International Journal of English Literature and Social Sciences
Vol-10, Issue-6; Nov-Dec, 2025

Peer-Reviewed Journal
Journal Home Page Available: https://ijels.com/
Journal DOI: 10.22161/ijels

A Quantitative Review of Ancient Chinese Book Digitization: Research Trends and Development Based on DeepSeek-V3 and BER Topic

Li Mingxia, Li Wenke

School of Publishing, University of Shanghai for Science and Technology, Shanghai 200093, China

Received: 07 Oct 2025; Received in revised form: 08 Nov 2025; Accepted: 14 Nov 2025; Available online: 19 Nov 2025

Abstract— *This study conducts a comprehensive quantitative analysis of the development and research trends of ancient Chinese book digitization over the past three decades. By employing advanced tools such as CiteSpace for bibliometric analysis, the BERTopic deep learning model for topic modeling, and the large language model DeepSeek-V3 for semantic trend interpretation, the research identifies the developmental stages of the field, eight major thematic areas, and emerging directions such as gamification, knowledge services, and AI model integration. Drawing upon a corpus of 562 peer-reviewed journal articles retrieved from the CNKI database, the study applies both statistical and semantic techniques to uncover key author clusters, topic evolution, and knowledge networks. This work contributes to the understanding of how digital and intelligent technologies are transforming the preservation, dissemination, and reuse of ancient texts, and provides theoretical support for the innovative transmission of traditional Chinese culture in the digital era.*

Keywords— ***Ancient books digitization; Deep learning; Large language model; Topic modeling; Digital humanities; Knowledge services***

I. INTRODUCTION

Ancient books are the crystallization of human wisdom in the age of print media and embody the essence and charm of traditional Chinese culture. With the rapid evolution of media technologies, there has been a profound transformation in reading habits, knowledge acquisition methods, and modes of information dissemination. As such, exploring how to preserve, disseminate, and repurpose ancient texts under digital and intelligent media conditions has become an essential path for achieving the creative transformation and innovative development of Chinese traditional culture.

In April 2022, the General Office of the CPC Central Committee and the State Council issued the 'Opinions on Promoting Ancient Book Work in the New Era', emphasizing the need to accelerate the transformation and utilization of ancient book resources by means such as excavating their contemporary value, enhancing accessibility, promoting digitization, and ensuring their widespread dissemination. Later that year, the National Plan for Ancient Book Collation and Publishing (2021–2035) was published, offering concrete guidance for the digitization of ancient texts.

The digitization of ancient books refers to the application of digital technology to restore, develop, and utilize classical literature resources. In China, this research began in the 1970s, progressing from basic text entry and indexing to the establishment of large-scale databases, and eventually advancing to deeper levels of digital processing through computational methods. This study integrates computer-based methodologies with the research of ancient texts. Using academic papers from the CNKI database published over the past thirty years, it combines bibliometric tools, deep learning models, and large language models to quantitatively analyze the current state of ancient book digitization and forecast its future development. The goal is to support the modernization and intelligent transmission of Chinese traditional culture.

II. RESEARCH METHODS AND TECHNICAL TOOLS IN THE FIELD OF ANCIENT BOOK DIGITIZATION IN CHINA

In China, bibliometric research on ancient book digitization has primarily relied on software tools such as CiteSpace to conduct quantitative analysis, generate knowledge maps, and extract thematic trends from journal databases. For example, Li Shiyu and colleagues employed CiteSpace to analyze both domestic and international research on ancient book digitization, integrating the Five-Source Theory and digital humanities frameworks to construct a research pathway that informs future development directions. Some scholars have focused on the digitization of traditional Chinese medicine (TCM) texts, using discipline life cycle theory along with tools like VOSviewer to identify developmental characteristics and thematic evolution patterns, as well as to predict future trends. Others, such as Wang Qiuyun, combined quantitative and qualitative methods to summarize the research scope, features, and trends in ancient book digitization, and proposed potential research directions for the field.

A review of existing literature reveals that scholars generally rely on traditional bibliometric tools like CiteSpace and VOSviewer for thematic analysis. These tools use built-in algorithms to extract keywords, conduct clustering, and analyze thematic trends. However, these methods are limited in capturing the deeper semantic content and core themes of the literature. To overcome this limitation, the introduction of deep learning models offers a more precise and scientifically grounded way to analyze textual content.

In this study, the deep learning model BERTopic is employed to significantly enhance both the depth and breadth of thematic extraction. BERTopic enables robust data-driven topic modeling, providing stronger empirical support for academic research. To further optimize the identification of thematic trends, this study incorporates the large language model DeepSeek-V3, developed by DeepSeek in 2024. DeepSeek-V3 is specifically optimized for Chinese natural language processing tasks and demonstrates superior capabilities in text comprehension, summarization, and information retrieval. Its powerful semantic extraction makes it particularly well-suited for trend identification in the field of ancient book digitization.

Additionally, this study utilizes the open-source network analysis software Gephi, which is cross-platform and JVM-based. Gephi supports exploratory data analysis, social network analysis, and biological network analysis through flexible data manipulation and layout options. Using Gephi's built-in community detection algorithms, this research maps the co-authorship network in the ancient book digitization field, revealing collaborative patterns among scholars and providing empirical evidence for understanding the knowledge production model and innovation dynamics in the field.

In summary, this study integrates CiteSpace bibliometric analysis, BERTopic topic modeling, DeepSeek-V3 large language model processing, and Gephi social network visualization to conduct a comprehensive and multifaceted analysis of 30 years of Chinese research literature on ancient book digitization. These digital humanities approaches enable a deeper understanding of research trends and future development trajectories in this domain.

III. DATA SOURCES AND DATA PROCESSING

3.1 Data Sources

This study draws data from the China National Knowledge Infrastructure (CNKI) database. Following the search strategy adopted by Fan Guihong and others, we used the following query: (Subject: 'Classics' AND 'Digitization') OR (Subject: 'Ancient Books' AND 'Digitization') OR (Subject: 'Ancient Books' AND 'Database'). The Boolean logic used was 'OR', the document type was limited to journal articles, and the sources were restricted to CSSCI and Peking University Core journals. The earliest publication meeting these criteria appeared in 1998; thus, the starting point for data collection was set to 1998, with an end date of May 20, 2025. A total of 891 papers were initially retrieved. After removing duplicates and irrelevant records through manual screening, 616 valid journal articles remained.

3.2 Data Processing

BERTopic is a topic modeling tool originally developed for English language texts. Therefore, prior to modeling, all Chinese texts were segmented using the Jieba Python library, which supports precise mode, full mode, and search engine mode. In this study, precise mode was chosen for its natural and accurate segmentation capabilities. Additionally, stopwords were removed using the Harbin Institute of Technology stopwords list to clean the corpus.

Some articles in the dataset lacked abstracts or keywords, which are essential for topic modeling. These records were excluded. After this filtering process, 562 valid documents with abstracts were retained for BERTopic modeling. These texts were further pre-processed and embedded using the 'gte-base-zh' model for Chinese text vectorization. Dimensionality reduction was performed using UMAP, followed by clustering with HDBSCAN. Finally, c-TF-IDF was applied to extract representative terms for each topic cluster.

This methodological pipeline ensures the accuracy and granularity of topic extraction and supports the

identification of thematic structures within the ancient book digitization research domain.

IV. RESEARCH FINDINGS

4.1 Annual Publication Volume Analysis: Developmental Stages and Characteristics

The number of academic publications over time is a key indicator of the developmental trajectory of a research field. By analyzing annual publication counts, we can gauge scholarly attention to the digitization of ancient books. Based on the year-by-year publication statistics (see Figure 1 in the original), the development of ancient book digitization in China can be categorized into four distinct stages:

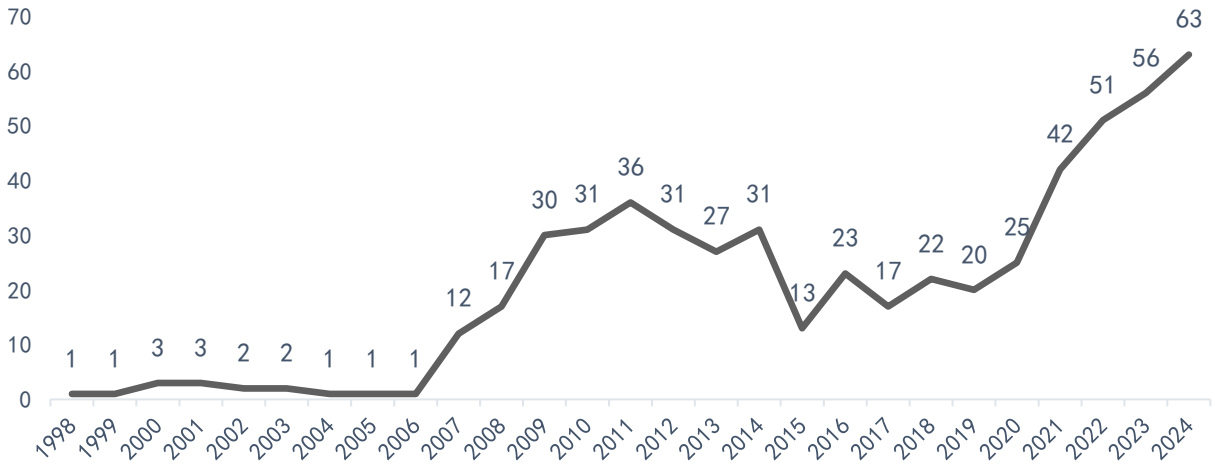

Fig.1. Annual Publication Volume on Ancient Book Digitization in China (1997–2024)

• Budding Stage (1998–2006):

During this period, research in the field was nascent, with relatively few publications. Topics primarily focused on theoretical discussions and general overviews, such as strategic development, trend analysis, and catalog construction. Most studies were conducted independently by individual scholars, and there was minimal inter-institutional collaboration. This phase was characterized by initial exploration and low network connectivity among researchers.

• Rapid Development Stage (2007–2014):

A notable surge in research activity followed the release of the 'Opinions on Strengthening Ancient Book Preservation' by the State Council in 2007. The number of annual publications increased steadily, peaking at 36 papers in a single year. This stage accounted for approximately 40% of all retrieved papers. Research themes expanded from theoretical discussions to include practical applications, such as digital standards for ancient books, case studies from abroad, and digital publishing. This period also witnessed the emergence of cross-disciplinary research involving traditional Chinese medicine databases and minority language texts. The maturation of digital technologies facilitated collaborative research groups, fostering diversification of research topics and deeper academic cooperation.

• Stabilization Stage (2015–2020):

From 2015 onwards, the number of annual publications declined slightly and stabilized at around 20 papers per year. Influenced by digital humanities and computational social science, the research increasingly incorporated interdisciplinary elements, including social network analysis, data mining, GIS visualization, and knowledge graphs. This phase reflected the field's growing integration with modern information technologies, consolidating prior achievements while also opening new research frontiers.

• Breakthrough Stage (2021–present):

Triggered by the issuance of the 'Opinions on Promoting Ancient Book Work in the New Era' in April 2022, the field entered a new phase of accelerated growth. By September 2024, the annual publication volume reached a historical high of 63 articles. Recent research emphasizes emerging technologies such as computational humanities and deep learning, marking a paradigm shift from content collation to knowledge extraction and service. In particular, studies on ancient TCM texts have deepened, and successful cases in literary and theatrical ancient book digitization offer valuable models. With ongoing policy support, research output in this field is expected to continue its upward

trajectory, further advancing the innovation and dissemination of ancient Chinese culture.

4.2 Author Productivity Analysis

Core authors play a central role in shaping the development of any academic field. Author productivity, especially the volume of publications by high-frequency contributors, serves as an indicator of academic influence and research leadership. Following the Price Law formula ($N = 0.749 \times \sqrt{Nmax}$, where Nmax is the highest frequency of publications by a single author), the threshold for core author productivity in this study is calculated to be approximately 4.17. Applying this threshold, authors with at least 4 publications are identified as core contributors in the field of ancient book digitization.

A total of 32 high-yield authors were identified, with the top 10 listed below. Among them, Wang Dongbo ranks first with 31 publications. His research primarily focuses on computational philology and digital humanities, and he has made substantial contributions to the integration of AI with ancient book studies. He is followed by Mao Jianjun, an early pioneer in the digitization of ancient books in China. Mao's work includes studies on electronic editions of ancient texts, copyright issues, and international digitization experiences.

Table 1. Publication Volume of Core Authors (Top 10)

Author	Number of Publications
Wang Dongbo	31
Mao Jianjun	23
Li Bin	15
Li Mingjie	12
Bao Ping	10
Zhang Wenliang	9
Chen Tao	7
Huang Shuiqing	7
Liu Jiangfeng	7
Chen Liping	6

These authors have played a pivotal role in defining the field's core research themes and have contributed significantly to advancing scholarly discourse on ancient book digitization.

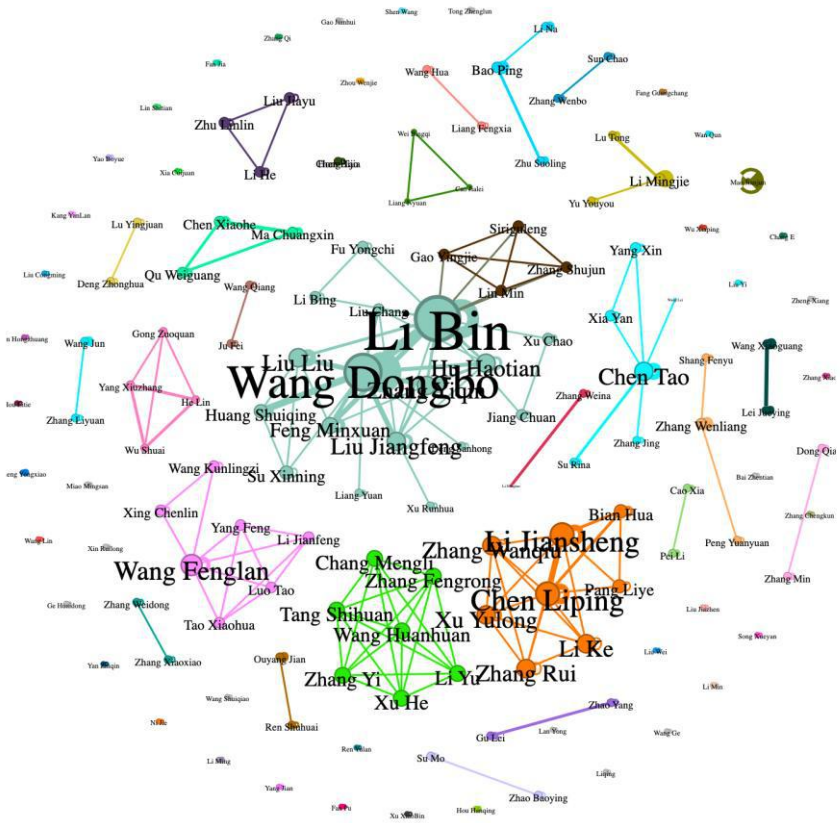

Fig.2. Co-authorship Network of Researchers in Ancient Book Digitization

4.3 Author Collaboration Network Analysis

Using Python, co-authorship data was extracted from the bibliographic metadata and structured into a co-occurrence matrix. This matrix was then imported into Gephi for visualization using the Fruchterman-Reingold algorithm to generate a network layout. The resulting author collaboration network comprises 143 nodes (authors) and 257 edges (collaborative links). The network exhibits an average degree of 3.594, average weighted degree of 9.832, a network diameter of 3, density of 0.025, modularity of 0.823, average clustering coefficient of 0.819, and average path length of 1.703.

These metrics suggest that, while no single dominant research cluster has yet emerged, collaboration among scholars is gradually increasing. Community detection using Gephi's modularity algorithm revealed 71 author clusters. To identify densely connected subgroups, a K-core filter (k=4) was applied, isolating the most cohesive co-authorship communities in the field.

Table 2. Top 5 Author Clusters in the Field of Ancient Book Digitization

Cluster	Author Cluster(s)	Research Field
Cluster1	Wang Dongbo, Li Bin, Liu Liu, Feng Minxuan, Zhang Yiqin, Huang Shuiqing, Liu Jiangfeng, Hu Haotian, Liu Chang, Xu Runhua, Li Bing, Liang Yuan, Jiang Chuan, Fu Yongchi, Su Xinning, Xu Chao, Deng Sanhong	digital humanities and knowledge services.
Cluster2	Li Jiansheng, Chen Liping, Bian Hua, Pang Liye, Zhang Wanqiu, Zhang Rui, Li Ke, Xu Yulong	TCM ancient texts and data minin
Cluster3	Wang Fenglan, Wang Kunlingzi, Xing Chenlin, Li Jianfeng, Yang Feng, Luo Tao, Tao Xiaohua	TCM databases and knowledge discovery
Cluster4	Chen Tao, Su Rina, Xia Yan, Zhang Jing, Yang Xin, Wang Lei	ancient book preservation and knowledge base construction
Cluster5	Tang Shihuan, Chang Mengli, Zhang Fengrong, Zhang Yi, Xu He, Li Yu	TCM Ancient Texts and Data Mining

These clusters reflect a field-wide orientation toward emerging research areas such as digital humanities, knowledge services, and intelligent text mining. Cluster 1, centered on Wang Dongbo and Li Bin, is particularly active in applying AIGC (AI-generated content) techniques to automatic summarization and named entity recognition in ancient texts. Clusters 2 and 5 both emphasize TCM-related studies but differ in geographical affiliation: Cluster 2 is based in Henan universities, while Cluster 5 is rooted in the Institute of Chinese Materia Medica.

Clusters 3 and 4 exhibit a hybrid focus that bridges traditional themes (e.g., preservation, cataloging) and modern technologies. These groups act as intermediaries linking historical scholarship with digital innovation. Overall, the evolving structure of author collaboration networks highlights increasing interdisciplinarity and the influence of geographic proximity, shared research interests, and technological integration.

4.4 Journal Publication Statistics: Dissemination Channels of Ancient Book Digitization Research

The distribution of publications across journals provides insights into the primary academic platforms and disciplinary orientations of a given research field. As illustrated in Figure 3 (in the original), research on ancient book digitization in China is predominantly published in journals related to library and information science (LIS) and publishing studies.

The predominance of LIS journals is attributable to the institutional affiliation of many ancient book resources—libraries and museums—and the involvement of library professionals in digitization initiatives. As a result, provincial and university libraries have become the principal research entities in this domain. Journals such as *Library and Information Service* and *Library Work and Study* are among the most prolific in terms of article volume.

On the other hand, the publication and dissemination of digitized ancient books has attracted attention from publishing scholars, particularly in relation to new publishing models under digital media conditions. Key publishing-related journals, such as *View on Publishing*, have also contributed significantly to the literature.

In summary, ancient book digitization research is primarily disseminated through LIS and publishing journals, reflecting the interdisciplinary nature of the field, which lies at the intersection of cultural heritage preservation, information science, and digital publishing.

4.5 Keyword Co-occurrence Analysis: Research Hotspots in Ancient Book Digitization

Keywords serve as concise summaries of a paper's thematic focus. Betweenness centrality was calculated to measure the influence of each keyword within the co-occurrence network. Keywords with centrality values above 0.1 were identified as critical nodes. As shown in Table 1 (original document), 'ancient book collation', 'digital humanities', 'database', and 'ancient book preservation' all exceeded this threshold, indicating their pivotal roles in the knowledge structure of the field.

The term 'digital humanities' emerged around 2013 and has since become a key concept in ancient book digitization studies. This interdisciplinary approach, also referred to as computational humanities or humanistic computation, involves applying digital tools to the analysis of cultural and historical texts. Early studies, such as those by Fan Jia, discussed the potential of text mining, visualization, and GIS in ancient book digitization, laying the groundwork for a paradigm shift toward data-driven cultural scholarship.

More recent high-centrality keywords include 'knowledge graph', 'digital humanities', 'knowledge base', 'text visualization', and 'artificial intelligence'. These terms signal a transition in the field toward computational frameworks and AI-assisted methods, which are driving the intelligent transformation of ancient book processing and enabling the creative reinvention of traditional Chinese culture in the digital era.

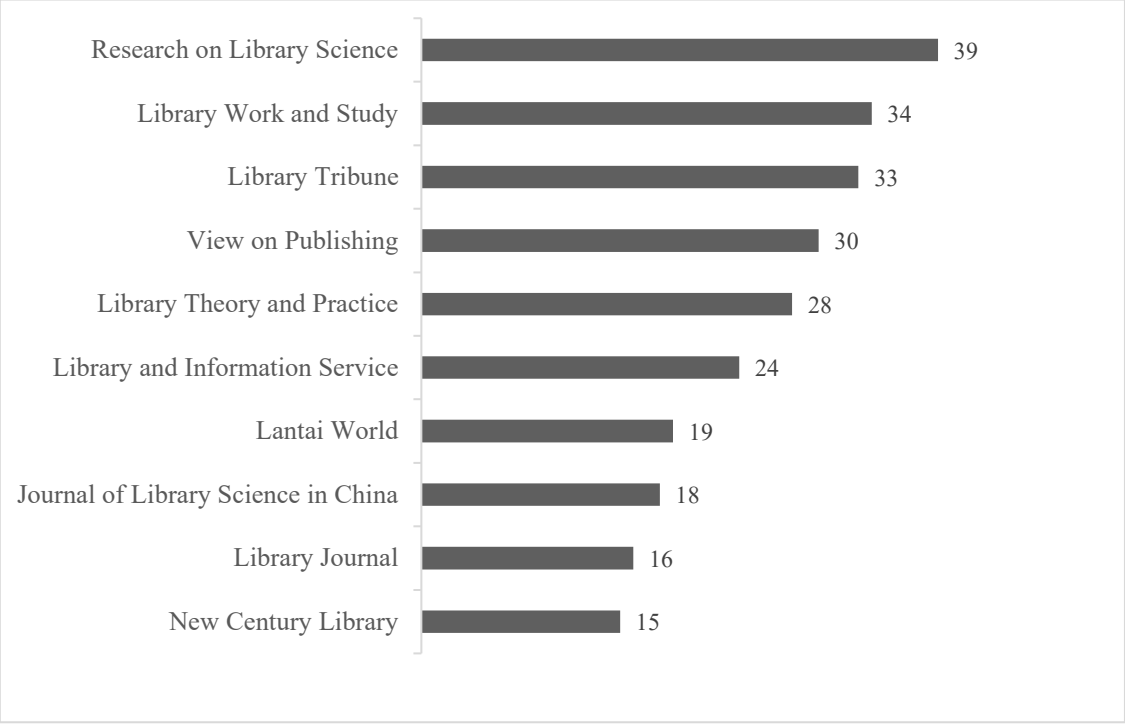

Fig.3. Journal Publication Statistics on Ancient Book Digitization (Top 10)

Table 3. Betweenness Centrality Metrics for High-Frequency Keywords

Frequency	Betweenness Centrality	Earliest Appearance	Keyword
96	0.44	2001	Ancient Books
92	0.39	2008	Digitalization
34	0.27	1998	Ancient Books Arrangement
64	0.25	2013	Digital Humanities
37	0.14	1998	Ancient Books Protection
19	0.10	2000	Database

4.6 Topic Modeling Analysis: Research Fields and Subfields in Ancient Book Digitization

Topic modeling is a widely used method in text mining to identify latent thematic structures within a corpus. In this study, pre-processing steps included word segmentation using Jieba and the removal of stopwords. Additionally, a customized dictionary based on high-frequency keywords identified by CiteSpace was added to Jieba to enhance segmentation accuracy.

Fig.4. Topic Term Scores

The BERTopic model was then used to extract topics from the corpus. The 'gte-base-zh' language model was employed to generate text embeddings. UMAP was applied for dimensionality reduction, followed by clustering using HDBSCAN. Finally, the c-TF-IDF model was used to extract representative terms for each topic.

The resulting topic-word distribution diagram (Figure 4 in the original) illustrates the top keywords associated with each topic. The model automatically grouped the research corpus into eight primary topics, each corresponding to a distinct research focus:

Topic 1: Ancient book databases, resource construction, and digital standards – reflects infrastructure building for digitization.

Topic 2: Digital publishing and knowledge service platforms – emphasizes publishing innovations and service-oriented dissemination.

Topic 3: Traditional Chinese medicine (TCM) texts and pharmacological data – focused on domain-specific digitization in medicine.

Topic 4: Classical philology, textual verification, and knowledge bases – illustrates the integration of traditional scholarship with computational frameworks.

Topic 5: Collation, preservation, and reuse methods – deals with digitization-driven approaches to text editing and conservation.

Topic 6: Deep learning, ancient text mining, and corpus processing – focuses on the use of AI in text analytics.

Topic 7: Cataloging, bibliographic standards, and metadata – addresses digital transformation in cataloging and bibliographic studies.

Topic 8: Digitization and preservation of ethnic minority ancient books – dedicated to inclusive and equitable access to diverse cultural resources.

This classification offers a comprehensive view of the current research landscape and illustrates the field's expansion into both traditional and emerging areas. The inclusion of AI, knowledge graphs, and semantic modeling indicates a shift toward a more intelligent and interconnected approach to ancient book digitization.

4.7 Evolutionary Trend Analysis

This study innovatively incorporates a large language model, DeepSeek-V3, to conduct evolutionary trend analysis, addressing the limitations of traditional bibliometric tools in deep semantic understanding and topic label generation. Using prompt engineering, DeepSeek-V3 was guided to perform automatic keyword labeling, which was then combined with semantic clustering and cosine similarity calculations to construct a topic evolution map for the field of ancient book digitization (see Figure 5 in the original).

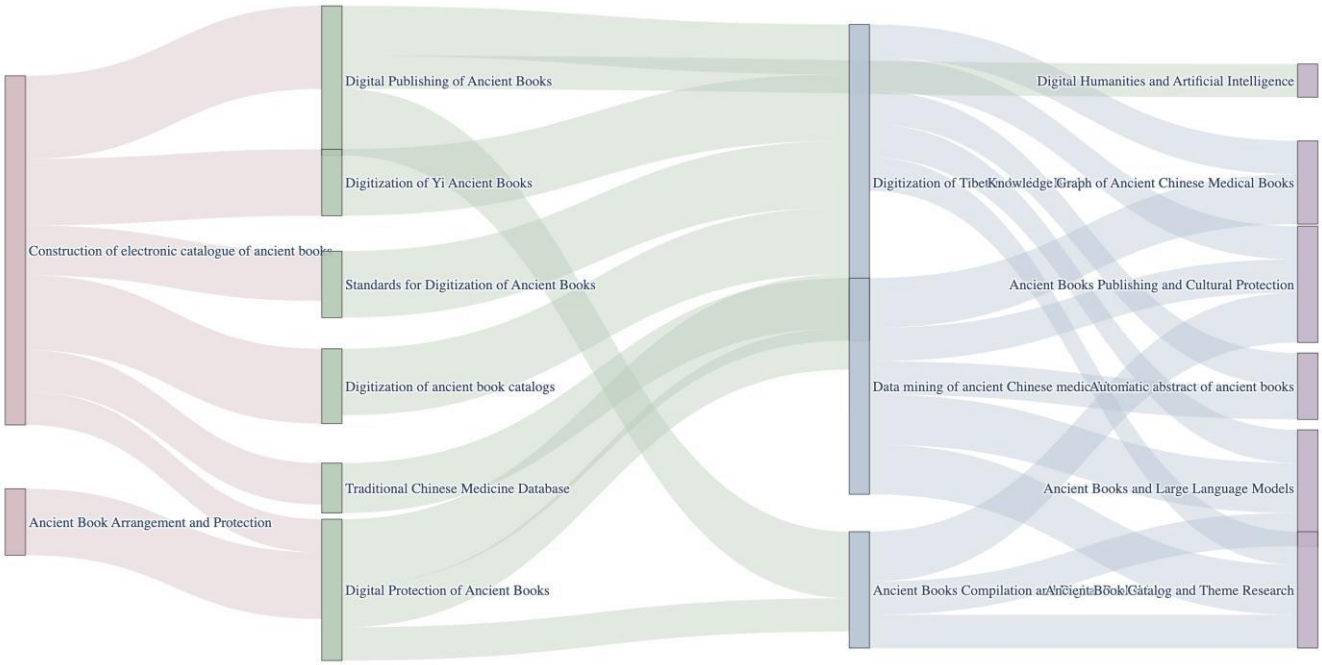

Fig.5. Temporal Evolution of Research Topics in Ancient Book Digitization

The analysis reveals a distinct multi-stage evolutionary trajectory that follows the pattern of 'Infrastructure Building – Technology Integration – Intelligent Applications – Knowledge Innovation.' This trajectory reflects a development paradigm characterized by the synergy between technological advancement and humanistic values.

• Stage 1 – Infrastructure Building (1998–2006):

This initial phase was focused on foundational tasks such as cataloguing, bibliographic registration, and basic protection of ancient texts. Efforts during this period laid the groundwork for subsequent digitization by compiling inventories and developing preservation techniques.

• Stage 2 – Technology Integration (2007–2014):

This stage was marked by the standardization of metadata, enhancement of OCR technology, and the development of structured annotation systems. Digitization projects began to adopt integrated technical frameworks and explored application-specific use cases.

• Stage 3 – Intelligent Applications (2016–2020):

The emergence of artificial intelligence led to the adoption of advanced technologies such as knowledge graph construction and natural language processing. These innovations enabled a paradigm shift in digital humanities, emphasizing data-driven and intelligent research methods.

• Stage 4 – Knowledge Innovation (2021–present):

This most recent phase focuses on deep knowledge extraction and smart publishing. Fueled by large language models and AI, the field has moved beyond technical advancements to emphasize the revitalization, dissemination, and creative reuse of ancient texts. This stage demonstrates a holistic integration of intelligent technologies with the broader goals of cultural preservation and innovation.

This evolutionary pathway demonstrates the continuity and progression of the field, where each stage builds upon the accomplishments of the previous one. The results highlight the increasing role of intelligent systems in enhancing the depth, scale, and impact of ancient book digitization research.

V. RESEARCH SUMMARY

5.1 Summary of the Current Research Status

This study employs cutting-edge artificial intelligence technologies—including the BERTopic deep learning model, the large language model DeepSeek-V3, and the Gephi network analysis tool—along with the CiteSpace bibliometric tool, to conduct a quantitative analysis of China's scholarly research on ancient book digitization over the past three decades. The analysis reveals key information about the field's developmental stages, research themes, and trend evolution.

Ancient book digitization in China began around the 1980s. Based on topic evolution patterns and research data, the study identifies four distinct stages of development:

• Early Stage:

The focus was on building digital catalogs and preserving ancient texts. Major initiatives included the development of bibliographic metadata and cataloging standards to facilitate indexing and access. For instance, CALIS (China Academic Library and Information System) developed an early national catalog system, comprising over 7 million bibliographic entries and 1.75 million authority records.

• Rapid Development Stage:

Starting in 2006, with greater institutional support and technological maturity, the field diversified. Minority language digitization projects (e.g., Tibetan and Mongolian texts) and digital publishing platforms gained traction. Examples include digital guidelines for Tibetan ancient texts and shared-resource models for Mongolian manuscripts.

• Stabilization Stage:

After 2015, the field began to embrace digital humanities. Research shifted from mere digitization to knowledge extraction, structured data generation, and visualization. Technologies such as knowledge graphs, entity recognition, and deep learning models enabled intelligent processing of complex textual materials. In TCM research, text mining was used to identify prescription patterns in classical works. Other projects used digital annotation to build structured historical knowledge bases (e.g., for *Zuo Zhuan*).

• Breakthrough Stage:

Since 2021, the field has seen a wave of innovation spurred by policy support and technical breakthroughs. Major initiatives include the digital reconstruction of *Yongle Dadian* and multimodal interactive publishing projects such as *Shan Hai Jing: Illustrated Guide to Mythical Creatures*, which uses AR, game engines, and voice synthesis. This stage emphasizes immersive engagement, knowledge enrichment, and intelligent user interaction.

Overall, ancient book digitization in China has moved from basic data entry to advanced knowledge discovery. The integration of digital humanities and AI has improved both the depth and scope of research. Going forward, it is crucial to expand public accessibility, encourage digital creativity, and adopt intelligent systems to unlock the cultural value of ancient texts in contemporary society.

5.2 Future Development Trends

The quantitative analysis of the current state of ancient book digitization in China reveals two major developmental directions:

(1) From the perspective of knowledge organization, there is a shift toward knowledge services and AI-driven discovery via large language models;

(2) From the perspective of user experience and learning, gamification is emerging as a key direction.

• Gamification Shift: Spatial Computing Enables Immersive Interaction

In 2023, Apple released a new MR headset, heralding the arrival of the spatial computing era. This technology bridges the physical and digital worlds, creating immersive, multi-sensory environments. In the context of ancient book digitization, spatial computing enables the creation of 3D models of texts, enhancing preservation, accessibility, and user interaction.

Using technologies such as optical scanning and image processing, ancient books can be digitally reconstructed in high fidelity, simulating tactile experiences and enabling VR-based reading and exploration. Additionally, spatial analytics and visualization techniques can reveal structural patterns within ancient texts. This immersive approach supports both scholarly inquiry and public engagement, offering new possibilities for interactive and educational applications.

• Large Language Model Shift: AI Chatbots Empower Intelligent Development

The emergence of ChatGPT in late 2022 sparked a global AI revolution. In ancient book digitization, large language models (LLMs) such as OpenAI's GPT and China's Xunzi series have demonstrated strong potential. These models can perform intelligent Q&A, contextual summarization, semantic understanding, translation, and annotation.

For instance, in 2023, the major project 'Cross-Language Knowledge Base Construction of Chinese Classical Texts' released the Xunzi LLM, specifically trained on ancient Chinese texts. This model addresses limitations in earlier tools and significantly improves interaction quality, accuracy, and accessibility. By providing a user-friendly conversational interface, AI chatbots help both experts and non-specialists access and interpret classical content.

• Knowledge Service Shift: Semantic Understanding and Knowledge Graph Construction

Knowledge services aim to deliver user-centered, context-aware support by extracting, organizing, and reassembling knowledge from large information pools. Technologies like semantic parsing and knowledge graph construction are increasingly applied in ancient book digitization.

Semantic understanding involves identifying key entities, relationships, and events in textual data. Knowledge graphs structure these elements into networks of interconnected information, enabling complex querying, logical inference, and smart visualization.

For example, in traditional Chinese medicine (TCM), researchers have used knowledge graphs to represent

treatment patterns and drug interactions derived from classical texts such as *Huangdi Neijing* and *Treatise on Cold Damage*. These applications demonstrate how intelligent technologies can transform static cultural content into dynamic knowledge systems, facilitating research, teaching, and innovation.

VI. RESEARCH LIMITATIONS AND FUTURE IMPROVEMENTS

This study relies primarily on CNKI journal data from the past 30 years, which does not fully capture the diversity of available information sources. Future research should consider integrating a broader range of data, including social media articles, newspaper reports, and broadcast content, to support more comprehensive topic modeling and public discourse analysis.

Additionally, although the use of large language models (LLMs) such as DeepSeek-V3 improves topic label generation, the accuracy of extracted labels still relies partly on human judgment. To address this, future studies could implement ensemble approaches, using multiple LLMs simultaneously to extract topic descriptors. By comparing and aggregating results across models, researchers may improve the reliability and representativeness of generated labels.

Furthermore, it is advisable to develop a standardized evaluation framework to objectively assess and compare the performance of different LLMs in topic labeling and semantic interpretation tasks. Such a framework would support more rigorous methodological design and reproducibility in digital humanities research.

By addressing these limitations, future studies on ancient book digitization can achieve greater methodological innovation and analytical precision. This, in turn, will contribute to the advancement of both theoretical frameworks and practical applications in the field.

REFERENCES

[1] Zhu, Ying. 2022. 'Opinions of the CPC Central Committee and the State Council on Promoting Ancient Book Work in the New Era.' The State Council of the People's Republic of China. https://www.gov.cn/zhengce/2022-04/11/content_5684555.htm.

[2] National Ancient Book Work Planning Group. 2022. 'National Plan for Ancient Book Work (2021–2035).' Journal of the National Library of China 31(6): 12.

[3] Li, Mingjie, Zhang, Xianke, and Chen, Mengshi. 2020. 'A Review of Ancient Book Digitization Research (2009–2019).' Library and Information Work 64(6): 130–137.

[4] Li, Shiyu, Zhang, Xiangxian, Shen, Wang, et al. 2023. 'Analysis of the Current Status and Research Paths of Ancient Book Digitization at Home and Abroad.' Modern Information 43(11): 4–20.

[5] Shen, Wang, Liu, Jiayu, and Li, He. 2022. 'Progress of Traditional Chinese Medicine Ancient Book Digitization in the View of Disciplinary Lifecycle.' Library and Information Work 66(22): 4–15.

[6] Wang, Qiuyun. 2021. 'Current Research and Development Trends of Ancient Book Digitization in China.' Library Research 2021(24): 9–14.

[7] Fan, Guihong, and Zhao, Chunyong. 2020. 'Research Fronts and Evolution Trends of Ancient Book Digitization Based on Knowledge Graphs.' Publishing Perspective 2020(11): 85–87.

[8] Loomis, GitHub Repository. 2023. 'Chinese Stopwords (HIT, Baidu).' https://github.com/loomis3632/stopwords.

[9] Wang, Yang, Xu, Shanshan, and Li, Chang. 2020. 'SVM-Based Model for Chinese Short Text Classification.' Application Research of Computers 37(2): 347–350.

[10] Chen, Yue, Chen, Chaomei, and Liu, Zeyuan, et al. 2015. 'Methodological Functions of CiteSpace.' Studies in Science of Science 33(2): 242–253.

[11] Deng, Jun, Ma, Xiaojun, and Bi, Qiang. 2014. 'Comparative Study of Social Network Tools: Ucinet and Gephi.' Information Theory and Practice 37(8): 133–138.

[12] Nie, Yaqing, Wu, Tingzhang, Wang, Ruoji, et al. 2023. 'Health Informatics Topic Mining Based on BERTopic.' Information Science: 1–25.

[13] Grootendorst, Maarten. 2022. 'BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure.' arXiv. http://arxiv.org/abs/2203.05794.

[14] Xu, Hanqing, and Teng, Guangqing. 2023. 'Opportunities and Challenges: Library Text Mining under AI Environment with BERTopic.' Information Science: 1–20.

[15] Zhu, Meihai, and Shi, Shu. 2023. 'Hotspots and Frontiers in Rural Industry Revitalization Research: A CiteSpace-Based Analysis.' Southern Agricultural Machinery 54(24): 91–95+100.

[16] Fan, Jia. 2013. 'The Connotation of "Digital Humanities" and the Deep Development of Ancient Books.' Library Research 2013(3): 29–32.

[17] Liu, Yishan, Wang, Yulin, and Li, Mingxin. 2017. 'Empirical Analysis on the Applicability of High-Frequency Word Threshold Methods in Word Frequency Analysis.' Digital Library Forum 2017(9): 42–49.

[18] Xu, Lihua. 2011. 'Reflections on the Digitization of Tibetan Ancient Books.' China Tibetology 2011(2): 153–158.

[19] Su, Rina. 2012. 'On the Construction of Mongolian Ancient Book Digitization.' Library and Information Work 2012(S2): 112–114.

[20] Hu, Qian. 2023. 'Content Production and Transition of Ancient Book Knowledge Service Platforms under the Background of Convergent Publishing.' Publishing and Distribution Research 2023(9): 44–49.

[21] Zhang, Tiange, Xin, Yijun, and Huang, Chaoyuan, et al. 2020. 'Medication Pattern Mining in Ancient Febrile Disease Texts.' Chinese Journal of Pharmacology and Clinics 36(2): 36–39.

[22] Li, Bin, Wang, Lu, and Chen, Xiaohe, et al. 2020. 'Text Annotation and Visualization of Ancient Literature from a Digital Humanities Perspective: A Case Study of the Zuo Zhuan Knowledge Base.' Journal of Academic Libraries 38(5): 72–80+90.

[23] Zhan, Xinhui. 2023. 'Imagination of Communication in the Age of Spatial Computing.' Young Journalist 2023(22): 14–16+23.

[24] Hong, Tao, and Chen, Bijia. 2022. 'Models and Scenarios of Knowledge Services in Ancient Book Digital Publishing.' Publishing Perspective 2022(24): 51–56.

[25] Li, Panfei, Zhang, Chuchu, and Li, Haiyan. 2023. 'Empowering the Inheritance and Innovation of TCM Ancient Book Knowledge with Technology.' Journal of Traditional Chinese Medicine 64(15): 1519–1524.