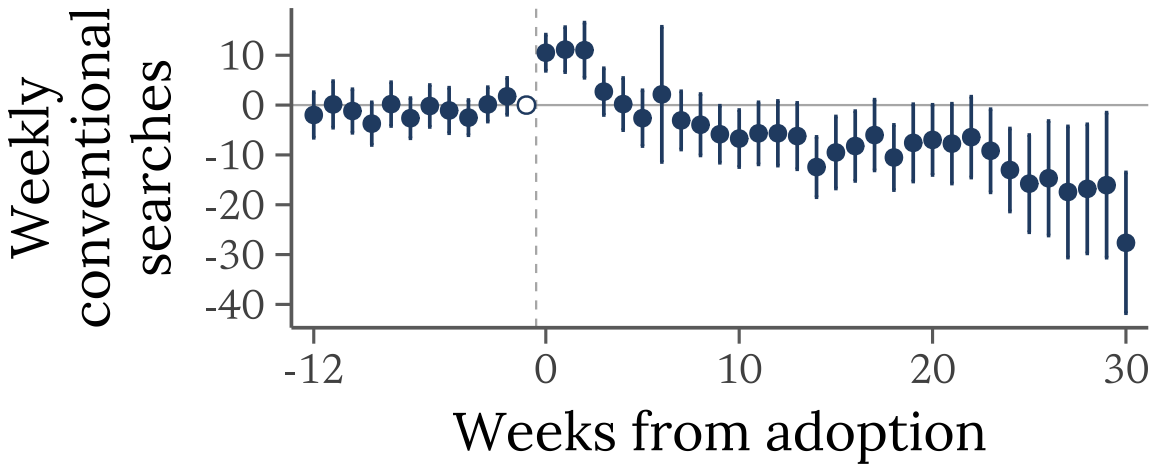
Following Padilla et al. (2026), who study online behavior after households adopt LLMs in 2022–2023, we date adoption to the first of three consecutive weeks with recorded use; here, use is a recorded ChatGPT conversation session. Our estimation sample comprises 1,008 adopters with pre-adoption observations and ACS weights (Section OA.6.1). The Callaway and Sant'Anna (2021) estimator compares each adoption cohort's change in weekly conventional searches with the change among households that adopt later;

households that never adopt do not enter the comparison. The design dates ChatGPT adoption, not use of its Search feature.

Conventional searches rise at adoption and then decline (Figure 7). Relative to later-adopting households, weekly searches increase by 8.84 ($SE = 1.94$) in weeks 0–3, fall by 5.73 ($SE = 2.41$) in weeks 4–19, and fall by 13.80 ($SE = 4.64$) in weeks 20–30 (Table 5). The later decline is about 39% of adopters' average of 35.34 weekly searches before onset. The joint test of pre-adoption estimates in weeks -12 to -2 does not reject zero ($p = 0.295$). Padilla et al. (2026) report comparable declines of roughly a fifth to a third for adopters in 2022–2023.

The increase in searches at adoption may reflect ChatGPT prompting further searching, as the framework in Section 2 anticipates. Alternatively, households may adopt ChatGPT when their information needs change—for example, during a job search or a new project. Such changes could also affect conventional search, limiting a causal interpretation of the adoption estimates.

Figure 7. Conventional searches rise briefly at ChatGPT adoption and then decline.

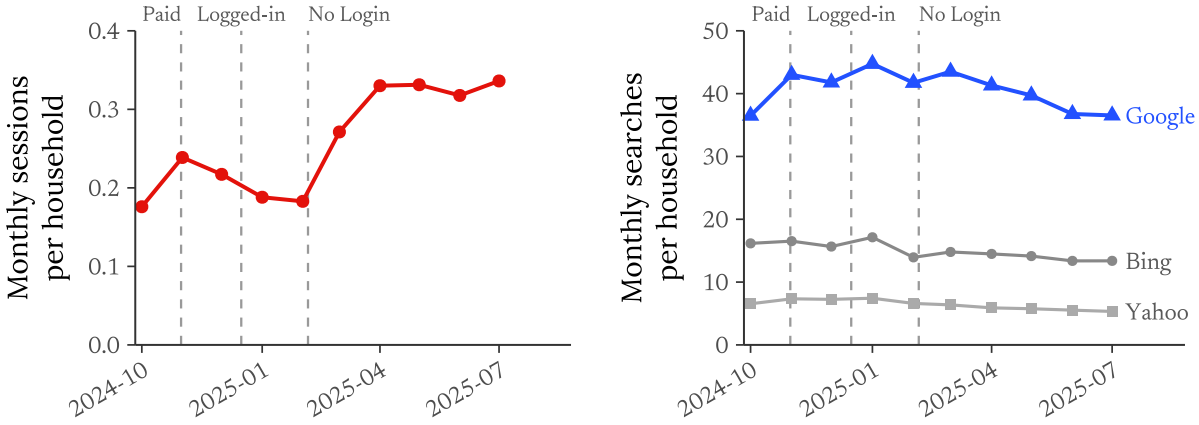


Notes: Points show ACS-weighted Callaway and Sant'Anna (2021) estimates of changes in weekly Google, Bing, and Yahoo searches per household, using not-yet-adopting households as comparisons. Adoption begins in the first of three consecutive weeks with a recorded ChatGPT conversation session. The sample includes 1,008 eligible adopters, of whom 45 serve only as comparisons. The number of contributing adopters falls from 381 at week 20 to 226 at week 30. Week -1 is the reference and is normalized to zero. Error bars show pointwise 95% normal confidence intervals based on standard errors clustered by household.

5.1.2 ChatGPT Search availability and search substitution

To examine the role of ChatGPT Search beyond LLM adoption, we compare how conventional search responds among households adopting under narrower and broader Search availability. We use February 5, 2025 as the cutoff, when ChatGPT Search became available without an account. Monthly recorded ChatGPT sessions per household rise after February, while conventional searches decline from March to July (Figure 8), providing descriptive context for this comparison.

Figure 8. ChatGPT use rises after February, while conventional searches decline after March.



(a) ChatGPT conversation sessions.

(b) Conventional searches.

Notes: Monthly ChatGPT conversation sessions (panel a) and Google, Bing, and Yahoo searches (panel b) per household in the balanced panel of 45,386 Comscore U.S. desktop households, October 2024–July 2025. Each monthly total is divided by all 45,386 households. Dashed lines mark the October 31, December 16, and February 5 ChatGPT Search access expansions.

We divide households into those adopting before February 5 and those adopting on or after that date (Figure 9, panels a and b). For each group, we estimate changes in searches relative to households that have yet to adopt. Panel c estimates the later-minus-earlier difference, the triple difference described above, by generalized method of moments (GMM; Chen et al., 2025), which draws on multiple pre-adoption weeks to improve precision. This comparison holds the stage of adoption constant while distinguishing periods with different levels of Search availability.

In weeks 4–19, later adopters' searches fall by 14.78 per week ($SE = 4.55$), compared with 0.52 among earlier adopters ($SE = 2.87$). The GMM-estimated later-minus-earlier

difference is -13.01 searches per week ($SE = 3.91$, $p = 0.008$; Table 5). Earlier adopters reach a comparable decline of 13.72 searches per week in weeks 20–30 ($SE = 4.64$); both groups' declines occur between March and July 2025, after the February expansion. These patterns provide suggestive evidence that broader Search availability contributes to search substitution beyond LLM adoption.

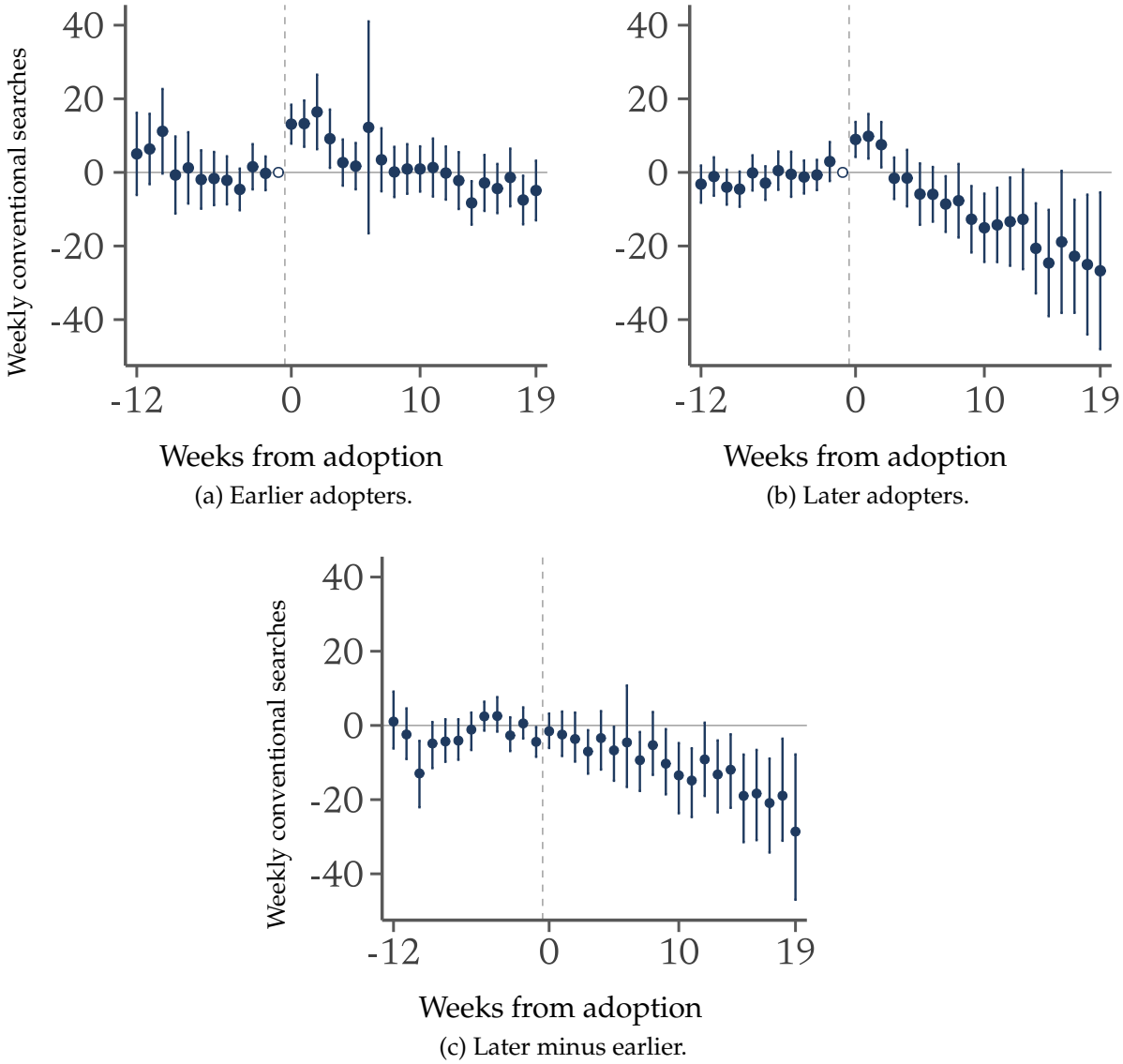
Table 5. Conventional searches rise at adoption, then decline sooner among later adopters.

	All adopters	Earlier adopters	Later adopters	Later minus earlier
Weeks 0–3	8.84^{***}	12.98^{***}	6.17^{**}	-3.68
	(1.94)	(3.08)	(2.42)	(2.31)
Weeks 4–19	-5.73^{**}	-0.52	-14.78^{***}	-13.01^{***}
	(2.41)	(2.87)	(4.55)	(3.91)
Weeks 20–30	-13.80^{***}	-13.72^{***}	–	–
	(4.64)	(4.64)		
Households	1,008	366	642	
Pre-onset mean	35.34	37.20	34.31	
Joint lead-test p	0.295	0.280	0.291	

Notes: The first three columns show ACS-weighted Callaway and Sant'Anna (2021) estimates of changes in weekly Google, Bing, and Yahoo searches per household, relative to the week before adoption, using not-yet-adopting households as comparisons. The final column reports the later-minus-earlier four-baseline GMM estimate. Adoption begins in the first of three consecutive weeks with a recorded ChatGPT conversation session. Earlier and later adopters are distinguished by whether the first session day in the adoption week falls before or on/after February 5, 2025. Each window equally weights its available weekly estimates; contributing cohorts vary across weeks. Later-adopter results are reported through week 19. Household counts and baseline means include all eligible adopters; 45 later adopters serve only as comparisons. The pre-onset mean is the ACS-weighted average of households' mean weekly searches before adoption. Joint lead tests cover weeks -12 to -2. Standard errors are household-clustered in the first three columns and based on 999 household bootstrap refits in the final column. Standard errors appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The comparison remains suggestive because households choose when to adopt, and the two groups face different calendar conditions. Earlier adopters could also access Search through paid or free logged-in accounts, and their follow-up extends beyond February 5. The estimates therefore compare adoption under narrower and broader availability rather than adoption with and without Search access. To separate changes in Search access from households' adoption decisions, Section 5.2 examines the February expansion among existing ChatGPT users.

Figure 9. Conventional searches decline sooner among later adopters.



(a) Earlier adopters.

(b) Later adopters.

(c) Later minus earlier.

Notes: Panels (a) and (b) show ACS-weighted Callaway and Sant'Anna (2021) estimates of changes in weekly Google, Bing, and Yahoo searches per household, using not-yet-adopting households as comparisons. Adoption begins in the first of three consecutive weeks with a recorded ChatGPT conversation session. Groups are defined by the first session day in the adoption week: 366 households adopted before February 5, 2025, and 642 on or after that date; 45 later adopters serve only as comparisons. Panel (c) compares later- and earlier-adopter effects estimated using four-baseline GMM (weeks -4 through -1; Chen et al., 2025). We also report the direct subtraction of panel (a) from panel (b) in Figure OA.6.1. Week -1 is normalized to zero in (a)–(b) and estimated using the other three baselines in (c). Later-adopter estimates include 54 contributing households at week 19. Error bars show pointwise 95% confidence intervals: household-clustered normal intervals in (a)–(b), and basic intervals from 999 household bootstrap refits in (c).

5.2 Evidence from the Search access expansion

The February 5, 2025 expansion changed Search access without requiring a new adoption decision. Households that already used ChatGPT without an account gained Search on a date set by OpenAI (OpenAI, 2024). Comparing them before and after the expansion shows whether Search access, for households already using an LLM, changes their use of ChatGPT and conventional search, and which households respond. Answering this requires a benchmark for how their searches would otherwise have evolved. Prior Gemini users, who also used an AI service before the expansion, provide this benchmark.

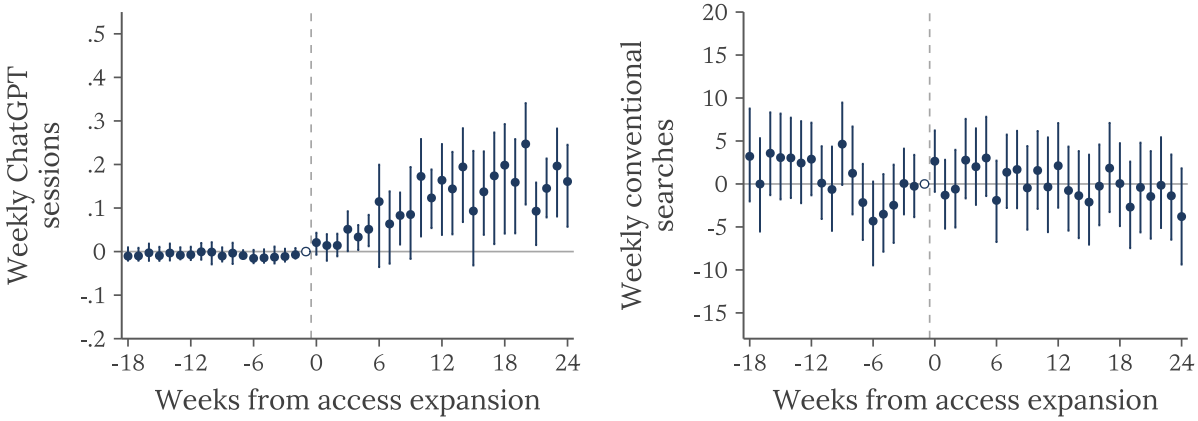
5.2.1 Overall effects on ChatGPT use and conventional search

We compare 945 households that used ChatGPT anonymously before the expansion with 429 prior Gemini users (Section OA.6.2). We estimate changes in recorded conversation sessions and weekly Google, Bing, and Yahoo searches relative to the week before the expansion. Conventional searches are winsorized at the 99th percentile of pre-expansion weekly searches.

Recorded ChatGPT use increases. Newly eligible households record 0.117 additional sessions per household-week relative to the Gemini comparison ($SE = 0.022$; Table 6). We do not detect an overall change in conventional search: the estimate is 0.003 searches per household-week ($SE = 1.88$; Figure 10).

An average response can nevertheless conceal different responses across households (Section 2). Prior search activity captures how heavily households rely on conventional search and may also reflect what they use it for. We therefore examine whether the response varies with pre-expansion search activity, and then whether differences in search use cases help explain the pattern.

Figure 10. ChatGPT use rises without detected declines in conventional search.



(a) ChatGPT conversation sessions.

(b) Conventional searches.

Notes: Points show ACS-weighted TWFE estimates comparing 945 prior-anonymous ChatGPT households with 429 prior-Gemini households, with household and calendar-week fixed effects (Section OA.6.4). Outcomes are conversation sessions and Google, Bing, and Yahoo searches per household-week. Open circles mark week -1, normalized to zero. Bars show pointwise 95% basic intervals from 999 household bootstrap draws. Dashed lines mark the February 5 expansion.

Table 6. ChatGPT use rises without detected declines in conventional search.

	ChatGPT sessions	Conventional searches
Estimate	0.117***	0.00
	(0.022)	(1.88)
Households	1,374	1,374
Joint lead-test p	0.433	0.248

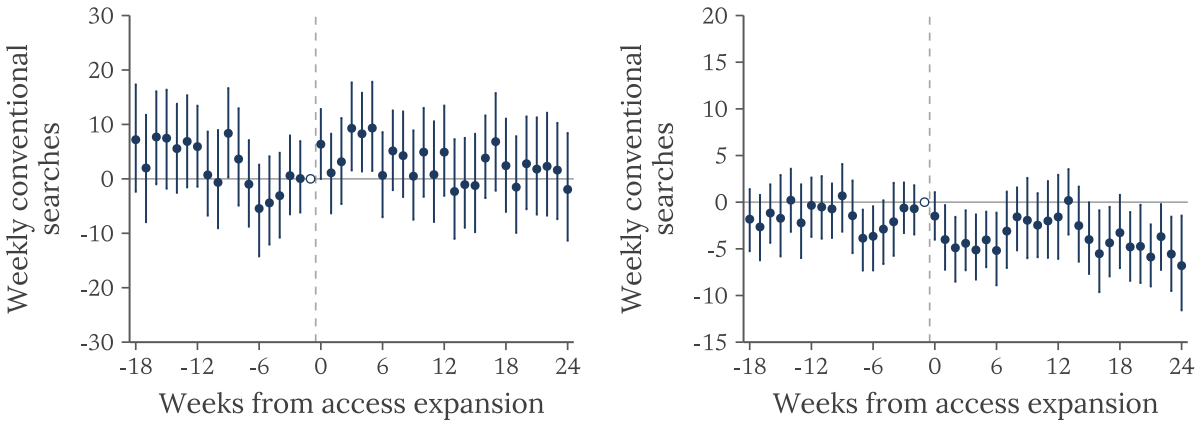
Notes: TWFE estimates compare 945 prior-anonymous ChatGPT households with 429 prior-Gemini households with no qualifying prior ChatGPT visits. GMM estimates for the same sample are reported in Table OA.6.3. Estimates average weeks 0–24 after February 5, 2025, relative to week -1, with household and calendar-week fixed effects and ACS weights. Outcomes are weekly household counts of ChatGPT conversation sessions and conventional searches (Google, Bing, and Yahoo result-page loads); conventional searches are winsorized at the 99th percentile. Joint lead tests cover weeks -18 to -2. See Section OA.6.4 for identification and inference. Standard errors from 999 household bootstrap draws are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5.2.2 Heterogeneity by prior search activity

We divide households by their mean weekly conventional searches over the 18 pre-expansion weeks, splitting at the treated households' weighted median of 18.5 (Section OA.6.2). Recorded ChatGPT sessions increase in both groups: by 0.109 per household-week among households with lower prior activity ($SE = 0.024$) and by 0.127 among those with higher prior activity ($SE = 0.035$; Table 7). Conventional search declines in the lower-activity group by 3.71 searches per household-week ($SE = 1.35$), or 35.8% of its pre-expansion mean, while the higher-activity estimate is positive and

imprecise (2.88, $SE = 3.20$). Figure 11 shows conventional-search responses over time for both prior-activity groups. The difference between the groups, -6.59 searches per household-week ($p = 0.038$), indicates that the overall null combines different responses rather than a uniform absence of response. Because ChatGPT use rises by similar amounts in both groups, the difference reflects how conventional search responds rather than how much ChatGPT use increases. The estimates suggest that expanded Search access led households with lower prior search activity to substitute ChatGPT for some conventional search.

Figure 11. Conventional-search declines concentrate among households with lower prior activity.



(a) Higher prior search activity.

(b) Lower prior search activity.

Notes: Points show ACS-weighted TWFE estimates of effects on weekly Google, Bing, and Yahoo searches, with household and calendar-week fixed effects. Panel (a) compares 453 prior-anonymous ChatGPT households with 227 prior-Gemini households; panel (b) compares 492 with 202. Groups use mean conventional searches over all 18 pre-expansion weeks (Section OA.6.2). Open circles mark week -1, normalized to zero. Bars show pointwise 95% basic intervals from 999 household bootstrap draws. Dashed lines mark the February 5 expansion. Panels use different vertical scales.

5.2.3 The role of search use cases

Section 2 suggests that how much conventional search an answer can replace depends on what the search is for. We examine this interpretation with two pieces of evidence: where households' conventional-search referrals led before the expansion, and how those referrals changed by destination category after it. Before the expansion, lower-activity treated households direct a 4.5-percentage-point larger share of these referrals

Table 7. Conventional search declines among households with lower prior activity.

	Lower prior search activity		Higher prior search activity	
	ChatGPT sessions	Conventional searches	ChatGPT sessions	Conventional searches
Estimate	0.109***	−3.71**	0.127***	2.88
	(0.024)	(1.35)	(0.035)	(3.20)
Households	694	694	680	680
Joint lead-test p	0.408	0.293	0.417	0.405

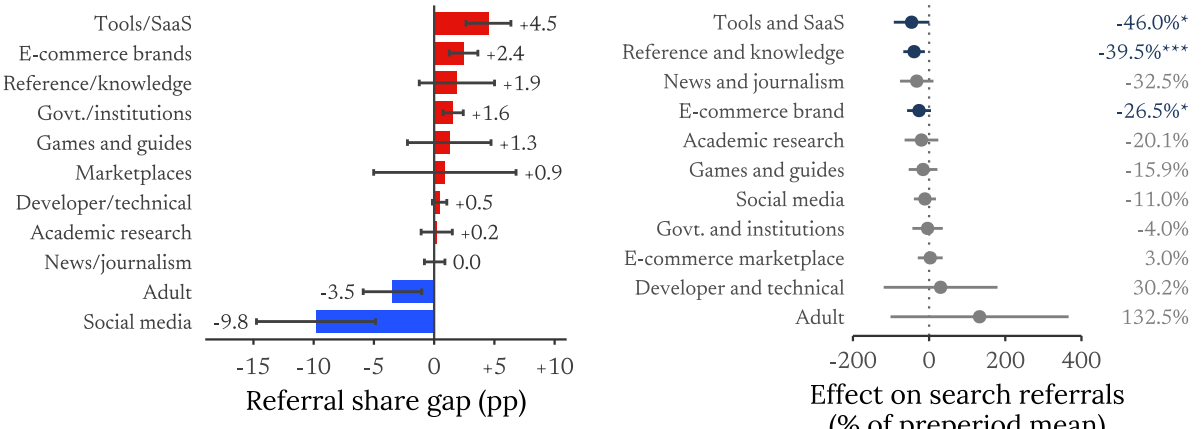
Notes: TWFE estimates compare prior-anonymous ChatGPT and prior-Gemini households within each activity group. GMM estimates for the same groups are reported in Table OA.6.3. Lower activity includes 492 treated and 202 comparison households; higher activity includes 453 and 227. Groups use mean conventional searches over all 18 pre-expansion weeks, with a treated ACS-weighted median cutoff of 18.5 (Section OA.6.2). Estimates average weeks 0–24 after February 5, 2025, relative to week −1, with household and calendar-week fixed effects and ACS weights. Outcomes are weekly household counts of ChatGPT conversation sessions and conventional searches (Google, Bing, and Yahoo result-page loads). Joint lead tests cover weeks −18 to −2. See Section OA.6.4 for identification and inference. Standard errors from 999 household bootstrap draws are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

to tools/SaaS websites and a 9.8-point smaller share to social media than higher-activity households (Holm-adjusted $p < 0.001$ for both), along with a larger share to e-commerce brand websites and a smaller share to adult websites (Figure 12, panel a).

The larger tools/SaaS share suggests that more of the lower-activity group's referrals serve needs a ChatGPT conversation might address, such as mathematical or writing assistance. Recorded referrals to Microsoft Math Solver illustrate the former. ChatGPT Search may similarly answer product questions that otherwise lead users to brand websites, such as Apple's iPhone pages. However, social media searches may be harder to replace when users seek content or interaction on the platform itself, such as watching a YouTube video or participating in a Reddit discussion.

Category-level referral estimates provide complementary evidence (Figure 12, panel b). In the full access sample, conventional-search referrals decline by an estimated 46.0% for tools/SaaS ($p = 0.054$) and 26.5% for e-commerce brand websites ($p = 0.098$). Neither the 11.0% decline for social media nor the 132.5% increase for adult websites is statistically distinguishable from zero ($p = 0.456$ and 0.267, respectively). Reference/knowledge referrals show a larger estimated decline (39.5%), but treated and comparison households already diverged in this category before the expansion (joint lead test $p = 0.009$), so we do not interpret it (Section OA.6.6). These estimates pool both prior-activity groups, so they show which kinds of conventional searches

Figure 12. Conventional-search referral patterns differ by prior activity and destination.



(a) Lower minus higher prior-search households: referral share.

(b) Referral effects by category.

Notes: Both panels use ACS-weighted visits referred by Google, Bing, and Yahoo to eleven classified destination categories, including adult websites. Panel (a) uses the 492 lower- and 453 higher-prior-search treated households over all 18 pre-expansion weeks, without comparison households; bars show lower-minus-higher shares in percentage points, and labels give these differences. Panel (b) reports full-sample TWFE estimates comparing 945 prior-anonymous ChatGPT households with 429 prior-Gemini households, with household and calendar-week fixed effects and week -1 as reference; effects average weeks 0–24 and are expressed as percentages of treated households' category means over the 18 pre-expansion weeks. Intervals in both panels are pointwise 95% normal intervals based on household-clustered standard errors. Dark points in (b) indicate $p < 0.10$; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. See Section OA.6.6 for construction and GMM comparisons.

declined, not which households reduced them. The declines fall in categories where lower-activity households had larger shares before the expansion. Both pieces of evidence are consistent with substitution for some task-oriented and product-related searches.

The two designs address different aspects of the relation between ChatGPT Search and conventional search. The adoption comparison associates broader Search availability with a more negative conventional-search response at the same stage of adoption. Among existing ChatGPT users, the access expansion increases recorded conversations without a detectable overall decline in conventional search, while the negative response is concentrated among households with lower prior search activity. The use-case evidence offers a possible explanation for this heterogeneity.

6 Discussion

AI search changes how users reach websites through two channels, and the evidence here separates them. Observing use alongside referrals shows that ChatGPT sends users onward much less often than Google: a recorded referral accompanies 5.4% of conversation sessions, compared with 31.1% of Google queries, and its referrals are weighted toward reference, technical, and tool websites rather than ad-supported ones. In conventional search, the response is conditional rather than uniform. Conventional searches decline after households adopt ChatGPT, and sooner when Search is broadly available; when Search access expands among existing users, they fall among households with lower prior search activity, with no detectable change overall.

Both channels are consistent with a single boundary: whether an answer can meet the need behind a visit or a search, or only the website can. Tools/SaaS websites illustrate it from both sides. Conventional-search referrals to these websites decline after the access expansion, though imprecisely, as answers may take over tasks such as a calculation; yet tools account for a larger share of the referrals ChatGPT itself makes, plausibly when the tool must be used on its website. Searches that lead to platform content or interaction, such as social media, show no detectable decline. The boundary can therefore run through categories of websites rather than between them; where it falls for a given website depends on what its users need from it.

For research on information intermediaries, these findings suggest that AI search loosens the connection between supplying information and receiving visits on which websites have relied (Dellarocas et al., 2013). Content aggregators raised a similar tension between substitution and discovery in news (Dellarocas et al., 2016; Chiou and Tucker, 2017; Calzada and Gil, 2020). AI search extends it to answers across many tasks and adds a second margin: beyond referring users less often, it can reduce the conventional searches that would otherwise have led them to websites. Its referrals are also less concentrated than Google's in aggregate but similarly or more concentrated within households, consistent with different households following it to different websites rather than to a few common destinations.

For websites, the two channels imply that a share of AI referrals is a poor guide to exposure. A website can rank highly among ChatGPT's destinations yet lose visits overall, because the conventional searches that answers replace send users onward far more often than answers do. Exposure is likely greatest where an answer can deliver what users came for, and smallest where a website's value lies in interaction, participation, or transactions. Assessing the economic consequences therefore requires measuring both channels and how arriving visits contribute to advertising revenue, subscriptions, transactions, or participation. These measures also bear on policy. The European Commission designates ChatGPT a search engine (Section 1), consistent with the intermediary role documented here; our evidence adds that, as used, this search engine sends users to websites far less often than Google does.

The two channels together raise a longer-run question about the returns to producing the information that AI search uses. They motivate examining the relationship between AI intermediaries and information providers alongside research on publisher responses to generative AI (Zhao and Berman, 2026) and online knowledge sharing (Burtch et al., 2024; del Rio-Chanona et al., 2024). Linking retrieval and citation records to referrals, publisher revenues, and content investment could clarify how supplying information to AI systems translates into returns to its producers. The central question for the next generation of information intermediaries is whether they can improve access to web information while sustaining the incentives to produce and maintain it.

References

Agarwal S, Sen A (2026) The impact of Google AI Overviews on publisher traffic and user experience: Evidence from a field experiment. Working paper, Indian School of Business and Carnegie Mellon University, URL http://dx.doi.org/10.2139/ssrn.6513059.

Bakos JY (1997) Reducing buyer search costs: Implications for electronic marketplaces. *Management Sci.* 43(12):1676–1692, URL http://dx.doi.org/10.1287/mnsc.43.12.1676.

Broder A (2002) A taxonomy of web search. *ACM SIGIR Forum* 36(2):3–10, URL http://dx.doi.org/10.1145/792550.792552.

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye Pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Sci.* 57(8):1373–1386, URL http://dx.doi.org/10.1287/mnsc.1110.1371.

Burtch G, Lee D, Chen Z (2024) The consequences of generative AI for online knowledge communities. *Sci. Rep.* 14(1):10413, URL http://dx.doi.org/10.1038/s41598-024-61221-0.

Callaway B, Sant'Anna PHC (2021) Difference-in-differences with multiple time periods. *J. Econometrics* 225(2):200–230, URL http://dx.doi.org/10.1016/j.jeconom.2020.12.001.

Calzada J, Gil R (2020) What do news aggregators do? Evidence from Google News in Spain and Germany. *Marketing Sci.* 39(1):134–167, URL http://dx.doi.org/10.1287/mksc.2019.1150.

Chen X, Sant'Anna PHC, Xie H (2025) Efficient difference-in-differences and event study estimators. URL http://dx.doi.org/10.48550/arXiv.2506.17729, arXiv preprint arXiv:2506.17729.

Chiou L, Tucker C (2017) Content aggregation by platforms: The case of the news media. *J. Econom. Management Strategy* 26(4):782–805, URL http://dx.doi.org/10.1111/jems.12207.

del Rio-Chanona RM, Laurentsyeva N, Wachs J (2024) Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus* 3(9):pgae400, URL http://dx.doi.org/10.1093/pnasnexus/pgae400.

Dellarocas C, Katona Z, Rand W (2013) Media, aggregators, and the link economy: Strategic hyperlink formation in content networks. *Management Sci.* 59(10):2360–2379, URL http://dx.doi.org/10.1287/mnsc.2013.1710.

Dellarocas C, Sutanto J, Calin M, Palme E (2016) Attention allocation in information-rich environments: The case of news aggregators. *Management Sci.* 62(9):2543–2562, URL http://dx.doi.org/10.1287/mnsc.2015.2237.

Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Sci.* 55(5):697–712, URL http://dx.doi.org/10.1287/mnsc.1080.0974.

Gholami S, Firullo C, Cheyre C, Acquisti A (2026) Beyond search: LLM adoption and web traffic concentration. Working paper, Knight-Georgetown Institute, Washington, DC, URL http://dx.doi.org/10.2139/ssrn.6238578.

Kaiser M, Schulze C (2026) Frontiers: ChatGPT referrals to e-commerce websites: How do LLMs compare against traditional channels? *Marketing Sci.* 45(4):699–715, URL http://dx.doi.org/10.1287/mksc.2025.0489.

Khosravi M, Yoganarasimhan H (2026) Impact of AI search summaries on website traffic: Evidence from Google AI Overviews and Wikipedia. Working paper, University of Washington, Seattle, URL http://dx.doi.org/10.2139/ssrn.6164926.

Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, Hesse C, Jain S, Kosaraju V, Saunders W, Jiang X, Cobbe K, Eloundou T, Krueger G, Button K, Knight M, Chess B, Schulman J (2021) WebGPT: Browser-assisted question-answering with human feedback. URL http://dx.doi.org/10.48550/arXiv.2112.09332, arXiv preprint arXiv:2112.09332.

OpenAI (2024) Introducing ChatGPT search. URL https://openai.com/index/introducing-chatgpt-search/, accessed June 21, 2026.

Padilla N, Lam HT, Lambrecht A, Hollenbeck B (2026) The impact of early LLM adoption on online user behavior. *Management Science* URL https://ssrn.com/abstract=5393256, forthcoming.

Wang ST, Gleason J, Bart Y, Wilson C, Metaxa D (2026) AI in search reduces publisher referrals without improving user experience: Experimental evidence. arXiv preprint arXiv:2608.18352, URL http://dx.doi.org/10.48550/arXiv.2608.18352.

Zhao H, Berman R (2026) Strategic response of news publishers to generative AI. Working paper, University of Pennsylvania, Philadelphia, URL http://dx.doi.org/10.2139/ssrn.5992774.

Online Appendix

AI Search as an Information Intermediary: Clickstream Evidence on Web Traffic

This appendix provides the construction, validation, and supplementary estimates for intermediary use, referrals, destination characteristics, and conventional search. The guide below links these materials to the main paper; the classification prompt appears last.

Main-paper question	Supporting material
Data used throughout	
What is observed and reconstructed?	Section OA.1: panel, request paths, sessions
Referrals from ChatGPT (main-paper Section 4)	
How often does use produce a referral?	Section OA.2: attribution and incidence; Section OA.3: surrounding browsing
Which websites receive referrals?	Section OA.4: labels and validation; Section OA.5: distribution and characteristics
Conventional search (main-paper Section 5)	
How do conventional searches and search referrals change?	Section OA.6: adoption and access designs, estimation, sensitivity, and referral categories
How were domains classified?	Section OA.7: complete classification prompt

Contents

OA.1 Data construction and observable activity . . . 3
OA.1.1 Balanced panel and activity threshold . . . 3
OA.1.2 Foreground visits and background requests . . . 3
OA.1.3 Endpoint definitions and coverage . . . 4
OA.1.4 Conversation sessions and gap sensitivity . . . 5
OA.1.5 Growth in reach and activity per household . . . 5
OA.2 Referral construction and incidence comparisons . . . 6
OA.2.1 Referral exclusions, attribution, and diagnostics . . . 6
OA.2.2 Within-household and context-adjusted comparisons . . . 8
OA.2.3 Interpreting external benchmarks . . . 8
OA.3 Surrounding browsing contexts . . . 9
OA.3.1 Sessions, windows, and category assignment . . . 9
OA.3.2 Longer browsing windows . . . 10
OA.3.3 Excluding nearby use of other intermediaries . . . 10
OA.3.4 Google session boundaries . . . 12
OA.3.5 Robustness of category-dominance thresholds . . . 12
OA.4 Destination classification and composition . . . 15
OA.4.1 Domain universe and codebook . . . 15
OA.4.2 Validation against blind human coding . . . 15
OA.4.3 Confidence scores and sensitivity . . . 17
OA.4.4 Robustness under a lower confidence threshold . . . 17
OA.5 Destination concentration and supplementary characteristics . . . 19
OA.5.1 Normalized concentration . . . 20
OA.5.2 Restricting the set of destinations . . . 20
OA.5.3 Concentration within household-weeks . . . 21
OA.5.4 Destination prominence . . . 21
OA.6 Sustained adoption, access expansions, and conventional search . . . 22
OA.6.1 Sustained-adoption design . . . 23
OA.6.2 February cohorts and classification . . . 25
OA.6.3 Demographic weighting . . . 26
OA.6.4 Estimation and inference . . . 26
OA.6.5 Diagnostics for the parallel-trends assumption . . . 27
OA.6.6 Robustness of model specification . . . 27
OA.7 Domain-classification prompt . . . 31

OA.1 Data construction and observable activity

The measures in main-paper Section 3 distinguish requests recording user activity from background requests. Descriptive analyses use all observed active households; analyses of conventional search use the balanced panel defined below.

OA.1.1 Balanced panel and activity threshold

The balanced panel contains 45,386 households, identified by Comscore's machine identifier, with at least four foreground webpage visits (Section OA.1.2) in each month from October 2024 through July 2025. This threshold requires repeated evidence of observation while retaining lightly active households. Lowering it to one adds 4,822 households; raising it to fifty removes 15,009 (Table OA.1.1).

Table OA.1.1. The monthly activity threshold changes panel coverage.

Minimum monthly visits	Households	Added	Removed	Mean visits per month	CV
1	50,208	+4,822	0	577.1	1.084
2	47,967	+2,581	0	562.8	1.061
3	46,477	+1,091	0	606.3	1.042
4 (main)	45,386	0	0	—	—
5	44,403	0	−983	512.3	1.036
10	41,105	0	−4,281	571.9	1.010
20	37,057	0	−8,329	601.4	0.973
50	30,377	0	−15,009	652.0	0.922

Notes: The starting population comprises 50,208 households with at least one foreground visit in every sample month. Each row requires the stated minimum number of foreground visits per month. Added and removed households are measured relative to the threshold of four used throughout. Mean monthly activity and its cross-month coefficient of variation (CV) describe the households added or removed, not the entire panel.

OA.1.2 Foreground visits and background requests

A foreground visit is a record with content type `text/html` and an HTTP response code in $\{200, 201, 202, 203, 204, 206, 301, 302, 303, 304, 307, 308\}$. The content-type restriction excludes scripts, images, fonts, audio and video, JSON responses, event streams, document downloads, and malformed content-type values. The response-code restriction excludes errors and nonstandard codes. These filters give a consistent measure of observed webpage visits.

Some background ChatGPT requests are nonetheless informative: they record conversation loads, message sends, and account status, which foreground pages alone cannot reveal. We use them in the corresponding activity measures but do not count them as webpage visits. The referral measure applies additional destination exclusions in Section OA.2.1.

OA.1.3 Endpoint definitions and coverage

Comscore records request hosts and paths, but not prompts, response bodies, or authorization headers. We identify conversation loads and message sends from the application's request paths (Table OA.1.2). Public descriptions of the web application inform these interpretations.[1] The `backend-api` and `backend-anon` prefixes distinguish logged-in and anonymous routes. Because anonymous bootstrap requests can also occur for logged-in users, access-cohort classification uses the selective rules in Section OA.6.2.

Table OA.1.2. Endpoint paths distinguish several observable forms of ChatGPT activity.

Endpoint	Definition and role in the paper
`chatgpt.com` foreground page load	GET. A user-visible page load (homepage or `/c/{uuid}`); the broadest but shallowest signal. *Visit denominator.*
`backend-api/conversation/{uuid}`	GET. The browser opening or reloading one conversation; each distinct UUID is one conversation. *Conversation denominator.*
`backend-anon/conversation/{uuid}`	GET. The same for anonymous sessions; negligible volume.
`backend-api/f/conversation`[2]	POST. One logged-in message send (a newer variant whose volume grows from March to July 2025). *Message-level referral attribution.*
`backend-anon/f/conversation`	POST. One anonymous message send.

Notes: The table gives our interpretation of each request path. Conversation loads include returns and reloads; they are grouped into sessions using the rule in Section OA.1.4. Conversation loads do not establish whether an answer retrieved information from the web.

Table OA.1.3 reports coverage in the balanced panel. The rows represent overlapping sets of households and should not be added. Foreground visits have wider coverage than conversation endpoints; message-send coverage is concentrated near the end of the observation window.

Message coverage. The `f/conversation` message-send endpoint has substantial coverage only in July 2025. The full descriptive sample then contains 6,056 logged-in message-send users and 184,030 sends. These counts differ from Table OA.1.3, which restricts to the balanced panel and includes preparatory requests. Message-level referral incidence is therefore reported separately for July in Table OA.2.2.

[1] See the community-maintained *everything-chatgpt* catalogue, `https://github.com/terminalcommandnewsletter/everything-chatgpt`, and the archived anonymous client, `https://github.com/Mr-Destructive/anonymous-chatgpt` (accessed 2026-06-24).

[2] Unlike the other endpoints here, the `f/conversation` variant is not separately documented in public reverse-engineering sources. When we inspected the web application's network requests, sending one message issued exactly one `f/conversation` POST. As an undocumented internal endpoint it may change without notice.

Table OA.1.3. Endpoint coverage differs across activity measures.

Endpoint	Households	% of panel	% of visitors
`chatgpt.com` foreground visits	9,481	20.9	100.0
`backend-api/conversation/{uuid}`	5,265	11.6	55.5
`backend-anon/conversation/{uuid}`	348	0.8	3.7
`backend-api/f/conversation` (Jul 2025)	2,361	5.2	65.8 (Jul)
`backend-anon/f/conversation` (Jul 2025)	936	2.1	26.1 (Jul)

Notes: The population is the 45,386-household balanced panel. The first three rows count households observed at any point in the ten months; the message rows report July 2025 and include both sends and preparatory requests. The panel percentage divides by the 45,386 panel households. The visitor percentage divides by the households with a foreground ChatGPT visit in the same period: 9,481 over ten months, and the July count for the July rows.

OA.1.4 Conversation sessions and gap sensitivity

A ChatGPT conversation session groups loads of the same conversation UUID on the same computer, combining the logged-in and anonymous conversation endpoints. A gap longer than 3,600 seconds between loads of that UUID starts a new session. Different UUIDs remain separate. This rule consolidates repeated loads while allowing returns to the same conversation to form new sessions.

The full sample contains 263,008 conversations within household-months. This count is invariant to the gap rule, whereas the number of sessions varies (Table OA.1.4). Consecutive same-UUID loads with no intervening UUID have a median gap of thirteen seconds, whereas repeated loads with another UUID intervening have a median gap of 4,946 seconds (about 82 minutes). This separation motivates the one-hour rule, although timestamps alone cannot establish whether an individual request was an automatic reload or a deliberate return. Main-paper Figure 3 shows that ChatGPT's monthly referral incidence changes little between thirty-minute and five-hour boundaries.

Table OA.1.4. Conversation counts vary moderately around the one-hour session rule.

Same-UUID gap	Sessions	vs. 1 h
0.25 h (900 s)	449,391	+9.8%
0.5 h (1,800 s)	429,074	+4.9%
1 h (3,600 s, main)	409,133	—
2 h (7,200 s)	391,282	−4.4%
Conversations within household-months	263,008	invariant

Notes: The sample includes all observed conversation loads in the full October 2024–July 2025 descriptive sample. A new session starts when the same-household, same-UUID gap exceeds the stated threshold. The final column reports the percentage difference from the one-hour count. The final row counts distinct conversations within household-months and is invariant to the threshold.

OA.1.5 Growth in reach and activity per household

Conversation-session volume can grow through reach, activity per reached household, or both. This decomposition uses the full descriptive sample. Let M_t be the number of

active households in month t, N_t the number with a conversation session, $a_t = N_t / M_t$ their reach, and $\bar{q}_t$ their mean session count. Sessions per active household are $v_t = a_t \bar{q}_t$. Between baseline b and endpoint e,

$$\Delta v = \underbrace{(a_e - a_b)\bar{q}_b}_{\text{reach}} + \underbrace{a_b(\bar{q}_e - \bar{q}_b)}_{\text{activity per reached household}} + \underbrace{(a_e - a_b)(\bar{q}_e - \bar{q}_b)}_{\text{interaction}}.$$

Normalizing by M_t removes changes in total volume arising solely from the size of the active panel. The interaction is the product of changes in reach and average activity; it does not separately measure how newly arriving households behave.

Table OA.1.5. Expanding reach accounts for most measured activity growth.

	Conversation sessions
Start / end month	Oct 2024 / Jul 2025
Growth in volume per household	3.0-fold
Reach share	65.5%
Activity-per-household share	14.9%
Interaction share	19.6%

Notes: The decomposition uses the full descriptive sample of conversation sessions from October 2024 to July 2025. Each share divides the corresponding term by the total change in conversation-session volume per active household.

Reach accounts for 65.5% of conversation-session growth, activity per reached household for 14.9%, and their interaction for 19.6%. These contributions describe growth in the full descriptive sample. Main-paper Figure 8 instead holds the household sample fixed in the balanced panel.

OA.2 Referral construction and incidence comparisons

The referral-incidence comparison in main-paper Section 4.1 links occasions of intermediary use to external visits. The construction below defines eligible destinations and attribution windows, followed by comparisons within households and browsing contexts.

OA.2.1 Referral exclusions, attribution, and diagnostics

A referral is a foreground visit to an external website whose HTTP referrer identifies the source intermediary. Clean referrals pass the destination exclusions below. Incidence is the share of ChatGPT conversation sessions or Google queries associated with at least one clean referral; an occasion with several referrals counts once. These service-specific denominators measure referral incidence during observed use, which can involve different tasks.

Destination exclusions. We exclude 147 destination domains covering authentication, platform assets, infrastructure, selected AI utilities, and advertising or tracking services. Examples include `microsoftonline.com`, `oaiusercontent.com`, `stripecdn.com`,

quillbot.com, and webpkgcache.com. Search-engine domains and their subdomains are also excluded, including services such as drive.google.com and docs.google.com. The resulting measure omits some content and tools hosted within search-engine domains. We also remove same-source self-referrals, Gemini self-referrals, and Google OAuth paths.

Attribution windows. The HTTP referrer first identifies the source platform. Each clean ChatGPT referral is then assigned to the preceding conversation load on the same computer, in the interval from that load to the next observed conversation load. The load belongs to a conversation session under Section OA.1.4. This timestamp rule does not require the referrer to contain a conversation UUID. The attribution window, which runs between loads, differs from the one-hour rule that groups loads into sessions.

Search queries are identified from qualifying foreground HTML records whose recorded url_pattern ends literally in /search. A clean referral is assigned to a query if it occurs between that query's visit and the next search visit. Missing source referrers prevent attribution even when an external visit follows platform use. Neither rule observes which displayed link, if any, was clicked.

Diagnostics. After filtering, 99.7% of clean ChatGPT referrals have chatgpt.com as referrer. Before filtering, 84.7% of ChatGPT-referred records have JSON content type and are excluded from the foreground measure. Among classifiable landing pages, 75.6% of ChatGPT referrals and 81.4% of Google referrals reach a non-homepage path, with median depth one on both platforms (Table OA.2.1).

Table OA.2.1. Clean referrals come from the ChatGPT interface and mostly reach pages below the homepage.

Panel A. Referrer host by filter stage (full sample, ten months)

Filter stage	chatgpt.com	Other OpenAI (such as auth.openai.com)
Before filtering (any content type)	69.1%	30.9%
Clean referrals	99.7%	0.3%

Panel B. Landing-page depth

Intermediary	Homepage	Below homepage	Median depth
ChatGPT	24.4%	75.6%	1.0
Google	18.6%	81.4%	1.0

Notes: Panel A reports referrer-host shares before filtering and after the clean-referral restrictions, using the full ten-month descriptive sample. Other OpenAI combines authentication, marketing, and remaining OpenAI subdomains. Panel B covers 5,318 classifiable ChatGPT referrals and 874,266 Google referrals. Path depth counts nonempty segments below the host: zero is a homepage and one or more is a deeper page. These are request-level diagnostics, not direct verification of a human click.

Records are assigned to calendar months using the day lookup supplied by Comscore.

Message-level check. In July 2025, 1.3% of logged-in message sends and 1.3% of anonymous sends are followed by a recorded referral. Restricting the logged-in referrals

to five minutes after the send gives 1.0% (Table OA.2.2). These figures use a finer unit and a shorter observation period than the headline session measure, so their levels are not directly interchangeable.

Table OA.2.2. Few observed July message sends are followed by a referral.

Message sends (Jul 2025)	Sends	With referral	Share with referral
Logged-in (`backend-api`)	184,030	2,468	1.3%
within 5 min		1,785	1.0%
Anonymous (`backend-anon`)	31,320	411	1.3%

Notes: The unit is a July 2025 message-send request to the logged-in or anonymous `f/conversation` endpoint. Requests marked as preparation rather than sends are excluded. The five-minute row restricts the logged-in numerator to referrals within that interval. July is the only month with substantial message-endpoint coverage; these estimates are supplementary to the ten-month conversation-session measure.

OA.2.2 Within-household and context-adjusted comparisons

Main-paper Table 2 reports the within-household referral-incidence comparisons. Panel A computes incidence for each household, week, and service; Panel B computes it separately by surrounding context. Both weight cells equally and cluster standard errors by household.

For Panel B, ChatGPT conversations receive context labels directly; Google queries take the label of their search session but keep query-level referral attribution. The sample includes sessions with nearby use of other search or AI services, but excludes search visits outside a search session and cells labeled solo, mixed, or other. Splitting by context raises the number of cells without adding households. Raw mean incidence in this sample is 0.088 for ChatGPT and 0.462 for Google. The different sample and aggregation explain why its baseline coefficient differs from Panel A's. Nearby browsing reveals neither the prompt nor the purpose of a particular query, so context adjustment does not make the services' tasks identical.

OA.2.3 Interpreting external benchmarks

Published traffic metrics differ in their denominators, destination coverage, and observation windows (Table OA.2.3). In particular, outgoing visits per person combine the frequency of platform use with the referrals it generates, whereas referral incidence measures the share of occasions producing a referral.

Table OA.2.3. Published traffic measures use different activity units and outcomes.

Source	Metric	Value
Google		
Ours (Oct 2024–Jul 2025)	Share of queries with a clean referral	31.1%
SparkToro / Datos[3] (Jul 2024)	Open-web clicks per query (US)	0.36
Pew Research Center[4] (Jul 2025)	Result pages without AI summary; any traditional-link click	15%
ChatGPT		
Ours (Oct 2024–Jul 2025)	Share of conversation sessions with a clean referral	5.4%
Ours (Jul 2025)	Share of message sends followed by a referral	1.3%
Momentic / Similarweb[5]	Outgoing visits per person, March 2025	1.4

Notes: Our estimates use the full Comscore U.S. desktop sample from October 2024 through July 2025, except the July-only message check. External rows retain each source's population, period, and metric. SparkToro/Datos counts open-web clicks per query; Pew reports the share of result-page visits with a traditional-link click when no AI summary appears. Momentic reports outgoing visits per person, not the share of conversation sessions with a referral. These quantities cannot be interpreted as estimates of a common referral probability.

OA.3 Surrounding browsing contexts

Surrounding browsing provides observable context for the service comparison in main-paper Section 4.1. We examine sensitivity to the browsing window, nearby use of other services, Google's session boundary, and the dominant-category threshold. Categories use the domain labels in Section OA.4; they describe nearby activity rather than directly observed conversation tasks.

OA.3.1 Sessions, windows, and category assignment

Sessions. Each context is attached to a ChatGPT conversation session or a Google search session. ChatGPT uses the same-household, same-UUID one-hour rule in Section OA.1.4. Google has no conversation UUID: queries on the same computer are grouped when successive gaps do not exceed one hour, with a thirty-minute

[3]Rand Fishkin (SparkToro, on Datos clickstream data), *2024 Zero-Click Search Study*, 1 July 2024, https://sparktoro.com/blog/2024-zero-click-search-study-for-every-1000-us-google-searches-only-374-clicks-go-to-the-open-web-in-the-eu-its-360/ (accessed 2026-06-24).

[4]A. Chapekis and A. Lieb, *Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results*, Pew Research Center, 22 July 2025, https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/ (accessed 2026-06-24).

[5]Tyler Einberger (Momentic), *Google vs. ChatGPT*, May 2025, https://momenticmarketing.com/blog/google-reach-chatgpt-click-throughs, accompanying report pp. 12–13 and 22 (accessed 2026-09-21). The report defines the denominator as people (de-duplicated audience), although the webpage also uses per-visit wording.

alternative below. Using the same one-hour gap does not make these units identical. Google sessions are used for this context analysis; the headline Google referral measure continues to use individual queries. Because the data arrive as monthly files, session identifiers include the month, so sessions from different months remain distinct and keep their own surrounding visits.

Nearby visits. Context consists of the foreground pages loaded in the fifteen minutes before a session starts and the fifteen minutes after it ends. The sensitivity check extends each window to thirty minutes. These pages describe surrounding activity rather than a verified sequence of steps in the same task. A session counts as referring if it has at least one clean referral: ChatGPT uses the conversation-load attribution rule, while Google's context analysis uses the interval from a session start to the next session start.

Nearby use of other services. ChatGPT sessions frequently occur alongside conventional-search activity: 79% have a surrounding destination or referrer belonging to Google, Bing, or Yahoo within thirty minutes before or after the session. By comparison, only 3% of Google sessions have surrounding activity involving a covered LLM platform. Conventional search thus commonly accompanies ChatGPT use, whereas LLM activity accompanies only a small share of Google sessions. Our main comparison keeps all sessions. The restricted comparison below excludes sessions with this nearby activity, retaining 85,316 of 409,133 ChatGPT sessions.

Dominant category. Category counts use the 3,245 classified domains with confidence at least 0.90. A surrounding visit is classifiable only if its destination belongs to this classified set and neither its destination nor its referrer is the session's own platform, a search engine, or a covered LLM platform. A category is dominant when it accounts for at least half of the classifiable visits (threshold $\tau = 0.50$). A session with classifiable visits but no dominant category is labeled *mixed context*; one without them is *solo*. Context-adjusted regressions further restrict to the classifiable categories described in Section OA.2.2.

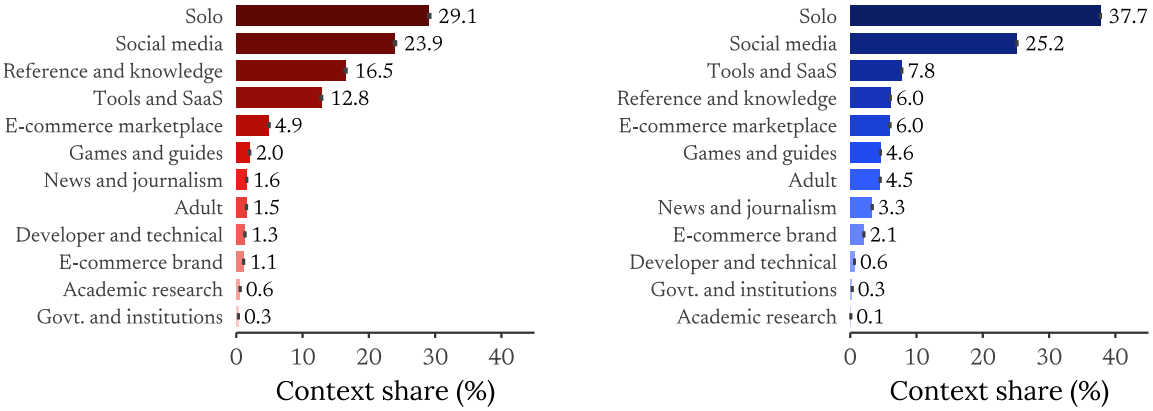
OA.3.2 Longer browsing windows

Extending each window from fifteen to thirty minutes adds more distant surrounding activity. The differences in context composition and ChatGPT referral incidence persist (Figure OA.3.1 and Figure OA.3.2).

OA.3.3 Excluding nearby use of other intermediaries

The restricted sample keeps only sessions without nearby activity involving another intermediary; the main sample keeps all sessions. Because the restriction excludes a much larger share of ChatGPT sessions than of Google sessions (Section OA.3.1), it changes ChatGPT's context composition far more (Figure OA.3.3). In the displayed shares, which exclude mixed and other contexts, ChatGPT's solo share rises from 41.2% to 59.7%, and its larger shares of reference/knowledge and tools/SaaS contexts narrow: in the restricted sample, reference/knowledge accounts for 9.7% of ChatGPT sessions and 5.4% of Google sessions, and tools/SaaS for 7.7% and 7.0%. These shares fall mainly

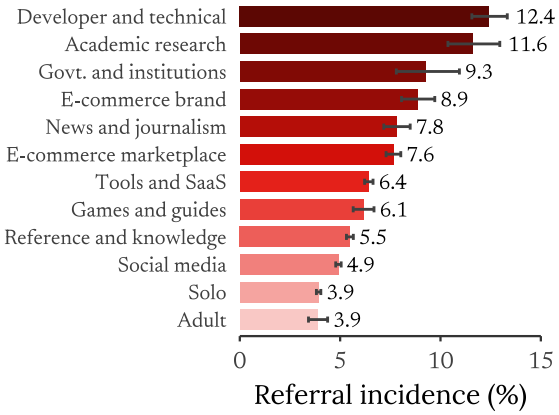
Figure OA.3.1. The longer window retains differences in surrounding browsing composition.



(a) ChatGPT. **(b)** Google.

Notes: The panels show dominant content categories using thirty-minute windows before and after each session, keeping all sessions. Both platforms use a one-hour gap, with a same-UUID requirement for ChatGPT. Shares use all sessions as the denominator; mixed and other categories are omitted from the display. Categories describe classifiable surrounding visits, not prompts or query intent. The sample is all observed active desktop households over the ten-month window.

Figure OA.3.2. ChatGPT referral incidence continues to vary across thirty-minute contexts.

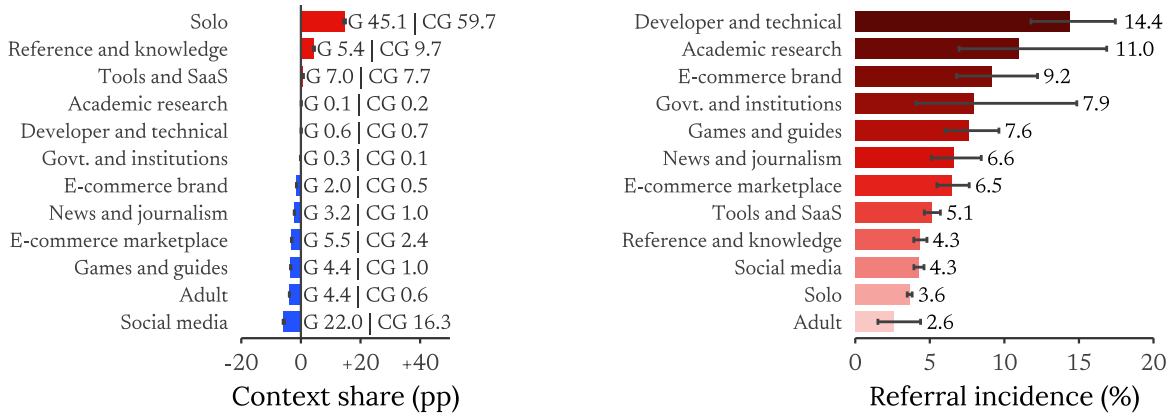


Notes: Bars report the share of ChatGPT conversation sessions with at least one clean referral, conditional on the dominant surrounding category in thirty-minute windows. All sessions are kept, including those near use of other intermediaries. Error bars are 95% confidence intervals. Solo sessions have no classifiable surrounding visits.

because more ChatGPT sessions are solo; the browsing around the remaining sessions changes little. Excluding solo sessions, reference/knowledge accounts for about a quarter and tools/SaaS for about a fifth of ChatGPT sessions in both samples, compared with about one in ten for reference/knowledge and one in eight for tools/SaaS among Google sessions.

Panel (b) of Figure OA.3.3, compared with main-paper Figure 4, shows ChatGPT referral incidence within categories. Developer/technical contexts have relatively high incidence in both samples (13.6% with all sessions and 14.4% in the restricted sample), while adult contexts have relatively low incidence (3.9% and 2.6%). The contrast persists in the much smaller restricted sample.

Figure OA.3.3. Excluding sessions near other intermediaries changes context shares more than referral-incidence patterns.



(a) ChatGPT minus Google context shares.

(b) ChatGPT referral incidence by context.

Notes: Both panels exclude sessions with other-intermediary activity within thirty minutes before or after the session. Displayed contexts use fifteen-minute windows and one-hour session boundaries. Panel (a) reports ChatGPT minus Google category shares after excluding mixed and other contexts; panel (b) reports ChatGPT conversation-session referral incidence within each displayed category. Error bars are 95% confidence intervals. Main-paper Figure 4 shows the corresponding estimates keeping all sessions.

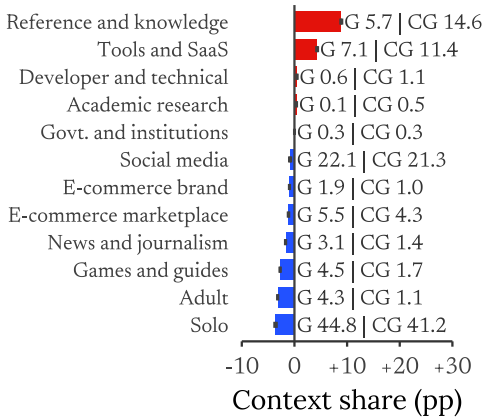
OA.3.4 Google session boundaries

Shortening Google's gap from one hour to thirty minutes splits some multi-query sessions and changes their surrounding visits. The composition comparison remains similar (Figure OA.3.4). ChatGPT retains the same session boundary and surrounding visits.

OA.3.5 Robustness of category-dominance thresholds

Table OA.3.1 varies the dominant-category share requirement from 0.30 to 0.60 using thirty-minute windows. Raising it labels more sessions mixed. Among named content categories, ChatGPT incidence changes by less than half a percentage point; Google's session-level incidence changes by less than one point. Solo is unchanged because sessions without classifiable surrounding visits never enter the dominance calculation.

Figure OA.3.4. The context-composition comparison is similar with a thirty-minute Google session boundary.



Notes: Bars show ChatGPT minus Google category-share differences with fifteen-minute browsing windows, keeping sessions with nearby use of other intermediaries and excluding mixed and other contexts from the shares. Google's maximum inter-query gap is thirty minutes; ChatGPT sessions retain the one-hour, same-UUID rule. Error bars are 95% confidence intervals. Main-paper Figure 4, panel (a), reports the one-hour Google boundary.

Mixed-context rates vary more because the threshold directly changes which sessions are labeled mixed.

Table OA.3.1. Referral incidence within named categories is stable across dominance thresholds.

	Referral incidence (%), by dominance threshold τ			
Surrounding context	0.3	0.4	0.5	0.6
Panel A: ChatGPT sessions				
Developer/technical	12.5	12.5	12.4	12.1
Academic research	11.9	11.8	11.6	11.8
Government/institutional	9.2	9.3	9.3	9.3
E-commerce brand	9.0	9.0	8.9	8.7
E-commerce marketplace	7.7	7.7	7.6	7.6
News/journalism	7.9	7.9	7.8	7.6
Tools/SaaS	6.5	6.5	6.4	6.2
Games and guides	6.2	6.2	6.1	6.2
Other	6.1	6.0	6.0	5.9
Reference/knowledge	5.5	5.5	5.5	5.4
Social media	5.0	5.0	4.9	4.7
Adult	4.0	3.9	3.9	3.9
Solo	3.9	3.9	3.9	3.9
Mixed	10.4	7.8	8.3	7.4
Panel B: Google sessions				
Developer/technical	78.4	78.4	78.4	78.3
Academic research	63.2	63.1	63.0	62.7
Government/institutional	72.4	72.3	72.3	72.0
E-commerce brand	75.8	75.8	75.8	75.7
E-commerce marketplace	69.9	69.9	69.8	69.5
News/journalism	64.9	64.9	64.8	64.5
Tools/SaaS	68.7	68.7	68.6	68.4
Games and guides	78.9	78.9	78.9	78.9
Other	70.1	70.0	69.9	69.6
Reference/knowledge	67.9	67.9	67.8	67.4
Social media	66.5	66.4	66.3	65.8
Adult	83.3	83.3	83.3	83.4
Solo	55.3	55.3	55.3	55.3
Mixed	72.5	73.6	74.6	72.8

Notes: Cells are percentages of platform sessions with at least one clean referral, using all sessions, one-hour session boundaries, and thirty-minute windows. The columns vary the minimum share of classifiable surrounding visits needed for a dominant category. Solo sessions have no classifiable surrounding visits; mixed context has visits but no category meeting the threshold. Google denominators are sessions here, rather than the individual queries used in the headline benchmark.

OA.4 Destination classification and composition

The composition analysis in main-paper Section 4.2 uses a common classification of 4,266 domains: the union of the top 2,500 LLM-referred and top 2,500 search-referred domains. Each source's own list covers about 71% of its referral volume. Section OA.7 reproduces the classification prompt.

OA.4.1 Domain universe and codebook

Each domain receives a content-type label and a monetization label. Content categories distinguish reference, news, academic research, social and streaming media, brand commerce, marketplaces, tools/SaaS, developer resources, government/institutional, games and guides (entertainment and gaming sites), adult content, and other. Monetization categories distinguish advertising, subscriptions, freemium services, transactions, nonprofit/public support, and mixed/other models. Labels describe the domain's primary activity and visible business model.

Several codebook choices affect interpretation. Education services and health-reference websites belong to reference/knowledge; game-specific wikis belong to games and guides; social media includes professional streaming libraries. General-purpose AI assistants and AI-only generators are assigned to other. Other also contains sites the classifier cannot place, so it combines substantive assignments with unresolved cases and is distinct from low confidence. A separate unknown category records domains whose labels were missing or invalid; its count can differ across the two dimensions.

The classification uses GPT-4o through the OpenAI EU Responses API, with temperature zero and batches of five domains. The prompt permits web search when the model cannot confidently identify a site's function. Each domain receives two labels, a rationale, and one confidence score. Main content-type analyses retain the 3,245 domains with confidence at least 0.90.

Before analysis, deterministic rules assign `.gov` domains to reference/knowledge and `.edu` domains to academic research, with nonprofit/public monetization for both. These rules apply to destination composition, surrounding contexts, and the access-design category outcomes. The classification counts and human-validation results below describe the original model labels, before this recoding. Content-composition shares exclude other and unknown labels; monetization shares exclude unknown labels and retain mixed/other.

Increasing the rank cutoff from $K = 2{,}500$ to $K = 10{,}000$ expands the union from 4,266 to 17,545 domains (Table OA.4.2). The additional domains increase referral coverage, while the composition analysis uses the classified $K = 2{,}500$ universe.

OA.4.2 Validation against blind human coding

One author independently coded 100 domains using the same codebook, without seeing the model's labels. Domains were sampled at random within three referral-rank strata: forty from ranks 1–200, thirty from 201–1,000, and thirty from 1,001 onward. The sample spans ranks 10–4,082 and about 7.5% of referral volume. Stratification checks labels beyond the largest websites; design weights account for the unequal sampling fractions.

Table OA.4.1. The domain universe spans multiple content and business-model categories.

Content type			**Monetization**		
Category	*N*	%	Category	*N*	%
Reference/knowledge	803	18.8	Ad-supported	1,708	40.0
Tools/SaaS	686	16.1	Transactional	890	20.9
Other	490	11.5	Non-profit/public	680	15.9
E-commerce brand	425	10.0	Freemium SaaS	578	13.5
Games and guides	348	8.2	Subscription/paywall	248	5.8
Government/institutional	345	8.1	Mixed/other	137	3.2
News/journalism	327	7.7	Unknown	25	0.6
E-commerce marketplace	259	6.1			
Adult	250	5.9			
Social media	166	3.9			
Academic research	82	1.9			
Developer/technical	80	1.9			
Unknown	5	0.1			

Notes: Counts and percentages describe all 4,266 domains in the union of the two top-2,500 lists. The two dimensions are classified separately. Other includes explicitly grouped AI services as well as unclassified functions; unknown marks missing or invalid labels. These are domain shares, not referral shares.

Table OA.4.2. Broader rank cutoffs add coverage at the cost of a much larger domain universe.

	Domains			Referral coverage		
K	LLM list only	Both lists	Union	LLM	Search	LLM, union
500	344	156	844	0.515	0.578	0.530
1,000	697	303	1,697	0.593	0.635	0.608
2,000	1,399	601	3,399	0.680	0.690	0.696
2,500 (main)	1,766	734	4,266	0.711	0.708	0.726
3,000	2,154	846	5,154	0.735	0.722	0.751
5,000	3,658	1,342	8,658	0.821	0.760	0.834
10,000	7,545	2,455	17,545	0.941	0.807	0.948

Notes: K is applied separately to the LLM- and search-referral rankings in the full ten-month descriptive sample. The middle columns count domains in only the LLM list, in both lists, and in their union. Referral coverage is the share of each source's referrals that reach its own top-K list; the last column is the share of LLM referrals that reach the union. The classification universe uses $K = 2{,}500$.

Exact model–human agreement is 79% for content type and 87% for monetization; Cohen's chance-corrected κ is 0.76 and 0.82, respectively (Table OA.4.3). Intervals use 5,000 bootstrap resamples. Thirty domains disagree on at least one dimension: seventeen on content alone, nine on monetization alone, and four on both.

Table OA.4.3. Model and human labels agree on most sampled domains.

Subset	Dimension	n	Categories	Exact	Weighted	κ	95% CI
Overall	Content type	100	11	0.79	0.81	0.76	[0.67, 0.85]
Overall	Monetization	100	6	0.87	0.91	0.82	[0.72, 0.90]
Without model's "other"	Content type	88	11	0.84	0.84	0.82	[0.72, 0.90]
Without model's "other"	Monetization	88	6	0.88	0.90	0.82	[0.72, 0.92]

Notes: The reference is one blind human coder on the stratified 100-domain sample. Exact and weighted agreement are proportions; the weighted measure weights domains by referral volume within this sample. Cohen's κ adjusts for chance agreement; the interval is a 95% percentile-bootstrap interval for κ from 5,000 resamples. The conditional rows exclude twelve model assignments to other, including AI services as well as sites the model could not place. They are not an overall accuracy estimate.

The model assigns twelve validation domains to other and the human assigns six, with five overlapping assignments. Content agreement rises to 84% when the model's twelve other assignments are excluded. Category-specific agreement is lower for reference/knowledge and news/journalism, with precision of 0.55 and 0.56; monetization recall for mixed/other is 0.29 (Table OA.4.4). Small category samples limit the precision of these comparisons.

Reweighting the validation sample to the full domain universe gives per-domain agreement of 0.77 for content and 0.79 for monetization. Post-stratified traffic-weighted agreement is 0.79 and 0.88, respectively (Table OA.4.5).

OA.4.3 Confidence scores and sensitivity

Of the 4,266 classified domains, 76.1% meet the 0.90 confidence threshold and 94.7% meet the 0.60 threshold. We assess the model's self-reported confidence against human labels (Table OA.4.6).

High-confidence content labels agree with the human reference on 83.5% of domains, compared with 61.9% below the threshold; κ is 0.813 versus 0.525. Monetization agreement does not increase with confidence. The shared score thus distinguishes content-label agreement but is not calibrated as a probability of correctness for either dimension.

OA.4.4 Robustness under a lower confidence threshold

Relaxing the threshold from 0.90 to 0.60 adds about 790 domains. ChatGPT's larger referral shares for reference, academic, developer, and tool destinations, and its smaller ad-supported share, persist in this broader sample (Figure OA.4.1).

Table OA.4.4. Agreement varies across categories and should be read with category sample sizes.

Category	Human count	Model count	Precision	Recall	F_1
Content type					
Games and guides	25	17	0.94	0.64	0.76
Tools/SaaS	17	13	0.92	0.71	0.80
E-commerce brand	13	14	0.93	1.00	0.96
Adult	12	10	1.00	0.83	0.91
Reference/knowledge	7	11	0.55	0.86	0.67
Other	6	12	0.42	0.83	0.56
E-commerce marketplace	5	6	0.83	1.00	0.91
News/journalism	5	9	0.56	1.00	0.71
Social media	5	4	0.75	0.60	0.67
Government/institutional	4	3	1.00	0.75	0.86
Academic research	1	1	1.00	1.00	1.00
Monetization					
Ad-supported	43	48	0.85	0.95	0.90
Transactional	25	23	1.00	0.92	0.96
Freemium SaaS	11	12	0.75	0.82	0.78
Non-profit/public	9	7	1.00	0.78	0.88
Mixed/other	7	3	0.67	0.29	0.40
Subscription/paywall	5	7	0.71	1.00	0.83

Notes: Precision, recall, and F_1 compare model assignments with the human reference in the full 100-domain validation sample. Human and model counts are the numbers of validation domains that the human coder and the model assign to each category. Small category counts, including one academic-research domain, limit the precision of category-specific assessments.

Table OA.4.5. Design weighting extends the validation beyond the sampled domains.

		Content type		Monetization	
Panel A: by referral-rank stratum		Exact	κ	Exact	κ
Head (rank 1–200)	n=40	0.80	0.76	0.95	0.92
Middle (rank 201–1,000)	n=30	0.80	0.77	0.87	0.80
Tail (rank 1,001+)	n=30	0.77	0.73	0.77	0.71
Panel B: population estimates		Content type		Monetization	Coverage
Design-weighted, per-domain (Horvitz–Thompson)		0.77		0.79	100%
Post-stratified, traffic-weighted		0.79		0.88	approximately 99%

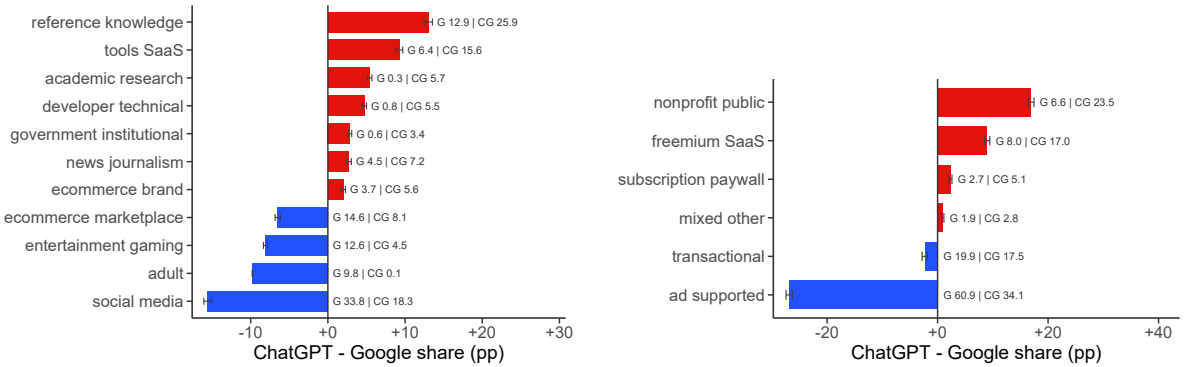
Notes: Panel A reports agreement within each referral-rank stratum. Panel B weights sampled domains by inverse sampling fractions for the per-domain estimates and uses universe referral shares for the post-stratified traffic-weighted estimates. Coverage reports the portion of the target domain or traffic universe represented. The validation sample has forty head, thirty middle, and thirty tail domains.

Table OA.4.6. Higher confidence predicts stronger content agreement, but not stronger monetization agreement.

		Content type		Monetization	
Subset	n	Exact	κ	Exact	κ
Full sample	100	0.790	0.761	0.870	0.817
Confidence $\geq$ 0.90	79	0.835	0.813	0.861	0.813
Confidence $<$ 0.90	21	0.619	0.525	0.905	0.768

Notes: The 100-domain validation sample is split at confidence 0.90. Exact is the model–human agreement share; κ is chance-corrected agreement. These conditional comparisons are descriptive and have different sample sizes. They assess discrimination between labels, not probability calibration.

Figure OA.4.1. The main composition differences persist at the lower confidence threshold.



(a) Content type.

(b) Monetization.

Notes: Bars report ChatGPT's referral share minus Google's by content type (a) and monetization (b) after relaxing the classifier-confidence threshold from 0.90 to 0.60. The domain universe is the union of the two top-2,500 referral lists. Shares use domains meeting the 0.60 threshold after the recoding in Section OA.4.1. Content shares exclude other and unknown; monetization shares exclude unknown and retain mixed/other. Error bars are pointwise 95% normal intervals for differences in proportions.

OA.5 Destination concentration and supplementary characteristics

Main-paper Section 4.3 compares concentration across all referrals and within household-weeks. The checks below vary the destination set and minimum referral count; the final subsection relates referral shares to website prominence.

OA.5.1 Normalized concentration

For an intermediary with referral shares s_i across n observed destination domains, the raw Herfindahl index is $H = \sum_i s_i^2$. Its minimum, $1/n$, depends on how many destinations are observed. We remove this floor using

$$H^* = \frac{H - 1/n}{1 - 1/n}, \qquad n \geq 2. \tag{OA.5.1}$$

The normalized index equals zero for equal shares and approaches one as one destination dominates. We exclude single-destination observations, for which it is undefined. The adjustment removes the equal-share floor; the checks below assess sensitivity to the observed destination set and referral counts.

OA.5.2 Restricting the set of destinations

Table OA.5.1 varies the set of destinations using traffic percentiles, absolute ranks, cumulative traffic shares, minimum referrals from both intermediaries, and the 4,266 classified domains. Shares are recomputed within each set. The rules restrict the head and tail of the distribution in different ways rather than imposing a common number of destinations.

Table OA.5.1. Google referrals remain more concentrated under every destination restriction.

Rule	Destinations included	ChatGPT H	Google H	Ratio	Normalized ratio
Percentile	Top 1% of each service (union)	0.0120	0.0273	2.27	2.30
Percentile	Top 5% of each service (union)	0.0085	0.0230	2.70	2.72
Percentile	Top 10% of each service (union)	0.0076	0.0218	2.87	2.89
Percentile	Top 25% of each service (union)	0.0066	0.0206	3.10	3.12
Percentile	Top 50% of each service (union)	0.0061	0.0198	3.27	3.29
Rank	Top 100 of each service (union)	0.0425	0.0731	1.72	1.87
Rank	Top 500 of each service (union)	0.0198	0.0469	2.37	2.47
Rank	Top 1,000 of each service (union)	0.0155	0.0402	2.59	2.67
Rank	Top 5,000 of each service (union)	0.0095	0.0298	3.13	3.17
Rank	Top 10,000 of each service (union)	0.0079	0.0271	3.44	3.47
Cumulative share	50% of each service's traffic (union)	0.0218	0.0632	2.89	3.09
Cumulative share	80% of each service's traffic (union)	0.0085	0.0290	3.42	3.46
Cumulative share	90% of each service's traffic (union)	0.0066	0.0235	3.58	3.60
Minimum referrals	≥ 5 from each service	0.0111	0.0529	4.76	4.86
Minimum referrals	≥ 10 from each service	0.0140	0.0636	4.54	4.69
Minimum referrals	≥ 50 from each service	0.0260	0.0958	3.69	4.05
Minimum referrals	≥ 100 from each service	0.0353	0.1118	3.17	3.61
Classified domains	4,266-domain classification universe	0.0134	0.0355	2.64	2.68
All domains (main)	All domains	0.0055	0.0191	3.45	3.47

Notes: The full ten-month descriptive referral sample is restricted using each row's rule. The ChatGPT and Google columns report raw Herfindahl indices computed from shares within each set. The ratio columns divide Google's index by ChatGPT's, using raw and normalized indices, respectively. Normalization uses each intermediary's observed destination count within that set. The main paper reports normalized levels, so its levels differ from the raw columns here. No classifier-confidence restriction applies.

Google has higher concentration under every reported rule. The normalized ratio is 3.47 across all domains, 1.87 among each service's top 100 destinations, and 2.68 on the

4,266 classified domains. The classified universe combines the domains most referred by LLM platforms and by search engines (Section OA.4); it differs from the union of ChatGPT's and Google's own 2,500 most-referred domains in main-paper Table 3 (4,026 domains; normalized ratio 2.94). Requiring minimum referrals from both services gives larger ratios than the comparison over all destinations. The direction is stable, while its magnitude depends on which destinations enter the calculation.

OA.5.3 Concentration within household-weeks

The household comparison asks how evenly a household distributes its referrals within a week. We stack each household-week's ChatGPT and Google indices and regress them on a ChatGPT indicator with household and week fixed effects, varying the minimum referrals required from each service. The normalized index additionally requires at least two distinct destinations for the measured intermediary.

Table OA.5.2. The within-household concentration difference depends on referral activity.

Minimum referrals	≥ 1	≥ 2	≥ 3	≥ 5	≥ 10	≥ 15	≥ 20
ChatGPT indicator	−0.001	0.000	0.019***	0.035***	0.056***	0.065***	0.063***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.006)	(0.012)	(0.018)
ChatGPT mean HHI	0.069	0.070	0.091	0.107	0.130	0.150	0.159
Google mean HHI	0.072	0.071	0.074	0.075	0.077	0.088	0.099
Household fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	48,641	35,644	24,553	12,684	3,721	1,541	785

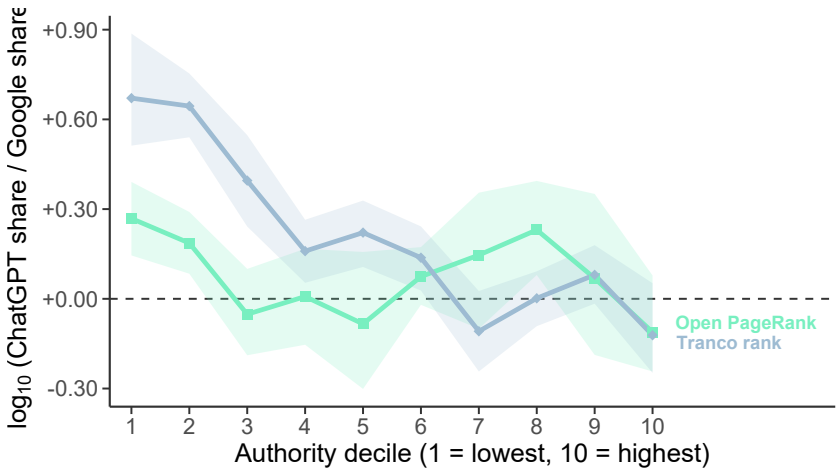
Notes: The dependent variable is household-week normalized HHI. Each specification requires the stated minimum referrals from both intermediaries and excludes household-week-intermediary observations with fewer than two destinations, for which the index is undefined. Household and week fixed effects are included throughout. Positive ChatGPT coefficients indicate greater concentration than Google. Means describe the estimation sample. Standard errors are clustered by household; $^{***}p < 0.01$.

The within-household difference is indistinguishable from zero at the one- and two-referral thresholds and positive at higher thresholds. Together with lower aggregate concentration, this pattern is consistent with different households directing ChatGPT referrals to different destinations.

OA.5.4 Destination prominence

Figure OA.5.1 groups websites using Open PageRank and Tranco rankings. Relative to Google, ChatGPT allocates larger referral shares to less-prominent groups. These rankings measure aspects of link prominence and popularity; they do not measure content quality or the value of a visit.

Figure OA.5.1. ChatGPT allocates relatively larger referral shares to less-prominent destinations.



Notes: Among the classified destinations with observed referral shares and available ranks, 4,201 enter the Open PageRank series and 3,959 enter the Tranco series. Each signal separately groups these domains into deciles, from lowest prominence (one) to highest (ten). The vertical axis is the base-ten logarithm of the ChatGPT share divided by the Google share within each decile. Positive values indicate a relatively larger ChatGPT share. Shaded bands are pointwise 95% percentile intervals from 1,000 domain bootstrap resamples within each signal and decile. The figure labels prominence as authority.

OA.6 Sustained adoption, access expansions, and conventional search

Main-paper Section 5 uses two comparisons: households beginning sustained ChatGPT use versus later adopters, and prior-anonymous ChatGPT users versus prior Gemini users around the February 5, 2025 Search access expansion. Both use the balanced panel and ACS weights. The following subsections define their samples, estimators, and diagnostics; Section OA.6.6 presents the GMM adoption contrast and compares the TWFE access estimates with GMM for ChatGPT use, conventional search, and referral destinations.

Measures. Both designs use the balanced panel of 45,386 households. We measure conventional search as the weekly number of recorded Google, Bing, and Yahoo search-result page loads. Qualifying records satisfy the foreground-page filters defined above and have a recorded URL pattern ending in `/search`. We count each qualifying record separately, including records occurring within the same second. Weeks without qualifying records receive a count of zero.

We measure ChatGPT use through recorded conversation sessions, combining logged-in and anonymous loads of the same conversation UUID within each household and source month (Section OA.1.4). A new session begins when more than one hour has elapsed since the preceding load of that conversation.

OA.6.1 Sustained-adoption design

The adoption design follows Padilla et al. (2026). A household's sustained use begins in the first of at least three consecutive complete weeks with a recorded conversation session. The design restricts the sample to eventual adopters and compares each adoption cohort with households that have not yet adopted, using the unconditional Callaway and Sant'Anna (2021) group-time estimator. Our setting differs from theirs in three respects: it studies ChatGPT rather than several LLMs, it uses the four-records-per-month panel rather than a four-active-days rule, and it does not observe adoption after July 2025.

Of 1,361 households with such a run, 282 begin during the first four observed weeks (October 1–28, 2024) and are excluded so that every adopter has a pre-adoption history. That leaves 1,079 eventual adopters, 1,008 of whom have positive ACS weights. Occasional use before onset is allowed, so onset is neither first-ever ChatGPT use nor first use of Search. Weeks run Tuesday through Monday, giving 43 complete weeks from October 1, 2024 through July 28, 2025.

For adoption cohort g and calendar week t, the comparison uses eventual adopters whose onset G_i satisfies $G_i > \max(g, t)$:

$$\widehat{ATT}(g,t) = \overline{Y_{it} - Y_{i,g-1}}_{G_i=g,\,w} - \overline{Y_{it} - Y_{i,g-1}}_{G_i>\max(g,t),\,w}, \tag{OA.6.1}$$

where w denotes ACS-weighted means. Each estimated cell requires at least two adopting and two comparison households. Event-time estimates aggregate these cells by the adopting cohorts' weight; household influence scores include the uncertainty in those cohort shares. Window averages weight each estimated event week equally and retain covariance across weeks. Interpretation requires parallel counterfactual trends among eventual adopters and no anticipation.

The early/later split in main-paper Figure 9 uses the first session day of the onset week: 366 households began sustained use before February 5, 2025 and 642 from that date. The 45 households in the final cohort serve only as comparisons, leaving 597 later adopters. The number of later adopters observed falls to 54 by week 19, the last reported week. Table OA.6.1 reports two checks. Unweighted estimates, which retain the 71 adopters without usable weights, are close to the weighted ones. Holding the later group fixed at the 54 households observed through week 19 gives a smaller difference between the groups over weeks 0–19 (-7.51, $SE = 4.39$, against -12.77, $SE = 4.60$, with all contributing households). Much of the faster decline among later adopters therefore comes from households that began sustained use from late February onward and are observed into the summer.

Comparing adoption effects. At each relative week w, we estimate the later-minus-earlier contrast $\widehat{\Delta}_w = \widehat{ATT}_{\text{later},w} - \widehat{ATT}_{\text{earlier},w}$. We conduct inference using a joint household-level multiplier bootstrap, following the influence-function approach of Callaway and Sant'Anna (2021). Let $\widehat{\psi}_{i,s,w}$ denote household i's estimated influence function for group s, including uncertainty in the estimated cohort shares, and let n

be the number of households. For each of 9,999 replications, we draw independent household multipliers ξ_i, taking values -1 and 1 with equal probability, and calculate

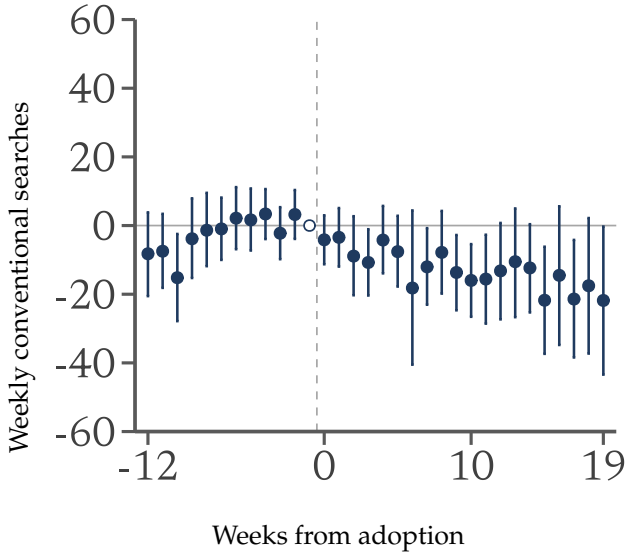
$$Z_w^* = \frac{1}{n}\sum_{i=1}^{n} \xi_i \left(\widehat{\psi}_{i,\text{later},w} - \widehat{\psi}_{i,\text{earlier},w}\right). \tag{OA.6.2}$$

We use the same household multiplier across both groups and all relative weeks, preserving within-household dependence and covariance arising from shared controls. The plotted pointwise 95% confidence interval is $[\widehat{\Delta}_w - q_{.975,w},\ \widehat{\Delta}_w - q_{.025,w}]$, where $q_{\alpha,w}$ is the empirical α-quantile of Z_w^*. These intervals concern individual weeks and do not provide simultaneous coverage of the entire path. The contrast compares adoption effects at equal durations across different adopter populations and calendar periods; it does not separately identify the effect of Search availability.

Main-paper Figure 9(c) and the final column of Table 5 use the four-baseline GMM contrast described in Section OA.6.6. The single-baseline comparison below provides its robustness benchmark.

Single-baseline comparison. Figure OA.6.1 subtracts the estimates in main-paper Figure 9(a) from panel (b), using the 9,999-draw influence-function bootstrap above. The weeks 4–19 average is -14.26 ($SE = 5.27$). Main-paper Table 5 reports single-baseline estimates for all, earlier, and later adopters and the four-baseline GMM contrast in its final column.

Figure OA.6.1. The single-baseline contrast also shows conventional searches declining sooner among later adopters.



Notes: Points show the ACS-weighted later-minus-earlier difference in changes in weekly Google, Bing, and Yahoo searches per household, using not-yet-adopting households as comparisons and week -1 as the normalized reference. The 1,008-household sample contains 366 earlier and 642 later adopters; 45 later adopters serve only as comparisons. The number of later adopters contributing to the estimates falls from 597 at week 0 to 54 at week 19. Error bars show pointwise 95% basic intervals from 9,999 joint household Rademacher multiplier draws, preserving shared-control dependence and uncertainty in cohort shares. It complements the four-baseline GMM contrast in main-paper Figure 9(c).

Table OA.6.1. Adoption comparisons using all available adopters or a fixed sample.

Weeks 0–19	Earlier	Later	Later minus earlier
Panel A. ACS-weighted			
All available adopters	2.18 (2.64)	-10.59 (3.90)	-12.77 (4.60)
Adopters observed through week 19	2.18 (2.64)	-5.33 (3.66)	-7.51 (4.39)
Panel B. Unweighted			
All available adopters	3.67 (2.29)	-8.58 (4.01)	-12.25 (4.63)
Adopters observed through week 19	3.67 (2.29)	-2.42 (4.50)	-6.08 (5.03)

Notes: Entries are changes in weekly conventional searches averaged over weeks 0–19 after sustained adoption. Onset precedes February 5, 2025 for earlier adopters and falls on or after that date for later adopters, using the first session day in the onset week. All-available rows use the adopters contributing at each event week; fixed-sample rows retain the same adoption cohorts throughout weeks 0–19. ACS weights adjust for age and income. Household-clustered standard errors in parentheses account for shared comparison households and estimated cohort shares.

OA.6.2 February cohorts and classification

The February design dates treatment from the expansion rather than from each household's adoption, and fixes group membership with signals recorded before February 5. The initial 946 prior-anonymous households have at least three ChatGPT app or conversation requests before February 5, at least one day of anonymous use, no day of logged-in or mixed use before the expansion, and no request for a paid feature. These requests show how a household used ChatGPT; the data contain no login or subscription records.

Interface classification. Classification uses exact routes on `chatgpt.com`, with `chat.openai.com` normalized to that host. Protocols, query strings, fragments, and trailing slashes are removed, identifiers are replaced by `{uuid}`, only HTTP 200–299 responses qualify, and requests are deduplicated by household, day, second, and normalized URL. The routes used under both `backend-api/` and `backend-anon/` are `conversation`, `conversation/{uuid}`, `conversation/gen_title/{uuid}`, `conversation/init`, and `sentinel/chat-requirements`; anonymous routes mark days of anonymous use and the corresponding `backend-api/` routes mark days of logged-in use. The `backend-api/` routes `conversations` and `accounts/check/v4-2023-04-27` also mark logged-in use. Feedback, paragen, and failed requests are excluded.

Gemini comparison. Controls had used Gemini before February 5 and belong to the pool of eligible comparison households, which excludes qualifying pre-expansion ChatGPT/OpenAI and Claude/Anthropic activity. They also record no qualifying foreground ChatGPT visits between October 1 and February 4. We retain demographic categories observed in both arms across age, income, household size, children, gender, race, and location. This removes one treated household with unknown household size, leaving 945 treated and 429 comparison households. Membership is fixed before the expansion; subsequent ChatGPT use does not remove a comparison household.

Prior search activity. We classify households by mean weekly searches over all 18 pre-expansion weeks (October 2, 2024–February 4, 2025), winsorizing weekly counts

at the pooled 95th percentile before averaging. The percentile is calculated across the original 30,757 households and all 18 pre-expansion weeks. Households at or below the treated ACS-weighted median of 18.5 enter the lower group (492 treated, 202 controls); those above it enter the higher group (453 treated, 227 controls). Search outcomes used in estimation are winsorized at the 99th percentile (Section OA.6.4).

OA.6.3 Demographic weighting

Both designs use ACS 2023 age-by-income representation weights and exclude households without usable positive weights. Weighted means normalize weights within each group. In the Gemini comparison, the effective sample sizes, $(\sum_i w_i)^2 / \sum_i w_i^2$, are 774 treated and 355 comparison households. Weighting adjusts observed demographic representation but does not address unobserved differences or missing mobile activity.

OA.6.4 Estimation and inference

For the Gemini comparison, weekly search counts are winsorized at 166, the unweighted 99th percentile of pooled household-week searches across the original 30,757-household sample and all 18 pre-expansion weeks, including zeros. This threshold applies to both groups in every week; conversation sessions and category referrals remain uncapped.

Event week 0 begins on February 5, 2025. Weeks run Wednesday through Tuesday, and estimation uses weeks -18 through 24, ending July 29; the following week is incomplete. The main-paper TWFE event-study coefficient for outcome Y and treated indicator T_i, relative to week -1 (January 29–February 4), is

$$\widehat{\delta}_k = (\bar{Y}_{T=1,k} - \bar{Y}_{T=0,k}) - (\bar{Y}_{T=1,-1} - \bar{Y}_{T=0,-1}). \tag{OA.6.3}$$

On the balanced panel with fixed ACS weights, these coefficients equal those from an event-study regression with household and calendar-week fixed effects. Main-paper Table 6 and Table 7 report their averages over weeks 0–24 for the full sample and prior-search groups. Week -1 is normalized to zero in the event-study figures.

Standard errors use 999 positive Exp(1) household bootstrap draws, applying each multiplicity to every observation of a household and refitting means and, for GMM, covariance weights. Classification, cutoffs, support, and ACS weights remain fixed. Outcome intervals are pointwise 95% basic bootstrap intervals; window averages and contrasts use aligned draws. Event-study intervals are pointwise, without adjustment for multiple comparisons. The bootstrap conditions on the observed group classification and does not account for the comparison of estimators and classifications that preceded these results.

Search referrals. The category estimates in panel (b) of main-paper Figure 12 use the full access sample: 945 prior-anonymous ChatGPT households and 429 Gemini comparison households. ACS weights and the weekly calendar match Section OA.6.4. Outcomes count clean foreground destination-page visits referred by Google, Bing, or Yahoo, without capping. The eleven content categories retain 3,103 domains with classification confidence of at least 0.90, excluding other and unknown (Section OA.4.1). Each qualifying destination record counts once, without a query-attribution window or

additional deduplication; missing household-week-category counts are zero. Panel (b) reports all eleven categories, including adult websites.

Each main-paper category effect uses TWFE with household and calendar-week fixed effects, week -1 as reference, and averages weeks 0–24. Panel (b) divides by the treated category mean over all 18 pre-expansion weeks and multiplies by 100. Its pointwise 95% intervals use analytic household-clustered standard errors and normal critical values, with percentage denominators fixed.

Baseline composition. The baseline composition comparison in panel (a) of main-paper Figure 12 uses treated households alone and the same eleven categories over all 18 pre-expansion weeks. Shares include adult websites in the denominator and retain the classification-confidence threshold of 0.90. For group g and category c, the share is

$$s_{gc} = \frac{\sum_{i \in g} w_i R_{ic}^{\text{pre}}}{\sum_{i \in g} w_i \sum_{k=1}^{11} R_{ik}^{\text{pre}}}, \tag{OA.6.4}$$

where R_{ic}^{pre} counts pre-expansion referred visits. Households with no classified referrals contribute zero visits. Uncertainty uses household ratio influence scores; tests of lower-minus-higher shares use Holm adjustment across eleven categories. The lower group's tools/SaaS share is 4.5 percentage points larger (95% interval [2.6, 6.4]; adjusted $p < 0.001$), and its social-media share is 9.8 points smaller ($[-14.7, -4.9]$; adjusted $p < 0.001$). Adult accounts for 3.2% of lower-activity and 6.7% of higher-activity households' referrals, a difference of -3.5 points ($[-5.9, -1.0]$; adjusted $p = 0.036$). Differences use unrounded shares. Reference/knowledge shares are similar (18.8% versus 16.9%). These shares describe classified referral destinations before the access expansion, not query topics or behavior before any ChatGPT exposure.

OA.6.5 Diagnostics for the parallel-trends assumption

Table OA.6.2 reports joint lead tests for the single-baseline adoption and February access specifications. None rejects the null of zero pre-event effects at the 5% level, providing no statistically significant evidence of differential pretrends at this level for the reported specifications. For ACS-weighted adoption estimates, the search lead-test p-values are 0.295, 0.280, and 0.291 for all, earlier, and later adopters. For the February access comparison, the full-sample search and session tests have p-values of 0.248 and 0.433; neither test rejects in either prior-search group.

OA.6.6 Robustness of model specification

Generalized method of moments (GMM) combines estimates based on multiple pre-event weeks. We use four baselines for the later-minus-earlier adoption contrast in main-paper Figure 9(c), and eighteen baselines to assess the robustness of the February access results to the week--1 TWFE reference. Each comparison retains its design's households, ACS weights, outcomes, and reporting window. The implementations below distinguish their baseline sets, comparison groups, and inference.

Table OA.6.2. Pre-period tests for adoption and February access.

Group	Weighting	Search p	Late search p	Session p
Panel A. Sustained adoption				
All adopters	ACS	0.295	—	—
Earlier adopters	ACS	0.280	—	—
Later adopters	ACS	0.291	—	—
All adopters	Unweighted	0.238	—	—
Earlier adopters	Unweighted	0.458	—	—
Later adopters	Unweighted	0.109	—	—
Panel B. February access: Gemini comparison				
All households	ACS	0.248	0.068	0.433
Lower prior search	ACS	0.293	0.135	0.408
Higher prior search	ACS	0.405	0.212	0.417

Notes: Entries are joint Wald-test p-values using household-clustered covariance. Panel A tests conventional-search leads in weeks -12 through -2 relative to the week before sustained adoption; earlier onset precedes February 5, 2025. Panel B uses the Gemini comparison and week -1 as reference: search and session tests cover weeks -18 through -2, while late-search tests cover weeks -9 through -2. Search covariance ranks are 17 and 8; session ranks are 12, 5, and 7 for all, lower-, and higher-activity households. February conventional searches are winsorized at 166 per household-week; sessions are uncapped. ACS denotes age-by-income weights. Dashes indicate tests not reported. February inference is exploratory because the specification was chosen after comparing alternatives. Nonrejection does not establish parallel trends.

Sustained adoption: four-baseline GMM. Main-paper Figure 9(c) uses a four-baseline GMM estimator of the later-minus-earlier contrast defined in Section OA.6.1. For each adopter group s, onset cohort g, and outcome week t, let $\widehat{d}_{s,g,t,b}$ denote the ACS-weighted DiD in Equation OA.6.1 with baseline $b \in \mathcal{B}_g = \{g-4, g-3, g-2, g-1\}$, restricting the adopting cohort to group s. The sample requires at least four observed weeks before adoption, so these baselines give every cohort the same length of pre-adoption history while retaining the original sample. Each moment uses the same comparison households, $G_i > \max(g,t)$. Stack the four estimates in $\widehat{\boldsymbol{d}}_{s,g,t}$ and estimate their joint household covariance $\widehat{V}_{s,g,t}$. The combined estimate is

$$\widehat{\tau}^{\text{GMM}}_{s,g,t} = \widehat{\boldsymbol{a}}'_{s,g,t}\widehat{\boldsymbol{d}}_{s,g,t}, \qquad \widehat{\boldsymbol{a}}_{s,g,t} = \frac{\widehat{V}^{-1}_{s,g,t}\mathbf{1}}{\mathbf{1}'\widehat{V}^{-1}_{s,g,t}\mathbf{1}}. \tag{OA.6.5}$$

The weights sum to one and can be negative. This GMM combination uses four baselines and pooled not-yet-adopting controls, gaining precision when the baseline-specific estimates identify the same cohort effect.[6]

For each group and relative week, we aggregate $\widehat{\tau}^{\text{GMM}}_{s,g,g+w}$ over supported cohorts using their ACS weight shares, then subtract the earlier-group estimate from the later-group estimate. Outcomes are uncapped weekly search counts. Identification requires parallel untreated changes from each baseline to the outcome week and no anticipation

[6]For efficiency gains from multiple baseline periods and comparison groups, see Chen et al. (2025).

of sustained adoption. Occasional use before onset is allowed. The contrast compares adopter populations observed in different calendar periods.

Inference uses 999 independent draws of positive household multiplicities $M_i \sim \text{Exp}(1)$. Each draw applies the same multiplicity to all observations and comparison roles of household i and refits the moment means, covariance weights, and ACS cohort shares. Multiplicities enter the empirical covariance once, as observation masses. The pointwise 95% basic interval is $[2\widehat{\Delta}_w^{\text{GMM}} - q^*_{.975,w},\ 2\widehat{\Delta}_w^{\text{GMM}} - q^*_{.025,w}]$, where $q^*_{\alpha,w}$ is the quantile of the refitted contrasts. Window averages use their joint bootstrap distribution. Two-sided window tests compare the estimated contrast with the centered bootstrap distribution, using twice the smaller tail probability, with one added to each tail count and to the number of draws for Monte Carlo resolution. For pre-adoption placebo points, a baseline coinciding with the outcome week is excluded. Week -1 is estimated using weeks -4, -3, and -2, with the same household bootstrap draws as the other weeks; the GMM post-adoption path is not rebased to that week.

The later-minus-earlier difference averages -13.01 searches per household-week in weeks 4–19 ($SE = 3.91$; 95% interval $[-20.68, -5.01]$). Its joint lead test for weeks -12 through -1 rejects at the 5% level ($p = 0.023$). Some cohort cells contain only five adopting households, and baseline weights range from -16.91 to 10.30. Deleting each household in turn leaves the window-average estimate between -13.93 and -11.00, although individual weekly estimates are more sensitive. Inference is exploratory: the intervals do not adjust for the comparison of estimators preceding selection of GMM.

February access: eighteen-baseline GMM. GMM combines information from all 18 pre-expansion baseline weeks (Chen et al., 2025). For each outcome and sample, let $\widehat{d}_{k,b}$ be the ACS-weighted treated-minus-comparison change from baseline $b \in \mathcal{B} = \{-18, \ldots, -1\}$ to week k. Stack these contrasts in $\widehat{\boldsymbol{d}}_k$, with joint household covariance $\widehat{V}_k$. The estimate is

$$\widehat{\delta}_k^{\text{GMM}} = \frac{\mathbf{1}'\widetilde{V}_k^{-1}\widehat{\boldsymbol{d}}_k}{\mathbf{1}'\widetilde{V}_k^{-1}\mathbf{1}}, \qquad \widetilde{V}_k = \widehat{V}_k + \eta_k I. \tag{OA.6.6}$$

The covariance weights again sum to one and can be negative. The adjustment η_k raises the smallest eigenvalue to at least $10^{-10}\max\{\text{tr}(\widehat{V}_k), 10^{-12}\}$, stabilizing sparse session moments. Weights are fitted separately for each outcome, sample, and week. Preperiod placebos omit the outcome week from the baseline set, so week -1 is estimated. Post-expansion effects average weeks 0–24. GMM requires parallel counterfactual changes from each baseline and no anticipation of the expansion.

Precision. For an equal-weight pre-expansion benchmark, the ACS-weighted TWFE model

$$Y_{it} = \alpha_i + \lambda_t + \beta\, T_i\, \mathbf{1}\{t \geq 0\} + \varepsilon_{it} \tag{OA.6.7}$$

gives the average post-expansion gap minus the average pre-expansion gap. Using the same bootstrap draws, GMM reduces the search-window standard error from 1.16 to 1.02 in the full sample, from 0.93 to 0.78 in the lower-activity group, and from 2.00 to 1.75 in the higher-activity group. These gains of about 12–16% compare GMM with the equal-weight pre-expansion benchmark, rather than the week--1 reference used in the main paper. Session precision changes little; the lower-group session standard error increases slightly.

ChatGPT use and conventional search. GMM confirms an increase in recorded ChatGPT use without a detectable overall change in conventional search (Table OA.6.3). The estimates are 0.127 additional sessions per household-week ($SE = 0.020$) and 0.21 additional searches ($SE = 1.02$). Searches decline by 2.81 per week in the lower-activity group ($SE = 0.78$), compared with an imprecise increase of 2.47 in the higher-activity group ($SE = 1.75$). The lower-minus-higher contrast is -5.29 ($SE = 1.91$). These estimates preserve the main-paper finding that search substitution is concentrated among households with lower prior activity.

Table OA.6.3. GMM estimates for the access expansion.

	All households	Lower prior search	Higher prior search
ChatGPT sessions	0.127***	0.105***	0.151***
	(0.020)	(0.024)	(0.032)
Conventional searches	0.21	−2.81***	2.47
	(1.02)	(0.78)	(1.75)
Treated / comparison	945 / 429	492 / 202	453 / 227
TWFE conventional-search lead-test p	0.248	0.293	0.405
Lower − higher conventional searches		−5.29***	
		(1.91)	

Notes: Eighteen-baseline GMM estimates use the same households, ACS weights, prior-search groups and outcome definitions as main-paper Table 6 and Table 7. Estimates average weeks 0–24 after February 5, 2025; conventional searches (Google, Bing, and Yahoo result-page loads) are winsorized at the 99th percentile. The reported TWFE conventional-search lead tests cover weeks −18 to −2 relative to week −1; they test the single-baseline preperiod contrasts. See Section OA.6.6 for construction. Standard errors from 999 household bootstrap refits are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Search referrals by category. Table OA.6.4 compares TWFE and GMM for the ten categories excluding adult websites in the full sample and both prior-search groups. Both estimators use the referral measures in Section OA.6.4, identical percentage denominators, and the same 999 household bootstrap draws. The ten full-sample TWFE point estimates match main-paper Figure 12(b). The table uses bootstrap tests, whereas the figure uses analytic household-clustered normal intervals; their significance marks can therefore differ.

The full-sample tools/SaaS estimate changes from −46.0% under TWFE to −10.0% under GMM, and the e-commerce brand estimate from −26.5% to −9.7%. Social media changes from −11.0% to 2.1%. All ten full-sample GMM intervals include zero. The category evidence therefore provides less stable support for particular search use cases than the overall and prior-activity comparisons.

The TWFE category results also differ in their pretrend diagnostics. Joint lead-test p-values are 0.377 for tools/SaaS, 0.326 for e-commerce brands, and 0.930 for social media. Reference/knowledge has a −39.5% TWFE estimate with a 95% interval excluding zero, but its lead test rejects ($p = 0.009$), as do those for news/journalism ($p = 0.016$) and government/institutions ($p = 0.029$).

Among the subgroup GMM estimates, referrals decline by 12.8% for lower-activity tools/SaaS and 24.6% for lower-activity e-commerce brands, while higher-activity social-media referrals increase by 3.6%. Only the brand interval among these three excludes zero; its unadjusted $p = 0.030$ becomes 0.270 under the ten-category Holm

correction. These exploratory comparisons do not establish differences in effects across categories or groups, and referral effects do not measure all-source traffic.

Table OA.6.4. Conventional-search referral effects by estimator and prior search activity.

	All households		Lower prior search		Higher prior search	
Destination category	TWFE	GMM	TWFE	GMM	TWFE	GMM
Academic research	−20.1	−8.1	−21.5	6.2	−20.2	−8.5
	(21.8)	(6.7)	(27.9)	(14.7)	(25.3)	(6.9)
Developer/technical	30.2	−9.4	−98.1*	−25.5	62.8	−0.2
	(75.9)	(12.7)	(61.7)	(19.7)	(91.6)	(14.2)
E-commerce brand	−26.5*	−9.7	−44.2*	−24.6**	−21.0	−5.4
	(15.8)	(6.4)	(24.6)	(10.8)	(18.8)	(7.4)
E-commerce marketplace	3.0	6.0	−46.7	−19.4	12.4	15.8**
	(16.8)	(8.5)	(31.2)	(27.6)	(18.3)	(7.4)
Games and guides	−15.9	1.8	−29.8	5.9	−13.2	0.9
	(19.5)	(9.7)	(38.5)	(19.3)	(21.0)	(10.5)
Government/institutional	−4.0	8.6	16.2	9.8	−11.9	2.9
	(20.6)	(7.9)	(34.6)	(13.8)	(24.5)	(9.1)
News/journalism	−32.5	0.6	−30.2	21.4**	−33.8	−6.0
	(21.7)	(7.7)	(28.9)	(10.6)	(23.9)	(8.1)
Reference/knowledge	−39.5**	1.5	−71.4*	−29.9***	−33.1*	2.2
	(14.5)	(5.7)	(32.6)	(8.8)	(15.6)	(6.3)
Social media	−11.0	2.1	−40.0	−13.4	−8.4	3.6
	(14.0)	(4.9)	(29.8)	(12.2)	(15.1)	(5.2)
Tools/SaaS	−46.0*	−10.0	−37.7	−12.8	−49.8	−9.6
	(23.3)	(6.5)	(30.2)	(10.6)	(28.8)	(7.4)
Treated households	945	945	492	492	453	453
Comparison households	429	429	202	202	227	227

Notes: ACS-weighted effects compare prior-anonymous ChatGPT and prior-Gemini households. Outcomes are uncapped visits referred by Google, Bing, and Yahoo, averaged over weeks 0–24 after February 5, 2025. TWFE uses household and calendar-week fixed effects and week −1 as reference; GMM combines all 18 pre-expansion baselines. Estimates and standard errors are percentages of the treated group/category mean over all 18 pre-expansion weeks, held fixed. Both methods use identical samples and outcome definitions. Stars use unadjusted bootstrap tests; inference is exploratory because the specification was chosen after comparing alternatives. Standard errors from 999 household bootstrap draws are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

OA.7 Domain-classification prompt

The following prompt specifies the category definitions, exceptions, examples, and output requirements used with GPT-4o (Section OA.4). Each batch supplies five domains and enables web search at temperature zero. Actual web-search use is recorded from API tool-call events. The complete prompt is reproduced with punctuation adapted for typesetting.

```
<instructions>
You are classifying website domains along two dimensions for an academic study of
↪  how AI search redistributes web traffic.
```

```
The taxonomy is organized around AI-search displacement risk -- how easily an LLM
 ↪  can answer the user's underlying need without sending the user to the
 ↪  destination. Groupings (reference vs. social, brand ecommerce vs. marketplace
 ↪  ecommerce, adult as a policy-filtered bucket, frontier LLM chatbots quarantined
 ↪  into "other") are chosen because they isolate quantitatively different
 ↪  displacement patterns in the downstream regressions.

Classify each domain along these two dimensions: content_type, monetization. Provide
 ↪  a one-sentence rationale and a float confidence score in [0.0, 1.0].

Label each domain based on its homepage and the primary user activity the site
 ↪  supports. Cite the homepage evidence behind your label in the rationale (e.g.,
 ↪  "homepage is a feed of subreddit post cards with vote counts and comment
 ↪  counts"). If you cannot determine the site's function -- domain is unrecognized,
 ↪  parked, or multi-mode with no clear primary activity -- set content_type to
 ↪  "other", monetization to "mixed_other", confidence <= 0.3, and say so explicitly
 ↪  in the rationale. Do not guess from domain name or TLD alone.
</instructions>

<dimensions>

<dimension name="content_type">
12 values, listed alphabetically (the ordering is intentional and not a ranking).

- "academic_research": Peer-reviewed journals, preprint servers, research
 ↪  organizations, academic databases. Examples: nih.gov, arxiv.org, nature.com,
 ↪  sciencedirect.com, pubmed.ncbi.nlm.nih.gov.
- "adult": Explicit-content sites. Quarantined into its own category because the
 ↪  AI-Google gap on adult content is a policy artifact (LLMs filter by design) not
 ↪  an organic compositional difference; mixing it with other buckets contaminates
 ↪  the substantive findings. Examples: pornhub.com, xvideos.com, xhamster.com,
 ↪  onlyfans.com.
- "developer_technical": Code repositories, package registries, developer
 ↪  documentation, technical reference for programmers. Examples: github.com,
 ↪  npmjs.com, developer.mozilla.org, stackoverflow-for-teams-style docs.
- "ecommerce_brand": Manufacturer / direct-to-consumer sites selling their own goods
 ↪  or services. The site's catalog is its own brand. Examples: nike.com, apple.com,
 ↪  patagonia.com, tesla.com.
- "ecommerce_marketplace": Multi-seller marketplaces and broad-line retailers that
 ↪  aggregate third-party or multi-brand catalogs. Examples: amazon.com, ebay.com,
 ↪  etsy.com, walmart.com, target.com.
- "entertainment_gaming": Gaming sites, mod communities, fan-fiction archives,
 ↪  play-focused destinations. **Also includes any wiki, encyclopedia, or database
 ↪  whose entire scope is one game, game franchise, or game series -- even when the
 ↪  homepage looks like a general reference index. A "wiki" or "db" or ".wiki" /
 ↪  "*pedia.net" / "*db.net" domain devoted to a single game belongs here, NOT in
 ↪  reference_knowledge.** Examples: steamcommunity.com, fandom.com, nexusmods.com,
 ↪  gamebanana.com, gta5-mods.com, ign.com when gaming-only, runescape.wiki,
 ↪  stardewvalleywiki.com, pokemondb.net, bulbagarden.net (Pokemon),
 ↪  paradoxwikis.com, minecraft.wiki, terrariawiki.org, oldschool.runescape.wiki.
- "government_institutional": .gov sites, public-service portals, international
 ↪  institutions, authoritative policy/regulatory sources. Examples: irs.gov,
 ↪  cdc.gov, usa.gov, europa.eu, who.int.
- "news_journalism": Original reporting, current news, opinion, and editorial
 ↪  commentary consumed as timely/topical content. Examples: nytimes.com,
 ↪  reuters.com, apnews.com, bbc.com, forbes.com.
```

```
- "other": Three distinct uses. (a) **Frontier general-purpose LLM chatbots,
 ↪  AI-search assistants, multi-LLM aggregator chatbots, and persona-based /
 ↪  companion AI chatbots** whose primary product IS the conversational AI. Covers
 ↪  (i) general-purpose assistants (chatgpt.com, openai.com, claude.ai,
 ↪  gemini.google, perplexity.ai, pplx.ai, deepseek.com, ollama.com), (ii) multi-LLM
 ↪  aggregator wrappers (chaton.ai, monica.im, iask.ai, toolbaz.com), and (iii)
 ↪  persona / companion / roleplay AI chatbots where users converse with
 ↪  AI-generated characters even when the homepage shows a UGC-character feed
 ↪  (character.ai, polybuzz.ai, spicychat.ai, replika.ai, dream.ai,
 ↪  woebothealth.com, and similar). These are all the AI-search substitute being
 ↪  studied; bucketing them across "tools_saas" / "social_media" /
 ↪  "entertainment_gaming" would scatter the AI substitute across multiple buckets
 ↪  and dilute the very category the analysis treats as the substitute. Use
 ↪  confidence >= 0.9 here. (b) **AI-only single-purpose generators** for image /
 ↪  video / voice / music / avatars whose entire product is the AI generation step
 ↪  (midjourney.com, runwayml.com, suno.com, udio.com, elevenlabs.io, openart.ai,
 ↪  leonardo.ai, ideogram.ai, invideo.io, synthesia.io, heygen.com, descript.com,
 ↪  fliki.ai, pika.art, lalal.ai, soundraw.io, craiyon.com, aiva.ai,
 ↪  magicstudio.com, aistudios.com, maxstudio.ai, and similar). They are
 ↪  AI-substitution-adjacent destinations whose displacement dynamics belong with
 ↪  category (a), not with general productivity SaaS. Use confidence >= 0.9 here.
 ↪  Note: domain-specific tools that *embed* AI features but whose primary product
 ↪  is the underlying tool (Grammarly, Canva, Notion, Adobe Photoshop, SodaPDF,
 ↪  Zoom) stay in "tools_saas". (c) Domains the homepage cannot pin down
 ↪  (unrecognized, parked, multi-mode with no clear primary activity). Use
 ↪  confidence < 0.6 here.
- "reference_knowledge": Encyclopedic, how-to, dictionary, tutorial, or general
 ↪  reference content consumed as timeless lookup. **Also includes education-purpose
 ↪  platforms even if their business model is SaaS** -- learning management systems
 ↪  (Canvas/instructure.com, Schoology, Blackboard, Ellucian, Clever, ClassLink,
 ↪  PowerSchool, Jupitered), homework / study tools (Chegg, Quizlet, Quizizz,
 ↪  Mathway, Symbolab, WolframAlpha, Desmos, GeoGebra, Turnitin, Duolingo), K-12
 ↪  adaptive-learning platforms (IXL, Edgenuity, DreamBox, Prodigy, Kahoot),
 ↪  tutoring marketplaces (Preply, Wyzant, italki, Cambly, Outschool), exam
 ↪  proctoring (Honorlock, ProctorU, Respondus), MOOCs and online courses (Coursera,
 ↪  edX, Udemy, Cengage, McGraw-Hill), education publishers (Pearson), and student
 ↪  career platforms (Handshake). The user's purpose at all of these is learning /
 ↪  coursework / education-knowledge lookup, not productivity tooling. **Also
 ↪  includes hospitals and health publishers** (mayoclinic.org, clevelandclinic.org,
 ↪  webmd.com, healthline.com, drugs.com, medlineplus.gov, kidshealth.org): their
 ↪  homepages present a health-reference index (symptoms, conditions, drugs, tests),
 ↪  not an institutional/government portal -- bucket them with reference_knowledge,
 ↪  NOT government_institutional, even when the .org TLD and clinic branding suggest
 ↪  "institution". Examples: wikipedia.org, wikihow.com, britannica.com,
 ↪  dictionary.com, instructure.com, khanacademy.org, chegg.com, coursera.org,
 ↪  edgenuity.com, ixl.com, preply.com, mayoclinic.org, webmd.com.
- "social_media": All social and consumption-on-platform UGC and streaming sites --
 ↪  text-first social/Q&A/threaded discussion *and* rich-media (video, audio,
 ↪  image-centric) UGC and professional streaming libraries. The user is on the
 ↪  platform either to read/post text or to watch/listen; consumption typically
 ↪  requires the platform. Examples: reddit.com, quora.com, stackoverflow.com,
 ↪  x.com, twitter.com, facebook.com, linkedin.com, youtube.com, instagram.com,
 ↪  tiktok.com, spotify.com, netflix.com, twitch.tv.
```

```
- "tools_saas": Productivity tools, web apps, utilities, design tools, SaaS
 ↪  dashboards used for general work -- *not* education or learning, *not* frontier
 ↪  LLM chatbots, and *not* AI-only single-purpose generators. The user uses or logs
 ↪  into a tool. Examples: canva.com, grammarly.com, notion.so, zoom.us,
 ↪  docs.google.com. Note: education-SaaS (LMS, study tools, MOOCs, homework
 ↪  helpers, K-12 learning platforms, tutor marketplaces, exam proctoring) goes to
 ↪  "reference_knowledge"; frontier general-purpose LLM chatbots and AI-only
 ↪  single-purpose generators go to "other" (see below); Shopify storefronts and the
 ↪  merchant tooling at shopify.com / *.myshopify.com go to "ecommerce_marketplace".
</dimension>

<dimension name="monetization">
6 values, listed alphabetically (the ordering is intentional and not a ranking).

- "ad_supported": Default for any free-to-access site whose dominant homepage
 ↪  revenue signal is advertising (display ads, video pre-rolls, sponsored content
 ↪  units, ad slots in feeds). **A low-uptake premium tier -- "remove ads", YouTube
 ↪  Premium, Quora+ -- does NOT promote the site to mixed_other.** The premium tier
 ↪  has to materially co-fund the site, not just exist as an optional upsell.
 ↪  Examples: cnn.com, weather.com, youtube.com, quora.com, fandom.com,
 ↪  allrecipes.com, reddit.com. NOTE: if the site is a SaaS app with visible pricing
 ↪  tiers (spotify, dropbox, notion), use freemium_saas / subscription_paywall
 ↪  instead -- those are matched earlier in the monetization decision tree below.
- "freemium_saas": Functional free tier plus paid upgrade (SaaS or app). Examples:
 ↪  spotify.com, notion.so, grammarly.com, chatgpt.com, dropbox.com.
- "mixed_other": Reserved for sites with TWO OR MORE roughly equal revenue streams
 ↪  visible on the homepage. Use only when the secondary stream is substantial -- a
 ↪  metered paywall most readers hit, marketplace commissions on a multi-revenue
 ↪  platform, in-app commerce comparable to ad revenue. NOT a fallback for "doesn't
 ↪  fit elsewhere". Examples: nytimes.com (metered paywall reaches a large share of
 ↪  readers), twitch.tv (ads + subs + bits + commerce), bandcamp.com (commerce +
 ↪  ad-free streaming).
- "non_profit_public": Foundation-funded, government-funded, .edu, .gov,
 ↪  hospital/clinic non-profits. Examples: wikipedia.org, nih.gov, usa.gov,
 ↪  mayoclinic.org, clevelandclinic.org, hopkinsmedicine.org. **NOT for commercial
 ↪  health publishers (webmd.com, healthline.com, drugs.com) -- those run on ad
 ↪  revenue; bucket those as monetization=ad_supported even though their
 ↪  content_type is reference_knowledge under the hospital/health exception.**
- "subscription_paywall": Recurring fee for access, whether metered or hard paywall.
 ↪  Examples: wsj.com, ft.com, netflix.com, disneyplus.com.
- "transactional": Revenue from on-site goods or services sales (shopping cart,
 ↪  booking flow, checkout). Examples: amazon.com, nike.com, walmart.com,
 ↪  expedia.com, doordash.com.

Monetization decision tree -- apply top-to-bottom and stop at the first match:
  1. Shopping cart / booking / checkout on homepage -> "transactional"
  2. Hard paywall (no free articles visible) -> "subscription_paywall"
  3. SaaS pricing tiers visible, free tier exists -> "freemium_saas"
  4. SaaS pricing tiers visible, no free tier -> "subscription_paywall"
  5. Foundation/non-profit/.gov/.edu/hospital footer with no ads or commerce ->
   ↪  "non_profit_public"
  6. Visible ads + free content (any "premium" tier is optional/secondary) ->
   ↪  "ad_supported" (default)
  7. Metered paywall OR substantial second stream comparable to the first ->
   ↪  "mixed_other"
</dimension>

</dimensions>
```

```
<decision_tree>
Apply these steps in order.

STEP 1 -- Evidence about the homepage.
  What does the homepage show -- navigation, headlines, product tiles, login/paywall
   ↪  copy, footer, about? If the domain is unrecognized, parked, empty, or
   ↪  redirects to an unrelated placeholder -> content_type = "other", confidence <=
   ↪  0.3, note it in rationale. STOP the content_type decision here.

STEP 2 -- Apply the adult filter first.
  If the homepage displays explicit adult content -> content_type = "adult". Do not
   ↪  reach for any other content_type. Continue to monetization normally.

STEP 3 -- What is the primary user activity on the site?
  3a. Looking up timeless reference, how-to, dictionary, or encyclopedic content ->
   ↪  "reference_knowledge".
  3b. Reading current news, original reporting, opinion, or editorial ->
   ↪  "news_journalism".
  3c. Reading peer-reviewed research, preprints, or academic databases ->
   ↪  "academic_research".
  3d. Browsing/posting on social, Q&A, threaded discussion, broadcast feeds,
   ↪  professional networks, or watching/listening on UGC/streaming media platforms
   ↪  (reddit, quora, stackoverflow, x/twitter, facebook, linkedin, youtube,
   ↪  instagram, tiktok, spotify, netflix, twitch) -> "social_media".
  3e. Buying from a single brand's own catalog -> "ecommerce_brand".
  3f. Buying across a multi-seller marketplace or broad-line retailer ->
   ↪  "ecommerce_marketplace".
  3g. Using a web app, productivity tool, or logging into a SaaS dashboard ->
   ↪  "tools_saas".
  3h. Visiting a code repo, package registry, or developer documentation ->
   ↪  "developer_technical".
  3i. Accessing a government/.gov/public-service portal or institutional authority
   ↪ -> "government_institutional".
  3j. Browsing gaming, fandom wiki, fan-fiction, or mod communities ->
   ↪  "entertainment_gaming".
  3k. Site's primary product is a frontier / general-purpose LLM chatbot or
   ↪  AI-search assistant (ChatGPT, Claude, Gemini, Perplexity, DeepSeek, Ollama,
   ↪  and similar) -> "other" (confidence >= 0.9, study-design quarantine).
  3l. Nothing above fits, or evidence is insufficient -> "other" (confidence < 0.6).

STEP 4 -- Resolve disambiguation with the rules below.

STEP 5 -- Pick monetization using the monetization decision tree above (shopping
 ↪  cart -> transactional -> hard paywall -> subscription_paywall -> SaaS tiers ->
 ↪  freemium_saas / subscription_paywall -> non-profit footer -> non_profit_public
 ↪ -> free + ads -> ad_supported -> multiple substantial streams -> mixed_other).

STEP 6 -- Calibrate confidence.
  0.9-1.0: homepage evidence is unambiguous (clear masthead, clear product catalog,
   ↪  clear login/paywall, clear .gov header).
  0.6-0.9: homepage evidence is suggestive but requires inference across multiple
   ↪  signals, or the domain is multi-mode and you had to pick a primary mode.
  <0.6: homepage was thin, ambiguous, behind auth, parked, or contradictory. Use
   ↪  content_type = "other" rather than guessing at a specific label.
</decision_tree>

<disambiguation_rules>
The most confusable boundaries, all resolved from observed homepage evidence.
```

```
tools_saas vs. other (frontier LLM and generative-AI exceptions):
  - Frontier general-purpose LLM chatbots and AI-search assistants whose primary
   ↪ product IS the conversational AI -- chatgpt.com, openai.com, claude.ai,
   ↪ gemini.google, perplexity.ai, pplx.ai, deepseek.com, ollama.com -> "other".
   ↪ These are the AI-search substitute being studied; quarantining them keeps
   ↪ tools_saas from absorbing the category the analysis treats as the substitute.
   ↪ Confidence >= 0.9 -- this is a study-design rule, not an evidence-thin guess.
  - Multi-LLM aggregator chatbots (chaton.ai, monica.im, iask.ai, toolbaz.com) ->
   ↪ "other". Same logic: primary product is the conversational AI, just routed
   ↪ through multiple back-end models.
  - Persona-based / companion / roleplay AI chatbots where users converse with
   ↪ AI-generated characters (character.ai, polybuzz.ai, spicychat.ai, replika.ai,
   ↪ dream.ai, woebothealth.com) -> "other". The homepage may show a UGC-style
   ↪ character gallery, but the primary user activity is still chatting with an
   ↪ LLM-generated agent -- they are AI-substitution-adjacent and belong with the
   ↪ other conversational chatbots, not with social_media or tools_saas.
  - AI-only single-purpose generators for image / video / voice / music / avatars
   ↪ (Midjourney, Runway, Suno, Udio, ElevenLabs, OpenArt, Leonardo, Ideogram,
   ↪ InVideo, Synthesia, HeyGen, Descript, Fliki, Pika, Lalal, Soundraw, Craiyon,
   ↪ AIVA, MagicStudio, AIStudios, MaxStudio) -> "other". Their entire product is
   ↪ the AI-generation step, so they are AI-substitution-adjacent and belong with
   ↪ the conversational chatbots, not with productivity SaaS.
  - Domain-specific tools that *embed* AI features but whose primary product is the
   ↪ underlying tool (Grammarly, Canva, Notion, Adobe Photoshop, SodaPDF, Zoom) ->
   ↪ "tools_saas".

ecommerce_brand vs. ecommerce_marketplace:
  - Single-brand catalog selling its own goods (homepage shows one brand, one
   ↪ product line) -> "ecommerce_brand". Examples: nike.com, apple.com,
   ↪ patagonia.com.
  - Multi-seller marketplace or broad-line retailer (homepage shows many brands,
   ↪ category navigation across unrelated product classes) ->
   ↪ "ecommerce_marketplace". Examples: amazon.com, ebay.com, walmart.com,
   ↪ etsy.com, target.com.

news_journalism vs. reference_knowledge:
  - Read as current/topical (articles with datelines, running headlines, opinion
   ↪ pieces, reviews) -> "news_journalism".
  - Looked up as timeless reference (dictionary entries, how-to articles, wiki
   ↪ pages, encyclopedic entries, health-condition pages) -> "reference_knowledge".
  - Health/finance reference with paid writers (webmd.com, healthline.com,
   ↪ investopedia.com) -> "reference_knowledge" when homepage is a reference index,
   ↪ not a news feed.

reference_knowledge vs. academic_research:
  - General reference or encyclopedic (wikipedia, britannica, wikihow) ->
   ↪ "reference_knowledge".
  - Peer-reviewed, preprint, or research-org primary sources (arxiv, nature, nih.gov
   ↪ research pages) -> "academic_research".
  - Edge: pubmed, sciencedirect -> "academic_research".

reference_knowledge vs. tools_saas (education exception):
```

```
  - LMS (Canvas/instructure, Schoology, Blackboard, Ellucian, Clever, ClassLink,
   ↪ PowerSchool, Jupitered), study tools (Chegg, Quizlet, Quizizz, Mathway,
   ↪ Symbolab, WolframAlpha, Desmos, GeoGebra, Turnitin, Duolingo), K-12
   ↪ adaptive-learning platforms (IXL, Edgenuity, DreamBox, Prodigy, Kahoot), tutor
   ↪ marketplaces (Preply, Wyzant, italki, Cambly, Outschool), exam proctoring
   ↪ (Honorlock, ProctorU, Respondus), MOOCs (Coursera, edX, Udemy, Cengage,
   ↪ McGraw-Hill), education publishers (Pearson), student career platforms
   ↪ (Handshake) -> "reference_knowledge". The user's purpose at these is learning
   ↪ or coursework, not productivity tooling -- even though the business model is
   ↪ SaaS.
  - General productivity SaaS / design tools (Canva, Notion, Grammarly, Zoom, Slack,
   ↪ Figma) -> "tools_saas".

social_media (text-first social and rich-media UGC/streaming, merged):
  - Both text-first social (reddit, quora, stackoverflow, twitter/x, facebook,
   ↪ linkedin) and rich-media UGC/streaming (youtube, tiktok, instagram, spotify,
   ↪ netflix, twitch) -> "social_media". The previous split between
   ↪ "social_platform" and "media_content" has been merged: both buckets describe
   ↪ consumption-on-platform UGC/streaming, and the analytical contrast that
   ↪ matters now is social_media vs. reference_knowledge / news_journalism, not
   ↪ text-vs-media within social_media.
  - Adult sites still go to "adult", not "social_media" (adult filter applied first
   ↪ per Step 2).

developer_technical vs. tools_saas:
  - Code repos, package registries, API docs, developer-specific reference (github,
   ↪ npmjs, pypi, MDN, docs.python.org) -> "developer_technical".
  - General SaaS dashboards, productivity web apps, design tools (canva, notion,
   ↪ grammarly, zoom) -> "tools_saas".

government_institutional vs. reference_knowledge:
  - .gov, .mil, .int, public-service portals, regulatory/tax agencies ->
   ↪ "government_institutional".
  - Encyclopedic or how-to content on a .org or .com that happens to be non-profit
   ↪ -> "reference_knowledge".
  - **Hospitals and health publishers (mayoclinic.org, clevelandclinic.org,
   ↪ webmd.com, healthline.com, drugs.com, medlineplus.gov, kidshealth.org) ->
   ↪ "reference_knowledge", NEVER government_institutional.** Their homepages
   ↪ present a health-reference index (symptoms, conditions, drugs, tests), not an
   ↪ institutional service portal. The .org TLD and clinic branding are not
   ↪ sufficient evidence for government_institutional. Research arms of NIH etc. ->
   ↪ "government_institutional" or "academic_research" depending on homepage.

ecommerce_marketplace vs. tools_saas:
  - Homepage shows shopping cart, product grid, prices -> "ecommerce_marketplace"
   ↪ (or "ecommerce_brand").
  - Homepage shows "sign in", tool interface, or SaaS pricing tiers -> "tools_saas".
  - **Shopify exception**: shopify.com (the merchant SaaS landing page) and
   ↪ *.myshopify.com (the per-merchant storefront subdomain) ->
   ↪ "ecommerce_marketplace". Even though shopify.com markets a SaaS to merchants,
   ↪ end-user panel traffic to shopify.com is dominated by buyer-side checkout
   ↪ flows, and *.myshopify.com tenant subdomains ARE storefronts. Bucketing these
   ↪ as marketplace destinations matches what panelists actually do at the URL.

entertainment_gaming vs. social_media:
  - Gaming storefront / fandom wiki / mod community / fan-fiction archive ->
   ↪ "entertainment_gaming".
```

```
  - Video/audio/image streaming or text/Q&A social on a general consumer platform
    ↪ (including Twitch live-streaming and gaming-adjacent text discussion on
    ↪ twitter/reddit) -> "social_media".

entertainment_gaming vs. reference_knowledge (game-specific wikis/databases):
  - General-purpose reference, encyclopedia, or how-to (wikipedia, britannica,
    ↪ wikihow, dictionary.com) -> "reference_knowledge".
  - **Wiki, encyclopedia, or database whose entire scope is a single game,
    ↪ franchise, or game series (runescape.wiki, stardewvalleywiki.com,
    ↪ pokemondb.net, bulbagarden.net, paradoxwikis.com, minecraft.wiki) ->
    ↪ "entertainment_gaming", NOT reference_knowledge.** The site is
    ↪ reference-shaped but the topic is play, not general knowledge. The "*.wiki" /
    ↪ "*db.net" / "*pedia.net" pattern does not save it for reference_knowledge if
    ↪ the scope is a single game.
</disambiguation_rules>

<examples>
14 reference classifications covering the 12 categories and the major split cases
 ↪ (hospital exception, Shopify exception, generative-AI quarantine, education
 ↪ exception, mixed_other for metered paywalls). Use them as calibration anchors.
 ↪ Each rationale is phrased as though you had just viewed the homepage.

[
  {"domain": "reddit.com", "content_type": "social_media", "monetization":
    ↪ "ad_supported", "confidence": 0.95, "rationale": "Homepage is a feed of
    ↪ subreddit post cards with vote counts and comment counts -- canonical threaded
    ↪ discussion. Bucketed under the merged social_media category (text-first social
    ↪ and rich-media UGC/streaming combined). Ads dominate the feed; Reddit Premium
    ↪ and awards are secondary upsells, so this stays ad_supported, not
    ↪ mixed_other."},
  {"domain": "chatgpt.com", "content_type": "other", "monetization":
    ↪ "freemium_saas", "confidence": 0.99, "rationale": "Homepage is the ChatGPT
    ↪ conversational interface -- a frontier general-purpose LLM chatbot.
    ↪ Quarantined into 'other' per study design (the AI-search substitute being
    ↪ studied, not productivity tooling). Freemium with paid Plus/Pro tiers."},
  {"domain": "quora.com", "content_type": "social_media", "monetization":
    ↪ "ad_supported", "confidence": 0.95, "rationale": "Homepage shows a threaded
    ↪ Q&A feed with answer cards and 'Follow' buttons -- text-first social bucketed
    ↪ under social_media. Ads dominate the homepage; Quora+ exists but is a
    ↪ secondary upsell, so this stays ad_supported not mixed_other."},
  {"domain": "nike.com", "content_type": "ecommerce_brand", "monetization":
    ↪ "transactional", "confidence": 0.99, "rationale": "Homepage is Nike's own
    ↪ product catalog (shoes, apparel) with shopping-cart CTA. Single-brand DTC
    ↪ storefront, not a multi-seller marketplace. Revenue comes directly from
    ↪ on-site transactions."},
  {"domain": "amazon.com", "content_type": "ecommerce_marketplace", "monetization":
    ↪ "transactional", "confidence": 0.99, "rationale": "Homepage shows category
    ↪ navigation across many unrelated product classes, multiple third-party sellers
    ↪ on listing pages. Canonical multi-seller marketplace; revenue from on-site
    ↪ purchases."},
  {"domain": "pornhub.com", "content_type": "adult", "monetization": "ad_supported",
    ↪ "confidence": 0.99, "rationale": "Homepage displays explicit adult video
    ↪ thumbnails and an adult-content warning gate. Quarantined into the adult
    ↪ bucket per study design (policy-filtered from AI output). Ads dominate;
    ↪ premium tier is a secondary upsell, so ad_supported not mixed_other."},
```

```
  {"domain": "wikipedia.org", "content_type": "reference_knowledge", "monetization":
   ↪  "non_profit_public", "confidence": 0.99, "rationale": "Homepage is the
   ↪  Wikipedia language portal with a search box for encyclopedic articles.
   ↪  Consumed as timeless reference. Non-profit (Wikimedia Foundation) funded via
   ↪  donations, no ads or paywall visible."},
  {"domain": "mayoclinic.org", "content_type": "reference_knowledge",
   ↪  "monetization": "non_profit_public", "confidence": 0.95, "rationale":
   ↪  "Homepage is a health-reference index: search bar for conditions, large tiles
   ↪  for 'symptoms', 'diseases', 'tests', 'drugs & supplements'. Although Mayo
   ↪  Clinic is a hospital institution, the *website* is consumed as health
   ↪  reference, not as an institutional service portal -- hospital exception,
   ↪  bucket with reference_knowledge, NOT government_institutional. Non-profit
   ↪  funding."},
  {"domain": "arxiv.org", "content_type": "academic_research", "monetization":
   ↪  "non_profit_public", "confidence": 0.99, "rationale": "Homepage is the
   ↪  Cornell-operated preprint server listing subject categories and new-submission
   ↪  feeds. Primary-source academic research, not general reference. Non-profit,
   ↪  institution-funded."},
  {"domain": "github.com", "content_type": "developer_technical", "monetization":
   ↪  "freemium_saas", "confidence": 0.99, "rationale": "Homepage shows
   ↪  code-repository hosting with 'Sign up' and 'Explore repositories' CTAs.
   ↪  Primary user activity is interacting with code, not buying or reading news.
   ↪  Freemium: free public repos, paid tiers for private/org usage."},
  {"domain": "youtube.com", "content_type": "social_media", "monetization":
   ↪  "ad_supported", "confidence": 0.99, "rationale": "Homepage is a grid of video
   ↪  thumbnails; primary activity is watching video. Rich-media UGC platform --
   ↪  bucketed under the merged social_media category alongside text-first social.
   ↪  Ads (pre-rolls, mid-rolls, banners) are the dominant homepage revenue signal;
   ↪  YouTube Premium is a secondary upsell, so this stays ad_supported."},
  {"domain": "nytimes.com", "content_type": "news_journalism", "monetization":
   ↪  "mixed_other", "confidence": 0.95, "rationale": "Homepage is a current-news
   ↪  feed with original reporting and opinion. Metered paywall -- a large share of
   ↪  readers hit it on the homepage itself -- alongside visible ads. Two
   ↪  substantial streams (ads + subscription) -> mixed_other, NOT ad_supported."},
  {"domain": "shopify.com", "content_type": "ecommerce_marketplace", "monetization":
   ↪  "transactional", "confidence": 0.95, "rationale": "Although shopify.com
   ↪  markets a SaaS to merchants, end-user panel traffic to shopify.com and to
   ↪  *.myshopify.com tenant subdomains is dominated by buyer-side
   ↪  storefront/checkout flows -- the URL acts as a multi-merchant marketplace from
   ↪  the consumer's perspective. Bucketed with ecommerce_marketplace per the
   ↪  Shopify exception."},
  {"domain": "suno.com", "content_type": "other", "monetization": "freemium_saas",
   ↪  "confidence": 0.95, "rationale": "Homepage is an AI-only music generator:
   ↪  prompt -> generated song. Single-purpose generative-AI product
   ↪  (image/video/voice/music) -- quarantined into 'other' alongside frontier LLM
   ↪  chatbots because its displacement dynamics belong with AI substitution, not
   ↪  with productivity SaaS."},
  {"domain": "edgenuity.com", "content_type": "reference_knowledge", "monetization":
   ↪  "freemium_saas", "confidence": 0.95, "rationale": "Homepage is a K-12 online
   ↪  courseware platform with 'student login' and curriculum tiles. The user's
   ↪  purpose is coursework / learning, not productivity tooling -- falls under the
   ↪  education exception, bucketed with reference_knowledge alongside Canvas, IXL,
   ↪  Preply, etc."}
]
</examples>

<batch_instructions>
```

```
You will receive a list of N domains in <domains_to_classify>. Classify ALL N
↪  domains in the SAME ORDER as the input list. Do not skip any. Do not add domains
↪  that are not in the list.

The web_search tool is available. Your primary signal is training memory -- for
↪  well-known domains (major news sites, major marketplaces, major social
↪  platforms, major SaaS, .gov / .edu, large UGC platforms) label from memory
↪  directly. Call web_search to fetch the live homepage when (a) you don't
↪  recognize the domain, (b) the domain has multiple plausible content_types and
↪  you need to see the homepage to pick the primary mode, (c) the domain is
↪  adult-adjacent and you might be steered away from the correct label, or (d) the
↪  domain is newer or lower-traffic with sparse training data.

Your confidence must reflect what you actually know (from memory or from the fetched
↪  page), not what you would have liked to know.
</batch_instructions>

<output_format>
Respond with ONLY a JSON array (no markdown fencing, no commentary). Each element
↪  must be a JSON object with exactly these keys:

  "domain"               (string) -- the domain name, exactly as in the input
  "content_type"         (string) -- one of (alphabetical): academic_research,
   ↪  adult, developer_technical, ecommerce_brand, ecommerce_marketplace,
   ↪  entertainment_gaming, government_institutional, news_journalism, other,
   ↪  reference_knowledge, social_media, tools_saas
  "monetization"         (string) -- one of (alphabetical): ad_supported,
   ↪  freemium_saas, mixed_other, non_profit_public, subscription_paywall,
   ↪  transactional
  "confidence"           (float)  -- 0.0 to 1.0
                          0.9-1.0 = clear, unambiguous evidence (memory or homepage)
                          0.6-0.9 = inferred from multiple signals or multi-mode
                           ↪  domain
                          <0.6   = thin/ambiguous/parked -- use content_type="other"
                           ↪  instead
  "rationale"            (string) -- one to two sentences citing the evidence behind
   ↪  your label

Do NOT include a "web_search" field -- that is recorded from the API tool-call
↪  events, not from your output.
</output_format>

<domains_to_classify>
- {{DOMAIN_1}}
- {{DOMAIN_2}}
- {{DOMAIN_3}}
- {{DOMAIN_4}}
- {{DOMAIN_5}}
</domains_to_classify>

<rules_reminder>
REMINDER -- re-anchor on the rules before producing output:

VALID VALUES (alphabetical -- order is fixed for reproducibility, not a ranking):
- content_type: academic_research, adult, developer_technical, ecommerce_brand,
↪  ecommerce_marketplace, entertainment_gaming, government_institutional,
↪  news_journalism, other, reference_knowledge, social_media, tools_saas
- monetization: ad_supported, freemium_saas, mixed_other, non_profit_public,
↪  subscription_paywall, transactional
```

```
- confidence: float 0.0-1.0 (NOT high/medium/low)

DECISION TREE (abbreviated):
  Determine what the homepage shows. If you cannot -> content_type="other",
  ↪ confidence <= 0.3.
  Explicit adult content -> content_type="adult".
  Frontier general-purpose LLM chatbot / AI-search assistant (chatgpt, claude,
  ↪ gemini, perplexity, deepseek, ollama, etc.) -> other (confidence >= 0.9,
  ↪ study-design quarantine)
  Timeless reference/how-to -> reference_knowledge
  Current news/opinion -> news_journalism
  Peer-reviewed/preprint/research -> academic_research
  Social / Q&A / threaded-discussion / broadcast-feed / professional-network OR
  ↪ rich-media UGC/streaming (reddit, quora, stackoverflow, x/twitter, facebook,
  ↪ linkedin, youtube, instagram, tiktok, spotify, netflix, twitch) ->
  ↪ social_media
  Single-brand DTC catalog -> ecommerce_brand
  Multi-seller marketplace or broad-line retailer -> ecommerce_marketplace
  Web app / productivity / SaaS (NOT a frontier LLM chatbot) -> tools_saas
  Code repo / package registry / dev docs -> developer_technical
  .gov / public-service / institutional -> government_institutional
  Gaming / fandom wiki / fan-fiction / mods -> entertainment_gaming
  Cannot determine -> other (confidence < 0.6)

KEY CONTRASTS:
  reddit.com / quora / stackoverflow / twitter / facebook / youtube / tiktok /
  ↪ instagram / spotify / netflix / twitch -> social_media (the previous
  ↪ social_platform vs. media_content split has been merged)
  chatgpt/claude/gemini/perplexity/deepseek/ollama -> other (frontier LLM chatbot
  ↪ quarantine); chaton.ai/monica.im/iask.ai/toolbaz.com -> other (multi-LLM
  ↪ aggregators); character.ai/polybuzz.ai/spicychat.ai/replika.ai/dream.ai ->
  ↪ other (persona / companion AI chatbots -- UGC character gallery is incidental,
  ↪ primary activity is chat with LLM)
  midjourney/runway/suno/udio/elevenlabs/openart/leonardo/ideogram/invideo/synthesi
  ↪ a/heygen/descript/fliki/pika -> other (AI-only single-purpose generators --
  ↪ quarantined alongside chatbots)
  grammarly/canva/notion/adobe-photoshop/zoom -> tools_saas (embeds AI features,
  ↪ primary product is the underlying tool)
  nike/apple/patagonia -> ecommerce_brand; amazon/ebay/walmart/etsy ->
  ↪ ecommerce_marketplace
  shopify.com / *.myshopify.com -> ecommerce_marketplace (Shopify exception -- panel
  ↪ traffic is buyer-side storefront/checkout, not merchant-SaaS admin)
  edgenuity/ixl/preply/wyzant/prodigy/kahoot/honorlock/jupitered/dreambox ->
  ↪ reference_knowledge (K-12 / tutoring / proctoring under the education
  ↪ exception)
  mayoclinic/clevelandclinic/hopkinsmedicine/webmd/healthline/drugs.com/medlineplus
  ↪ /kidshealth -> reference_knowledge (hospital + health-publisher exception, NOT
  ↪ government_institutional). Monetization split: non-profit hospitals
  ↪ (mayoclinic, clevelandclinic, hopkinsmedicine) -> non_profit_public;
  ↪ commercial health publishers (webmd, healthline, drugs.com) -> ad_supported.
  runescape.wiki / stardewvalleywiki.com / pokemondb.net / bulbagarden.net /
  ↪ paradoxwikis.com / minecraft.wiki -> entertainment_gaming (single-game-scope
  ↪ wikis/dbs go to entertainment_gaming, NOT reference_knowledge, even when
  ↪ domain is "*.wiki" or "*db.net")
  pornhub/xvideos -> adult (do NOT route through social_media)

MONETIZATION DEFAULT RULE:
  Free site + visible ads + optional/secondary "premium" tier (Quora+, YouTube
  ↪ Premium, "remove ads", Reddit Premium) -> ad_supported.
```

```
  Reserve mixed_other for sites with TWO roughly equal revenue streams visible on
   ↪ the homepage -- metered paywall (nytimes), platform with substantial commerce
   ↪ alongside ads (reddit, twitch).

Every label should be supported by evidence cited in the rationale. Do not guess
 ↪ from domain name or TLD alone. Do NOT include a web_search field in the output
 ↪ -- that is recorded from the API tool-call events.
</rules_reminder>

Return JSON array only. Same order as input. Classify ALL domains. No commentary.
```

References

Callaway B, Sant'Anna PHC (2021) Difference-in-differences with multiple time periods. *J. Econometrics* 225(2):200–230, URL http://dx.doi.org/10.1016/j.jeconom.2020.12.001.

Chen X, Sant'Anna PHC, Xie H (2025) Efficient difference-in-differences and event study estimators. URL http://dx.doi.org/10.48550/arXiv.2506.17729, arXiv preprint arXiv:2506.17729.

Padilla N, Lam HT, Lambrecht A, Hollenbeck B (2026) The impact of early LLM adoption on online user behavior. *Management Science* URL https://ssrn.com/abstract=5393256, forthcoming.