ARTICLE

Check for updates

# Evol project: a comprehensive online platform for quantitative analysis of ancient literature

Jun Wang[1,2,3], Siyu Duan[1,2 ✉], Binghao Fu[1,2], Liangcai Gao[4] & Qi Su[2,3,5]

Quantitative cultural studies have witnessed a surge with the rapid development of computer technology in recent years. Since ancient literature constitutes a long-time-span repository for human culture, with quantitative methods and ancient texts, scholars can study the genesis and progression of human history and society across historical epochs from digital perspectives. Nevertheless, traditional humanities scholars often lack the requisite technical skills, creating a demand for interactive platforms. This paper introduces the Evol platform—an online tool designed for the quantitative analysis of ancient literature. Equipped with various analysis functions and visualization tools, the Evol platform allows users to quantify literary documents through intuitive online interaction. Using this platform, we investigated three cases of cultural evolution in ancient Chinese history: (1) the changing attitude of the government towards nomadic ethnic groups; (2) the formulation and propagation of an allusion phrase related to the Battle of Muye; (3) the influence of the Book of Changes across diverse cultural domains. By showcasing cases across diverse semantic units and topics, Evol demonstrates its potential in providing efficient and low-cost experimental tools catering to the realms of culturomics, history, and philology.

[1] Department of Information Management, Peking University, Beijing, China. [2] Center for Digital Humanities, Peking University, Beijing, China. [3] Institute for Artificial Intelligence, Peking University, Beijing, China. [4] Wangxuan Institute of Computer Science, Peking University, Beijing, China. [5] School of Foreign Languages, Peking University, Beijing, China. ✉email: duansiyu@pku.edu.cn

## Introduction

Quantitative methods are increasingly important in humanities and social sciences studies. In the past decade, text-based quantitative analysis has gained remarkable progress. Scholars have leveraged various computational methods, ranging from traditional statistics (Lansdall-Welfare et al., 2017; Newberry et al., 2017; Alshaabi et al., 2021; Newberry and Plotkin, 2022) to deep neural networks (Garg et al., 2018; Kozlowski et al., 2019; Giulianelli et al., 2020), to investigate sociocultural issues. Since ancient literature constitutes a long-time-span repository for human culture, by employing quantitative methods and ancient texts, scholars can investigate the genesis and progression of human history and society from digital perspectives. However, given that many humanities scholars lack computer skills to process and analyze data from scratch, there is a growing need for tools that enable scholars to conduct quantitative analysis in an interactive and intuitive manner without requiring extensive technical knowledge.

Computer technology and machine intelligence have furnished scholars with potent tools for quantitative studies. Google N-gram Viewer[1] (Michel et al., 2011) is an online computing platform for diachronic n-grams frequency, giving rise to a wave of studies for culturalomics, including linguistic nuances (Perc, 2012), psychology changes (Greenfield, 2013), and conceptual history (Oishi et al., 2013). However, quantitative text analysis can be studied at various semantic units, including words, phrases, sentences, and documents. Since Google N-gram Viewer focuses on the phrase-level analysis function for literature in the last two centuries, it may not adequately address the various analytical needs for diverse semantic units and long-time-span human history. Therefore, there arises a demand for an online platform with comprehensive analysis tools tailored to extensive temporal contexts. Evol, our solution, is designed to meet this need.

Evol[2] is a comprehensive data analysis platform for literary works. A screenshot of this platform is shown in Fig. 1. Its large-scale built-in corpus equipped with diverse analysis functions enables users to explore the cultural phenomena of interest. The Evol platform presents an innovative solution for quantitative cultural analysis that caters to various domains, including culturomics, history, philology, etc. Its efficiency in offering quick-start experiences for quantitative cultural analysis appeals to both novices and enthusiasts, facilitating profound explorations within this scholarly domain.

This paper presents the technical framework and potential application scenarios of the Evol platform. First, we described the processing pipeline for corpus building, including data collection, labelling, and pre-processing, along with the rationale and methodology that underpin these steps. Next, we introduced the functional modules, including the analysis modules for hierarchical text reuse, word co-occurrence, diachronic n-gram, frequency count, browsing, and retrieval. These modules collectively form a multi-perspective and multi-level framework for text-based cultural analysis. Finally, we presented three case studies conducted on the Evol platform at the levels of word, phrase, and document respectively, discussing three cultural evolution issues: (1) the changing attitude of the Chinese government towards seven nomadic ethnic groups in 1500 years; (2) the formulation and propagation of an allusion phrase related to the Battle of Muye; (3) the influence of *the Book of Changes* across diverse cultural domains. These cases demonstrate Evol's potential in quantitative cultural studies. The concluding section discusses the challenges, limitations, and prospects of the Evol project, providing a valuable reference for digital humanities scholars undertaking similar endeavors.

## Methods

In this section, we will introduce the design and implementation of the Evol platform, including its data building and functional modules. It offers valuable experiences and references for academic teams interested in embarking on similar projects. A schematic diagram of the technical framework is shown in Fig. 2.

**Corpus building**. In this section, we will introduce the methods and technologies employed in the corpus building of the Evol project, including data collection, labelling, and pre-processing. We applied some existing toolkits, as well as deep neural network



**Fig. 1 Screenshot of the platform homepage.** It displays the platform's functions (left column), corpus catalog (middle column), and literature information (right column).
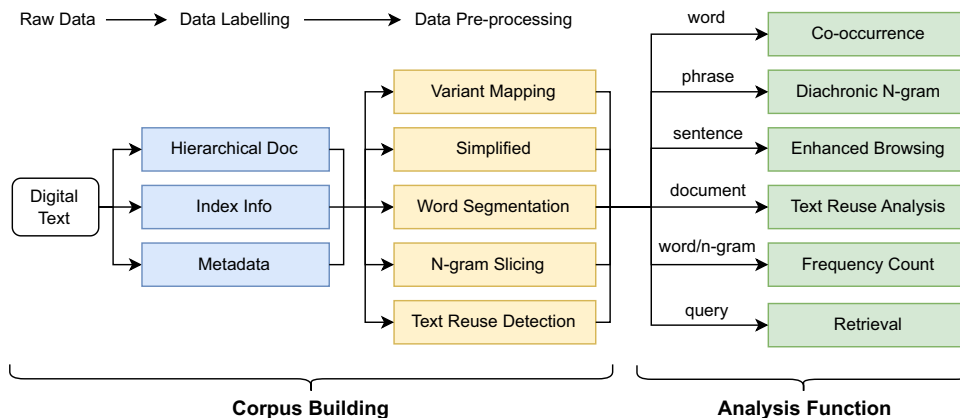
**Fig. 2 Technical framework of the platform.** It illustrates the progression from raw data in digital text form, through labelling (in blue) and pre-processing (in yellow), to the development of various analytical functions (in green).

algorithms to process data. The development of platform functions is based on this information-rich corpus.

*Data collection.* Throughout millennia of human history, the study of ancient cultural phenomena often involves the examination of data spanning several centuries, imposing requirements on the temporal scope and content density of the corpus. The ancient Chinese literary corpus, spanning over two millennia, stands as a voluminous repository encapsulating societal, historical, and cultural aspects. With its extensive chronological breadth, this corpus inherently lends itself to amenable text-based cultural analysis.

The platform incorporates a built-in corpus of ancient Chinese literature spanning over 2000 years, including almost all the classics before the surge in the volume of literature due to the popularity of woodblock printing since the Tang Dynasty (618–907), as well as some selected classics thereafter: 133 types of ancient classics from various fields; Twenty-Four Histories, *Zizhi Tongjian* (资治通鉴), *Continuation of Zizhi Tongjian*, and 15 other history books; The large-scale anthology, *Quan shang gu san dai Qin Han San guo Liu chao wen* (全上古三代秦汉三国六朝文), including over 15,000 articles. These texts encompass diverse cultural facets, including philosophy, history, religion, etc, and the amount is constantly growing. All these digital books are collected from the Internet.

*Data labelling.* Raw digital text is unstructured data that is insufficient for the development of interactive analysis platforms, requiring further labelling. When designing the corpus structure, we considered two factors: each type of data should have a unique identifier, and the storage form needs to be convenient for developers to check and modify. With these two principles, we prepared three kinds of labelled data: document data, index data, and metadata.

Document data: The document data is derived from the digitized books collected. We organized these books in a hierarchical structure based on punctuation and manual processing, starting from the top with books and moving down to chapters (articles), paragraphs, sentences, and clauses. The document data is stored in JSON format, with each level of the hierarchy assigned a unique identifier.

Index data: Relevant background information is crucial for text-based cultural analysis. We compiled each type of background information into a single XLSX file, including the index of time, people, and catalog.

- **Time Index**. Due to the sparse distribution of ancient literature on the timeline, and the need to annotate publication years for a large volume of text, a timestamp based on the Common Era is not suitable for this data. As an alternative, we adopted the dynasty-level timestamp. Certainly, if scholars intend to develop a comparable platform for a more recent collection of texts that is more extensive and includes readily available time data, utilizing Common Era-based timestamps is undeniably the better choice. In our system, the time index table contains the dynasty names in Chinese history, with each dynasty assigned a unique identifier in chronological order. All time-related information on the platform is indexed from this table.
- **People Index**. The people index table contains the historical figures involved in the platform, with each person assigned a unique identifier. When using author and editor information, it will be indexed from this table.
- **Catalog Index**. The catalog index table contains the built-in catalog of the platform. It should be noted that the catalog classification of ancient Chinese literature is a long-term debate (Li et al., 2021). In practice, we observed that the traditional classification is inconvenient for a quantitative analysis platform and is unfriendly for the future expansion to multiple languages. Therefore, considering the analysis functions of the Evol platform, we designed a hierarchical catalog based on the topics of books.

In the Appendices, we provided the temporal and category distribution of the current version of our corpus.

Metadata: We manually labelled the English title, author, editor, publication time, recording time, and catalog of each book. During labelling, we referred to the corresponding index data. We uniformly stored the metadata of all books in an XLSX file, which can be easily checked and modified by the platform administrator.

*Data pre-pre-processing.* To ensure acceptable interaction response on the online platform, we processed the data in advance. The pre-processing is conducted on four levels, i.e., character, word, n-gram, and sentence.

Variant mapping and simplified: Currently, both simplified and traditional characters are used in different regions of China. In addition, the written form of ancient Chinese employs traditional characters, some of which have variants. To accommodate the

linguistic idiosyncrasies of diverse regions and map the variants, we used OpenCC[3] to transform all textual data into a consistent simplified format and store this copy. This approach enables the system to facilitate user queries in both simplified and traditional character forms.

Word segmentation: Chinese is a character-based language, but the smallest semantic unit is the word. Chinese text analysis often needs to be conducted at the word level, making word segmentation a common procedure in text processing. To optimize the online response time, we conducted word segmentation on the corpus and pre-counted the word frequency at the chapter (article) level. The segmentation tool is Jiayan[4]. This data is used in the co-occurrence analysis module and word count module.

N-gram slicing: Performing real-time n-gram slicing for n-gram frequency statistics can be computationally expensive. Therefore, we conducted n-gram slicing (1–4 grams) on the whole corpus in the pre-processing stage and pre-counted the frequency at the chapter (article) level. This data is used in the n-gram count module.

Text reuse detection: The consumption of real-time calculation of text reuse detection at the document level is unacceptable in online services. As a solution, we pre-detected them in the pre-processing stage. We applied the latest approach for text reuse detection in ancient Chinese (Duan et al., 2023), which uses pre-trained deep learning models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019) and contrastive learning (Gao et al., 2021) to get personalized text similarity models without supervision, thereby detecting text reuses within ancient Chinese literature through sentence embedding similarity. Currently, we have identified over 14 million text reuses in this corpus, and this number is continuously increasing. We saved the results as index IDs in the JSON file, corresponding to the document data. The system accesses this data in text reuse browsing and analysis.

**Function design**. There are already many digital collections of ancient Chinese literature publicly available online, such as Daizhige[5], which is composed of ancient text collections; Erudition[6], which features photocopies of some books, and Jihe[7], which also provides access to ancient inscriptions and rubbings. However, most of these platforms only offer functions such as full-text search and online reading. Additionally, there are digital platforms that provide value-added functions such as named entity annotations (Shidianguji[8]), character relationship discovery (CSAB[9]), and text reuse linkages (Sturgeon, 2019) (Ctext[10]). Nevertheless, none is tailored to accommodate the demands of quantitative analysis over large volumes of textual data. The Evol platform, however, emerges as a solution to fulfill this requirement.

The Evol platform offers a range of multi-perspective analysis functions and visualization tools which are equipped with various analysis algorithms specially tailored for cultural analysis purposes. At the word level, the co-occurrence analysis module explores the context of words. At the phrase level, the diachronic n-gram module assesses the usages of specific phrases through frequency changes over time. At the sentence and document level, Evol incorporates distinctive modules for text reuse analysis: millions of text reuse sentences were pre-detected with deep learning models for enhanced browsing, and document-level intertextual connections among literature are hierarchically displayed in the text reuse module. Besides, foundational text analysis functions like text retrieval and frequency count have also been further enhanced in the Evol platform. Its interface

supports both Chinese and English, catering to the needs of users worldwide, without requiring a programming background.

*Hierarchical text reuse for intertextual analysis*. When creating literary works, humans deliberately or inadvertently reuse texts from others, resulting in innate intertextual networks within the literature. These intertextual networks provide traceable evidence for the dissemination and evolutionary trajectory of human ideas. As a manifestation of intertextual relationships, text reuse serves as quantitative evidence for various cultural studies themed on the similarity (Sturgeon, 2018b; Burns et al., 2021), influence (Büchler et al., 2013; Forstall et al., 2014), and evolution (Hartberg and Wilson, 2017; Duan et al., 2023) of literary works. The feasibility of this text analysis approach has been validated across different languages, including Latin (Coffee et al., 2012b), French (Ganascia et al., 2014), English (Smith et al., 2013), and ancient Chinese (Sturgeon, 2018a). The Evol platform employed the text reuse technique to effectively quantify and visually represent the instances of text reuse, thereby facilitating the identification and exploration of potential cultural phenomena.

Real-time detection of text reuse is a time-consuming and resource-demanding task. Therefore, we undertook the task of pre-detecting text reuse within the corpus in advance and incorporated the results into the platform. Unlike platforms that provide online services for text reuse browsing (Sturgeon, 2019) (Ctext) or retrieval (Coffee et al., 2012a) (Tesserae[11]), Evol is equipped with hierarchical and multi-perspective tools. Within this module, users can select a collection of literature based on their interests. The platform generates statistical results and visualizes text reuse relations at the levels of book, chapter, and sentence, respectively. A schematic diagram is shown in Fig. 3.

At the book level, the platform displays selected books as an interactive intertextual network. The nodes of the network are individual books, while the edges represent reused sentences between two books. The total count of text reuse for each book in the selected collection is displayed above. By clicking the corresponding edges, users can explore reused sentences between any two books.

At the chapter level, the platform displays the text reuse distribution of a target book using a rectangular tree diagram. The area of each rectangle represents text reuse frequency between the corresponding chapter of the target book and the related book. By clicking on a rectangle, users can access the corresponding reused sentences.

At the sentence level, the platform assists in identifying the frequently reused sentences. It sorts the sentences of the target book based on their reuse frequency. For each sentence, the platform visualizes the diachronic change of its reuse frequency with a line graph and shows its reused books.

The pairs of text reuse sentences are entirely generated through the computation of the deep neural network model applied in the system. In the upcoming version, a user feedback function could be implemented, and validation from experts will be sought to further enhance the platform's capabilities.

*Word co-occurrence visualization for contextual analysis*. Texts are entwined with the broader contextual fabric to convey human ideas. The semantics or referential meaning of a word may undergo changes in different literature, which can be discerned through the distribution of its co-occurring words. Word co-occurrence refers to the frequency at which different words appear together within a certain context, which is widely applied in word-level cultural studies (Wijaya and Yeniterzi, 2011; Moeller et al., 2018). The Evol platform offers a co-occurrence analysis module that shows connections between words from a contextual perspective. A screenshot of this function is shown in
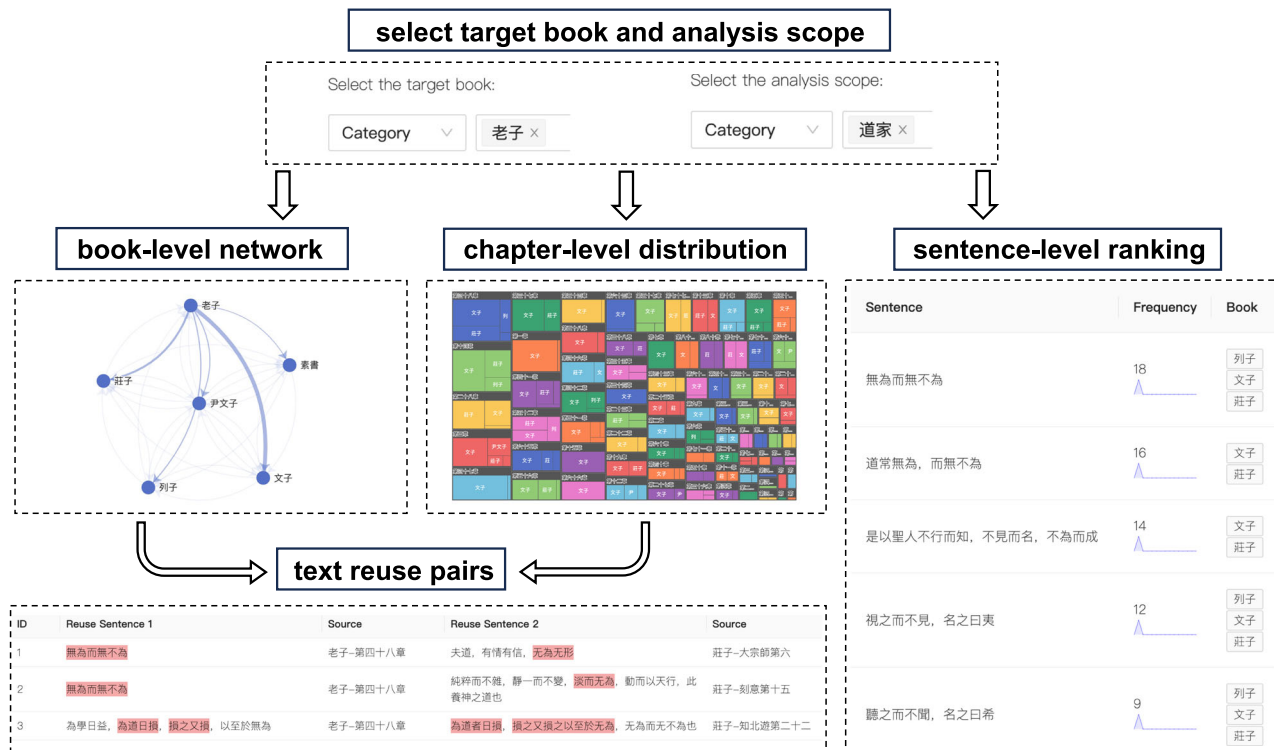
**Fig. 3 Screenshots of the text reuse analysis function.** Hierarchical visualized results and the raw data of text reuse can both be accessed.

Fig. 4a. The process of the co-occurrence function involves three steps:

1. **Retrieval**. The user inputs a textual query, and then the system locates this query in the corpus and extracts its context. The context range has three levels: paragraph, chapter, and book.
2. **Statistics**. The system performs word frequency statistics on the retrieved context.
3. **Visualization**. The statistical results are visually presented through word cloud diagrams.

After these processes, the platform produces analysis results in two forms: word cloud diagrams for interactive visualization of statistical outcomes, and text retrieval results for the query. The number of co-occurrence words can be adjusted from 10 to 100. To optimize computational efficiency, pre-processed word segmentation data is used in this function, and precomputed word frequency results are employed for chapter and book levels. In addition to viewing the visualization results online, users can also download the complete co-occurrence statistics in XLSX format. This enables more in-depth customization for further investigation.

For instance, in Fig. 4b–d, we illustrated the co-occurrence vocabulary of the word '国 (nation)' in keystone works of three philosophical schools. The divergence in views on righteousness and benefit between Confucianism and Mohism is shown in their co-occurrence frequency: Confucianism places more emphasis on righteousness, while Mohism slightly leans towards benefit. The highest frequency of the word '人 (person)' implies the individualistic characteristic of Taoism.

This module can be utilized in the study of diverse lexical categories. For instance, it enables the investigation of specific historical individuals by inputting their names, the examination of corresponding historical events by inputting relevant keywords, and the exploration of the evolution of philosophical concepts in academic discourse.

*Diachronic N-gram Vicissitude on the 2000-year timeline.* Human language is constantly changing with the evolution of society, giving rise to new vocabulary while phasing out old ones. These transformations manifest in the usage frequency of texts, often mirrors specific cultural phenomena of a particular historical epoch. The Evol platform provides a diachronic statistical function for n-grams, which allows users to investigate the frequency change of n-grams across different eras through line charts. A screenshot of this function is shown in Fig. 5. Diachronic n-gram analysis was widely promoted by the Google N-gram Viewer (Michel et al., 2011). Although this method has been applied to some sociocultural studies within modern Chinese corpora (Zeng and Greenfield, 2015; Hamamura and Xu, 2015), there is no such tool for ancient Chinese literature. Considering the distinct characteristics of ancient Chinese literature, we designed this module with several adaptations.

- **Dynasty-level timeline**. The timestamp of books allows the selected text collection to be sorted over a timeline, facilitating the observation of temporal fluctuations. Due to the sparseness and ambiguity of time information in ancient Chinese literature, an AD-year-based timeline is not suitable for data visualization. Therefore, we adopted the dynasty-level timeline instead.
- **Two timestamp kinds**. It should be noted that some Chinese ancient books were published long after the era they depicted or were written. For example, most history books were published after the era that it recorded; Some authors' anthologies, such as the *Collected Works of Tao Yuanming* (陶淵明集), were compiled and published by scholars in later generations. Considering this particularity of ancient literature, we employed two kinds of timelines, one is the publication time, and the other is the recording time. The timestamp of each book was labelled manually in the corpus-building stage. Both results on the publication timeline and the recording timeline will be visualized during analysis.
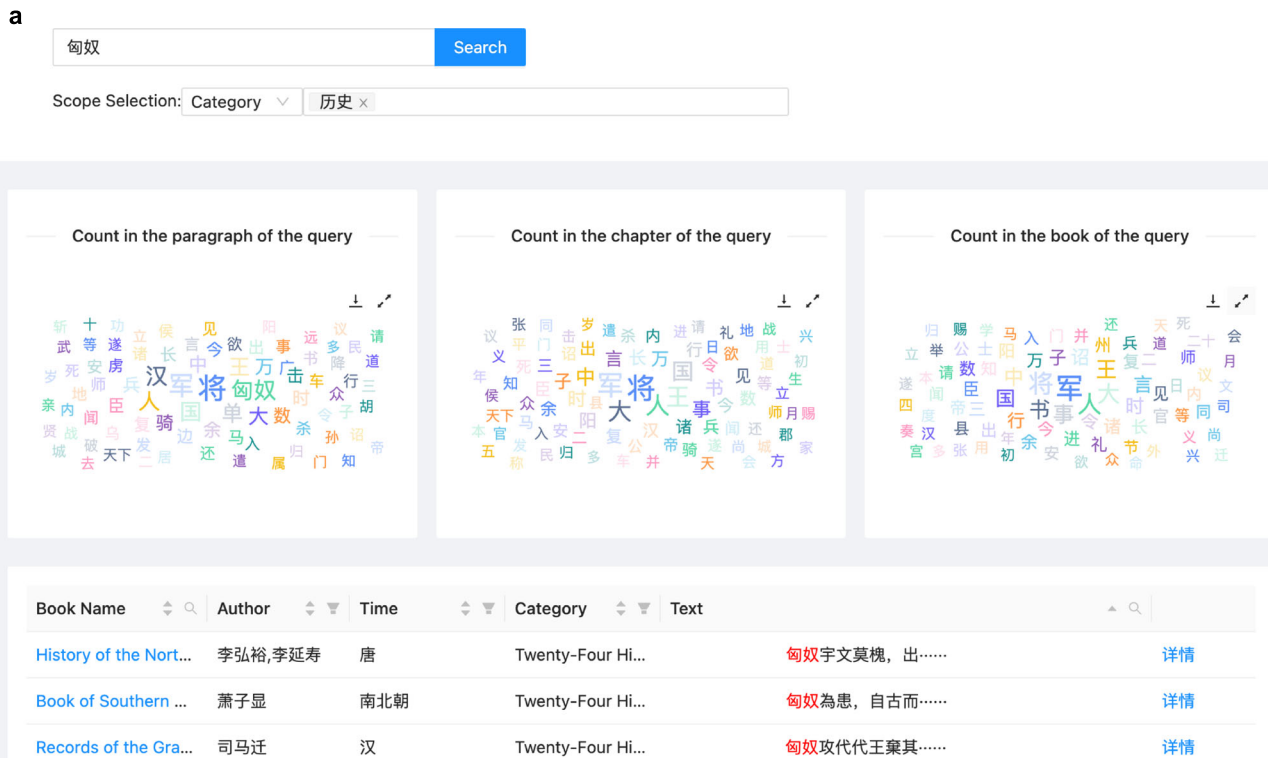
**Fig. 4 Screenshot of the co-occurrence analysis function. a** By inputting the user query and selecting the literature scope, the platform returns co-occurrence word clouds at three levels. Co-occurrence vocabulary of word '国 (nation)' in keystone works of Confucianism (**b**), Mohism (**c**), and Taoism (**d**).

- **Customized scope**. Unlike the Google N-gram Viewer system[1] which takes the entire dataset to calculate the frequency, Evol provides users with the flexibility to define their desired scope of exploration by enumerating the titles, specifying the categories, or delimiting the timespan.
- **Calculation rule**. The frequency of each n-gram in a specific dynasty is a ratio, where the numerator is the number of occurrences, and the denominator is the total number of characters in the selected set for that period. The frequency result can be a combination of multiple n-grams. For instance, as shown in Fig. 5a, users may input '皇帝 (emperor) ＋ 陛下 (your majesty) ＋皇上 (his majesty)' to amalgamate multiple variants of the word 'emperor'.

In the example illustrated in Fig. 5a, a transition between two distinct self-references of an emperor can be observed, coinciding with a pivotal historical event: the introduction of the new self-reference term 'zhen (朕)' by Qin Shi Huang during the Qin Dynasty. In Fig. 5b, c, we showed the diachronic changes of the phrase '礼乐 (ritual and music)' in historical and philosophical books, respectively. From the legendary period to the Han Dynasty, the frequency of 'ritual and music' in these two types of

literature appears to exhibit contrasting trends. This corresponds to the political upheavals in the Spring and Autumn and Warring States periods, marked by the collapse of ritual and music, and the endeavors of philosophers to restore them.

Like the Google N-gram Viewer system, this method is not immune to errors arising from polysemy. Given the characteristics of Chinese as a character-level language, these types of errors are further exacerbated. As a result, the n-gram module based on ancient Chinese literature is better suited for observing the changing usage of proper nouns and phrases. And the sparsity of ancient texts leads to non-smooth results on the timeline. These would introduce limitations to the diachronic n-gram module. Nevertheless, as shown in the cases, this function still produces meaningful results.

*Frequency count for semantic units*. Owing to its generalizability and interpretability, frequency statistics have been applied in textual analysis for quite some time. Counting semantic units within the text remains a fundamental step in quantitative cultural research. The platform provides frequency statistics functions for two types of semantic units: words and n-grams. Screenshots of these two functions are shown in Fig. 6. Once a

### Setting



### Frequency Changes – Recording time
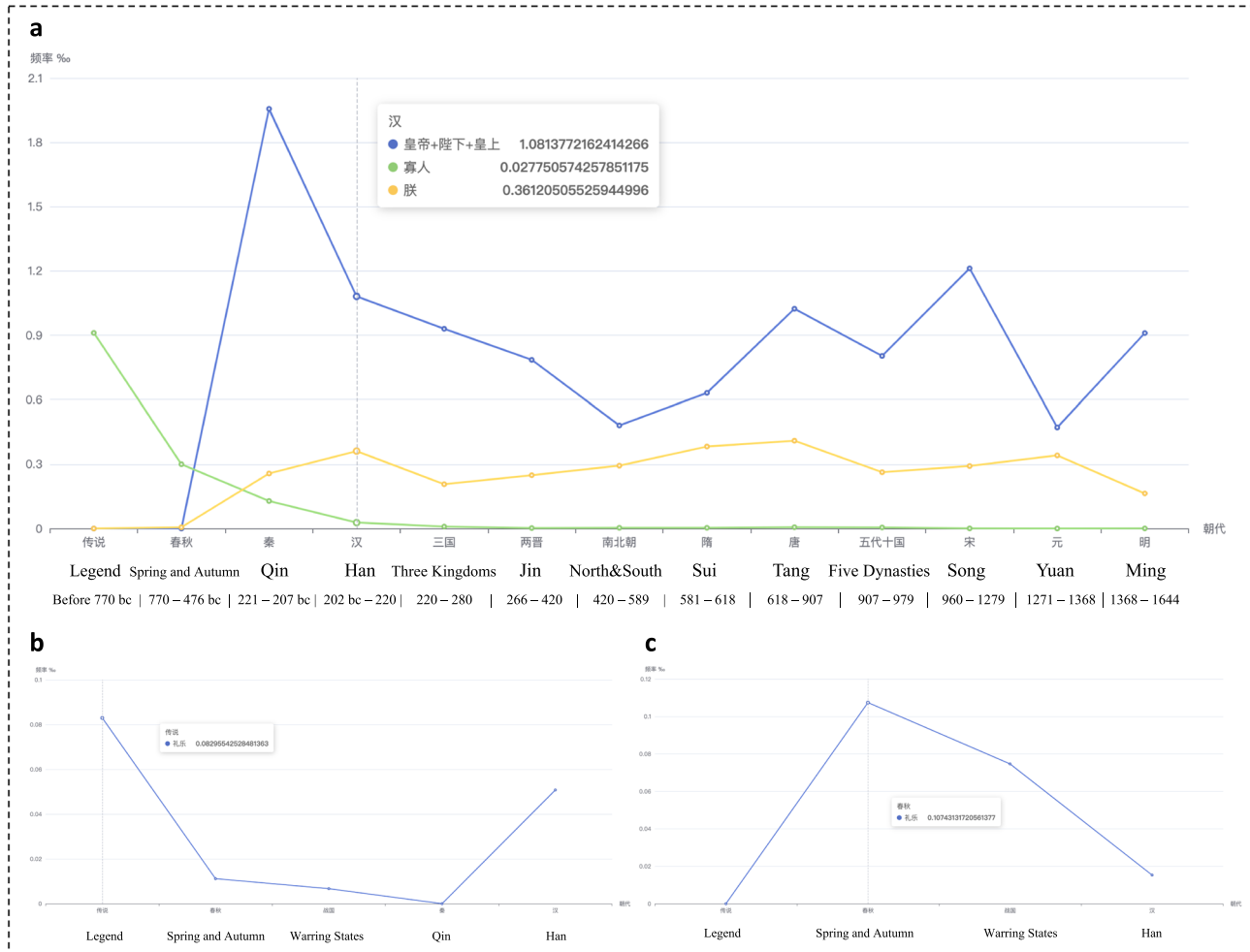


**Fig. 5 Screenshots of the diachronic n-gram function.** By inputting the user query and selecting the literature scope, the platform will return the frequency changes of n-grams on two timelines. The figure presents the outcomes sorted by the recording timeline. **a** Diachronic frequency changes of variants and self-references of an emperor. Diachronic frequency changes of phrase '礼乐 (ritual and music)' in historical books (**b**) and philosophical books (**c**) from the legendary period to the Han Dynasty.



**Fig. 6 Screenshots of the frequency count function.** By selecting the literature scope, the platform will return the count and frequency results of words or n-grams.

user selects a literature collection, the platform performs frequency statistics of words or n-grams within the selected collection.

The output statistical results include a sorted frequency count and ratio. The frequency ratio is the ratio of the frequency count to the total number of characters in the selected collection. This
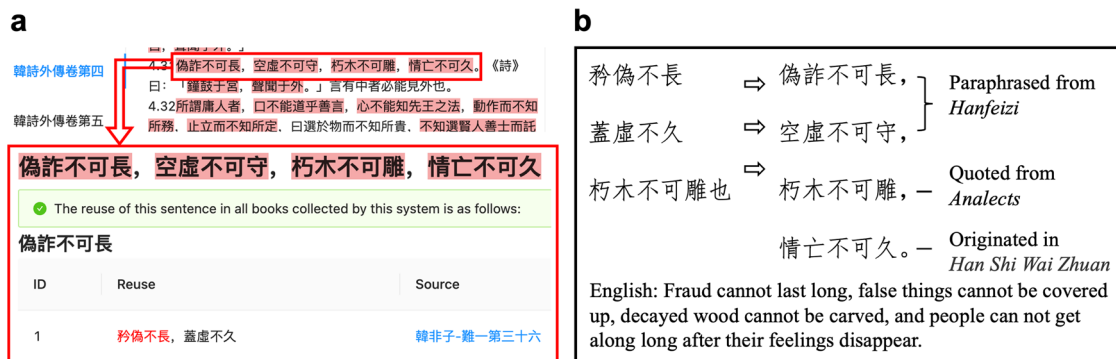
**Fig. 7 Enhanced browsing function for text reuse. a** Find text reuses by clicking the sentence highlighted in red. **b** A summary of text reuse results for a sentence in *Han Shi Wai Zhuan*. a is screenshots of the platform; b is handmade.

module can introduce a built-in Chinese dictionary to filter words in the dictionary. A stopword-based filtering is available, too, by which n-grams and words composed of stopwords will be filtered out. The platform offers a default stopwords list that users can further modify. To ensure a quick response, we pre-processed the corpus by segmenting it into words and slicing it into n-grams, and then pre-counted their frequency in each chapter.

*Enhanced browsing with text reuse linkage.* Data analysis and visualization afford scholars novel insights into the multifaceted dimensions of textual data. However, it is imperative to underscore that meticulous engagement with the primary source texts remains essential, particularly in cultural studies. The platform incorporates a corpus of ancient Chinese literature, which can be browsed online with an enhanced reading function for text reuse exploration. A schematic diagram for text reuse browsing is shown in Fig. 7a. When the user enables the text reuse button, the reused sentences will be highlighted in red. By clicking on a specific sentence, users can find similar sentences in the corpus. This function helps humanities scholars investigate the spread and evolution of texts through simple online interaction, saving time for humanities scholars to search in massive literature (He et al., 2004).

For instance, the reuses of a sentence in *Han Shi Wai Zhuan* (韩诗外传, 200 BC–130 BC) are summarized in Fig. 7b, suggesting that four parts of this sentence have different origins. The first two clauses were quoted from *Hanfeizi* (韩非子, 280 BC–233 BC) and bear semantic similarity to the original text but differ in characters. The third clause was quoted from the *Analects* (论语, 551 BC–479 BC) and shares identical characters except for the particle. The origin of the fourth clause remains unattributed, presumably original to the author of *Han Shi Wai Zhuan*. This observation reveals the complexity of text evolution, wherein authors selectively retain, succeed, and develop new content when dealing with predecessors' texts. Owing to deep neural networks, these different patterns of text reuse have been detected and built into the platform.

*Enhanced text retrieval.* Text retrieval is a basic function of the digital library. The Evol platform is equipped with a series of enhanced features to further process and display search results. A screenshot is shown in Fig. 8.

- Customizable corpus scope. Users can freely select the search scope, in which statistical functions will perform.
- Customizable search targets. Users can specify the type of search target, including book titles, chapter titles, author names, and full text.
- Fuzzy search. Fuzzy search allows for matches with a certain degree of difference. The edit distance of non-stopword characters is applied to limit the degree of fuzziness.

- Secondary search. The secondary search is performed in the context of the first search results. The context scope includes sentences, paragraphs, chapters, and books.
- Temporal visualization. Displays the frequency count (with bar chart) and frequency ratio (with line chart) of the user query in different dynasties.
- Category visualization. Displays the frequency count of the user query among different categories with a pie chart.
- Sorting. The results can be sorted by metadata such as time, book, author, and category.

## Results

Cultural evolution studies hold significance in elucidating how human society has developed into its contemporary configuration. They pertain to the transformation exhibited by diverse cultural constituents, encompassing language, value systems, societal structures, etc (Bernhardt, 1999; Yi et al., 2018). Scrutinizing cultural evolution helps fathom the principles and mechanisms that have underpinned the genesis and progression of human history and society. With Evol, users can get text analysis results through simple online interactions, which helps to start a cultural study with minimal cost. This section showcases the efficacy and potential of Evol by presenting several case studies of cultural evolution analysis on three levels: word, phrase, and document, demonstrating the effectiveness of this system in culture studies.

**Word-level evolution: attitude towards nomadic ethnic groups**. Ancient China grappled with diverse nomadic ethnic groups over thousands of years (Fei, 2017), engaging in interactions encompassing warfare, intermarriage, and diplomacy (Barfield, 1989; Di Cosmo, 2002). China's attitudes towards these foreign ethnic groups fluctuated over different periods. We employed co-occurrence analysis to investigate the evolving relationships between the ancient Chinese government and nomadic ethnic groups. By inputting the names of these ethnic groups into the co-occurrence analysis function and selecting history books of their activity period, we can find their co-occurrence words, which can reflect the attitude of the Chinese government.

Two examples are showcased in Fig. 9a, b, both are paragraph-level co-occurrences. Within the co-occurring word cloud of the Xiongnu (匈奴), a nomadic ethnic group active during the Han dynasty, numerous negative terms related to war, such as '杀 (kill)' and '死 (death)' are prominent. In contrast, during the Yuan dynasty, when the Mongols replaced the Han Nationality regime to become China's rulers, the co-occurring vocabulary was predominantly associated with political affairs, with fewer negative terms.
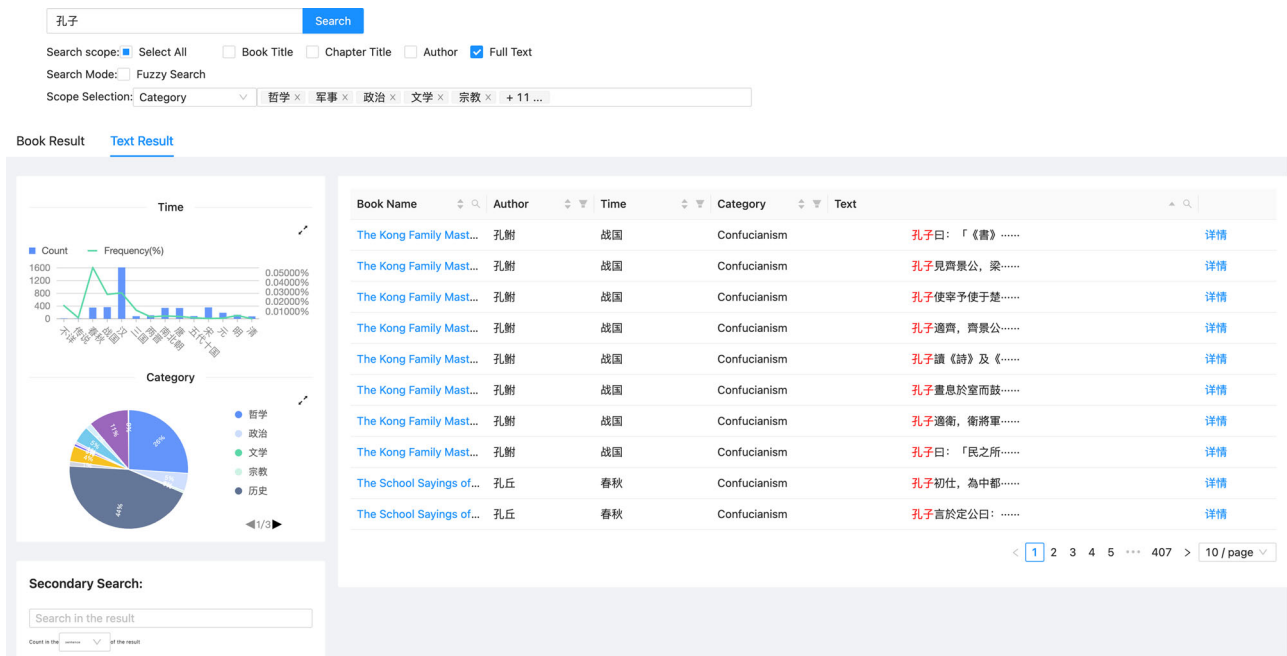
**Fig. 8 Screenshot of the text retrieval function.** Additional statistics, filtering, and visualization can be applied.
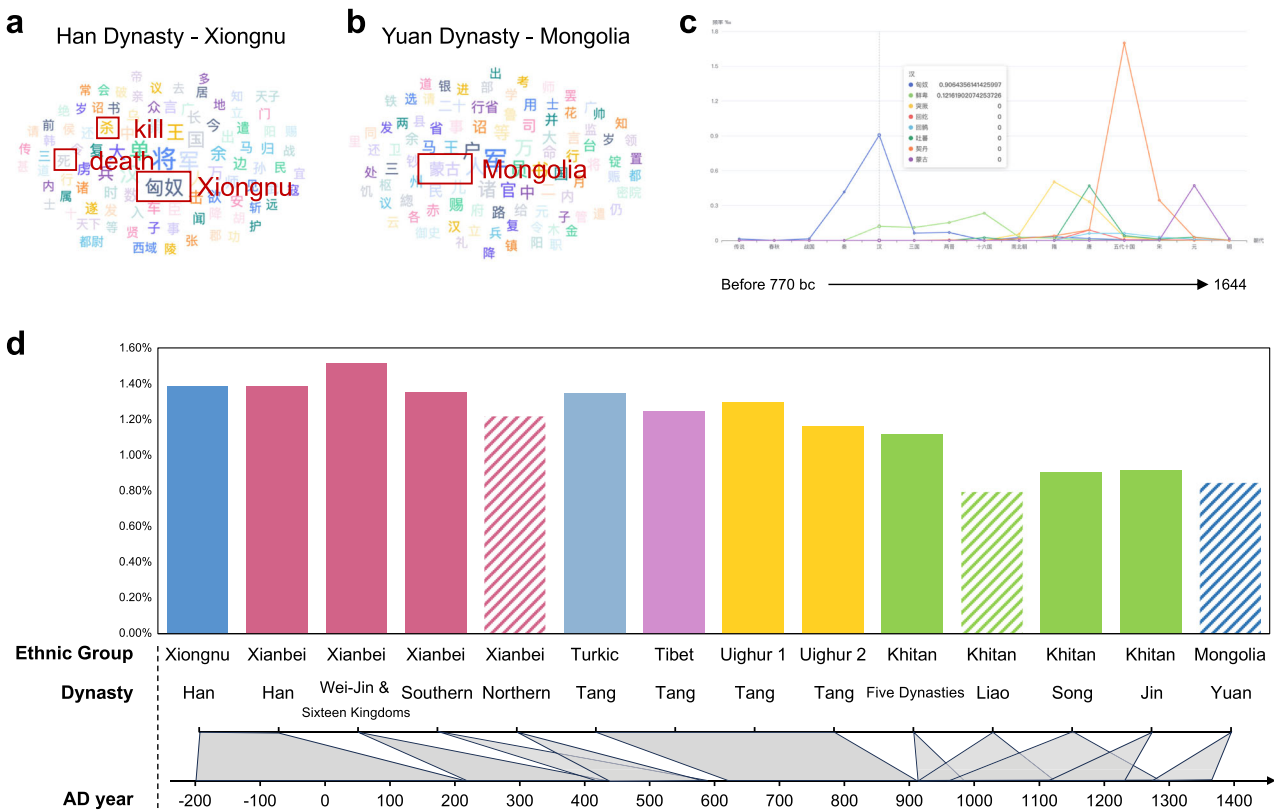


**Fig. 9 Evolution of negative sentiment towards nomadic ethnic groups in ancient Chinese history. a** Co-occurring vocabulary of Xiongnu in history books recording the Han dynasty. **b** Co-occurring vocabulary of Mongolia in history books recording the Yuan dynasty. **c** Frequency changes of nomadic ethnic group names in ancient Chinese history books. **d** Evolution of the negative sentiment contained in the co-occurring vocabulary of foreign ethnic group names, calculated within ancient Chinese history books. Each ethnic group is represented by a distinct color. The striped bars indicate that that ethnic group established the regime recorded by these historical texts. The bottom of the images is labelled with the corresponding years in the Common Era for each dynasty. a, b, c are screenshots of the platform; d is handmade.

To examine these changes across a broader timespan, we shifted our focus to a more diverse range of ethnic groups. The diachronic frequency of various ethnic groups is depicted in Fig. 9c, and we employed this result, in conjunction with historical common sense, to determine the active periods of each group. We utilized the platform to compute and download the co-occurring results for seven ethnic groups across 1500 years. For each case, we selected the top 300 words, excluding the names of the ethnic groups. Each word underwent scoring using a sentiment classification model for classical Chinese[12], which outputs probabilities for five sentiment degrees ranging from extremely negative to extremely positive. Finally, for each case, the sentiment scores were summed with word frequency weighting. We examined the average probability of extremely negative sentiment in co-occurring vocabulary, as displayed in Fig. 9d.

From the results, we can quantitatively observe changes in ancient China's attitudes toward these nomadic ethnic groups:

- Given that the Xianbei established the regimes of the Northern dynasties and the Khitan established the Liao dynasty, from the result it can be observed that, for a specific nomadic ethnic group, compared to contemporaneous historical records, the historical records of their own regime show an obvious lower level of negative sentiment toward that ethnic group. This confirms the validity of measuring hostility based on the negative sentiment of co-occurrence words.

- Overall, there is a declining trend in extremely negative sentiments towards nomadic ethnic groups. We guess that over time, China gradually eased its relations with these nomadic ethnic groups. This is consistent with the mainstream view in Chinese ethnic studies, which suggests that the mainstream trend in ethnic relations throughout Chinese history has been the increasing closeness among different ethnic groups (Weng, 1984). One exception is the Xianbei in the 2nd to 4th century, hostilities were reinforced compared to the Han dynasty.

- During the Tang dynasty, the Turkic elicited the strongest negative sentiments, followed by the Uighur and Tibet. This suggests an unfavorable relationship between the Turkic and the Tang regime. Notably, within the Uighur ethnic group, the variant Uighur1 (回纥, used before 788 AD) exhibits stronger negative sentiment compared to Uighur2 (回鹘, used after 788 AD), implying a gradual reduction in hostility towards the Uighur during the Tang Dynasty.

- From the Five Dynasties and Ten Kingdoms period to the Song dynasty, the antagonism towards the Khitan people diminished. In the same period of the Liao dynasty, two other regimes, the Song and Jin dynasties, demonstrated similar levels of hostility towards the Khitan.

**Phrase-level evolution: formulation and propagation of allusion.** Around 1046 BC, King Wu of Zhou launched an attack on King Zhou of Shang, leading to the downfall of the Shang Dynasty and the establishment of the Zhou Dynasty. This historical event, the Battle of Muye, has been extensively mentioned in literature over the next 3000 years. While commonly referred to as '武王伐纣 (King Wu attacked Zhou)', there are various textual variations of this event. Since the system has built-in millions of text reuse pairs, the enhanced browsing module allows users to find sentence reuses and variations across various literary works. In this section, we used this module to investigate the variants of this allusion. We gathered three frequently used

phrases (武王伐纣,武王克殷,武王克商), consolidated similar texts, and removed irrelevant ones. As a result, we obtained 281 texts describing the event, with 48 main variations and the frequency distribution is illustrated in Fig. 10a. Variants of different constituents within sentences were separately analyzed, including subject (shown in Fig. 10b), object (shown in Fig. 10c), and predicate (shown in Fig. 10d).

To examine the temporal changes of variant usage, we utilized diachronic n-gram functionality to assess the frequency changes of the top five most frequently used variants. The timeline spans from the Spring and Autumn periods to the Northern and Southern Dynasties periods (excluding the brief Qin dynasty), and our built-in corpus nearly comprehensively encompasses literary works created during this interval. The results are displayed in Fig. 10e. Notably, 'King Wu attacked Zhou' was not the initially predominant phrase. It was absent in books authored during the Spring and Autumn periods, while other variants were present. The usage of 'King Wu attacked Zhou' began to surface in the Warring States period and subsequently became the mainstream narrative form over the following thousand years.

To be more specific, we inputted the top three predicate variants of this phrase into a diachronic n-gram system, and the results are displayed in Fig. 10f. Notably, the usage frequency of 'ke (克)' and 'zhu (诛)' remained relatively stable, while the employment of 'fa (伐)' exhibited significant fluctuations. During the Spring and Autumn and Warring States periods, 'fa (伐)' was considerably more prevalent than the other two words. However, after this time, its usage gradually decreased and became close to the other two words. Nevertheless, the fixed phrase 'King Wu attacked Zhou' persisted as the mainstream narrative form in subsequent eras. We supposed that the fixed phrase 'King Wu attacked Zhou' emerged and gained widespread usage during the Warring States period, and in later epochs, even as the word 'fa (伐)' progressively dwindled in written usage, the fixed phrase for this allusion persisted and was not readily supplanted.

**Document-level evolution: spread across diverse cultural domains.** Yi-ology, a renowned divination school in ancient China, owes its origin to the Book of Changes (易经, around 1000 BC) (Wilhelm et al., 1967) and had a significant impact on Chinese culture (Smith et al., 2014). As the Book of Changes covers various aspects of Yi-ology, its chapters differ in popularity. Intertextuality is often used to measure the influence of texts in quantitative literary criticism, and text reuse is an effective approximation method (Büchler et al., 2013). By employing the text reuse analysis module, we can determine the number of similar sentence pairs between documents and view their distribution visually.

In this section, we used this module to evaluate the popularity of the Book of Changes. We retrieved the text reuse results of the Book of Changes in three literature collections: literature of Yi-ology, twenty-four histories (93 BC–1739), and literature of pre-Qin and Han dynasties (Legend period–220) (see Appendices for literature selection). The screenshots of their chapter-level distributions are shown in Fig. 11. In each screenshot, a rectangular section of the same color represents a chapter from the Book of Changes. Within this rectangular section, multiple smaller rectangles represent different books. The size of these smaller rectangles signifies the number of intertextual pairs between that chapter of the Book of Changes and those books.

For different chapters of the Book of Changes, we gauged their popularity based on the number of reused sentences. It can be observed that in all three screenshots, the largest rectangular sections correspond to the summary chapters of the Book of Changes, Xi Ci
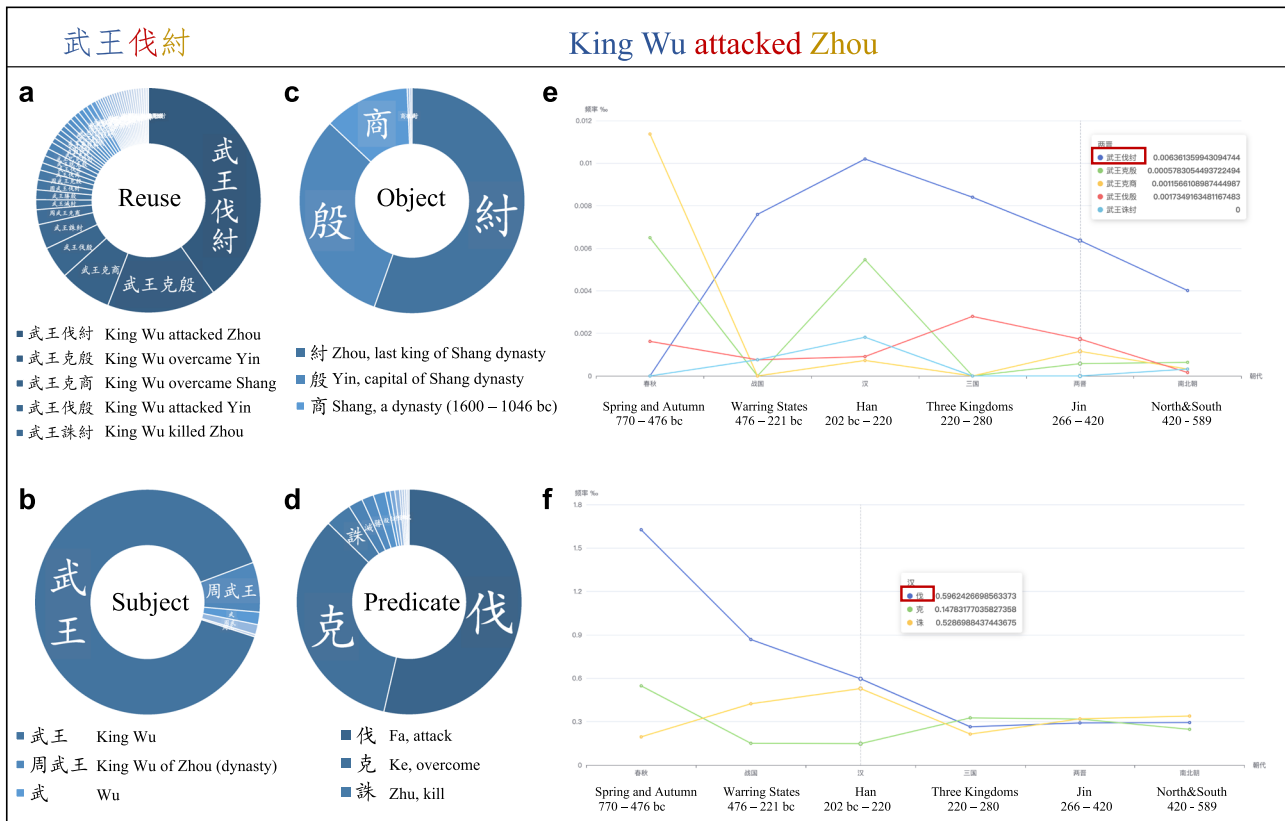
**Fig. 10 Quantitative results related to 'King Wu attacked Zhou'. a** Variation distribution of 'King Wu attacked Zhou'. **b** Variation distribution of the subject 'King Wu'. **c** Variation distribution of the object 'Zhou'. **d** Variation distribution of the predicate 'fa (attack)'. **e** Changes in the top 5 high-frequency variations of 'King Wu attacked Zhou' across ancient Chinese literature spanning over 1000 years. **f** Changes in the top 3 high-frequency variations of the predicate across ancient Chinese literature spanning over 1000 years. **a**, **b**, **c**, **d** are handmade; **e**, **f** are from screenshots of the platform.
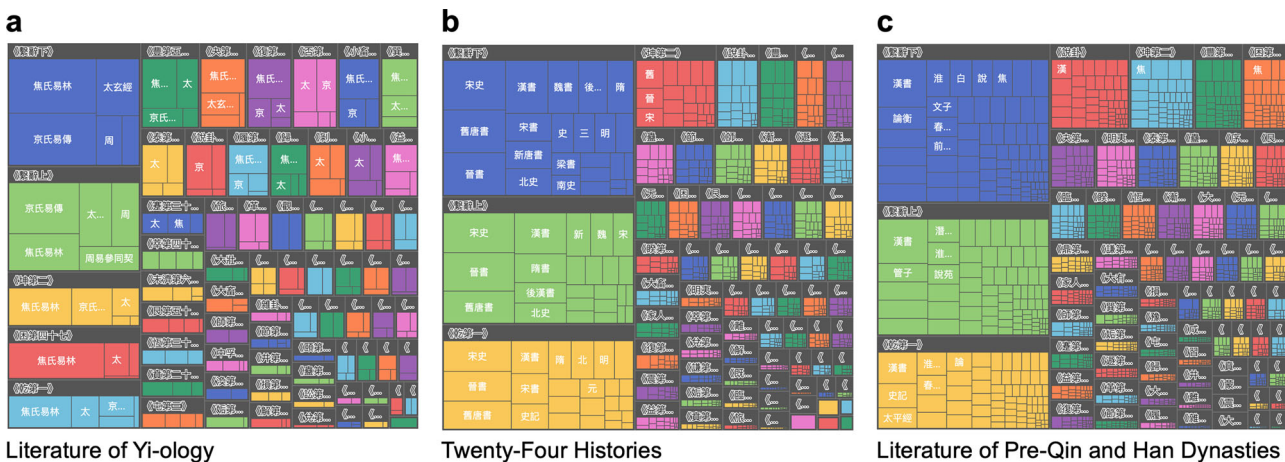


**Fig. 11 Screenshots of chapter-level text reuse distributions among three literature collections.** Different colors represent separate chapters of *the Book of Changes*. The size of each rectangle represents the amount of text reuse. **a** Text reuse distributions among the literature of Yi-ology. **b** Text reuse distributions among twenty-four histories. **c** Text reuse distributions among the literature of Pre-Qin and Han Dynasties.

(系辞), which stands out with the most frequently reused sentences among all three collections, confirming its expected prominence.

We also noticed that the chapter-level distribution of reused frequency varies across the three literature collections. There are clear differences in intertextuality distribution within and outside disciplines: In Fig. 11a, the distribution of sizes among rectangles of different colors is relatively uniform. However, in Fig. 11b, c, there is a greater disparity in the sizes of rectangles. This indicates that in Yi-ology literature which corresponds to Fig. 11a, the

reused sentences, an indicator of its influence, are more uniformly distributed among the chapters of *the Book of Changes*. While in other literature which corresponds to Fig. 11b, c, its influence is concentrated on popular chapters.

## Discussion

With a commitment to dismantling the technical barriers between computer science and humanities fields, the Evol

platform aims to provide an open and convenient online interactive experience without requiring programming skills. However, developing such an online computing platform for a large-scale corpus is a challenging task. Many practical online service issues need to be taken into consideration, including balancing computing consumption, response time, data transmission, and user experience. Being a non-profit academic institution, reasonable designs allow it to offer free services to ordinary users.

Although the quantitative results provided by the Evol platform give sound suggestions on many cultural issues, certain limitations persist. Notably, the corpus has not yet covered all ancient Chinese texts, which could potentially render some analytical outcomes incomplete or inconclusive. Besides, while the statistics results from the platform can serve as evidence for microscopic semantic change and language evolution, it may face challenges when addressing macroscopic cultural issues, such as community-level and society-level studies. In such cases, the computed results may not sufficiently support an entire research inquiry but are better suited for initial exploration and providing supplementary evidence. Most importantly, reference to traditional humanities research is indispensable. Quantitative research should not be regarded as a replacement for traditional humanities research; instead, the two approaches complement and mutually support each other. Quantitative research provides quantitative evidence to substantiate the conclusions drawn from traditional humanities inquiries. Equally, traditional humanities research contributes the essential theoretical framework and necessary interpretation required for the completion of quantitative research.

The Evol platform is a beginner-friendly tool tailored to perform quantitative cultural analysis on large-scale ancient corpora. Moving forward, we plan to expand the corpus as well as extend its functionality to accommodate multi-language data, thereby catering to a broader spectrum of users within the academic community.

### Data availability

### Notes
1 Google N-gram Viewer https://books.google.com/ngrams/
2 http://evolution.pkudh.xyz/
3 OpenCC https://github.com/BYVoid/OpenCC
4 Jiayan https://github.com/jiaeyan/Jiayan
5 Daizhige http://148.66.58.42/
6 Erudition http://dh.ersjk.com/
7 Jihe http://www.ancientbooks.cn/
8 Shidianguji https://www.shidianguji.com/
9 CSAB https://csab.zju.edu.cn/
10 Ctext https://ctext.org/
11 Tesserae https://tesseraev3.caset.buffalo.edu/
12 Guwen Sent https://huggingface.co/ethanyt/guwen-sent

### References

Alshaabi T, Adams JL, Arnold MV et al. (2021) Storywrangler: a massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. Sci Adv 7(29):eabe6534. https://doi.org/10.1126/sciadv.abe6534

Barfield TJ (1989) The perilous frontier: nomadic empires and China, 221 BC to AD 1757. Blackwell, London

Bernhardt, K, 1999. Women and property in China, 960–1949. Stanford University Press. http://www.sup.org/books/title/?id=318

Büchler M, Geßner A, Berti M et al. (2013) Measuring the influence of a work by text re-use. Bull Inst Class Stud 122:63–79. http://www.jstor.org/stable/44216323

Burns PJ, Brofos JA, Li K, et al. (2021) Profiling of intertextuality in Latin literature using word embeddings. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4900–4907. https://aclanthology.org/2021.naacl-main.389

Coffee N, Koenig JP, Poornima S et al. (2012a) The Tesserae Project: intertextual analysis of Latin poetry. Lit Linguist Comput 28(2):221–228. https://doi.org/10.1093/llc/fqs033

Coffee N, Koenig JP, Poornima S et al. (2012b) Intertextuality in the digital age. Trans Am Philological Assoc 1974:383–422. http://www.jstor.org/stable/23324457

Devlin J, Chang M-W, Lee K et al. (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186. https://aclanthology.org/N19-1423

Di Cosmo N (2002) Ancient China and its enemies: the rise of nomadic power in East Asian history. Cambridge University Press. https://doi.org/10.1017/CBO9780511511967

Duan S, Wang J, Yang H et al. (2023) Disentangling the cultural evolution of ancient China: a digital humanities perspective. Humanit Soc Sci Commun 10:310. https://doi.org/10.1057/s41599-023-01811-x

Fei X (2017) The formation and development of the Chinese nation with multi-ethnic groups. Int J Anthropol Ethnol 1:1–31. https://doi.org/10.1186/s41257-017-0001-z

Forstall C, Coffee N, Buck T et al. (2014) Modeling the scholars: detecting intertextuality through enhanced word-level n-gram matching. Digit Scholarsh Humanit 30(4):503–515. https://doi.org/10.1093/llc/fqu014

Ganascia J-G, Glaudes P, Del Lungo A (2014) Automatic detection of reuses and citations in literary texts. Lit Linguist Comput 29(3):412–421. https://doi.org/10.1093/llc/fqu020

Gao T, Yao X, Chen D (2021) SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910. https://aclanthology.org/2021.emnlp-main.552

Garg N, Schiebinger L, Jurafsky D et al. (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci USA 115(16):E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Giulianelli M, Del Tredici M, Fernández R (2020) Analysing lexical semantic change with contextualised word representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3960–3973. https://aclanthology.org/2020.acl-main.365

Greenfield PM (2013) The changing psychology of culture from 1800 through 2000. Psychol Sci 24(9):1722–1731. https://doi.org/10.1177/0956797613479387

Hamamura T, Xu Y (2015) Changes in Chinese culture as examined through changes in personal pronoun usage. J Cross-Cult Psychol 46(7):930–941. https://doi.org/10.1177/0022022115592968

Hartberg YM, Wilson DS (2017) Sacred text as cultural genome: an inheritance mechanism and method for studying cultural evolution. Relig Brain Behav 7(3):178–190. https://doi.org/10.1080/2153599X.2016.1195766

He Z, Zhu G, Fan S (2004) Parallel passages from pre-Han and Han texts series. The Chinese University Press, Hong Kong

Kozlowski AC, Taddy M, Evans JA (2019) The geometry of culture: analyzing the meanings of class through word embeddings. Am Socio Rev 84(5):905–949. https://doi.org/10.1177/0003122419877135

Lansdall-Welfare T, Sudhahar S, Thompson J et al. (2017) Content analysis of 150 years of British periodicals. Proc Natl Acad Sci USA 114(4):E457–E465. https://doi.org/10.1073/pnas.1606380114

Li W, Wang F, Wang J (2021) Exploring the classification of traditional chinese bibliographies through interactive visualization. 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 246–249. https://doi.org/10.1109/JCDL52503.2021.00071

Liu Y, Ott M, Goyal N et al. (2019) Roberta: a robustly optimized bert pretraining approach. https://doi.org/10.48550/arXiv.1907.11692

Michel JB, Shen YK, Aiden AP et al. (2011) Quantitative analysis of culture using millions of digitized books. Science 331(6014):176–182. https://doi.org/10.1126/science.1199644

Moeller J, Ivcevic Z, Brackett MA et al. (2018) Mixed emotions: network analyses of intra-individual co-occurrences within and across situations. Emotion 18(8):1106–1121. https://doi.org/10.1037/emo0000419

Newberry MG, Plotkin JB (2022) Measuring frequency-dependent selection in culture. Nat Hum Behav, 1–8. https://doi.org/10.1038/s41562-022-01342-6

Newberry MG, Ahern CA, Clark R et al. (2017) Detecting evolutionary forces in language change. Nature 551(7679):223–226. https://doi.org/10.1038/nature24455

Oishi S, Graham J, Kesebir S et al. (2013) Concepts of happiness across time and cultures. Personal Soc Psychol Bull 39(5):559–577. https://doi.org/10.1177/0146167213480042

Perc M (2012) Evolution of the most common English words and phrases over the centuries. J R Soc Interface 9(77):3323–3328. https://doi.org/10.1098/rsif.2012.0491

Smith DA, Cordell R, Dillon EM (2013) Infectious texts: modeling text reuse in nineteenth-century newspapers. 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 2013, pp. 86–94, https://doi.org/10.1109/BigData.2013.6691675

Smith JK, Bol PK, Adler JA et al. (2014) Sung dynasty uses of the I Ching. Princeton University Press

Sturgeon D (2018a) Unsupervised identification of text reuse in early Chinese literature. Digit Scholarsh Humanit 33(3):670–684. https://doi.org/10.1093/llc/fqx024

Sturgeon D (2018b) Digital approaches to text reuse in the early Chinese corpus. J Chin Lit Cult 5(2):186–213. https://doi.org/10.1215/23290048-7256963

Sturgeon D (2019) Chinese text project: a dynamic digital library of premodern Chinese. Digit Scholarsh Humanit 36(Supplement_1):i101–i112. https://doi.org/10.1093/llc/fqz046

Vaswani A, Shazeer N, Parmar N et al. (2017) Attention is all you need. Advances in neural information processing systems, 30. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Weng D (1984) 论中国民族史. Ethno-National Studies (民族研究) (4):1–8

Wijaya DT, Yeniterzi R (2011) Understanding semantic change of words over centuries. In: Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web. pp. 35–40. https://doi.org/10.1145/2064448.2064475

Wilhelm R, Baynes CF, Jung CG (1967) The I Ching or Book of Changes. Bollingen series, 19

Yi X, van Leeuwen B, van Zanden JL (2018) Urbanization in China, ca. 1100–1900. Front Econ China 13(3):322–368. https://doi.org/10.3868/s060-007-018-0018-9

Zeng R, Greenfield PM (2015) Cultural evolution over the last 40 years in China: using the Google Ngram Viewer to study implications of social and political change for cultural values. Int J Psychol 50(1):47–55. https://doi.org/10.1002/ijop.12125

## Acknowledgements

## Author contributions
Jun Wang initially conceived and managed the project and wrote the paper. Siyu Duan designed and participated in the development of the platform, analyzed the data, and wrote the paper. Binghao Fu participated in the development of the platform and revised the paper. Liangcai Gao participated in the development of the platform and revised the paper. Qi Su supervised the development of the platform and revised the paper.

## Competing interests
The author(s) declare no competing interests.

## Ethical approval
Ethical approval was not required as the study did not involve human participants.

## Informed consent
Informed consent was not required as the study did not involve human participants.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-02763-6.

**Correspondence** and requests for materials should be addressed to Siyu Duan.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.