SCIENCE CHINA
Information Sciences

• REVIEW •

May 2026, Vol. 69, Iss. 5, 151101:1–151101:25
https://doi.org/10.1007/s11432-025-4816-6

A literature review of literature reviews in pattern analysis and machine intelligence

Penghai ZHAO[1], Xin ZHANG[1], Jiayue CAO[1], Ming-Ming CHENG[1,2], Jian YANG[1] & Xiang LI[2,1*]

[1]*Tianjin Key Laboratory of Visual Computing and Intelligent Perception, College of Computer Science, Nankai University, Tianjin 300350, China*
[2]*Nankai International Advanced Research Institute (SHENZHEN FUTIAN), Shenzhen 518045, China*

Received 1 April 2025/Revised 4 September 2025/Accepted 2 November 2025/Published online 30 March 2026

Abstract The rapid growth of research in pattern analysis and machine intelligence (PAMI) has rendered literature reviews essential for consolidating and interpreting knowledge across its many subfields. In this work, we present a comprehensive tertiary analysis of PAMI reviews along three complementary dimensions: (i) identifying structural and statistical regularities in existing surveys; (ii) developing quantitative strategies that help researchers navigate and prioritize within the expanding review corpus; and (iii) critically assessing emerging AI-generated review systems. To support this study, we construct RiPAMI, a large-scale database containing more than 3000 review articles, and combine narrative synthesis with statistical analysis to capture structural and content-level features. Our analyses reveal distinctive organizational patterns as well as persistent gaps in current review practices. Building on these insights, we propose practical, article-level strategies for indicator-guided navigation that move beyond simple citation counts. Finally, our evaluation of state-of-the-art AI-generated reviews indicates encouraging advances in coherence and organization, yet also highlights enduring weaknesses in reference retrieval, coverage of recent work, and the incorporation of visual elements. Together, these findings provide both a critical appraisal of existing review practices and a forward-looking perspective on how AI-generated reviews can evolve into trustworthy, customizable, and transformative complements to traditional human-authored surveys.

Keywords AI-for-Research, literature review, umbrella study, AI-generated review, bibliometrics

Citation Zhao P H, Zhang X, Cao J Y, et al. A literature review of literature reviews in pattern analysis and machine intelligence. Sci China Inf Sci, 2026, 69(5): 151101, https://doi.org/10.1007/s11432-025-4816-6

1 Introduction

In both natural and artificial systems, entropy exhibits an intrinsic tendency to increase over time, gradually driving structures toward disorder. A comparable phenomenon occurs in the realm of human knowledge. As research outputs grow rapidly, the knowledge system accumulates scattered and overlapping information, which can lead to redundancy and inefficiency in creating new insights. To counter this trend, literature reviews function as an essential organizing mechanism. Much like gravity that shaped the early universe from dispersed particles into coherent structures, reviews gather fragmented research findings and arrange them into a meaningful whole. A literature review therefore serves not only to demonstrate familiarity with the academic work on a given topic but also to position that work within a broader context. By compiling and synthesizing the most relevant publications, it provides a clear and comprehensive overview of a field and supports the efficient advancement of knowledge.

Nearly every field features its own body of literature reviews, particularly in the rapidly evolving domain of pattern analysis and machine intelligence. As a foundational technology and research direction for AI, it spans multiple areas such as image classification [1–4], image segmentation [5–8], object detection [9–12], and natural language processing [13–15]. As reported by the AI Index Report [16], there has been a striking surge in artificial intelligence publications, soaring from 200000 in 2010 to nearly 500000 by 2021. This exponential increase has subsequently led to a proliferation of related literature reviews. This trend is clearly illustrated in Figure 1, which shows a marked increase in the annual publication of reviews, underscoring the growing prevalence of literature reviews in the field.

The increasing number of literature reviews across various AI subfields has raised several concerns. First, from the authors' perspective, the characteristics of literature reviews in the pattern analysis and machine intelligence (PAMI)

* Corresponding author (email: xiang.li.implus@nankai.edu.cn)

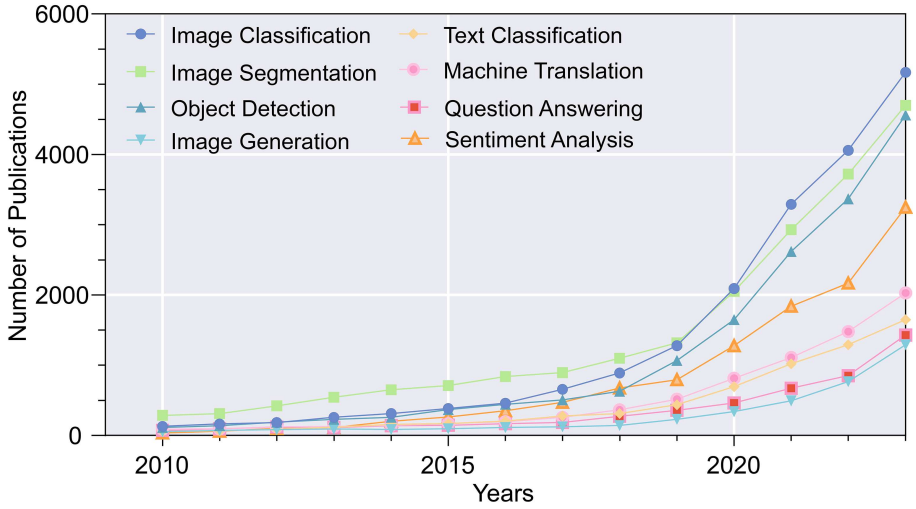

Figure 1 (Color online) Annual publication trends of literature reviews in the field of PAMI. A notable rising trend can be observed since 2015, reflecting an increasing scholarly focus and growing recognition of the importance of review articles in synthesizing the state of research in PAMI. The data were collected from the Google Scholar search engine.

field remain underexplored, including aspects such as structural conventions, statistical patterns, and adherence to established review methodologies. Second, researchers face difficulties navigating the rapidly expanding body of literature reviews, as the sheer volume within the same field makes it harder to identify which reviews are most relevant to their research objectives. Lastly, with the growing prevalence of AI-generated reviews, it becomes crucial to assess their benefits, limitations, and overall reliability compared to traditional human-authored reviews.

1.1 Research question and contribution

To provide a coherent exploration of literature reviews in the PAMI field, this study is organized around three interrelated research questions that progress from current understanding, to practical strategies, and finally to future prospects.

• **RQ1:** *What structural conventions and statistical patterns characterize literature reviews in the PAMI field?* To address this question, we construct the RiPAMI database containing more than 3000 review papers and apply both narrative synthesis and statistical analysis. Specifically, we analyze section structures, authorship patterns, reference and citation distributions, as well as visual element usage. Using LLM-based information extraction, we further quantify six representative content features (e.g., taxonomy, PRISMA, benchmark) and investigate their prevalence across sub-fields and temporal trends. This yields a comprehensive mapping of structural and statistical practices in PAMI reviews.

• **RQ2:** *Given these observed patterns, how can quantitative bibliometric indicators be designed and applied to help researchers efficiently navigate and select among the growing number of reviews?* To answer this question, we design four novel, real-time, article-level, and field-normalized bibliometric indicators: the Topic Normalized Citation Success Index ($TNCSI$), the Impact Evolution Index (IEI), the Reference Quality Measurement (RQM), and the Review Update Index (RUI). We validate these indicators using the RiPAMI database, examining their mathematical properties, interpretability, and correlations with review impact. We further illustrate their practical utility by showing how they support indicator-guided navigation, enabling researchers to efficiently screen and prioritize reviews beyond simple citation counts.

• **RQ3:** *When applied to emerging AI-generated reviews, what strengths and weaknesses can be observed compared with human-authored surveys, and what do these observations imply for their reliability and practical utility?* To answer this question, we conduct a systematic evaluation of several state-of-the-art automated review generation systems. Our analysis considers their pipeline design (e.g., intention analysis, retrieval, synthesis, report generation), reference selection strategies, structural organization, and use of visual elements. The results reveal clear patterns: recent systems are able to generate coherent and reasonably organized reviews, sometimes enriched with figures or taxonomies, yet they still suffer from critical shortcomings. These include a tendency to over-rely on highly cited but outdated references, limited capacity to recognize and integrate very recent work, and insufficient incorporation of explanatory visuals or appraisal criteria, etc. Such findings not only highlight the current gap between automated and human-authored reviews, but also underscore the importance of improving reliability,

transparency, and customization if AI-generated reviews are to become practically useful in scholarly practice.

In summary, we present a literature review of literature reviews and discuss the common concerns faced by existing reviews in the PAMI field. To the best of our knowledge, there has been limited scholarly attention to addressing these considerations. All the data and code framework used in this paper are publicly available at https://sway.cloud.microsoft/2TXEuPuNlDKEmC9p.

1.2 Organization of the paper

The remainder of the paper is organized as follows. Section 2 briefly introduces the criteria for literature screening, methods for the RiPAMI construction, and formulas for metric calculations. Section 3 integrates narrative synthesis with statistical examination, providing both a qualitative account of structural and content patterns and a quantitative mapping of meta-data features across existing surveys. Section 4 further discusses how the proposed metrics can assist in efficiently selecting the proper reviews. The characteristics of human-authored versus AI-generated literature reviews are discussed in Section 5. In Section 6, we explore the challenges and future prospects of literature reviews. Finally, the paper concludes in Section 7.

2 Methodology

2.1 Scope

This study examines all fourteen types of review articles as classified in Grant's comprehensive typology [17], including narrative review, systematic review, and state-of-the-art review. While this research is technically a tertiary or "umbrella" study, we opted for the more commonly understood term "literature review" to improve clarity and acceptance within the PAMI field. In addition, although the quantity of reviews is considerably smaller compared to that of normal papers, it remains impractical to analyze reviews within every field. Therefore, this paper will focus only on reviews in the PAMI field.

2.2 Literature selection criteria for narrative reviews

To support the analysis in Subsection 3.1, we establish specific criteria for selecting literature included in the narrative review. The search process draws on Google Scholar, Semantic Scholar, IEEE Xplore, and ScienceDirect, covering publications from major publishers such as IEEE, Elsevier, Springer, and others.

The filtering procedure follows three steps. First, we compile 106 search keywords derived from the scopes of leading journals and conferences in the PAMI domain, including terms such as *speech recognition*, *optical character recognition*, and *self-supervised learning*. Second, we require that a paper's title contains either "survey" or "review," and that the chosen keyword appears in its title or abstract. Finally, we exclude preprints not published in journals or conference proceedings, ensuring that all selected articles are peer-reviewed.

These criteria provide a focused and high-quality set of review articles, which serve as the basis for the narrative analysis presented in Subsection 3.1.

2.3 Literature selection criteria for statistical analysis

The immense number of literature reviews in the PAMI field emphasizes the need for approaches beyond manual selection, analysis, and synthesis. Therefore, to ensure a comprehensive and accurate analysis of literature reviews in the PAMI field, we develop an automated process that simulates the manual selection procedure to construct the RiPAMI database.

2.3.1 *Data source*

Reliable data sources for analyzing extensive reviews are fundamentally important. Based on the means of data acquisition and storage, existing scientific scholar data sources may be classified into two main categories: web-based and snapshot-based sources. Web-based source data refers to the meta-data that can be retrieved from the data provider in real-time with the use of the web crawler or the API (e.g., Semantic Scholar, arXiv, CrossRef). Such an approach would only consume a small amount of storage on the local machine, but it needs to query meta-data every single time. On the contrary, the offline snapshot consumes a larger amount of storage space but eliminates the need for frequent API queries. Since the snapshot is a mirror image of relevant papers before a specific time, its data remain unchanged over time compared to the API-based sources. As a result, the snapshot ensures a

Table 1 Comparisons between various sources. "Counts Only" in the citations column means that the source only records the citation counts, while "Complete" signifies that the data source provides a complete list of citations.

Database	Title & Authors	Venue	Abstract	Citations	References	Source types	Free
arXiv	✓	×	✓	×	×	API-based	✓
CrossRef	✓	✓	✓	Counts Only	✓	API-based	✓
Google Scholar	✓	✓	✓	Complete	×	Crawler-based	✓
IEEE Xplore	✓	✓	✓	Counts Only	×	API-based	✓
Semantic Scholar	✓	✓	✓	Complete	✓	API-based	✓
Web of Science	✓	✓	✓	Complete	✓	API-based	×
Scopus	✓	✓	✓	Complete	✓	API-based	×
arXiv Data File	✓	×	✓	×	×	Snapshots	✓
CrossRef Data File	✓	✓	✓	Counts Only	✓	Snapshots	✓
RiPAMI (Ours)	✓	✓	✓	Complete	✓	Snapshots	✓

consistent dataset in all of the experiments, avoiding issues of irreproducibility caused by changes in the provided web-retrieving service.

Table 1 compares various commonly used data sources. However, none of these sources is perfect. For example, arXiv is a free distribution service and an open-access archive for millions of scholarly articles in various fields. However, the arXiv data source fails to provide the publication venue, citations, and references. Semantic Scholar seems promising, but it suffers from a lower update rate than arXiv and a narrower search scope than Google Scholar. Google Scholar is an online search engine that indexes scholarly literature from a wide range of disciplines. Although Google Scholar uses powerful automated programs to retrieve files for inclusion in its search results, it still faces challenges in many areas. Beel [18–20] argues that Google Scholar places a high weight on citation counts in its ranking algorithm and has therefore been criticized for exacerbating the Matthew effect [21]. Moreover, the citation counts displayed on Google Scholar are subject to manipulation by complete nonsense articles indexed on Google Scholar (e.g., citations from AI-generated pre-print papers published on arXiv should have been ignored). Therefore, a promising engineering solution is to leverage the strengths of various approaches to overcome the weaknesses of each approach, as will be detailed in the next subsection.

2.3.2 *Database construction*

To ensure reproducible experiments and prevent overburdening the server, we construct an SQL-based database dubbed RiPAMI (Reviews in Pattern Analysis and Machine Intelligence, pronounced as /ri:pæmi/). This database stores information related to the paper, such as title, abstract, date of publication, venue, citation counts, and reference details.

As illustrated in Figure 2, we implement a structured process to construct the RiPAMI database, ensuring that the selected articles are review papers relevant to the pattern recognition field. The process begins with searching with keywords derived from the scopes of relevant journals and conferences (check Supplementary Material for the entire keywords list). These keywords are then formatted into query strings for the arXiv API as follows: (ti:"review" OR ti:"survey") AND (ti:"keyword.lower()" OR abs:"keyword.lower()"). This structure specifically retrieves articles whose titles or abstracts include both the terms "review" or "survey" and the specified keywords. Additionally, we apply regular expressions to check that the abstracts of returned articles include the relevant keywords, ensuring that only appropriate review articles enter the RiPAMI database. To avoid potential copyright and licensing issues, papers are retrieved and downloaded using the arXiv's API. Calls to the API are made by means of HTTP requests to a certain URL. The responses will be parsed and stored in an SQL-based database. Through this process, we collect a total of 4106 samples. To further refine our database, we employ a double-checking phase. Papers are first preliminarily classified using a GPT-based filter, followed by a manual validation step in which four junior annotators conduct the initial screening and two advanced annotators provide the final confirmation to ensure the removal of false positives. Lastly, we supplement the remaining 3099 entries with additional meta-data, including the publication date, citation counts, and reference details. As mentioned in Subsection 2.3.1, we suggest enriching the meta-data of papers by leveraging a combination of disparate data sources. Considering the potential legal risks of crawling to obtain academic data from Google Scholar, the Semantic Scholar API was employed to obtain additional meta-data, such as citation and reference details, which are not provided by the arXiv API. The final SQL-based snapshot encompasses a wide range of data, including ID, title, publication date, citations, and more, facilitating efficient data retrieval and statistical analysis.

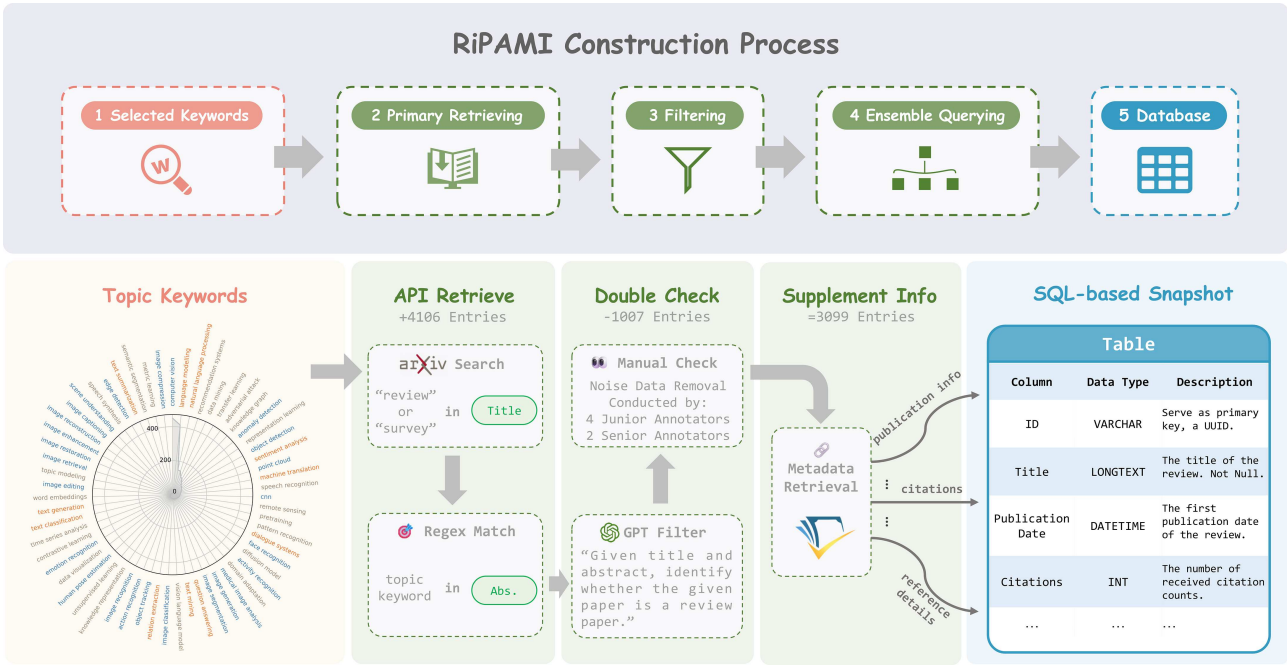

Figure 2 (Color online) A schematic overview of the database construction process. From initial keyword selection to the final SQL-based RiPAMI snapshot, three key steps are implemented to ensure the data in RiPAMI are clean, accurate, and reliable. For simplicity, the polar diagram shows only a subset of the keywords used for retrieval. Since citation counts fluctuate over time, the retrieval date for related information is set to October 2024.

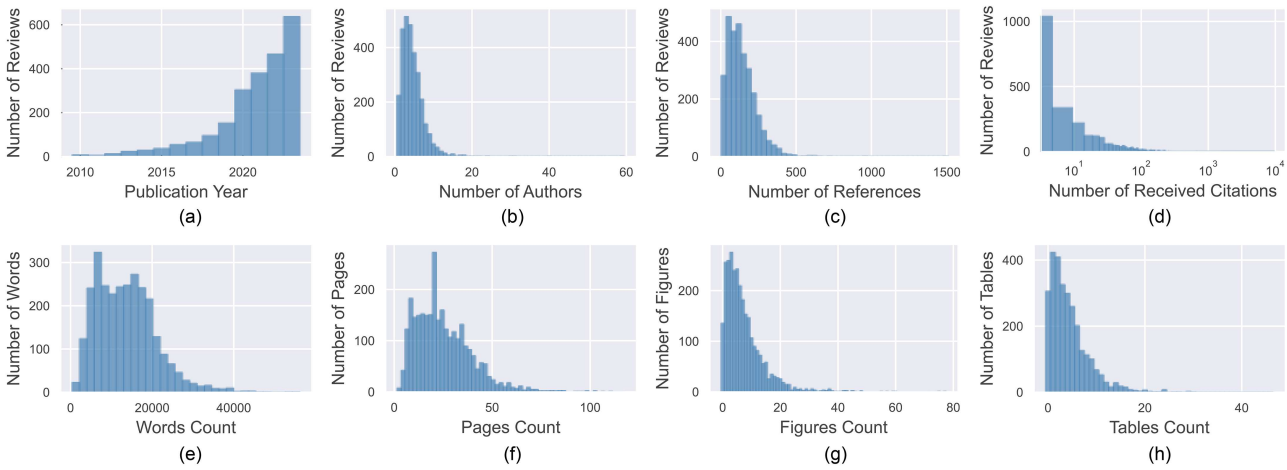

Figure 3 (Color online) Statistics of the RiPAMI database. The dataset spans a wide range of publication years, citation impact, and reference counts, providing a representative overview of literature reviews in the PAMI field. (a) Distribution of publication years; (b) distribution of the number of authors; (c) distribution of the number of references; (d) distribution of received citations; (e) distribution of word counts; (f) distribution of page counts; (g) distribution of figures; (h) distribution of tables.

2.3.3 *Meta-data statistics of RiPAMI*

The database consists of more than 3000 literature reviews from a variety of sources, publication years, and fields. To elucidate the characteristics of the RiPAMI database, we conduct a statistical analysis and plot the result in Figure 3.

Years of publication. Figure 3(a) presents the distribution of publication years for literature reviews in the RiPAMI database. A clear upward trend in the number of reviews over time is evident, mirroring the pattern observed in Figure 1. Both figures show a steady increase, with a notable surge between 2019 and 2020. This trend reflects a strong alignment between our sample data and the original dataset in terms of statistical consistency.

Number of the authors. A review paper typically aims to provide a comprehensive overview of a specific topic. Involving multiple authors with diverse expertise could enhance the depth and breadth of the review. Figure 3(b)

Table 2 Statistical meta-data summary of RiPAMI.

Attribute	Max	Min	Mean	Median	Mode
Pub year	2024	2010	2021	2022	2024
Authors	59	1	4.61	4	3
Refs	1700	1	138.19	120	66
Cites	11578	0	78.48	11	0
Words	57883	393	13613.57	12974	5903
Pages	118	2	25.24	22	20
Figs	77	0	7.42	6	3
Tabs	47	0	4.38	3	1

indicates that the majority of reviews are written by fewer than 10 authors.

Number of the references. The number of references in a survey paper may influence its credibility and reliability. As shown in Figure 3(c), the distribution of literature review references in the RiPAMI database follows a log-normal pattern.

Number of citations. We count the citations of papers in RiPAMI and plot them in Figure 3(d). A power law distribution of received citations could be found. This phenomenon where a small subset of papers receives the majority of citations is sometimes referred to as the "Matthew Effect" or the "Pareto principle", as reported in [21, 22].

Length of literature reviews. The words count and the number of pages of a review article can partially reflect its depth and breadth. In general, more detailed and comprehensive reviews tend to contain a higher word count and longer pages, thus requiring more time and effort to conduct and comprehend. As shown in Figures 3(e) and (f), review articles in the PAMI field typically range from 5000 to 20000 words and span 10 to 40 pages. At an average reading speed of 240 words per minute [23], reading a full review would take over 20 min.

Number of the visual elements. Visual elements refer to visual representations of information, such as images and charts. They provide a clear presentation of data, simplify complex content, and enhance reader comprehension. By employing LLMs for information extraction, we may ascertain the number of figures and tables each review article contains, as depicted in Figures 3(g) and (h). It can be observed from our analysis that the majority of review articles contain fewer than 10 figures or tables.

Further statistical details, such as maximum (Max), minimum (Min), mean, median, and mode, are available in Table 2.

2.4 Bibliometric indicators for reviews

Literature reviews reflect the state of their respective fields, highlighting current research priorities and offering a glimpse into emerging trends. Despite their significance, systematic methods for evaluating literature reviews remain underdeveloped. By introducing quantitative tools, bibliometric methods offer a promising solution, enabling researchers to objectively analyze literature reviews and uncover key patterns and trends. Bibliometrics is a research field focused on the use of quantitative methods and statistical analysis to evaluate various aspects of scholarly publications [24]. These methods and analysis, including widely recognized metrics like citation counts and impact factors, offer researchers valuable insights in their daily scientific work, aiding them in making prompt and informed decisions. However, most existing bibliometric methods suffer from several limitations including unfair comparison, misuse, and manipulation as introduced in the "San Francisco Declaration on Research Assessment" (DORA) [25] and the Leiden Manifesto [26]. Notably, many of these metrics were not specifically designed for evaluating literature reviews, further limiting their applicability in this context.

In light of these limitations, we propose two new impact indicators, $TNCSI$ and IEI, along with two quality indicators, RQM and RUI.

2.4.1 *TNCSI*

Most existing bibliometric methods face challenges in conducting cross-disciplinary comparisons due to variations between fields [27–30]. Although a few methods offer the capability for cross-disciplinary comparison, they either require manual assignment of field-specific keywords or involve complex and costly retrieval to calculate metrics. These drawbacks have, to some extent, hindered individual researchers from adopting these methods, thereby limiting the broader application of the metrics.

To fairly compare the external influence of literature reviews across different fields within an acceptable cost, we propose guiding large language models (LLMs) to generate the topic key phrase and then calculating the success

index of the current review within the same topic (see Appendix for details). The $TNCSI$ estimates the impact of research publications within an LLM-generated topic by normalizing citation counts to a scale between 0 and 1. The formula for calculating the $TNCSI$ is

$$TNCSI = \int_0^{\text{citeNum}} \lambda e^{-\lambda x}\, dx, x \geqslant 0, \tag{1}$$

where λ is the scale parameter obtained through maximum likelihood estimation. Detailed steps for estimating λ and constructing the probability density function are provided in the Supplementary Material.

The $TNCSI$ demonstrates favorable mathematical properties and interpretability. First, the $TNCSI$ algorithm employs maximum likelihood estimation to convert the probability mass function into a probability density function. This process ensures that, in theory, the $TNCSI$ differentiates between papers with distinct citation counts, avoiding the assignment of identical values to them. Second, the $TNCSI$ possesses physical significance, representing the probability that a specific paper's citation count surpasses that of any other paper on the same topic. For example, a paper with a $TNCSI$ of 0.5 means it has more citations than half of the papers within the same topic. Furthermore, the calculation of $TNCSI$ is computationally efficient, as it avoids requiring a complete or precise ranking of all related papers. Instead, it provides a reliable estimation of the probability that the given paper surpasses others in terms of citation count within the same topic.

2.4.2 *IEI*

Imagine a scenario where two papers, A and B, receive the same number of citations. The number of new citations per month for A remains steady, whereas the number of new citations for B grows exponentially. In this context, while acknowledging the importance of Paper A, it is generally assumed that Paper B holds a greater reference value. Analyzing the popularity or citation trends of the literature may help researchers stay informed about the latest developments and identify potential areas for future research.

The IEI is defined as the average slope across l distinct points on a $n = l - 1$ degree Bézier curve representing the citation trend over time. The formula for calculating IEI_{L_l} is

$$IEI_{L_l} = \sum_{a=0}^{l-1} \frac{(y_i/x_i)}{l}, \tag{2}$$

where l represents the number of months observed. x_i and y_i are the components of the tangent vector at the a-th point on the Bézier curve, indicating the magnitude along the x- and y-axes, respectively. More details can be found in the Supplementary Material.

With the guidance of the IEI, researchers may further discern the various literature reviews within the same field that exhibit close citation counts. A higher IEI indicates that an increasing number of studies reference the review, signaling its growing influence and attention. Therefore, in practical applications, researchers may prioritize the study of more promising reviews, as indicated by higher positive IEI values, to mitigate redundancy and improve efficiency.

2.4.3 *RQM*

A literature review, in its essence, cannot fabricate insights from a void. It fundamentally relies on the substance of existing references. Without a solid foundation of credible and high-quality sources, a literature review may lack the necessary building blocks to construct a meaningful analysis or argument. These sources provide the empirical evidence and theoretical context that ground the review, making the role of references indispensable in the creation of a substantial literature review.

The RQM incorporates both the quality and timeliness of references by modifying the Gompertz function, which exhibits a sigmoidal growth pattern—slow at the beginning and end, with a rapid increase in the middle phase:

$$RQM = 1 - e^{-\beta \cdot e^{-(1-ARQ)\cdot S_{mp}}}. \tag{3}$$

In this equation, β represents the shift parameter, ARQ denotes the average reference quality, while S_{mp} refers to the median semester count of the reference age, indicating the time from the publication date of the references to the issuance of the review (which will be detailed in the Supplementary Material.

Beyond simply assessing the average impact of references, the RQM integrates references' timeliness and its influence on review quality. This metric addresses the tendency among authors to rely heavily on classic studies,

which may no longer fully align with current research trends. Additionally, the adjustment coefficient in RQM is calibrated through statistical and optimization algorithms, tailored to the unique characteristics of each research discipline. This data-driven approach minimizes reliance on heuristic parameter settings and increases RQM's adaptability across various academic disciplines.

2.4.4 *RUI*

The RUI refers to the measure of the extent to which a literature review is required to be updated due to the iteration of technology, theory, etc. The index is related to both the literature itself and the research interests of the topic. Generally, a high update index suggests that a literature review is in need of an immediate update. Conversely, a lower update index implies that few advances have been made to the investigated field, and the review is still up-to-date.

The RUI quantifies the necessity of updating a literature review by combining two indicators: the coverage difference ratio (CDR) and the review aging degree (RAD). The formula for calculating the RUI is

$$RUI = p \cdot CDR + q \cdot RAD, \tag{4}$$

where, p and q are weighting coefficients, set to 10 and 5, respectively. The 2:1 ratio ensures that CDR plays a dominant role, while the specific magnitudes primarily enhance numerical readability rather than altering the substantive balance between the terms. Detailed definitions and calculations of the CDR and RAD are provided in the Supplementary Material.

The proposed RUI integrates both the popularity of the research field and the natural aging of individual reviews. It reflects how reviews require updating over time while taking into account the pace of progress within the field. By balancing these factors, RUI offers real-time insights into how urgently a review needs updating. This urgency may arise from rapid developments in the field or the natural aging of the review's content. RUI moves beyond the simplistic reliance on publication dates, providing a more contextualized assessment of its timeliness.

2.5 Information extraction for reviews

Given that this study aims to identify the common features of reviews within the field, solely relying on manual inspection would be both impractical and resource-intensive. To further investigate the features of reviews in the PAMI field, we employed LLM-based information extraction techniques to analyze the review content and parse the response to structured results. The proposed approach converts unstructured text into structured data, facilitating further analysis and interpretation. Details regarding the information extraction method can be found in the Supplementary Material.

3 Review of reviews in PAMI

This section provides an integrated overview of the PAMI field by combining narrative synthesis and statistical analysis, aiming to identify common features of existing reviews and to uncover quantitative patterns across various sub-fields, thereby laying the foundation for subsequent in-depth discussion.

3.1 Narrative reviews in various sub-fields in PAMI

In this subsection, we offer a subjective analysis of reviews from diverse fields. Through this narrative approach, we aim to examine the compositional characteristics shared by reviews across various fields. It should be noted that only a subset of fields is discussed here for illustrative purposes, while the complete list of covered sub-fields is provided in the Supplementary Material.

3.1.1 *Computer vision*

Computer vision is one of the most popular subfields of pattern recognition and machine intelligence. This section will focus on the literature reviews within the realm of computer vision.

Image classification refers to the task of assigning a label or a category to an input image, which is one of the most renowned tasks in the field of computer vision. Rawat et al. [31] explored the development and advancements of deep convolutional neural networks (CNNs) in the field of image classification. Ref. [31] covered the historical context, their role in the deep learning renaissance, and the notable contributions and challenges faced in recent years. It highlights the remarkable progress of CNNs in image classification, while also acknowledging the

ongoing research efforts to address challenges and provide recommendations for future exploration. Schmarje et al. [32] provided a comprehensive survey on semi-, self-, and unsupervised learning methods for image classification. The survey compares and analyzes 34 different methods based on their performance and commonly used ideas, highlighting the trends and research opportunities in the field. Through comprehensive analysis, the authors reveal the potential of semi-supervised methods for real-world applications and identify challenges such as class imbalance and noisy labels. Furthermore, the paper emphasizes the importance of combining different techniques from various training strategies to improve overall performance. In addition to CNNs, there exist alternative techniques for image classification. Ref. [33] presented a comprehensive analysis of support vector machines (SVM) in image classification. It discusses various techniques that can enhance classification accuracy and highlights its advancements. Liu et al. [34] investigated more than 100 different visual Transformers comprehensively in three fundamental CV tasks, including classification, detection, and segmentation. They also propose a taxonomy to categorize various transformers into six groups.

Object detection entails identifying and localizing objects of interest within an image or video. Liu et al. [35] offered a comprehensive survey on the advancements in deep learning-based generic object detection. This paper discusses an extensive range of issues, including detection frameworks, taxonomies, feature depiction, training strategies, and evaluation metrics. Though there have been significant advancements in generic object detection, the detection of small objects, which focuses on identifying objects with a small size, still presents challenges. The review conducted by Cheng et al. [10] investigates 181 papers, constructs two large-scale datasets (SODA-D and SODA-A), and evaluates the performance of mainstream small object detection methods. Object detection demonstrates the utility and effectiveness across multiple domains. Li et al. [12] and Litjens et al. [36] investigated numerous methods and applications of object detection in remote sensing and medical image analysis, respectively, showing that these methods have the flexibility to be applied in various scenarios and meet different needs.

Image segmentation is the process of dividing an image into meaningful and distinct regions to facilitate analysis and understanding. As described earlier, Minaee et al. [6] proposed a taxonomy for image segmentation methods which divides models into 11 categories. In addition to the taxonomy, the authors evaluate the quantitative performances of various methods on popular benchmarks. The paper also identifies open challenges and proposes promising research directions for future advancements in deep-learning-based image segmentation. Given that most image segmentation algorithms heavily rely on expensive pixel-level annotations, interest in weakly supervised image segmentation methods has increased. Ref. [37] surveyed label-efficient deep image segmentation methods. According to the study, weakly supervised segmentation approaches can be categorized into four hierarchical types, ranging from no supervision to inaccurate supervision. The authors investigated each of these four methods in separate sections, highlighting the strategies used to bridge the gap between weak supervision and dense prediction. Image segmentation techniques have a wide range of applications in the field of medical image processing, as introduced in [7, 38–40].

3.1.2 *Natural language processing*

Acclaimed as the jewel of the artificial intelligence crown, natural language processing (NLP) stands as a pivotal domain within the field of PAMI. Here, we provide a further discussion on several popular NLP research directions.

Named entity recognition (NER) involves identifying and classifying named entities in text, such as person names, organizations, locations, and dates. The survey by Li et al. [41] begins by introducing NER resources, including tagged NER corpora and off-the-shelf NER tools. Then, authors categorize existing studies based on a taxonomy that considers distributed representations for input, context encoder, and tag decoder. The paper surveys representative methods for applying deep learning in various NER tasks, and provides a valuable reference for designing deep learning-based NER models. NER serves as the foundation technique for various natural language applications, such as relation extraction [42] and knowledge graph [43]. Due to the linguistic variance of different languages, NER methods may also vary from language to language. Surveys about various language-specific NER could be found in [44–46].

Sentiment analysis focuses on determining the sentiment or emotion expressed in text, such as positive, negative, or neutral. Yadav et al. introduced the process of gathering and analyzing people's opinions and sentiments from various sources such as social media platforms and blogs in their paper [47]. The paper evaluates and compares different approaches used in sentiment analysis, with a focus on supervised machine learning methods like Naive Bayes and SVM algorithms. The common application areas of sentiment analysis and the challenges involved in accurately interpreting sentiments are also reported. Yue et al. [48] categorized and compared a large number of techniques and methods from three different perspectives: task-oriented, granularity-oriented, and methodology-oriented. It also explores different types of data and advanced tools for research, highlighting their strengths and

limitations.

Language modeling involves training models to understand and generate human language. In early attempts, recurrent neural networks achieved desirable performance and wide application at that time, despite some shortcomings. Ref. [49] specifically focused on RNNs and long short-term memory (LSTM) cells. The authors highlight the limitations of traditional RNNs and emphasize the significance of LSTM in handling long-term dependencies. They discuss various LSTM cell variants and their performance on different characteristics and tasks. Furthermore, the paper also categorizes LSTM networks into two major types: LSTM-dominated networks which optimize connections between inner LSTM cells, and integrated LSTM networks which incorporate advantageous features from various components. Recently, LLMs have drawn widespread attention. LLMs demonstrate significant performance improvements and unique abilities such as in-context learning, setting them apart from smaller-scale models. By investigating more than 600 studies, Zhao et al. [50] conducted a comprehensive review of the recent advancements in LLMs. The authors discuss the evolution of language modeling techniques, from statistical models to neural models, and highlight the emergence of pre-trained language models as a powerful approach in NLP tasks. The survey focuses on LLMs with a parameter scale exceeding 10 billion and explores four key aspects: pre-training, adaptation tuning, utilization, and capacity evaluation. The paper also presents available resources for developing LLMs and discusses important implementation guidelines. Overall, this survey serves as an up-to-date and valuable reference for researchers and engineers interested in the field of LLMs.

3.1.3 *Others*

Reviews in other popular sub-fields will be investigated in this section.

Zhou et al. [51] covered the evolution of pre-trained foundation models from BERT to ChatGPT and highlighted their significance as parameter initializations for downstream tasks. Ref. [51] explored popular pre-trained foundation models in text, image, and graph modalities, discussing their components, pre-training methods, and advancements thoroughly. Ref. [51] also addressed topics including model efficiency, compression, security, and privacy, while offering valuable insights into scalability, logical reasoning ability, and cross-domain learning. Another survey paper [52] presented a comprehensive review of the state-of-the-art in self-supervised recommendation (SSR). Ref. [52] proposed an exclusive definition of SSR and developed a taxonomy that categorizes existing SSR methods into four categories: contrastive, generative, predictive, and hybrid. It further introduces an open-source library called SELFRec, which incorporates a wide range of SSR models and benchmark datasets. Through rigorous experiments and empirical comparison, Ref. [52] derived significant findings related to the selection of self-supervised signals for enhancing recommendation. The conclusion highlights the limitations and outlines future research directions in the field of self-supervised recommendation.

3.2 Popular structure of reviews in PAMI

The review's structure is alternatively referred to as the framework of the review. It is the outline that authors consider when starting to conduct the survey. Generally, a well-designed structure is considered essential for enhancing a paper's readability and helping readers grasp its core concepts and knowledge. Such a framework typically includes foundational components such as the abstract, introduction, methodology, discussion, and conclusion, each serving a distinct communicative function. To investigate how these elements are actually organized, we conducted a large-scale analysis of section titles across our collected corpus of survey papers. We first identified a representative set of section keywords (e.g., Introduction, Methods, Experiments, Results, Discussion, Future Work, Conclusion) and extracted their occurrences. For each keyword, we normalized its position within the document (ranging from 0 at the beginning to 1 at the end) and then aggregated these values across all papers. The resulting distributions, illustrated with violin plots in Figure 4, reveal clear structural anchors: "Introduction" consistently appears at the very beginning, while "Conclusion" is concentrated at the end. Expository sections such as "Background" and "Preliminaries" occur early; "Related Work" and "Taxonomy" appear in the early-to-middle part; and "Methods" and "Dataset" are most often located around the middle. Empirical sections ("Experiments", "Results", "Evaluation") tend to cluster in the latter half, whereas reflective components ("Application", "Discussion", "Challenges", "Future") are predominantly placed near the end. More detailed information on the algorithm can be found in the Supplementary Material.

To gain a deeper understanding of structural practices, we applied information extraction techniques to the RiPAMI database. Using an ensemble query approach based on LLMs, we automatically identified six representative content features: (1) Taxonomy—whether the review explicitly outlines a taxonomy of methods; (2) PRISMA—whether it specifies inclusion/exclusion criteria akin to PRISMA guidelines [53]; (3) Preliminary—whether it provides background knowledge in a distinct section; (4) Benchmark—whether it presents quantitative comparisons

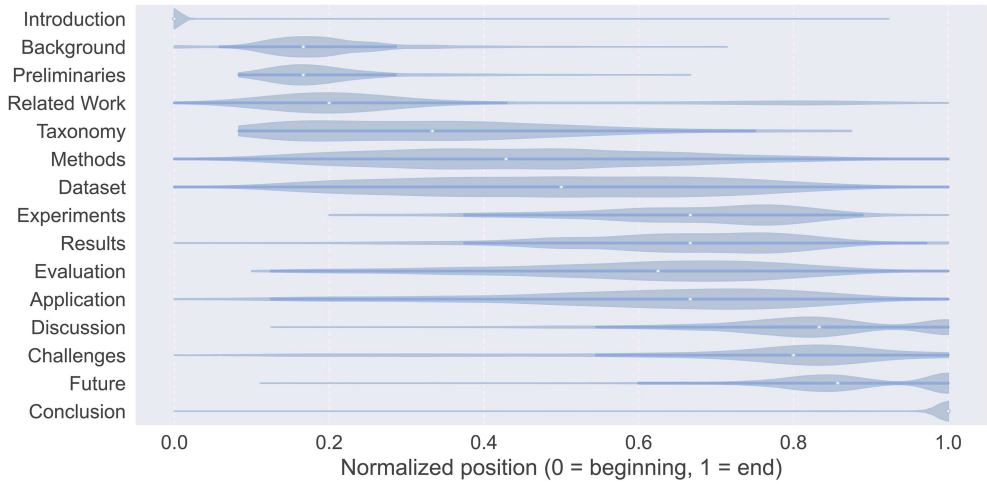

Figure 4 (Color online) Distributions of normalized section positions across survey papers. Core sections such as Introduction and Conclusion exhibit highly consistent placement at the beginning and end, while methodological, empirical, and reflective components show broader variation across the document body.

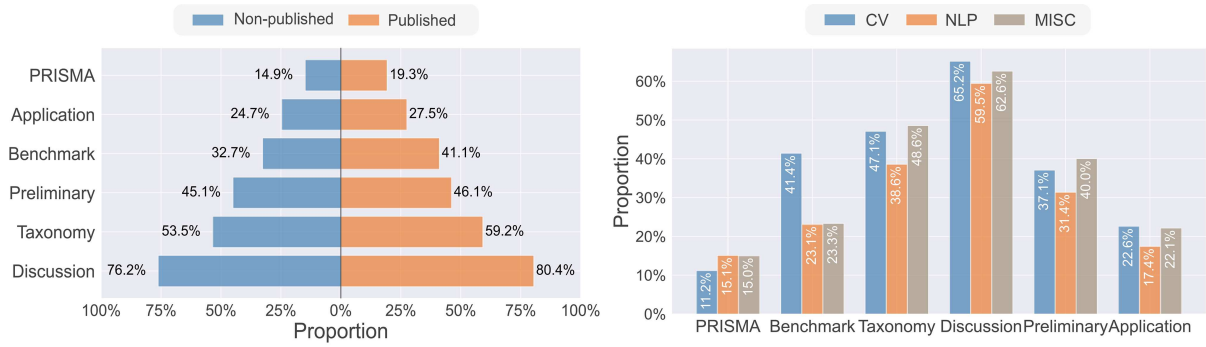

Figure 5 (Color online) Statistical analysis of review features in the RiPAMI database using LLM-based information extraction. "Published" refers to peer-reviewed papers, whereas "Non-published" denotes preprints without peer review. CV, NLP, and MISC correspond to the fields of computer vision, natural language processing, and miscellaneous domains, respectively.

across methods; (5) Application—whether it discusses real-world applications; and (6) Discussion—whether it addresses challenges, limitations, or future directions. Each feature was recorded in binary format (0/1) for statistical analysis (see Supplementary Material for extraction details).

The left panel of Figure 5 presents the overall prevalence of the six identified features. A majority of reviews include a dedicated discussion section on challenges and future directions, reflecting the academic expectation that surveys should not only summarize existing work but also identify open problems and potential avenues for research. In contrast, fewer than 20% of reviews in the PAMI field adopt PRISMA-like methodological criteria, regardless of publication venue, revealing substantial room for more rigorous systematic practices. The right panel of Figure 5 highlights domain-specific differences: reviews in computer vision (CV) show a pronounced emphasis on benchmarking, whereas those in smaller or cross-disciplinary areas (MISC) more frequently incorporate preliminaries.

Figure 6 illustrates temporal trends. All six features have increased steadily over the past decade, with the inclusion of "Discussion" and "Preliminary" sections showing the sharpest rise. Particularly noteworthy is that the adoption of PRISMA-style reporting has grown by more than 10% between 2015 and 2024, pointing to a gradual shift toward more structured and methodologically robust reviews within the PAMI community.

Building on the preceding analysis, we now discuss the four major components commonly observed in survey papers: the abstract, the beginning part, the middle part, and the ending part.

3.2.1 *Abstract*

The abstract is widely regarded as the most important part of a paper, as it provides a concise summary of the core content, enabling readers to quickly grasp the main points and conclusions. While the level of detail may vary, a structured abstract improves clarity and reproducibility by dividing the content into clearly labeled sections, each

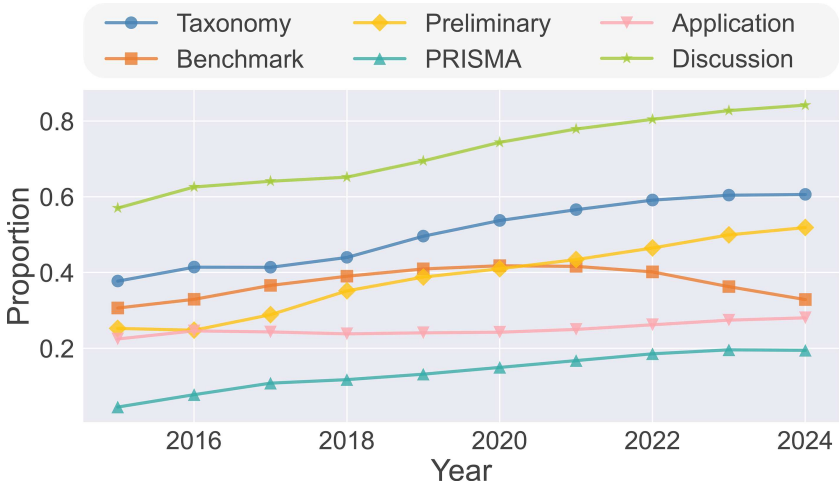

Figure 6 (Color online) Trends in the proportions of review features over time. The proportion of most features has been rising progressively over the years. For improved visual clarity, Gaussian smoothing is applied to the raw data.

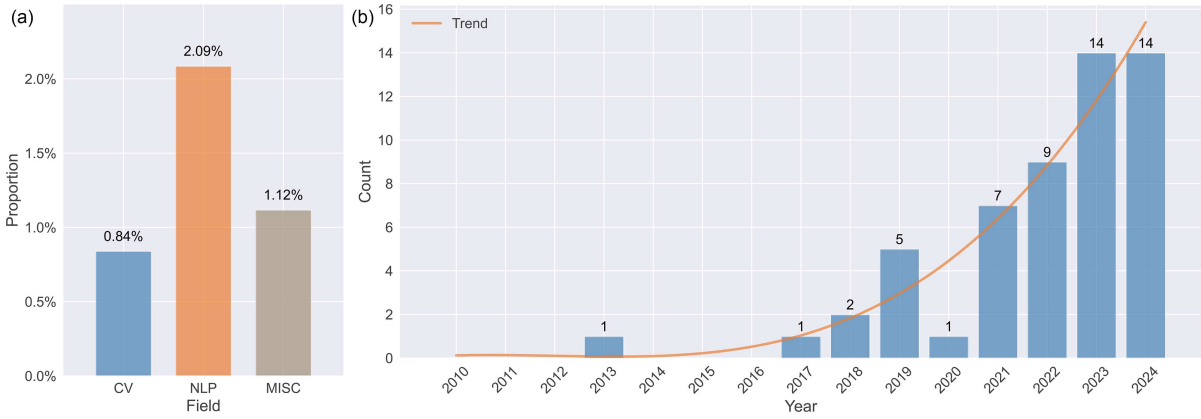

Figure 7 (Color online) Proportion and trend of structured abstracts. (a) The bar chart shows the proportion of structured abstracts across different fields (CV, NLP, MISC), while (b) the line chart illustrates the trend in the number of structured abstracts over the years from 2010 to 2024.

fulfilling a specific communicative function. Importantly, international reporting standards such as the PRISMA guidelines [53] explicitly recommend the use of structured abstracts in systematic reviews and meta-analyses, as they enhance transparency and facilitate information retrieval [54, 55]. A typical structured abstract consists of sections such as Background, Objective, Methods, Results, Conclusion/Discussion, and sometimes Limitations. For instance, Sariyanidi et al. [56] structured their abstract by introducing the background on facial affect, highlighting persistent challenges in the field, and then summarizing their methods, results, and discussion.

We further instruct the LLM to assess whether an abstract adheres to the structured format outlined by PRISMA guidelines (detailed in the Supplementary Material). As shown in Figure 7(a), the proportion of structured abstracts in the PAMI field remains relatively low, suggesting that this field has yet to fully embrace the structured abstract format. However, Figure 7(b) illustrates a clear upward trend in the number of abstracts meeting the structured abstract requirements, with a steady increase observed year on year. This highlights a growing adoption of standardized abstract formats in recent research.

3.2.2 *The beginning part*

Similar to the research article, the introduction section of the literature review usually includes contextualizing the research topic, identifying the knowledge gap, defining the scope and objectives, and laying the foundation knowledge for wider audiences. "Introduction" usually lies at the very beginning of the paper. Most reviews first introduce the definition of the topic to acquaint readers with a basic understanding of the field. For instance, the term "Named Entity Recognition" may be unfamiliar to many scholars outside the field of natural language processing. A clever introduction might provide the origins of the terminology and inform the readers what is named

entity recognition, as has been done in [57]. A brief definition of the field in the introduction is also welcomed for some relatively popular fields; e.g., the literature review [31] examines the concept of image categorization in the first sentence.

In addition to contextualizing the research topic, many introductions also highlight the existing research and identify gaps or challenges in the current understanding of the topic. It sets the stage for the literature review by explaining why the research being reviewed is necessary and what gaps are expected to be filled. Wang et al. [58] emphasized that while there are comprehensive surveys of self-supervised learning in computer vision, there is a lack of a similar overview specifically tailored to the remote sensing community.

Readers need to evaluate if the article is worth reading before investing more time in it. A statement of the scope or contributions of the article is a way for readers to quickly assess its relevance. In the first section, Wangkhade et al. [59] provided us with several important contributions including analyzing well-known technologies, proposing taxonomies of approaches, and summarizing benefits and challenges of sentiment analysis.

Preliminaries and problem formulations are also popular with readers, as they both provide profound background knowledge. Given that the Gumbel-max involves considerable mathematical concepts and calculations, a "Preliminaries" section in the review of the Gumbel-max trick [60] serves the purpose of providing background information and basic understanding related to the Gumbel-max trick.

For systematic reviews, a clear statement regarding the review methodology (e.g., literature search strategy and criteria for inclusion and exclusion) is essential [53, 61]. Such a statement helps to minimize potential bias and enables other researchers to replicate the study's findings. Memon et. al. [62] dedicated a section to detailing their review process, including subsections on the review protocol, inclusion and exclusion criteria, search strategy, study selection process, quality assessment criteria, and data extraction and synthesis, ensuring transparency and rigor in their methodology.

3.2.3 *The middle part*

The middle part of a survey paper, also known as the review part, presents a detailed examination of relevant research studies, methodologies, findings, and theories related to the chosen theme. Beyond mere description, it synthesizes information from the literature to provide a cohesive and integrated understanding of the topic. This synthesis not only aligns disparate works but also evaluates their contributions and interrelations. It often includes comparative analysis, highlighting similarities and differences in approaches, and may identify limitations in the existing research that warrant further investigation.

Across existing surveys in the PAMI field, different organizational strategies can be observed. Some reviews arrange the discussion by grouping methods according to their technical characteristics, thereby creating typologies that help readers compare families of approaches. For instance, Minaee et al. conducted a comprehensive survey on deep learning-based image segmentation models [6], grouping them into ten categories such as fully convolutional models, encoder-decoder-based models, and attention-based models. This mode of presentation highlights the breadth of methodological options, but can be less accessible for readers who are not yet familiar with the field.

Other surveys are organized around specific research problems or tasks, with each section devoted to a challenge and the solutions proposed for it. A review on egocentric vision hand analysis by Bandini et al. [63] illustrates this approach: the study is divided into sections such as hand segmentation, hand detection, and hand identification, with subsections that further explore issues like robustness to illumination changes or lack of pixel-level annotations. This problem-driven framing can be especially useful in application-oriented areas, though it may offer less systematic elaboration of theoretical underpinnings.

In many cases, authors blend these strategies. A review may be structured by tasks or challenges, while within each task methods are further grouped and compared. Guo et al. [64], for example, divided their survey into task-based sections but introduced similar approaches group by group within each task. Such a blended organization can provide both problem-driven context and methodological coverage, giving readers multiple perspectives on the same body of work.

Overall, the middle part of a survey is where synthesis and comparison take place, and where authors draw together diverse contributions into a structured narrative. Whether arranged around methods, challenges, or a combination of both, these organizational choices illustrate how surveys strive to balance comprehensiveness, accessibility, and clarity in guiding readers through complex literature.

3.2.4 *The ending part*

The subsequent ending part serves as a succinct conclusion to the reviewed studies. Within this section, the authors prefer to sum up the key findings to answer the question in the beginning part. It usually highlights both

the advantages and disadvantages of the reviewed literature to conclude research gaps and suggest potential avenues for future research.

One of the main objectives of the concluding part is to provide a concise summary of the key insights derived from the reviewed studies and to emphasize their significance and relevance to the research topic. In some literature reviews, however, the analysis and synthesis are so extensive that the main body spans a considerable number of pages. For instance, Liu et al. [35] presented a survey whose main body covers 51 double-column pages, posing a challenge for readers who may not have the time to read it thoroughly. Fortunately, the study also provides a concise summary, which enables readers to quickly grasp the essential content and navigate to sections of particular interest.

When discussing further in the ending part, the authors may attempt to identify gaps and point out future directions through the analysis of existing literature. Some of the studies present the gaps and future directions separately. Hussain et al. [65] first discussed major challenges of multi-view video summarization such as lack of synchronization, instability of camera, and crowded scenes individually. Where there are challenges, there are future research directions. In addition to challenges, the authors also provide recommendations and future directions from various perspectives including models, benchmark datasets, and agent-based MVS. Conversely, some studies choose to combine discussions of current issues with emerging research trends, as seen in the 'Future Directions' section of [66].

For literature reviews delving into pragmatic methodologies, the inclusion of an applications section is quite fitting. Ref. [67] offered a comprehensive review of various surveys on convolutional neural networks, dedicating a distinct section to the typical applications of 1-D, 2-D, and multidimensional CNNs. Similarly, Ref. [68] illustrated the deployment of few-shot learning techniques across disciplines like computer vision, natural language processing, and reinforcement learning.

4 Indicator-guided navigation for literature reviews

As the number of published reviews continues to grow, it has become common for multiple surveys to cover similar areas, attempt to answer analogous research questions, and often arrive at comparable conclusions. This overlap creates information redundancy and, to some extent, reduces research efficiency. Human scholars frequently rely on academic search engines such as Google Scholar, Semantic Scholar, and Web of Science, prioritizing high-ranking results typically sorted by citation counts to identify a select number of reviews for closer examination. However, these rankings are sometimes perceived as biased or even manipulated [69, 70], and an over-reliance on citation counts can intensify the Matthew effect, which may ultimately hinder sustainable academic development within the field.

Building on the indicators introduced in Section 2, this section demonstrates how they can be applied to enable indicator-guided navigation. Such navigation not only helps researchers more effectively evaluate and select reviews, but also provides a principled mechanism for AI-for-Research systems that use human-authored surveys as starting points for automated outline generation and knowledge synthesis. In this way, the indicators serve as a bridge between methodological rigor and practical strategies for navigating the expanding corpus of reviews.

4.1 Appraising reviews across various topics

A fundamental need may lie in the assessment of reviews across various fields, as most reviews are primarily aimed at readers who are relatively new to a field and may lack an in-depth understanding of the field. For instance, a researcher working in a popular field like object detection may start exploring a less familiar subfield such as optical character recognition (a subfield of computer vision). When encountering a review with only 100 citations, the researcher might dismiss its importance, unaware that in a niche area, 100 citations can actually indicate significant influence. This could lead to an oversight of critical insights in the subfield, simply due to differing citation norms between areas. The proposed $TNCSI$ offers an effective solution by automatically identifying the field of a given article and dynamically estimating its normalized impact based on the field and its current citation count. Illustrative examples are presented in Table 3.

4.2 The role of reference quality in selecting reviews

As the saying goes, *'You can't make something out of nothing'*. The same principle applies to literature reviews: a well-crafted review is typically built upon a thorough analysis of high-quality references. To further explore the relationship between the quality of references in literature reviews and the review's impact, we conduct a visual

Table 3 Example use cases of $TNCSI$. A similar number of citations (Cites) in various topics may lead to notably different $TNCSI$ values.

Title	Cites	Topic	$TNCSI$↑
Involvement of machine learning for breast cancer image classification: a survey [71]	104	Breast cancer image classification	0.79
Class-incremental learning: a survey [72]	99	Class-incremental learning	0.61
A systematic survey of prompt engineering on vision-language foundation models [73]	100	Prompt engineering	0.94
End-to-end speech recognition: a survey [74]	100	End-to-end speech recognition	0.59

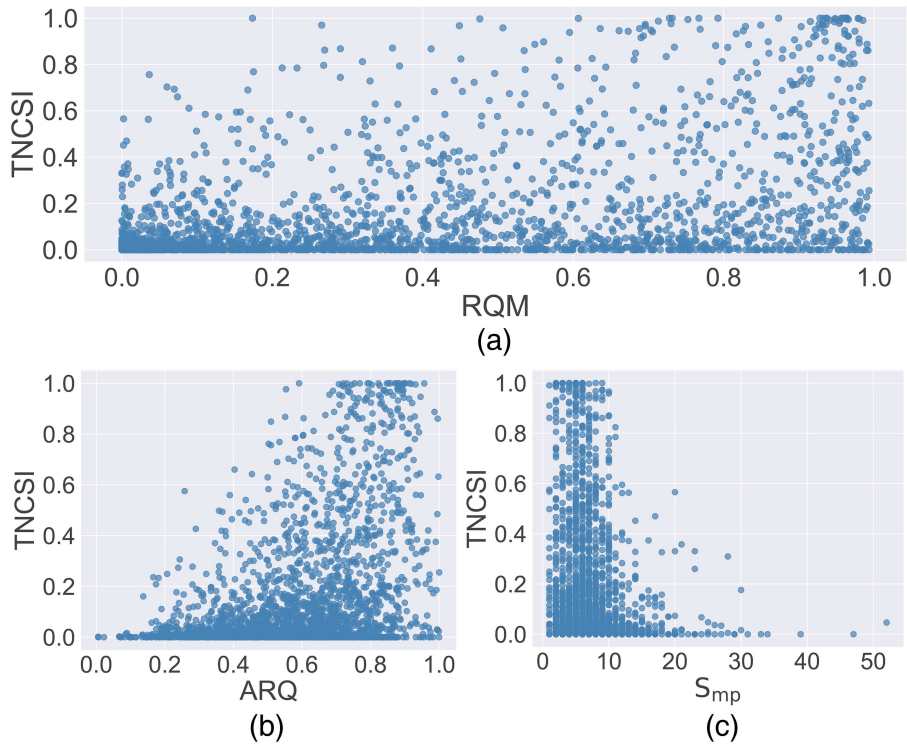

Figure 8 (Color online) Visualization of the relationship between reference quality and $TNCSI$. Each point corresponds to a review paper. The $TNCSI$ shows positive correlations with RQM (a) and ARQ (b), but a negative correlation with S_{mp} (c).

Table 4 Correlation between reference quality and $TNCSI$.

Metric	Pearson	Pearson p-value	Spearman	Spearman p-value
ARQ	0.43	6.99×10^{-40}	0.42	1.11×10^{-37}
S_{mp}	−0.27	6.29×10^{-16}	−0.32	2.21×10^{-21}
RQM	0.53	1.34×10^{-62}	0.51	1.86×10^{-57}

analysis of the distributions of ARQ (representing the quality of references), S_{mp} (representing the recency of references), and the corresponding $TNCSI$ (representing the scholar impact) for reviews published after 2020, as shown in Figure 8.

Figure 8(a) highlights the correlation between RQM and $TNCSI$. The scatter plot indicates a notable trend where lower RQM values are generally associated with lower $TNCSI$ values, with a significant concentration of points observed in the range of $TNCSI$ between 0.0 and 0.2. Figure 8(b) demonstrates that reviews citing higher-quality references are more likely to achieve higher academic impact. Generally, it is difficult for reviews with an ARQ below 0.5 to reach a $TNCSI$ higher than 0.5. Furthermore, Figure 8(c) shows that reviews with a median reference age exceeding 10 semesters (i.e., 5 years) tend to exhibit relatively limited impact.

To further quantify the relationship between reference features and cumulative academic impact, we present the Pearson and Spearman correlation coefficients between various indicators and $TNCSI$ in Table 4. Both ARQ and RQM show moderate positive correlations with $TNCSI$, with Pearson and Spearman coefficients around 0.43 and 0.53, respectively. These correlations are highly significant, as indicated by p-values close to zero. In contrast, S_{mp} exhibits a negative correlation, as indicated by Pearson and Spearman coefficients of −0.27 and −0.32, respectively.

Table 5 Example use cases of RQM. A higher RQM reflects higher scholarly impact (ARQ) and a younger age (S_{mp}) of the references.

Title	$ARQ\uparrow$	$S_{mp}\downarrow$	$RQM\uparrow$
A survey on video diffusion models [75]	0.72	2	0.94
Video diffusion models: a survey [76]	0.83	3	0.95
A survey on multimodal large language models [77]	0.83	1	0.99
Multimodal Large Language Models: A Survey [78]	0.69	5	0.65
A survey on benchmarks of multimodal large language models [79]	0.52	2	0.85
A survey on evaluation of multimodal large language models [80]	0.19	2	0.63

Table 6 Example use cases of IEI. A positive IEI indicates that the paper is gaining increasing attention.

Title	Cites	$TNCSI\uparrow$	$IEI\uparrow$
Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks [81]	565	0.99	−1.41
A survey on knowledge distillation of large language models [82]	40	0.30	1.10
A review of practical ai for remote sensing in earth sciences [83]	29	0.11	−0.16
Change detection methods for remote sensing in the last decade: a comprehensive review [84]	23	0.23	0.40

Table 7 Example use cases of RUI. RUI quantifies the urgency of updating reviews across different fields.

Title	Year	Topic	$RUI\downarrow$
A review of object detection based on deep learning [85]	2020	Object detection	24.10
A comprehensive review of object detection with deep learning [86]	2023	Object detection	6.72
Time-series forecasting with deep learning: a survey [87]	2020	Time-series forecasting	29.09
Long sequence time-series forecasting with deep learning: a survey [88]	2023	Time-series forecasting	6.44

These statistically validated results demonstrate the importance of reference quality in influencing the academic impact of reviews.

Considering that the quality of references exhibits a moderately strong and significant positive correlation with the scholarly impact of a review, RQM can be regarded as an intuitive metric for efficiently gauging the average quality of references. In particular, as observed from Figure 8, reviews with an RQM greater than 0.8, an ARQ within the range of 0.7 to 0.9, and an S_{mp} not exceeding 10 are more likely to achieve higher academic impact. Illustrative examples of RQM are provided in Table 5.

4.3 Tracking citation momentum with *IEI*

Although $TNCSI$ enables cross-domain comparisons, it fails to track changes in citation trends. In such cases, the IEI can offer additional insight. A higher IEI suggests that the paper's rate of gaining new citations has been increasing over time, reflecting a growing level of recognition within the academic community. Conversely, an IEI below zero implies that the paper's citation rate has been declining recently. Table 6 provides examples of IEI applications. In some cases, selecting a paper with a higher IEI may be preferable to choosing one with a higher $TNCSI$, as it reflects a more recent consensus among researchers.

4.4 Assessing the need for updating reviews

Literature review inevitably ages, yet the rate of aging varies significantly across fields. The RUI allows us to gauge the degree of aging of a review. Typically, fields with higher publication rates undergo faster advancements (indicated by a higher RUI), meaning that reviews in these areas tend to age more quickly. Conversely, in more established fields with a lower RUI, the aging process is correspondingly slower. Compared to relying solely on publication time to assess recency, RUI provides a more accurate estimation. Examples illustrating the exploitation of the RUI are provided in Table 7.

4.5 Insights into field trends via *IEI* and *RUI*

As outlined in Subsections 2.4.2 and 2.4.4, IEI and RUI are indicators used to evaluate changes in citation trends and the recency of individual articles. By utilizing these two indicators, we generate scatter plots depicting the temporal distribution of IEI and RUI values for review papers within the domains of recommendation systems and language modeling (as illustrated in Figure 9).

In the left panel of Figure 9, we observe a rapid growth trend in the development of surveys within the field of "language modeling" since 2023, which is reflected by the increasing number of surveys with positive IEI values.

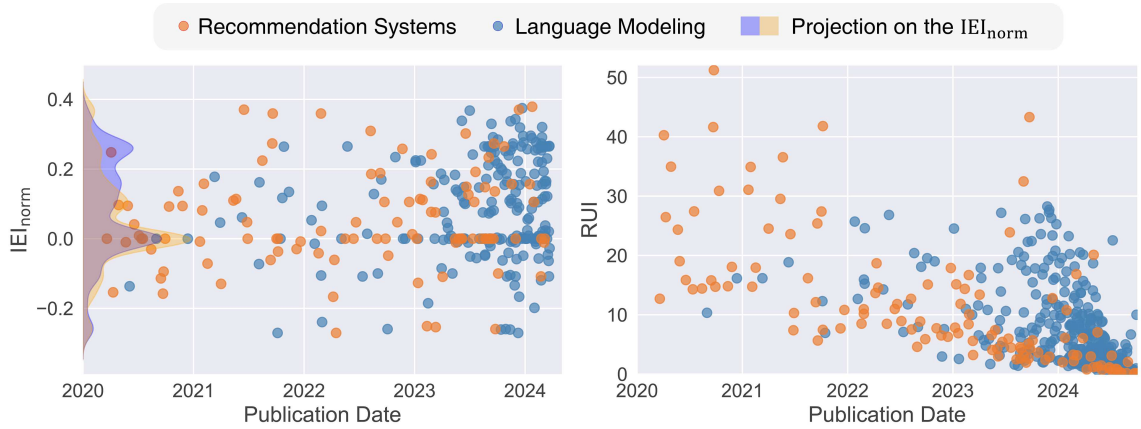

Figure 9 (Color online) Temporal distribution of IEI and RUI in surveys for recommendation systems and language modeling. Higher IEI and RUI values in language modeling over the past two years indicate a trend of intensified development. Although surveys in popular fields tend to attract more academic attention, they also require authors to make greater efforts to keep the content up-to-date.

The right panel of Figure 9 further supports this finding. Since 2023, the median RUI for language modeling surveys has noticeably exceeded that of surveys in the recommendation systems field. As RUI is positively correlated with the volume of publications within a field, a higher RUI suggests a larger publication volume. While this trend underscores the swift progress in the language modeling domain, it also highlights the growing burden on authors to produce and maintain high-quality surveys in such a rapidly evolving and highly competitive area.

By analyzing the statistical characteristics of numerous review articles' IEI and RUI, we can uncover developmental patterns across various fields. These patterns provide valuable insights for gauging the momentum of progress within each field and advancing exploration in emerging areas.

4.6 Integrated use of the proposed indicators

In general, researchers are encouraged to apply a set of metrics, rather than relying solely on citation counts to select review articles. In practice, reviews with a $TNCSI$ close to 1 are often prioritized, as this suggests that the review is cited more frequently than most other papers in the same field, indicating widespread esteem within the field. However, supplementary indicators can provide further insights: the IEI captures shifts in citation trends, the RQM evaluates the quality of references, and the RUI reflects how recently the review may require updating. Leveraging a diverse set of indicators allows for the efficient filtering of numerous similar reviews while mitigating the limitations inherent in single-metric citation reliance.

Although the proposed metrics provide an intuitive numerical indication, it is important to emphasize that no existing metric, including the proposed ones, can truly capture the intrinsic value of a review (particularly for systematic review). The proposed impact or quality indicators should be regarded as supplementary instruments rather than definitive measures of quality. It is our contention that the true value of a review resides in its capacity to address the specific research questions and meet the needs of its intended audience. More examples of integrated indicators usage and further explanations regarding the fair use of metrics are detailed in the Supplementary Material.

5 AI-generated literature reviews

5.1 Overview of AI-generated literature reviews

Traditionally, literature reviews have been manually conducted by researchers who analyze and synthesize scholarly sources to provide an overview of the current state of knowledge in a specific field. Recently, with the advancement of AI technologies, there has been a growing interest in leveraging artificial intelligence techniques, especially LLMs, to automate or assist in the generation of the literature review [89–93]. Typically, users are simply required to indicate their area of research interest, and the system will then automatically generate a literature review.

The automated creation of a literature review is a multidisciplinary endeavor that integrates knowledge from various fields. It relies not only on artificial intelligence technologies but also requires the merging of knowledge from other fields such as data science, bibliometrics, and database engineering. These components are instrumental in extracting, storing, and synthesizing relevant information from vast amounts of literature. Early attempts in

AI-generated literature reviews involve training language models, e.g., LLaMA [94], on a large corpus of academic papers, research articles, and other scholarly content. These models can then be used to generate coherent and contextually relevant text based on prompts or queries related to a specific research topic. However, given that a fully trained or fine-tuned language model has no access to the latest scholarly advances, these models would not generate a literature review that includes the latest scholarly materials. To overcome this limitation, most advanced AI-generated literature review systems adopt a structured pipeline consisting of four main steps: intention analysis, knowledge retrieval, synthesis, and report generation. The generation of a literature review begins with intention analysis, where the system identifies the research goals and scope. On this basis, the core process consists of knowledge retrieval and synthesis, which are closely interwoven rather than strictly sequential. Retrieval broadens the evidence base by applying criteria such as keywords, publication dates, and citation counts, and by collecting materials not only from academic publications but also from sources such as blogs, tutorials, and open-source repositories. Synthesis complements this step by analyzing the collected resources, identifying relationships among them, and organizing the information into coherent themes. By reinforcing each other, retrieval and synthesis ensure that the system produces a comprehensive and well-grounded review. Finally, the process culminates in report generation, where the integrated content is transformed into a structured and readable document.

Beyond their technical implementation, AI-generated literature reviews are often viewed as a promising tool for reducing researchers' time and effort. However, concerns persist regarding both the reliability and the ethical implications of such content. Zybaczynska et al. [95] observed that current AI systems frequently fail to deliver substantial, accurate information or critical judgment. Similarly, Elali et al. [96] highlighted the risks of fabrication and falsification in AI-generated research, emphasizing their potentially serious consequences for the scientific community. Despite these challenges, there have been notable demonstrations of AI-assisted reviews reaching publication. For instance, Aydın et al. used large language models such as ChatGPT and Google Bard to generate reviews in areas including digital twins in healthcare [97] and the metaverse [98]. Even so, many academic publishers continue to adopt a cautious stance. Leading journals such as *Nature* and *Science*, for example, explicitly prohibit the listing of ChatGPT or other automated tools as paper authors. We contend, nevertheless, that AI-generated reviews—when not intended for direct publication—can still provide substantial value. In particular, they may serve as an efficient aid for researchers seeking to keep pace with rapid developments in fast-evolving fields.

5.2 Current state of AI-generated literature review

In this subsection, our investigation delves into various efforts that have been made in the field of employing AI techniques to generate literature reviews.

5.2.1 *Comparison of recent AI-generated literature review systems*

Positioned as a significant direction in AI-for-Research, automated review generation systems have garnered substantial attention across the research and technology communities. Early explorations—such as Paper Digest [99], Jenni AI [100], and various GPT-store plugins [101, 102]—highlight the feasibility of leveraging large language models to generate literature reviews with minimal user input. Building upon this foundation, recent efforts (see Figure 10) have introduced more systematic and scalable approaches that emphasize structured pipelines, citation fidelity, and quality-controlled synthesis. For instance, hierarchical catalogue generation for literature review foregrounds outline construction as a first-class task, releasing a large-scale dataset and semantics-plus-structure metrics that supply supervision for catalogue-guided synthesis [103]. Building on this foundation, AutoSurvey [91] pioneers an end-to-end pipeline—seed retrieval and outline planning, subsection drafting by specialized LLMs, global integration, and iterative evaluation—demonstrating feasibility at scale. To scale capacity and preserve citation fidelity, SurveyGO [104] employs the LLM×MapReduce-V2 "convolutional stacking" architecture and uses an internal benchmark to ensure structural rigor and precise citations. Focusing on workflow quality, SurveyX [105] adopts a two-phase Preparation→Generation design that blends offline/online retrieval with an AttributeTree preprocessing pipeline and a re-polishing stage, yielding measurable gains in content quality and citation accuracy. SurveyForge [106] narrows the human-AI gap via heuristic-guided outline construction and a memory-driven scholar navigation agent (SANA), coupled with temporal- and citation-aware re-ranking to improve reference reliability and downstream drafting. Emphasizing user control and multimodality, InteractiveSurvey [107] supports outline- and section-wise RAG synthesis and produces reports enriched with structural diagrams and extracted figures/tables. From a reliability perspective, SciSage [108] introduces a citation-aware multi-agent framework that "reflects while writing," critiquing drafts at outline, section, and document levels and re-ranking multi-source retrieval by recency and citation signals to strengthen coherence and references.

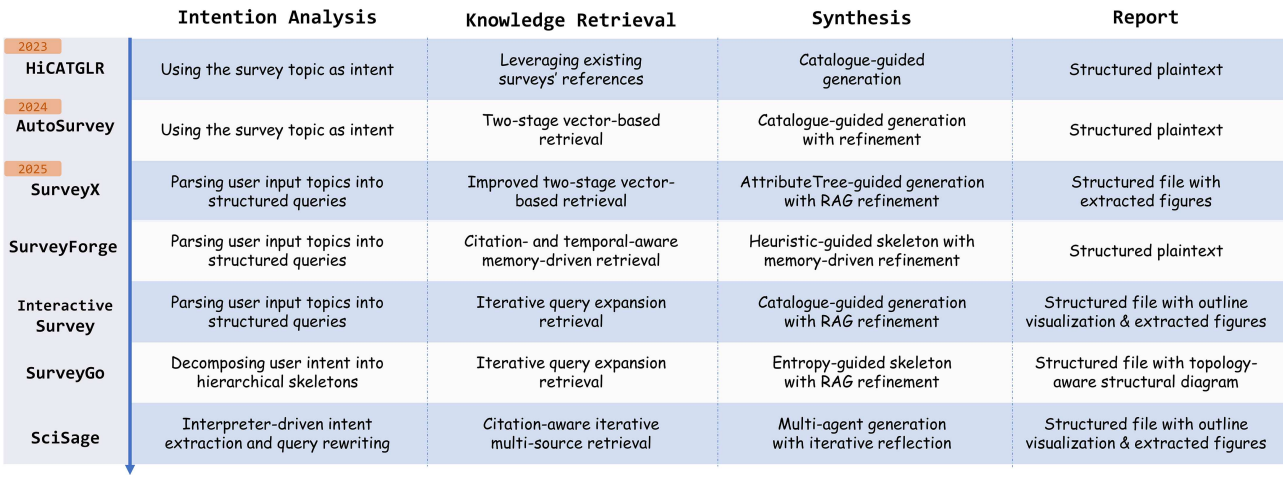

Figure 10 (Color online) Comparative overview of recent survey generation systems. The figure contrasts representative methods (HiCAT-GLR [103], AutoSurvey [91], SurveyX [105], SurveyForge [106], InteractiveSurvey [107], SurveyGo [104], SciSage [108]) across four key stages: intention analysis, knowledge retrieval, synthesis, and report generation. The timeline highlights their evolution from 2023 to 2025.

It is not difficult to observe that, as technologies evolve over time, automated survey systems have gradually matured across the four core components of intention analysis, knowledge retrieval, synthesis, and report generation. Compared to early explorations, many recent systems now produce reviews that are markedly more coherent, informative, and credible. The generated outputs exhibit more reasonable structural organization, richer visual elements such as extracted or synthesized figures and diagrams, and more accurate and verifiable citations. These improvements collectively enhance the readability and reliability of AI-generated reviews, allowing such systems to move beyond proof-of-concept prototypes and begin offering practical value in assisting researchers with staying abreast of emerging fields.

5.2.2 *Human-authored vs. AI-generated reviews*

To highlight the differences in underlying design principles, we provide a descriptive comparison between human-authored and AI-generated reviews. Table 8 [17, 91, 99, 100, 102, 104, 109, 110] shows that nearly all automatic summarization systems fail to incorporate visual elements. Furthermore, most AI-generated review systems do not include explicit appraisal criteria, making it difficult for users to discern how quality control is enforced and thereby reinforcing the perception of these systems as black boxes.

In Figure 11, we further present a set of scatter plots depicting the references' quality of (a) reviews published in several academic journals and (b) reviews generated by various AI systems. To allow sufficient time for the references selected by the AI-generated review system to accumulate academic impact, a two-month interval is set between obtaining the reference list and calculating their $TNCSI$. The horizontal axis represents the reference age, and the vertical axis corresponds to the proposed $TNCSI$. The color and size of the scatter points indicate the IEI of the references. A positive IEI value signifies a gradually increasing citation trend, resulting in warm-colored scatter points. The larger the size of the scatter point, the greater the value of IEI. Conversely, when the IEI value is negative, the scatter points are cool-colored, and the size decreases as the value decreases. A closer examination of Figure 11 reveals a distinct distribution among human-authored and AI-generated depicted reviews in their respective scatter plots. Human-authored reviews exhibit clustered dots in the upper left corner, indicating a trend toward referencing more recent and influential sources (i.e., citing newly published studies with limited influence at the time of writing). Conversely, AI systems predominantly cite established, high-impact papers, creating a relative sparsity in the upper left corner. We argue that this discrepancy stems from seasoned researchers' ability to assess a paper's value through a diverse, fine-grained, and content-based manner, rather than relying solely on citation counts. This allows them to render more precise judgments about the potential significance of a paper. In contrast, most existing AI-based review systems depend exclusively on citation counts as their primary filtering metric, limiting their capacity to identify high-quality recent publications as references. We also note the presence of a potential boundary line for the IEI metric in Figure 11(b), which is likely a result of the quality control mechanisms in certain automated review systems.

Overall, despite continuous technological advances, AI-generated review systems still face several critical limitations. Most systems continue to rely primarily on similarity-based retrieval, with limited attention to the quality

Table 8 Comparison between human-authored and AI-generated literature reviews. "V.E." indicates the presence of visual elements.

Review	Auto-level	V.E.	SALSA analysis [17]
Human [109]	Manually	✓	A typical narrative review aims to offer valuable insights in visual prompt learning. The selection criteria for references are not specified, but it is clear that each reference has been thoroughly appraised. Conduct an in-depth synthesis and offer insightful analysis.
Jenni [100]	Semi-automated	×	Automated searching and appraising references are not supported. The generation process relies on user interaction, and only a narrative description is provided. Provide little to no synthesis or analysis.
ChatPaper [110]	Automated	×	Retrieving relevant literature from arXiv based on multiple LLM-generated keywords without an appraisal step. Each section contains plain descriptions of the related reference. Provide little to no synthesis or analysis.
PaperDigest [99]	Automated	×	Searching papers with the user-specified keyword. It seems to appraise the quality of references with private criteria. The generated content is more like a concise summary. Provide little to no synthesis or analysis.
askyourpdf [102]	Automated	×	No official explanations are found for how to retrieve references and appraise the quality of the literature. The generated review includes a brief analysis and description of the related concept, current state of development, and gaps. Provide little to no synthesis or analysis.
AutoSurvey [91]	Automated	×	Narrative review without any visual elements. Although no formal appraisal framework is utilized, the approach incorporates semantic verification and multi-LLM evaluation. Generated reviews include narrative descriptions of key contributions, research trends, and gaps. Provide little to no synthesis or analysis.
SurveyGo [104]	Automated	✓	No official explanations are found for how to retrieve references and appraise the quality of the literature. The generated review exhibits a clear outline structure. Offer limited synthesis and analysis.

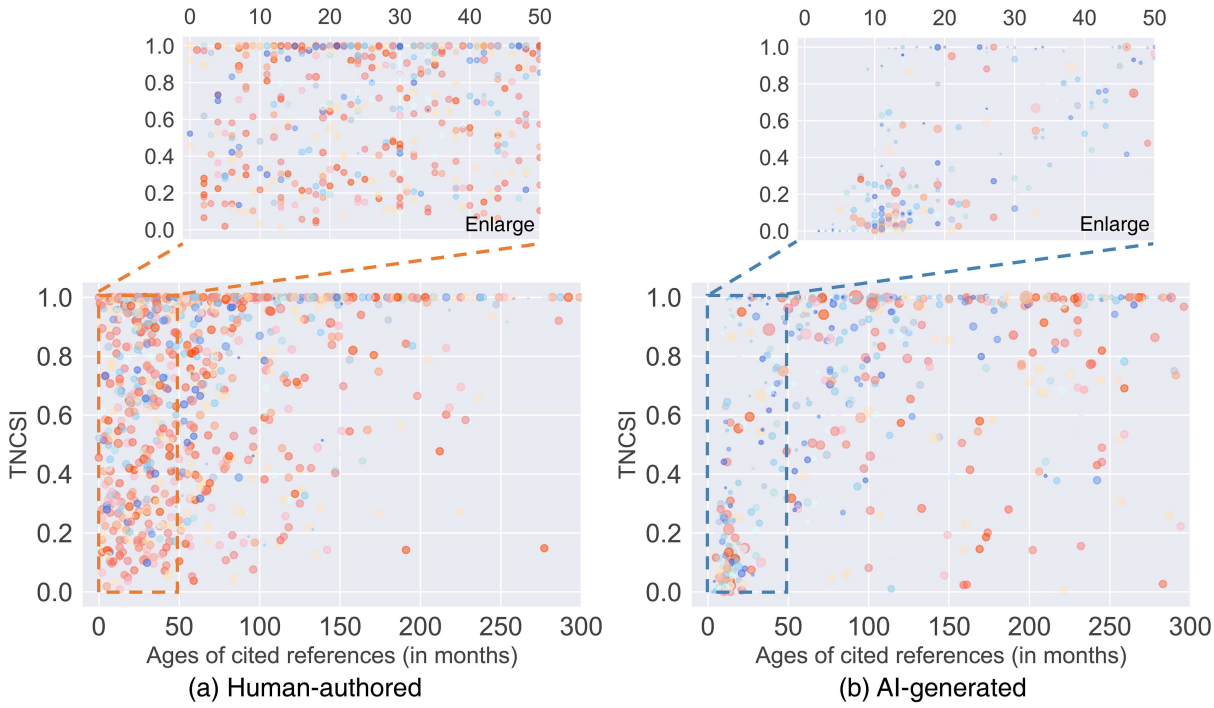

Figure 11 (Color online) Visualization of references' quality of human-authored and AI-generated literature reviews. (a) illustrates the quality-age distribution of references in human-written reviews, while (b) depicts the quality-age distribution of references in AI-generated reviews. It can be observed that the TNCSI of references less than one year old in AI-generated reviews is significantly lower than that in human-authored reviews. This suggests that current AI-generated review systems struggle to appraise the academic value based on the article content.

of the retrieved literature. Although citation counts are sometimes used as a proxy, they are domain-dependent,

temporally biased, and systematically disadvantage newly published work, making them an imperfect signal at best. In addition, current AI systems remain underdeveloped in terms of personalization: they seldom elicit or model users' specific research questions, and thus struggle to generate tailored surveys or guide users toward customized choices. Instead, they often imitate human-written surveys that prioritize coverage, sacrificing the customization that AI could uniquely provide. Even when visual elements are included, they are typically extracted directly from source papers rather than designed as polished, interpretive illustrations (e.g., classification diagrams, timelines, or methodological overviews). Moreover, automated approaches have yet to demonstrate the capability to robustly produce field-level artifacts such as reproducible benchmark tables or domain-specific taxonomies that systematically capture and compare the state of the art. Taken together, these limitations suggest a growing need to explore more reliable, transparent, and intent-aligned approaches to AI-generated literature reviews.

5.3 Toward reliable, transparent, and intent-aligned AI-generated reviews

Although the capabilities of AI-generated review systems are steadily improving, their adoption in practice remains modest. This may be partly due to ongoing challenges related to reliability, transparency, and the lack of intent-aware customization.

A key factor limiting the reliability of current systems lies in the lack of mechanisms for evaluating the credibility and quality of referenced literature. Many models retrieve papers based solely on lexical or embedding similarity, without regard to publication venue, peer-review status, or citation history—signals that are often crucial in scholarly contexts. Enhancing reliability may require the integration of explicit citation-quality management components that filter or re-rank literature based on such criteria [111, 112]. Furthermore, deeper exploitation of LLMs' reasoning capabilities [113] could allow for more substantive synthesis across sources, moving beyond surface-level summaries toward analytical, comparative, and even critical review writing. Emerging techniques, such as diffusion-based scientific visualization [114], may also contribute to reliability by enabling the automated generation of polished, interpretable visual elements.

Transparency remains a persistent challenge. Most existing systems function as black boxes: users are seldom informed about how references are retrieved, filtered, or retained. The lack of interpretable intermediate steps or explainable selection criteria undermines trust and limits rigorous evaluation. In contrast, greater transparency in AI-generated reviews can markedly strengthen users' confidence, as it enables them to better understand and assess the reliability of the outputs. To foster both trust and accountability, AI review systems should disclose their retrieval logic, inclusion and exclusion criteria, and synthesis provenance in a manner that is accessible to both technical and non-technical audiences. Such transparent pipelines not only facilitate critical appraisal but also make it easier to identify errors, biases, and omissions in the generated content.

Beyond reliability and transparency, one of the most overlooked opportunities of AI-generated reviews may lie in intent-aligned customization. Unlike human-authored reviews—which, once published, must serve a broad and heterogeneous readership—AI systems can adapt outputs dynamically to specific user goals. Depending on the research purpose, a system could generate a concise state-of-the-art overview for practitioners, a mapping review to highlight gaps in the literature, or a full systematic review to support meta-analysis. By tailoring outputs to user intent, such systems can deliver reviews that are more concise, targeted, and practically useful, with flexibility in analytic depth, coverage, and timeliness. This shift toward purpose-driven review generation underscores a unique advantage of AI: not simply replicating established formats, but reimagining the practice of literature reviewing around individual needs and research contexts.

6 Challenges and the future of the literature review in PAMI

Challenges and future opportunities for both human-authored and AI-generated reviews are discussed in this section.

6.1 Challenges

Information overload. While our proposed indicators help alleviate information overload to some extent, several challenges remain unresolved. First, as scientific research continues to expand rapidly, the volume of literature review is expected to grow significantly. Collecting and screening a vast number of publications in these emerging fields will likely lead to incomplete searches. Both human authors and AI systems are required to address the challenge of identifying the most relevant references from the extensive body of published literature. In addition to incomplete searches, the research community should consider exploring ways to avoid redundant efforts. For

example, the first two review papers [115, 116] on the segment anything model [117], uploaded to arXiv just two days apart, shared at least 70 overlapping references. Avoiding such duplication not only helps reduce the waste of resources (despite both papers offering unique insights) but also minimizes information redundancy.

Lag and obsolescence. Scientific knowledge is constantly evolving, yet the process of compiling reviews often struggles to keep up with the latest research advancements—particularly when created by human authors and subject to peer review. Comprehensive reviews are often delayed relative to the emergence of new research areas. For instance, when Mamba [118] started gaining traction, no in-depth reviews were available during its initial months. This delay reflects the time-intensive nature of gathering, evaluating, and synthesizing an ever-expanding body of literature, along with the requirements of the peer-review process.

6.2 Future

Long-term support literature review. The concept of long-term support (LTS) for literature reviews offers a practical solution for maintaining relevance and timeliness in the fast-paced environment of academic research. Unlike traditional reviews, which are generally published once and remain static, an LTS literature review would involve regular updates to include new research findings and advancements in the field. This approach is especially valuable when structured frameworks or taxonomies have been developed, as they serve as foundational tools in their domain and require continuous refinement to remain useful. While such updates may be curated by the original authors to preserve rigor and intent, automated or AI-assisted mechanisms could also play a critical role in monitoring new publications, suggesting revisions, and integrating fresh evidence. In this way, the field benefits from both the continuity of established classifications and the responsiveness afforded by ongoing, partly automated maintenance.

Intent-aligned and customized AI-generated literature review. Human-authored reviews are usually designed to be as comprehensive as possible, aiming not only for broad coverage of existing work but also to serve a heterogeneous readership—from newcomers seeking orientation, to experienced researchers looking for synthesis, and practitioners focusing on applications. In contrast, AI-generated reviews are not constrained by this pursuit of exhaustiveness; instead, they can be tailored to align with specific user intentions and deliver the most relevant literature for given purposes. This ability to generate customized reviews highlights AI's potential to complement traditional human-authored surveys by providing focused, intent-driven perspectives. Future development may further enhance this customization by integrating precise content-based evaluations and visual elements, thereby improving both the relevance and readability of AI-generated reviews.

7 Conclusion

This study provides the first large-scale tertiary analysis of literature reviews in the PAMI field and yields several key conclusions.

First, our structural investigation of more than 3000 reviews uncovers highly consistent organizational anchors—such as the placement of introductions and conclusions—yet also reveals systematic variation in and limited adoption of methodological reporting standards. Features like benchmarking, preliminaries, and structured abstracts are gaining momentum, but overall practice remains uneven across subfields and publication types. These findings underscore both the strengths and the blind spots of current review-writing conventions.

Second, our proposed bibliometric indicators ($TNCSI$, RQM, IEI, and RUI) demonstrate concrete value in navigating the rapidly growing body of surveys. They help normalize impact across domains, highlight the role of reference quality, track citation momentum, and identify the timeliness of reviews. Together, these measures provide researchers with practical tools for screening overlapping surveys and reducing over-reliance on raw citation counts. They also supply principled signals for future AI-for-Research systems that may incorporate human-authored reviews as seeds for automated synthesis.

Third, our comparative assessment of AI-generated and human-authored reviews points to both progress and limitations. While recent AI systems show improved coherence, organization, and even integration of visual elements, they remain constrained by over-reliance on citation counts, weak personalization, and difficulty in capturing very recent advances. Human experts still excel at nuanced appraisal of new contributions, but AI systems hold promise as intent-driven complements that can deliver customized perspectives for different user needs.

We release our code framework for meta-data retrieval, indicator computation, and statistical analysis. Although designed for PAMI, the approach is readily extensible to other domains, and we hope it will serve as a foundation for more reliable, transparent, and purpose-driven review practices across research communities.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62576177, 62361166670), Shenzhen Science and Technology Program (Grant Nos. QNXMB20250701090801002, JCYJ20250604184027034, JCYJ20240813114237048), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515011435), Fundamental Research Funds for the Central Universities (Grant No. 070-63253222), Tianjin Key Laboratory of Visual Computing and Intelligent Perception (VCIP), "Science and Technology Yongjiang 2035 Key Technology Breakthrough Plan Project (Grant No. 2025Z053), and Chinese Government-Guided Local Science and Technology Development Fund Projects (Scientific and Technological Achievement Transfer and Transformation Projects) (Grant No. 254Z0102G). Computation was supported by the Supercomputing Center of Nankai University (NKSC).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen L, Li S, Bai Q, et al. Review of image classification algorithms based on convolutional neural networks. Remote Sens, 2021, 13: 4712
2 Masana M, Liu X, Twardowski B, et al. Class-incremental learning: survey and performance evaluation on image classification. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 5513–5533
3 Machado G R, Silva E, Goldschmidt R R. Adversarial machine learning in image classification: a survey toward the defender's perspective. ACM Comput Surv, 2021, 55: 1–38
4 Mai Z, Li R, Jeong J, et al. Online continual learning in image classification: an empirical survey. Neurocomputing, 2022, 469: 28–51
5 Mazurowski M A, Dong H, Gu H, et al. Segment anything model for medical image analysis: an experimental study. Med Image Anal, 2023, 89: 102918
6 Minaee S, Boykov Y Y, Porikli F, et al. Image segmentation using deep learning: a survey. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 3523–3542
7 Siddique N, Paheding S, Elkin C P, et al. U-Net and its variants for medical image segmentation: a review of theory and applications. IEEE Access, 2021, 9: 82031–82057
8 Hao S, Zhou Y, Guo Y. A brief survey on semantic segmentation with deep learning. Neurocomputing, 2020, 406: 302–321
9 Zaidi S S A, Ansari M S, Aslam A, et al. A survey of modern deep learning based object detection models. Digital Signal Processing, 2022, 126: 103514
10 Cheng G, Yuan X, Yao X, et al. Towards large-scale small object detection: survey and benchmarks. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 13467–13488
11 Qian R, Lai X, Li X. 3D object detection for autonomous driving: a survey. Pattern Recognition, 2022, 130: 108796
12 Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogrammetry Remote Sens, 2020, 159: 296–307
13 Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv, 2024, 56: 1–40
14 Hao T, Li X, He Y, et al. Recent progress in leveraging deep learning methods for question answering. Neural Comput Applic, 2022, 34: 2765–2783
15 Malik M, Malik M K, Mehmood K, et al. Automatic speech recognition: a survey. Multimed Tools Appl, 2021, 80: 9411–9457
16 Maslej N, Fattorini L, Brynjolfsson E, et al. The AI index 2023 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, 2023
17 Grant M J, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libraries J, 2009, 26: 91–108
18 Beel J, Gipp B. Google Scholar's ranking algorithm: an introductory overview. In: Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), Rio de Janeiro (Brazil), 2009. 230–241
19 Beel J, Gipp B. Academic search engine spam and Google Scholar's resilience against it. J Electron Publishing, 2010, doi: 10.3998/3336451.0013.305
20 Beel J, Gipp B. On the robustness of Google Scholar against spam. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 2010. 297–298
21 Merton R K. The Matthew effect in science. Science, 1968, 159: 56–63
22 Brzezinski M. Power laws in citation distributions: evidence from Scopus. Scientometrics, 2015, 103: 213–228
23 Brysbaert M. How many words do we read per minute? A review and meta-analysis of reading rate. J Mem Language, 2019, 109: 104047
24 Broadus R N. Toward a definition of "bibliometrics". Scientometrics, 1987, 12: 373–379
25 Jordan R. The San Francisco declaration on research assessment. Biology Open, 2013, 2: 533–534
26 Hicks D, Wouters P, Waltman L, et al. Bibliometrics: The Leiden Manifesto for research metrics. Nature, 2015, 520: 429–431
27 Trueger N S, Thoma B, Hsu C H, et al. The altmetric score: a new measure for article-level dissemination and impact. Ann Emergency Med, 2015, 66: 549–553
28 Purkayastha A, Palmaro E, Falk-Krzesinski H J, et al. Comparison of two article-level, field-independent citation metrics: field-weighted citation impact (FWCI) and relative citation ratio (RCR). J Informetrics, 2019, 13: 635–642
29 Tong S C, Chen F Y, Yang L Y, et al. Novel utilization of a paper-level classification system for the evaluation of journal impact: an update of the CAS Journal Ranking. Quant Sci Stud, 2023, 4: 960–975
30 Tong S, Chen F, Yang L, et al. Novel utilization of a paper-level classification system for the evaluation of journal impact: an update of the CAS Journal Ranking. Quantitative Sci Studies, 2023, 4: 960–975
31 Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neural Computation, 2017, 29: 2352–2449
32 Schmarje L, Santarossa M, Schroder S M, et al. A survey on semi-, self- and unsupervised learning for image classification. IEEE Access, 2021, 9: 82146–82168
33 Chandra M A, Bedi S S. Survey on SVM and their application in image classification. Int J Inf Tecnol, 2021, 13: 1–11
34 Liu Y, Zhang Y, Wang Y, et al. A survey of visual transformers. IEEE Trans Neural Netw Learn Syst, 2024, 35: 7478–7498
35 Liu L, Ouyang W, Wang X, et al. Deep learning for generic object detection: a survey. Int J Comput Vis, 2020, 128: 261–318
36 Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis. Med Image Anal, 2017, 42: 60–88
37 Shen W, Peng Z, Wang X, et al. A survey on label-efficient deep image segmentation: bridging the gap between weak supervision and dense prediction. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 9284–9305
38 Qureshi I, Yan J, Abbas Q, et al. Medical image segmentation using deep semantic-based methods: a review of techniques, applications and emerging trends. Inf Fusion, 2023, 90: 316–352
39 Xun S, Li D, Zhu H, et al. Generative adversarial networks in medical image segmentation: a review. Comput Biol Med, 2022, 140: 105063
40 Ma J, Chen J, Ng M, et al. Loss Odyssey in medical image segmentation. Med Image Anal, 2021, 71: 102035
41 Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng, 2020, 34: 50–70
42 Nasar Z, Jaffry S W, Malik M K. Named entity recognition and relation extraction. ACM Comput Surv, 2021, 54: 1–39

43 Al-Moslmi T, Gallofre Ocana M, Opdahl A L, et al. Named entity extraction for knowledge graphs: a literature overview. IEEE Access, 2020, 8: 32862–32881
44 Liu P, Guo Y, Wang F, et al. Chinese named entity recognition: the state of the art. Neurocomputing, 2022, 473: 37–53
45 Weegar R, Pérez A, Casillas A, et al. Recent advances in Swedish and Spanish medical entity recognition in clinical texts using deep neural approaches. BMC Med Inform Decis Mak, 2019, 19: 1–4
46 Qu X, Gu Y, Xia Q, et al. A survey on Arabic named entity recognition: Past, recent advances, and future trends. 2023. ArXiv:2302.03512
47 Yadav A, Vishwakarma D K. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev, 2020, 53: 4335–4385
48 Yue L, Chen W, Li X, et al. A survey of sentiment analysis in social media. Knowl Inf Syst, 2019, 60: 617–663
49 Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation, 2019, 31: 1235–1270
50 Zhao W X, Zhou K, Li J, et al. A survey of large language models. 2023. ArXiv:2303.18223
51 Zhou C, Li Q, Li C, et al. A comprehensive survey on pretrained foundation models: a history from Bert to ChatGPT. 2023. ArXiv:2302.09419
52 Yu J, Yin H, Xia X, et al. Self-supervised learning for recommender systems: a survey. IEEE Trans Knowl Data Eng, 2024, 36: 335–355
53 Page M J, McKenzie J E, Bossuyt P M, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 2021, 372: n71
54 Kostoff R N, Hartley J. Open letter to technical journal editors regarding structured abstracts: this letter proposes that structured abstracts be required for all technical journal articles. J Inf Sci, 2002, 28: 257–261
55 Beller E M, Glasziou P P, Altman D G, et al. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med, 2013, 10: e1001419
56 Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell, 2014, 37: 1113–1133
57 Nadeau D, Sekine S. Named entities: recognition, classification and use. Lingvisticae Investigationes, 2007, 30: 3–26
58 Wang Y, Albrecht C M, Braham N A A, et al. Self-supervised learning in remote sensing: a review. IEEE Geosci Remote Sens Mag, 2022, 10: 213–247
59 Wankhade M, Rao A C S, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev, 2022, 55: 5731–5780
60 Huijben I A M, Kool W, Paulus M B, et al. A review of the Gumbel-max trick and its extensions for discrete stochasticity in machine learning. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 1353–1371
61 Kitchenham B, Pretorius R, Budgen D, et al. Systematic literature reviews in software engineering—a tertiary study. Inf Software Tech, 2010, 52: 792–805
62 Memon J, Sami M, Khan R A, et al. Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). IEEE Access, 2020, 8: 142642
63 Bandini A, Zariffa J. Analysis of the hands in egocentric vision: a survey. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 6846–6866
64 Guo Y, Wang H, Hu Q, et al. Deep learning for 3D point clouds: a survey. IEEE Trans Pattern Anal Mach Intell, 2020, 43: 4338–4364
65 Hussain T, Muhammad K, Ding W, et al. A comprehensive survey of multi-view video summarization. Pattern Recognition, 2021, 109: 107567
66 Yang L, Jiang H, Song Q, et al. A survey on long-tailed visual recognition. Int J Comput Vis, 2022, 130: 1837–1872
67 Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst, 2022, 33: 6999–7019
68 Lu J, Gong P, Ye J, et al. A survey on machine learning from few samples. Pattern Recognition, 2023, 139: 109480
69 Ibrahim H, Liu F, Zaki Y, et al. Google Scholar is manipulatable. 2024. ArXiv:2402.04607
70 Martín-Martín A, Thelwall M, Orduna-Malea E, et al. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics, 2021, 126: 871–906
71 Nahid A A, Kong Y. Involvement of machine learning for breast cancer image classification: a survey. Comput Math Methods Med, 2017, 2017: 1–29
72 Zhou D W, Wang Q W, Qi Z H, et al. Class-incremental learning: a survey. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 9851–9873
73 Gu J, Han Z, Chen S, et al. A systematic survey of prompt engineering on vision-language foundation models. 2023. ArXiv:2307.12980
74 Prabhavalkar R, Hori T, Sainath T N, et al. End-to-end speech recognition: a survey. IEEE ACM Trans Audio Speech Lang Process, 2024, 32: 325–351
75 Xing Z, Feng Q, Chen H, et al. A survey on video diffusion models. ACM Comput Surv, 2025, 57: 1–42
76 Melnik A, Ljubljanac M, Lu C, et al. Video diffusion models: a survey. 2024. ArXiv:2405.03150
77 Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. 2023. ArXiv:2306.13549
78 Wu J, Gan W, Chen Z, et al. Multimodal large language models: a survey. In: Proceedings of 2023 IEEE International Conference on Big Data (BigData), 2023. 2247–2256
79 Li J, Lu W. A survey on benchmarks of multimodal large language models. 2024. ArXiv:2408.08632
80 Huang J, Zhang J. A survey on evaluation of multimodal large language models. 2024. ArXiv:2408.15769
81 Wang L, Yoon K J. Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 3048–3068
82 Xu X, Li M, Tao C, et al. A survey on knowledge distillation of large language models. 2024. ArXiv:2402.13116
83 Janga B, Asamani G, Sun Z, et al. A review of practical AI for remote sensing in earth sciences. Remote Sens, 2023, 15: 4112
84 Cheng G, Huang Y, Li X, et al. Change detection methods for remote sensing in the last decade: a comprehensive review. Remote Sens, 2024, 16: 2355
85 Xiao Y, Tian Z, Yu J, et al. A review of object detection based on deep learning. Multimed Tools Appl, 2020, 79: 23729–23791
86 Kaur R, Singh S. A comprehensive review of object detection with deep learning. Digital Signal Processing, 2023, 132: 103812
87 Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Phil Trans R Soc A, 2021, 379: 20200209
88 Chen Z, Ma M, Li T, et al. Long sequence time-series forecasting with deep learning: a survey. Inf Fusion, 2023, 97: 101819
89 de la Torre-López J, Ramírez A, Romero J R. Artificial intelligence to automate the systematic review of scientific literature. Computing, 2023, 105: 2171–2194
90 Tian Y, Gu X, Li A, et al. Overview of the NLPCC2024 shared task 6: scientific literature survey generation. In: Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing, 2024. 400–408
91 Wang Y, Guo Q, Yao W, et al. Autosurvey: large language models can automatically write surveys. In: Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024
92 Lai Y, Wu Y, Wang Y, et al. Instruct large language models to generate scientific literature survey step by step. In: Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing, 2024. 484–496
93 Hu Y, Li Z, Zhang Z, et al. Hireview: hierarchical taxonomy-driven automatic literature review generation. 2024. ArXiv:2410.03761
94 Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. 2023. ArXiv:2302.13971
95 Zybaczynska J, Norris M, Modi S, et al. Artificial intelligence-generated scientific literature: a critical appraisal. J Allergy Clin Immunol-Pract, 2024, 12: 106–110
96 Elali F R, Rachid L N. AI-generated research paper fabrication and plagiarism in the scientific community. Patterns, 2023, 4: 100706
97 Aydın Ö, Karaarslan E. OpenAI ChatGPT generated literature review: Digital twin in healthcare. In: Emerging Computer Technologies

2. İzmir: İzmir Akademi Dernegi, 2022
98 Aydin Ö. Google bard generated literature review: metaverse. J AI, 2023, 7: 1–14
99 Paper Digest. Paper digest AI-powered research platform. https://www.paperdigest.org/review/. 2023
100 Altum Inc. Jenni AI-your AI research assistant. https://jenni.ai/. 2023
101 Seamlees. Seamless-AI literature review tool for scientific research. https://seaml.es/. 2023
102 BlockTechnology Oü. AI-powered literature review generator. https://askyourpdf.com/tools/literature-review-writer. 2023
103 Zhu K, Feng X, Feng X, et al. Hierarchical catalogue generation for literature review: a benchmark. 2023. ArXiv:2304.03512
104 Wang H, Fu Y, Zhang Z, et al. Llm×mapreduce-v2: entropy-driven convolutional test-time scaling for generating long-form articles from extremely long resources. 2025. ArXiv:2504.05732
105 Liang X, Yang J, Wang Y, et al. Surveyx: academic survey automation via large language models. 2025. ArXiv:2502.14776
106 Yan X, Feng S, Yuan J, et al. Surveyforge: on the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. 2025. ArXiv:2503.04629
107 Wen Z, Cao J, Wang Z, et al. Interactive survey: an LLM-based personalized and interactive survey paper generation system. 2025. ArXiv:2504.08762
108 Shi X, Kou Q, Li Y, et al. Scisage: a multi-agent framework for high-quality scientific survey generation. 2025. ArXiv:2506.12689
109 Wang J, Liu Z, Zhao L, et al. Review of large vision models and visual prompt engineering. Meta-Radiology, 2023, 1: 100047
110 kaixindelele. Chatpaper. https://github.com/kaixindelele/ChatPaper. 2023
111 de Winter J. Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. Scientometrics, 2024, 129: 2469–2487
112 Zhao P, Xing Q, Dou K, et al. From words to worth: newborn article impact prediction with LLM. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 1183–1191
113 Guo D, Yang D, Zhang H, et al. Deepseek-r1: incentivizing reasoning capability in LLMs via reinforcement learning. 2025. ArXiv:2501.12948
114 Chang Y, Feng Y, Sun J, et al. Sridbench: benchmark of scientific research illustration drawing of image generation model. 2025. ArXiv:2505.22126
115 Zhang C, Liu L, Cui Y, et al. A comprehensive survey on segment anything model for vision and beyond. 2023. ArXiv:2305.08196
116 Zhang C, Puspitasari F D, Zheng S, et al. A survey on segment anything model (SAM): vision foundation model meets prompt engineering. 2023. ArXiv:2306.06211
117 Kirillov A, Mintun E, Ravi N, et al. Segment anything. 2023. ArXiv:2304.02643
118 Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. 2023. ArXiv:2312.00752