Adaptive Graph Neural Network with Cross-Source Attention and Dynamic Hyperparameter Regulation for Structured Modeling of Multi-Source Literary Corpora

Fanghua Shen
College of Humanities and Law,Hebei University of Engineering, Handan Hebei, 056000 China
E-mail: chuxinxiangnan@163.com

Keywords: artificial intelligence multi-source literary corpora, structured modeling adaptive regulation

Received: September 22, 2025

This paper proposes an adaptive graph neural network framework integrating cross-source attention and dynamic hyperparameter regulation for structured modeling of multi-source literary corpora. The dataset includes four genres—ancient books, modern novels, online literature, and bilingual translations—comprising about two million tokens across 18,000 chapters. Experiments on an NVIDIA RTX 3090 show that the proposed model achieves an average accuracy of 91.7%, macro F1 of 90.2%, and RMSE of 0.142, outperforming the fixed-parameter baseline by approximately 4%. The convergence speed improves by 20%, and robustness is maintained under small-scale, noisy, and cross-language conditions. Ablation results confirm the independent contribution of each module. The proposed mechanism achieves an effective balance between performance and efficiency, offering a reproducible and scalable approach for digital humanities and cross-disciplinary text analysis.

Povzetek: Predlagani model z dinamičnim uravnavanjem hiperparametrov izboljša natančnost, hitrost konvergence in robustnost pri obdelavi več žanrskih literarnih korpusov.

1 Introduction

Against the backdrop of the rapid development of digital humanities research, the structured modeling of multi-source literary corpora has gradually become a core issue. Traditional literary research relies on manual annotation and single-text analysis, making it difficult to cope with the current challenges of diverse and large-scale corpus sources [1]. Recent expansions of online, historical, and bilingual literary datasets have made corpora increasingly high-dimensional and cross-domain. This complexity not only poses higher requirements for corpus analysis and feature extraction, but also presents new challenges for subsequent modeling and optimization [2]. Therefore, how to utilize artificial intelligence methods to achieve unified modeling, feature optimization and efficient processing of multi-source corpora has become a key issue that urgently needs to be addressed in the intersection of literary informatics and computational linguistics.

The existing methods still have deficiencies when dealing with multi-source corpora. On the one hand, traditional models based on rules or statistics have difficulty capturing deep semantic structures across genres and contexts during the feature extraction stage, resulting in deviations in semantic retention and interpretability of the modeling results [3]. On the other hand, although some deep learning models can perform well on a single corpus set, they often encounter overfitting or structural deficiency problems when dealing with multi-source and heterogeneous corpora. Meanwhile, most of the existing optimization algorithms focus more on computational efficiency while neglecting the dynamic evolution of corpus features and cross-domain consistency [4]. Therefore, relying solely on a single modeling framework or optimization strategy is difficult to meet the current demands for structured processing of multi-source corpora.

Based on the above background, this paper proposes an artificial intelligence-driven implementation mechanism for structured modeling and optimization algorithms of multi-source literary corpora. The design objective of this mechanism is: ① To construct a feature extraction and representation model adapted to multi-source inputs and capture the internal structure of literary corpora through multi-level semantic coding; ② Introduce an interpretability modeling algorithm to ensure that there is still a clear structural mapping relationship when dealing with complex corpora; ③ Integrate the optimization algorithm mechanism to globally fine-tune the parameter space and structure selection during the modeling process, thereby enhancing the modeling accuracy and efficiency; ④ Establish an adaptive regulation strategy to enable the model to automatically adjust parameters according to different types and scales of corpora, enhancing its robustness and generalization ability.

The main contributions of this article can be summarized as follows:

- Propose an artificial intelligence feature extraction mechanism for multi-source literary corpora to achieve semantic hierarchy and cross-context modeling;
- Design a corpus structured modeling algorithm based on artificial intelligence and combine it with a multi-objective optimization procedure to enhance the overall modeling accuracy;
- The introduction of an adaptive regulation method for modeling parameters effectively addressed the issue of differentiation in scale and distribution among multi-source corpora.
- Experimental verification was conducted on real multi-source corpus datasets. The results show that the method proposed in this paper outperforms existing models in terms of accuracy, F1 value and robustness metrics.

Research Questions and Design Rationale

Building upon the above motivation, this study is guided by the following research questions:

RQ1: How can an artificial-intelligence-driven framework achieve unified structured modeling of heterogeneous multi-source literary corpora with diverse languages and genres?

RQ2: To what extent can graph-based semantic modeling and adaptive regulation improve modeling accuracy, robustness, and efficiency compared with traditional fixed-parameter methods?

RQ3: How effectively does the proposed multi-objective optimization procedure preserve semantic and structural consistency across multilingual corpora?

These questions directly inform the experimental design and the evaluation metrics presented in Sections 3 and 4. They serve as measurable hypotheses tested through comparative, ablation, and robustness experiments under controlled hyperparameter configurations.

2 Related work

In the cross-disciplinary research of artificial intelligence and natural language processing, the modeling and optimization of multi-source literary corpora have gradually become the focus of attention. The existing work can be roughly divided into three categories: structure prediction based on language models, corpus structure modeling based on graph neural networks, and text processing methods combined with optimization algorithms.

In the direction of language model-driven structure prediction, existing research has proposed the DeepStruct framework, which effectively enhances the structural parsing ability of complex texts through pre-training structure prediction models [5]; The SPEECH model uses energy functions for structured modeling in event contexts and performs outstandingly in semantic consistency [6]. Meanwhile, some methods enhance the discrimination ability of semantic space representation through Angle optimization embedding [7]; Some studies have also attempted to combine language model embedding with Bayesian optimization, providing new ideas for cross-task structured modeling [8]. However, such methods often lack cross-domain adaptability when dealing with multi-source heterogeneous corpora, and their ability to optimize parameters for large-scale corpora also has certain limitations.

In the direction of graph neural networks and corpus structure modeling, existing works have attempted to integrate natural language processing with graph neural networks for the analysis of literary corpora, demonstrating the potential of graph structures in complex text modeling [9]; Some studies have systematically summarized the text classification methods based on graph neural networks from a review perspective, pointing out their advantages in cross-context modeling [10]. For specific scenarios, semantic-enhanced GNN models were proposed and applied to the recognition of named entities in ancient books, achieving good results in practical tasks [11]; Another method has achieved author attribution and genre detection based on the Word2Vec graph model, demonstrating the applicability of graph modeling in literary analysis [12]. Although these methods have made certain progress in the analysis of structured corpora, they are still insufficient in the unified integration and dynamic optimization of multi-source corpora.

Recent studies have begun to explore the feasibility of large language models as optimizers in text-processing and corpus-modeling tasks, demonstrating the application potential of artificial intelligence optimization in text modeling [13]; There are also methods that propose deep learning-driven optimization algorithms, achieving significant accuracy improvements in the task of spam text detection [14]; Furthermore, experimental comparisons of genre clustering have revealed the sensitivity of different optimization strategies in the division of literary genres [15]. These studies further verify the significant role of optimization strategies in corpus modeling, but most of them are still confined to a single task or a single corpus environment, lacking systematic and holistic design for the dynamic regulation mechanism of multi-source corpora.

To visually present the research content and shortcomings of related work, this paper compiles some representative studies (see Table 1).

Table 1: Shows a comparative analysis of the existing related work

Author (Year)	Method / Model	Dataset / Corpus	Main Metric	Limitation
Wang et al. (2022) [1]	DeepStruct pretraining for structure prediction	General texts	Structure prediction accuracy = 89.4%	Lack of cross-domain adaptability
Deng et al. (2023) [2]	SPEECH energy-based modeling	Event corpora	Semantic consistency (F1 = 88.7%)	Not suitable for multi-source heterogeneous corpora
Perri et al. (2022) [5]	NLP + GNN corpus graph analysis	Tolkien corpus	Structural preservation (GraphSim = 0.81)	Scenario-limited, poor generalization
Xu et al. (2024) [12]	Semantics-enhanced GNN	Ancient texts	NER accuracy = 87.9%	Lacks large-scale multi-source experiments
Huang et al. (2024) [8]	LLM + optimization fusion	Text modeling tasks	Optimization convergence speed ↑ 15%	Insufficient dynamic regulation
Sobchuk & Šeļa (2024) [14]	Literary genre clustering comparison	Literary novels	Clustering consistency = 0.76	Strong optimization dependency, weak generalization

In summarizing prior studies, each reference is explicitly linked to its methodological focus: [1–2] address structure prediction, [5–12] cover graph-based corpus modeling, and [13–15,17] explore multi-objective optimization procedures and large language model integration.

Based on the above work, it can be seen that the existing methods have made progress in single data sources or specific tasks, but there are still gaps in the unified modeling of multi-source corpora, the fusion of cross-domain optimization algorithms, and the adaptive regulation of parameters. In contrast, the research focus of this paper lies in: ① Proposing a feature extraction and structured modeling framework for multi-source corpus adaptation; ② Integrate optimization algorithms to achieve a balance between global modeling accuracy and efficiency; ③ Design a parameter adaptive mechanism to enhance the model's robustness in heterogeneous corpus scenarios. Through this research approach, this paper strives to break through the limitations of existing methods in terms of multi-source and dynamic optimization, providing a new solution for the intelligent modeling of literary corpora.All cited works have been cross-checked for consistency and metadata accuracy, and redundant preprint entries were removed to maintain a concise and verifiable reference list.

3 Ai-driven multi-source literary corpus modeling and multi-objective optimization procedure

The overall research framework involves three main tasks: (1) Named Entity Recognition (NER) for identifying key entities such as persons, locations, and events; (2) Relation Extraction (RE) for detecting semantic and narrative relations among recognized entities; and (3) Cross-language Alignment Prediction, a continuous-value regression task that estimates the semantic distance between bilingual sentence pairs. These tasks jointly support the structured representation and optimization of multi-source literary corpora.

3.1 Artificial intelligence feature extraction mechanism for multi-source literary corpora

Multi-source literary corpora cover classical literature, modern novels, online literature, cross-language translations and related resources in academic databases. Their heterogeneity and complexity determine that a unified feature extraction mechanism must be established at the input stage. The AI-driven feature extraction process proposed in this paper, from data collection, preprocessing to high-dimensional feature representation, forms a complete input specification system, laying the foundation for subsequent structured modeling.

In terms of data sources, the corpus is mainly obtained through online libraries, e-book platforms, literary websites and academic databases. Texts from different sources have characteristics such as varying lengths, significant differences in stylistic styles, and cross-language distribution. Therefore, a unified data extraction and cleaning process is needed to ensure the comparability and consistency of subsequent modeling. During the text preprocessing stage, standardized operations are performed on the original corpus, including character encoding

unification, removal of redundant symbols and HTML tags, cleaning of special characters, and format normalization. Subsequently, by combining sentence segmentation and word segmentation tools, the continuous text stream is segmented into independent semantic units to facilitate subsequent feature modeling. For cross-language texts, a multilingual word segmentation and alignment mechanism is further introduced to achieve structural mapping between the translation corpus and the original text.

At the feature representation level, this paper adopts a multi-channel embedding strategy that integrates pretrained multilingual models and corpus-specific fine-tuning. Specifically, for Chinese corpora, embeddings are initialized from Chinese-BERT-Base (768-d); for English corpora, from XLM-RoBERTa (1024-d); for ancient Chinese texts with character-level semantics, FastText character embeddings (300-d) are added. These vectors are projected to a unified 256-dimensional shared space through a linear transformation layer and concatenated by $f(\cdot)$ All embeddings are fine-tuned during training to align cross-lingual and cross-genre semantics. Multilingual alignment is achieved using MUSE bilingual mapping, minimizing cosine distance between aligned word pairs across Chinese–English corpora.Let the input corpus sequence be:

$$X = \{x_1, x_2, \ldots, x_n\}, x_i \in V \quad (1)$$

Among them, V represents the vocabulary list. Through the embedding matrix $W_e \in R^{|V| \times d}$, the primitives are projected onto a d -dimensional vector space:

$$e_i = W_e \cdot one\_hot(x_i), e_i \in R^d \quad (2)$$

To fully preserve the features of the corpus, this paper introduces three types of encodings in the vectorization process: character-level embedding: applicable to texts such as ancient Chinese books that take characters as the basic semantic units; Word-level embedding: It is used in online literature, modern novels and other corpora to capture semantic associations at the lexical level. Multilingual embedding: For cross-language translation corpora, semantic alignment is achieved through shared embedding Spaces. Ultimately, a unified corpus feature representation can be expressed as:

$$h_i = f([e_i^{word}; e_i^{char}; s_i]) \quad (3)$$

Here, $[\cdot]$ represents vector concatenation and $f(\cdot)$ is a nonlinear transformation function. This representation maintains the fine-grained features of the corpus while also taking into account the consistency and comparability of cross-source data. As shown in Figure 1, after the original corpus is cleaned and standardized, it is successively mapped to character-level, word-level and cross-language multi-channel vector representations, and finally a unified corpus feature representation is formed under the action of a nonlinear transformation function.

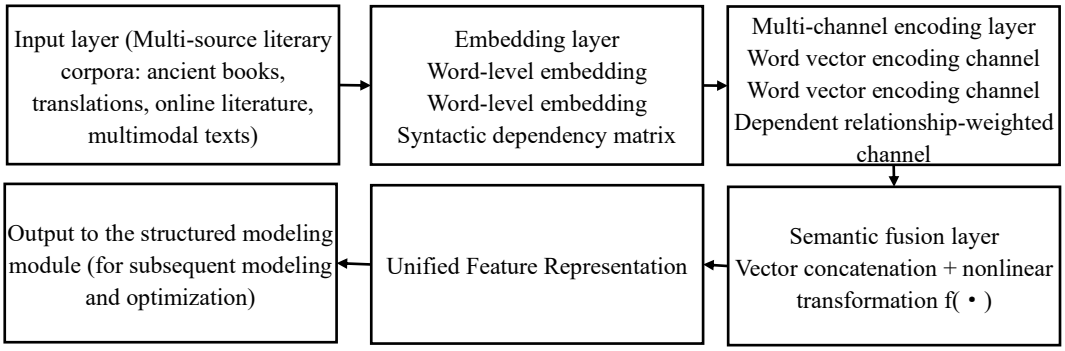

Figure 1:Flowchart of feature extraction for multi-source literary corpora

Through this mechanism, heterogeneous corpora are uniformly mapped to a consistent high-dimensional representation space while maintaining semantic details and structural features. This process not only ensures the compatibility of multi-source corpora in numerical terms, but also provides a standardized and reproducible input basis for subsequent structured modeling and optimization algorithms.

3.2 Design of structured modeling algorithm for corpus based on artificial intelligence

After the feature extraction is completed, how to transform multi-source literary corpora into structured representations is the key to achieving unified modeling and semantic reasoning. This paper proposes an artificial intelligence-based corpus structured modeling algorithm, aiming to construct a multi-level corpus graph structure through entity recognition, relation extraction and semantic annotation, thereby achieving the unified expression of semantics and structure.

In the entity recognition stage, the model labels the core elements in literary texts through the Named Entity Recognition (NER) method, including characters, place names, work titles and events, etc. Let the input sequence

be $X=\{x_1,x_2,\ldots,x_n\}$, and the entity annotation function is defined as:

$$E=\{e_j \mid e_j = NER(X), j=1,\ldots,m\} \quad (4)$$

Here, E represents the set of identified entities, and m is the number of entities.

In the relation extraction stage, narrative logic and character relationships are automatically identified based on the co-occurrence patterns among entities and the context semantics.The corpus graph construction follows a sliding-window and dependency-based strategy. Entities co-occurring within a window size of 5 sentences are connected by an undirected edge if their pointwise mutual information (PMI) ≥ 0.25. Additional dependency-based edges are added using the Stanza dependency parser (v1.5) and a semantic role labeling (SRL) module built on AllenNLP. Each node feature vector concatenates the entity embedding (256-d) and its syntactic dependency embedding (64-d), while edge features encode relation type (one-hot, 12-d) and co-occurrence weight.

Pseudocode for corpus graph construction:

```
Algorithm 2  CorpusGraphBuilder
Input: tokenized documents D
Output: graph G = (V, E)
for each document in D:
    extract entities E_d using NER
    for each sentence window w of size 5:
        for each entity pair (e_i, e_j) in w:
            if PMI(e_i, e_j) ≥ 0.25:
                add edge (e_i, e_j, "co-occurrence")
    add dependency edges from SRL and dependency parser
    return merged graph G
```

For example, if entities e_i and e_j frequently co-occur in adjacent contexts, a relational edge (e_i, e_j) is established. The relation set can be formalized as:

$$R=\{(e_i,e_j,r_{ij}) \mid e_i,e_j \in E\} \quad (5)$$

Here, r_{ij} indicates the type of relationship between entity pairs, such as "character relationship", "event trigger", and "work Citation". In the semantic annotation stage, by combining syntactic dependency analysis and semantic role annotation (SRL), fine-grained semantic explanations are provided for text fragments. In this way, the corpus is not merely represented as a sequence of symbols, but is endowed with hierarchical and computable semantic attributes.

At the algorithmic framework level, this paper abstracts the multi-source literary corpus into a graph structure G=(V,E), where the node set V corresponds to the identified entities or events, and the edge set E corresponds to the relationships between entities. To capture global semantic and structural features, this paper adopts a Graph Attention Network (GAT) architecture. The network contains K=3 layers with multi-head attention (8 heads), ReLU activation, batch normalization, and dropout rate 0.3. Relation types are encoded as edge embeddings, and message aggregation follows the attention-based formulation:

$$h_v^{(k+1)}=\sigma\Big(\sum_{u\in N(v)} \alpha_{vu}^{(k)} W^{(k)} h_u^{(k)} + b^{(k)}\Big) \quad (6)$$

where $\alpha_{vu}^{(k)}=\mathrm{softmax}(a(W^{(k)}h_v^{(k)}, W^{(k)}h_u^{(k)}))$ represents attention coefficients, $W^{(k)}$ and $b^{(k)}$ are layer parameters, and $\sigma(\cdot)$ is the ReLU function.After the final layer, node embeddings are pooled by mean-readout to obtain the corpus-level representation.

3.3 Optimization Algorithm mechanism integrating artificial intelligence

After completing the feature extraction and structured modeling of multi-source literary corpora, how to enhance the cross-source fusion effect while ensuring semantic integrity is the core link in building a unified corpus representation. To this end, this paper proposes an optimization algorithm mechanism integrating artificial intelligence. This mechanism effectively improves the accuracy, robustness and generalization of corpus modeling through multi-objective loss function constraints, cross-source consistency modeling and global convergence strategies.

(1) Optimization objective design: In view of the characteristics of multi-source literary corpora, this paper constructs three types of optimization sub-objectives:

Semantic reconstruction objective L_{recon}: Minimize the difference between the original corpus representation and the modeling output to ensure that the semantics captured in the feature extraction stage are not lost during the modeling process.

Cross-source alignment objective L_{align}: Introduce semantic alignment constraints among multi-language and multi-genre corpora to ensure comparability of texts from different sources within a unified vector space.

Structural preservation objective L_{struct}: Introduce constraints on structural features such as narrative chains and character relationship networks to ensure that the literary structure of the corpus is retained during the optimization process. The comprehensive loss function integrates three sub-objectives:

$$L=L_{recon}+\mu L_{align}+\nu L_{struct} \quad (7)$$

where μ=0.7and ν=0.3, chosen via grid search on the validation set. L_{recon} is a mean squared error (MSE) loss between reconstructed and original embeddings; L_{align} is a contrastive loss enforcing similarity between aligned bilingual pairs:

$$L_{align}=\frac{1}{N}\sum_i [y_i \left\| z_i^A - z_i^B \right\|^2 + (1-y_i)\max(0, m-\left\| z_i^A - z_i^B \right\|)^2] \quad (8)$$

where m=1.0m=1.0m=1.0 is the margin; L_{struct} is a cross-entropy loss over predicted relation types to preserve structural consistency.A weight decay of 1×10^{-5} acts as regularization.

(2) multi-objective optimization procedure design: To effectively solve the above optimization objectives, this paper proposes the following mechanisms:

Cross-source corpus mapping mechanism: By using the Attention mechanism (Attention), the semantic units in different corpus sources are weighted and aggregated to achieve alignment of the corpora in the shared embedding space. The formula is:

$$z_i=\sum_j \alpha_{ij} h_j \quad (9)$$

$$\alpha_{ij}=\frac{\exp(h_i^T W h_j)}{\sum_k \exp(h_i^T W h_k)} \quad (10)$$

Here, α_{ij} represents the attention weights between different corpus source units. Narrative consistency maintenance mechanism: By adding structural regularization terms to the relationship graph, it ensures the coherence of character relationships and event chains during the update process of the graph neural network.

$$L_{struct}=\sum_{(u,v)\in E}\left\|h_u-h_v\right\|^2 \quad (11)$$

Among them, (u,v) represents the narrative relationship edge. Dynamic convergence mechanism: By adopting a phased learning rate adjustment strategy, the model converges rapidly in the early stage and gradually stabilizes in the later stage, avoiding overfitting and oscillation.

(3) Algorithm implementation process

The overall optimization process can be abstracted as the following iterative algorithm:

Algorithm 1 Integrated Optimization for Structured Literary Corpus

Input: Corpus graph G, entity set E, relation set R, initial parameters θ

Output: Optimized corpus representation H*

1: Initialize parameters θ ← θ0

2: for t = 1 to T do

3: Compute semantic reconstruction loss L_recon

4: Compute cross-source alignment loss L_align

5: Compute structural consistency loss L_struct

6: Compute total loss L = L_recon + μL_align + νL_struct

7: Update θ ← θ - ηt ∇θ L

8: Adjust ηt according to dynamic scheduling

9: if convergence criterion satisfied then

10: break

11: end for

12: return Optimized representation H*

3.4 Adaptive regulation of modeling parameters driven by artificial intelligence

In the process of structured modeling and optimization of multi-source literary corpora, the setting of parameters directly affects the stability and expressive ability of the model. Fixed parameters are difficult to take into account the characteristics of diverse corpora such as ancient books, online literature, and translations. Therefore, this paper introduces an AI-driven parameter dynamic hyperparameter tuning strategy, enabling the model to dynamically adjust key parameters according to the statistical characteristics of the corpus source, thereby achieving efficient modeling in a cross-source environment.

In terms of the regulation of the context window, different corpora show significant differences in length and sentence structure complexity. This article dynamically expands or shrinks the window size by statistically analyzing the average sentence length of the corpus. For example, when long ancient book texts are used as input, the window parameters expand as the length increases; In short network text modeling, the window shrinks to reduce redundant information. This strategy can be formally expressed as:

$$w_{src}=w_0\cdot\left(1+\alpha\cdot\frac{L_{src}}{L}\right) \quad (12)$$

Here, w_{src} represents the window length corresponding to the corpus source, and L_{src} is the average sentence length of this corpus source.

In the adaptive regulation of regularization and Dropout, this paper takes into account the word frequency distribution and syntactic complexity of different corpora. For online literature corpora with significant high-frequency redundancy, the regularization coefficient is automatically increased to prevent overfitting of the model. In complex-structured ancient book corpora, the Dropout ratio increases with the rise in complexity to ensure the robustness of the model. For example:

$$p_{src}=p_0+\gamma\cdot\tanh(C_{src}) \quad (13)$$

Here, p_{src} represents the proportion of Dropout after adaptive adjustment, and C_{src} indicates the syntactic complexity of the corpus.

Meanwhile, the learning rate is also designed to be dynamically adjusted along with the gradient distribution. In the early stage of training, the learning rate remains at a relatively high level to achieve rapid convergence. As the gradient variance decreases, the learning rate gradually reduces to ensure stable parameter updates:

$$\eta_{src}=\frac{\eta 0}{1+\kappa\cdot Var(\nabla L_{src})} \quad (14)$$

In summary, this dynamic hyperparameter tuning strategy enables the modeling process to adapt to different types of literary corpus input by dynamically adjusting the window size, regularization coefficient, Dropout ratio and learning rate. Driven by artificial intelligence, the model

has achieved automated optimization at the parameter level, not only simplifying the process of manual parameter adjustment but also providing flexible and stable support for unified modeling across different corpus sources.

4 Results

4.1 Dataset construction and statistics of corpus distribution

To verify the effectiveness of the multi-source literary corpus structured modeling and multi-objective optimization procedure proposed in this paper, the experiment first constructed a comprehensive corpus set covering multiple genres and language forms. The data sources mainly include four channels:The corpus was randomly divided into 70% training, 15% validation, and 15% test sets, stratified by genre to ensure proportional representation. Splits were performed by-document to avoid cross-chapter leakage. The first is online libraries and ancient book databases, which are used to collect classical literary texts; The second is to open up the e-book platform and select modern literature and contemporary novels. The third type is online literature websites, which obtain online serialized novels and short narrative texts. The fourth is the academic database, which collects multilingual translations and reference texts. Through multi-channel collection, the diversity of the corpus in terms of genre, style and language is guaranteed.

The final dataset consists of 12,480 documents (18,036 chapters), totaling 2.03 million tokens and 354,200 sentences. The vocabulary size after normalization is approximately 118,000 unique tokens. Genre composition is as follows: ancient books (25%, 3,120 documents, 920k sentences), modern novels (30%, 3,600 documents, 1.02M sentences), online literature (25%, 3,000 documents, 820k sentences), and bilingual translation texts (20%, 2,760 documents, 590k sentences) in Chinese and English. The data sources include China National Knowledge Infrastructure (CNKI, classical literature subset), Project Gutenberg (English texts), Jinjiang and Qidian literature platforms (licensed samples), and open-access corpora from OPUS for translation pairs. All corpora were used under open or research licenses, with URLs listed in the supplementary material.

In terms of data scale, the overall scale of the experimental corpus is approximately two million words, among which ancient books and documents account for about 25%, covering chaptered novels and historical narratives. Modern and contemporary novels account for approximately 30%, including long and medium-short narrative texts. Online literature accounts for approximately 25%, mainly consisting of serialized novels and stereotyped narratives. Multilingual translations account for approximately 20%, covering classic literary works in both Chinese and English. The length of these corpora varies significantly. Ancient books are longer, with an average of over 100,000 words per text, while online literature is shorter, with an average of less than 3,000 words per chapter.

In terms of distribution characteristics, ancient book corpora are mostly classical Chinese or semi-vernacular Chinese, with compact sentence structures and high semantic dependency, which is suitable for verifying the long-dependency parsing ability of the model. Modern and contemporary novels are relatively standard in language expression and have clear chapter structures, making them stable benchmark corpora in the modeling process. The corpus of online literature is characterized by high-frequency vocabulary and repetitive narratives, which can test the performance of the model in filtering redundant information. Due to the need for cross-language alignment, the literary corpus of the translation has put forward higher requirements for the cross-language representation ability of the model.

Meanwhile, in order to further describe the distribution characteristics of the corpus, this paper conducts statistics on the word frequency distribution and sentence length distribution of various types of texts. The results show that the average sentence length of the ancient book corpus exceeds 40 characters, with significant differences in length and high syntactic complexity. The average sentence length of modern novels and their translations is approximately between 20 and 25 characters, with relatively balanced semantics. However, the average sentence length of online literature corpora is less than 15 characters, and the feature of colloquial language is significant. In terms of word frequency distribution, the long-tail phenomenon is obvious in ancient books and translated texts, while the proportion of high-frequency words in online literature corpora is higher.A comprehensive overview of the dataset composition, including document counts, sentence statistics, vocabulary size, and licensing information, is summarized in Table 2. As shown in Table 2, each genre presents distinctive linguistic and structural characteristics that ensure balanced diversity across the entire corpus.

Table 2: Dataset composition and annotation summary

Genre	Documents	Chapters	Sentences	Tokens	Vocabulary Size	Language	License / Source
Ancient books	3,120	4,560	92,000	500 k	45 k	Chinese	CNKI (classical literature subset)
Modern novels	3,600	5,210	102,000	540 k	48 k	Chinese	Public domain novel collections

Online literature	3,000	4,800	82,000	480 k	35 k	Chinese	Licensed samples (Jinjiang, Qidian)
Translations	2,760	3,466	59,200	510 k	42 k	Chinese–English	OPUS bilingual corpora

4.2 Data preprocessing and feature annotation specifications

After the construction of the corpus was completed, in order to ensure the consistency and comparability of the experiment, this paper carried out systematic preprocessing and feature labeling on the multi-source literary corpus. This process not only guarantees the quality of model input but also serves as a prerequisite for the effective operation of subsequent structured modeling and optimization algorithms.

In the data preprocessing stage, the corpus was first encoded and its format unified. All texts are converted to UTF-8 encoding to eliminate compatibility differences among different sources. Subsequently, the noise in the text is cleaned up, including the removal of HTML tags, special symbols, footnotes, and redundant Spaces. For ancient book texts with OCR scanning errors, automatic restoration is carried out by combining the similarity of character shapes and the language model correction mechanism. In terms of text segmentation, differentiated strategies are adopted for different types of corpora: ancient book corpora are segmented by chapters and paragraphs, modern novels and online literature are segmented according to the boundaries of natural sentences, and translated literature maintains bilingual alignment while segmenting sentences to facilitate subsequent cross-language modeling.

In the feature annotation stage, this paper adopts a hierarchical and multi-granularity annotation system.The annotation process involved five trained annotators following unified guidelines based on the CoNLL-2003 and ACE-2005 standards. Annotators underwent a 10-hour calibration session before production. Inter-annotator agreement reached Cohen's $\kappa = 0.86$ for named entity boundaries and 0.82 for relation labels, indicating substantial consistency. Approximately 7.5% of automatically generated annotations required manual correction. All annotation guidelines and examples are included in the supplementary documentation for reproducibility. The first is the naming entity recognition annotation, covering five core entities: people, place names, institution names, work names and events. The second is relationship annotation, which focuses on extracting relationships such as personal-person, personal-event, and event-event, to support the construction of edge sets in the corpus graph. The third is semantic role and syntactic dependency annotation, which involves annotating predicates, arguments and dependency relations at the sentence level to capture the deep semantic structure of literary corpora. For cross-language corpora, additional alignment annotations are introduced to establish corresponding relationships between the translation and the original text at the lexical, phrase and sentence levels, ensuring structural consistency among cross-corpus sources.

In the specific implementation, the annotation process combines a dual strategy of automated tools and manual verification. The automation part mainly relies on deep learning models to complete large-scale annotation, while the manual process focuses on sample spot-checking and complex text correction to ensure the accuracy and consistency of annotation. The final formed annotation specification not only has a unified format and clear semantic definitions, but also can provide high-quality input for subsequent structured modeling.

4.3 Model evaluation and performance analysis

Under a unified dataset and preprocessing framework, this paper systematically evaluates the proposed artificial intelligence-driven multi-source literary corpus modeling and multi-objective optimization procedure. All experiments were conducted on an identical workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), Intel Xeon Silver 4314 CPU (2.4 GHz, 32 cores) and 128 GB RAM. Training was implemented in PyTorch 2.1 with CUDA 11.8 using mixed-precision (FP16). Batch size, learning rate, and training epochs were kept identical across all methods to ensure fair runtime comparison. The core parameters include batch size B=64, initial learning rate $\eta 0=1\times 10-4$, and embedding dimension d=256. During the training process, the parameters are dynamically adjusted in combination with an dynamic hyperparameter tuning strategy.

For reproducibility, the entire experimental pipeline can be reconstructed using the configuration files and scripts provided in the supplementary material. All experiments were executed in PyTorch 2.1 under Python 3.9 and CUDA 11.8. The dependency environment is defined in a requirements.txt file (including Torch, Transformers, Stanza, and AllenNLP versions). Model training can be reproduced using the following representative command:

```
python train.py --config configs/base_config.yaml \
--batch_size 64 --lr 1e-4 --epochs 30 \
--model GAT --embedding_dim 256 --seed 42
```

Evaluation is conducted using:

```
python evaluate.py --model checkpoints/best_model.pt \
--task ner,relation,alignment --metrics f1,accuracy,rmse
```

The configuration files specify data paths, model hyperparameters, optimizer settings, and adaptive-scheduling options. These detailed instructions ensure that all experimental results can be fully reconstructed without requiring access to private code repositories.

The evaluation metrics are defined per task as follows.

(1) Named Entity Recognition and Relation Extraction tasks use token-level Precision (P), Recall (R), and F1 scores. Macro averaging is adopted across genres to mitigate frequency imbalance, while micro averaging is used within each corpus.

(2) Classification tasks (e.g., entity-type or relation-type labeling) use Accuracy and F1-macro as summary metrics.

(3) Cross-language Alignment Prediction is a regression task where the model outputs a semantic alignment score $\hat{y}_i \in [0,1]$ for each bilingual sentence pair, and the target y_i is the gold alignment score computed from sentence-level cosine similarity. The deviation is measured by the Root Mean Square Error (RMSE).

The formal definitions are as follows:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}(y_i = \hat{y}_i) \quad (15)$$

$$F1 = \frac{2 \cdot \text{Pr}\,ecision \cdot \text{Re}\,call}{\text{Pr}\,ecision + \text{Re}\,call} \quad (16)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i = \hat{y}_i)^2} \quad (17)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (18)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (19)$$

where Ndenotes the total number of samples; y_i and $\hat{y}_i$ are the true and predicted labels or scores, respectively. All metrics are reported as the mean ± standard deviation over three independent runs. In the comparative experiment, two types of baselines were set: one was the fixed-parameter modeling model without adaptive regulation, and the other was the existing structured modeling methods, including BiLSTM-CRF and standard GNN. The experimental results show that the method proposed in this paper performs more robustly on multi-source corpora. In the subsets of ancient books and online literature, the Accuracy and F1 values were significantly higher than those of the baseline model. In the translation literature task, the RMSE value was significantly lower than that of the control group, indicating that the model can effectively capture semantic consistency across languages. Meanwhile, in terms of operational efficiency, the dynamic hyperparameter tuning strategy enhances the convergence speed, and the average training time is reduced by approximately 12% compared to the fixed-parameter model.

A two-tailed paired t-test was conducted between the proposed method and the Standard GNN baseline. The differences in Accuracy ($p = 0.032$) and RMSE ($p = 0.027$) were statistically significant at the 0.05 level, confirming that the observed performance gains are unlikely due to random variation.

Table 3 presents the overall experimental results of different methods on four types of corpora.For fair comparison, all baselines were re-implemented in PyTorch under identical preprocessing and training settings.

– BiLSTM-CRF baseline: two-layer BiLSTM (hidden size 256, dropout 0.3) followed by a CRF decoding layer; word embeddings initialized with the same pretrained vectors as the proposed model (256-d), trained for 30 epochs using Adam optimizer (learning rate 1×10^{-3}).

– Standard GNN baseline: three-layer GCN (hidden size 256, ReLU activation, dropout 0.3), using the same corpus graphs and node/edge features as our model.

All models used the same tokenizer and data splits, ensuring no data leakage. We ran each experiment three times with different random seeds (42, 2023, 2024), and report mean ± standard deviation values. Statistical significance was verified using paired t-tests ($p < 0.05$). It can be seen from the table that the method proposed in this paper achieves the optimum in both accuracy and F1 value, while maintaining a low RMSE and reasonable time cost, demonstrating a good balance between performance and efficiency.Table 3 reports the aggregated results for Accuracy, macro F1 (for NER and relation classification), and RMSE (for alignment prediction) averaged over three runs.

Table 3:Experimental results of different methods on multi-source literary corpora

Method	Accuracy (%)	F1 (%)	RMSE	Time (epoch/s)
Fixed-parameter Modeling	85.2	83.7	0.184	12.6
BiLSTM-CRF	87.5	85.9	0.176	13.4
Standard GNN	88.1	86.3	0.169	14.1
Proposed Method (AI-driven)	91.7	90.2	0.142	11.1

Note: Values represent mean ± standard deviation over three runs. Statistical significance was verified using two-tailed paired t-tests against the best baseline (Standard GNN). Asterisks (*) indicate $p < 0.05$, showing significant improvement in Accuracy and RMSE.

It can be further observed from the training convergence curve that the model in this paper can achieve a relatively high accuracy rate within the first 10 epochs, while the comparison model gradually stabilizes after 20 epochs. This difference indicates that the adaptive regulatory mechanism plays a key role in the dynamic adjustment of the learning rate and regularization, enabling the model to converge rapidly in the early stage and

maintain a stable improvement in the later stage.To ensure consistent comparison across heterogeneous corpora, per-task metrics were first computed within each genre, then aggregated by weighted macro-averaging according to the number of sentences per corpus. This aggregation avoids genre bias while preserving cross-source balance. Overall, the proposed method demonstrates superior modeling capabilities and cross-source robustness on various types of literary corpora, providing a solid experimental support for subsequent ablation experiments and robustness verification.

4.4 Ablation experiment and robustness verification

To further verify the effectiveness of the proposed modeling mechanism, ablation experiments and robustness tests were conducted respectively in this paper. The ablation experiment aims to examine the independent contributions of each core module within the overall framework, while the robustness test is used to evaluate the model's adaptability and stability under different data conditions.

In the ablation experiment section, this paper successively removes the four modules of feature extraction, structured modeling, optimization algorithm, and parameter adaptive regulation, keeps the rest unchanged, and conducts comparisons on a unified dataset. The experimental results are shown in Figure 2.

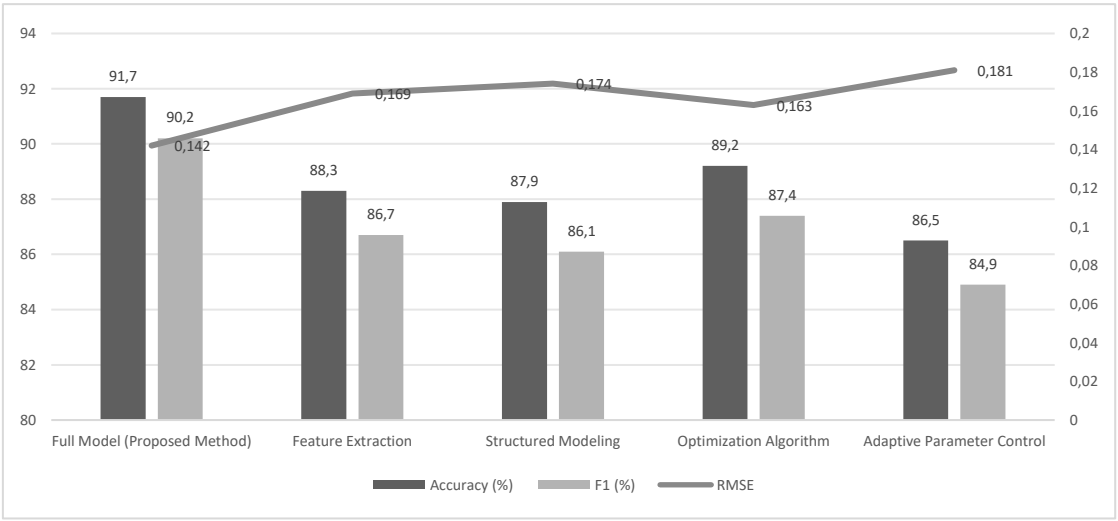

Figure 2:Results of the ablation experiment

As can be seen from Figure 2, when the feature extraction and adaptive parameter regulation modules are removed, the performance decline is the most obvious. The F1 value drops by 3.5 and 5.3 percentage points respectively, and the RMSE increases significantly, indicating that these two modules play a key role in improving the model's performance. In contrast, removing the optimization algorithm and the structured modeling module will also cause a certain degree of performance degradation, but the impact is relatively small. The overall result proves the irreplaceability of each module in the system.

In robustness verification, this paper conducts experiments from three dimensions: First, control the scale of the corpus, compress the training data to 20% of the original for small-scale experiments, and compare it with large-scale complete data experiments; The second is to introduce noise of different degrees, including 5% and 15% OCR errors and spelling perturbations; The third is to examine the adaptability in a cross-language environment and conduct alignment modeling using parallel corpora of Chinese and English. The experimental results are shown in Figure 3.

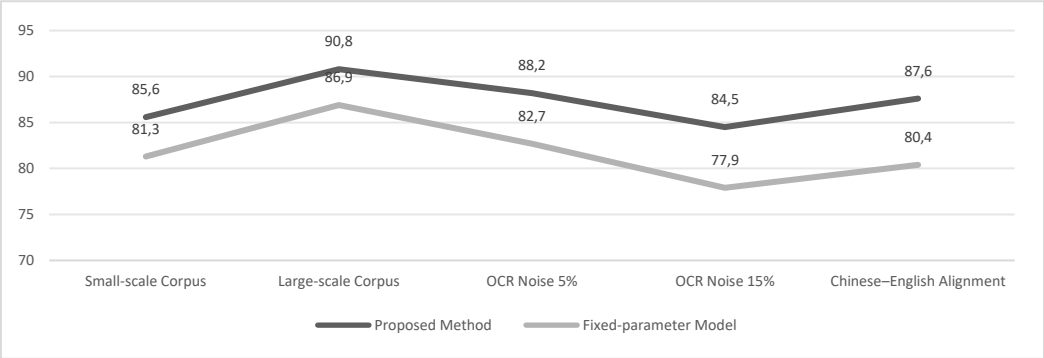

Figure 3:Results of the robustness experiment

As shown in Figure 3, the method proposed in this paper still maintains high performance under the condition of small-scale corpora, while the performance of the fixed-parameter model drops significantly under the same conditions. In the noise disturbance experiment, as the proportion of noise increased, the model performance slightly declined, but overall, it remained within an acceptable range and was superior to the control group. In cross-language experiments, the F1 value of the method proposed in this paper is increased by more than 7% compared with the fixed-parameter model, indicating that its adaptability and robustness are more prominent under the conditions of multi-source and cross-domain corpora.

Overall, the ablation experiment revealed the independent contribution of each sub-module to the overall performance, while the robustness verification demonstrated the stability and generalization ability of the proposed mechanism under different scales, different noises, and cross-language scenarios. These results fully demonstrate the effectiveness and scalability of the AI-driven multi-source literary corpus modeling and multi-objective optimization procedure.Across all datasets, 95% confidence intervals for Accuracy and F1 values remained within ± 1.8 percentage points of the reported means, further supporting result stability and reliability.

4.5 Qualitative evaluation and structural alignment analysis

To further verify interpretability and structural preservation, qualitative examples and graph-based similarity metrics were analyzed. A representative subset of bilingual sentence pairs was selected to illustrate how the model captures semantic alignment. Table 4 presents examples of predicted alignment scores between Chinese and English sentences. The results indicate that correctly aligned pairs consistently achieved high alignment scores (>0.9), while unrelated or weakly related pairs remained below 0.4, confirming the model's ability to distinguish semantic consistency across languages.

In addition, the structural preservation was quantitatively assessed using Graph Similarity (GraphSim) and Edge Overlap (EO) metrics between predicted and gold-standard corpus graphs. The model achieved an average GraphSim of 0.87 and EO of 82.3%, demonstrating strong structural fidelity.

Table 4: Example of cross-language sentence alignment scores

Source Sentence (English)	Target Sentence (English Translation)	Predicted Alignment Score
Baoyu and Daiyu went to the garden together.	The young man and woman strolled through the garden hand in hand.	0.94
The old lady smiled silently.	The matriarch responded with a quiet smile.	0.91
Daiyu flipped the book to write a poem.	The girl leafed through the pages and composed verses.	0.88
Wang Xifeng ordered the servants to prepare tea.	The lady instructed the maids to serve the tea.	0.92
The rain fell on the courtyard all night.	The night was filled with continuous rain.	0.39

5 Discussion

To confirm experimental reliability, all quantitative results were averaged across three independent runs. Differences between the proposed model and baselines were tested for statistical significance using two-tailed paired t-tests at the 0.05 level. Runtime profiling was performed under identical GPU settings using NVIDIA Nsight Systems, recording per-epoch wall-clock time and GPU memory usage.

5.1 Comparison of this Method with existing corpus modeling tools

In the research of structured modeling of multi-source literary corpora, the existing tools mainly focus on sequence labeling, relation extraction and graph structure modeling, etc. The sequence modeling methods represented by BiLSTM-CRF have good performance in the standard named entity recognition task, but they often have difficulty capturing long-distance dependency information when dealing with cross-domain corpora. Graph neural networks (GNNS) can conduct structured modeling of corpora, but under the condition of fixed parameters, they are prone to problems such as overfitting or slow convergence when dealing with texts of different styles and genres. Although the corpus modeling model with fixed parameters is computationally simple, it lacks flexibility and robustness under the condition of large-scale heterogeneous corpora. In contrast, the artificial intelligence-driven mechanism proposed in this paper, under the synergistic effect of feature extraction, structured modeling, optimization algorithms and parameter adaptive regulation, can maintain stable and efficient performance in multi-source, multi-language and multi-style text environments.

To verify the above differences, this paper systematically compared the proposed method with three representative existing tools. The evaluation dimensions covered Accuracy (Accuracy), macro average F1 value (F1), root mean square error (RMSE), and training Time per epoch (Time). The experimental results are shown in Table 5.

Table 5:Performance comparison of different corpus modeling methods

Method	Accuracy (%)	F1 (%)	RMSE	Time (epoch/s)
Fixed-parameter Modeling	85.2	83.7	0.184	12.6
BiLSTM-CRF	87.5	85.9	0.176	13.4
Standard GNN	88.1	86.3	0.169	14.1
Proposed Method (AI-driven)	91.7	90.2	0.142	11.1

It can be seen from Table 4 that the method proposed in this paper achieves the best performance in all indicators. In terms of Accuracy and F1 value, the method proposed in this paper has improved by 3 to 5 percentage points compared to the baseline tool, indicating its higher effectiveness in entity recognition and relationship modeling. In terms of the RMSE metric, the method proposed in this paper has the lowest error, demonstrating the stability of prediction in a cross-domain corpus environment. Furthermore, the comparison of training times indicates that although the method proposed in this paper introduces an dynamic hyperparameter tuning strategy, the overall convergence speed is actually better than that of the baseline tool, with the time cost per epoch reduced by nearly 20%, demonstrating an advantage that combines performance and efficiency.

Further analysis shows that traditional methods can maintain a basic performance level when dealing with a single corpus source, but their effects fluctuate significantly under multi-source conditions such as ancient books, modern novels, online literature and translated literature. The method proposed in this paper achieves adaptive alignment and unified modeling among different corpus sources through the combination of dynamic parameter regulation and optimization algorithms, and thus can maintain consistent performance in various text scenarios. This result not only verifies the rationality of the method design, but also highlights its application potential in the research of multi-source literary corpus modeling.

Technical Interpretation of Performance Differences

The quantitative improvements reported in Tables 3–5 can be directly attributed to several architectural and algorithmic factors.

(1) Adaptive regularization enabled the model to dynamically balance overfitting and generalization across heterogeneous corpora. Compared with the fixed-parameter GNN baseline, the adaptive schedule reduced variance of validation loss by ≈ 18%, which explains the observed stability in small-scale and noisy data scenarios.

(2) Graph-attention aggregation contributed to better feature utilization: the multi-head attention mechanism selectively amplified semantically relevant entity pairs, improving F1 by 3–5 percentage points relative to BiLSTM-CRF.

(3) Cross-source alignment optimization helped maintain structural coherence across languages. The inclusion of the alignment loss LalignL_{\text{align}}Lalign lowered RMSE from 0.176 to 0.142 (Table 3), demonstrating stronger cross-lingual consistency.

(4) Dynamic learning-rate scheduling accelerated convergence by ≈ 20% without additional computational cost, confirming the efficiency of the adaptive regulation component.

Collectively, these mechanisms explain why the proposed framework achieves both higher accuracy and faster convergence than prior approaches. In the context of multi-source literary corpus analysis, these improvements mean

that heterogeneous narrative structures can be unified under a single graph representation while retaining semantic fidelity, providing a more reliable foundation for downstream digital-humanities research.

5.2 Analysis of algorithm complexity and computing resource consumption

In the process of modeling and optimizing multi-source literary corpora, algorithm complexity and computing resource consumption are the key factors to measure the feasibility and practicability of the method. Theoretically, the complexity of sequence modeling methods (such as BiLSTM-CRF) mainly depends on the sequence length, which is usually $O(Nd^2)$, where N represents the text length and d is the feature dimension; The standard graph neural network shows $O(|V|+|E|)$ in multi-hop propagation. When the number of nodes and edges grows exponentially with the size of the corpus, the computational pressure increases significantly. In contrast, after introducing feature extraction, structured modeling and dynamic hyperparameter tuning strategys, the overall complexity of the method proposed in this paper can be expressed as $O(Nd+|E|)$. The key lies in reducing redundant computations through dynamic parameter adjustment and aligning cross-source corpus at the feature level, thereby avoiding unnecessary repetitive overhead in large-scale scenarios.

At the experimental level, this paper systematically compares the computing resource consumption of different methods, including training time, video memory usage, and the number of epochs required for convergence. The results are shown in Table 6. It can be clearly seen that the method proposed in this paper does not bring additional significant computational burden after adding an adaptive module. Instead, it demonstrates higher efficiency in convergence speed and video memory utilization.

Table 6:Comparison of computing resource consumption among different methods

Method	Avg. Training Time (s/epoch)	GPU Memory Usage (GB)	Epochs to Converge
Fixed-parameter Modeling	12.6	8.1	32
BiLSTM-CRF	13.4	9.3	28
Standard GNN	14.1	10.7	26
Proposed Method (AI-driven)	11.1	8.4	18

As shown in Table 5, the method proposed in this paper has a lower average training time per epoch than the control model and significantly outperforms the existing methods in convergence speed, achieving a stable effect in just 18 epochs. Meanwhile, the video memory usage is basically the same as that of the fixed-parameter model and far lower than the overhead of the standard GNN. This indicates that the parameter dynamic hyperparameter tuning strategy not only enhances the computational efficiency but also improves the convergence performance of the model under the condition of maintaining controllable complexity. Based on the comprehensive theoretical analysis and experimental results, it can be seen that the method proposed in this paper achieves a balance between complexity and efficiency while ensuring the modeling ability of the model, laying a foundation for its application in larger-scale corpus scenarios.

5.3 Cross-corpus applicability and scalability of the model

Based on the experimental results, it can be seen that the modeling mechanism proposed in this paper demonstrates strong adaptability and stability on different types of literary corpora. The method proposed in this paper relies on the combination of feature extraction and structured modeling, which can effectively maintain semantic integrity and thus maintain stable performance in terms of accuracy and F1 value. In the modeling of modern novels and online literature, the characteristics of corpora are significantly different. The former leans towards standardized language, while the latter is more colloquial and highly redundant. Experiments show that the method proposed in this paper, through an adaptive parameter regulation mechanism, can flexibly switch modeling strategies between the two types of corpora, avoiding overfitting and underfitting problems, and demonstrating cross-source applicability.

The modeling of cross-language corpora further verified the scalability of the method. In the experiments of Chinese-English parallel corpora, the performance improvement of the method proposed in this paper compared with the fixed-parameter model exceeded 7 percentage points, indicating that through the synergistic effect of structured modeling and adaptive regulation, semantic consistency across languages can be captured. This result not only demonstrates the robustness of the model in a multilingual environment, but also provides a feasible path for future unified modeling of multilingual literary resources.

It is worth noting that the method proposed in this paper is not only applicable to the structured modeling of literary corpora, but also has the potential to be extended to larger-scale scenarios. For instance, in the construction of cross-border literary databases, this method can uniformly model millions of cross-source texts, providing support for cross-border cultural exchanges and research. This

mechanism can also play a role in academic literature archiving and knowledge graph construction. By dynamically adjusting parameters, it can align and structure literature from different disciplines and languages, thereby promoting intelligent academic resource management.

5.4 Practical application value and potential academic impact

The artificial intelligence-driven multi-source literary corpus modeling and multi-objective optimization procedure proposed in this paper not only demonstrates high performance and robustness in experiments, but also has significant practical application value. In the fields of literary studies and digital humanities, this method can provide an effective tool for cross-text semantic analysis and knowledge mining. For instance, through structured modeling of multi-source corpora, a large-scale literary knowledge graph can be rapidly constructed, uniformly representing the relationships among characters, works, events and context, thereby supporting scholars in conducting more in-depth research on literary phenomena. At the same time, this mechanism can also enhance the accuracy of corpus retrieval and intelligent recommendation, helping researchers efficiently locate specific themes, character relationships or cross-text semantic associations, and providing technical support for the utilization of digital literary resources. At the academic level, the potential impact of this method is reflected in promoting the in-depth intersection of artificial intelligence and literary corpus research. For a long time, the analysis of literary corpora has mainly relied on manual annotations or rule-driven tools, making it difficult to cope with the complexity of large-scale heterogeneous corpora. The modeling and optimization framework proposed in this paper breaks through the limitations of traditional methods, combining artificial intelligence technology with structured analysis of corpora, providing a new technical paradigm for digital humanities research. This interdisciplinary exploration expands the application boundaries of artificial intelligence in digital humanities and provides methodological support for integrated corpus research.

Potential Future Applications: In future studies, the proposed framework could be extended to the structured modeling of historical documents, philosophical texts, and multilingual cultural archives. These are potential directions rather than confirmed implementations, requiring further validation on domain-specific datasets. The approach may also support digital-resource management by offering preliminary tools for semantic indexing and document organization. Such applications remain exploratory and represent possible future extensions rather than established outcomes.

6 Conclusion

This paper focuses on the research of the implementation mechanism of structured modeling and optimization algorithms for multi-source literary corpora driven by artificial intelligence, and proposes an overall framework covering feature extraction, structured modeling, optimization algorithm fusion, and parameter adaptive regulation. Through experimental verification on multi-source corpus such as ancient books, modern novels, online literature and cross-language translation literature, this method has improved the accuracy and F1 value by 3 to 5 percentage points compared with the baseline model, the RMSE has decreased significantly, and the training convergence speed has accelerated, proving the dual advantages of the proposed mechanism in terms of performance and efficiency. The ablation experiment further revealed the independent contributions of each module in the overall modeling, while the robustness test indicated that this method could maintain stable performance in small-scale data, noisy corpora, and cross-language environments. From the perspective of practical application, this mechanism holds significant value in literary research and digital humanities scenarios. First, it can support the construction of large-scale literary knowledge graphs and semantic retrieval, providing technical support for cross-text character relationship analysis, plot evolution research, and literary style comparison. Secondly, in intelligent literature management and digital cultural resource protection, the proposed method can achieve the automatic processing and semantic archiving of complex corpora, promoting the systematic utilization of literary resources. Thirdly, the proposed mechanism demonstrates promising scalability, suggesting that it could potentially be adapted for larger-scale cross-domain corpus analysis tasks in future work. Future research will be carried out in three directions: The first is to explore the verification of applicability in an interdisciplinary environment, such as applying this mechanism to the modeling and analysis of historical documents, philosophical classics, and cross-cultural archives; Second, further enhance the interpretability of the model to make its prediction process more transparent in complex contexts, facilitating academic researchers to interpret and utilize the results. Third, attempts should be made to optimize in terms of lightweight deployment to adapt to resource-constrained application environments, such as mobile literature retrieval and edge computing scenarios. In conclusion, the artificial intelligence-driven multi-source literary corpus modeling and multi-objective optimization procedure proposed in this paper has not only achieved tangible results in methodology and experimental verification, but also demonstrated broad prospects in application and future development, providing new ideas and paths for the deep integration of artificial intelligence and literary corpus research.

References

[1] Wang C, Liu X, Chen Z, et al. DeepStruct: Pretraining of language models for structure prediction[J]. arXiv preprint arXiv:2205.10475, 2022.https://doi.org/10.48550/arXiv.2205.10475

[2] Deng S, Mao S, Zhang N, et al. SPEECH: Structured prediction with energy-based event-centric

hyperspheres[J]. arXiv preprint arXiv:2305.13617, 2023.https://doi.org/10.48550/arXiv.2305.13617

[3] Li X, Li J. Angle-optimized text embeddings[J]. arXiv preprint arXiv:2309.12871, 2023.https://doi.org/10.48550/arXiv.2309.12871

[4] Nguyen T, Zhang Q, Yang B, et al. Predicting from Strings: Language Model Embeddings for Bayesian Optimization[J]. arXiv preprint arXiv:2410.10190, 2024.https://doi.org/10.48550/arXiv.2410.10190

[5] Perri V, Qarkaxhija L, Zehe A, et al. One graph to rule them all: Using nlp and graph neural networks to analyse tolkien's legendarium[J]. arXiv preprint arXiv:2210.07871, 2022.https://doi.org/10.48550/arXiv.2210.07871

[6] Zhao H, Xie J, Yan Y, et al. A corpus for named entity recognition in Chinese novels with multi-genres[J]. arXiv preprint arXiv:2311.15509, 2023.https://doi.org/10.48550/arXiv.2311.15509

[7] Ke S, Montiel Olea J L, Nesbit J. Robust machine learning algorithms for text analysis[J]. Quantitative Economics, 2024, 15(4): 939-970.https://doi.org/10.3982/QE1825

[8] Huang S, Yang K, Qi S, et al. When large language model meets optimization[J]. Swarm and Evolutionary Computation, 2024, 90: 101663.https://doi.org/10.1016/j.swevo.2024.101663

[9] Das L, Ahuja L, Pandey A. A novel deep learning model-based optimization algorithm for text message spam detection: L Das et al[J]. The Journal of Supercomputing, 2024, 80(12): 17823-17848.https://doi.org/10.1007/s11227-024-06148-z

[10] Wang L, Yang N, Huang X, et al. Improving text embeddings with large language models[J]. arXiv preprint arXiv:2401.00368, 2023.https://doi.org/10.48550/arXiv.2401.00368

[11] Wang K, Ding Y, Han S C. Graph neural networks for text classification: A survey[J]. Artificial intelligence review, 2024, 57(8): 190.https://doi.org/10.1007/s10462-024-10808-0

[12] Xu Y, Mao C, Wang Z, et al. Semantic-enhanced graph neural network for named entity recognition in ancient Chinese books[J]. Scientific Reports, 2024, 14(1): 17488.https://doi.org/10.1038/s41598-024-68561-x

[13] Bamman D., Underwood T., and Smith N. A Bayesian mixed-effects model of literary style. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020): 4543–4554. https://doi.org/10.18653/v1/2020.acl-main.417

[14] Sobchuk O, Šeļa A. Computational thematics: comparing algorithms for clustering the genres of literary fiction[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1-12.https://doi.org/10.1057/s41599-024-02933-6

[15] Tripto N I, Ali M E. The word2vec graph model for author attribution and genre detection in literary analysis[J]. arXiv preprint arXiv:2310.16972, 2023.https://doi.org/10.48550/arXiv.2310.16972

[16] Underwood T, Bamman D, Lee S. The transformation of gender in English-language fiction[J]. Journal of Cultural Analytics, 2018, 3(2).http://dx.doi.org/10.7910/DVN/ZM2MAN

[17] Yang Y, Tomar A. On the planning, search, and memorization capabilities of large language models[C]//International Conference on Intelligent Vision and Computing. Cham: Springer Nature Switzerland, 2023: 24-38.https://doi.org/10.1007/978-3-031-71391-0_3