# Adverse Selection in the AI Data Commons

Kai Zhu*

**Abstract**

Generative AI depends on high-quality web content, yet no market compensates its producers. We document adverse selection in this AI data commons: facing a binary opt-out choice, the highest-quality producers exit first, degrading the remaining commons. Studying media and news sites at scale, we find a steep quality–blocking gradient: high-factual outlets block at nearly six times the rate of low-factual sources, with misinformation sources remaining most accessible. Publishers strategically target training crawlers while blocking search crawlers at lower rates. Event studies trace this divergence to the introduction of AI-specific opt-out mechanisms, and simulations show the quality degradation stems from the compositional pattern of exit. Current opt-out regimes produce adverse selection by design, with direct consequences for the quality of AI-mediated consumer information; market mechanisms that properly compensate content producers can substantially reverse the degradation.

**Keywords:** adverse selection, AI data commons, misinformation, media quality, robots.txt, data governance

*Bocconi University. Email: kai.zhu@unibocconi.it.

# 1 Introduction

Generative AI derives its power from training on high-quality web content. Yet no systematic market compensates the producers of that content: although licensing deals between AI companies and select publishers have begun to emerge, most content producers receive nothing for their contributions to training corpora (Kim et al., 2025). Without a functioning market, producers face a binary choice—permit AI crawlers to access their content, or block them via the `robots.txt` protocol. The critical question is not how many producers opt out, but *which ones*—and whether the pattern of exit systematically degrades the quality of what remains. We call this pool of accessible web content the *AI data commons*—a shared resource that, like other commons (Ostrom, 1990), is vulnerable to degradation when individual exit decisions are uncoordinated. If the highest-quality producers exit first, the result is a lemons problem (Akerlof, 1970): adverse selection operating through composition rather than price, with direct consequences for the quality of AI-mediated consumer information.

We study this problem in the media and news sector, where the stakes for information quality are highest. Media organizations produce original, fact-checked, and editorially curated content—the kind of high-quality human-generated text on which AI systems depend for factual grounding. When such content is replaced by lower-quality or synthetic alternatives, model outputs degrade (Shumailov et al., 2024). Media content also underpins the advertising-funded ecosystem: as AI-powered summarization threatens to reduce click-through traffic to original sources, the business model that produces quality journalism comes under pressure (Jain and Qian, 2021). If the highest-quality outlets withdraw while misinformation sources remain freely accessible, the result is algorithmic exclusion (Tucker, 2024)—degraded AI outputs propagating through every summary, search result, and recommendation that draws on this commons.

The `robots.txt` protocol is the *de facto* mechanism for managing AI crawler access. It is advisory and carries no binding legal force, though major training crawlers—GPTBot, ClaudeBot—generally respect blocking directives (Kim et al., 2025). AI companies deploy

*training crawlers* (GPTBot, Google-Extended, ClaudeBot) that permanently incorporate content into model weights, and *search crawlers* (OAI-SearchBot, PerplexityBot) that index content for AI-powered search. The economic distinction is critical: training extracts value from content without compensation, while search can drive traffic back to the original source—giving publishers sharply different incentives to block each type. The most significant event during our study period was OpenAI's GPTBot announcement (August 2023), which included explicit opt-out guidance. The *New York Times v. OpenAI* lawsuit (December 2023) and Cloudflare's one-click AI blocking tool (July 2024) reinforced the trend.

Using 9,611 media and news sites rated by Media Bias/Fact Check (MBFC) and a 45-month panel, we document six findings. First, a stark quality–blocking gradient: high-factual outlets block AI crawlers at 47%, compared to 8% for low-factual sources, a pattern consistent across all four quality measures. Second, MBFC-flagged questionable sources block at just 6%, versus 35% for credible sources, a 5.7× *misinformation asymmetry*. Third, blocking exhibits ideological sorting: centrist outlets block at the highest rates (51%) while political extremes remain most accessible, implying that AI training corpora are becoming ideologically skewed toward the poles.

Fourth, event studies trace the emergence of the quality–blocking gradient to OpenAI's GPTBot announcement (August 2023), which made AI-specific blocking salient and actionable; near-zero pre-announcement gradients gave way to sharp, persistent amplification, with post-period coefficients five to six times larger in magnitude. Fifth, high-quality publishers strategically target training crawlers over search crawlers—45% versus 32%—with training coefficients 38% larger across quality measures, consistent with protecting content assets while maintaining AI search visibility. Sixth, calibrated simulations show that the quality degradation of available media training data—a 24.3% decline in high-factual content share—is accounted for entirely by the compositional pattern of blocking: the same overall blocking rate applied randomly produces zero quality decline.

We contribute to three literatures. First, *data markets*: we provide the first systematic

evidence of adverse selection in AI training data. Work on data market failures has examined pricing inefficiencies and data externalities (Acemoglu et al., 2022; Bergemann and Bonatti, 2024); we identify a setting where the lemons problem operates through composition rather than pricing—quality-dependent exit degrades the commons (Goldfarb and Tucker, 2019). Research on privacy polarization shows opt-out decisions cluster among specific populations, degrading representativeness (Miklós-Thal et al., 2024); we find an analogous pattern among quality-sensitive publishers (Goldfarb and Que, 2023; Lefrere et al., 2025). Supply-side evidence from image platforms confirms the behavioral mechanism (Abeillon et al., 2024).

Second, *misinformation and media*: we identify a new upstream channel through which AI systems may amplify misinformation—not by choosing unreliable content, but because high-quality sources have systematically withdrawn from training data. The misinformation literature has studied how false news spreads and how platform design shapes the information environment (Vosoughi et al., 2018; Allcott and Gentzkow, 2017); we show that adverse selection tilts the training corpus toward unreliable content before any model is trained, a supply-side mechanism distinct from the demand-side dynamics previously documented.

Third, *content strategy and data governance*: training data composition is *endogenous* to the governance regime—opt-out mechanisms, even when individually rational, produce lemons equilibria by design. The training-versus-search differentiation reveals that publishers make strategic content distribution decisions—blocking value extraction while permitting value creation—not passively reacting to AI. Parallel evidence from image data confirms that copyright-based filtering similarly distorts training datasets (Shcherbakov et al., 2025). Statutory licensing emerges as welfare-superior to both pure opt-out regimes and uncompensated exceptions (Peukert, 2025); market mechanisms that properly compensate content producers could realign incentives and reverse the compositional degradation.

The remainder of the paper is organized as follows. Section 2 develops the theoretical framework and derives testable predictions. Section 3 describes the data. Section 4 presents the empirical results. Section 5 reports the simulation analysis. Section 6 discusses

implications and concludes.

## 2 Theory

We develop a simple model of adverse selection in AI training data to formalize the mechanism and derive testable predictions. Full propositions with proofs are in Online Appendix A; here we describe the key logic and predictions.

Content producers indexed by quality $q$ make a binary decision: permit AI crawlers to access their content or block. Each producer weighs two forces: a *visibility benefit* from remaining in the AI ecosystem (inclusion in AI-powered search, recommendations, and training data increases traffic and brand recognition) and an *appropriation cost* from AI training (the expected loss from uncompensated content incorporation into model weights, including competitive substitution risk and intellectual property concerns).

Two assumptions drive the result. First, the appropriation cost is strictly increasing and convex in quality—higher-quality producers face larger and accelerating costs because their content is more valuable, more costly to produce, and more susceptible to competitive substitution by AI-generated summaries. Second, the visibility benefit increases with quality, but more slowly than the appropriation cost: visibility benefits face diminishing returns as AI-driven traffic reaches a ceiling, while appropriation costs accelerate with the value of the content at risk. Because the net cost of remaining in the commons is strictly increasing in quality, a unique threshold $q^*$ exists such that all producers above it block and all below permit access. The average quality of the remaining commons falls strictly below the population mean: adverse selection. Since high-quality producers also produce more content, adverse selection imposes a volume penalty on top of the compositional penalty—the "double penalty" quantified in the simulation.

The training-versus-search distinction sharpens this logic. Training crawlers impose strictly greater appropriation costs than search crawlers at every quality level, and this gap

grows with quality: training involves permanent, non-attributed incorporation into model weights, whereas search indexing provides traffic benefits analogous to traditional search engines. The training-blocking threshold is therefore lower than the search-blocking threshold, and there exists a range of producers who block training crawlers while permitting search crawlers—precisely those who find it worthwhile to protect their intellectual property from model training while maintaining AI search visibility.

Events that increase the salience of appropriation costs—such as the GPTBot announcement, which made opt-out salient and actionable for the first time—shift the cost function upward, lowering the blocking threshold. Because the cost shift is at least as large for higher-quality producers (whose content is more visibly at risk), the producers who switch from permitting to blocking are drawn from the top of the remaining quality distribution, amplifying the quality–blocking gradient at the moment the shock hits.

The model generates four predictions tested in the empirical analysis:

**P1 (Quality–Blocking Gradient):** Higher-quality producers are more likely to block (Section 4.1).

**P2 (Training–Search Gap):** The training–search blocking gap widens with quality (Section 4.5).

**P3 (Event-Driven Amplification):** A salience shock amplifies the quality–blocking gradient (Section 4.4).

**P4 (Corpus Degradation):** Quality-dependent blocking degrades the remaining commons, with degradation increasing in the blocking rate (Section 5).

# 3 Data

Our analysis uses two complementary samples of MBFC-rated media sites. The *cross-section*, observed as of February 2026, comprises 9,611 sites—all MBFC-rated media domains—with blocking status obtained via live scraping of each site's `robots.txt` file, providing maximum

statistical power for cross-sectional analyses. The *panel* comprises 7,002 of these sites tracked monthly from June 2022 to February 2026 via HTTP Archive crawls (272,458 site-month observations across 45 months), enabling temporal analysis. The panel is a subset of the cross-section; the remaining 2,609 sites lack consistent HTTP Archive coverage across the panel window. The panel is unbalanced; balanced-panel robustness checks confirm all results (Online Appendix F.2).

We integrate six external data sources, organized around three dimensions. For *blocking behavior*, *HTTP Archive* provides monthly crawls of millions of websites, recording `robots.txt` files; we parse these to construct the panel, identifying which of 23 distinct AI crawlers each site blocks and classifying crawlers as training-purpose (GPTBot, Google-Extended, ClaudeBot) or search-purpose (OAI-SearchBot, PerplexityBot). For *editorial quality*, *Media Bias/Fact Check (MBFC)* is an independent media-monitoring organization that provides human-validated ratings of factual reporting (six levels from "very low" to "very high"), editorial credibility (three levels), political orientation (five categories from left to right), and questionable-source flags (conspiracy, fake news, propaganda, pseudoscience, satire). For *web authority*, *Majestic Million* provides Trust Flow scores (0–100) based on backlink quality; *Open PageRank* provides web graph centrality scores; *Tranco* is a research-oriented top-sites ranking aggregating multiple popularity lists to reduce manipulation risk; and the *Wikipedia API* determines whether each site has a dedicated article, serving as an independent notability signal.

The cross-section includes 4,794 high-factual, 1,016 mostly-factual, 2,528 mixed-factual, and 1,273 low-factual sites; 1,984 carry MBFC questionable-source flags. Cross-sectional analyses (quality–blocking regressions, misinformation asymmetry, simulation calibration) use the full 9,611 sites; panel analyses (temporal dynamics, event studies, training vs. search) use the 7,002-site panel.[1]

We employ four complementary quality measures, enabling triangulation across differ-

[1]Online Appendix B provides data source details, descriptive statistics, per-bot blocking rates, and PCA construction.

ent quality dimensions. Two are ordinal MBFC ratings: *MBFC factual reporting* (six-point scale, coded 1–6) and *MBFC credibility* (three-point scale, coded 1–3), both reflecting expert editorial judgment about content quality. Two are composite indices constructed via principal component analysis, each standardized to mean zero and unit variance so that coefficients are interpretable as the effect of a one-standard-deviation increase: the *domain quality index* (first PC of Trust Flow, Open PageRank, and inverse Tranco rank—objective web-graph metrics independent of editorial content) and the *content quality index* (first PC of MBFC factual, MBFC credibility, and Wikipedia notability—editorial assessment augmented by a notability signal). The two composite indices are conceptually distinct: the domain quality index is an "outside-in" measure reflecting how the web treats the site, while the content quality index is an "inside-out" measure of what the site produces. All quality metrics are measured prior to or independently of AI blocking decisions, avoiding reverse causality.

Our primary dependent variable, $\text{Block}_i$, equals one if a site blocks at least one AI crawler in the February 2026 cross-section. We exclude CCBot (Common Crawl) from all blocking measures because its widespread blocking predates AI-specific concerns and reflects bandwidth management and general scraping aversion rather than deliberate AI opt-out decisions. Content producers may also protect their content through IP blocking, paywalls, or reducing content production; we focus on `robots.txt` because it provides the cleanest signal—directly observable, standardized across sites, and targeting specific AI crawlers by name, enabling the training-versus-search distinction central to our analysis. Overall, 28.9% of MBFC-rated media sites block at least one AI bot. We also examine $\text{BlockTrain}_i$ (blocks any training bot) and $\text{BlockSearch}_i$ (blocks any search bot) as separate outcomes, available for both the cross-section and the panel, to analyze selective blocking behavior. GPTBot is the most-blocked individual crawler at 32.3%, followed by ClaudeBot (30.0%) and Google-Extended (28.1%); per-bot rates are detailed in Online Appendix B.2.

# 4 Results

## 4.1 The Quality–Blocking Gradient

We begin with model-free evidence of the quality–blocking gradient (P1). Figure 1 presents blocking rates by MBFC factual reporting level and credibility rating across the full 9,611-site cross-section. The pattern is striking in both dimensions: blocking rises monotonically with quality, from 8% among low-factual sources to 47% among high-factual outlets. The credibility dimension tells the same story—low-credibility sites block at just 7%, while high-credibility outlets block at 45%.

Cross-sectional regressions with category fixed effects confirm the gradient across all four quality measures (full table in Online Appendix C.1). A one-unit increase in MBFC factual reporting is associated with a 13.4 percentage point increase in blocking probability ($p < 0.01$); MBFC credibility yields 19.7 pp per level ($p < 0.01$). A one-standard-deviation increase in the content quality index is associated with a 17.6 pp increase ($p < 0.01$); the domain quality index yields 7.8 pp ($p < 0.01$). Four distinct measures—two based on editorial assessment, one combining editorial and notability signals, and one based entirely on web-graph and traffic metrics—all produce significant positive coefficients, ruling out the possibility that the gradient is an artifact of any single measurement approach.

A natural concern is that the gradient reflects site *size* (and associated resources) rather than quality per se, since large sites have IT teams to implement blocking. We test this directly by estimating quality–blocking coefficients within site-size terciles (Online Appendix F.3); the gradient persists among small, medium, and large sites, ruling out a pure resources explanation. Additional robustness checks including logit models, balanced-panel restrictions, and placebo permutation tests (1,000 random shuffles of quality ratings, permutation $p < 0.001$) confirm these results (Online Appendix F.1, F.2, F.6).

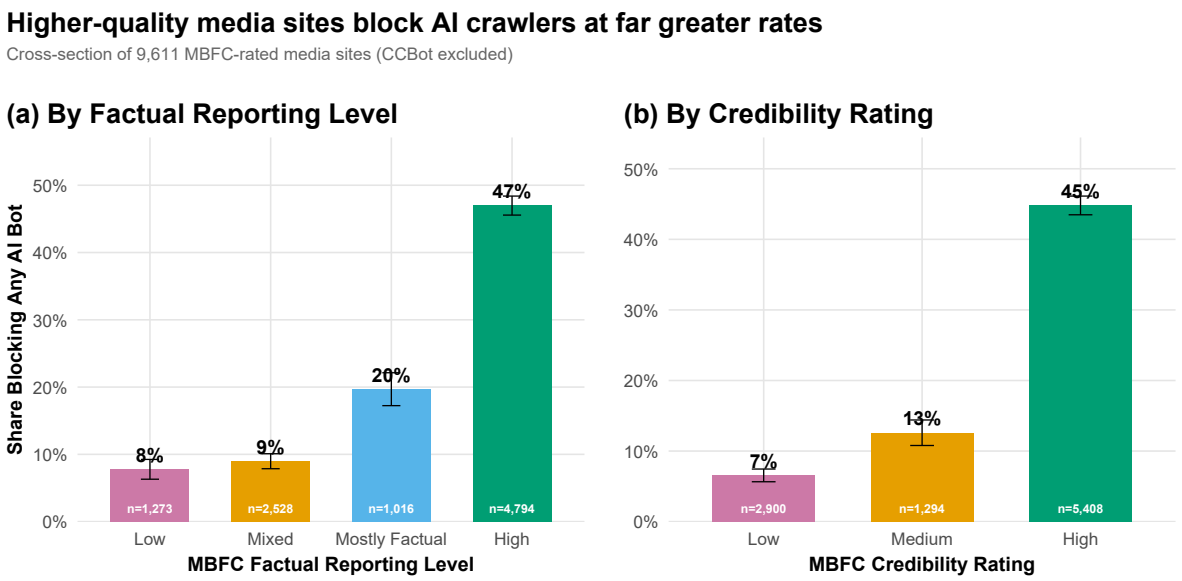



Figure 1: The quality–blocking gradient in media sites. Blocking rises monotonically from 8% (low-factual) to 47% (high-factual) and from 7% (low-credibility) to 45% (high-credibility). $N = 9{,}611$ MBFC-rated media sites, cross-section.

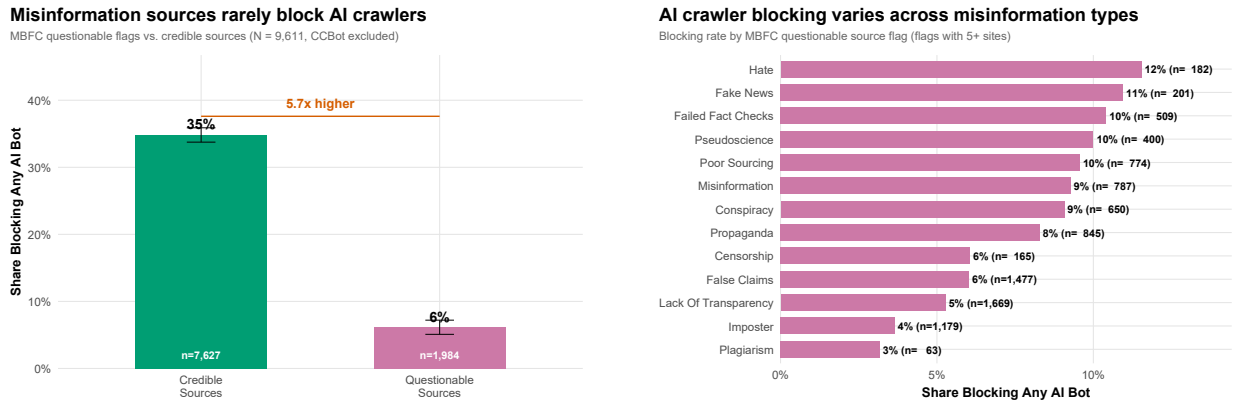
## 4.2 The Misinformation Asymmetry

The most consequential manifestation of adverse selection is what we term the *misinformation asymmetry*: MBFC-flagged questionable sources—sites identified as promoting conspiracy theories, fake news, propaganda, or pseudoscience—block AI crawlers at dramatically lower rates than credible sources. Figure 2 quantifies this gap. Panel (a) shows that credible media outlets block at 34.8%, compared to just 6.1% for questionable sources—a 5.7× difference. This implies that for every questionable source that blocks, nearly six credible sources do the same, systematically shifting the composition of accessible training data toward unreliable content.

Panel (b) disaggregates questionable sources by specific MBFC flag type. Blocking rates are uniformly low across all misinformation categories, ranging from 3% (plagiarism, imposter sites) to 12% (hate speech, fake news). Even the highest-blocking misinformation category falls far below the 35% average for credible sources. This consistency across mis-

information types indicates that the asymmetry reflects a systematic pattern: sources that traffic in misinformation—regardless of type—have little incentive to restrict AI access to their content.

Regression analysis (Online Appendix C.4) shows that conditional on the content quality index, questionable-source status has no independent effect on blocking probability. The misinformation asymmetry is thus a downstream consequence of adverse selection in the quality dimension rather than an independent phenomenon.




(a) Credible vs. questionable sources

(b) By questionable-source flag type

Figure 2: The misinformation asymmetry. (a) Credible outlets block at 5.7× the rate of questionable sources. (b) All misinformation flag types exhibit low blocking rates, from 3% to 12%. $N = 9{,}611$ sites; 1,984 flagged as questionable.

## 4.3 Ideological Sorting

Beyond the model's predictions, political orientation independently predicts blocking behavior. Figure 3 presents blocking rates by MBFC political orientation among the 4,650 sites with bias ratings. The pattern is sharply non-linear: center outlets block at the highest rate (51.1%), followed by right-center (48.8%) and left-center (39.1%). Left-leaning outlets block at 23.7%, while right-leaning outlets block at just 10.8%. This inverted-U shape—with political extremes remaining most accessible to AI crawlers—implies that AI training data is becoming ideologically sorted, with moderate, centrist reporting withdrawing while content from both poles remains.

Regressions controlling for both quality indices with category fixed effects (Online Appendix C.5) show that content quality explains much of the political gradient—centrist outlets tend to be higher quality—but a meaningful residual gradient persists even after controlling for both quality dimensions. Left outlets block 23 pp below center ($p < 0.01$) and right outlets 19 pp below center ($p < 0.01$) after quality controls, suggesting that organizational characteristics correlated with political orientation—such as revenue model, institutional norms, or legal strategy—also shape blocking decisions.

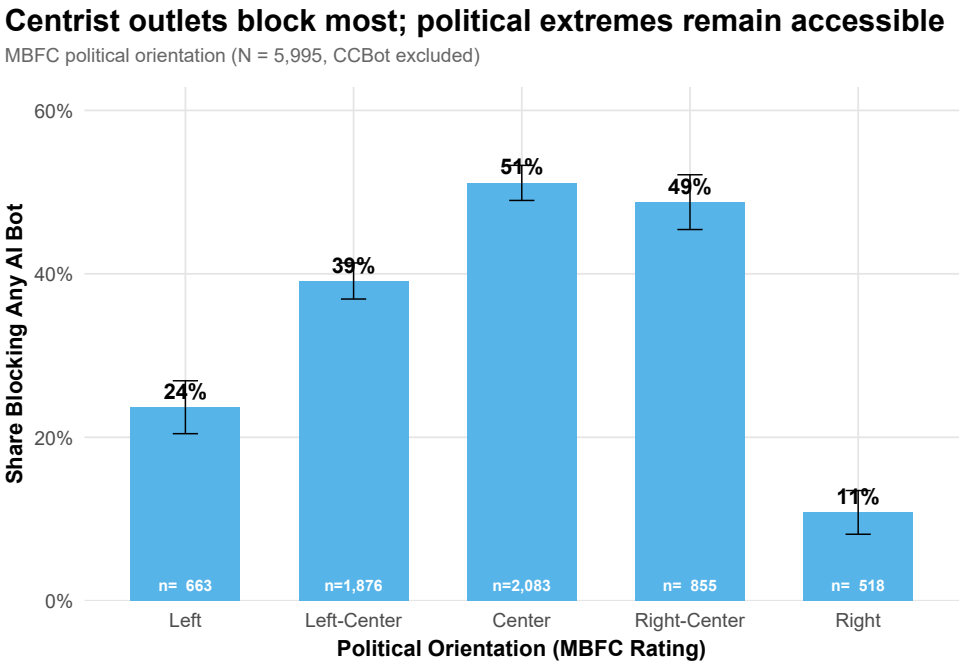



Figure 3: AI crawler blocking by political orientation. Center outlets block at the highest rates (51%); political extremes remain most accessible (right: 11%). $N = 4{,}650$ sites with MBFC bias ratings.

## 4.4 Temporal Dynamics

The cross-sectional evidence establishes the current landscape of adverse selection; we now test P3 by examining *when* these patterns emerged. Across the full 45-month panel, blocking rates were uniformly low prior to the GPTBot announcement (August 2023)—fewer than 10% of sites blocked any AI crawler, with no meaningful quality gradient. The an-

nouncement triggered rapid differential adoption: high-factual sites moved first and fastest, reaching over 50% blocking within 12 months, while low-factual sites remained below 10% (Online Appendix D). The gradient has continued to widen after the initial shock rather than plateauing.

To formally test when the quality–blocking gradient emerged, we estimate event studies interacting quality with monthly event-time dummies relative to the GPTBot announcement (full results in Online Appendix D):

$$\text{Block}_{it} = \sum_{\tau \neq \tau_0} \beta_\tau \, \text{Quality}_i \times \mathbf{1}[t = \tau] + \gamma_c + \delta_t + \varepsilon_{it},$$

where $\tau_0$ is the reference period (earliest month), $\gamma_c$ are category fixed effects, $\delta_t$ are month fixed effects, and standard errors are clustered at the site level. The GPTBot announcement provides a plausibly exogenous shock to the salience and feasibility of AI-specific blocking; the event study tests whether the gradient amplified at that moment. All four measures show the same pattern: small pre-announcement coefficients followed by sharp amplification. Post-period coefficients are five to six times larger in magnitude than pre-period coefficients across all measures, with bootstrap confidence intervals excluding unity. The consistency across content-based measures (factual, credibility, content quality index) and web-authority measures (domain quality index) confirms robustness. Rambachan and Roth (2023) sensitivity analysis confirms that post-period effects remain significant even under conservative assumptions about pre-trend extrapolation. Detrended event studies that subtract the pre-period linear trend confirm the structural break persists. Site fixed effects specifications yield strikingly similar patterns, ruling out time-invariant confounders.

Panel regressions exploiting the full 272,458 site-month panel confirm the gradient (Online Appendix C.2). Quality coefficients range from 0.046 (domain quality index) to 0.135 (MBFC credibility), all significant at $p < 0.01$. A site fixed effects specification that replaces time-invariant quality levels with a quality $\times$ post-GPTBot interaction provides additional

evidence: the interaction term is strongly positive (0.049, $p < 0.01$), confirming that the gradient widened after the introduction of AI-specific crawlers, net of all time-invariant site characteristics. The post-2023 subsample yields even stronger coefficients, indicating that adverse selection has *intensified* over time rather than attenuating.

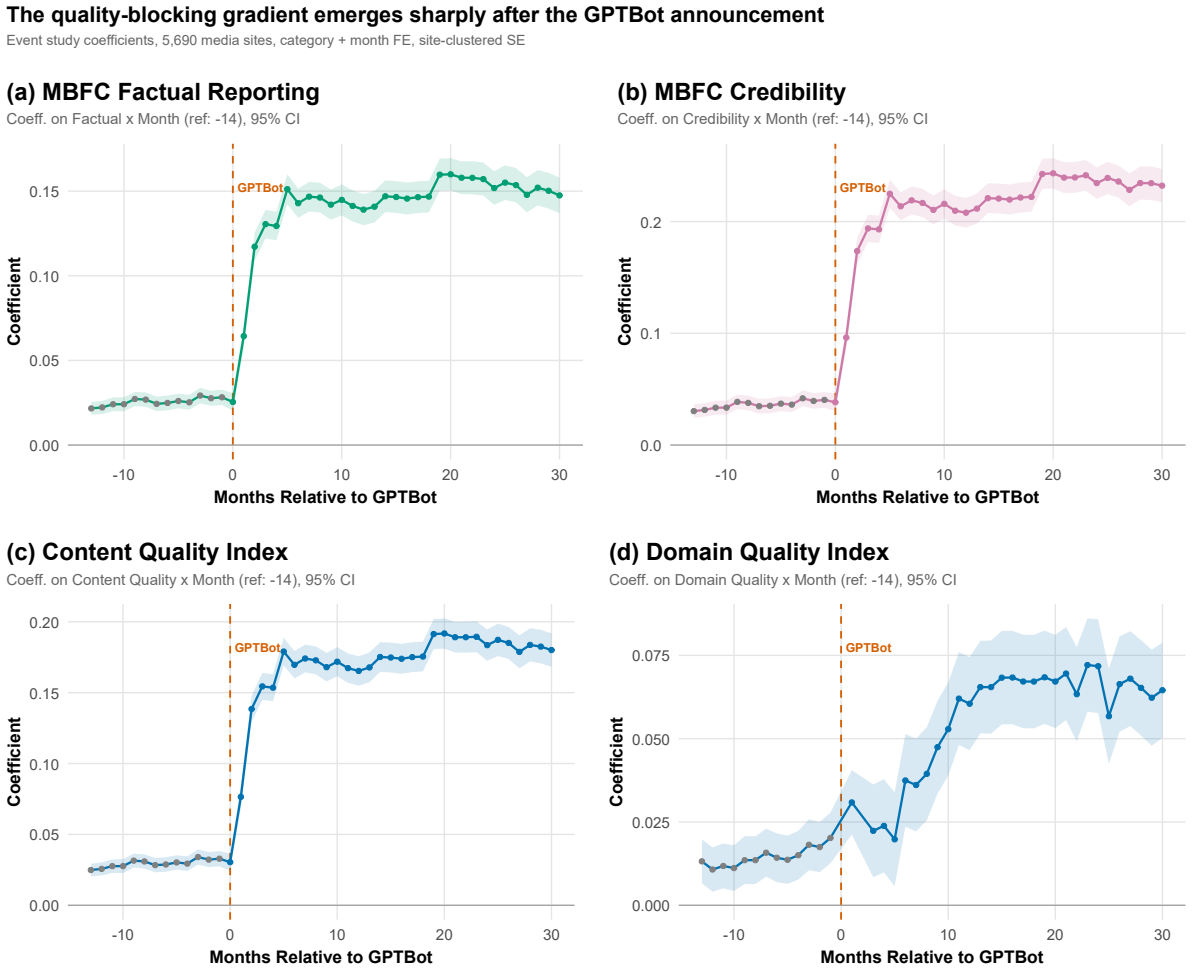



Figure 4: Event study: quality–blocking gradient before and after the GPTBot announcement (August 2023). Each panel plots $\beta_\tau$ coefficients from the interaction of a quality measure with monthly event-time dummies. Pre-announcement coefficients are economically small; post-announcement coefficients are five to six times larger. $N = 272{,}458$ site-months, 7,002 sites.

## 4.5 Training versus Search

The training-versus-search distinction provides a direct test of P2: if appropriation costs drive blocking, publishers should target training crawlers—which permanently incorporate content into model weights—more aggressively than search crawlers, which can drive traffic back. Among high-factual sites, 45.2% block training bots compared to 32.4% blocking search bots—a 12.8 percentage point gap that widens monotonically with quality (Figure 5).

Panel regressions confirm the pattern across all four quality measures (Online Appendix C.3): training coefficients are approximately 38% larger than search coefficients. This is not blanket technology rejection; it is targeted protection of content assets while maintaining AI search visibility. Publishers differentiate between crawlers that extract value (training) and those that can create value (search)—a strategic content distribution decision consistent with the model's prediction that training appropriation costs exceed search costs, especially for high-quality producers.

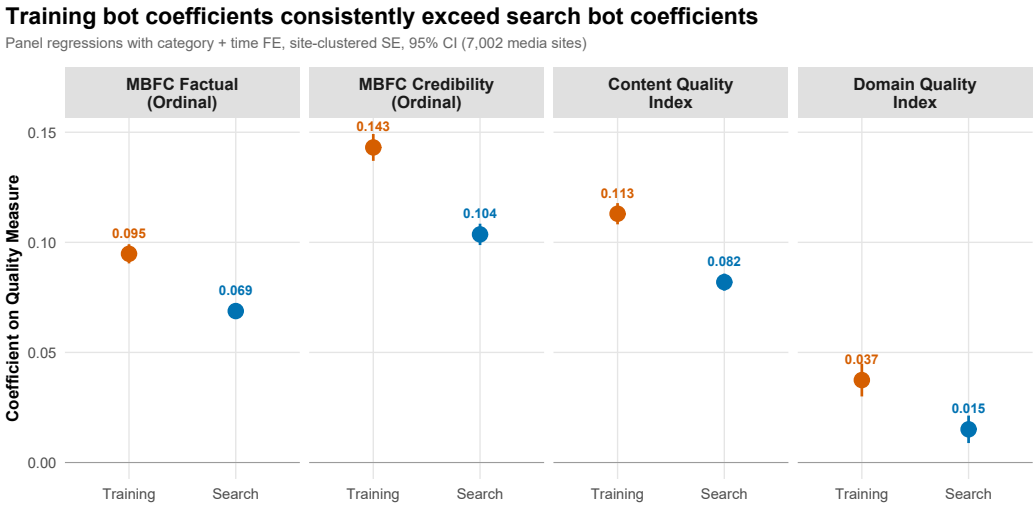



Figure 5: Training vs. search bot blocking coefficients across quality measures. Training coefficients are approximately 38% larger than search coefficients, confirming that high-quality publishers selectively target training crawlers. $N = 272{,}458$ site-months, 7,002 sites.

# 5 Simulation

The empirical results demonstrate *which* sites block; simulation tests P4 by translating these site-level decisions into corpus-level quality consequences. A site-level blocking decision affects corpus-level quality through two channels: the quality of the content withdrawn and the volume of tokens that content represents. We model 10,000 simulated media sites, each assigned to one of four MBFC factual tiers (Low, Mixed, Mostly Factual, High) with empirical proportions from the cross-section. Each tier has three calibrated parameters: a blocking rate equal to the observed tier-level mean, a quality score based on normalized MBFC factual ordinal, and a content volume drawn from a lognormal distribution centered on the tier-specific mean Trust Flow.[2]

Table 1 reports results across five counterfactual scenarios averaged over 1,000 Monte Carlo replications. The key comparison is between current observed blocking and random blocking—both block at the same overall rate (approximately 41%), but random blocking applies this rate uniformly across tiers. Current observed blocking produces a 24.3% decline in high-factual content share and a 9.7% decline in token-weighted average quality, while random blocking produces zero quality decline. The difference is stark: random blocking removes content proportionally across tiers, leaving composition unchanged, while quality-dependent blocking systematically depletes the corpus of its most reliable content. *The entire quality degradation is attributable to adverse selection, not to blocking per se.*

The licensing deal scenario, which halves blocking rates for High and Mostly Factual tiers, recovers most of the loss (7.0% vs. 24.3% high-factual decline), demonstrating that targeted interventions can substantially reverse the adverse selection dynamic. The universal blocking scenario—applying the High-tier rate uniformly—confirms that even at elevated blocking levels, uniform blocking barely degrades quality. The problem is the asymmetry, not the level.

[2]High-factual sites produce approximately 11× the estimated token volume of low-factual sources. Calibration parameters are computed from the 5,782 cross-sectional sites with complete domain quality data; see Online Appendix E.1 for details.

Figure 6 illustrates the mechanism directly: both distributions are generated at the *same* overall blocking rate, but quality-dependent blocking shifts the distribution leftward, depleting the right tail of high-factual content. This leftward shift *is* the adverse selection effect.

The volume dimension reveals a *double penalty*. Current observed blocking and random blocking withdraw nearly identical shares of total token volume (41.3% and 41.1%), yet only current observed blocking degrades quality. Adverse selection thus imposes both a quantity penalty (shared by any blocking regime) and a compositional penalty (unique to quality-dependent blocking).

Dynamic projections to 2028 show continued quality degradation across deceleration, baseline, and acceleration scenarios, with the high-factual content share projected to fall to approximately one-third of available media training tokens by late 2028 under the baseline scenario (Online Appendix E.2).

Table 1: Media Training Data Quality Under Five Blocking Scenarios

| Scenario | Quality Decline (%) | High-Factual Decline (%) | Volume Decline (%) |
|---|---|---|---|
| No blocking | 0.0 | 0.0 | 0.8 |
| Current observed | 9.7 | 24.3 | 41.3 |
| Random blocking | 0.0 | 0.0 | 41.1 |
| Licensing deal | 2.7 | 7.0 | 22.4 |
| Universal blocking | 0.0 | 0.0 | 55.6 |

*Notes:* Tier-based simulation calibrated from 5,782 MBFC-rated media sites. N=10,000 simulated sites, 1,000 replications. CCBot excluded. Quality and high-factual declines relative to no-blocking baseline. Volume decline is the percentage of total token volume withdrawn by blocking. Random blocking applies the overall average rate (41%) uniformly across all tiers. Licensing deal halves blocking rates for High and Mostly Factual tiers. Universal blocking applies the High-tier rate (56%) to all tiers.

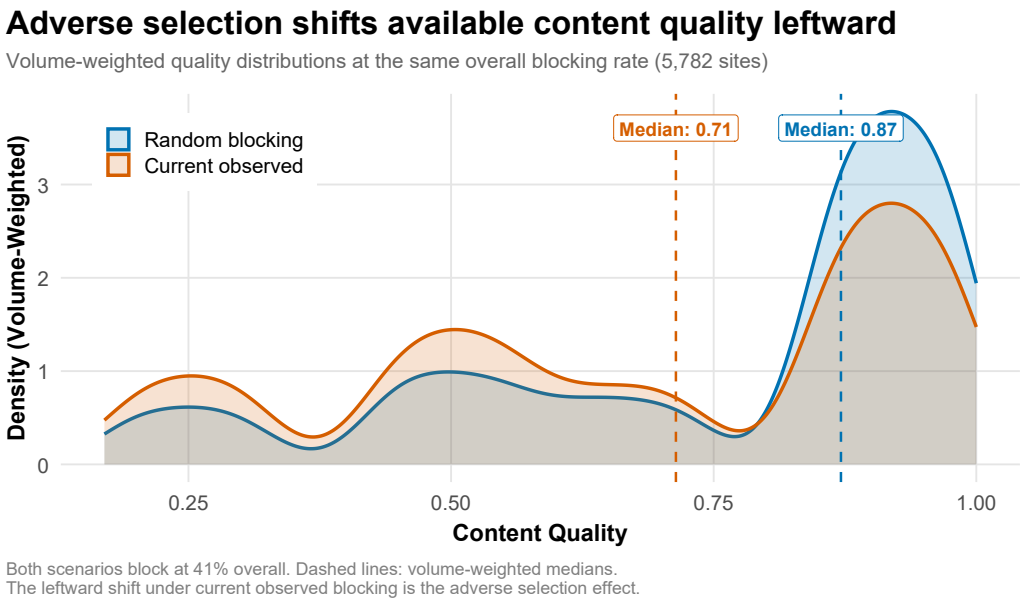



Figure 6: Quality distribution of available media content under random vs. current observed blocking. Both scenarios block at the same overall rate. The leftward shift under quality-dependent blocking is the adverse selection effect.

# 6 Discussion

Our findings establish that the AI data commons suffers from adverse selection. The quality–blocking gradient is steep (47% vs. 8%) and consistent across four independent quality measures, including an MBFC-independent domain quality index (P1). The gradient emerged sharply following the GPTBot announcement, with post-event coefficients five to six times larger than pre-period values (P3). Publishers strategically differentiate between crawlers that extract value and those that can create it—blocking training crawlers at approximately 38% higher rates than search crawlers (P2). Calibrated simulations confirm that the entire quality degradation of the training corpus is attributable to this compositional pattern: the same blocking rate applied randomly produces zero quality decline (P4). Beyond these predictions, the misinformation asymmetry—a 5.7× gap between credible and questionable sources—and ideological sorting further characterize the distortion.

From a market design perspective, this pattern reveals a distinctive form of informa-

tion asymmetry. Unlike traditional adverse selection where product quality is unobservable to buyers, here it is the *sellers* who possess superior information about the value of their content—and this private information drives their exit decisions (Akerlof, 1970). The `robots.txt` mechanism, designed as a simple access management tool, has become the vehicle through which high-quality producers leave the market. The commons does not merely shrink as publishers opt out; it becomes systematically less factual, less credible, and more ideologically skewed—a compositional market failure with direct consequences for the consumer information environment (Goldfarb and Que, 2023). This pattern extends beyond media: audits of image training data show that copyright-based filtering similarly introduces systematic bias, concentrating content in narrow semantic clusters (Shcherbakov et al., 2025). Adverse selection in training data provides the upstream mechanism for algorithmic exclusion (Tucker, 2024): AI systems trained on degraded corpora produce systematically less factual outputs not because they choose misinformation, but because the credible alternatives have withdrawn. Because the misinformation asymmetry is a downstream consequence of quality-dependent blocking, any intervention that addresses the quality–blocking gradient would simultaneously ameliorate the misinformation problem.

These results point to specific design principles for data governance. The training-versus-search differentiation reveals that publishers block value extraction while permitting value creation—licensing frameworks should preserve this distinction. Our simulation provides quantified evidence: the licensing deal scenario, which halves blocking rates for high-quality tiers, recovers most of the quality loss (7.0% vs. 24.3% high-factual decline), demonstrating that targeted interventions can substantially reverse the adverse selection dynamic. Formal welfare analysis confirms that statutory licensing is welfare-superior to both pure opt-out regimes and uncompensated exceptions (Peukert, 2025). Collective licensing mechanisms could lower transaction costs, while compensation frameworks such as data dividends could realign incentives. More broadly, opt-out mechanisms produce lemons equilibria by design—regulatory frameworks including the EU AI Act should account for adverse selection in

training data composition.

Several limitations merit acknowledgment. First, `robots.txt` compliance is voluntary and varies across crawlers, though major training crawlers generally respect blocking (Kim et al., 2025). To the extent that high-quality publishers also pursue additional protection strategies (IP blocking, paywalls), our measured adverse selection is a conservative lower bound. Second, our simulation simplifies blocking into a four-tier structure; sensitivity analysis confirms robustness (Online Appendix E.5), but quantitative magnitudes should be interpreted as illustrative. Third, we document degradation of training data inputs, not model output consequences. Future research should measure downstream effects on AI factual accuracy, evaluate whether licensing agreements reduce blocking among participating publishers, and study cross-country variation under differing intellectual property regimes.

The fundamental tension is not between AI companies and content producers—it is the absence of a functioning market between them. Without compensation, the `robots.txt` opt-out mechanism replaces price signals with a binary choice, and the result, as we have documented, is textbook adverse selection. Our evidence shows that producers respond to incentives: they make strategic, differentiated blocking decisions, and even partial compensation substantially reverses quality degradation. Establishing market mechanisms that properly compensate content producers for their contributions to AI training—whether through licensing frameworks, collective bargaining, or regulatory mandates—is not merely desirable but necessary. The quality of the AI-mediated information environment depends on it.

# References


Abeillon, D., T. Haese, M. Kaiser, C. Peukert, and K. Staub (2024). Strategic behavior and AI training data. Working paper, HEC Lausanne.

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics 14*(4), 218–256.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics 84*(3), 488–500.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*(2), 211–236.

Bergemann, D. and A. Bonatti (2024). Data, competition, and digital platforms. *American Economic Review 114*(8), 2553–2595.

Goldfarb, A. and V. F. Que (2023). The economics of digital privacy. *Annual Review of Economics 15*, 267–286.

Goldfarb, A. and C. Tucker (2019). Digital economics. *Journal of Economic Literature 57*(1), 3–43.

Jain, S. and K. Qian (2021). Compensating online content producers: A theoretical analysis. *Management Science 67*(11), 7075–7090.

Kim, T., K. Bock, C. Luo, A. Liswood, C. Poroslay, and E. Wenger (2025). Scrapers selectively respect robots.txt directives: Evidence from a large-scale empirical study. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*.

Lefrere, V., L. Warberg, C. Cheyre, V. Marotta, and A. Acquisti (2025). Does privacy regulation harm content providers? a longitudinal analysis of the impact of the GDPR. *Management Science*. Forthcoming.

Miklós-Thal, J., A. Goldfarb, A. Haviv, and C. Tucker (2024). Digital hermits. *Marketing Science 43*(4), 697–708.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Peukert, C. (2025). The economics of copyright and AI: Empirical evidence and optimal policy. In-depth analysis PE 778.859, European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs.

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *The Review of Economic Studies 90*(5), 2555–2591.

Shcherbakov, V., I. Dalaud, and C. Peukert (2025). AI needs better data than the law allows. Working paper, University of Lausanne.

Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal (2024). AI models collapse when trained on recursively generated data. *Nature 631*, 755–759.

Tucker, C. (2024). Algorithmic exclusion: The fragility of algorithms to sparse and missing data. Report, Brookings Institution.

Vosoughi, S., D. Roy, and S. Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151.

# Online Appendix to:

# Adverse Selection in the AI Data Commons

Kai Zhu*

This online appendix provides the full theoretical model, detailed data documentation, complete regression tables, event study analyses, simulation details, and robustness checks for the main paper.

## Contents

**A Theoretical Model** **4**

A.1 Setup . . . . . 4

A.2 The Blocking Decision . . . . . 4

A.3 Training vs. Search Crawlers . . . . . 5

A.4 Event-Driven Amplification . . . . . 6

A.5 Empirical Predictions . . . . . 7

**B Data Details** **8**

B.1 Data Sources . . . . . 8

B.2 Per-Bot Blocking Rates . . . . . 9

B.3 Quality Measure Construction . . . . . 9

B.4 Descriptive Statistics . . . . . 10

*Bocconi University. Email: kai.zhu@unibocconi.it.

**C Additional Results** **12**

C.1 Cross-Sectional Regressions . . . 12

C.2 Panel Regressions . . . 12

C.3 Training vs. Search Bot Regressions . . . 12

C.4 Misinformation Regression . . . 17

C.5 Political Orientation Regressions . . . 17

C.6 Alternative Explanations . . . 18

**D Event Study Details** **19**

D.1 Specification . . . 19

D.2 Event Study Combined Figure . . . 19

D.3 Pre-Trend Tests . . . 20

D.4 Pre/Post Magnitude Comparison . . . 21

D.5 Rambachan and Roth Sensitivity Analysis . . . 22

D.6 Detrended Event Study . . . 22

D.7 Site Fixed Effects Event Study . . . 23

**E Simulation Details** **24**

E.1 Tier Calibration . . . 24

E.2 Quality Trajectory . . . 25

E.3 Composition Trajectory . . . 26

E.4 Factual Trajectory to 2028 . . . 26

E.5 Sensitivity Analysis . . . 27

**F Robustness** **28**

F.1 Logit Models . . . 28

F.2 Balanced Panel . . . 30

F.3 Size Heterogeneity . . . 30

F.4 Political Sample Selection . . . 32

F.5 Blocking Intensity . . . . . 32
F.6 Placebo Test . . . . . 35
F.7 August 2024 Data Repair . . . . . 36

# A Theoretical Model

## A.1 Setup

Consider a continuum of content producers indexed by quality $q \in [0, 1]$, distributed according to a continuous CDF $F(q)$ with positive density everywhere. Each producer makes a binary decision: permit AI crawlers to access their content ($b_i = 0$) or block ($b_i = 1$). Content volume $v(q)$ is strictly increasing in quality, reflecting the empirical regularity that high-quality publishers produce more content.[1]

## A.2 The Blocking Decision

Each producer weighs two forces. First, a *visibility benefit* $t(q) \geq 0$ from remaining in the AI ecosystem—inclusion in AI-powered search results, recommendation systems, and training data increases traffic and brand recognition. Second, an *appropriation cost* $c(q) \geq 0$ from AI training—the expected loss from having content incorporated into model weights without compensation, including competitive substitution risk and intellectual property concerns.

We assume:

1. $c(q)$ is strictly increasing and convex: higher-quality producers face larger and accelerating appropriation costs, because their content is more valuable, more costly to produce, and more susceptible to substitution by AI-generated summaries.
2. $t(q)$ is weakly increasing but with $t'(q) < c'(q)$ for all $q$: while visibility benefits also increase with quality, they do so more slowly than appropriation costs—visibility benefits face diminishing returns as AI-driven traffic reaches a ceiling, while appropriation costs accelerate with the value of the content at risk.
3. $\pi(0) = c(0) - t(0) < 0$ and $\pi(1) = c(1) - t(1) > 0$: the lowest-quality producers find it worthwhile to remain in the commons (negligible appropriation cost relative to

[1] In our data, high-factual media sites produce approximately 11× the estimated token volume of low-factual sources, as proxied by Trust Flow.

visibility benefits), while the highest-quality producers prefer to block (appropriation cost dominates).

A producer blocks if and only if $c(q) > t(q)$. Define $\pi(q) \equiv c(q) - t(q)$, the net cost of remaining in the commons.

**Proposition 1** (Adverse Selection)**.** *Under assumptions 1–3, there exists a unique threshold $q^* \in (0,1)$ such that all producers with $q > q^*$ block and all producers with $q < q^*$ permit access. The average quality of the remaining commons, $\bar{q}_{avail} = \mathbb{E}[q \mid q < q^*]$, is strictly below the population mean $\bar{q} = \mathbb{E}[q]$.*

*Proof.* Since $\pi(q)$ is strictly increasing and continuous (by assumptions 1 and 2), and $\pi(0) < 0$ while $\pi(1) > 0$ (by assumption 3), by the intermediate value theorem $\pi$ crosses zero exactly once at some $q^* \in (0,1)$. All producers with $q > q^*$ have $\pi(q) > 0$ and block; all with $q < q^*$ permit access. The average quality of available content is $\bar{q}_{\text{avail}} = \mathbb{E}[q \mid q < q^*] < \mathbb{E}[q] = \bar{q}$ because the distribution is truncated from above. □

**Corollary 1** (Volume Penalty)**.** *If content volume $v(q)$ is strictly increasing, the share of total content volume withdrawn by blocking exceeds the share of producers who block: adverse selection imposes a volume penalty on top of the compositional penalty.*

*Proof.* Blocked producers have $q > q^*$ and thus above-population-average content volume: $\mathbb{E}[v(q) \mid q > q^*] > \mathbb{E}[v(q)]$, since $v$ is strictly increasing and the blocked set is the upper tail of the quality distribution. The volume share of blocked content, $(1 - F(q^*)) \cdot \mathbb{E}[v(q) \mid q > q^*]/\mathbb{E}[v(q)]$, therefore exceeds the producer share $1 - F(q^*)$. □

## A.3 Training vs. Search Crawlers

Content producers face two types of AI access. *Training crawlers* incorporate content into model weights, creating a training appropriation cost $c_T(q)$. *Search crawlers* index content for AI-powered search, providing direct visibility benefits but at a lower appropriation cost $c_S(q)$. We add two assumptions:

4. $c_T(q) > c_S(q)$ for all $q > 0$: training use creates strictly greater appropriation cost because it involves permanent, non-attributed incorporation into model weights, whereas search indexing provides traffic benefits analogous to traditional search engines.
5. $c'_T(q) > c'_S(q)$ for all $q$: training appropriation costs grow faster with quality than search appropriation costs, because the IP risk from permanent incorporation accelerates more steeply with content value than the risk from search indexing.

**Proposition 2** (Training vs. Search Blocking)**.** *Under assumptions 1–5, the training-blocking threshold satisfies $q_T^* < q_S^*$: more producers block training crawlers than search crawlers. There exists an interval $[q_T^*, q_S^*]$ of producers who block training crawlers while permitting search crawlers.*

*Proof.* Define $\pi_T(q) \equiv c_T(q) - t(q)$ and $\pi_S(q) \equiv c_S(q) - t(q)$. By assumption 4, $\pi_T(q) > \pi_S(q)$ for all $q > 0$, so $\pi_T$ crosses zero at a lower threshold: $q_T^* < q_S^*$. Producers with $q \in [q_T^*, q_S^*]$ have $\pi_T(q) > 0$ but $\pi_S(q) < 0$, so they block training crawlers while permitting search crawlers. By assumption 5, $\pi'_T(q) - \pi'_S(q) = c'_T(q) - c'_S(q) > 0$, so the gap $\pi_T(q) - \pi_S(q)$ widens with quality. □

*Remark.* In a stochastic extension where each producer faces an idiosyncratic cost shock, blocking becomes probabilistic. The deterministic thresholds $q_T^* < q_S^*$ imply that the blocking probability gap $\Pr(\text{block training} \mid q) - \Pr(\text{block search} \mid q)$ is positive and increasing in $q$ when $c'_T(q) > c'_S(q)$, as tested in the empirical analysis.

## A.4 Event-Driven Amplification

Events that increase the salience of appropriation costs—such as the GPTBot announcement—shift the cost function upward.

**Proposition 3** (Event-Driven Amplification)**.** *If a salience shock shifts the appropriation cost from $c(q)$ to $\tilde{c}(q) = c(q) + \Delta(q)$ with $\Delta(q) > 0$ and $\Delta'(q) \geq 0$ (the shock is weakly larger for higher-quality producers), then the new threshold satisfies $\tilde{q}^* < q^*$. The producers who*

*switch from permitting to blocking have quality in* $[\tilde{q}^*, q^*]$*—the top of the pre-shock available distribution—so their average quality exceeds the pre-shock commons mean.*

*Proof.* The new net cost is $\tilde{\pi}(q) = \pi(q) + \Delta(q)$. Since $\Delta(q) > 0$, $\tilde{\pi}(q) > \pi(q)$ for all $q$. At the old threshold, $\tilde{\pi}(q^*) = \pi(q^*) + \Delta(q^*) = \Delta(q^*) > 0$, so $q^*$ now blocks. Since $\tilde{\pi}$ is strictly increasing (as $\tilde{\pi}'(q) = \pi'(q) + \Delta'(q) > 0$) and $\tilde{\pi}(0) = \pi(0) + \Delta(0) < 0$ for sufficiently small $\Delta(0)$ (or by assumption 3 if $\Delta(0)$ is small relative to $|\pi(0)|$), there exists a unique $\tilde{q}^* < q^*$ where $\tilde{\pi}(\tilde{q}^*) = 0$. The switching set $[\tilde{q}^*, q^*]$ lies at the top of the pre-shock available distribution (since all previously available producers had $q < q^*$), so the average quality of switching producers exceeds $\mathbb{E}[q \mid q < q^*]$. □

This proposition formalizes the prediction that a salience shock causes the highest-quality producers among those still in the commons to exit first, amplifying the quality–blocking gradient at the moment the shock hits.

## A.5 Empirical Predictions

The model generates four testable predictions:

**P1 (Quality–Blocking Gradient):** Higher-quality content producers are more likely to block AI crawlers (Proposition 1).

**P2 (Training–Search Gap):** The training-blocking threshold is lower than the search-blocking threshold, and there exists a range of producers who selectively block training while permitting search, with the gap widening with quality (Proposition 2).

**P3 (Event-Driven Amplification):** A salience shock (GPTBot announcement) shifts $c(q)$ upward, amplifying the quality–blocking gradient, with the highest-quality producers among those still in the commons exiting first (Proposition 3).

**P4 (Corpus Degradation):** Quality-dependent blocking degrades both the count-weighted and volume-weighted average quality of the remaining commons below the population mean, with degradation increasing in the blocking rate (Proposition 1, Corollary 1).

# B Data Details

## B.1 Data Sources

We integrate six external data sources to construct a comprehensive picture of media site quality and AI blocking behavior.

Table 1: Data Sources

| **Source** | **Variables** | **Coverage** |
|---|---|---|
| HTTP Archive | `robots.txt` directives, monthly | 718K sites, 45 months |
| MBFC (Media Bias/Fact Check) | Factual, credibility, bias, flags | 9,611 media sites |
| Majestic Million | Trust Flow (link quality, 0–100) | Top 1M domains |
| Open PageRank | PageRank score (web centrality) | 7,197 media sites |
| Tranco | Popularity ranking (top 1M) | 5,046 media sites |
| Wikipedia API | Article existence (notability) | All 9,611 sites queried |

*HTTP Archive* conducts monthly crawls of millions of websites, recording `robots.txt` files. We parse these to identify which of 23 AI crawlers each site blocks, classifying crawlers as training-purpose (GPTBot, Google-Extended, ClaudeBot) or search-purpose (OAI-SearchBot, PerplexityBot).

*Media Bias/Fact Check (MBFC)* provides human-validated ratings: factual reporting (six levels), editorial credibility (three levels), political orientation (five categories), and questionable-source flags. MBFC ratings are determined pre-treatment and are not confounded by blocking behavior.

*Majestic Million* provides Trust Flow scores (0–100) based on backlink quality. *Open PageRank* provides web graph centrality scores. *Tranco* aggregates multiple popularity lists into a research-oriented ranking. *Wikipedia API* determines whether each site has a dedicated article.

All quality metrics are measured prior to or independently of AI blocking decisions.

## B.2 Per-Bot Blocking Rates

Figure 1 presents per-bot blocking rates across 22 AI crawlers (excluding CCBot). GPTBot is the most-blocked at 32.3%, followed by ClaudeBot (30.0%) and Google-Extended (28.1%). Training crawlers are systematically blocked at higher rates than search crawlers.

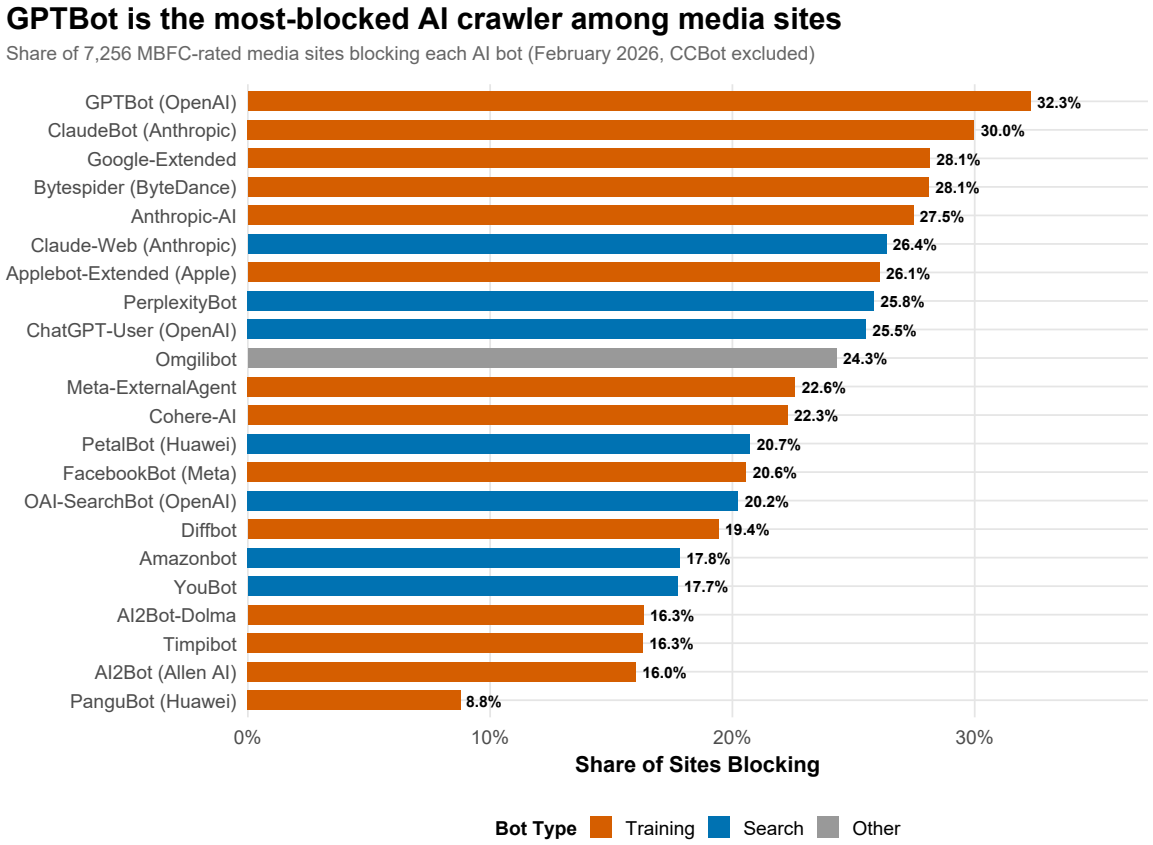



Figure 1: Per-Bot AI Crawler Blocking Rates Among MBFC-Rated Media Sites

## B.3 Quality Measure Construction

The *content quality index* is constructed via PCA of three standardized variables: MBFC factual reporting (ordinal, 1–6), MBFC credibility (ordinal, 1–3), and Wikipedia notability (binary). The first principal component loads positively on all three inputs.

The *domain quality index* is the first principal component of Trust Flow, Open PageRank, and inverse Tranco rank—objective web-graph and traffic metrics independent of editorial

content.

Both indices are standardized to mean zero and unit variance. Missing values are handled via median imputation when at least two of three inputs are non-missing; observations with fewer than two non-missing inputs receive a missing index value.

## B.4 Descriptive Statistics

Table 2 reports summary statistics for the full cross-section of 9,611 MBFC-rated media sites.

Table 2: Cross-Sectional Descriptive Statistics: MBFC-Rated Media Sites

| Variable | N | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|---|
| **Panel A: Full Sample** | | | | | | |
| Blocks any AI bot | 9611 | 0.289 | 0.453 | 0.000 | 0.000 | 1.000 |
| Blocks training bots | 7002 | 0.353 | 0.478 | 0.000 | 0.000 | 1.000 |
| Blocks search bots | 7002 | 0.267 | 0.442 | 0.000 | 0.000 | 1.000 |
| Domain quality index | 5782 | 0.000 | 1.000 | -2.115 | -0.170 | 4.598 |
| Content quality index | 9611 | 0.000 | 1.000 | -1.933 | 0.447 | 1.636 |
| MBFC factual (1–6) | 9611 | 3.940 | 1.247 | 1.000 | 4.000 | 6.000 |
| MBFC credibility (1–3) | 9602 | 2.261 | 0.893 | 1.000 | 3.000 | 3.000 |
| Trust Flow (0–100) | 5758 | 7.436 | 1.136 | 5.481 | 7.367 | 12.383 |
| Open PageRank | 7197 | 4.181 | 0.826 | 2.280 | 4.340 | 7.620 |
| Tranco rank | 5046 | 195588.443 | 227572.316 | 11.000 | 103080.500 | 996730.000 |
| Wikipedia notable | 9611 | 0.004 | 0.064 | 0.000 | 0.000 | 1.000 |

Table 3 disaggregates key variables by domain quality tercile. The table is displayed in landscape orientation due to the number of variables.

Table 3: Descriptive Statistics by Quality Tercile

| Quality Tercile | N | Blocks any AI bot | Blocks training bots | Blocks search bots | MBFC factual (1–6) | MBFC credibility (1–3) | Content quality index | Domain quality index | Trust Flow (0–100) | Open PageRank | Wikipedia notable |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Low | 1928 | 0.313 | 0.297 | 0.236 | 3.831 | 2.258 | -0.048 | -0.974 | 6.264 | 3.877 | 0.000 |
| Medium | 1927 | 0.447 | 0.423 | 0.334 | 4.361 | 2.607 | 0.370 | -0.156 | 7.330 | 4.416 | 0.000 |
| High | 1927 | 0.474 | 0.470 | 0.337 | 4.497 | 2.721 | 0.499 | 1.131 | 8.701 | 4.987 | 0.021 |

*Note:*
Cross-section of 9,611 MBFC-rated media sites (February 16, 2026). Quality terciles based on domain quality index.

# C Additional Results

## C.1 Cross-Sectional Regressions

Table 4 reports cross-sectional quality–blocking regressions with category fixed effects. All four quality measures yield strongly significant positive coefficients: MBFC factual reporting (0.134, $p < 0.01$), MBFC credibility (0.197, $p < 0.01$), content quality index (0.176, $p < 0.01$), and domain quality index (0.078, $p < 0.01$).

## C.2 Panel Regressions

Table 5 reports panel regressions exploiting the full 272,458 site-month panel. The quality–blocking gradient is robust to category and month fixed effects, site fixed effects with quality $\times$ post-GPTBot interactions, and post-2023 subsample restrictions. The site FE interaction specification confirms that the gradient widened *within* sites after GPTBot, not merely across sites.

## C.3 Training vs. Search Bot Regressions

Table 6 presents panel regressions with training-bot and search-bot blocking as separate dependent variables. The training coefficient consistently exceeds the search coefficient by approximately 38% across content-based quality measures, confirming that publishers strategically target training crawlers over search crawlers—especially high-quality publishers.

Table 4: Cross-Sectional Quality–Blocking Gradient: Media Sites

| | (1) Factual | (2) Credibility | (3) Content Q | (4) Domain Q |
|---|---|---|---|---|
| MBFC factual (ordinal) | 0.134*** | | | |
| | (0.003) | | | |
| MBFC credibility (ordinal) | | 0.197*** | | |
| | | (0.005) | | |
| Content quality index | | | 0.176*** | |
| | | | (0.004) | |
| Domain quality index | | | | 0.078*** |
| | | | | (0.006) |
| Num.Obs. | 9609 | 9600 | 9609 | 5780 |
| R2 | 0.136 | 0.151 | 0.151 | 0.025 |
| R2 Adj. | 0.135 | 0.150 | 0.151 | 0.024 |
| FE: category | X | X | X | X |

* p <0.1, ** p <0.05, *** p <0.01
Cross-section (N = 9,611).

Table 5: Panel Models: Quality and AI Crawler Blocking (Media Sites)

| | Factual | Credibility | Content Q | Domain Q | Site FE | Post-2023 |
|---|---|---|---|---|---|---|
| MBFC factual (ordinal) | 0.094*** | | | | | |
| | (0.003) | | | | | |
| MBFC credibility (ordinal) | | 0.131*** | | | | |
| | | (0.004) | | | | |
| Content quality index | | | 0.108*** | | | |
| | | | (0.003) | | | |
| Domain quality index | | | | 0.046*** | | 0.061*** |
| | | | | (0.004) | | (0.005) |
| Quality x Post-GPTBot | | | | | 0.049*** | |
| | | | | | (0.005) | |
| Post-GPTBot | | | | | 0.299*** | |
| | | | | | (0.005) | |
| Num.Obs. | 272458 | 272331 | 272458 | 239534 | 239515 | 163615 |
| R2 | 0.167 | 0.158 | 0.165 | 0.131 | 0.616 | 0.040 |
| R2 Adj. | 0.167 | 0.158 | 0.165 | 0.130 | 0.606 | 0.039 |
| FE: category | X | X | X | X | | X |
| FE: date | X | X | X | X | | X |
| FE: site_id | | | | | X | |

* p <0.1, ** p <0.05, *** p <0.01
Col 6: post-Aug 2023 subsample. SE clustered at site level.

Table 6: Training vs.\ Search Bot Blocking by Media Quality

| | Train: Quality | Search: Quality | Train: Factual | Search: Factual | Train: Cred. | Search: Cred. | Train: Content Q | Search: Content Q |
|---|---|---|---|---|---|---|---|---|
| Domain quality index | 0.037*** | 0.015*** | | | | | | |
| | (0.004) | (0.003) | | | | | | |
| MBFC factual (ordinal) | | | 0.095*** | 0.069*** | | | | |
| | | | (0.002) | (0.002) | | | | |
| MBFC credibility (ordinal) | | | | | 0.143*** | 0.104*** | | |
| | | | | | (0.003) | (0.002) | | |
| Content quality index | | | | | | | 0.113*** | 0.082*** |
| | | | | | | | (0.002) | (0.002) |
| Num.Obs. | 239534 | 239534 | 272458 | 272458 | 272331 | 272331 | 272458 | 272458 |
| R2 | 0.162 | 0.130 | 0.211 | 0.170 | 0.211 | 0.170 | 0.214 | 0.172 |
| FE: category | X | X | X | X | X | X | X | X |
| FE: date | X | X | X | X | X | X | X | X |

* p <0.1, ** p <0.05, *** p <0.01
MBFC-rated media sites only.

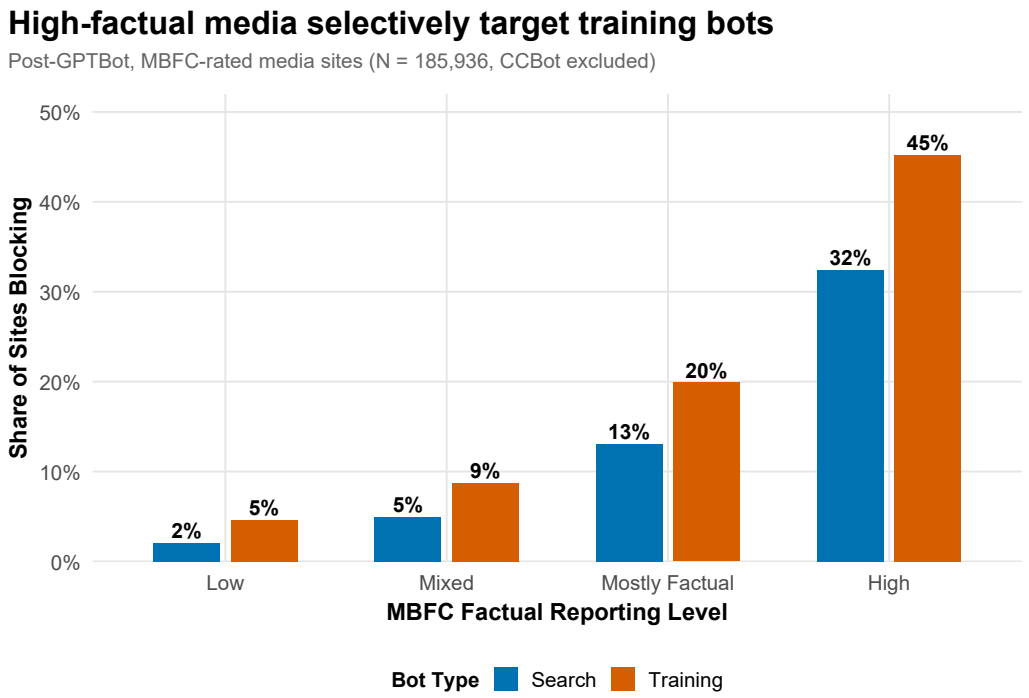



Figure 2: Training vs. Search Bot Blocking by Factual Level

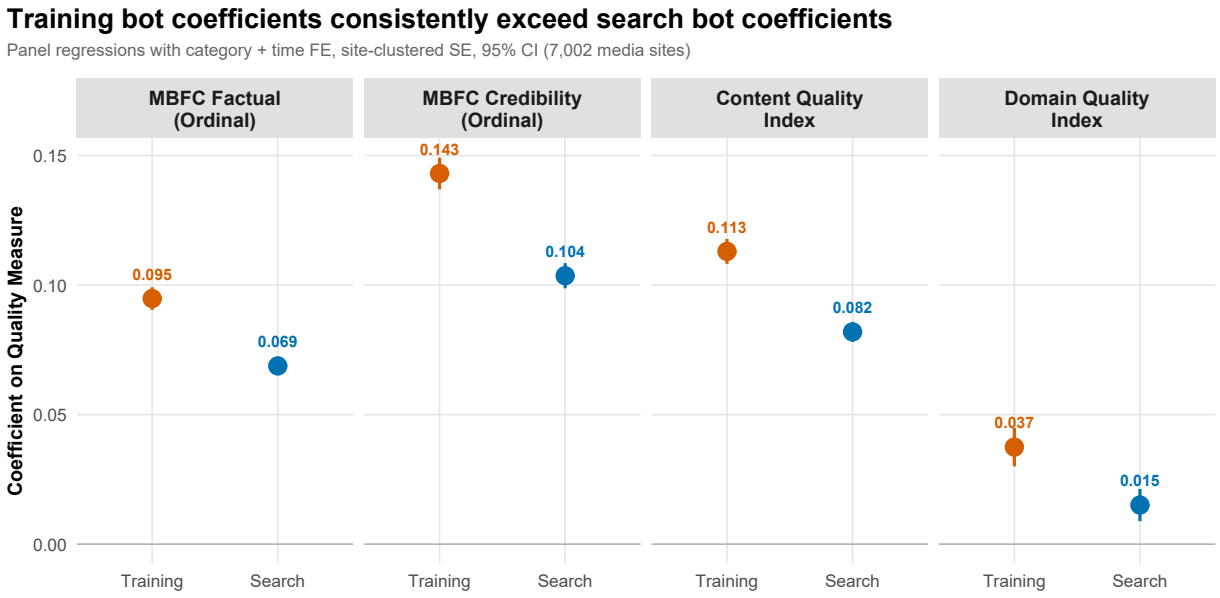



Figure 3: Training vs. Search Bot Coefficients Across Quality Measures

Table 7: Misinformation Asymmetry: Questionable vs.\ Credible Sources

| | (1) Raw gap | (2) +Domain Q | (3) +Content Q | (4) +Both +Cat FE |
|---|---|---|---|---|
| (Intercept) | 0.348*** | 0.438*** | 0.292*** | |
| | (0.005) | (0.007) | (0.005) | |
| Questionable source | -0.287*** | -0.298*** | -0.013 | 0.064** |
| | (0.011) | (0.022) | (0.014) | (0.026) |
| Domain quality index | | 0.064*** | | 0.032*** |
| | | (0.006) | | (0.006) |
| Content quality index | | | 0.173*** | 0.209*** |
| | | | (0.006) | (0.008) |
| Num.Obs. | 9611 | 5782 | 9611 | 5780 |
| R2 | 0.066 | 0.053 | 0.151 | 0.145 |
| FE: category | | | | X |

* p <0.1, ** p <0.05, *** p <0.01
Content quality: PCA(MBFC factual, credibility, Wikipedia).

## C.4 Misinformation Regression

The following table shows that conditional on the content quality index, questionable-source status has no independent effect on blocking probability. In column (1), the raw gap is $-0.287$ ($p < 0.01$); adding the content quality index in column (3) reduces this to $-0.013$ (insignificant). With both quality indices and category FE in column (4), the coefficient flips to $+0.064$ ($p < 0.05$). The misinformation asymmetry is thus a downstream consequence of adverse selection in the quality dimension rather than an independent phenomenon.

## C.5 Political Orientation Regressions

Table 8 reports regressions examining political orientation and blocking across four specifications that progressively add quality controls. All political orientation categories block significantly less than the omitted center category ($p < 0.01$). The right–center gap attenu-

Table 8: Political Bias and AI Crawler Blocking

| | (1) Bias only | (2) +Domain Q | (3) +Content Q | (4) +Both +Cat FE |
|---|---|---|---|---|
| (Intercept) | 0.626*** | 0.623*** | 0.440*** | |
| | (0.012) | (0.012) | (0.018) | |
| Left | -0.354*** | -0.352*** | -0.234*** | -0.230*** |
| | (0.024) | (0.024) | (0.026) | (0.026) |
| Left-Center | -0.173*** | -0.177*** | -0.159*** | -0.163*** |
| | (0.017) | (0.017) | (0.017) | (0.017) |
| Right-Center | -0.101*** | -0.102*** | -0.045** | -0.045** |
| | (0.022) | (0.022) | (0.022) | (0.022) |
| Right | -0.483*** | -0.467*** | -0.206*** | -0.186*** |
| | (0.028) | (0.028) | (0.035) | (0.035) |
| Domain quality index | | 0.044*** | | 0.048*** |
| | | (0.007) | | (0.007) |
| Content quality index | | | 0.234*** | 0.236*** |
| | | | (0.018) | (0.018) |
| Num.Obs. | 4650 | 4650 | 4650 | 4650 |
| R2 | 0.085 | 0.091 | 0.117 | 0.125 |
| FE: category | | | | X |

* p <0.1, ** p <0.05, *** p <0.01
Content quality: PCA(MBFC factual, credibility, Wikipedia).

ates with quality controls (from −0.101 to −0.045), suggesting quality explains part of the ideological gradient, but left (−0.230) and right (−0.186) remain strongly negative even after controlling for both quality indices and category FE.

## C.6 Alternative Explanations

Several alternative mechanisms could generate a quality–blocking gradient:

**Resources vs. quality.** Large, well-resourced sites have IT teams to implement blocking, and size correlates with quality. We test this by estimating quality–blocking coefficients

within site-size terciles (Appendix F.3); the gradient persists in all terciles, ruling out a pure resources explanation.

**Revenue model.** Subscription-based sites (typically high-quality) may have more to lose from AI summarization. Revenue model data is unavailable, but revenue model is a *mechanism* through which quality affects blocking—consistent with the adverse selection interpretation.

**Legal strategy.** Publishers involved in AI-related lawsuits may block preemptively. The gradient holds across thousands of sites, most not involved in litigation.

# D Event Study Details

## D.1 Specification

Our event study estimates:

$$\text{Block}_{it} = \sum_{\tau \neq \tau_0} \beta_\tau \, \text{Quality}_i \times \mathbf{1}[t = \tau] + \gamma_c + \delta_t + \varepsilon_{it}$$

where $\text{Quality}_i$ is the quality measure, $\tau$ indexes months relative to the GPTBot announcement (August 2023), $\tau_0$ is the reference period (earliest month), $\gamma_c$ are category fixed effects, and $\delta_t$ are month fixed effects. Standard errors are clustered at the site level.

## D.2 Event Study Combined Figure

Figure 4 presents event studies for all four quality measures. All four panels show the same pattern: near-zero pre-GPTBot coefficients followed by a sharp, persistent increase after August 2023. The consistency across content-based (factual, credibility, content quality index) and web-authority (domain quality index) measures rules out measurement artifacts.

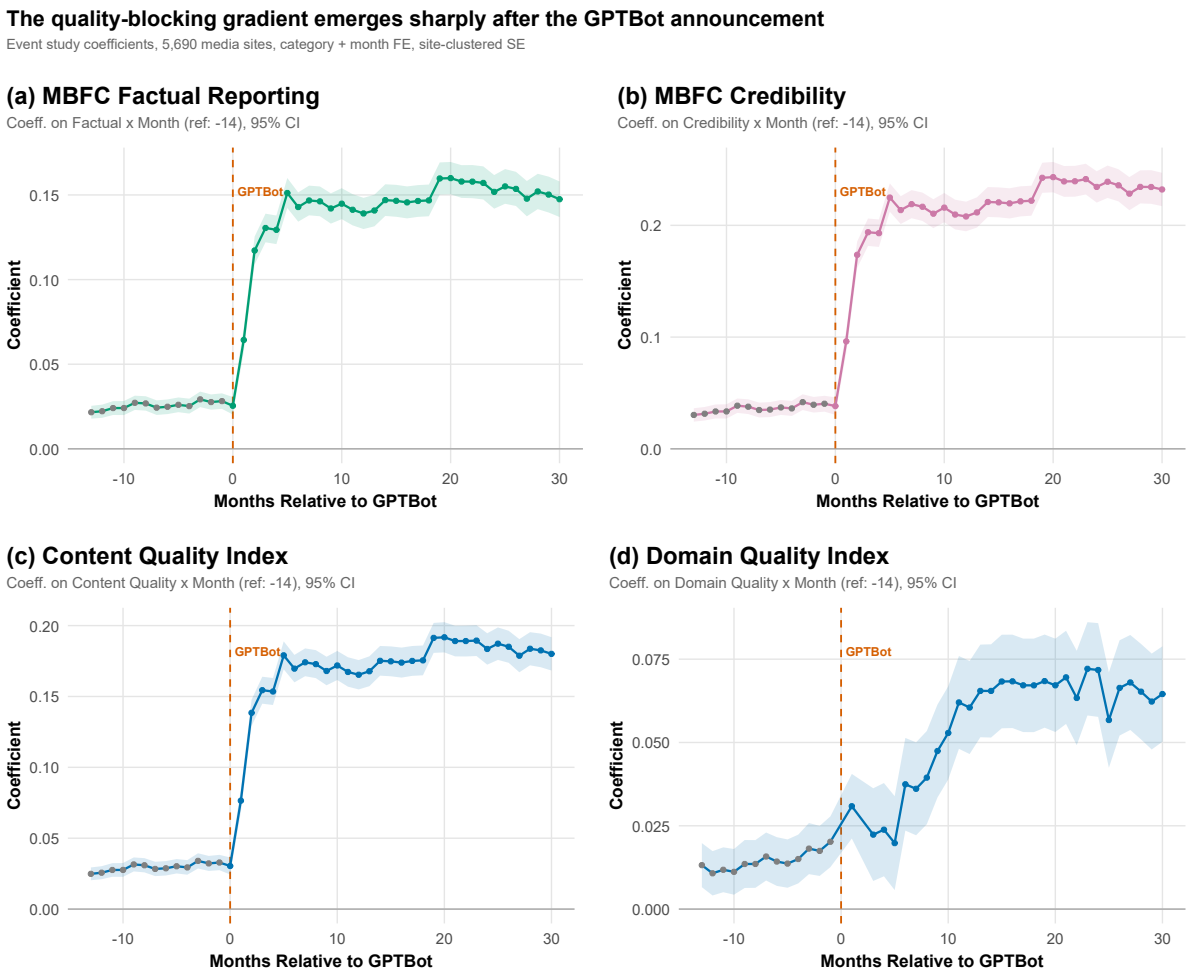



Figure 4: Event Study: Quality–Blocking Gradient Before and After GPTBot Announcement

## D.3 Pre-Trend Tests

Table 9 reports joint $F$-tests of all pre-GPTBot event-study coefficients. All four $F$-tests reject the null at conventional significance levels. However, the pre-period coefficients are economically small relative to post-announcement coefficients—approximately five to six times smaller in magnitude.

Table 9: Joint F-Tests of Pre-Trend Coefficients

| Quality Measure | F-statistic | p-value | Pre-period months |
|---|---|---|---|
| MBFC Factual | 15.38 | 0e+00 | 13 |
| MBFC Credibility | 13.17 | 0e+00 | 13 |
| Content Quality Index | 14.78 | 0e+00 | 13 |
| Domain Quality Index | 3.10 | 1e-04 | 13 |

*Note:*
Wald test of joint significance of all pre-GPTBot event-study coefficients. Category and month fixed effects included. Standard errors clustered at site level.

## D.4 Pre/Post Magnitude Comparison

Table 10 compares mean absolute values of event-study coefficients before and after GPTBot. Post-period coefficients are five to six times larger, with 95% bootstrap confidence intervals excluding unity. This magnitude difference is the primary basis for our temporal analysis: while a small pre-trend exists, the GPTBot announcement produced a discrete, large amplification of the quality–blocking gradient.

Table 10: Pre- vs. Post-Period Coefficient Magnitudes

| Quality Measure | Mean \|Pre\| | Mean \|Post\| | Ratio | 95% CI |
|---|---|---|---|---|
| MBFC Credibility | 0.0361 | 0.2120 | 5.9 | [5.3, 6.4] |
| Content Quality Index | 0.0295 | 0.1676 | 5.7 | [5.1, 6.2] |
| MBFC Factual | 0.0255 | 0.1401 | 5.5 | [5.0, 5.9] |
| Domain Quality Index | 0.0145 | 0.0562 | 3.9 | [3.3, 4.5] |

Mean absolute value of event-study coefficients before and after GPTBot announcement. Ratio: post/pre magnitude. 95% CI from 1,000 bootstrap draws.

## D.5 Rambachan and Roth Sensitivity Analysis

We apply the Rambachan and Roth (2023) sensitivity analysis framework, which constructs confidence sets under assumptions about the maximum deviation of post-treatment trends from pre-treatment trends. Figure 5 shows that the confidence interval for the first post-GPTBot effect remains bounded away from zero for $\bar{M}$ values up to at least 1, indicating that results are robust even if post-treatment trends deviate substantially from pre-treatment trends.

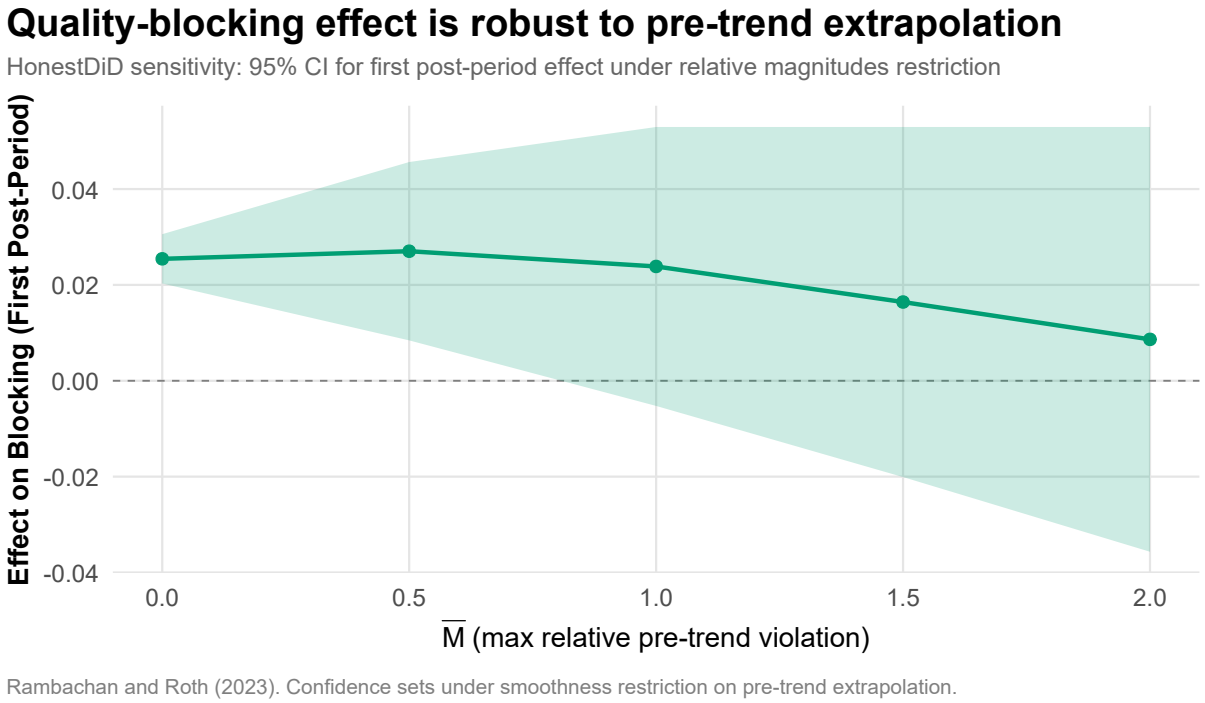



Figure 5: HonestDiD Sensitivity Analysis: Robustness to Pre-Trend Extrapolation

## D.6 Detrended Event Study

Figure 6 presents detrended event studies that subtract the pre-period linear trend from all event-study coefficients. If the pre-trend were driving the post-period results, detrending would eliminate the post-GPTBot increase. Instead, the structural break at GPTBot persists clearly after detrending, confirming that the post-period amplification is not an extrapolation

of pre-existing trends.

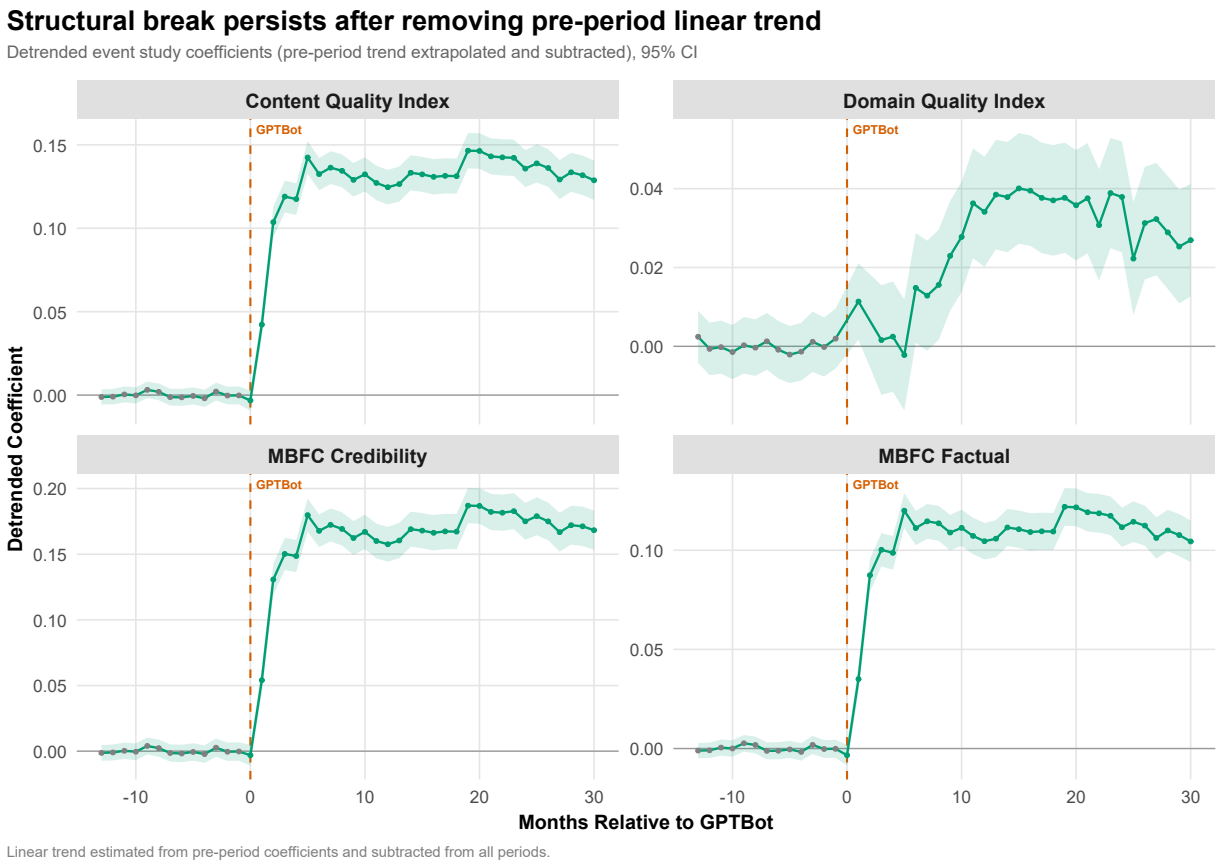



Figure 6: Detrended Event Study: Structural Break Persists After Removing Pre-Period Linear Trend

## D.7 Site Fixed Effects Event Study

Figure 7 overlays site FE coefficients on category FE coefficients. The site FE specification absorbs all time-invariant site characteristics, providing a more demanding identification. The patterns are strikingly similar across both specifications, confirming that results are not driven by compositional differences between sites in different categories or by time-invariant confounders.

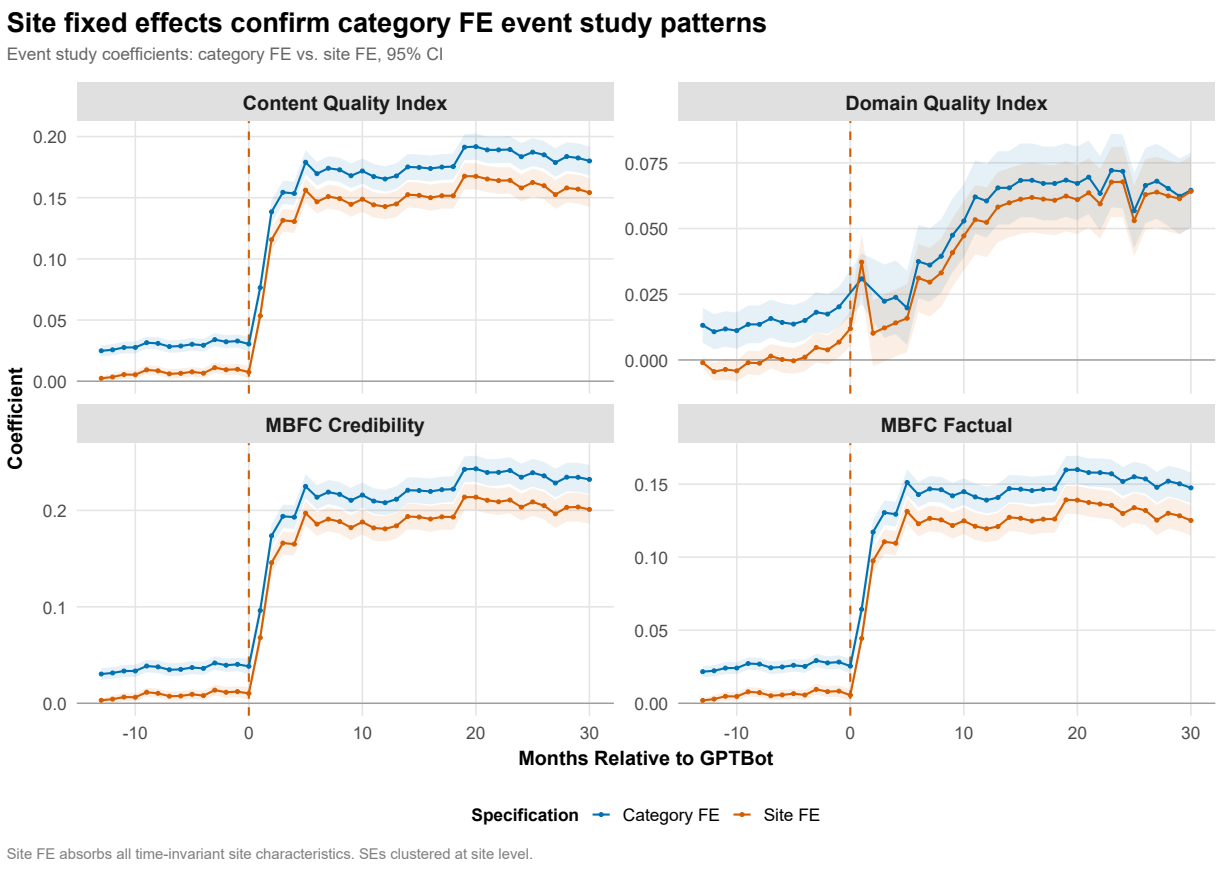



Figure 7: Event Study: Category FE vs. Site FE Specifications

# E Simulation Details

## E.1 Tier Calibration

Table 11 reports the empirical parameters used to calibrate the four-tier simulation model. Each tier corresponds to an MBFC factual reporting level, with blocking rates, quality scores, and content volumes drawn directly from the cross-sectional data. High-factual sites block at 55.6% and produce approximately 11× the token volume of low-factual sites (proxied by Trust Flow), creating the "double penalty" described in the main text.

Table 11: Simulation Tier Calibration Parameters

| Factual Tier | Sites | Share (%) | Blocking Rate (%) | Quality Score | Mean Trust Flow |
|---|---|---|---|---|---|
| Low | 600 | 10.4 | 9.0 | 0.25 | 6.7 |
| Mixed | 969 | 16.8 | 15.6 | 0.50 | 7.4 |
| Mostly Factual | 642 | 11.1 | 29.3 | 0.67 | 7.8 |
| High | 3,571 | 61.8 | 55.6 | 0.92 | 7.5 |

*Note:*
Calibrated from 5,782 MBFC-rated media sites. Blocking rates are cross-sectional means (February 2026). Quality scores are normalized MBFC factual ordinal values. Volume proxied by Trust Flow.

## E.2 Quality Trajectory

Figure 8 traces the projected quality and composition trajectory of available media training data through 2028 under three scenarios: deceleration (blocking growth slows), baseline (linear extrapolation of current trends), and acceleration (blocking growth continues at 1.5× current rate). Under the baseline scenario, token-weighted average quality declines by 24.5% from June 2022 levels, and high-factual content share falls from approximately 50% to approximately one-third of available tokens.

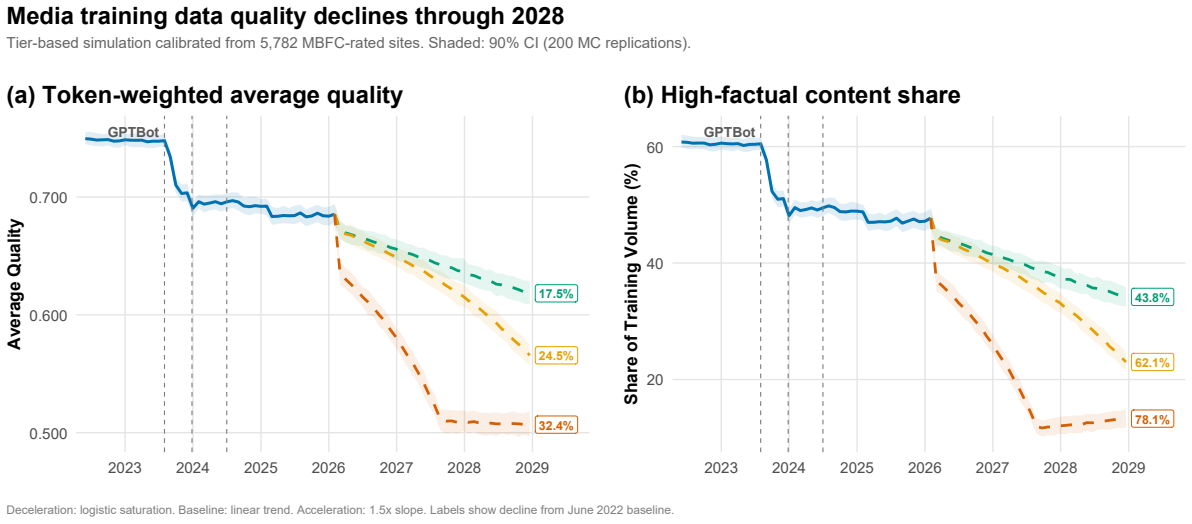



Figure 8: Media Training Data Quality and Composition Trajectory Through 2028

## E.3 Composition Trajectory

Figure 9 shows the projected share of high-quality versus low-quality media content in the available training corpus. High-quality content's share of available training volume declines steadily as high-factual outlets block at accelerating rates, while low-quality content's share increases proportionally—a direct consequence of the adverse selection dynamic.

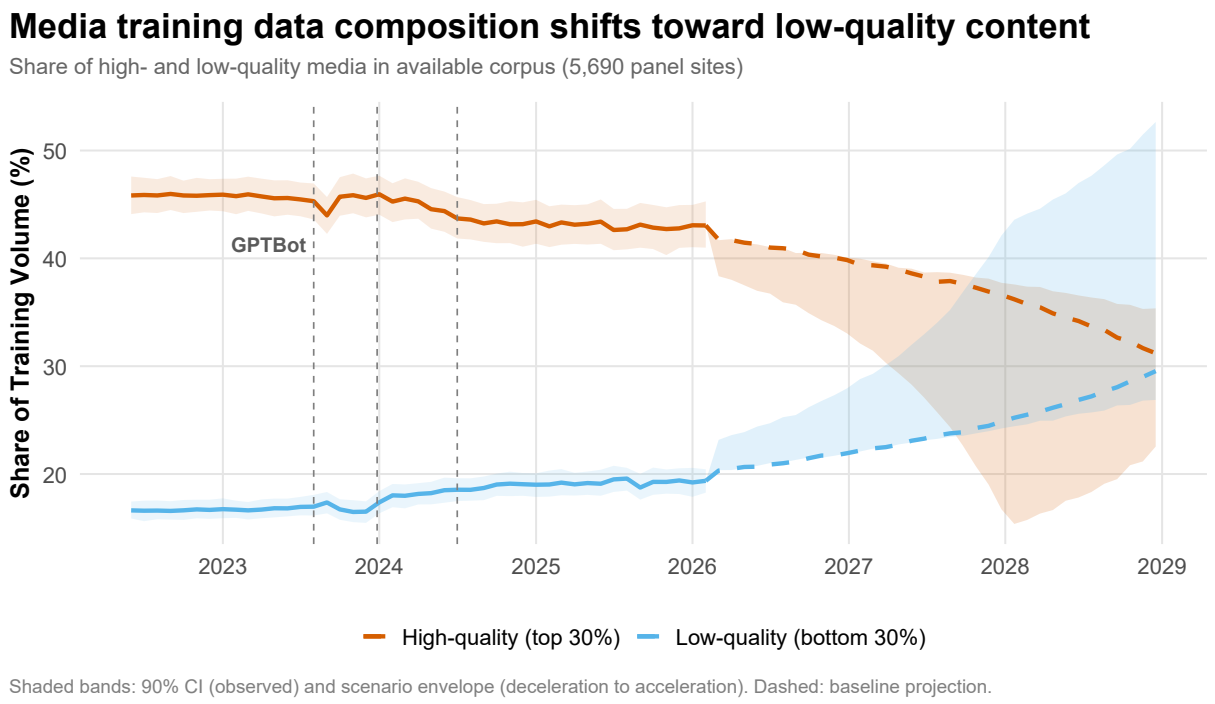



Figure 9: Projected Composition of Available Media Training Data by Factual Tier

## E.4 Factual Trajectory to 2028

Figure 10 presents the projected blocking trajectory by MBFC factual level, showing the divergence between high-factual and low-factual blocking rates through 2028. The gap widens across all three scenarios, with high-factual sites projected to reach 60–80% blocking by late 2028 while low-factual sites remain below 25%.

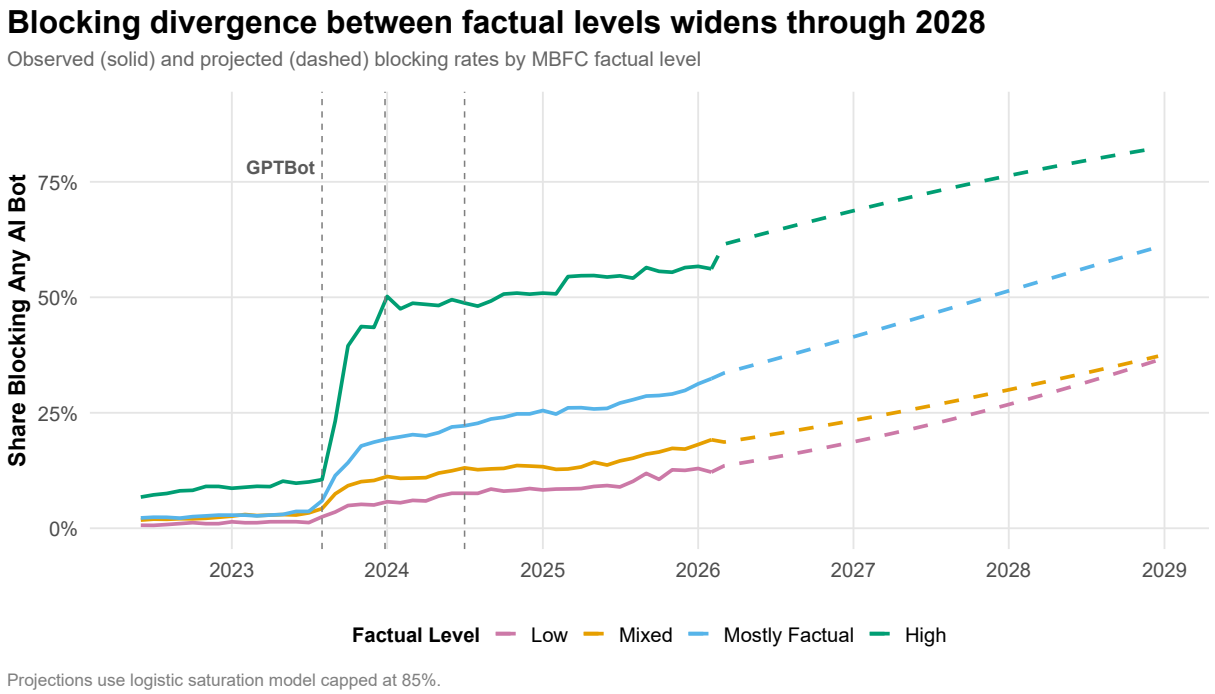



Figure 10: Projected High-Factual Content Share Through 2028

## E.5 Sensitivity Analysis

Figure 11 examines robustness of the simulation results. Panel (a) varies the blocking intensity multiplier from 0 (no blocking) to 2× (double current rates), showing that high-factual share decline is approximately linear for moderate intensities but exhibits diminishing returns at high intensities as the pool of high-factual sites that can still block is depleted. Panel (b) compares random versus selective blocking at each overall rate—the vertical gap between the two curves at any blocking rate is the pure adverse selection effect, isolating the quality degradation attributable to the compositional pattern of blocking rather than blocking *per se*.

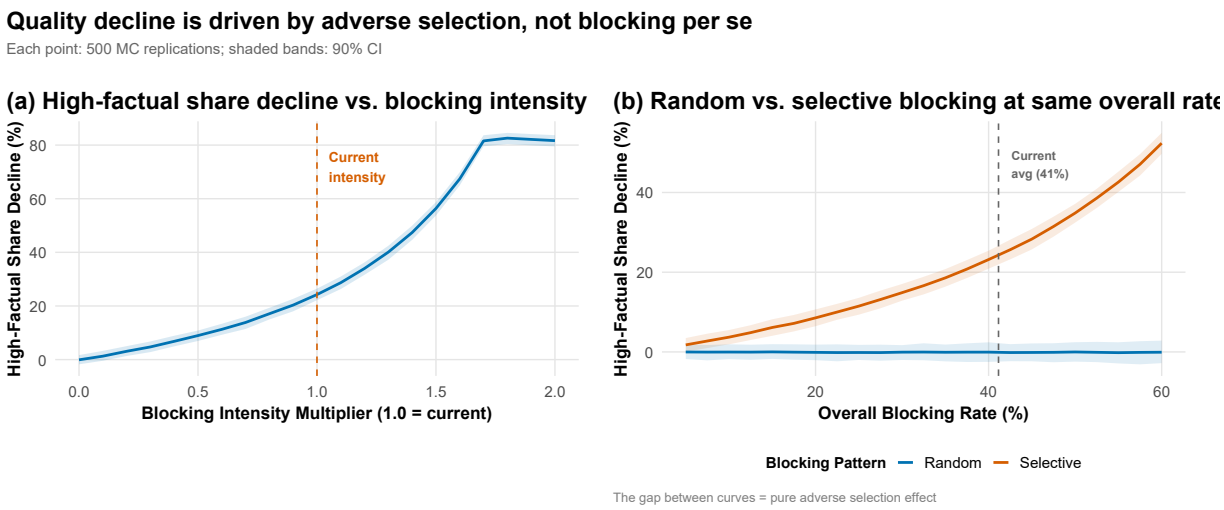



Figure 11: Sensitivity Analysis: Quality Decline and Adverse Selection Decomposition

# F Robustness

## F.1 Logit Models

Table 12 presents logit estimates (columns 5–8) alongside the LPM (columns 1–4). Sign, significance, and relative ordering are identical across all four quality measures, confirming that the linear probability model provides an adequate approximation.

Table 12: Logit vs.\ LPM: Cross-Sectional Quality–Blocking Gradient

| | (1) LPM | (2) LPM | (3) LPM | (4) LPM | (5) Logit | (6) Logit | (7) Logit | (8) Logit |
|---|---|---|---|---|---|---|---|---|
| MBFC factual (ordinal) | 0.134*** | | | | 0.888*** | | | |
| | (0.003) | | | | (0.027) | | | |
| MBFC credibility (ordinal) | | 0.197*** | | | | 1.311*** | | |
| | | (0.005) | | | | (0.039) | | |
| Content quality index | | | 0.176*** | | | | 1.162*** | |
| | | | (0.004) | | | | (0.034) | |
| Domain quality index | | | | 0.078*** | | | | 0.325*** |
| | | | | (0.006) | | | | (0.028) |
| Num.Obs. | 9609 | 9600 | 9609 | 5780 | 9606 | 9597 | 9606 | 5777 |
| R2 | 0.136 | 0.151 | 0.151 | 0.025 | 0.133 | 0.147 | 0.147 | 0.018 |
| FE: category | X | X | X | X | X | X | X | X |

* p <0.1, ** p <0.05, *** p <0.01
MBFC-rated media sites, cross-section.

## F.2 Balanced Panel

Table 13 restricts the sample to sites observed in all 45 monthly crawls, eliminating concerns about compositional changes from entry and exit. Coefficients are stable relative to the full panel, confirming that attrition does not drive the quality–blocking gradient.

## F.3 Size Heterogeneity

Figure 12 shows that the quality coefficient is positive and significant within all three site-size terciles (small, medium, and large), ruling out a pure "resources" explanation where only large sites with IT teams implement blocking. Table 14 presents the interaction model, which shows a negative factual × large interaction ($-0.042$, $p < 0.01$), indicating that the gradient is actually *stronger* among smaller sites.

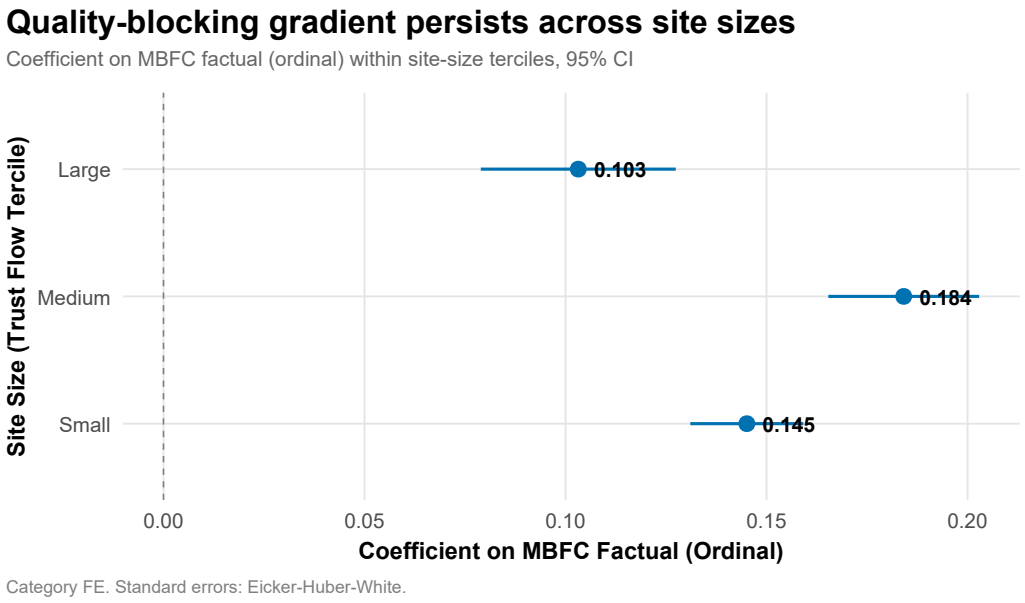



Figure 12: Quality–Blocking Coefficient by Site Size Tercile

Table 13: Balanced Panel Robustness: Quality–Blocking Gradient

| | (1) Factual | (2) Credibility | (3) Content Q | (4) Domain Q |
|---|---|---|---|---|
| MBFC factual (ordinal) | 0.110*** | | | |
| | (0.004) | | | |
| MBFC credibility (ordinal) | | 0.164*** | | |
| | | (0.005) | | |
| Content quality index | | | 0.131*** | |
| | | | (0.004) | |
| Domain quality index | | | | 0.041*** |
| | | | | (0.005) |
| Num.Obs. | 202680 | 202590 | 202680 | 202680 |
| R2 | 0.190 | 0.187 | 0.192 | 0.131 |
| FE: category | X | X | X | X |
| FE: date | X | X | X | X |

* p <0.1, ** p <0.05, *** p <0.01
DV: blocks any AI bot (excl. CCBot). Category + month FE. Site-clustered SEs.

Table 14: Size Heterogeneity: Quality $\times$ Site Size Interaction

| | (1) Main Effect | (2) With Interaction |
|---|---|---|
| MBFC factual (ordinal) | 0.154*** | 0.145*** |
| | (0.005) | (0.008) |
| Medium sites | | -0.098* |
| | | (0.053) |
| Large sites | | 0.258*** |
| | | (0.062) |
| Factual $\times$ Medium | | 0.039*** |
| | | (0.012) |
| Factual $\times$ Large | | -0.042*** |
| | | (0.014) |
| Num.Obs. | 5756 | 5756 |
| R2 | 0.130 | 0.139 |
| FE: category | X | X |

* p <0.1, ** p <0.05, *** p <0.01
Size terciles based on Trust Flow. MBFC-rated media sites.

## F.4 Political Sample Selection

Table 15 compares quality–blocking coefficients across the full sample, bias-rated subsample, and non-bias-rated subsample. The coefficient is positive and significant in all three samples (0.134, 0.168, and 0.049 respectively), confirming that the quality–blocking gradient is not an artifact of MBFC's bias rating coverage.

## F.5 Blocking Intensity

Table 16 confirms the quality gradient extends to the intensive margin: higher-quality sites block more bots, not just any bot. Columns 1–4 report Poisson regressions with the count of AI bots blocked as the dependent variable; columns 5–8 report OLS with blocking share. Figure 13 shows the mean number of bots blocked rising from 2.5 (mixed-factual) to 13.7 (high-factual).

Table 15: Political Sample Selection: Quality Gradient by Bias Rating Coverage

| | (1) All Sites | (2) Bias-Rated | (3) Non-Bias-Rated |
|---|---|---|---|
| MBFC factual (ordinal) | 0.134*** | 0.168*** | 0.049*** |
| | (0.003) | (0.008) | (0.004) |
| Num.Obs. | 9609 | 5995 | 3614 |
| R2 | 0.136 | 0.066 | 0.040 |
| FE: category | X | X | X |

* p <0.1, ** p <0.05, *** p <0.01
Col 1: all MBFC sites. Col 2: sites with bias ratings. Col 3: sites without.

Table 16: Blocking Intensity: Count (Poisson) and Share (OLS)

| | (1) Factual | (2) Credibility | (3) Content Q | (4) Domain Q | (5) Factual | (6) Credibility | (7) Content Q | (8) Domain Q |
|---|---|---|---|---|---|---|---|---|
| MBFC factual (ordinal) | 0.603*** | | | | 0.028*** | | | |
| | (0.004) | | | | (0.001) | | | |
| MBFC credibility (ordinal) | | 0.905*** | | | | 0.041*** | | |
| | | (0.006) | | | | (0.002) | | |
| Content quality index | | | 0.802*** | | | | 0.037*** | |
| | | | (0.005) | | | | (0.002) | |
| Domain quality index | | | | 0.119*** | | | | 0.010*** |
| | | | | (0.004) | | | | (0.002) |
| Num.Obs. | 9606 | 9597 | 9606 | 5777 | 9609 | 9600 | 9609 | 5780 |
| R2 | 0.099 | 0.107 | 0.108 | 0.005 | 0.043 | 0.046 | 0.047 | 0.004 |
| FE: category | X | X | X | X | X | X | X | X |

* p <0.1, ** p <0.05, *** p <0.01
MBFC-rated media sites, cross-section.

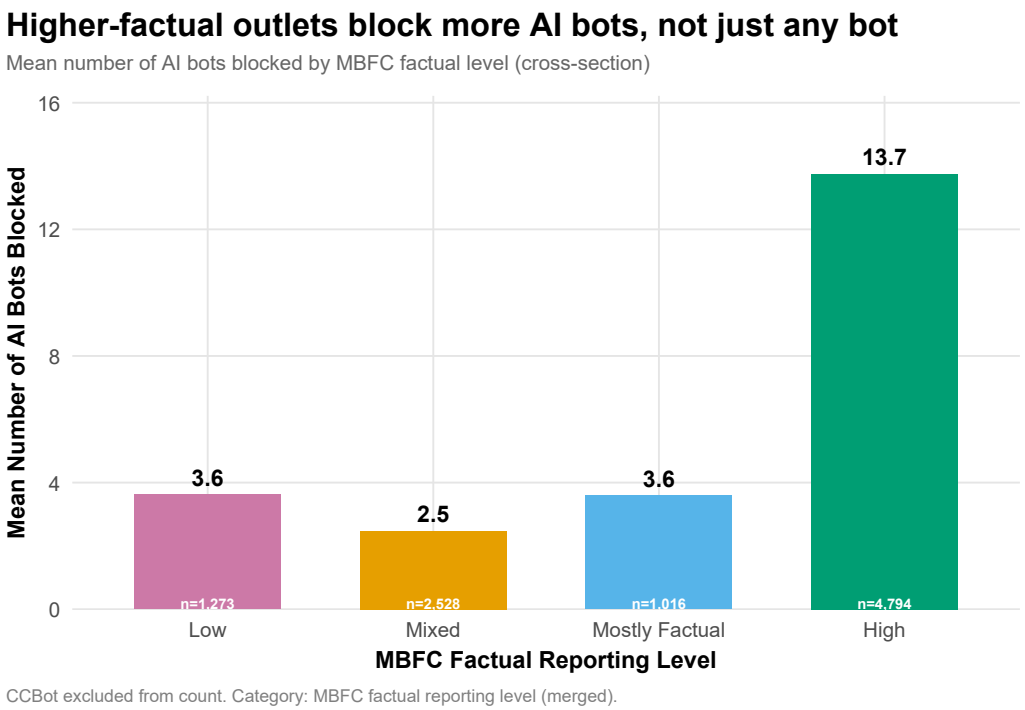



Figure 13: Mean Number of AI Bots Blocked by MBFC Factual Level

## F.6 Placebo Test

Figure 14 shows the distribution of placebo coefficients from 1,000 random shuffles of MBFC factual ratings. The observed coefficient falls far in the right tail of the permutation distribution (permutation $p < 0.001$), confirming that the quality–blocking gradient reflects a genuine association rather than a statistical artifact.

Table 17: August 2024 Data Repair Sensitivity

| | (1) Full Panel | (2) Excl. Aug 2024 |
|---|---|---|
| MBFC factual (ordinal) | 0.102*** | 0.102*** |
| | (0.003) | (0.003) |
| Num.Obs. | 239534 | 234274 |
| R2 | 0.185 | 0.187 |
| FE: category | X | X |
| FE: date | X | X |

* p <0.1, ** p <0.05, *** p <0.01
Col 1: full panel with Aug 2024 repair. Col 2: Aug 2024 excluded entirely.

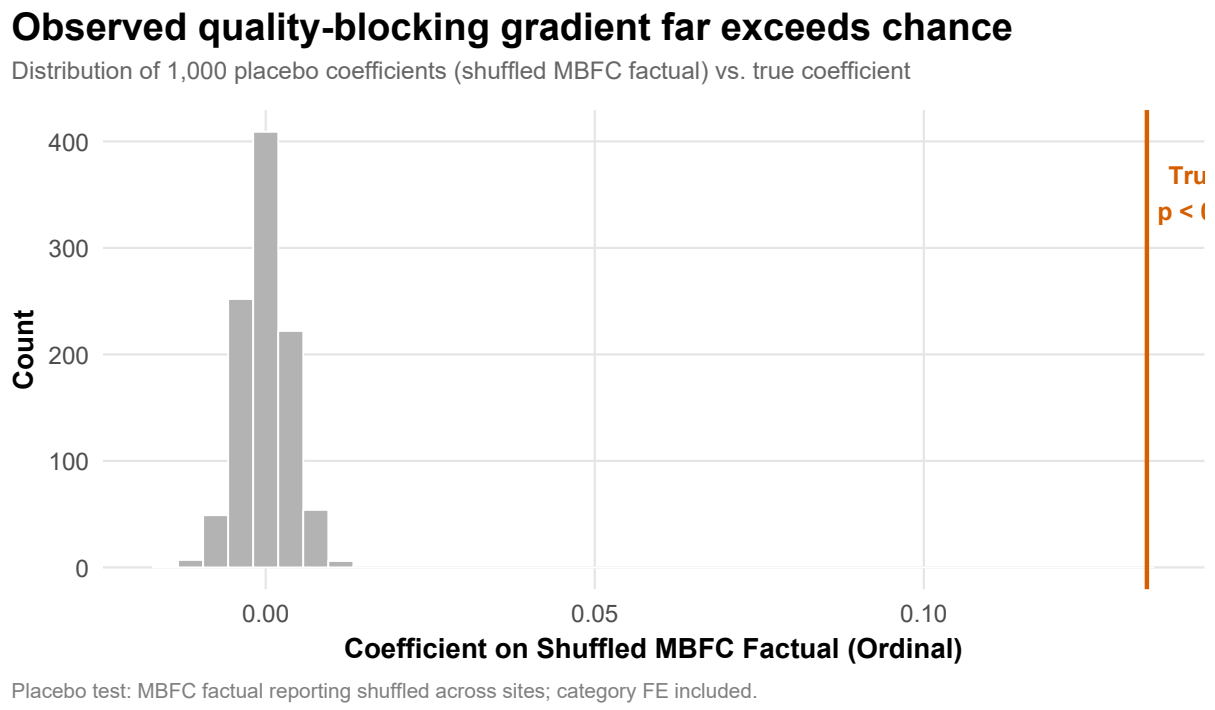



Figure 14: Placebo Test: Shuffled Quality Assignments

## F.7 August 2024 Data Repair

The August 2024 HTTP Archive crawl experienced partial data collection issues. Table 17 shows that excluding August 2024 entirely produces virtually identical estimates to the full sample, confirming that the data repair does not affect results.

# References

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *The Review of Economic Studies 90*(5), 2555–2591.