GIIM: A Graph Information Integration Method for Chinese-Kazakh CLIR
Supplementary Material

Author Name1 AN1@SAMPLE.COM
Author Name2 AN2@SAMPLE.COM
Author Name3 AN3@SAMPLE.COM
Author Name4 AN4@SAMPLE.COM
Author Name5 AN5@SAMPLE.COM
Address

1. Dataset Details

In order to provide sufficient data support for cross-language information retrieval (CLIR) between Chinese and Kazakh, we constructed CKIRD, a Chinese-Kazakh information retrieval dataset, via translation. Compared to the document-level retrieval dataset CLIRMatrix[1], CKIRD focuses on passage-level ranking, thus enabling a more fine-grained evaluation of the proposed GIIM framework's generalization ability across languages and tasks. We further performed filtering on neighboring entities, as high-quality entity information plays a crucial role in the effectiveness of information aggregation in the GIIM framework. The followings provide detailed descriptions of the CKIRD dataset, the process of obtaining and filtering neighboring entities, and the overall data preprocessing procedures.

Table 1: Statistical Comparison of CLIRMatrix and CKIRD Datasets

Dataset / **Attribute**	**CLIRMatrix**			**CKIRD**		
	Train	Test	Dev	Train	Test	Dev
Chinese Queries	5,087	1,088	387	4,653	1,000	298
Annotated query-document pairs	508,700	108,800	38,700	9,531	1,748	518
Avg. Relevant Docs/Passages	10.57	10.74	11.31	1.19	1.41	1.41
Candidate Docs/Passages	-	100	100	-	100	100
Relevance Label	Multi-level			Binary		
Task Type	Document Retrieval			Passage Retrieval		

1.1. CKIRD

To further enrich the experimental scenarios, we constructed a new passage retrieval dataset, CKIRD, based on the Kazakh open-domain question-answering dataset KazQAD[2]. we con-

1. https://www.cs.jhu.edu/ shuosun/clirmatrix/
2. https://github.com/IS2AI/KazQAD

structed CKIRD through translation, specifically: (1) Queries in KazQAD were translated from Kazakh to Chinese using Google Translate; (2) Similar to CLIRMatrix, we constructed 100 candidate documents for each query in the validation and test sets, where positive samples are passages labeled as 1, and negative samples were randomly selected from the remaining passages to complete the 100 candidates. We re-partitioned the dataset, resulting in CKIRD containing 825,309 passages, with 4,663, 300, and 1,000 queries in the training, validation, and test sets, respectively. The relevance labels between queries and passages are binary (0 for irrelevant, 1 for relevant), as illustrated in Figure 1. This dataset provides additional challenges and application scenarios for passage-level retrieval tasks.

query	passage	label
中国古代第一位皇帝是谁?	Цинь Шихуанди. 22 жасында билікке келген Ин Чжэн 17 жыл бойы бытыраңқылықта өмір сүрген Қытай жерін бір орт-қа бағындыру жолында күрес жүргізді. Б.з.б. 221 ж. Қытайдағы 6 патшалықты өзіне бағындырып, ұлы билеуші — Цинь Шихуанди — “\"Циньнің бірінші императоры\"” деген атпен бір орталыққа бағынған біртұтас Қытай мемлекетін құрды…	1
	Цинь Шихуанди. б.з.б. 214 ж. юэ тайпалары Аулак мемлекетінің әскерлерімен бірлесе отырып, Цинь армиясын талқандады. Дегенмен, Цинь әскерлері Намвьет және Аулак мемлекетінің солт…	0
哈萨克斯坦科学院何时成立？	Қазақ Ғылым Академиясы. "Қазақ Ғылым Академиясы – Қаныш Сәтбаевтың «Пионер» журналының 1946 жылы 1-санында (12-бет) жарияланған мақаласы. Талым өзінің сөзін балаларға арнай отырып, оларды 1946 жылдың 1 маусымы күні ел астанасы Алматы қаласында тұңғыш Қазақстан Ғылым Академиясы құрылғандығымен құттықтайды…	1
	Ермұхан Бекмаханұлы Бекмаханов. 1946-1947 жылдарда Қазақ КСР Ғылым академиясында жаңадан құрылған Тарих, археология және этнография институты директорының ғылыми жұмыс жөніндегі орынбасары, 1947 жылдан бастап, өмірінің соңына дейін, ягни 1966 жылғы мамырдың алтысына дейін Қазақ мемлекеттік университетінде өзі ұйымдастырған Қазақстан тарихы кафедрасын басқарды…	0
	…	

Figure 1: Examples from the CKIRD Dataset.

1.2. Data Preprocessing

To ensure data quality and diversity, we performed detailed preprocessing steps on the CLIRMatrix and CKIRD datasets. Specifically: (1) Entity annotation and matching. We manually annotated all Chinese queries in CLIRMatrix and CKIRD with corresponding entities and retrieved their Chinese, Kazakh, and English labels and descriptions via the *Wikidata API*[3]. Due to the imbalance in corpora across languages, missing Chinese or Kazakh information was supplemented by translating English information using *Google Translate*[4]. Queries that could not be matched with entities were removed. The final statistics for CLIRMatrix and CKIRD are shown in Table 1. (2) Neighboring entity filtering. We used the BGE[5] embedding model to compute the similarity between Chinese queries and their neighboring entities in Chinese and selected the top 10 most similar neighboring entities as supplementary information. Statistical results show that the average number of neighboring entities in the CLIRMatrix dataset is 7.70, while in the CKIRD dataset, it is 8.87. Through these preprocessing steps, we ensured the quality, diversity, and reliability of the data, laying a solid foundation for subsequent model evaluation.

2. Pseudocode for Graph Information Integration

In this supplementary material, we provide the detailed pseudocode of the **Graph Information Integration** process used in our model, as shown in Algorithm 1. The pseudocode clarifies the construction of node features, adjacency matrix, and the graph convolution operations applied in the proposed framework.

Additionally, we present an example adjacency matrix for the case where the number of neighboring entities is 2, as illustrated in Figure 2. This example helps demonstrate the connection patterns and the structural design of the adjacency matrix, which are critical for effectively integrating multi-source graph information.

	v_{qd}	$v'_{e^s_0}$	$v'_{e^s_1}$	$v'_{e^s_2}$	$v'_{e^t_0}$	$v'_{e^t_1}$	$v'_{e^t_2}$
v_{qd}	1	1	0	0	1	0	0
$v'_{e^s_0}$	1	1	1	1	1	0	0
$v'_{e^s_1}$	0	1	1	0	0	1	0
$v'_{e^s_2}$	0	1	0	1	0	0	1
$v'_{e^t_0}$	1	1	0	0	1	1	1
$v'_{e^t_1}$	0	0	1	0	1	1	0
$v'_{e^t_2}$	0	0	0	1	1	0	1

Figure 2: adjacency matrix example.

3. https://www.wikidata.org/w/api.php
4. https://translate.google.com/
5. https://huggingface.co/BAAI/bge-base-zh-v1.5

Algorithm 1 Graph Information Integration

Require: Query-document vector: v_{qd}; Source language aligned entity vectors: $\boldsymbol{v}'_{e_i^s}(i = 0, 1, \ldots, N)$; Target language aligned entity vectors: $\boldsymbol{v}'_{e_i^t}(i = 0, 1, \ldots, N)$; Number of GCN layers: l

Ensure: Knowledge vector: v_{kg}

Node feature construction:

$\boldsymbol{X} \leftarrow [\boldsymbol{v}_{qd}{}^T; \{\boldsymbol{v}'_{e_i^s}{}^T\}_{i=0}^N; \{\boldsymbol{v}'_{e_i^t}{}^T\}_{i=0}^N]^T$ ▷ Stack vectors to form node matrix

Adjacency matrix $\boldsymbol{A}$ construction:

Connect v_{qd} to $v'_{e_0^s}$ and $v'_{e_0^t}$

Connect $v'_{e_i^s} \leftrightarrow v'_{e_i^t}$ for $i = 0, 1, ..., N$

Connect $v'_{e_0^s}$ to all $v'_{e_i^s}$ for $i = 1, ..., N$

Connect $v'_{e_0^t}$ to all $v'_{e_i^t}$ for $i = 1, ..., N$

$\tilde{\boldsymbol{A}} \leftarrow \boldsymbol{A} + \boldsymbol{I}$ ▷ Add self-loops to adjacency matrix

Graph convolution operations:

for $k = 0$ to $l - 1$ **do**

  $\tilde{\boldsymbol{D}} \leftarrow \mathrm{Diag}(\sum_j \tilde{\boldsymbol{A}}_{ij})$ ▷ Calculate degree matrix

  $\boldsymbol{X} \leftarrow \mathrm{ReLU}(\tilde{\boldsymbol{D}}^{-1/2} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-1/2} \boldsymbol{X} \boldsymbol{W}_k)$ ▷ Graph convolution

end

Knowledge vector generation:

$v_{kg} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i^{(l)}$ ▷ Average pooling of node representations

return v_{kg}

3. Implementation Details

The experiments of GIIM are initialized with the mBERT pre-trained weights provided by HuggingFace. During training, both the GCN and MLP modules are configured with two layers. To enhance model performance, mBERT is fine-tuned, and the key hyperparameters such as learning rates, temperature coefficient, and loss weight coefficient are tuned on the validation set, with detailed values shown in Table 2. In addition, the number of neighboring entities (N), batch size, and the number of training epochs are carefully set to ensure the stability of the model across different data scales.

4. Detailed Experimental Results of Parameter Sensitivity Analysis

In this section, we present comprehensive sensitivity analysis results on key parameters, including the number of neighboring entities (N) and the contrastive loss weight (λ), across both CLIRMatrix and CKIRD datasets. The heatmaps in Figure 3 reveal that the model achieves robust performance when the number of neighboring entities N is in the range of $[3, 5]$ and the contrastive loss weight λ is with in $[0.3, 0.4]$. Notably, the best retrieval performance on both datasets is observed when $N = 4$ and $\lambda = 0.3$, suggesting that the GIIM framework is stable and effective across different parameter configurations and retrieval scenarios.

Table 2: GIIM Experimental Settings

Parameter	Value	Note
mBERT fine-tuning learning rate (l_{r1})	1×10^{-5}	None
Other module learning rate (l_{r2})	1×10^{-3}	Including GCN, MLP, etc.
Contrastive loss temperature coefficient (τ)	0.05	None
Contrastive loss weight coefficient (λ)	0.3	Range $[0.0, 0.9]$
Number of neighboring entities (N)	4	Range $[1, 7]$
Batch size	16	None
Epochs	16	None

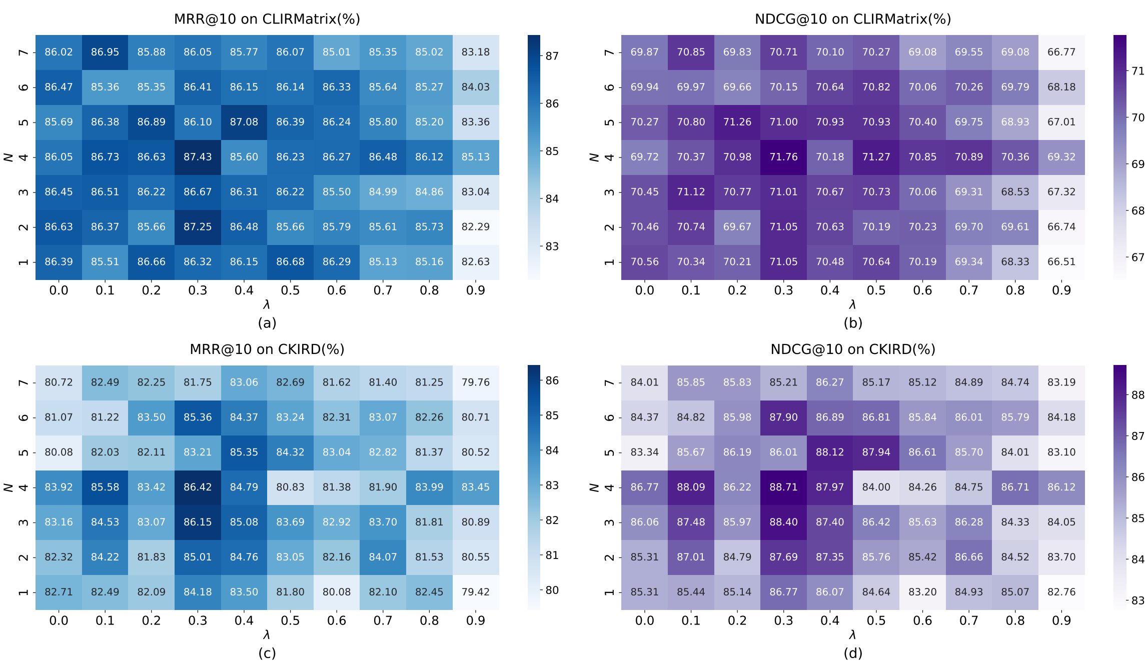

Figure 3: Heatmaps of model performance with varying numbers of neighboring entities (N) and contrastive learning weights (λ) on CLIRMatrix and CKIRD datasets. (a) and (b) show MRR@10 and NDCG@10 scores on CLIRMatrix, respectively; (c) and (d) show the corresponding results on CKIRD. The optimal performance is consistently observed when $N = 4$ and $\lambda = 0.3$, indicating the model's stability across different retrieval tasks. Darker color indicates higher performance.