



# Semi-automating the Scoping Review Process: Is it Worthwhile? A Methodological Evaluation

Shan Zhang<sup>1</sup> · Chris Palaguachi<sup>2</sup> · Marcin Pitera<sup>2</sup> · Chris Davis Jaldi<sup>3</sup> · Noah L. Schroeder<sup>1</sup> · Anthony F. Botelho<sup>1</sup> · Jessica R. Gladstone<sup>2</sup>

Accepted: 30 October 2024 / Published online: 9 November 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024, corrected publication 2024

## Abstract

Systematic reviews are a time-consuming yet effective approach to understanding research trends. While researchers have investigated how to speed up the process of screening studies for potential inclusion, few have focused on to what extent we can use algorithms to extract data instead of human coders. In this study, we explore to what extent analyses and algorithms can produce results similar to human data extraction during a scoping review—a type of systematic review aimed at understanding the nature of the field rather than the efficacy of an intervention—in the context of a never before analyzed sample of studies that were intended for a scoping review. Specifically, we tested five approaches: bibliometric analysis with VOSviewer, latent Dirichlet allocation (LDA) with bag of words, *k*-means clustering with TF-IDF, Sentence-BERT, or SPECTER, hierarchical clustering with Sentence-BERT, and BERTopic. Our results showed that topic modeling approaches (LDA/BERTopic) and *k*-means clustering identified specific, but often narrow research areas, leaving a substantial portion of the sample unclassified or in unclear topics. Meanwhile, bibliometric analysis and hierarchical clustering with SBERT were more informative for our purposes, identifying key author networks and categorizing studies into distinct themes as well as reflecting the relationships between themes, respectively. Overall, we highlight the capabilities and limitations of each method and discuss how these techniques can complement traditional human data extraction methods. We conclude that the analyses tested here likely cannot fully replace human data extraction in scoping reviews but serve as valuable supplements.

**Keywords** Pedagogical agent · Machine learning · Natural language processing · Topic modeling · Clustering · Scoping review

It has long been recognized that there are various approaches to research synthesis (Cooper, 1988; Grant & Booth, 2009). For example, in the educational sciences,

meta-analyses seek to quantitatively synthesize the efficacy of interventions or the relationships between variables, while systematic reviews may qualitatively describe aspects of pedagogical intervention design or efficacy. Although these methods are well-known and quite common in the field, we wish to draw attention to the fact that meta-analysis and systematic review are not the only systematic methods of reviewing the literature (Grant & Booth, 2009). Rather, they are families of reviewing approaches, both of which fall under the broader classification of systematically-conducted review methods.

One characteristic of nearly all systematic review methods<sup>1</sup> is that they can be time-consuming and require a lot of human effort between study screening, data extraction, and data analysis (Chernikova et al., 2024; Wang & Luo, 2024). This presents a conundrum for the research synthesist. We want to provide a comprehensive review while also ensuring that it can be completed within a reasonable period of time. For example, it is not uncommon in the field to screen thousands of abstracts before narrowing down a final sample of less than 100 studies that qualify for the analysis. While one may think the time-consuming work is done there, the experienced reviewer knows that it is not. Data must then be extracted from those qualifying studies in a manner that is transparent, reliable, and able to be analyzed. Moreover, all the details of this process must be rigorously recorded to enable transparent reporting aligned with the PRISMA standards (Page et al., 2021). Taken together, the systematic review process can be quite the time-consuming endeavor.

It is not surprising then that research is on-going to help reviewers screen studies more quickly. For example, Campos et al. (2024) examined the efficacy of a natural language processing screening tool during the title and abstract screening process. Similarly, Chernikova et al. (2024) used machine learning to help identify studies that may be appropriate for their analysis.

While this work is without a doubt important for moving the field forward, there is a complementary line of research around the use of different methods for extracting data from studies more quickly. This is the area of work in which the present study is situated, and based on our own work and recent preprints and articles we have read, we believe that the rapid innovation in large-language models (LLMs) and machine learning (ML) techniques may dramatically increase the number of studies in this area. For example, Wang and Luo (2024) created an application that uses LLMs to extract educationally-relevant data from studies. Our own preliminary work indicated that various LLMs, even the free versions, were able to extract a variety of types of data relatively reliably, although the accuracy of such data extraction varied depending on what types of data were being extracted and which LLM was used (Jaldi & Schroeder, 2024). However, a caveat to this is that LLM abilities and drawbacks are constantly changing. Given this rapidly changing landscape, for the purposes of this study, we sought to use more well-established methodologies that are well known in the field. Specifically, we investigate how, and to what extent, bibliometric analysis and various machine learning methods can help us extract data from study titles, keywords,

---

<sup>1</sup> Except perhaps the rapid review, which is designed to be completed more quickly.

author information, year of publication, and abstracts to provide unique insights and save time during the scoping review process.

## Scoping Reviews

As noted, there are various styles of systematically-conducted reviews, each with its own specific purpose and methodology (see Grant & Booth, 2009). A common scenario a research synthesist may find themselves in is wanting to better understand a body of literature they may not be intimately familiar with. As a consequence, at the onset of their interest, it is not clear to them if the literature supports, for example, a meta-analysis. Moreover, the landscape of the literature in the area may be unclear. Perhaps some quick database searches produced a variety of related studies, but they were about very different topics or expressed quite different perspectives on the issue. A potential solution to this problem is the *scoping review*. Scoping reviews can be a critical part of the larger systematic review process, highlighting the nature of the field, where there is room for further systematic synthesis, and where additional empirical research is needed (Arksey & O'Malley, 2005; Munn et al., 2018; Tricco et al., 2016, 2018). While scoping reviews as a methodology have existed for some time (Arksey & O'Malley, 2005) and there is PRISMA guidance for how they should be reported (Tricco et al., 2018), they are less common in educational fields. Despite this, educational researchers have noticed the method and there has recently been a rise in scoping reviews published, particularly around educational technologies (Blair et al., 2021; Ranieri et al., 2022; Schroeder et al., 2023; Sperling et al., 2024; Su & Yang, 2022; Zhang et al., 2024b).

However, scoping reviews are inherently limited in the types of information they provide, often being viewed as more “descriptive” than they are “analytical.” As an example, scoping reviews typically do not examine the efficacy of an intervention, but rather may seek to describe the types of research investigating the efficacy of an intervention (Arksey & O'Malley, 2005; Tricco et al., 2018). In short, a scoping review will provide quite a different dataset and understanding than say, a meta-analysis, on a similar topic.

Like all systematic reviews, scoping reviews tend to be time-intensive endeavors because they require a systematic literature search, thorough documentation, the screening of hundreds or thousands of studies, and the subsequent data extraction and thematic analysis (Tricco et al., 2018). Given the time investment required to complete a methodologically-strong scoping review, some scholars may be reluctant to conduct such work, perhaps fearing a challenging publication landscape for scoping reviews since they do not synthesize the efficacy of an intervention, but rather describe the body of evidence around such an intervention. Given these circumstances, we questioned whether any methods from data analytics or machine learning could help produce similar results as human-coded scoping review data and improve the efficiency.

## **Analytics and Machine Learning Methods in Educational Sciences**

Data analytics and machine learning methods have contributed to our understanding in the field of education for some time. For example, previous studies have used common techniques such as bibliometric analysis (Maphosa & Maphosa, 2023; Merigó & Yang, 2017; Merigó et al., 2015), clustering (Antonenko et al., 2012; Zhang et al., 2024a), and topic modeling (Chen et al., 2021; Maphosa & Maphosa, 2023), among other methodologies. There are a vast number of different techniques that exist, however not all are useful to those conducting a scoping review. Specifically, some consider semantic context in their analysis, whereas others do not. Some are interpretable, whereas others are “black box” methods. Finally, we have found that it is easy to overgeneralize the results of some analyses. For example, it is critical to examine *what the algorithm does* and *how the algorithm works* in order to truly understand and interpret the results without over-generalization and misrepresentation. With these types of considerations in mind, we began our exploration of how analytics and machine learning can be used as part of the scoping review process.

## **The Present Study**

To set the stage for the methodologies explored in this paper, it is important to understand the contextual foundation that initiated the inquiry. We were interested in understanding the body of literature around the use of pedagogical agents (PAs) on learners’ motivation, also known as motivational agents (Siegle et al., 2023). Part of our team was familiar with this body of research, but another portion was not. Our literature searches located a considerable number of studies, many more than were initially anticipated, yet our timeline to complete the project remained the same. As researchers, we found ourselves in a paradox. We understood the value of scoping reviews in the systematic review and meta-analysis process, yet the time investment involved was a challenge given the timeframe we were working with.

Given this scenario, we turned to the literature to see if there were analytical techniques that could be used to extract similar data from studies as a scoping review would produce. To this end, we investigated the use of bibliometric analyses, topic modeling approaches (latent dirichlet allocation (LDA) and BERTopic) and various clustering techniques, all of which we anticipated could provide *some* of the information typically produced from or during a scoping review.

In this paper, we strive to make methodological contributions to the research synthesis literature. We first introduce and then critically examine various data analytical techniques, identifying what we learned about our never-before analyzed a sample of studies from each. We highlight what these methods tell us as research synthesists (and importantly, do *not* tell us) and make suggestions for when each may be helpful. We also comment on the reproducibility of these

analyses and the importance of reproducibility during the systematic review process. In sum, we contribute a series of analyses describing the field of motivational agents and provide a methodological contribution to the literature by investigating to what extent various analytical techniques can potentially reduce the need for human data extraction in the scoping review process.

To this end, we present the following sections. In the general methods, we describe the methods used across all methods examined, including our systematic literature review and study screening process, as aligned with the PRISMA standards (Page et al., 2021). In Part I, we utilize bibliometric analysis, followed by a general introduction to the machine learning methods and the necessary pre-processing of data in Part II. In Part III, we use LDA to perform topic modeling, and Part IV demonstrates how different types of clustering with different feature extraction methods can be used in this context. Part V concludes our methodological exploration by describing the use of BERTopic. Finally, Part VI presents a general discussion and the implications of this work.

## General Methods

As noted, this study contextualizes the use of various analytic methods through the context of an actual scoping review that had not been analyzed. The purpose of the study was to describe the nature of the body of evidence around the use of PAs and their influence on learners' motivation. With that in mind, we conducted the following systematic literature search and screening process.

### Literature Search

On September 15, 2023, we conducted a comprehensive literature search across eight databases: MEDLINE with Full Text, APA PsycInfo, Academic Search Complete, Education Research Complete, Eric, CINAHL Plus with Full Text, ACM Digital Library, and Web of Science. The search string was formulated to capture studies involving learners interacting with PAs and measuring motivation-related constructs: (“virtual human”\* OR “embodied agent”\* OR “virtual character”\* OR “pedagogical agent”\* OR “conversational agent”\* OR “motivational agent”\*) AND (motivat\* OR self-efficacy OR self-confidence OR ability belief\* OR self-concept OR interest\* OR engag\* OR value\* OR util\* OR “sense of belonging” OR belong\*).

Initially, the search yielded 2478 studies. After removing duplicates, 1650 studies remained for abstract screening. Table 1 presents the results of the database search and the database-building process.

### Study Screening

The study screening process was divided into two phases. Figure 1 provides an overview of the process.

**Table 1** Overview of the database search and duplicate removal process

Database	Results
Medline with Full Text	253
APA PsychInfo	225
Academic Search Complete	211
Education Research Complete	111
ERIC	80
CINAHL Plus with Full Text	75
ACM	489
Web of Science	922
Articles cited in Baylor (2011)	44
Articles included in Guo and Goh (2015) motivation analysis	10
Articles included in Heidig and Clarebout (2011)	26
Articles included in Schroeder and Adesope (2014)	15
Articles included in Wang et al. (2023) (note, some could not be located, the 17 reflects located studies)	17
Total	2,478
Duplicates removed	823
Records could not be located in English	5
Total for Phase I Screening	1650

## Inclusion and Exclusion Criteria

For a study to be included in the review, it must meet the following criteria:

- Must have at least one condition with a visible PA, which must be a virtual character rather than a video, rendering, or image of an actual human.
- Must compare two conditions, either a PA to a no PA group or a PA group to another PA group.
- Must collect either quantitative or qualitative data.
- Must measure some aspect of students' motivation (motivation, self-efficacy, confidence beliefs, ability beliefs, self-concept, interest, engagement, value, utility, etc.).
- Must be an empirical study with primary data.
- Must be published in English.
- Must be publicly accessible via databases or inter-library loan requests.

Studies were excluded based on the following criteria:

- The study examined the use of avatars (representations of self).

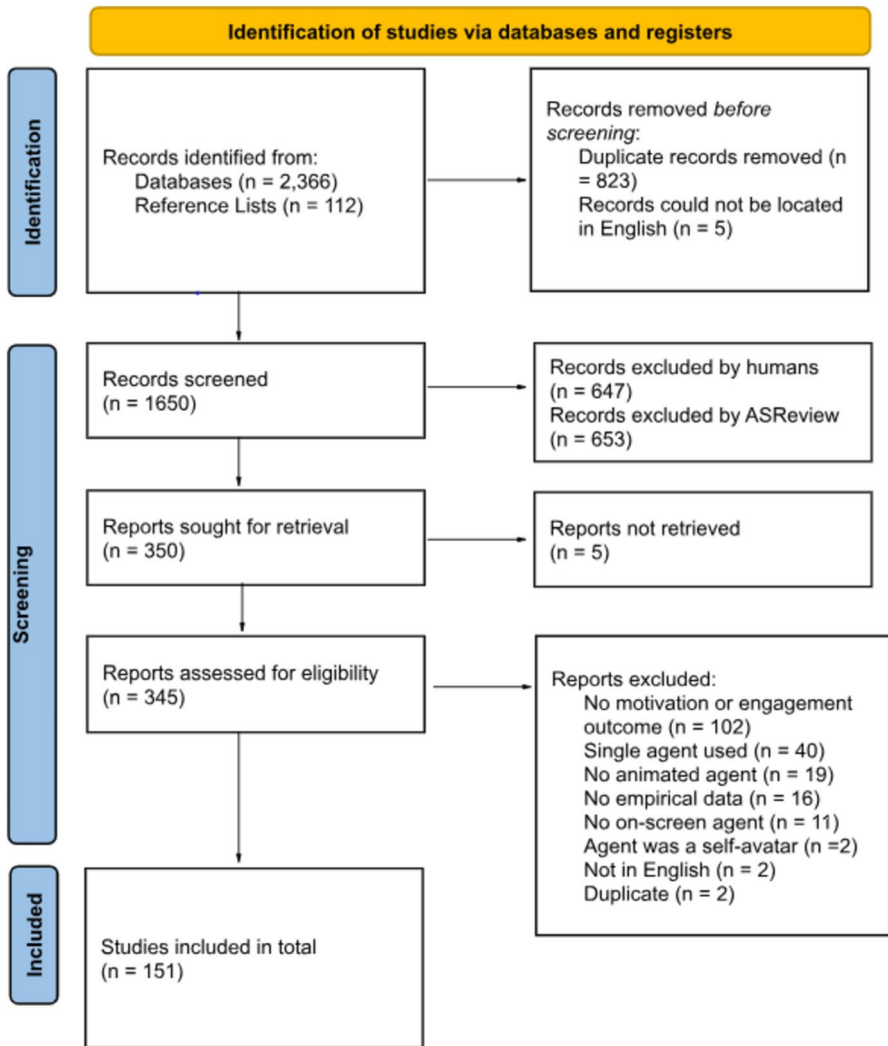


Fig. 1 PRISMA diagram, adapted from Page et al. (2021)

## Phase I Screening

During the first screening phase, the titles and abstracts of the studies were reviewed to determine if they met the inclusion criteria. All identified studies were imported into ASReview (<https://asreview.nl/>), a screening tool that employs natural language processing to rank studies by relevance. We trained the system using ten studies, with five considered relevant (Baylor & Plant, 2005; Domagk & Niegemann, 2005; Plant et al., 2009; Rosenberg-Kima et al., 2007, 2010) and five deemed irrelevant based on our inclusion and exclusion criteria (Gottsacker et al., 2022; Kolodkin

et al., 2012; Koprinkova-Hristova et al., 2013; Sukthankar & Sycara, 2011; Wang et al., 2021). Sentence-BERT (SBERT) was used for feature extraction, with logistic regression as the classification method, “maximum” as the query strategy, and dynamic resampling (doubling) to maintain balance.

ASReview was used for title and abstract screening, informed by findings from a recent study which indicated that screening 60% of abstracts could identify 95% of included studies using logistic regression and SBERT (Campos et al., 2024). We applied this approach to our database, screening 60% of the studies. To validate this, we examined the longest sequence of consecutively irrelevant studies, finding 126 such studies, accounting for 7.58% of our sample. Based on Campos et al. (2024), a recall rate exceeding 95% corresponds to 7% consecutively irrelevant data. Therefore, we concluded that screening 60% of our database would likely identify 95% of relevant studies.

In the abstract screening phase, 1650 studies were imported into ASReview. Out of these, 647 studies were excluded by one author and 653 studies were excluded by ASReview, leaving 345 studies for full-text screening.

## Phase II Screening

In the second phase, the same author who conducted the abstract screening independently reviewed the full texts of the remaining 345 studies. This resulted in 151 studies meeting the inclusion criteria and being analyzed in the review. The references for all studies included in the review are available in Supplementary Materials 1.

## Data Cleaning

Many data science and machine learning methods require data to be “cleaned” before analysis. We implemented a comprehensive data cleaning process to ensure the quality and consistency of the abstracts and to remove any potential noise for further analysis. Initially, all abstracts were stripped of specific copyright notices, author notes, and other non-essential text using a series of regex-based replacements. For instance, text such as “[Copyright &y& Elsevier]” and similar phrases were systematically removed. Following this, we focused on eliminating any numbers, unusual symbols, and extraneous punctuation to ensure that the abstracts were standardized and ready for analysis.

## Data Availability

All of the data used for this paper, and the Python code for the machine learning methods, are available on OSF: [https://osf.io/5tb2y/?view\\_only=1bd4b35df6f2423f871d4eda523e0277](https://osf.io/5tb2y/?view_only=1bd4b35df6f2423f871d4eda523e0277).



## Part I—Bibliometric Analysis

When entering new areas of research or conducting a scoping review, it is essential to thoroughly understand the prominent authors, key lines of inquiry, and the overall structure of the field. This is particularly important in rapidly expanding areas like Educational Psychology and Learning Sciences, where aligning with influential works and current trends is crucial for advancing the discipline. To achieve this, scholars often turn to bibliometric analysis. Bibliometric analysis is a widely recognized method for synthesizing and navigating complex research landscapes (Donthu et al., 2021). For instance, bibliometric analysis has been used to map whole fields such as educational psychology (Hassan et al., 2024).

Bibliometric analysis not only describes the studies within a sample but also reveals their interconnections (Yu et al., 2020). It allows researchers to track trends and major shifts within a field, such as those caused by global events like COVID-19 (Li & Jiang, 2021). This approach provides critical insights into publication frequency, relationships between key publications and keywords, and collaborative author networks (Slimi & Carballido, 2023). For example, Maphosa and Maphosa (2023) used keyword analysis, topic modeling, and author co-citation analysis to uncover evolving trends in artificial intelligence research in higher education.

In this study, we utilized VOSviewer, a powerful bibliometric mapping tool developed by Van Eck and Waltman (2010). VOSviewer visualizes bibliometric maps based on the co-occurrence of words in titles and abstracts, as well as the co-citation of authors. Through these analyses, we sought to deepen our understanding of the intellectual structure and research dynamics within our sample. While bibliometric analysis represents a different methodological approach than a scoping review, we feel that it can add important context that a scoping review may otherwise not uncover. As such, bibliometric analysis represented our starting point to begin semi-automating the scoping review process for helping us understand the field.

### Bibliometric Analysis Methods

Publication data was extracted from all the 151 included studies into a CSV file. This publication dataset included information such as authors, year of publication, title, abstract, and keywords. This dataset was then uploaded onto VOSviewer. For this analysis, we used features from VOSviewer to analyze the abstract, title, keywords, and author co-citation co-occurrences.

For the analysis of abstracts and titles, VOSviewer requires users to select several thresholds prior to conducting the analysis. These thresholds included the minimum number of occurrences of a term and the number of terms to be selected. The minimum occurrence threshold determines how frequently a term must appear to be included in the analysis, with higher thresholds applying more stringent criteria. The number of terms to be selected adjusts the criteria for the number of words used in the analysis, with VOSviewer defaulting to include 60% of the most relevant terms. Following the bibliometric analysis conducted by Maphosa and Maphosa (2023), we used VOSviewer's default values, recommended for filtering the most relevant

terms. Starting with 3745 unique words, we applied a minimum occurrence threshold of 10, leaving 67 words. The second threshold, including 60% of the most relevant terms, resulted in 40 words for clustering in this analysis.

To analyze the keyword and title data, we reformatted and relabeled the data, replacing the original abstract data with keyword data and labeling the new column as the abstract title. Similar to the previous analysis, threshold values were selected before conducting the analysis. Starting with 1302 unique words, we applied a minimum occurrence threshold of 10, which left 40 words. The second threshold, including 60% of the most relevant terms, resulted in 18 words for clustering in this analysis.

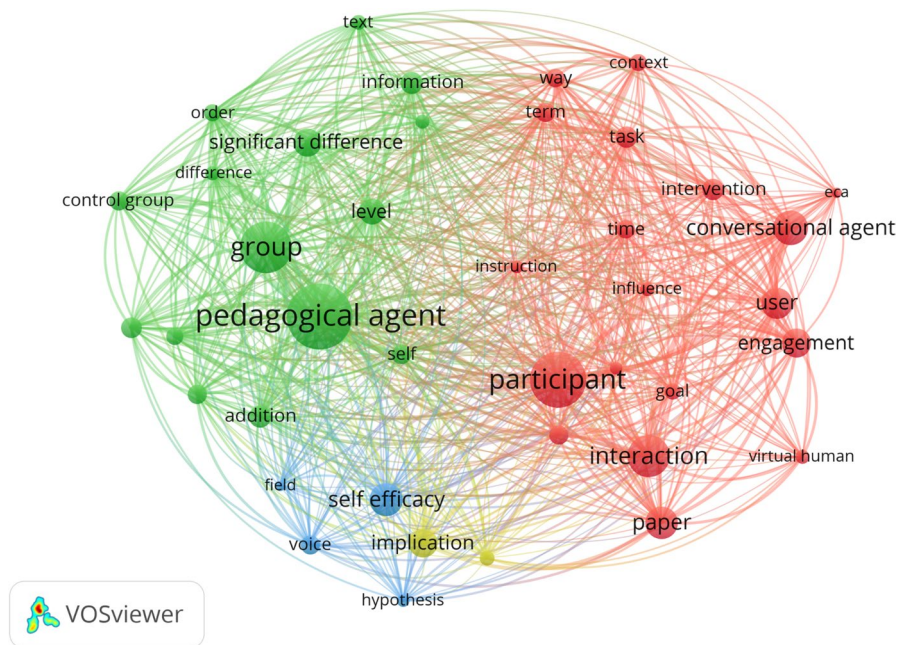
To analyze the author co-citation co-occurrence, it is important to make sure that there is a consistent formatting of the authors' names. For example, Baylor, A. L., Baylor, Amy L., and Baylor, A. are present in the dataset but are all the same author. Data reformatting and cleaning were necessary to address variations in authors' names across publications. After cleaning the dataset, we identified a total of 393 unique authors. Similar to the abstract and title analysis, several threshold values must be selected before conducting the author co-citation analysis. These thresholds included the minimum number of documents of an author and the number of authors to be selected. To analyze the relationship of all the authors in this review, the minimum document threshold for an author was set to 1. This allowed all 393 authors to be included in the analysis. The second threshold allowed us to make adjustments to the total number of authors to be included. The default value by VOSviewer was set to 393, and we left it in the default as we wanted to see all the authors' connections. The last threshold displayed all authors and the document and total link strength. We noticed that IEEE was included as an author due to publication data extraction. We confirmed that this was not an author or group of researchers and IEEE was not included in the study. This resulted in 392 authors being included in this co-citation analysis.

## Bibliometric Analysis Results

We first explored the patterns found in the study abstracts and titles. In Fig. 2, we see an overview of the 40 words left from filtering the terms, leaving 4 clusters of terms.

In this visualization, we can see each cluster denoted by different colors. Additionally, we can see the terms with the highest occurrences. High occurrence clusters are defined by the size of the term within the cluster. In the green cluster, the biggest term is "pedagogical agent" with a recurrence of 61. In the red cluster, the biggest term is "participant" with a recurrence of 52. In the blue cluster, the biggest term is "self-efficacy" with a recurrence of 31. In the yellow cluster, the biggest term is "implication" with a recurrence of 27. Although we did see a high frequency of words related to this review (pedagogical agent, self-efficacy, motivation, etc.), the other words in clusters did not necessarily create interpretable themes, making this analysis of terms used in the abstract and title difficult to interpret.

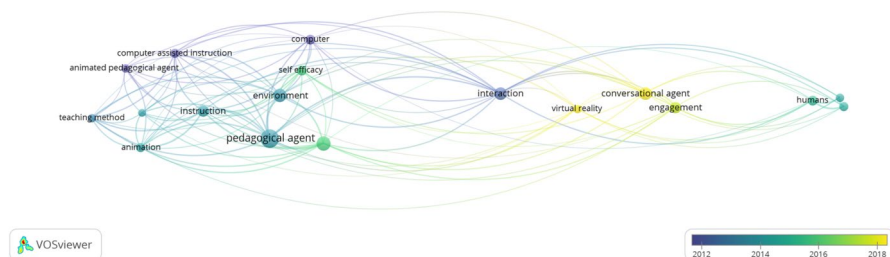
Additionally, we analyzed the keywords and title data alongside the year of publication to look at trends of words used over time. In this analysis, VOSviewer



**Fig. 2** Abstract and title word clusters

averages the year the term or word is used in a publication. In this review, the studies ranged from 1998 to 2023, however, due to the way VOSviewer averages year of publication for terms, for this analysis the studies ranged from 2011 to 2019. Since we are interested in exploring the patterns of keywords, we explored the highest occurrence of keywords (size of clusters) and the time keywords are present. In Fig. 3, we see the 2 clusters where the highest occurrence terms are “pedagogical agent” and “interaction.”

As noted, we also sought to explore the shifts in keywords over time. For example, we noticed that many early studies focused on design or the environment (left side of Fig. 3; 2011–2014, e.g., animated pedagogical agent, animation, and teaching method). Focusing on keywords like self-efficacy and motivation, we see that



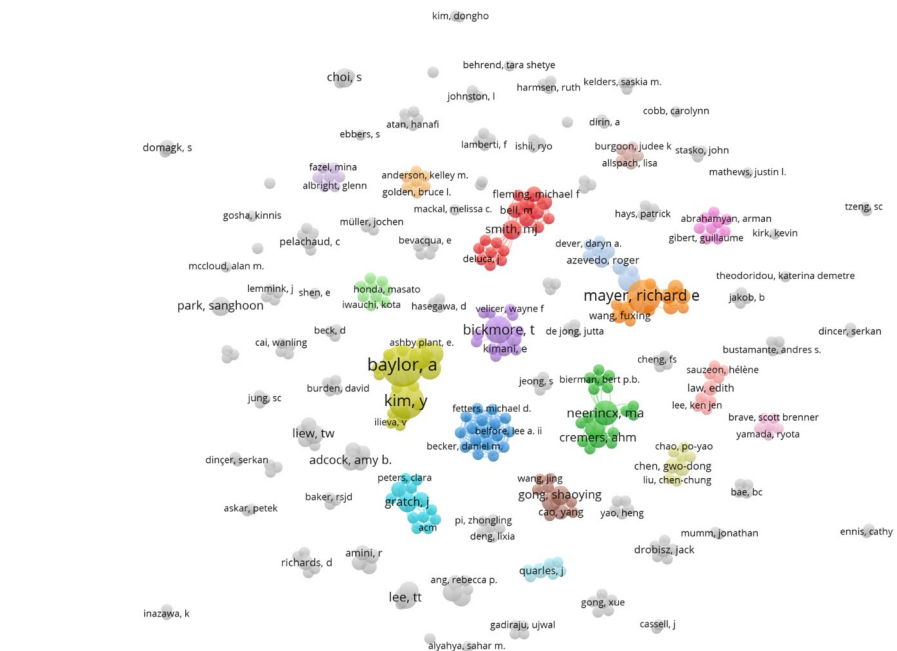
**Fig. 3** Keyword and title word clusters overtime

these studies are averaging towards the end of 2015. When focusing on the keyword engagement, on average these studies are around mid-2017. These patterns are interesting as they can help researchers better understand how the field shifts their focus or the words they use. Another interesting pattern is looking at the average year the keyword pedagogical agent was used compared to conversational agents. In our analysis, we see that many of the earlier studies (end of 2013) used keywords like “pedagogical agent,” and the later studies (end of 2019) used “conversational agent.”

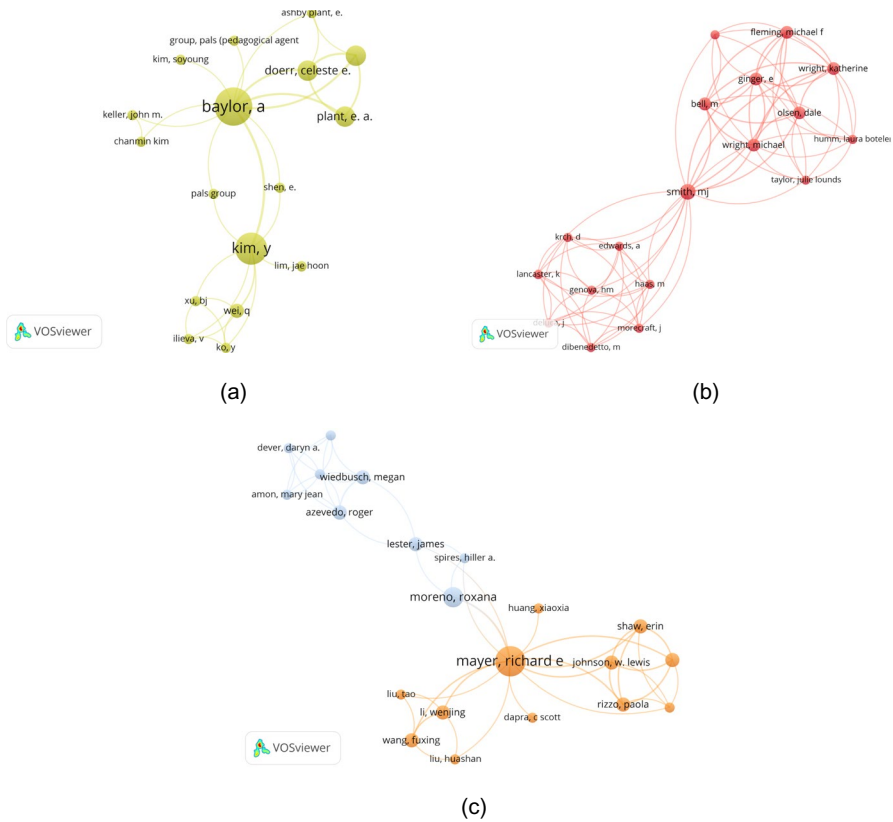
For author co-citation, we explored the different relationships between the authors in this review. We wanted to know who is working with each other and how they are contributing to this specific work. In Fig. 4, we can see an overview of all 392 authors.

The size of the clusters corresponds to the number of documents the authors have contributed to in this review. While the term mapping (abstract, title, and keywords) features were not particularly useful in our case, the co-citation author mappings effectively visualize all the authors included in this review and their interrelationships. In Fig. 5a–c, we highlighted several clusters of authors with the highest document co-occurrence (article count) and total link strength (co-author connections).

The authors within these clusters who have the highest number of documents and total link strength include Amy Baylor, Michael J. Smith, and Richard E. Mayer. This helps to highlight the research teams and cross-collaboration of authors within the field.



**Fig. 4** Overview of authors



**Fig. 5** Research group clusters by author co-occurrence

## Bibliometric Analysis Discussion

Our bibliometric analysis highlights its utility for reviewers seeking a quick and user-friendly method to analyze patterns in word usage and author connections across articles (Yu et al., 2020). VOSviewer proved to be an effective tool for identifying key researchers and research teams, providing valuable insights into the collaborative networks within the field. This capability is particularly beneficial for researchers new to a research area, as it offers a method to begin identifying key lines of systematic research. In addition, keyword and title analysis illustrated shifts in focus and terminology within the field, highlighting significant changes over time. Despite its strengths, our analysis revealed limitations when using abstract and title data. While the high frequency of certain words confirmed the interests of our team, the clusters themselves were not as informative as anticipated.

To summarize, we found that the bibliometric analysis allowed us insights into who the collaborative research teams in the area are and how keywords have changed over time. However, we did not gain any insights into the *types* of work happening

in the field, which is a key aspect of scoping reviews. As such, we returned to the literature for alternative analyses that may get us closer to our goal. We needed something that would help us identify the topics of studies so we could better understand our sample.

## Part II—Introduction to Machine Learning Approaches

Following the bibliometric analysis using VOSviewer, we located three machine learning techniques which could provide plausible alternatives for learning about the nature of our sample. We explored three methodologies to extract and analyze topics and themes from textual data: LDA, clustering (we used various methods but settled on hierarchical clustering, as explained later), and BERTopic (see Fig. 6). All analyses were run in Python.

### Data Preprocessing

The following preprocessing steps were applied to normalize the textual data, which is common when conducting natural language processing (NLP). First, we converted all text to lowercase to eliminate potential case-sensitivity issues. Next, we tokenized the text, splitting it into individual words, or tokens, which represent independent linguistic components (e.g., contractions like “I’m” were divided into “I” and “m”). We then performed lemmatization, which reduces words to their base forms to remove linguistic differences based on tense and other forms (e.g., “learns,” “learning,” and “learned” were mapped to “learn” and “better” was reduced to “good”). Lastly, we removed punctuation and stopwords. Stopwords are frequent but generally do not carry significant meaning in the context of these analyses, such as “the,” “and,” “a” and “of.” These steps enhanced the text data’s readiness for analysis and prepared it for clustering and topic modeling in subsequent processes.

## Part III—LDA

As noted, analyzing the abstracts using bibliometric analysis lacked contextual meaning. Consequently, we were unable to identify distinct and meaningful topics or themes within the literature. One strategy to address this is topic modeling (Sandhiya et al., 2022), with a common method being that of LDA. LDA is an unsupervised machine learning technique that can help examine the main topics of a collection of textual data by analyzing word probability distributions across documents. It is particularly effective for handling large-scale written essays in a cost-efficient and timely manner, making it a popular choice in educational data mining. It has been widely used to extract topics from a large amount of abstracts (Sperandeo et al., 2020), teachers’ reflections (Zhang et al., 2024a), analyze rich textual dialogues from Massively Open Online Course discussion forum posts (Ezen-Can et al., 2015), identify distinct themes in course enrollments (Motz et al., 2018),

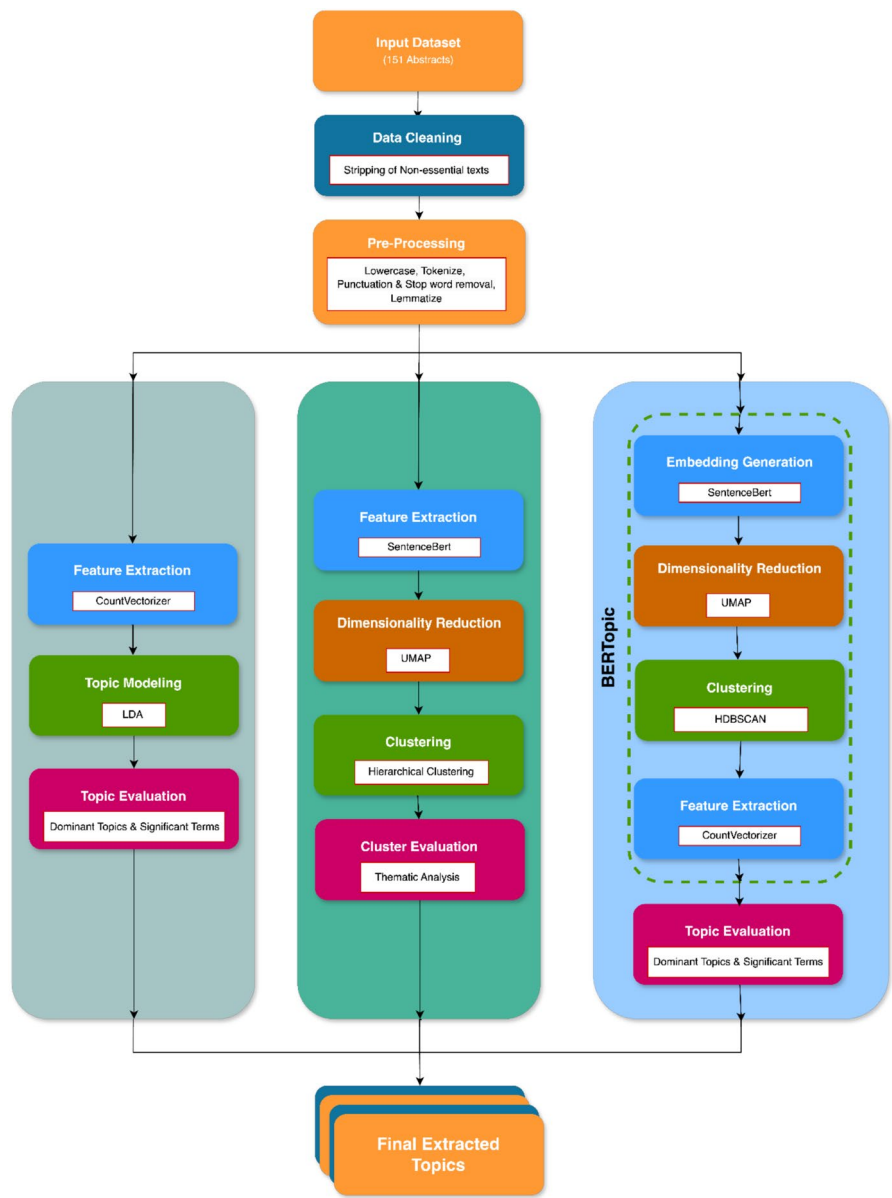


Fig. 6 Methodological framework overview

describe underlying topics from educational leadership research literature (Wang et al., 2017), extract topics from open-ended questions in teacher self-assessments (Buenaño-Fernandez et al., 2020), and reveal distinct topics in the dynamic relationship between artificial intelligence and higher education (Maphosa & Maphosa, 2023). We sought to explore how LDA can be used to inform our scoping review.



## LDA Methods

To prepare our textual data (study abstracts) for LDA, we employed the Bag-of-Words (BoW) for feature extraction. BoW represents a document as a collection of its words, with each unique word considered a feature, and the document is represented as a vector of word frequencies. It is a widely used method due to its simplicity and effectiveness in quantifying word presence in text data.

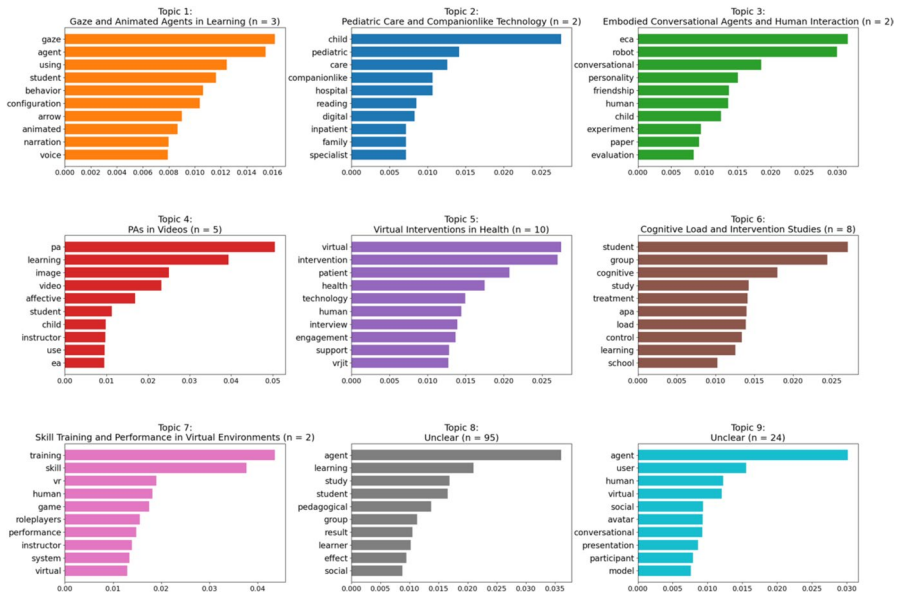
Once the textual data was transformed into BoW vectors, we applied the LDA model to identify latent topics within the corpus. To determine the optimal number of topics, we computed the coherence score for topic numbers ranging from 2 to 15. This range was chosen to balance granularity and interpretability. A lower number of topics might oversimplify the data, missing important nuances, while a higher number might result in overly specific or fragmented topics that are harder to interpret. The coherence score (ranging from 0 to 1) is a crucial metric because it measures the degree of semantic similarity between high-scoring words in a topic, ensuring that the topics identified are both meaningful and interpretable. High coherence scores indicate that the words within a topic co-occur frequently and are contextually related, making the topics more understandable and relevant to the corpus. It is important to note that while LDA does not directly account for semantic similarity, the coherence scores used to assess its output can reflect how semantically related the words are in a topic.

Once we finalized the optimal number of topics by selecting the ones associated with the highest coherence score, we printed out each topic along with the top 10 associated keywords and visualized the distribution of abstracts across various topics over time to highlight research trends. To interpret each topic, inductive thematic analysis was applied to generate code and themes based on the most frequent 10 keywords.

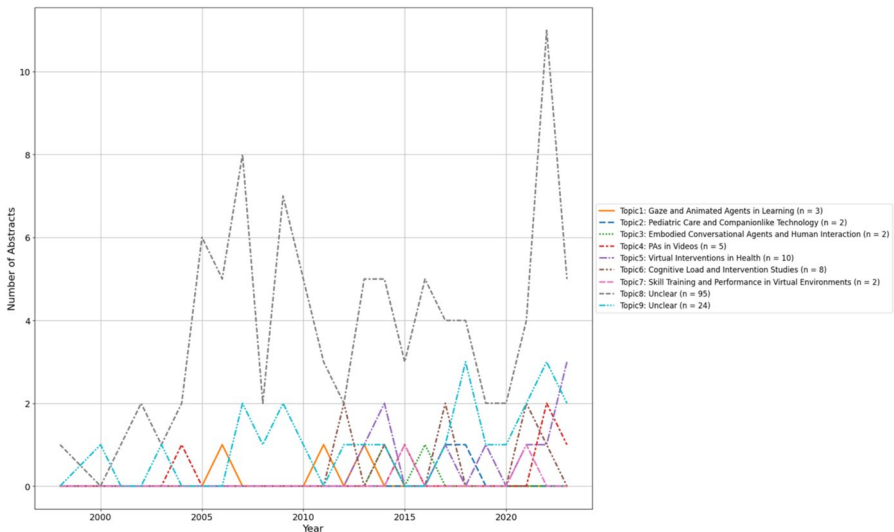
## LDA Results

Overall, we identified nine topics with a coherence score of 0.34, which is considered a moderate score (Krishnan, 2023). Figure 7 visually represents the results of the topic modeling analysis using LDA. Two researchers reviewed the top 10 keywords associated with each topic and collaborated to determine the main topics. The identified themes in our analysis range from specific topics like “Gaze and Animated Agents in Learning,” “Embodied Conversational Agents and Human Interaction,” “Skill Training and Performance in Virtual Environments,” and “Pediatric Care and Companion-like Technology” to broader categories such as “Virtual Interventions in Health,” “PA in Videos” and “Cognitive Load and Intervention Studies,” with varying numbers of abstracts included. Each theme represents a unique aspect of the research corpus, highlighting the diversity of topics covered. Notably, most topics include less than 10 abstracts, and two “Unclear” topics contain the most abstracts ( $n = 95$  and  $n = 24$ ).





**Fig. 7** Topic word distribution ( $n$  = abstracts)



**Fig. 8** Topic distribution over years

Similar to the bibliometric analysis, we can look at how topics or terms shift over time. In Fig. 8, we can see all 9 topics displayed across publication dates.

The line graph illustrates the distribution of abstracts across various topics over time, spanning from the year 2000 to 2022. Each colored line represents a distinct

topic, with the  $y$ -axis indicating the number of abstracts and the  $x$ -axis showing the year. The most significant trend is observed in Topic 8, labeled as “Unclear,” which consistently shows the highest number of abstracts across the years, with a noticeable peak around 2020. This suggests a substantial portion of the abstracts could not be clearly categorized into the predefined topics, and this was not constrained to either newer or older studies.

## LDA Discussion

Overall, the results from LDA reveal the predominant keywords associated with each topic. For example, the topic “Gaze and Animated Agents in Learning” shows a strong emphasis on terms such as “agent,” “gaze,” and “attention,” indicating a focus on non-verbal cues and interactive agents to enhance learning experiences. By extracting the top keywords associated with each topic, we have a general understanding of the focus areas within our studies, how many studies fall within each focus area, and how topics change over time. These patterns were not able to be observed from the bibliometric analysis.

However, it is very clear that LDA has limitations for the research synthesist. First, with our sample the vast majority of studies (79%) fell within the “Unclear” topics. This implies that LDA worked well for identifying specific or narrowly used keywords, however it was unable to meaningfully categorize the vast majority of our studies that may use more “generic” descriptions in their abstract. Second, the model primarily relies on word frequency and co-occurrence patterns, which may not fully capture the semantic context of the texts. This can make it challenging to interpret topics based solely on the top keywords, potentially limiting the depth of understanding of the topics and trends. This could explain our large “unclear” topics.

In summary, from LDA we learned that we have a number of narrow, specialized topics in our sample of studies. However, our purpose is to understand our *entire* sample, not only a small portion of it. Accordingly, we sought an analysis that can utilize other information rather than simply word co-occurrence patterns.

## Part IV—Clustering Techniques

Since LDA relies on word frequency and co-occurrence patterns without capturing semantic context, we next turned to clustering algorithms and the feature extraction methods available for them (for an introduction, see Antonenko et al., 2012). Our team began with  $k$ -means clustering since it is a common method to explore and reveal patterns of the underlying structure of the data, making it a good initial approach for exploratory data analysis. First, we used TF-IDF as the feature extraction method. However, our results were not satisfying in that they were not making notable progress toward our goals beyond what we achieved with LDA. We next tried  $k$ -means clustering with SBERT because we felt that adding semantic context to the feature extraction may prove useful. SBERT was viewed as an appealing feature extraction method because it effectively captures semantic similarities between

sentences, providing high-quality embeddings that have been successfully used in various NLP tasks (Reimers & Gurevych, 2019). However, again, we found that *k*-means did not provide us a particularly fruitful result.<sup>2</sup> We realized that we not only wanted to see the clusters of studies and understand clear themes from each cluster, but how they were related to one another as well.

We thus turned to hierarchical clustering with SBERT for feature extraction. Hierarchical clustering is an unsupervised clustering learning method used to hierarchically group observations or samples in a dataset according to similarity, building a tree structure to show clustering relationships at different levels. This would allow us to visually see how groups of studies are related. We hoped that between the visual diagram and the analysis capturing semantic meaning within abstracts would lead to more interpretable themes in our data.

### Hierarchical Clustering with SBERT Methods

We applied hierarchical clustering to 151 abstracts to explore the data structure and identify similarities between studies. This method reveals natural groupings and uncovers multi-level patterns and these groupings provide a detailed understanding of how different research works are related.

First, we loaded the SentenceTransformer model “all-mpnet-base-v2” (Hugging Face, n.d.) and encoded the abstracts to generate embeddings. An embedding is a numerical representation of a sentence in a continuous vector space. Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) is a pre-trained language model developed by Google that uses transformers to understand the context of a word in search queries by considering both the words before and after it. SBERT (Reimers & Gurevych, 2019) is a modification of the BERT model that generates semantically meaningful sentence embeddings, allowing for efficient comparison of sentences. SBERT uses a pre-trained BERT model to convert sentences into dense vectors, capturing their semantic meaning. Subsequently, we applied dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP). UMAP is a dimension reduction technique that can be used for visualizations and for general non-linear dimension reduction. It is particularly effective for this purpose as it preserves the local and global structure of the data and it has been previously demonstrated to work well with embeddings (McInnes et al., 2020).

We then used an agglomerative strategy, which is a type of hierarchical clustering, that starts by treating each data point as its own cluster and then iteratively merges the closest pairs of clusters until a single cluster remains. To determine the optimal number of clusters, we combined this with silhouette score analysis, a method that measures how similar each data point is to its own cluster compared to other clusters. Higher silhouette scores indicate better-defined clusters. We then employed Euclidean distance as the distance metric, which calculates the straight-line distance

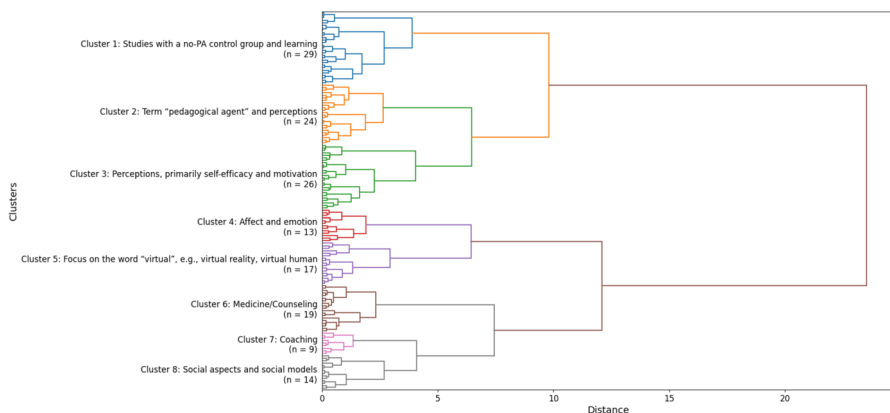
<sup>2</sup> The code and results for all clustering analyses are available on OSF [https://osf.io/5tb2y/?view\\_only=906db97a71434f6bbec865b1e6fe89aa](https://osf.io/5tb2y/?view_only=906db97a71434f6bbec865b1e6fe89aa).

between points in a multi-dimensional space. Ward's minimum variance method was used as the linkage criterion meaning we merged clusters in a way that minimizes the total within-cluster variance within all clusters while splitting samples in a hierarchy of groupings. Finally, we used the “maxclust” parameter in the fcluster function from the SciPy library. This divided the dataset into the final number of clusters. This approach allows us to evaluate the clustering quality by computing silhouette scores for different numbers of clusters, ultimately identifying the configuration with the highest silhouette score. In addition, we visualized hierarchical clustering using a dendrogram, to help us better understand the tree structure and hierarchical relationships between clusters. This visualization was enhanced with distinct colors for each cluster, providing a clear and interpretable representation of the clustering results. To interpret each cluster and evaluate its structures, thematic analysis was applied to generate code and themes based on abstracts within each cluster (Fig. 9).

### Hierarchical Clustering with SBERT Results

As shown in Fig. 9, we identified eight clusters with a moderate silhouette score of 0.44. The number of abstracts within each cluster ranges from 9 to 29, with Clusters 7 and 1 having the most abstracts. The dendrogram visually represents relationships between the clusters.

To evaluate the clusters and examine the themes for each, an inductive thematic analysis was conducted. A researcher with extensive experience in PA research reviewed all the abstracts within each cluster (blinded to the dendrogram) and identified themes in each cluster (see Fig. 9). A second researcher also reviewed and agreed with all themes. Interestingly, despite being blinded to the dendrogram when creating the themes for each cluster, these themes align very well with the structure of the dendrogram, where themes from Cluster 2 and Cluster 3, for example, are more related than Cluster 1, following the hierarchy.



**Fig. 9** The dendrogram for hierarchical clustering with 8 clusters ( $n$  = number of abstracts)

## Hierarchical Clustering with SBERT Discussion

One should note that there are many methods from which to choose when clustering data, and there are similarly a variety of feature extraction methods used for clustering. We provided our line of analytical reasoning, moving from TF-IDF and *k*-means clustering, to *k*-means clustering with SBERT, to our final arrival at SBERT with hierarchical clustering. However, it is worth noting that during the review process of this article, a reviewer made us aware of a published method of semi-automating scoping reviews, published by Mozgai et al. (2023). We have replicated their approach<sup>2</sup> and analyzed the results, however we believe the results from hierarchical clustering with SBERT are more promising for our purposes, which is why we have chosen to report it in the main body of the text. For clarity, the method by Mozgai et al. (2023) uses SPECTER, a document-level transformer model, which is specifically trained for review contexts as the specialized feature extraction method, in combination with *k*-means clustering (Cohan et al., 2020). SPECTER is conceptually similar to SBERT, which is used for transforming sentences into dense vectors to capture semantic similarities, but is particularly tailored for tasks involving review data, and consequently the results were similar to our results when SBERT was combined with *k*-means clustering. While interesting in its own way, similar to our results from SBERT with *k*-means clustering, we did not feel that the results from SPECTER moved us closer to the goal we sought to achieve.

In contrast, as demonstrated in our analysis, hierarchical clustering with SBERT can effectively categorize large sets of data into distinct, yet related, themes by capturing the semantic context and relationships within the text. In the case of our data, it provided richer and more nuanced representations of studies than LDA or *k*-means clustering with different feature extractions (TF-IDF, SBERT, or SPECTER). A key feature we found particularly advantageous of hierarchical clustering, as opposed to *k*-means, is that it provides a clear structural organization through a dendrogram. The dendrogram reveals the relationships between different clusters at various levels of granularity. This allows for a more nuanced understanding of data patterns by showing how clusters merge or split at different similarity thresholds. It offers flexibility in choosing the number of clusters post-analysis (see Fig. 9). This is in contrast to *k*-means clustering, which requires pre-specifying the number of clusters. With our dataset, we feel that hierarchical clustering with SBERT provided an efficient “starting point” for further qualitative analysis since it was able to identify patterns and trends within the data. Our results enabled us to focus on specific clusters of interest, reducing the time and effort required to manually sift through large volumes of abstracts.

However, hierarchical clustering has limitations. For example, it cannot determine the optimal number of clusters without additional validation methods and it may struggle with scalability for very large datasets (Vayansky & Kumar, 2020). Additionally, it relies heavily on the quality of the input data and the chosen distance metric, which can impact the accuracy and interpretability of the resulting clusters (Rüdiger et al., 2022). While hierarchical clustering excels at revealing relationships between data points and themes, it cannot provide probabilistic topic distributions

for each document, a strength of topic modeling that helps in understanding the mixture of themes within individual texts.

In summary, hierarchical clustering with SBERT addressed the key limitations of LDA for our use case in that we were able to cluster all of our studies into interpretable, meaningful themes. This allowed us to understand the high-level nature of the field quickly. However, we began to question if there was a way to combine the benefits of topic modeling with the benefits of hierarchical clustering with SBERT.

## Part V—BERTopic

Our search for a potential analysis to combine the benefits of topic modeling and clustering led us to an analysis set called BERTopic. In the literature, we found that BERTopic has proven useful in various text mining tasks for extracting meaningful insights from large corpora (Abuzayed & Al-Khalifa, 2021; Egger & Yu, 2022; Maphosa & Maphosa, 2023; Sánchez-Franco & Rey-Moreno, 2022).

BERTopic uses sentence embeddings, allowing for it to retain both contextual and semantic understanding from the documents. This gives it an advantage over LDA and is similar to using SBERT. In particular, this advantage is reflected in topic coherence and topic relevance metrics where it directly outperforms LDA (Abuzayed & Al-Khalifa, 2021). Moreover, a potential advantage over hierarchical clustering is its use of an alternate clustering approach: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which does not require specification of the number of clusters beforehand, allowing for a more accurate topic distribution shape (Grootendorst, 2022). It also uses class-based Term-Frequency Inverse Document Frequency (C-TF-IDF) for topic representations, which takes into account word frequency in relevance to the cluster (Grootendorst, 2022). This helps highlight words that are unique and significant within their respective clusters, identifying more distinct topics compared to the traditional BoW method utilized by LDA. BERTopic also provides options for additional fine-tuning through the use of additional representation models which may allow for better interpretability of topics.

## BERTopic Methods

We used BERTopic to perform topic modeling. By embedding text into high-dimensional vectors using BERT, BERTopic effectively retains semantic information, allowing for more accurate and coherent topic identification. We chose SBERT as the embedding generation method as it has been proven to build effective embeddings in other BERTopic use cases and for consistency in comparison with the previous method (Reimers & Gurevych, 2019). For dimensionality reduction, we used UMAP; it preserves more of the local and global features compared to other well-known methods for reducing dimensionality like Principal Component Analysis (McInnes et al., 2020).

Unlike hierarchical clustering where we had to determine the optimal number of clusters, the HDBSCAN clustering algorithm dynamically determines the optimal number of clusters based on the density of the data. It also differentiates itself from hierarchical clustering through its identification of outliers outside of finalized topics. It leaves these outliers as unclassified in order to, ideally, have a more cohesive collection of clusters that it generates. Clustering was conducted using HDBSCAN with the leaf clustering method, which produces more clusters, including smaller, less significant ones. This was done in order to have the maximum breakdown of topics for the highest potential of interpretation as well as to more closely reflect the amount of clusters generated in previous methods.

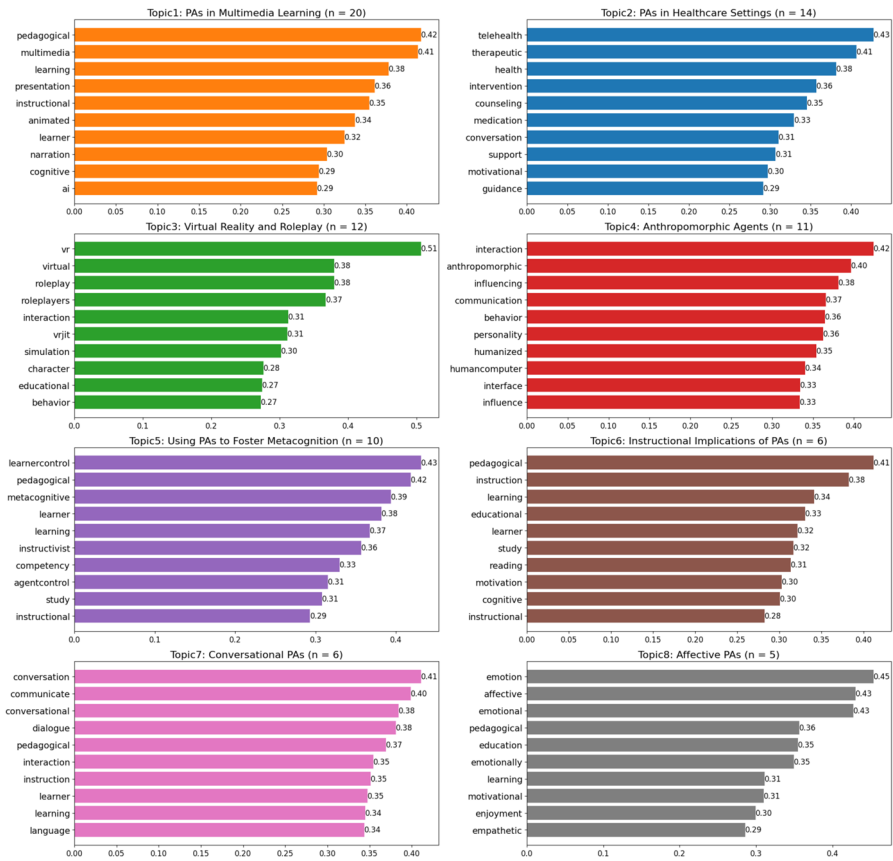
As noted, BERTopic applies C-TF-IDF to refine the topic representations. The class-based variation takes an additional step further by considering word frequency across the entire class rather than across individual documents which adjusts the keywords to topic relevance. In addition to these we also applied KeyBERTInspired, which utilizes the semantic relationship between keywords and the set of documents within each topic to further refine the topic representations. These techniques do not change the clusters, but they do aim to allow for better thematic interpretation with the top keywords having more contextual and semantic relevance to the topics. We also explored various other topic representation techniques that can be found in the Supplemental Materials 2 and 3.

## BERTopic Results

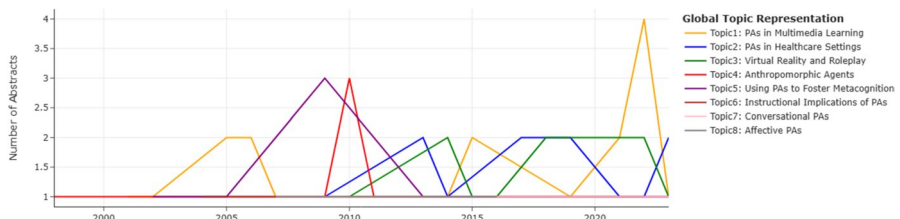
One method of representing the topics was by creating a visualization of the top keywords. As illustrated in Fig. 10, the top 10 keywords based on word score were identified for each topic ( $n=8$ ). We then used these to conduct thematic analysis and generated themes for topics.

In Fig. 10, we were able to identify the themes and abstract count of each topic generated. Nearly half of the documents ( $n=67$ , 44%) were determined by the clustering algorithm to be outliers, indicating noise or documents that did not fit well into any identified topic. These documents are not displayed in this figure as they do not have their own topic and do not have keywords generated. The number of abstracts across the topics ranged from 5 to 20, with the topic of “PA’s in Multimedia Learning” having the most and three topics sitting close together at the bottom: “Affective PAs,” “Conversational PAs,” and “Instructional Implications of PAs”. Outside of the top topic being more broad, the rest of the topics seem to concern particular niches within the field.

Similar to the previous analyses, we created a “Topics Over Time” graph in Fig. 11 to visualize the temporal distribution and evolution of these topics throughout the dataset. Few topics were notable except that anthropomorphic agents spiked during 2010 and PAs in multimedia learning saw a sharp increase in 2022, whereas other topics were more consistent over time.



**Fig. 10** Topic word scores ( $n$ =abstracts)



**Fig. 11** Topics over time

## BERTopic Discussion

The findings from using BERTopic provided an overview of the topic structure within the corpus, highlighting key areas of research and their relative prevalence. The use of HDBSCAN's leaf method ensures that even subtle and nuanced topics



are captured, while setting aside studies that do not fit within the cluster shapes. However, this may be viewed as either beneficial or detrimental: studies were not forced into topics they do not really belong in, but consequently you can also be left with studies that do not fall into an easily interpreted topic. With 67 (44%) documents unclassified in our sample, that is a substantial portion of the corpus BERTopic did not assign to a topic. BERTopic does offer options to reduce the number of unclassified topics through a variety of methods, but this would undermine the purpose of utilizing HDBSCAN.

Beyond the clustering and distribution of documents across topics, BERTopic offers topic keywords, similar to LDA. After all, BERTopic is a topic modeling approach, but it prioritizes keywords that are more contextually relevant to their respective themes than may be found in LDA. This can be advantageous for scoping reviews as it provides meaningful and rapid insights, potentially eliminating the need for more in-depth inductive thematic analysis if the studies are classified well and the topics are interpretable.

However, BERTopic also comes with some limitations. BERTopic does not account for multiple topics in a single document (in our case, a study abstract), which renders it unable to represent the nuance of topics within documents. It is also important to take into account that although BERTopic does utilize context in the distribution of topics with the use of SBERT and other sentence transformers, the topic representations are generated with C-TF-IDF. We attempt to improve these through the use of other representation models which may slightly improve interpretability, but results may vary with accuracy of document representation and interpretability.

In summary, BERTopic revealed a number of topics, but many studies were unclassified (44% of our sample). Therefore, similar to LDA, the results were not particularly helpful for those seeking alternatives for human coding during scoping reviews.

## Part VI—General Discussion

We began this study motivated by a single question: can we use analyses to reduce the amount of time humans spend coding when conducting scoping reviews? We investigated the impact of using different machine learning and data analytical approaches during the scoping review process. Consequently, it is important to briefly describe the implications of each analysis towards this goal before discussing the implications for others conducting scoping reviews.

The bibliometric analyses revealed specific clusters of collaborating authors, providing insights into systematic lines of research within the field. Interestingly, the analysis also indicated that relatively few research groups were actively working in this area. This enabled us to quickly identify and conduct background reading on the major lines of research. Such a step is often crucial when identifying potential moderators for meta-analyses or determining key variables for data extraction in systematic reviews (Hansen et al., 2022).

Compared to other methods employed in this study, bibliometric analysis was the easiest to set up, requiring minimal computer programming skills. This simplicity may explain why researchers increasingly adopt this method, as it offers valuable insights into research groups and publication trends. Despite these advantages, the bibliometric analysis struggled to uncover meaningful themes in abstract clusters, prompting us to explore alternative methods such as topic modeling with LDA for more detailed thematic exploration. LDA revealed a number of distinct topics within studies and categorized the studies into these respective topics. However, a substantial number of studies (79% of our sample) were assigned to topics that were not easily interpretable by our research team, and other researchers have encountered this issue as well (Lee et al., 2017; Maier et al., 2018; Yang et al., 2016). Analyzing the topics produced, we felt LDA was better for highlighting discrete, narrow areas of research than broader, coherent topics. For example, LDA was able to categorize a small number of studies as being around gaze, pediatric care, or cognitive load; topics that were not identified through other analytic techniques we tried. However, for our purposes of trying to understand the field as a whole rather than small niches within it, this analysis was not particularly useful.

We next explored clustering techniques. After exploring k-means clustering with a few different feature extraction methods, we felt that adding semantic context to the analysis and including a visual diagram of the relationships between clusters may facilitate interpretation, so we tested the use of hierarchical clustering with SBERT. This analysis produced much more interpretable clusters of studies from which we could determine distinct themes, and these themes were, in a sense, validated by a visual diagram that demonstrated the relationships among clusters. Moreover, all studies were assigned to a cluster. This analysis seems quite beneficial for categorizing all studies into existing clusters while maintaining a broad enough interpretation of a cluster to avoid the extremely precise and narrow topics identified by LDA. We suspect that accounting for semantic context via SBERT was an important component to the success we found through hierarchical clustering. This analysis helped us identify that there were likely enough studies for us to conduct a meta-analysis. In fact, we have pursued that study and located more than 50 studies that met our inclusion criteria. As such, we can view the clustering result as a rather conservative indicator in this case, but a great starting point none-the-less. Importantly, while human interpretation of themes may not always be reproducible, clustering results are if the proper code accommodations are prepared. For an example of a replicable analyses, please see our code in the OSF repository<sup>2</sup>.

Finally, we questioned if adding semantic context to topic modeling and using clustering via BERTopic would further clarify topics within the sample of studies. However, BERTopic produced results in which 44% of our studies were unclassified, and the issue of a substantial number of unclassified studies (also called outliers) has occurred in other studies as well (Egger & Yu, 2022). This leads us to believe that BERTopic may not be a great fit for this type of work if the goal is to understand the full corpus of studies rather than identify small topics within it. That said, the top words identified using BERTopic were more contextually relevant and coherent than those found from LDA, which is consistent with the results of another recent study (Abuzayed & Al-Khalifa, 2021). In addition, a smaller number

of studies were unclassified. Hence, if topic modeling is desired, BERTopic may provide a more useful approach than LDA in some contexts, especially as it did not try to force unlike studies into a topic.

In summary, we were able to identify various themes within the literature in our sample, and the different analyses helped us identify the size and scope of each. For example, from our hierarchical clustering with SBERT we decided to pursue a meta-analysis. We feel that using this approach was successful in our sample in the sense that the clustering revealed a conservative estimate of the number of studies in the area, enabling our further work without the need for human data extraction in the scoping review process. Another substantial implication of our analyses is that the bibliometric analyses helped us to identify who key author networks are in the field, which allowed us to begin focused reading of systematic lines of literature in the area quickly. Together, we feel that the bibliometric analysis and the hierarchical clustering with SBERT were the most impactful analyses for us. However, depending on the depth of the research questions in the scoping review, we argue these two analyses alone may not replace human data extraction, as discussed in the next section.

## Implications

After completing this exploration, we have a number of concrete implications for practice. Namely, when taking on a scoping review, the author must first identify what their *actual goal* is. While this seems obvious, the purpose of a scoping review is to understand the *nature* of the field (Tricco et al., 2018). We entered this study with this broad understanding. However, we quickly realized that this framing of the question did not actually move us specifically closer to what one may want to know. For example, if one wishes to identify specific topics within studies, LDA or BERTopic may be appropriate, as long as the possibility of uninterpretable topics or unclassified studies is acceptable. Meanwhile, hierarchical clustering with SBERT may be beneficial if one needs to understand how studies cluster together, are related, and the quantity of studies in each cluster. For our purposes, we found hierarchical clustering with SBERT to be the closest to what we had actually wanted to know about our sample. Upon reflection, what we truly wanted to know about our sample was whether there was an appropriate amount of literature to support a systematic review or meta-analysis. While a full scoping review would have provided us additional context, we found it was not necessary in our case. Similarly, if someone just wants to know the major research lines in the area, we found that bibliometric analysis of author co-occurrence was very helpful for that specific task.

A key question we must now face is: did we really complete a scoping review, and did these analyses reduce the need for human data extraction? Although the various methods differed in the information they provided us and we feel that we better understand our dataset than when we began, we argue that we did not complete a true scoping review with these methods. Nor do we feel that these methods alone are appropriate for conducting a true scoping review. Scoping reviews typically involve understanding more depth than only the major themes or topics in the

field as derived from abstracts. Consequently, we believe these methods may help those seeking to understand what themes are present in the literature and identify the potential for conducting a systematic review or meta-analysis, more than those aiming to conduct a *true*, in-depth scoping review. However, if one simply wants to understand high level trends in the field and they truly believe that information may be located within abstracts as opposed to requiring the full-text, then perhaps these methods may be a useful alternative. For example, we were primarily interested in knowing if enough studies existed for a meta-analysis, and our analyses clearly indicated that it was likely. We would argue however, that in that case one is not conducting a true scoping review, similar to how we do not believe we have conducted a true scoping review.

One advantage of these methods as opposed to purely human interpretation and categorization of abstracts into categories is in regards to reproducibility. While humans are fallible and may not interpret themes the same way on different occasions, between individuals, or based on the same criteria when reviewing abstracts, all the methods we focus on here are reproducible. We have included the necessary components of code to ensure that no matter how many times we ran the results, we would get the same result. This is one area in which the researcher trying to identify or categorize studies by abstract may find it particularly helpful to use algorithms in combination with human interpretation. That is, the algorithm can be used to identify the main topics (LDA) or clusters (clustering) and sometimes relationships between clusters (as with hierarchical clustering), and then humans can make their interpretations based on this defined set of criteria rather than ad-hoc criteria they may use when reviewing and categorizing based on potentially hundreds of abstracts.

Consequently, our study revealed to us just how important it is to truly understand what these analytic methods are doing, how they are doing it, and what the results mean. Table 2 describes the methods used in this study, what they do, how they do it, and what the results may mean on a conceptual level. It is important that the systematic reviewer interested in these methods really parse apart these factors when deciding if bibliometrics or machine learning might be able to replace human data extraction in their study. For example, if one were interested in understanding the types of PAs used in our studies, these analyses likely cannot replace human data extraction. Similarly, if one aimed to understand the research designs, participant populations, or how outcomes were measured, these analyses are unlikely to produce the desired results. However, if one wishes to know approximately how many studies there are around various topics in the field, hierarchical clustering or topic modeling may be a good starting point.

Finally, a key consideration is the question of do the methods we describe here save time? This should be examined from two different perspectives. First, we feel that these methods certainly save time compared to human data extraction for a true scoping review. For example, a review of resource use in the production of reviews found that extracting data from studies, such as key results and research design, took one person 41–65 min per study (Nussbaumer-Streit et al., 2021). Obviously, this timeframe will depend on the coding form, the transparency of the manuscript from which data is being extracted, and the individual's familiarity with the field, but

**Table 2** Summary of methods used and how they may contribute to scoping reviews in the future

Analysis	What each does in the context of our dataset (study abstracts)	How each works in the context of our dataset and methodological choices	What the results mean (strictly speaking)	What each does not tell us (from a scoping review perspective)
Bibliometric analysis	Identifying patterns or trends associated with title, keyword, abstract, and author data	Analyzes and creates network visualizations for word and citation co-occurrences	Identify publishing authors and research teams in the field, along with keyword networks	In-depth and contextual details about frameworks, study areas, research gaps, or methodological approaches
	LDA (using BoW)	Identifying latent topics with associated keywords within a text dataset	Analyzes the distribution of words and their co-occurrences across the abstracts and assigning probabilities to words for each topic and to topics for each abstract	Identify a set of “topics”, each is defined by a set of keywords
Hierarchical clustering with SBERT	Groups similar abstracts into clusters based on semantic similarity	Uses sentence embeddings from SBERT to represent documents, followed by agglomerative hierarchical clustering with Ward’s minimum variance method	Identifies hierarchical structure and relationships between clusters	In-depth thematic analysis within each cluster, contextual details about the development of themes
	BERTopic	Identifies clusters using semantic similarity and generates representations for those clusters based on keywords	Sentence embeddings, dimensionality reduction, and HDBScan to identify clusters; C-TF-IDF and KeyBERTInspired to generate representations	In-depth thematic analysis within each cluster, contextual details about the development of themes

based on our experiences with extracting data for scoping reviews (Schroeder et al., 2023; Zhang et al., 2024b), we feel this is a reasonably accurate estimate for our use case. Meanwhile, if one were familiar with machine learning methods, implementing the methods as we have here simply requires applying the template code (which we have provided via OSF) to their dataset: a process that takes only minutes. Even for the inexperienced, we do not see it as likely that implementing the bibliometric and machine learning methods as we have here would be more time-consuming than extracting all the data as one would during a true scoping review. This highlights the crux of the problem: are you conducting a true scoping review, in which our belief is that these methods may be nice supplements but not replacements for human data extraction, or are you simply trying to understand the quantity and nature of major themes in the field? If the latter, we still believe these methods to be more time-efficient than manual review and categorization of abstracts by a human, especially as the number of abstracts can be quite high. For example, in this study we had 151 abstracts to categorize.

To summarize, we do believe the methods we have described can make notable contributions to scoping reviews, but we do not see them as replacements for human data extraction in a true scoping review. Further, we believe the methods to be more time-efficient than having a human reading abstracts and categorizing them to try and quantify and classify the extent to which any particular theme is present in the literature.

## Conclusion

It is well-known that systematic reviews can be very time consuming to complete (Chernikova et al., 2024; Wang & Luo, 2024). While research around automated screening tools for speeding up the process of reviewing studies for potential inclusion is ongoing (Campos et al., 2024; Chernikova et al., 2024), our study focused on ways to increase the speed of the data extraction process. Extracting data from studies for the purposes of a scoping review can be a time-consuming endeavor, yet it often needs to be done to understand the nature of a field. We posed the question, can bibliometric analyses or machine learning techniques replace human data extraction for scoping reviews? Based on the results of our analyses, we concluded that it depends on if one is trying to accomplish a true scoping review, or if they actually wish to understanding the broader themes within the field. In other words, it is of the utmost importance that the reviewer understands what they actually want to learn from their work, rather than focusing first on the type of review they are conducting. While hierarchical clustering with SBERT worked well for showing us the high-level trends in the field, we believe that these analyses are not viable replacements for human data extraction in a true scoping review. Rather, they can be viewed as complementary. In particular, for those new to an area of research, hierarchical clustering with SBERT and bibliometric analyses may highlight productive reading paths more quickly if the goal is a high-level understanding of the field rather than a detailed understanding. In other words, whether or not analytics are a viable alternative to human

data extraction is a matter of grain size. If you are interested in broadly scoped, thematic questions (e.g., What areas of research exist within this field?) analytics may be helpful, whereas finer-grained questions (e.g., What types of PAs are used in studies?) are, with respect to the techniques tested here, better left to human data extraction.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10648-024-09972-0>.

**Funding** This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant DRL-2229612.

## Declarations

**Conflict of Interest** The authors declare no competing interests.

## References

- Abuzayed, A., & Al-Khalifa, H. (2021). Bert for Arabic topic modeling: An experimental study on bertopic technique. *Procedia Computer Science*, 189, 191–194. <https://doi.org/10.1016/j.procs.2021.05.096>
- Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, 60(3), 383–398. <https://doi.org/10.1007/s11423-012-9235-8>
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. <https://doi.org/10.1080/1364557032000119616>
- Baylor, A. L. (2011). The design of motivational agents and avatars. *Educational Technology Research and Development*, 59(2), 291–300. <https://doi.org/10.1007/s11423-011-9196-3>
- Baylor, A. L., & Plant, E. A. (2005). Pedagogical agents as social models for engineering: The influence of appearance on female choice. *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, 125, 65–72.
- Blair, C., Walsh, C., & Best, P. (2021). Immersive 360° videos in health and social care education: A scoping review. *BMC Medical Education*, 21(1), 590. <https://doi.org/10.1186/s12909-021-03013-y>
- Buenaño-Fernandez, D., González, M., Gil, D., & Luján-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An lda topic modeling approach. *IEEE Access*, 8, 35318–35330. <https://doi.org/10.1109/ACCESS.2020.2974983>
- Campos, D. G., Fütterer, T., Gfrörer, T., Lavelle-Hill, R., Murayama, K., König, L., Hecht, M., Zitzmann, S., & Scherer, R. (2024). Screening smarter, not harder: A comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. *Educational Psychology Review*, 36(1), 19. <https://doi.org/10.1007/s10648-024-09862-5>
- Chen, H., Wang, X., Pan, S., & Xiong, F. (2021). Identify topic relations in scientific literature using topic modeling. *IEEE Transactions on Engineering Management*, 68(5), 1232–1244. <https://doi.org/10.1109/TEM.2019.2903115>
- Chernikova, O., Stadler, M., Melev, I., & Fischer, F. (2024). Using machine learning for continuous updating of meta-analysis in educational context. *Computers in Human Behavior*, 156, 108215. <https://doi.org/10.1016/j.chb.2024.108215>
- Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. Preprint retrieved from <https://arxiv.org/abs/2004.07180>
- Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, 1(1), 104–126. <https://doi.org/10.1007/BF03177550>



- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint retrieved from <https://arxiv.org/abs/1810.04805>
- Domagk, S., & Niegemann, H. (2005). Pedagogical agents in multimedia learning environments: Do they facilitate or hinder learning? In *Towards sustainable and scalable educational innovations informed by the learning sciences* (pp. 654–657). IOS Press.
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, top2vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology*, 7, 886498. <https://doi.org/10.3389/fsoc.2022.886498>
- Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, (pp. 146–150). <https://doi.org/10.1145/2723576.2723589>
- Gottsacker, M., Norouzi, N., Schubert, R., Guido-Sanz, F., Bruder, G., & Welch, G. (2022). Effects of environmental noise levels on patient handoff communication in a mixed reality simulation. *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. <https://doi.org/10.1145/3562939.3565627>
- Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. Preprint retrieved from <https://arxiv.org/abs/2203.05794>
- Guo, Y. R., & Goh, D.H.-L. (2015). Affect in embodied pedagogical agents: Meta-analytic review. *Journal of Educational Computing Research*, 53(1), 124–149. <https://doi.org/10.1177/0735633115588774>
- Hansen, C., Steinmetz, H., & Block, J. (2022). How to conduct a meta-analysis in eight steps: A practical guide. *Management Review Quarterly*, 72(1), 1–19. <https://doi.org/10.1007/s11301-021-00247-4>
- Hassan, W., Martella, A. M., & Robinson, D. H. (2024). Identifying the most cited articles and authors in educational psychology journals from 1988 to 2023. *Educational Psychology Review*, 36(3), 1–25.
- Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, 6(1), 27–54. <https://doi.org/10.1016/j.edurev.2010.07.004>
- Hugging Face. (n.d.). sentence-transformers/all-mpnet-base-v2 [Transformer model]. Hugging Face, Inc. Retrieved July 26, 2024, from <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>
- Jaldi, C. D., & Schroeder, N. L. (2024). *Large language models for systematic review data extraction*. Association for Educational Communications and Technology Conference.
- Kolodkin, A., Boogerd, F. C., Plant, N., Bruggeman, F. J., Goncharuk, V., Lunshof, J., Moreno-Sanchez, R., Yilmaz, N., Bakker, B. M., Snoep, J. L., Balling, R., & Westerhoff, H. V. (2012). Emergence of the silicon human and network targeting drugs. *European Journal of Pharmaceutical Sciences*, 46(4), 190–197.
- Koprinkova-Hristova, P., Oubbati, M., & Palm, G. (2013). Heuristic dynamic programming using echo state network as online trainable adaptive critic. *International Journal of Adaptive Control & Signal Processing*, 27(10), 902–914.
- Krishnan, A. (2023). Exploring the power of topic modeling techniques in analyzing customer reviews: A comparative analysis. Preprint retrieved from <https://arxiv.org/abs/2308.11520>
- Lee, T. Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., & Findlater, L. (2017). The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105, 28–42. <https://doi.org/10.1016/j.ijhcs.2017.03.007>
- Li, J., & Jiang, Y. (2021). The research trend of big data in education and the impact of teacher psychology on educational development during COVID-19: A systematic review and future perspective. *Frontiers in Psychology*, 12, 753388.
- Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2–3), 93–118. <https://doi.org/10.1080/19312458.2018.1430754>




- Maphosa, V., & Maphosa, M. (2023). Artificial intelligence in higher education: A bibliometric analysis and topic modeling approach. *Applied Artificial Intelligence*, 37(1), 2261730. <https://doi.org/10.1080/08839514.2023.2261730>
- McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. Preprint retrieved from <https://arxiv.org/abs/1802.03426>
- Merigó, J. M., & Yang, J.-B. (2017). A bibliometric analysis of operations research and management science. *Omega*, 73, 37–48. <https://doi.org/10.1016/j.omega.2016.12.004>
- Merigó, J. M., Mas-Tur, A., Roig-Tierno, N., & Ribeiro-Soriano, D. (2015). A bibliometric overview of the journal of business research between 1973 and 2014. *Journal of Business Research*, 68(12), 2645–2653. <https://doi.org/10.1016/j.jbusres.2015.04.006>
- Motz, B., Busey, T., Rickert, M., & Landy, D. (2018). Finding topics in enrollment data. *International Educational Data Mining Society*.
- Mozgai, S., Kaurlooto, C., Winn, J., Leeds, A., Heylen, D., Hartholt, A., & Scherer, S. (2023). Machine learning for semi-automated scoping reviews. *Intelligent Systems with Applications*, 19, 200249. <https://doi.org/10.1016/j.iswa.2023.200249>
- Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18(1), 143. <https://doi.org/10.1186/s12874-018-0611-x>
- Nussbaumer-Streit, B., Ellen, M., Klerings, I., Sfetcu, R., Riva, N., Mahmić-Kaknjo, M., ... & Gartlehner, G. (2021). Resource use during systematic review production varies widely: a scoping review. *Journal of clinical epidemiology*, 139, 287–296.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. <https://doi.org/10.1136/bmj.n71>
- Plant, E. A., Baylor, A. L., Doerr, C. E., & Rosenberg-Kima, R. B. (2009). Changing middle-school students' attitudes and performance regarding engineering with computer-based social models. *Computers & Education*, 53(2), 209–215. <https://doi.org/10.1016/j.compedu.2009.01.013>
- Ranieri, M., Luzzi, D., Cuomo, S., & Bruni, I. (2022). If and how do 360° videos fit into education settings? Results from a scoping review of empirical research. *Journal of Computer Assisted Learning*, 38(5), 1199–1219. <https://doi.org/10.1111/jcal.12683>
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese bert-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Rosenberg-Kima, R. B., Plant, E. A., Doerr, C. E., & Baylor, A. L. (2010). The influence of computer-based model's race and gender on female students' attitudes and beliefs towards engineering. *Journal of Engineering Education*, 99(1), 35–44. <https://doi.org/10.1002/j.2168-9830.2010.tb01040.x>
- Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., & Doerr, C. E. (2007). The importance of interface agent visual presence: Voice alone is less effective in impacting young women's attitudes toward engineering. In Y. De Kort, W. IJsselstein, C. Midden, B. Eggen, & B. J. Fogg (Eds.), *Persuasive Technology* (vol. 4744, pp. 214–222). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-540-77006-0\\_27](https://doi.org/10.1007/978-3-540-77006-0_27)
- Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLoS ONE*, 17(4), e0266325. <https://doi.org/10.1371/journal.pone.0266325>
- Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing*, 39(2), 441–459. <https://doi.org/10.1002/mar.21608>
- Sandhiya, R., Boopika, A. M., Akshatha, M., Swetha, S. V., & Hariharan, N. M. (2022). A review of topic modeling and its application. In *Handbook of Intelligent Computing and Optimization for Sustainable Development* (pp. 305–322). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119792642.ch15>
- Schroeder, N. L., & Adesope, O. O. (2014). A systematic review of pedagogical agents' persona, motivation, and cognitive load implications for learners. *Journal of Research on Technology in Education*, 46(3), 229–251. <https://doi.org/10.1080/15391523.2014.888265>

- Schroeder, N. L., Romine, W. L., & Kemp, S. E. (2023). A scoping review of wrist-worn wearables in education, and career opportunities. *Computers and Education Open*, 5, 100154. <https://doi.org/10.1016/j.caeo.2023.100154>
- Siegle, R. F., Schroeder, N. L., Lane, H. C., & Craig, S. D. (2023). Twenty-five years of learning with pedagogical agents: History, barriers, and opportunities. *TechTrends*, 67(5), 851–864. <https://doi.org/10.1007/s11528-023-00869-3>
- Slimi, Z., & Carballido, B. V. (2023). Systematic review: AI's impact on higher education - learning, teaching, and career opportunities. *TEM Journal*, 1627–1637. <https://doi.org/10.18421/TEM123-44>
- Sperandeo, R., Messina, G., Iennaco, D., Sessa, F., Russo, V., Polito, R., Monda, V., Monda, M., Messina, A., Mosca, L. L., Mosca, L., Dell'Orco, S., Moretto, E., Gigante, E., Chiacchio, A., Scognamiglio, C., Carotenuto, M., & Maldonato, N. M. (2020). What does personality mean in the context of mental health? A topic modeling approach based on abstracts published in PubMed over the last 5 years. *Frontiers in Psychiatry*, 10. <https://doi.org/10.3389/fpsyt.2019.00938>
- Sperling, K., Stenberg, C.-J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open*, 6, 100169. <https://doi.org/10.1016/j.caeo.2024.100169>
- Su, J., & Yang, W. (2022). Artificial intelligence in early childhood education: A scoping review. *Computers and Education: Artificial Intelligence*, 3, 100049. <https://doi.org/10.1016/j.caeai.2022.100049>
- Sukthankar, G., & Sycara, K. (2011). Activity recognition for dynamic multi-agent teams. *ACM Transactions on Intelligent Systems and Technology*, 3(1). <https://doi.org/10.1145/2036264.2036282>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Kastner, M., Levac, D., Ng, C., Sharpe, J. P., Wilson, K., Kenny, M., Warren, R., Wilson, C., Stelfox, H. T., & Straus, S. E. (2016). A scoping review on the conduct and reporting of scoping reviews. *BMC Medical Research Methodology*, 16(1), 15. <https://doi.org/10.1186/s12874-016-0116-4>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garrity, C., ... Straus, S. E. (2018). Prisma extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- van Eck, N. J., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. <https://doi.org/10.1007/s11192-009-0146-3>
- Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582. <https://doi.org/10.1016/j.is.2020.101582>
- Wang, Y., Bowers, A. J., & Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of eqa articles from 1965 to 2014. *Educational Administration Quarterly*, 53(2), 289–323. <https://doi.org/10.1177/0013161X16660585>
- Wang, Y., Gong, S., Cao, Y., Lang, Y., & Xu, X. (2023). The effects of affective pedagogical agent in multimedia learning environments: A meta-analysis. *Educational Research Review*, 38, 100506. <https://doi.org/10.1016/j.edurev.2022.100506>
- Wang, X., & Luo, G. (2024). *Metamate: Large language model to the rescue of automated data extraction for educational systematic reviews and meta-analyses*. <https://doi.org/10.35542/osf.io/wn3cd>
- Wang, J., Li, X., Pan, L., & Zhang, C. (2021). Parametric 3d modeling of young women's lower bodies based on shape classification. *International Journal of Industrial Ergonomics*, 84. <https://doi.org/10.1016/j.ergon.2021.103142>
- Yang, K., Cai, Y., Chen, Z., Leung, H., & Lau, R. (2016). Exploring topic discriminating power of words in latent dirichlet allocation. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2238–2247). The COLING 2016 Organizing Committee. <https://aclanthology.org/C16-1211>
- Yu, Y., Li, Y., Zhang, Z., Gu, Z., Zhong, H., Zha, Q., Yang, L., Zhu, C., & Chen, E. (2020). A bibliometric analysis using VOSviewer of publications on covid-19. *Annals of Translational Medicine*, 8(13), 816. <https://doi.org/10.21037/atm-20-4235>
- Zhang, S., Li, H., Li, H., Botelho, A. F., & Israel, M. (2024a). Investigating the dynamic change of pre-and in-service teachers' experiences, attitudes, and perceptions through CS autobiography using topic modeling. In *Proceedings of the 17th international conference on educational data mining* (pp. 921–926). <https://doi.org/10.5281/zenodo.12729999>
- Zhang, S., Jaldi, C. D., Schroeder, N. L., & Gladstone, J. R. (2024b). Pedagogical agents in K-12 education: a scoping review. *Journal of Research on Technology in Education*, 1–28. <https://doi.org/10.1080/15391523.2024.2381229>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

**Shan Zhang<sup>1</sup> · Chris Palaguachi<sup>2</sup> · Marcin Pitera<sup>2</sup> · Chris Davis Jaldi<sup>3</sup> · Noah L. Schroeder<sup>1</sup>  · Anthony F. Botelho<sup>1</sup> · Jessica R. Gladstone<sup>2</sup>**

✉ Noah L. Schroeder  
schroedern@ufl.edu

<sup>1</sup> University of Florida, Gainesville, FL 32611, USA

<sup>2</sup> University of Illinois Urbana-Champaign, Champaign, IL 61820, USA

<sup>3</sup> Wright State University, Dayton, OH 45435, USA