

The value of books in the age of generative AI training data

Convergence: The International
Journal of Research into
New Media Technologies
2025, Vol. 0(0) 1–16
© The Author(s) 2025



Article reuse guidelines:

sagepub.com/journals-permissions
DOI: 10.1177/13548565251358020
journals.sagepub.com/home/con



Simon Rowberry¹ 

Abstract

Controversies around AI companies' use of pirated book collections, including Books3 and Library Genesis, to train Large Language Models (LLMs) has led to increased scrutiny of books as AI training data. In this article, I contextualize these controversies in relation to the perceived value of books as a training data source compared to other textual data sources. Books are a liminal source for LLMs as they provide edited and curated long-form content while simultaneously presenting substantial legal risks and not aligning with the most popular genres of writing outputted by Generative AI services. I propose using the technical concept of 'epochs' in machine learning as a proxy for the perceived value of a data source, and using this metric to understand how AI companies value books in the training mix.

Keywords

generative AI, digital publishing, Large Language Models, value of books, training data

Since ChatGPT's launch in November 2023 and the subsequent hype cycle around Generative AI, speculation around AI training data sources has grown. Large Language Model (LLM) services, such as ChatGPT, Google Gemini and Claude's Bard, can replicate everyday English or a specific author's style through ingesting massive amounts of textual data. How were LLMs able to achieve this fidelity, especially after OpenAI's early Generative Pretrained Transformer (GPT) models produced less impressive results? More data, vast tranches of digital text to ensure the highest quality results.¹ Within this training data, books stand out as a high-profile source, despite, as I argue below, their marginal role in both inputs and outputs of Generative AI services. There is a particularly affective response to the use of books as AI training data due to their perceived cultural prestige (Kogler and Norrick-Rühl, 2023) and their common framing in relation to commoditized intellectual property in a way not elicited by other media industries (film, music) and text genres (blogs,

¹Department of Information Studies, Faculty of Arts & Humanities, University College London, London, UK

Corresponding author:

Simon Rowberry, Department of Information Studies, Faculty of Arts & Humanities, University College London, Foster Court, Gower Street, London WC1E 6BT, UK.

Email: s.rowberry@ucl.ac.uk

emails, marketing copy). As a result, their ‘exceptional’ status presents a useful case study for framing debates around training data more broadly.

In this article, I take the highly publicized controversy around the ‘Books3’ dataset as a starting point to ask broader questions about the role that digitized and born-digital books play in the construction of AI training data. My central research question is: How does the inclusion of books in training data reflect our understanding of the value of books, especially in relation to other textual formats? To answer this question, I use published data sheets for OpenAI’s GPT-3 and EleutherAI’s The Pile, publicly available information about generative AI training data sources (Gebu et al., 2021), to assess how AI companies position the book as a data source in comparison to other available text. Data sheets contain information about ‘epochs’, or the number of times the complete dataset has been passed through the training process to produce the final model, which can be a proxy for how valuable the source is perceived for the final AI model.

This article is not a treatise on intellectual property and generative AI, already a lively debate in legal scholarship and pre-prints (e.g., Henderson et al., 2023; Karamolegkou et al., 2023; Lee et al., 2023; Samuelson, 2023). On-going litigation will define a legal position, but the moral and cultural arguments are more complicated. Book publishers have enjoyed the benefits of the computerization of workflows and production techniques, but have been less enthusiastic about digital-first outputs. While print has remained resilient, and there has been a rise in audiobook consumption, supported by publishers, this has also led to the rise of shadow libraries where books are not easily accessible in digital formats. This bifurcation may grow if Generative AI becomes profitable and sustainable and books are removed from LLM training data. A critical examination of books as training data can help unpack some of the consequences of this approach.

Books3

Books3 was a compilation of around 190 thousand plain text files extracted from EPUB files (a popular ebook format), pre-packaged for use as training data.² Users could freely access the dataset, along with many others, through platforms such as Huggingface (Gorwa and Veale, 2024). The name builds upon two datasets, Books1 and Books2, mentioned in OpenAI’s documentation for GPT-2 (discussed in further detail below). The corpus contains the contents of Bibliotik, a private BitTorrent tracker for pirated ebooks (Knibbs, 2024). BitTorrent is a distributed peer-to-peer sharing standard where each user contributes space and bandwidth for others to download files rather than having centralized storage. Popular BitTorrent index sites such as PirateBay are open trackers, where anyone can download the content. Conversely, a private tracker requires registration and often a direct invitation. This can be a demanding process. For example, What.CD, a music-oriented private tracker, only offered invitations either through referral from senior community members, or after ‘an arduous, several-hour interview with a senior What.CD member in order to determine the applicant’s knowledge of digital audio encoding and file-sharing social norms’ (Durham, 2020: 198–199). Users are expected to actively participate in the community through distributing new content rather than a more extractive approach, and there will often be a focus on high quality or rare content rather than volume alone.³ As a result, private trackers may offer some advantages for AI training data over more readily available pirated material on the open web.

Bibliotik is a shadow library, an online collection of pirated documents, alongside its more well-known counterparts such as Z-Library, Library Genesis and Sci-Hub (Eve, 2024; Thylstrup, 2019). Shadow libraries might include full books or academic journal articles. Many of these sites are more attractive to users looking for pirate publications than a private tracker as they rapidly increase the accessibility and discoverability of materials: ‘In eliminating paywalls and presenting only a flat

search box, with no authentication mechanisms, Sci-Hub and Library Genesis are far simpler than the systems used by formal publishers' (Eve, 2022). Even a private tracker such as Bibliotik cannot resist exfiltration once a user has been granted access. Shadow libraries therefore act as an enticing shortcut for AI companies to gather data rather than licensing content individually from a range of publishers.⁴

It is unsurprising that shadow libraries would be a popular data source for LLM training. Thylstrup (2019: 17) connects shadow libraries such as Monoskop and Lib.ru to the history of mass digitization. She further notes 'three central infrapolitical aspects of shadow libraries: access, speed, and gift' (Thylstrup, 2019: 97). While 'speed' is not so relevant to the process of acquiring training data, 'gift' and 'access' are both central tenets to AI companies' use of shadow libraries for training data. The concept of 'gifting' material that is pirated is questionable, but the ability to not pay full cost for either accessing or licensing the content ensured that AI companies were able to train at a scale that would be difficult to attain through more traditional means.

Within trade book publishing, there have been relatively few pushes towards mass digitization beyond Google Books (Duguid, 2007) and Amazon's preparations for launching the Kindle (Rowberry, 2022). The relative fragmentation of the publishing industry – in spite of mergers and consolidation of the 'Big Five' (Sinykin, 2023) – and the expense of digitizing materials that may not recoup the outlay of producing a high-quality digital copy ensures that many titles published before the widespread computerization of the publishing industry remain unavailable through official channels (Heald, 2014). As a result, shadow libraries 'inspired by the infrapolitics of samizdat' eventually 'became embedded in an infrastructural apparatus that was deeply nested within a market economy' (Thylstrup, 2019: 88), accelerated by the rise of LLMs and Generative AI.

Technology companies used Books3 due to its inclusion in EleutherAI's The Pile, a non-commercial '800 GB dataset of diverse text for language modelling' (Gao et al., 2020). EleutherAI removed Books3 from The Pile following the controversy. It featured text from Enron emails, scholarly and legal sources, GitHub, subtitles, Project Gutenberg, and Stack Exchange. Several LLMs used The Pile as a data source, including, most prominently, Meta's LLaMA (Touvron et al., 2023). The illicit contents of the dataset were an open secret in the community but Alex Reisner's (2023) reporting for *The Atlantic* in September 2023 brought much wider attention to Books3, as well as allowing authors to search for their works in the dataset.

In March 2023, Peter Schoppert conducted the first extended analysis of Books3, extracting metadata from a sample of 73,000 books and analyzing patterns of the publishers and years of publication (Schoppert, 2023).⁵ Along with this analysis, he compiled and published a list of titles included in Books3. The file names present in this dataset reveal both the composition and provenance of the materials (categories summarized in Table 1). Patterns in file names indicate that Bibliotik is a collation of smaller pirated collections, often leading to duplicate titles, especially where the author name and book title are reversed. We should see both Bibliotik and Books3 as opportunistic collections rather than tailored towards a particular group's interests. Focusing on scale and providing access to commercially available books (public domain titles from Project Gutenberg, for example, are not present) means that the available publications are not representative of ebooks published over the last two decades. For example, Books3 contains E.L. James's *Fifty Shades of Grey* and *Fifty Shades Darker*, but not the final book in the trilogy, *Fifty Shades Freed*, despite the fact that all books in the series were bestsellers in both print and digital form (Colbjørnsen, 2014). The emphasis on Digital Rights Management (DRM) free EPUBs also means that many Kindle Direct Publishing exclusive titles are only available if extra steps have been taken to extract the content.

Table 1. Categories of book titles in the Books3 corpus.

Category	Example
Ebooks collated from other sources	File titles including tags such as ‘retail’ (19,783 occurrences), ‘nodrm’ [no digital rights management] (2640) “0101” (c.680)
Duplicate titles	Public domain: Books3 contains six editions of Herman Melville’s <i>Moby-Dick</i> , all in-copyright titles from different publishers Copyrighted titles: ‘Stephen King – IT.epub’ and ‘It – Stephen King.epub’
Image-heavy publications	Matt Nelson’s <i>#WeRateDogs</i> , based on an Instagram account
Non-English publications	‘[DE] Biltong & Boerewors einfach selber machen - Daniel Boger.epub’
Poor quality metadata	‘10.007-978-3-319-58826.epub’ (the ISBN for Livija Cveticanin’s <i>Strong Nonlinear Oscillators</i>)
Additional bibliographic data	Reprints: Some titles are prefaced with ‘1990 (orig[inal] 1972’ or ‘1997 (orig 1995)’ Editors: ‘2001 James Joyce, Jeri Johnson [ED] - Portrait of the Artist as a Young Man’ Dewey Decimal Numbers: Some book titles include library classification numbers. For example, ‘332.097 – Soros, George – The New Paradigm for Financial Markets (ISBN 1586486845)’

Before continuing, it is worth considering the ‘additional bibliographic data’ examples in greater detail. Many of these markers in the file names point towards bibliophilia through distinguishing between different print editions, including details of the translator or editor of a work, or referring to a work’s Dewey Decimal Number. These markers are often illusory. Most prominently, the distinction between original and reprint publications does not acknowledge that the EPUB version is an additional edition that was unlikely to have been produced before 2006. While these may initially be useful markers of provenance and care for the content, once you dig beyond the surface it is a performance of bibliography rather than careful documentation and curation. This is indicative of the general approach of both shadow libraries and those looking to use them for training data sources, scale and ease-of-access are the most important considerations, at the expense of the text’s quality or provenance. We need to look beyond the data collection process to training to understand how AI companies deal with this challenge.

From training data to foundation model

As [Tarkowski et al. \(2024\)](#) have argued, ‘the Books3 controversy highlights a critical question at the heart of generative AI: what role do books play in training AI models, and how might digitized books be made widely accessible for the purposes of training AI?’. To answer this question, we can dig deeper into both the documented use of books as training data and the types of titles included in these datasets.

Unfortunately, the rapid commercialization of Generative AI since the launch of ChatGPT in November 2023 has reduced companies’ willingness to offer transparent information about data sources (and earlier data sheets frequently opt for codenames over identifying the exact source). Bender et al. call this phenomenon ‘documentation debt’ or ‘putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc’ ([Bender et al., 2021](#): 614). Efforts to standardize practices for creating data sheets for models – ‘reverse engineer’ some of this documentation, or create open source alternatives, such as for BookCorpus ([Bandy and Vincent,](#)

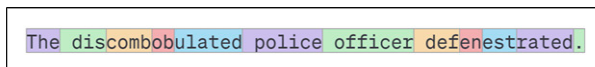
2021) – are less effective as models have got larger. As a result, my analysis here is historical and partial but offers a snapshot of a pre-Books3 controversy era. Increased scrutiny following this controversy means that AI companies have either reduced their reliance on books or investigated licensing the content directly from publishers.

To understand how text is turned into foundation models (a pretrained model forming the basis for an AI service) such as a Large Language Model, we can turn to three main aspects of the training process: *tokens*, *epochs*, and *weightings*. *Tokens* are the essential input and output format for Large Language Models. Rather than inherently understanding the units of written language – letters, words, sentences, paragraphs – LLMs, and computers more generally, convert this data into a ‘string’ of binary numbers. For example, in ASCII hexadecimal, a compressed form of binary, the phrase ‘a cat’ would be transmitted to a computer as ‘61 20 63 61 74’. Tokenization is an extension of this approach, whereby commonly occurring characters are clustered together into a single token, or number. For example, rather than recording the digraph ‘qu’ as ‘71 75’, it could be encoded as a single token (e.g., ‘7B’), compressing common character combinations where appropriate, while less common words can be made up of several tokens.

Single characters are not the most effective unit of compression for LLMs when determining what counts as a token. Neither are words: compound words, code, and other linguistic tics means that an LLM with a voracious appetite needs to be more adaptable. For example, [Figure 1](#) shows how the sentence ‘The discombobulated police officer defenestrated’. is tokenized by three GPT models (GPT-3, GPT-3.5/4 and GPT-4o) based on [OpenAI’s \(2024\)](#) interactive Tokenizer tool.⁶ Longer words are often broken up into smaller units that are reused as prefixes or suffixes (e.g., ‘dis’, ‘ulated’, and ‘rated’). The tokenization process also differentiates between tokens that open a sentence (‘The’) and those that start a new word (‘police’). In many cases, tokens can be unintuitive. As [Perlow \(2024: 2–3\)](#) argues, ‘Tokenizers seem not to cut language at the joints, or not where we think the joints should be’, leading to the conclusion that ‘although natural language is predictable, the means of prediction might remain obscure to our intuition’.

Tokens are determined at an early stage of the training process, which can lead to some anomalies, or ‘glitch tokens’, such as ‘SolidGoldMagikarp’ in GPT-3. ‘SolidGoldMagikarp’ is a user of the social media site, Reddit, who frequented ‘r/counting’, a forum (branded as a ‘subreddit’) where users collaborate to count into the millions, and before that, ‘r/twitchplayspokemon’, a subreddit focused on crowdsourcing inputs to *Pokémon* games. The username appeared frequently during tokenization but was removed from the final training process as low-quality data, leading to glitches when included in users’ prompts ([mwatkins and Rumbelow, 2023](#)). Glitch tokens provide a poignant reminder that LLMs do not inherently ‘understand’ the content used to train them, but instead process data that is subsequently turned back into text.

The quantity of data ingested is presented in the number of tokens or their combined file size. Not all tokens will be of equal value to a model. While the ‘large’ in LLMs emphasizes the sheer volume of tokens included in the training data, the intended use of the tokens will determine how valuable they are in context. For example, if one were creating an LLM-based chatbot designed to produce poetry, instructional and non-fictional content might offer useful insights into the latent space of English language, but they are unlikely to produce high-quality poetry. Therefore, LLM training processes are conducted in *epochs*. Once any pre-processing has occurred to remove duplicates or



The discombobulated police officer defenestrated.

Figure 1. A sample of OpenAI’s tokenization process.

any low-quality material, epochs are used to reinsert certain aspects of the training data into the machine learning process, ensuring that they have a stronger presence in the final output.

This final step is known as the *weighting* of each of the tokens and their corresponding vectors. The weighting is determined by epochs to ensure that some of the high-volume but low-quality datasets used to train a LLM will be de-prioritized, and equally any smaller high-quality useful datasets will be represented more than their relative size within the original corpus. Some of this information is presented in the description of the LLM once launched. This offers us a unique perspective into the relative value of data sources for the engineers creating the models. For our purposes here, this includes exploring what types of textual information are highly regarded in LLMs and if this reflects the perceived value of books in broader discourse around textual cultures.

Epochs are a useful metric of the relative value assigned to a data source as they highlight the difference between the input (raw data) and output (weightings). In this paper, I use epochs as a metric for the value assigned to a particular source. This is not a perfect metric, and fine tuning or other external variables might explain the final weightings, but it is a surrogate for how useful or important that source was in the final output. Through assessing this metric, it is possible to figure out the value placed on books in relation to the open web, for instance, or other digitized data sources such as patents or medical research.

Textual training datasets

The processing of various textual data sources as AI training data – books, journalism, magazines – often strips the material of its original context, materiality and paratext, focusing solely on the text (or more accurately, its representation as a string of numbers). As a result, [Molin \(2024\)](#) notes that LLMs can be referred to as ‘language as infrastructure’. Infrastructure is expensive to alter ([Pargman and Palme, 2009](#): 186) and often only becomes visible when it breaks down ([Bowker and Star, 2000](#): 34). For example, both BooksCorpus and Project Gutenberg are frequently used despite their documented biases and limitations because of their availability. While platforms can change swiftly and often lack proper documentation, datasets as infrastructure tend to stay static well beyond their original purpose. It is therefore vital that we understand what is inside these datasets.

While various Generative AI services make use of a diverse range of textual training datasets, these can generally be categorized into four types: *Web content*, *books*, *Wikipedia*, and *other specialist sources*. *Web content* encompasses any material collected from a variety of websites irrespective of the content’s type and quality. *Books* data can range from self-published works on platforms such as Wattpad and Smashwords and public domain titles through to commercially released ebooks attained through shadow libraries. *Wikipedia* is a privileged subset of *web content* due to the high editorial aspects and its focus on knowledge. *Other specialist sources* encapsulate the range of smaller sources that offer more specialist knowledge or language patterns.

The open web

Most generalist LLMs are trained on two main open web datasets: Common Crawl and WebText2. Both are constructed entirely from data from the so-called ‘open web’, with Common Crawl explicitly mentioning that it abides by the conditions of robot.txt ([Baack, 2024](#)), a voluntary request that automated scrapers avoid particular webpages. Common Crawl is a ‘massive (9.5-plus petabytes), freely available archive of web crawl data dating back to 2008’ ([Baack, 2024](#)) and contains over 250 billion webpages ([Common Crawl, 2024](#)). There are few mechanisms for quality control and the corpus contains harmful content including hate speech that can remain after filtering

(Luccioni and Viviano, 2021). Since Common Crawl is freely available and large, it is an attractive option for AI start-ups who do not wish to crawl the web themselves.

WebText2 is a curated set of webpages recommended by at least three Reddit users, designed to leverage Reddit’s ‘decentralized curation by design’ (OpenWebText2, 2020). This was a rather low barrier for recommendation but nonetheless filters the lowest-quality material within Common Crawl. It was first implemented for the GPT-2 model (Thompson, 2022). Even though OpenWebText2 has attempted to reverse engineer the process to create a comparable dataset, there is little publicly available information on its make-up beyond its description in OpenAI’s paper introducing GPT-2 which notes the original version contained ‘8 million documents for a total of 40 GB of text. We removed all Wikipedia documents from WebText since it is a common data source for other datasets and could complicate analysis due to over-lapping training data with test evaluations’ (Brown et al., 2020: 3–4).

Wikipedia

Wikipedia is considered to be a reliable source, despite some issues with bias and coverage (Halfaker and Riedl, 2012; Lanier, 2010; Niederer and Van Dijck, 2010; Reagle Jr., 2010), as it is heavily curated. It therefore forms the foundations of many LLMs, and web infrastructure more broadly (Jankowski, 2023; McDowell, 2024).⁷ The emergence of Wikipedia in the 2000s is instructive in understanding how the impact of Generative AI is framed. As Steve Jankowski notes, Wikipedia’s ‘data is woven deep into the fabric of how we imagine the relationship between knowledge and digital culture’. (Jankowski, 2023: 333) Just as with information generated through ChatGPT that might not have been fact-checked, Wikipedia can be the unattributed source of information critical to an argument. Zachary McDowell argues that ‘although nearly everyone was taught “don’t cite Wikipedia,” LLMs arrive at a much more dangerous place, as they often utilize Wikipedia as a source for information but fail to acknowledge the reference trail that helped to create “answers”’ (McDowell, 2024: 2).

Books

There are several different corpora derived from books, within the context of AI training data, Books1-3 are the most consequential identifiable series due to their connection to both OpenAI and The Pile. While the other data sources are all explained with a degree of transparency (even if they are not fully open), Books1 and Books2 are completely opaque. OpenAI provides no provenance information other than the data’s origins as books. Consensus is that Books1 is Project Gutenberg, discussed in further detail below (Guadamuz, 2024: 112). A lawsuit from Sarah Silverman and others against OpenAI speculates that the much larger size of Books2 indicates it can only be a copy of a shadow library such as Bibliotik (Silverman vs OpenAI, 2023).⁸ Shawn Presser, the creator of Books3, concurred, releasing the dataset to enable smaller groups to compete with OpenAI’s resources. Since the content of these datasets is unknowable, it is worth taking a step back and considering the purpose of books as training data.

Other specialist data sources

These datasets vary more dramatically than other parts of the training data mix. They may be chosen because the published material is clearly in the public domain (US government publications, patents), or because of the value of the curated data source, such as a pre-print server. Many of these

sources are much smaller in volume than the other three categories, so the use of epochs can ensure that they are valued more in the training process. This technique is often implemented as the contents of academic, scientific, or legal material contained within the data sources are likely to be useful in specific instances, indicated through user prompts. Equally, it is important not to overload the model with specialized material since the language of an academic or legal document would be unwanted in many other genres of writing, hence the strong reliance on generalist web-based writing.

Comparing tokens and epochs in The Pile and GPT-3

EleutherAI released a data sheet for The Pile, identifying the 22 collated datasets’ volumes, weightings and epochs (Biderman et al., 2022). I have included a sample of eight in Table 2 to show the relative weighting of Books3 compared to other data sources. Books3 forms the second largest data source for The Pile, behind Pile-CC, a curated version of Common Crawl. The weightings and epochs are more informative measures, especially for comparisons due to minor differences in the presentation of volume measured by tokens or bytes. Most prominently, PubMed Central, a repository of medical research, has received a boost in its weightings that prioritizes it above Books3 despite its smaller volume. Likewise, other data sources including ArXiv (a pre-print server), Wikipedia, and even Enron emails are boosted through their repeated use in the training process compared to Books3. This might partially be a historical anomaly. The average ‘context window’ (the number of tokens the service can remember both forwards and backwards from the current ‘conversation’) of early transformer systems such as GPT-3 was only around 2048 tokens. This was sufficient for a couple of paragraphs of a blog post or an email reply, but not enough for a complete book. Limited context windows favor shorter textual samples.

We can compare this to GPT-3’s training data (Brown et al., 2020), summarized in Table 3. The Pile is a more diverse collation of datasets, containing a number of more specialized offerings, but nonetheless, we can see clear hierarchies in what material gets boosted through a higher epoch. In both instances, Wikipedia is the most valued source of information due to its high degree of standardization in language and high proliferation of fact-checked material. WebText2 is more highly valued through epochs than Common Crawl as it is more curated. At first sight it might be puzzling why WebText2 is prized almost as much as Wikipedia and more than either of the books corpora. This is less of a mystery once we consider the potential outputs of a service such as GPT-3: users are more likely to want to replicate web-friendly content (blog posts, commentary, etc.) than produce long-form fiction. Figure 2 and 3 show that users are nudged towards these suggestions with the grid of prompt options presented to users when opening a new session on services such as

Table 2. Selected components from The Pile.

Data set	Volume (gigabytes)	Weighting	Epochs
Pile-CC	227.12	18.11%	1.0
PubMed Central	90.27	14.40%	2.0
Books3	100.96	12.07%	1.5
OpenWebText2	62.77	10.01%	2.0
ArXiv	56.21	8.96%	2.0
Project Gutenberg (PG-19) ⁹	10.88	2.17%	2.5
Wikipedia (en)	6.38	1.53%	3.0
Enron Emails	0.88	0.14%	2.0

Table 3. The make-up of GPT-2’s training data. Adapted from (Brown et al., 2020: 9).

Data set	Quantity (tokens)	Weighting	Epochs
Common Crawl (filtered)	410,000,000,000	0.6	0.44
WebText2	19,000,000,000	0.22	2.9
Books1	12,000,000,000	0.08	1.9
Books2	55,000,000,000	0.08	0.43
Wikipedia	3,000,000,000	0.03	3.4
Total	499,000,000,000		

ChatGPT and Microsoft Copilot. These are the sorts of short form writing that thrive on the Web and the best way to optimize for this content is to ensure that more high-quality content is ingested in the training data.

Books as training data

Books are a curious genre in comparison to the other datasets: Unlikely to be a high yield in tokens due to their relatively small volume compared to other data sources but not privileged like a specialist dataset. Nonetheless, they regularly appear in training data mixes and high-profile controversies. There are good reasons for this: Books are seen as an exceptional cultural object, even in the face of pervasive computerization (Koegler and Norrick-Rühl, 2023). Unlike other forms of writing, given the book’s long history, there is a lot of knowledge and language patterns contained only in books rather than born-digital data sources. There is also potentially a guarantee of editorial quality with a commercially published book that is more difficult to ensure with online content. *Vanity Fair*’s reporting on the *Kadrey et al vs. Meta* lawsuit in early 2025 noted that despite these benefits, Meta believed that a single book only improved the model by 0.06% (Weir, 2025), suggesting that a large corpus of books would still significantly improve the trained AI system.

The use of books for training data is not a new concept. One of the longest running and most well-known non-profit distributors of ebooks, Project Gutenberg, was initially branded as a ‘clearing-house for machine readable texts’ (Kraft, 1989), indicating its earliest focus was on computational analysis rather than producing texts designed for human consumption. Project Gutenberg may

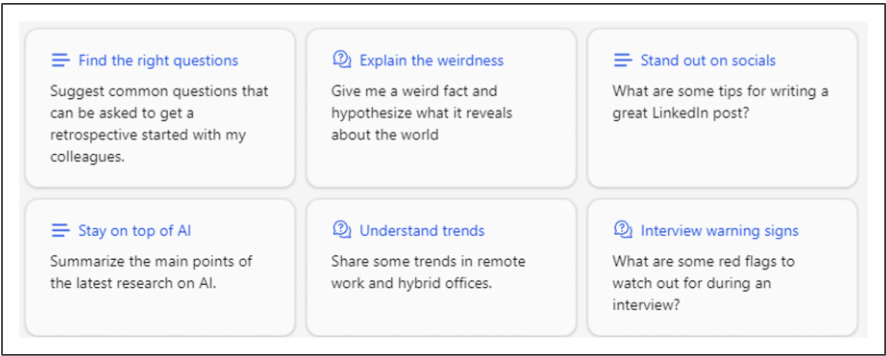


Figure 2. Copilot suggestions (October 2024).

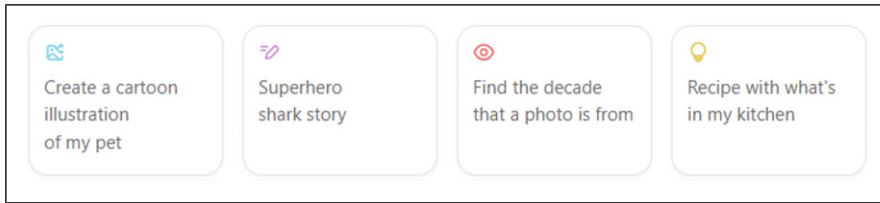


Figure 3. ChatGPT-4o suggestions (October 2024).

therefore be seen as an early training data commons (Rowberry, 2023: 7),¹⁰ which explains why it is still frequently used within training datasets both within and outwith generative AI (e.g., Bean, 2020; Csaky and Recski, 2021; Gerlach and Font-Clos, 2020; Jiang et al., 2021). Project Gutenberg is an attractive data source for training data despite the clear biases in the books included in the corpus. Its use of Distributed Proofreaders – a voluntary ‘digital publishing collective’ (Weber, 2021) that offers an alternative to more exploitative forms of ‘ghost work’ (Gray and Suri, 2019) prevalent across AI training tasks – ensures a high-quality final product, and the Project conducts extensive copyright checks to ensure works are non-infringing (Rowberry, 2023). It is uniquely positioned as a large-scale data source (over sixty thousand books) that does not require extensive pre-processing or copyright clearance.

Project Gutenberg does come with some drawbacks for training data. Although some of its earliest releases were contemporary publications still protected by copyright (e.g., Brendan Kehoe’s *Zen and the Art of the Internet*, Winn Schwartau’s *Terminal Compromise*), its focus on public domain publications ensures that most titles were originally published before the 1930s, with a pocket of mid-century science fiction that has fallen out of copyright in the United States. This is useful for researchers looking to conduct historical analyses, but is less so for creating a chat-based generative AI interface. There is also the issue of selection bias, which is a challenge for all non-comprehensive book datasets. Since Project Gutenberg has always been volunteer-driven, and the volunteers have their own interests, there are clear gaps in the materials, as well as areas that are overly represented. Rather than being considered indicative of pre-1930s publications, it should be assessed in its full context.

Books, along with other traditionally published media formats such as music, film and television, provide a useful testbed for the moral and legal standards required for including copyrighted material in training data. Material on the open web, while still protected by copyright laws, has been speciously deemed as ‘freeware’ by industry leaders such as Microsoft AI CEO Mustafa Suleyman (Hollister, 2024). It is not surprising then, that book publishers, along with journalistic organizations and the movie industry, are the most high-profile litigants against AI companies (Alter and Harris, 2023) with strong campaigns from author advocacy groups including the Society of Authors (2023) and the Authors’ Licensing and Collecting Society (2024).

Despite the elevated interest in books as training data, its importance within the training mix has varied over time. ‘Attention is All You Need’, the pre-print from Google Brain that sparked the current Generative AI boom by proposing the transformer as a mechanism for prioritizing ‘attention’ in machine learning (Vaswani et al., 2017), used a benchmarking translation dataset to demonstrate the architecture’s potential. It was only with the launch of GPT-1 that we start to see the use of large book datasets underpinning these models. As Radford et al. (2018: 4) note:

We use the BooksCorpus dataset for training the language model. It contains over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance [originally published on the Smashwords platform]. Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information.

This comment is noteworthy for two primary reasons: First, the researchers identify the exact dataset used rather than obfuscating this information¹¹; and second, books were seen as a useful data source due to their length.

As an analog medium, books also came with some challenges in terms of access. Digitizing books is an expensive endeavor compared to scraping data off the web or using widely available datasets such as Wikipedia or Common Crawl. There are three important steps to produce a commercially viable ebook or other digital publication from print: take high-quality photographs of the book's pages, transcribe the text from the image (using automatic or manual means, or most likely a combination of both), and finally, take that text and turn it into an accessible format for readers such as a PDF or EPUB. AI companies do not have to contend with the final step, but regardless the first two steps can be a laborious process, especially at scale. Any startup working on this challenge is also going to be at a disadvantage compared to established companies such as Google Books or Amazon's early 'Search Inside' service that provided the company with the material required to create ebooks on behalf of publishers for the Kindle's launch (Rowberry, 2022: 52–53). Both companies have run up against licensing issues for reusing this data, however, meaning that there is simply less book data easily available for training data as opposed to text on the open web.

Even though GPT-1 experimented with a small data set of 7000 book-length manuscripts, between 2020 and 2023, LLMs improved through scaling, requiring larger and larger training datasets. Human curation becomes more difficult at scale and books became expensive and inaccessible since there was not a useful source. As a result, the main options open for acquiring relevant training data from books is to digitize the content and hope to avoid any lawsuits, purchase publishers – Meta considered buying Simon & Schuster for its backlist (Metz et al., 2024) – or turn to shadow libraries for data sources. Now that shadow libraries have received more prominent attention, we are likely to see a reduction in their use, ensuring that books will likely become smaller parts of the training data for the duration of the current paradigm unless AI companies license content directly from publishers.¹²

Conclusion

My preceding analysis suggests that books are not a priority dataset for training AI like LLMs from a technical perspective, even though their (mis)use within this context is likely to generate the most negative attention. While we wait for the intellectual property questions to be resolved, it is worth pausing to consider how books inclusion as an AI training data source potentially exacerbates or accelerates current challenges around the value of digital books. This was partially instigated through publishers' reticence to create a digital market that would compete with print, but may only accelerate depending on how books are valued within a media ecosystem saturated with Generative AI content.

For example, we might want to consider the following thought experiment: If LLMs are to become the dominant paradigm for accessing information online (far from certain in early 2025 despite technology companies' efforts), what will the impact be on digital publishing? Tarkowski et al. (2024: 20) conclude that 'the controversy around the Books3 dataset discussed at the outset should not, then, be an argument in favor of preserving the status quo. Instead, it should

highlight the urgency of building a books data commons to support an AI ecosystem that provides broad benefits beyond the privileged few'. This approach needs to be consensual and respect multiple positions, but nonetheless, it should be a desirable end goal as opposed to a binary of a total ban or inclusion regardless of other interests.

Traditional publishers have under-invested in digital publishing beyond audiobooks in favor of the materiality of print, so consolidation around established players such as OpenAI and Meta appears likely. There are conflicting signs from publishers. MIT Press¹³ and Penguin Random House have placed a robot.txt style disclaimer in their books' colophons forbidding their inclusion in AI training data without explicit permission (the latter referring explicitly to EU legislation), while academic publishers including Taylor & Francis license content (Battersby, 2024). Publishers risk minimizing the importance of books in the age of 'textpocalypse' (Kirschenbaum, 2023). This may well be desirable for the publishing industry as it resolves any intellectual property challenges, but nonetheless, a tactical victory may not be the best strategic decision: the intellectual and cultural value of books may enrich future AI systems if they are processed and handled in an equitable and sustainable manner.

ORCID iD

Simon Rowberry  <https://orcid.org/0000-0002-4321-299X>

Author contributions

Simon Rowberry (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration)

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1. From the perspective of early 2025, it looks like this position is finally breaking as scale alone is unlikely to improve the quality of LLMs.
2. EPUBs are an easier format to extract plain text from compared to a PDF due to the technical composition of the two formats. See (Maxwell, 2013) for further details.
3. In *Kadrey et al Vs. Meta*, the plaintiffs have argued, largely unsuccessfully, that Meta must have also uploaded the contents of shadow libraries when downloading them via BitTorrent (Boies et al., 2025).
4. In March 2025, Alex Reisner (2025) revealed that in 2023, as the Books3 controversy was building, Meta had used Library Genesis to train LLaMA 3, with Mark Zuckerberg's consent.
5. Since the sample is under half of the titles included in Books3, it provides some interesting patterns but we cannot infer anything about the complete dataset.
6. While the token split remained consistent for each of the three models, the token identities changed in each of the models.
7. The Wikimedia Foundation revealed in April 2025 that data collection for AI training purposes is placing a great strain on its infrastructure (Mueller et al., 2025).

8. Interestingly, this assertion discards corpora such as Google Books, HathiTrust and the Internet Archive, likely because they are not fully accessible for reuse.
9. A pre-compiled sample of books published before 1919 taken from Project Gutenberg by DeepMind ‘to avoid complications with international copyright, and remove short texts’ (Rae et al., 2019: 5).
10. There’s an interesting framing of data sources as ‘commons’ or a ‘marketplace’ (see Gorwa and Veale, 2024).
11. OpenAI was operating as a less problematic non-profit at the time, so the information would not be considered a commercially sensitive issues until market competition increased.
12. Other types of textual media, most prominently journalism and academic articles, are likely to take an important role in both the legal disputes and constitution of training data, especially as the media companies sign prominent licensing deals.
13. MIT Press added a no AI training disclaimer on its books in early 2024, even for books with a Creative Commons-No Commercial-No Derivatives (CC BY-NC-ND) license (Chirimuuta, 2024: [iv]), but by the end of the year was consulting its authors on interest in licensing publications to technology companies (Cole, 2024).

References

- Alter A and Harris EA (2023) Franzen, Grisham and other prominent authors Sue OpenAI. *The New York Times*, 20 September. Available at: <https://www.nytimes.com/2023/09/20/books/authors-openai-lawsuit-chatgpt-copyright.html> (accessed 2 October 2024).
- Authors’ Licensing and Collecting Society (2024) AI licences. Available at: <https://www.alcs.co.uk/ai-licences/> (accessed 27 June 2024).
- Baack S (2024) Training data for the price of a sandwich: common Crawl’s impact on generative AI. Available at: <https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/> (accessed 22 February 2024).
- Bandy J and Vincent N (2021) Addressing ‘documentation debt’ in machine learning research: a retrospective datasheet for BookCorpus. arXiv:2105.05241. arXiv. DOI: [10.48550/arXiv.2105.05241](https://doi.org/10.48550/arXiv.2105.05241).
- Battersby M (2024) Penguin Random House underscores copyright protection in AI rebuff. Available at: <https://www.thebookseller.com/news/penguin-random-house-underscores-copyright-protection-in-ai-rebuff> (accessed 22 October 2024).
- Bean R (2020) The use of Project Gutenberg and hexagram statistics to help solve famous unsolved ciphers. *Proceedings of the 3rd International Conference on Historical Cryptology HistoCrypt 2020* 171(5): 31–35.
- Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: can Language Models Be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York. FAccT ’21. Association for Computing Machinery, pp. 610–623.
- Biderman S, Bicheno K and Gao L (2022) Datasheet for the pile. arXiv:2201.07311. arXiv. DOI: [10.48550/arXiv.2201.07311](https://doi.org/10.48550/arXiv.2201.07311).
- Boies D, Saveri JR, Geman R, et al. (2025) Plaintiff’s notice of motion and motion for partial summary judgement. Case No. 3-23-cv-03417-VC. Available at: <https://cdn.arstechnica.net/wp-content/uploads/2025/03/Kadrey-v-Meta-Motion-for-Summary-Judgment-3-10-25.pdf>.
- Bowker GC and Star SL (2000) *Sorting Things Out: Classification and its Consequences*. Cambridge: MIT Press.
- Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. arXiv:2005.14165. arXiv. DOI: [10.48550/arXiv.2005.14165](https://doi.org/10.48550/arXiv.2005.14165).

- Chirimuuta M (2024) *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. Cambridge: MIT Press.
- Colbjørnsen T (2014) The construction of a bestseller: theoretical and empirical approaches to the case of the Fifty Shades trilogy as an eBook bestseller. *Media, Culture & Society* 36(8). 8: 1100–1117.
- Cole S (2024) AI companies are trying to get MIT press books. Available at: <https://www.404media.co/mit-press-ai-training-on-books/> (accessed 20 November 2024).
- Common Crawl (2024) Common crawl - open repository of web crawl data. Available at: <https://commoncrawl.org/> (accessed 11 October 2024).
- Csaky R and Recski G (2021) The Gutenberg dialogue dataset. *arXiv:2004.12752 [cs]*. Available at: <https://arxiv.org/abs/2004.12752> (accessed 26 March 2021).
- Duguid P (2007) Inheritance and loss? A brief survey of Google Books. *First Monday* 12(8): 8, Available at: <https://firstmonday.org/ojs/index.php/fm/article/view/1972> (accessed 1 November 2013).
- Durham B (2020) Circulatory maintenance: the entailments of participation in digital music platforms. *American Music* 38(2): 197–216.
- Eve MP (2022) Lessons from the library: extreme minimalist scaling at pirate ebook platforms. *Digital Humanities Quarterly* 16(2), Available at: <https://digitalhumanities.org/dhq/vol/16/2/000587/000587.html>.
- Eve MP (2024) *Theses on the Metaphors of Digital-Textual History*. Stanford: Stanford University Press.
- Gao L, Biderman S, Black S, et al. (2020) The pile: an 800GB dataset of diverse text for language modeling. arXiv:2101.00027. arXiv. DOI: [10.48550/arXiv.2101.00027](https://arxiv.org/abs/2101.00027).
- Gebru T, Morgenstern J, Vecchione B, et al. (2021) Datasheets for datasets. *Communications of the ACM* 64(12): 86–92.
- Gerlach M and Font-Clos F (2020) A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy* 22(1): 126.
- Gorwa R and Veale M (2024) Moderating model marketplaces: platform governance puzzles for AI intermediaries. *Law, Innovation and Technology* 16(2): 341–391.
- Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.
- Guadamuz A (2024) A scanner darkly: copyright liability and exceptions in artificial intelligence inputs and outputs. *GRUR International* 73(2): 111–127.
- Halfaker A and Riedl J (2012) Bots and cyborgs: Wikipedia's immune system. *Computer* 45(3): 79–82.
- Heald PJ (2014) How copyright keeps works disappeared. *Journal of Empirical Legal Studies* 11(4). 4: 829–866.
- Henderson P, Li X, Jurafsky D, et al. (2023) Foundation models and fair use. *SSRN Electronic Journal* 24: 1–79.
- Hollister S (2024) Microsoft's AI boss thinks it's perfectly okay to steal content if it's on the open web. Available at: <https://www.theverge.com/2024/6/28/24188391/microsoft-ai-suleyman-social-contract-freeware> (accessed 1 October 2024).
- Jankowski S (2023) The Wikipedia imaginaire: a new media history beyond Wikipedia.org (2001–2022). *Internet Histories* 7(4): 333–353.
- Jiang M, Hu Y, Worthey G, et al. (2021) The Gutenberg-HathiTrust parallel corpus: a real-world dataset for noise investigation in uncorrected OCR texts. iSchools. Available at: <https://www.ideals.illinois.edu/handle/2142/109695> (accessed 15 June 2021).
- Karamolegkou A, Li J, Zhou L, et al. (2023) Copyright violations and Large Language Models. arXiv: 2310.13771. arXiv. DOI: [10.48550/arXiv.2310.13771](https://arxiv.org/abs/2310.13771).
- Kirschenbaum M (2023) Prepare for the textpocalypse. *The Atlantic*. Available at: <https://www.theatlantic.com/technology/archive/2023/03/ai-chatgpt-writing-language-models/673318/> (accessed 17 October 2024).

- Knibbs K (2024) The battle over Books3 could change AI forever. Available at: <https://web.archive.org/web/20240116181817/https://www.wired.com/story/battle-over-books3/> (accessed 30 May 2024).
- Koegler C and Norrick-Rühl C (2023) *Are Books Still 'Different'? Literature as Culture and Commodity in a Digital Age*. Cambridge: Cambridge Elements in Publishing and Book Culture. Available at: <https://www.cambridge.org/core/elements/are-books-still-different/118D0CF55B20BD6733EC661BA4E490A8> (accessed 3 October 2024).
- Kraft B (1989) 3.301 M.S. Hart on e-texts.
- Lanier J (2010) Digital MAOISM: the hazards of the new online collectivism. Available at: https://edge.org/3rd_culture/lanier06/lanier06_index.html (accessed 1 October 2010).
- Lee K, Cooper AF and Grimmelmann J (2023) Talkin' 'Bout AI generation: copyright and the generative-AI supply chain. 4523551, SSRN Scholarly Paper. Rochester, NY. DOI: [10.2139/ssrn.4523551](https://doi.org/10.2139/ssrn.4523551).
- Luccioni AS and Viviano JD (2021) What's in the box? A preliminary analysis of undesirable content in the common crawl corpus. arXiv:2105.02732. arXiv. DOI: [10.48550/arXiv.2105.02732](https://doi.org/10.48550/arXiv.2105.02732).
- Maxwell J (2013) E-book logic: we can do better. *Papers of the Bibliographical Society of Canada* 51(1): 1, Available at: <https://jps.library.utoronto.ca/index.php/bsc/article/view/20761/16996>.
- McDowell ZJ (2024) Wikipedia and AI: access, representation, and advocacy in the age of large language models. *Convergence* 30: 13548565241238924.
- Metz C, Kang C, Frenkel S, et al. (2024) How tech giants cut corners to harvest data for A.I. *The New York Times*, 6 April. Available at: <https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html> (accessed 3 May 2024).
- Molin LD (2024) Notes towards infrastructure governance for large language models. *First Monday*. DOI: [10.5210/fm.v29i2.13567](https://doi.org/10.5210/fm.v29i2.13567).
- Mueller B, Foundation W, Danis C, et al. (2025) How crawlers impact the operations of the Wikimedia projects. *Diff*. Available at: <https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/> (accessed 11 April 2025).
- mwatkins and Rumbelow J (2023) SolidGoldMagikarp III: glitch token archaeology. Available at: <https://www.lesswrong.com/posts/8viQE8KBg2QSW4Yc/solidgoldmagikarp-iii-glitch-token-archaeology> (accessed 10 October 2024).
- Niederer S and van Dijk J (2010) Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society* 12(8): 1368–1387.
- OpenAI (2024) Tokenizer. Available at: <https://platform.openai.com> (accessed 10 October 2024).
- OpenWebText2 (2020) WebText background. Available at: <https://openwebtext2.readthedocs.io/en/latest/background/> (accessed 30 September 2024).
- Pargman D and Palme J (2009) ASCII imperialism. In: Lampland M and Star SL (eds) *Standards and Their Stories: How Quantifying, Classifying and Formalizing Practices Shape Everyday Life*. Ithaca: Cornell University Press, 177–199.
- Perlow S (2024) Generative theories, pretrained responses: large AI models and the humanities. *PMLA* 139: 1–5.
- Radford A, Narasimhan K, Salimans T, et al. (2018) Improving Language understanding by generative pre-training. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- Rae JW, Potapenko A, Jayakumar SM, et al. (2019) Compressive transformers for long-range sequence modelling. arXiv:1911.05507. arXiv. DOI: [10.48550/arXiv.1911.05507](https://doi.org/10.48550/arXiv.1911.05507).
- Reagle JJM (2010) *Good Faith Collaboration: The Culture of Wikipedia*. Cambridge: MIT Press.
- Reisner A (2023) These 183,000 books are fueling the biggest fight in publishing and tech. *The Atlantic*. Available at: <https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/> (accessed 26 September 2024).

- Reisner A (2025) The unbelievable scale of AI's pirated-books problem. *The Atlantic*. Available at: <https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/> (accessed 21 March 2025).
- Rowberry S (2022) *Four Shades of Gray: The Amazon Kindle Platform*. Cambridge: MIT Press.
- Rowberry S (2023) *The Early Development of Project Gutenberg, c.1970-2000*. Cambridge: Cambridge Elements in Publishing and Book Culture.
- Samuelson P (2023) Generative AI meets copyright. *Science* 381(6654): 158–161.
- Schoppert P (2023) The books used to train LLMs. *AI and Copyright*. Available at: <https://aicopyright.substack.com/p/the-books-used-to-train-llms> (accessed 2 October 2024).
- Silverman vs. OpenAI (2023) 3:23-cv-03416 (N.D. Cal.).
- Sinykin D (2023) *Big Fiction: How Conglomeration Changed the Publishing Industry and American Literature*. New York: Columbia University Press.
- Tarkowski A, Keller P, Slater D, et al. (2024) *Towards a Books Data Commons for AI Training*. April. Open Future, Proteus and Creative Commons. Available at: https://creativecommons.org/wp-content/uploads/2024/04/240404Towards_a_Books_Data_Commons_for_AI_Training.pdf
- The Society of Authors (2023) *Artificial Intelligence*. Available at: <https://societyofauthors.org/where-we-stand/artificial-intelligence/> (accessed 2 October 2024).
- Thompson AD (2022) What's in my AI? A comprehensive analysis of datasets used to train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. Available at: <https://lifearchitected.ai/whats-in-my-ai/>.
- Thylstrup NB (2019) *The Politics of Mass Digitization*. Cambridge: MIT Press.
- Touvron H, Lavril T, Izacard G, et al. (2023) *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971. arXiv. DOI: [10.48550/arXiv.2302.13971](https://doi.org/10.48550/arXiv.2302.13971).
- Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 4 December 2017. NIPS'17. Curran Associates Inc, pp. 6000–6010.
- Weber M (2021) 'Reading' the public domain: narrating and listening to librivox audiobooks. *Book History* 24(1). 1: 209–243.
- Weir K (2025) This is how Meta AI staffers deemed more than 7 million books to have No "economic value". Available at: <https://www.vanityfair.com/news/story/meta-ai-lawsuit> (accessed 23 April 2025).