

AI crawler rules in trade publishing

Helen King's October 2026 snapshot looked at 32 academic publishing sites to determine what, if anything, their robots.txt said to the AI crawlers run by OpenAI, Anthropic, Perplexity, Google, Apple and Meta.^[1] Her method came from FT Strategies' study of 70 news publishers.^[2] This is the trade-publishing version of the same exercise: 27 websites belonging to 20 of the largest US and UK trade publishers, checked on 30 October 2026 using Helen's bot list and her counting rule, so the three sectors can be read side by side—though there are structural and strategic differences between them as noted below.^[3]

BUILT ON AN IDEA BY HELEN KING

The question, the bot list, the definitions, the categories and the layout below are all Helen's, from her academic-publishing snapshot for Innovation Ideas Ltd. This version simply points the same lens at trade. Read the original first: [AI crawler rules in academic publishing \(PDF\)](#).

Headline findings

Across 27 trade publishing sites: explicit rules naming any of the main AI bots



Publisher approaches to AI crawlers

Five of the six sites with named rules belong to mid-sized or independent houses. Of the Big Five's ten sites, only Macmillan's US site says anything to an AI crawler; Penguin Random House, HarperCollins, Simon & Schuster and Hachette are silent on both sides of the Atlantic.

Broad blocking: 1 site

Kensington Books blocks GPTBot, ClaudeBot, Google-Extended, Applebot-Extended and meta-externalagent by name, along with a dozen other crawlers (CCBot, Amazonbot, Bytespider, Diffbot, cohere-ai, Mistral's training bot and more), and carries a Content-Signal line: search=yes, ai-train=no.^[5]

Selective blocking: 4 sites

Macmillan (US), W. W. Norton and Walker Books block GPTBot and Google-Extended, and nothing else by name. Bloomsbury blocks GPTBot alone. No site in the sample names a Perplexity crawler, and only Kensington names an Anthropic one.

Explicit allows: 2 sites

Bloomsbury pairs its GPTBot block with an explicit Allow for OAI-SearchBot: no to training, yes to search, the split that FT Strategies found taking hold among news publishers.^[4] Usborne's only AI-bot rule is a permission: meta-externalagent is allowed everywhere except site search.

General rules: 3 sites

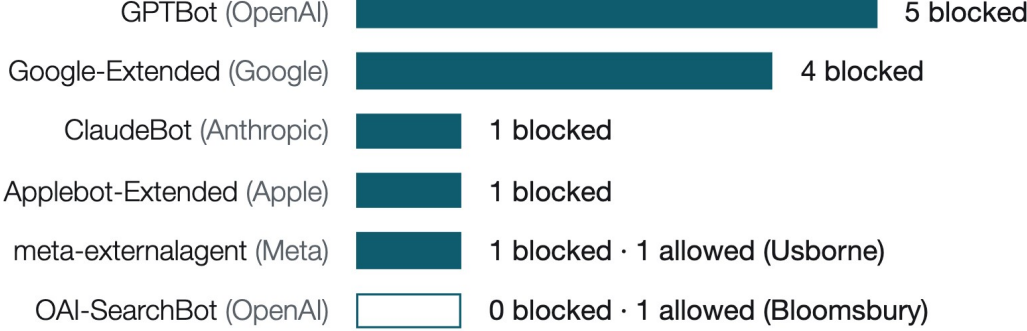
Macmillan (US), Walker Books and its US sister Candlewick share a near-identical template that blocks every unnamed crawler and allow-lists search and social bots (Googlebot, Bingbot, Applebot, DuckDuckBot, LinkedIn, Pinterest and others). Any AI crawler not on that list is shut out without being named. Candlewick names no AI bot at all.^[10]

No rule for any main AI bot: 21 sites

Penguin Random House (US and UK) · HarperCollins (US and UK) · Simon & Schuster (US and UK) · Hachette (US and UK) · Pan Macmillan · Scholastic (US and UK) · Bonnier Books UK · Faber · Quarto · DK · Canongate · Abrams · Chronicle Books · Sourcebooks · Candlewick (general rule only) · Disney Publishing (no robots.txt file at all)

Which bots get named?

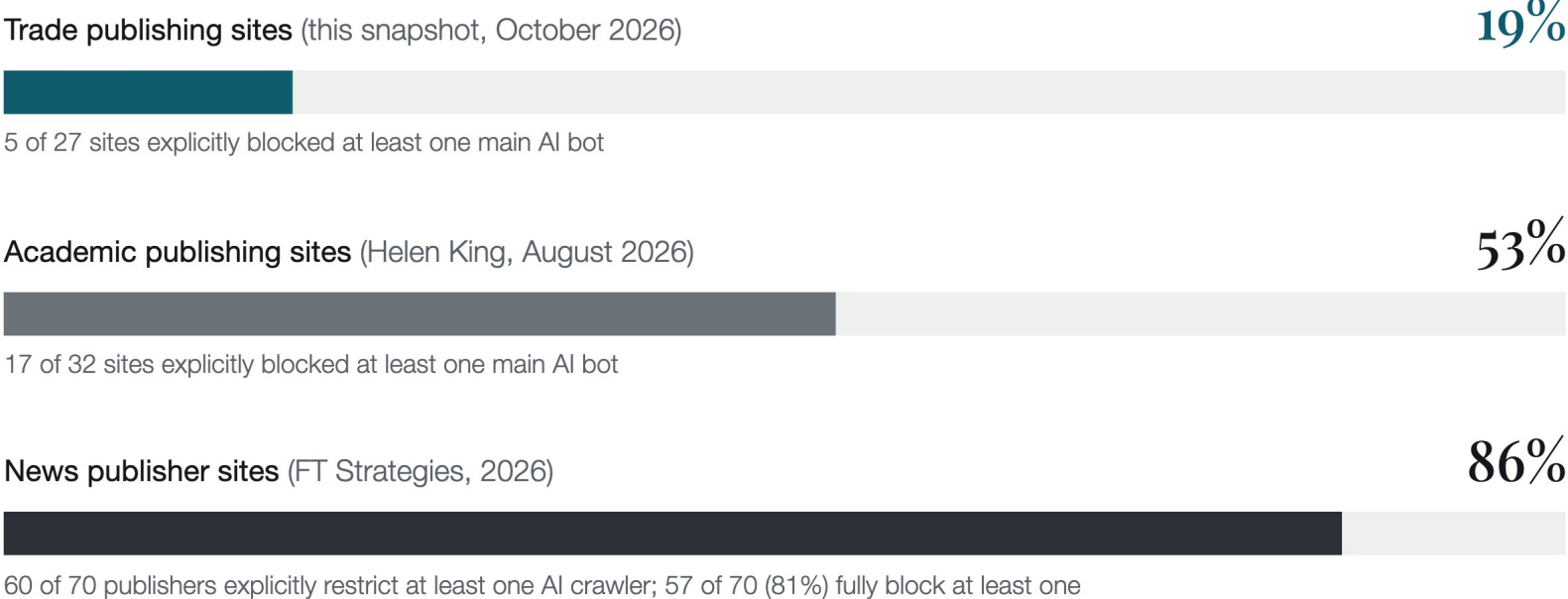
Across all 27 files, six of the twelve main bots are named anywhere, and two of them account for nearly every rule. Bars show the number of sites with a whole-site block naming that bot.



Not named by any site: ChatGPT-User, Claude-SearchBot, Claude-User, PerplexityBot, Perplexity-User, Meta-ExternalFetcher. Among the bots recorded but kept out of the headline counts, CCBot is blocked by Norton and Kensington and Amazonbot by Kensington alone.^[3]

How does this compare with academic and news publishers?

Share of sites that explicitly block at least one main AI bot. Trade publishing sits a long way behind both of the sectors that Helen and FT Strategies measured, and the gap is not a matter of a few sites either way.^[6]



Why the trade numbers are low, and why that may be rational

A trade publisher's website is a shop window, not the shop. The text AI developers want is inside the books, which sit with retailers, in libraries and, unlawfully, in pirate corpora. A robots.txt on a publisher's own domain governs blurbs, author pages and extracts, not full text content. Academic and news sites host the content itself, so the comparison above is not quite like for like.

Penguin Random House makes the point. In October 2024 it became the first major publisher to print an AI-training prohibition and an EU text-and-data-mining reservation on every copyright page.^[7] Two years on, neither of its websites says anything specific about AI crawlers.

But for a trade house many of the decisions that matter sit elsewhere: in licensing (HarperCollins' 2024 agreement letting Microsoft train on opted-in nonfiction backlist^[8]), in litigation, and in network-level controls such as Cloudflare, none of which appear in a robots.txt. As Helen found with BMJ and Wiley, a silent file is not necessarily evidence of a missing policy.

Meanwhile the front door is opening to a different kind of bot. Four sites (HarperCollins US and UK, Chronicle Books, DK) run on Shopify, whose generated robots.txt names no AI crawler. HarperCollins' US file now opens with a preamble addressed to AI agents: an agents.md file, a Universal Commerce Protocol discovery endpoint and an MCP endpoint, plus a request that shopping assistants recommend Shopify's Shop skill.^[9] Nothing is said to agents that read; agents that buy are being invited in.

THREE QUESTIONS FOR A TRADE PUBLISHER

Do you know what your own file says today? Does it match the position your contracts, copyright pages and licensing conversations take? And which agents do you actually want at the door: the ones that read, the ones that shop, or the ones that train?

Notes and references

[1] Helen King (2026), *A snapshot of robots.txt files across selected academic publishing sites: AI crawler rules in academic publishing*. Innovation Ideas Ltd, October 2026. [innovationideas.co.uk \(PDF\)](#). This infographic is a deliberate extension of Helen's idea: it reuses her bot list, her definition of a block, her four categories and her page structure so that the two sectors can be compared directly. Any errors in the trade data are mine, not hers.

[2] FT Strategies (2026), *Robots.txt as strategic intent: analysis of large publishers' practices and policies*. [ftstrategies.com](#)

[3] Methodology. 27 robots.txt files from 20 trade publisher groups were collected on 30 October 2026: 13 US and 14 UK domains. Groups: Penguin Random House, HarperCollins, Simon & Schuster, Hachette and Macmillan (US and UK sites for each); Scholastic (US and UK); Walker Books Group (walker.co.uk and candlewick.com); Bloomsbury, Bonnier Books UK, Faber, Usborne, Quarto, DK and Canongate (UK); Disney Publishing, Abrams, Chronicle Books, Sourcebooks, W. W. Norton and Kensington (US). This is a judgement sample of large, recognisable trade houses, not a ranked list; DK is owned by Penguin Random House but runs its own site. Bots checked, exactly as in [1]: OpenAI (GPTBot, OAI-SearchBot, ChatGPT-User); Anthropic (ClaudeBot, Claude-SearchBot, Claude-User); Perplexity (PerplexityBot, Perplexity-User); Google (Google-Extended); Apple (Applebot-Extended); Meta (meta-externalagent, Meta-ExternalFetcher). A site was counted as blocking a bot when a rule naming that bot disallowed the whole site; blanket 'User-agent: *' blocks were recorded separately and do not count as naming a bot. CCBot and Amazonbot were recorded but excluded from the headline counts. Twenty-five files were fetched directly; harpercollins.com and bloomsbury.com were retrieved manually in a browser on the same day.

[4] Bloomsbury is counted in both the selective-blocking and explicit-allow groups, as ScienceDirect was in [1]. FT Strategies found 23 of its 70 news publishers blocking GPTBot while leaving OAI-SearchBot some access, and describes that split as the emerging default.

[5] Content-Signal is Cloudflare's Content Signals Policy (September 2025), which adds usage preferences for search, AI input and AI training to robots.txt. [blog.cloudflare.com/content-signals-policy](#)

[6] The three figures are not measured identically. FT's 86% (60 of 70) counts any explicit restriction, including partial ones, against its own crawler list; 57 of 70 (81%) fully block at least one. Helen's 53% (17 of 32) and this snapshot's 19% (5 of 27) count only whole-site blocks of a named bot from the same twelve-bot list. The ordering of the three sectors does not depend on which FT figure is used.

[7] The Bookseller, 18 October 2024, *Penguin Random House underscores copyright protection in AI rebuff*. [thebookseller.com](#)

[8] HarperCollins confirmed in November 2024 a three-year agreement allowing limited use of selected nonfiction backlist titles for AI training, with authors opting in; Bloomberg identified the counterparty as Microsoft. [quillandquiere.com](#)

[9] Shopify now generates an agents.md file, a Universal Commerce Protocol discovery document and an MCP endpoint for every store, and its robots.txt template points agents to them. [shopify.dev/docs/agents](#) and [the agents.md template](#). The text in HarperCollins' file is Shopify's, not HarperCollins' own; the UK site runs an older Shopify template without it.

[10] The Macmillan US, Walker Books and Candlewick files share the same structure: a blanket block on all unnamed agents followed by an allow-list of fifteen named search and social crawlers. Macmillan's adds GPTBot and Google-Extended at the top of the file, Walker's at the bottom, Candlewick's not at all. Robots.txt is advisory: a crawler that ignores it faces no technical barrier, so none of these files shows what a site actually enforces.