npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01669-z

Multi-dimensional intelligent reorganization and utilization of knowledge in 'Biographies of Chinese Thinkers'

Check for updates

Jiangfeng Liu[1,2], Zhiyuan Liu[1,2], Yu Shen[1,2], Ran Zhang[1,2], Ningyuan Song[1,2], Jialong Liu[1,3] & Lei Pei[1,2] ✉

Biographical texts often fail to fully showcase their rich semantic knowledge due to traditional narrative modes and knowledge gaps between authors and readers. A multidimensional knowledge reorganization framework for biographical texts involves semantic description, fine-grained knowledge extraction, and knowledge reorganization applications. Based on ontology theory, a core conceptual model for biographical texts was established, employing GPT-4 and BERT for entity recognition. Knowledge reorganization strategies were proposed for key application scenarios and validated through case visualizations. A conceptual model for biographical texts was constructed. Significant enhancement of tag corpora was achieved through LLMs and the RoBERTa-BiLSTM-CRF model, achieving optimal fine-tuning in NER. Strategies based on temporal-spatial transformation, social network analysis, and thematic evolution were proposed, culminating in a knowledge graph centered on "Character-Works-Ideas". Based on methods proposed by us, issues in semantic description and knowledge extraction of biographical texts have been effectively resolved, enhancing the application value of biographical resources.

Biography review texts, combining biographical narratives with evaluative commentary, encompass the Biographical master's life, social relationships, and viewpoints, and are authored by seasoned humanities scholars. These texts integrate the richness of biographical historical material with the critical insights of commentary, requiring the author to comprehensively master information about the subject and blend professional knowledge with critical thinking. However, the academic and professional nature of biography review texts often makes it challenging for general readers to grasp their content readily. Due to high knowledge barriers and linear narrative structures, it is difficult for ordinary readers to understand the content of these texts deeply.

As an interdisciplinary research field, digital humanities (DH) merges computer technology with traditional humanities research to offer new methods for optimizing biographical texts' organization and internal knowledge structure[1]. These technologies, through techniques like text quantification and data mining[2], significantly lower the barriers to reading comprehension and enhance the efficiency of knowledge application. They breathe new life into traditional humanities resources and, through knowledge reorganization and visualization technologies, intuitively display the rich semantic information within biographies, thereby enhancing user experience. This study uses the *Series of Biographies of Chinese Thinkers* as an example to construct a core concept model specifically for biographies, refining the type definitions and processing standards for biographical events. By leveraging large language models (LLMs), large models are used to generate sequential annotation corpora, and the study realizes entity recognition and knowledge triple generation for biographies of thinkers. For the main application scenarios of biographical texts, knowledge reorganization schemes are designed, followed by example verification and visual analysis.

Digital Humanities originated from humanities computing and is a new interdisciplinary research field that involves multiple disciplines such as computer science, literature, history, and sociology. It represents the convergence and integration of information technology and the humanities[3]. With technological advancement, its concept has continuously expanded. Besides the disciplinary perspective, DH is also viewed as a general solution, methodology, or a series of applications. For example, Unsworth defines it as a representative practice and modeling approach[4]. In summary, this study defines Digital Humanities as a new discipline that integrates

[1]Data Intelligence and Cross-Innovation Laboratory, Nanjing University, Nanjing, China. [2]School of Information Management, Nanjing University, Nanjing, China. [3]Data Management Innovation Research Centre, Nanjing University, Nanjing, China. ✉e-mail: plei@nju.edu.cn

multidisciplinary theories and technologies, primarily based on computer technology, to acquire, analyze, and apply common knowledge in the humanities through digital computation.

Biographies, as a unique type of biographical text, are considered a highly valuable resource in historical scholarship due to their rich historical documents and ancient texts. It chronicles historical figures and events and encapsulates expert insights, making it a treasure trove for in-depth exploration of historical and academic significance. In the wave of digitalization, the excavation and semantic relationship revelation of ancient resources like critical biographies have become research hotspots. The mutual corroboration of different historical resources enhances the comprehensiveness and diversity of research, highlighting the growing importance of specialized historical resources such as biographies. This makes them a key element in studying and utilizing historical resources. Based on biographical text data, ontology models are constructed[5,6], biographies are analyzed as linked data[7,8], biography entities and relationships between entities are extracted[9], and a biographical knowledge graph is constructed[10]. Entities are then linked to external datasets[11], reorganizing knowledge about the biographies to achieve the goal of knowledge utilization, which is a common practice in biographical text processing. Tamper explored how to use named entity recognition and network analysis for individual and group research in biographies[12]. Zhang et al. proposed a framework for automated biography construction, using commonsense knowledge to enhance entity linking, improving the precision and recall of entity extraction[13].

The intelligent processing of ancient texts forms the basis for research in historical resource excavation. The significant differences in language expression styles between ancient texts and modern writings present a major obstacle to the intelligent processing of ancient texts, reading promotion, and knowledge services. Numerous scholars are researching different levels of knowledge granularity and various knowledge extraction tasks in processing ancient texts, ultimately constructing digital humanities service application platforms[14]. Ancient languages preserve past cultures and histories. However, researching ancient languages is full of challenges, requiring experts to tackle a series of text-based tasks, from deciphering lost languages to restoring damaged inscriptions and identifying the authors of literary works. While technology-aided research has long supported the study of ancient texts, recent advancements in artificial intelligence and machine learning have reshaped the scale and detail of analysis in the humanities[15]. The fundamental operation in digital humanities research is the digitization of ancient books, which primarily includes cataloging via computers, image scanning, and text recognition (ancient book OCR), text correction, format calibration, and so on[16]. The challenge in this area mainly arises from the computer recognition issues caused by the large number of variant characters in ancient texts. After digitization, intelligent processing and semantic content understanding of ancient book texts are necessary, which specifically includes sentence segmentation and punctuation, word segmentation and part-of-speech tagging[17], named entity recognition[18,19], relation extraction[20], sentiment analysis[21], text classification[22], event recognition[23], summarization[24,25], translation[26], knowledge graph construction[27], and knowledge-based question answering. The primary technical approaches can be broadly categorized into rule-based methods, machine learning-based methods (such as CRF and SVM), and deep learning-based methods. Among them, deep learning-based methods can be further divided into traditional deep learning models (including CNN, RNN, LSTM, GRU), pre-trained language models (including BERT, GPT, T5, and their derived models like RoBERT, ALBERT, domain-specific models like SikuBERT, guwenBERT), and large language models (such as ChatGPT, Claude, Gemini, Qwen, and Baichuan). Since the release of ChatGPT by OpenAI at the end of 2023, many scholars have applied large language models to social science research[28] and conducted related evaluations[29]. In the field of digital humanities, scholars have also applied large language models[30] to tasks such as constructing cultural heritage knowledge-based question-answering systems[31], named entity recognition in historical documents[32], and ancient text translation[33].

Data annotation, as a labor-intensive task, is costly, but high-precision annotated datasets are crucial for improving the performance of supervised machine learning models. Large-scale language models, represented by GPT-4, offer new opportunities for the automation of data annotation. Thanks to their simplicity, ease of use, and speed, LLMs are applied in text processing fields, including data annotation and classification, sentiment analysis, and critical discourse analysis[34]. Regarding the entire process of data annotation, existing research includes using LLMs for data annotation, evaluating annotation results from LLMs, and training models using LLM-annotated results to learn text features[35]. In terms of automatic annotation, Ming et al. proposed AutoLabel, an automated text data annotation method combining LLMs and Active Learning[36]. Tang et al. proposed PDFChatAnnotator, a multimodal data annotation tool that involves Human–LLM collaboration[37]. For quality assessment, Li compared the quality differences between crowdsourced annotation and LLM annotation and proposed a Crowd-LLM hybrid label aggregation method, validating its effectiveness[38]. Kim et al. proposed a Human–LLM cooperative annotation method where the explanations generated by LLMs can provide additional information as validator models, helping people better comprehend LLM-generated labels[39]. The task of information extraction aims to extract structured knowledge from natural language to meet specific research needs. In recent years, large-scale language models have shown excellent capabilities in executing and understanding complex text processing tasks[40], and many scholars have done extensive work on information extraction tasks based on LLMs[41], including named entity recognition[42,43], relation extraction, event extraction, and proposing frameworks for general information extraction using LLMs[44,45]. Regarding better utilization of large language models for information extraction, existing research can be broadly divided into three stages in terms of experimental processes: prompt engineering at the input stage, training and fine-tuning of large models in the middle stage, and quality control and evaluation at the output stage. Among these, the training and fine-tuning of large models are costly and not the preferred option for general information extraction using large models. In prompt engineering, ChatExtract, a prompt framework guide, has been proposed for extracting structured information from text using LLMs[46]. This framework mainly includes two stages: the first stage involves simple classification to filter out input texts without the required information, and the second stage involves distinguishing single-valued from multi-valued information, encouraging negative responses, using appropriately redundant prompts, contextual prompts, and strict output formats to facilitate efficient information extraction[46]. The ChatIE framework suggests converting zero-shot learning tasks into a two-stage multi-turn questioning and answering task using LLMs, and experiments have verified its impressive performance on multiple datasets of three tasks: named entity recognition, relationship extraction, and event extraction[47]. Beyond quantitative text analysis, some scholars have focused on using LLMs for text coding to replace traditional human coding for qualitative research[48]. In natural language generation tasks, LLMs exhibit output coherence and naturalness significantly superior to traditional methods, demonstrating the efficiency and versatility of large language models. However, they also have limitations, including high usage costs, strong dependency on training data, and inevitable biases. Future research will focus on improving model efficiency, reducing resource consumption, enhancing multimodal integration capabilities, improving robustness and fairness, and strengthening small-sample learning capabilities in low-resource environments to further expand their application scenarios and social value. In this study, we focus on using LLMs to assist with data annotation and help extract specific entities from text and determine relationships between entities. Drawing on existing research, we propose a prompt framework tailored to this field.

Digital humanities, as an emerging trend in the humanities, have attracted significant attention since their inception, making notable progress in theoretical construction, technological application, and platform development. Historical resources, in particular, have become a popular area for development and utilization. Against this backdrop, the application demands for historical resources in conjunction with digital humanities

technologies are becoming increasingly prominent. Although genres such as chronological biographies, local chronicles, and poetry have already been incorporated into digital humanities practices, there remains ample room for expanding the research on biographies text. This is mainly reflected in: (1) *Data resources*. Existing biographical and critical studies often provide a broad overview while neglecting detailed exploration, leading to an imbalance between "biography" and "critique". They tend to either focus heavily on life narratives or fall into subjective evaluations, lacking objective depth. It is necessary to refine the granularity of data and balance the content of evaluations. (2) *Research methods*. Biographical studies still primarily rely on traditional content analysis methods. There is insufficient application of text processing techniques and a lack of methodological diversity, which limits the depth and breadth of content exploration and affects the popularization and deepening of research. There is a need to enhance the application of digital humanities methods. (3) *Practical application*. The research outcomes on biography evaluations are limited and vary in quality, failing to fully explore the unique characteristics and language styles of the texts. They lack targeted application scenarios. There should be an effort to innovate the application scenarios and methods of organizing content based on the characteristics of the texts, aiming to achieve multidimensional knowledge reorganization.

In summary, the proposed issues to be addressed are as follows: (1) What are the core concepts and attributes contained in biographical knowledge, and what are the relationships between these concepts? (2) How to accurately describe events in biographical texts to achieve standardized and large-scale construction? (3) How can the core concept entities and relationships of biographical texts be obtained conveniently and efficiently? (4) What are the scenarios for the reorganization of knowledge in biographical texts, and what corresponding reorganization schemes exist?

The significance and value of this research lie in integrating multidisciplinary knowledge, starting from fine-grained data to explore units and structures of knowledge, thereby enhancing the theoretical and practical depth of biographical studies. By leveraging ontology theory, named entity recognition, and visualization techniques, the research deepens the exploration of knowledge organization and application in biographical texts, promoting the effective utilization of historical and humanistic resources. The research not only experiments with the extraction and application of biographical resources but also explores the integration of humanistic resources and digital technology, providing methodological support for digital humanities projects and demonstrating the practical application value of knowledge reorganization.

The innovative aspects of the research include: firstly, the construction of core conceptual models and event description schemes for biographical studies to achieve deep semantic descriptions of biographical content, addressing the lack of unified modeling in previous research. Secondly, combining natural language generation and natural language understanding approaches, the study proposes a method based on the LLMs paradigm, utilizing large language models for generating biographical label corpora. By comparing different BERTology models, the research effectively balances cost and accuracy in building high-quality label corpora. Lastly, for different biographical application scenarios, through the reinterpretation of biographical content and the reorganization of knowledge units, the study achieves efficient application of biographical knowledge.

Methods

Research design

This paper aims to construct a multi-dimensional knowledge reorganization framework suitable for biographical texts. By describing, acquiring, and organizing the knowledge within these works, we intend to achieve its extraction and application. The research encompasses the semantic description of biographical texts, fine-grained knowledge extraction, and the application of knowledge reorganization (see research framework in Fig. 1). First, based on the characteristics of the biographical genre, a top-level conceptual model is constructed to clarify core concept attributes and their relationships, alongside the classification and description of biographical events. Second, using the *Series of Biographies of Chinese Thinkers* as a basis, LLMs is utilized to enhance labeled corpora. By comparing the fine-tuning effects of different models, fine-grained recognition of biographical texts and the construction of knowledge triples are achieved. Lastly, tailored knowledge reorganization schemes are designed for different application scenarios, and a website platform is built for case visual analysis.

Ontology

The concept of ontology, originating from the ontology in ancient Greek philosophy, explores the essential characteristics of entities and their interrelationships, serving as the foundation for understanding the world and human existence. In the fields of computer science and artificial intelligence, ontology technology primarily focuses on operational definitions and is widely used for the standardized representation of conceptual models. Ontologies are classified based on the degree of formalization and domain dependence into highly informal, structurally informal, semi-formal, and strictly formal ontologies[49], as well as top-level ontologies, domain ontologies, task ontologies, and application ontologies[50]. The basic components of an ontology include classes, relationships, functions, axioms, and instances[51], formally described through languages such as XML, RDF, and OWL. Construction methods include manual, semi-automatic, and automatic approaches, with manual and semi-automatic methods like METHONTOLOGY[52] and the seven-step method[53] being widely adopted due to their structured and iterative construction processes. Ontology construction is an iterative process, and tools like Protégé effectively support the development and maintenance of ontologies.

Integrating LLMs for textual knowledge extraction

The "White Paper on Artificial Intelligence Generated Content (AIGC)" defines AIGC as a type of content production method and a collection of technologies that encompass multi-modal information such as text, image, and voice. A Prompt in LLMs serves as an instruction that guides the model to generate the desired content. Prompt Engineering involves processing input text information according to specific templates to restructure tasks, thereby fully leveraging the knowledge production capabilities of language models. Together with the traditional "pre-training and fine-tuning" approach, Prompt Engineering constitutes one of the two fundamental paradigms of pre-trained models.

This paper addresses the issue of obtaining a labeled corpus in sequence labeling using prompt engineering methods. By continuously adjusting the AI roles and dialog content, iterating and refining the model's prompts, and controlling the model's input and output, it achieves mass production of the labeled corpus. Given the limitations of generative AI in handling arithmetic induction and symbolic reasoning tasks, resulting in inaccurate entity position information[54], the CoNLL labeling task is split into two steps: "determining entity boundaries and types" and "calculating entity positions and label conversion." A large language model is used to annotate entity boundaries and types in the original text, and a post-processing script is written to convert to BIO format. Using this method, three types of typical large language models are selected, and 1000 sentences of "basic corpus" are used as samples to compare the accuracy and completeness of each model in generating sequence labels. Specifically, the process requires first specifying appropriate prompts to guide the large model in sequence labeling and then constructing a Python program for information extraction and format conversion of the output content.

Prompt composition

(1) Role setting: "Data Annotator";
(2) Task objective: Sequence labeling, Named Entity Recognition (NER);
(3) Task details: The 7 types of entities to be annotated are: Characters' names (PER), Location names (LOC), Organizations (ORG), Time expressions (TIM), Era names (EMP), Official position (OFI), and works (BOK);
(4) Output format: The sequence labeling method is ENAMEX, which uses XML format to directly add entity tags to the given text. Control the model's input and output, where the Input is a piece of raw text with

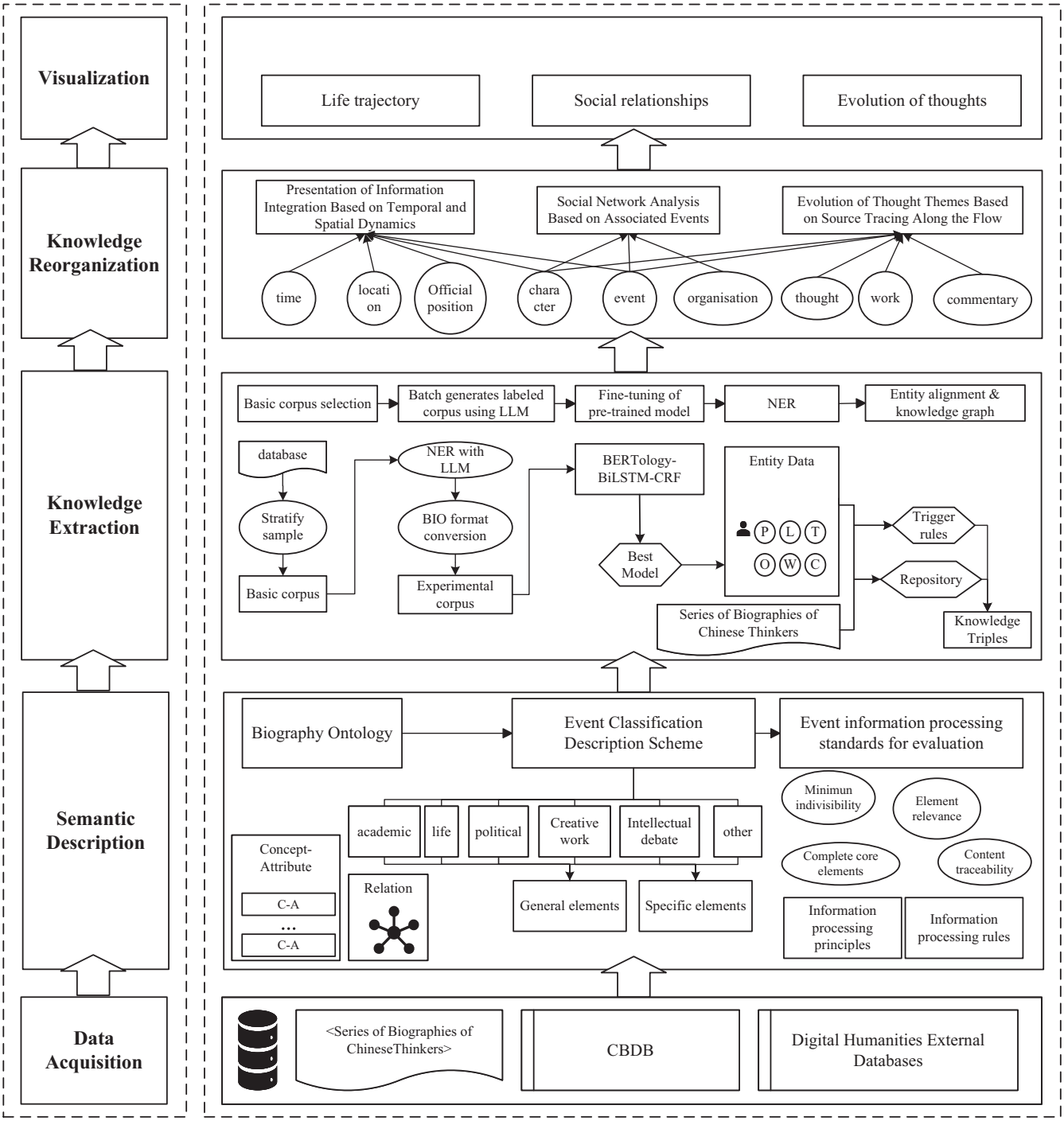

Fig. 1 | Research organization framework. Research framework Figure depicts the research framework of this paper, including data preprocessing, semantic description, knowledge extraction, knowledge reorganization, and visualization.

length n, and the Output is the resulting text after entity annotation. The example is as follows:

Input: $T_1T_2T_3T_4T_5T_6T_7T_8T_9T_{10}\ldots\ldots T_n$

Output: $T_1T_2T_{3<Type\text{-}1>}T_{4<\/Type\text{-}1>}T_5T_{6<Type\text{-}2>}T_{7<\/Type\text{-}2>}T_8T_9T_{10}\ldots\ldots T_n$

(5) Examples: Provide several annotated examples for the model to mimic, enhancing the model's understanding of the task.

(6) Iterative optimization: Improve the instruction based on the sample annotation results.

Post-processing of model output

(1) **Entity attribute extraction**: Extract entity information from the annotated text in the output, including entity name, start position, end position, and entity type.

{text:$T_1T_2T_3T_4\ldots\ldots T_n$***, result:****[{name,start,end,type}, …,{name,start, end,type}]}*

(2) **Format conversion**: Convert the results from the previous step to the CoNLL2002 format, using the "BIO" tagging scheme. Tags are classified as "B–X", "I–X", or "O", where X represents the entity type,

"B–X" indicates that the character belongs to an entity of type X and is the start character of that entity, "I–X" indicates that the character belongs to an entity of type X but is not the start character, and "O" indicates that the character does not belong to any entity.

To improve the accuracy of entity recognition in biographical texts and reduce recycling costs, this study adopts a "pre-training and fine-tuning" paradigm based on natural language understanding (NLU) to construct a biographical entity extraction model, utilizing LLMs-generated corpora for fine-tuning. The BERTology-BiLSTM-CRF (Fig. 2) is used as the baseline model. First, by comparing the semantic representation capabilities of various BERTology pre-trained models, the word vectors of each model are obtained. These word vectors are then input into a bidirectional LSTM layer for feature extraction and semantic encoding. Finally, the CRF layer performs feature decoding to output the predicted sequence labels.

The BERTology series models, including RoBERTa, SikuBERT, etc., are derivatives of the BERT model. They leverage the masked language model (MLM) and bidirectional transformer[55] structure to effectively capture semantic features between words and between sentences. The Transformer employs a self-attention mechanism, calculates word associations through a weight coefficient matrix, and optimizes word vector representations. The bidirectional LSTM (BiLSTM) achieves bidirectional sequence feature extraction by stacking forward and backward LSTM layers, addressing the limitations of unidirectional LSTM. The CRF, a commonly used model for sequence labeling, combines the advantages of the Maximum Entropy Model and the Hidden Markov Model to effectively solve the label bias problem, thereby improving the accuracy of sequence labeling.

The model's performance in single-entity recognition is evaluated using Precision, Recall, and F1-score. The overall performance is assessed using micro average, macro average, and weighted average metrics.

Results

Semantic description scheme of biographies of Chinese thinkers

To achieve multi-source knowledge reorganization of biographical texts, this study designs a core concept model to standardize such texts' semantic descriptions. Biographical texts primarily involve detailed narration and evaluation of events in a character's life, with their core structure comprising an orderly sequence of events centered around the character. Given the significant differences in data sources, resource types, and degrees of structuration of biographical texts, this model aims to unify the semantic descriptions of these texts, thereby supporting the integration and application of knowledge.

The core concepts include characters, events, locations, times, organizations, official positions, works, thoughts, and commentaries (Fig. 3). Specifically, characters are the subjects described, events are records of activities related to these characters, and locations and times provide spatial and temporal context for the events. Organizations and official positions reflect the background of social activities, while works

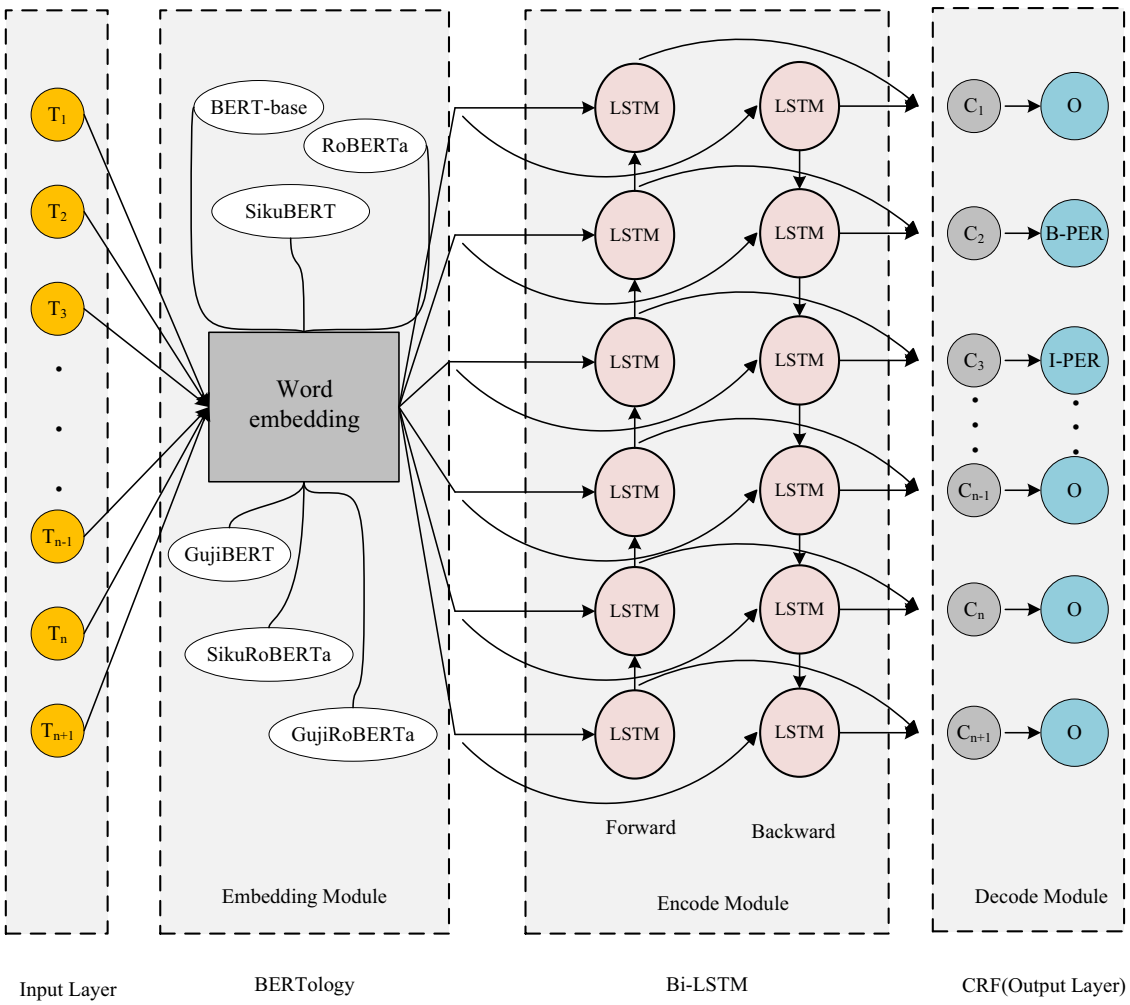

Fig. 2 | BERTology-BiLSTM-CRF. The BERTology-BiLSTM-CRF diagram describes the structure of the model, including the input layer, text vectorization (BERT), coding layer (Bi-LSTM), decoding layer (CRF, output layer).

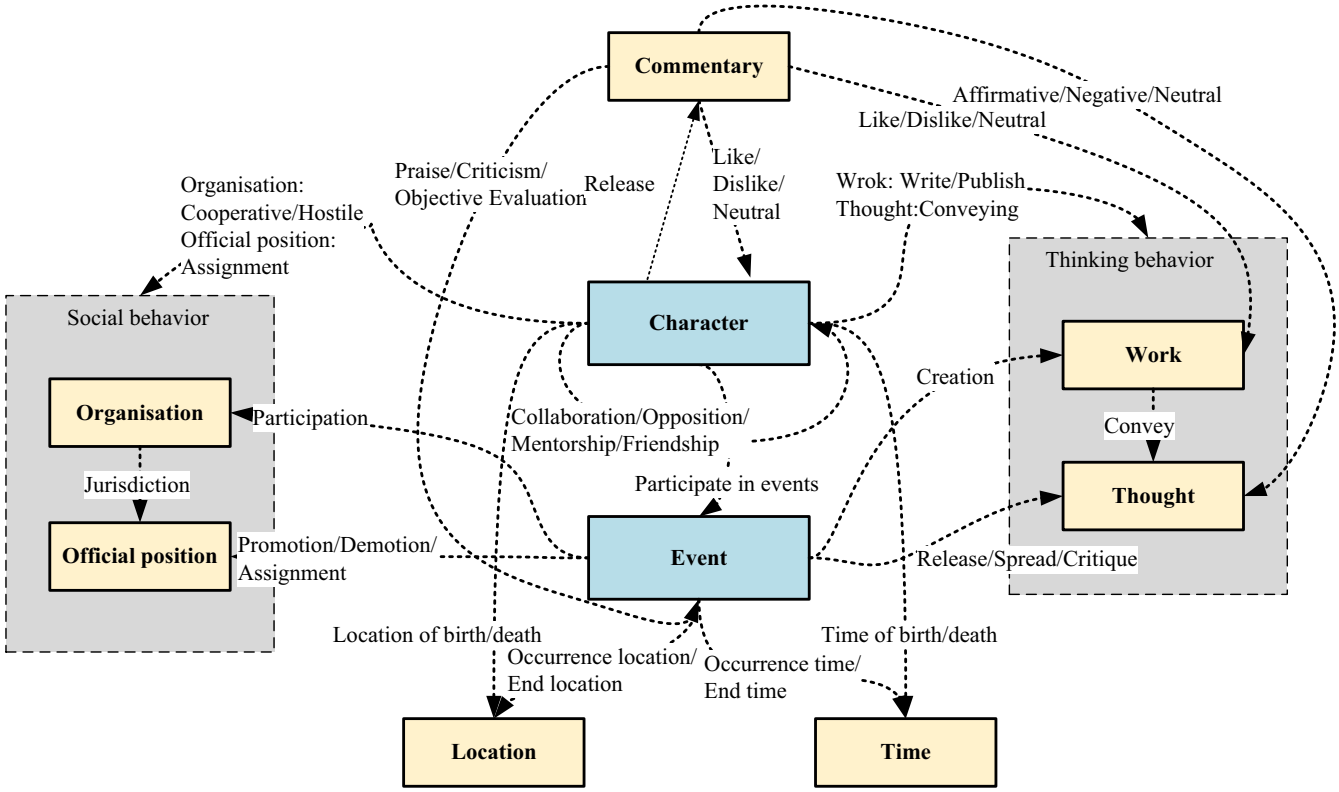

Fig. 3 | Diagram of the core conceptual model of biographical texts. Diagram of the core conceptual model of biographical texts demonstrates the seven core concepts of biographical knowledge organization and their relationship to characters and events.

Table 1 | Key data attributes of the core conceptual entities of the biographical text

Core concepts	Data attributes	Meaning
Character	Name Gender Pseudonym(字号)	Main Subject and Related Figures
Event	Background Content	Historical activities associated with individuals
Location	Ancient-place Modern-place SpatialCoorinate	Including administrative divisions and natural locations
Time	Period Chronology-time Emperor era(年号纪年)	Ancient Calendar Systems and the Gregorian Calendar
Organization	Group -name Group -nature	organizations, including administrative institutions, clans/ political groups, and academic schools of thought
Official position(职官)	Position	Positions within the Bureaucratic System
Work(著作)	Name	Published or unpublished works
Thought(思想)	Category	Thoughts or propositions of a person
Commentary(评论)	Content	Evaluation of the characters, events, and their impact

Table of key data attributes of the core conceptual entities of the biographical text demonstrates the core knowledge attributes of the conceptual entities of the biographical text and their interpretation.

and thoughts exhibit the intellectual products of the characters. Commentaries involve evaluations of the characters and events. Moreover, this model also includes norms for event types, descriptive methods, and information processing.

The model is constructed using the strategy of reusing existing generic ontology models, concerning the ICOM's conceptual reference model (CIDOC-CRM)[49] (V7.2.3) model, which has demonstrated significant capabilities in the fusion, exchange, and integration of heterogeneous cultural heritage information resources. Data consistency and operability are ensured by adapting its generic concepts, attributes, and relational framework. During the attribute definition process, the practicality and accuracy of the model are further improved by drawing on the methodology of Zhang et al.[56] in dealing with Jixia(稷下) character entities (Table 1).

The core conceptual model of the biographical text elucidates nine major core concepts along with their attributes and defines the relationships between these concepts (Table 2). The concept of "Character" is the core of the model, associated with all other concepts, and is primarily depicted through events. The concepts of "organization-official position" and "work-thought" respectively reflect the intrinsic patterns of social and intellectual activities. "Commentary", as a unique concept, refers to the degree of acceptance of events, works, and thoughts.

To accurately represent events in biographical texts, it is essential to design a framework for event types and descriptions within such texts. This framework formalizes the biographical section into an ordered list of events, supported by two dimensions: time and location, to organize the biographical content. Events serve as the primary carriers within biographical accounts, and their description is at the core of knowledge representation. Biographical events involve various entities and naturally exhibit classification characteristics. Six major categories of biographical events are defined: Academic, Political, Life, Creative Work, Intellectual Debate, and Others, which are further subdivided into 20 subcategories (Fig. 4). For example, the "Academic" category includes subcategories such as education and research visits; the "Political" category involves appointments and resignations; the "Life" category describes actions such as residing and visiting friends; the "Creative Work" category records literary creation activities; and the "Intellectual Debate" category includes activities such as debates and other forms of intellectual exchanges.

Table 2 | Core concept relationship for biographical texts

Initial entity	Target entity	Entity relationship type
Character	Character	Co-operation/hostility/mastery/friendship
	Event	Participation events
	Time	Time of birth/death
	Location	Place of birth/death
	Organization	Co-operating/hostile organization
	Official Position	Position held
	Work	Write/publish
	Thought	Communicate
	Commentary	Publish
Event	Time	Start/end time
	Location	Where it happens/ends
	Organization	Participation
	Official Position	Promoted/relegated/assumed
	Work	Creation
	Thought	Publish/disseminate/criticize
Organization	Official Position	Set up
Work	Thought	Communicate
Commentary	Character	love/hate/neutral
	Event	Appreciate/criticism/objective evaluation
	Work	Affirmative/negative/neutral
	Thought	Affirmative/negative/neutral

Table of core concept relationship for biographical texts demonstrates the core concepts and types of relationships in critical texts.

To systematically describe events, we summarize the "general elements" and "specific elements" of events. The general elements.

(Table 3) include character, time, location, type, text, and source, providing a basic framework for the event. The specific elements (Table 4) pertain to particular knowledge elements of subcategories of events. For instance, the academic school in academic events and the titles of official positions in political events are crucial for distinguishing different types of events.

By combining the general elements and specific elements of events, the event description method proposed in this paper can not only generalize the common features of events but also handle special cases in small sample events. This provides structured technical support for identifying and representing anomalous features in rare events, enabling the batch-structured representation of evaluative biographical events, thereby enhancing the in-depth understanding and application of evaluative biographical texts.

In the practical application of biographical texts, the event texts do not necessarily explicitly cover all the aforementioned general elements. There are also irregularities in the specific methods of expression. Potential issues may include, but are not limited to, a single event containing multiple sub-events, vague or missing information regarding time and location, and unclear references to characters. Therefore, to better construct a semantic description scheme for biographical events, this paper defines the principles and rules for processing biographical event information, forming a standardized approach to information handling.

Four fundamental processing principles: minimum indivisibility, complete core elements, element relevance, and content traceability, collectively constitute the framework for handling biographical events.

1. Minimum indivisibility principle: This ensures that each event serves as the smallest atomic unit, characterized by a definite time and location.

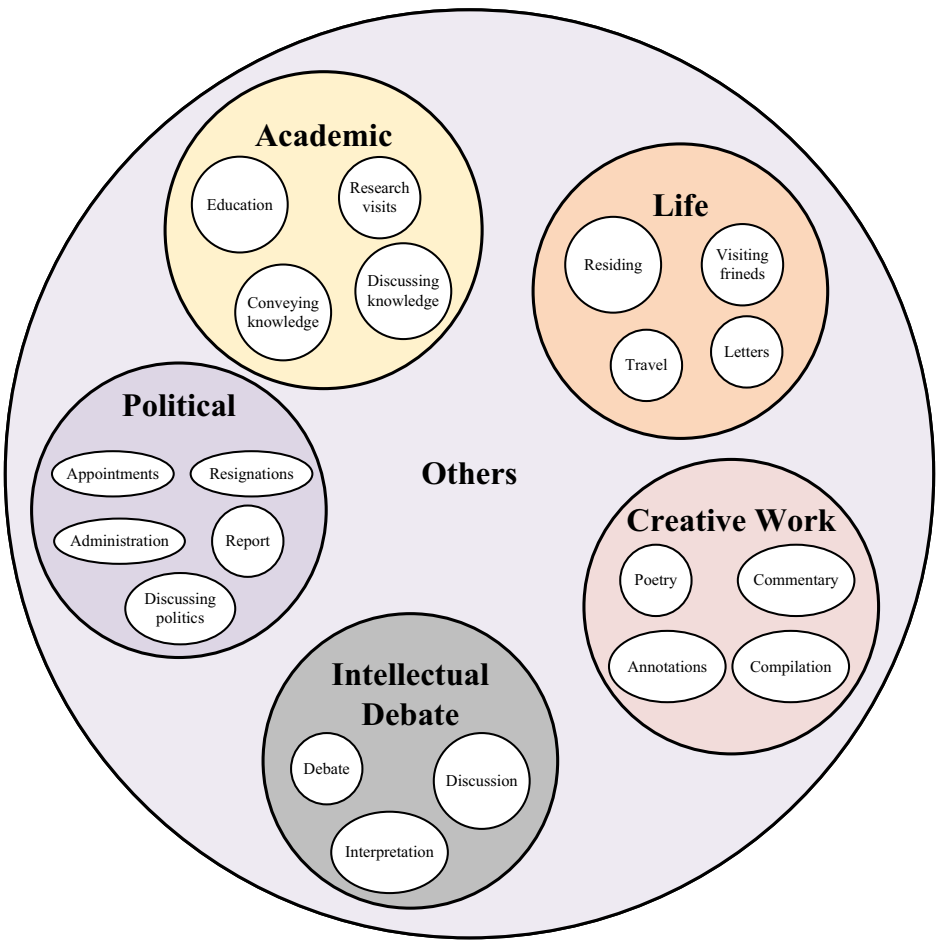

Fig. 4 | Classification system of biographical text events. Classification system of biographical text events demonstrates the specific types of events that biographical events contain and their respective attribute types.

Composite events should be decomposed into multiple sub-events to maintain the independence and completeness of event descriptions.

2. Complete core elements principle: Each event must encompass six core elements: character, time, place, type, text, and source, ensuring comprehensive and structurally complete event descriptions.

3. Element relevance principle: This emphasizes the interconnections between events, constructing an ordered event chain through element relevance, such as the co-occurrence of characters, chronological progression, spatial shifts, or changes in official positions.

4. Content traceability principle: Every event should be verifiable through one or more segments of the biographical text, ensuring the authenticity and traceability of events.

To standardize event descriptions and address issues such as the absence of event subjects and unclear time and location, specific operational methods have been developed to ensure that the handling process is based on reliable references. Table 5 provides common problems and solutions for 9 core entities, while other less frequent issues are to be resolved flexibly according to four main principles.

Knowledge extraction of biographical text integrated with LLMs

To explore the semantic description scheme of biographical evaluations, this study uses the *Series of Biographies of Chinese Thinkers* as the experimental corpus. Initially, a basic corpus is constructed through small-scale text sequence annotation and multiple rounds of verification. Subsequently, prompt engineering and LLMs are employed to compare the performance of different generative AI models. The best-performing model is then selected for large-scale corpus expansion. Using a BERTology-BiLSTM-CRF model, fine-tuning is conducted with the LLMs-generated corpus. The optimal recognition model is determined by comparing the performance of different models. Finally, combining named entity recognition results with text features, a batch construction of biographical knowledge triplets is performed, providing a data foundation for knowledge reorganization.

Given the complex structure and linguistic diversity of biographical texts, this study selects seven types of entities for recognition: characters' names, location names, time, era names, organizations, official positions, and works (Table 6). Biographical texts blend modern and classical Chinese, presenting a variety of linguistic models and expression forms, which manifest in entity recognition as issues such as polysemy and omitted references. The characters' names entity category encompasses various appellations for individuals, the time entity includes different calendrical systems, and the place names entity is influenced by historical changes. Organizations and official positions entities are linked to political activities and serve as key information sources for political event studies; the works entity reflects the thoughts of characters and is of significant value for evaluative research.

The research is divided into two parts: the acquisition of annotated biography entity corpora and the construction of an entity extraction model. First, a large-scale entity-labeled corpus is generated using LLMs, and the CoNLL format is employed to enhance the clarity and generalizability of the data. Second, a secondary pre-training and fine-tuning of the BERTology pre-trained model is conducted to meet the specific needs of biographical texts, deepening the understanding of the language style and knowledge structure of biographies, and supporting research on knowledge reorganization. This entity recognition strategy improves the processing efficiency of biographical texts while taking into account the characteristics and research requirements of biographical texts. It effectively supports the description of the basic attributes of biographical events and the study of characteristic entities, such as dynastic changes, official demotions and migrations, and the dissemination of ideas.

We obtained the photocopied PDF files of the relevant books from the publisher who published the set of books, commissioned a third-party company to carry out OCR recognition processing, and removal of abnormal characters, and then we organized the staff to carry out manual proofreading, and finally obtained the XML format files of the relevant books in the finishing process. In our experiment, we need to convert the XML format to BIO format, see Table 7.

To accommodate the particularities of biographical texts and obtain high-quality labeled corpora, we adopt a "few-shot learning" strategy. This approach involves testing multiple LLMs on a refined basic corpus to select the optimal model for large-scale corpus generation. Key steps include selecting the basic test corpus and choosing the appropriate LLM. The basic corpus is selected from the Series of Biographies of Chinese Thinkers, covering multiple eras and academic fields to ensure representativeness and

Table 3 | General elements for biographical events

Common elements	Meaning
Character	Characters involved
Time	Occurrence/End Time
Location	Place of occurrence/end
Category	Type of event
Text	Text describing the event.
Source	Name of the work contained in the event

The General elements for the biographical events table show the main event elements included in biographical events and their interpretations.

Table 4 | Special elements for biographical events

Special elements	Meaning	Subclasses of involving events
Academic Themes	Academic events involving topics	Studying/visiting/lecturing/lecturing
Academic Works	Academic events involving works	Study/visit/lecture/thesis
Academic Genres	Academic events involving genres	Study/visit/lecture/presentation
Political Themes	Political events involving topics	Governance/discussion/presentation
Name of Official Position	Officials involved in appointments and dismissals	Appointment/removal from Office
Name of Institution	Organizations involved in appointments and dismissals	Appointment/Dismissal
Sonnet Texts	Contents of sonnets	Presentation
Route	Routes of traveling/visiting friends	Visiting friends/traveling
Content of Letters	Contents of letters	Correspondence
Object Books	Names of books compiled or annotated	Notes/compilation
Evaluation Content	Critical texts	Commentary
Ideological Issues	Topics of thought	Debate/conversation/interpretation

The special elements for the biographical events table show the special elements for biographical events with their interpretations and the subordinate events included.

Table 5 | Rules for handling information on biographical events

Entity	Question description	Question example	Prescription
Character	Missing information related to the character (e.g., place of origin, font size, etc.)	In this year, when Huiyuan was five years old, although Fan Xuan's age at that time cannot be determined, he must have been an adult.	Based on the supplementary information from CBDB, if there is none, the relevant attribute information will be left blank.
	Characters not directly related to the subject appear in the description of the event.	"Sun Suifang stated in his book: The situation of the Sun family after they entered Guangdong is more thoroughly and meticulously examined in Professor Luo Xianglin's book 'An Examination of the Origins of the National Father's Family Affairs'(《国父家事源流考》)"	Distinguish between direct and indirect relationships in terms of personal relationships; indirect relationships are not established. In this instance, Luo Xianglin and Sun Yat-sen do not establish a direct relationship.
Time	Missing time-related information	In the sixteenth year of King Cheng of Chu(楚), Duke Huan of Qi(齐桓公) led the vassal lords to invade Chu, reaching Zhaoling. The state of Chu sent Qu Wan(屈完) to negotiate an alliance, and the armies withdrew.	Trace the events in chronological order, identifying the nearest preceding or following events with clear time information. For example, the cessation of hostilities by Duke Huan of Qi follows the "invasion of Chu" event.
	The information about the time is vague, and it is impossible to get the exact time of the event.	In the second year of Zhenghe(征和) (91 BC) in spring, Emperor Che Liu(刘彻) of Han divided the prime minister's office into two departments: left and right.	Use the period from vague temporal expressions as the possible time information for the event and note that it is inferred. For instance, the second year of Zhenghe in spring should be denoted as a period.
	There is a difference in the way of describing the same event between ancient and modern times.	In the fourth year of Yongguang(永光) (40 BC) during summer, on the day of Jiaxu(甲戌) in the sixth month, a fire broke out at the east watchtower of Emperor Xuan's temple(宣帝庙). Four days later, on the last day of the month (Wuyin(戊寅)), a solar eclipse occurred.	Convert the time based on an ancient-to-modern timeline to unify it on the same scale. Unify the chronological years, reign years, and Common Era (CE) years.
Location	Location-related information is missing	After Gaozu(高祖) pacified Chang'an(长安), he acquired old objects from Heng.	By default, use the location of the preceding event. For example, the location of the old instruments in the case is identified as Chang'an.
	The location information is vague, and it is impossible to get the event's exact location.	When Cao Cao(曹操) withdrew his troops from Liucheng, he did not return by the original route but instead took the southern route, moving toward the southwest.	Use the approximate area of a vague location as the possible event location and note accordingly. In this case, "southwest of Liucheng" is a vague location, so the specific cities in the southwest of Liucheng are considered possible ranges.
	Differences in naming styles between ancient and modern locations within the same event.	His original surname was Chen(陈), and his ancestral home was Yingchuan(颍川) (present-day Xuchang, Henan Province).	Translate the location into a standardized name according to the "Ancient and Modern Place Names Comparison Table."(《古今地名对照表》) This case constructs the mapping between Yingchuan and Xuchang, converting as needed.
Event	Multiple times in the event	For example, Ruan Ji (210–263) of the Seven Sages of the Bamboo Grove(竹林七贤).	Split it into two events. For example, the case should be split into a birth event and a death event.
	Multiple consecutive locations within the same event, are difficult to disentangle.	In the third year of Yuanhe(元和), Chong(充) moved his family, was summoned to Yangzhou, governed Danyang, Jiujiang, and Lujiang, and later entered to become Zhizhong.	An event mapping to multiple location elements. The event in the case includes multiple locations that Wang Chong passed through.
Official Position	Official information is partially omitted.	However, Huang Degong(黄得功) stationed troops in Chu (滁)and He(和), stationed in Luzhou(庐州); Liu Zeqing(刘泽清) stationed in Huaiyang(淮阳); Gao Jie(高杰) stationed in Xusu(徐泗); Liu Liangzuo(刘良佐) stationed in Fengshou(凤寿).	Complete based on the context, if undetermined, leave blank.
Commentary	Commentary information is missing from documentary sources	He did not agree with Han Yu's(韩愈) evaluation of Xunzi as being "largely good with minor flaws."(大醇而小疵)	Complete based on external information, if undetermined, leave blank.

Table of rules for handling information on biographical events shows the specific rules to be followed for some of the problems encountered when handling information on biographical events for different entities.

diversity. Through stratified sampling, 1000 pieces of biographical review text are selected and manually annotated using the BIO tagging method (Table 7). In terms of model selection, the tagging effects of three LLMs (gpt-3.5-turbo-0125, gpt4-turbo-preview-0125, and qwen-turbo) are compared.

The *Series of Biographies of Chinese Thinkers* is a collaborative work by multiple scholars, covering over 270 thinkers from the Pre-Qin period to the Republic of China, and spanning disciplines such as literature, history, and philosophy. This series provides ample corpus material for this study. In the selection process, comprehensive coverage of different eras, disciplines, and ideological schools is ensured. The basic corpus undergoes multiple rounds of cross-validation to ensure annotation accuracy, and descriptive statistics of its textual features help establish an initial understanding of the distribution of biographical entities.

After organizing and annotating 1000 sentences from the biography corpus, a "basic corpus" containing 48,421 characters was constructed and subjected to multiple rounds of cross-validation to ensure annotation accuracy. The sentence length of the basic corpus Fig. 5—Left) mainly ranges between 10-60 characters, accounting for more than 50%, with the highest number of sentences falling in the 20-40 character range. The entity distribution (Fig. 5—Right) results show that the most frequent entity is Characters (PER), totaling 1600, followed by Works (BOK) with 512. In contrast, the entities for era names (EMP), Time (TIM), and Official

Table 6 | Entity type definition & example

Entity type	Tag	Example
Characters' names	PER	Special attention should be given to the development of regulated verse (lüshi) by [Cao Pi]. (特别要重视的是[曹丕]对于七律韵诗的发展。)
Location names	LOC	Dou Jiande led his troops south from [Leshou] (present-day [Xian County], [Hebei]). (窦建德从[乐寿] (今[河北][献县]) 引兵南下。)
Time	TIM	In [356 BCE], the State of Qin, under the leadership of Duke Xiao and with the assistance of Shang Yang, began significant reforms. (秦国在[公元前356年], 秦孝公任用商鞅开始进行变法。)
Era names	EMP	In [the sixteenth year of the Zhenguan era] (643 CE), Wei Zheng was already gravely ill. ([贞观十六年] (643), 魏徵已重病在身。)
Organizations	ORG	"The states of [Zhao], [Wei], and [Han] have all perished; they are but old states now." ("[赵]、[魏]、[韩]皆亡矣, 其皆故国矣。")
Official Positions	OFI	There has never been an [imperial censor (formal title of a dynastic official)] as lenient. ([御史大夫]未有及宽者也。)
Works	BOK	It is recorded in the ["Zuo Zhuan, the eighteenth year of Duke Xiang"] that Shi Kuang sang the Southern Wind. ([《左传·襄公十八年》]记载师旷歌南风。)

Table of entity type definition & example shows the core entity types, abbreviation labels, and example sentences.

Table 7 | Examples of entity labeling

Example 1		Example 2		Example 3	
八	B-TIM	而	O	服	B-PER
月	I-TIM	王	B-PER	虔	I-PER
,	O	充	I-PER	,	O
洛	B-LOC	对	O	字	O
州	I-LOC	荀	B-PER	子	B-PER
都	B-OFI	子	I-PER	慎	I-PER
督	I-OFI	的	O	,	O
张	B-PER	《	B-BOK	河	B-LOC
亮	I-PER	天	I-BOK	南	I-LOC
为	O	论	I-BOK	荥	B-LOC
刑	B-ORG	》	I-BOK	阳	I-LOC
部	I-ORG	只	O	人	O
尚	B-OFI	字	O	。	O
书	I-OFI	未	O		
,	O	提	O		
参	O	。	O		
预	O				
朝	O				
政	O				
。	O				

Notes:
1: 八月, 洛州都督张亮为刑部尚书, 参预朝政。
In August(八月), Zhang Liang(张亮), governor(都督) of Luozhou(洛州), became the Minister of Justice(刑部尚书) and participated in the imperial government.
2: 而王充对荀子的《天论》只字未提.
Wang Chong(王充), on the other hand, said nothing about Xunzi's(荀子)'Treatise on Heaven'(天论).
3: 服虔, 字子慎, 河南荥阳人。
Fuqian (服虔), with the character Zishen(子慎), was a native of Xingyang(荥阳), Henan(河南) Province.
The Entity Annotation Example table shows how to annotate entities sequentially using the BIO annotation method.

position (OFI) are relatively sparse, reflecting the specific style and entity sparsity of the text.

The data annotation was conducted using the large model with prompts as shown in Table 8 (Full Version Section in Attachment S1), where the last row labeled "text" is used to call the text to be annotated. Performance comparison of the models (Fig. 6) revealed that gpt4-turbo-preview-0125 outperformed gpt-3.5-turbo-0125 and qwen-turbo in both individual and overall annotation metrics across seven types of entities, demonstrating the best overall performance. Particularly in recognition of character and work entities, the number of annotations by gpt4-turbo-preview-0125 closely matched the actual number, indicating high accuracy and recall rates. In comparison, gpt-3.5-turbo-0125 performed better in annotating character, location, organization, and time entities, while qwen-turbo performed better in era names, official positions, and work entities. For organization, time, era names, and official positions official position entities, gpt4-turbo-preview-0125 exhibited higher recall rates but lower accuracy, resulting in approximately 1.5–2.5 times the correct annotation numbers (Fig. 7).

Comprehensive results show that gpt4-turbo-preview-0125 outperforms the other two models in terms of annotation accuracy and overall metrics for various entities. This model performs particularly well in annotating character (PER) and work (BOK) entities, whereas it exhibits a characteristic of "high recall rate, low accuracy rate" in categories such as organization (ORG), time (TIM), era names (EMP), and official position (OFI). It is important to note that for LLMs, Prompt strategies and experimental corpora may impact the performance of entity annotation. Hence, the aforementioned results reflect the performance under the current experimental settings and do not fully represent the model's general knowledge annotation capabilities. Moreover, the construction and annotation process of the basic corpus strictly adheres to scientific sampling and annotation principles to ensure data representativeness and annotation quality. By analyzing the basic corpus, this study not only improves the processing efficiency of biographical texts but also provides a solid foundation for subsequent knowledge reorganization research.

After comparing the three LLMs, the best-performing gpt4-turbo-preview-0125 was selected as the main model for corpus expansion. Through multiple rounds of manual verification and correction, a total of 92,119 sentences were generated as the "experimental corpus." The static statistical comparison between this experimental corpus and the basic corpus (Fig. 8) shows that the number of entities and sentence volume expanded by approximately 90 times. This proves that LLM technology is an efficient means of obtaining a large amount of high-quality labeled text, significantly improving entity annotation efficiency and enhancing corpus diversity, thus providing solid support for fine-tuning biographical entity recognition models.

The method of generating a labeled corpus by LLM has the advantages of a low threshold and high convenience, as the required labeled corpus can be obtained through simple prompt writing and post-processing. Moreover, it is highly efficient in distributed computing, has strong portability, and is suitable for a wide range of scenarios. However, its complex "black box" nature reduces the interpretability of the model and increases the uncertainty of the output. The cost of use increases with the improvement in model performance, and cost-benefit ratios need to be considered in large-scale applications. Additionally, when using LLM, attention must be paid to data privacy protection and compliance issues to ensure a comprehensive management system.

To evaluate the performance of the entity extraction model, this study employed a stratified sampling method, dividing the filtered "experimental

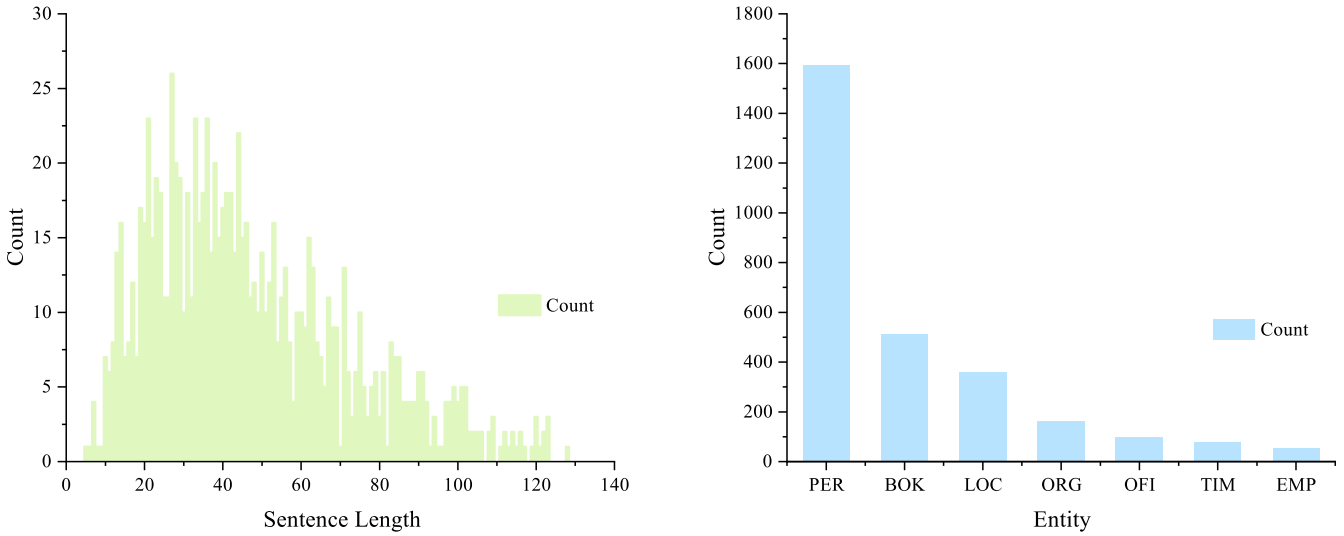

Fig. 5 | Histogram of sentence length and entity distribution of the basic corpus. Histogram of sentence length and entity distribution of the basic corpus is divided into two parts, showing the sentence length and the number of entities of each type in the corpus.

Table 8 | Prompt (reduced version)

Function	Text
"role"	"system",
"content"	"You are a data annotation expert.",
"role"	"user",
"content"	""" Ignore all previous instructions. Please complete the following named entity recognition task. I will input a text of length n characters, and you will return the result in the format of output:dict, which contains the following two items:
	1.text (i.e., input) 2.modified_text (Annotate all entities in the text. For each entity in the text, enclose it in XML format. For example, '苏轼' is a person name entity, so your annotation should look like '<PER > 苏轼</PER > '. You only need to annotate the following 7 entity types: Person Name: PER, Location Name: LOC, Organization Name: ORG, AD Year: TIM, Ancient Chinese Emperor Era Name: EMP, Official Title: OFI, Book Title: BOK.) Here are some examples: ……

Table of prompt shows the prompt used for large language models in this paper for entity recognition.

corpus" into training, testing, and validation sets in a 7:2:1 ratio. Based on the BERTology-BiLSTM-CRF architecture, we fine-tuned six pre-trained models and selected the optimal model based on the performance on the validation set and the change in the composite loss function on the test set. The specific results are shown in Fig. 9.

On the test set, the RoBERTa-BiLSTM-CRF model exhibited the best overall performance (F1-score: 89.15%, Precision: 89.33%), while the BERT-based BERT-BiLSTM-CRF model demonstrated the highest recall rate (Recall: 90.28%). In contrast, the SikuBERT, SikuRoBERTa, GujiBERT, and GujiRoBERTa models, which were pre-trained on specific corpora, showed average performance in entity recognition on the biographical text, reflecting the influence of the mixed classical and vernacular features of the text on the models.

Further analysis revealed that these pre-trained models exhibited high initial F1-scores (over 80%) during the early iterations, with RoBERTa showing the most prominent initial performance. Through trend analysis over the first 20 iterations (Fig. 10), the iterative efficiency of the RoBERTa-BiLSTM-CRF model was found to be superior to other models, while models represented by SikuRoBERTa and GujiRoBERTa exhibited lower iterative efficiency. This may be due to the mismatch between their pre-trained corpora and the characteristics of the biographical text.

In terms of single entity recognition, the RoBERTa-BiLSTM-CRF model performs the best for character, organization, and official position entities with F1 scores of 91.71%, 69.56%, and 68.43%, respectively. On the other hand, the BERT-BiLSTM-CRF model excels in recognizing location, time, and work entities with F1 scores of 79.25%, 89.16%, and 99.57%, respectively. The GujiRoBERTa-BiLSTM-CRF model shows the best performance in recognizing era names with an F1 score of 91.21%, which may be attributed to the large amount of ancient book era name information included in its pre-training text (Table 9).

There are differences in initial performance and iterative efficiency among the six BERTology-BiLSTM-CRF models. The RoBERTa-BiLSTM-CRF model exhibits the best initial performance, possibly due to its pre-training content being oriented towards traditional Chinese texts, while the SikuRoBERTa model shows the lowest performance. For single entity recognition, RoBERTa-BiLSTM-CRF performs best in recognizing character, organization, and official position entities. The BERT-BiLSTM-CRF excels in recognizing location, time, and work entities, whereas GujiRoBERTa excels in the recognition of era names, which may be related to its pre-training texts. Overall, the RoBERTa-BiLSTM-CRF model demonstrates the best comprehensive entity recognition effect on biographical texts, while models like GujiBERT and SikuBERT perform moderately.

According to the results of the ablation experiments (see Tables 10 and 11), there are significant differences in the performance of different models on the named entity recognition (NER) task. Considering the three metrics: weighted average, micro average, and macro average, GujiRoBERTa performs the best, with F1 scores of 89.35%, 89.14%, and 82.92%,

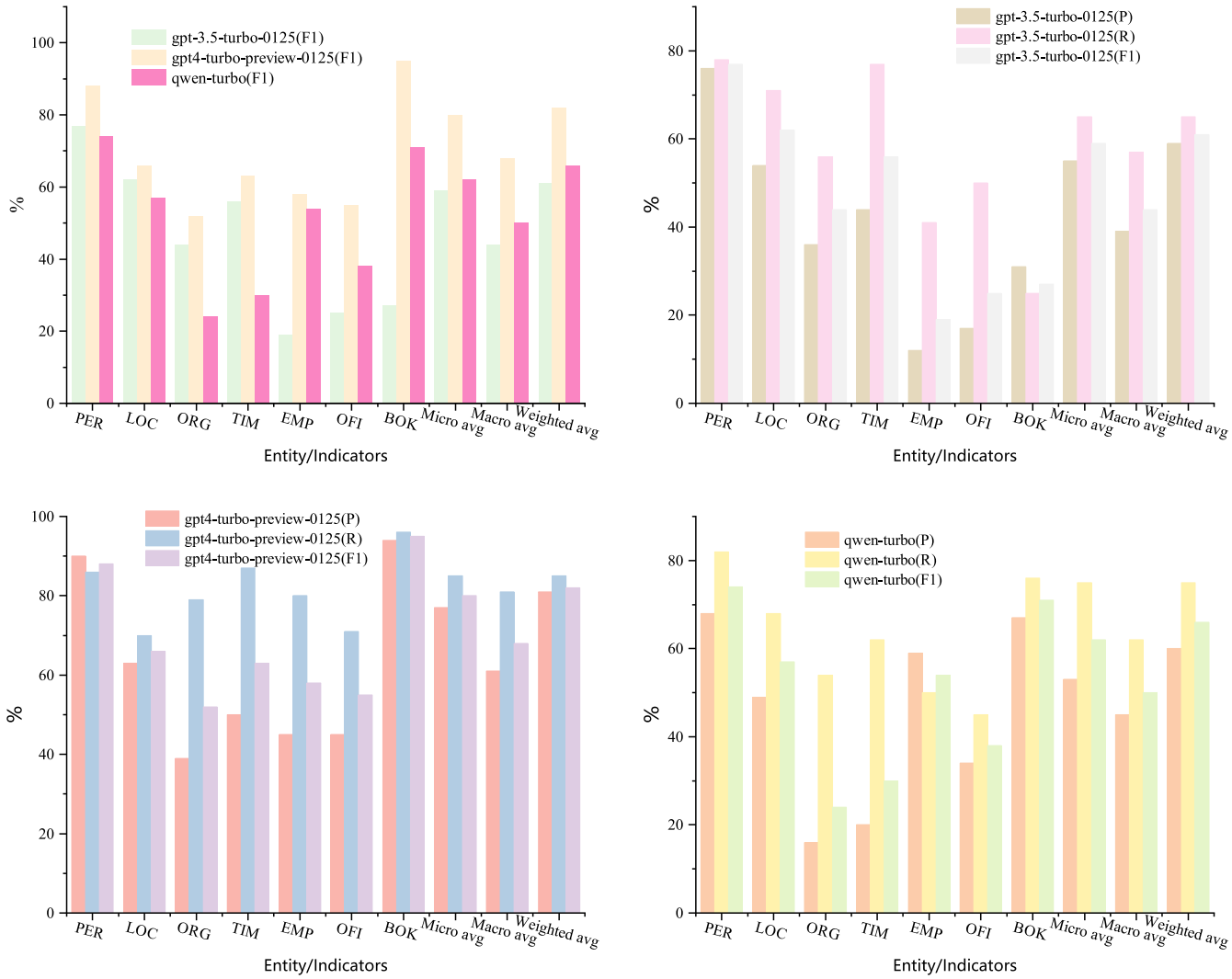

Fig. 6 | LLM's annotation performance on 'basic corpus'. Figure LLM's annotation performance on 'basic corpus' demonstrates the annotation performance of different big language models on the basic corpus set.

respectively, demonstrating excellent adaptability to the majority of entity classes in the NER task. GujiBERT's overall performance is close to that of GujiRoBERTa, but slightly inferior, which may be attributed to the advantage of the RoBERTa architecture in handling contextual information.

Among other models, BERT-base-Chinese matches GujiRoBERTa in the Macro Average metric (F1 = 82.92%), indicating its stable performance in small category entity recognition tasks. However, SiKuBERT and SiKuRoBERTa have lower F1 scores across all metrics compared to other models, possibly due to a mismatch between their pre-training objectives or data distribution and the specific task.

Based on the aforementioned experimental results, the introduction of LSTM-CRF further enhanced the performance of strong foundational models (e.g., GujiRoBERTa). This improvement is particularly notable in the Weighted Average and Micro Average metrics, indicating that LSTM-CRF plays a significant role in enhancing the classification capability of the NER task, especially in modeling complex contexts and improving the precision of boundary decisions. In summary, the architecture combining GujiRoBERTa and LSTM-CRF demonstrates the best overall performance in the NER task, proving its broad applicability in both majority and small category entity recognition.

To address the issue of polysemy (one word with multiple meanings) or synonymy (multiple words with the same meaning) in entity recognition, this study adopts a knowledge-based linking-based entity alignment method. For character entities, this method utilizes external knowledge bases, primarily CBDB, for disambiguation (see Fig. 11). Names with a unique Code are considered unambiguous entities; otherwise, multiple elements, such as the era and place of origin in the context, are matched for alignment. The handling approach for entities like works and official positions is similar.

Entities in biographical texts are closely related to relationships, manifested as entity attributes or relationships, constituting SPO (Subject–Predicate–Object) triples. Considering the complexity of extracting triples directly from texts, a relationship extraction strategy based on trigger rules and a biographical knowledge base is constructed. Trigger rules identify common "entity relationships" or "entity properties" by inducing high-frequency trigger words (such as place of origin and age) in the context of entities, thereby generating knowledge triples. For texts without explicit trigger rules, a relationship extraction scheme based on the biographical knowledge base is established. Through word segmentation and part-of-speech tagging, keywords related to event types are extracted from the biographical texts using c-TF-IDF, constructing a relationship model for relation extraction.

By combining these two methods, this study generates triples from the results obtained by the biographical entity extraction model, forming nine conceptual sets, including character, time, place, era names, official position, organization, work, thought, and commentary, as well as the relationships between these sets. This supports the knowledge reorganization of biographical corpora.

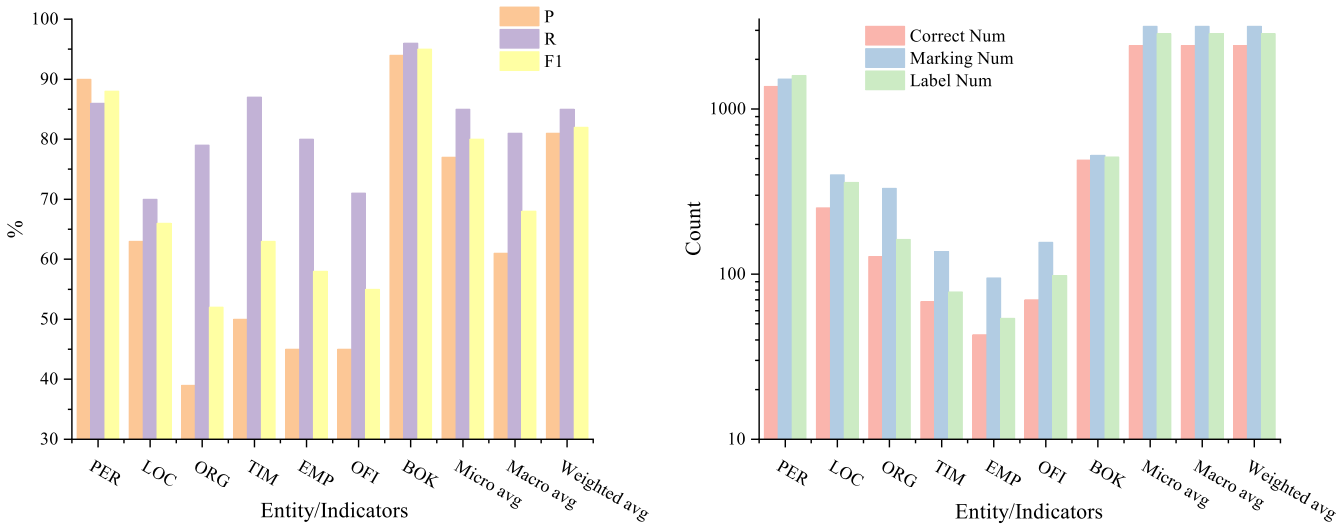

Fig. 7 | Gpt4's annotation performance on 'basic corpus'. Figure Gpt4's annotation performance on 'basic corpus' demonstrates the annotation performance of the GPT4 model on the basic corpus set.

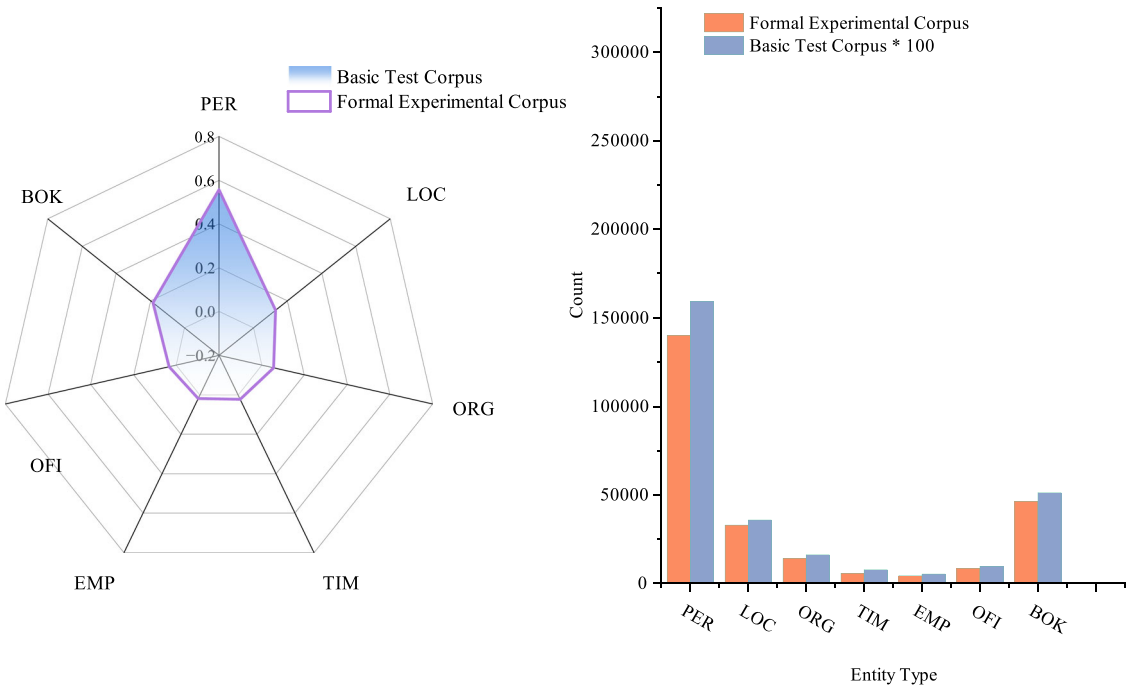

Fig. 8 | Statistics of the formal experimental corpus and the basic test corpus extended by LLMs. Statistical maps of the formal and basic corpora show the distribution of the number of entities in both corpora.

Applications of biographical text knowledge reorganization

This chapter starts from the application scenarios of biographical knowledge reorganization, reconstructs the major tasks involved in the content of biographies based on the needs of biography reading, and verifies reorganization using corpus fragments from the *Series of Biographies of Chinese Thinkers* through semantic description schemes and fine-grained knowledge extraction results. Finally, it constructs the "Semantic Publishing Platform for Chinese Traditional Thought and Culture" to provide visualization and reading services.

The biographical text knowledge reorganization scenarios involve the wisdom crystallized by experts and scholars in the study of historical figures. Because the authors reference a large number of ancient books during the writing process, and many thought processes are not reflected in the text, readers may feel confused while reading. To solve this problem, it is necessary to delve into the application scenarios of biographies, recombine the core knowledge units, lower the reading threshold, and improve readability.

The three main application scenarios that urgently need knowledge reorganization in biographical texts include: reviewing the life trajectory of the biographer, recreating the social relationships of the biographer, and evaluating the evolution of ideological schools.

For the scenario of reviewing the life trajectory of the biographer, an information combination scheme based on temporal and spatial transitions is proposed (Fig. 12). This scheme formalizes the life trajectory of the biographer into an ordered event list centered on the biographer, advancing related events of the character through changes in time and spatial shifts to achieve a multi-dimensional description of their life trajectory. Additionally, to describe the official career of individuals, an official position list mapped with temporal and spatial data is constructed. This scheme overcomes the

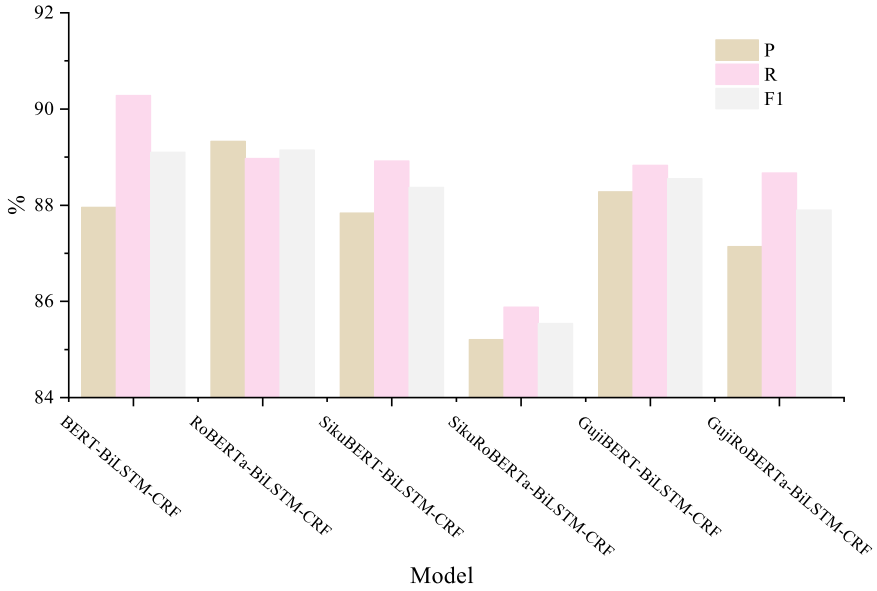

Fig. 9 | Effectiveness of fine-tuning pre-trained models. The pre-trained model effect plot demonstrates the performance of the BERTology class of models on entity recognition.

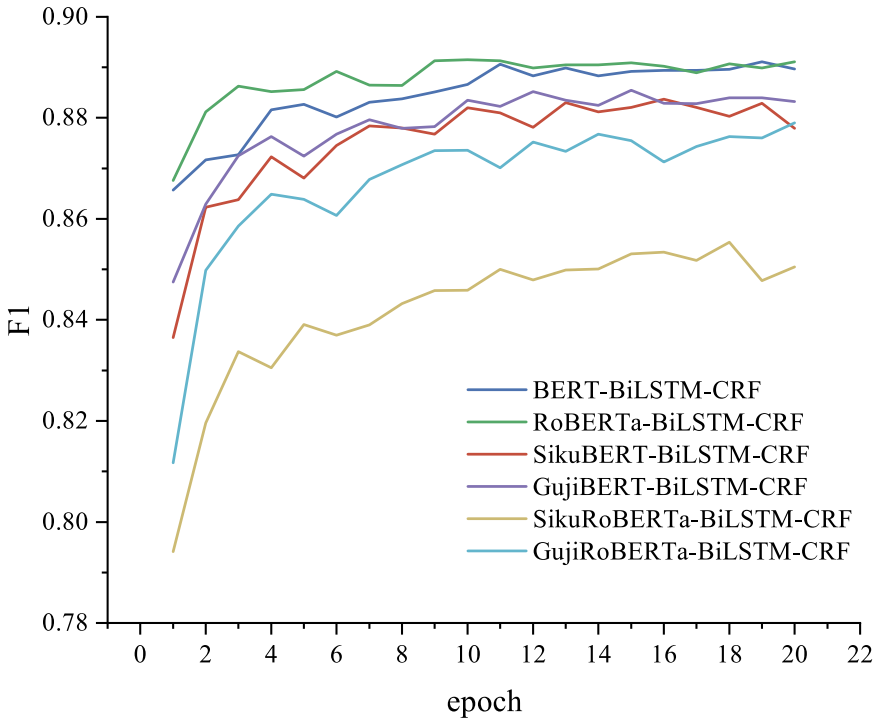

Fig. 10 | F1 score curve of the BERTology-BiLSTM-CRF model concerning the number of iterations. The BERTology-BiLSTM-CRF performance graph shows the entity recognition performance (F1 values) of different BERTlogy-BiLSTM-CRF models when trained for different epochs.

Table 9 | Optimal models for single entity recognition

Entity type	Optimal model	F1 (%)
PER	RoBERTa-BiLSTM-CRF	91.71
LOC	BERT-BiLSTM-CRF	79.25
ORG	RoBERTa-BiLSTM-CRF	69.56
TIM	BERT-BiLSTM-CRF	89.16
EMP	GujiRoBERTa-BiLSTM-CRF	91.21
OFI	RoBERTa-BiLSTM-CRF	68.43
BOK	BERT-BiLSTM-CRF	99.57

Table of optimal models for single entity recognition shows the optimal recognition models and the corresponding performance F1 values for different types of entities.

limitations of traditional biographical texts that rely on a single-time narrative through technologies like GIS (geographic information system) and human-computer interaction, achieving an all-encompassing overview of character events.

For scenarios that reconstruct the social relations of the biography's subject, a social network analysis scheme based on associated events is proposed (Fig. 13). This scheme constructs the social relationships between the biography's subject and others, as well as organizations, using specific events as links, and formally represents these relationships as a multimodal social network. By employing methods such as social network analysis and knowledge graph techniques, this scheme weaves together complex relationships centered around events, thereby enhancing the clarity of relationship definitions and organizational efficiency.

Table 10 | Ablation experiment

Entity	Best BERTology-BiLSTM-CRF	F1(BERTology-BiLSTM-CRF) (%)	Best BERTology	F1(BERTology) (%)
BOK	BERT-BiLSTM-CRF	99.57	RoBERTa	99.45
EMP	GujiRoBERTa-BiLSTM-CRF	91.21	RoBERTa	89.47
LOC	BERT-BiLSTM-CRF	79.25	GujiRoBERTa	78.74
OFI	RoBERTa-BiLSTM-CRF	68.43	GujiRoBERTa	65.63
ORG	RoBERTa-BiLSTM-CRF	69.56	BERT	69.48
PER	RoBERTa-BiLSTM-CRF	91.71	GujiBERT	92.30
TIM	BERT-BiLSTM-CRF	89.16	GujiBERT	88.19

Table of ablation experiment shows how removing part of the model structure affects the entity recognition performance.

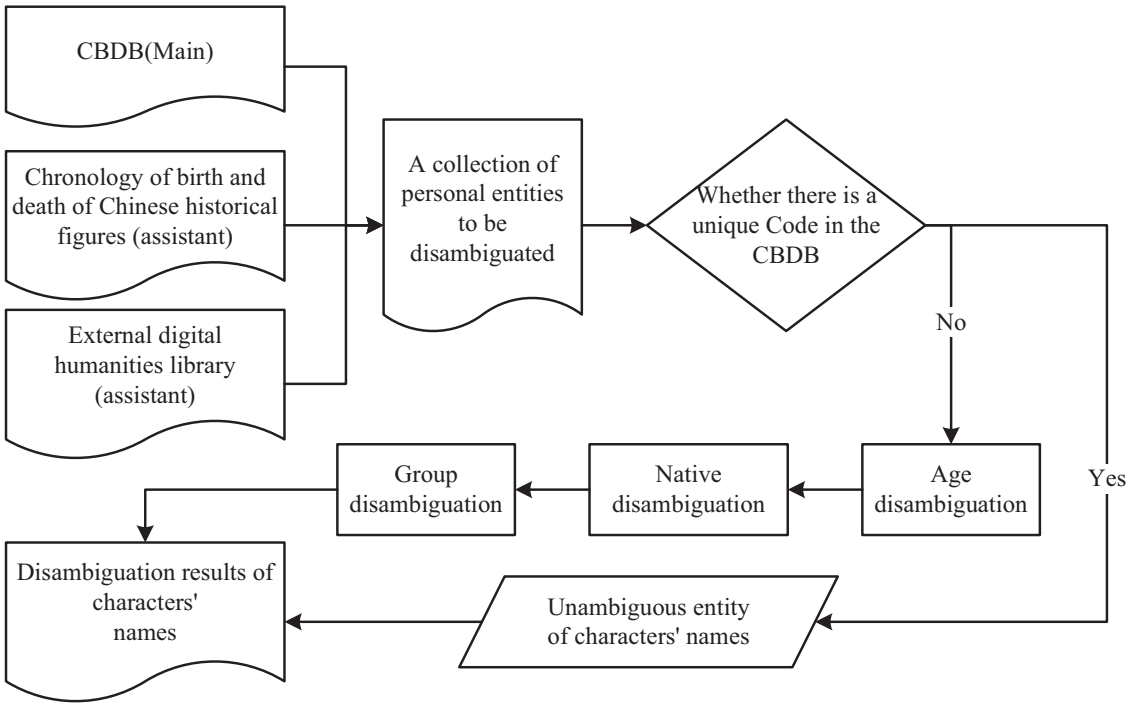

Fig. 11 | Alignment framework for name-entities on biographical texts. The framework diagram for knowledge alignment of commentary text demonstrates the process including the use of external knowledge bases such as CBDB in conjunction with information such as character biographies to help aid in the disambiguation and knowledge alignment of commentary entities.

For the transformation and evaluation scenario of ideological schools, this research proposes a theme evolution method based on source tracing along the flow (Fig. 14). This scheme focuses on the ideological schools formed by the same type of subjects across different biographical texts. By utilizing technologies such as topic modeling and time series analysis, the scheme examines the dynamic changes and historical status of thematic ideas. This research aims to help readers establish a comprehensive understanding and assess the historical significance of the character and their thoughts in the context of their current reading.

The above three reorganization schemes reflect the knowledge reorganization approach of this paper's biographical semantic description in different application scenarios. The first two schemes emphasize the biography ("biography" aspect), characterized by event-centered reorganization, while the third scheme emphasizes evaluation ("evaluation" aspect), with a focus on the core thoughts and commentary. It is hoped that these schemes will enhance the readability and depth of understanding of biographical texts and promote the efficient organization and presentation of biographical content.

Based on the potential application scenarios and related plans of the aforementioned knowledge reorganization, the project team has developed the Semantic Publishing Platform for Traditional Chinese Thought and Culture (V1.0). (As the system is under development and has not yet been released to the public, if you are interested you can send an email to the author asking for the specific URL of the current version.) This platform uses the Series of Biographies of Chinese Thinkers as its data foundation, relying on books, characters, and thoughts to present related historical events and subsequent evaluations, aiming to explore the thought knowledge mining, life records, and posthumous evaluations of thinkers from multiple perspectives.

The book module aims to provide readers with smart reading and semantic knowledge analysis services, comprising four pages, including book, biography, reading, and knowledge entity pages. Clicking a knowledge entity on the right side of the reading page navigates to the knowledge entity page, where one can see in which biographies and pages the knowledge entity appears, and link directly to the corresponding page for reading.

The character module aims to visualize the life of the biographies' subjects through geographic visualization, showing the positions of characters in different historical periods. Clicking different locations reveals the historical location names and displays the historical events at those locations along a timeline, helping readers better understand the life of the figure and get an intuitive understanding of their rise and fall from a geographical perspective.

The thought module extracts the core ideas of each subject and displays them through dynamic word clouds, providing search options by keywords, proposer, discipline attributes, or strokes. Clicking a single thought leads to the thought detail page, which includes the thought's proper name, discipline type, proposer, proposing era, source, background introduction, book index (linking to the book reading page where the thought appears), related recommendations (similar or related concepts), and knowledge graph (showing relationships between thoughts and characters).

The knowledge graph module constructs a knowledge graph around related entities and relationships with characters, works, and thoughts as the mainlines, supporting keyword searches and different layout formats. It allows navigation to related character details and thought details pages. The character details interface includes basic information (proper name, courtesy name, dynasty, birth and death date, birthplace, ideological schools, place of origin, representative works, character's profile), representative thoughts (categorized by discipline), life events (historical events), knowledge graph ("figure-work-thought" knowledge graph), representative works (linking to related works pages), and related recommendations (related biographical books).

Table 11 | Performance of NER using only BERT architecture

Entity	Precision	Recall	F1	
weighted avg	87.66%	91.08%	89.28%	BERT
weighted avg	87.62%	90.67%	89.08%	RoBERTa
weighted avg	87.71%	91.14%	89.34%	GujiBERT
weighted avg	87.68%	91.15%	89.35%	GujiRoBERTa
weighted avg	87.14%	91.01%	88.98%	SikuBERT
weighted avg	87.19%	90.90%	88.95%	SikuRoBERTa
micro avg	87.10%	91.08%	89.05%	BERT
micro avg	87.12%	90.67%	88.86%	RoBERTa
micro avg	87.15%	91.14%	89.10%	GujiBERT
micro avg	87.21%	91.15%	89.14%	GujiRoBERTa
micro avg	86.61%	91.01%	88.75%	SikuBERT
micro avg	86.56%	90.90%	88.68%	SikuRoBERTa
macro avg	79.79%	86.52%	82.92%	BERT
macro avg	80.09%	85.84%	82.79%	RoBERTa
macro avg	79.79%	86.41%	82.87%	GujiBERT
macro avg	80.06%	86.15%	82.92%	GujiRoBERTa
macro avg	79.56%	86.35%	82.73%	SikuBERT
macro avg	79.26%	86.34%	82.54%	SikuRoBERTa

The NER performance table for the BERTology models demonstrates the entity recognition performance of the BERTology models alone after removing the BiLSTM-CRF structure.

Discussion

The biographical texts, as a specialized historical resource, contain rich domain knowledge and expert insights. Due to their high knowledge barriers and complex semantic information, their usage is primarily limited to academic research and has not been widely accessible to the public. To address this issue, this study uses the *Series of Biographies of Chinese Thinkers* as an experimental corpus and constructs a multi-dimensional knowledge reorganization framework encompassing three levels: semantic description, fine-grained knowledge extraction, and knowledge reorganization application.

Firstly, based on ontology theory and the characteristics of biographical commentaries, we constructed an ontology model containing nine core concepts to clarify the main objects of knowledge reorganization. Biographical events were categorized into six major categories and 20 subcategories, and a combined method of general and specific elements was used to solve the description problem. Four processing principles and their specific norms were defined. Secondly, focusing on entity recognition in biographical commentaries, we extracted seven types of entity concepts and constructed relational triples using rule-based and knowledge-based methods, transforming them into structured knowledge. A two-step knowledge extraction scheme combining generative AI and traditional natural language understanding models was proposed to balance the accuracy of entity recognition and the workload of knowledge extraction. Various large language models were compared for prompt annotation effects, and fine-tuning outcomes with BERTology models were evaluated.

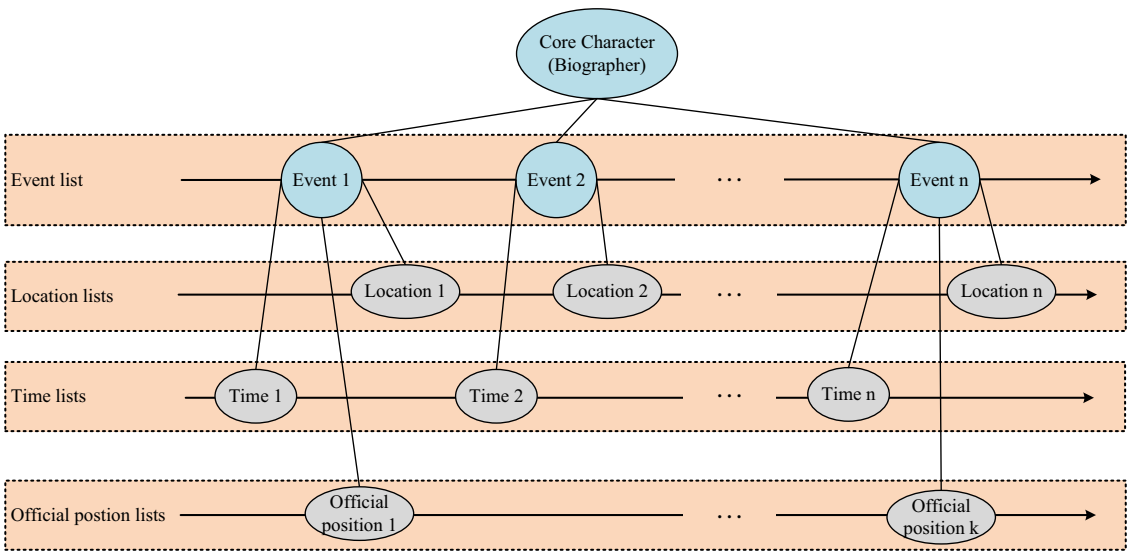

Fig. 12 | Presentation of information integration based on temporal and spatial dynamics. Presentation of information integration based on temporal & spatial dynamics demonstrates a logical framework for information aggregation from a spatio-temporal perspective.

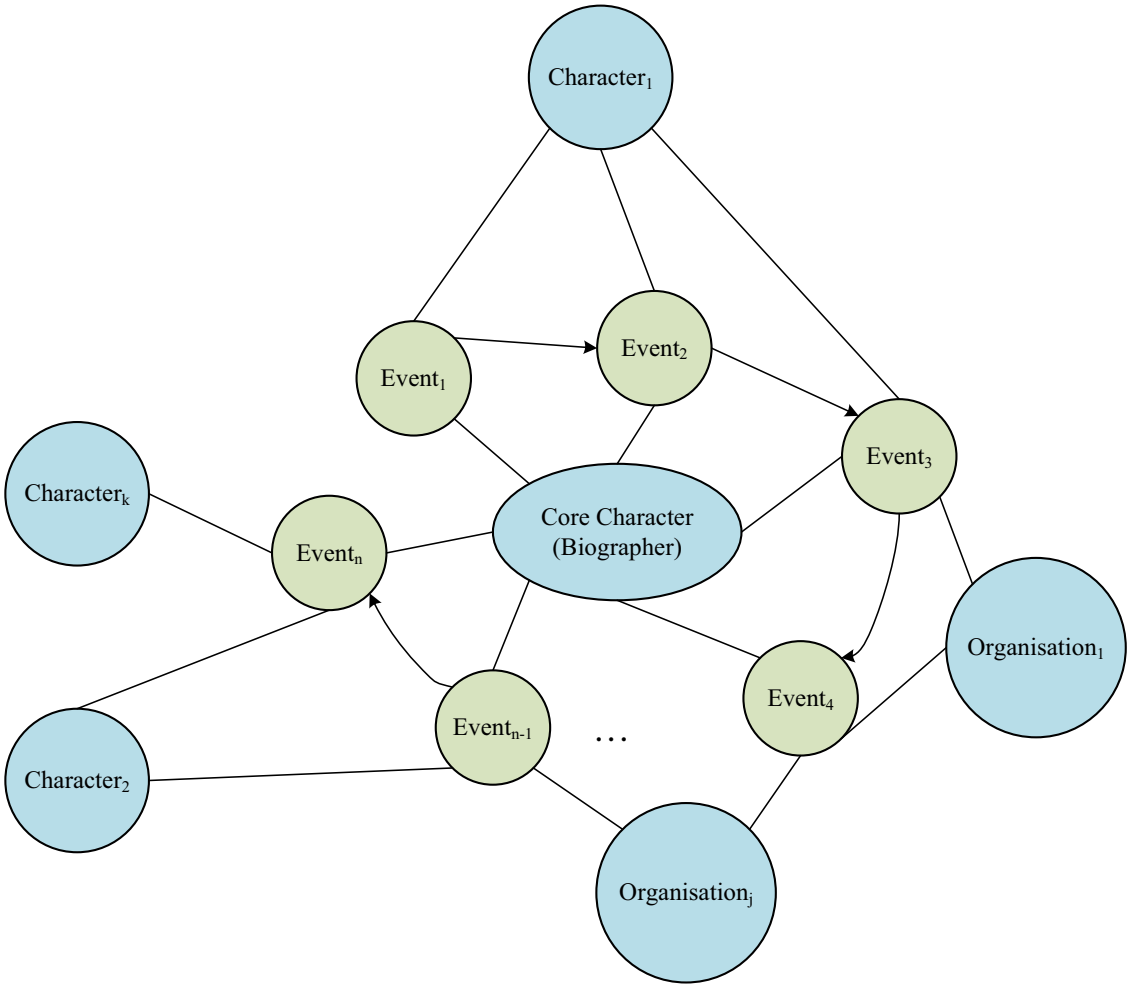

Fig. 13 | Social network analysis based on associated events. Social network analysis based on associated events shows how to perform social network analysis of people based on associations between events.

The optimal model, a combination of gpt4-turbo-preview-0125 and RoBERTa-BiLSTM-CRF, was selected for entity recognition, and triples were generated based on trigger rules and the knowledge base. Finally, specific application requirements were addressed by refining three major application scenarios and constructing corresponding knowledge reorganization schemes, including the presentation of spatiotemporal information, social network analysis, and thematic evolution of thoughts. On this basis, an initial semantic publishing platform for traditional Chinese intellectual culture, centered on character-works-thoughts, was constructed using a knowledge graph as the knowledge organization tool. This platform provides semantic publishing and intelligent reading services, achieving the visualization of thinkers' lives, categorization of thoughts, and knowledge association analysis.

The multi-dimensional knowledge reorganization framework proposed in this paper not only realizes the public value of biographical texts and enhances their visibility among the general public but also serves as an effective attempt at the efficient application of biographical corpus from a digital humanities perspective. It has significant research value and practical implications. The theoretical and practical validity of the multi-dimensional knowledge reorganization framework for biographical texts has been verified. Future considerations include further expanding the range of core concepts covered by the constructed biographical ontology model, delving deeper into the domain-specific applications of large language models in the knowledge extraction process, and extending the range of potential practical fields for biographical knowledge reorganization applications to develop more extensive and systematic applications.

One of the primary limitations is the framework's dependency on the specific structure and content of the experimental corpus. Its adaptability to different genres, cultural contexts, and historical periods remains to be fully tested. Biographies with more narrative styles or those from non-Chinese contexts may present challenges in ontology alignment and semantic representation due to differing conceptual frameworks. Additionally, the use of large language models (LLMs) for entity recognition and relationship extraction, while effective, presents scalability issues. The computational demands and occasional inaccuracies in recognizing low-frequency entities highlight the need for further refinement, especially when applying the framework to larger, more diverse datasets.

To address these challenges, future research could focus on expanding the ontology model to cover a broader range of core concepts and relationships, potentially integrating domain-specific ontologies. Refining the event classification scheme through testing with diverse biographical corpora will also enhance its generalizability. Advancements in domain-specific applications of LLMs are crucial. Developing pre-trained models tailored to biographical texts and employing active learning strategies could improve accuracy and efficiency. Exploring resource-efficient alternatives will make the framework more accessible for smaller research projects. Extending the framework's practical applications is essential for maximizing its societal impact. Potential initiatives include integration into digital archives, museum exhibits, and educational platforms, providing enriched access to historical and cultural knowledge. Collaborations with historians, educators, and technology developers will further refine its usability and relevance.

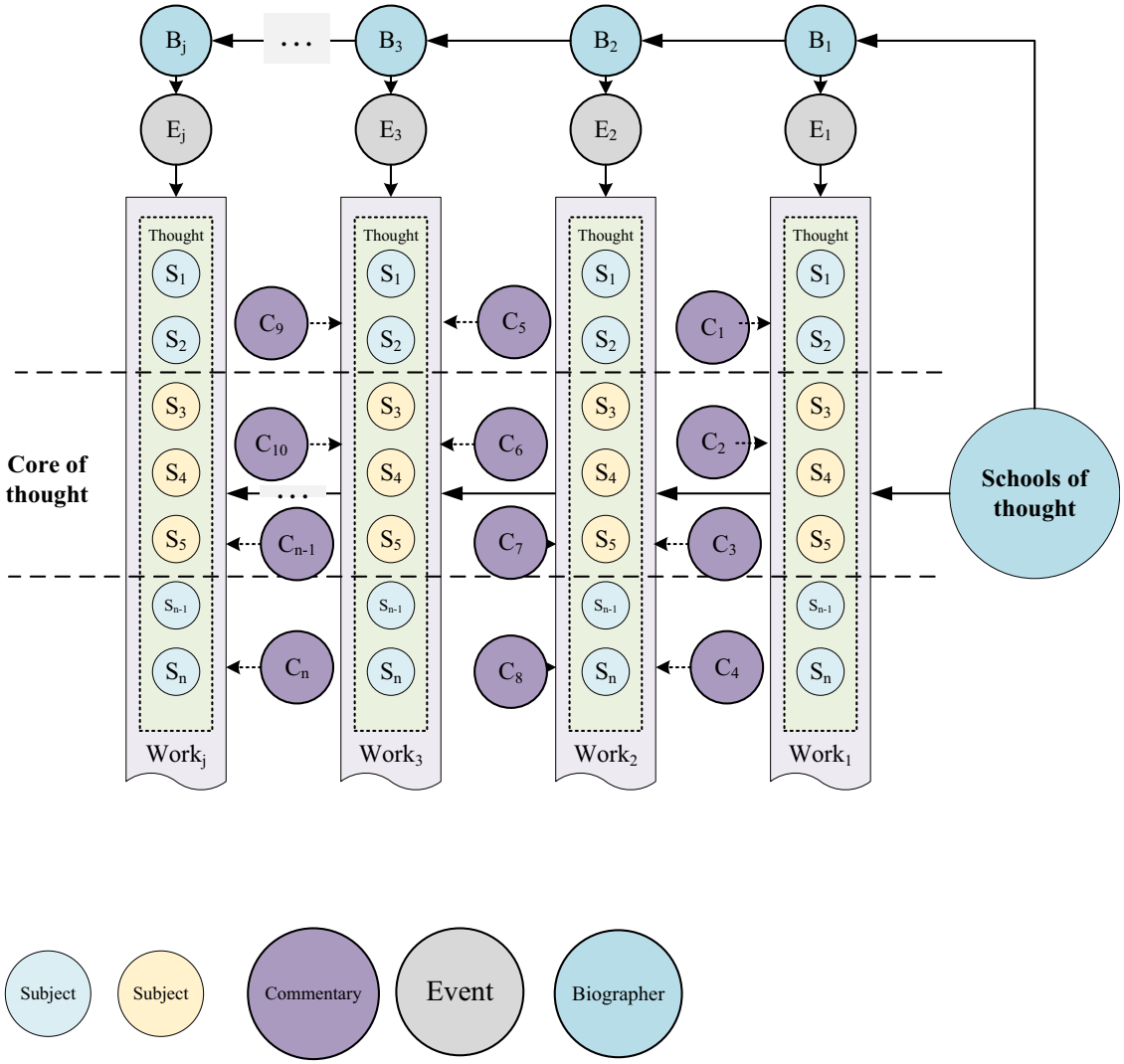

Fig. 14 | Evolution of thought themes based on source tracing along the flow. Evolution of thought themes based on source tracing along the flow demonstrates a logical framework for how to trace the variation of thought themes from the source flow of ideas.

Data availability

Data is provided within the manuscript or Supplementary Information Files.

Received: 3 October 2024; Accepted: 28 March 2025;
Published online: 05 May 2025

References

1. Liu, W. & Ye, Y. Exploring technical system and theoretical structure of digital humanities. *J. Libr. Sci. China* **43**, 32–41 (2017).
2. Burrows S. *Digital Humanities*. (Oxford University Press, Oxford, UK, 2021).
3. Lin, S. W. From humanities computing to digital humanities: the transformation of concepts and research methods. *Libr. Trib.* **39**, 12–20 (2019).
4. Unsworth J. *What is Humanities Computing and What Is Not? Defining Digital Humanities*. (Routledge, London, UK, 2013).
5. Leskinen, P., Tuominen, J., Heino, E. & Hyvönen, E. An ontology and data infrastructure for publishing and using biographical linked data. in *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II): Co-Located with 16th International Semantic Web Conference (ISWC 2017)*. (CEUR, Vienna, Austria, 2017) pp 15–26.
6. Adhi, B. P. & Widodo, W. Development of biography information system based on semantic web using biography ontology: requirement phase. *J. Phys. Conf. Ser.* **1402**, 66063 (2019).
7. Tamper, M., Leskinen, P., Hyvönen, E., Valjus, R. & Keravuori, K. Analyzing biography collections historiographically as Linked Data: Case National Biography of Finland. *Semant. Web* **14**, 385–419 (2023).
8. Tamper, M. et al. editors. *Digital Heritage Progress in Cultural Heritage: Documentation, Preservation, and Protection*. (Cyprus, Nicosia; Springer International Publishing, Cham, 2018) pp 125–137.
9. Plum, A., Ranasinghe, T., Jones, S., Orasan, C. & Mitkov, R. Biographical semi-supervised relation extraction dataset. in: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Association for Computing Machinery, New York, USA, 2022) pp 3121–3130.
10. Rantala, H., Hyvönen, E. & Tuominen, J. Finding and explaining relations in a biographical knowledge graph based on life events: case BiographySampo. in: *CEUR Workshop Proceedings*. (CEUR, Aachen, Germany, 2023).
11. Leskinen, P., Hyvönen, E. & Tuominen, J. Analyzing and visualizing prosopographical linked data based on biographies. *Biographical Data in a Digital World 2017*. (CEUR, Linz, Austria, 2018) pp 39–44.
12. Tamper M., Leskinen P. & Hyvönen E. Visualizing and analyzing networks of named entities in biographical dictionaries for digital

humanities research. in: (ed Gelbukh A.) *Computational Linguistics and Intelligent Text Processing*. (Cham, Budapest, Hungary; Springer Nature, Switzerland, 2023) pp 199–214.
13. Tianlei, Z., Xinyu, Z. & Mu, G. KeEL: knowledge enhanced entity linking in automatic biography construction. *J. China Univ. Posts Telecommun.* **22**, 57–71 (2015).
14. Chen, C.-M. & Chang, C. A Chinese ancient book digital humanities research platform to support digital humanities research. *Electron. Libr.* **37**, 314–336 (2019).
15. Sommerschield, T. et al. Machine learning for ancient languages: a survey. *Comput. Linguist* **49**, 703–747 (2023).
16. Chen, L. Digitalization of Ancient Books and construction of classical knowledge repository from the perspective of digital humanities. *J. Libr. Sci. China* **48**, 36–46 (2022).
17. Chiu, T., Lu, Q., Xu, J., Xiong, D. & Lo, F. PoS Tagging for classical chinese text. in: (eds Lu Q., Gao H. H.). *Chinese Lexical Semantics*. (Springer International Publishing, Virtual Event, Cham, 2015) pp 448–456.
18. Kogkitsidou, E. & Gambette P. Normalisation of 16th and 17th century texts in French and geographical named entity recognition. in: *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*. New York, USA: Association for Computing Machinery. pp 28–34 (2020).
19. Yan, C., Wang, R. & Fang, X. SEN: A subword-based ensemble network for Chinese historical entity extraction. *Nat. Lang. Eng.* **29**, 1043–1065 (2023).
20. Li, B. et al. Few-shot relation extraction on ancient Chinese documents. *Appl. Sci.* **11**, 12060 (2021).
21. Martins, M. D. J. D. & Baumard, N. How to develop reliable instruments to measure the cultural evolution of preferences and feelings in history? *Front. Psychol.* **13**, 786229 (2022).
22. Liang, J. Research on the classification algorithms for the classical poetry artistic conception based on feature clustering methodology. in: *Proceedings of the 2015 International Conference on Electrical, Computer Engineering and Electronics*. (Atlantis Press, Jinan, China, 2015) pp 423–427.
23. Borenstein, N., da Silva Perez, N. & Augenstein, I. Multilingual event extraction from historical newspaper adverts. in: (eds Rogers A., Boyd-Graber J., Okazaki N.). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. (Association for Computational Linguistics, Toronto, Canada, 2023) pp 10304–10325.
24. Wijayanti, R., Khodra, M. L., Surendro, K. & Widyantoro, D. H. Learning bilingual word embedding for automatic text summarization in low resource language. *J. King Saud. Univ.* **35**, 224–235 (2023).
25. Elmenshawy, M. A., Hamza, T. & El-Deeb, R. Automatic Arabic text summarization (AATS): a survey. *J. Intell. Fuzzy Syst.* **43**, 6077–6092 (2022).
26. Lutskiv A. & Lutsyshyn R. Corpus-based translation automation of adaptable corpus translation module. in: (eds Sharonova, N., Lytvyn V., Cherednichenko O., Kupriianov Y., Kanishcheva O., Hamon T., et al.). in: *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021) Volume I: Main Conference*. (CEUR, Lviv, Ukraine, 2021) pp 511–527.
27. Chen, T. et al Text classification for fault knowledge graph construction based on CNN-BiLSTM. in: *2023 35th Chinese Control and Decision Conference (CCDC)*. (IEEE, Yichang, China, 2023) pp 2727–2732.
28. Xu, R. et al. AI for social science and social science of AI: a survey. *Inf. Process. Manag.* **61**, 103665 (2024).
29. Bao, T. & Zhang, C. Z. Performance evaluation of ChatGPT on Chinese Information Extraction—an empirical study bythree typical extraction tasks. *Data Anal. Knowl. Discov* **7**, 1–16 (2023).
30. Riemenschneider, F., & Frank, A. Exploring Large Language Models for Classical Philology. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Vol 1, 5181–15199.(Association for Computational Linguistics., Toronto, Canada, 2023)
31. Xu, L., Lu, L., Liu, M., Song, C. & Wu, L. Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Herit. Sci.* **12**, 118 (2024).
32. González-Gallardo, C.-E. et al. Yes but. *Can ChatGPT Identify Entities in Historical Documents? ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. (ACM/IEEE, Santa Fe, USA, 2023) pp 184–189.
33. Verma, S., Gupta, N., AB, C. & Chauhan, R. A novel framework for ancient text translation using artificial intelligence. *Adv. Distrib. Comput. Artif. Intell. J.* **11**, 411–425 (2022).
34. Törnberg, P. *How to Use Large-language Models for Text Analysis*. (SAGE Publications Ltd., London, UK, 2024).
35. Tan, Z. et al. Large language models for data annotation: a survey. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2402.13446 (2024).
36. Ming, X., Li, S., Li, M., He, L. & Wang, Q. AutoLabel: automated textual data annotation method based on active learning and large language model. in: (eds Cao C., Chen H., Zhao L., Arshad J., Asyhari T., Wang Y.). *Knowledge Science, Engineering and Management*. (Springer Nature, Singapore, 2024) pp 400–411.
37. Tang, Y., Chang, C.-M. & Yang, X. PDFChatAnnotator: a human-LLM collaborative multi-modal data annotation tool for PDF-format catalogs. in: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. (Association for Computing Machinery, New York, USA, 2024) pp 419–430.
38. Li, J. A comparative study on annotation quality of crowdsourcing and LLm via label aggregation. in: *ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, Seoul, Korea, 2024) pp 6525–6529.
39. Wang, X., Kim, H., Rahman, S., Mitra, K. & Miao, Z. Human-LLM collaborative annotation through effective verification of LLM labels. in: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. (Association for Computing Machinery, New York, USA, 2024) pp 1–21.
40. Thelwall, M. ChatGPT for complex text evaluation tasks. *J. Assoc. Inf. Sci. Technol.* https://doi.org/10.1002/asi.24966 (2024).
41. Xu, D. et al. Large language models for generative information extraction: a survey. *Frontiers of Computer Science* **18**, 186357 (2024).
42. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
43. Hu, Y. et al. Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Med. Inform. Assoc.* **31**, 1812–1820 (2024).
44. Wang, X. et al. InstructUIE: multi-task instruction tuning for unified information extraction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2304.08085 (2023).
45. Guo Y., et al. Retrieval-augmented code generation for universal information extraction. in: (eds Wong D. F., Wei Z., Yang M.) *Natural Language Processing and Chinese Computing*. (Springer Nature, Singapore, 2025) pp 30–42.
46. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).
47. Wei, X. et al. ChatIE: zero-shot information extraction via chatting with ChatGPT. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.10205 (2024).
48. Tai, R. H. et al. An examination of the use of large language models to aid analysis of textual data. *Int. J. Qual. Methods* **23**, 16094069241231168 (2024).
49. Moraitou, E., Aliprantis, J., Christodoulou, Y., Teneketzis, A. & Caridakis, G. Semantic bridging of cultural heritage disciplines and tasks. *Heritage* **2**, 611–630 (2019).
50. Guarino N. & Giaretta P. *Ontologies and Knowledge Bases*. (IOS Press, Amsterdam, Netherlands, 1995) pp 25–32.

51. Gómez-Pérez, A. & Benjamins, R. Overview of knowledge sharing and reuse components: ontologies and problem-solving methods. in: *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99) Workshop KRR5: Ontologies and Problem-Solving Methods: Lesson Learned and Future Trends*. (Morgan Kaufmann Publishers Inc., Stockholm, Sweden, 2003).
52. Fernández-López, M., Gómez-Pérez, A. & Juristo Juzgado, N. METHONTOLOGY: from ontological art towards ontological engineering. in: *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series | AAAI-97 Spring Symposium Series | 24–26 March 1997 | Stanford University, EEUU. Stanford University, EEUU: Facultad de Informática (UPM)*. (EEEU, Palo Alto, USA, 1997) pp 33–40.
53. Filho, H. P. P. *Ontology Development 101: A Guide to Creating Your First Ontology*. (Stanford University, Stanford, USA, 2001) pp 1–25.
54. Qian, J., Wang, H., Li, Z., Li, S. & Yan, X. Limitations of language models in arithmetic and symbolic induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Vol. 1, 9285–9298 (Association for Computational Linguistics, Toronto, Canada, 2023)
55. Vaswani, A et al. *Attention is All you Need*. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1706.03762 (2023).
56. Zhang, Y. H., Bai, R. J., Zhang, Y. J., Geng, Z. D. & Wang, Z. M. Knowledge reconstruction and spatiotemporal visualization of Jixia characters from the perspective of digital humanities. *Digital Libr. Forum* **19**, 1–12 (2023).

Acknowledgements

This research was supported by the major project of the National Social Science Foundation of China [24&ZD190], the youth project of the Jiangsu Provincial Social Science Foundation [23TQC006], and the project of the Postgraduate Research & Practice Innovation Program of Jiangsu Province [KYCX24_0111]. We thank Shu Jia from Nanjing University Press for providing data support for this study.

Author contributions

Jiangfeng Liu devised the study plan and led the writing of the article; Z.L. conducted experiments and wrote the article. Y.S. and R.Z. performed data collection and processing. N.S. led the development of the related website platform. Jialong Liu communicated and coordinated the finalization of the internal structure of the platform. L.P. reviewed the articleand supervised the whole process.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s40494-025-01669-z.

Correspondence and requests for materials should be addressed to Lei Pei.