

Selective Participation in the AI Data Commons

Which publishers restrict AI crawlers, and for which uses?

Kai Zhu
Bocconi University

NBER Summer Institute 2026
July 2026

Web access is an endogenous AI input

Value

AI relies on **high-quality web content**.

Gunasekar et al. 2023; Longpre et al. 2023

Web access is an endogenous AI input

Value

AI relies on **high-quality web content**.

Gunasekar et al. 2023; Longpre et al. 2023

Price

Most access remains **uncompensated**.

Examples: OpenAI–Reddit \$60M; OpenAI–Associated Press.

Web access is an endogenous AI input

Value

AI relies on **high-quality web content**.

Gunasekar et al. 2023; Longpre et al. 2023

Price

Most access remains **uncompensated**.

Examples: OpenAI–Reddit \$60M; OpenAI–Associated Press.

Choice

Publishers can restrict **crawler access**.

Network blocking exists; `robots.txt` is standardized and observable.

Web access is an endogenous AI input

Value

AI relies on **high-quality web content**.

Gunasekar et al. 2023; Longpre et al. 2023

Price

Most access remains **uncompensated**.

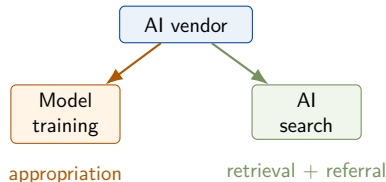
Examples: OpenAI–Reddit \$60M; OpenAI–Associated Press.

Choice

Publishers can restrict **crawler access**.

Network blocking exists; robots.txt is standardized and observable.

ONE VENDOR, DISTINCT USES



Web access is an endogenous AI input

Value

AI relies on **high-quality web content**.

Gunasekar et al. 2023; Longpre et al. 2023

Price

Most access remains **uncompensated**.

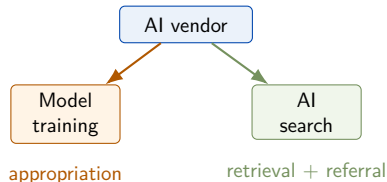
Examples: OpenAI–Reddit \$60M; OpenAI–Associated Press.

Choice

Publishers can restrict **crawler access**.

Network blocking exists; robots.txt is standardized and observable.

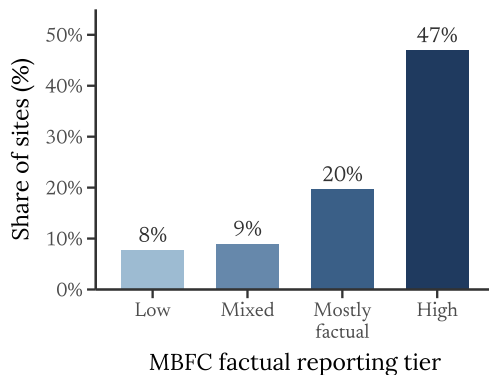
ONE VENDOR, DISTINCT USES



Research question: Who withdraws under uncompensated access—and from which AI uses?

Setting: 9,611 news and media publishers.

The accessible pool is selected on quality



47%

high-factual outlets restrict at least one AI crawler

8%

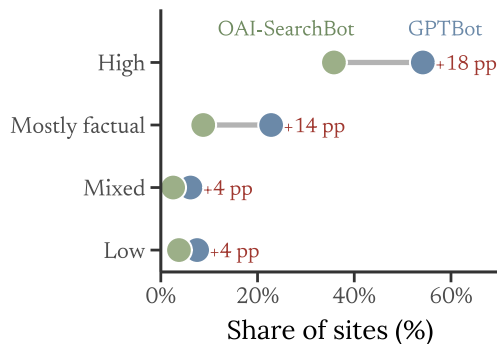
low-factual outlets do the same

Nearly a sixfold gap

Further composition evidence: Misinformation sources and ideological extremes are relatively more accessible.

Publishers distinguish training from search

Same publisher and vendor; only the intended AI use changes.



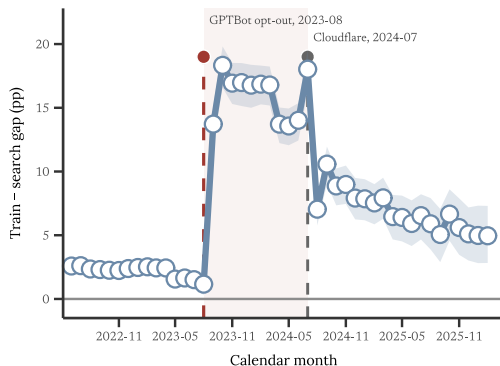
TRAINING MINUS SEARCH
+18 pp
at the high-factual tier

+4 pp
at the low-factual tier

**The use-specific wedge
grows with quality.**

Interpretation: Higher-quality publishers are more likely to preserve search-like access while restricting training, consistent with a distinction between appropriation and referral.

Implementation frictions reshape use-specific access



AUGUST 2023

GPTBot is named

Salience makes a training crawler easy to target.

Gap opens to about 18 pp

JULY 2024

Cloudflare one-click tool

Broad crawler-class blocking becomes cheap.

Search restrictions catch up

Salience opens the wedge; lower implementation costs compress it.

Implications for market design

A voluntary access regime reveals three distinct choices.

WHO?	Publisher selection	Higher-quality outlets withdraw at sharply higher rates.
WHAT?	Use selection	The same publisher treats training and search differently.
HOW?	Implementation friction	Salience and one-click tools change which choices are expressed.

Participation: A market can return high-quality content to the AI data commons.

Contracting: Unbundle search from training so prices and terms can vary by use.