SCIENCE CHINA
Information Sciences

• Supplementary File •

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[1,2*]

[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*

Appendix A Details of the proposed indicators.

In this section, we provide further insights into the proposed indicators. As presented in Tab. A1, various previous efforts have aimed to tackle the challenge of comparing metrics across fields. For instance, Paper [1] analyzes two well-known field-independent citation metrics at the article level: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). This study suggests that both metrics perform similarly in normalizing citations across research domains. However, these metrics require predefined field categorizations (such as the Scopus All Science Journal Classification) and may not adequately assess papers in emerging subfields. The Field Normalized Citation Success Index (FNCSI), introduced in Refs. [2, 3], is defined as the likelihood that a paper published in Journal A is cited more frequently than a randomly selected paper from Journal B. Although FNCSI offers robustness, it should be noted that it depends on predefined topic keywords and is specifically suited for journal-level assessments.

As the Leiden Manifesto [4] states: metrics should "*Account for variation by field in publication and citation practices*". We develop the concept of impact indicators, which is a measure used to gauge the impact of a certain paper in its field. The reason for using the term "indicator" rather than "metric" is we believe that the impact of a certain paper cannot be fully and accurately measured. While metrics like citation counts, h-index, or journal impact factors could indicate a paper's influence within the academic community, they fail to capture all aspects of its impact. For instance, a research paper might lead to significant advancements in theory, methods, or understanding in its field, none of which would necessarily be reflected in academic metrics. Similarly, a paper might contribute novel concepts or techniques that become influential over time but are not initially evident in citation counts. Hence, the term "indicator" is favored, as it suggests a signal without asserting to encompass the entirety of the academic paper's contribution.

Metric	Assessing Level	Normalized	Pre-defined Keywords Free
Citation Counts	A/J	×	✓
Impact Factor	J	×	×
FNCSI [3]	J	Field and Value	×
CiteScore [5]	J	×	×
SNIP [6]	J	Field	×
FWCI [7]	A	Field	×
RCR [8]	A	Field	×
TNCSI(*Ours*)	A/J	Field and Value	✓

Table A1 Metrics for evaluating scholar impact of papers: "A" and "J" stand for article-level and journal-level. Field-normalized signifies that the metric can be utilized across fields. Value normalized indicates that the range of the value is between 0 and 1.

Appendix A.1 TNCSI

$TNCSI$ is a metric used to assess the academic impact of articles across various disciplines. The fundamental distinction of $TNCSI$ from other metrics lies in its independence from the predefined keyword. The requirement for a predefined keyword

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

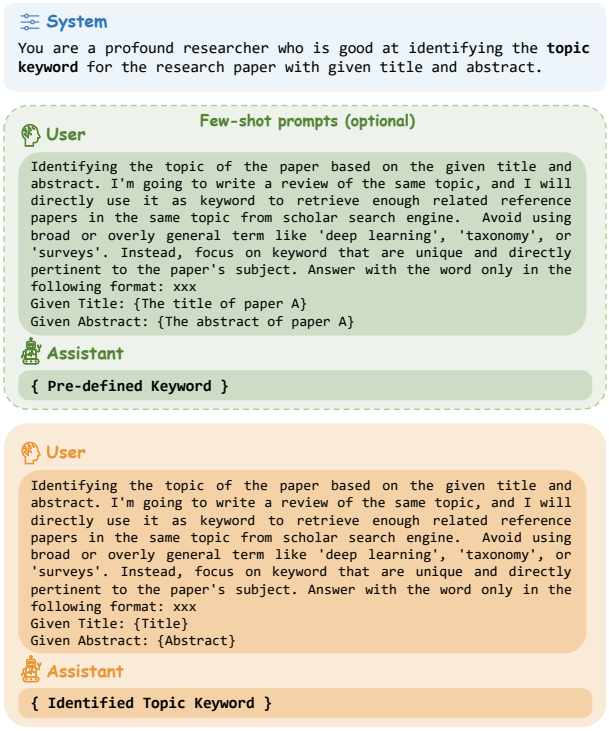

Figure A1 Conceptual illustration of topic keyword generation process: Few-shot prompting may enhance the response quality of the large language model.

is a primary reason that most bibliometric indicators face challenges in facilitating cross-disciplinary comparisons. To address such a limitation, we propose adopting ChatGPT [9] (gpt-3.5-turbo-0125) to generate the topic keyword and retrieve related papers according to the keyword. As illustrated in Fig. A1, we ask ChatGPT to identify the most representative topic keyword with the paper's title and abstract. ChatGPT is one of the most advanced and influential LLMs in the field of natural language processing [10]. Equipped with state-of-the-art language understanding capabilities, ChatGPT has revolutionized the way we interact with AI-powered conversational systems by simply setting "system", "user", and "assistant" roles. The "system" role sets the conversation's behavior and initial context. It provides instructions to guide the assistant's responses. The "user" role represents the individual interacting with ChatGPT, who inputs messages to the assistant. The "assistant" role is the ChatGPT model itself which would respond based on the provided instructions and user input.

We count the citations of papers with the help of Semantic Scholar API. For each paper, we retrieved up to $k = 1000$ relevant sources using the API. Based on the selected k papers and the corresponding citation counts p_c for each paper p, we can calculate the discrete citation frequency distribution of the k papers in a certain topic. We may further consider the distribution as a probability mass function:

$$P(X = x) = \frac{\text{Citation}_x}{k}, \tag{A1}$$

where $Citation_x$ represents the number of papers with x citations. The k papers related to the topic and their meta-data could be retrieved using a pre-defined topic keyword through online scholar search engines or API, such as Semantic Scholar, CrossRef, or Google Scholar.

Considering that the number of papers with x citations generally follows an exponential decay, we utilize the maximum likelihood estimation method to fit $P(X = x)$ and obtain the probability density function (PDF) of a continuous exponential decay distribution:

$$f(x) = \lambda e^{-\lambda x}, x \geqslant 0. \tag{A2}$$

Next, we will calculate $TNCSI$ by performing a definite integral on $f(x)$.

$$TNCSI = \int_0^{\text{citeNum}} f(x)\, dx, x \geqslant 0, \tag{A3}$$

where $f(x)$ represents the probability density at the value x, and λ is the results obtained from the maximum likelihood estimation, representing the scale parameter controlling the scaling. Finally, the definite integral of $f(x)$ over the interval $[0, citeNum]$ gives us the desired $TNCSI$:

In most scenarios, the generated topic word is sufficient to meet expectations, which can be further used as the keyword to retrieve papers from online scholarly search engines. Optionally, one can set "System", "User", and "Assistant" roles

NO.	User Prompt Content	Few-shot	NED↓
1	Please analyze the title and abstract provided below and identify the main topic or central theme of the review paper. Focus on key term and the overall subject matter to determine the primary area of research or discussion.The output should be formatted as following: xxx	×	0.75
2	Given title and abstract, please provide the searching key phrase for me so that I can use it as keyword to search highly related papers from Google Scholar or Semantic Scholar. Please avoid responding with overly general keyword such as deep learning, taxonomy, or surveys, etc. Answer with the words only in the following format: xxx	×	0.40
3	Identifying the topic of the paper based on the given title and abstract. Avoid using broad or overly general term like 'deep learning', 'taxonomy', or 'surveys'. Instead, focus on keyword that is unique and directly pertinent to the paper's subject. Answer with the word only in the following format: xxx	×	0.36
4	Identifying the topic of the paper based on the given title and abstract. So that I can use it as keyword to search highly related papers from Semantic Scholar. Avoid using broad or overly general term like 'deep learning', 'taxonomy', or 'surveys'. Instead, focus on keyword that is unique and directly pertinent to the paper's subject. Answer with the word only in the following format: xxx	×	0.32
5	Identifying the topic of the paper based on the given title and abstract. I'm going to write a review of the same topic and I will directly use it as keyword to retrieve enough related reference papers in the same topic from scholar search engine. Avoid using broad or overly general term like 'deep learning', 'taxonomy', or 'surveys'. Instead, focus on keyword that are unique and directly pertinent to the paper's subject. Answer with the word only in the following format: xxx	×	0.29
6	Identifying the topic of the paper based on the given title and abstract. I'm going to write a review of the same topic and I will directly use it as keyword to retrieve enough related reference papers in the same topic from scholar search engine. Avoid using broad or overly general term like 'deep learning', 'taxonomy', or 'surveys'. Instead, focus on keyword that are unique and directly pertinent to the paper's subject. Answer with the word only in the following format: xxx	✓	**0.28**

Table A2 Effectiveness of prompt engineering: The few-shot approach slightly improves the performance of topic keyword extraction.

before the final query to improve response quality and create more tailored interactions with the ChatGPT. In other words, a few-shot user-assistant pair prompts the ChatGPT with context on topic granularity. For example, a paper about the classification of irises by an improved CNN may have different perspectives. Some researchers focus more on the algorithm of the improved CNN, while others may be interested in classifying irises. Such ambiguity would likewise make it difficult for ChatGPT to identify the most representative topic keyword as expected. However, this could be addressed by providing the context which consists of (1) the identical prompt template with the replaced title and abstract as user input, and (2) the expected topic keyword as assistant output. By default, we adopt a well-known paper [11] as an example to guide models to generate the topic keyword for all papers.

Similar to other LLMs, ChatGPT adapts its responses based on user-provided natural language prompts. However, natural language prompts can be ambiguous, and different prompts may lead the model to respond with varying quality. Thus, we follow the practices of LLM prompt engineering and carefully design the prompt to optimize the desired output. To determine the optimal prompt, we construct a dataset by manually annotating the topic keywords of 201 papers from various domains published later than the ChatGPT being trained and then compare the performance of multiple prompts on this dataset. The normalized edit distance [12] is adopted to measure the similarity between the GPT-generated keyword and our annotated keyword, where a lower value indicates a higher quality of the prompt. As can be seen from Tab. A2, some of the designed prompts achieve decent NED scores for papers in various domains.

It is noteworthy that the proposed indicator exhibits a certain degree of robustness with respect to the choice of keywords. This robustness stems from the underlying mechanisms of modern search engines, which rely primarily on semantic similarity rather than strict keyword matching. As a result, even when the keywords differ, the retrieved literature remains largely consistent. In Table A3, we report the KL divergences of citation distributions obtained from searches with different but

semantically related keywords. The results show that the average KL divergence remains below 0.1, indicating that the metric is highly stable under variations in keyword phrasing.

Anchor Term	Comparison Terms	Avg. KL
Object Detection	Target Detection, Object Localization	0.080
Face Detection	Facial Landmark Detection, Face Recognition	0.052
Image Classification	Object Categorization, Visual Classification	0.086
Pose Estimation	Human Pose Detection, Body Pose Estimation	0.194
Semantic Segmentation	Scene Segmentation, Class-Level Segmentation	0.049
Generative Adversarial Networks	GAN, Adversarial Networks	0.014
Object Tracking	Visual Tracking, Moving Object Detection	0.075
Image Super-Resolution	Image Enhancement, Super-Resolution Imaging	0.014
Action Recognition	Activity Recognition, Gesture Recognition	0.090
3D Reconstruction	Three-Dimensional Reconstruction, Scene Reconstruction	0.010
Pedestrian Detection	Person Localization, Human Detection	0.042
Vehicle Detection	Car Detection, Automobile Detection	0.113
Text-to-Image Generation	Visual Synthesis, Image Generation from Text	0.020
Neural Style Transfer	Style Transfer, Image Transformation	0.147
Speech Recognition	Voice Recognition, Speech-to-Text	0.044
Object Recognition	Object Identification, Visual Recognition	0.076
Medical Imaging	Radiology Imaging, Clinical Imaging	0.011
Robotic Vision	Machine Vision, Automated Vision	0.022
Data Augmentation	Data Synthesis, Synthetic Data Generation	0.179
Human-Computer Interaction	User Interface Interaction, HCI	0.005
Scene Graph Generation	Object-Relationship Graph, Contextual Graphs	0.161
Overall	—	0.071

Table A3 Average KL divergence of synonym groups: The metric is computed in a single direction with the first term of each group as the anchor, and the "Comparison Terms" column lists the alternative expressions. These values are generally low, indicating robustness under keyword variations.

To provide a clear understanding of the proposed quality indicators, we render the graphical representations of $TNCSI$ in Fig. A2 left panel. As can be seen in Fig. A2, $TNCSI$ equals the area under the probability density function curve, which is fitted with the use of maximum likelihood estimation. In the right panel, $TNCSI$ is contrasted with the percentile-based citation indicator, which corresponds to the empirical cumulative distribution function (ECDF). Unlike the smooth curve of $TNCSI$, the ECDF is discrete, meaning that different citation counts may share the same percentile value. This discreteness can introduce unavoidable errors in certain tasks such as network training [13].

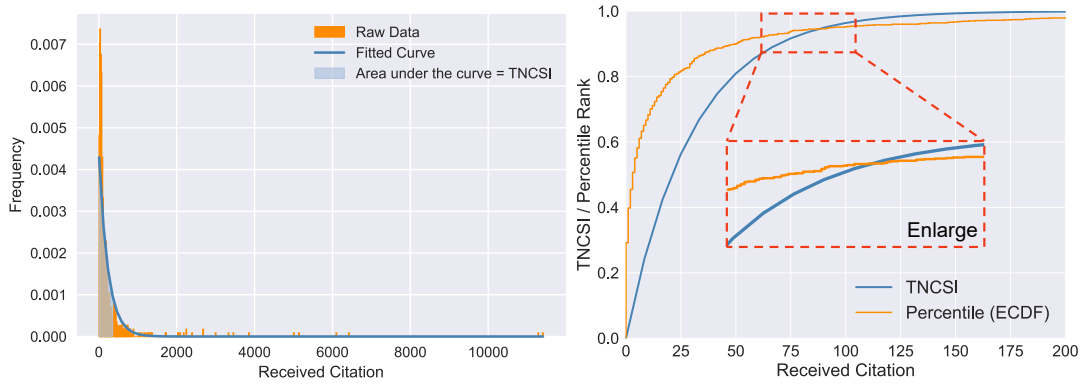

Figure A2 Illustration of the proposed quality indicators: The $TNCSI$ is defined as the area under the fitted probability density function curve estimated via maximum likelihood, resulting in a smooth cumulative representation. In contrast, the percentile-based citation indicator corresponds to the empirical cumulative distribution function (ECDF), which is discrete and may assign the same percentile value to different citation counts.

Appendix A.2 IEI

The IEI is an indicator used to measure changes in citation trends over a given period. A direct method for evaluating changes in citation trends is to apply polynomial fitting to the scatter data and calculate the sum of the derivatives at each point. However, in practice, polynomial fitting methods tend to be sensitive to outliers when applied to data lacking discernible distribution patterns, potentially compromising their robustness significantly. Consequently, the numerical values obtained may not accurately reflect the underlying citation trend. In contrast to these series analysis-based methods, we propose a morphological theory-grounded Impact Evolution Index (IEI), which converts the citation trend into a clear and interpretable numerical value.

To calculate the IEI, we first need to obtain the distribution of citation counts over time since publication. Once the citation data is retrieved, we may create a sequence Seq_{citation} about the number of citations. The $i \in \{0, 1, 2..., l\}$ item in the sequence $Seq_{\text{citation}}[i]$ represents the number of received citations in the i_{th} month after the publication, where l represents the number of months allocated for trend observation. Then, a sequence Seq_{time} of the same length as Seq_{citation} is generated by enumerating from 0 to $l-1$, Typically, the minimum recommended value for l is 6 or higher. This ensures that the data used for the analysis is adequately representative and the results are reliable. We match the items at the same positions in Seq_{time} and Seq_{citation} to determine a set of discrete coordinates $\{(Seq_{\text{time}}[i], Seq_{\text{citation}}[i])\}$, which serve as the control points for shaping the Bézier curve. A Bézier curve is a mathematical representation of smooth curves commonly used in computer graphics, image editing, and design software. The curve starts at the first control point and ends at the last control point, while the intermediate control points influence the curvature and direction of the curve. The number of control points determines the degree $n = l - 1$ of the curve.

$$C(t) = \sum_{i=0}^{n} B_{i,n}(t) P_i, \tag{A4}$$

$$B_{i,n}(t) = \binom{n}{i} (1-t)^{n-i} t^i, t \in [0, 1], \tag{A5}$$

where $B_{i,n}(t)$ represents the coefficient of the Bézier curve at a given parameter value t, which determines the position along the curve ($t = 0$ means the start and $t = 1$ means the end). $\binom{n}{i}$ is the binomial coefficient, also known as "n choose i". It represents the number of ways to choose i elements from a set of n elements. P_i stands for the i_{th} control point of the curve.

Given the continuity of the Bézier curve, we can compute its derivative as follows:

$$C'(t) = n \cdot \sum_{i=0}^{n-1} B_{i,n-1}(t) \cdot (P_{i+1} - P_i)\,. \tag{A6}$$

The tangent vector C'_a at the a-th point on the Bézier curve is further given by Eq. (A7).

$$\begin{aligned} C'_a &= n \cdot \sum_{i=0}^{n-1} B_{i,n-1}(\frac{a}{n}) \cdot (P_{i+1} - P_i)\,, \\ &= (x_a, y_a),\ a = 0, 1, \cdots, n, \end{aligned} \tag{A7}$$

where x_a and y_a are components of the vector, representing its magnitude along the x and y axes respectively.

Finally, the IEI_{L_l} can be obtained by averaging the slope of $l = n + 1$ distinct points on the curve. Moreover, different months may contribute differently to the IEI. For instance, if we desire closer months to have a greater impact, we can achieve this by adjusting the weighting coefficients, w_a, of the slope at different points to calculate their weighted averages

(See in Eq. (A8)). In addition, the instantaneous trend could be regarded as the slope of the last month in the sequence. It can be obtained by setting $w_n = 1$ and the other weighting coefficients to 0. We denote the IEI focused on the last month among the latest l months (excluding the current month) as IEI_{I_l}, as depicted in Eq. (A9).

$$IEI_{W_l} = \sum_{a=0}^{n} \frac{w_a(y_a/x_a)}{n+1}, \tag{A8}$$

$$IEI_{I_l} = C'(1) = n \cdot (P_n - P_{n-1}). \tag{A9}$$

Note that the value of l can be configured flexibly to meet actual demands. In general, the longer the period being analyzed, the more stable the citation trend becomes. We usually prefer to analyze the most recent 6 months of citations when constructing a Bézier curve of degree 5 and calculate IEI_{L_6}, IEI_{W_6}, and IEI_{I_6} based on the curve.

The illustration of the IEI is shown in Fig. A3. In Fig. A3, the horizontal axis labeled 0 to 5 inversely denotes the months prior to the current month, with 0 representing 6 months ago and 5 denoting the previous month. From the figure it can be observed that, compared with the fifth-degree polynomial fitting, the IEI avoids overfitting and produces a smoother and more stable trend. In contrast to the first- or third-degree polynomial fittings, which either oversimplify the dynamics or introduce spurious inflection points, the IEI achieves a balanced representation that captures the overall trajectory while maintaining robustness against local noise.

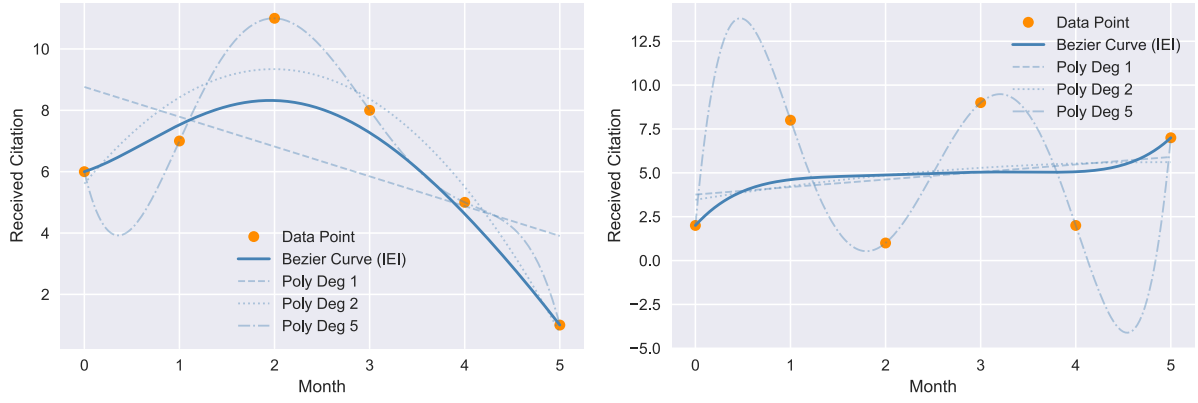

Figure A3 Visualization of the proposed IEI: IEI is the average of the derivatives of each control point on the Bézier curve.

Appendix A.3 RQM

The quality of references in a literature review is a multifaceted concept involving various aspects, such as credibility, relevance, breadth, and depth. Quantitatively assessing these elements is challenging. To begin with, precisely defining the relevance of scientific literature is difficult; both co-citation analysis [14] and similarity in paper embeddings have inherent conceptual limitations. For instance, while the concept of breadth in a literature review can theoretically be measured as the ratio of cited references to the total number of relevant references, accurately calculating this ratio is challenging. Identifying all relevant references through keyword searches or citation networks remains elusive.

Given the challenges in quantifying reference quality through the direct factors mentioned, we instead consider an indirect indication of reference quality using $TNCSI$. Here, $TNCSI$ serves as a quality indicator based on collective user assessments, effectively capturing a statistical summary of numerous researchers' evaluations of these direct factors for a given paper. Timeliness also matters. An up-to-date literature review ensures that the latest advancements, developments, and perspectives within a field are incorporated. This temporal relevance enhances the accuracy and effectiveness of research outcomes by reflecting the current state of knowledge. By emphasizing the currency of references, we can assess how thoroughly the literature review integrates the most recent research and developments in the field.

To account for both the quality and timeliness of cited references, we propose modifying the Gompertz function to model the reference quality . The Gompertz function is characterized by its sigmoidal, which indicates a slow growth rate at the start and end of a time period, with more rapid growth in the middle phase. This pattern is often observed in natural phenomena, such as species dynamics [15], tumor growth [16], etc.

The value of RQM is determined by the average reference quality (ARQ), the shift parameter β, and the median age of references S_{mp}. The calculation procedures of ARQ are as follows: The first step is the extraction of the cited reference list. For most publications, their reference lists could be provided by Semantic Scholar API. For a small number of reviews, the reference list provided by Semantic Scholar may contain errors. In this case, although there are powerful computer vision-based algorithms [17,18] available for extracting the reference list within PDFs, our requirements are relatively simple and can be effectively met by relying on the heuristic algorithms or ChatGPT. More specifically, for literature review with a relatively fixed citation format, we can use the PDFMiner [19] to read text from PDF files and use heuristic rules to match citations. Alternatively, the text can be analyzed using ChatGPT to extract in-text citations. The second step is similar to the calculation of the $TNCSI$, where the ChatGPT and a well-designed prompt (as presented in Fig. A1) are utilized to obtain the topic keyword of the review. Next, we calculate the $TNCSI$ for each reference in the list. To conserve

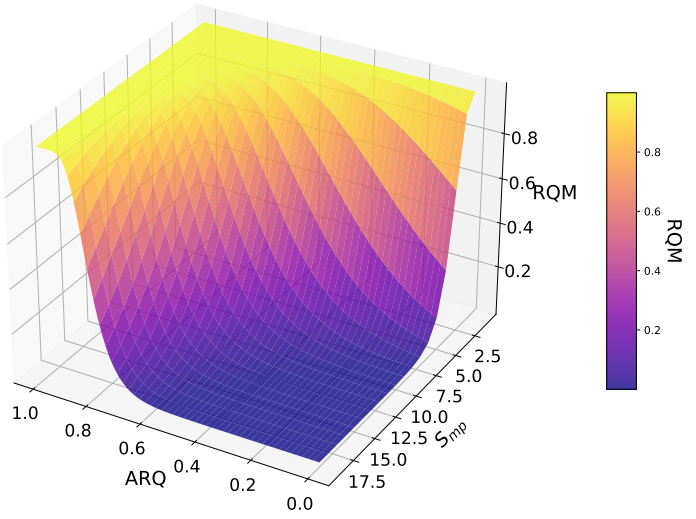

Figure A4 3D visualization of the proposed RQM

computational resources, we avoid using the ChatGPT to generate keywords for each reference. Instead, the $TNCSI$ of all cited literature is calculated using a sharing topic keyword. Finally, the coverage can be further calculated in Eq. (A10):

$$ARQ = \frac{\sum_{i=1}^{N_R} TNCSI(Ref_i)}{N_R}, \tag{A10}$$

where $TNCSI(\cdot)$ refers to the $TNCSI$ value of the i_{th} cited reference, and N_R stands for the number of the reference. In certain instances, it has been noted that calculating $TNCSI_s$ for each cited literature is also reasonable. However, this paper primarily emphasizes the current impact of the cited references, hence the utilization of $TNCSI$ in this context.

The shift parameter β can be set empirically or obtained statistically. For statistical estimation, we first examine the distribution of S_{mp} and ARQ of the RiPAMI database. The results indicate that (1) the S_{mp} of over 50% of the reviews falls within the [5,10] interval, where we denote the lower boundary as l_s and the upper boundary as r_s; (2) the average of ARQ is approximately 0.6, denoted as $\overline{ARQ}$. Then, the problem of asserting for β is reconceptualized as an optimization problem. As shown in Eq. (A11), the objective here is to identify the value of β that maximizes the integral of the absolute value of $RQM'(S_{mp}; \beta, \overline{ARQ})$ over the range l_s to r_s, subject to the constraints of $\overline{ARQ} = 0.6$. Such an approach would endow RQM with a more discriminative nature. It should be noted that different fields may result in distinct values of β. For this study, the β has been set as 5 for all fields.

$$\beta_{\text{opt}} = \arg\max_{\beta} \left(\int_{l_s}^{r_s} \left| RQM'(S_{mp}; \beta, \overline{ARQ}) \right| \, dS_{mp} \right). \tag{A11}$$

The range of RQM extends from 0 to 1, where values closer to 1 signify a higher quality of the referenced literature. As illustrated in Fig. A4, when the ARQ of a paper remains constant, an increase in the variable S_{mp} will lead to a decrease in the RQM value. Conversely, when S_{mp} remains constant, a higher ARQ will elevate the RQM value.

Appendix A.4 RUI

To evaluate the RUI, we may start with the coverage of references before and after publication. This coverage ratio can, to some extent, indicate the extent to which a review requires updating within its field. However, as mentioned earlier, accessing the coverage of a review is difficult. Fortunately, this problem is subtly avoided in calculating the ratio of relevant papers before and after publication. Assuming that the ratio of references containing the topic keyword in the title to all references is R_k, the total number of relevant articles can be estimated by dividing the number of articles containing those keywords retrieved from a search engine by R_k. Note that R_k generally remains consistent before and after the publication, the Coverage Difference Ratio (CDR) can then be calculated in Eq. (A12). The theoretical value range of CDR is greater than 0 to positive infinity. When the CDR of a review equals 1, it indicates that the current field has yielded new publications sufficient to constitute half of the literature referenced in the review.

$$CDR = \frac{N_{pc} \cdot R_k}{R_k \cdot N_{mp}} = \frac{N_{pc}}{N_{mp}}, \tag{A12}$$

where N_{mp} and N_{pc} denote the number of relevant literature from the median publication date of the cited references to the publication date of the review, and from the publication date of the review to the current time, respectively.

In addition, similar to the inevitable process of biological aging, literature reviews also undergo a gradual aging process throughout time. Such passage of time bestows upon literature reviews increasing aging progress, where the degree of aging

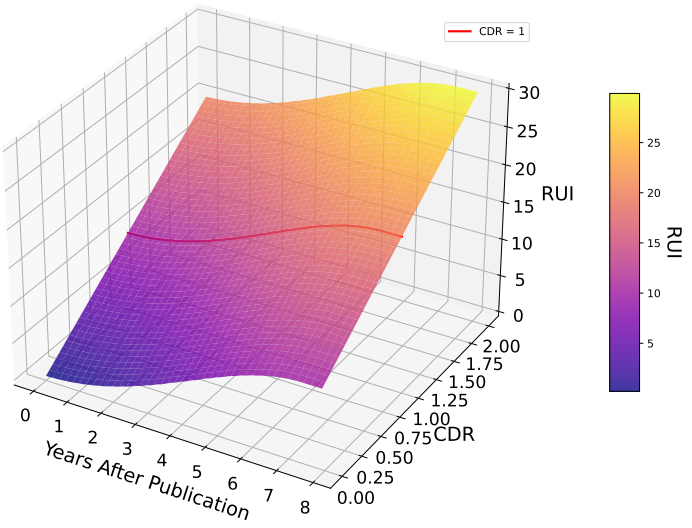

Figure A5 3D visualization of the proposed RUI

can be conceptualized as a normalized value of the academic impact already achieved. To further explore the aging of reviews in the field of PAMI, we conducted a statistical analysis of the yearly number of newly received citations for reviews published between 2015-2017 in RiPAMI. In contrast to earlier findings, however, the distribution of received citations of reviews over time follows a t-distribution rather than a log-normal distribution of the regular paper, as previously reported in Ref. [20–22]. Due to the insufficient duration of published sample data, the observation of citation-time trend curves is incomplete. Therefore, we conducted a three-degree polynomial fitting on the limited 6-year citation trend data and transformed the positive segment of the fitted curve into a PDF. To obtain the corresponding cumulative distribution function (CDF), we employed the cumulative trapezoidal numerical integration method for an approximate estimation. Thus, the Review Aging Degree (RAD) is given by:

$$RAD(M_{pc}) = \int_0^{M_{pc}/12} (px^3 + qx^2 + rx + s)\, dx, \tag{A13}$$

where M_{pc} denotes the duration in months from the publication of the review to the present, $p = -0.003$, $q = 0.001$, $r = 0.1267$, $s = 0.0129$ are the coefficients obtained by polynomial fitting. Please note that the integral symbol used here is for illustrative purposes only. The strict mathematical definition involves the accumulation of discrete trapezoidal areas.

Finally, the RUI could be obtained by a weighted summation of CDR and RAD. A sculptural visualization of the RUI's contours is crystallized in Fig. A5.

Appendix A.5 Ensuring fair use and avoiding metric misinterpretation

No metric is perfect. Researchers should recognize the limitations of metrics and avoid their misuse or misinterpretation. All citation-based metrics, in particular, are subject to various biases, such as those influenced by the Matthew effect, which can inadvertently contribute to the persistence and dissemination of inaccuracies within academic literature [23].

For TNCSI. $TNCSI$ is primarily used to assess the cumulative impact of literature reviews and should not be used to evaluate the quality of literature reviews, especially those published in different years.

For IEI. IEI solely represents citation trend changes over a specific past period and should not be extrapolated to predict future citation trends solely.

For RQM. As mentioned earlier, due to practical and conceptual constraints, RQM considers only the average impact and timeliness of references, without accounting for the relevance of the references to the investigated topic. Consequently, there is a potential for manipulating this metric by citing influential references unrelated to the topic to inflate the RQM score.

For RUI. As its name suggests, the RUI is a metric used to assess the extent to which a review requires updating. It is not designed to quantify the degree of obsolescence of the perspectives or conclusions within the review.

Additionally, it is important to note that all the metrics are designed for article-level analysis of reviews and should not be used to infer the development of an entire field based on the metrics of a single article.

Appendix A.6 Alignment with the Leiden Manifesto

Table A4 shows the alignment of our proposed four metrics with the specific guidelines of the Leiden Manifesto. The metrics—$TNCSI$, IEI, RQM, and RUI—demonstrate strong adherence to most principles outlined in the manifesto. Specifically, all four metrics align with principles emphasizing the integration of quantitative evaluation with qualitative

Principle	*TNCSI*	*IEI*	*RQM*	*RUI*
Quantitative evaluation should support qualitative, expert assessment.	✓	✓	✓	✓
Measure performance against the research missions of the institution, group or researcher.	×	×	×	×
Protect excellence in locally relevant research.	N/A	N/A	N/A	N/A
Keep data collection and analytical processes open, transparent and simple.	✓	✓	✓	✓
Allow those evaluated to verify data and analysis.	✓	✓	✓	✓
Account for variation by field in publication and citation practices.	✓	✓	✓	✓
Base assessment of individual researchers on a qualitative judgement of their portfolio.	N/A	N/A	N/A	N/A
Avoid misplaced concreteness and false precision.	✓	✓	✓	✓
Recognize the systemic effects of assessment and indicators.		✓		
Scrutinize indicators regularly and update them.	N/A	N/A	N/A	N/A

Table A4 Alignment of Leiden Manifesto principles with proposed metrics: "✓" indicates alignment with the principle, "×" indicates non-alignment, and "N/A" indicates principles not applicable.

expert assessment, transparency in data collection and analysis, verification of data and analysis by those evaluated, accounting for field-specific variations, and avoiding misplaced concreteness and false precision.

The metrics do not align with the principle of measuring performance against the research missions of the institution, group, or researcher, as indicated by the "×" in the corresponding row. We would like to point out that this reflects a deliberate trade-off made to balance engineering complexity. While more fine-grained measurements of performance would indeed enhance the objectivity and fairness of the metrics, they would inevitably increase the difficulty of data acquisition. In the future, with advancements in academic search engines and breakthroughs in natural language understanding, this issue is expected to be effectively addressed.

Appendix A.7 Illustrative applications of the indicators

Table. A5 provides more examples of how the four proposed indicators further assist researchers in refining their review selections. Compared to relying solely on titles, citation counts, and publication dates, the proposed metrics provide more information from various perspectives to assist researchers in review selections.

Appendix B Information extraction techniques

Appendix B.1 Word counts

Word counting differs from character counting, where methods relying solely on regular expressions often yield imprecise results due to the complexity of word segmentation rules. To achieve precise word counting, we utilize the Python NLTK library for word tokenization and exclude all tokens representing non-alphabetic elements. As a result, the presented word count only includes alphabetic words, excluding numbers, contractions, and any special symbols.

Appendix B.2 Visual elements counts

The method employed for counting visual elements extends beyond simply relying on the number of layout elements labeled as "Image" in the PDF (e.g., "LTImage" tag in the PDFMiner). Counting layout elements is prone to inaccuracies, as vector graphics formats like SVG and WMF can encapsulate multiple sub-bitmaps within a single graphic, leading to further errors in the counting process. To address this issue, we utilize LLM-based information extraction techniques to identify the captions associated with all images and tables in the review. The total number of visual elements is determined based on the caption count. As shown in Fig. B1, the first step involves initial filtering of the PDF document to identify chunks containing key terms such as "Fig" and "Tab." The document is first segmented into non-overlapping text chunks, each

Table A5 Comparison of various reviews with the proposed indicators and metrics.

Meta-Data					Evaluation			
Title	Year	Cites	Refs	Topic	TNCSI↑	IEI↑	RQM↑	RUI↓
Object Detection with Deep Learning: A Review [24]	2018	3533	252	deep learning-based objection detection	1.0	-7.51	0.96	99.45
Few-shot Object Detection: a Survey [25]	2022	34	70	few-shot object detection	0.26	-0.58	0.68	23.49
Recent Few-shot Object Detection Algorithms: A Survey with Performance Comparison [26]	2022	19	186	few-shot object detection	0.16	-0.69	0.66	25.47
Recent progresses on object detection: a brief review [27]	2019	38	110	object detection	0.03	0.08	0.36	43.15
A Survey of Modern Deep Learning Based Object Detection Models [28]	2021	619	113	object detection	0.41	-0.60	0.47	16.39
A Survey on Curriculum Learning [29]	2021	446	148	curriculum learning	0.92	1.09	0.14	14.25
Automatic Text Summarization Methods: A Comprehensive Review [30]	2022	36	102	automatic text summarization	0.36	0.24	0.01	6.25
Review of Automatic Text Summarization Techniques & Methods [31]	2020	181	117	automatic text summarization	0.90	0.32	0.06	22.10
Graph Self-supervised Learning: A Survey [32]	2021	453	183	graph self-supervised learning	0.99	-2.85	0.95	164.52
Self-supervised Learning on Graphs: Contrastive, Generative, or Predictive [33]	2021	207	139	graph self-supervised learning	0.90	-1.02	0.94	97.67
A Review of Deep-learning-based Medical Image Segmentation Methods [34]	2021	436	112	medical image segmentation	0.75	-3.11	0.26	23.36
Biomedical Image Segmentation: A Survey [35]	2021	20	157	biomedical image segmentation	0.13	-0.28	0.00	10.11
A Survey of Methods, Datasets, and Evaluation Metrics for Visual Question Answering [36]	2021	33	211	visual question answering	0.15	0.67	0.12	13.63
From Image to Language: a Critical Analysis of Visual Question Answering (VQA) Approaches, Challenges, and Opportunities [37]	2023	8	321	visual question answering	0.04	0.52	0.13	3.52
A Survey on Vision Transformer [38]	2020	1537	326	vision transformer	1.0	-0.13	0.94	1246.32
Transformers in Vision: A Survey [39]	2021	1959	285	transformers in computer vision	1.0	-6.41	0.98	456.40
A Survey of Visual Transformers [40]	2021	248	244	visual transformers	0.82	-0.18	0.97	115.19
A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking [41]	2023	18	99	efficient vision transformers	0.17	0.15	0.79	12.62
A Survey on Graph Diffusion Models: Generative AI in Science for Molecule, Protein and Material [42]	2023	37	151	graph diffusion models	0.56	-0.31	0.95	7.29
A Survey on Audio Diffusion Models: Text to Speech Synthesis and Enhancement in Generative AI [43]	2023	54	141	audio diffusion models	0.40	0.53	0.46	49.41
Diffusion Models: a Comprehensive Survey of Methods and Applications [44]	2022	921	394	diffusion models	0.91	-1.26	0.72	30.98
Geometric Deep Learning on Molecular Representations [45]	2021	234	208	molecular representations	0.93	1.43	0.79	23.56
Ensemble Deep Learning in Bioinformatics [46]	2020	194	116	ensemble bioinformatics	0.72	1.27	0.30	16.36
A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends [47]	2023	38	308	self-supervised learning	0.11	-0.11	0.31	11.21
Self-Supervised Learning: Generative or Contrastive [48]	2020	1341	184	self-supervised learning	0.98	-6.96	0.89	79.72
A Comprehensive Survey on Segment Anything Model for Vision and Beyond [49]	2023	61	223	segment anything model	0.95	0.06	0.97	166.69
A Survey on Visual Mamba [50]	2024	20	101	visual mamba	0.25	-2.17	0.95	45.37
A Survey on Large Language Model Based Autonomous Agents [51]	2023	692	193	LLM-based autonomous agents	1.0	-0.85	0.97	261.74
Retrieval-Augmented Generation for Large Language Models: A Survey [52]	2023	770	229	retrieval-augmented generation	1.0	6.11	0.97	101.36

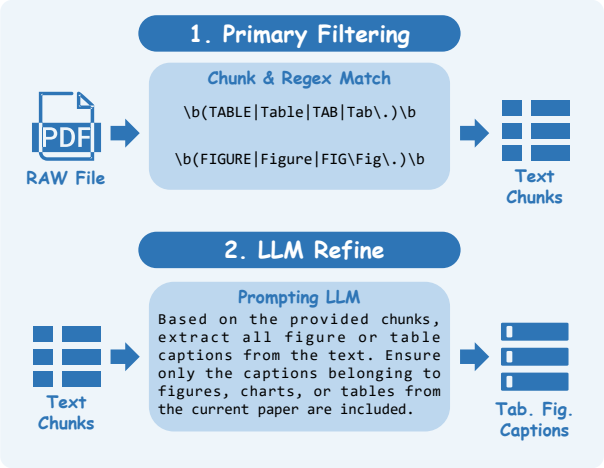

Figure B1 LLM for figure and table caption extraction: This LLM RAG-based approach allows for the retrieval of figure and table captions at an acceptable cost.

with a maximum length of 400 characters, using the newline character ('\n') as the delimiter. Next, regular expressions are applied within each chunk to filter and retain only those containing the key terms for subsequent analysis. Finally, we utilize the LLM with Retrieval-Augmented Generation (RAG) enabled to analyze these chunks and parse the response into a list of figure and table captions. This method allows for precise extraction of figure and table captions while avoiding the additional costs of full-document processing by the LLM.

Appendix B.3 Review feature extraction

This paper performs a statistical analysis of six features across more than 3,000 review samples, providing insights into factors such as compliance with PRISMA standards and the presence of application sections. Similar to visual element counting, these features are identified using LLM RAG-based information extraction methods.

Our proposed method for extracting review features is depicted in Fig. B2. The process begins with document analysis and recognition. Due to the inherent structure of PDF files, the machine-readable text order often diverges from the natural reading sequence. Additionally, elements such as captions and superscripts can introduce noise, disrupting semantic continuity during direct text extraction. To overcome these challenges, we propose utilizing the Nougat model [17] to process document images and convert them into structured text. This structured text enables efficient extraction of the Table of Contents (TOC) and content from individual sections.

We adopt a cost-effective approach during RAG by limiting the need for the LLM to process the entire text. In simple terms, specific sections of the article are selectively provided to the LLM to generate targeted responses. Alongside the title and abstract, the extracted introduction section and TOC are provided to guide the LLM in identifying whether the authors propose a new taxonomy or include a dedicated section for inclusion and exclusion criteria (indicating adherence to PRISMA guidelines) in the reviewed literature. To assess whether the authors discuss preliminaries, applications, or future challenges, we propose relying exclusively on TOC information, further minimizing costs. For benchmarking, the LLM is instructed to analyze extracted captions (illustrated in Fig. B1) to determine whether the authors have performed quantitative benchmarking of existing methods. In addition, the LLM is guided to judge whether the abstract follows structured-abstract standards [53]. All responses generated by the LLM are formatted into a structured 0/1 schema and stored in the RiPAMI database.

Appendix B.4 Position statistics of section titles

To analyze the structural characteristics of reviews, we compute the normalized positions of section titles that match a predefined set of keywords (e.g., *introduction*, *conclusion*). The procedure follows three steps.

First, the keyword list is standardized by removing empty entries, converting to lowercase, and eliminating duplicates while preserving the original order. Second, for each paper, all section titles are retrieved, and the total number of titles is used to normalize their positions within the document. Specifically, the index of each title is divided by $(n-1)$, where n denotes the number of titles in the paper, yielding a value between 0 (beginning of the document) and 1 (end of the document). Finally, each title is checked against the keyword list in a case-insensitive substring matching manner, allowing one title to be counted under multiple keywords if applicable.

This method produces, for each keyword, a distribution of normalized positions across the entire corpus, which forms the basis for our structural analysis.

Appendix C The relationship between review features and TNCSI

In this study, we dedicate our efforts to analyzing and synthesizing six key types of content features in review articles within the PAMI field. It is important to emphasize that the presence or absence of these features in an article should not be

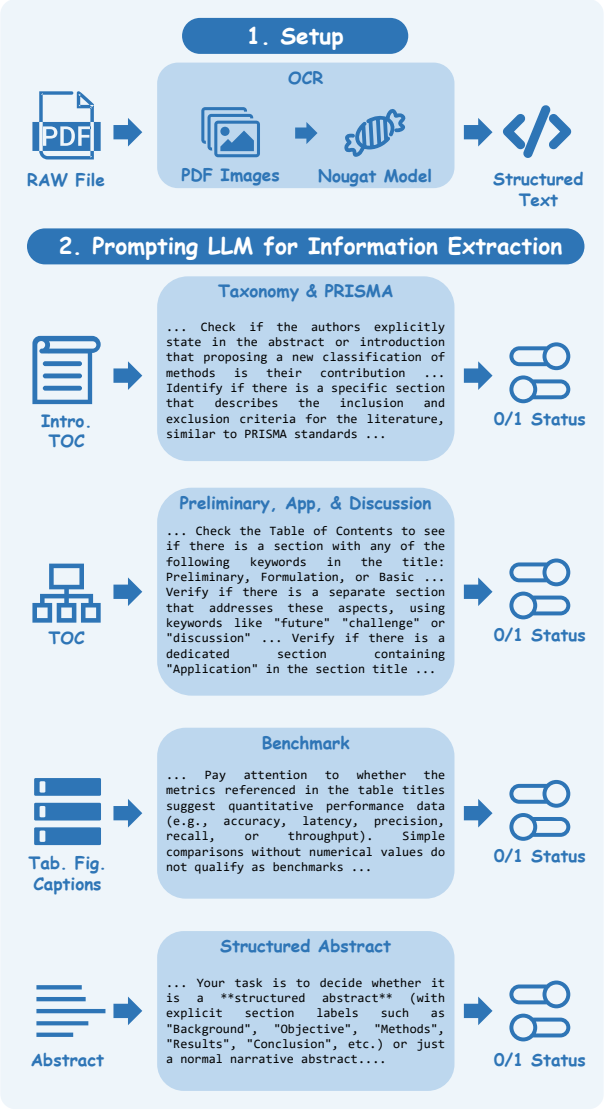

Figure B2 LLM for Review Features Extraction: By utilizing OCR technology [17], we extract structured text and input the introduction section (Intro.), table of contents (TOC), and captions of visual elements into the LLM to derive review features.

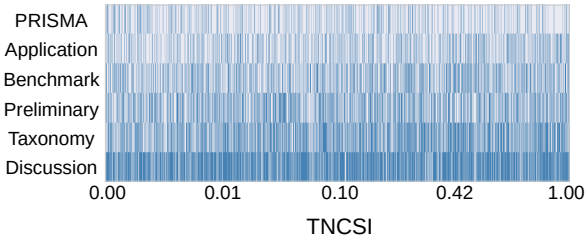

Figure C1 Analysis of six review content features and their association with $TNCSI$: Gene expression heatmap illustrating the independence between six review content features (PRISMA, Application, Benchmark, Preliminary, Taxonomy, and Discussion) and $TNCSI$. Darker shades indicate a match with the feature, corresponding to a value of 1.

interpreted as a direct measure of its overall quality or potential academic value. As shown in Fig. C1, there is no apparent correlation between the distribution of these six features and $TNCSI$.

Appendix D Complete list of sub-fields in CV, NLP, and MISC

For transparency and reproducibility, we provide a complete catalogue of subfields across the three major domains, namely CV, NLP, and MISC, as summarized in Table D1.

Appendix E Code framework: PyBiblion

This study involved extensive coding and engineering practices during the processes of dataset construction and statistical analysis. To support this work, we developed an open-source code library, PyBiblion, designed for bibliometric and statistical analysis. As its name suggests, PyBiblion is a Python-based library designed for bibliometric and statistical analysis, drawing from the Greek root "biblion", meaning "book" or "document" to emphasize its focus on scholarly literature.

PyBiblion offers several core advantages, one of which is lazy loading technology. Unlike most existing frameworks, PyBiblion implements lazy loading for paper information. This means that no network communication occurs during the initialization of an instance object; instead, data is only loaded when it is explicitly accessed. This mechanism helps to reduce the number of requests, alleviating server-side communication pressure and fostering a more equitable usage environment. The second advantage of PyBiblion lies in its user-friendliness. Extensive engineering optimizations have been implemented to integrate information retrieval and metric computation seamlessly. This design enables users to compute metrics such as the proposed $TNCSI$ and RQM with a single line of code, offering a highly efficient and intuitive experience. In addition, PyBiblion integrates numerous practical features, including database support, multithreading execution, visualization tools, and statistical analysis capabilities, further enhancing its functionality and versatility.

Table D1 Comprehensive subfield catalogue of CV, NLP, and MISC domains

Domain	**Subfields**
CV	action detection, action recognition, activity detection, activity recognition, anomaly detection, boundary detection, CNN, computer vision, depth estimation, edge detection, emotion recognition, face detection, face recognition, facial recognition, gesture analysis, gesture recognition, hand gesture recognition, handwriting recognition, human activity recognition, human detection, human pose estimation, image captioning, image classification, image clustering, image compression, image editing, image enhancement, image generation, image inpainting, image matching, image quality assessment, image recognition, image reconstruction, image restoration, image retrieval, image segmentation, image-based localization, instance segmentation, medical image analysis, medical image segmentation, object detection, object tracking, optical character recognition, person re-identification, point cloud, saliency detection, salient object detection, scene segmentation, scene understanding, super-resolution, superpixels, video object segmentation, video processing, video summarization, video understanding, visual question answering, visual tracking
NLP	dialogue modeling, dialogue systems, document analysis, document analysis and recognition, document clustering, document layout analysis, document retrieval, language modeling, language modelling, machine translation, named entity disambiguation, named entity recognition, natural language processing, question answering, relation extraction, sentiment analysis, sentiment classification, text classification, text clustering, text generation, text mining, text summarization, text-to-image generation, text-to-speech conversion, text-to-speech synthesis
MISC	adversarial attack, audio classification, biometric authentication, biometric identification, contrastive learning, data mining, data visualization, diffusion model, domain adaptation, graph mining, knowledge graph, knowledge representation, machine learning interpretability, meta-learning, metric learning, multi-label classification, pattern matching, pattern recognition, pre-training, pretraining, prompt learning, recommendation systems, recommender systems, remote sensing, representation learning, self-supervised learning, semantic segmentation, signature verification, speech emotion recognition, speech enhancement, speech recognition, speech synthesis, speech-to-text conversion, time series analysis, time series forecasting, topic detection, topic modeling, transfer learning, unsupervised learning, vision language model, word embeddings, zero-shot learning

References

1 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). Journal of Informetrics, 2019, 13: 635–642
2 Shen Z, Tong S, Chen F, et al. The utilization of paper-level classification system on the evaluation of journal impact. arXiv e-prints, 2020, pages arXiv–2006
3 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 Science Studies, 2023, pages 1–16
4 Hicks D, Wouters P, Waltman L, et al. Bibliometrics: the leiden manifesto for research metrics. Nature, 2015, 520: 429–431
5 Teixeira da Silva J A. Citescore: Advances, evolution, applications, and limitations. Publishing Research Quarterly, 2020, 36: 459–468
6 Moed H F. Measuring contextual citation impact of scientific journals. Journal of informetrics, 2010, 4: 265–277
7 FWCI. Field-weighted citation impact. https://libguides.usc.edu.au/researchmetrics/researchmetrics-field-weighted-citation-impact. Accessed 25 December 2023
8 Hutchins B I, Yuan X, Anderson J M, et al. Relative citation ratio (rcr): a new metric that uses citation rates to measure influence at the article level. PLoS biology, 2016, 14: e1002541
9 OpenAI. Chatgpt: optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022. Accessed 25 December 2023
10 Zhao W X, Zhou K, Li J, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023
11 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2020
12 Yujian L, Bo L. A normalized levenshtein distance metric. IEEE transactions on pattern analysis and machine intelligence, 2007, 29: 1091–1095
13 Zhao P, Xing Q, Dou K, et al. From words to worth: Newborn article impact prediction with llm. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, 2025. 1183–1191
14 Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 1973, 24: 265–269
15 Bruce R C. Application of the gompertz function in studies of growth in dusky salamanders (plethodontidae: Desmognathus). Copeia, 2016, 104: 94–100
16 Vaghi C, Rodallec A, Fanciullino R, et al. Population modeling of tumor growth curves and the reduced gompertz model improve prediction of the age of experimental tumors. PLoS computational biology, 2020, 16: e1007178
17 Blecher L, Cucurull G, Scialom T, et al. Nougat: Neural optical understanding for academic documents, 2023
18 Cheng H, Zhang P, Wu S, et al. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15138–15147
19 PDFMiner. PDFMiner.six. [Online; accessed 14-Nov-2023]
20 Matricciani E. The probability distribution of the age of references in engineering papers. IEEE Transactions on Professional Communication, 1991, 34: 7–12
21 Egghe L, et al. Citation age data and the obsolescence function: Fits and explanations. Information Processing & Management, 1992, 28: 201–217
22 Moed H F. Statistical relationships between downloads and citations at the level of individual documents within a single journal. Journal of the American Society for Information Science and Technology, 2005, 56: 1088–1097
23 Tatsioni A, Bonitsis N G, Ioannidis J P A. Persistence of contradicted claims in the literature. JAMA, 2007, 298: 2517–26
24 Zhao Z Q, Zheng P, Xu S t, et al. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 2019, 30: 3212–3232
25 Antonelli S, Avola D, Cinque L, et al. Few-shot object detection: A survey. ACM Computing Surveys (CSUR), 2022, 54: 1–37
26 Liu T, Zhang L, Wang Y, et al. Recent few-shot object detection algorithms: A survey with performance comparison. ACM Transactions on Intelligent Systems and Technology, 2023, 14: 1–36
27 Zhang H, Hong X. Recent progresses on object detection: a brief review. Multimedia Tools and Applications, 2019, 78: 27809–27847
28 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
29 Wang X, Chen Y, Zhu W. A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44: 4555–4576
30 Sharma G, Sharma D. Automatic text summarization methods: A comprehensive review. SN Computer Science, 2022, 4: 33
31 Widyassari A P, Rustad S, Shidik G F, et al. Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences, 2022, 34: 1029–1046
32 Liu Y, Jin M, Pan S, et al. Graph self-supervised learning: A survey. IEEE Transactions on Knowledge and Data Engineering, 2023, 35: 5879–5900
33 Wu L, Lin H, Tan C, et al. Self-supervised learning on graphs: Contrastive, generative, or predictive. IEEE Transactions on Knowledge and Data Engineering, 2021
34 Liu X, Song L, Liu S, et al. A review of deep-learning-based medical image segmentation methods. Sustainability, 2021, 13: 1224
35 Alzahrani Y, Boufama B. Biomedical image segmentation: a survey. SN Computer Science, 2021, 2: 1–22
36 Sharma H, Jalal A S. A survey of methods, datasets and evaluation metrics for visual question answering. Image and Vision Computing, 2021, 116: 104327
37 Ishmam M F, Shovon M S H, Mridha M, et al. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. arXiv preprint arXiv:2311.00308, 2023
38 Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45: 87–110
39 Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey. ACM computing surveys (CSUR), 2022, 54: 1–41
40 Liu Y, Zhang Y, Wang Y, et al. A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems, 2023, pages 1–21
41 Papa L, Russo P, Amerini I, et al. A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking. arXiv preprint arXiv:2309.02031, 2023
42 Zhang M, Qamar M, Kang T, et al. A survey on graph diffusion models: Generative ai in science for molecule, protein and material. arXiv preprint arXiv:2304.01565, 2023
43 Zhang C, Zhang C, Zheng S, et al. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. arXiv preprint arXiv:2303.13336, 2023, 2

44 Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 2023, 56: 1–39
45 Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nature Machine Intelligence, 2021, 3: 1023–1032
46 Cao Y, Geddes T A, Yang J Y H, et al. Ensemble deep learning in bioinformatics. Nature Machine Intelligence, 2020, 2: 500–508
47 Gui J, Chen T, Cao Q, et al. A survey on self-supervised learning: Algorithms, applications, and future trends. arXiv preprint arXiv:2301.05712, 2023
48 Liu X, Zhang F, Hou Z, et al. Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering, 2021, 35: 857–876
49 Zhang C, Liu L, Cui Y, et al. A comprehensive survey on segment anything model for vision and beyond. arXiv preprint arXiv:2305.08196, 2023
50 Zhang H, Zhu Y, Wang D, et al. A survey on visual mamba. Applied Sciences, 2024, 14: 5683
51 Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432, 2023
52 Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2023
53 Beller E M, Glasziou P P, Altman D G, et al. Prisma for abstracts: reporting systematic reviews in journal and conference abstracts. PLoS medicine, 2013, 10: e1001419