ZIN

Studia Informacyjne
Information Studies

VOL. 62 2024 NO. 2A(124A)
p-ISSN 0324-8194
e-ISSN 2392-2648

WDIiB UW

SBP
STOWARZYSZENIE
BIBLIOTEKARZY
POLSKICH

REDAKCJA | EDITORIALS

Redaktor naczelny | *Editor in Chief:* Małgorzata Kisilowska-Szurmińska
Redaktor tematyczny, z-ca redaktora naczelnego | *Co-editor in Chief:* Anna Mierzecka
Sekretarz redakcji | *Secretary:* Magdalena Paul
Redaktor językowy – język polski | *Philological editor (Polish):* Magdalena Paul
Redaktor językowy – język angielski | *Philological editor (English):* Zuzanna Wołodko
Redaktor statystyczny | *Statistical editor:* Anna Grzecznowska
Redaktor techniczny i korekta | *Technical editor and proofreading:* Marta Lach

RADA REDAKCYJNA | EDITORIAL BOARD

Wiesław Babik (*Uniwersytet Jagielloński*)
Peter A. Bruck (*Research Studio, Austria*)
Laurence Favier (*Université de Lille, France*)
Birger Hjørland (*University of Copenhagen, Denmark*)
Michèle Hudon (*University of Montreal, Canada*)
Bruno Jacobfeuerborn (*Deutsche Telecom, Germany*)
Tibor Koltay (em., *Eszterházy Károly Egyetem University, Hungary*)
Kazimierz Krzysztofek (*Uniwersytet SWPS*)
Dariusz Kuźmina (Przewodniczący | Chair) (*Uniwersytet Warszawski*)
Elena Maceviciute (*University of Borås, Sweden*)
Krystyna Matusiak (*University of Denver, USA*)
Mieczysław Muraszkiewicz (*Politechnika Warszawska, Uniwersytet w Białymstoku, Filia w Wilnie*)
Widad Mustafa El Hadi (*Université de Lille, France*)
David Nicholas (*CIBER, UK; Tomsk State University, Russia*)
Henryk Rybiński (*Politechnika Warszawska*)
Barbara Stefaniak (em., *Uniwersytet Śląski*)
Elżbieta Stefańczyk (*SBP*)
Tomasz Szapiro (*Szkoła Główna Handlowa w Warszawie*)
Joseph T. Tennis (*University of Washington, USA*)
Jadwiga Woźniak-Kasperek (*Uniwersytet Warszawski*)
Elżbieta B. Zybert (em., *Uniwersytet Warszawski*)

RADA KONSULTACYJNA | CONSULTING BOARD

Hanna Batorowska, em., UP
Sabina Cisek, UJ
Piotr Gawrysiak, PW
Ewa Głowacka, UMK
Małgorzata Góralska, UWr
Mirosław Górny, UAM
Elżbieta Gondek, em., UŚ
Artur Jazdon, UAM
Katarzyna Materska, UŚ
Marek Nahotko, UJ
Piotr Nowak, UAM
Zbigniew Osiński, UMCS
Diana Pietruch-Reizes, UJ
Maria Próchnicka, UJ
Arkadiusz Pulikowski, UŚ
Remigiusz Sapa, UJ
Jadwiga Sadowska, em., UwB
Marta Skalska-Zlat, em., UWr
Stanisław Skórka, UP
Marzena Świgoń, UWM
Jacek Tomaszczyk, UŚ

Wersja elektroniczna (http://ojs.sbp.pl/index.php/zin) jest wersją pierwotną czasopisma.
Zawartość czasopisma jest indeksowana w CEJSH, CEEOL, CSA, LISTA,
Knowledge Organization Literature, PBN, PBB.

The paper version is the original version of the journal.
The content of the journal is indexed in CEJSH, CEEOL, CSA, LISTA,
Knowledge Organization Literature, PBN, PBB.

Pełne teksty roczników 1972–2023 znajdują się w Archiwum SBP: http://www.ac.sbp.pl/;
pełne teksty roczników 2009–2023 znajdują się na Platformie OJS
Czasopism Naukowych SBP: http://ojs.sbp.pl/index.php/zin/issue/archive

ZAGADNIENIA INFORMACJI NAUKOWEJ
Studia Informacyjne

ISSUES IN INFORMATION SCIENCE
Information Studies

VOL. 62 2024 NO. 2A(124A)
p-ISSN 0324–8194
e-ISSN 2392–2648

WDIiB UW

SBP
STOWARZYSZENIE
BIBLIOTEKARZY
POLSKICH

WYDAWNICTWO
NAUKOWE
I EDUKACYJNE
SBP

Warszawa 2024

ISSUES IN INFORMATION SCIENCE – INFORMATION STUDIES

The core purpose of *Issues in Information Science – Information Studies (Zagadnienia Informacji Naukowej – Studia Informacyjne,* ZIN – *Studia Informacyjne)* is to provide a forum for the dissemination of scientific papers and research results in the field of information science and other disciplines which analyze social and technological aspects of various information-related activities performed by contemporary communities. Moreover, the journal is to disseminate critical reviews and summaries of new publications in the field of information science and reports from important conferences discussing contemporary information problems.

We publish papers in Polish or English. For each paper a set of metadata is provided: an abstract and keywords in both languages) as well as author's bio and contact information.

The subtitle of the journal – *Information Studies* – emphasizes the interdisciplinary nature of its subject profile covering a broad spectrum of issues studied by various academic disciplines and professional activity domains related to access to resources of recorded information and knowledge and the use of these resources by contemporary man and society. Other subjects to be covered by ZIN – *Information Studies* involve: (1) theoretical ponderings on the practice of information-related activities performed by various communities, (2) the results of research on the conditions influencing those activities and ways of improving methods and tools employed for the activities in question, (3) the methodology of information science research, information science history and education concerning the information science. The subject profile of ZIN – *Information Studies* covers, among else, the issues of:

- information services in institutions of science, culture, business, education and administration,
- information and knowledge management,
- traditional and online scholarly communication,
- information and knowledge organization,
- metadata theory and practice,
- Web 2.0,
- Semantic Web,
- information architecture,
- information websites usability,
- digital humanities,
- human-computer interaction,
- natural language processing,
- information retrieval,
- use of information and behavior of the information users,
- social response to modern information technologies,
- culture of information,
- information, digital and media skills,
- information policy,
- information ethics.

ZIN – *Information Studies* is addressed to: (1) information science teachers and lecturers, researchers and students, (2) practitioners of information-related activities who analyze methods and tools used to implement those activities in various domains and organizational environments, (3) politicians and donators related to information activities in various domains. The journal content may also be of some interest to teachers, students and researchers in other disciplines of science which deal with various aspects of information existence and use in the contemporary world.

ZIN – *Information Studies* is included in the list of journals scored by Polish Ministry of Science and Higher Education and indexed by: Central European Journal in Social Sciences and Humanities (CEJSH), Central and Eastern European Online Library (CEEOL), Library and Information Science and Technology Abstracts (LISTA), Polish Bibliography of Book Studies (PBB), WorldCat and Polish Scholarly Bibliography (PBN). The journal is registered in the European Reference Index for the Humanities (ERIH Plus).

ZAGADNIENIA INFORMACJI NAUKOWEJ – STUDIA INFORMACYJNE

Głównym celem *Zagadnień Informacji Naukowej – Studiów Informacyjnych* (ZIN – *Studia Informacyjne*) jest zapewnienie forum dla rozpowszechniania artykułów naukowych i wyników badań z zakresu nauki o informacji (informatologii) oraz innych dyscyplin, w których podejmowane są analizy społecznych i technologicznych aspektów działalności informacyjnej prowadzonej w różnych sferach współczesnego życia społecznego. Czasopismo służyć ma również rozpowszechnianiu krytycznych recenzji i omówień publikacji z tego zakresu oraz problemowych sprawozdań z ważnych konferencji poświęconych współczesnym problemom informacyjnym.

Publikujemy artykuły w językach polskim i angielskim. Każdy artykuł posiada zestaw metadanych: abstrakt i słowa kluczowe (w obu językach) oraz nota biograficzna autora i dane do kontaktu z nim.

Czasopismo adresowane jest zarówno do czytelnika polskiego jak i zagranicznego, publikujemy artykuły zarówno w języku polskim jak i angielskim. Podtytuł czasopisma – *Studia Informacyjne* – podkreśla interdyscyplinarny charakter jego profilu tematycznego, który obejmuje szeroki zakres problemów podejmowanych przez dyscypliny akademickie i dziedziny działalności zawodowej związane z zapewnianiem dostępu do utrwalonych zasobów informacji i wiedzy oraz ich wykorzystywaniem przez współczesnego człowieka i współczesne społeczeństwo. Czasopismo publikuje też artykuły prezentujące teoretyczną refleksję o praktycznej działalności informacyjnej prowadzonej w różnych dziedzinach i obszarach życia społecznego, a także wyniki badań służących poznaniu różnych uwarunkowań tej działalności oraz doskonaleniu jej metod i narzędzi. Na łamach ZIN publikowane są także artykuły poświęcone metodologii badań informatologicznych, historii nauki o informacji oraz edukacji w zakresie nauki o informacji. Profil tematyczny półrocznika ZIN – *Studia Informacyjne* obejmuje m.in. problematykę:

- usług informacyjnych w instytucjach nauki, kultury, biznesu, edukacji i administracji,
- zarządzania informacją i wiedzą,
- komunikacji naukowej i cyfrowej komunikacji naukowej,
- organizacji informacji i wiedzy,
- teorii i praktyki metadanych,
- zagadnień Web 2.0,
- zagadnień Sieci Semantycznej,
- architektury informacji,
- projektowania użytecznych serwisów informacyjnych,
- humanistyki cyfrowej,
- interakcji człowiek – komputer,
- przetwarzania języka naturalnego,
- wyszukiwania informacji,
- wykorzystywania informacji i zachowań informacyjnych użytkowników,
- społecznej recepcji nowoczesnych technologii informacyjnych,
- kultura informacji,
- kompetencji informacyjnych i cyfrowych,
- polityki informacyjnej,
- etyki informacyjnej.

Zagadnienia Informacji Naukowej – Studia Informacyjne adresowane są do wykładowców, badaczy i studentów nauki o informacji, a także praktyków działalności informacyjnej, krytycznie analizujących metody i narzędzia jej realizacji w różnych środowiskach dziedzinowych i organizacyjnych oraz polityków i donatorów działalności informacyjnej w różnych dziedzinach. Lektura czasopisma może też zainteresować wykładowców, studentów i badaczy innych dyscyplin, które zajmują się różnymi aspektami funkcjonowania informacji we współczesnym świecie.

Zagadnienia Informacji Naukowej znajdują się na liście czasopism punktowanych Ministerstwa Nauki i Szkolnictwa Wyższego. Czasopismo jest indeksowane w bazach: Central European Journal in Social Sciences and Humanities (CEJSH), Central and Eastern European Online Library (CEEOL), Library and Information Science and Technology Abstracts (LISTA), Polska Bibliografia Bibliologiczna (PBB), WorldCat, Polska Bibliografia Naukowa (PBN). Czasopismo jest zarejestrowane w European Reference Index for the Humanities (ERIH Plus).

Contents

From the Editors (Magdalena Paul Szałkowska) 7

Marzena Świgoń
GenAI in scholarly communication in light of interviews with humanists, theologians, and artists in the early stages of their careers in 2023 and 2024 9

Monika Krakowska, Magdalena Zych
The art of prompt engineering as an old/new form of dialogic information seeking using artificial intelligence models 23

Len Krawczyk, Łukasz Iwasiński, Mateusz Szymański
Data scientists in the scientific literature: LDA topic modelling on the semantic scholar database 47

Arkadiusz Pulikowski
The application of the ChatGPT language model for automatic generation of structured abstracts 75

Call for Papers 86

From the Editors

When „AI" was chosen as the word of the year by the Collins Dictionary and „artificial intelligence" was named the phrase of the year by the linguists' committee in the 2023 Word of the Year competition under the patronage of the Polish Language Council, intense public debates were already underway regarding the impact of AI on education, science, art, work, and other areas of life. Despite the passage of time and a certain degree of acclimatisation to new technologies based on artificial intelligence, these discussions remain ongoing. In this issue, we invite you to engage in a scholarly debate on the relationship between information science, artificial intelligence, and non-human actors interacting with people (such as bots, algorithms, autonomous systems, etc.).

Given the central theme of this issue, we could not resist asking one of the chatbots for its perspective on the matters of interest to us. Here is ChatGPT 4o's response to the question: „In your opinion, what is the significance of artificial intelligence and non-human actors for the discipline of information science? What would you like to convey to the readers—specialists in the field of information science?"

> **Artificial intelligence and non-human actors: a new reality for information science**
> Contemporary information science stands on the threshold of a fundamental transformation driven by the development of artificial intelligence (AI) and the emergence of non-human actors. This is a moment when we must perceive information not only as a resource but also as an element of a dynamic ecosystem in which both humans and autonomous technological systems participate.
>
> **The contribution of AI and non-human actors to information science**
> AI provides tools that enable researchers and practitioners to explore information on an unprecedented scale. Through automation of data processing, semantic modelling, and machine learning, non-human actors support and sometimes surpass human capabilities in analysing complex datasets. Their role extends beyond mere tools; they are becoming active participants in information management processes, with their influence increasingly felt in academic research, education, and everyday life.
>
> **New challenges and research questions**
> - Ethics and responsibility: How can we manage the growing autonomy of AI systems to mitigate risks associated with manipulation, bias, or privacy violations?
> - The definition of an information actor: Can a non-human entity be considered a legitimate participant in information processes?
> - Human-machine interaction: What communication models should be developed to ensure effective and ethical collaboration between humans and non-human systems?
>
> **A vision for the future**
> Researchers specialising in information science have a unique opportunity—and responsibility—to shape how AI and non-human actors are integrated into society. This is not merely about technological advancement but also about understanding its implications for human experience, knowledge, and values. Together, we can create a future in which technology enhances our ability to understand and manage information while preserving human autonomy and ethical foundations.

Leaving it to our Readers to assess the originality and relevance of the chatbot's response, we invite you to explore the articles prepared by representatives of information science and related disciplines.

The topic of artificial intelligence and non-human actors in the context of information science has generated significant interest among authors. Consequently, we encourage you to explore this special issue, which presents articles on artificial intelligence in the context of information science published in English. We encourage everyone to explore the scientific articles that have been prepared by the Authors.

For the first article, Marzena Świgoń conducted interviews with young Polish scientists from humanities, theology, and arts and gathered their opinions on the impact of GenAI on various aspects of scholarly communication. Monika Krakowska and Magdalena Zych aimed to address a gap in research on integrating information retrieval models with prompt engineering. Jagiellonian University scholars' paper synthesises theoretical and practical considerations of dialogic communication with artificial intelligence. In turn, Len Krawczyk, Łukasz Iwasiński and Mateusz Szymański focused on the representation of data scientists in scientific literature. The authors used Latent Dirichlet Allocation (LDA) topic modelling to the resources available within the Semantic Scholar API. Finally, Arkadiusz Pulikowski prepared a qualitative study to evaluate the usefulness of the ChatGPT language model in generating structured abstracts for academic publications.

We also invite everyone to familiarise themselves with the themes of the upcoming issues planned for 2025. The first will focus on research into the affective aspects of information behaviour, the emotions that influence or accompany them, and issues of human information well-being. The second will be dedicated to scholarly communication, academic libraries, bibliometrics, and the broader field of data science.

On behalf of the Editorial Board
Magdalena Paul Szałkowska

GenAI in scholarly communication in light of interviews with humanists, theologians, and artists in the early stages of their careers in 2023 and 2024[1]

Marzena Świgoń
ORCID: 0000-0003-3600-8349
Institute of Journalism and Social Communication
Faculty of Humanities
University of Warmia and Mazury in Olsztyn

Abstract

Purpose: This paper presents opinions on the impact of GenAI on various aspects of scholarly communication in light of interviews with Polish representatives in three fields: humanities, theology, and arts.
Methods: The longitudinal interviews were conducted in two rounds, in the spring of 2023 and 2024 (National Science Centre project No 2022/45/B/HS2/00041), with early career researchers in disciplines such as linguistics, literary studies, history, archeology, philosophy, polish language studies, science of culture and religion, art sciences, theology, music, film, and fine arts.
Results: A significant difference was found between the two rounds of interviews concerning Polish respondents' experience using GenAI. Although the interest in GenAI increased during the year, it only involved preliminary exploration.
Value: The results can be used in future comparative analyses, both time-related and among respondents in various fields and countries. The extract of results from the Polish interviews described here also contributes to the international analysis of Harbingers 3.

Keywords

ChatGPT. Early career researchers. GenAI. Generative artificial intelligence. Harbingers. Humanists. Scholarly communication. Scientists-artists. Theologians.

Text received on 1st of July 2024.

---

[1] This research was funded in whole or in part by the National Science Centre in Poland, grant number 2022/45/B/HS2/00041. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

ZIN 2024, 62(2), 9–22
DOI: 10.36702/zin2024.02.07

1. Introduction

Generative Artificial Intelligence (GenAI) denotes a set of AI tools that generate text, images, and video from prompts. Their development was made possible by improving deep neural networks, including LARGE LANGUAGE MODELS (LLM). Technology companies like OpenAI, Microsoft, and Google are developing their own GenAI models: ChatGPT, Copilot, Gemini, Midjourney, and DALL-E.

Although GenAI has been in development for several decades, it was not until ChatGPT was made available in the autumn of 2022 that interest increased in the changes that such tools could induce in scholarly communication, i.e. in activities taken by researchers at various stages of their work – from seeking information through using it to publishing and sharing.

The state of the art on this issue is changing dynamically with the development of technology and its increasing use by users. The following is only an outline of the literature review, as there are already in-depth works on the subject (e.g., Herman et al., 2024).

It is still not known what changes GenAI can bring about in these activities from seeking to sharing, as few empirical studies have been conducted in this field, and literature reviews dominate among those publications (Herman et al., 2024; Conroy, 2023; Imran & Almusharraf, 2023; Hosseini et al., 2023; Van Noorden & Perkel, 2023). Publications in information science, including on the ethics of AI, are also worth noting (Floridi, 2023; Capurro, 2020).

The international team of the "Harbingers of Change" projects has observed signs of changes in scholarly communication for nearly ten years, first in the context of open science and popularising social media (Harbingers 1 – 2016–2019), then in the context of the pandemic (2020–2022), and now following the popularisation of GenAI (2023).

What is the potential of GenAI, and will it play a significant role in research and teaching? Is GenAI going to change the scholarly communication system, and if it is, in what areas? Literature on the subject (Herman et al., 2024) mentions possible applications of AI at practically every stage of creative work. They are called, for example, copilots, co-researchers, co-authors or assistants (Ansari, Ahmad, & Bhutta, 2023; Conroy, 2023; Dwivedi et al., 2023; Imran & Almusharraf, 2023; Zielinski et al., 2023), and the number of their users is increasing fast, for example ChatGPT currently has over 180 million users (Duarte, 2024).

At the same time, studies of GenAI use in the academic environment showed both the advantages and disadvantages of these modern tools (Van Noorden & Perkel, 2023; Jie et al., 2023; Fui-Hoon Nah et al., 2023; Susarla et al., 2023; Salah et al., 2023; Nordling, 2023; European Research Council, 2023). Some features included boosting creative processes, helping write and edit texts, formulating thoughts and refining text, minimising language barriers, and – possibly – reviewing. Particular hopes were pinned on help in monotonous, tedious work done by scientists, e.g., in analysing large data sets and literature reviews. On the other hand, the threat of the ChatGPT output unreliability, its so-called "hallucinations," has been reported. Instant responses and a set of references on a prescribed topic must be verified by humans, which takes time and effort. Scientists around the world are now testing GenAI tools.

One should look for signs of changes in scholarly communication by observing beginner researchers in various fields and countries open to technological changes. That is the object of the latest Harbingers 3 project. The Polish study conducted among humanists, theologians, and artists-scientists as part of the NCN project (no. 2022/45/B/HS2/00041), initiated as the Polish continuation of earlier Harbingers 1 and 2 projects (Świgoń, Nicholas, 2023), is currently also an essential part of the international Harbingers 3 analysis (Nicholas et al., 2024).

2. The aim, research questions, methods, and organisation of the study

This paper aims to show the respondents' initial experience and contact with AI and present their opinions on the impact of artificial intelligence development on scholarly communication. Due to the novelty of the subject matter, this was an explorative study (the first interviews were conducted three months after open access to ChatGPT was provided).

The interviewees' responses on AI are part of the output of longitudinal interviews with representatives of three fields, humanities, theology, and arts, on scholarly communication, which were conducted within a project financed by NCN. This paper describes the results of interviews conducted by the author in spring 2023 and spring 2024.

The respondents were at the early stages of their careers, i.e., they were either preparing their doctoral dissertations or they had already obtained the title of doctor – not later, however, than seven years before (they did not have the title of doctor habilitated at the start of the project). Twenty-five humanists (14 representatives of literary studies, linguistics, philosophy, history, Polish language studies, arts, culture and religion studies, and archaeology), theologians (5), and artists-scientists (6 representatives of music, fine arts, and conservation of works of art, as well as film and theatre) took part in the first round of interviews. Twenty-two people were included in the second round.

The study methodology is similar in all Harbingers projects (Świgoń, Nicholas, 2023). The interviews are recorded, and their transcripts are subsequently prepared and supplemented by email. The respondents' statements were analysed qualitatively (Gioia, 2022; Gioia, 2023; Marying, 2000). Although the methodology has already been described in two articles on the ZIN (Świgoń 2023ab) pages, it is worth repeating a few details. The interviews lasted about an hour and a half and were conducted via Skype. Respondents were able to complete their statements by email. Both transcription and coding were carried out manually. The analysis results were illustrated with quotations; all answers given to questions analysed in this study are deposited in an open data repository RepOD.

The first questionnaire for interviews conducted with Polish humanists and artists in spring 2023 (under the NCN grant) contained one general (open-ended) question about AI as one of the external factors impacting scientists' work. However, the 2024 questionnaire contained a whole set of questions. The questions had been prepared in cooperation with an international team of researchers and based on the latest literature on the subject. In this way, the Polish project concerning humanists and artists became an essential part of the international comparative analysis Harbingers 3, as it was the first time that researchers in these fields had been taken into consideration in Harbingers' projects, focused on natural and mathematical sciences (Świgoń, Nicholas, 2023; Nicholas et al., 2024). Since the Polish humanists, theologians, and artists were grouped as Arts&Humanities in Harbingers 3, their opinions had to be presented in greater detail than in the international analysis.

The research questions in this analysis were formed as:

- What is the difference between the respondents' opinions on the role of AI in scholarly communication in spring 2023 and spring 2024?
- How do the respondents perceive the impact of AI on various components of scholarly communication, from information search through its evaluation to sharing, in light of the responses given in spring 2024?

3. Results

3.1. Results of the 2023 interviews

In spring 2023, i.e., several months after ChatGPT was launched (November 2022), a great majority of the respondents (18 out of 25) claimed that maybe in the future, but "for now," neither ChatGPT nor any other AI tool had an impact on their scientific, teaching or artistic activity. This subject did not attract much interest during the interviews that focused on other external factors that impacted scholarly communication, e.g., ministerial reforms, the pandemic, or the outbreak of the war in Ukraine (Świgoń 2023a, 2023b).

Several respondents with some experience using ChatGPT (and/or other tools) emphasised the opportunities and threats arising from the development of artificial intelligence. They saw opportunities in preparing literature reviews, searching for information, and linguistic verification. The threats they saw included chat abuse by students, which is why they announced that they would replace written tests with oral tests as a form of students obtaining credits. They also predicted AI to be less creative than humans. The following quotes illustrate these opinions.

> *AI development is valuable and reasonable as long as it makes searching for source information easier; this applies to all tools for searching online journals. I find any significant interference by artificial intelligence in my discipline harmful.* [history]
>
> *I believe it is a great tool, maybe not in writing papers, but obviously not in philosophy. It is a good tool for quick proofreading.* [philosophy]
>
> *(...) I see a threat that practising science may lose its „flavour." We will talk to a computer rather than another person (...). AI does not pose any threat in my field. It cannot conclude or formulate thoughts with potential for development.* [theology]
>
> *(...) I worry about students, and I think there is a risk of them switching off the thinking process. Because when you treat this tool as a mental process assistant, it will definitely make research work more efficient. For now, I am analysing artificial intelligence from the moral theology perspective and thinking about the moral aspects I can deal with in this context.* [theology]

Scientists-artists working with students expressed interesting opinions about ChatGPT and the tools used to create graphics and videos.

> *A lecturer must make students aware that AI generates random images, and one must change their thinking. Creativity requires a different approach now, seeking new ways, which is the only method for making an artist competitive in AI. Artificial intelligence is a professional threat to artists as it is great for generating images in simple fields of art, e.g., in making illustrations. (...) I teach my students how to protect their work against AI (...).* [fine arts]
>
> *(...) Although I used to fight plagiarism, the student's contribution is now impossible to determine – written papers prepared to win them credit have lost any sense. It is similar to giving students assignments that involve taking a series of photographs (...). It is only a matter of time before artificial intelligence's contribution to generating a photo-realistic image becomes unrecognisable. As a filmmaker, I am not afraid of artificial intelligence yet. The mechanism of filmmaking is largely conservative (...). Artificial intelligence will not revolutionise work on a film set as dynamically as in other fields.* [film]

To conclude this section, one should mention a linguist respondent involved in research projects related to artificial intelligence, which concerned *(...) its practical use: (...) I used my philological knowledge (of the Polish language), and (...) I developed complex linguistic rules for machine processing.*

3.2. Results of the 2024 interviews

One year after the first interviews, no respondents were without contacts, even sporadic or individual, with ChatGPT or other GenAI tools. The responses are characterised below in the same order as in the interview questionnaire.

The first part of the questions concerned general **experience or contact with AI** in their private life and at work.

The extensive descriptive responses, which included the names of applications, programs, and tools, indicated that the respondents understood the concept of GenAI.

ChatGPT appeared in every response as a specific "initiator" of the whole phenomenon and one of the prompts in the interview. Frequent mentions of the name testified to the young scientists' curiosity, although the extent of the phenomenon exploration varied within the group – from an individual attempt to systematic use. As far as the latter is concerned, it was not about writing scientific texts but official letters, social media posts, or editing grant applications.

Regarding texts, Polish humanists, theologians, and artists often mentioned the names of various translators, including Google and DeepL, as well as programs for checking spelling and style, e.g., Grammarly, Jasnopis or an umbrella name of LLM (Large Language Models). They also mentioned the names of various GenAI tools, including those used for working with images, e.g., Midjourney, DALL-E, Adobe, Photoshop, Runaway, Google Genie, and Canva. They provided assistants' names in popular search engines: Google Bard and Microsoft Copilot. Only some admitted to buying a subscription when asked about paid versions of various programs. In general, they used free versions.

The rapid development of AI is perceived as a positive thing by this respondent group, although with little enthusiasm. Hope was expressed for such an improvement of those tools that would result in workload reduction, although this was not the case at this stage – in light of responses given in spring 2024. Apart from individual mentions of text editions – casual rather than purely scientific – the respondents saw a specific obligation to learn new programs. The respondents mention learning new tools during classes together with students.

The respondents also expressed their concerns associated with distrust and the need to verify the responses generated by AI, which requires extra effort and time. A possible increase in the number of cases of plagiarism was also mentioned. It was suggested that the standards and legislative solutions applicable to AI use in various areas of life and work should be altered to eliminate fraud threats and protect personal data. Generally, nearly all respondents were more or less concerned about the unknown role that AI may play in the future.

The differences in perception of the AI development dynamics between spring 2023 and spring 2024 are appropriately illustrated by what an artist said:

> *A year ago, we were not aware of the potential of AI. Many technology startups have been established since then. Vast amounts of money have been invested in them. When I last wrote about it, I pointed out that AI coped very well with photographs and images. However, it had problems with video, as it disregarded multiple aspects before and after a frame, which is why the resulting film was of low resolution and full of artefacts. Interestingly, the paper is still in print, and its topicality has been lost.* [film]

When asked whether their opinions were typical of their professional and private circles, they generally responded that they were. However, there were extreme individual responses, both more enthusiastic than their friends and indicative of more considerable resistance against technology novelties.

Only one artificial intelligence researcher was among the respondents in the first round of interviews, but several in the second round mentioned including this subject matter in their research.

The following are the responses to the other questions concerning AI from the interview questionnaire, which were more detailed compared to the initial part of the interview.

The respondents were asked about AI's impact on research **reputation** and its positive (enhanced productivity) and negative (lack of transparency) implications. [Does AI have any implications for research reputation? It may include positive (e.g., enhanced productivity) and negative implications (lack of transparency).] Most respondents pointed to various implications, and six (out of 22) had no opinion.

Among the positive implications, the respondents mentioned more effective detection of plagiarism, reduction of tedious work, more attractive classes with students, acceleration of research processes, extending the scope of research, more effective verification of information, support in obtaining and analysing data and minimising any research errors.

Negative implications included: verification of the text originality is more complex, lack of possibility of verifying the actual extent to which AI supported the researcher's work, a decrease in reliability of scientific research, an increase in the number of publications at the expense of their quality, relaxing of minds and weakening of critical thinking, creating works of art and science with no individual style or character, the risk of work on unreliable data from AI, lack of procedures and control of AI use in scientific research, unauthorised use of accomplishments of other researchers and artists, text unification as a result of machine translations.

The impact of an **AI function in search engines** on the results was relatively small in this group of respondents. A majority (18 out of 22) responded that either they did not see any significant changes compared to the first interview of spring 2023 or they knew nothing of search support by AI tools. One person said that she had switched off the AI function in Bing. There were only four comments with positive reflections. These respondents described the search engines' work as "much faster" and "more intuitive" and AI's suggestions concerning publications on a prescribed topic as "useful." Artists mentioned the fast development of graphic programs during the year between the interviews.

> *It is now much faster to search for information via search engines with AI and built-in chat features than a year ago. (...) Adobe databases also have their graphics support browsers. I feel that the cosmic difference compared to when we first spoke, especially about Photoshop, filling in backgrounds, and various other elements, is now a potent, time-saving tool.* [fine arts]

Another question concerned the impact of AI on **sharing** research results and presenting them in a more accessible form [Does AI have any implications for connecting and/or research sharing? For instance, summarising/translating research into more accessible formats]. Unfortunately, most respondents had no opinion about it, probably due to insufficient experience. Those who answered this question pointed to the benefits of automatic translation of texts in foreign languages, i.e., the advantages of various translators and spellcheckers. Some respondents mentioned text abstracts with AI tools and identifying keywords.

There was a question in the questionnaire about **identifying** the AI contribution to text generation [What would make them suspect that published material was AI generated?] Nearly all the respondents (20 out of 22) answered it. Apart from hints/prompts, such as inconsistent writing style, lack of personalisation, content too good to be true, and inaccurate/lack of proper citations, the respondents also mentioned the following: repeating information within one issue, errors concerning the essence of an issue (or explanations provided in a way it is done in encyclopedias, without using the terminology usually used by experts); disordered style; verbosity (formally correct texts, but without any sense); superficiality (too general, insufficiently nuanced considerations); no traces of the author's personality; inconsistent or contradictory citations; wrong paraphrasing (can be recognised if one knows the original text); proper content, but generalised and presenting the same conclusions in a loop.

Moreover, respondents-artists pointed to features of images that could suggest the use of AI, such as uneven texture, incorrect lighting, lack of natural defects and random elements, images that are too perfect, or the opposite-ones with apparent errors.

Other questions were also associated with the quality of information. They are concerned about the risk of **text "production,"** even whole periodicals of a low scientific value [Do they believe that the AI-associated potential for rapid production of low-quality scientific articles brings about a decline in the overall quality of research output, indeed, facilitated the growth of predatory journals and papermills?]. Nearly all respondents in the three fields admitted that the risk was real. However, their responses were like forecasts rather than an account of their experience because many had not encountered the issue in their field. Individual responses to the opposite effect suggested that AI capabilities were overrated, for example, in generating philosophical or high value scientific texts.

The respondents (representatives of various disciplines, not only of arts) brought up the issue of fake texts, images, and videos. They talked about "animating" images, replacing voice, and generating photographs in a specific style. One of the artists said that *we will either learn to conduct multi-step verification or trust what we see because once it is posted online, its verification will not be possible.* A respondent from culture and religion studies mentioned the need *for proper self-education and for teaching it in student classes.*

These considerations were supplemented with a question about other risks associated with AI use in the context of **ethics** [Do they think AI is raising any other issues of integrity and ethics?]. Here, the respondents focused on examples provided in the question, i.e., deepfakes, job displacement, and autonomous systems. They pointed to the need for introducing new legislation for state-controlled procedures for data verification and counteracting fake information. They suggested education at all levels (children, adolescents, and adults). It was also proposed that perpetrators of AI-related abuse should be punished.

The issue of publishing in science started with a question about changing the policy concerning publication **authorship** associated with AI development [Are these policies – authorship and publishing – changing because of 'AI' becoming (sort of) another author?]. The Polish respondents in the three fields admitted that nothing had changed in this regard and/or that they had heard nothing about it. Some respondents even argued that AI could not and would not be regarded as a text co-author.

The next question was about AI's help in **publishing productivity** [Has 'AI' helped publishing productivity? For instance, has 'AI' been used to expedite and/or make more efficient and/or improve the writing process of grant proposals and/or publications? Has 'AI' been used to locate suitable journals to publish in according to the manuscript's title or abstract?]. All responses but one were negative. Although they disagreed with the term "productivity" in the humanities context, they admitted that their experience with using AI tools was insufficient to give an opinion on such help. One person admitted to preparing article abstracts using AI. Several respondents talked about the systematic use of translators and programs to check spelling and style.

The question about AI for **summarising** articles [Used 'AI' as a tool for summarising scientific articles/extracting critical information from complex texts to facilitate a literature review] was also answered negatively in most cases, with only three respondents admitting to using ChatGPT for this purpose. However, they believe these tools are unreliable as they leave out key points. In other words, in light of the spring 2024 responses, the AI capabilities in summarising texts were overrated or insufficiently explored in the group under study.

Humanists did not have much experience in AI to formulate research questions or hypotheses [Used 'AI' to detect gaps in knowledge to locate a topic for new research and to construct hypotheses]. Only one person mentioned AI being used to *determine work directions (...) to bring some order to further steps*, which did not give the expected results.

According to earlier Harbingers' studies, when looking for **journals to publish their findings**, the respondents considered factors such as IF, prestige in the discipline, time needed for publication, open access, and indexation in international databases. The present question was whether AI changed these factors in any way [Will 'AI' change their relative ratings or introduce any new factors?]. It turned out that it did not – none of the respondents mentioned any new circumstances. Therefore, these choice factors remain the most important. According to one of the respondents, some editorial boards and institutions have asked for some time for certificates that a text was not generated with AI.

The interview questionnaire also brought up the issue of the impact of AI on **reviewing** [What do you think an AI-based peer-review should be capable of doing if it is to replace the current system? For instance, speeding up a review, using an automated reviewer, or a post-publication peer review?]. The respondents admitted, in general, that they had not thought about it as they had not had to face this problem. Some of them thought that reviewers should not use AI in any way as it cannot replace a human. On the other hand, others thought it would be possible if AI tools were improved, and a human would make the final decision. Those respondents who predicted that AI would be used for preparing reviews in the future mentioned potential advantages such as a quick review, better ability to recognise plagiarism, more insight, help in editing a review text, and increased objectivity of a text evaluation. A respondent from philosophy added: In this exploitation-based system of reviewing articles for very expensive periodicals for free, I *favour replacing humans with AI tools.*

The question about using AI to propagate scientific research [Is 'AI' capable of facilitating/enhancing their outreach activities?] brought only answers in the categories of "I do not know" and "I have no experience in this regard." One person mentioned using ChatGPT to write posts on social media, which was adapted to various user age groups.

The penultimate question in the interview questionnaire concerned AI as a transformational force in the scholarly communication system [Will AI be a transformational force? If so, in what ways? What will be the advantages and disadvantages of the transformations that will take place?] Most responders gave an affirmative answer (only 5 out of 22 people gave a negative one or were unsure). The advantages of such transformation included better and more in-depth analyses compared to human abilities, making work easier by analysing large bodies of data and performing other automatic activities, quick and effective text translation, quicker information circulation, expanding the information range, quicker information selection, reducing the size of student groups, the emergence of new scientific disciplines. Its disadvantages included the unreliability of AI-generated content, lack of responsibility for errors, data theft, violation of privacy and other abuses, information quality deterioration due to fake news and disinformation, the abundance of low-quality papers, lulling researchers into a sense of security and making them indolent, and the need for greater regulation and reform in science.

The following statement illustrates the predicted appreciation of human skills:

> *Most people will buy cheap, low-quality books for children, written and illustrated by AI and published in many copies. However, a minority will buy expensive books written and illustrated by original and high-quality artists. That is a rhetorical question: Which children benefit more from reading their books? However, this applies only to humanities and arts* [Polish language studies]

The last question in the 2024 interview concerned a deepening inequality in access to AI tools [Will the use of AI exacerbate existing disparities and inequalities, with people with access to AI-based tools speeding up their publication processes]. As in the previous question, the respondents likely agreed with this hypothesis (7 out of the 22 answered *no/I do not know*). It was pointed out that individuals with higher technical capabilities and access to paid AI tools can gain an advantage, be more productive, write better texts, and achieve their goals with less effort. Some humanists did not expect inequalities arising from GenAI use in their speciality to deepen because it was based on sources that have yet to be digitised. Consequently, it requires traditional work methods.

4. Conclusion

This paper presents the first experiences and opinions of beginner humanists, theologians, and artists-scientists on the impact of GenAI on scholarly communication. It was one of the first studies in these disciplines worldwide.

A considerable difference was observed in perception of the subject matter between the first and the second rounds of the longitudinal interviews. Only several respondents in the spring 2023 interviews admitted that such programs as ChatGPT and other AI tools had an impact on scientific, artistic, and teaching activities, and a majority of the respondents (out of 25 individuals in three fields) claimed to not see such an impact. Most respondents talked about the absence of any impact, or they added “for now,” i.e., predicting that this could change. The situation was different after a year, i.e., in spring 2024, when all the respondents in all three disciplines had had contact with this issue, although to a various extent. Several specific questions (based on the literature of the subject analysis) related to the impact of AI on various aspects of scholarly communication were answered in a way that indicated on one hand curiosity – but not excitation – about the subject matter, but on the other – an initial phase of recognising and exploring new tools. Of the topics proposed for discussion, the most significant response was obtained about AI’s impact on a researcher’s reputation and so-called “productivity,” described in the context of opportunities and threats. The other aspects of scholarly communication, such as authorship, searching, or sharing/propagation, provoked several responses indicative of the lack of knowledge and experience among the group of humanists, theologians, and artists-scientists under study.

Despite the significant changes in the GenAI between spring 2023 and spring 2024 and, consequently, potential changes in opinions soon, one can still hope that this study provided interesting comparative material for future analyses, including comparisons as part of the international Harbingers 3 project in various fields and countries and future studies concerning scholarly communication.

Open Data

Świgoń, M.(2024). *Polscy badacze na wczesnym etapie kariery w humanistyce, teologii i sztuce – o generatywnej sztucznej inteligencji w komunikacji naukowej* [Data set]. RepOD. https://doi.org/10.18150/MLUX8U.

Funding

This research was funded in whole or in part by the National Science Centre in Poland, grant number 2022/45/B/HS2/00041. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

References

Ansari, A.N., Ahmad, S. & Bhutta, S.M. (2023). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. *Education and Information Technologies*, 29, 11281–11321. doi: 10.1007/s10639-023-12223-4.

Capurro, R. (2020). The age of Artificial Intelligences: A personal reflection. *International Review of Information Ethics*, 28.

Conroy, G. (2023). Scientists used ChatGPT to generate an entire paper from scratch — but is it any good? *Nature*, 619, 443–444. doi: 10.1038/d41586-023-02218-z

Duarte, F. (2024). Number of ChatGPT users (June 2024) [online]. *Exploding Topics*. Retrieved from: https://explodingtopics.com/blog/chatgpt-users.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 1026242. doi: 10.1016/j.ijinfomgt.2023.102642.

European Research Council (2023). *Foresight: Use and impact of Artificial Intelligence in the scientific process* [online]. https://erc.europa.eu/sites/default/files/2023-12/AI_in_science.pdf.

Floridi, L. (2023). *The Ethics of Artificial Intelligence: principles, challenges, and opportunities*. Oxford Academic. https://doi.org/10.1093/oso/9780198883098.001.0001.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304. doi: 10.1080/15228053.2023.2233814.

Gioia, D. (2021). A systematic methodology for doing qualitative research. *Journal of Applied Behavioral Science, 57*(1), 20–29.

Herman, E., Nicholas, D., Abrizah, A., Revez, J., Rodríguez Bravo, B., Swigon, M., Clark, D., Xu, J., Watkinson, A. (2024). The impact of AI on the post-pandemic generation of early career researchers: What we know or can predict from the published literature. *Learned Publishing, 37*(4), e1623. doi: 10.1002/leap.1623.

Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., MacDonald, N., Holmes, K.L. & Kho, A. (2023). An exploratory survey about using ChatGPT in education, healthcare, and research. *Plos One, 18*(10), e0292216. doi: 10.1371/journal.pone.0292216.

Imran, M., & Almusharraf, N. (2023). Analysing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. doi: 10.30935/cedtech/13605.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. doi: 10.1145/3571730.

Magnani, G. & Gioia, D. (2023). Using the Gioia Methodology in international business and entrepreneurship research. *International Business Review, 32*(2), 102097.

Mayring, P. (2000). Qualitative Content Analysis [28 paragraphs]. *Forum QualitativeSozialforschung / Forum: Qualitative Social Research, 1*(2), Art. 20, http://nbn-resolving.de/urn:nbn:de:0114-fqs0002204.

Nicholas, D., Swigon, M., Clark, D., Abrizah, A., Revez, J., Herman, E., Rodrigues-Bravo, B., Xu, J., Watkinson, A. (2024). The impact of generative AI on the scholarly communications of early career researchers: An international, multi-disciplinary study. *Learned Publishing, 37*(4), e1628. doi: 10.1002/leap.1628.

Nordling, L. (2023). How ChatGPT is transforming the postdoc experience. *Nature, 622*, 655–657. doi: 10.1038/d41586-023-03235-8.

Salah, M., Al Halbusi, H., & Abdelfattah, F. (2023). May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. *Computers in Human Behavior: Artificial Humans, 1*(2). doi: 10.1016/j.chbah.2023.100006.

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus effect of Generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research, 34*(2), 399–408. doi: 10.1287/isre.2023.ed.v34.n2.

Świgoń, M. (2023a). Early career researchers in art, humanities and theology at the time of external crises – in light of the first round of longitudinal interviews of spring 2023. *Zagadnienia Informacji Naukowej, 61*(1), 52–70.

Świgoń, M. (2023b). How do early career researchers perceive success in their fields? – report on interviews with humanists, theologians and scientists-artists in Poland. *Zagadnienia Informacji Naukowej, 61*(2), 32–49. doi: 10.36702/zin2023.02.02.

Świgoń, M., Nicholas, D. (2023). Scholarly communication of early career researchers – organisation of international research with the participation of Poles and its Polish follow-up. *Media – Kultura – Komunikacja społeczna, 19*, 133–148. doi: 10.31648/mkks.8595.7008.

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature, 621*, 672–675. doi: 10.1038/d41586-023-02980-0.

Zielinski, C., Winker, M.A., Aggarwal, R., Ferris, L.E., Heinemann, M., Lapeña, J.F., Pai, S.A., Ing, E., Citrome, L., Alam, M., Voight, M., Habibzadeh, F. (2023). *Chatbots, Generative AI, and scholarly manuscripts. WAME recommendations on Chatbots and Generative Artificial Intelligence in relation to scholarly publications* [online]. WAME. [May 31, 2023]. Retrieved from: https://wame.org/page3.php?id=106.

Generatywna sztuczna inteligencja w komunikacji naukowej w świetle wywiadów z humanistami, teologami i artystami na wczesnym etapie kariery w 2023 i 2024 roku

Cel: W artykule przedstawiono opinie na temat wpływu GenAI na różne aspekty komunikacji naukowej w świetle wywiadów z przedstawicielami trzech dziedzin: humanistyki, teologii oraz dziedziny sztuki.
Metody: Wywiady podłużne odbyły się w dwóch turach – wiosną 2023 i 2024 roku (projekt NCN nr 2022/45/B/HS2/00041) z badaczami z następujących dyscyplin: lingwistyka, literaturoznawstwo, historia, archeologia, filozofia, polonistyka, nauki o kulturze i religii, nauki o sztuce, teologia, sztuki muzyczne, sztuki filmowe oraz plastyczne.
Rezultaty: Stwierdzono dużą różnicę w doświadczeniach w korzystaniu z GenAI przez polskich respondentów pomiędzy dwiema turami wywiadów; chociaż po roku wzrosło zainteresowanie GenAI, nadal było to raczej dopiero rozpoznawanie i wstępna eksploracja zjawiska.
Wartość: Wyniki mogą być wykorzystane do przyszłych analiz porównawczych, zarówno w czasie, jak i wśród respondentów z różnych dziedzin oraz krajów. Opisany tu fragment wyników z polskich wywiadów jest także wkładem do międzynarodowej analizy Harbingers 3.

Słowa kluczowe

ChatGPT. GenAI. Generatywna sztuczna inteligencja. Humaniści. Komunikacja naukowa. Naukowcy-artyści. Początkujący naukowcy. Sygnały zmian. Teologowie.

MARZENA ŚWIGOŃ, Profesor nauk społecznych, dyrektor Instytutu Dziennikarstwa i Komunikacji Społecznej Uniwersytetu Warmińsko-Mazurskiego w Olsztynie, redaktor naczelna czasopisma „Media-Kultura-Komunikacja Społeczna" (Wydaw. UWM); członkini Komitetu Naukoznawstwa Polskiej Akademii Nauk; publikacje: M. Świgoń (2015). Dzielenie się wiedzą i informacją (Wydaw. UWM); M. Świgoń (2013). Personal Knowledge and Information Management – conception and exemplification. "Journal of Information Science", 39(6); M. Świgoń (2012). Zarządzanie wiedzą i informacją. (Wydaw. UWM).

Contact details:
marzena.swigon@uwm.edu.pl
Uniwersytet Warmińsko-Mazurski w Olsztynie
Instytut Dziennikarstwa i Komunikacji Społecznej
ul. Kurta Obitza 1, 10-725 Olsztyn

The art of prompt engineering as an old/new form of dialogic information seeking using artificial intelligence models

Monika Krakowska
ORCID: 0000-0002-2724-9880
Magdalena Zych
ORCID: 0000-0001-9770-3674
Faculty of Management and Social Communication
Jagiellonian University in Krakow

Abstract

Purpose/Thesis: The article synthesises theoretical and practical considerations of dialogic communication with artificial intelligence, focusing on established information retrieval models. It explores the interdisciplinary nature of information behaviour research and the evolution of retrieval models.
Approach/Methods: A qualitative methodology incorporated critical literature analysis and a case study using ChatGPT to search scientific literature.
Results and conclusions. The analysis revealed interdependencies between traditional and modern models, emphasising cognitive and exploratory aspects of information retrieval.
Research limitations: Focuses on specific prompt engineering models and a singular case study.
Practical Implications: Understanding established models is crucial for developing prompt engineering.
Originality/Value: This study addresses a gap in research on integrating information retrieval models with prompt engineering.

Keywords

Artificial Intelligence (AI). ChatGPT. Conversational Information Retrieving. Dialogic Approach. Information Seeking Model. Information Searching. Prompt Engineering (PE)

Text received on 14th of October 2024.

1. Introduction

This article addresses the issue of modelling conversation and conducting a dialogue between humans and machines. Despite the assumption that new technologies and continuously evolving interaction patterns between humans and computers are being developed, the reality is apparent: this is merely an extension and

ZIN 2024, 62(2), 23–46
DOI: 10.36702/zin2024.02.08

refinement of existing communication patterns in information science and the dialogical model of information retrieval in the relationship between humans and computers.

Dialogue systems mimic human conversation, ranging from simple chatbots to more complex components like video game NPCs, which interact with players through dialogue. These systems function as decision trees, where user input triggers specific responses (Pisarski, 2024, p.230). Conversational systems can be categorised into two types: tool systems, designed for efficiency and accuracy, and anthropomorphic systems, which simulate human-like interaction to foster emotional engagement. The latter plays a crucial role in information-seeking models, aiding users in discovering and constructing context through intuitive exploration (Chen J. et al., 2024; Zhou et al., 2024).

Integrating patterns derived from interaction in dialogue systems and information-seeking behaviour will undoubtedly facilitate the evolution of conversational systems beyond their current role as mere information-seeking tools. They must and will act as intermediaries in dynamic, user-centred information landscapes. The conversation influences the user's understanding and experience of the information sought in these landscapes.

The prompt engineering literature highlights clear correlations, focusing on developing and optimising prompts to utilise language models effectively. User experience principles, like Nielsen's heuristics, should be considered when designing conversational systems, particularly in aligning system language with the user's context. Defining the task, formality, and specialist terms is essential for reliability (McNulty, 2024), and prompts must be user-friendly (Springs, 2024). Language, tone, and style should ensure inclusivity, with frameworks like Persona or Audience Persona Patterns helping to tailor systems to user profiles (Corral, 2023). Challenges include linguistic flexibility, representing diverse users, and avoiding over-generalisation. Prompt engineering is closely linked to AI literacy, requiring an understanding of how inputs affect outputs (Bates, 2024; Lund, 2023). Information and digital literacy are crucial when managing generative AI (Zhang, 2024), especially to avoid AI hallucinations—false outputs triggered by tricky queries (Pisarski, 2024). Strategies like handling missing data (Ruksha, 2024; Srinivasan, 2024) can reduce hallucinations but require advanced AI knowledge. Tools like ChatGPT allow users to conduct complex analyses without expertise, raising concerns about safety and reliability. The ACRL's information literacy framework emphasises critical evaluation of AI results (Hall & McKee, 2024), especially in AI-assisted library services where librarians need GenAI skills. Controlled vocabulary use, personality trait analysis via lexical tools like DesPrompt (Wen et al., 2023), information extraction from documents (Yuan et al., 2023), and recommender systems design (Zixuan Yi et al., 2024) underscore the integration of information science with prompt engineering.

There is growing interest in information retrieval principles of relevance, completeness, and accuracy. To elicit creative responses, prompts must be clear and unambiguous and encourage open-ended replies (McNulty, 2024; Springs, 2024). Moreover, prompts must be pertinent to the interaction. They facilitate discourse, aligning with user objectives. Well-crafted prompts are key to meaningful interactions in dialogue systems. This article concerns the pivotal elements of prompt engineering and their interconnection with selected models in information behaviour and retrieval.

2. Conceptual framework

Information seeking and searching is part of the broader spectrum of human information behaviour. Information seeking is an intentional search for relevant information that addresses specific needs (Case & Given, 2016; Cisek, 2017; Savolainen, 2017). The stages and activities involved in this process are also subordinate to broader information behaviour, occurring in various contexts and interactions between humans and technology, such as computers, search systems, or AI models (Krakowska, 2022). Establishing a theoretical basis for understanding the interdependence of information-seeking patterns and prompt engineering is crucial for evaluating the hypothesis.

2.1. Information seeking, searching and retrieving

It is essential to understand comprehensively the processes involved in information retrieval, particularly concerning their interaction with artificial intelligence models and the principles that underpin them. That should be contextualised within broader information-seeking models, which have been extensively researched. The term "retrieval" is often used imprecisely, so clarifying and differentiating the terminology is important. Information behaviours involve multidirectional activities related to sources, channels, and information systems. These include recording, seeking, interpreting, and using information (Fidel, 2011; Ford, 2015). These behaviours can be categorised into intentional and iterative processes, including information seeking, searching, and retrieving, aimed at solving problems and tasks effectively.

In the context of human-computer interactions, information exploratory searching is a crucial aspect of information behaviour. This process includes several activities related to information seeking. The compilation, verification and resolution of queries comprise this process. Exploratory searching is an integrated learning process. Learning is the process of acquiring, comparing and integrating knowledge.

This search is exploratory. Accretion involves enhancing and structuring knowledge through various processes (Marchionini, 2006). This exploratory model addresses complex problems and enhances cognitive capacities in individuals through symbiotic relationships between humans and computers, which identify information landscapes. This exploratory mode of information retrieval engages with resources to foster new knowledge (White & Roth, 2009; Materska, 2020).

Information seeking is the cognitive effort to gain insight, involving searching for accurate information, questioning, or scanning the environment (Thani & Hashim, 2011). It is a form of problem-solving that includes identifying, interpreting, and evaluating information with potential repetition (Marchionini, 1989).

In contrast, information searching focuses on acquiring specific information from particular sources, often online. That involves query formulation and evaluating the relevance of results, with both observable system actions and unobservable cognitive processes at play (Bawden & Robinson, 2012; Spink & Cole, 2006; Spink & Dee, 2007).

Information retrieval, a subset of searching, extracts information using retrieval systems on databases or web resources. It synchronises queries with search terms and results but does not consider user context or the complexity of their information needs (Lin, 2017). Maria Próchnicka defines information retrieval as a multidirectional interaction that transforms the user's knowledge asymmetry by generating complete and relevant information based on the user's needs (Próchnicka, 2004). She highlights that key components in modelling information retrieval behaviour are system performance and individual cognitive features, which shape how queries are formulated and processed. This dialogue with the system integrates and modifies knowledge while addressing or generating new information needs (Próchnicka, 2004; 2001; 2002). Query formulation and result evaluation are tied to problem identification, information extraction, and scanning (Ellis, 1989; 1992; Bates, 1989; Marchionini, 2006). Interactive information retrieval, involving communication between the user, intermediary, and system, supports both conversational and dialogic models of information seeking and links to conversational prompt engineering (Ingwersen, 1992).

Conversational models, which view dialogue as the foundation of human-computer interaction, have long been developed in information science, particularly within information behaviour research. Foundational models by Garry Marchionini and Nicholas Belkin in the 1990s identified key stages in the information retrieval process. Later models, such as those by Marcia Bates (1990; 1999), David Ellis (1989; 1992), Maria Próchnicka (2004), and Reijo Savolainen (2016; 2019), have evolved alongside advancements in information systems and AI. These earlier models form the basis of modern prompt engineering, reflecting expressive and creative information-seeking approaches (Fredrick, 2024; Zhang, 2023). Prompt engineering, a form of query-based interaction, uses patterns learned from natural language to interpret human input, highlighting the fluid and adaptive nature of human-AI communication, including how search queries are refined based on results.

Belkin's concept posits that human interaction with text is active, as individuals seek, engage with, and interpret texts to make meaning and achieve goals (Belkin & Cool, 1993; Belkin & Marchetti & Cool, 1993). This interaction is part of information-seeking behaviour, where people search for texts or advice to resolve knowledge gaps. Garry Marchionini's (1995) information-seeking model outlines eight stages: recognising, comprehending, selecting, formulating, executing, examining, extracting, reflecting, reiterating, and ceasing. The model shows how information is extracted and integrated with existing knowledge. Marchionini also identifies three types of browsing – directed, semi-directed, and undirected – alongside factors that influence search behaviour, such as the searcher, task, system, and context. Belkin and Marchionini's models underpin the dialogical model of information seeking, correlating human cognitive, informational, and emotional processes with system engagement. Figure 1 shows the proposed integrated scheme based on these two models. This dialogue clarifies knowledge gaps, refines queries, and generates new knowledge through interactions with the system.

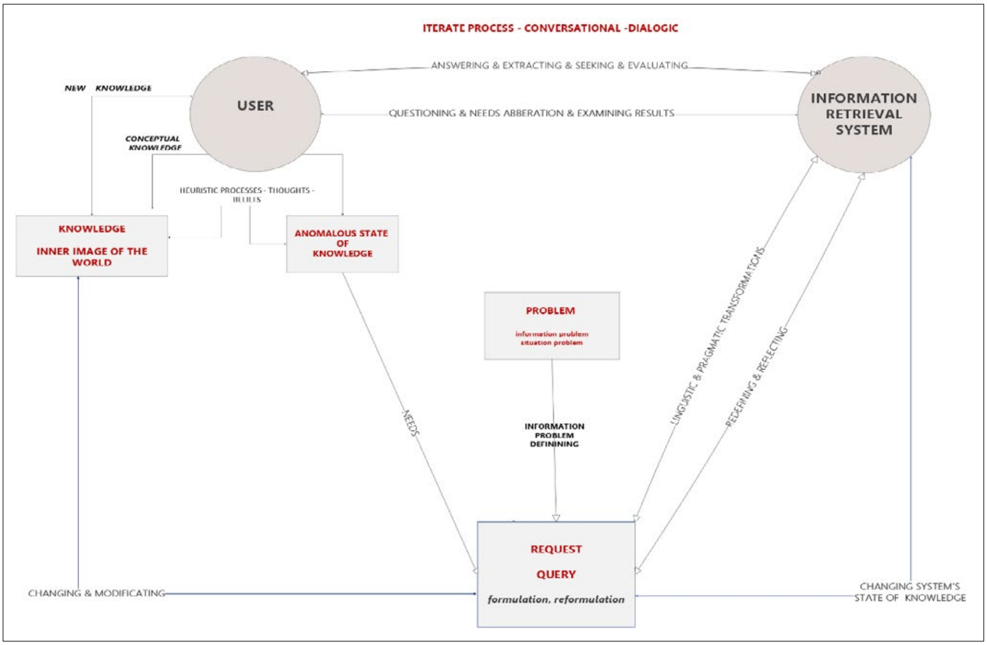

ITERATE PROCESS - CONVERSATIONAL -DIALOGIC
ANSWERING & EXTRACTING & SEEKING & EVALUATING
NEW KNOWLEDGE
USER
QUESTIONING & NEEDS ABBERATION & EXAMINING RESULTS
INFORMATION RETRIEVAL SYSTEM
CONCEPTUAL KNOWLEDGE
HEURISTIC PROCESSES - THOUGHTS - BELIEFS
KNOWLEDGE
INNER IMAGE OF THE WORLD
ANOMALOUS STATE OF KNOWLEDGE
PROBLEM
information problem
situation problem
LINGUISTIC & PRAGMATIC TRANSFORMATIONS
REDEFINING & REFLECTING
NEEDS
INFORMATION PROBLEM DEFININING
REQUEST
QUERY
formulation, reformulation
CHANGING & MODIFICATING
CHANGING SYSTEM'S STATE OF KNOWLEDGE

Fig. 1. Proposed model of information-seeking models
based on Nicholas Belkin & Garry Marchionini
(Belkin & Marchetti, Cole, 1993; Belkin et al., 1995; Marchionini, 1995).

Source: self-authored.

Information-seeking involves engaging with texts to interpret and solve problems, often within systems like IR systems. Prompts can guide this process, framing the user's query to help refine their search. Dialogic information seeking, where users and systems interact in an ongoing exchange, allows for dynamic prompts and responses that adapt to users' evolving needs, improving the efficiency of finding relevant information.

2.2. Prompts – definitions

In the context of generative artificial intelligence (GenAI), a prompt is a means of communication between a user and a Large Language Model (LLM) that enables the user to guide the model appropriately when generating a response (McNulty, 2024; Springs, 2024; Srinivasan, 2024). The message the user provides may be in the form of written text, an oral message, and visual and audiovisual objects. Similarly, the generated response may present itself in the form of linear written text or multimodal text.

GenAI, a branch of artificial intelligence (AI), is based on machine learning (ML) generative LLM models, a subset of artificial neural networks (ANN). These models generate content from user prompts (Sahoo et al., 2024; Zhang, 2024), with examples including ChatGPT, Midjourney, DALL-E, and Microsoft Copilot (Akakpo, 2024; Hassani & Silva, 2024; McNulty, 2024). Key terminology associated with GenAI includes "prompt engineering" (PE), "prompt tuning," and "prompt design." PE, the most general term, optimises user-AI communication, whether manual or automated (Huang et al., 2024; Mudadla, 2024), combining elements from AI, linguistics, and user experience (Lo, 2023). It guides language model predictions without altering model weights (Srinivasan, 2024) and focuses on designing prompts for optimal AI interpretation (Greyling, 2023; Lund, 2023). PE is considered an essential skill that merges language, logic, and creativity (Springs, 2024; Zhou, 2023), forming part of both AI and information literacy (Lund, 2023; McNulty, 2024). Some also regard PE as an art form (Frederick, 2024), with variations like "prompt answer engineering" (Huang et al., 2024) and "GPT engineering" (Springs, 2024) specific to ChatGPT prompts.

Prompt design, a key element of prompt engineering (PE), involves creating and optimising user instructions in natural language to elicit specific AI responses (Mudadla, 2024). PE also includes prompt tuning, which focuses on crafting prompt templates for particular tasks and learning selected parameters (Srinivasan, 2024). However, models using prompt tuning tend to be less stable and specialised in specific tasks with limited general knowledge (Shi et al., 2024; Spathis & Kawsar, 2024). "Prompt writing" outside the academic context involves creating inputs to guide AI outputs (McNulty, 2024). Additionally, "prompt pattern" outlines the structured components of a prompt for generating coherent and relevant text

(Marques et al., 2024), while “prompt template” is a static format where variables can be substituted (Greyling, 2023; Vogel, 2024).

2.3. Overview of how to write prompts: elements, writing style, typologies

The general rules for writing prompts are undergoing a period of adjustment as AI models and their capabilities continue to evolve. Some rules may appear logically inconsistent if viewed as a universal solution applicable to all AI models and tasks. For instance, negation, typically discouraged (Ruksha, 2024), is nevertheless observed in Negative Prompting (Aryani, 2023).

The typical components of a prompt include an agenda that provides task context (McNulty, 2024; Springs, 2024; Srinivasan, 2024), instructions describing the task (Springs, 2024), a trigger offering specific examples for the AI to develop (Springs, 2024; Srinivasan, 2024), and the format for the response, including handling exceptions (Springs, 2024; Srinivasan, 2024). This list is illustrative and not exhaustive, as no definitive set of prompt components exists. Frameworks like AUTOMAT, CLEAR, CO-STAR, and RICCE offer guidance on structuring prompts, with AUTOMAT being the most comprehensive. AUTOMAT includes seven elements: defining the AI’s role, audience, action, output format, style, handling exceptions, and setting topic boundaries (Vogel, 2024). Precision in prompt language is crucial, balancing clarity with flexibility to allow for AI creativity (Lynch et al., 2023; McNulty, 2024). Simplicity in language and breaking complex tasks into steps improve results (Srinivasan, 2024; Vogel, 2024). Avoiding biased terms and using proper punctuation further enhance the effectiveness of prompts (Patel, 2024; Warraich, 2024).

Academic and professional literature identifies various approaches to writing prompts, such as strategies, methods, techniques, patterns, templates, and formulas, though their organisation lacks consistency (Ruksha, 2024). In AI models like ChatGPT, communication can occur via a user interface or API (Sufi, 2024). Prompts are also categorised as hard or soft. Written in natural language, hard prompts are static and associated with prompt design, often using templates (Greyling, 2023). Soft prompts, generated through prompt tuning, adapt to data and the model, undergoing a learning process (Greyling, 2023; Zixuan Yi et al., 2024).

Furthermore, prompts can be classified according to the number of examples provided in the message. That is referred to as N-Shot prompting (Corral, 2023). This perspective distinguishes between Zero-Shot, One-Shot, and Few-Shot Prompting based on the number of examples provided. However, complex prompts may involve multiple variants at different stages of interaction with GenAI.

In addition to the aforementioned typologies, K. Ruksha’s (2024) proposal is worthy of note. It applies to the entire field of PE and distinguishes between (1) Single Prompt Techniques, which approach aims to obtain a maximally useful response with a single, optimised user query; (2) Multiple Prompt Techniques: this approach combines different prompts and assumes that a dialogue with the AI will be carried out through successive iterations; (3) the application of external communication tools in interaction with the LLM, including RAG and ReAct.

In the corpus of literature analysed, references to Chain of Thought (CoT) were most common (see Figure 2). Indeed, this is one of the most well-known ways, and it is combined especially with various variants of N-Shot-Prompting. Various advanced prompts and further variants of prompt writing are also based on CoT. Due to its versatility and popularity, CoT was used in the case study. For a more detailed description of the CoT in relation to the scientific literature search task and the information behaviour models analysed, please see the case study section.

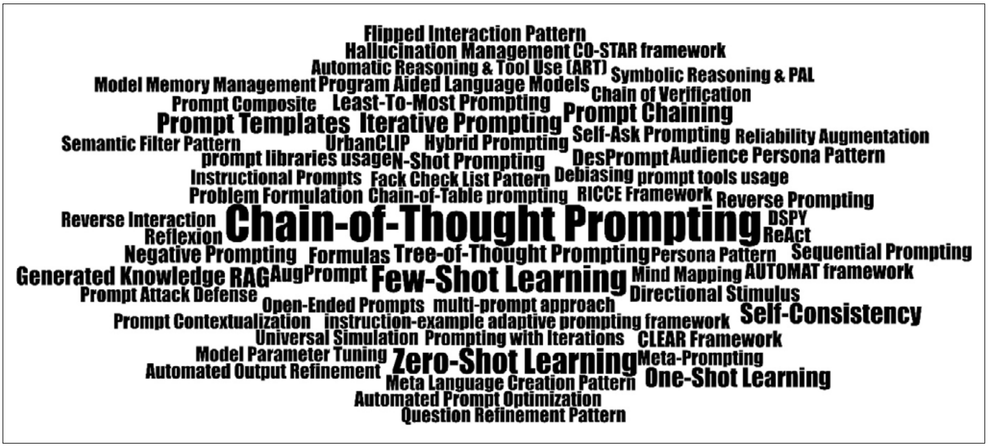


Fig. 2. MAXQDA code cloud showing specific ways to write prompts in PE.

Source: self-authored.

Given the numerous approaches to writing prompts, each with distinct characteristics in response generation, choosing one that aligns with the specific problem and the GenAI model's capabilities is essential. Possible GenAI prompts include: web page analysis (Hall & McKee, 2024), content classification (Chen, S. et al., 2024; Huang et al., 2024; Marques et al., 2024; Shi et al., 2024; Song et al., 2023), social media content creation (Hall & McKee, 2024), evaluation criteria development (Hall & McKee, 2024), data analysis and visualisation (Hall & McKee, 2024; Hassani & Silva, 2024; Sufi, 2024), learning materials development (Hall & McKee, 2024), forecasting (Hassani & Silva, 2024), idea generation and brainstorming (Hall & McKee, 2024), narrative generation (Lynch et al., 2023), synthetic dataset creation (Litake et al., 2024; Lynch et al., 2023; Sufi, 2024), keyword identification (Zhang, 2024), literature search assistance (Zhang, 2024), reasoning (Chen, S. et al., 2024), requirements elicitation (Marques et al., 2024), content review (Hall

& McKee, 2024; Marques et al., 2024), academic article hint searching (Zhang, 2024), sentiment analysis (Lynch et al., 2023), and programming code writing (Hall & McKee, 2024; Marques et al., 2024).

Concurrently, the academic literature identifies categories of problems for which GenAI models are considered inadvisable. Such categories of problems include, but are not limited to, the following: the analysis of private information or data (Hall & McKee, 2024); the analysis of current events (Hall & McKee, 2024); the citation of sources (Hall & McKee, 2024); and the analysis and visualisation of data. Furthermore, the analysis of complex data (Hall & McKee, 2024), as well as fact-checking (Hall & McKee, 2024), is problematic, as evidenced by the difficulty of answering tricky questions from the examples provided by Mariusz Pisarski (2024, p. 234). Some of the discouraged tasks are inconsistent with the list of problems that GenAI was used to solve. That is particularly relevant in the context of forecasting, which involves the analysis of current events and the visualisation of complex data. Furthermore, caution should be exercised when searching for scientific literature and citing sources due to the hallucinatory nature of GenAI models.

3. Research goals

The primary purpose of this article (A1) was to identify and describe the theoretical and practical foundations of PE in information science, focusing on well-established dialogic/conversational information retrieval models. The premise of this paper is that information science has long been concerned with issues currently being applied to the development of PE. A particular and most obvious area is human-machine interaction, mainly through dialogue/conversational systems. An additional aim (A2) was to identify possible further areas of commonality between information science and PE and, consequently, to identify well-established achievements in information science from which PE researchers and practitioners can benefit and other research fields for joint development within both disciplines.

Two principal research questions have been identified as particularly relevant to the aforementioned objectives. Research question RQ1 is linked to main objective A1 and consists of four extended subquestions (RQ1.1–RQ1.4). RQ2 is derived from the secondary research aim A2.

- RQ1. What are the relations between the dialogic/conversational models of information seeking well established in information science and the way prompts are written in PE?
  - RQ1.1. What are examples of dialogic/conversational information-seeking models established in information science? What elements do these models comprise? How does the human-machine conversation work in these models?

  - RQ1.2 What are the ways of writing prompts in PE? What elements can prompts consist of? What does it mean for a human to communicate with a GenAI model using prompts? What are the limitations of prompt communication with GenAI? In which problem situations are ChatGPT-type GenAI models used?
  - RQ1.3. What are the similarities between the well-established dialogic/conversational models of information seeking in information science and human communication with GenAI via prompts?
  - RQ1.4. What are the differences between the dialogic/conversational models of information seeking well established in information science and human communication with GenAI using prompts?
- RQ2. Beyond human-machine communication through dialogic/conversational models of information seeking/retrieval, are there common research areas between information science and PE? If so, what are these areas?

Given the ongoing development of GenAI models, the research questions relate to the current state of PE, understood as 2024 and the ChatGPT 4o model.

4. Methodology

The article presents the results of a qualitative strategy (Nowell et al., 2017) comprising a critical literature review (Cisek, 2010) and a case study.

A critical literature review was conducted to determine the extent of literature on PE and possible links to information science and information behaviour. In order to collect relevant academic and industry-related literature, a preliminary literature search was carried out on September 3rd 2024 to identify the topic area and obtain an initial list of keywords using Semantic Scholar (AI-based) and the Medium service (https://medium.com/). Keyword searches were used: a) in the case of Medium, the search was performed from within Google using an instruction consisting of a site command to search within the domain and the keyword prompt (site:medium.com prompt), and b) in Semantic Scholar, the term prompt was searched. 510000 results were extracted from the Semantic Scholar database, sorted by relevance, and the first 10 pages of the results table were examined. The Semantic Scholar material was only used to gain an initial understanding of PE terminology.

Subsequently, on September 8th, 2024, a search was conducted in the LISTA database to extend the search to information science publications. The search consisted of three phases. The first phase of the search used the phrase: (DE 'INFORMATION-seeking behaviour' OR DE 'INFORMATION needs' OR DE 'INFORMATION-seeking strategies' OR 'information seeking' OR 'information literacy' OR 'information behaviour' OR 'information seeking' OR 'information literacy' OR DE 'INFORMATION literacy') AND ('prompt engineering' OR 'prompting'

OR 'generative AI'), resulting in 26 publications. The search term 'prompt engineer*' was used in the second phase of the LISTA search, yielding 21 publications. In the third phase, the combination of the phrases 'prompt* tuning' OR 'prompt* method*' OR 'prompt* technique*' OR 'prompt design' OR 'prompt* pattern*' OR 'prompt AI' OR 'prompt artist*' yielded 14 results. A total of 59 results were included in the literature analysis and critique: a). from LISTA searches (within stages 1, 2, and 3 and after elimination of duplicates): 40 publications; b) from Medium: 19 professional articles.

All retrieved publications were then uploaded into MAXQDA. The use of MAXQDA streamlined the analysis and critique of the literature as it allowed for the efficient collection of sequential readings in a single environment and allowed for the highlighting and coding of content related to the subject of the study. The analysis sought to answer the following questions:

(1) How are prompts and related terms defined?
(2) How are prompts written (general principles, elements and types of prompts, limitations of PE)?
(3) Who is involved in writing prompts, and what competencies are associated with PE?
(4) What are the links between PE and information science?
(5) What are the areas of application of prompts outside information science?
(6) What tasks are PE used to solve?
(7) In what problem situations is PE not recommended?

These seven questions were assigned to the MAXQDA codebook (229 codes were obtained).

5. Case study: collecting scientific literature on a given topic (research on immersion in the virtual reality (VR) environment)

In the case study, a scientific literature search task was selected. The objective was to identify publications that were relevant to the problem situation. That is one of the tasks for which prompts have been employed (see, for example, Zhang, 2024). The challenge lies in both the complexity of the process of selecting appropriate literature that meets certain content and formal criteria and the limitations of LLM in the form of hallucination. In the case described here, the problem was related to scientific publications in immersion research in VR environments.

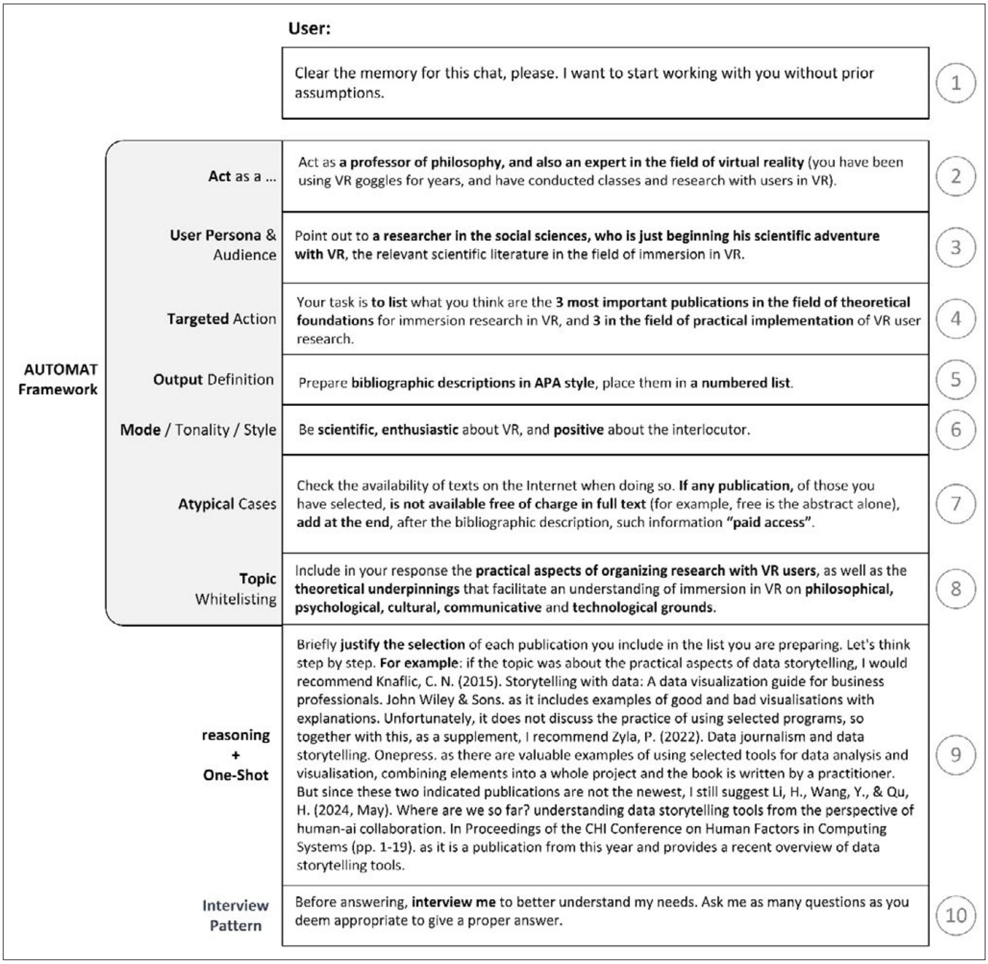
The conversation with ChatGPT (model 4o) was initiated by clearing memory (see excerpt one in Fig. 3). After consideration of the available options for formulating prompts, it was determined that the Multiple Prompt approach would be most suitable. It captures the complexity of the user's interaction with information systems by assuming that it is possible to converse with the GenAI model based on successive iterations, through which the essence of the information need – for both the user and the system – is more clearly captured.

The initial prompt emphasised the need to justify the selection of literature, prompting the use of Chain-of-Thought (CoT) to ensure the model provides a sequential rationale for each step (Ruksha, 2024). CoT can also be combined with multiple examples in instructional contexts, aiding in developing complex prompts. By clarifying the response formulation process, CoT enhances the transparency and coherence of LLM responses (Sahoo et al., 2024). Additionally, literature analysis (see Figure 2) shows that CoT is widely used for prompt writing. It is a key method for exploring links between modern prompt techniques and dialogic information retrieval models.

In the case study presented here, a request for justification was employed, along with an illustration of a potential approach to selecting subsequent publications. Additionally, a reminder was provided to act following the sequential execution of steps. Typically, the CoT is explained with relatively straightforward mathematical examples. However, the essence of the task at hand is more complex. Nevertheless, it can still be described with a CoT by breaking the task down into smaller scopes (identifying thematic areas such as theoretical foundations and practical considerations of the research (excerpt three in Fig. 3) and listing specific aspects to specify the topic (8)), presenting the expected way of justification and encouraging step-by-step thinking (9).

In order to represent the problem situation as accurately as possible, an AUTOMAT framework was employed (see excerpts 2-8). That allowed a detailed description of the thematic scope of the search to be provided, the user's initial competence to be defined (3), and a specific response format to be indicated (5, 7). Moreover, it was assumed that the information need may be dynamic and not fully conscious or explicit in the user's mind. Therefore, an approach was adopted to enable ChatGPT to facilitate the user in articulating the essence of the information need more effectively. Accordingly, an interview pattern (Cangiano, 2023) was employed, through which the GenAI model was instructed to conduct an information interview with the user and incorporate the extracted responses into the final information product generated by the AI. This approach was initiated with the command "interview me" (10). Consequently, ChatGPT guided the user with a series of questions to enhance comprehension of the problem situation (see excerpt 11 in Fig. 4).

In the dialogue shown in Figures 3 and 4, ChatGPT provided three outcomes with justifications and availability of data for "Theoretical Foundations" and "Practical Implementation of Virtual Reality (VR) User Research." No publications were fabricated, and the answers followed the specified outcome format (5). Figure 5 displays an example publication in APA format, with a numbered list and accessibility information. While the DOI link did not redirect to the full text, free access was confirmed via Google Scholar. A brief justification for the suggestion was included. Additionally, ChatGPT offered a section on “Practical Aspects for Organising Research with VR Users,” covering lab setup, think-aloud protocols, observations, surveys, and data recording. The full chat is accessible at: https://chatgpt.com/share/66fe483d-47b0-8009-8090-ca499a1a2ed9 (accessed October 4th, 2024).



① Clear the memory for this chat, please. I want to start working with you without prior assumptions.

② Act as **a professor of philosophy, and also an expert in the field of virtual reality** (you have been using VR goggles for years, and have conducted classes and research with users in VR).

③ Point out to **a researcher in the social sciences, who is just beginning his scientific adventure with VR**, the relevant scientific literature in the field of immersion in VR.

④ Your task is **to list** what you think are the **3 most important publications in the field of theoretical foundations** for immersion research in VR, and **3 in the field of practical implementation** of VR user research.

⑤ Prepare **bibliographic descriptions in APA style**, place them in **a numbered list**.

⑥ Be **scientific, enthusiastic** about VR, and **positive** about the interlocutor.

⑦ Check the availability of texts on the Internet when doing so. **If any publication,** of those you have selected, **is not available free of charge in full text** (for example, free is the abstract alone), **add at the end**, after the bibliographic description, such information **“paid access”**.

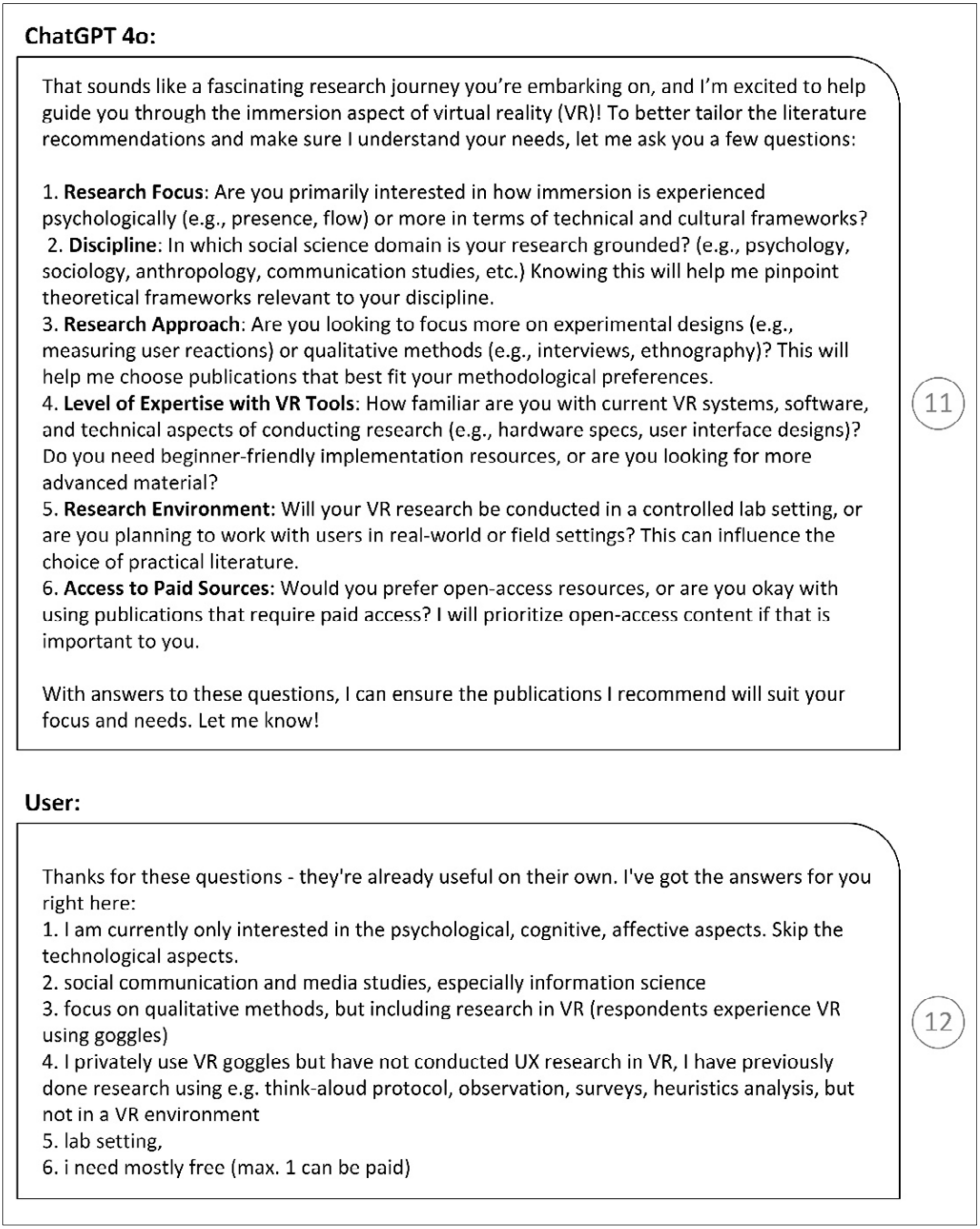
⑧ Include in your response the **practical aspects of organizing research with VR users**, as well as the **theoretical underpinnings** that facilitate an understanding of immersion in VR on **philosophical, psychological, cultural, communicative** and **technological grounds**.

⑨ Briefly **justify the selection** of each publication you include in the list you are preparing. Let's think step by step. **For example**: if the topic was about the practical aspects of data storytelling, I would recommend Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons. as it includes examples of good and bad visualisations with explanations. Unfortunately, it does not discuss the practice of using selected programs, so together with this, as a supplement, I recommend Zyla, P. (2022). Data journalism and data storytelling. Onepress. as there are valuable examples of using selected tools for data analysis and visualisation, combining elements into a whole project and the book is written by a practitioner. But since these two indicated publications are not the newest, I still suggest Li, H., Wang, Y., & Qu, H. (2024, May). Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-19). as it is a publication from this year and provides a recent overview of data storytelling tools.

⑩ Before answering, **interview me** to better understand my needs. Ask me as many questions as you deem appropriate to give a proper answer.

Fig. 3. Initial prompt structure (ChatGPT 4o).

Source: self-authored.



ChatGPT 4o:

That sounds like a fascinating research journey you're embarking on, and I'm excited to help guide you through the immersion aspect of virtual reality (VR)! To better tailor the literature recommendations and make sure I understand your needs, let me ask you a few questions:

1. **Research Focus**: Are you primarily interested in how immersion is experienced psychologically (e.g., presence, flow) or more in terms of technical and cultural frameworks?
2. **Discipline**: In which social science domain is your research grounded? (e.g., psychology, sociology, anthropology, communication studies, etc.) Knowing this will help me pinpoint theoretical frameworks relevant to your discipline.
3. **Research Approach**: Are you looking to focus more on experimental designs (e.g., measuring user reactions) or qualitative methods (e.g., interviews, ethnography)? This will help me choose publications that best fit your methodological preferences.
4. **Level of Expertise with VR Tools**: How familiar are you with current VR systems, software, and technical aspects of conducting research (e.g., hardware specs, user interface designs)? Do you need beginner-friendly implementation resources, or are you looking for more advanced material?
5. **Research Environment:** Will your VR research be conducted in a controlled lab setting, or are you planning to work with users in real-world or field settings? This can influence the choice of practical literature.
6. **Access to Paid Sources:** Would you prefer open-access resources, or are you okay with using publications that require paid access? I will prioritize open-access content if that is important to you.

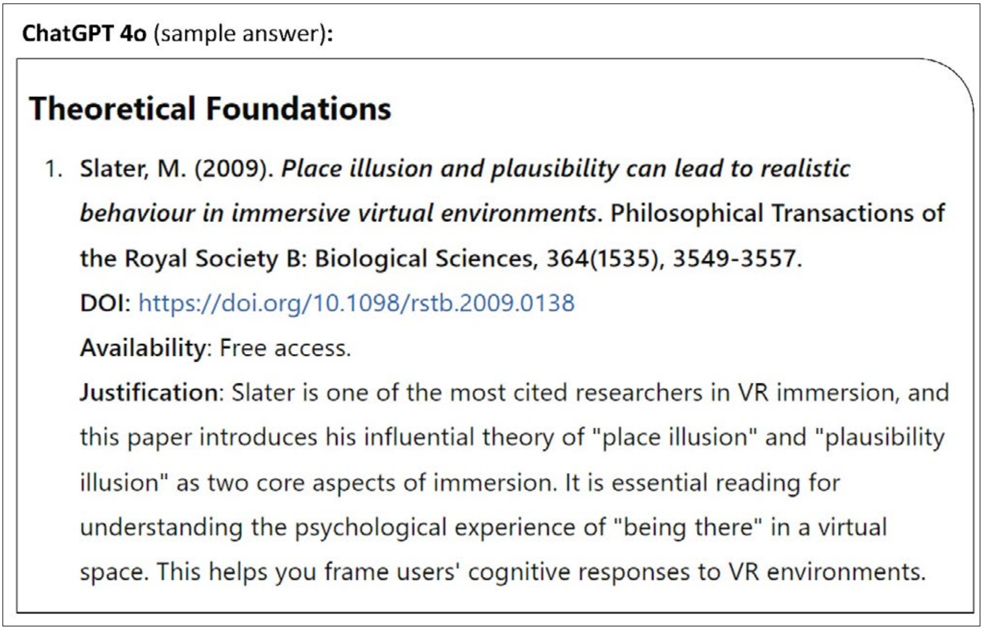
With answers to these questions, I can ensure the publications I recommend will suit your focus and needs. Let me know!

User:

Thanks for these questions - they're already useful on their own. I've got the answers for you right here:
1. I am currently only interested in the psychological, cognitive, affective aspects. Skip the technological aspects.
2. social communication and media studies, especially information science
3. focus on qualitative methods, but including research in VR (respondents experience VR using goggles)
4. I privately use VR goggles but have not conducted UX research in VR, I have previously done research using e.g. think-aloud protocol, observation, surveys, heuristics analysis, but not in a VR environment
5. lab setting,
6. i need mostly free (max. 1 can be paid)

Fig. 4. ChatGPT user interview (a follow-up, ChatGPT 4o).

Source: self-authored.



ChatGPT 4o (sample answer):

Theoretical Foundations

1. **Slater, M. (2009).** ***Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments*****. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3549-3557.**
   DOI: https://doi.org/10.1098/rstb.2009.0138
   Availability: Free access.
   Justification: Slater is one of the most cited researchers in VR immersion, and this paper introduces his influential theory of "place illusion" and "plausibility illusion" as two core aspects of immersion. It is essential reading for understanding the psychological experience of "being there" in a virtual space. This helps you frame users' cognitive responses to VR environments.

Fig. 5. A single publication record from the ChatGPT response (ChatGPT 4o).

Source: self-authored.

6. Discussion

Examining dialogic information-seeking models reveals apparent interconnections with established conversational information-seeking and retrieval frameworks (RQ1). The dialogic process conducted with ChatGPT, alongside using the CoT and AUTOMAT frameworks, highlights how reasoning and query refinement are iteratively managed. This process supports transparent interactions with the system and facilitates the retrieval of relevant responses. The steps taken by the system were clarified, including the formulation of answers and the breakdown of tasks and information problems into more detailed stages. This approach effectively addressed the user's knowledge gaps, allowing the concretisation of information needs and the creation of new knowledge. By applying filters and narrowing information retrieval pathways, queries were precisely specified, directing knowledge retrieval. The system's reasoning modes were clarified, resulting in a tailored interaction. The system encouraged ongoing dialogue, enabling users to submit reusable requests regarding queries, topics, and analytical approaches. That led to a flexible, multifaceted conversational process focused on information seeking and retrieval. The case study highlights the AI model's ability to deepen discussions and precisely redirect literature searches, as demonstrated in the Interview Pattern phase (excerpt 12 in Figure 4). It is crucial to verify the information provided by ChatGPT, as hallucinations may occur even if they were absent initially. Techniques like providing sources in a specified format with active links or DOIs have been implemented to aid in information verification. The completeness of ChatGPT's results is a separate issue and will be explored in future studies.

The AUTOMAT framework refers to Belkin's and Marchionini's models of information extraction processes (see Fig. 1) and incorporates virtually the same components and processes. It also highlights the numerous activities that the user and the system must undertake to clarify, redefine, and transform the user's information needs expressed through the questions asked, representing a continuous, iterative, planned and creative way of extracting relevant answers from the system. The request to clear memory during the start of the dialogue with the system (see excerpt one in Fig. 3) became the basis for removing previously accumulated knowledge, pre-remembered assumptions, the way the user usually formats the result, bridging the user's cognitive bias and leaving the generator's image of the world unconstrained by conditions previously suggested by the user. The heuristic processes undertaken during the recognition of the information need, the user's conceptual state of knowledge, highlighted by Nicholas Belkin and Garry Marchionini, were taken into account during stages 2 and 3 (see Fig. 3). The objective was to prompt the system to envisage the user and the circumstances under which the problem is to be solved, for example, by utilising virtual reality goggles. Subsequently, a persona was constructed, delineating the context of functioning and the performance of a particular task. This information constituted the problem statement. Clarifying the context and problem situation proved instrumental in defining the task correctly and concretely, as well as fostering awareness of the inherent complexity of the information need. That formed the basis for developing a reformulation of requests and queries per the AUTOMAT framework. The concretisation of the task to be performed by the system (see excerpts 4 and 5 in Fig. 3) facilitated the clarification of queries and the linguistic and pragmatic transformation of the user's thoughts and conceptual knowledge about the information need and information problem. The anthropomorphic format of the dialogic process of information retrieval in relation to the mode or style of the expected replica of the system (see excerpt 6 in Fig. 3) considered the processes of understanding how the user communicates with the machine and how the machine can also take into account its modes of communication. That allowed the construction of the affective and cognitive context of the situation. The "Atypical cases stage" (see excerpt seven in Fig. 3) correlates with the extracting, iterating and reflecting processes, whereby the user's information need is further specified, thus influencing the expected actions of the system in relation to the extracted results.

Part 8 of the framework (see Fig. 3) defines the complexity of the information problem and helps reframe queries based on the user's knowledge. A detailed description of the information need drives the system to perform the search. In the Reasoning + One-Shot trials (see excerpt nine in Fig. 3), methods for information retrieval were refined, especially given ChatGPT's limitations with negation. Users guide the system on how to explain, reason, and deduce, which aids in clarifying their knowledge, identifying gaps, controlling the retrieval process, and evaluating results. The concretisation and iterative formulation of queries during the conversation serve to comprehensively alter the knowledge of both the user and the system. This results in a change of state of knowledge, the creation of new knowledge, and new ways of information seeking.

The final phase of the prompt framework invites dialogue to detail the information problem and needs (see excerpt 10 in Fig. 3). The interaction between the user and ChatGPT refined the extraction of relevant results. ChatGPT's questions (see excerpt 11 in Fig. 4) helped adjust the information needs, clarify the problem's context, and guide the user in forming new queries. Questions about the research focus, approach, and user competence (persona) were crucial for understanding the issue and modifying the system's retrieval pattern. The responses (see Part 12 in Fig. 4) helped define the persona, search domain, and extract results. The iterative process of questioning, answering, and seeking, based on Belkin's and Marchionini's models, was integrated into prompt engineering formats like AUTOMAT, CoT, and Interview Pattern. These elements highlight the parallels between prompt engineering and established information-seeking models. The contemporary framework addresses need definition, query formulation, and reasoning refinement, incorporating cognitive processes into the AUTOMAT framework.

7. Conclusions and limitations

The analysis irrefutably revealed the components, modes of reasoning and conversation, processes and relationships between the dialogic/conversational models of information retrieval that are firmly established in information science and the way of writing prompts (RQ1). The selected models from information science are not recent, but they are indisputably well-established, described and used in further analyses. They provide a strong foundation for developing information retrieval models during prompt engineering. Furthermore, they align perfectly with existing, contemporary schemes and frameworks developed for writing prompts. This paper outlines the various dialogue models developed in information science and how a conversation is established between humans and computers (RQ1.1). We have identified the most important requirements, including prompt writing, limitations, and the meaning of this relation in human-computer communication, especially with GenAI and its application in ChatGPT (RQ1.2). The case study and conversation with a user demonstrated how contemporary information retrieval formats and selected models from the field of information science can be linked and how they differ (RQ1.3 and RQ 1.4). The models analysed and compared account for the anomalous state of knowledge, the iterative nature of the process of acquiring information, the dialogic nature of human-computer interaction, the refinement of the information problem by reformulating queries, their processing dynamics by the user and the system, and the creative nature of knowledge creation based on extracting, needs abating, and examining results.

It is important to acknowledge the limitations of the study. The study focused on specific models and methods of prompt writing. The rationale behind the selection is clear, but additional options could be considered in subsequent research projects. The same is true concerning the selection of the specific problem situation. It would be wrong to extrapolate the example from the case study to all potential scenarios related to information retrieval and using prompts. The description of the state of PE presented in the article pertains to the ChatGPT 4o model, which constituted the latest publicly available model at the time of the research. Subsequent GenAI models will undoubtedly offer new or enhanced functionality. This article has demonstrated the links between information science and PE in their current state of development. In light of potential future GenAI models, an update in the form of further research is necessary.

References

Akakpo, M. G. (2024). Skilled for the future: Information literacy for AI use by university students in Africa and the role of librarians. *Internet Reference Services Quarterly*, *28*(1), 19–26. doi: 10.1080/10875301.2023.2280566.

Aryani, A. (2023). 8 types of prompt engineering. *Medium,* [12.10.2024], https://medium.com/@amiraryani/8-types-of-prompt-engineering-5322fff77bdf.

Bates, M. E. (2024). Voices of the searchers. *Computers in Libraries*, *44*(2), 35–35.

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science, 50*(12), 1043–1050.

Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing & Management*, *26*(5), 575–591.

Bawden, D., & Robinson, L. (2011). Individual differences in information-related behaviour: What do we know about information styles? In: A. Spink, J. Heinström (eds.). *New Directions in Information Behaviour. Library and Information Science,* 1. London, UK: Emerald, 282–300.

Belkin, N. J., & Cool, C. (1993). The concept of information seeking strategies and its use in the design of information retrieval systems. In: *AAAI Spring Symposium on Case-Based Reasoning and Information Retrieval,* [12.10.2024], https://cdn.aaai.org/Symposia/Spring/1993/SS-93-07/SS93-07-001.pdf.

Belkin, N.J., Marchetti, P.G., & Cool, C. (1993). BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing & Management*, *29*(3), 324–344.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, *9*(3), 379-395.

Cangiano, A. (2023). *Prompt Engineering for Everyone*. Cognitive Class, [12.10.2024], https://cognitiveclass.ai/courses/prompt-engineering-for-everyone.

Case, D.O., & Given, L.M. (2016). *Looking for information. A survey of research on information seeking, needs, and behavior* (4th edition). Bingley: Emerald Group Publishing Limited.

Chen, J., Guo, F., Ren, Z., Li, M., & Ham, J. (2024). Effects of anthropomorphic design cues of chatbots on users' perception and visual behaviors. *International Journal of Human-Computer Interaction*, *40*(14), 3636-3654.

Chen, S., Li, Y., Lu, S., Van, H., Aerts, H. J. W. L., Savova, G. K., & Bitterman, D. S. (2024). Evaluating the ChatGPT family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, *31*(4), 940–948. doi: 10.1093/jamia/ocad256.

Cisek, S. (2010). Metoda analizy i krytyki piśmiennictwa w nauce o informacji i bibliotekoznawstwie w XXI wieku. *Przegląd Biblioteczny*, 78(3), 273–284, [12.10.2024], https://ruj.uj.edu.pl/xmlui/bitstream/handle/item/58573/cisek_metoda_analizy_i_krytyki_pismiennictwa_2010.pdf?sequence=1&isAllowed=y.

Cisek, S. (2017). Zachowania informacyjne. W: A. Żbikowska-Migoń, M. Skalska-Zlat (red.), *Encyklopedia książki*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, 2, 643–647.

Corral, M. J. (2023). Prompt patterns you should know. *Medium,* [12.10.2024], https://medium.com/@corraljrmiguel/21-prompt-patterns-you-should-know-636c931bba2a.

Ellis, D. (1992). The physical and cognitive paradigms in information retrieval research. *Journal of Documentatio*n, *48*(1), 45–64.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, *45*(3), 171–212.

Fidel, R. (2011). Approaches to investigating information interaction and behaviour. In: I. Ruthven, D. Kelly (eds.), *Interactive information seeking, behaviour and retrieval.* Facet Publishing, 61–76.

Ford, N. (2015). *Introduction to information behaviour.* Facet Publishing.

Frederick, D. E. (2024). Prompt engineering – a disruption in information seeking? *Library Hi Tech News, 41*(3), 1–5. doi: 10.1108/LHTN-03-2024-0037.

Greyling, C. (2023). Prompt tuning, hard prompts & soft prompts. *Medium,* [12.10.2024], https://cobusgreyling.medium.com/prompt-tuning-hard-prompts-soft-prompts-49740de6c64c.

Hall, B., & McKee, J. (2024). An early or somewhat late ChatGPT guide for librarians. *Journal of Business & Finance Librarianship, 29*(1), 58–69. doi: 10.1080/08963568.2024.2303944.

Halvorson, O. H. (2024). Innovation and responsibility: Librarians in an era of generative AI, inequality, and information overload. *SLIS Student Research Journal, 13*(2), 4–9.

Hassani, H., & Silva, E. S. (2024). Predictions from generative Artificial Intelligence models: towards a new benchmark in forecasting practice. *Information*, *15*(6), 291. doi: 10.3390/info15060291.

Hattatoglu, D. F. (2023). Prompt engineering formulas for ChatGPT and other language models. *Academy Team,* [12.10.2024], https://medium.com/academy-team/prompt-engineering-formulas-for-chatgpt-and-other-language-models-5de3a922356a.

Huang, W., Wang, J., & Cong, G. (2024). Zero-shot urban function inference with street view images by prompting a pre-trained vision-language model. *International Journal of Geographical Information Science*, *38*(7), 1414–1442. doi: 10.1080/13658816.2024.2347322.

Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham, [12.10.2024], http://peteringwersen.info/publications/0060_ingwersen_iri.pdf.

Krakowska, M. (2022). *Zachowania informacyjne człowieka w kontekście zjawiska epistemicznej bańki informacyjnej: propozycja nowej koncepcji*. Wydawnictwo Uniwersytetu Jagiellońskiego w Krakowie.

Lin, S.-J. (2017). Human-computer interaction research in information retrieval. In: J.D. McDonald, M. Levine-Clark (eds.), *Encyclopedia of Library and Information Sciences* (4th ed.). CRC Press. doi: 10.1081/E-ELIS4.

Litake, O., Park, B. H., Tully, J. L., & Gabriel, R. A. (2024). Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, *31*(6), 1404–1410. doi: 10.1093/jamia/ocae081.

Lo, L. S. (2023). The art and science of prompt engineering: A new literacy in the information age. *Internet Reference Services Quarterly*, *27*(4), 203–210. doi: 10.1080/10875301.2023.2227621.

Lund, B. (2023). The prompt engineering librarian. *Library Hi Tech News*, *40*(8), 6–8. doi: 10.1108/LHTN-10-2023-0189.

Lynch, C. J., Jensen, E. J., Zamponi, V., O'Brien, K., Frydenlund, E., & Gore, R. (2023). A structured narrative prompt for prompting narratives from Large Language Models: Sentiment assessment of ChatGPT-generated narratives and real tweets. *Future Internet*, *15*(12), 375. doi: 10.3390/fi15120375.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science, 40*(1), 54–66.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge University Press.

Marchionini, G. (2006). Toward human-computer information retrieval. *ASIST Bull.* June/July, [12.10.2024], https://asistdl.onlinelibrary.wiley.com/doi/10.1002/bult.2006.1720320508.

Marques, N., Silva, R. R., & Bernardino, J. (2024). Using ChatGPT in software requirements engineering: A comprehensive review. *Future Internet, 16*(6), 180. doi: 10.3390/fi16060180.

Materska, K. (2020). „Exploratory search" – kłopoty architekta informacji z terminologią. *Annales Universitatis Paedagogicae Cracoviensis Studia ad Bibliothecarum Scientiam Pertinentia, 18*, 213–233. doi: 10.24917/20811861.18.13.

McNulty, N. (2024). Prompt writing for generative AI: Mastering techniques. *Medium,* [12.10.2024], https://medium.com/@niall.mcnulty/prompt-writing-for-generative-ai-mastering-techniques-d2cea15ad754.

Mudadla, S. (2024). Difference between prompt design and prompt engineering? *Medium,* [12.10.2024], https://medium.com/@sujathamudadla1213/difference-between-prompt-design-and-prompt-engineering-2a90af3d1e2a.

Nicieja, K. (2023). Beyond prompts. *Medium.* https://medium.com/@kamil/beyond--prompts-1e7499fef910.

Nowell, L.S., Norris, J.M., White, D.E., & Moules, N.J. (2017). Thematic analysis: striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16,* 1–13.

Patel, K. (2024). Mastering prompt engineering with the RICCE Framework. *Medium,* [12.10.2024], https://medium.com/@keyur.shubham2014/mastering-prompt-engineering-with-the-ricce-framework-9ab53fc45b2f.

Pisarski, M. (2024). *Figury obecności w cyfrowych mediach: od hipertekstu do sztucznej inteligencji.* Universitas. ISBN 978-83-242-4006-7.

Próchnicka, M. (2004). *Człowiek i komputer: dialogowy model wyszukiwania informacji.* Wydawnictwo Uniwersytetu Jagiellońskiego.

Próchnicka, M. (2000). Modelowanie procesu wyszukiwania informacji. W: M. Kocój (red.), *Biblioteka i informacja w komunikowaniu: Jubileusz 25-lecia studiów Bibliotekoznawstwa i Informacji Naukowej w Uniwersytecie Jagiellońskim*, 116–124, [12.10.2024], https://ruj.uj.edu.pl/xmlui/handle/item/59041.

Próchnicka, M. (2001). Interfejsy dialogowe do systemów wyszukiwania informacji. *Zagadnienia Informacji Naukowej, 78*(2), 71–80, [12.10.2024], https://ruj.uj.edu.pl/xmlui/handle/item/65752.

Ruksha, K. (2024). Prompt engineering: classification of techniques and prompt tuning. *The Modern Scientist,* [12.10.2024], https://medium.com/the-modern-scientist/prompt-engineering-classification-of-techniques-and-prompt-tuning-6d4247b9b64c.

Sahoo, S. S., Plasek, J. M., Xu, H., Uzuner, Ö., Cohen, T., Yetisgen, M., Liu, H., Meystre, S., & Wang, Y. (2024). Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association, 31*(9), 2114–2124. doi: 10.1093/jamia/ocae074.

Savolainen, R. (2016). Information seeking and searching strategies as plans and patterns of action: A conceptual analysis. *Journal of Documentation, 72*(6), 1154–1180.

Savolainen, R. (2017). Information need as trigger and driver of information seeking: a conceptual analysis. *Aslib Journal of Information Management, 69*(1), 2–21.

Savolainen, R. (2019). Modeling the interplay of information seeking and information sharing: A conceptual analysis. *Aslib Journal of Information Management, 71*(4), 518–534.

Shi, S., Hu, K., Xie, J., Guo, Y., & Wu, H. (2024). Robust scientific text classification using prompt tuning based on data augmentation with L2 regularisation. *Information Processing & Management, 61*(1), N.PAG-N.PAG. doi: 10.1016/j.ipm.2023.103531.

Song, C., Cai, F., Zheng, J., Zhao, X., & Shao, T. (2023). AugPrompt: Knowledgeable augmented-trigger prompt for few-shot event classification. *Information Processing & Management, 60*(4), N.PAG-N.PAG. doi: 10.1016/j.ipm.2022.103153.

Spathis, D., & Kawsar, F. (2024). The first step is the hardest: Pitfalls of representing and tokenising temporal data for large language models. *Journal of the American Medical Informatics Association, 31*(9), 2151–2158. doi: 10.1093/jamia/ocae090.

Spink, A., & Cole, C. (ed.). (2006). *New directions in human information behavior*. Springer.

Spink, A., & Dee, C. (2007). Cognitive shifts related to interactive information retrieval. *Online Information Review, 31*(6), 845–860. doi: 10.1108/14684520710841801.

Springs. (2024). Prompt engineering: Examples and best practices. *Medium,* [12.10.2024], https://medium.com/@springs_apps/prompt-engineering-examples-and-best-practices-82b1da724643.

Srinivasan, K. (2024). What is prompt? What is prompt engineering & prompt tuning? *IceApple Tech Talks*, [12.10.2024], https://medium.com/iceapple-tech-talks/what-is-prompt-what-is-prompt-engineering-prompt-tuning-3fe628d97cf5.

Sufi, F. (2024). Addressing data scarcity in the medical domain: A GPT-based approach for synthetic data generation and feature extraction. *Information, 15*(5), 264. doi: 10.3390/info15050264.

Thani, R. A., & Hashim, L. (2011). Information needs and information seeking behaviors of social science graduate students in Malaysia. *International Journal of Business and Social Science, 2*(4), 137-143, [12.10.2024], https://ijbssnet.com/journals/Vol._2_No._4;_March_2011/16.pdf.

Vogel, M. (2024). The perfect prompt: A prompt engineering cheat sheet. *The Generator,* [12.10.2024], https://medium.com/the-generator/the-perfect-prompt-prompt-engineering-cheat-sheet-d0b9c62a2bba.

Warraich, S. (2024). Choosing the right prompt: A practical guide to different prompting styles. *Medium,* [12.10.2024], https://medium.com/@simrwaraich/choosing-the-right-prompt-a-practical-guide-to-different-prompting-styles-8e6d6a498a96.

Wen, Z., Cao, J., Yang, Y., Wang, H., Yang, R., & Liu, S. (2023). DesPrompt: Personality--descriptive prompt tuning for few-shot personality recognition. *Information Processing & Management, 60*(5), N.PAG-N.PAG. doi: 10.1016/j.ipm.2023.103422

White, R.W., & Roth, R.A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 1*(1), 1–98. doi: 10.2200/s00174ed1v01y200901i.

Yuan, C., Cao, Y., & Huang, H. (2023). Collective prompt tuning with relation inference for document-level relation extraction. *Information Processing & Management,* 60(5). doi: 10.1016/j.ipm.2023.103451.

Zhang, B. (2023). Prompt engineers or librarians? An exploration. *Medical Reference Services Quarterly, 42*(4), 381–386. doi: 10.1080/02763869.2023.2250680.

Zhang, L. (2024). Exploring generative AI with Chatgpt for possible applications in information literacy instruction. *Journal of Electronic Resources Librarianship, 36*(1), 64–69. doi: 10.1080/1941126X.2024.2306058.

Zhou, H., Li, M., Xiao, Y., Yang, H., & Zhang, R. (2024). LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction. *Journal of the American Medical Informatics Association, 31*(9), 2010–2018. doi: 10.1093/jamia/ocae147.

Zhou, Y. (2023). Prompt design patterns: mastering the art and science of prompt engineering. *Generative AI Revolution,* [12.10.2024], https://medium.com/

generative-ai-revolution-ai-native-transformation/prompt-design-patterns-mastering-the-art-and-science-of-prompt-engineering-d3c7eb659bac.

Zixuan Yi, Ounis, I., & Macdonald, C. (2024). Contrastive graph prompt-tuning for cross--domain recommendation. *ACM Transactions on Information Systems, 42*(2), 1–28. doi: 10.1145/3618298.

Sztuka prompt engineering jako stara/nowa forma dialogowego poszukiwania informacji z wykorzystaniem modeli sztucznej inteligencji

Abstrakt

Cel/Teza: Artykuł syntetyzuje teoretyczne i praktyczne rozważania na temat komunikacji dialogowej z sztuczną inteligencją, koncentrując się na uznanych modelach wyszukiwania informacji. Bada interdyscyplinarny charakter badań nad zachowaniami informacyjnymi oraz ewolucję modeli wyszukiwania.

Koncepcja/Metody badań: Zastosowano metodologię jakościową, obejmującą krytyczną analizę literatury oraz studium przypadku wykorzystujące ChatGPT do wyszukiwania literatury naukowej.

Wyniki i wnioski: Analiza ujawniła współzależności między tradycyjnymi a nowoczesnymi modelami, podkreślając poznawcze i eksploracyjne aspekty wyszukiwania informacji.

Ograniczenia badań: Skoncentrowano się na specyficznych modelach prompt engineering oraz jednym studium przypadku.

Zastosowania praktyczne: Zrozumienie uznanych modeli jest kluczowe dla rozwoju prompt engineering.

Oryginalność/Wartość poznawcza: Niniejsze badanie wypełnia lukę w badaniach nad integracją modeli wyszukiwania informacji z prompt engineering.

Słowa kluczowe:

ChatGPT. Konwersacyjne wyszukiwanie informacji. Model wyszukiwania informacji. Podejście dialogowe. Prompt engineering (PE). Sztuczna inteligencja (AI). Wyszukiwanie informacji.

MONIKA KRAKOWSKA is an assistant professor with a postdoctoral position at the Institute of Information Science, Jagiellonian University, Krakow, Poland. Her research interests include interdisciplinary approach in human information behaviour, user experience methodology, information experiencing in different contexts and spaces, including immersive ones. Selected publications: Doświadczanie informacji (information experience) jako kategoria zachowań informacyjnych człowieka: próba eksplikacji (2023), Zachowania informacyjne człowieka w kontekście zjawiska epistemicznej bańki informacyjnej: propozycja nowej koncepcji (2022), Diagnozowanie przestrzeni informacyjnej (information space) z perspektywy zachowań informacyjnych człowieka (2022).

MAGDALENA ZYCH, PhD, is a research assistant at the Institute of Information Studies, Jagiellonian University, Krakow, Poland. Her research focuses on user experience, data storytelling,

and methodology of information user research. Major publications: Big data-driven investigation into the maturity of library research data services (RDS) (Journal of Academic Librarianship, 2023), Education for business analysts in Poland (Education for Information, 2020), Przekaz symboliczny i podejście User Experience na przykładzie serwisów internetowych teatrów krakowskich (Zagadnienia Informacji Naukowej, 2017).

Contact details:
monika.krakowska@uj.edu.pl
magdalena.zych@uj.edu.pl
Uniwersytet Jagielloński
Wydział Zarządzania i Komunikacji Społecznej
Instytut Studiów Informacyjnych
ul. prof. Stanisława Łojasiewicza 4, 30-348 Kraków

Data scientists in the scientific literature: LDA topic modelling on the semantic scholar database

Len Krawczyk
ORCID: 0009-0001-5450-6475
College of Inter-area Individual Studies
in the Humanities and Social Science
University of Warsaw

Łukasz Iwasiński
ORCID: 0000-0003-2126-7735
Faculty of Journalism, Information and Book Studies
University of Warsaw

Mateusz Szymański
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

Abstract

Purpose/Thesis: This paper explores the representation of data scientists in scientific literature. It aims to answer the following questions: How has the number of publications on data scientists evolved over time? How are papers regarding data scientists distributed over different fields of study? In what context are data scientists represented in the scientific literature?
Approach/Methods: The authors used Latent Dirichlet Allocation (LDA) topic modelling to the resources available within the Semantic Scholar API.
Results and conclusions: There has been an increase in the number of publications on data scientists since 2008. A robust connection between data scientists and information technology, as well as biomedical research, was found. Little literature discusses data scientists in a sociocultural context.
Originality/Value: To our knowledge, no studies have been devoted to the representation of data scientists in scientific literature. The research may contribute to the conceptualisation of this notion.

Keywords:

Data Science. Latent Dirichlet Allocation. Semantic Scholar. Text Mining. Topic Modeling.

Text received on 14th of October 2024.

ZIN 2024, 62(2), 47–74
DOI: 10.36702/zin2024.02.09

1. Introduction

The rapid increase in the amount of data produced globally requires new forms of data management to derive value from it. Extracting knowledge from extensive databases demands skilled professionals capable of creating statistical models to uncover structured and unstructured data patterns. These professionals are commonly referred to as data scientists. However, due to the relative novelty of this phenomenon, the term 'data scientist' does not yet have a fixed definition (Hazzan et al., 2023). Given the rapid evolution of data science as a profession, definitions and roles continue to shift, reflecting its dynamic nature and widespread influence across diverse domains. Usually, data science is defined as a multidisciplinary field (Cleveland, 2001). Data scientists are typically proficient in applying statistical, analytical, and machine-learning techniques to draw insights from data (Donoho, 2017; Ho et al., 2019), often intending to create value in a commercial context (Reyes & Felipe, 2018).

In scientific literature, data scientists are primarily treated as a professional group (Espinoza & Gellegos, 2019). Efforts to define data scientists often involve analysing their skills and qualifications by examining quantitative data from various sources, such as job offers (Ho et al., 2019) and heterogeneous sources (Ismail & Zainal Abidin, 2016; Coelho Da Silveira et al., 2020). There is a scarcity of qualitative research on data scientists, although a few studies do exist (Pereira, Cunha, & Fernandes, 2020; Żulicki, 2022; Lowrie, 2017). Despite their undeniable impact on everyday life (Śledziewska & Włoch, 2020) and the broader scientific community (Hazzan & Mike, 2023), there is limited research on data scientists themselves outside the commercial context.

Big data is an essential factor not only in today's global economy but also in knowledge production (Krumholz, 2014; Priestley & McGrath, 2019). Data scientists wield powerful tools with uncertain implications (Boyd & Crawford, 2012) that have the potential to reshape the world. Therefore, we believe it is crucial to explore this topic further to understand better who shapes modern knowledge and how science reflects dynamic global changes. This paper aims to examine the representation of data scientists in scientific literature. It also strives to explore associations between data science and other fields of study, which may contribute to the conceptualisation of this term. To achieve this, we have employed a data scientist's toolkit, including text mining techniques and Latent Dirichlet Allocation (LDA) topic modelling, to analyse a vast repository of scholarly data accessible through the Semantic Scholar API. This approach leverages both computational power and theoretical insight, providing a robust framework for capturing and analysing the complex web of themes and relationships embedded within the literature on data science.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for large collections of discrete data, especially text corpora. In the original paper written by Blei et al. (2003) on Latent Dirichlet Allocation, the authors trained the model on several datasets for different purposes, including text corpora. The main text data sets used for training and evaluating LDA included, among others, scientific abstracts from the C. elegans community, which contained 5,225 documents. The use of scientific abstracts as part of the evaluation and demonstration of LDA's capabilities paper does illustrate its suitability for analysing scientific abstracts by uncovering latent themes or topics within a large collection of text documents. The advantage of using LDA is the fact that it is a powerful technique for unsupervised analysis, making it one of the most extensively used text-mining tools in research on scholarly data (Thakur & Kumar, 2022) and currently a recognised scientometric tool in library and information sciences (Lamba & Madhusudhan, 2018a; Lamba & Madhusudhan, 2018b; Miyata et al., 2020; Sugimoto et al., 2011). This has been reflected in numerous studies using this method for thematic clustering of scientific articles in multidisciplinary literature research (Anupriya & Karpagavalli, 2015; Griffiths & M. Steyvers, 2004) as well within domain-specific context such as information communication technologies (Lim & Maglio, 2018; Liu et al., 2016; Cortez et al., 2018; Chen, Wang & Lu, 2016), biomedical sciences (Ebrahimi, Dehghani & Makkizadeh, 2023; García et al., 2020; Yoon & Suh, 2019; Zou, 2018), management (Cho et al., 2017; Joo et al., 2018; Moro et al., 2015) environmental sciences (Chang et al., 2021; Dayeen et al., 2020; Jeon et al., 2018; Syed et al., 2018). It was used for the classification of scientific papers, as well as for finding patterns of rhetorical moves (Louvigne et al., 2013).

Using LDA to analyse data on the representation of data scientists in the scientific literature is a fitting approach, especially considering the multidisciplinary nature of data science, as it spans fields such as statistics, computer science, machine learning, business intelligence, and domain-specific applications like healthcare, finance, and social sciences. LDA is particularly well-suited to uncover hidden topics across large corpora of text, making it practical for identifying the diverse themes and sub-disciplines present in the literature that may not be immediately apparent through manual analysis.

Given that we utilise the Semantic Scholar API to access vast amounts of scholarly data, LDA's ability to handle large data sets makes it a suitable choice. With its broad scope and frequent updates, data science literature can be overwhelming to classify and interpret manually, but LDA allows for scalable and automated topic identification.

By applying LDA, we can discover latent themes that may connect data science to other fields of study. Doing so allows us to understand better how data science interacts with, influences, and is influenced by other fields. LDA for this type of analysis is appropriate because it leverages the model's strengths in identifying latent topics within large, multidisciplinary datasets. It captures data science's complex and interconnected nature, providing valuable insights into its influence, development, and conceptualisation in scholarly discourse. To our best knowledge, no such research has been published.

2. Research objectives and methodology

Our research questions are as follows:

- Q1. How has the number of publications on data scientists evolved over time?
- Q2. How are papers regarding data scientists distributed over different fields of study?
- Q3. In what context are data scientists represented in the scientific literature?

To answer these questions, we used text mining. It is a collection of techniques designed to recognise patterns within unstructured and semi-structured textual data. It aims to uncover previously undiscovered knowledge (Fan et al., 2006). Exploring patterns in the scientific literature often involves topic modelling. The fundamental concept behind topic modelling revolves around developing a probabilistic generative model for a collection of textual documents. In topic modelling, documents are conceived as blends of topics, where each topic represents a probability distribution across words (Thakur & Kumar, 2022). Our methodology involves performing topic modelling on scientific abstracts to identify topics that can be discerned within scientific literature on data scientists. Automated methods, of course, come with inherent limitations. An evident drawback is the lack of control over the quality of the data being analysed. The potential for incorporating unsuitable data into the analysed dataset is ever-present due to the nature of automated data extraction methods, particularly when dealing with extensively unstructured resources. To circumvent the complexities associated with data and feature extraction from online sources, we utilised the Semantic Scholar database, which can be accessed through the Semantic Scholar Academic Graph API (S2AG). The documents were also automatically gathered, but the architecture of Semantic Scholar facilitates further processing by design (Kinney et al., 2023).

Semantic Scholar is based on an advanced data processing system that consistently acquires documents and metadata from various sources. Semantic Scholar collaborates with over 50 publishers, data providers, and aggregators, integrating content from more than 500 academic journals, university presses, and scholarly societies worldwide. Notable partners include the Association for Computational Linguistics, ACM, arXiv, BioOne, bioRxiv, BMJ Journals, University of Chicago Press, CiteSeerX, Clinical Trials Transformation Initiative, DBLP, De Gruyter, Frontiers, HAL, HighWire, IEEE, Karger, medRxiv, Microsoft, Papers With Code, Project MUSE, PubMed, SAGE Publishing, Science, Scientific.Net, SciTePress, Springer Nature, SPIE, SSRN, Taylor & Francis Group, MIT Press, The Royal Society Publishing, Wiley, and Wolters Kluwer. These partnerships enhance the discoverability of scholarly content and provide valuable insights into how researchers engage with academic materials (Semantic Scholar, n.d.).

This system extracts text and metadata, standardises and clarifies details such as authors, institutions, and venues, categorises the subject area of each paper, produces a textual overview of its significant findings, and carries out additional functions. The Semantic Scholar database encompasses over 200 million articles, approximately 80 million authors, and around 550,000 publication venues (Kinney, 2023). This breadth of content renders the database extensive and provides comprehensive coverage of scientific resources.

We requested access to the Semantic Scholar API key. Although we were granted access to make up to 100 requests per second, downloading a dataset of the scale we were targeting – potentially up to 200 million entries – posed significant logistical and temporal challenges. Specifically, at this rate, it would take approximately 23 days of continuous, uninterrupted data requests to retrieve the entire corpus. This limitation highlights several practical issues, including the risk of network interruptions or API service limitations, which could lead to incomplete data collection or require retries, further extending the retrieval timeline. Moreover, handling such a large volume of data presents challenges regarding data storage capacity, processing power, and data management during analysis. Given these constraints, we focused on defining a more targeted dataset using specific keywords and limiting the number of entries retrieved.

With the vast volume of available literature, a focused keyword approach allowed us to create a manageable and thematically relevant corpus while preserving analytical depth. Our initial approach involved employing the keyword "data scientist" as a search query, as manual checks indicated that the volume of results for "data scientist" was the same as for "data scientists." We recognise that keyword dependence may inadvertently exclude some related studies. However, given that data science remains a relatively novel and niche topic, we decided to download the 10,000 most relevant entries for the keyword "data scientist" per year. We believe that the specificity of "data scientist" minimises ambiguity, enabling a more focused analysis aligned with the study's objectives. Thus, while keyword dependence may introduce bias, it also reveals valuable insights into the disciplinary contexts, research focus, and evolving engagement with data science across various fields. We gathered publications spanning from January 2005 to August 2023. The obtained database contained 188,066 entries for further analysis.

The following inclusion criteria for entries to the corpus were established:

(1) An entry must have a non-empty abstract.
(2) An entry must contain the phrase „data scientist" in the title or the abstract.
(3) An entry must be associated with a publication venue in some way – the

field venue or publication_venue has to be non-empty.

(4) An entry must be recognised as written in English.

Feature extraction of the data was provided by Semantic Scholar. We decided to select the following features for the analysis and filtering: paperId (an identifier of a paper), title, abstract, fieldsOfStudy, publicationTypes, publicationVenue (an identifier of a journal), venue (journal name) and year of publication. After filtering, duplicates were removed from the corpus.

For the analysis and processing, we used Python language with specific libraries. For text pre-processing and analysis, the following libraries were used:

(1) re for text cleaning,
(2) nltk for tokenisation and stop-words cleaning,
(3) spacy for lemmatisation,
(4) wordcloud for data visualisation,
(5) langdetect for language detection.

For LDA analysis, we used:

(1) re for text cleaning,
(2) nltk for tokenisation and stop-words cleaning,
(3) spacy for lemmatisation,
(4) wordcloud for data visualisation,
(5) langdetect for language detection.

The pipeline for analysis was taken in the following steps:

(1) Database acquisition from Semantic Scholar API (188 066 most relevant entries to keyword "data scientist").
(2) Filtering by inclusion criteria (1–3) mentioned above, performed on lowercased abstracts and titles in order to gather all relevant data (but the abstracts were saved with capitalisation for further analysis).
(3) Language identification and filtering out non-English publications.
(4) Lowercasing abstracts to avoid distinguishing words with the same meaning.
(5) Word filtering:
    a. Removing "-" in the middle of words to preserve words so they would not be treated as separate tokens.
    b. Removing one-character words.
    c. Removing numbers and special characters.
    d. Removing stop-words (most used words in English) to exclude words that occur most frequently and create unnecessary noise in the data.
(6) Lemmatization – aggregating various grammatical forms of a word to treat it as a single entity, denoted by the word's lemma or its base form as found in a dictionary.
(7) Removing extra stop-words (data, science, etc.) to eliminate highly common words often associated with data scientists, which could introduce unwanted noise.

(8) Counting total word occurrences to determine other potential stop-words and create a word cloud.
(9) Tokenization (splitting text into separate words).
(10) n-gram counting (n = 2,3) to uncover the most common bigrams and trigrams.
(11) Joining meaningful n-grams as a single token to preserve tokens with separate meanings, eg., machine learning – machinelearning.
(12) Token filtering to avoid noise in the data:
  a. Removing tokens that occurred less than five times in the corpus.
  b. Token has to occur in at least 3 unique documents.
  c. The token has to be longer than 3 characters or be included in a list of meaningful tokens (such as ml, nlp, or ai).
(13) Deleting duplicate entries from the database to avoid using the identical article metadata in the analysis more than once.
(14) Performing LDA with sklearn.
(15) Visualizing the results with pyLDAvis.
(16) Iterative experimenting with the number of clusters to collaboratively find the number most suitable for interpretation.
(17) Qualitative interpretation of the topics.

3. Results

Through the process of filtering the initial database by removing entries with empty abstracts or empty publication venues, we gathered a collection of 76,817 scholarly metadata files. We then applied an additional condition, requiring each entry to contain the phrase “data scientist” in either the abstract or the title, resulting in a refined database of 2,239 entries. As Semantic Scholar categorises its entries into fields of study, the distribution of documents within the corpus we created for this research is depicted in Table 1.

Table 1. Document count by field of study in relation to the volume of the whole corpus.

field of study	**# of docs with the phrase “data scientist”**	**total # of docs in the corpus**	**% of docs with the phrase “data scientist”**
Computer Science	1654	43034	3.84
Medicine	339	25355	1.34
Not Assigned	299	7031	4.25
Mathematics	91	3578	2.54
Engineering	86	2415	3.56
Sociology	47	1466	3.21

field of study	**# of docs with the phrase “data scientist”**	**total # of docs in the corpus**	**% of docs with the phrase “data scientist”**
Business	31	1295	2.39
Biology	29	5637	0.51
Psychology	29	2685	1.08
Political Science	25	1411	1.77
Physics	16	1229	1.30
Geography	14	1571	0.89
Geology	3	550	0.55
Economics	3	544	0.55
Art	2	87	2.30
Materials Science	1	375	0.27
History	1	264	0.38
Philosophy	0	78	0.00

Note: A paper may have one, none or multiple fields of study assigned. Source: self-authored.

We encountered the overrepresentation of Computer Science publications in our corpus. It was expected as it reflects the historical and disciplinary roots of data science, primarily anchored in computational and technical domains. Conversely, fewer publications in such fields of studies as Social Studies, Geology and Philosophy may stem from different terminologies or less frequent engagement with the explicit term “data scientist”.

A significant proportion of publications in our corpus belong to the domain of Computer Science (1654), comprising 3.84% of the total publications in this field. The field of Medicine follows in terms of the number of publications, though with a significantly lower document count (339). This trend is unsurprising, given the widespread integration of technical advancements and AI solutions in medical research and diagnostics (Lai et al., 2021). However, it is important to note that the database may exhibit a bias due to the potential underrepresentation of papers from other fields of study. This bias could be attributed to Semantic Scholar’s original focus as a database for computer science, geoscience, and neuroscience, which only expanded to include biomedical literature starting in 2017 (Fricke, 2018). Despite this, the proportion of entries containing the term “data scientist” in relation to the entire corpus remains relatively low (1.34%) compared to fields like Mathematics (91 publications/2.54%), which, despite having fewer total publications, demonstrates better relative representation. Engineering has 86 publications, representing 1.08%. Other fields exhibit very low numbers (<50) of abstracts or titles containing the phrase “data scientist”. This distribution of documents across various fields reveals a scarcity of research on data scientists within socio-economic and socio-cultural contexts. It highlights a significant research gap in this area.

Numbers of publications per year are displayed in Table 2. The first publication involving data scientists in the abstract or title was published in 2008. The number of publications started to rise gradually in 2012, which seems related to the beginning of "the era of big data" which started in this period (Floridi, 2014). Interest in the subject was the highest in 2020. In 2022, the number of publications noticeably decreased, but in August 2023, there were 237 publications, and it is reasonable to expect that it has risen by the end of the year.

Table 2. Number of publications per year.

Year	#
2008	1
2009	3
2010	2
2011	4
2012	11
2013	39
2014	52
2015	85
2016	133
2017	188
2018	243
2019	308
2020	356
2021	349
2022	255
2023	237

Source: self-authored.

By performing the steps mentioned in the "Research objectives and methodology" section, we acquired outcomes displaying word cloud visualising the 50 most frequent words (Figure 1), 25 of which are presented in Figure 2. As anticipated, the most prevalent term prior to lemmatisation and word filtration is "data", followed by "scientist". To enhance the relevance of the analysis, we decided to exclude these words by generating additional customised stop words.

Another step in the analysis was lemmatising the vocabulary and performing *n*-gram counting to uncover prevalent co-occurring phrases. This process offered valuable insights by helping to create supplementary tokens that encapsulate meaningful expressions. The most common bigrams often refer to specific phenomena, such as artificial intelligence or machine learning, and therefore, they should not be split into separate tokens. We created a list of such phrases as specific tokens. The results of *n*-gram counting are shown in Figures 3 to 5.

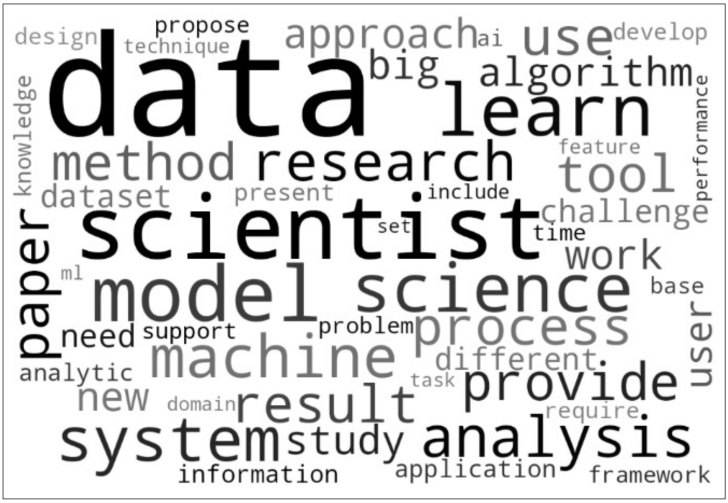


Figure 1. Word cloud of 50 most common words.

Source: self-authored.

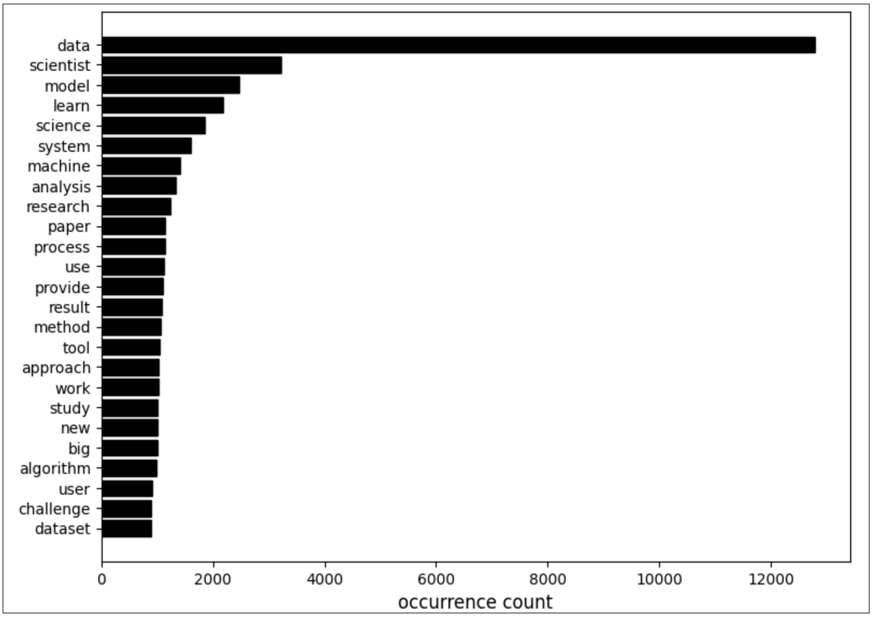


Figure 2. 25 most common words.

Source: self-authored.

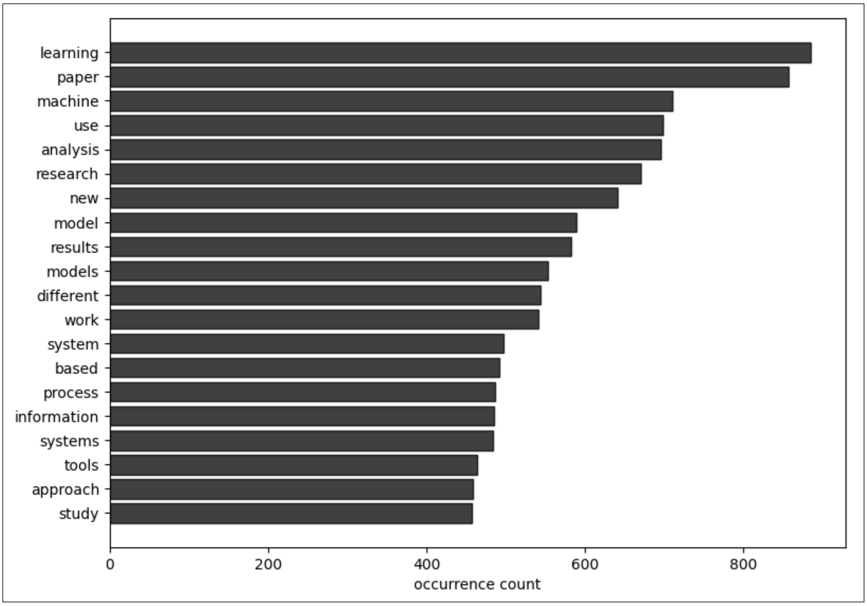


Figure 3. The 20 most common unigrams.

Source: self-authored.

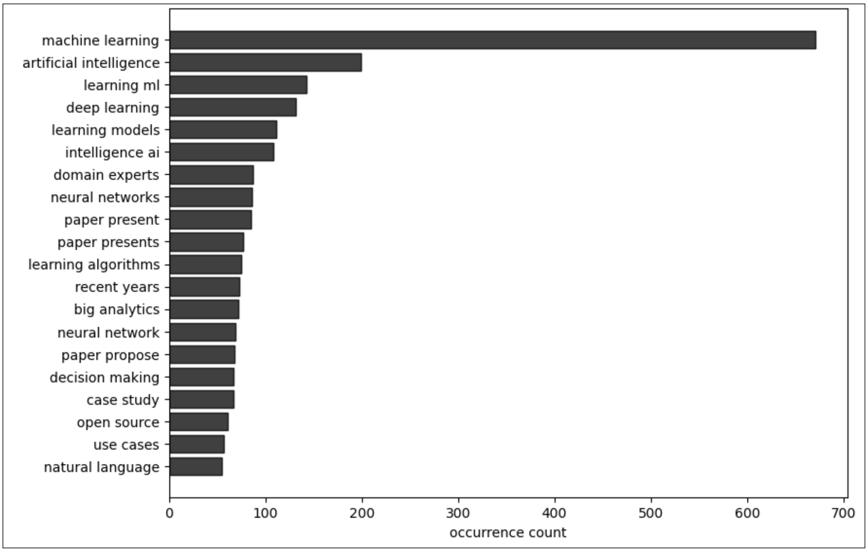


Figure 4. The 20 most common bigrams.

Note: Some bigrams include acronyms such as "AI" or "ML," which may not appear naturally in the text but are instead artefacts of automatic preprocessing.

Source: self-authored.

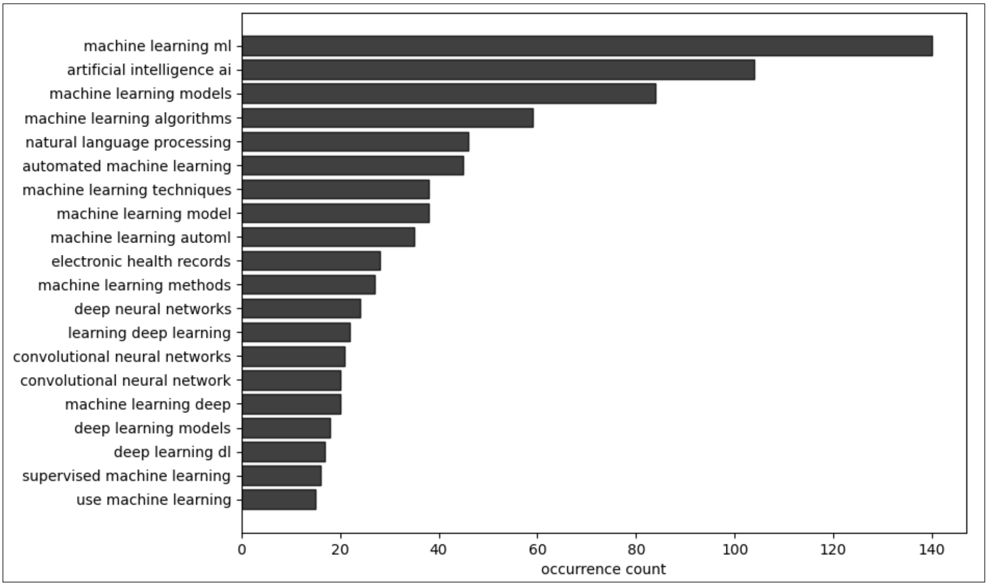


Figure 5. The 20 most common trigrams.

Note: Some trigrams include acronyms such as "AI" or "ML," which may not appear naturally in the text but are instead artefacts of automatic preprocessing.

Source: self-authored.

To make LDA work effectively, careful token filtering is necessary. We decided that a token should appear at least 5 times in the corpus to be worth considering. Also, for a token to be relevant, it has to appear in at least 3 documents, has to be longer than 3 characters or be part of the list of specific, meaningful short expressions, such as "ml" (*machine learning*), "ai" (*artificial intelligence*) or "nlp" (*natural language processing*). This approach aimed to filter out unnecessary acronyms while keeping the meaningful ones.

Following token filtering, we proceeded to perform Latent Dirichlet Allocation topic clustering. LDA is a generative probabilistic model that defines a topic as a distribution of words. Within this framework, each document in the corpus is a mixture of topics, and each topic is a mixture of words from the entire corpus vocabulary. More precisely, for each topic, a non-negative probability is assigned to each word from the vocabulary, and each document is a convex combination of topics (Blei et al., 2003).

The number of topics (clusters) is chosen arbitrarily. When selecting the optimal number of clusters or (topics) researchers have a range of quantitative and qualitative methods at their disposal, depending on the character of the problem. David Blei, author of the LDA original paper, states in another article, "The standard for selecting a solution is not so much accuracy as a *utility: Does the model* *simplify the data in a way that is interpretable, passes tests of internal and external validity, and is useful for further analysis?*" (Blei & Lafferty, 2009). This highlights that practical interpretability and usefulness should often take precedence over rigid accuracy metrics. Therefore, determining the optimal number of clusters for this type of study relies on the researcher's qualitative assessment rather than a prescribed heuristic (Wiedemann, 2016). However, interpretations must be approached cautiously, relying on subject-area specialists on the team (DiMaggio et al., 2013). Because we are a multidisciplinary team comprising three individuals with diverse backgrounds, including two researchers with experience in human and social sciences and a machine learning student with professional data science expertise, we adopted a process of collaborative, iterative experimentation to determine the number of clusters. Through this process, we arrived at a selection of 20 clusters, a decision that emerged as the most harmonious fit with the dataset's content, demonstrating a coherent and meaningful structure. This choice reflects both qualitative and data-driven considerations, ensuring a robust and insightful interpretation of the data.

The coverage *C(t)* of a topic *t* is defined as follows:

$$C(t) = \frac{\sum_d (d) \cdot p(t \mid d)}{\sum_{t'} \sum_d (d) \cdot p(t' \mid d)}$$

where $|d|$is a document length (in tokens) and $p(t|d)$ is a measure of assignment of a document d to a topic t. This measures how large a portion of documents in a corpus is captured by the topic. The LDA model allows for the adjustment of the term's relevance, which can help synthesise the idea behind a topic. Siever & Shirley (2014) defined the relevance as follows:

Let ϕ_{kw} denote the probability of term $w \in \{1, ..., V\}$ for topic $k \in \{1, ..., K\}$, where V denotes the number of terms in the vocabulary, and let p_w denote the marginal probability of term w in the corpus. The relevance of term w to topic k given a weight parameter λ(where $0 \leq \lambda \leq 1$) is defined as:

$$r(\mathrm{w}, \mathrm{k}, \lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log\left(\frac{\phi_{kw}}{p_w}\right)$$

where λ determines the weight given to the probability of term w under topic k relative to its lift, which is the ratio of a term's probability within a topic to its marginal probability across the corpus. Setting $\lambda = 1$ results in ranking terms in decreasing order of their topic-specific probability, and setting $\lambda = 0$ ranks terms by their lift (Siever & Shirley, 2014).

We decided to include results for $\lambda = 1$ and $\lambda = 0.5$. The value of represents a balance between words with a high probability of occurrence in the topic (which may also appear frequently in other topics) and words that are more distinctive to the chosen topic. This approach can be particularly advantageous when the words with the highest probability are overly general, making it challenging to uncover the underlying theme of a topic.

As a result of our experimentation with a number of clusters, we uncovered 20 clusters (topics), which we labelled and assigned to 5 different categories. Some topics fit more than one category. One topic (Graphical Data and Security) was excluded from the analysis because of non-coherent words and weak coverage (2.4%). Results of LDA topics modelling on 20 clusters and words λ =1 and λ =0.5 are presented in Table 3.

Table 3. LDA topics, relevant words and coverage in the corpus.

topic summary	**words with λ = 1**	**words with λ =0.5**	**coverage**
Big Data Analytics in Healthcare This topic is focused on big data research in healthcare, the application of AI models for knowledge discovery in medical contexts and its challenges.	big, research, clinical, health, analytics, analysis, methods, tools, ai, challenges, studies, use, information, knowledge, healthcare, researchers, including study, business, social	clinical, big, health, studies, research, analytics, healthcare, challenges, medicine, social media, care, including, researchers, methods, review, risk, tools, scientific, ai, artificial intelligence	11.6%
Systems, Databases and Scalability This topic is focused on large-scale systems, databases, and addresses issues related to scalability.	systems, query, system, analysis, processing, large, analytics, users, queries, time, graph, distributed, database, different, performance, python, algorithms, applications, user, big	query, queries, graph, distributed, processing, database, python, execution, systems, large, spark, languages, users, system, exploration, scalable, parallel, interactive, analytics, apache	11.1%
ML: Classification and Prediction This topic is focused on machine learning models, feature selection and engineering, prediction and classification.	models, machine learning, model, feature, prediction, algorithms, results, features, accuracy, dataset, learning, methods, system, process, different, predictive, datasets, ml, performance, classification	models, feature, prediction, accuracy, classifiers, machine learning, features, predictive, model, dataset, selection, algorithms, feature engineering, classification, classifier, regression, results, learning, random, schema	7.9%

topic summary	words with λ = 1	words with λ =0.5	coverage
ML: Automation and Pipelines This topic is focused on machine learning automation, pipeline development, and documentation.	ml, machine learning, automl, model, system, pipelines, process, pipeline, time, models, code, approach, systems, learning, automated, performance, documentation, solutions, tools, support	ml, automl, machine learning, pipelines, pipeline, documentation, automation, automated machine, hyperparameter, ml models, sales, automated, model, code, tuning, drilling, metalearning, cleaning, system, automate	6.1%
Privacy Preserving The topic is focused on privacy concerns, data analysis, and the application of technology to manage sensitive information.	esearch, privacy, social, big, analysis, information, paper, work, network, management, based, applications, storage, networks, access, technologies, methods, use, datasets, questions	privacy, social, network, big, research, storage, networks, tensor, information, differential, privacy preserving, journalists, sensitive, paper, work, secure, analysis, qualitative, management, topological	5.7%
Deep Learning and Image Classification The topic is focused on deep learning, including image classification, (convolutional) neural networks, and performance evaluation.	deeplearning, model, machine learning, analysis, images, classification, datasets, models, performance, accuracy, neural network, techniques, methods, based, image, paper, approaches, tasks, dataset, neural networks	deep learning, images, neural network, image, convolutional, classification, neural networks, segmentation, deep, datasets, accuracy, trained, speech, imaging, encoding, model, bias, machine learning, performance, chat gpt	5.6%
Algorithms and Statistical Methods The topic is focused on algorithms, mathematical tools and probabilistic analysis, including machine learning methods.	algorithm, model, learning, machine learning, results, paper, based, matrix, analysis, work, dataset, different, process, approach, study, time, experiments, techniques, methods, method	matrix, matrices, algorithm, markov, reduction, experiments, stochastic, kernel, breast, linear, probability, projection, india, scenario, educational, estimation, learning, summaries, transition, dimension	5.1%
COVID Pandemic The topic is focused on the COVID-19 pandemic, including health analysis and disease detection in the context of AI.	covid, pandemic, health, detection, different, study, people, mining, use, model, ai, work, analysis, public, approach, based, results, important, information, process	covid, pandemic, detection, professions, coronavirus, screening, healthy, health, pregnant, seizure, people, vaccine, said, spread, mining, infectious, population, chest, interventions, covidnet	4.8%

topic summary	words with λ = 1	words with λ =0.5	coverage
Education and Skill Development The topic is focused on education, skill development, research projects, and programs to enhance learning.	students, research, education, skills, learning, university, paper, course, project, training, information, programs, help, statistics, student, model, provide, need, article, fairness	students, education, course, university, student, skills, courses, curriculum, research, graduate, universities, project, programs, programme, teaching, fairness, college, institutions, learning, statistics	4.8%
Profession, Job Requirements and Roles The topic is focused on job requirements, technology utilisation, and the roles of engineers in creating technological solutions in companies.	ai, job, engineers, software, systems, design, technologies, development, companies, artificial intelligence, use, analysis, work, technology, need, requirements, big, research, roles, process	ai, job, engineers, software, companies, technologies, artificial intelligence, roles, design, designers, software engineering, fair, jobs, requirements, trust, systems, transport, technology, development, company	4.5%
Notebooks and Programming Methods The topic is focused on code notebooks, programming tools, and explainability in computational analysis.	notebooks, methods, research, notebook, programming, different, model, python, explainability, computational, jupyter, software, design, models, framework, development, code, explanations, systems, tools	notebooks, notebook, explainability, jupyter, programming, explanations, toolkit, python, serverless, serving, book, explanation, computational, adaptive, readers, methods, coding, metrics, software, pruning	4.5%
Quality Assessment, Effectiveness and Transparency The topic is focused on quality assessment in various applications, using algorithms and statistical methods in the context of challenges and transparency.	quality, different, systems, models, machine learning, use, algorithms, methods, model, challenges, approach, chapter, paper, time, statistical, real, transparency, solutions, problems, management	quality, chapter, trading, traffic, transparency, book, volatility, periodic, feminism, taxonomy, army, production, forecasts, concerns, real, discussed, road, regularisation, coherent, reader	4.4%

topic summary	words with λ = 1	words with λ =0.5	coverage
Tools for Business Analytics The topic is focused on tools and technologies' applications in business. It covers big data analytics, business needs, and collaboration within organisations.	tools, business, big, analytics, different, process, paper, analysis, research, need, organisations, work, challenges, learning, information, collaboration, article, social, use, knowledge	blockchain, business, tools, organisations, collaboration, unstructured, big, content, analytics, theories, organisational, competencies, centre, big analytics, shared, alternatives, face, article, spreadsheets, behaviour	4.4%
Healthcare Informatics and Patient Care The topic is focused on healthcare informatics, patient care, medical data, and digital solutions in the field.	healthcare, clinical, research, health, medical, care, informatics, patients, systems, information, system, big, patient, medicine, knowledge, development, computer, digital, group, team	healthcare, informatics, clinical, medical, care, patients, health, insurance, medicine, vehicle, group, national, radiation, patient, research, biomedical, nursing, translational, nurses, collaborations	3.7%
Health Information Management and Privacy The topic is focused on health information management, privacy concerns in healthcare systems.	health, information, results, management, privacy, development, model, framework, skills, research, digital, use, analysis, AI, design, patient, training, performance, platform, study	centres, phishing, privacy, residents, firm, health, digital, composition, patient, plant, management, skills, ehealth, radiology, aiml, physicians, twin, personalised, leadership, safety	3.7%
Stock Prices Forecasting The topic is focused on business analytics, forecasting stock prices, data visualisation, including predictive modelling.	model, training, models, based, business, machinelearning, paper, framework, digital, visualisation, process, stock, analytics, problems, approach, big, research, solutions, platforms, prediction	model, training, stock, prices, bidaml, lstm, oil, tweets, models, arima, forecasting, rmse, auto, firms, business, sports, geosparkviz, cnns, digital, timeseries	2.9%
Biomedicine: Cancer and Molecular Data Analysis The topic is focused on computational analysis of cancer, genes and other molecular data using computational frameworks.	code, framework, package, available, features, https, analysis, cancer, pipeline, cell, best, opensource, time, dataset, complex, networks, computational, graph, results, brain	package, cell, code, https, embeddings, cancer, molecular, brain, tpot, expression, cells, antipatterns, variants, genes, gene, framework, intensity, pvldb, pipeline, embedding	2.9%

topic summary	words with λ = 1	words with λ =0.5	coverage
Graphical Data and Security (Excluded) This topic was excluded from the analysis because of non-coherent words.	model, users, analysis, system, security, information, packages, function, common, paper, values, results, table, object, functions, user, ggplot, approach, graphical, code	ggplot, packages, graphical, security, plotting, table, cyber, object, columns, iris, function, scripts, graphics, plot, textual, cray, plugin, flood, values, plots	2.4%
Ethics The topic is focused on ethics, environmental impact, reporting, and the societal role of technology.	ethics, ethical, climate, field, statistics, impact, issues, society, different, reporting, big, work, biodiversity, computing, social, community, interdisciplinary, need, develop, cloud	ethics, climate, ethical, biodiversity, reporting, official, society, athletes, carbon, epistemic, impacts, accreditation, session, civiliser, literacy, exercise, whatif, mlai, heat, arise	2.2%
IoT, Devices and Malware Analysis The topic is focused on IoT, benchmark devices and malware analysis.	IoT, malware, programming, etal, based, use, visual, devices, statistical, interactive, metadata, users, time, scheme, need, provide, approach, dataset, user, prototype	malware, iot, scheme, labelling, etal, compliance, array, chart, visual, devices, recovery, prototype, spam, plotly, expressions, programming, benchmarks, citation, skill, metadata	2.2%

Source: self-authored.

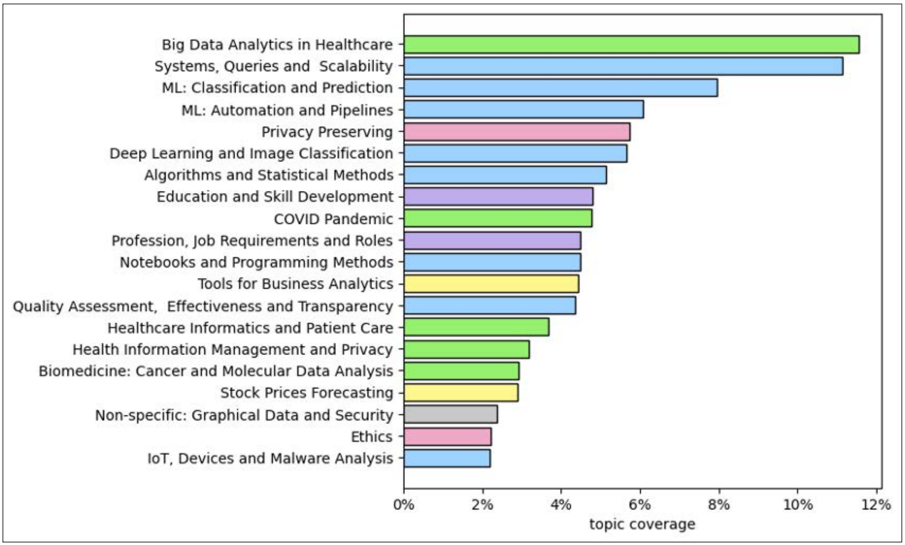


Figure 6. The coverage of topics.

Source: self-authored.

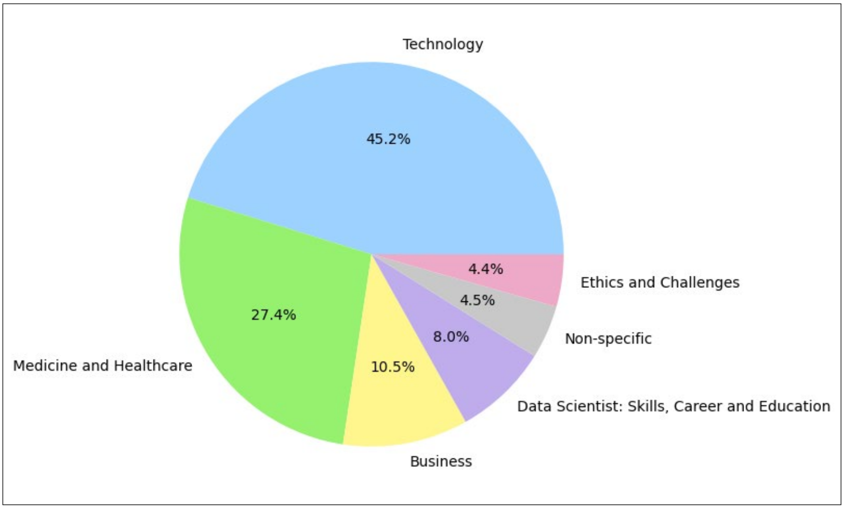


Figure 7. The coverage of topics grouped into categories.

Source: self-authored.

The topic categories are as follows:

(1) Information Technology (8/20 topics):
- Systems, Databases and Scalability,
- Machine Learning: Classification and Prediction,
- Machine Learning: Automation and Pipelines,
- Deep Learning and Image Classification,
- Algorithms and Statistical Methods,
- Notebooks and Programming Methods,
- Quality Assessment, Effectiveness and Transparency,
- IoT, Devices and Malware Analysis.

The Information Technology category covers 8 of the 20 topics and delves into various crucial technological aspects shaping data science. Systems, Databases, and Scalability relate to architecture issues essential to the work of data scientists. Topics such as Machine Learning: Classification and Prediction and Machine Learning: Automation and Pipelines highlight issues related to artificial intelligence. Topics like Deep Learning, Image Classification and Algorithms, and Statistical Methods explore the advanced techniques used in data analysis. Notebooks and Programming Methods emphasise the role of programming tools, while Quality Assessment, Effectiveness, and Transparency touch on issues related to data analysis and algorithms challenges. Lastly, IoT, Devices, and Malware Analysis showcases how data science extends into emerging fields, revealing its versatile applications.

(2) Medicine and Healthcare (Topics 5/20):
- Big Data Analytics in Healthcare,

  - COVID Pandemic,
  - Healthcare Informatics and Patient Care,
  - Health Information Management and Privacy,
  - Biomedicine: Cancer and Molecular Data Analysis.

The Medicine and Healthcare category, covering 5 out of the 20 topics, delves into critical aspects at the intersection of data science and healthcare. Big Data Analytics in Healthcare and Healthcare Informatics and Patient Care underscore how data-driven insights enhance healthcare delivery. Health Information Management and Privacy refer to the sensitive realm of protecting patients' data. The analysis uncovered that data scientists had an impact on the COVID-19 pandemic and vaccine development. The category also includes Biomedicine: Cancer and Molecular Data Analysis, spotlighting data scientists' contribution to understanding complex diseases and their diagnostics. This category showcases the profound influence data scientists have on improving healthcare outcomes. It also emphasises that data scientists make significant contributions to medicine-related research. The pivotal role of data science in biomedical research, facilitated by artificial intelligence tools, significantly enhances knowledge production and scientific advancement in this field (Lai et al., 2021). Nonetheless, this progress is accompanied by many ethical concerns and challenges, particularly privacy and security (Krumholz, 2014).

(3) Ethics and Challenges (Topics 4/20):
  - Ethics,
  - Privacy Preserving,
  - Health Information Management and Privacy,
  - Quality Assessment, Effectiveness and Transparency.

The Ethics and Challenges category, comprising 4 of the 20 topics, delves into essential dimensions of ethical considerations within data science. Topics like Ethics relate to ethical dilemmas that arise when handling data. Privacy Preserving underscores the importance of safeguarding individual privacy while extracting insights from data. Quality Assessment, Effectiveness, and Transparency also highlight the ongoing pursuit of maintaining data science practices' quality, effectiveness, and transparency. This category highlights data science practitioners' ethical and practical challenges in pursuing responsible and impactful data-driven decision-making supported by interpretable algorithms. This is coherent with numerous ethical issues addressed toward big data in various contexts, such as data management (Nair, 2020), health research (Rothstein, 2015) and privacy and security preservation (Joshi, 2020).

(4) Business (Topics 3/20):
  - Tools for Business Analytics,
  - Stock Prices Forecasting,
  - Profession, Job Requirements and Roles.

The Business category, encompassing 3 out of the 20 topics, delves into key aspects of data science within a business context. Topics such as Tools for Business Analytics stress the significance of data-driven tools in shaping business decisions. Stock Prices Forecasting explores the application of data science in predicting financial market trends. Additionally, Profession, Job Requirements, and Roles shed light on the evolving landscape of data science roles within the business sphere and their role in teams and collaboration with domain experts. This category showcases how data science is leveraged to inform business strategies, which aligns well with studies assessing the profound impact of data science on business (Mishra, 2021).

(5) Data Scientist: Skills, Career and Education (Topics 2/20):
   - Education and Skill Development,
   - Profession, Job Requirements and Roles.

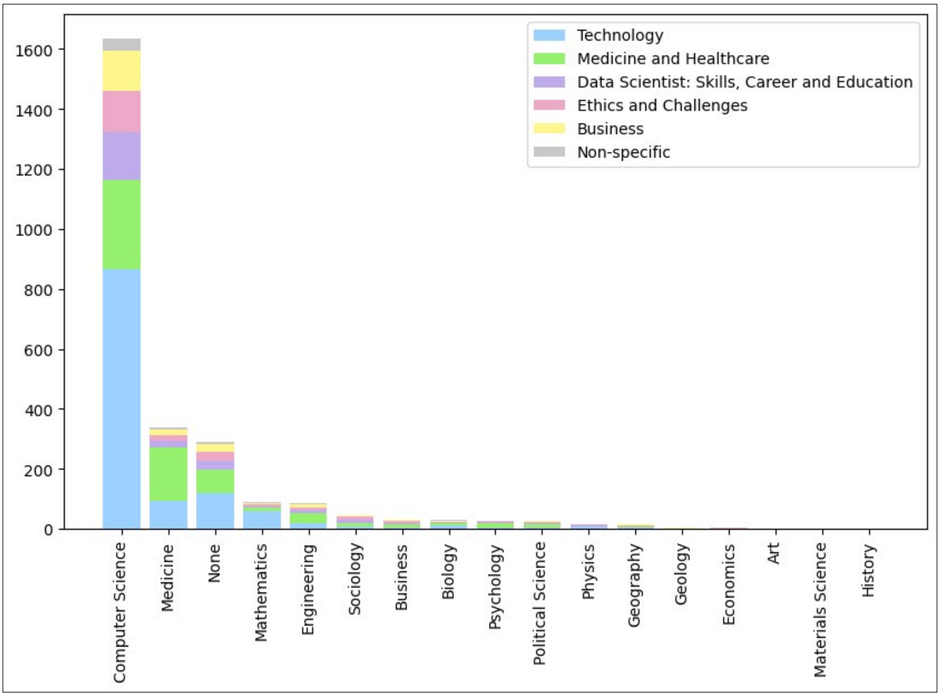


Figure 8: Topic categories in different fields of studies. Since each topic is a mixture of topics, the y-axis represents a weighted combination of the number of occurrences and topic shares.

Source: self-authored.

The Data Scientist: Skills, Career, and Education category encompasses 2 of 20 topics. Topics such as Education and Skill Development put on the spot the pathways and skills required for a data science career. Profession, Job Requirements, and Roles delve into the dynamic roles and evolving requirements within the data science profession. As was said before, professionals in this field significantly impact business, and employment in this field is rising (Mishra, 2021). This category offers insights directly into the social world of data scientists, regarding them as distinct subjects of study.

After grouping topics into categories, we investigated their overlap with the documents' fields of study as a form of external validation for the number of clusters. The results were satisfactory, as the distribution of topics reasonably matched the assigned fields of study. Topics related to Technology covered over half of the publications categorised under Computer Science, indicating strong alignment. Similarly, publications within the field of Medicine predominantly covered topics related to Health and Medicine. Furthermore, Technology-related topics showed substantial relative coverage in the field of Mathematics, whereas Engineering appeared to be more heavily influenced by topics related to Medicine and Health.

4. Study limitations

While extensive and comprehensive, using a corpus built from the Semantic Scholar database presents several limitations that must be considered when interpreting results. While Semantic Scholar extensively covers many disciplines, some fields or subfields may be less thoroughly represented. Moreover, recommendations and suggestions offered by Semantic Scholar based on artificial intelligence may reflect algorithmic biases, potentially favouring certain types of content.

One key limitation in constructing our database is keyword dependence, as the corpus relies on the search term "data scientist". This approach may inadvertently exclude relevant studies that use synonymous or related terms, thus introducing bias. This leads to challenges with generalizability, as findings may not extend well to broader or related areas of data science, particularly those using different terminology or less common phrases. Finally, evolving terminology poses a challenge, as the meaning and context of terms like "data scientist" have likely changed from 2005 to 2023, potentially affecting the interpretation of trends and roles captured in the corpus. These limitations underscore the need for caution and contextual awareness when analysing such data.

The method used for analysis also comes with limitations. The "bag of words" approach used in Latent Dirichlet Allocation has limitations due to its simplification of text data. This approach does not consider the order or arrangement of words in a document, which can impact the interpretation of topics and overlook nuances in meaning as it treats each word as independent of its neighbouring words, disregarding the inherent sequential or contextual information present in natural language. This can lead to a loss of meaning, as words' meanings often depend on their surrounding context. In this approach, each document is represented solely by the frequency of words, ignoring other valuable features like sentence structure, document length, or other text features.

There are several problems with analysing abstracts with LDA. Abstracts typically have limited text length, often consisting of only a few sentences or paragraphs. Due to this brevity, statistical methods like LDA can be susceptible to noise, resulting in accidental words or proper names being incorrectly attributed as highly relevant to a topic by the model. These terms may be coincidental or have low overall significance for the meaning of the entire document. Another consideration is that in cases where the corpus of abstracts is limited, rare or emerging topics might not have enough occurrences to generate coherent topics in LDA, resulting in their underrepresentation. Also, abstracts tend to follow specific language patterns, making them relatively homogeneous. This homogeneity can lead to LDA identifying topics aligned with generic scientific discourse rather than capturing more specific content.

To enhance the quality and relevance of the results, conducting a coverage comparison with alternative databases would be beneficial.

5. Summary

The analysis has shown a consistent upward trajectory in the number of publications centred on data scientists since 2008. The peak was observed in 2020, with the total number of publications being $n = 356$. With a minor regression observed in 2022 ($n = 255$), data scientists are still an area of interest in scientific literature, reaching 237 publications in August 2023.

A plethora of publications regarding data scientists reside within the domain of Computer Science ($n = 1654$). The second field is Medicine (339). A substantial portion of the corpus entries (299) lacks ascribed fields of study, constituting the third most prevalent category in the subject. Mathematics, Engineering, and other fields show a modest presence, while other disciplines exhibit minimal representation.

The study unveils an extensive landscape of literature that delves into the Information Technology category. Also, a distinct link between data scientists' works and the realm of biomedical research was found. This connection can be observed through various subfields, such as cancer genomics, patient data management, and the data-driven response to the COVID-19 pandemic. Moreover, a robust connection between data scientists and the business sector is evident. The documents within the corpus address an array of themes, ranging from data science applications in business intelligence to the roles of data scientists in teams in work environments. Ethical dilemmas and challenges arising from the proliferation of big data are prominently featured within the literature concerning data scientists. These discussions encompass concerns regarding data privacy preservation, spanning diverse contexts, including medical domains. Furthermore, code quality and algorithm transparency deliberations contribute to this ethical discourse. However, only two thematic categories make data scientists a central subject of study. They revolve around their professional roles and job requirements. There is also a thread regarding the courses and training for data scientists.

Generally, little literature discusses data scientists in a sociocultural context, with only a small number of publications within the field of Sociology ($n = 47$) and a lack of distinct topics on the subject. We consider this a striking gap in the literature because we believe it is important to study data scientists as social actors, given how much they shape knowledge and decision-making in various areas, such as medicine and business, as seen in this study. Another issue making data scientists an interesting subject in social sciences is the ethical implications of data collection, analysis, and use, which are critical in today's digital age. Sociology can explore how data scientists navigate ethical dilemmas related to privacy, consent, and bias, contributing to discussions on responsible data practices and regulations. Understanding their role can provide insights into how data-driven decisions shape societal dynamics and structures. They also create models that predict and explain human behaviour based on data patterns. Studying their methodologies can shed light on the underlying assumptions and biases that influence these models, thereby enhancing our understanding of how human behaviour is quantified and analysed.

References

Anupriya, P., & Karpagavalli, S. (2015). LDA-based topic modelling of journal abstracts. In *Proceedings of the 2015 International Conference on Advanced Computing and Communication Systems* (pp. 1–5). IEEE.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi: 10.1080/1369118X.2012.678878.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71–94). Taylor & Francis.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Cho, K.-W., Bae, S.-K., & Woo, Y.-W. (2017). Analysis of topic trends and topic modelling of KSHSM Journal Papers using text mining. *The Korean Journal of Health Service Management*, *11*(4), 213–224. doi: 10.12811/kshsm.2017.11.4.213.

Chen, J., Wang, T. T., & Lu, Q. (2016). THC-DAT: A document analysis tool based on topic hierarchy and context information. *Library Hi-Tech*, *34*, 64–86.

Coelho Da Silveira, C., Marcolin, C. B., Da Silva, M., & Domingos, J. C. (2020). What is a data scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias*, *8*(1), 25–39. doi: 10.5585/iptec.v8i1.17263.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69*(1), 21–26. doi: 10.1111/j.1751-5823.2001.tb00477.x.

Cortez, P., Moro, S., Rita, P., King, D., & Hall, J. (2018). Insights from a text mining survey on Expert Systems research from 2000 to 2016. *Expert Systems, 35*(3), e12280. doi: 10.1111/exsy.12280.

Cho, K.-W., Bae, S.-K., & Woo, Y.-W. (2017). Analysis of topic trends and topic modelling of KSHSM Journal Papers using text mining. *The Korean Journal of Health Service Management, 11*(4), 213–224. doi: 10.12811/kshsm.2017.11.4.213.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. doi: 10.1080/10618600.2017.1384734.

Dayeen, F. R., Sharma, A. S., & Derrible, S. (2020). A text mining analysis of the climate change literature in industrial ecology. *Journal of Industrial Ecology, 24*(2), 276–284. doi: 10.1111/jiec.12998.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modelling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics, 41*(6), 570–606. doi: 10.1016/j.poetic.2013.08.004.

Ebrahimi, F., Dehghani, M., & Makkizadeh, F. (2023). Analysis of Persian bioinformatics research with topic modelling. *BioMed Research International, 2023*(1), 3728131. doi: 10.1155/2023/3728131.

Espinoza Mina, M. A., & Gallegos Barzola, D. D. P. (2019). Data scientist: A systematic review of the literature. In M. Botto-Tobar, G. Pizarro, M. Zúñiga-Prieto, M. D'Armas, & M. Zúñiga Sánchez (eds.), *Technology Trends* (Vol. 895, pp. 476–487). Springer International Publishing.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49*(9), 76–82.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Fricke, S. (2018). Semantic scholar. *Journal of the Medical Library Association: JMLA, 106*(1).

García, D., Massucci, F. A., Mosca, A., Rafols, I., Rodrıguez, A., & Vassena, R. (2020). Mapping research in assisted reproduction worldwide. *Reproductive BioMedicine Online, 40*(1), 71–81. doi: 10.1016/j.rbmo.2019.10.013.

Hazzan, O., & Koby, M. (2023). Data science as a research method. In O. Hazzan & M. Koby (eds.), *Guide to Teaching Data Science* (pp. 121–135). Springer International Publishing.

Ho, A., Nguyen, A., Pafford, J. L., & Slater, R. (2019). A data science approach to defining a data scientist. *SMU Data Science Review, 2*(3).

Ismail, N. A., & Zainal Abidin, W. (2016). Data scientist skills. *IOSR Journal of Mobile Computing & Application, 3*(4), 52–61. doi: 10.9790/0050-03045261.

Joo, S., Choi, I., & Choi, N. (2018). Topic analysis of the research domain in knowledge organization: A latent Dirichlet allocation approach. *Knowledge Organization, 45*(2), 170–183. doi: 10.5771/0943-7444-2018-2-170.

Jeon, H. J., Kim, D. Y., Han, K. J., Han, D. W., Son, S. W., & Lee, C. M. (2018). An analysis of indoor environment research trends in Korea using topic modelling: Case study on abstracts from the journal of the Korean Society for Indoor Environment. *Journal of Odor and Indoor Environment*, *17*(4), 322–329. doi: 10.15250/joie.2018.17.4.322.

Joshi, M. V. (2020). Security/privacy issues and challenges in big data. *International Research Journal of Engineering and Technology*, *07*(06).

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., ... & Weld, D. S. (2023). The Semantic Scholar open data platform. *arXiv*. https://arxiv.org/pdf/2301.10140.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, *33*(7), 1163–1170.

Lai, Y., Kankanhalli, A., & Ong, D. (2021). Human-AI collaboration in healthcare: A review and research agenda. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (2021).

Lamba, M., & Madhusudhan, M. (2018a). Metadata tagging of library and information science theses: Shodhganga (2013-2017). In *Beyond the Boundaries of Rims and Oceans: Globalizing Knowledge with ETDs*.

Lamba, M., & Madhusudhan, M. (2018b). Application of topic mining and prediction modelling tools for library and information science journals. In M. R. Murali Prasad, A. Munigal, R. Nalik, M. Madhusudhan, & G. Surender Rao (eds.), *Library Practices in Digital Era* (pp. 395–401). BS Publications.

Lim, C., & Maglio, P. P. (2018). Data-driven understanding of smart service systems through text mining. *Service Science*, *10*(2), 154–180. doi: 10.1287/serv.2018.0208.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modelling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608. doi: 10.1186/s40064-016-3252-8.

Luna-Reyes, L. F. (2018). The search for the data scientist: Creating value from data. *ACM SIGCAS Computers and Society*, *47*(4), 12–16. doi: 10.1145/3243141.3243145.

Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society*, *4*(1), 2053951717700925. doi: 10.1177/2053951717700925.

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The human side of big data: Understanding the skills of the data scientist in education and industry. In *Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 503–512). IEEE.

Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: Topic modelling and visualization. *Scientometrics*, *125*(1), 665–687. doi: 10.1007/s11192-020-03657-5.

Nair, S. R. (2020). A review on ethical concerns in big data management. *International Journal of Big Data Management*, *1*(1), 8–25.

Pereira, P., Cunha, J., & Fernandes, J. P. (2020). On understanding data scientists. In *Proceedings of the 2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 1–5). IEEE.

Priestley, J. L., & McGrath, R. J. (2019). The evolution of data science: A new mode of knowledge production. *International Journal of Knowledge Management*, *15*(2), 97–109. doi: 10.4018/IJKM.2019040106.

Rothstein, M. A. (2015). Ethical issues in big data health research: Currents in contemporary bioethics. *Journal of Law, Medicine & Ethics*, *43*(2), 425–429.

Semantic Scholar. (n.d.). *Publisher partners* [online]. Retrieved from: https://www.semanticscholar.org/about/publishers [11.11.2024].

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70).

Śledziewska, K., & Włoch, R. (2020). *Gospodarka cyfrowa. Jak nowe technologie zmieniają świat.* Warsaw University Press. doi: 10.31338/uw.9788323541943.

Syed, S., Borit, M., & Spruit, M. (2018). Narrow lenses for capturing the complexity of fisheries: A topic analysis of fisheries science from 1990 to 2016. *Fish and Fisheries*, *19*(4), 643–661. doi: 10.1111/faf.12280.

Thakur, K., & Kumar, V. (2022). Application of text mining techniques on scholarly research articles: Methods and tools. *New Review of Academic Librarianship*, *28*(3), 279–302.

Wiedemann, G. (2016). *Text mining for qualitative data analysis in the social sciences*. Springer Vs.

Yoon, J. E., & Suh, C. J. (2019). Research trend analysis by using text-mining techniques on the convergence studies of AI and healthcare technologies. *Journal of Information Technology Services*, *18*(2), 123–141.

Żulicki, R. (2022). *Data science: Najseksowniejszy zawód XXI wieku w Polsce. Big data, sztuczna inteligencja i PowerPoint*. Wydawnictwo Uniwersytetu Łódzkiego.

Zou, C. (2018). Analyzing research trends on drug safety using topic modelling. *Expert Opinion on Drug Safety*, *17*(6), 629–636. doi: 10.1080/14740338.2018.1458838.

Data scientists w literaturze naukowej: modelowanie tematyczne LDA w bazie danych Semantic Scholar

Abstrakt

Cel/Teza: Niniejszy artykuł analizuje reprezentację *data scientists* (specjalistów ds. analizy danych) w literaturze naukowej. Celem jest odpowiedź na następujące pytania: Jak zmieniała się liczba publikacji na temat *data scientists* na przestrzeni lat? Jak publikacje dotyczące *data scientists* są rozproszone w różnych dziedzinach nauki? W jakim kontekście *data scientists* są przedstawiani w literaturze naukowej?
Koncepcja/Metody badań: Zastosowano modelowanie tematów metodą utajonej alokacji Dirichleta (LDA) do zasobów dostępnych w ramach API Semantic Scholar.
Wyniki i wnioski: Od 2008 roku obserwuje się wzrost liczby publikacji na temat *data scientists*. Odkryto silny związek pomiędzy *data scientists* a technologią informacyjną oraz badaniami biomedycznymi. Niewiele publikacji porusza temat *data scientists* w kontekście społeczno-kulturowym.
Oryginalność/Wartość poznawcza: Zgodnie z naszą wiedzą, dotychczas nie prowadzono badań poświęconych reprezentacji data scientists w literaturze naukowej. Przeprowadzone badanie może przyczynić się do konceptualizacji tego pojęcia.

Słowa kluczowe

Data science. Eksploracja tekstu. Modelowanie tematyczne. Semantic Scholar. Utajona alokacja Dirichleta.

LEN KRAWCZYK, osoba absolwencka kierunku filozofia nowych mediów na Uniwersytecie Śląskim, obecnie studiuje socjologię cyfrową na Uniwersytecie Warszawskim. Wiceprzewodnicząca Koła Naukowego Socjologii Cyfrowej. Stypendysta ministra za wybitne osiągnięcia naukowe. Publikowała w czasopiśmie Psychiatria Danubina.

ŁUKASZ IWASIŃSKI, absolwent kierunku Business and Technology w International Faculty of Engineering na Politechnice Łódzkiej (tytuł magistra inżyniera – 2006 r.) oraz socjologii na Uniwersytecie Łódzkim (tytuł magistra – 2007 r., stopień doktora – 2013 r.). Pracuje jako adiunkt w Katedrze Informatologii Wydziału Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Wykładał m.in. na Uniwersytecie Łódzkim (na kierunkach socjologia oraz dziennikarstwo i komunikacja społeczna), Politechnice Łódzkiej (na kierunku organizacja i zarządzanie), w Społecznej Akademii Nauk (w ramach programu MBA). Autor książki „Socjologiczne dyskursy o konsumpcji" (2016) oraz kilkudziesięciu rozdziałów w monografiach zbiorowych i artykułów z dziedziny socjologii, kulturoznawstwa i informatologii.

MATEUSZ SZYMAŃSKI, absolwent studiów magisterskich z matematyki na Uniwersytecie Śląskim, obecnie student kierunku machine learning na Uniwersytecie Warszawskim. Pracuje jako data scientist i programista Python. Publikował w czasopiśmie Psychiatria Danubina.

Contact details:
mm.krawczy10@student.uw.edu.pl
l.iwasinski@uw.edu.pl
Uniwersytet Warszawski
Wydział Dziennikarstwa Informacji i Bibliologii
ul. Nowy Świat 69, 00-927 Warszawa

The application of the ChatGPT language model for automatic generation of structured abstracts

Arkadiusz Pulikowski
ORCID: 0000-0003-1807-8642
Institute of Culture Studies
Faculty of Humanities
University of Silesia in Katowice

Abstract

Purpose/Thesis: The research aimed to evaluate the usefulness of the ChatGPT language model in generating structured abstracts for academic publications.
Approach/Methods: The methodology was qualitative. The study analysed ten articles from the journal Zagadnienia Informacji Naukowej—Studia Informacyjne (5 in Polish, 5 in English). ChatGPT version 4o was used to generate structured abstracts of the selected articles, then compared with the original abstracts to assess whether ChatGPT provided the required information in each section.
Results and conclusions: ChatGPT demonstrated strong capabilities in analysing and summarising documents to create abstracts for scientific publications in the field of information science. The language model performed well for both languages, with only two abstracts exhibiting significant issues in specific sections.
Originality/Value: The study showed the potential of language models, such as ChatGPT, in generating structured abstracts for bibliographic and full-text databases and as a complement to the researcher's workshop.

Keywords:

AI-generated abstracts. ChatGPT. Language models. Structured abstracts.

Text received on 15th of September 2024.

---

1. Introduction

The release of the ChatGPT tool by OpenAI in 2023 was undoubtedly a groundbreaking event, with an increasing influence on many areas of human activity. The capabilities offered by ChatGPT and similar language models were also quickly recognised within the scientific community. The number of publications employing artificial intelligence (AI) in research across various fields is proliferating. Document summarisation is among the model's key features in natural language

ZIN 2024, 62(2), 75–86
DOI: 10.36702/zin2024.02.10

processing. The author decided to leverage this capability to create structured abstracts automatically.

Structured abstracts consist of clearly labelled sections (e.g., Background, Purpose, Methods, Results, Conclusions), which help present information clearly and consistently to readers. While these components are also present in traditional abstracts, they are not explicitly labelled or organised similarly. Headings ensure authors follow a standardised format, reducing the risk of omitting essential elements. Key components such as the purpose, methods, results, and conclusions are generally expected in high-quality abstracts, whether traditional or structured. However, a structured format makes the abstract clearer, easier to read, and easier to search (Pulikowski, 2020, p. 25–26).

The research aims to evaluate the usefulness of the ChatGPT language model for automatically generating structured abstracts. A comparative analysis method was employed to verify the model's utility for abstract generation by comparing author-generated structured abstracts with those generated by ChatGPT. The study utilised research papers published in *Zagadnienia Informacji Naukowej – Studia Informacyjne* (*Issues of Information Science – Information Studies*).

2. Previous studies

Among the numerous publications discussing the benefits and risks of using artificial intelligence in scientific articles, a small but rapidly growing group focuses on abstract generation. A leading topic in this area is the comparison of abstracts generated by language models with original abstracts written by authors, specifically in terms of similarity and distinguishability (blind tests, automatic AI detection, linguistic accuracy, ethical considerations). Most of the publications come from the medical sciences.

The latest papers on the currentl versions of ChatGPT suggest that it has significant potential in generating scientific abstracts, with studies indicating varying levels of quality and accuracy compared to human-written counterparts. In a comparative analysis, original abstracts outperformed those generated by ChatGPT 3.5 and 4.0 in terms of quality; however, ChatGPT-generated abstracts were found to be more readable (Cheng et al., 2023; Hwang et al., 2024). Gravel et al. (2024) report similar findings, noting that while ChatGPT 4.0 does not produce abstracts of higher quality than those written by researchers, it is a valuable tool to help researchers improve the quality of their abstracts. Additionally, experienced reviewers needed help differentiating between AI-generated and human-written abstracts, indicating that ChatGPT can produce convincing content (Holland et al., 2024; Stadler et al., 2024). Despite some concerns regarding hallucinations and inaccuracies, ChatGPT's ability to summarise and generate concise abstracts suggests it could be a valuable tool in medical research (Hake et al., 2024).

3. Methodology

The study utilised publications in the field of information science, both in Polish and English, which appeared in the journal *Zagadnienia Informacji Naukowej – Studia Informacyjne* during the 2022–2023 period. A total of 10 research papers were selected – 5 in Polish and 5 in English. The purpose of juxtaposing publications in two languages was to assess whether ChatGPT would handle natural language processing equally well in both cases. Table 1 presents descriptions of the articles and assigned identifiers, which will be used primarily to discuss the study's results.

The selected articles' PDF files were downloaded from the journal's website (http://ojs.sbp.pl/index.php/zin), and the original abstracts were subsequently removed. The research employed the ChatGPT language model, version 4o (Omni), as it was the only model at the time of the study (early June 2024) capable of analysing attached text files. This unique functionality, combined with ChatGPT's widely recognised expertise in language processing, made it the optimal choice for the research.

Table 1. List of publications used in the study.

ID	Article title and author	ZIN No.
PL1	*Analiza struktury leksykalnej tytułów drapieżnych czasopism* Białka N.	2022, 60 (1)
PL2	*Tagowanie zdjęć portretowych w serwisie Instagram* Kosik N.	2022, 60 (1)
PL3	*Budowa i charakterystyka Korpusu Polskich Czasopism Naukoznawczych* Kulczycki E., Mena Y. A. Z., Krawczyk F.	2023, 61 (2)
PL4	*Walki informacyjne w paradygmacie ekosystemów informacyjnych* Materska K.	2023, 61 (1)
PL5	*Modele dojrzałości systemów informacyjnych na przykładzie bibliotek cyfrowych i serwisów danych badawczych* Nahotko M.	2022, 60 (1)
EN1	*Information literacy and information behaviour of disadvantaged people in the COVID-19 pandemic. Case study of beneficiaries of the charitable foundation* Kisilowska-Szurmińska M., Paul M., Piłatowicz K.	2023, 61 (1)
EN2	*Information technology maturity and acceptance models integration: the case of RDS* Nahotko M.	2023, 61 (1)

ID	Article title and author	ZIN No.
EN3	*Research on digital culture (cyberculture) – knowledge domain analysis based on bibliographic data from the Web of Science database* Osiński Z.	2023, 61 (1)
EN4	*Full-Text Search in the Resources of Polish Digital Libraries* Pulikowski A.	2022, 60 (2)
EN5	*How do early career researchers perceive success in their fields? Report on interviews with humanists, theologians, and scientists-artists in Poland* Świgoń M.	2023, 61 (2)

Source: self-authored.

The subject of the analysis was the abstracts generated by ChatGPT based on the attached files containing articles. For each document, ChatGPT was tasked with creating a structured abstract consisting of four sections, corresponding to the sections required by the journal's editorial board as mandatory: Purpose/Thesis, Approach/Methods, Results and Conclusions, and Originality/Value. The abstract generated by the language model was compared with the author's original abstract to determine whether it contained the expected information in each respective section. In cases of uncertainty, the full version of the article was consulted. That was often necessary, as ChatGPT frequently selected information different from the article's author for the individual sections of the abstract. The aim of the comparison was not, as in other studies, to assess whether the AI-generated abstract could be distinguished from a human-written one but rather to examine whether language models could be used to generate informative abstracts, particularly for bibliographic and full-text databases or reference management software. Under this assumption, grammatical and stylistic correctness is of secondary importance but remains relevant for ensuring the accurate and easy comprehension of the text.

The prompt for the articles in Polish was as follows:

> „Zapoznaj się z artykułem naukowym w załączonym pliku i na jego podstawie napisz abstrakt w języku polskim, nie dłuższy niż 200 słów, składający się z czterech akapitów zatytułowanych: Cel badań, Metody badań, Wyniki i wnioski z badań, Wartość poznawcza badań"

In turn, for the articles in English:

> „Read the scientific article in the attached file and, based on it, write an abstract in English, no longer than 200 words, consisting of four paragraphs entitled: Purpose of the research, Research methods, Results and conclusions of the research, Cognitive value of the research"

In both prompts, all labelled sections of the abstract included the word "research" (in Polish: "badań") to enhance the precision of ChatGPT's response. The 200-word limit for the generated abstracts was established based on the analysis of the length of the authors' abstracts. This data is presented in Table 2, together with the word count of abstracts generated by the language model. As can be observed, despite the clearly specified word limit in the prompt, ChatGPT exceeded the 200-word threshold in three cases: PL4, EN3, and EN4.

Table 2. Number of words in authors' abstracts and generated by ChatGPT.

	PL1	PL2	PL3	PL4	PL5	EN1	EN2	EN3	EN4	EN5
Author	177	117	102	109	116	200	229	119	199	157
ChatGPT	185	195	153	203	141	198	173	203	237	142

Source: self-authored.

The prompts for individual articles were entered in new chat sessions to ensure they were not interpreted as related within a single thread. The context was specified within ChatGPT's custom instructions settings to tailor the responses better. In the section "What should ChatGPT know about you to provide better responses?" the following was entered: "I am a university professor. My field of expertise is information science". Meanwhile, in the section "How would you like ChatGPT to respond?" it was specified: "I prefer concise responses written in scientific language".

4. Study results

The research results were positively surprising in only two cases. ChatGPT generated abstract sections where the content was incorrect or significantly insufficient (PL5 – purpose, EN3 – results). No significant differences were observed in the informational content between abstracts generated for publications in Polish and English. ChatGPT performed equally well in both cases. It is also worth noting that none of the generated abstracts contained fabricated information, a phenomenon known as "hallucination". The research results are presented in Table 3.

Table 3. Correctness of individual sections of abstracts generated by ChatGPT.

	PL1	PL2	PL3	PL4	PL5	EN1	EN2	EN3	EN4	EN5
Purpose	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
Methods	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Results	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Value	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Source: self-authored.

In some abstracts, sections generated by the language model provided more detailed information than those written by the authors of the publications. That occurred most often when the description generated by ChatGPT was longer than the author's version. However, this issue needs to be revised for a definitive evaluation, and thus, the author decided not to include it in the research results. It is much more feasible to determine whether each abstract section contains the appropriate information than to evaluate which of the two correctly formulated sections is superior.

Apart from the informational content, it is also important to consider the language of the abstracts generated by ChatGPT. The texts were written clearly and comprehensibly, employing a typical academic style. The stylistic and grammatical correctness of the abstracts is of a high standard. That can be easily assessed in the following example for the EN5 article. A comparison of all author-written and language model-generated abstracts used in the study has been made available in the Zenodo repository (Pulikowski, 2024).

EN5. How do early career researchers perceive success in their fields? Report on interviews with humanists, theologians, and scientists-artists in Poland [Marzena Świgoń, ZIN 2023, 61 (2)]

Purpose/Thesis

Author: This paper presents the Polish researchers' opinions on success perception in science and examines the role of bibliometric and altmetric indices/metrics. It also provides respondents' statements on how their work is evaluated at Polish universities.

ChatGPT: This research explores how early career researchers in the humanities, theology, and arts in Poland perceive success. It examines factors associated with reputation and success, including bibliometric and altmetric indicators, aiming to address gaps in previous studies focused on sciences and social sciences.

Approach/Methods

Author: The data come from in-depth interviews (spring 2023) with early career researchers in Linguistics, Literary Studies, History, Philosophy, Polish studies, Theology, Arts, Music, Plastic/Fine Arts, and Film.

ChatGPT: The study uses longitudinal interviews, with the first round in spring 2023. Interviews were partly structured, recorded, and transcribed for thematic analysis. The responses were manually coded using a heuristic approach and a standardised thematic framework.

Results and conclusions

Author: The respondents' attachment to indices used in the periodical evaluation of scholars at Polish universities, such as the number of publications, ministerial score points, and external grants, was observed. The respondents were critical of citations as the determinant of success and generally not very interested in altmetrics.

ChatGPT: Early career researchers prioritise academic publications, grants, international cooperation, and recognizability. Citations and altmetrics are relevant but less significant. Arts researchers emphasise non-numerical success measures more than those in humanities and theology.

Originality/Value

Author: The importance of this research (National Science Centre grant No 2022/45/B/HS2/00041), which is the Polish continuation of the two international projects (2016–2022), lies mainly in providing a platform to researchers in the arts and humanities, which were not included in earlier studies devoted to the sciences and social sciences.

ChatGPT: This research highlights the unique challenges of early career researchers in humanities, theology, and arts in Poland. It emphasises diverse success indicators beyond bibliometrics, informing policies to support their development and recognition.

Since ChatGPT performed very well in processing Polish while generating abstracts for publications in this language, an additional test was conducted to assess how it would handle generating Polish abstracts based on English articles. This capability could be particularly useful for users of bibliographic or full-text databases. To evaluate ChatGPT's performance in this context, the research was repeated for English publications (EN1–EN5), using the prompt designed for Polish publications. The results presented in Table 4 show that changing the language of the generated abstract did not affect its accuracy. It can be assumed that ChatGPT may be equally effective in many other languages it supports. However, the level of support for those languages may vary depending on linguistic complexity and data availability, making this an assumption that requires further verification.

Table 4. Correctness of Polish abstracts generated
by ChatGPT based on English publications

	EN1	EN2	EN3	EN4	EN5
Purpose	✓	✗	✓	✓	✓
Methods	✓	✓	✓	✓	✓
Results	✓	✓	✓	✓	✓
Value	✓	✓	✓	✓	✓

Source: self-authored.

5. Study limitations

When analysing the presented results, it is important to consider the research's limitations. First and foremost, it should be noted that ChatGPT exhibits significant variability in the responses it generates. The answers are produced dynamically, incorporating an element of randomness, which means that even when the same question is repeated, the response may be formulated slightly differently – similar in information content, but potentially better or worse. The study shows that the response will still be correct in most cases. Additionally, the model evolves, continuously improving, with new and significantly modified versions being released periodically. All of these factors contribute to a variable and dynamic environment.

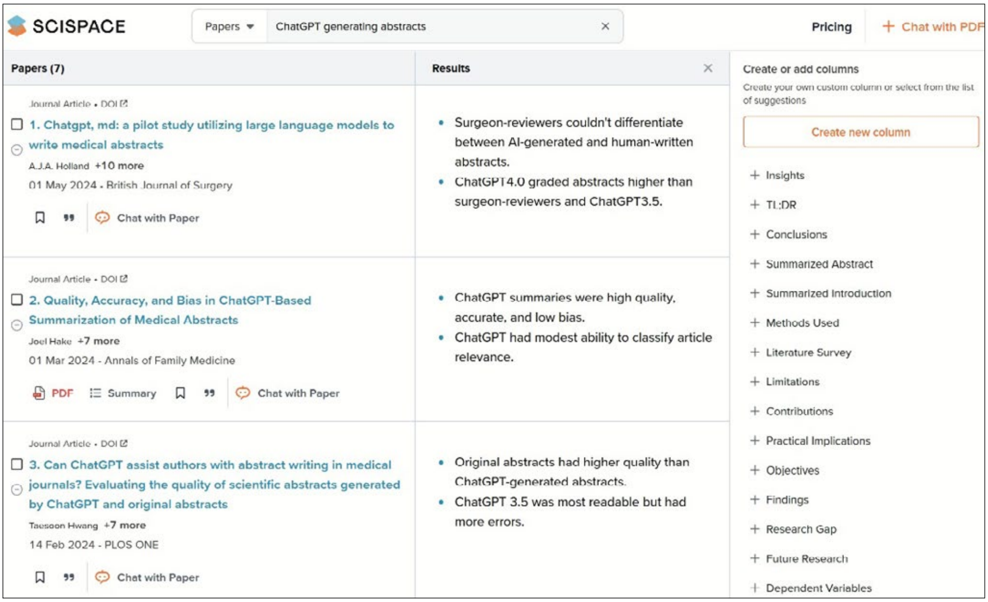
Another important limitation to consider when analysing the research results is the focus on a single journal from one discipline – information science – as well as the small number of publications included in the study – 10 in total. Even within the same discipline, there is no certainty that the results would be equally satisfactory for publications with a higher level of content complexity.

Finally, it is important to mention the subjectivity involved in evaluating the correctness of the generated abstracts. Although the Author made every effort to ensure that the assessment was reliable, it cannot be ruled out that another representative of the discipline, using slightly different criteria, might evaluate the abstracts generated by ChatGPT differently.

6. Conclusions

ChatGPT demonstrated strong capabilities in analysing and summarising documents to create abstracts for scientific publications in information science. The research conducted, along with other studies mentioned in the 'Previous Studies' section, confirms that language models can be successfully used to automatically create structured abstracts, particularly for bibliographic and full-text databases, thereby expanding the existing functionalities of these systems.

A good example of a service that already utilises the capabilities of language models is Scispace (https://typeset.io). It allows users to expand the list of retrieved publications with additional columns containing automatically generated short descriptions based on predefined headings modelled after structured abstract sections (Figure 1). In addition, users can create their custom headings using the "Create new column" button and freely engage in conversations about selected publications using the "Chat with Paper" option.



Papers (7)	Results
Journal Article • DOI 1. Chatgpt, md: a pilot study utilizing large language models to write medical abstracts A.J.A. Holland +10 more 01 May 2024 • British Journal of Surgery	• Surgeon-reviewers couldn't differentiate between AI-generated and human-written abstracts. • ChatGPT4.0 graded abstracts higher than surgeon-reviewers and ChatGPT3.5.
Journal Article • DOI 2. Quality, Accuracy, and Bias in ChatGPT-Based Summarization of Medical Abstracts Joel Hake +7 more 01 Mar 2024 - Annals of Family Medicine	• ChatGPT summaries were high quality, accurate, and low bias. • ChatGPT had modest ability to classify article relevance.
Journal Article • DOI 3. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts Taesoon Hwang +7 more 14 Feb 2024 - PLOS ONE	• Original abstracts had higher quality than ChatGPT-generated abstracts. • ChatGPT 3.5 was most readable but had more errors.

Figure 1. Adding columns to the list of retrieved publications in the Scispace service

Source: https://typeset.io.

In addition to application in bibliographic and full-text databases, language models such as ChatGPT can be effectively used to analyse scientific publications and as a source of inspiration for creating author-written abstracts. When analysing articles, it is possible to generate summaries of a specified length (e.g., 200 or 300 words) containing sections tailored to individual needs and written in the user's chosen language. Users can also ask language models for further clarification or elaboration on specific topics. In the case of personal publications, the abstract writing process can be enhanced by generating one or several versions of summaries using the language model, which can serve as a source of inspiration for further work or to improve an already written abstract.

References

Cheng, S., Tsai, S., Bai, Y. et al. (2023). Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: Cross-sectional study. *Journal of Medical Internet research*, *25*, e51229. doi: 10.2196/51229.

Gravel, J., Dion, C., Fadaei Kermani, M. et al. (2024). Will ChatGPT-4 improve the quality of medical abstracts? *medRxiv*, 2024–02. doi: 10.1101/2024.02.09.24302591.

Hake, J., Crowley, M., Coy, A., et al. (2024). Quality, accuracy, and bias in ChatGPT-based summarisation of medical abstracts. *The Annals of Family Medicine*, *22*(2), 113–120. doi: 10.1370/afm.3075.

Holland, A., Lorenz, W., Cavanaugh, J., et al. (2024). ChatGPT, MD: A pilot study utilising Large Language Models to write medical abstracts. *British Journal of Surgery*, *111*(5), znae122.039. doi: 10.1093/bjs/znae122.039.

Hwang, T., Aggarwal, N., Khan, P. Z., et al. (2024). Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *Plos One*, *19*(2). doi: 10.1371/journal.pone.0297701.

Pulikowski, A. (2020). The relation between the structure of abstracts in LIS and anthropology journals and their rank. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, *58*(1), 24–39. doi: 10.36702/zin.645.

Pulikowski, A. (2024). The comparison of structured abstracts generated by the ChatGPT language model with the author's original abstracts derived from research papers [Data set]. Zenodo. doi: 10.5281/zenodo.13765256.

Stadler, R. D., Sudah, S. Y., Moverman, M. A., et al. (2024). Identification of ChatGPT-generated abstracts within shoulder and elbow surgery poses a challenge for reviewers. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. doi: 10.1016/j.arthro.2024.06.045.

Zastosowanie modelu językowego ChatGPT do automatycznego generowania ustrukturyzowanych abstraktów

Abstrakt

Cel/teza: Badanie miało na celu ocenę użyteczności modelu językowego ChatGPT do generowania ustrukturyzowanych abstraktów publikacji naukowych.
Koncepcja/metody badań: Badania miały charakter jakościowy. Analizie poddano 10 artykułów z czasopisma Zagadnienia Informacji Naukowej – Studia Informacyjne, 5 w języku polskim i 5 w języku angielskim. Korzystając z modelu językowego ChatGPT w wersji 4o, dla każdego artykułu wygenerowano ustrukturyzowane abstrakty, które następnie porównywano z abstraktami autorskimi w celu sprawdzenia, czy zawierają poprawne informacje w poszczególnych sekcjach.
Wyniki/wnioski: ChatGPT potwierdził duże możliwości w zakresie analizy i streszczania dokumentów w celu tworzenia abstraktów publikacji naukowych z zakresu informacji naukowej. Model językowy poradził sobie równie dobrze z publikacjami w języku angielskim i polskim. Tylko w przypadku dwóch abstraktów wykryto błędnie wygenerowaną treść pojedynczych sekcji.
Oryginalność/wartość poznawcza: Badanie pokazało potencjał modeli językowych, takich jak ChatGPT, w tworzeniu ustrukturyzowanych abstraktów, zarówno na potrzeby bibliograficznych i pełnotekstowych baz danych, jak i jako uzupełnienie warsztatu badacza.

Słowa kluczowe

ChatGPT. Generowanie abstraktów przez AI. Modele językowe. Ustrukturyzowane abstrakty.

ARKADIUSZ PULIKOWSKI, profesor uczelni w Instytucie Nauk o Kulturze na Wydziale Humanistycznym Uniwersytetu Śląskiego w Katowicach. Zainteresowania badawcze: wyszukiwanie informacji, zachowania informacyjne, infometria, digitalizacja informacji. Wybrane publikacje: Modelowanie procesu wyszukiwania informacji naukowej. Strategie i interakcje (Katowice, 2018), Searching for LIS scholarly publications: a comparison of search results from Google, Google Scholar, EDS, and LISA (Journal of Academic Librarianship, 2021, współaut. A.Matysek), The Relation between the structure of abstracts in LIS and anthropology journals and their rank (Zagadnienia Informacji Naukowej, 2020).

Contact details:
arkadiusz.pulikowski@us.edu.pl
Uniwersytet Śląski w Katowicach
Wydział Humanistyczny
Instytut Nauk o Kulturze
ul. Bankowa 11, 40–007 Katowice

CALL FOR PAPERS #2

SCIENCE COMMUNICATION,
ACADEMIC LIBRARIES,
RESEARCH LIBRARIES,
DATA SCIENCE, BIBLIOMETRICS

SUBMISSION DEADLINE:
15TH SEPTEMBER 2025

WYDAWNICTWO
NAUKOWE
I EDUKACYJNE
SBP