

BioMedJImpact: A Comprehensive Dataset and LLM Pipeline for AI Engagement and Scientific Impact Analysis of Biomedical Journals

Ruiyu Wang
Emory University
Atlanta, Georgia, USA
jonathan.wang@emory.edu

Yuzhang Xie
Emory University
Atlanta, Georgia, USA
yuzhang.xie@emory.edu

Xiao Hu
Emory University
Atlanta, Georgia, USA
xiao.hu@emory.edu

Carl Yang
Emory University
Atlanta, Georgia, USA
j.carlyang@emory.edu

Jiaying Lu
Emory University
Atlanta, Georgia, USA
jiaying.lu@emory.edu

Abstract

Assessing journal impact is central to scholarly communication, yet existing resources rarely capture how collaboration and artificial intelligence (AI) research jointly shape venue prestige in biomedicine. We present BioMedJImpact, a large-scale, biomedical-oriented dataset built from 1.74 million PubMed Central articles across 2,744 journals. BioMedJImpact integrates bibliometric indicators, collaboration features, and an LLM-derived AI engagement rate, defined as the proportion of AI-related articles within each journal-year. Specifically, AI engagement rate is extracted through a reproducible three-stage LLM pipeline. We analyze how collaboration intensity and AI engagement rate jointly influence scientific impact across two temporal subsets (2016–2019, 2020–2023). Two main patterns emerge: journals with larger author teams tend to have higher citation impact, while AI engagement rate is positively associated with Impact Factor only in the 2019 subset. To validate the LLM pipeline for deriving the AI engagement rate, we conduct human evaluation, confirming substantial agreement in AI relevance detection and consistent subfield classification. Together, BioMedJImpact provides both a comprehensive dataset at the interface of biomedicine and AI and a validated framework for scalable, content-aware scientometric analysis. Code and dataset are available at <https://github.com/JonathanWry/BioMedJImpact>.

CCS Concepts

• **Information systems** → **Data mining**; • **Computing methodologies** → *Natural language processing*; • **Applied computing** → *Life and medical sciences*.

Keywords

LLM for feature extraction, journal scientific impact analytics, sequential prompt engineering

1 Introduction

The scientific impact of journals plays a central role in academic communication, influencing research visibility, funding allocation, and institutional evaluation [6]. In biomedicine, metrics such as the impact factor (IF), citation counts, and journal ranking (e.g., JCR quartiles, SCImago Journal Rank) serve as key decision-making

tools for authors, institutions, and funders [4]. These indicators are sensitive to changes in the research landscape. For example, many general medical journals such as *The New England Journal of Medicine* and *The Lancet* saw sharp spikes in IF in 2021 due to the COVID-19 publication surge, followed by a return toward pre-pandemic levels in 2022 [9]. Meanwhile, artificial intelligence (AI) has increasingly transformed biomedical research [16, 23], reshaping how research is conducted and evaluated, and potentially altering traditional scientific impact indicators.

Existing open datasets on scientific impact (e.g., AMiner [19], DBLP [12], and Microsoft Academic Graph [22]) have greatly advanced large-scale analyses of scholarly networks and citation behavior; DBLP focuses on computer science, while AMiner and MAG provide broad, cross-disciplinary coverage. However, these existing resources are not designed specifically for the biomedical domain, and they lack the granularity needed to capture AI’s influence within biomedical research. To tackle these problems, we leverage multi-source data to build a new dataset on biomedical journals’ scientific impact, called BioMedJImpact. BioMedJImpact integrates three major categories of features: (a) Bibliometric indicators (e.g., impact metrics, citation counts), (b) Collaboration indicators (e.g., author diversity, institutional diversity), and (c) AI-related indicators (e.g., AI engagement rate, AI subfield distribution) that quantify the presence and distribution of AI-related research across journals. Specifically, bibliometric indicators are sourced from the Journal Citation Reports (JCR) and CiteFactor; collaboration indicators are derived from the PubMed Central Open Access subset (PMC-OA) metadata, capturing author and institutional structures; and the AI-related indicators are derived from article abstracts using a large language model (LLM)-based pipeline, enabling us to identify AI-related publications and their associated subfields. For downstream analysis, we construct two temporal subsets: **BioMedJImpact-2019** (2016–2019) and **BioMedJImpact-2023** (2020–2023).

In total, we build BioMedJImpact, a comprehensive dataset for analyzing biomedical journals’ scientific impact, consisting of 2,744 journals. We derive 55 features covering bibliometric, collaboration, and AI-related indicators. Using PMC-OA, we matched journals by source and year and identified 1,740,112 papers, processed through our LLM-based pipeline, yielding an overall AI engagement rate of 3.77%. Collaboration intensity—especially larger and more diverse

author teams—is consistently associated with higher citation impact, while AI engagement rate is associated with Impact Factor only in BioMedJImpact-2019. Overall, **BioMedJImpact** combines LLM-derived semantic indicators with traditional bibliometric and collaboration features to support scalable analysis of how content and collaboration shape scientific impact.

2 Related Work

2.1 Scientific Impact Modeling

Scientific impact has long been a core concern in scientometrics and information science, offering insight into how scholarly influence accumulates and providing practical tools for evaluating research quality, allocating funding, and guiding publication strategies [6]. Among various indicators, citations remain the primary quantitative signal, forming the basis of a family of *citation-based indicators* used to assess journals, authors, and individual papers [2]. At the journal level, Garfield’s journal impact factor formalized citation aggregation as a venue-level indicator [7], while the JCR quartile scheme (Q1–Q4) situates journals within disciplinary hierarchies, providing a coarse yet actionable measure of prestige for authors, editors, and institutions [10]. Despite well-documented limitations including field-normalization challenges and citation skewness, these citation-based indicators remain interpretable benchmarks that correlate with long-term scientific attention [21].

Beyond citation-based indicators, research has explored how various factors contribute to scientific impact. Collaborative indicators such as team size and co-authorship networks have been shown to correlate with citation influence across disciplines [11]. Studies also indicate that thematic and linguistic content learned from titles and abstracts encodes meaningful cues of scholarly influence [3].

2.2 Large Language Model-Based Feature Extraction

Alongside traditional feature extraction methods, recent advances in LLMs have fundamentally changed how features can be mined from scientific papers. Conventional feature extraction pipelines often rely on task-specific supervised models for named-entity, relation, and event extraction [20]. By contrast, LLMs enable prompt-based extraction that can screen documents for topical relevance, identify domain-specific terms, and map those terms to controlled taxonomies. Recent surveys document strong zero-shot and few-shot performance of LLMs for generative information extraction across domain-general tasks including named-entity, relation, and event extraction [25].

In biomedical corpora, LLMs have been applied to instruction following information extraction across core tasks including named-entity recognition, relation extraction, and procedure extraction [1, 24]. In concrete biomedical applications, LLMs have likewise demonstrated practical utility. In radiology, the RadEx framework provides a modular architecture for developing information-extraction systems from free-text radiology reports, supporting both generative and encoder-based models [17]. Similarly, LLMs have demonstrated their feasibility, accuracy, and efficiency for large-scale study design elements (PICO) extraction from clinical abstracts in PubMed [16].

3 Dataset Construction

3.1 Multisource Integration for Initial Dataset Construction

In this study, we curate BioMedJImpact, a comprehensive journal-level dataset for large-scale analysis of biomedical journal impact and AI engagement, by integrating data from: (i) the *PubMed Central Open Access subset* (PMC-OA) [14] for full-text and metadata of biomedical articles, (ii) *Journal Citation Reports* (JCR) [10] for journal bibliometric records including historical impact metrics and citation information, and (iii) the *Directory of Open Access Journals* (DOAJ) [13] for journal-level open-access policies and publication practices. Based on these resources, we assemble 17 per-journal, per-year indicators (see Table 2 for details). Among all data sources, PMC-OA serves as the core foundation of BioMedJImpact. It provides full-text and metadata for **4,298** biomedical journals. After matching these journals with available bibliometric records from JCR, we retain **2,744** journals for downstream analysis of content, citation patterns, and collaboration indicators. Of these, **1,694** journals are indexed in the DOAJ, enabling the integration of open-access policies and publication practices into the dataset.

Table 1: Summary of statistics for the BioMedJImpact-2019 and BioMedJImpact-2023 subsets.

Statistic \ Sub Dataset	-2019	-2023
# Journals	1367	2685
# Journals with Quartile	1243	2321
Percentage of Q1 journals	57.2%	46.92%
# Journals with IF	1367	2685
Avg IF	3.43	3.35
Std IF	3.12	4.01
# Journals with Total Cites_3Y	1247	2314
Avg Total Cites (3Y)	27263	33018
Std Total Cites (3Y)	116497	132542

3.1.1 Dataset/Year Split. To facilitate downstream modeling and isolate temporal effects, we partition the unified dataset into two temporal subsets: **BioMedJImpact-2019** (2016–2019) and **-2023** (2020–2023). Within each subset, we focus on three commonly used journal-level targets: *Impact Factor*, *Quartile*, and *Total Cites (3Y)*. These targets are widely used in academic assessment systems: Impact Factor reflects short-term citation influence, Quartile indicates a journal’s relative standing within its subject category, and Total Cites (3Y) captures sustained citation accumulation. The temporal split accounts for structural shifts in publishing and citation dynamics during the COVID-19 period, reducing confounding from pandemic-related disruptions. For each subset, we retain only journals with valid IFs in the subset’s target year. After filtering, **BioMedJImpact-2019** contains 1,367 journals and **BioMedJImpact-2023** contains 2,685 journals, with over 90% of journals including quartile rankings and citation-based metrics. Table 1 summarizes the retained sets.

3.1.2 Bibliometric Indicators. BioMedJImpact integrates bibliometric indicators from multiple publicly available sources. Historical journal indicators are collected from JCR hosted on ResearchGate¹

¹<https://www.researchgate.net/>

(2016–2024), and missing values were supplemented using CiteFactor². Extracted fields include journal title, ISSN/EISSN, subject category, quartile ranking (Q1–Q4), impact factor, and total citations. Policy attributes are integrated via cross-referencing with the DOAJ [13], which provides publication delay (in weeks), author copyright-retention status, and article processing charges. All sources are harmonized by ISSN/EISSN as unique identifiers, with fuzzy title matching applied for unresolved cases.

3.1.3 Collaboration Indicators. To characterize author collaboration patterns, we process full-text XML archives from PMC-OA [14]. PMC-OA provides rich article-level metadata, including author affiliations, article types, and author-supplied keywords. From each article, we extract the number of distinct institutions and participating countries using both structured tags (<institution>, <country>) and fallback string-pattern matching when such tags are absent.

We additionally define a *cross-country collaboration rate* as the proportion of articles with author affiliations spanning multiple countries. These metrics enable standardized comparisons of institutional and international collaboration intensity across disciplines and temporal spans.

3.1.4 Dataset Overview. After deriving the bibliometric and collaboration indicators, we conduct dataset characteristic analytics. Figure 1(a,b) depict the temporal distribution of journal quartiles from 2016 to 2023. The overall quartile composition remains stable, with the proportion of Q1 journals exhibiting a modest decrease between 2021 and 2022 (also reflected in the increased Shannon equitability index). Figure 1(c,d) show longitudinal trends in author and institutional collaboration. The median number of authors per paper ranges from 6.2 to 6.9, and no substantial differences in collaboration patterns are observed across Q1–Q4 journals.

3.2 LLM-Based Feature Enrichment for Journal AI Engagement

To capture how journals engage with the rapidly growing field of AI, we extend BioMedJImpact beyond traditional bibliometric and collaboration indicators by incorporating features extracted through an LLM-based pipeline. These content-level signals quantify each journal’s AI engagement and, when combined with structural and collaboration features, yield a more comprehensive dataset. The pipeline produces AI-related article shares, which are integrated into the full feature set detailed in Table 2.

3.2.1 LLM-Based Content Analysis on AI. To quantify AI involvement, we annotate PMC-OA article abstracts with an LLM to estimate journal-level AI engagement (Figure 2). We implement a three-step LLM pipeline with *Gemma-3-12B* [8]:

- (1) **Relevance Filtering Gate:** Each abstract is screened by an LLM prompt to determine whether it is relevant to artificial intelligence or machine learning. Abstracts containing phrases such as “deep learning-based model,” “AI-assisted diagnosis,” or “neural network training” are labeled as AI-relevant. Non-technical mentions (e.g., “intelligent design”) are filtered out.
- (2) **Keyword Extraction and Subfield Mapping:** Abstracts identified as AI-relevant are then processed by a second LLM prompt that simultaneously (1) extracts AI-related

Table 2: Summary of features. Feature γ_{-1} , γ_{-2} , and γ_{-3} denote covariates from one to three years prior to the prediction year.

Feature Group	Features Included
Bibliometric Indicators	• Impact Factor γ_{-1}, γ_{-2}, γ_{-3} • Quartile γ_{-1}, γ_{-2}, γ_{-3} • Total Cites (3Y) γ_{-1}, γ_{-2}, γ_{-3} • Total References γ_{-1}, γ_{-2}, γ_{-3} • Publication Count γ_{-1}, γ_{-2}, γ_{-3} • Publication Delay (in weeks) • Author Copyright Retention • Article Processing Charges • Subject Category
Collaboration Indicators	• Avg. Authors γ_{-1}, γ_{-2}, γ_{-3} • Std. Authors γ_{-1}, γ_{-2}, γ_{-3} • Author Quartiles Q_{25}, Q_{50}, $Q_{75} \times \gamma_{-1}$, γ_{-2}, γ_{-3} • Avg. Institutions γ_{-1}, γ_{-2}, γ_{-3} • Std. Institutions γ_{-1}, γ_{-2}, γ_{-3} • Institution Quartiles Q_{25}, Q_{50}, $Q_{75} \times \gamma_{-1}$, γ_{-2}, γ_{-3} • Cross-country collaboration rate γ_{-1}, γ_{-2}, γ_{-3}
AI-Related Indicators	• AI Engagement rate γ_{-1}, γ_{-2}, γ_{-3}

keywords (e.g., CNN, reinforcement learning) and (2) maps each abstract to one or more predefined AI subfields, including *Natural Language Processing*, *Computer Vision*, *Learning Algorithms*, *Knowledge Representation*, *Search methodologies*, and *Distributed AI*. This integrated keyword–subfield reasoning step enables consistent subfield assignment and supports downstream analysis of AI research themes.

- (3) **Validation Gate:** A secondary verification prompt re-evaluates extracted keywords to confirm their alignment with AI subfields and removes ambiguous or noisy terms (e.g., “training session”). This ensures semantic precision and minimizes false positives in downstream statistical analyses.

From this pipeline, we derive the *AI engagement rate*, defined as the proportion of articles within each journal–year that are classified as AI-related: $E_{j,t} = N_{j,t}^{AI} / N_{j,t}^{\text{total}}$, where $N_{j,t}^{AI}$ is the number of AI-related abstracts in journal j during year t , and $N_{j,t}^{\text{total}}$ is the total number of abstracts published by that journal in the same year.

3.2.2 AI Engagement Patterns. We summarize AI-related publication trends in Figure 3, based on LLM-annotated journal content. AI activity is most prominent in multidisciplinary areas linked to computer science and mathematics, with additional strong engagement in imaging-focused fields such as *Radiology and Medical Imaging*, and *Neuroscience*. In terms of dataset coverage, AI engagement indicators are available for 902 journals in 2019, including 631 with full three-year AI coverage. By 2023, this expands to 1,010 journals, with 803 fully covered over three years. Collaboration features show similarly broad availability, present in 1,095 journals in 2019 and 979 in 2023.

4 Correlation Analysis

To examine how journal features shape biomedical scientific impact, we fit linear mixed-effects (LME) models [15] with subject-category random intercepts, which account for persistent disciplinary differences while estimating within-field associations. In practice, the random-effect variance occasionally collapses to approximately zero or becomes non-identifiable. In such cases, we refit the same

²<https://www.citefactor.org>

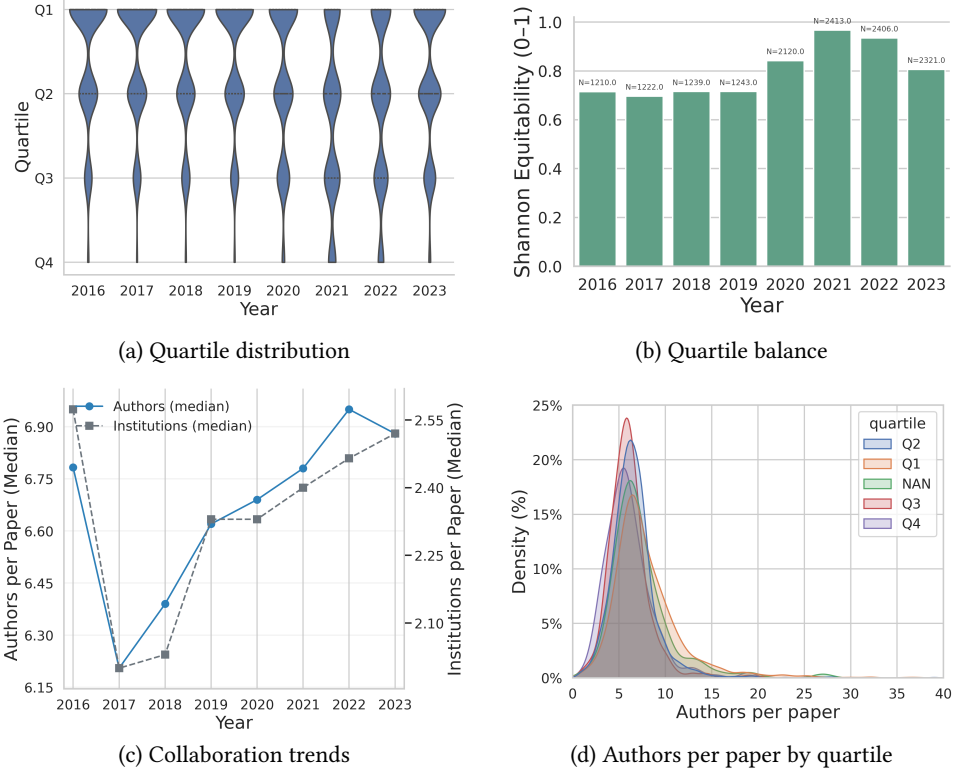


Figure 1: Dataset overview of quartile and collaboration indicators over time.

Table 3: LME/OLS correlation analysis for BioMedJImpact-2019. Reported significance uses Bonferroni-adjusted p-values within each model (* denotes $p_{adj} < 0.05$, ** denotes $p_{adj} < 0.01$, and *** denotes $p_{adj} < 0.001$).

Target	Variable	Coef.	Std. Err.	95% CI (L-H)	Signif.
Impact Factor	Avg_Authors_2016	0.882	0.092	[0.701, 1.063]	***
	AI_Engagement_Rate_2018	19.38	4.84	[9.89, 28.88]	**
Total Cites	Publication_Count_2016	-146.31	15.12	[-175.94, -116.68]	***
	Publication_Count_2018	136.93	14.31	[108.89, 164.97]	***
	Publication_Count_2017	108.13	19.45	[70.00, 146.25]	***
	Total_Refs_2018	-0.775	0.185	[-1.137, -0.413]	***
	Author_Copyright_Retention	-30673.59	7468.64	[-45311.86, -16035.31]	**
	Std_Institutions_2017	10995.30	3305.07	[4517.48, 17473.12]	*

specification using ordinary least squares (OLS), retaining the identical set of fixed effects.

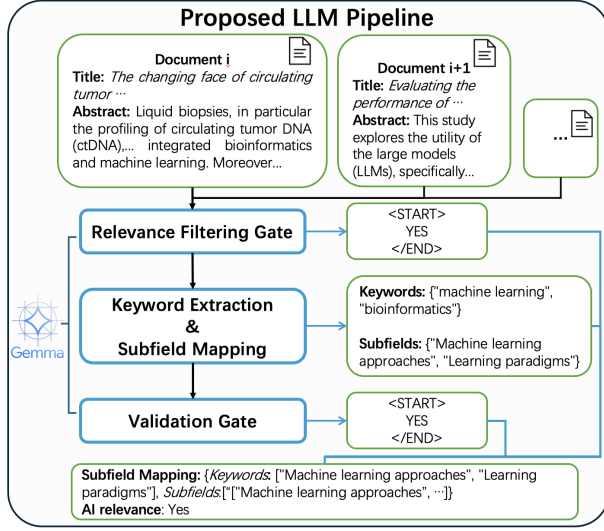
Because each model includes multiple fixed-effect coefficients, we correct for multiple comparisons *within each fitted model* using a Bonferroni adjustment over all fixed effects (excluding the intercept). Unless otherwise stated, statistical significance throughout this section refers to Bonferroni-adjusted p-values [5] (p_{adj}), with * $p_{adj} < 0.05$, ** $p_{adj} < 0.01$, and *** $p_{adj} < 0.001$. Tables 3 and 4 report only coefficients that remain significant after this correction.

For **Impact Factor**, collaboration intensity remains positively associated after multiple-testing correction. In BioMedJImpact-2019, the average number of authors per paper (Avg_Authors_2016) is strongly associated with higher impact ($\beta = 0.882$, $p_{adj} < 0.001$). AI engagement is also significant in BioMedJImpact-2019: a higher AI publication share in 2018 (AI_Engagement_Rate_2018) predicts

higher impact ($\beta = 19.38$, $p_{adj} < 0.01$). In BioMedJImpact-2023, Avg_Authors_2021 remains positively associated with impact ($\beta = 0.534$, $p_{adj} < 0.001$), whereas the corresponding AI engagement variables do not remain significant after Bonferroni correction. For **Total Cites**, temporal effects in publication volume persist after correction. In BioMedJImpact-2019, publication counts from 2016–2018 and reference volume in 2018 (Total_Refs_2018) remain significant, indicating that recent production and referencing practices correlate with subsequent citation totals. Policy and collaboration structure also matter: Author_Copyright_Retention is negatively associated with citations ($\beta = -3.07 \times 10^4$, $p_{adj} < 0.01$), while institutional diversity in 2017 (Std_Institutions_2017) is positive ($p_{adj} < 0.05$). In BioMedJImpact-2023, citations are most strongly associated with publication count and authorship variability, while Author_Copyright_Retention remains strongly negative.

Table 4: LME/OLS correlation analysis for BioMedJImpact-2023. Reported significance uses Bonferroni-adjusted p-values within each model (* denotes $p_{adj} < 0.05$, ** denotes $p_{adj} < 0.01$, and *** denotes $p_{adj} < 0.001$).

Target	Variable	Coef.	Std. Err.	95% CI (L-H)	Signif.
Impact Factor	Avg_Authors_2021	0.534	0.119	[0.300, 0.767]	***
	Publication_Count_2022	114.31	21.43	[72.31, 156.32]	***
Total Cites	Author_Copyright_Retention	-55034.58	12306.28	[-79154.44, -30914.72]	***
	Std_Authors_2020	3369.13	769.05	[1861.82, 4876.44]	***
	Std_Authors_2022	3018.81	880.83	[1292.42, 4745.20]	*
	Std_Authors_2021	-3672.30	1164.23	[-5954.14, -1390.45]	*

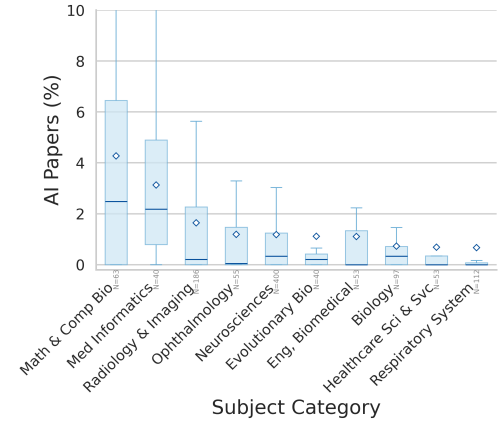
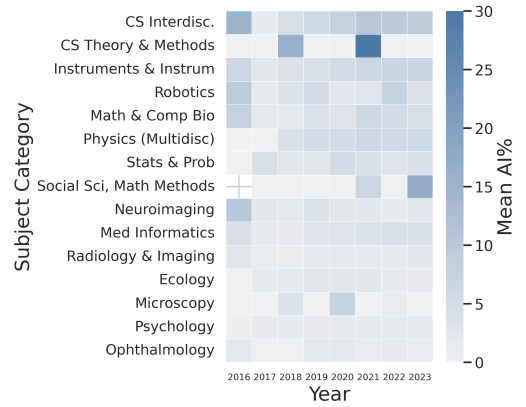
**Figure 2: Overview of our LLM pipeline for AI engagement rate analysis.** Step 1 determines whether the input document is AI-relevant. Step 2 extracts AI terms and maps them to a taxonomy. Step 3 further reduces ambiguity and false positives.

For **Quartile**, no fixed-effect coefficient remains significant under the Bonferroni correction in either period, suggesting quartile rank is less robustly explained by the included lagged features than continuous impact outcomes under this conservative correction.

5 Evaluation on LLM-based Feature Extraction

5.1 Quality Evaluation on Select Journal Categories

To visualize the semantic patterns detected by our model, we generate word clouds from validated AI-related keywords (Figure 4). For each subject category, keywords from AI-relevant articles are aggregated and weighted by relative frequency, providing an overview of dominant methodological themes. Clear differences emerge across disciplines. Math and Computational Biology journals feature a mix of classical machine learning and deep learning, with frequent mentions of *neural networks* and *random forests*. Mentions of *convolutional* and *graph neural networks* reflect applications to structured biological data. Radiology and Imaging is dominated by deep learning and computer-vision terms such as *convolutional networks* (CNN), *U-Nets*, and *segmentation*, reflecting its focus on supervised image analysis, while tree-based and linear models appear far less often. In contrast, Healthcare Science and Services emphasizes

**(a) AI% by subject category****(b) AI% by year and subject category****Figure 3: AI engagement rate patterns, where the rate denotes the proportion of AI-related articles within each journal-year.** Panel (a): Top-10 subject categories ranked by pooled mean AI% across all journal-year records; boxes show distributions, diamonds show means. Panel (b): Top-15 categories ranked by year-normalized mean AI%—the mean across journals per category-year, then averaged over years.

interpretable, clinically aligned models such as *logistic regression* and *random forests*, consistent with the needs of electronic health records and decision-support systems. Across fields, ML and DL form shared foundations, but their form reflects domain-specific data: CNNs dominate imaging, statistical and representation learning shape biology, and interpretable models underpin healthcare.

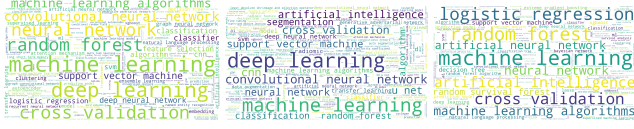
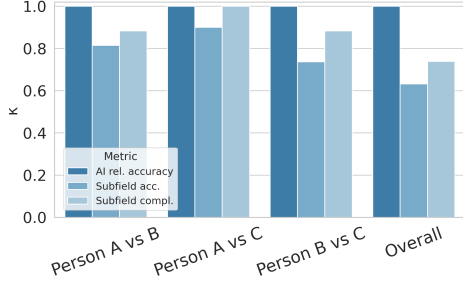
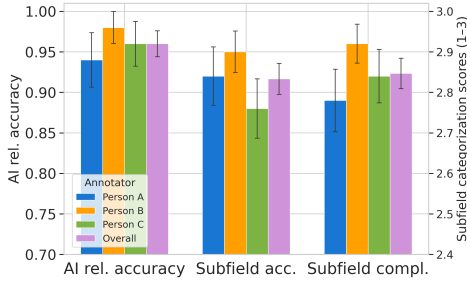


Figure 4: Subject category-specific word clouds of validated AI subfield keywords. From left to right: (a) Math and Computational Biology, (b) Radiology and Imaging, and (c) Healthcare Science and Services. Word size reflects frequency of extracted AI concepts within each journal subset.



(a) Agreement (κ)



(b) Mean scores with Std. error

Figure 5: Human evaluation results. (a) Pairwise Cohen's κ for each annotator pair and metric, with an overall bar showing Fleiss' κ across all three annotators. (b) Per-annotator and overall mean scores with standard errors: AI-relevance accuracy (left axis, 0–1) and subfield accuracy/completeness (right axis, 1–3).

These trends should be viewed cautiously. We classify AI subfields using the ACM CCS system [18], which, though widely used, is not tailored to biomedical research and may group diverse work into broad categories. Although supplemented with keyword extraction, we prioritized accurate category assignment over comprehensive coverage, so specialized methods may be underrepresented.

5.2 Human Evaluation of LLM Annotations

We evaluate the LLM pipeline on 100 stratified PMC-OA articles: 50 predicted AI-positive and 50 predicted AI-negative. Three author-team members (A, B, and C) participate in the study. We split articles into four sets of 25: one annotated by all raters and three assigned to a single rater, yielding 50 annotations per rater and 25 shared articles per pair. Each article is evaluated along three dimensions:

- **AI relevance accuracy:** whether the abstract involves artificial intelligence or machine learning.

- **Subfield accuracy (1-3):** how accurately LLM-identified AI subfields or keywords reflect the content based on the ACM CCS system (1 = incorrect, 2 = partially correct, 3 = completely correct).
- **Subfield completeness (1-3):** whether extracted AI subfields cover key AI-related technical aspects in the abstract based on ACM CCS system (1 = insufficient, 2 = partial, 3 = fully complete).

Annotators first evaluated AI relevance using only the title and abstract and were blinded to LLM relevance prediction. After submitting the relevance judgment, they were shown LLM-generated keywords and subfield assignments to evaluate subfield accuracy and completeness. We evaluate the reliability of LLM-generated annotations using κ statistics, which account for agreement beyond chance. Pairwise agreement is measured using Cohen's κ , while Fleiss' κ is reported for the 25 abstracts annotated by all three annotators. Since downstream subfield judgments are only meaningful when AI content is correctly detected, if the LLM misclassifies AI relevance, corresponding subfield accuracy and completeness scores are set to 1 by design. All Cohen's κ and Fleiss' κ statistics were computed on the 25 articles annotated by all three raters.

Figure 5(a) shows that AI relevance annotation reaches substantial agreement across annotators ($\kappa > 0.8$), demonstrating consistent human judgments of LLM's AI relevance detection. Subfield accuracy and completeness show moderate to substantial agreement ($\kappa > 0.6$), indicating higher subjectivity in assessing technical precision and coverage while reflecting consistent interpretation. Figure 5(b) reports individual and overall scores for each annotation dimension. AI relevance accuracy remains high across annotators, while slight variations in subfield completeness indicate stricter criteria among some annotators. Overall, the LLM-based pipeline demonstrates strong performance across evaluation dimensions, although author involvement may introduce confirmation bias.

6 Conclusion

We introduced **BioMedJImpact**, a large-scale biomedical dataset built from over 1.7 million PMC-OA articles across 2,700 journals, integrating bibliometric, collaboration, and LLM-derived AI indicators to understand how biomedical publishing evolves in the AI era. Using this dataset, we examine how collaboration and AI engagement shape scientific impact. We show that greater collaboration intensity is associated with higher citation impact. We found that AI-related article share is positively associated with Impact Factor in BioMedJImpact-2019 but not in BioMedJImpact-2023 after multiple-testing correction. The LLM-based annotation pipeline is validated through human evaluation, confirming substantial agreement in AI relevance detection and consistent subfield classification. Overall, **BioMedJImpact** offers a comprehensive dataset at the intersection of biomedicine and AI and a scalable methodology for content-aware scientometric analysis.

Ethical and Responsible Use. This study analyzes biomedical articles from the PubMed Central Open Access Subset through PMC-authorized bulk-access services for non-commercial research. We only release journal-level aggregates and derived indicators, rather than full texts or patient-level records, and follow applicable article licenses and PMC terms. The resulting resource is intended for scientometric research only.

References

- [1] Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu, and Zhe He. 2025. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digital Medicine* 8, 1 (2025), 166.
- [2] Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? A review of studies on citing behavior. *J. Doc.* 64, 1 (2008), 45–80.
- [3] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-Level Representation Learning Using Citation-Informed Transformers. In *Proceedings of the 58th ACL*. 2270–2282.
- [4] Peng Dong, Marie Loh, and Adrian Mondry. 2005. The "impact factor" revisited. *Biomed. Digit. Libr.* 2 (2005).
- [5] Olive Jean Dunn. 1961. Multiple Comparisons Among Means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.
- [6] Eugene Garfield. 1955. Citation indexes for science: A new dimension in documentation through association of ideas. *Science* 122, 3159 (1955), 108–111.
- [7] Eugene Garfield. 2006. The History and Meaning of the Journal Impact Factor. *JAMA* 295, 1 (2006).
- [8] Gemma Team. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786* (2025). arXiv:2503.19786 [cs.CL]
- [9] Sang-Jun Kim. 2024. Explosive increase and decrease in articles, citations, impact factor, and immediacy index during the COVID-19 pandemic: a bibliometric study. *Sci. Ed.* 11, 2 (2024), 107–113.
- [10] Anna Krampl. 2019. Journal Citation Reports. *J. Med. Libr. Assoc.* 107, 2 (2019), 278–280.
- [11] Vincent Larivière, Yves Gingras, Cassidy R. Sugimoto, and Andrew Tsou. 2015. Team Size Matters: Collaboration and Scientific Impact since 1900. *Journal of the Association for Information Science and Technology* 66, 7 (2015), 1323–1332. doi:10.1002/asi.23266
- [12] Michael Ley. 2002. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *SPIRE 2002*. Springer, 1–10.
- [13] Heather Morrison. 2017. Directory of open access journals (DOAJ). *Charlest. Advis.* 18, 3 (2017), 25–28.
- [14] National Library of Medicine. 2003. PMC Open Access Subset [Internet]. <https://pmc.ncbi.nlm.nih.gov/tools/openftlist/>. [cited 11/5/2025].
- [15] José C. Pinheiro and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY.
- [16] Tim Reason, Julia Langham, and Andy Gimblett. 2024. Automated Mass Extraction of Over 680,000 PICO from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study. *Pharm. Med.* 38, 5 (2024), 365–372.
- [17] Daniel Reichenpfader, Jonas Knupp, André Sander, and Kerstin Denecke. 2026. RadEx: A Framework for Structured Information Extraction from Radiology Reports Based on Large Language Models. In *dHealth 2026: Proceedings of the 20th Health Informatics Meets Digital Health Conference (Studies in Health Technology and Informatics, Vol. 335)*. 47–52. doi:10.3233/SHTI260053
- [18] Bernard Rous. 2012. Major update to ACM's computing classification system. *Commun. ACM* 55, 11 (2012), 12–12.
- [19] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-Miner: Extraction and Mining of Academic Social Networks. In *ACM SIGKDD*. 990–998.
- [20] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5784–5789. doi:10.18653/v1/D19-1585
- [21] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.
- [22] Kuansan Wang, Iris Shen, Charles Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft Academic Graph: When experts are not enough. *Quant. Sci. Stud.* 1, 1 (2020), 396–413.
- [23] Yuzhang Xie, Hejie Cui, Ziyang Zhang, Jiaying Lu, Kai Shu, Fadi Nahab, Xiao Hu, and Carl Yang. 2025. KERAP: A Knowledge-Enhanced Reasoning Approach for Accurate Zero-shot Diagnosis Prediction Using Multi-agent LLMs. In *Proceedings of AMLA*.
- [24] Yuzhang Xie, Xu Han, Ran Xu, Xiao Hu, Jiaying Lu, and Carl Yang. 2025. HypKG: Hypergraph-Based Knowledge Graph Contextualization for Precision Healthcare. In *International Semantic Web Conference*. Springer, 328–348.
- [25] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large Language Models for Generative Information Extraction: A Survey. *Frontiers of Computer Science* 18, 6 (2024), 186357. doi:10.1007/s11704-024-40555-y