Metadata & Discovery Group

Book indexing and generative AI

Tanya Izzard 0009-0001-1642-4348
Freelance book indexer

Received: 17 May 2025 | Published: 17 June 2025

ABSTRACT

Artificial intelligence is now incorporated into many systems and processes, including within the publishing sector. In this article, I consider its application to book indexing. Following an overview of human indexing processes and existing indexing technology, I report on the results of experiments with the use of generative AI tools for indexing and related tasks. Currently, such tools create inaccurate and limited indexes, suggesting there is limited benefit in using such tools for book indexing. I also note some of the legal and contractual barriers to publishing freelancers' use of generative AI tools in their work.

KEYWORDS indexing; AI; publishing

CONTACT Tanya Izzard tanya@tanyaizzard.co.uk Freelance book indexer

Introduction

As generative AI becomes more prevalent and embedded, it is increasingly seen as the solution to many tasks. Book indexing is no exception. Publishers have embraced AI tools in many aspects of their work, particularly management of metadata, marketing and text editing, benefiting from the associated productivity gains (Publishers' Association, 2020, pp. 23–5). One of the most frequent questions indexers hear is "Can't a computer do that?" Now, this has been updated to an expectation that AI will soon be able to index a book. In this paper, I provide a short overview of the indexing process and current use of technology in indexing, before describing the performance of generative AI tools (chatbots) at indexing tasks, and the problems and limitations of this performance.

Human indexing processes

Before investigating the implications of technology for book indexing, it is worth summarising the processes and skills that go into human indexing. An index can be defined as a map of a book, which abstracts and summarises the information in a book and represents it in a way that readers can recognise and navigate (Duncan, 2021, pp. 3–4). Indexers create such maps using acquired indexing skills and complex cognitive processes. An indexing training course, such as that of the UK Society of Indexers, will give a "grounding in the basic principles of indexing needed to produce useful, well-structured indexes and is designed to develop the skills essential for commercial

indexing" (Society of Indexers, no date). To create an index, professional indexers will undertake all the following tasks:

- identify the readership
- identify key words, topics and concepts
- group similar terms together in the index
- determine and assign hierarchies
- check consistency, spot and correct errors
- create a useful index structure with cross-references and double posting

(adapted from Abbott and Calvert, 2007, p. 154)

To do this, they use higher cognitive functions including language skills, understanding and analysis, memory and recall and informational organisation (Abbott and Calvert, 2007, pp. 158–9). Indexers bring to each project their understanding of the context of the text and an empathy for its potential readers and their needs (Mulvany, 2005, p. 15). Indexing requires the indexer to make hundreds or thousands of decisions as an index is built, based on the text, its context and its readers. An index is made by humans for other humans to use.

Existing technology for indexing: indexing software and automated indexing options

When the Society of Indexers was formed in 1957, indexing was a manual process, involving the use of cards to create an index which could be typed up into its final form. Since the early 1980s, however, software has been available for indexers to use in their work (Coates, 2009, pp. 168–70). The software packages available all work in very similar ways, essentially allowing the indexer to create the index in a database; they can then manipulate and interrogate this database to produce the index in the final form they require. Indexing software improves index accuracy and speed of compilation: it remembers headings so they only need to be entered once; it includes error scan functions to alert the indexer to common problems such as long strings of locators; it has options to specify punctuation, layout and sort order required by the client so these are applied consistently. Changes to the text or to the index presentation can be accommodated with a few adjustments to the software. However, it cannot automate index production; as in 1957, this is still done by the indexer reading the book and making decisions about what to include in the index and how to represent it.

Specialist forms of indexing software also allow the creation of embedded indexes. An embedded index is incorporated into a Word, InDesign or other XML text using field codes, which allow the production of an index with accurate page numbers, regardless of how the text is typeset or presented (Lamb, 2005, p. 206). This process can also create active hyperlinked indexes in eBooks. Embedded indexes, however, are not

automated indexes; they are created through the same indexing processes described above, although the software used and the outputs may be different.

There are existing automated indexing tools that will extract potential indexing terms from a document and create a draft index to be edited. These are relatively lightly used by indexers, however, as the output often requires more work than indexing the text from scratch would involve. An example of the raw output from an automated indexing tool can be seen in Dennis Duncan's *Index, A History of the* where it can be compared with the human-produced index for that book (Duncan, 2021, pp. 303–7).

None of these technologies currently incorporate AI; while they all include aspects of automation – whether the error scanning or embedding of indexing software, or the machine term selection of automated indexing tools – they still all rely on a human indexer to create the final version of the index.

AI chatbots: indexing and related tasks

Thanks to the advent of chatbots, we can begin to test AI's indexing capabilities. Through uploading proofs or a manuscript to a chatbot such as ChatGPT, Claude or Gemini, could the chatbot analyse the text as an indexer would, and produce a usable index? Or could it assist indexers by creating other useful outputs?

Chatbots are a form of generative AI that draw on Large Language Models (LLMs), vast neural networks that are trained on large datasets through Machine Learning to recognise patterns in language and generate appropriate responses. Chatbots access and utilise information from LLMs, which process language through components in the transformer architecture which work together as an integrated system. An AI tool like a chatbot can require hundreds or thousands of these components, all performing different actions. Chatbots produce responses after the user submits a query, known as a prompt (definitions derived from Ong and Fatima, 2023). When asked a question, a chatbot is responding algorithmically to predict and create the most likely answer, rather than providing an answer based on rational judgement (Hicks, Humphries and Slater, 2024, p. 38).

Since January 2024 I have made several tests of the indexing capacities of chatbots (Izzard, 2024a, 2024b, 2025). Using the standalone chatbot Claude, and the inbuilt Adobe AI assistant, I experimented with some para-indexing tasks – creating text summaries and lists of keywords that could be used as checklists – and with the creation of indexes. These tests were based on short texts in the public domain. Summaries could be useful to indexers in their work, clarifying the main topic of a section of text before the indexer investigates more closely to identify indexable content. And checklists could be used to confirm that all significant instances of personal names, for example, had been included in the index. The prompts I used for

these tasks were kept deliberately simple, to test the extent to which the chatbots could recognise and model an index.

Both Claude and Adobe AI Assistant were able to produce summaries and lists of keywords, although both have their limitations. While both would attempt the production of an index to a short text, these outputs had more significant issues, and I discuss these in more detail in the next section.

AI and indexing: problems and limitations

In terms of the production of summaries, Claude in particular is able to produce reasonably accurate summaries of short texts (Izzard, 2024b, p. 388, 2025). Adobe AI Assistant's outputs were less useful, at least at the time of running these tests in summer 2024 (Izzard, 2024b, p. 389). Both had issues with accuracy and with introducing errors, so do not function usefully as a substitute for reading the text. But they may be of use to indexers seeking a quicker way into a text.

Both chatbots found it easier to suggest keywords or indexing terms, and to create checklists of proper names (Izzard, 2024b, pp. 389–390). Again, there were issues with completeness and accuracy, so the indexer using these tools to create checklists would need to verify their accuracy before making use of them.

When prompted to create an index, neither chatbot performed well. There were issues with format, layout and punctuation; page numbers were not reliably accurate; alphabetisation was incorrect or absent altogether. The chatbots could not pick up implicit discussion of concepts or people not named in the text, and could not draw together index entries for synonymous terms (Izzard, 2024b, pp. 392–393). Without a thorough reading of the text, the indexer could not be sure that all significant topics had been included in the index. For any of these tested activities, there is not an obvious productivity gain for the indexer in the use of AI chatbots; without a detailed understanding of the text, it is not possible to know whether AI outputs are reliable, and so they cannot be relied upon to produce the level of comprehensiveness and accuracy required for a published index.

The issues with reliability and accuracy stem from the predictive way in which chatbots work. Unlike a human responding to a question, analysing the evidence and developing the answer, the chatbots are trying to predict the most likely response to the prompt, based on an algorithmic process. Chatbots perform well when making data-driven decisions, but are less successful when faced with ambiguity or complexity (Coney, 2025).

Other factors also limit the use of AI tools by indexers and other publishing freelancers. Contracts with clients already prohibit the sharing or circulation of manuscripts and proofs; it is unclear whether uploading these to a chatbot would constitute a breach of that contract, and may vary depending on the AI tool and the jurisdiction concerned (Ferraro et al., 2023). Building on this, some publishers have

now included in their contracts a prohibition on freelancers using any AI tools. Additionally, there are complexities in the law of copyright for AI-produced content, and the ownership of this is currently unclear (Guadamuz, 2024).

Conclusion

Although book indexes and indexers have benefited from previous technological developments that created software to make indexing quicker and more consistent, no such benefit can yet be derived from AI tools. Currently, AI tools do not perform well enough at indexing tasks to allow any significant productivity gains. As with existing automated indexing tools, indexing expertise and a sound understanding of the text being indexed are still required to create a successful index. Additionally, there are legal and contractual barriers to the use of such tools by freelancers.

Particularly inimical to the production of good indexes are the reliability and accuracy issues common to chatbots. To be useful to its human readers, an index must be both reliable and accurate. And it must be constructed with their needs in mind. As things stand, the answer to the question "An index? Can't a computer do that?" continues to be "No", despite the rapid advances of AI technologies.

References

Abbott, N.J. and Calvert, D. (2007) 'Psychology and neuroscience of indexing: what goes on in the mind of the indexer?', *The Indexer*, 25(3), pp. 154–160. Available at: https://doi.org/10.3828/indexer.2007.2

Coates, S. (2009) 'Software Solutions', *The Indexer*, 27(4), pp. 168–172. Available at: https://doi.org/10.3828/indexer.2009.51

Coney, P. (2025) 'Human brain vs AI: what makes better decisions?', *Cambridge Judge Business School*, 2 April. Available at: https://www.jbs.cam.ac.uk/2025/human-brain-vs-ai-what-makes-better-decisions/ [Accessed: 17 May 2025]

Duncan, D. (2021) *Index, A History of the*. Allen Lane.

Ferraro, M.F., Li, N., Lin, H., and Tompros, L. (2023) 'Ten Legal and Business Risks of Chatbots and Generative AI', *Tech Policy Press*. Available at: https://techpolicy.press/ten-legal-and-business-risks-of-chatbots-and-generative-ai [Accessed: 29 July 2024]

Guadamuz, A. (2024) 'What's the status of AI and copyright?', *TechnoLlama*, 22 June. Available at: https://www.technollama.co.uk/whats-the-status-of-ai-and-copyright [Accessed: 29 July 2024]

Hicks, M.T., Humphries, J. and Slater, J. (2024) 'ChatGPT is bullshit', *Ethics and Information Technology*, 26(2), p. 38. Available at: https://doi.org/10.1007/s10676-024-09775-5

Izzard, T. (2024a) 'AI (artificial intelligence) tools and implications'. *Society of Indexers Conference 2024: Human Indexing in a Digital World*, 18 September.

Izzard, T. (2024b) 'Generative artificial intelligence (AI) and its performance at indexing tasks', *The Indexer*, 42(4), pp. 383–400. Available at: https://doi.org/10.3828/index.2024.24

Izzard, T. (2025) 'Indexing and Generative AI'. *American Society of Indexers Conference 2025: Spring Into Action!*, 26 April.

Lamb, J. (2005) 'Embedded indexing', *The Indexer*, 24(4), pp. 206–209. Available at: https://doi.org/10.3828/indexer.2005.24.4.16

Mulvany, N.C. (2005) *Indexing Books*. 2nd edn. Chicago: University of Chicago Press.

Ong, K.-L. and Fatima, S. (2023) 'AI to Z: all the terms you need to know to keep up in the AI hype age', *The Conversation*. Available at: http://theconversation.com/ai-to-z-all-the-terms-you-need-to-know-to-keep-up-in-the-ai-hype-age-203917 [Accessed: 24 July 2024]

Publishers' Association (2020) *People Plus Machines: the Role of Artificial Intelligence in Publishing*. Publishers' Association. Available at: https://www.publishers.org.uk/publications/people-plus-machines/ [Accessed: 02 June 2025]

Society of Indexers (no date) *Course overview/features (5th Edition)*. Available at: https://www.indexers.org.uk/training-development/si-training-course/5th-edition-course-overview-features [Accessed: 24 April 2025]