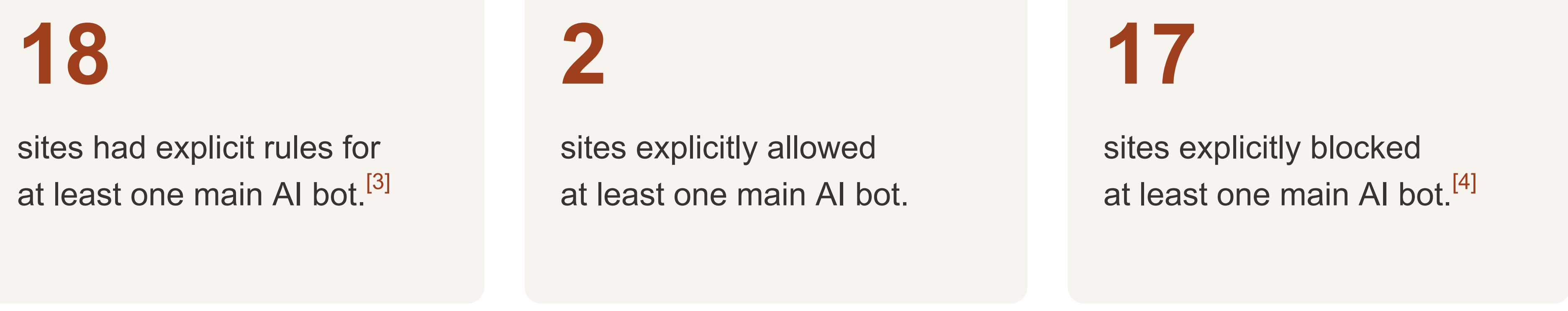


AI crawler rules in academic publishing

Inspired by FT Strategies' 2026 analysis of AI crawler rules across major news publishers,^[1] this snapshot applies a similar approach to academic publishing, and finds that the 32 selected sites vary widely in their robots.txt rules for AI bots from providers including OpenAI, Anthropic, Google, Perplexity and Meta.^[2]

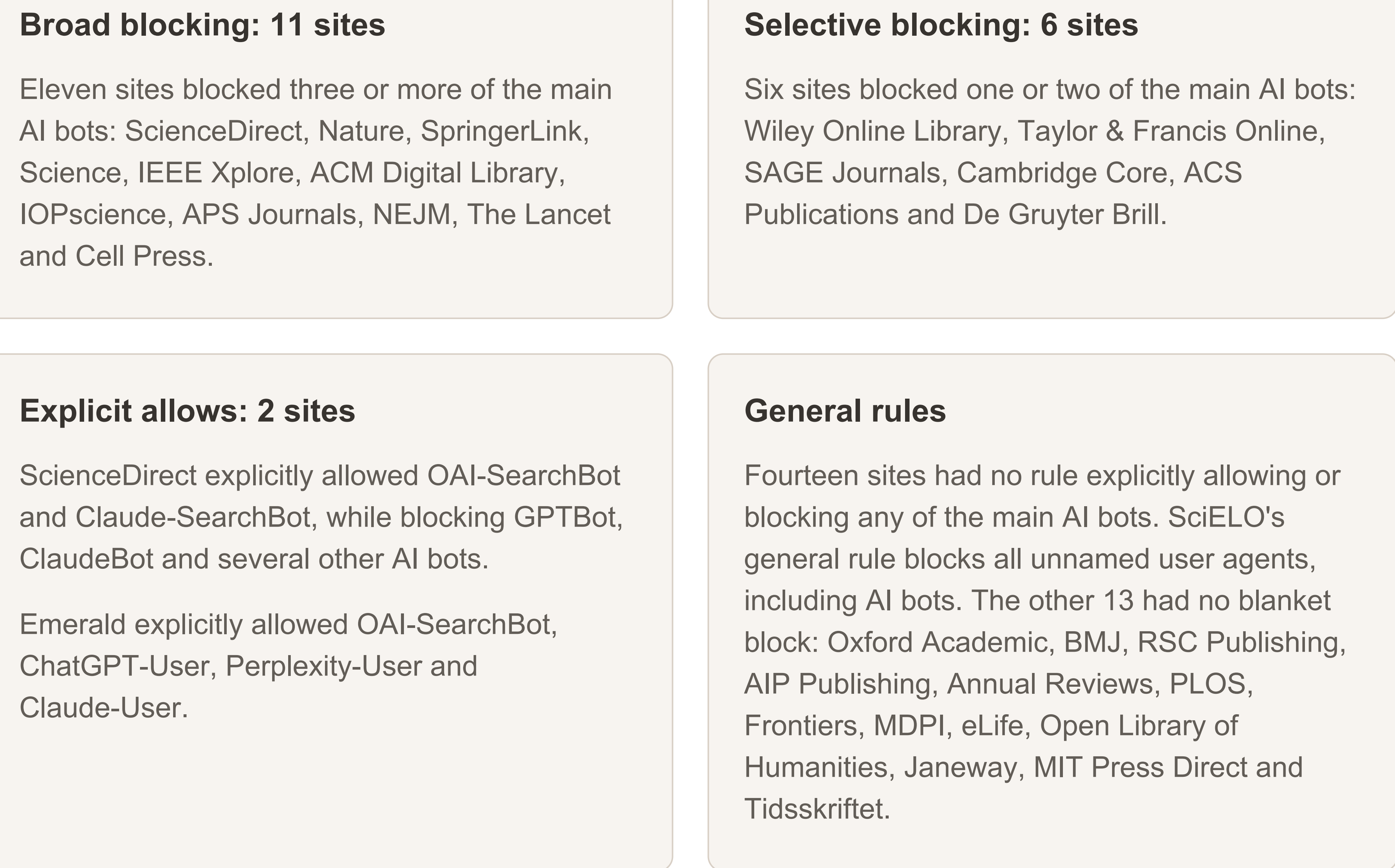
Headline findings

Across 32 academic publishing sites: explicit rules allowing or blocking AI bots.



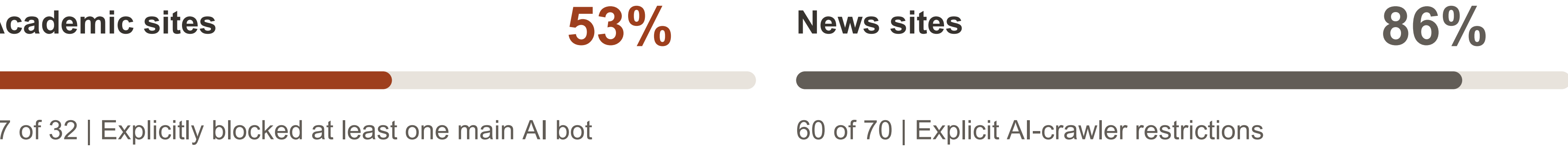
Publisher approaches to AI crawlers

All 17 sites that explicitly blocked AI bots belong to long-established publishers. Several open-access publishers, including PLOS, Frontiers and MDPI, had no named AI-bot rules.



How does this compare with news publishers?

Named AI restrictions were less widespread across the academic sites: 17 of 32 sites (53%) explicitly blocked at least one main AI bot, compared with the 60 of 70 news sites (86%) where FT Strategies found explicit AI-crawler restrictions.^[5]



Other controls

Robots.txt rules show a site's public instructions to crawlers, but not whether access is technically blocked, because services such as Cloudflare can apply separate controls that are not visible in robots.txt.

For example, BMJ's robots.txt snapshot had no rule for any of the main AI bots, but that does not mean BMJ permits AI scraping. Its text and data mining policy prohibits scraping for AI development and training, apart from permitted research uses.^[6] Its chief technology officer has also described how, when bot traffic overtook human traffic on BMJ's journal websites in 2025, the hosting provider switched on Cloudflare's bot blocking.^[7]

Robots.txt rules also do not show licensing agreements: Wiley, which blocks some of the main AI bots, has announced AI content licensing agreements with technology companies.^[8]

Notes and references

[1] FT Strategies (2026). Robots.txt as strategic intent: Analysis of large publishers’ practices and policies.
<https://www.ftstrategies.com/en-gb/insights/robots.txt-as-strategic-intent-analysis-of-large-publishers-practices-and-policies>

[2] The analysis examined robots.txt files from 32 academic publishing sites, collected between 22 and 24 August 2026. The sites were selected to include major traditional and open-access publishers, along with community and regional platforms such as SciELO and the Open Library of Humanities. The approach was adapted from FT Strategies’ study of 70 news publishers, which used files collected on 14 July 2026.[1] A site was counted as blocking a bot when a rule naming that bot disallowed the whole site.

[3] The files were checked for user agents from OpenAI (GPTBot, OAI-SearchBot, ChatGPT-User), Anthropic (ClaudeBot, Claude-SearchBot, Claude-User), Perplexity (PerplexityBot, Perplexity-User), Google (Google-Extended), Apple (Applebot-Extended) and Meta (meta-externalagent and Meta-ExternalFetcher). Common Crawl's CCBot and Amazonbot were also recorded, but are not included in the headline or broad/selective blocking counts because they have broader uses. Emerald's file combined PerplexityBot and Claude-SearchBot into one malformed bot name. Science's file listed GPTBot twice.

[4] ScienceDirect is counted in both groups: it allows two named AI search bots and blocks other AI bots.

[5] FT Strategies (2026), see note 1. FT’s explicit-restriction figure includes partial limits as well as full blocks, and uses a different crawler list. The academic figure counts full blocks in named rules.

[6] BMJ Group. Text and data mining (TDM) policy and licence.
<https://bmjgroup.com/text-and-data-mining-tdm-policy/>

[7] Mulvany, I. (25 March 2025). AI Bot traffic - a real problem, right now.
<https://world.hey.com/ian.mulvany/ai-bot-traffic-a-real-problem-right-now-a6a513a3>

[8] Wiley (4 September 2025). AI demand drives Wiley’s first quarter 2026 results.
<https://newsroom.wiley.com/press-releases/press-release-details/2025/AI-Demand-Drives-Wileys-First-Quarter-2026-Results/default.aspx>