Selective Participation in the AI Data Commons

Kai Zhu

Bocconi University

Abstract

AI systems rely on web content, but access is increasingly governed by voluntary crawler restrictions. We study this access regime among news and media publishers, where content quality matters for factual and civic information. We provide large-scale evidence that publisher access is sharply selective: high-factual outlets restrict AI crawlers at 47%, versus 8% among low-factual outlets. Related composition results show that misinformation sources and outlets at the political extremes become relatively more accessible. A further finding is use-specific selectivity: higher-quality publishers are more likely to restrict training while preserving search-like access, consistent with training offering weaker attribution and referral benefits than search. The evidence shows that publishers are not rejecting AI access wholesale—they distinguish uses. Our findings imply that AI data governance should move beyond binary opt-out toward use-specific access design.

Keywords: AI data governance, AI data commons, AI crawlers, `robots.txt`, selective access, news publishers, training versus search, misinformation

1 Introduction

Generative AI systems do not draw on the web as a fixed, neutral, fully open data commons. They draw on a changing access environment in which content producers can declare which AI uses they tolerate. Whether that environment retains reliable content matters: what compliant crawlers can access may shape downstream answers, summaries, and search results. Yet producers receive no systematic compensation when their work is crawled for AI. This paper asks how they respond to that bargain — who withdraws, and from which uses.

We study this question in the media and news sector. News publishers produce fact-checked, time-sensitive, and editorially curated content, so their participation shapes the quality composition of declared AI access. Their incentives are two-sided. Incorporating content into model training can appropriate its value without attribution or referral. Search- and discovery-oriented AI access can also substitute for publisher traffic, but it can preserve indexing, citation, and audience flow. The same publisher may therefore rationally restrict training use while allowing search-like use.

Our empirical object is declared access: whether a site's `robots.txt` disallows named AI crawlers. The protocol is public, standardized, and crawler-specific — OpenAI's GPTBot, which may gather training data, is distinct from OAI-SearchBot, which supports search[1] — but economically blunt: a publisher can allow or deny a crawler, with no prices, contracts, or compensation. Declared restrictions are not realized exclusion from model training: compliance is voluntary, though major AI crawlers generally respect directives (Kim et al., 2025). They are a standardized statement of which named crawlers a site chooses to restrict.

Our evidence combines three measurement files built on the universe of Media Bias/Fact Check (MBFC) rated news domains: a live site-level cross-section in February 2026, a monthly panel of directives from HTTP Archive captures running from 2022 to 2026, and a separate live scrape that reads restrictions for 23 named crawlers per site. We mea-

[1] OpenAI's publisher documentation distinguishes the functions of GPTBot and OAI-SearchBot.

sure publisher quality from several independent angles—editorial ratings of factual reporting and credibility, and web-presence measures of publisher prominence—so the same patterns emerge across genuinely different aspects of quality.

We document three key results. First, declared restrictions are quality-selective. High-factual outlets restrict at least one named AI crawler at 47%, compared with 8% among low-factual outlets, and the same gradient appears for credibility ratings, a content-quality index, and a web-presence-based domain-prominence index: more credible and more established publishers are more likely to restrict. Two further consequences follow. The *misinformation asymmetry*: MBFC-flagged questionable sources restrict at only 6.1%, against 34.8% among unflagged sites, making flagged content more represented among sites declaring no restriction. And *ideological sorting*: center outlets restrict most, while outlets at the political poles remain the most accessible.

Second, the selectivity is use-specific. Restricting search while allowing training is nearly nonexistent: 8 sites, about 0.1% of those with training/search outcomes; by contrast, 8.8% restrict training only and 26.6% restrict both. We make the margin explicit with a direct estimand — the gap between a publisher's training and search restrictions: higher quality predicts a larger gap in every specification we estimate. The same pattern holds within OpenAI's bot family, holding vendor identity fixed: among the same publishers, 54.2% of high-factual outlets block GPTBot while 35.8% block OAI-SearchBot, and the within-publisher gap rises with quality.

Third, the timing of restrictions is both use-specific and aligned with two events that changed the options publishers faced. After OpenAI's GPTBot announcement in August 2023 made named-crawler opt-out a salient, documented option, the high-minus-low MBFC factual-rating gap in training restrictions opened first — 36.4 percentage points by December 2023, against 19.4 for search — and the training-over-search differential reached about 18.0 points during the 2023–2024 salience window. As Cloudflare's one-click AI-blocking tool appeared in July 2024 and substantially lowered the technical barrier to blocking, search

restrictions caught up and the differential narrowed to 5.0 points by February 2026.

The paper first contributes to emerging work on AI data access and the AI data commons. Existing research documents declining consented access and greater restriction by reputable sites (Longpre et al., 2024; Steinacker-Olsztyn et al., 2025; Bouchaud and Ramaciotti, 2025). We show that selective participation has two layers: selection by quality and selection by use. Higher-quality publishers restrict more, while publishers that differentiate uses disproportionately restrict training and preserve search-like access. This use-specific layer is our more focused contribution: it makes participation an economically meaningful choice across uses rather than a binary decision. Our framework links both layers to quality-dependent net costs and divergent training/search benefits.

The paper's second contribution is to digital economics and access governance. Selective opt-out can make observed digital data unrepresentative (Goldfarb and Tucker, 2019; Miklós-Thal et al., 2024), and algorithmic systems inherit the gaps (Tucker, 2024). We show that participation concerns not only who permits access but which uses they permit, connecting the evidence to publisher strategy (Zhao and Berman, 2026; Abeillon et al., 2024), blocking agency (Liu et al., 2025), and compensated-access institutions (Archer et al., 2026; Peukert, 2025). Because directives differentiate training from search, access composition under use-specific rules is a natural object for market design.

The remainder of the paper is organized as follows. Section 2 presents the framework, Section 3 describes the data, Section 4 presents the results, and Section 5 discusses their implications.

2 Theory

A compact framework organizes the evidence; Appendix A15 states the formal version. A publisher with quality q decides whether to allow crawler use $u \in \{T, S\}$, where T denotes training-oriented access and S search- or discovery-oriented access. Allowing use u carries

an appropriation or substitution cost $C_u(q)$ and a visibility or referral benefit $B_u(q)$. Let κ denote the implementation hurdle for expressing a crawler-specific restriction. The publisher restricts use u when the net cost of allowing access exceeds that hurdle:

$$R_u(q) = \mathbf{1}\{C_u(q) - B_u(q) > \kappa\}.$$

If the net cost of allowing crawler access rises with quality over the observed range, higher-quality publishers restrict more. This requires no informational asymmetry about quality; it requires only that opt-out be selective with respect to quality.

The training/search distinction adds the use-specific prediction. Training access can incorporate content into model weights and capabilities, typically without attribution or referral traffic, so its benefit term is small for most publishers. Search-like access can also substitute for visits, but it preserves discovery, indexing, and citation, so its benefit term is larger. The local condition

$$\frac{\partial}{\partial q}\left[C_T(q) - B_T(q)\right] > \frac{\partial}{\partial q}\left[C_S(q) - B_S(q)\right] \geq 0$$

implies that the latent net-access cost rises faster for training than for search. If the training net cost is also weakly higher over the observed range, the restriction thresholds are ordered: a region of publishers restricts training while allowing search, and search-only restriction is rare. We summarize the use-specific margin with the direct estimand

$$D \;=\; R_T - R_S \;\in\; \{-1, 0, 1\},$$

the training-minus-search restriction gap. $D = 1$ means a publisher restricts training but not search; $D = -1$ the reverse; $D = 0$ means the two uses are treated identically. The framework predicts a positive quality gradient in D over the interior range represented in

the data and that $D = -1$ is rare.[2]

The framework also disciplines what timing evidence can show. OpenAI's GPTBot announcement (August 2023) made named-crawler opt-out a salient, documented option; high-quality publishers, with more at stake, could be expected to move first. Cloudflare's one-click AI-blocking tool (July 2024) lowered implementation costs and applied broadly across AI crawlers.[3] A common reduction in that hurdle predicts more restriction for both uses, but does not determine whether the training-only region widens or narrows. The later narrowing of the observed training-minus-search factual-rating differential reflects search restrictions catching up to training restrictions. Crawler-blind tooling and rising search-substitution concerns are consistent with that pattern, but the timing evidence does not identify either mechanism or either event's causal effect.

Table 1 distinguishes the empirical objects associated with three access regimes.

Table 1: Access regimes and empirical objects

Access regime	Publisher choice	Empirical implication	What the evidence establishes
Binary opt-out	Allow or restrict crawler access	Restrictions are quality-selective if net access costs rise with quality	Higher-quality media restrict far more often (Section 4.1)
Use-specific access	Choose training and search access separately	D rises with quality; search-only restrictions are rare	Search-only restrictions are nearly absent; D rises with all quality measures (Section 4.2)
Compensated access	Allow use under contract or price	Participation may rise if compensation offsets appropriation costs	Not estimated; observed states identify the access margin a contract could target (Table 5)

Notes. The first two rows yield testable predictions; the third motivates a descriptive accounting exercise, not a welfare estimate. The paper does not estimate prices, reservation values, or licensing effects.

[2]The deterministic threshold model produces an intermediate training-only region, not a globally increasing binary D. With smooth heterogeneity the observable probability gap can rise locally, but if both restriction probabilities eventually approach one it must return toward zero; Appendix A15 states the exact conditions.

[3]Cloudflare's July 3, 2024 announcement describes the default one-click control for blocking AI bots.

The empirical analysis tests two predictions and examines one descriptive temporal pattern:

P1: Quality-selective declared restrictions. More credible, more factual, and more established publishers are more likely to restrict AI crawler access.

P2: Use-specific access. The training-minus-search gap D rises with quality, and search-only restrictions are rare.

T1: Descriptive timing pattern. Did the training-access gradient rise earlier during the 2023–2024 salience window, followed by a later rise in search restrictions?

3 Data

Our evidence combines three measurement files built on the universe of MBFC-rated media domains; Appendix A1.1 accounts for every sample and denominator. The first is a live site-level scrape of each domain's `robots.txt` in February 2026, covering 9,611 domains and supplying the primary declared-restriction outcome. The second is a monthly panel of directives from HTTP Archive captures, June 2022 to February 2026: 7,002 sites and 272,458 site-months. The panel is smaller because a site enters only when the archive's monthly crawls capture its `robots.txt`; it is unbalanced, with each site contributing the months in which it is captured. The panel supplies the aggregate training/search outcomes and monthly timing evidence. The third file is a separate February 2026 live scrape at the site-bot level, read for 23 named crawlers per successfully fetched site and used for the paired-bot and site-bot fixed-effects comparisons.

Declared access in `robots.txt` is a useful but incomplete measure of publisher participation. Declarations are public, standardized, crawler-specific, and archived monthly, permitting comparisons across agents and time. But compliance is voluntary; publishers also use paywalls, IP blocking, bot detection, and private licensing that we do not observe; and a

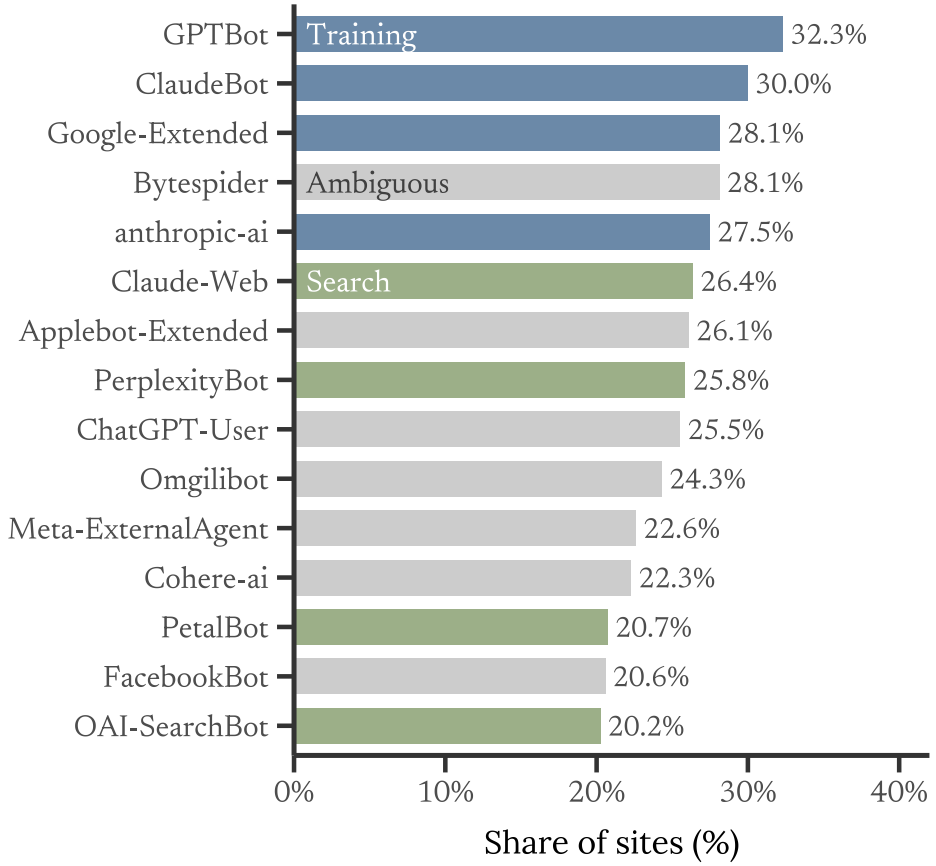
directive does not establish exclusion from any training corpus. We therefore interpret it as an observable access policy toward named crawlers, not a complete measure of preferences or realized AI use.

The primary cross-sectional outcome equals one if a site restricts any of 131 tracked AI-related crawler tokens. The retained live-scrape file contains this site-level aggregate, which includes CCBot, Common Crawl's long-running general-purpose crawler.[4] For use-specific outcomes we classify crawlers by function.[5] The clean taxonomy is a narrow, curated classification by predominant documented function—four training and five search/discovery tokens—rather than a claim that every crawler has one immutable use. Thirteen mixed or insufficiently documented tokens enter only the broad-taxonomy robustness; Appendix A6 documents the provider evidence and all 23 crawlers, and Appendix A10 shows both the sensitivity to broader labels and the leave-one-crawler-out results. Figure 1 previews the raw landscape: training-oriented crawlers dominate the top of the ranking—a first, model-free glimpse of the use-specificity that Section 4.2 formalizes.

[4]Because the 131-token file does not retain its bot components, we do not imply that CCBot can be subtracted from it. In the separate 7,256-site scrape with complete indicators for 23 crawlers, the any-restriction factual gradient is 0.163 both with and without CCBot; only nine sites change classification (Appendix A3).

[5]The monthly panel's aggregate outcomes BlockTrain and BlockSearch follow the HTTP Archive's labels (three training crawlers: GPTBot, Google-Extended, ClaudeBot; two search crawlers: OAI-SearchBot, PerplexityBot) and exclude CCBot.

Figure 1: Training-oriented crawlers are the most-restricted crawlers.



Note. Share of the 7,256 sites with successful February 2026 `robots.txt` fetches restricting each plotted crawler; lower-rate rows are omitted for readability. In-bar labels mark the crawler classes of the clean taxonomy; training-oriented crawlers occupy the top of the ranking, led by GPTBot at 32.3%. CCBot is omitted from this display and from the use-specific baskets; its exact inclusion sensitivity is in Appendix A3.

We organize the publisher measures into two families. The editorial/content family begins with MBFC factual reporting (six ordinal steps) and credibility (three steps);[6] a binary Wikipedia-page match adds encyclopedic standing. Their first principal component forms the content-quality index. The separate web-presence/establishment family uses $\log(1 +$ Majestic referring subnets), Open PageRank, and inverse-log Tranco rank, $-\log(1 + \text{rank})$,

[6]MBFC describes its source ratings as human-coded assessments. Its factual-reporting score combines fact-check record, sourcing, transparency, and one-sidedness/omission; its political-orientation score combines economic ideology, social ideology, straight-news balance, and editorial bias; and its credibility rating combines factual reporting, bias, traffic/longevity, and press-freedom adjustments. We use these ratings as structured third-party source metadata and triangulate them with web-presence measures; see Appendix A2.

to form the domain-prominence index. Referring subnets measure network reach rather than Majestic's proprietary Trust Flow score.

The headline models use four operationalizations: the two raw MBFC ratings and the two composite indices. The first three are related rather than independent replications: factual reporting and credibility correlate 0.90, and both dominate the content-quality index. The domain-prominence index is more distinct, correlating 0.24–0.25 with the editorial ratings and 0.26 with the content-quality index (exact transforms, loadings, scaling, and correlations in Appendix A2). Agreement across the two families is therefore informative without separating editorial quality from publisher establishment. Our preferred interpretation is that more credible and more established publishers restrict more.

Table 2: Publisher measures: scales, construction, and interpretation

Measure and scale	Construction and interpretation
DIRECT MEASURES: MBFC EDITORIAL RATINGS	
MBFC factual reporting *Ordinal, 1–6*	Human-coded editorial assessment of factual performance
MBFC credibility *Ordinal, 1–3*	Human-coded assessment of broader source credibility
CONSTRUCTED MEASURES: COMPOSITE INDICES	
Content-quality index *Standardized*	First principal component of factual reporting, credibility, and Wikipedia-page match
Domain-prominence index *Standardized*	First principal component of referring subnets, PageRank, and inverse Tranco rank; captures web reach and establishment

Notes. The table maps the constructs used in the headline regressions. Exact provenance, coverage, transformations, missing-data rules, loadings, and correlations appear in Appendix A2.

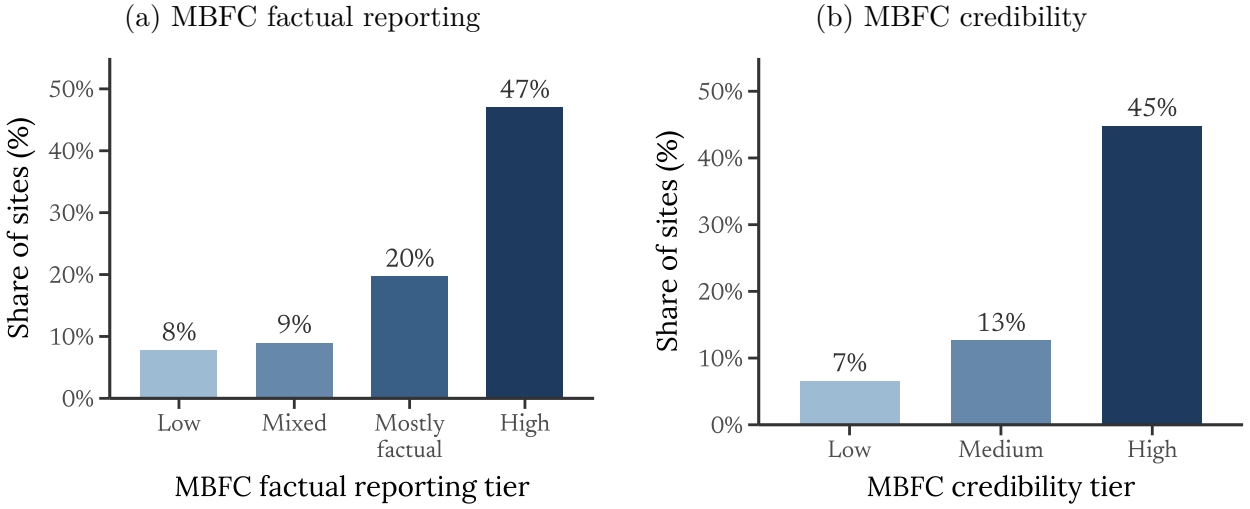
4 Results

4.1 Result 1: Quality-selective declared restrictions

Quality gradient. Figure 2 presents the paper's first descriptive fact. Declared restrictions rise monotonically with factual reporting (panel (a)), from 8% among low-factual sites

to 47% among high-factual outlets, and with credibility (panel (b)), from 7% to 45%. Regression estimates confirm the gradient for the four operationalizations in Table 2: 0.134 per factual step, 0.197 per credibility step, 0.176 per SD of content quality, and 0.078 per SD of domain prominence (category fixed effects; Table 3). On their common sample, both composite indices remain positive jointly: 0.197 for content quality and 0.032 for domain prominence (Appendix A2). This does not isolate editorial quality from publisher establishment, but shows independent associations with both measured dimensions. The gradient also holds in the balanced panel of 4,778 sites observed in all 45 months (coefficients 0.041–0.157; Appendix A5). Finally, among HTTP-200 fetches the factual coefficient rises from 0.134 to 0.164, so the full-sample retrieval convention is conservative (Appendix A4).

Figure 2: Higher-quality media restrict AI crawlers far more often.



Note. Outcome is an indicator for restricting at least one of 131 tracked AI-related crawler tokens in the February 2026 live cross-section; the retained aggregate includes CCBot. $N = 9,611$ MBFC-rated media sites; sites without a retrievable `robots.txt` are coded as declaring no restriction. Exact CCBot sensitivity in the separate 23-bot scrape is in Appendix A3.

Table 3: The restriction gradient holds across four publisher measures

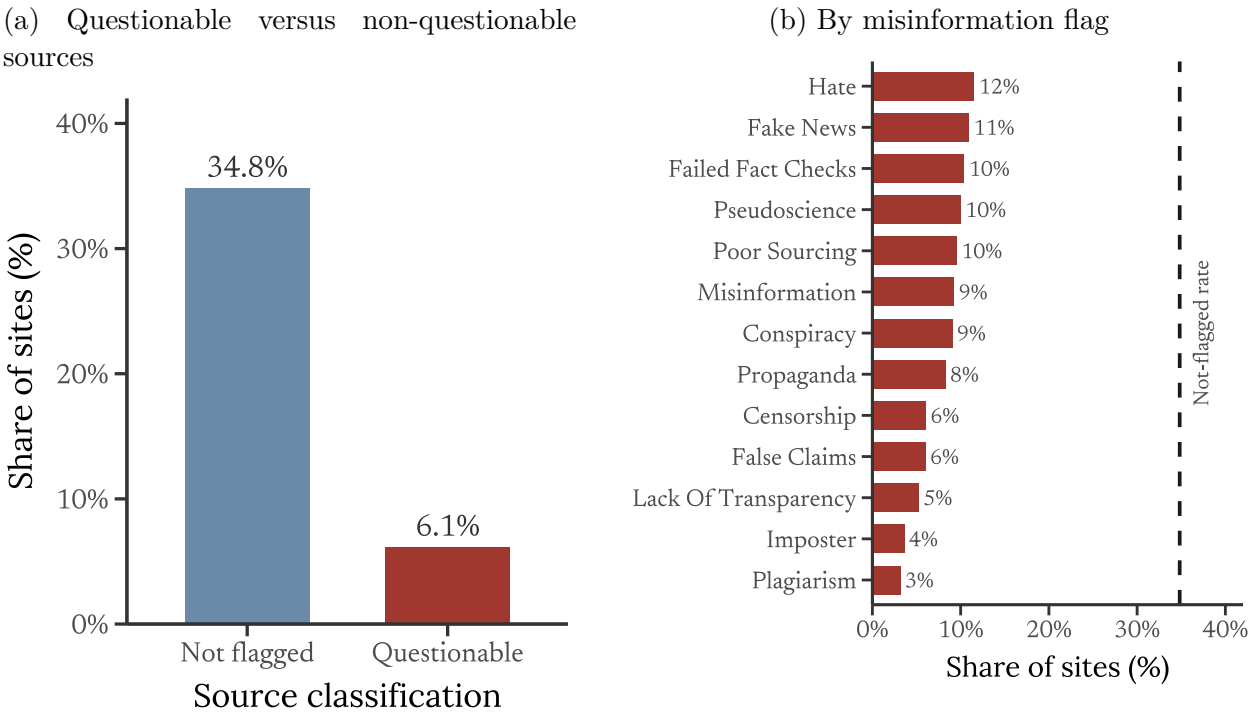
Quality measure	Estimate	SE	N	R2
MBFC factual reporting	0.134***	(0.003)	9,611	0.136
MBFC credibility	0.197***	(0.004)	9,602	0.151
Content-quality index	0.176***	(0.004)	9,611	0.151
Domain-prominence index	0.078***	(0.006)	5,782	0.025

Notes: Linear probability models. Outcome is an indicator for restricting at least one of the 131 tracked AI-related crawler tokens in the February 2026 live cross-section (N = 9,611); this retained site-level aggregate includes CCBot. All models include category fixed effects with HC1 robust standard errors. MBFC coefficients are per one ordinal step; index coefficients are on the full-cross-section SD scale. The separate 23-bot scrape permits exact CCBot inclusion/exclusion and is reported in Appendix A3. Significance: * p<0.10, ** p<0.05, *** p<0.01.

The misinformation asymmetry. The gradient does not just rank publishers; it changes the composition of sites declaring no restriction. The starkest pattern is what we term the *misinformation asymmetry*: MBFC-flagged questionable sources—sites identified as promoting conspiracy theories, fake news, propaganda, or pseudoscience—restrict AI crawlers at dramatically lower rates than sites without an MBFC questionable-source flag. Panel (a) of Figure 3 quantifies the gap: 6.1% among the 1,984 flagged sources versus 34.8% among unflagged sites. Panel (b) disaggregates by flag type: rates range from 3% to 12%, all far below the unflagged rate. The comparison group is not limited to high-credibility outlets. The raw asymmetry largely reflects quality composition: flagged sources are concentrated at low quality, where restrictions are rare.[7]

[7]Within the fixed 5,782-site sample with both quality indices, the unadjusted questionable-source coefficient of -0.327 reverses sign and becomes positive (0.064) once both indices and category fixed effects are added (Appendix A13.1).

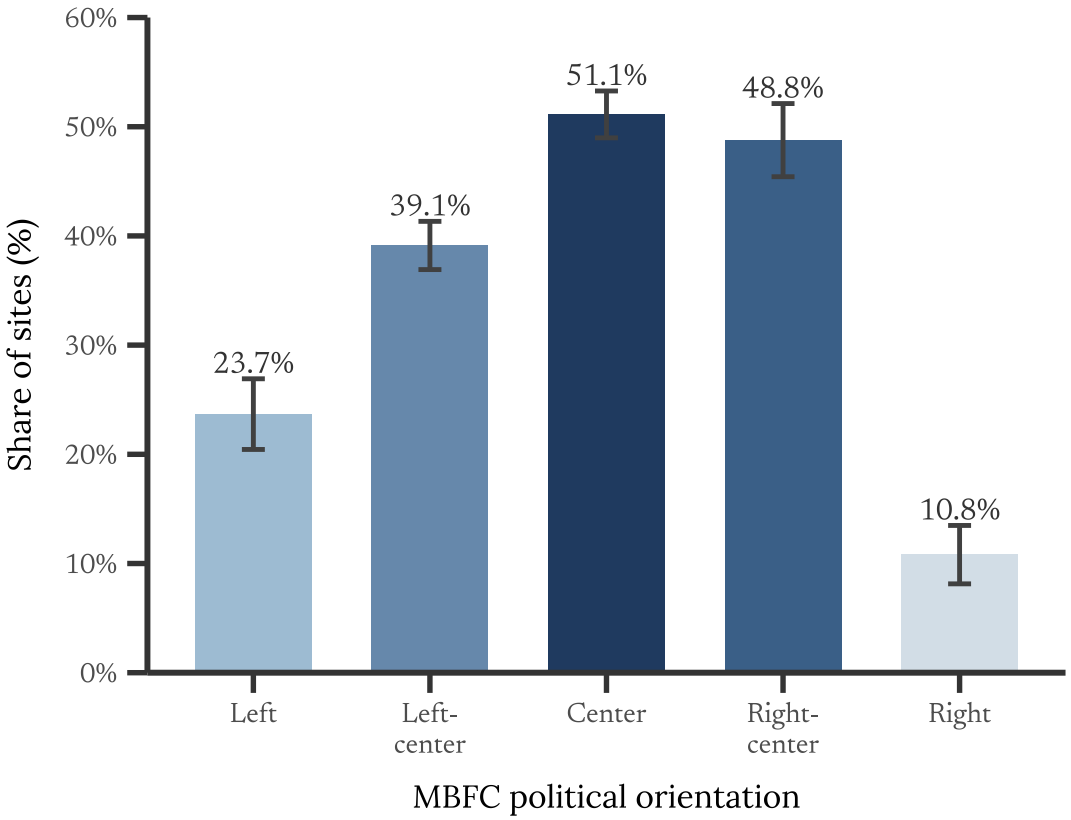
Figure 3: Questionable sources declare fewer AI-crawler restrictions.



Note. Outcome is the February 2026 indicator for restricting at least one of 131 tracked AI-related crawler tokens, including CCBot; full 9,611-site cross-section. Questionable sources are the 1,984 sites carrying MBFC flags. Panel (b) shows flag types with at least 30 sites (a site can carry several flags); the dashed line marks the rate among sites without a questionable-source flag. Uniformly low restriction across flags tilts the declared-accessible pool toward flagged content.

Ideological sorting. Political orientation shows a sharply non-linear cross-sectional pattern. Among the 5,995 sites with left-to-right MBFC ratings (Figure 4), center outlets declare restrictions at the highest rate (51.1%), followed by right-center (48.8%) and left-center (39.1%); left outlets restrict at 23.7% and right outlets at 10.8%. Thus centrist reporting is less represented, and both poles more represented, in the observed declared-accessible layer. Quality explains part of the pattern, but left and right gaps persist after controlling for both quality indices and site category (Appendix A13.1). These are conditional associations, not changes over time or direct effects of ideology.

Figure 4: Declared AI-crawler restrictions peak among centrist outlets.



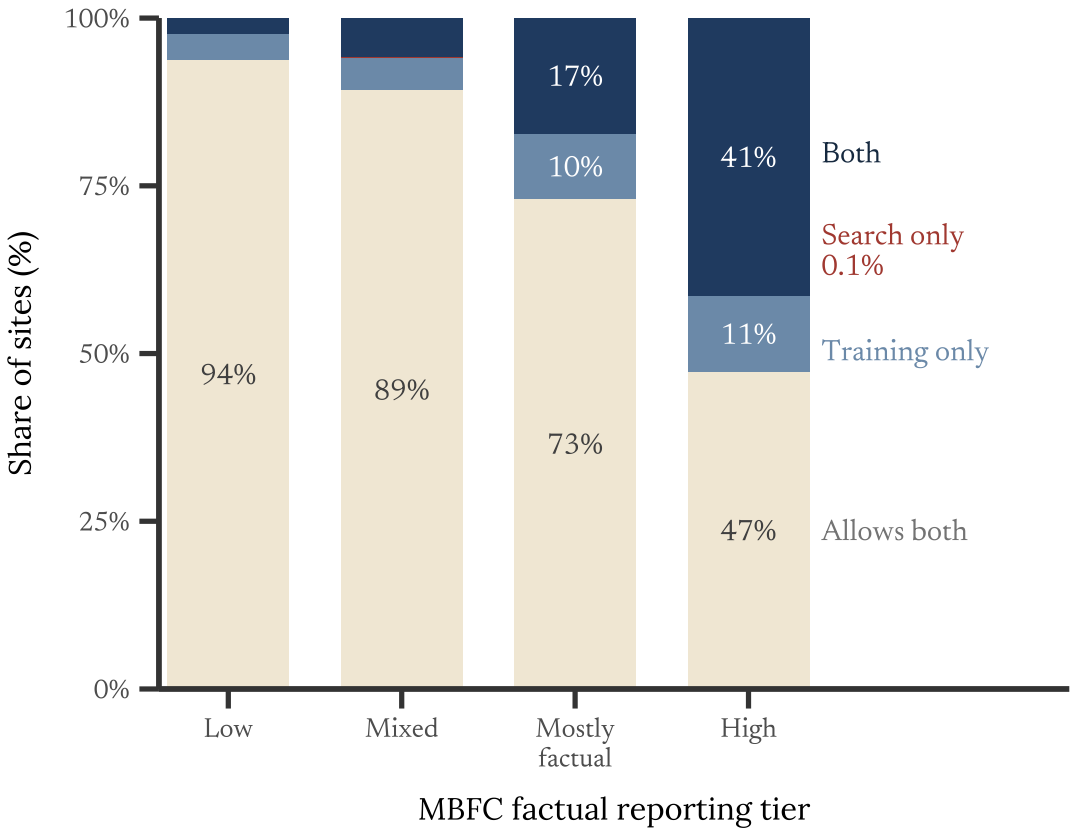
Note. Bars are February 2026 rates of restricting at least one of 131 tracked AI-related crawler tokens, including CCBot; vertical lines are 95% confidence intervals. Sample is 5,995 sites with left-to-right MBFC ratings. Quality controls attenuate but do not eliminate the descriptive pole gaps (Appendix A13.1).

4.2 Result 2: Selective participation is use-specific

Access states: publishers distinguish uses. If publishers objected to AI bots only as a class, training and search restrictions would move together; if they distinguish appropriation from discovery, restrictions should concentrate on training. Figure 5 classifies sites with training/search outcomes into four states by factual tier, and the tier patterns are stark. Among high-factual publishers, 11.2% restrict training only and 41.4% restrict both, while search-only restrictions sit at 0.1%; among low-factual publishers, 93.9% allow both. The conditional probabilities sharpen the point (Appendix A7): conditional on restricting

anything, only 0.3% of sites are search-only — 8 sites in total — and conditional on allowing search, 12.0% of sites restrict training, rising to 19.1% among high-factual publishers. A decomposition of the quality gradient in any-restriction shows it is built entirely from training-only and joint states. When publishers differentiate, they almost always restrict training at least as much as search.

Figure 5: Search-only restrictions barely exist at any quality level.

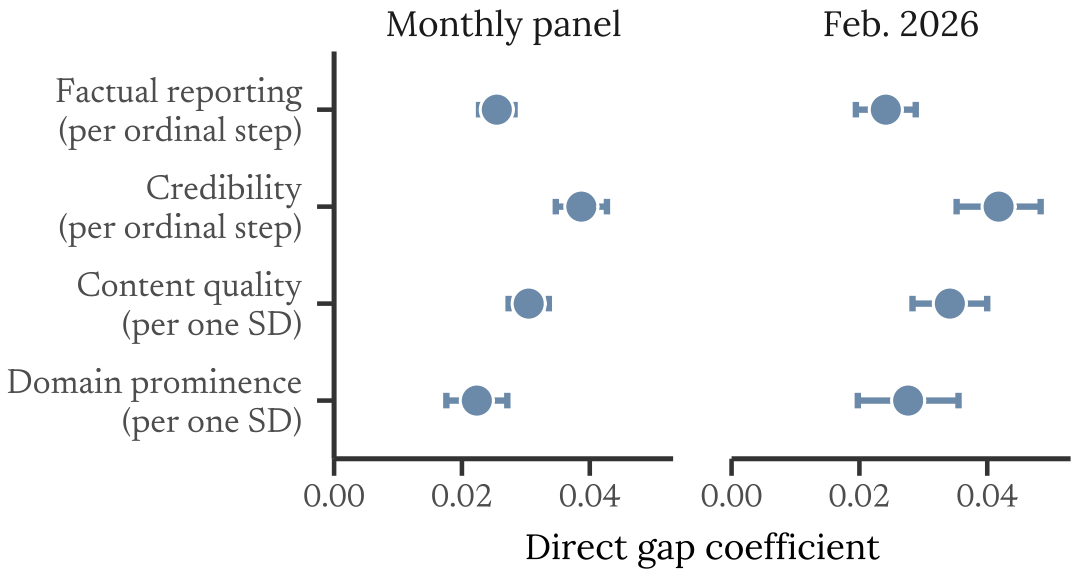


Note. Sample is the 7,002-site cross-section with nonmissing training/search outcomes in February 2026 (CCBot excluded). Percentages are within factual tier. Exact counts, conditional probabilities, and the gradient decomposition are in Appendix A7.

The direct training-minus-search gap. The core evidence makes the use-specific margin an explicit estimand: $D_{it} = \text{BlockTrain}_{it} - \text{BlockSearch}_{it}$. Regressing D on quality asks whether the same publisher restricts training more than search. Because D is a within-site difference, site characteristics that shift both outcomes equally cancel; use-specific awareness

or crawler salience need not. Figure 6 plots the direct D-regression estimates. In the monthly panel, with category and month fixed effects, one factual step predicts a 0.025 larger gap (training coefficient 0.093 versus search coefficient 0.068); one SD of content quality predicts 0.030 (training 0.112 versus search 0.081); credibility and domain prominence give 0.039 and 0.022. The February 2026 cross-section repeats the pattern (0.024 per factual step; 0.034 per SD of content quality). All eight direct-gap estimates are positive with $p < 0.01$ (Appendix A8).[8]

Figure 6: Higher quality predicts restricting training over search.



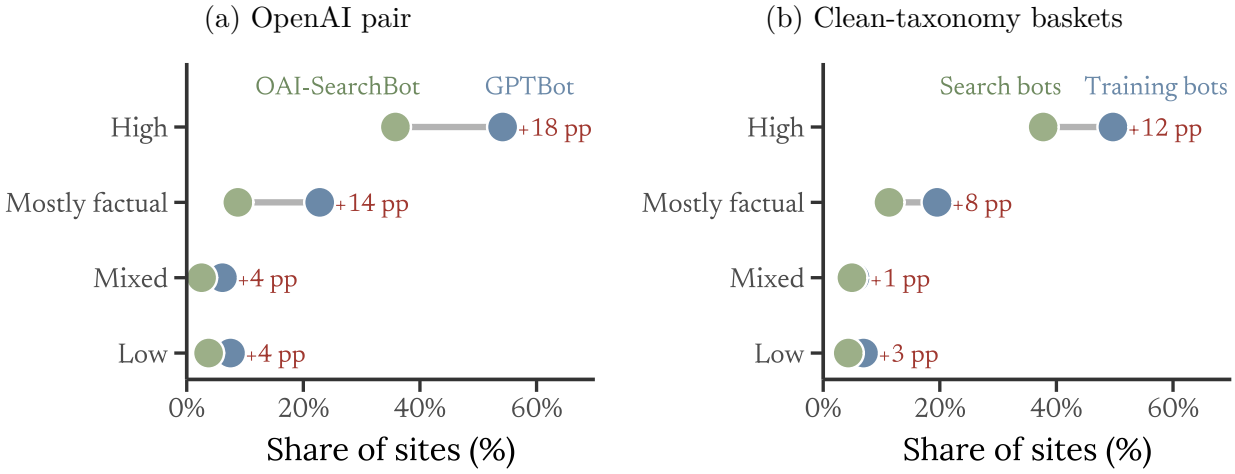
Note. Points are coefficients from direct linear models of D = BlockTrain − BlockSearch on one publisher measure at a time; bars show 95% confidence intervals for those direct-gap coefficients. The monthly panel uses 272,458 site-months from 7,002 sites, category and month fixed effects, and site-clustered standard errors. The February 2026 cross-section uses 7,002 sites, category fixed effects, and HC1 standard errors. MBFC coefficients are per ordinal step. Content quality and domain prominence denote the standardized content-quality and domain-prominence indices. Full estimates are in Appendix A8.

Paired bots: same publisher, same vendor. The aggregate baskets bundle bot function with vendor identity and salience, so we go to the bot level, where the cleanest comparison is OpenAI's explicitly named pair. Among the same 7,256 publishers, 54.2% of

[8] On a common sample the D coefficient equals the training coefficient minus the search coefficient by OLS linearity; the gap model supplies the standard error and test for that difference, plus a site-level outcome.

high-factual outlets block GPTBot (training) while 35.8% block OAI-SearchBot (search), a within-publisher gap of 18.4 percentage points that shrinks to about 4 points among low-factual outlets (panel (a) of Figure 7). In regressions of the within-site pair difference on quality, the OpenAI gap rises 0.053 per factual step and 0.070 per SD of content quality; the clean-taxonomy pair difference rises 0.037 and 0.051 (panel (b) of Figure 7). The corresponding OpenAI site-bot fixed-effects estimates are 0.053 and 0.071: the factual estimates agree at displayed precision, while the content-quality estimates are 0.070 and 0.071 under the two specifications (Appendix A9). Their close alignment shows that the result is not an artifact of how the paired averages are formed. The Anthropic pair (ClaudeBot versus Claude-Web) shows the same sign with much smaller magnitudes (0.004 per factual step), consistent with that pair's murkier functional split, and we treat it only as robustness.

Figure 7: Within-publisher training/search gaps rise with quality.



Note. Successful February 2026 `robots.txt` fetches (7,256 sites). Points are shares of publishers restricting each training and search crawler by factual tier; annotations are the within-tier training-minus-search gaps in percentage points. Panel (a): GPTBot versus OAI-SearchBot, the cleanest documented functional pair. Panel (b): means over the narrow clean-taxonomy baskets. The smaller and less cleanly documented Anthropic pair (ClaudeBot versus Claude-Web), taxonomy provenance, and site-bot fixed-effects models are in Appendix A9 and Appendix A6.

Interpretation. Three layers of evidence—site-level access states, site-month gap models, and within-vendor pairs—show that publishers use crawler-specific directives to differentiate training from search-like access. The pattern is consistent with concern about appropriation

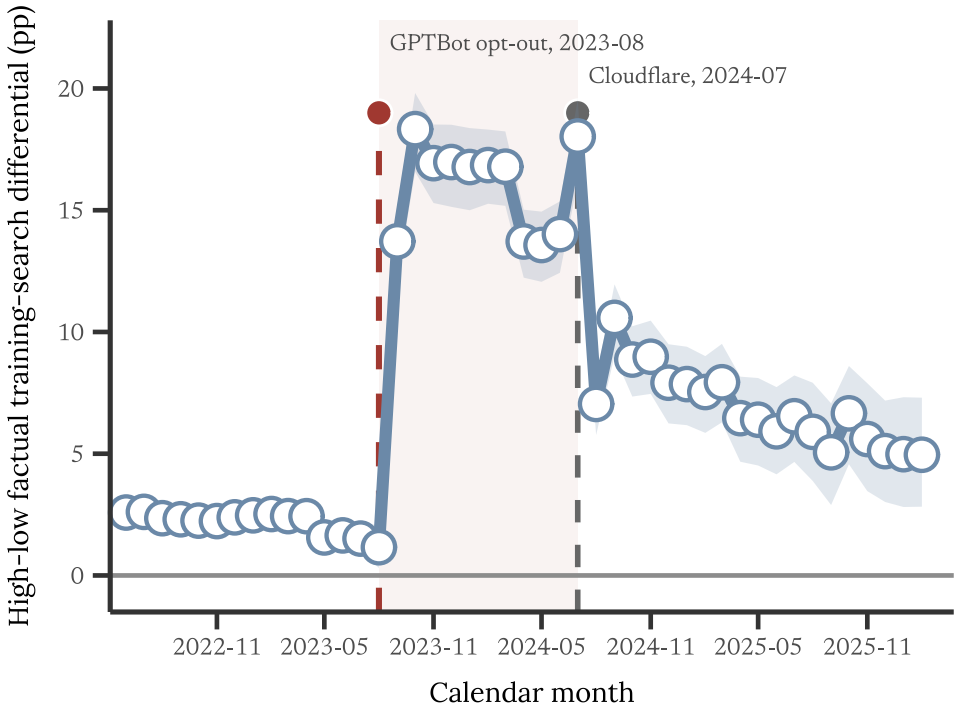
alongside benefits from discovery, but does not reveal willingness to pay or contract.[9] It instead establishes an observed use-specific margin that licensing and pay-per-crawl research can investigate.

4.3 Result 3: The timing of restrictions is use-specific

The use differential over time. Figure 8 tracks the timing of use-specific selectivity directly. For visual transparency it plots the difference between the high-minus-low MBFC factual-rating gap in training restrictions and the corresponding factual-rating gap in search restrictions, with a site-level bootstrap band. Month-specific regressions of D on continuous factual reporting and content quality show the same broad trajectory (Appendix A12.1). Panel composition does not generate the pattern: among the 4,778 sites observed in all 45 months, the monthly differential correlates 0.988 with the full series and differs by only 0.8 percentage points on average. The balanced-panel differential is 18.8 points in December 2023, 18.6 in July 2024, and 4.2 in February 2026 (Appendix A5). Before August 2023 the differential was small but not zero (about 2 percentage points; the average pre-period monthly factual gradient is 0.008 per ordinal step), which we report as-is: high-factual publishers differentiated slightly even before AI crawling became salient. The differential then rose after OpenAI's GPTBot announcement and stayed between 13.6 and 18.3 points through mid-2024. Its later narrowing coincides with Cloudflare's July 2024 launch of a one-click, crawler-blind AI-blocking tool and reflects search restrictions moving toward the training restrictions high-factual publishers had already adopted. This sequence is consistent with lower technical barriers broadening restriction, but the clustered timing does not identify that mechanism. By February 2026 the differential had narrowed to 5.0 points.

[9] Bot labels bundle function, salience, and rollout timing. The gradient survives dropping GPTBot entirely (0.032 per factual step) and remains positive among sites exhibiting at least one named-bot restriction (Appendix A10). These checks reduce, but cannot eliminate, explanations based on awareness or taxonomy.

Figure 8: The training–search factual-rating differential rose and later narrowed.



Note. Monthly panel of 7,002 sites. The series is the difference between the high-minus-low MBFC factual-rating gap in training restrictions and the corresponding factual-rating gap in search restrictions; it does not combine all four publisher measures. The band is a 95% interval from a site-level bootstrap (500 replications), and the shaded window marks the salience period. The red pin marks August 2023 (named-crawler documentation); the grey pin marks July 2024 (one-click blocking tools). Month-specific continuous factual-reporting and content-quality gradients of D are in Appendix A12.1. The evidence is descriptive: ChatGPT diffusion, litigation, and blocking tools were clustered in the same window, so no single event is identified as the cause.

Levels behind the differential. The underlying training and search restriction rates explain why the differential narrows (Table 4; period averages in Appendix A12). Among high-factual outlets, training restrictions averaged 2.4% before August 2023 and 44.6% after; search restrictions moved from 0.2% to 32.0%. Among low-factual outlets the post-period averages were only 4.5% and 2.0%. The resulting high-minus-low training gap reached 36.4 points in December 2023 when the search gap was 19.4; by July 2024 the two gaps were 42.3 and 24.3; by February 2026 the search gap had largely caught up (45.3 versus 40.4). Thus the narrowing does not mean that training restrictions fell: training restrictions rose earlier, while search restrictions later spread alongside one-click, crawler-blind blocking. The sequence

motivates instruments that can express the observed use distinction, without attributing the change to any one tool.

Table 4: The training factual-rating gap rose earlier than the search factual-rating gap

Month	Training gap	Search gap	Train - search
2022-06	2.6	0.0	2.6
2023-07	2.3	0.7	1.5
2023-08	1.9	0.8	1.2
2023-12	36.4	19.4	17.0
2024-07	42.3	24.3	18.0
2026-02	45.3	40.4	5.0

Notes: Monthly panel of 7,002 sites; outcomes are the aggregate training- and search-basket restriction indicators. Entries are percentage-point differences between high- and low-MBFC-factual-rating site shares in the stated month. The final column is computed from the underlying unrounded gaps and can differ by 0.1 point from subtracting the displayed components. The table describes the observed sequence and does not identify any event's causal effect.

5 Discussion

We document three key facts about participation in the AI data commons. Declared AI-crawler restrictions among media publishers are quality-selective; the selectivity is use-specific, with higher-quality publishers restricting training over search on every measured margin; and the use-specific pattern rose around the GPTBot announcement before narrowing as search restrictions caught up. Publishers are not rejecting AI wholesale: with a coarse instrument, they differentiate access by use.

These facts produce a compositional consequence analogous to adverse selection, although the mechanism differs from the canonical hidden-quality setting (Akerlof, 1970). Classic adverse selection concerns hidden quality and which goods trade; here the evidence concerns selection into declared access, not hidden information. Higher-quality publishers disproportionately declare restrictions, especially for training, so reliable journalism is underrepresented in the declared-accessible pool. We do not observe realized model-training or retrieval

composition.

Table 5 makes the observed composition differences concrete. Among all 7,002 sites, 56.3% are high-factual; under search rules the share is 45.0%; under training rules it is 41.3%, while the questionable share rises from 15.1% to 22.0%. The 614 training-only restrictors are 72.0% high-factual and almost never questionable. Their directives preserve search while restricting training: this is the observed margin a compensated-training offer could target, but the declarations do not show acceptance or price.

Table 5: Observed content composition differs across declared-access margins

Accessible pool	Sites	High-factual	Low-factual	Questionable	Left/right pole
All sites	7,002	56.3%	12.1%	15.1%	19.1%
Search-accessible	5,134	45.0%	16.1%	20.1%	25.2%
Training-accessible	4,528	41.3%	17.6%	22.0%	27.5%
Licensing margin	614	72.0%	5.2%	5.7%	13.3%

Notes: Descriptive recalculation of observed February 2026 access states among the 7,002 sites with training/search outcomes; it is not an estimate of licensing take-up, prices, willingness to contract, or welfare. Search-accessible: declares no restriction on search crawlers; Training-accessible: no restriction on training crawlers; Licensing margin: restricts training while allowing search — the observed access margin to which a compensated-training offer could be directed. Left/right-pole shares are computed among bias-rated sites in each pool (N = 5,224; 3,431; 2,895; 542, respectively).

Two directions follow. First, test whether access composition shifts AI answers, citations, or factuality. Second, use the observed differentiation to motivate measurement of reservation values and participation elasticities, and evaluation of pay-per-crawl, collective licensing, and use-tiered contracts.

The broader lesson is that access governance shapes participation. Under the current regime, declared restrictions are concentrated among credible and established publishers and, within publisher, on training use. Future access institutions should therefore distinguish uses and evaluate whether compensation can preserve high-quality participation rather than assume that binary opt-out is sufficient.

Funding and Competing Interests

The authors declare no competing interests. This blinded review copy omits any funding-source details that could identify the authors. Any applicable funding disclosure can be provided separately to the journal through the submission system and final title page.

References

Abeillon, Diane, Tobias Haese, Maximilian Kaiser, Christian Peukert, and Kevin Staub. 2024. "Strategic Behavior and AI Training Data." Working paper, HEC Lausanne.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84 (3): 488–500.

Archer, Richard, Soheil Ghili, and Nima Haghpanah. 2026. "Pay-Per-Crawl Pricing for AI: The LM-Tree Agent." *arXiv preprint*.

Bouchaud, Paul, and Pedro Ramaciotti. 2025. "Web Crawler Restrictions, AI Training Datasets & Political Biases." *arXiv/HAL preprint*.

Goldfarb, Avi, and Catherine Tucker. 2019. "Digital Economics." *Journal of Economic Literature* 57 (1): 3–43.

Kim, Taein, Karstan Bock, Claire Luo, Amanda Liswood, Chloe Poroslay, and Emily Wenger. 2025. "Scrapers Selectively Respect robots.txt Directives: Evidence from a Large-Scale Empirical Study." *Proceedings of the ACM Internet Measurement Conference*.

Liu, Enze, Elisa Luo, Shawn Shan, Geoffrey M. Voelker, Ben Y. Zhao, and Stefan Savage. 2025. "Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers." *Proceedings of the ACM Internet Measurement Conference*.

Longpre, Shayne, et al. 2024. "Consent in Crisis: The Rapid Decline of the AI Data Commons." *NeurIPS Datasets and Benchmarks Track*.

Miklós-Thal, Jeanine, Avi Goldfarb, Avery Haviv, and Catherine Tucker. 2024. "Digital Hermits." *Marketing Science* 43 (4): 697–708.

Peukert, Christian. 2025. "The Economics of Copyright and AI: Empirical Evidence and Optimal Policy." *European Parliament Policy Department Analysis* PE 778.859.

Steinacker-Olsztyn, Nicolas, Devashish Gosain, and Ha Dao. 2025. "Is Misinformation More Open? A Study of robots.txt Gatekeeping on the Web." Accepted at *The ACM Web Conference 2026*.

Tucker, Catherine. 2024. "Algorithmic Exclusion: The Fragility of Algorithms to Sparse and Missing Data." *Brookings Institution Report*.

Zhao, Hangcheng, and Ron Berman. 2026. "Strategic Response of News Publishers to Generative AI." *SSRN working paper*.

Supplementary Appendix

Selective Participation in the AI Data Commons

Contents

A1 Samples, outcomes, and coefficient interpretation **2**
A1.1 Sample flow . . . 2
A1.2 Coefficient interpretation . . . 4
A2 Quality-measure documentation **4**
A3 CCBot robustness **8**
A4 Fetch-status robustness **9**
A5 Balanced-panel robustness **10**
A6 Crawler taxonomy **11**
A7 Access states and conditional probabilities **12**
A8 Direct training-minus-search gap models **14**
A9 Paired-bot and site-bot fixed-effects evidence **16**
A10 Taxonomy and salience robustness **19**
A11 Heterogeneity in the use-specific gap **21**
A12 Timing detail **22**
A12.1 Event-time gaps by use . . . 23
A13 Secondary composition implications **27**
A13.1 Conditional regressions . . . 28
A13.2 Are the secondary patterns use-specific? . . . 29
A14 Composition scenarios: construction detail **30**
A15 Formal framework **31**

Purpose and roadmap

This appendix documents the measurement, robustness checks, and detailed estimates behind the manuscript. A Python build reads only the frozen project data and generates every table, figure-data file, and headline-number macro; R scripts render the final graphics. One generated macro file supplies headline values to all three documents, preventing numerical drift.

The sections follow the paper's three results. For Result 1 they document samples, outcomes, quality measures, CCBot treatment, and panel composition. For Result 2—the paper's core—they report access states, direct gap models, paired-bot and fixed-effects estimates, and taxonomy and salience checks. For Result 3 they provide period and event-time detail. The appendix closes with secondary composition results and the formal framework.

A1 Samples, outcomes, and coefficient interpretation

A1.1 Sample flow

Table 1 accounts for every sample and denominator. The February 2026 live cross-section records whether each site's `robots.txt` restricts any of 131 tracked AI-related crawler tokens, including CCBot, Common Crawl's long-running general-purpose crawler; only that site-level aggregate is retained. Unretrievable files (9,611 minus 6,304 sites) are coded as declaring no restriction. The unbalanced monthly panel comes from HTTP Archive captures and supplies the training/search outcomes; its February 2026 cross-section is the same 7,002 sites. A separate successful-fetch scrape supplies complete indicators for 23 named crawlers and permits crawler-specific checks.

Table 1: Samples and denominators used in the paper

Sample	Sites	Observations	Used for	Notes
Full MBFC media cross-section	9,611	9,611	Primary quality gradient; secondary implications	February 2026 live scrape; any of 131 tracked AI-related crawler tokens, including CCBot; 3,307 sites without a retrievable robots.txt are coded as declaring no restriction
Domain-prominence estimation sample	5,782	–	Domain-prominence regressions; conditional models	At least two of log referring subnets, Open PageRank, and inverse-log Tranco rank are observed
Questionable-flagged subset	1,984	–	Misinformation implication	MBFC questionable-source flag
Political-orientation subsample	5,995	–	Political-orientation implication	MBFC bias rated left to right; non-political MBFC labels excluded
Political + both quality indices	4,650	–	Quality-controlled political regressions	Political-orientation subsample with both composite indices observed
Monthly HTTP Archive panel	7,002	272,458	Timing; training/search outcomes; panel gap models	June 2022 to February 2026 (45 months); unbalanced. Credibility complete case: 6,999 sites/272,331 observations; domain-prominence complete case: 5,690/239,534
Training/search cross-section	7,002	–	Access states; conditionals; composition; cross-sectional gap models	Panel sites with February 2026 training/search outcomes (identical to panel site set); credibility complete case: 6,999; domain-prominence complete case: 5,690
Balanced panel subset	4,778	215,010	Quality-gradient and use-specific timing robustness	Sites observed in all 45 months; domain-prominence complete case: 4,504 sites/202,680 observations
Per-bot live scrape	7,256	166,888	Site-bot fixed effects; paired bot gaps; taxonomy robustness	HTTP-200 fetches in the February 2026 per-bot scrape; 23 crawlers per site; a cross-section, not a site-bot-month panel

Notes: Every table and figure reports one of these denominators. The cross-section outcome and the per-bot file are two separate February 2026 scrape runs with different fetch stacks and retry behavior, hence their different success counts (6,304 versus 7,256). The monthly training/search outcomes come from HTTP Archive captures, which is why the training/search sample equals the panel site set.

A1.2 Coefficient interpretation

Outcomes are declared `robots.txt` restrictions, not realized exclusion from AI systems. Linear-probability coefficients read as percentage-point changes in the probability of declaring a restriction. MBFC factual reporting (six ordinal steps) and MBFC credibility (three ordinal steps) enter per ordinal step. Both PCA scores are standardized in the full cross-section. After the site-level scores are mapped into the monthly panel, they are standardized once over the full panel's site-month observations; cross-sectional, site-bot, and balanced-panel subsamples retain their parent-file scale rather than being re-standardized within each regression. Cross-sectional models include site-category fixed effects with HC1 robust standard errors; panel models include category and month fixed effects with standard errors clustered by site; site-bot models include site and bot fixed effects with standard errors clustered by site. The direct use-specific outcome is $D = \text{BlockTrain} - \text{BlockSearch} \in \{-1, 0, 1\}$.

A2 Quality-measure documentation

Table 2 defines each measure, its scale, construction, and coverage. MBFC's public methodology describes its current source ratings as human evaluations: reviewers inspect a minimum of 10 headlines and 5 full stories per source, use manual and keyword searches and comparisons with verified sources, and conduct the scoring themselves rather than relying on automated scoring.[1] For sources reviewed or re-reviewed under the current system, factual reporting is a weighted score over failed fact checks (40%), sourcing (25%), transparency (25%), and one-sidedness/omission (10%). The left-right orientation score is a weighted composite of economic ideology (35%), social ideology (35%), straight-news balance (15%), and editorial bias (15%). MBFC's credibility rating combines factual reporting, bias, traffic/longevity, and country press-freedom adjustments into high, medium, or low credibility categories. MBFC also describes political bias as inherently subjective and its methodology

[1] Media Bias/Fact Check, "Methodology," last updated February 5, 2026, https://mediabiasfactcheck.com/methodology/.

as its own rather than a universally validated scientific scale. We therefore treat the ratings as structured third-party source metadata, not ground truth, and use the independently constructed web-presence index as a triangulation check.

Two features matter for interpretation. First, MBFC ratings are a February 2026 snapshot applied retrospectively to the panel, so timing analyses assume ratings are stable over the 45-month window. Second, the four operationalizations are not interchangeable: the editorial measures (factual reporting, credibility) correlate 0.90 with each other and 0.98 with the content-quality index that they dominate, while the domain-prominence index—built from log Majestic referring subnets, Open PageRank, and inverse-log Tranco rank—correlates only 0.24–0.25 with the editorial ratings and 0.26 with the content-quality index (Table 3). Log referring subnets measure network reach rather than Majestic's proprietary Trust Flow score; inverse-log Tranco rank is $-\log(1+\text{rank})$, so higher values indicate greater web prominence. The measures therefore triangulate a common gradient across related editorial measures and a distinct prominence family, while still not separating editorial quality from established-publisher status. The manuscript's preferred reading is accordingly that more credible and more established publishers restrict more.

Table 2: Quality-measure definitions and coverage

Measure	Scale	Construction	Coverage
MBFC factual reporting	Ordinal 1-6 (very-low to very-high)	Media Bias/Fact Check editorial rating; February 2026 snapshot applied to all periods	9,611
MBFC credibility	Ordinal 1-3 (low to high)	Media Bias/Fact Check credibility rating; February 2026 snapshot	9,602
Content-quality index	Full-cross-section mean 0, SD 1	First principal component of standardized MBFC factual reporting, MBFC credibility, and a binary Wikipedia-page match; median imputation when at least two of the three inputs are non-missing, otherwise missing	9,611
Domain-prominence index	Full-cross-section mean 0, SD 1	First principal component of standardized log(1 + Majestic referring subnets), Open PageRank, and -log(1 + Tranco rank); median imputation when at least two inputs are non-missing, otherwise missing	5,782

Notes: Coverage counts are non-missing sites in the 9,611-site February 2026 cross-section. MBFC ratings are a February 2026 snapshot; the panel analysis therefore assumes ratings are stable over the 45-month window. Both PCA scores are standardized in the full cross-section. When mapped into the panel they are re-standardized once over the full 272,458 site-month file; robustness subsamples retain their parent-file scale rather than being re-standardized row by row. Regression tables report MBFC coefficients per ordinal step and identify the relevant index scale.

Table 3: Pairwise correlations among the four publisher measures

Measure	Factual	Credibility	Content Q	Domain prom.
MBFC factual	1.00	0.90	0.98	0.24
MBFC credibility	0.90	1.00	0.98	0.25
Content quality	0.98	0.98	1.00	0.26
Domain prominence	0.24	0.25	0.26	1.00

Notes: Pairwise-complete correlations in the 9,611-site media cross-section. The content-quality index is the first principal component of MBFC factual reporting, MBFC credibility, and a binary Wikipedia-page match; the domain-prominence index is the first principal component of log(1 + Majestic referring subnets), Open PageRank, and -log(1 + Tranco rank). The editorial measures are strongly correlated with one another, while the domain-prominence index captures a distinct web-presence dimension, so the measures triangulate rather than duplicate one another.

Table 4: Content quality and domain prominence jointly predict declared restrictions

Quality measure	Separate model	Joint model
Content quality (SD)	0.206*** (0.006)	0.197*** (0.006)
Domain prominence (SD)	0.078*** (0.006)	0.032*** (0.007)

Notes: Linear probability models on the common 5,782-site complete-case sample. The outcome is the February 2026 indicator for restricting at least one of 131 tracked AI-related crawler tokens, including CCBot. Separate models include one index at a time; the joint model includes both. All models include site-category fixed effects and HC1 robust standard errors. Cells report coefficients with standard errors in parentheses; both indices use the full-cross-section SD scale. Significance: * p<0.10, ** p<0.05, *** p<0.01.

Because the indices capture related but distinct dimensions, Table 4 estimates them separately and jointly on their common 5,782-site sample. In the joint model, the content-quality association is 0.197 per SD and the domain-prominence association is 0.032 per SD; both are positive and statistically precise. Domain prominence attenuates more when content quality enters, but retains an independent association. This check supports the paper's "credible and established" interpretation without treating either coefficient as a pure causal effect.

Table 5 documents both index constructions and independently recomputes them from their stated inputs. The content-quality index loads almost entirely on the two MBFC ratings, with a negligible contribution from the binary Wikipedia-page match. The domain-prominence index is the first principal component of standardized log referring subnets, Open PageRank, and inverse-log Tranco rank. For both indices, missing inputs are median-imputed when at least two of the three inputs are present, and observations with fewer than two non-missing inputs receive a missing index value. The independent scores correlate 1.000 with the analysis indices, establishing equality up to the documented sign and normalization.

Table 5: Index construction documentation

Component	Value	Detail
MBFC factual reporting (z)	0.707	PC1 loading
MBFC credibility (z)	0.707	PC1 loading
Wikipedia notability (z)	0.033	PC1 loading
Variance explained by PC1	63.5%	
Correlation with analysis index	1.000	Independent recomputation check
Log referring subnets (z)	0.600	Domain-prominence PC1 loading
Open PageRank (z)	0.563	Domain-prominence PC1 loading
Inverse-log Tranco rank (z)	0.568	Domain-prominence PC1 loading
Variance explained by domain PC1	85.2%	
Correlation with analysis domain index	1.000	Independent recomputation check

Notes: Top panel: first-principal-component loadings for the content-quality index used in the analysis (inputs standardized; median imputation when at least two of three inputs are present), its explained variance, and an independent recomputation check. Bottom panel: the same documentation for the domain-prominence index using standardized log(1 + Majestic referring subnets), Open PageRank, and -log(1 + Tranco rank), with the same missing-data rule. Correlations of 1.000 show equality up to the documented sign and normalization.

A3 CCBot robustness

CCBot is Common Crawl's long-running general-purpose crawler. Its restrictions can predate named AI-crawler governance and can reflect bandwidth, archival, or general anti-scraping concerns, making it a useful sensitivity case rather than a clean training/search classification. The full 9,611-site live cross-section tracks 131 tokens and retains only their site-level any-restriction aggregate, which includes CCBot; it would therefore be incorrect to describe that file as if CCBot could be subtracted after aggregation. Table 6 instead performs the exact comparison in the separate 7,256-site scrape with complete indicators for 23 crawlers. Removing CCBot changes the any-restriction classification for only nine sites, and the factual-reporting coefficient is 0.163 either way. CCBot remains outside the use-specific baskets and is labeled "legacy" in the crawler taxonomy below.

Table 6: CCBot inclusion robustness

Outcome			Restriction rate	Factual coefficient	SE	N	R2
23-bot CCBot	outcome,	including	39.7%	0.163***	(0.004)	7,256	0.170
22-bot CCBot	outcome,	excluding	39.6%	0.163***	(0.004)	7,256	0.170

Notes: Exact sensitivity check in the separate February 2026 per-bot scrape, restricted to 7,256 sites with successful fetches and complete indicators for all 23 crawlers. The first row defines any restriction over all 23; the second removes CCBot before aggregation. Only 9 sites are classified differently. Linear probability models include category fixed effects and HC1 robust standard errors; coefficients are per one ordinal step of MBFC factual reporting. The full 9,611-site live cross-section tracks 131 tokens and retains only the site-level aggregate, which includes CCBot; it is therefore not presented as if CCBot could be subtracted from that file. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A4 Fetch-status robustness

The primary cross-sectional outcome conservatively codes a site without a retrievable `robots.txt` as declaring no restriction. Because successful retrieval is itself positively related to factual reporting, we report the restriction gradient among HTTP-200 fetches as well as the full-sample estimate. Table 7 shows that the factual-reporting coefficient increases from 0.134 in the full sample to 0.164 among successful fetches; the successful-fetch indicator itself rises 0.050 per factual step. Thus differential retrieval attenuates rather than creates the headline quality gradient.

Table 7: Robustness to robots.txt fetch status

Specification	Factual coefficient	SE	N	R2
Primary coding: all sites	0.134***	(0.003)	9,611	0.136
Restriction outcome: HTTP-200 only	0.164***	(0.004)	6,304	0.177
Successful-fetch outcome: all sites	0.050***	(0.004)	9,611	0.018

Notes: Linear probability models with category fixed effects and HC1 robust standard errors. The first two rows use the declared-restriction outcome; the primary full-sample convention codes sites without a retrievable robots.txt as declaring no restriction, while the second row restricts the sample to HTTP-200 fetches. The third row uses successful fetch as the outcome. Coefficients are per one ordinal step of MBFC factual reporting. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A5 Balanced-panel robustness

Because the monthly panel is unbalanced, changing site availability could in principle affect both the estimated quality gradient and the use-specific timing pattern. Table 8 restricts estimation to the 4,778 sites observed in all 45 months, with category and month fixed effects and site-clustered standard errors. The gradient remains positive and precise for all four operationalizations, with coefficients ranging from 0.041 to 0.157.

Table 8: Balanced-panel quality-gradient robustness

Quality measure	Estimate	SE	N	Sites
MBFC factual	0.108***	(0.003)	215,010	4,778
MBFC credibility	0.157***	(0.005)	214,920	4,776
Content quality	0.127***	(0.004)	215,010	4,778
Domain prominence	0.041***	(0.005)	202,680	4,504

Notes: Balanced panel of 4,778 sites observed in all 45 months. Linear probability models use the panel blocked-any outcome excluding CCBot, category and month fixed effects, and site-clustered standard errors. MBFC coefficients are per ordinal step; index coefficients retain the full-panel site-month SD scale. Significance: * p<0.10, ** p<0.05, *** p<0.01.

Table 9 replicates the principal timing result on the same balanced panel. Across all 45 months, its difference between the high-minus-low MBFC factual-rating gaps in training and search correlates 0.988 with the full series, with a mean absolute difference of 0.8 percentage points. The balanced differential rises from a pre-August 2023 average of 2.4 points to 18.8 in December 2023 and 18.6 in July 2024, then narrows to 4.2 by February 2026. The opening-and-catch-up pattern therefore does not reflect sites entering or leaving the monthly panel.

Table 9: Balanced-panel replication of the use-specific timing result

Timing point	Full panel (pp)	Balanced panel (pp)	Balanced - full (pp)
Pre-Aug. 2023 mean	2.2	2.4	+0.2
December 2023	17.0	18.8	+1.8
July 2024	18.0	18.6	+0.5
February 2026	5.0	4.2	-0.8

Notes: Entries are the high-minus-low factual differential in D = BlockTrain - BlockSearch, in percentage points. The full panel contains 7,002 sites and 272,458 site-months; the balanced panel contains 4,778 sites observed in all 45 months. The pre-period row is the mean of the monthly differentials before August 2023. Across all 45 months, the full and balanced series correlate 0.988, with a mean absolute difference of 0.8 percentage points. The replication is descriptive and is not interpreted as identifying a causal event effect.

A6 Crawler taxonomy

Table 10 lists all 23 named crawlers, their source labels, narrow clean-taxonomy classifications, panel membership, and restriction rates. "Clean" denotes a curated basket based on predominant documented function, not an immutable single use. The OpenAI pair is clearest: GPTBot controls potential training use, whereas OAI-SearchBot governs ChatGPT search visibility; Perplexity likewise states that PerplexityBot is for search rather than foundation-model training. Google-Extended instead mixes future training and grounding, and the frozen Claude-Web token predates Anthropic's current ClaudeBot/Claude-SearchBot/Claude-User split.[2] Thirteen mixed or insufficiently documented tokens enter only the broad robustness. The panel uses three training and two search labels. Because unequal baskets can create mechanical differences, the aggregate gap is the estimand but the one-to-one OpenAI pair is the cleanest comparison. The per-bot file is cross-sectional, not a site-bot-month panel.

[2]Provider documentation checked July 2026: OpenAI publisher FAQ; Google crawler documentation; Anthropic crawler documentation; and Perplexity crawler documentation.

Table 10: Crawler taxonomy used in the bot-level analyses

Bot	Scraper label	Clean taxonomy	Panel role	Blocking rate
GPTBot	training	Training	Training	32.3%
CCBot	legacy	Excluded (legacy)	–	30.7%
ClaudeBot	training	Training	Training	30.0%
Google-Extended	training	Training	Training	28.1%
Bytespider	training	Ambiguous (excluded)	–	28.1%
anthropic-ai	training	Training	–	27.5%
Claude-Web	search	Search/discovery	–	26.4%
Applebot-Extended	training	Ambiguous (excluded)	–	26.1%
PerplexityBot	search	Search/discovery	Search	25.8%
ChatGPT-User	search	Ambiguous (excluded)	–	25.5%
Omgilibot	other	Ambiguous (excluded)	–	24.3%
Meta-ExternalAgent	training	Ambiguous (excluded)	–	22.6%
Cohere-ai	training	Ambiguous (excluded)	–	22.3%
PetalBot	search	Search/discovery	–	20.7%
FacebookBot	training	Ambiguous (excluded)	–	20.6%
OAI-SearchBot	search	Search/discovery	Search	20.2%
Diffbot	training	Ambiguous (excluded)	–	19.4%
Amazonbot	search	Ambiguous (excluded)	–	17.8%
YouBot	search	Search/discovery	–	17.7%
Ai2Bot-Dolma	training	Ambiguous (excluded)	–	16.3%
Timpibot	training	Ambiguous (excluded)	–	16.3%
Ai2Bot	training	Ambiguous (excluded)	–	16.0%
PanguBot	training	Ambiguous (excluded)	–	8.8%

Notes: The 23 named crawlers in the February 2026 per-bot scrape (7,256 sites with successful fetches). Blocking rates are shares of sites disallowing the crawler. The clean taxonomy is a narrow, curated classification based on each crawler's predominant documented function; it is not a claim that crawler functions are immutable or perfectly separable. Tokens with mixed or insufficiently documented roles are excluded from clean baskets but enter the broad-taxonomy robustness. The panel-basket column shows the aggregate monthly HTTP Archive definitions (three training bots; two search bots). The per-bot file is a February 2026 cross-section, not a site-bot-month panel.

A7 Access states and conditional probabilities

Table 11 reports the exact counts behind the manuscript's access-state figure. Search-only restrictions are nearly absent (8 of 7,002 sites), while 614 sites restrict training only and 1,860 restrict both.

Table 11: Use-specific access states by factual-reporting tier

Factual tier	Sites	Allows both	Training only	Search only	Both
Low	847	93.9% (795)	3.8% (32)	0.0% (0)	2.4% (20)
Mixed	1,490	89.3% (1,330)	4.8% (71)	0.3% (4)	5.7% (85)
Mostly Factual	722	73.1% (528)	9.6% (69)	0.1% (1)	17.2% (124)
High	3,943	47.3% (1,867)	11.2% (442)	0.1% (3)	41.4% (1,631)

Notes: Cross-section of 7,002 MBFC-rated media sites with nonmissing training/search outcomes in February 2026. Cells report row percentages with counts in parentheses. Training and search are aggregate bot-type outcomes used in the analysis; CCBot is excluded.

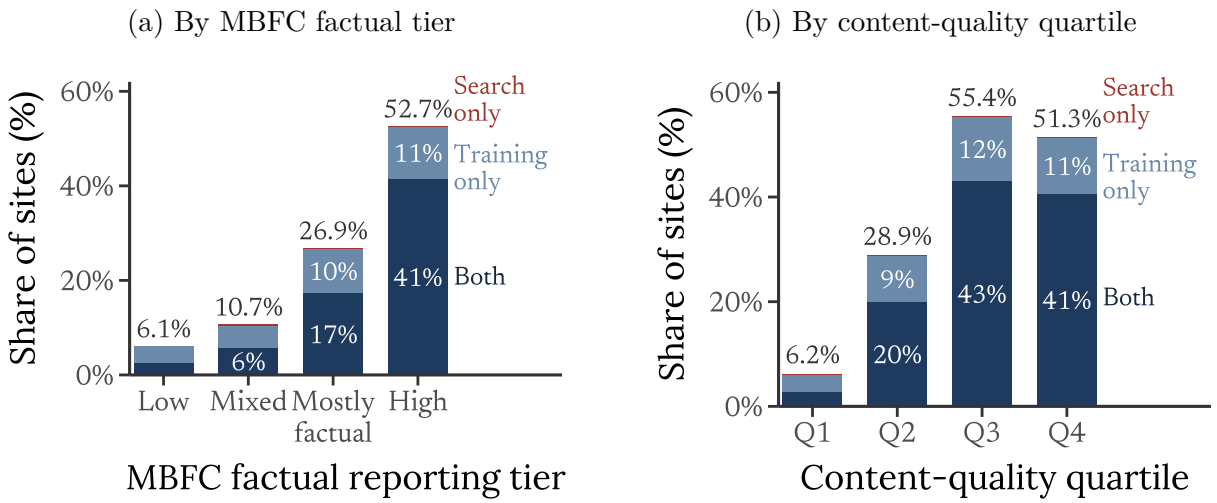
Table 12 reports the conditional probabilities. Conditional on allowing search, 12.0% of sites restrict training (19.1% among high-factual sites); conditional on restricting training, 75.2% also restrict search; conditional on restricting anything, only 0.3% are search-only. Figure 1 decomposes the quality gradient in any-restriction into its three restricting states, by factual tier (panel (a)) and by content-quality quartile (panel (b)): the gradient is built from training-only and joint restrictions, with search-only contributing essentially nothing at any quality level.

Table 12: Conditional use-specific access probabilities

Sample	Pr(restrict training | allows search)	Pr(restrict search | restricts training)	Pr(search-only | any restriction)
All sites	12.0% (N=5,134)	75.2% (N=2,474)	0.3% (N=2,482)
Low factual	3.9% (N=827)	38.5% (N=52)	0.0% (N=52)
Mixed factual	5.1% (N=1,401)	54.5% (N=156)	2.5% (N=160)
Mostly factual	11.6% (N=597)	64.2% (N=193)	0.5% (N=194)
High factual	19.1% (N=2,309)	78.7% (N=2,073)	0.1% (N=2,076)

Notes: February 2026 cross-section of 7,002 sites with nonmissing training/search outcomes (CCBot excluded). Each cell reports a conditional probability of declared restriction with its conditioning-set size. Search-only restrictions are nearly absent: conditional on restricting anything, almost all sites restrict training at least as much as search.

Figure 1: Decomposition of the any-restriction gradient into access states.



Note. Sample is the 7,002-site training/search cross-section. Bars stack the shares restricting both uses, training only, and search only; the labels above each column are the total any-restriction rates. Content-quality groups are rank-based quartiles computed within the sample, with site ID used only to break ties in the discrete composite score.

A8 Direct training-minus-search gap models

Table 13 reports the full direct-gap estimates for all four quality measures, in the monthly panel and the February 2026 cross-section, together with the separate training and search coefficients from the same samples. All eight gap coefficients are positive with $p < 0.01$. On a common sample the gap coefficient equals the training coefficient minus the search coefficient by OLS linearity; the gap specification contributes a correct standard error for that difference and an outcome that nets out site-level blocking propensity by construction.

Table 13: Direct training-minus-search gap models

Sample	Quality measure	Training	Search	Gap D	SE(D)	N	Clusters
Monthly panel	MBFC factual	0.093***	0.068***	0.025***	(0.001)	272,458	7,002
Monthly panel	MBFC credibility	0.141***	0.102***	0.039***	(0.002)	272,331	6,999
Monthly panel	Content quality	0.112***	0.081***	0.030***	(0.002)	272,458	7,002
Monthly panel	Domain prominence	0.038***	0.015***	0.022***	(0.002)	239,534	5,690
Feb. 2026 cross-section	MBFC factual	0.155***	0.130***	0.024***	(0.002)	7,002	
Feb. 2026 cross-section	MBFC credibility	0.237***	0.195***	0.042***	(0.003)	6,999	
Feb. 2026 cross-section	Content quality	0.207***	0.172***	0.034***	(0.003)	7,002	
Feb. 2026 cross-section	Domain prominence	0.086***	0.058***	0.028***	(0.004)	5,690	

Notes: Outcome D equals the training-restriction indicator minus the search-restriction indicator, so positive coefficients mean higher quality predicts restricting training over search within the same site-month (panel) or site (cross-section). Panel models include category and month fixed effects with site-clustered standard errors; cross-section models include category fixed effects with HC1 standard errors. By OLS linearity the D coefficient equals the training coefficient minus the search coefficient on the same sample; the D column adds a correct standard error for that difference. MBFC coefficients are per ordinal step. Index coefficients use the panel-wide site-month scale in panel rows and the full-cross-section scale in cross-sectional rows. Significance: * p<0.10, ** p<0.05, *** p<0.01.

Table 14 condenses the use-specific evidence across levels of aggregation—panel outcomes, site-bot fixed effects, and the within-OpenAI comparison—for the two headline quality measures.

Table 14: Use-specific access evidence

Analysis	Outcome/comparison	Quality measure	Estimate	N
Aggregate site-month panel	Training outcome	MBFC factual	0.093	272,458 site-months
Aggregate site-month panel	Search outcome	MBFC factual	0.068	272,458 site-months
Aggregate site-month panel	Training - search gap (D)	MBFC factual	0.025	272,458 site-months
Aggregate site-month panel	Training outcome	Content quality	0.112	272,458 site-months
Aggregate site-month panel	Search outcome	Content quality	0.081	272,458 site-months
Aggregate site-month panel	Training - search gap (D)	Content quality	0.030	272,458 site-months
Site-bot FE	Clean training vs search bots	MBFC factual	0.037***	65,304 site-bots
Site-bot FE	Clean training vs search bots	Content quality	0.051***	65,304 site-bots
Within OpenAI	GPTBot vs OAI-SearchBot	MBFC factual	0.053***	14,512 site-bots
Within OpenAI	GPTBot vs OAI-SearchBot	Content quality	0.071***	14,512 site-bots

Notes: Panel rows are linear probability models on the 272,458-site-month panel (7,002 site clusters) with category and month fixed effects and site-clustered standard errors; the gap row uses the direct outcome D = training - search; panel-row stars are omitted because the corresponding full models with standard errors are reported in the gap-model table. Site-bot rows use successful February 2026 robots.txt fetches with site and bot fixed effects and site-clustered standard errors; coefficients are training-bot x quality interactions with stars from those models. MBFC coefficients are per ordinal step. Content quality uses the panel-wide site-month scale in panel rows and the full-cross-section scale in site-bot rows. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A9 Paired-bot and site-bot fixed-effects evidence

Table 15 reports regressions of within-site paired differences—GPTBot minus OAI-SearchBot, ClaudeBot minus Claude-Web, and mean clean-training minus mean clean-search—on quality. Table 16 reports the corresponding site-bot fixed-effects models, which absorb site and bot intercepts and cluster by site; Table 17 isolates the within-vendor versions. The paired-difference regressions add category fixed effects, while the fixed-effects models absorb site and bot intercepts. Because every successfully fetched site has complete data on all 23 crawlers, the estimates agree closely: the OpenAI factual and both

clean-basket coefficients are identical to three decimals, and the OpenAI content-quality estimates differ only in the last reported decimal. The OpenAI pair is the cleanest because provider documentation explicitly separates potential training from search visibility. The smaller Anthropic estimates are robustness only because the frozen Claude-Web token has a less stable functional interpretation.

Table 15: Paired bot gap models: same publisher, training versus search crawler

Pair	Quality measure	Estimate	SE	N
OpenAI: GPTBot - OAI-SearchBot	MBFC factual	0.053***	(0.003)	7,256
OpenAI: GPTBot - OAI-SearchBot	Content quality	0.070***	(0.003)	7,256
Anthropic: ClaudeBot - Claude-Web	MBFC factual	0.004**	(0.002)	7,256
Anthropic: ClaudeBot - Claude-Web	Content quality	0.008***	(0.002)	7,256
Clean taxonomy: mean training - mean search	MBFC factual	0.037***	(0.002)	7,256
Clean taxonomy: mean training - mean search	Content quality	0.051***	(0.002)	7,256

Notes: Site-level regressions on successful February 2026 robots.txt fetches (7,256 sites; 23 bots per site). The outcome is the within-site difference between blocking the training-oriented crawler(s) and blocking the search-oriented crawler(s) of the same vendor or clean taxonomy. Models include category fixed effects with HC1 standard errors. MBFC coefficients are per ordinal step; content quality is on the full-cross-section SD scale. Positive estimates mean higher-quality publishers restrict training relatively more than search within the same publisher. Significance: * p<0.10, ** p<0.05, *** p<0.01.

Table 16: Site-bot fixed-effects evidence on use-specific access

Taxonomy	Quality measure	Estimate	SE	N	Sites
Clean taxonomy	MBFC factual	0.037***	(0.002)	65,304	7,256
Clean taxonomy	MBFC credibility	0.060***	(0.002)	65,250	7,250
Clean taxonomy	Content quality	0.051***	(0.002)	65,304	7,256
Clean taxonomy	Domain prominence	0.038***	(0.003)	42,948	4,772
Broad taxonomy	MBFC factual	0.003***	(0.001)	152,376	7,256
Broad taxonomy	MBFC credibility	0.008***	(0.001)	152,250	7,250
Broad taxonomy	Content quality	0.006***	(0.001)	152,376	7,256
Broad taxonomy	Domain prominence	0.010***	(0.002)	100,212	4,772

Notes: Unit is site-bot in the February 2026 live scrape, restricted to successful robots.txt fetches (HTTP 200). Each estimate is from a separate linear probability model with site and bot fixed effects, clustered by site. The coefficient is Training bot x quality. The clean taxonomy uses GPTBot, Google-Extended, ClaudeBot, and anthropic-ai as training bots and OAI-SearchBot, PerplexityBot, YouBot, Claude-Web, and PetalBot as search/discovery bots. The broad taxonomy follows the media scraper's broader labels. MBFC interactions are per ordinal step; index interactions retain the full-cross-section SD scale. Significance: * p<0.10, ** p<0.05, *** p<0.01.

Table 17: Within-vendor training versus search comparisons

Comparison	Quality measure	Estimate	SE	N	Sites
OpenAI: GPTBot vs OAI-SearchBot	MBFC factual	0.053***	(0.003)	14,512	7,256
OpenAI: GPTBot vs OAI-SearchBot	Content quality	0.071***	(0.003)	14,512	7,256
Anthropic/Claude: ClaudeBot vs Claude-Web	MBFC factual	0.005**	(0.002)	14,512	7,256
Anthropic/Claude: ClaudeBot vs Claude-Web	Content quality	0.008***	(0.002)	14,512	7,256

Notes: Unit is site-bot in the February 2026 live scrape, restricted to successful robots.txt fetches (N site-bot observations and site clusters reported per row). Each estimate is from a separate linear probability model with site and bot fixed effects and site-clustered standard errors. The coefficient is training-oriented bot x quality; MBFC factual is per ordinal step and content quality retains the full-cross-section SD scale. OpenAI is the cleanest comparison because GPTBot and OAI-SearchBot are explicitly named training and search crawlers from the same vendor. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A10 Taxonomy and salience robustness

The training/search comparison bundles basket composition, vendor identity, salience, and rollout timing. Table 18 shows that the broad source-label gradients remain positive and precise but are roughly an order of magnitude smaller than the clean-basket estimates; classification matters, so the paper gives greatest weight to the OpenAI pair. Both vendor pairs are positive, although Anthropic's magnitudes are small. Blocking-share and blockers-only sensitivities also remain positive, but both condition on outcomes from the same bot file and are descriptive checks rather than causal awareness controls. In Table 19, every clean-basket estimate remains positive and significant when one crawler is removed, including GPTBot (0.032 per factual step).

Table 18: Taxonomy and salience robustness for the use-specific gap

Specification	Quality measure	Estimate	SE	N
Clean taxonomy (baseline)	MBFC factual	0.037***	(0.002)	7,256
Clean taxonomy (baseline)	Content quality	0.051***	(0.002)	7,256
Broad taxonomy (media-scraper labels)	MBFC factual	0.003***	(0.001)	7,256
Broad taxonomy (media-scraper labels)	Content quality	0.006***	(0.001)	7,256
Within-vendor: OpenAI pair	MBFC factual	0.053***	(0.003)	7,256
Within-vendor: OpenAI pair	Content quality	0.070***	(0.003)	7,256
Within-vendor: Anthropic pair	MBFC factual	0.004**	(0.002)	7,256
Within-vendor: Anthropic pair	Content quality	0.008***	(0.002)	7,256
Clean, controlling for overall blocking share	MBFC factual	0.025***	(0.002)	7,256
Clean, controlling for overall blocking share	Content quality	0.037***	(0.002)	7,256
Clean, sites blocking at least one of 23 bots	MBFC factual	0.035***	(0.008)	2,879
Clean, sites blocking at least one of 23 bots	Content quality	0.063***	(0.009)	2,879

Notes: Site-level regressions of paired training-minus-search blocking gaps on quality, February 2026 per-bot scrape (category fixed effects; HC1 standard errors). The broad taxonomy uses the media scraper's own training/search labels (14 training bots, 7 search bots); the clean taxonomy is the narrower curated classification described in Appendix A6. The blocking-share control and blockers-only restriction are descriptive sensitivities constructed from the same bot outcomes, not causal adjustments. MBFC coefficients are per ordinal step; content quality is on the full-cross-section SD scale. Significance: * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 19: Leave-one-bot-out robustness for the clean-taxonomy gap

Excluded bot	Quality measure	Estimate	SE	N
GPTBot (training)	MBFC factual	0.032***	(0.002)	7,256
GPTBot (training)	Content quality	0.045***	(0.002)	7,256
Google-Extended (training)	MBFC factual	0.041***	(0.002)	7,256
Google-Extended (training)	Content quality	0.056***	(0.002)	7,256
ClaudeBot (training)	MBFC factual	0.037***	(0.002)	7,256
ClaudeBot (training)	Content quality	0.051***	(0.002)	7,256
anthropic-ai (training)	MBFC factual	0.037***	(0.002)	7,256
anthropic-ai (training)	Content quality	0.052***	(0.003)	7,256
OAI-SearchBot (search)	MBFC factual	0.036***	(0.002)	7,256
OAI-SearchBot (search)	Content quality	0.050***	(0.002)	7,256
PerplexityBot (search)	MBFC factual	0.044***	(0.002)	7,256
PerplexityBot (search)	Content quality	0.060***	(0.002)	7,256
YouBot (search)	MBFC factual	0.032***	(0.002)	7,256
YouBot (search)	Content quality	0.045***	(0.002)	7,256
Claude-Web (search)	MBFC factual	0.045***	(0.002)	7,256
Claude-Web (search)	Content quality	0.061***	(0.003)	7,256
PetalBot (search)	MBFC factual	0.027***	(0.002)	7,256
PetalBot (search)	Content quality	0.036***	(0.002)	7,256

Notes: Each row re-estimates the clean-taxonomy training-minus-search gap regression after removing one named crawler from its basket (category fixed effects; HC1 standard errors; February 2026 per-bot scrape). The gradient is not driven by any single crawler, including GPTBot. MBFC coefficients are per ordinal step; content quality is on the full-cross-section SD scale. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A11 Heterogeneity in the use-specific gap

Table 20 explores whether use-specific differentiation varies across observable site attributes. Category-level attributes (original-content and content-producer indicators) are collinear with the category fixed effects in this media-only sample and are therefore excluded. The

questionable-source interaction is negative under both quality specifications: flagged sources exhibit a smaller use differential than otherwise non-questionable sources at the same measured quality. Interactions with log Tranco rank and log referring subnets are close to zero and statistically indistinguishable from zero. These are descriptive moderation patterns, not mechanism tests, and the manuscript builds no claim on them.

Table 20: Heterogeneity in the use-specific gap

Moderator	Quality measure	Quality	Mod.	Interaction	SE	N
MBFC questionable	Content quality	0.035***	-0.119***	-0.088***	(0.019)	7,002
MBFC questionable	MBFC factual	0.024***	0.100***	-0.048***	(0.009)	7,002
log Tranco rank (z)	Content quality	0.021***	-0.019***	-0.002	(0.005)	5,014
log Tranco rank (z)	MBFC factual	0.013***	-0.020	-0.000	(0.004)	5,014
Referring subnets (log, z)	Content quality	0.022***	0.022***	0.001	(0.004)	5,667
Referring subnets (log, z)	MBFC factual	0.014***	0.027*	-0.001	(0.003)	5,667

Notes: Site-level regressions of D = training-minus-search restriction gap on quality, a moderator, and their interaction, in the February 2026 training/search cross-section (N per model reported; category fixed effects; HC1 standard errors). Continuous moderators are standardized; category-level moderators (original-content and content-producer indicators) are excluded because they are collinear with the category fixed effects in this media-only sample. These are descriptive moderation patterns, not causal mechanism tests. MBFC coefficients are per ordinal step; content quality is on the full-cross-section SD scale. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A12 Timing detail

Table 21 reports pre- and post-August 2023 averages of any-crawler, training, and search restrictions by factual tier. Table 22 reports period averages of the direct gap D by tier for three periods: before August 2023, the salience window through July 2024, and the later catch-up period.

Table 21: Training and search restrictions before and after August 2023

Period	Factual tier	Any AI crawler	Training	Search	Training-search gap
Pre-Aug. 2023	Low	1.7%	0.0%	0.0%	0.0
Pre-Aug. 2023	Mixed	8.1%	0.2%	0.0%	0.2
Pre-Aug. 2023	Mostly Factual	2.7%	1.0%	0.1%	0.8
Pre-Aug. 2023	High	9.3%	2.4%	0.2%	2.3
Aug. 2023-Feb. 2026	Low	7.5%	4.5%	2.0%	2.5
Aug. 2023-Feb. 2026	Mixed	16.1%	8.7%	4.9%	3.8
Aug. 2023-Feb. 2026	Mostly Factual	22.4%	19.7%	12.9%	6.8
Aug. 2023-Feb. 2026	High	48.0%	44.6%	32.0%	12.6

Notes: Monthly panel of 7,002 sites (272,458 site-months), June 2022 through February 2026. Outcomes are declared-restriction indicators (any named AI crawler, training basket, search basket; CCBot excluded); cells are averages over site-months within factual tier and period, in percent. The period split is descriptive; August 2023 is used as a salient marker for named AI-crawler opt-out rather than as an isolated causal shock.

Table 22: Average training-minus-search gap D by factual tier and period

Period	Low	Mixed	Mostly Factual	High
Pre-Aug. 2023	0.0	0.2	0.8	2.3
Aug. 2023-Jul. 2024	1.4	3.4	7.3	16.1
Aug. 2024-Feb. 2026	3.3	4.0	6.6	10.3

Notes: Monthly site panel (272,458 site-months; 7,002 sites). Entries are period averages of D = training-restriction indicator minus search-restriction indicator, in percentage points, by MBFC factual tier. The periods are descriptive: August 2023 marks the salient start of named AI-crawler opt-out and July 2024 the arrival of one-click blocking tools; neither is treated as an isolated causal shock.

A12.1 Event-time gaps by use

Table 23 reports selected months of the high-minus-low MBFC factual-rating gaps in training and search restrictions separately; Figure 2 plots the two series with approximate confidence bands, and Figure 4 shows the underlying levels by factual tier (training in panel (a), search in panel (b)). The training gap opened first (36.4 versus 19.4 points in December 2023) and the search gap largely caught up by February 2026 (45.3 versus 40.4). Figure 3 plots month-

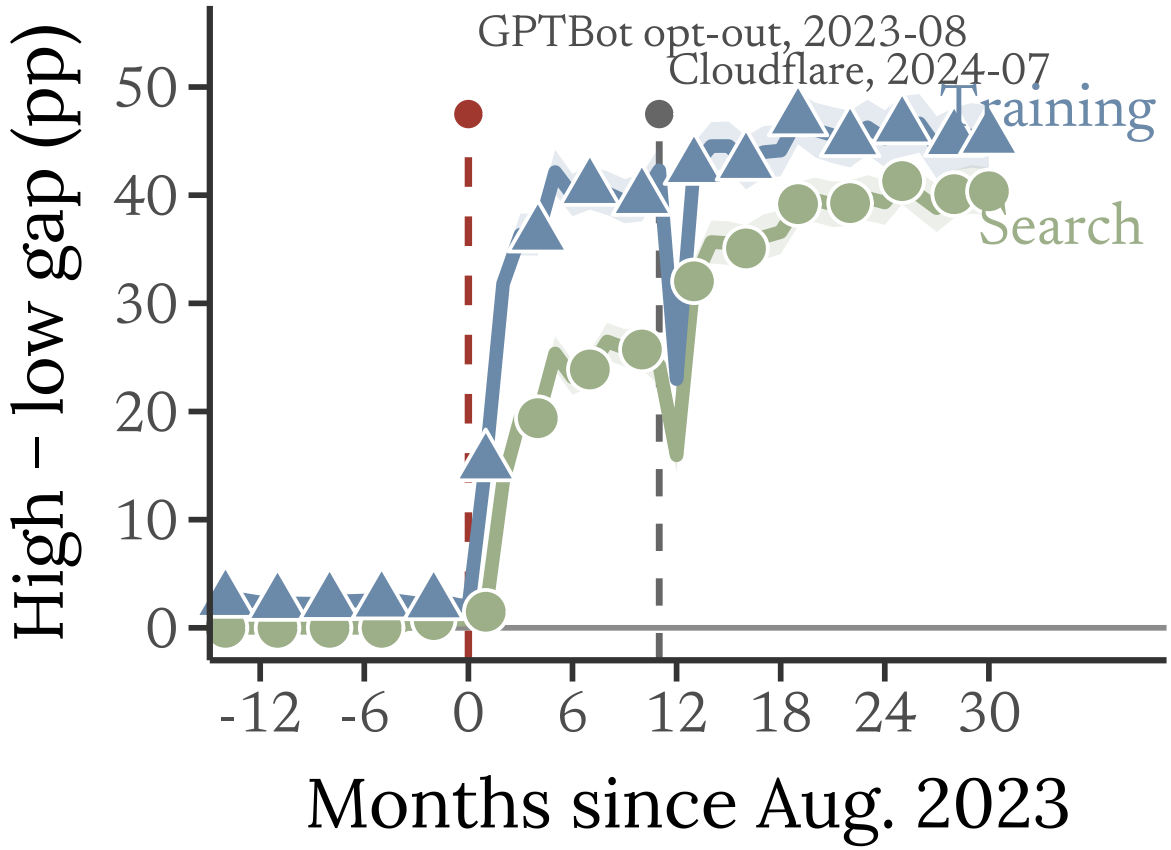
specific regression gradients of D on continuous factual reporting and content quality, also exported as `r1_table_monthly_gap_gradients.csv`; the average pre-period factual gradient is 0.008 per ordinal step—small but not zero, which the manuscript reports explicitly. All timing evidence is descriptive: ChatGPT diffusion, named-crawler documentation, litigation, and blocking tools were clustered in the same window.

Table 23: Selected high-minus-low MBFC factual-rating gaps in training and search restrictions

Month	Event month	Training gap	Search gap	Train - search
2022-06	-14	2.6	0.0	2.6
2023-07	-1	2.3	0.7	1.5
2023-08	0	1.9	0.8	1.2
2023-12	4	36.4	19.4	17.0
2024-07	11	42.3	24.3	18.0
2026-02	30	45.3	40.4	5.0

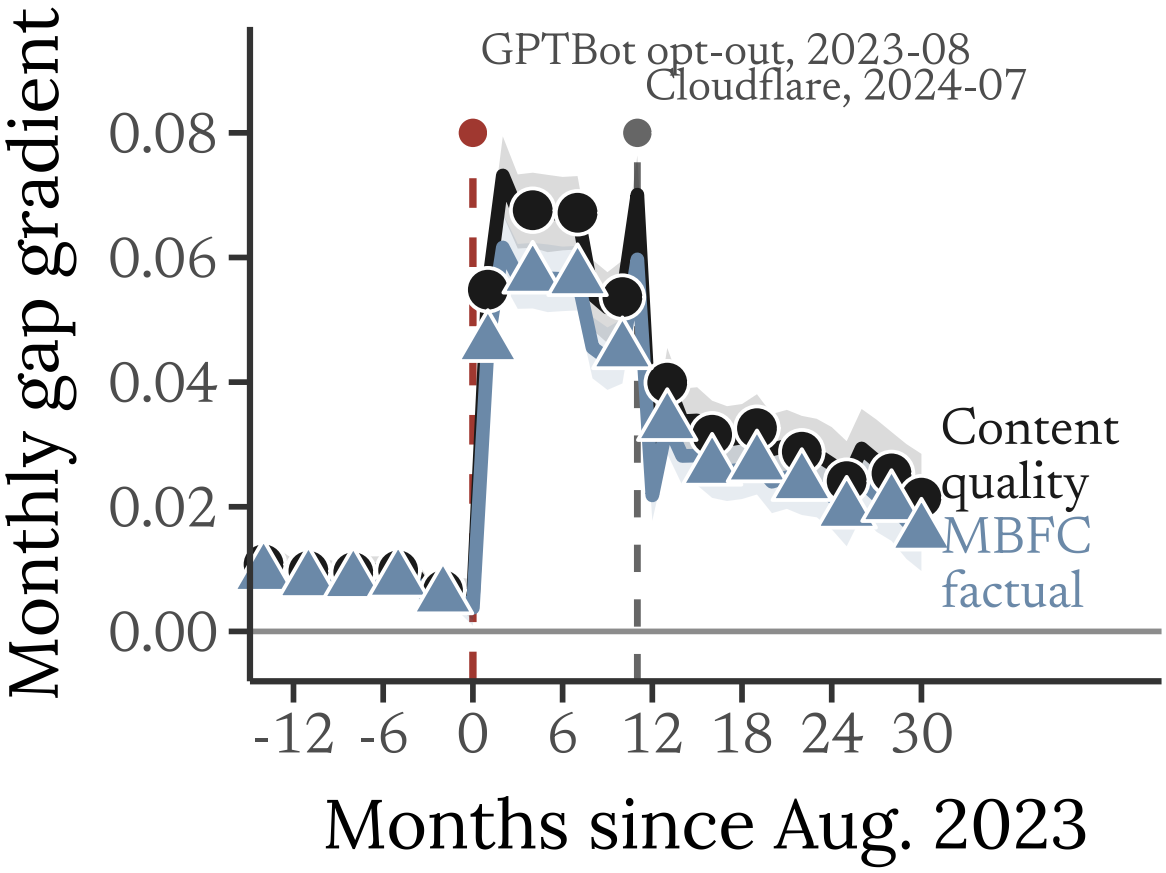
Notes: Monthly panel of 7,002 sites (272,458 site-months); outcomes are the aggregate training- and search-basket restriction indicators. Entries are percentage-point differences between high- and low-MBFC-factual-rating sites. The training-minus-search column is computed from the underlying unrounded gaps, so it can differ by 0.1 point from subtracting the displayed components. Event month 0 is August 2023. The table supports the descriptive event-time figure and is not interpreted as a causal event-study estimate.

Figure 2: Event-time high-minus-low factual-rating gaps in training and search restrictions.



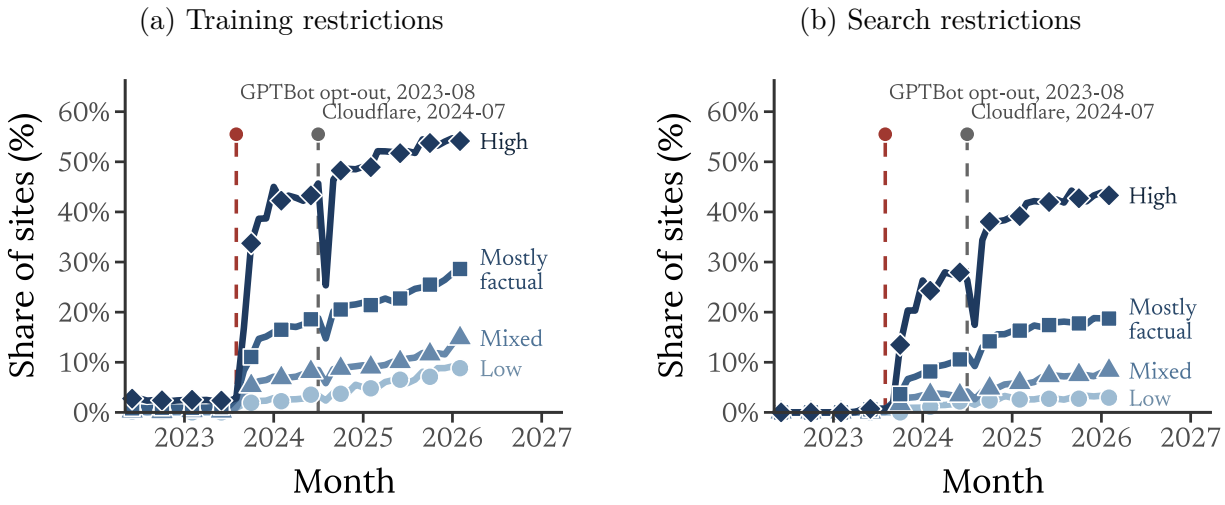
Note. Monthly panel of 7,002 sites. Bands are approximate 95% intervals for differences in proportions. Month 0 is August 2023; the grey pin marks July 2024.

Figure 3: Monthly gradients of the direct gap on quality.



Note. Monthly coefficients from regressing $D = \text{BlockTrain} - \text{BlockSearch}$ on quality—MBFC factual reporting (per ordinal step) and the content-quality index (on the full-panel site-month SD scale), estimated separately—with category fixed effects and HC1 intervals. Month 0 is August 2023; the grey pin marks July 2024.

Figure 4: Training and search restriction levels by factual tier.



Note. Monthly panel of 7,002 sites; monthly shares restricting the aggregate training and search baskets, by MBFC factual tier.

A13 Secondary composition implications

Table 24 reports the raw rates: questionable sources restrict at 6.1% versus 34.8% among sites without an MBFC questionable-source flag, and center outlets restrict most among the politically rated. The comparison group is "not flagged," not a subset restricted to high-credibility outlets.

Table 24: Secondary implications: misinformation and political orientation

Panel	Group	Sites	Blocking rate
Questionable-source status	Not MBFC-questionable	7,627	34.8%
Questionable-source status	MBFC questionable	1,984	6.1%
Political orientation	Left	663	23.7%
Political orientation	Left-center	1,876	39.1%
Political orientation	Center	2,083	51.1%
Political orientation	Right-center	855	48.8%
Political orientation	Right	518	10.8%

Notes: Outcome is the February 2026 indicator for restricting at least one of the 131 tracked AI-related crawler tokens (including CCBot); cells report group restriction rates with group sizes. Questionable-source rows use the full 9,611-site cross-section; political-orientation rows use the 5,995 sites with left-to-right MBFC ratings. The comparison group is all sites not carrying an MBFC questionable-source flag; it is not restricted to high-credibility outlets. These raw patterns are interpreted as downstream composition implications of quality-selective opt-out, not as independent causal mechanisms.

A13.1 Conditional regressions

Table 25 shows how the gaps respond to quality controls. To make the questionable-source ladder comparable, every row uses the same 5,782 sites with both composite indices. Within that fixed sample, the unadjusted coefficient is -0.327; it becomes 0.064 with both indices and category fixed effects—a statistically significant sign flip. Conditional on measured quality and category, flagged sources restrict more than otherwise non-questionable sources. This does not erase the full-sample raw rate gap; it shows that the gap is a composition pattern rather than an additional tendency of flagged sites to welcome AI crawling. Political-orientation gaps attenuate with quality controls but do not vanish, so we treat political orientation as residual heterogeneity rather than pure composition.

Table 25: Conditional secondary-implication regressions

Pattern	Specification	Estimate	SE	N
Questionable source	(1) Unadjusted, common sample	-0.327***	(0.016)	5,782
Questionable source	(2) + Content quality	0.067***	(0.018)	5,782
Questionable source	(3) + Both quality indices	0.064***	(0.018)	5,782
Questionable source	(4) + Both + category FE	0.064***	(0.018)	5,782
Left vs center	(1) Unadjusted	-0.274***	(0.020)	5,995
Left vs center	(4) + Both + category FE	-0.230***	(0.025)	4,650
Left-center vs center	(1) Unadjusted	-0.120***	(0.016)	5,995
Left-center vs center	(4) + Both + category FE	-0.163***	(0.017)	4,650
Right-center vs center	(1) Unadjusted	-0.024	(0.020)	5,995
Right-center vs center	(4) + Both + category FE	-0.045**	(0.022)	4,650
Right vs center	(1) Unadjusted	-0.403***	(0.018)	5,995
Right vs center	(4) + Both + category FE	-0.186***	(0.030)	4,650

Notes: Linear probability models on the February 2026 cross-section; the outcome is restriction of at least one of the 131 tracked AI-related crawler tokens (including CCBot). Questionable-source rows hold the estimation sample fixed at the sites with both quality indices (N reported per row), so the ladder is comparable across specifications; row (1) is therefore an unadjusted common-sample coefficient, not the full-sample raw rate difference. Political-orientation unadjusted rows use the sites with left-to-right MBFC ratings; controlled rows additionally require both quality indices; center is the omitted category. Quality controls are the standardized content-quality and domain-prominence indices; specification (4) adds category fixed effects. HC1 standard errors. The questionable-source gap is largely explained by quality composition and turns positive once quality is controlled: conditional on measured quality and category, flagged sources restrict more than otherwise non-questionable sources. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A13.2 Are the secondary patterns use-specific?

Table 26 asks whether the composition patterns extend to the use-specific margin D. The answer is not uniform. The questionable-source coefficient is small and statistically indistinguishable from zero with the content-quality control, but negative and significant with the factual-reporting control. Relative to center outlets, three of the four political-orientation coefficients are statistically significant and range in sign. These conditional associations are

descriptive and specification-dependent; they do not establish direct effects of misinformation status or ideology. The manuscript therefore makes no independent claim from this table.

Table 26: Are the secondary patterns use-specific?

Specification	Term	Estimate	SE	N
D on Content quality + questionable	Questionable source	-0.007	(0.008)	7,002
D on Content quality + questionable	Content quality	0.033***	(0.004)	7,002
D on MBFC factual + questionable	Questionable source	-0.024***	(0.007)	7,002
D on MBFC factual + questionable	MBFC factual	0.020***	(0.003)	7,002
D on content quality + political orientation	Left vs center	-0.039***	(0.014)	5,224
D on content quality + political orientation	Left-center vs center	-0.058***	(0.010)	5,224
D on content quality + political orientation	Right-center vs center	0.032**	(0.016)	5,224
D on content quality + political orientation	Right vs center	-0.028	(0.019)	5,224

Notes: Outcome is D = training-restriction indicator minus search-restriction indicator in the February 2026 training/search cross-section (category fixed effects; HC1 standard errors; center is the omitted political-orientation category). The secondary patterns are conditional-composition checks, not separate causal mechanisms. Significance: * p<0.10, ** p<0.05, *** p<0.01.

A14 Composition scenarios: construction detail

The manuscript's composition table recalculates the composition of the observed February 2026 access states in the 7,002-site training/search sample under three access rules: the full population benchmark, sites accessible under search rules (BlockSearch = 0), and sites accessible under training rules (BlockTrain = 0); the licensing-margin row describes the 614 training-only restrictors. By construction the search-rules pool exceeds the training-rules pool by the number of training-only restrictors net of search-only restrictors. Political-pole shares are computed among sites with left-to-right MBFC ratings, whose counts the table reports. The exercise involves no behavioral model, no token or volume weighting, and no projection; it is bookkeeping on observed declarations that shows which content sits at each

access margin.

A15 Formal framework

This section states the formal version of the manuscript's framework and records the conditions that discipline its predictions.

Setup. A publisher has quality q distributed on $[\underline{q}, \bar{q}]$. For each use $u \in \{T, S\}$ (training, search/discovery), let

$$g_u(q) = C_u(q) - B_u(q)$$

be the net cost of allowing access: $C_u(q)$ is the appropriation or substitution cost and $B_u(q)$ is the visibility, referral, or discovery benefit. Let $\kappa \geq 0$ be the implementation hurdle for learning about and expressing a use-specific `robots.txt` restriction. The publisher restricts use u when the avoided net access cost exceeds that hurdle:

$$R_u(q) = \mathbf{1}\{g_u(q) > \kappa\}.$$

Using one κ for both uses is a normalization; use-specific hurdles shift the thresholds without changing the ordering logic below.

A1 (quality-increasing net costs). g_T and g_S are continuous and strictly increasing in q on the observed range: higher-quality content is more valuable to appropriate and more exposed to substitution.

A2 (use ordering). On the observed range, $g_T(q) \geq g_S(q)$, with $g_T'(q) > g_S'(q) \geq 0$ wherever the derivatives exist. Training access is largely unattributed, while search-like access can preserve referral and discovery value; these benefit differences, together with any cost differences, make the net cost of training access weakly higher and more quality-sensitive.

Assumption A2 attaches visibility benefits to search-like use; nothing in the framework assumes that inclusion in training data generates traffic. The pointwise ordering is stated explicitly because a slope inequality alone does not order the two restriction thresholds.

Proposition 1 (Threshold structure)**.** *Under A1, the set of publishers restricting use u is an upper interval. If $g_u(\underline{q}) < \kappa < g_u(\bar{q})$, there is a unique interior threshold q_u^* satisfying $g_u(q_u^*) = \kappa$, and $R_u(q) = 1$ if and only if $q > q_u^*$. If the hurdle lies outside that range, either all or no publishers restrict that use.*

Proof. Continuity gives an interior crossing when the hurdle lies between the endpoint values, and strict monotonicity makes that crossing unique. Above the crossing $g_u(q) > \kappa$; below it $g_u(q) < \kappa$. The boundary cases follow directly. □

Proposition 2 (Use-ordered thresholds)**.** *Under A1–A2, $R_T(q) \geq R_S(q)$ for every q. When both thresholds are interior, $q_T^* \leq q_S^*$; publishers with $q \in (q_T^*, q_S^*]$ restrict training while allowing search, and no publisher restricts search while allowing training. Hence $D(q) = R_T(q) - R_S(q)$ is nonnegative and equals one on the intermediate region.*

Proof. If $R_S(q) = 1$, then $g_S(q) > \kappa$. Assumption A2 implies $g_T(q) \geq g_S(q) > \kappa$, so $R_T(q) = 1$. The threshold ordering and characterization of the intermediate region then follow from Proposition 1. □

Adding a common publisher shock to both g_T and g_S smooths restriction probabilities while preserving their ordering. Use-specific shocks permit the rare $D = -1$ observations seen in the data. Under arbitrary shocks, however, a globally increasing $\Pr(D = 1 \mid q)$ does not follow mechanically: the empirical prediction is local, over the range in which the training probability rises faster than the search probability.

Remark 1 (Bounded outcomes)**.** *The latent gap $g_T(q) - g_S(q)$ can widen globally in q, but the observable probability gap is bounded. If both restriction probabilities converge to one at sufficiently high quality, their difference must eventually return toward zero; the analogous*

statement holds at the low end if both converge to zero. A widening observed gap is therefore a local prediction for an interior range, not a global monotonicity claim. The observed use-specific restriction rates remain well inside the unit interval.

Proposition 3 (Salience and tooling changes)**.** *Consider a common reduction in the implementation hurdle from κ to $\kappa - \delta$, with $\delta > 0$. Restrictions weakly expand for both uses and the ordering $R_T(q) \geq R_S(q)$ is preserved. If both post-change thresholds remain interior and distinct, the training-only region persists. A common hurdle reduction alone does not determine whether that region widens or narrows. Search catches up only under an additional change that lowers the search threshold more than the training threshold—for example, a relatively larger rise in search substitution cost or a tool that expands search restrictions disproportionately.*

Proof. Lowering κ weakens each restriction inequality and therefore weakly expands each restricting set. Assumption A2 continues to imply pointwise ordering. The final statements follow from comparing the changes in the two thresholds. □

Remark 2 (Not a lemons market)**.** *Nothing above requires the AI developer to be uninformed about q; source identity and measured quality are largely observable. The selection operates through participation: higher-quality publishers withdraw from uncompensated access, so the accessible pool becomes less representative even though no information is hidden. The framework is an organizing device for selective participation under a coarse access technology, not a welfare or equilibrium model.*