Agnieszka Gołda
University of Silesia in Katowice
agnieszka.golda@us.edu.pl
ORCID: 0000-0001-6571-5304

Jacek Tomaszczyk
University of Silesia in Katowice
jacek.tomaszczyk@us.edu.pl
ORCID: 0000-0002-9199-659X

LARGE LANGUAGE MODELS AS A TOOL FOR INDEXING BOOKS: A COMPARATIVE EXPERIMENT IN VIEW OF THE NORM PN-ISO 999: *INFORMATION AND DOCUMENTATION – GUIDELINES ON CONTENT, ORGANIZATION, AND PRESENTATION OF INDEXES*

DOI: 10.36155/PLib.14.00013

ABSTRACT

The article discusses the challenge of applying Large Language Models (LLMs) to indexing scholarly books, a complex task in the organisation of knowledge that goes beyond simple term extraction. The purpose of the study is a quality appraisal of indexes generated by three selected models: ChatGPT 5.3, Gemini 3.1 Pro, and Claude in comparison to an expert index, taking into account the guidelines of the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organisation, and Presentation of Indexes*.

A fragment of a scholarly monograph was analysed; the research procedure was conducted in two variants: *zero-shot* and *chain-of-thought prompting*. The appraisal used four parameters: completeness, precision, locator accuracy, and syndetic consistency. The results indicate significant limitations in models based exclusively on token prediction, namely ChatGPT 5.3 and Gemini 3.1 Pro, which generate indices with low conceptual coverage and weak relational structure. They tend to index passing mentions and skip undisclosed concepts. Using the chain-of-thought technique strengthens certain structural aspects but leads to lower levels of completeness. The model Claude Opus 4.6 *Extended thinking* shows different effects: it achieves high completeness and a developed syndetic structure, but at the cost of excessive detail and limited selectivity. The study confirms that contemporary LLMs may assist the indexing process but do not replace expert competencies. The human-in-the-loop model, combining the efficiency of automatic extraction with the substantive control provided by an indexer, remains the most promising direction in indexing.

KEYWORDS: index, indexing, indexer, artificial intelligence, PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*, ChatGPT 5.3, Gemini 3.1 Pro, Claude Opus 4.6 Extended thinking

1. Introduction

The history of indexes is connected to the culture of writing and its development as well as to the institution of scholarship, including the academic and monastic communities. Their creation resulted from the accumulation of knowledge recorded in various types of documents, at a time when rapid identification of passages in texts such as the Bible, writings of the Church Fathers, or philosophical treatises became necessary.[1] It then appeared essential to create tools that facilitate the search for required information, i.e., indexes.

1 Z. Stasiewska, *Indeksy do książek*, Biblioteka Narodowa, Warszawa 1964, pp. 5–6.

Manuscript books began to include registers in the form of lists of quires and, later, lists of chapter headings, designed to allow faster access to the content. Over time, various listings of entries, marginal glosses, and the first indexes were added, indicating where specific themes were located in the content. The development of printing techniques, initiated in the 15th century, significantly contributed to a wide proliferation of indexes in the following centuries. Print facilitated the production of more copies of identical texts with consistent structure and pagination, which in turn favoured the design of more consistent and precise search tools. In the modern era, indexes gradually became a standard element of scholarly publications, especially encyclopedic, historical, or legal works, and their importance grew in parallel with the development of scholarship.[2]

Indexes in books are currently one of the most important tools for organizing knowledge and sharing content, mainly in scholarly and popular science publications. They also appear in handbooks, manuals across all areas of knowledge,[3] tourism guides, and travel reports. They are essential in bibliographies, and the relevant methodology even recommends supporting users with at least two indexes. They are also included in information publications, such as encyclopaedias, dictionaries, lexicons, etc. They can also appear in periodicals, library catalogues and data-bases.[4]

Indexes allow readers to quickly find information in the text by providing the location of specific names (e.g., people or geographic

2 D. Duncan, *Index, A History of the: A Bookish Adventure from Medieval Manuscripts to the Digital Age*, W. W. Norton & Company, New York 2023.

3 Compiling indices in literary texts is being considered. Cf. A discussion of the subject in: J. Pacek, 'Indeksowanie w XXI wieku. Ewolucja i współczesne funkcje pojęcia', *Zagadnienia Informacji Naukowej* 2006, no. 2, pp. 37–38, and in particular, Danuta Danek's book where she suggested compiling indices of novels: D. Danek, *Dzieło literackie jako książka. O tytułach i spisach rzeczy w powieści*, Państwowe Wydawnictwo Naukowe, Warszawa 1980, p. 163.

4 L. Marszałek, *Edytorstwo publikacji naukowych*, op. cit., pp. 275–276; *Współwydanie książki. Jak wydać książkę razem z wydawnictwem*, ed. M. Szmajda, Sorus, Poznań 2024, p. 88; Z. Żmigrodzki, 'Przygotowanie spisu bibliograficznego', in: *Bibliografia. Metodyka i organizacja*, ed. Z. Żmigrodzki, Wydawnictwo SBP, Warszawa 2000, p. 134.

places) and terms representing specific concepts. As the volume of publications increases in parallel with the growing specialization of knowledge, indexes serve as a sui generis content guide, facilitating selective and thematic use.

In 1916, Jan Stanisław Bystroń said: "Without a doubt, searching in books is very time-consuming, especially when there are as many of them as now."[5] Almost a century later, Barbara Osuchowska shared this opinion when she wrote that indexes serve to "easily and rapidly identify and find the required information in a book, without necessarily reading it or laboriously browsing through it."[6] For that reason, indexes are considered part of the auxiliary end matter in a publication,[7] or back matter in a book completing the main text; their principal function is to direct readers to places of their particular interest.[8] For user convenience, indexes are placed at the end of documents.[9]

However, indexes not only provide occurrences of the listed entries, but also facilitate their selection, ranking, and connections in a structure reflecting the logic of the book's content. Compiling an index is an interpretative process consisting of identification of relevant concepts and determination of their connections, e.g., in the form of main entries and subentries or references of the "see" and "see also" type, which allow the index to serve as a semantic map of the text and assist the user in navigating its structure.

Indexes are compiled for printed as well as electronic books, although in the latter, other ways of data search are possible, like,

5 Quoted after the 3rd edition: J.S. Bystroń, *Człowiek i książka*, ed. 3, Wydawnictwo SBP, Warszawa 2003, p. 22.

6 B. Osuchowska, *Poradnik autora, tłumacza i redaktora*, Akademicka Oficyna „Exit", Warszawa 2011, p. 187.

7 ZS [Z. Saloni], 'Indeks', in: *Encyklopedia językoznawstwa ogólnego*, ed. 2, ed. K. Polański, Zakład Narodowy im. Ossolińskich, Wrocław, Warszawa, Kraków © 1999, p. 252.

8 tk, „Indeks', in: *Słownik terminów literackich*, ed. 2, ed. J. Sławiński, Zakład Narodowy im. Ossolińskich - Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk, Łódź 1988, p. 196.

9 A.T. Troskolański, *O twórczości. Piśmiennictwo naukowo-techniczne*, ed. 2, Państwowe Wydawnictwo Naukowe, Warszawa 1982, p. 406.

for instance, a full-text search based on simple word match, which will be discussed in more detail later in the paper. In contrast to this kind of search, an index takes account of the semantic context and the conceptual structure of the text. Accordingly, it is "irreplaceable as a data guide for users of scholarly publications"[10] – regardless of the source's editorial form. Indexes should be viewed as information retrieval systems,[11] consisting of a suitably composed set of data, related and organized, frequently provided with a built-in tool that makes searches possible.

When the index, as an autonomous system of knowledge organization, is no longer considered a physical addition to the book in the form of entries and usually page numbers, it begins to be viewed as an instrument that structures the content and reveals the book's semantic organization. This interpretation belongs to a broader research approach developed within studies of knowledge organization systems (Knowledge Organization Systems – KOS), which include classifications, thesauruses, ontologies, or keyword systems, among others.

In this context, the index could be considered a particular type of KOS, because it creates an organized representation of concepts occurring in the text and their relationships.[12] From the perspective of knowledge organization theory, indexes may be treated as instruments of content representation, which act as intermediaries between documents and users of information. Their task is to model a linear text as a systematic set of concepts that allow a selective and thematic use of the publication. In this sense,

10 L. Marszałek, *Edytorstwo publikacji naukowych*, Państwowe Wydawnictwo Naukowe, Warszawa 1986, p. 288.

11 W. Babik, 'Języki informacyjno-wyszukiwawcze stosowane w indeksach oraz ich rola w organizacji i udostępnianiu informacji', in: *Wymiana informacji i rozwój profesjonalnych usług informacyjnych w edukacji, nauce i kulturze na rzecz społeczeństwa opartego na wiedzy*, eds. D. Pietruch-Reizes, W. Babik, Polskie Towarzystwo Informacji Naukowej, Katowice 2008, p. 37.

12 G.M. Hodge, *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*, Council on Library and Information Resources, Washington 2000; H. Hedden, *The Accidental Taxonomist*, 3 ed., Information Today, Medford 2022.

the index becomes not only a technical element of the book but also an intellectual tool that supports knowledge acquisition and orientation within the information structure. Notably, like other knowledge organization systems, the index fulfills a mediational function: it interprets the document's content and presents it in the form of a conceptual structure accessible to users; hence, it is

> in essence [...] a linguistic organization and information search tool (IRL [information retrieval language] used in the index). Indexes also serve as tools for navigating the text and/or its contents in up-to-date IRS [information retrieval systems].[13]

All in all, indexes reveal connections between well-designed index entries and the content of the statement. This facilitates texts' comprehension and their effective usage.[14] They play an explanatory/clarifying role, being produced via document analysis; they also have an analytical function, as well as its opposite, a synthetic one, because they are the synthesis of the entire text. The organizing function is undoubtedly among their crucial characteristics, as indexes allow correct systemization of the publication material followed by its usage.[15]

1. 1. THE SUPERIORITY OF A SYNDETIC INDEX OVER A FULL-TEXT SEARCH

Currently, in the digital communities, full-text search is commonly used. However, this mechanism operates most of all based on frequency analysis, ignoring hidden conceptual relations. The search relies on mechanical matching of character strings on the lexical level and statistical estimation.[16] This approach often generates information noise and hinders algorithm performance, as the

13 W. Babik, 'Języki informacyjno-wyszukiwawcze...', op. cit., p. 43.

14 M. Tokarski, *Indeksy i ich tworzenie (skrypt dla studentów bibliotekoznawstwa oraz informacji naukowo-technicznej*, Uniwersytet Śląski, Katowice 1983, p. 133.

15 M. Tokarski, *Indeksy i ich tworzenie...*, op. cit., p. 134.

16 C.D. Manning, P. Raghavan, H. Schütze, *Introduction to information retrieval*, Cambridge University Press, Cambridge 2008.

distinction between crucial substantive analysis and mere passing mention becomes harder. In the classical theory of editing, the passing mention is considered the basic criterion of entry selection.[17] In addition, a search based exclusively on word frequency is rarely able to identify concepts hidden in the text. In this context, the superiority of an index,[18] called for the purpose of this paper, syndetic, becomes clear. It functions as an intentionally designed tool for knowledge organization and navigation. The term *syndetic* (from ancient Greek *syndetikos* – cohesive) designates a structure beyond an ordinary, linear register of words. It is a system, within which selected concepts are intentionally connected via a net of semantic relationships, revealed through entries and subentries (ranking relations) and normalized full and partial references representing accordingly relations of equivalence (synonymy) and association.[19]

The decisive superiority of the syndetic index over full-text algorithms is due to the fact that the process of its creation has an analytical character, and not exclusively mechanical. A professional indexer[20] conducts an interpretative semantic analysis of text, which requires a cognitive effort to define phenomena and find hidden relations between individual concepts.[21] A syndetic structure results in links between entries created by a thick network of full and supplemental references and, in addition, explicitly indicates connections between selected terms and contents, which

17 N.C. Mulvany, *Indexing Books*, 2 ed., The University of Chicago Press, Chicago, London 2005.

18 For the needs of this paper, we define index as an orderly set of terms representing the contents of the document which indicate where specific issues occur and which mirrors their conceptual structure. It is a tool for analysis and representation, and even interpretation of the contents, and not only a mechanical list of words occurring in the text. See Ibidem; H.H. Wellisch, *Indexing from A to Z*, H. W. Wilson, New York 1995; D.M. Stauber, *Facing the Text: Content and Structure in Book Indexing*, Cedar Row Press, Eugene 2004.

19 F.W. Lancaster, *Indexing and abstracting in theory and practice*, 3 ed., Facet Pub., London 2003.

20 This term 'Indexer' is ambiguous, because it can refer to the person compiling an index or to a computer program used to create indexes. Here, it means the person, a specialist who compiles indexes.

21 J. Pacek, 'Indeksowanie w XXI wieku...', op. cit., pp. 32–49.

in turn allows grouping of scattered information and eliminates the risk of repeating synonymic records under many different entries.[22] These operations make the index relatively autonomous, offering users an orderly map of logical connections that a full-text search alone cannot capture.

1. 2. FORMULATION OF A RESEARCH PROBLEM

Rapid development of Large Language Models (LLMs) has marked a new stage in the evolution of automatic natural language processing systems (Neuro-Linguistic Programming, NLP). Tools based on generative artificial intelligence (AI) show unprecedented efficiency in automatic data extraction, including advanced named entity recognition, as well as rapid semantic categorization of large volumes of text. This progress has fueled technological discourse about the possibility of fully automating analysis and knowledge organization processes, often bypassing fundamental architectural limitations of these models, including their probabilistic rather than cognitive nature.[23]

In the organization of knowledge studies, we must differentiate between a mechanical, even very advanced, extraction of concepts and the intellectual process of constructing a correct conceptual representation. While LLMs can handle frequency analysis, their ability to generate coherent, purposeful, and hallucination-free networks of syndetic relationships remains an area requiring in-depth verification. This is why we have to ask a basic empirical question: is generative artificial intelligence currently able to independently generate indexes of a quality and usefulness comparable to what qualified indexers can do? Although this issue has been discussed in the most recent professional studies, and their results indicate the necessity of efficiency assessment of artificial

22 PN-ISO 999: *Informacja i dokumentacja – Wytyczne dotyczące zawartości, struktury i prezentacji indeksów*.

23 E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York 2021, pp. 610–623.

intelligence used in complex, multi-stage tasks performed by indexers.[24] Doubts exist as to the statistical models' ability to correctly recognize hidden concepts, usage of coherent terminology within indexes, including equalization of synonyms and creation of non-trivial associative links which determine the value of an index.[25] As until now no studies were conducted on sources in Polish, finding an answer to this question became urgent, because possibilities offered by LLMs evolve in an unusually dynamic manner. This research challenge requires a comparative analysis of the results obtained in the process of applying two different methods of index creation: traditional indexing performed by experts and fully automatized entry generation by artificial intelligence. For the purpose of this paper, we formulated an initial hypothesis that while contemporary language models constitute a useful tool for term extraction, architectural limitations result in a lower quality than what is expected from professional indexing, which imposes the use of the *human-in-the-loop* (HITL) approach, i.e., an AI model that requires substantive control by an indexer.

2. Differences between Human Cognition and LLM Text Processing

Indexes are compiled as a result of a process, which cannot be considered purely mechanical, as was indicated above: "A well prepared index is not only a 'dry' alphabetical list; its creation requires from the author a significant intellectual effort, naming certain contents, phenomena, as well as ability to find out relationships between concepts, and often, also to state such relationships."[26] This process is called indexing.

In the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organisation, and Presentation of Indexes*, we find detailed

24 T. Izzard, 'Generative artificial intelligence (AI) and its performance at indexing tasks', *The Indexer* 2024, 42(4), pp. 383–400.

25 E. Bartmess, M.R. Combs, 'LLM-generated book indexes: Can they replace professionally created indexes?', *The Indexer* 2025, 43(4), pp. 327–348.

26 J. Pacek, 'Indeksowanie w XXI wieku…', op. cit., p. 38.

descriptions of the successive stages in the indexer's work. The proper performance of tasks includes:

a) identifying and localizing data significant within the material being indexed;
b) distinguishing between data related to an issue and a passing mention about a theme;
c) excluding short mentions of an issue if they do not include anything of significance for potential users;
d) analysing concepts discussed in the document with the objective of creating strings of entries;
e) ensuring a proper choice of terms used in the index from the point of view of users' needs, i.e., allowing them to rapidly:
   1) determine whether information about a given issue is present or absent in an unknown text;
   2) find information about a remembered position in a known or partially known text;
   3) identify the right documents in a collection.
d) indicating relations between concepts;
e) grouping together information related to an issue scattered because of the document's or the set of documents layout;
f) combining entries and determiners into positions [...];
g) directing users via cross-references from terms not selected as index entries to terms selected as entries; j) collating positions in a systematic and useful order.[27]

Indexer's creativity, efficiency, and analytic abilities determine whether the index will be useful and correspond to the real needs of the publication's readers. Before the indexers begin their task, they determine what type of index is required for a specific document and what is its purpose. They also decide about the language of the index – if the publication includes texts in various languages, the index must be translated. Next, based on an analysis of the document, they extract one- or multi-word index entries, i.e.,

27 PN-ISO 999: *Informacja i dokumentacja – Wytyczne dotyczące zawartości, struktury i prezentacji indeksów*, pp. 7–8.

identify and localize information relevant for the users and leave out irrelevant data. Entries are extracted from the main text, but occasionally from other elements of the contents, except the title page, table of contents, dedication, summaries, and abstracts. The indexer must also perform tasks such as changing the grammatical form of certain entries, standardizing and, at times, completing index entries. For instance, guidelines recommend using transliteration or transcription and names in their original or Polish version. In indexes, it is particularly important to avoid synonymous notations, and at the same time to link entries connected by a net of references – complementary (to related information) and cross-references (to relevant information).[28]

After completing these activities, the indexer constructs index entries in alphabetical, nested alphabetical, substantive (logical), or other arrangements, then adds numerical indicators, such as page numbers, paragraphs, chapters, and bibliographical positions, that allow them to be located in publications. At the editing stage, index entries are often printed in a smaller font than the main text and in a multi-column layout, which saves space in printed volumes. The basic difference between human cognition and text processing by LLMs comes down to different paradigms of data analysis. While a professional indexer performs an interpretative, context-aware semantic assessment of the document, transformer-based models use probabilistic mechanisms to predict successive tokens and word vector spaces. In automated systems, such as those using Retrieval-Augmented Generation (RAG), data search is most often based on computing the distance between vectors in multidimensional space, using measures such as cosine similarity. [29]

This approach, while effective from the computing point of view, reduces natural language to statistical structures. In consequence,

28 PN-ISO 999: *Informacja i dokumentacja – Wytyczne dotyczące zawartości, struktury i prezentacji indeksów*, p. 7.

29 X. Dong, W. Zhu, H. Wang, X. Chen, P. Qiu, R. Yin, Y. Su, Y. Wang, *Talk Before You Retrieve: Agent-Led Discussions for Better RAG in Medical QA*, https://arxiv.org/pdf/2504.21252 [Accessed March 16, 2026].

classical algorithms perform very well in extracting lexical units based on frequency analysis (co-occurrence), but as a rule they have significant limitations in identifying hidden relations. In a typical process of creation of an autonomous knowledge organization system, the indexer deliberately ignores statistical noise, focusing on building structured links (hierarchical, synonymic, associative), that often are not a direct consequence of the text continuity, but rather from a broader knowledge of the discipline and from predicting the user's needs.[30]

In addition, applying standard language models based exclusively on token prediction to independent generation of complex syndetic networks is impeded by architectural barriers. Research on using LLMs in ontology-building processes indicates that while algorithms can generate large lists of potential entries, they rely on identifying statistical correlations rather than a thorough understanding of the analyzed domain.[31] It often leads to the phenomenon of hallucination, i.e., generating seemingly correct links from the syntactic perspective, but erroneous from the point of view of a rigorous and precise controlled vocabulary. These restrictions, compounded by context window limits, make language models unable to independently maintain the structural coherence of indexes, requiring intervention by an indexer who controls and corrects what digital tools produce.

The limited capacity of a context window and the probabilistic nature of models become particularly problematic when processing large documents, such as scholarly monographs. In such situations, LLM algorithms tend to generalize, a tendency called probabilistic inertia. This phenomenon marginalizes or overrides author-specific terminology in favor of higher-frequency concepts in the model's general training corpus. In addition, efficiency assessments of LLMs

30 B. Hjørland, 'Semantics and Knowledge Organization', *Annual Review of Information Science and Technology* 2007, vol. 41, iss. 1, pp. 367–405.

31 B. Lewandowska-Tomaszczyk, G. Pawłowski, 'Testing ChatGPT on terminology generation, definitions: translation, and ontology creation in German, English and Polish', *Research in Language* 2025, 23(1), pp. 320–340.

on long texts reveal a significant decrease in attention mechanisms in the middle of the analyzed document. This deficit, known in the scholarly literature as the lost-in-the-middle effect, results in serious informational omissions and uneven coverage of the indexed material.[32] The literature has identified these shortcomings as a barrier to fully automated approaches to creating exhaustive and precise book indexes. However, this situation is changing due to the most recent evolution of AI architecture towards agentic workflows, which try to compensate the inherent weaknesses of probabilistic models through implementation of internal adaptive thinking and the ability to use external computing tools.

3. Research Methodology

In contemporary scholarship – even considering only Polish publications – issues related to defining indexes, their types, and functions, are being researched in such disciplines as literary studies,[33] linguistics,[34] editing,[35] bibliology,[36] library science,[37] information

32 N.F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, 'Lost in the middle: How language models use long contexts', *Transactions of the Association for Computational Linguistics* 2024, no. 12, pp. 157–173.

33 tk, 'Indeks', in: *Słownik terminów literackich*, ed. 2, ed. J. Sławiński, Zakład Narodowy im. Ossolińskich – Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk, Łódź 1988, p. 196; M. Krassowski, *Leksykon terminów literackich*, ed. 2, Twój Styl, Warszawa 1996, p. 95; M. Król, G. Krupiński, H. Sułek, *Nowy słownik terminów literackich*, Krakowskie Wydawnictwo Naukowe, Kraków 2009, p. 109.

34 ZS [Z. Saloni], 'Indeks', in: *Encyklopedia językoznawstwa ogólnego*, ed. 2, ed. K. Polański, Zakład Narodowy im. Ossolińskich, Wrocław, Warszawa, Kraków © 1999, p. 252.

35 *Zasady typowego opracowania edytorskiego*, ed. L. Marszałek, Panstwowe Wydawnictwo Naukowe, Warszawa 1962, p. 76; 'Indeks', in: *Słownik wydawcy*, ed. Barbara Kalisz, Wydawnictwo Naukowo-Techniczne, Warszawa 1997, p. 77; B. Osuchowska, *Poradnik autora, tłumacza i redaktora*, Akademicka Oficyna „Exit", Warszawa 2011, p. 187.

36 'Indeks', in: *Encyklopedia wiedzy o książce*, ed. A. Birkenmajer, B. Kocowski, J. Trzynadlowski, Zakład Narodowy im. Ossolińskich, Wrocław, Warszawa, Kraków 1971, pp. 991–992.

37 O.U. [O. Ungurian], 'Indeks', in: *Encyklopedia współczesnego bibliotekarstwa polskiego*, ed. K. Głombiowski, B. Świderski, H. Więckowska, Zakład Narodowy im. Ossolińskich – Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk 1976, pp. 134–135; 'Indeks', in: *Podręczny słownik bibliotekarza*, ed. G. Czapnik, Z. Gruszka, H. Tadeusiewicz, Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, Warszawa 2011, p. 117.

science,[38] and within an already established tradition, as can be seen in the main information guides of these disciplines. There are also studies, which present and assess use of various LLM tools in generating indexes for books and other scholarly materials compared to indexes compiled by indexers. For instance, research conducted by Elizabeth Bartmess and Michele R. Combs in 2025 and 2026,[39] Tanya Izzard in 2025[40] as well as Bruno Cardoso Greco, Paula Regina Dal'Evedove and Albino Januário Nhacuongue – from the same period.[41] Because until now, no verification of LLM efficiency on Polish language sources was ever performed, it became the goal of the present study to compare the quality of indexes generated by three popular LLMs – ChatGPT, Gemini and Claude, to an expert-compiled index of a scholarly text. The research material is a fragment of a multi-author scholarly monograph *Bibliotekarstwo*, edited by Anna Tokarska (Warsaw 2013), provided with an expert-compiled index. *Moduł IV. Organizowanie dostępu do informacji*, 73 pages long and containing over 29 000 words was selected for analysis. This volume fits the context window of contemporary LLMs, which allows to limit mistakes resulting from memory overload and, at the same time, to reveal the models' real architectural limitations. The multi-author character of the selected text is an additional factor which increases the degree of difficulty.

38 'Indeks', in: *Słownik terminologiczny informacji naukowej*, ed. M. Dembowska, Zakład Narodowy im. Ossolińskich – Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk 1979, p. 51; 'Indeks', in: *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, ed. B. Bojar, Wydawnictwo SBP, Warszawa 2002, p. 83.

39 E. Bartmess, M.R. Combs, 'LLM-generated book indexes: Can they replace professionally created indexes?', *The Indexer* 2025, 43(4), pp. 327–348; E. Bartmess, M.R. Combs, 'Can the current generation of large language models (LLMs) produce an adequate book index?', *The Indexer* 2026, 44(1), pp. 35–48, https://doi.10.3828/index.2026.4 [Accessed June 08, 2026].

40 T. Izzard, 'Book indexing and generative AI', *Catalogue & Index* 2025, iss. 211, pp. 3–8, https://journals.cilip.org.uk/catalogue-and-index/article/view/746/680 [Accessed June 08, 2026].

41 Greco Bruno Cardoso, Dal'Evedove Paula Regina, Nhacuongue Albino Januário, 'O potencial dos LLM (Large Language Models) na indexação automática', *ISKO Brasil* 2025, 8, https://isko.org.br/ojs/index.php/iskobrasil/article/view/97 [Accessed June 08, 2026].

Work with this sort of material requires standardization of terms representing the same concepts used by different authors.[42] Reference standard (*ground truth*) is provided by a sub-index, generated from the original index of the book and composed exclusively from those entries and sub-entries that were created by the indexer and refer to the analysed fragment of the publication.

Paid models: ChatGPT 5.3, Gemini 3.1 Pro, and Claude Opus 4.6 Extended thinking were used in the research. Each of the systems was tested twice. In the first approach, *zero-shot*, models were ordered to generate a full index based on a single prompt and material in the form of the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*, without providing examples of a correctly constructed index, i.e., without demonstrating the expected output structure:

Zero-shot

Assume the role of a professional indexer of scholarly books. Your task is to compile a full index for the attached material (a tekst.pdf file). Base your semantic analysis and the construction of entries, references, and locators exclusively on the principles provided in the norm PN-ISO 999 (norma.pdf file). Present the completed index as an alphabetical, structured list of index entries along with corresponding locators.

The second test was conducted according to the structured approach, *chain-of-thought prompting*, which divides the index generation process in successive stages: extraction of terms, their semantic grouping, structuring references and attribution of the final locators.

Prompt *chain-of-thought prompting*

Assume the role of a professional indexer of scholarly books. Your task is to compile a full index for the attached material (a tekst.pdf file). Base your semantic analysis and the construction of entries, references, and locators exclusively on the principles provided in the norm PN-ISO 999 (norma.pdf file).

42 E. Bartmess, M.R. Combs, 'LLM-generated book indexes...', pp. 327–348.

Perform the algorithmic task, step by step, presenting the results of each of the stages below, and at the end, generate the final result:

Stage 1. Extraction and selection of potential terms. Analyze the text and select concepts eligible for indexing strictly relying on criteria for entry selection described in the attached norm PN-ISO 999. Write out a working list of selected concepts.

Stage 2. Normalization and semantic reconciliation. Analyse the list from Stage 1. Apply directives from the norm PN-ISO 999 to link synonyms and terminological variants. Select the preferred forms of the main entries and briefly justify the most important decisions.

Stage 3. Building a syndetic structure. Determine relations between normalized concepts. Construct a hierarchy including entries and sub-entries and design a network of references (of the type: see and of the type: see also), rigorously observing structuring principles defined in the norm PN-ISO 999. Present a working draft of this structure.

Stage 4. Attribution of locators. Attribute page numbers to each element in the structure, making sure that criteria for indicator attribution fully conform to the attached norm PN-ISO 999.

Final result: after implementing the four analytical stages discussed above, at the end of your answer, present the final completed index in the form of a structured, alphabetical list of index entries along with locators and references.

Both research variants maintained the same task content and an identical input material, including the analyzed text fragment and the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*. The difference between the variants consisted exclusively in the instructions: in the *zero-shot* approach, the task was formulated as a single command, but in chain-of-thought prompting, the indexing process was broken into stages.

The two-variant procedure serves a diagnostic purpose. The *zero-shot* approach allows to estimate to what degree the models are independently able to reconstruct the indexing procedurę based on the general instruction and source material. On the other hand, the approach based on the sequence of tasks or *chain-of-thought*

prompting, allows to verify whether the breakdown of the indexing process into operations improves the quality of the generated index. A comparison of the two variants' results will then reveal whether the limitations observed in automated indexing are due to competence (connected to representation of knowledge by the model) or procedure (result from the way of formulating tasks).

The generated indexes were then evaluated according to only four key parameters used in studies of data search systems:

1. completeness (*recall*) – the degree of inclusion of concepts figuring in the expert-compiled index;
2. precision – the percentage of entries displaying a real navigational value allowing to estimate the models' tendency to index passing mentions;
3. locator accuracy – verification of correct attribution of page numbers to index entries and identification of locator hallucination cases;
4. syndetic consistency – evaluation of the correct index construction, including ranking relations (entries and sub-entries) and references.

Applying these four parameters leads to a multi-aspect evaluation of indexes generated by language models, including an adequate representation of contents, as well as correctness of construction and precision of localization.

4. The Results of the Experiment

4.1. *ZERO-SHOT* APPROACH

An assessment of LLMs selected for study (at this stage, we will present results obtained using two of the three tools, namely, ChatGPT 5.3 and Gemini 3.1 Pro, the last one – Claude Opus 4.6 Extended thinking – requires a separate description, which is provided later in the paper) in the *zero-shot* variant confirms structural limitations of the architecture based exclusively on token prediction. Comparative analysis of the generated entries and the benchmark adopted for the needs of the study reveals the basic difference between a cog-

nitive understanding of the text and its statistical processing. Table 1 shows a summary comparison of the four elements of assessment.

TABLE 1. QUANTITATIVE DATA BREAKDOWN FOR ZERO-SHOT INDEXES IN COMPARISON WITH EXPERT-COMPILED INDEX

Indicator	Expert-compiled index	ChatGPT 5.3	Gemini 3.1 Pro
1. Completeness			
Number of the main entries	77	62	55
Expert benchmark coverage	—	9 (11,7%)	23 (29,9%)
2. Precision			
Entries absent in expert benchmark	—	53	32
3. Locator accuracy			
Page range hallucination	—	occur	sporadically
4. Syndetic consistency			
References in total	42	3	5
Including "see also"	42	0	0
Including "see" (references)	—	3	5
Sub-entries	yes (expanded)	sporadic	isolated

Source: provided by the authors.

Completeness

Expert benchmark coverage is very low in both models. ChatGPT 5.3 identified correctly only 9 of 77 expert-compiled entries (11,7%), and Gemini 3.1 Pro – 23 of 77 (29,9%). Gaps in concepts requiring abstract thinking are particularly evident. The term "e-exclusion" and "information literacy" were precisely identified by the expert (p. 428–429). ChatGPT 5.3 omitted both issues, confirming its inability to identify concepts expressed implicitly in the text. On the other hand, Gemini 3.1 Pro captured them under different labels ("digital exclusion" and "informational education"), showing some capacity for semantic abstraction but also raising doubts about terminological standardization. The lack of a developed search engine taxonomy in both machine-generated indexes is most significant. Expert-compiled index includes 9 entries related to different search engine types (global, local, national, semantic, universal, specialized, WWW pages, scholarly, Internet), which create a coherent concept tree. None of the models extracted these terms, which suggests that the models are unable to distinguish the taxonomy implicitly present in the text.

Precision

While a professional indexer is able to distinguish a key concept from a term used only for historical or anecdotal background, AI models index them as terms of equal weight. The phenomenon of excessive extraction of passing mentions occurs in both models but in a different way. ChatGPT 5.3 includes hardware and historical terminology in the index. Such entries, as "Hollerith, Herman", "Moore's law", "Blu-ray Disc", "CD", "IBM", "gopher" or "WAIS", played in the text of the monograph only an auxiliary function. Gemini 3.1 Pro on the other hand, generates social and mobile space noise because entries, such as: "Facebook", "Wikipedia", "smartphones", "tablets" or "QR codes", also do not merit separate index entries in the context of the analyzed monograph. Indexer, intent on providing the index with navigational value, was right to leave out these passing mentions.

It should also be said, that the distinct nature of the historical hardware noise generated by ChatGPT 5.3, compared to the media-mobile noise found in Gemini 3.1 Pro, may reflect differences in training data composition or preferences due to both models' architecture, known in the literature as inductive biases.[43] At the same time, Gemini 3.1 Pro better recognized specific library systems (ALEPH, LIBRA, MAK, MATEUSZ, MOL, PATRON, PROLIB, SOWA), which constitute a relevant component of expert-compiled indexes, resulting in a smaller number of missed entries (32 of 55) in comparison with ChatGPT 5.3 (53 of 62).

Locator accuracy

Lack of understanding of the structure causes page range hallucination, connected to contribution of large, several-page-long ranges to very narrow concepts. For instance, ChatGPT 5.3 attrib

43 LLM's inductive bias lies in a systematic tendency to represent, process or generate data in a specific manner, resulting from its architecture, learning mechanisms and computing construction, not directly from the contents of training data. C.M. Rytting, D. Wingate, *Leveraging the Inductive Bias of Large Language Models for Abstract Textual Reasoning*, 2021, https://arxiv.org/abs/2110.02370 [Accessed June 08, 2026].

utes to processors (CPU) and to memory RAM the 407–412 page range, while these data are located precisely on specific pages. Gemini 3.1 Pro shows a smaller tendency to this type of phenomenon, although it also makes occasional mistakes.

Syndetic consistency

The greatest shortcoming of both models is the almost non-existent syndetic apparatus. The expert-compiled index includes 42 references, all of the kind "see also", responsible for a dense network of semantic relations between concepts (e.g., "Aggregator → see also Internet search engine"; "ALEPH → see also Integrated library system"). Meanwhile, ChatGPT 5.3 generated 3 references, and Gemini 3.1 Pro – 5, all exclusively of the "see" type (simple equivalence references, e.g., "e-books *see* electronic books"), and none of the models generated a "see also" reference revealing a semantic relationship between related concepts. This is a qualitative, not only quantitative difference, because references "see also", discussed in the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*, attest to an index' navigational usefulness and to the understanding of the field's semantic structure. Flattening the hierarchical structure is an equally important problem. While an indexer constructs a logical concept tree (e.g., main entry "Internet search engine" with subentries: document assessment, typology; or "Subject headings vocabulary" with subentries: KABA, MeSH), ChatGPT 5.3 generated a flattened structure with separate independent entries for "CPU" and "RAM", instead of placing them logically below the superior entry. Gemini 3.1 Pro also designs index entries without grouping them below the main entry, even though it managed better ranking for "library systems."

Recapitulation

Research results explicitly demonstrate that LLMs, in their zero-shot approach, cannot meet the requirements presented in the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*. They generate expanded but dysfunctional lists of terms that lack semantic depth

and hierarchical organization of knowledge. However, Gemini 3.1 Pro shows higher indicators than ChatGPT 5.3 in all four measures of assessment: higher coverage of expert benchmark (29,9% compared to 11,7%), a smaller number of missed entries (32 compared to 53), a less pronounced tendency to locator hallucinations, as well as a higher number of references (5 compared to 3). Nevertheless, even the better model covers less than one-third of expert-compiled entries and does not generate a single "see also" reference, which confirms the need for more advanced indexing methods described later in the paper.

4.2. *CHAIN-OF-THOUGHT PROMPTING* APPROACH

Applying the *chain-of-thought prompting* technique, which forces the model to perform multi-stage reasoning before generating an answer, produces a different effect from the *zero-shot* variant. Comparative analysis reveals a paradoxical dynamic: both models generate significantly fewer entries, but of a higher structural quality. A summary tally of this research procedure is provided in Table 2.

TABLE 2. QUANTITATIVE DATA BREAKDOWN FOR CHAIN-OF-THOUGHT PROMPTING INDEXES IN COMPARISON WITH EXPERT-COMPILED INDEX

Indicator	Expert-compiled index	ChatGPT 5.3	Gemini 3.1 Pro
1. Completeness			
Number of the main entries	77	28 (in *zero-shot: 62*)	21 (in *zero-shot: 55*)
Expert benchmark coverage	—	5 (6,5%) (in *zero-shot: 9; 11,7%*)	11 (14,3%) (in *zero-shot: 23; 29,9%*)
2. Precision			
Entries absent in expert benchmark	—	23/28 (in *zero-shot: 53/62*)	10/21 (in *zero-shot: 32/55*)
3. Locator accuracy			
Page range hallucinations	—	smaller (occur in *zero-shot*)	sporadically (in *zero-shot: sporadically*)
4. Syndetic consistency			
References "see"	—	0 (*in zero-shot: 3*)	5 (in *zero-shot: 5*)
References "see also"	42	0 (*in zero-shot: 0*)	2 (*in zero-shot: 0*)
Sub-entries	expanded	14 (*in zero-shot: sporadic*)	2 (*in zero-shot: isolated*)

Source: compiled by the authors.

Completeness

The most evident effect of applying *chain-of-thought prompting* technique is the reduction of the number of generated entries. In ChatGPT 5.3, 62 entries were produced in *zero-shot* approach, and 28 in the second application, i.e., 34 index entries less. An identical phenomenon occurred in Gemini 3.1 Pro – a decrease from 55 to 21 entries. This reduction, however, is not a quality selection – next to the elimination of informational noise, the models also eliminated a part of entries correctly identified in the earlier variant. Expert benchmark coverage fell from 11,7% to 6,5% for ChatGPT 5.3 (from 9 to 5 entries) and from 29,9% to 14,3% for Gemini 3.1 Pro (from 23 to 11 entries).

It is worth emphasizing that Gemini 3.1 Pro, using the *chain-of-thought prompting* approach, lost half of the entries that were correctly identified earlier, including "NUKAT", "PATRON", "Google", and "new media". On the other hand, from the ChatGPT 5.3 index, entries such as "databases", "electronic libraries", and "data search" disappeared. Taxonomy of data search engines – 9 types in the expert-generated index – was still invisible.

Precision

While the *chain-of-thought prompting* technique did not improve expert benchmark coverage, it reduced information noise. The number of entries missed in the expert-generated index decreased from 53 to 23 for ChatGPT 5.3 and from 32 to 10 for Gemini 3.1 Pro. ChatGPT 5.3 in the *chain-of-thought prompting* mode also eliminated most of the earlier hardware-historical noise (among others, disappeared "Blu-ray Disc", "Hollerith", "Moore's law", "IBM", "WAIS"), although it introduced new entries ("Pinakes", "DNS", "FTP", "telnet"). Gemini 3.1 Pro, on the other hand, eliminated media-mobile noise ("Facebook", "smartphones", "tablets", "Wikipedia"), leaving only 10 entries from the benchmark.

Noise reduction confirms that sequential reasoning allows the models to select a certain level of passing mentions. It is not, however a targeted selection, because models eliminate erroneous as well as correct entries, which suggests a general narrowing of at-

tention rather than a real evaluation of index concepts' navigational values.

Locator accuracy

In locator accuracy, the *chain-of-thought prompting* mode shows a slight improvement. ChatGPT 5.3 still generates expanded page ranges, although on a smaller scale than in the *zero-shot* mode. Gemini 3.1 Pro, on the other hand, maintains a level similar to the earlier variant, with minor page range hallucinations. This phenomenon was not eliminated using the multi-stage reasoning technique.

Syndetic consistency

It is precisely in the dimension of syndetic consistency that the technique of *chain-of-thought prompting* brings the most interesting and diverse effects. Both models demonstrate the possibility of shaping index structure, but each does it in a different manner. ChatGPT 5.3 went from very flattened structure to developed hierarchical trees with sub-entries. In the *chain-of-thought prompting* mode, it generated 14 sub-entries grouped under the main entries, such as: "computer" (building, CPU, RAM, mass memory), "libraries" (automation, computerization, information systems), "information retrieval systems" (indexing information, coding information), "library catalogue" (central, OPAC). It is a qualitative change compared to the *zero-shot* mode, where the same concepts figured as different and independent entries. At the same time, the discussed tool did not generate in the *chain-of-thought prompting* mode references "see" and "see also" (in the *zero-shot* mode, there were 3 references "see").

Gemini 3.1 Pro adopted a different strategy. Instead of creating hierarchical trees (it generated only 2 sub-entries: "electronic documents → with local access / with remote access"), it focused on constructing a syndetic apparatus. For the first time in the entire experiment, references of the "see also" type appeared: "library catalogues → see also OPAC" and "library systems → see also integrated library systems." There are only 2 such references, in contrast to 42 in the expert-generated index, but their very presence shows the

ability to understand semantic relations between related concepts, an ability absent in the remaining variants.

It is important to underline the difference between the two models' strategies. ChatGPT 5.3 in the *chain-of-thought prompting* mode prefers internal ranking (relation entry – sub-entry), and Gemini 3.1 Pro prefers an external network (references between entries). Both strategies are components of a full syndetic apparatus required by the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*, but neither model connects them in a coherent whole.

Recapitulation

The *chain-of-thought prompting* technique produces a paradoxical effect in both models: it enforces deeper processing, which results in the appearance of structural elements (sub-entries in ChatGPT 5.3, references "see also" in Gemini 3.1 Pro), but at the same time limits the scope of analysis. Models process more thoroughly but generate fewer index entries. Expert benchmark coverage decreases to, respectively, 6,5% and 14,3%, i.e., values that make such indexes unhelpful from the user's point of view. These results confirm that the *zero-shot* and *chain-of-thought prompting* approaches do not meet the requirements of the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*, which provides the reason for applying the agency architectures described later in the paper.

4.3. CHANGE OF PERSPECTIVE

The indexes generated by Claude Opus 4.6 Extended thinking (Anthropic) represent a qualitative departure from the results presented thus far; for that reason, they should be described separately. However, before presenting quantitative data, we should note a phenomenon visible exclusively in the model's reasoning trace. Claude Opus 4.6 Extended thinking in the mode *zero-shot*, without any methodological instruction, adopted a quasi-agency strategy: it independently read and interpreted the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and*

Presentation of Indexes, then analyzed the source text page after page, planned the index structure, and only then began its compilation. None of the remaining models behaved this way; ChatGPT 5.3 and Gemini 3.1 Pro generated indexes without an introductory phase. This difference is reflected in the quantitative results. A full tally of all LLM variants in the context of the four evaluation criteria is presented in Table 3.

TABLE 3. QUANTITATIVE DATA BREAKDOWN FOR ALL VARIANTS IN COMPARISON WITH EXPERT-COMPILED INDEX

Indicator	Expert-compiled index	ChatGPT 5.3 *zero-shot*	Gemini 3.1 Pro *zero-shot*	Claude Opus 4.6 Extended thinking *zero-shot*	Claude Opus 4.6 Extended thinking *chain-of-thought prompting*
1. Completeness					
Main entries	77	62	55	194	173
Sub-entries	expanded	sporadic	isolated	117	87
2. Precision					
Entries absent in the expert benchmark index	—	53/62	32/55	see Tab. 4	see Tab. 4
3. Locator accuracy					
Page range hallucination	—	occur	sporadically	—	—
4. Syndetic consistency					
References "see"	—	3	5	5	11
References "see also"	42	0	0	16	20
5. Methodological elements					
Two-column layout	no	no	no	yes	yes

Source: compiled by the authors.

Completeness

Claude Opus 4.6 Extended thinking in the *zero-shot* mode generated 194 main entries and 117 sub-entries, in the *chain-of-thought prompting* mode – 173 main entries and 87 sub-entries. For comparison: the expert-compiled index includes 77 main entries, ChatGPT 5.3 generated 62, and Gemini 3.1 Pro – 55. Claude Opus 4.6 Extended thinking not only covers a decisive majority of expert-generated entries, but also creates elaborate hierarchical trees with subentries not found in the remaining models in the *zero-shot* mode.

Significantly, Claude Opus 4.6 Extended thinking, as the only program in zero-shot mode, identified the full taxonomy of search engines (global, local, national, semantic, universal, specialized, scholarly, WWW pages), which did not occur in ChatGPT 5.3 or Gemini 3.1 Pro in either variant.

Syndetic consistency

Claude Opus 4.6 Extended thinking generated 16 "see also" references in zero-shot mode and 20 in chain-of-thought prompting mode, corresponding to 38% and 48% of the syndetic apparatus of the expert benchmark (42 references), respectively. For comparison, ChatGPT 5.3 and Gemini 3.1 Pro in the *zero-shot* mode did not generate a single "see also" reference, and Gemini 3.1 Pro in the chain-of-thought prompting mode generated only 2. The difference between Claude Opus 4.6 Extended thinking and the other models appears not only quantitative but, most of all, qualitative – as it seems to grasp the fact that an index requires a network of semantic relations between concepts.

Methodological elements

The presence of elements going beyond a list of index entries constitutes a different phenomenon. Claude Opus 4.6 Extended thinking in both research variants added an introductory note explaining conventions used in the index (page range, priority rules, abbreviations of references) and applied a double-column layout with indents for determiners - according to the recommendations of the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes* relative to the typographic layout of printed indexes. The other models did not generate any of these elements.

Expert assessment

Experts further evaluated the index generated by Claude Opus 4.6 Extended thinking in the zero-shot variant using color-coding (green for entries common with the expert benchmark; blue for entries added by a tool recognized as justified; red for entries present in the expert benchmark but omitted by the model). The authors formulated 35 comments categorized in Table 4.

TABLE 4. CLASSIFICATION OF EXPERT COMMENTS TO CLAUDE OPUS 4.6 EXTENDED THINKING (ZERO-SHOT) INDEX

Comment category	Quantity	Examples
Excessively detailed and indexing of passing mentions	18	"are not discussed, only mentioned"; "leave out all names of social media, because they are given only as examples in the chapter"; "a too detailed entry, as computer software is available"
Selection and formulation of entries	8	"in the original: FBC" [not "Federacja Bibliotek Cyfrowych"]; „a better entry: convergent book"; "this entry does not reflect the users' real informational needs"
Lack of formal consistency	3	"shows inconsistency in the usage of singular and plural forms in the index entries"; "in some entries, acronyms are provided, in some other entries, acronyms are expanded"
Duplicates and repetitions	3	"double entry, a reference is sufficient"; "repeated entry"
Omissions (missing entries)	3	"no entry for Biblioteka Aleksandryjska 406"; "no entry for karty przezierne 406"

Source: compiled by the authors.

Most of the comments describe the excessive detail generated by the tools (18 of 35 comments). Claude Opus 4.6 Extended thinking indexes terms, which play only an illustrative role in the text of the monograph – the authors point out that the names of social media services (Facebook, Flickr, YouTube) "are quoted only as examples of certain library services." Some Internet tools (gopher, BBS) "are only mentioned in context" and are not discussed. This is the kind of phenomenon identified in ChatGPT 5.3 and Gemini 3.1 Pro, with the difference that in Claude Opus 4.6 Extended thinking it appears next to correct entries rather than replacing them.

The second-largest category (8 comments) concerns the choice and formulation of entries. The authors point out cases in which Claude Opus 4.6 Extended thinking used a full name instead of an acronym ("in the original: FBC", not "Federacja Bibliotek Cyfrowych [Federation of Digital Libraries]") or proposed an entry less precise than possible ("a better entry: convergent book"). These comments show that the model understands the topic but lacks expert intuition about terminological conventions accepted in the academic community. Lack of formal consistency (3 comments) occurs in mistaken use of plural and singular in the entries, as well as in inconsistent treatment of acronyms. Some entries consist of an acronym only, others of a full name, and others of the full name with the acronym in brackets. Doublets (3 comments)

are cases where the same concept occurs twice in different forms. Both types of mistakes indicate a lack of consistency control that indexers naturally perform using working memory across the entire index.

The authors of the papers consider the missing index entries particularly interesting (3 comments). While having generated 194 main entries, Claude Opus 4.6 Extended thinking omitted, among others, "Biblioteka Aleksandryjska [Library of Alexandria]" (p. 406) and "transparent cards" (p. 406). The authors note a paradox: the model introduced "Kallimach z Cyreny [Callimachus of Cyrene]" ("a too detailed entry"), but omitted "Biblioteka Aleksandryjska," the main and more recognizable concept.

A Case of Recognized Superiority of the Model

In one case, the authors recognized that a solution proposed by Claude Opus 4.6 Extended thinking was better. The comment applies to the structure of the catalogue entries. While the expert-compiled index ranked "Katalog" as the main entry and the individual catalogue types as sub-entries, the model treated each catalogue type (scholarly, WWW pages, central) as a main entry. The authors stated that: "after reviewing the chapter, Claude's approach seems more suitable – establishing a typology of catalogues was not the goal of the study, it was rather to determine which tools should be used to search for information." This case shows that a model can reflect the source text's intent more accurately than an expert approach based on established conventions.

Claude Opus 4.6 Extended thinking: *zero-shot versus chain-of-thought*

In contrast with ChatGPT 5.3 and Gemini 3.1 Pro, where the *chain-of-thought prompting* technique caused a significant decrease of the number of entries (34 fewer entries in both models), in Claude Opus 4.6 Extended thinking the decrease is minor: 11% fewer main entries (from 194 to 173) and almost 26% fewer sub-entries (from 117 to 87). At the same time, the quality of the syndetic apparatus improved: the number of "see also" references increased from 16 to 20, and "see" references from 5 to 11. The *chain-of-thought prompting*

technique functions as theoretically intended: it improves understanding without excessively limiting the scope of analysis.

Recapitulation

The results from Claude Opus 4.6 Extended thinking confirm the shift in perspective associated with automated indexing. While ChatGPT 5.3 and Gemini 3.1 Pro do not fully reflect the meaning – they do not understand sufficiently to be able to identify key concepts – Claude Opus 4.6 Extended thinking displays an opposite problem: an excessive level of detail. The model understands the text, recognizes the conceptual structure, builds a syndetic apparatus, and applies the norm PN-ISO 999: Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes, but still cannot fully distinguish a central concept from passing mentions. Expert analysis indicates that over half of the comments concern this particular mistake. However, it is a corrigible mistake because reducing the number of correct entries is easier than creating them from scratch.

5. Research Limitations and Methodological Considerations

An objective interpretation of the results of the presented experiment requires pointing out methodological boundary conditions that could have influenced the final form of the generated indexes. Asymmetry of the cognitive context between man and machine constitutes the basic limitation of research. A professional indexer compiling the index had access to the entire monograph (over 700 pages). This allowed the indexer to construct a global model of the book's contents, identify its main content, and recognize the connections between chapters. The language models operated exclusively on an isolated, 73-page-long fragment of the book (pp. 405–477).

While this procedure favoured, to a certain degree, the human approach, it was a technological necessity. Loading the models with the full text of the book would, in most of the tested systems, have exceeded the text window limit or intensified the phenom-

enon of significant loss of attention in the central part of the long publication. Consequently, a reliable evaluation of the reasoning mechanism and the application of the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes* would have become more difficult.

The difference in the infrastructure available to the various models is a significant factor that needs to be noted. ChatGPT 5.3 and Gemini 3.1 Pro were tested only in conversational mode, in which the model generated answers directly in the chat window. Claude Opus 4.6 Extended thinking, on the other hand, operated in an environment facilitating the use of computer tools – it was able to read files, create documents in DOCX, and format output data according to editorial rules, as confirmed by the presence of a two-column layout, impossible to generate in conversational mode. This, however, does not change the quality evaluation of the concept extraction itself and the building of the syndetic apparatus, which are independent of the output format.

Another issue is that language models are non-deterministic by nature, and the same prompt may produce different results on subsequent start-ups. The study discussed in this paper is based on a single run for each variant (model × prompting technique). In a rigorous approach, we would have to perform multiple runs and record measures of the central tendency and variability. However, we consciously accepted this restriction because of the high computational cost of a full evaluation (six experimental variants, each requiring a manual expert estimate) and because of the study's exploratory nature and purpose: evaluating qualitative benchmarks, not estimating statistical parameters.

We dispensed with fine-tuning LLM parameters, such as temperature or frequency penalties, which allow steering the balance between the algorithm's predictability and creativity. The study was also constrained by the subject of evaluation: not isolated language models *sensu stricto*, but their practical implementations accessed through conversational interfaces. This means the results could have been influenced not only by the model's architecture,

but also by application layers added by the service provider, such as system instructions, file organization and processing mechanisms, or result-formatting strategies. In consequence, the study does not compare solely "pure" LLM models, but rather the integrated environments of their practical use. However, this limitation does not weaken the cognitive value of the experiment, because the study aimed to capture the effectiveness of the tools under conditions of actual use by a researcher, editor, or librarian, rather than in a laboratory setting where a model is isolated from its operational layer. The indicated limitations do not undermine the validity of formulated conclusions. The differences observed between the models are so distinct that they would be difficult to explain using only stochastic fluctuations or asymmetry in the test conditions. However, they provide a reason to plan further, more controlled studies, as discussed in the final part of the paper.

6. Discussion and Conclusions

The conducted experiment allows us to formulate a thesis that there is no single form of "AI indexing", and the quality of the results depends fundamentally on the model's architecture and the employed prompt technique. The results produce a three-level maturity hierarchy.

Token-based models in *zero-shot* mode (ChatGPT 5.3, Gemini 3.1 Pro) are at the first level. Indexes generated by these models are characterized by a semantic deficit: low expert benchmark coverage (11,7% and 29,9%), a lack of "see also" references, a flat, non-hierarchical structure, and excessive indexing of passing mentions. These models treat the text as an unstructured sequence of tokens and extract frequently occurring lexical units without understanding their role in the book's structure.

Multi-stage reasoning (chain-of-thought prompting) occupies the second level; it triggers a paradoxical dynamic: shaping the indexes' structural elements at the expense of coverage. ChatGPT 5.3, in the *chain-of-thought prompting* stage, generated 14 sub-entries (compared to sporadic generation in zero-shot mode), but bench-

mark coverage fell to 6,5%. Gemini 3.1 Pro, in the *chain-of-thought prompting* mode, produced 2 "see also" references, but the coverage fell from 29,9% to 14,3%. The *chain-of-thought prompting* technique leads to deeper processing but, at the same time, narrows the scope of attention of the analyzed models – they think better but about fewer concepts.

The agency architecture (Claude Opus 4.6 Extended thinking), placed at the third level, shifts the problem paradigm. Claude Opus 4.6 Extended thinking, in the *zero-shot* (here, spontaneously) and *chain-of-thought prompting* mode, adopted a multi-stage strategy: consulted the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes*, sequentially analysed the text, planned the structure, and only then proceeded to generate the index. As a result, it produced 194 main entries, 117 sub-entries, and 16 "see also" cross-references. Here, a semantic deficit gives way to excessive detail – a qualitatively different issue, much easier to correct.

The research results directly impact how the role of the indexer is perceived. The norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes* assumes that indexes are created by human intellectual analysis – selection, ranking, and construction of semantic relations. The experiment confirms that none of the tested models meets these requirements on its own. Even Claude Opus 4.6 Extended thinking, the best of the tested tools, generates an index that requires expert intervention: elimination of excessive details, better selection and formulation of entries, and repair of formal inconsistencies.

At the same time, the results suggest that the expert's role is transforming. In the traditional model, an indexer creates indexes from scratch, serving as an analyst, an architect, and an executor. In the model supported by AI, its role shifts towards an editor and reviewer, evaluating the accuracy of entries generated by the machine, reducing their excessive number, correcting formal inconsistencies, and adding what was missing. It is a human-in-the-loop (HITL) model in which the machine provides raw mate-

rial with high conceptual coverage, and the human indexer adds expert precision. Significantly, in one case described earlier, the machine's solution was recognized as superior to the human expert's, suggesting that this relationship is not one-way.

This is a relevant conclusion from the study's perspective. Manual indexing of a monograph is a days- or weeks-long process. Index generation by Claude Opus 4.6 Extended thinking takes minutes, and its expert revision takes hours. If further research confirms this, the proportion of the indexer's workload would change fundamentally.

The experiment also provides observations relevant to a wider discussion on the nature of language processing by LLMs. The observed three-tier differentiation of results suggests that the ability to build knowledge structures (and not only to generate text) is an emergent feature and appears only in more complex system configurations which include advanced reasoning mechanisms. The analysis of errors and mistakes made by LLM tools is particularly interesting. ChatGPT 5.3 and Gemini 3.1 Pro do not distinguish implicitly expressed concepts or *build semantic relationships* between them. Claude Opus 4.6 Extended thinking distinguished the same concepts and built relationships, but is unable to distinguish passing mentions from the central concept fully. This is a limitation of higher-level judgment, as it assumes understanding of text as an initial condition. The gap between these two kinds of deficit is greater than could be expected based on numerical indicators alone.

The exploratory nature of the study described in the paper opens up possibilities for further research.

1° It is appropriate to experiment on the entire monograph using models with expanded context windows, as it would eliminate the asymmetry of the cognitive context. 2° Multiple runs (e.g., 5–10 for each variant) would allow us to estimate the variability of the results and to answer the question of how stable the observed differences are. 3° A promising direction would be testing agency systems that have a defined workflow by using different stages to

analyze norms, concept extraction, hierarchy building, consistency verification, and formatting instead of relying on the model's spontaneous strategy. 4° *Fine-tuning* performed on a corpus of expert indexes could allow the models to assimilate terminological conventions and selectivity norms, as their absence is the main source of expert comments in the discussed study. Finally, it would be worthwhile to extend the analysis to other forms of knowledge organization (e.g., thesauri, taxonomies, ontologies), to verify whether the observed maturity hierarchy is universal or specific to indexing. Automated generation of indexes according to the norm PN-ISO 999: *Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes* remains a task without a solution, but the research conducted for this paper shows that the distance between a machine and a human being becomes smaller and its nature evolves. The problem is not that AI fails to understand the text, but that it understands too much without being able to select what is relevant expertly. This shortcoming can be corrected, and the pace of architectural changes in Large Language Models suggests that the next generation of models may significantly reduce it. For the time being, the most effective strategy remains a human-machine collaboration – specifically, an AI-generated index that an expert has edited.

BIBLIOGRAPHY

Greco Bruno Cardoso, Dal'Evedove Paula Regina, Nhacuongue Albino Januário, 'O potencial dos LLM (Large Language Models) na indexação automática', *ISKO Brasil* 2025, 8, https://isko.org.br/ojs/index.php/iskobrasil/article/view/97 [Accessed June 08, 2026].

Babik Wiesław, 'Języki informacyjno-wyszukiwawcze stosowane w indeksach oraz ich rola w organizacji i udostępnianiu informacji', in: *Wymiana informacji i rozwój profesjonalnych usług informacyjnych w edukacji, nauce i kulturze na rzecz społeczeństwa opartego na wiedzy*, eds. Diana Pietruch-Reizes, Wiesław Babik, Polskie Towarzystwo Informacji Naukowej, Katowice 2008, pp. 37–43.

Bartmess Elizabeth, Combs Michele R., 'LLM-generated book indexes: Can they replace professionally created indexes?', *The Indexer* 2025, 43(4), pp. 327–348.
Bartmess Elizabeth, Combs Michele R., 'Can the current generation of large language models (LLMs) produce an adequate book index?', *The Indexer* 2026, 44(1), pp. 35–48, https://doi.10.3828/index.2026.4 [Accessed June 08, 2026].
Bender Emily M., Gebru Timnit, McMillan-Major Angelina, Shmitchell Shmargaret, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York 2021, pp. 610–623.
Bystroń Jan Stanisław, *Człowiek i książka*, ed. 3, Wydawnictwo SBP, Warszawa 2003.
Danek Danuta, *Dzieło literackie jako książka. O tytułach i spisach rzeczy w powieści*, Państwowe Wydawnictwo Naukowe, Warszawa 1980.
Duncan Dennis, *Index, A History of the: A Bookish Adventure from Medieval Manuscripts to the Digital Age*, W. W. Norton & Company, New York 2023.
Hedden Heather, *The Accidental Taxonomist*, 3 ed., Information Today, Medford 2022.
Hjørland Birger, 'Semantics and Knowledge Organization', *Annual Review of Information Science and Technology* 2007, vol. 41, iss. 1, pp. 367–405.
Hodge Gail M., *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*, Council on Library and Information Resources, Washington 2000.
'Indeks', in: *Encyklopedia wiedzy o książce*, ed. Aleksander Birkenmajer, Bronisław Kocowski, Jan Trzynadlowski, Zakład Narodowy im. Ossolińskich, Wrocław, Warszawa, Kraków 1971, pp. 991–992.
'Indeks', in: *Słownik wydawcy*, ed. Barbara Kalisz, Wydawnictwo Naukowo-Techniczne, Warszawa 1997, p. 77.
'Indeks', in: *Podręczny słownik bibliotekarza*, ed. Grzegorz Czapnik, Zbigniew Gruszka, Hanna Tadeusiewicz, Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, Warszawa 2011, p. 117.

'Indeks', in: *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, ed. Bożenna Bojar, Wydawnictwo SBP, Warszawa 2002, p. 83.

'Indeks', in: *Słownik terminologiczny informacji naukowej*, ed. Maria Dembowska, Zakład Narodowy im. Ossolińskich – Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk 1979, p. 51.

'Indeksator', in: *Słownik wydawcy*, ed. Barbara Kalisz, Wydawnictwo Naukowo-Techniczne, Warszawa 1997, p. 78.

Izzard Tanya, '*Book indexing and generative AI', Catalogue & Index* 2025, iss. 211, pp. 3–8, https://journals.cilip.org.uk/catalogue-and-index/article/view/746/680 [Accessed June 08, 2026].

Izzard Tanya, 'Generative artificial intelligence (AI) and its performance at indexing tasks', *The Indexer* 2024, 42(4), pp. 383–400.

Krassowski Maciej, *Leksykon terminów literackich*, ed. 2, Twój Styl, Warszawa 1996.

Król Marek, Krupiński Grzegorz, Sułek Henryk, *Nowy słownik terminów literackich*, Krakowskie Wydawnictwo Naukowe, Kraków 2009.

Lancaster Frederick Wilfrid, *Indexing and abstracting in theory and practice*, 3rd ed., Facet Pub., London 2003.

Lewandowska-Tomaszczyk Barbara, Pawłowski Grzegorz, 'Testing ChatGPT on terminology generation, definitions: translation, and ontology creation in German, English and Polish', *Research in Language* 2025, 23(1), pp. 320–340.

Liu Nelson F., Lin Kevin, Hewitt John, Paranjape Ashwin, Bevilacqua Michele, Petroni Fabio, Liang Percy, 'Lost in the middle: How language models use long contexts', *Transactions of the Association for Computational Linguistics* 2024, no. 12, pp. 157–173.

Manning Christopher D., Raghavan Prabhakar, Schütze Hinrich, *Introduction to information retrieval*, Cambridge University Press, Cambridge 2008.

Marszałek Leon, *Edytorstwo publikacji naukowych*, Państwowe Wydawnictwo Naukowe, Warszawa 1986.

Mulvany Nancy C., *Indexing Books*, 2 ed., The University of Chicago Press, Chicago, London 2005.

O.U. [Ungurian Olgierd], ‘Indeks’, in: *Encyklopedia współczesnego bibliotekarstwa polskiego*, ed. Karol Głombiowski, Bolesław Świderski, Helena Więckowska, Zakład Narodowy im. Ossolińskich – Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk 1976, pp. 134–135.

Osuchowska Barbara, *Poradnik autora, tłumacza i redaktora*, Akademicka Oficyna „Exit”, Warszawa 2011.

Pacek Jarosław, ‘Indeksowanie w XXI wieku. Ewolucja i współczesne funkcje pojęcia’, *Zagadnienia Informacji Naukowej* 2006, no. 2, pp. 32–49.

PN-ISO 999: Information and documentation – Guidelines on Content, Organization, and Presentation of Indexes.

Rytting Christopher Michael, Wingate David, *Leveraging the Inductive Bias of Large Language Models for Abstract Textual Reasoning*, 2021, https://arxiv.org/abs/2110.02370 [Accessed June 08, 2026].

Stasiewska Zofia, *Indeksy do książek*, Biblioteka Narodowa, Warszawa 1964.

Stauber Do Mi, *Facing the Text: Content and Structure in Book Indexing*, Cedar Row Press, Eugene 2004.

tk, ‘Indeks’, in: *Słownik terminów literackich*, ed. 2, ed. Janusz Sławiński, Zakład Narodowy im. Ossolińskich – Wydawnictwo, Wrocław, Warszawa, Kraków, Gdańsk, Łódź 1988, p. 196.

Tokarski Michał, *Indeksy i ich tworzenie (skrypt dla studentów bibliotekoznawstwa oraz informacji naukowo-technicznej*, Uniwersytet Śląski, Katowice 1983.

Troskolański Adam T., *O twórczości. Piśmiennictwo naukowo-techniczne*, ed. 2, Państwowe Wydawnictwo Naukowe, Warszawa 1982.

Wellisch Hans H., *Indexing from A to Z*, H. W. Wilson, New York 1995.

Współwydanie książki. Jak wydać książkę razem z wydawnictwem, ed. Mirosław Szmajda, Sorus, Poznań 2024.

Xuanzhao Dong, Wenhui Zhu, Hao Wang, Xiwen Chen, Peijie Qiu, Rui Yin, Yi Su, Yalin Wan, *Talk Before You Retrieve: Agent-Led Discussions for Better RAG in Medical QA*, https://arxiv.org/pdf/2504.21252 [Accessed March 16, 2026].

Zasady typowego opracowania edytorskiego, ed. Leon Marszałek, Państwowe Wydawnictwo Naukowe, Warszawa 1962.

ZS [Saloni Zygmunt], 'Indeks', in: *Encyklopedia językoznawstwa ogólnego*, ed. 2, ed. Kazimierz Polański, Zakład Narodowy im. Ossolińskich, Wrocław, Warszawa, Kraków cop. 1999, p. 252.

Żmigrodzki Zbigniew, 'Przygotowanie spisu bibliograficznego', in: *Bibliografia. Metodyka i organizacja*, ed. Zbigniew Żmigrodzki, Wydawnictwo SBP, Warszawa 2000, pp. 125–147.

Translated by Dr. Elżbieta Olechowska