

CHisAgent: A Multi-Agent Framework for Event Taxonomy Construction in Ancient Chinese Cultural Systems

Xuemei Tang¹ Chengxi Yan^{2*} Jianghang Gu^{1*} Chu-Ren Huang¹

¹The Hong Kong Polytechnic University

²Renmin University of China

xuemeitang00@gmail.com, 20218113@ruc.edu.cn

{jinghang.gu, churen.huang}@polyu.edu.hk

Abstract

Due to the long temporal span and highly diverse expressions in historical documents, large language models (LLMs) still struggle with historical and cultural reasoning in Chinese history. Although taxonomic structures can effectively organize historical knowledge and facilitate knowledge-rich applications, existing historical taxonomies remain limited due to the high cost of manual construction. To address this limitation, we propose **CHisAgent**, a multi-agent LLM framework for automated historical taxonomy induction in ancient Chinese history. CHisAgent explicitly decomposes taxonomy construction into three role-specialized stages: a bottom-up *Inducer* that derives an initial hierarchy from raw historical corpora, a top-down *Expander* that introduces missing intermediate concepts using LLM world knowledge, and an evidence-guided *Enricher* that integrates external structured historical resources to ensure faithfulness. Extensive reference-free and reference-based evaluations show that the collaboration among these agents improves taxonomy coverage and hierarchical completeness, while further analysis suggests that the induced taxonomy facilitates cross-cultural alignment of historical concepts. Code and experimental data are available at <https://github.com/tangxuemei1995/CHisAgent>.

1 Introduction

Recently, LLMs have achieved strong performance in real-world understanding tasks such as emotion understanding (Belay et al., 2025) and scene planning (Sun et al., 2025). However, several works suggest that these models exhibit limited capacity for historical and cultural reasoning, particularly in non-English cultural contexts (Hauser et al., 2024; Uraierlertprasert et al., 2024; Tang et al., 2024; Ghaboura et al., 2025; Bu et al., 2025). Moreover, even when provided with explicit historical texts,

models often struggle to capture deeper cultural knowledge beyond surface-level patterns (Tang et al., 2024), as such knowledge is structured hierarchically and exhibits diverse expressions.

Taxonomies provide a structured abstraction of domain concepts through hierarchical “is-a” relations (Shang et al., 2020), serving as an effective way to organize knowledge and facilitate model understanding. Compared with knowledge graphs and ontologies, taxonomies require less complex structural and semantic modeling, making them more scalable for organizing large-scale historical and cultural knowledge across long temporal spans (Franz and Thau, 2010). Moreover, taxonomies can serve as a semantic foundation for constructing knowledge graphs and ontologies (Pan et al., 2025), and support downstream tasks such as historical information retrieval, cultural knowledge alignment, and historically grounded reasoning in LLMs.

Previous works have explored taxonomies in specific historical domains, such as Italian public administration (Pandolfo, 2015) and traditional Chinese time systems (Wang et al., 2024). However, historical documents are characterized by large scale, complex narratives, and highly diverse expression styles (Chang et al., 2021; Tang et al., 2024). As a result, existing historical taxonomies remain limited in scale and cultural coverage because manual taxonomy construction is labor-intensive (Bordea et al., 2016). Recent studies have shown that LLMs can effectively support automated taxonomy induction and expansion (Chen et al., 2023; Lahiri et al., 2025; Straková et al., 2023; Ghamlouch and Alam, 2025), offering new possibilities for large-scale historical taxonomy construction.

Therefore, in this work, we propose a multi-agent LLM framework for historical event taxonomy construction based on the *Twenty-Four Histories*, a canonical collection of official Chinese

*Corresponding Author

historical records spanning ancient to late imperial China. We adopt an event-centered taxonomy design, as events and their roles serve as fundamental units in historical narrative understanding (Goy et al., 2015; Li et al., 2020). In this framework, LLM agents are assigned distinct roles across three stages. The framework first induces an initial taxonomy from raw historical corpora in a bottom-up manner, then expands it through top-down generalization using LLM world knowledge, and finally refines it by incorporating corpus-derived evidence and external structured historical resources to ensure faithfulness. This framework significantly accelerates the construction of historical knowledge organizations.

Our main contributions are summarized as follows:

- We propose **CHisAgent**, the multi-agent LLM framework specifically designed for large-scale historical event taxonomy induction, which decomposes the task into induction, expansion, and completion stages with explicitly role-specialized agents.
- We construct a large-scale event taxonomy for ancient Chinese history and culture based on the *Twenty-Four Histories*, spanning over 3,000 years and covering major political, military, diplomatic, and social events.
- We conduct extensive evaluations under both reference-free and reference-based settings, demonstrating strong coverage and alignment with source texts, and further experiments validate its effectiveness in cross-cultural alignment.

2 Related Work

2.1 Automated Taxonomy Induction

Taxonomy construction generally consists of multiple stages, including entity extraction, hierarchy induction, taxonomy expansion, and evaluation. Each stage is traditionally labor-intensive and costly, which has motivated extensive exploration into automated methods (Shen et al., 2018; Chen et al., 2021; Lee et al., 2022). Recent LLMs have been increasingly applied to automate multiple stages of taxonomy construction. For example, Zeng et al. (2024) proposed an in-context learning framework (Chain-of-layer) for taxonomy induction from a given entity set, evaluated through

ancestor-descendant matching against existing taxonomies using precision, recall, and F1. Shen et al. (2025) reformulated entity set expansion, taxonomy expansion, and seed-guided taxonomy construction as sibling and parent identification tasks, and trained instruction-tuned LLMs to address these tasks. Similarly, Zeng et al. (2025) proposed CodeTaxo, which represents taxonomies as executable Python programs for LLM-based taxonomy expansion. Their evaluation uses accuracy and Wu & Palmer similarity to measure node assignment correctness and structural semantic similarity.

In addition, LLM-driven approaches have been applied to taxonomy construction in specific domains such as engineering, social media, and research fields (Aggarwal et al., 2026; Lahiri et al., 2025; Zhang et al., 2025).

Multi-agent framework. Building on single-model approaches, recent studies assign LLMs specialized roles and allow them to collaborate with other agents or human experts to induce and expand taxonomies. For instance, Li et al. (2025) proposed CLIMB, which converts clustering outputs into canonical occupational nodes and refines the hierarchy via iterative generator–evaluator interaction. Their evaluation is indirect, relying on LLM-based annotation performance to measure label consistency, coverage, and usability rather than gold-standard taxonomies. Kargupta et al. (2025) introduced TaxoAdapt for scientific literature, a multidimensional adaptive framework that uses LLMs to adjust taxonomy width and depth based on topical distributions. It first generates a coarse taxonomy and then iteratively expands lower-level structures with LLM-guided rules. Evaluation combines automatic and human assessment of hierarchical validity, coherence, and coverage at both node and structure levels.

2.2 Taxonomy Induction in Historical and Cultural Domains

Prior work in cultural heritage has explored ontology-based knowledge modeling for historical artifacts. For example, a CIDOC CRM-aligned ontology (Zhang, 2026) was developed for Chinese painting collections (CCACP), integrating metadata standards to support knowledge graph construction and cultural analysis. In the context of Chinese historical knowledge, Wang et al. (2024) proposed a time ontology for representing traditional Chinese dating systems, focusing primarily on temporal structures. Beyond ontologies,

several studies have constructed domain-specific taxonomies for historical applications. Wei et al. (2023) manually developed an event taxonomy based on the Twenty-Four Histories for event extraction, while Tang et al. (2025) designed a relation taxonomy to support structured historical relation extraction. However, existing taxonomies are relatively coarse-grained and task-oriented, with limited coverage of fine-grained conceptual distinctions. The lack of effective methods for processing large-scale historical documents has hindered further development, while the emergence of LLMs provides new opportunities for automatic knowledge organization.

3 Methodology

Our framework consists of three LLM-based agent modules: the *Inducer*, the *Expander*, and the *Enricher*, designed to address key challenges in historical event taxonomy construction.

Historical documents pose three major challenges in taxonomy construction. First, they often exhibit condensed expressions and stylistic variation, making event boundary detection and trigger identification difficult. Second, historical event taxonomies derived from corpora are often incomplete, with missing intermediate concepts and irregular hierarchical structures due to uneven coverage and limited structural reasoning ability. Third, historical understanding frequently requires external contextual knowledge, while LLM-generated content may suffer from semantic drift and insufficient grounding.

To address these challenges, we first introduce the **Inducer**, which performs bottom-up event hierarchy induction using a multi-extractor design with an iterative refinement loop, targeting the difficulty of event boundary detection and implicit structure extraction. However, the resulting hierarchy may remain incomplete or structurally inconsistent. Therefore, the **Expander** introduces missing intermediate concepts to improve hierarchical completeness and coherence. Finally, to ensure historical faithfulness and reduce semantic drift, the **Enricher** integrates corpus evidence with external structured historical knowledge for verification and grounding.

3.1 Inducer

The Inducer performs event extraction, domain classification, and bottom-up taxonomy induction

in a collaborative agent framework.

We first construct the corpus from the *Twenty-Four Histories*. We extract major chapter types, including “傳” (Biographies), “誌” (Treatises), “書” (Records), and “紀” (Annals), from each historical document, resulting in 192 chapters and 584,096 characters (details in Appendix A). Then extractor agents identify event instances and event types from each chapter. We use multiple extractors to improve coverage. This yields an event set $E = \{e_1, \dots, e_M\}$. We find that LLM-based extraction is sufficiently reliable for constructing an event type pool, as further analyzed in Appendix B.

A classifier agent assigns each event type to a domain $d \in D$ based on event descriptions and contextual evidence. Event types are aggregated within each domain as $E_d = \{e_1, \dots, e_m\}$. The domains are refined through clustering and expert validation, covering eight categories including Politics, Military, Diplomacy, Society, Ritual, Nature, and Individual. The full description of each domain is provided in Appendix C. The number of event types in each domain is presented in Appendix D.

Inducing the Initial Taxonomy. Within each domain, we construct the taxonomy using a bottom-up iterative process with two specialized LLM roles. Three Generator LLMs merge semantically related event types to form higher-level nodes, while a Merger LLM performs deduplication and concept unification. The refined results are then fed back to each Generator to produce the next level of abstraction. This iterative process continues until the number of high-level event categories within a domain falls below ten, which is empirically chosen based on preliminary experiments to balance abstraction and interpretability. In practice, this stopping criterion is typically reached after two or three merging rounds, resulting in an initial taxonomy.

Each node is denoted as $n_{d,l,i}$, where d is the domain, l the hierarchy level, and i the node index. The taxonomy for domain d is defined as $\mathcal{T}_d = (\mathcal{N}_d, \mathcal{H}_d)$, where $\mathcal{N}_d$ denotes nodes with definitions and $\mathcal{H}_d$ denotes parent-child relations. As shown in Figure 2, we use “Diplomacy” as an example of the initial taxonomy produced by the Inducer (partial).

3.2 Expander

The Expander addresses structural incompleteness in the initial taxonomy induced by the Inducer, which is caused by loss of intermediate abstrac-

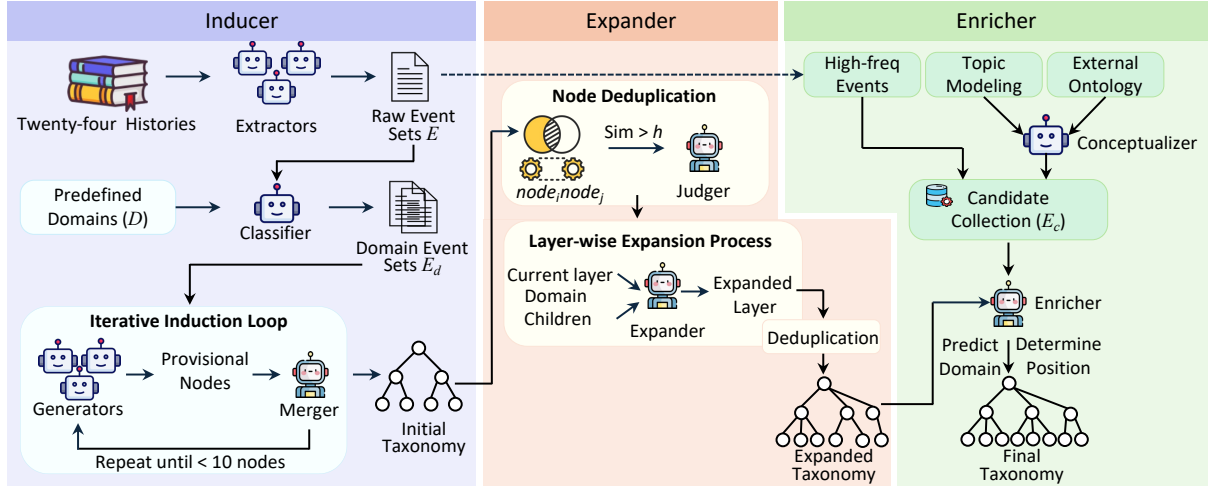


Figure 1: Illustration of the multi-agent CHisAgent framework, consisting of three key stages: the Inducer, the Expander, and the Enricher.

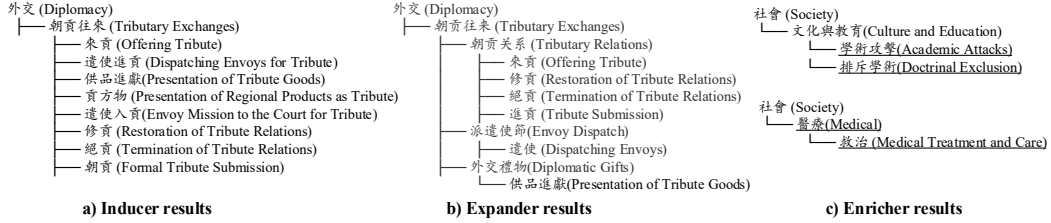


Figure 2: Illustration of different stages. The Inducer and Expander use an example from the Diplomacy domain, while the Enricher uses the Society domain. Underlined nodes are introduced by the Enricher.

tions during bottom-up aggregation. It performs layer-aware top-down refinement to improve hierarchical completeness within each domain.

The expansion process consists of node deduplication, hierarchical expansion, and post-expansion cleaning.

Node Deduplication. To reduce redundancy, we perform both within-domain and cross-domain deduplication. For any node pair (n_i, n_j) , we compute semantic similarity $s_{ij} = \text{Sim}(n_i, n_j)$. Nodes with $s_{ij} > h$ are considered redundant, and a judge LLM selects a representative node.

Layer-wise Hierarchical Expansion. We then perform top-down expansion to recover missing or underrepresented conceptual structures. For each layer l , the model receives current nodes, their children, and domain context, and identifies (i) missing sibling categories, (ii) missing intermediate abstractions, and (iii) inconsistent parent-child assignments. The Expander then inserts new nodes and reassigns children to enforce hierarchical consistency. As shown in Figure 2, the Expander adds a new layer to the Diplomacy domain by inserting intermediate nodes to refine the structure of lower-level nodes.

Post-expansion Cleaning. Since new nodes

may introduce redundancy or semantic overlap, a final global deduplication step is applied to ensure structural compactness.

3.3 Enricher

The LLM-based **Enricher** improves the completeness and corpus alignment of the expanded taxonomy while preserving faithfulness to historical evidence. It follows a two-step design to recover event concepts that may be omitted during the *Inducer* and *Expander* stages.

Candidate Event Collection. We construct a comprehensive candidate event set E_c from three complementary sources. (1) *High-frequency events* extracted from the initial event set E , where event types with frequency $f > 5$ are selected as frequent candidates (E_{freq}). (2) *Topic-based events* derived from the corpus using BERTopic (Groendorst, 2022), yielding latent topics $\text{Topics} = \{t_1, \dots, t_k\}$ that capture potential missing mid- and high-level event structures (Tang et al., 2026). Detailed topic statistics are provided in Appendix E. (3) *Ontology-derived relations* collected from external historical knowledge bases, including the Chinese Biographical Database (CBDB) and the historical relation ontology proposed by Tang et al.

(2025), forming a relation set $R = \{r_1, \dots, r_n\}$.

Since topics and relations are not directly aligned with event-type representations, an LLM-based **Event Conceptualizer** transforms $Topics$ and R into explicit event-type candidates, producing E_{Topics} and E_R . The full candidate set is: $E_c = E_{freq} \cup E_{Topics} \cup E_R$. We then perform global deduplication over E_c using semantic similarity filtering to remove redundant event types.

Positioning Candidate Events. For each candidate event $e \in E_c$, the Enricher first predicts its most likely domain d based on event semantics and domain descriptions. It then checks redundancy with existing nodes in the current domain taxonomy $\mathcal{T}_d$. If no duplication is detected, e is inserted into an appropriate position within $\mathcal{T}_d$, yielding the enriched taxonomy.

The Enricher improves corpus alignment and representational fidelity by incorporating previously missing but evidence-supported event concepts from external sources. As shown in Figure 2, it augments the taxonomy with additional nodes derived from external knowledge.

4 Experiments

4.1 Experimental Setup

Based on preliminary experiments, we assign different LLMs to specialized roles across stages according to their relative task performance. In the extraction and induction stage, DeepSeek-V3 and Qwen3 are used as extractor agents, while DeepSeek-V3, Qwen3, and GPT-4o are used as generator agents. GPT-4o further serves as the domain classifier. GPT-5 serves as the merger role. In the expansion stage, GPT-5 functions as the judge for resolving semantic duplicates and also acts as the expander for generating higher-level taxonomy nodes. In the enrichment stage, GPT-4o is used as the event conceptualizer, while GPT-5 serves as the Enricher to integrate candidate events into the taxonomy. The API versions and costs of all models are provided in Appendix F.

For all semantic similarity computations, taxonomy nodes and their definitions are encoded using text-embedding-3-small, and cosine similarity is applied for comparison. Semantic duplication is identified using a similarity threshold of 0.6, above which node pairs are considered redundant; this threshold is determined through manual evaluation.

4.2 Baselines

We compare our framework with four categories of taxonomy construction methods.

Traditional Taxonomy Induction Methods.

We reproduce two representative non-LLM taxonomy induction approaches: HiExpan and HiSetExpan (Shen et al., 2018).

Single-Agent LLM Methods. We design three prompting-based taxonomy construction settings using GPT-4o, DeepSeek-V3, and Qwen3:

LLM + Domains, where the model generates taxonomies from predefined domains in a top-down manner;

LLM + Event Corpus, where the model induces taxonomies bottom-up from fine-grained event types;

LLM + Event Corpus + Domains, where both domain knowledge and event types are provided for intermediate taxonomy induction.

Multi-Agent Methods. We further compare with two recent multi-agent taxonomy construction frameworks: Chain-of-Layer (CoL) (Zeng et al., 2024) and TaxoAdapt (Kargupta et al., 2025).

Human-Crafted Taxonomy. We include CHED (Wei et al., 2023) as a human-authored reference taxonomy derived from the *Twenty-Four Histories*.

Implementation details of all baselines are provided in Appendix G.

4.3 Evaluation Metrics

We evaluate the constructed taxonomies using two types of metrics, reference-free and reference-based, covering structural and semantic aspects. The detailed calculation procedures for all metrics are provided in Appendix H.

Reference-free Metrics are as follows.

Path Granularity (Kargupta et al., 2025). This metric assesses whether a node $(n_{d,l,i})$ is semantically more specific than its parent. The score (0 or 1) is assigned by GPT-4o.

CSC (Wulschleger et al., 2025). A structural consistency metric that evaluates whether distances between nodes align with their semantic relations. Higher scores indicate stronger alignment.

Coverage Rate. This metric measures coverage of historical events in the *Twenty-Four Histories*, extracted from non-overlapping chapters.

Reference-Based Metrics. These metrics compare the constructed taxonomy with the human-curated taxonomy of CHED (Wei et al., 2023),

Table 1: Evaluation results of taxonomies generated by different methods.

Categories	Methods	Settings	Path Gran. $\uparrow$	CSC $\uparrow$	Cover. Rate $\uparrow$	Node Recall $\uparrow$	Novelty $\uparrow$	Significance $\uparrow$
Human	CHED	-	80.25	0.1432	45.39	-	-	-
Single-agent	Qwen3	Domain	95.83	0.2566	35.68	43.33	59.62	-0.0197
	DeepSeek-V3		95.40	0.1366	39.87	43.33	54.74	-0.0318
	GPT-4o		98.96	0.2973	30.82	53.33	53.85	-0.0326
	Qwen3	Corpus	93.55	0.1213	46.23	48.89	59.22	0.0174
	DeepSeek-V3		76.56	0.1556	41.29	42.22	44.93	-0.0211
	GPT-4o		100.00	0.3931	27.47	33.33	49.15	-0.0952
	Qwen3	Corpus + Domain	88.75	0.2356	33.67	36.67	59.09	-0.0379
	DeepSeek-V3		79.80	0.1485	61.81	53.33	66.99	0.0769
	GPT-4o		97.50	0.2590	37.52	48.89	36.36	-0.0331
Non-agent	HiExpan	Corpus + Domain	62.80	0.1896	42.21	42.22	74.50	0.1098
	HiSetExpan		83.33	0.2801	28.14	25.56	71.82	0.0515
Multi-agent	CoL	Domain	86.39	0.1895	44.44	52.17	80.92	-0.1424
		Corpus + Domain	80.99	0.2479	43.72	46.67	50.69	0.0104
	TaxoAdapt	Corpus + Domain	70.98	0.2075	<u>74.17</u>	<u>60.00</u>	<u>69.21</u>	0.3927
	CHisAgent	Corpus + Domain	91.54	0.2784	75.13	68.89	77.70	<u>0.1887</u>

Table 2: Ablation results of different components in our framework.

Method	Components			Path Gran.	CSC	Cover. Rate	Node Recall	Novelty	Significance
	Inducer	Expander	Enricher						
CHisAgent	✓	×	×	90.45	0.2761	71.61	61.11	77.22	0.1310
	✓	✓	×	94.35	0.2464	72.36	65.56	77.59	0.1667
	✓	✓	✓	91.54	0.2784	75.13	68.89	77.70	0.1887

capturing different aspects of deviation from the reference.

Node Recall. Measures the proportion of human taxonomy nodes captured by the constructed taxonomy under a semantic similarity threshold.

Novelty (Kaplan et al., 2022). Quantifies the proportion of newly introduced nodes not present in the human taxonomy. Higher values may indicate either true extensions or over-generation beyond the reference.

Significance (Kaplan et al., 2022). Measures the difference in granularity compared to the human taxonomy. Positive values indicate finer-grained structures, while negative values indicate coarser ones.

We note that reference-based metrics should be interpreted jointly, as they reflect different aspects of coverage and deviation from the human taxonomy rather than uniformly better or worse quality.

4.4 Overall Experimental Results

The evaluation results are presented in Table 1. We first observe that single-agent generation methods tend to achieve higher Path Granularity and CSC scores than multi-agent methods, as they construct the entire taxonomy in a unified process. Traditional taxonomy induction methods show no clear advantage in complex systems with large-scale event type spaces. In contrast, multi-agent methods provide a more explicit decomposition of the taxonomy construction process, leading to improved structural coverage and better alignment with ref-

erence taxonomies. In particular, our method achieves the best results in terms of Coverage Rate and Node Recall. TaxoAdapt obtains the highest Significance score, indicating that it produces the most fine-grained taxonomy relative to the human reference. Chain-of-Layer (Domain) achieves the best performance on the Novelty metric. Statistics of the generated taxonomies are provided in Appendix I.

4.5 Ablation Study

Table 2 presents an ablation study evaluating the contributions of the three agent modules: Inducer, Expander, and Enricher. Using only the Inducer yields the lowest Node Recall and Coverage Rate, indicating that the induced taxonomy remains relatively coarse-grained and incomplete. Adding the Expander substantially improves Path Granularity and Node Recall, showing its effectiveness in constructing deeper hierarchical structures. The slight decrease in CSC suggests that maintaining semantic consistency becomes more challenging as the taxonomy expands. Further incorporating the Enricher improves Coverage Rate, Node Recall, and Significance, indicating that it effectively supplements missing concepts and enriches historical content. Overall, the results show that the three modules play complementary roles in improving taxonomy quality.

4.6 Domain Analysis

We further compare taxonomies generated by GPT-4o (Corpus + Domain), CoL (Corpus + Domain), TaxoAdapt, and our method across eight domains using the radar charts in Figure 3.

Figure 3(a) reports Coverage Rate, measuring how well each method captures event types extracted from other chapters of the *Twenty-Four Histories*. Most methods obtain higher coverage in Politics, Military, and Individual categories, while TaxoAdapt additionally performs well in Society. This metric is partially influenced by the quality and recall of LLM-based event extraction, which may introduce domain-specific biases.

In addition, Figure 3(b) reports Novelty, reflecting the introduction of nodes absent from the human taxonomy. All methods generate the largest number of novel nodes in Individual, Society, and Politics. Our method shows particularly high novelty in Individual and Economy-Livelihood, whereas GPT-4o and TaxoAdapt produce more novel nodes in Society. In contrast, CoL exhibits a relatively balanced novelty distribution across domains.

Additional analyses of parent-child relationship rationality and recalled node distributions are provided in Appendix J.

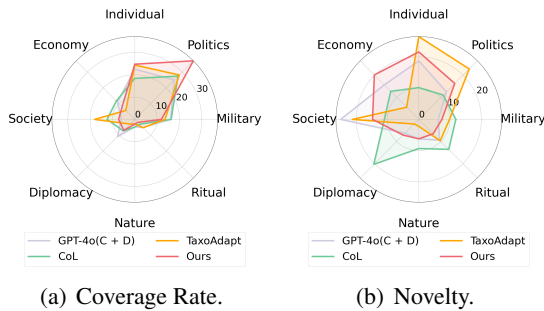


Figure 3: Radar chart of evaluation metrics of different methods across various domains. GPT-4o (C + D) denotes GPT-4o (Corpus + Domain), and “Economy” represents “Economy-Livelihood”.

4.7 Human Evaluation

Following Zhang et al. (2025), we conduct a human evaluation to assess the quality of taxonomies generated by different methods (GPT-4o (Corpus + Domain), CoL (Corpus + Domain), TaxoAdapt, and Ours). The evaluation focuses on semantic clarity, structural coherence, and semantic distinction, and considers five complementary perspectives as follows.

Node Clarity. Evaluators judge whether the

node’s semantics are clear and whether the node can appropriately represent an event type in ancient Chinese cultural system. Scores range from 0 to 5.

Hierarchical Coherence. Evaluators assess whether the hierarchical structure from the root to the leaf is logically coherent. Scores range from 0 to 5.

In-domain Orthogonality. Evaluators measure the degree of semantic independence among nodes within the same domain, assessing whether taxonomy concepts exhibit minimal semantic overlap and redundancy. Scores range from 0 to 5.

Cross-domain Orthogonality. Evaluators assess whether nodes assigned to different domains are semantically distinguishable from each other. Scores range from 0 to 5.

Completeness. Evaluators judge whether each complete taxonomy sufficiently covers the conceptual space of Chinese historical events. Scores range from 0 to 5.

Details of node sampling and evaluation guidelines are provided in Appendix K.

We employ three independent evaluators with expertise in Chinese history. After two rounds of discussion, the inter-rater agreement demonstrates moderate-to-substantial consistency across evaluators, with a weighted Cohen’s kappa of 0.6982, a Spearman correlation of 0.5722, and an ICC(2,1) score of 0.6992, indicating generally reliable human judgments. Following prior work, the final scores were obtained by averaging annotator ratings. Human evaluation results are presented in Figure 4. TaxoAdapt achieves the highest scores in Clarity and Completeness, but obtains relatively lower Cross-domain Orthogonality, suggesting semantic overlap across domains despite extensive node generation. In contrast, GPT-4o (Corpus + Domain) attains the highest Cross-domain Orthogonality, indicating strong separation between categories, but exhibits lower completeness. Overall, our method achieves a more balanced performance across all metrics, maintaining high levels of completeness, semantic clarity, and structural orthogonality, which aligns closely with human judgments. A more detailed analysis of model error patterns is provided in section 4.11.

4.8 Human Evaluation of Event Extraction

Since extraction errors may propagate to subsequent taxonomy construction stages, we further evaluate the quality of the Extractor agents through a targeted human evaluation. We evaluate the event

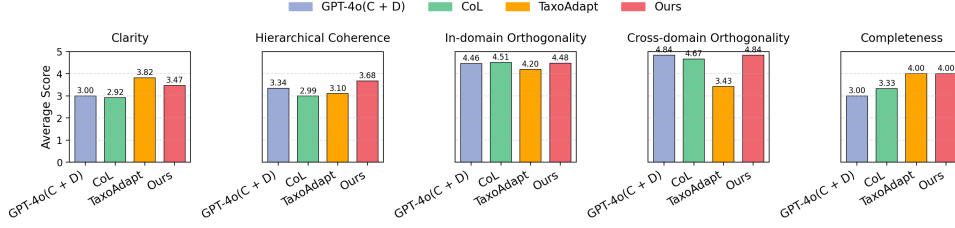


Figure 4: Average human evaluation scores of taxonomies generated by different methods across five dimensions.

Table 3: Human evaluation results for event extraction by different Extractor agents.

Extractor	Extracted	Correct	Missed	Precision	Recall
Qwen3	242	195	101	0.8058	0.6588
DeepSeek-V3	237	197	120	0.8312	0.6215

extraction results of Qwen3 and DeepSeek-V3 on 10 Classical Chinese historical passages. Two annotators independently judged whether each extracted event was correct and identified events missed by the Extractor agents. Due to the limited availability of large-scale annotated document-level event extraction datasets for Classical Chinese historical texts, we adopt human evaluation rather than directly comparing with existing document-level event extraction methods that require task-specific training data.

The two annotators achieved an overall agreement of 0.83 on event correctness. To provide a conservative evaluation, an extracted event is considered correct only when both annotators agree that it is valid; otherwise, it is counted as an extraction error. Table 3 summarizes the results. Both extractors reliably identify major historical events. Qwen3 achieves broader coverage with higher recall (0.6588), whereas DeepSeek-V3 adopts a more conservative extraction strategy, achieving higher precision (0.8312). Overall, the two extractors obtain comparable F1 scores.

We further analyze the missed and incorrect extractions. Most missed cases involve fine-grained sub-events within long sentences rather than completely overlooked events. The main sources of missed events include: (1) *granularity differences*, where models extract a main event while annotators identify additional fine-grained actions; (2) *multi-event sentence structures*, where multiple coordinated actions are present but models identify only the dominant event; and (3) *implicit event expressions*, which require contextual inference rather than explicit trigger words.

False positives mainly arise from overly broad

Table 4: Robustness analysis of Path Granularity evaluation using different LLM evaluators.

Methods	GPT-4o	GLM-5.2
GPT-4o (Corpus)	100.00	98.28
HiExpan (Corpus+Domain)	62.80	57.81
HiSetExpan (Corpus+Domain)	83.33	83.33
CoL (Corpus+Domain)	80.99	71.53
TaxoAdapt (Corpus+Domain)	70.98	70.62
CHisAgent (Corpus+Domain)	91.54	82.67

interpretations of frequent trigger words, such as 賜” (grant/bestow), 聞” (hear/learn of), 誅” (execute/punish), 遣” (send/dispatch), and “曰” (say/state), as well as from treating background or consequence descriptions as independent events. These errors highlight the challenges inherent in Classical Chinese historical texts, including ambiguous event boundaries, ellipsis, and complex narrative structures.

4.9 Robustness of LLM-based Evaluation

To examine whether the evaluation results are sensitive to the choice of LLM evaluator, we conduct an additional robustness analysis using GLM-5.2 ($T = 0$), which belongs to a different model family from the GPT models used in our original evaluation. We evaluate Path Granularity using the same evaluation protocol and prompts as in the original experiments. The results are shown in Table 4.

Although the absolute scores vary somewhat across evaluators, the relative performance patterns remain largely consistent. In particular, GPT-4o (Corpus) and CHisAgent achieve the highest scores, while HiExpan and TaxoAdapt obtain lower scores. This consistency across different model families suggests that the comparative conclusions are not strongly dependent on a specific LLM evaluator or its evaluation preferences.

This cross-family evaluation provides additional evidence that our evaluation findings are robust to evaluator choice, mitigating concerns about potential evaluator bias arising from using models from the same family.

4.10 Cross-Cultural Generalization across East Asian Historical Corpora

In this section, we evaluate whether the induced taxonomies generalize across historical corpora within the East Asian cultural sphere. We consider representative records from Japan, Korea, and Vietnam, including the *Nihon Shoki* (日本書紀)¹, the *Joseon Sillok* (朝鮮實錄)², and the *Dai Viet Su Ky Toan Thu* (大越史記全書)³.

For each corpus, we randomly sample several chapters and extract events using Qwen3 and DeepSeek-V3. As all corpora are written in classical Chinese, we assume comparable extraction quality across datasets. The resulting number of event types in each corpus is approximately 300–400. We report the Coverage Rate of each taxonomy on these corpora in Table 5. With the exception of CHED, all methods achieve their lowest coverage on the Vietnamese corpus, suggesting greater divergence from Chinese-centered taxonomies than the Korean and Japanese records.

Case Study. We further conduct a fine-grained analysis of tributary interactions as a representative diplomatic event type. Although tribute relations appear across East Asian polities, the induced taxonomy reveals systematic differences in their narrative organization. Specifically, as shown in Figure 2, tribute events are mapped to separate nodes for incoming tribute (來貢) and outgoing tribute (進貢), capturing different narrative perspectives.

We manually verify all extracted tribute-related event samples in each corpus. Then we observe that events from Chinese and Japanese corpora predominantly align with the incoming tribute (來貢) perspective, Vietnamese records more frequently align with outgoing tribute (進貢), and Korean records exhibit a mixed pattern. These findings demonstrate the importance of perspective-aware taxonomy nodes for cross-cultural comparison, allowing semantically similar practices to be aligned while preserving differences in historical narration.

4.11 Error Analysis

We conduct a quantitative error analysis based on the human evaluation results (Section 4.7). Among the sampled nodes, 32% received low scores (lower than 4) for hierarchical coherence, making structural inconsistency the dominant source of errors.

Table 5: Cross-corpus Coverage Rate of taxonomies generated by different methods.

Methods	Korea	Japan	Vietnam
CHED	43.73	53.90	50.42
CoL	29.74	29.96	18.78
TaxoAdapt	75.13	74.04	56.71
CHisAgent	75.51	80.50	60.55

This suggests that the most common error type is hierarchical over-segmentation, where the model introduces near-synonymous intermediate nodes without clear conceptual distinctions. Examples include “任用 (Appointment) → 任命 (Nomination)” and “辯解 (Justification) → 解釋 (Explanation)”. These errors suggest that the model may confuse lexical similarity with true hierarchical subsumption, leading to overly fine-grained or weakly differentiated taxonomic structures. Other error types are provided in Appendix L.

5 Conclusion

In this study, we propose CHisAgent, a multi-agent framework for constructing a large-scale taxonomy of ancient Chinese historical events. The taxonomy construction process is decomposed into three stages, Inducer, Expander, and Enricher, implemented by specialized LLM agents for hierarchical induction, semantic expansion, and taxonomy refinement. Experimental results demonstrate that our framework produces taxonomies with strong semantic clarity, hierarchical coherence, and broad coverage across multiple historical domains. We further show that the induced taxonomy generalizes to East Asian historical corpora. It supports cross-cultural alignment of semantically related historical events, while preserving differences in narrative perspective. Beyond taxonomy construction, our work highlights the potential of structured event taxonomies for historical knowledge organization and AI-assisted cultural understanding. In future work, we plan to integrate the taxonomy into a historical event retrieval system to facilitate more accurate event exploration and knowledge access. Our analysis also reveals several remaining challenges, including hierarchical over-segmentation, mixed ontological criteria, and cross-domain semantic overlap. Future work will explore ontology refinement under stronger logical constraints, temporal and causal event modeling, and validation across multilingual historical corpora.

Limitations

This work has several limitations.

¹https://en.wikipedia.org/wiki/Nihon_Shoki

²<https://hanchi.ihp.sinica.edu.tw/mqlc/hanjishilu>

³<https://ctext.org/wiki>

First, although we employ multiple LLMs to reduce the bias introduced by any single model, LLM-based taxonomy construction inevitably inherits biases from model pretraining. In particular, modern language models may project contemporary conceptual structures onto historical contexts or introduce hallucinated abstractions when expanding sparse historical evidence.

Second, the performance of current LLMs on Classical Chinese remains limited. Classical Chinese texts are highly concise, context-dependent, and often ambiguous, posing challenges for accurate semantic interpretation. As a result, the generated taxonomy may still contain errors or omissions, and human intervention is occasionally required to ensure correctness.

Third, while the *Twenty-Four Histories* constitute a highly authoritative and representative corpus of ancient Chinese historiography, they primarily reflect official and elite perspectives. Certain aspects of ancient Chinese culture, such as folk practices, regional traditions, and informal social dynamics, are underrepresented, limiting the completeness of the resulting taxonomy.

Fourth, our event-centered taxonomy design, while effective for modeling historical narratives and social processes, may not fully capture cultural phenomena that evolve gradually or are not naturally event-driven, such as philosophical traditions or long-term ideological shifts.

Fifth, our framework relies on LLM-based agents for hierarchical induction, expansion, and refinement, which introduces inherent stochasticity into the construction process. As a result, running the framework multiple times may yield slightly different taxonomy structures. While our experiments show that the overall high-level organization and major event categories remain stable across runs, variations may still occur in fine-grained node formation and hierarchical branching. A systematic study of structural variance and stability across multiple runs is left for future work.

Finally, although we employ both automatic metrics and human evaluation, these evaluation protocols have inherent limitations. Structural metrics cannot fully capture cultural appropriateness, and human evaluation is constrained in scale and may be influenced by annotators' backgrounds.

Ethics Statement

The human evaluations in this study were conducted by internal members of the research team (excluding the authors). All participants are research assistants and graduate students with expertise in the relevant domain. Participation was entirely voluntary, and all participants were informed of the study's objectives and the nature of the tasks. No personal, sensitive, or identifiable information was collected during the evaluation.

Beyond the human evaluation process, our work aims to facilitate the organization and exploration of historical knowledge through automated taxonomy construction. However, automatically generated taxonomies may amplify biases present in historical sources or models, oversimplify complex historical events, or be misused without appropriate expert verification. To mitigate these risks, we emphasize that the generated taxonomies are intended as research aids and should not be treated as authoritative historical interpretations or replacements for expert judgment.

Acknowledgments

This work was supported by the General Research Fund (GRF) of the Research Grants Council (RGC) of Hong Kong (Grant No. 15609725), the National Natural Science Foundation of China (Grant No. 72204258), and The Hong Kong Polytechnic University under the project "Evaluating the Syntax-Semantics Knowledge in Large Language Models" (Grant No. P0055270).

References

- Tanay Aggarwal, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2026. [Large language models for scholarly ontology generation: An extensive analysis in the engineering field](#). *Information Processing & Management*, 63(1):104262.
- Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. [CULEMO: Cultural lenses on emotion - benchmarking LLMs for cross-cultural emotion understanding](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.
- Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. [Semeval-2016 task 13: Taxonomy extraction evaluation \(texeval-2\)](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–11, San Diego, California. Association for Computational Linguistics.

- 2016), page 1081–1091, San Diego, California. Association for Computational Linguistics.
- Fan Bu, Zheng Wang, Siyi Wang, and Ziyao Liu. 2025. [An investigation into value misalignment in llm-generated texts for cultural heritage](#). *IEEE Transactions on Emerging Topics in Computational Intelligence*, page 1–15.
- Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. 2021. [Time-aware Ancient Chinese text translation and inference](#). In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 1–6, Online. Association for Computational Linguistics.
- Boqi Chen, Fandi Yi, and Dániel Varró. 2023. [Prompting or fine-tuning? a comparative study of large language models for taxonomy construction](#). In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, page 588–596.
- Catherine Chen, Kevin Lin, and Dan Klein. 2021. [Constructing taxonomies from pretrained language models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4687–4700, Online. Association for Computational Linguistics.
- Nico Franz and David Thau. 2010. [Biological taxonomy and ontology development: scope and limitations](#). *Nature Precedings*, page 1–1.
- Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. [Time travel: A comprehensive benchmark to evaluate llms on historical and cultural artifacts](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, page 23627–23641, Vienna, Austria. Association for Computational Linguistics.
- Zeinab Ghamlouch and Mehwish Alam. 2025. [Enriching Taxonomies using Large Language Models](#). In *Proceedings of 28th European Conference on Artificial Intelligence - ECAI 2025 (Demo Track)*, Bologna, Italy.
- Anna Goy, Diego Magro, and Marco Rovera. 2015. [Ontologies and historical archives: A way to tell new stories](#). *Applied Ontology*, 10(3–4):331–338.
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#). *arXiv preprint arXiv:2203.05794*.
- Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and R. M. del Rio-Chanona. 2024. [Large language models’ expert-level global history knowledge benchmark \(hist-llm\)](#). *Advances in Neural Information Processing Systems*, 37:32336–32369.
- Angelika Kaplan, Thomas Kühn, Sebastian Hahner, Niko Benkler, Jan Keim, Dominik Fuchß, Sophie Corallo, and Robert Heinrich. 2022. [Introducing an evaluation method for taxonomies](#). In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, EASE ’22*, page 311–316, New York, NY, USA. Association for Computing Machinery.
- Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. 2025. [Taxoadapt: Aligning llm-based multidimensional taxonomy construction to evolving research corpora](#). (arXiv:2506.10737). ArXiv:2506.10737 [cs].
- Avishek Lahiri, Yufang Hou, and Debarshi Kumar Sanyal. 2025. [TaxoAlign: Scholarly taxonomy generation using language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30191–30211, Suzhou, China. Association for Computational Linguistics.
- Dongha Lee, Jiaming Shen, Seongku Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. [Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters](#). In *Proceedings of the ACM Web Conference 2022, WWW ’22*, page 2819–2829, New York, NY, USA. Association for Computing Machinery.
- He Li, Linlin Zhu, Wang Shen, Xingye Du, Shuhe Guan, and Jun Deng. 2020. [Research on knowledge organization and visualization of historical events in the republic of china era](#). *Library Trends*, 69(1):138–163.
- Nan Li, Bo Kang, and Tijl De Bie. 2025. [Building data-driven occupation taxonomies: A bottom-up multi-stage approach via semantic clustering and multi-agent collaboration](#). (arXiv:2509.15786). ArXiv:2509.15786 [cs].
- Huitong Pan, Qi Zhang, Mustapha Adamu, Eduard Dragut, and Longin Jan Latecki. 2025. [Taxonomy-driven knowledge graph construction for domain-specific scientific applications](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, page 4295–4320, Vienna, Austria. Association for Computational Linguistics.
- Laura Pandolfo. 2015. [Stole: A reference ontology for historical research documents](#). In *DC@ AI* IA*, pages 13–18.
- Chao Shang, Sarthak Dash, Md. Faisal Mahbub Chowdhury, Nandana Mihindukulasooriya, and Alfio Gliozzo. 2020. [Taxonomy construction of unseen domains via graph-based cross-domain knowledge transfer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2198–2208, Online. Association for Computational Linguistics.
- Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. 2019. [Setexp: Corpus-based set expansion via context feature selection and rank ensemble](#). (arXiv:1910.08192). ArXiv:1910.08192 [cs.CL].

- Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T. Vanni, Brian M. Sadler, and Jiawei Han. 2018. [Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion](#). In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 2180–2189, New York, NY, USA. Association for Computing Machinery.
- Yanzhen Shen, Yu Zhang, Yunyi Zhang, and Jiawei Han. 2025. [A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3208–3220, Vienna, Austria. Association for Computational Linguistics.
- Jana Straková, Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. [Extending an event-type ontology: Adding verbs and classes using fine-tuned llms suggestions](#). In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, page 85–95, Toronto, Canada. Association for Computational Linguistics.
- Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, and Soujanya Poria. 2025. [Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 14199–14214, Vienna, Austria. Association for Computational Linguistics.
- Xuemei Tang, Zekun Deng, Qi Su, Hao Yang, and Jun Wang. 2024. [Chisiec: An information extraction corpus for ancient chinese history](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 3192–3202, Torino, Italia. ELRA and ICCL.
- Xuemei Tang, Zekun Deng, Jun Wang, and Qi Su. 2025. [Chisnere: a premodern chinese corpus with named entity and relation annotation](#). *Digital Scholarship in the Humanities*, 40(2):617–638.
- Xuemei Tang, Linxu Wang, and Jun Wang. 2026. [Language model collaboration for relation extraction from classical chinese historical documents](#). *Information Processing & Management*, 63(1):104286.
- Norawit Urailetprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. [Sea-vqa: Southeast asian cultural context dataset for visual question answering](#). In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, page 173–185, Bangkok, Thailand. Association for Computational Linguistics.
- Linxu Wang, Jun Wang, and Tong Wei. 2024. [Using ontology to model time description in historical chinese texts](#). *Digital Scholarship in the Humanities*, 39(1):423–429.
- Congcong Wei, Zhenbing Feng, Shutan Huang, Wei Li, and Yanqiu Shao. 2023. [CHED: A cross-historical dataset with a logical event schema for classical Chinese event detection](#). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 875–888, Harbin, China. Chinese Information Processing Society of China.
- Pascal Wulfschleger, Majid Zarharan, Donnacha Daly, Marc Pouly, and Jennifer Foster. 2025. [No gold standard, no problem: Reference-free evaluation of taxonomies](#). (arXiv:2505.11470). ArXiv:2505.11470 [cs].
- Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. [Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples](#). (arXiv:2402.07386). ArXiv:2402.07386 [cs].
- Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. 2025. [Code-Taxo: Enhancing taxonomy expansion with limited examples via code language prompts](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4131–4144, Vienna, Austria. Association for Computational Linguistics.
- Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2025. [LLMTaxo: Leveraging large language models for constructing taxonomy of factual claims from social media](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19627–19641, Vienna, Austria. Association for Computational Linguistics.
- Ziyi Zhang. 2026. [Conceptual model construction for the a comprehensive collection of ancient chinese paintings based on metadata and ontology](#). *npj Heritage Science*, 14(1):63.

A Dataset

We select the version of *Twenty-Four Histories* are published by Zhonghua Book Company (中華書局). Table 6 reports the character counts of sampled chapters from each book in the *Twenty-Four Histories*, which serve as a proxy for corpus size across historical periods. We open all data at <https://anonymous.4open.science/r/ACL-BC1E/>.

B Extractor Evaluation

We randomly sample 400 event instances from the outputs of the extractors and manually evaluate the extraction performance. The results show that the extraction achieves an accuracy of 85%.

The most common errors include non-event summaries (e.g., entity or identity descriptions), behavioral attribute descriptions, and cases where multiple events are merged into a single expression (e.g., “A and B” format), which is not suitable for higher-level event abstraction. Overall, factual extraction errors account for less than 5% of the total samples.

C Domain Description

- **Politics:** Encompasses events related to state governance, power structures, policy formulation and implementation, as well as judicial and legal affairs.
- **Military:** Encompasses events related to warfare, military operations, strategic planning, and army organization and management.
- **Diplomacy:** Includes inter-state political and cultural interactions conducted through official channels, such as envoy missions, tribute relations, and peace negotiations.
- **Society:** Covers social activities, cultural practices, customs, and collective events involving groups, organizations, or institutions beyond the individual level.
- **Ritual:** Refers to ceremonial and ritual activities conducted by ancient states or royal families.
- **Economy–Livelihood:** Covers events related to economic production, resource distribution, daily subsistence, and state finance.
- **Nature:** Includes natural phenomena, related events, and their impacts on human society.

- **Individual:** Covers events related to individual life trajectories, identity attributes, personal behaviors, and social relationships.

Table 6: Character statistics of sampled chapters from the *Twenty-Four Histories*.

Historical Book	Characters
<i>Records of the Grand Historian</i> (史記)	39,267
<i>Book of Han</i> (漢書)	37,100
<i>Book of Later Han</i> (後漢書)	26,039
<i>Records of the Three Kingdoms</i> (三國志)	14,170
<i>Book of Jin</i> (晉書)	32,849
<i>Book of Song</i> (宋書)	46,604
<i>Book of Southern Qi</i> (南齊書)	25,738
<i>Book of Liang</i> (梁書)	37,690
<i>Book of Chen</i> (陳書)	20,338
<i>Book of Wei</i> (魏書)	29,581
<i>Book of Northern Qi</i> (北齊書)	11,885
<i>Book of Zhou</i> (周書)	41,639
<i>Book of Sui</i> (隋書)	31,472
<i>History of the Southern Dynasties</i> (南史)	9146
<i>History of the Northern Dynasties</i> (北史)	11311
<i>Old Book of Tang</i> (舊唐書)	32,175
<i>New Book of Tang</i> (新唐書)	19,132
<i>Old History of the Five Dynasties</i> (舊五代史)	14,405
<i>New History of the Five Dynasties</i> (新五代史)	10,845
<i>History of Song</i> (宋史)	17,861
<i>History of Liao</i> (遼史)	14,869
<i>History of Jin</i> (金史)	23,591
<i>History of Yuan</i> (元史)	18,297
<i>History of Ming</i> (明史)	18,092
Total	584,096

D Event Type Distribution Across Domains

We report the event type distribution across domains in Table 7.

Table 7: Distribution of extracted event types across different domains.

Domain	Number of Event Types
Politics	1681
Military	690
Diplomacy	168
Society	710
Ritual	202
Economy–Livelihood	283
Nature	119
Individual	1023

E Clustering Topics

Following prior work by Tang et al. (2026), which manually constructs a relation ontology for classical Chinese documents using document-derived topics as the organizational basis, we leverage topic clustering from the *Twenty-Four Histories* to support taxonomy construction.

Table 8: Clustered historical topics extracted from the *Twenty-Four Histories*.

Topic ID	Topic Label
0	戰爭行動 (Military Actions)
1	節令時序 (Seasonal Calendar)
2	政務管理 (Administrative Governance)
3	詔令文書 (Imperial Decrees)
4	晉升調任 (Promotion and Transfer)
5	天文觀測 (Astronomical Observation)
6	皇室活動 (Royal Activities)
7	外交朝貢 (Diplomacy and Tribute)
8	地理 (Geography)
9	紀年 (Chronology)
10	自然災害 (Natural Disasters)
11	科舉考試 (Imperial Examinations)
12	言行紀錄 (Speech and Deeds)
13	登基繼位 (Accession and Enthronement)
14	年號 (Reign Titles)
15	音樂樂器 (Music and Instruments)
16	干支 (Ganzhi Cycles)
17	赦免公告 (Amnesty Decrees)
18	祭祀禮儀 (Rituals and Sacrifices)
19	帝王稱謂 (Imperial Titles)
20	邊疆防務 (Frontier Defense)
21	官員任命 (Official Appointments)
22	巡察 (Inspection)

We first segment all documents in the corpus using an ancient Chinese word segmenter⁴. Each paragraph is then treated as a basic textual unit and encoded with BERTopic. Based on multiple experimental trials and manual inspection, we set the number of topics to 30. For each topic, we extract the top 30 representative words. These words are subsequently summarized into a single concept word or phrase using GPT-4o, yielding a compact semantic label for each topic. Finally, we apply human validation to ensure that each topic label is semantically coherent and appropriate for the historical context, resulting in 23 retained topics (Table 8).

F API Version and Cost

We use the following model versions in our experiments: GPT-4o (gpt-4o-2024-11-20), DeepSeek-V3 (DeepSeek-V3-0324), Qwen3 (qwen-plus-2025-07-28), and GPT-5 (gpt-5-2025-08-07). For all models except GPT-5, the temperature is set to 0.

The majority of token consumption is attributed to Qwen3, DeepSeek-V3, and GPT-4o, which are primarily used in the extractor stage for processing historical corpora and generating event sentences, event types, and trigger words.

⁴<https://github.com/tangxuemei1995/Ancient-chinese-segmenter>

In contrast, GPT-5 is mainly used in the Expander and Enricher stages, which involve significantly fewer tokens due to their more structured generation process.

The total API cost of the entire pipeline is approximately \$100.

G Baselines

We compare our framework with representative taxonomy construction methods spanning traditional non-agentic approaches, single-agent LLM prompting settings, multi-agent frameworks, and human-crafted taxonomies.

Traditional Non-Agentic Methods. We reproduce two representative taxonomy induction methods: HiExpan (Shen et al., 2018) and HiSetExpan (Shen et al., 2018).

For HiExpan, we construct domain-specific seed taxonomies from our taxonomy and iteratively expand taxonomy hierarchies through both depth expansion and width expansion. Width expansion enlarges sibling sets using skip-patterns, embedding similarity, and sibling consistency, while depth expansion discovers child concepts under existing taxonomy nodes. Compared with the original setting, we adapt the method to Classical Chinese corpora and replace REPEL with BGE-M3 embeddings.

HiSetExpan is a baseline method proposed by Shen et al. (2018). HSetExpan differs from HiExpan in that it iteratively applies SetExpan (Shen et al., 2019) at each taxonomy level. For each lower-level node, the method attaches it to the parent node with the highest parent-child similarity score.

Single-Agent LLM Methods. We design three prompting-based taxonomy construction settings using GPT-4o, DeepSeek-V3, and Qwen3.

- **LLM + Domains.** The model is given only predefined domains and generates the taxonomy in a top-down manner by progressively producing fine-grained event types.
- **LLM + Event Corpus.** The model is provided with fine-grained event types extracted from the corpus and induces the taxonomy bottom-up through iterative clustering and abstraction.
- **LLM + Event Corpus + Domains.** The model is provided with both predefined domains and event-type corpus, and induces in-

intermediate taxonomy levels by integrating domain knowledge with event types.

Multi-Agent Methods. We further compare with two recent multi-agent taxonomy construction frameworks.

Chain-of-Layer (Zeng et al., 2024) constructs taxonomies in a top-down manner through iterative layer-wise entity selection. Following the original framework, we first build a candidate event-type pool from automatically generated event inventories, then iteratively generate and select intermediate taxonomy nodes layer by layer using GPT-4o.

TaxoAdapt (Kargupta et al., 2025) adaptively expands taxonomy width and depth according to the distribution of event types assigned to each node. Following the original setting, we use predefined domains and event-type inventories extracted from the corpus. GPT-4o first generates pseudo taxonomy labels, after which event types are assigned through classification-based matching. Width or depth expansion is then dynamically triggered according to node coverage.

Human-Crafted Taxonomy. We include the CHED taxonomy (Wei et al., 2023) as a human-crafted reference taxonomy. Curated from the *Twenty-Four Histories* for historical event extraction, it provides a coarse-grained manually designed event hierarchy.

H Metrics

Path Granularity. We use GPT-4o to score each parent-child pair, assigning 1 if the child is more specific than the parent and 0 otherwise. Finally, we compute the proportion of pairs that satisfy this condition.

CSC. CSC is a structure-level metric that quantifies the alignment between the taxonomy’s hierarchical structure and the semantic similarity of concepts. Specifically, it measures the extent to which concept pairs that are farther apart in the taxonomy are also semantically less similar, while closer nodes exhibit higher semantic similarity.

$$W_{n_a, n_b} = \frac{2 \cdot \text{lca}(p(n_a), p(n_b))}{|p(n_a)| + |p(n_b)|} \quad (1)$$

where $p(n_a)$ denote the path from the root concept to a target concept n_a . Furthermore, let $\text{lca}(p(n_a), p(n_b))$ denotes the depth of the least common ancestor of the paths $p(n_a)$ and $p(n_b)$.

$$\text{CSC} := \tau(W_{n_a n_b}, S_{n_a n_b}) \quad (2)$$

where τ denotes Kendall rank correlation, and $S_{n_a n_b}$ represents the semantic similarity between concept representation. We encode each node using text-embedding-3-small and compute cosine similarity to obtain S .

Coverage Rate. This metric evaluates how well the generated taxonomy can organize knowledge from the *Twenty-Four Histories*. We sample two chapters from each book (excluding those used during experiments), and use Qwen3, DeepSeek-V3, and GPT-4o to extract event samples (text + event type). After deduplication, we compute the cosine similarity between each event type and every node in the taxonomy. If the similarity is above 0.6, we consider that event type covered. The final coverage rate is the proportion of event types that are covered. All nodes are encoded using text-embedding-3-small.

Node Recall. This metric measures how many nodes in the human taxonomy ($\mathcal{T}$) are recalled by the generated taxonomy ($\mathcal{C}$). We encode each node in both taxonomies using text-embedding-3-small and compute cosine similarity between node pairs. If the similarity exceeds 0.6, we consider the human node recalled.

Novelty. Novelty quantifies the proportion of completely new classes or categories in a taxonomy that are not present in any previous taxonomies. For a generated taxonomy ($\mathcal{C}$) and a human taxonomy ($\mathcal{T}$), it is defined as:

$$\text{Novelty}(\mathcal{C}, \mathcal{T}) = \frac{\sum_{c \in \mathcal{C}} \text{new}(\mathcal{C}, \mathcal{T}, c)}{|\mathcal{C}|} \in [0, 1] \quad (3)$$

Here, $\text{new}(\mathcal{C}, \mathcal{T}, c) = 1$ if class (c) is entirely new compared to taxonomy ($\mathcal{T}$), and 0 otherwise. A higher value indicates that the taxonomy introduces more novel information.

To determine whether a class c is new relative to the human taxonomy, we use the same cosine-similarity procedure as in Node Recall. If a node in the generated taxonomy cannot find any human node with similarity greater than 0.6, we consider it a novel node.

Significance. We measure the significance of a taxonomy by its ability to provide fine-grained categorization of events. To compare a generated taxonomy ($\mathcal{C}$) with a human-curated taxonomy ($\mathcal{T}$), we adopt classification delta, defined as:

$$\text{class. delta}(\mathcal{C}, \mathcal{T}, R) = \frac{|\sim_{\mathcal{C}}| - |\sim_{\mathcal{T}}|}{|R|} \quad (4)$$

where (R) is the set of event types extracted by LLMs during the Coverage Rate evaluation and successfully matched by both taxonomies.

Here, $(|\sim_{\mathcal{C}}|)$ denotes the number of distinct taxonomy nodes in $(\mathcal{C})$ that events in $(\mathcal{R})$ are assigned to, $(|\sim_{\mathcal{T}}|)$ denotes the corresponding number of distinct nodes in the human taxonomy $(\mathcal{T})$.

A positive classification delta indicates that events in (R) are distributed across more nodes in $(\mathcal{C})$ than in $(\mathcal{T})$, implying that $(\mathcal{C})$ provides a more fine-grained categorization. A delta of zero indicates that $(\mathcal{C})$ matches the granularity of the human taxonomy, while a negative value suggests that $(\mathcal{C})$ is coarser and assigns events to fewer categories.

I Structural Statistics of Taxonomies

The statistics of the taxonomies generated by different method are reported in Table 9. We report the maximum depth, average depth, and the average number of branches per parent node. We find that taxonomies generated by LLMs in a single run, namely Qwen3, DeepSeek-V3, and GPT-4o, tend to exhibit relatively simple structures, characterized by shallower depths. In contrast, the human-crafted taxonomy and the multi-agent method produce structures that are generally deeper and more hierarchical.

J Additional Domain Analysis

Figure 5(a) presents parent-child relationship rationality. Our method and GPT-4o consistently achieve relatively high scores across all domains, whereas CoL and TaxoAdapt show larger performance fluctuations, particularly in Ritual and Diplomacy. Figure 5(b) illustrates the distribution of recalled nodes. Different methods exhibit distinct preferences: our method mainly recalls nodes from Politics and Military; TaxoAdapt focuses more on Military, Society, Nature, and Individual; and CoL concentrates on Military, Society, and Diplomacy.

K Human Evaluation

Node Clarity. For each taxonomy, we randomly sample 50 nodes and ask evaluators to judge whether the node’s semantics are clear and whether the node can appropriately represent an event type in ancient Chinese cultural system. Scores range from 0 to 5.

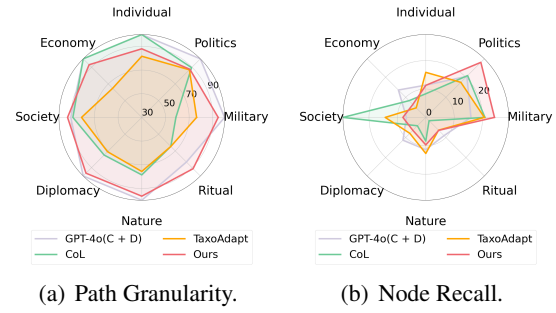


Figure 5: Radar chart of evaluation metrics of different methods across various domains. GPT-4o (C + D) denotes GPT-4o (Corpus + Domain), and “Economy” represents “Economy-Livelihood”.

Hierarchical Coherence. For each taxonomy, we randomly sample 50 nodes together with their full paths, and evaluators assess whether the hierarchical structure from the root to the leaf is logically coherent. Scores range from 0 to 5.

In-domain Orthogonality. For each taxonomy, we randomly sample 50 nodes and retrieve their top 5 most similar nodes within the same domain. Evaluators then assess whether the semantics of the sampled node overlap with those similar nodes. If no semantic overlap exists, the node receives a score of 5; if there is overlap with one node, the score is 4; and so on.

Cross-domain Orthogonality. The procedure is similar to in-domain orthogonality, except that the top 5 similar nodes are retrieved from other domains. Evaluators judge whether cross-domain semantic overlap exists.

Completeness. Evaluators assess whether a taxonomy sufficiently covers the conceptual space of Chinese historical events, considering its overall breadth, depth, and coverage of important event types and conceptual categories. The evaluation is conducted at the taxonomy level to better reflect the practical usefulness and comprehensiveness of the constructed taxonomy. Scores range from 0 to 5.

All human evaluation guidelines are presented in Figure 8.

L Error Analysis

We conduct a quantitative error analysis based on the human evaluation results (Section 4.7). Specifically, we analyze the proportion of nodes that received low ratings across different evaluation dimensions:

- **Node Clarity:** 1/50 sampled nodes received a

Table 9: Structural statistics of taxonomies generated by different methods.

Models	Strategies	Nodes	Leaves	Max. Dep.	Avg. Dep.	Branch
Qwen3	Domain	112	72	3	3	3
DeepSeek-V3		103	63	3	3	2.72
GPT-4o		112	72	3	3	3
Qwen3	Corpus	113	72	3	3	3
DeepSeek-V3		74	53	3	3	4
GPT-4o		60	40	4	4	3.05
Qwen3	Corpus + Domain	96	56	3	3	2.5
DeepSeek-V3		214	180	3	3	7.62
GPT-4o		96	58	3	3	2.67
HiExpan	Corpus + Domain	208	157	6	4.52	4.65
HiSetExpan	Corpus + Domain	118	86	4	3.81	4.58
CoL	Domain	540	440	6	3.74	5.75
	Corpus + Domain	160	108	6	3.67	3.27
TaxoAdapt	Corpus + Domain	1109	860	6	4.93	4.59
CHED	Human	100	67	5	2.73	3.52
CHisAgent	Corpus + Domain	483	283	6	3.93	2.43

score lower than 4.

- **Hierarchical Coherence:** 16/50 sampled nodes received a score lower than 4.
- **In-domain Orthogonality:** 10/50 sampled nodes were judged to exhibit semantic redundancy (lower than 5) with other nodes in the same domain.
- **Cross-domain Orthogonality:** 4/50 sampled nodes exhibited semantic overlap (lower than 5) across domains.

Completeness evaluates the overall coverage of the taxonomy rather than individual nodes and therefore is not suitable for node-level proportion analysis.

These results indicate that while our method maintains strong semantic clarity, the primary limitations lie in hierarchical organization and intra-domain redundancy.

To further illustrate these error types, we provide representative examples below:

L.1 Hierarchical Coherence Errors

The main limitation is over-segmentation and redundant intermediate layers.

In some cases, the model introduces overly fine-grained or semantically redundant intermediate nodes, resulting in weak hierarchical justification. For example:

- 政治 (Politics) → 人事任免 (Personnel Appointment) → 任用 (Appointment) → 任命 (Nomination) → 職務調整 (Position Adjustment)

- 政治 (Politics) → 政務與決策 (Governance and Decision-making) → 決策裁定 (Decision Ruling) → 決策 (Decision)
- 個人 (Individual) → 行為活動 (Behavioral Activities) → 辯解 (Justification) → 解釋 (Explanation)
- 個人 (Individual) → 行為活動 (Behavioral Activities) → 允諾 (Promise) → 同意 (Agreement)

In these examples, parent and child nodes are nearly synonymous, and the additional layers do not introduce meaningful semantic distinctions. This suggests that the model occasionally over-expands the hierarchy without sufficient semantic differentiation.

L.2 In-domain Orthogonality Errors

To further analyze semantic redundancy, we examined sampled nodes flagged during in-domain orthogonality evaluation. Among the 10 sampled cases with semantic overlap, only one node received a low final score (2/5), while most cases received 4/5, indicating mild rather than severe redundancy.

Several overlapping cases occur within ritual-related categories, such as:

- 祭祀与皇室活動 (Ritual and Imperial Activities)
- 祭祀活動實施 (Ritual Implementation)
- 定期祭祀 (Regular Ritual)
- 祈禳祭祀 (Supplicatory Ritual)
- 祭祀供品与物品 (Ritual Offerings)

- 儀式執行 (Ceremonial Execution)

These nodes differ along multiple dimensions (function, frequency, material components, and execution process). However, the taxonomy occasionally mixes these classification criteria, leading to *partial semantic overlap*. This suggests that the model does not always consistently align subtype expansion along a single ontological dimension.

L.3 Cross-domain Orthogonality Errors

We further examine cross-domain semantic overlap cases. Among the sampled nodes, only one case received a low final score (2/5), while most received 4/5, indicating that severe cross-domain redundancy is rare.

We identify a recurring pattern: institutional abstraction across domains.

Certain nodes involve institutional or regulatory structures that naturally exist across multiple domains. For example:

- 外交 (Diplomacy) → 報告制度 (Reporting Policies)
- 政治 (Politics) → 制度与政策 (Institutional Reform / Policies)

Although categorized under different domains, these nodes share a higher-level abstraction (“institutional mechanism”), which leads to partial semantic proximity. This reflects the inherent cross-domain nature of institutional governance structures.

M Case Study

In this section, we add detailed running examples to illustrate the role of each stage, Inducer, Expander, and Enricher, in a step-by-step manner, as described below.

As shown in Figure 6, induced taxonomy (after Inducer) are at left. In the induced taxonomy, all tribute-related actions were placed at the same level. This resulted in a mixture of: relational states (e.g., Restoration of Tribute Relations, Termination of Tribute Relations), operational actions (e.g., Envoy Mission to the Court for Tribute), and ritual gift exchanges (e.g., Presentation of Tribute Goods). After applying the Expander: The Expander introduces three intermediate categories: 朝貢關係 (Tributary Relations) (relational dimension), 使節派遣 (Envoy Dispatch) (operational dimension),

外交禮物(Diplomatic Gifts) (ritual/material exchange dimension). This restructuring improves sibling-level semantic coherence and clarifies the institutional logic of tribute diplomacy. The example demonstrates that the Expander does not merely add nodes, but reorganizes heterogeneous siblings into conceptually consistent substructures.

We further illustrate structural refinement in the “Individual – Emotional” domain in Figure 6. Here, emotional states of different polarity were placed at the same level, resulting in a flat and heterogeneous structure.

After applying the Expander, the Expander introduces a polarity-based abstraction (Positive Emotions vs. Negative Emotions), enforcing conceptual consistency between parent and child nodes. This reduces semantic heterogeneity and enhances structural clarity.

To illustrate the Enricher stage, we provide examples where important event types were absent from the induced taxonomy and were subsequently introduced through external knowledge and corpus signals. As shown in Figure 7, For the case 1, the initial taxonomy lacked categories capturing intellectual conflicts. However, historical records frequently public denunciations, doctrinal exclusion, suppression of teaching, factional academic disputes. The Enricher identifies and incorporates these discourse-level events, adding 學術攻擊 (Academic Attacks) and 排斥學說(Doctrinal Exclusion) under the cultural domain. These nodes capture dynamics of intellectual power and doctrinal boundary-making that were previously unrepresented.

In the second case, the initial taxonomy did not include any medical-related events. However, historical documents frequently describe actions such as diagnosis, prescribing medicine, and treatment outcomes. The Enricher introduces “醫療(Medical)-救治 (Medical Treatment and Care)” to improve coverage of everyday social practices.

This example demonstrates that the Enricher performs semantic coverage expansion rather than structural reorganization.

N Taxonomy Example

Figure 9 illustrates representative portions of the core taxonomy for the diplomacy and society domains.

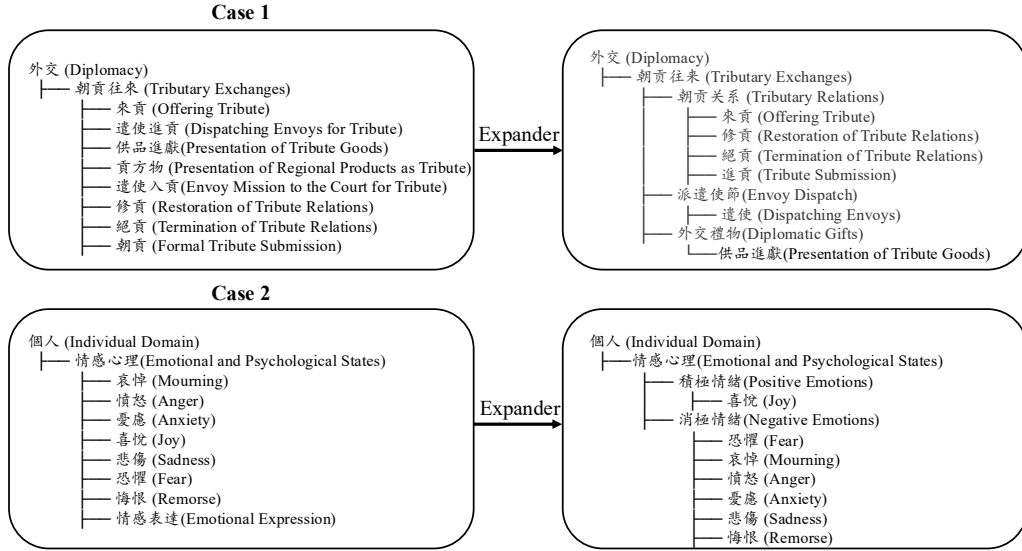


Figure 6: Expander examples in Diplomacy and Individual domains.



Figure 7: Enricher examples in Society domains.

O Sensitivity to Similarity Thresholds

To examine the sensitivity of the proposed methods to the similarity threshold, we evaluate all methods under five threshold settings, ranging from 0.5 to 0.7. The results are reported in Table 10. Although the absolute values of Coverage, Node Recall, and Novelty vary with different thresholds, the relative performance trends across methods remain largely consistent. In particular, CHisAgent and TaxoAdapt consistently outperform CoL in terms of Coverage and Node Recall across different threshold settings, while achieving a favorable balance with Novelty. These results demonstrate that the comparative performance of the methods is robust to the choice of similarity threshold. Based on this analysis, we retain 0.6 as the default threshold for our experiments.

Table 10: Performance comparison under different similarity thresholds.

Methods	Threshold	Coverage	Node Recall	Novelty
CoL	0.50	79.48	67.78	36.81
	0.55	58.12	55.11	43.75
	0.60	43.72	46.67	50.69
	0.65	34.84	28.89	70.83
	0.70	19.10	26.67	81.94
TaxoAdapt	0.50	98.66	85.56	44.69
	0.55	80.15	81.11	61.04
	0.60	74.17	60.00	69.21
	0.65	65.58	57.78	85.10
	0.70	51.76	44.44	92.55
CHisAgent	0.50	97.24	91.11	55.63
	0.55	88.11	73.33	68.05
	0.60	75.13	68.89	77.70
	0.65	66.16	45.56	87.82
	0.70	40.03	36.67	91.03

P Human Evaluation of Event Classification

To independently assess the reliability of the Classifier agent, we conducted a human evaluation of its domain assignments. We randomly sampled **500 events** classified by GPT-4o and asked two annotators with expertise in Chinese historical texts to assign each event to one or more of eight predefined domains: Military (事), Politics (政治), Diplomacy (外交), Society (社), Individual (人), Economy and Livelihood (民生), Nature (自然), and Ritual (祭祀). Annotators were allowed to assign multiple domains when an event involved multiple aspects.

The two annotators achieved substantial agree-

1. Clarity Evaluation Criteria

Overall goal:

Assess whether each taxonomy node label is clear, accurate, unambiguous, and easy to understand, and whether it is appropriate within the current domain.

Scoring Criteria:

0 – Unclear / Incorrect

Label and definition are confusing, contradictory, or meaningless; the concept cannot be identified.

1 – Heavily Context-Dependent

Label is unclear and the definition is incomplete or vague; understanding requires strong reliance on context or prior domain knowledge.

2 – Partially Clear but Ambiguous

General meaning can be inferred, but boundaries are unclear; multiple interpretations are possible.

3 – Clear but Misclassified

Label and definition are clear and unambiguous, but the node does not fit the current domain or taxonomy.

4 – Clear but Suboptimal Fit

Label and definition are clear and correct, but the domain fit is not ideal (e.g., scope slightly too broad or narrow).

5 – Clear, Precise, and Well-Fitted

Label is concise and accurate; definition is complete and unambiguous; the node fits the domain perfectly.

2. Hierarchical Coherence

Overall goal:

Assess whether the taxonomy structure follows a clear coarse-to-fine hierarchy, with semantically consistent and well-formed parent-child relationships.

Scoring Criteria:

0: Hierarchy is chaotic; parent-child relations are confused or unusable.

1: Severe hierarchical errors or abrupt level jumps.

2: Multiple nodes are placed at inappropriate hierarchical levels.

3: Generally reasonable structure with minor hierarchical issues.

4: Parent-child relations are mostly coherent and clearly organized.

5: Hierarchy is fully coherent, precise, and strictly semantically aligned.

3. In-domain Orthogonality

Overall goal:

Assess whether nodes within the same domain are conceptually independent and have clear boundaries.

Evaluation Method:

Reviewers assess overlap by asking: *Does node X overlap with other nodes (A–E)?*

Semantic overlap is judged by pairwise comparison between X and each of A–E.

If node X already exists, determine whether nodes A, B, C, D or E are still necessary.

A parent-child relationship is **not** considered overlap.

Scoring Criteria:

1: Complete overlap

0: No overlap

4. Cross-domain Orthogonality

Overall goal:

Assess whether nodes from different domains overlap conceptually, i.e., whether nodes in different domains express similar or identical meanings.

Evaluation Method:

Reviewers assess overlap by asking: *Does node X overlap with nodes in other domains (A–E)?*

Semantic overlap is judged by pairwise comparison between X and each of A–E.

If node X already exists, determine whether nodes A, B, C, D or E are still necessary.

Unlike in-domain evaluation, a parent-child relationship is **considered overlap**, making this criterion stricter.

Scoring Criteria:

1: Complete overlap

0: No overlap

5. Completeness

Overall goal:

Assess whether the taxonomy sufficiently covers the major aspects of the domain and maintains a balanced structure.

Scoring Criteria:

0: Extremely poor coverage; many key topics are missing.

1: Major thematic areas are severely missing.

2: Some obvious gaps or clear imbalance across branches.

3: Good coverage with minor gaps or slight imbalance.

4: Comprehensive coverage with a largely balanced structure.

5: Fully comprehensive coverage, appropriate depth, and highly balanced branches.

Figure 8: Guidelines for Human Evaluation.

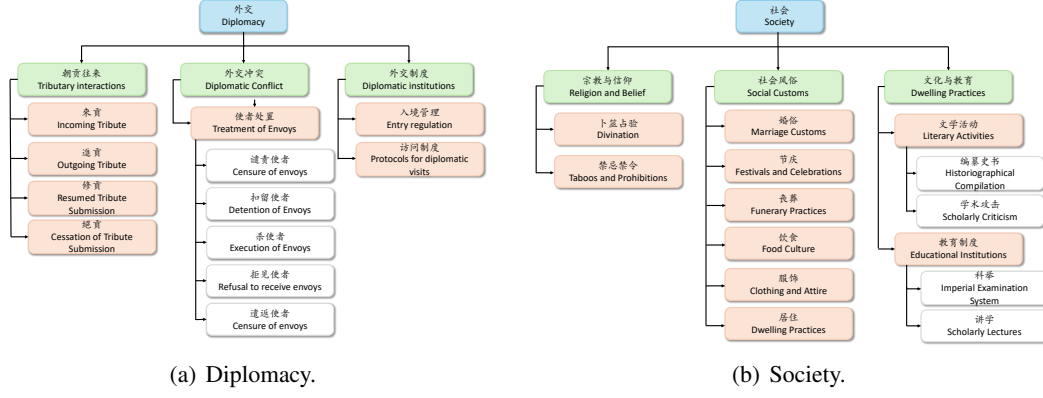


Figure 9: Taxonomy (partial) Example for Diplomacy and Society Domain.

ment despite the inherent ambiguity of historical domain classification. The exact label-set agreement was 59.6%, with a mean Jaccard similarity of 0.71 and a macro-averaged Cohen’s κ of 0.63. The per-domain κ values ranged from 0.30 for Society to 0.81 for Military, indicating reasonable consistency across most domains.

We constructed a human consensus gold standard using the following procedure. When both annotators agreed on the primary domain, that domain was adopted as the consensus label. When their label sets had exactly one overlapping domain, the overlapping domain was adopted. Cases without a clear consensus ($N = 86$, 17.2%) were marked as disputed and excluded from strict evaluation, resulting in 414 events with reliable consensus labels.

Compared with this human-annotated gold standard, the Classifier agent achieved a strict accuracy of 68.4% (283/414) on the consensus subset and a weighted F1 of 0.71. We additionally report a lenient accuracy of 79.6% (398/500), which considers a prediction correct when the predicted domain falls within the union of domains assigned by the two annotators. This metric accounts for legitimate cross-domain interpretations and shows that many predictions remain within the range of human-acceptable classifications.

Error analysis indicates that most classification errors occur at the boundary between Politics (政治) and Individual (人). Examples include “民” (showing benevolence toward the people), “毒酒” (granting poisoned wine), and “科及第” (passing the imperial civil service examination). Such cases reflect genuine ambiguity between state-level affairs and individual actions in historical narratives, rather than straightforward classification failures.

These results provide an independent human-verified assessment of the Classifier agent. GPT-4o is used as the operational classifier in our pipeline, while the human evaluation serves as an external assessment of its classification reliability.

Q Prompts

All prompts used in the experiments are shown in Figures 10–17. The original prompts are translated into English to facilitate understanding.

Prompt Template for Extractor

You are a researcher specializing in information extraction from Classical Chinese texts and are constructing a historical event ontology based on the *Twenty-Four Histories*.

You will be given a passage written in Classical Chinese, taken from the *Twenty-Four Histories*. Your task is to segment the text according to its event-level semantic structure, extract semantically complete historical events or sub-events, and output the results in **JSON format**.

1. Event Definition

An **event** refers to an action, state, or change described in the text. It is typically associated with a **trigger** (the core verb or keyword) and may involve multiple **arguments** or attributes, such as time, location, persons, objects, purpose, or outcome.

A complete event does not need to include all possible elements, but it must contain at least a clear action verb or state.

When segmenting the text, please base your decisions on **semantic units**, rather than relying solely on punctuation or sentence length.

2. Output Requirements

Please return the results as a **JSON array** in the following format:

```
[
  {
    "id": 1,
    "text": "original Classical Chinese sentence",
    "event type": "a concise summary of the event type (e.g., accession, campaign,
    enfeoffment, migration)",
    "trigger": "the core verb or keyword that triggers the event"
  }
]
```

•] **text**: Preserve the original Classical Chinese sentence exactly; do not translate or rewrite it.

•**event_type**: Summarize the event type using a **single noun or verb phrase** that directly describes the event; do not use adjectives.

•**trigger**: Explicitly identify the verb or keyword that triggers the event, extracted directly from the original text.

Now, please process the following content and output the results in the format specified above:

Figure 10: Prompt 1 – Extractor Prompt Example

Prompt Template for Classifier

Please assign the following event-type terms extracted by a large language model (LLM) from the *Twenty-Four Histories* to the most appropriate domain(s).

A term may belong to multiple domains.

The domains are:

{domain list}

Output **only** the assignment results. Ensure that **every term is assigned**. Do **not** add any explanations.

Return the result in **JSON format**, as shown below:

```
{
  "Military": [...],
  "Economy & Livelihood": [...],
  "Nature": [...],
  "Individual": [...],
  "Ritual": [...],
  "Politics": [...],
  "Diplomacy": [...],
  "Society": [...]
}
```

Now, please analyze and assign the following terms:

Figure 11: Prompt 2 – Classifier Prompt Example

Prompt Template for Generator

You are a taxonomy construction expert participating in the development of subcategories under the “{domain} Events” category.

The current upper-level path is:

“Event” → “{domain} Event”

At present, no subcategories have been defined under this category.

Based on the fine-grained event type list provided below, determine whether these event types can be covered by one or more **abstract sub-event types**. If so, please do the following for each sub-event type:

• **Assign an appropriate category name:**

- Use a **single word only** (do not use multi-word expressions, phrases, or combined forms such as “X and Y”);
- Prefer the most concise term possible;
- You may select the most representative term from the fine-grained event types;
- The category name **must not duplicate the domain name** (e.g., if the domain is “Ritual,” the sub-event type cannot also be named “Ritual”).

• **Assign the fine-grained event types** to the corresponding sub-event type(s).

• **Provide a concise and clear definition** that accurately characterizes the events covered by each sub-event type.

Please output the results strictly in the following **JSON format**:

```
{  
  "sub-event type1:XXX": ["fine-grained event type 1", "fine-grained event type 2"],  
  "sub-event type1:XXX_definition": "",  
  "sub-event type2:XXX": ["fine-grained event type 1", "fine-grained event type 2"],  
  "sub-event type2:XXX_definition": ""  
}
```

The list of fine-grained event types is as follows:

Figure 12: Prompt 3 – Generator Prompt Example

Prompt Template for Merger

Below are the **taxonomy** constructed by three LLMs for the **{domain}** of the *Twenty-Four Histories*. These systems are built by abstracting upward from fine-grained, low-level event types. Please follow the steps below:

Use GPT-4o's taxonomy as the primary reference:

1. Deduplication

First, deduplicate the nodes in GPT-4o's **taxonomy**, including nodes that are semantically highly similar or redundant.

2. Uniqueness constraint

Ensure that each child node belongs to only one parent node; the same child node must not appear under multiple parent nodes.

Node supplementation:

3. Supplement from other taxonomies

Select appropriate nodes from the other taxonomies to supplement GPT-4o's results.

4. Coverage constraint

Add nodes only when GPT-4o's taxonomy shows clear coverage gaps.

5. Uncertainty constraint

Do not add nodes that are ambiguous or semantically unclear.

The goal is to make GPT-4o's taxonomy more complete while maintaining a clear and well-structured hierarchy.

Bottom-level node constraint:

6. No new leaf nodes

All bottom-level (leaf) nodes must already exist in at least one of the threetaxonomies. Introducing entirely new leaf nodes is not allowed.

Final check:

7. Final deduplication

Deduplicate the final taxonomy again, keeping only one node for semantically highly similar entries.

After completing these steps, output the optimized taxonomy.

The output structure must be consistent with GPT-4o's taxonomy format and must be in **JSON format**.

Figure 13: Prompt 4 – Merger Prompt Example

Prompt Template for Node Judger

You are an **taxonomy optimization assistant**. I will provide a set of taxonomy nodes from different domains that are semantically very close (i.e., synonym clusters). These nodes may have duplicated parent nodes or highly similar meanings.

Your tasks are:

Select **one node and its path** as the final canonical node to retain.

Choose the **most appropriate domain** for this node.

Output the **complete retained path** exactly as given (do not modify the original path).

If a second-level or third-level node name **duplicates a domain name in another domain** (for example, "Diplomacy" appears both as a top-level domain and as a lower-level node in another domain), then:

Retain the node where it functions as a **top-level domain**;

Delete or ignore the duplicated lower-level nodes appearing in other domains.

List the remaining node names as **synonyms**, preserving the **full original paths** of the merged nodes.

Output requirements:

Strict **JSON format**

Output JSON only

The output must contain the following fields:

canonical_term: the retained node name (the last level of the path)

path: the complete retained path

domain: the selected domain

synonyms: other node names (optional)

Example output:

```
{
  "canonical_term": "Submission",
  "path": ["Military", "Surrender and Allegiance", "Submission"],
  "domain": "Military",
  "synonyms": ["Diplomacy - Allegiance and Vassalage - Submission"]
}
```

Input nodes:

Figure 14: Prompt 5 – Node Judger Prompt Example

Prompt Template for Expander

You are an taxonomy structure optimization assistant. Our ultimate goal is to construct an event taxonomy for The Twenty-Four Histories of Chinese historical literature. The events in this taxonomy must be clearly defined, representative, and hierarchically well-structured.

I will provide you with the following information:

The domain name and description

The current $\{l\}$ -level node and its complete structure, including:

The name of the $\{l\}$ -level node

The definition of the $\{l\}$ -level node

All existing child nodes and their definitions (including $\{l+1\}$ -level and deeper nodes)

Other $\{l\}$ -level nodes in the same domain and their definitions

Please optimize the $\{l\}$ -level node structure by following the steps below.

Step 1: Analyze and Generate $\{l+1\}$ -level Node Dimensions

Analyze the core properties of the $\{l\}$ -level node based on its nature and definition.

When generating $\{l+1\}$ -level nodes, you must strictly follow these rules:

Each $\{l+1\}$ -level node must represent exactly one distinct event type. The boundaries between $\{l+1\}$ -level nodes must be clear, and different event natures must not be mixed.

It is strictly forbidden to use compound forms such as "X and Y" or similar constructions as $\{l+1\}$ -level node names. Only a single word is allowed. This point is extremely important.

For example, nodes such as "Removal and Demotion of Office" or "Departure and Retirement" are not allowed as $\{l+1\}$ -level nodes.

If existing child nodes reflect different event natures, you must split them into multiple independent $\{l+1\}$ -level nodes.

Step 2: Attach Atomic Events

Attach all existing child nodes (atomic events) under the generated $\{l+1\}$ -level nodes as appropriately as possible.

Each atomic event may be assigned to only one parent node and must not be duplicated.

If certain atomic events do not fit under any generated $\{l+1\}$ -level node, you may:

Consider promoting them directly to new $\{l+1\}$ -level nodes (use this option with caution and do not introduce additional higher-level nodes); or

Move them to a more appropriate $\{l\}$ -level node, if they belong elsewhere.

Preserve the original atomic event names and definitions exactly, without adding any extra fields.

Ensure that the child nodes under each $\{l+1\}$ -level node are semantically clear and non-redundant.

Output Requirements

The output must be in JSON format, and the structure must remain consistent with the original taxonomy.

Do not add any extra fields such as `_new` or `third_layer_nodes`.

Example format:

```
{
  "{l}-level Node Name": {
    "definition": "...",
    "Generated {l+1}-level Node 1": {
      "definition": "...",
      "Atomic Event 1": {"definition": "..."},
      "Atomic Event 2": {"definition": "..."}
    },
    "Generated {l+1}-level Node 2": {
      "definition": "...",
      "Atomic Event 3": {"definition": "..."}
    }
  }
}
```

Figure 15: Prompt 6 – Expander Prompt Example

Prompt Template for Conceptualizer

You are an expert specializing in knowledge extraction and ontology construction for The Twenty-Four Histories. I will provide a set of Relation Ontology entries extracted from a knowledge graph built on The Twenty-Four Histories. These relations describe semantic connections among people, events, places, institutions, and other entities recorded in the historical texts.

Your tasks are as follows:

1. Identify relations with event characteristics:
2. That is, relations that represent historical actions, processes, changes, or decisions, rather than static attributes, identity assignments, or geographic/administrative affiliations.
3. Transform the identified relations into Event Ontology entries:
4. Each event entry should function as an independent event-type node with a clear action- or process-oriented meaning.
5. Recommend an appropriate Event Domain for each event, including but not limited to the following categories:

{domain list}

Output format requirement:

Please return the results in structured JSON format, as illustrated below:

```
[
  {
    "original_relation": "Appointment of officials",
    "event_name": "Official Appointment Event",
    "suggested_domain": "Politics"
  },
  {
    "original_relation": "Launching a war",
    "event_name": "War Initiation Event",
    "suggested_domain": "Military"
  }
]
```

Figure 16: Prompt 7 – Conceptualizer Prompt Example

Prompt Template for Enricher

Stage1: You are an expert in constructing event ontologies based on the Twenty-Four Histories, with extensive knowledge of event types and classification systems in ancient Chinese historical records.

Please determine whether a given topic is suitable to be included as an “event” in a Twenty-Four Histories event ontology.

Evaluation criteria:

An “event” should involve elements such as time, actor(s), an action, and an outcome.

Even if the topic refers to an institutional or administrative activity, it should still be considered an event if there is a clear actor and a describable action.

If the topic merely denotes a time reference, object, attribute, or document type (i.e., lacks an action), it should not be treated as an event.

You may refer to high-frequency event types commonly found in the Twenty-Four Histories, such as inspections, imperial tours, royal progresses, and diplomatic visits.

Please strictly follow the JSON format below in your output:

```
{
  "topic": "{topic}",
  "is_event": "Yes/No",
  "domain": "(If 'Yes', specify a domain from {domain list}; if 'No', leave this field empty)",
  "reason": "Briefly explain the rationale for your judgment"
}
```

Stage2: You are an expert in constructing event ontologies for the Twenty-Four Histories, with deep familiarity with event types and classification systems in ancient Chinese historical records.

You are given an event topic and the domain to which it belongs. You also have access to the complete hierarchical ontology structure of that domain, including second-level, third-level, and deeper nodes.

Please complete the following tasks:

Based on the semantics of the event topic, select the most appropriate position for it within the ontology structure:

The event may be placed at the same level as an existing third-level node, or as a child of an existing third- or fourth-level node;

You must choose a position strictly within the provided set of existing child nodes.

Provide a definition for the new node;

Give a brief and clear justification explaining why the event should be placed at this position rather than under other candidate nodes;

If there are secondary (less optimal) candidate nodes, list them with explanations.

Please strictly follow the JSON format below:

```
{
  "Optimal path of the node": "Level-1 node > Level-2 node > ... > Most specific node",
  "Definition of the new node": "Definition of this node",
  "Justification": "Why the event is placed here rather than under other child nodes",
  "Alternative paths": [
    {
      "Node name": "Name of the alternative node",
      "Justification": "Why this node could also be considered"
    }
  ]
}
```

Additional notes:

Ensure the output JSON is complete and parsable;

Do not introduce nodes that do not exist in the ontology.

Figure 17: Prompt 8 – Enricher Prompt Example