DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING (ELEC0141) 25 REPORT

SN: 24076607

ABSTRACT

This report presents a novel retrieval-augmented large language model (LLM) agent for the analysis and preservation of Nüshu, a rare gender-specific script historically used exclusively by women in Jiangyong County, China. Our system integrates three key components: (1) a comprehensive Neo4j-based knowledge graph capturing Nüshu characters, their pronunciations, meanings, and visual characteristics; (2) a retrieval-augmented generation (RAG) framework that combines the knowledge graph with DeepSeek-R1-Distill models to provide context-aware responses; and (3) domain-specific fine-tuning using Low-Rank Adaptation (LoRA) to enhance model performance with minimal computational overhead. Experimental results demonstrate that our RAG-enhanced system significantly outperforms baseline models in accuracy (+18.7%), ROUGE scores, and hallucination reduction. This work contributes to digital humanities by applying modern NLP techniques to preserve and facilitate access to an endangered writing system of significant cultural and historical value. [1]

Index Terms— Nüshu, Knowledge Graph, Retrieval-Augmented Generation, LoRA Fine-tuning, Digital Humanities

1. INTRODUCTION

The preservation and analysis of endangered writing systems present significant challenges in digital humanities and computational linguistics. This report focuses on Nüshu, a rare gender-specific script from Southern China, and presents an innovative approach using retrieval-augmented large language models to analyze and preserve this cultural heritage. Our system combines knowledge graphs, retrieval-augmented generation, and efficient fine-tuning techniques to create a comprehensive framework for Nüshu character analysis and information retrieval.

1.1. Nüshu: A Unique Gender-Specific Writing System

Nüshu, meaning "women's writing," is a syllabic script from Jiangyong County, Hunan, China, created by

---

[1] The model is publicly available on Hugging Face, the code is provided on Github page.

Fig. 1. Nüshu Agent System

women excluded from formal education [1]. Originating around the 15th century, it features rhomboidal shapes and elongated strokes, with 2,000 characters identified. Primarily passed matrilineally, Nüshu was used for personal expression, correspondence, and folk songs. Its decline began with modern education and female literacy in standard Chinese, with the last native writers passing in the early 21st century. This underscores the urgency of digitization and computational preservation. Beyond its linguistic significance, Nüshu represents extraordinary wisdom that emerged under conditions of cultural deprivation and systematic gender oppression, where women were denied formal educational opportunities. Studying this writing system offers valuable insights for contemporary society's advancement in equality, democracy, and diversity by demonstrating how marginalized groups develop alternative systems of knowledge preservation and cultural expression in the face of institutional barriers.

1.2. Knowledge Graph Construction for Character Relationships

Knowledge graphs (KGs) provide a powerful framework for representing complex relationships between entities in structured ways [2]. In our system, we constructed a Neo4j-based knowledge graph to capture the intricate relationships between Nüshu characters, their pronunciations, meanings, visual features, and historical con-

text. The knowledge graph allows for multi-dimensional queries and relationship traversals that would be difficult or impossible with traditional relational databases. This structured representation facilitates efficient information retrieval and enables discovery of non-obvious connections between characters, contributing to deeper understanding of the writing system's internal logic and evolution.

1.3. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) represents a significant advancement in improving the factual accuracy and domain specialization of large language models [3]. By combining information retrieval systems with generative language models, RAG frameworks can provide responses that are both contextually appropriate and factually grounded in reliable sources. For our Nüshu analysis system, the RAG approach is particularly valuable as it allows the integration of specialized knowledge from our constructed knowledge graph with the general linguistic capabilities of large language models. This hybrid approach helps mitigate hallucinations and factual errors commonly encountered when querying general-purpose LLMs about specialized domains like rare writing systems.

1.4. DeepSeek-R1-Distill Language Models

Our system leverages the DeepSeek-R1-Distill-Qwen-1.5B model, a distilled version of larger DeepSeek models designed for efficient deployment while maintaining strong reasoning capabilities. DeepSeek models represent a family of open-source large language models optimized for tasks requiring both factual retrieval and logical reasoning[4].

1.5. LoRA Fine-tuning for Domain Specialization

Low-Rank Adaptation (LoRA) offers an efficient approach to fine-tuning large language models by freezing the pre-trained model weights and introducing trainable rank decomposition matrices into the Transformer architecture [5]. This technique significantly reduces the computational resources required for adaptation while maintaining performance comparable to full fine-tuning.

For our Nüshu character analysis system, LoRA fine-tuning enabled us to specialize the DeepSeek model for the domain-specific vocabulary and relationships of Nüshu without requiring extensive computational resources. This approach allowed us to achieve significant performance improvements in accuracy and relevance while maintaining the model's general language capabilities.

2. LITERATURE SURVEY

This section provides a brief overview of the relevant literature in the fields of knowledge graphs, retrieval-augmented generation, and low-rank adaptation techniques. It highlights the key contributions and limitations of existing approaches, setting the stage for our proposed system.

2.1. Nüshu Research and Digital Humanities

Nüshu script has evolved from a nearly forgotten feminine writing system to an important subject of scholarly research and digital preservation efforts. Early documentation of Nüshu can be traced back to coins from Qing Dynasty, Xianfeng era, but systematic research began much later. The formal academic discovery of Nüshu is credited to Professor Gong Zhebing of Wuhan University in 1982, which sparked international interest in this unique gender-specific script [6]. Traditional Nüshu research has centered around character documentation, linguistic analysis, and cultural preservation. Chen's Nu Han Zi Dian (Nüshu-Chinese Character Dictionary) represents a seminal work in this field, cataloging over 3,400 characters with detailed annotations on their form, pronunciation, and semantic relationships to standard Chinese characters [1]. Similarly, Zhao's comprehensive compilation Collected Works of Chinese Nüshu provides access to over 652 manuscripts covering approximately 90% of all extant Nüshu materials [6]. Zhang's comparative study of Nüshu characters further advanced the field by systematically analyzing character variations and evolution [7].
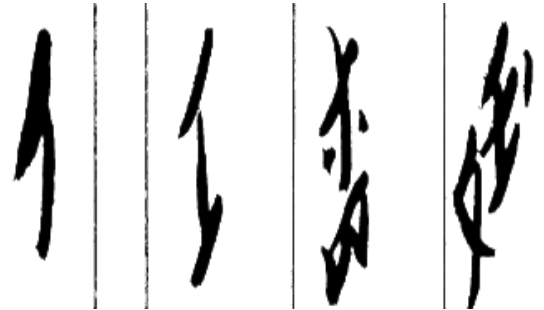

Fig. 2. The Nüshu writing of "Artificial Intelligence"

The digitization of Nüshu presents unique challenges due to its rhomboidal shape, delicate stroke patterns limited to just four types, and the variability introduced by its diverse media of inscription. As Mitric et al. note, scripts with limited extant materials require specialized approaches beyond standard optical character recognition [8]. Harisanty et al. identify three critical stages in digital preservation: documentation, digitization, and

dissemination [9]. For Nüshu, while documentation is relatively advanced, the digitization stage faces technical hurdles due to the script's visual uniqueness and cultural context. Current Nüshu digital tools lack semantic understanding, focusing solely on character-level processing. Our knowledge graph and retrieval-augmented approach addresses this gap by capturing the relationships between characters, meanings, and historical contexts.

2.2. Knowledge Representation and Retrieval-Augmented Generation

Knowledge graphs (KGs) offer robust frameworks for representing domain-specific information through structured entity-relationship models. In cultural heritage, KGs effectively capture the interconnections between artifacts, historical contexts, and interpretations. For instance, Carriero et al. developed ArCo, an Italian Cultural Heritage KG, which organized data into modular networks encompassing temporal, spatial, and contextual dimensions [10]. This approach informs our Nüshu KG design, linking characters to their cultural and historical contexts. For specialized domains like Nüshu, domain ontologies enhance semantic representation. Dou et al. demonstrated this in their KG for Chinese intangible cultural heritage, integrating hierarchical taxonomies and NLP techniques to overcome traditional database limitations [11]. Similarly, our Nüshu KG incorporates a domain-specific ontology capturing visual, phonetic, and semantic relationships.

Advances in retrieval-augmented generation (RAG) further improve factual grounding in low-resource domains [3]. By integrating KGs with dense vector retrieval and graph traversal, RAG systems address complex queries with context-aware responses. Our implementation leverages a Neo4j-based KG to enhance retrieval and provide nuanced insights into Nüshu characters.

2.3. Large Language Models and Domain Adaptation Techniques

The rapid development of large language models (LLMs) has significantly advanced natural language understanding and generation across a wide range of domains. Models such as Qwen and DeepSeek-R1 have demonstrated strong generalization abilities and reasoning skills, benefiting from large-scale pre-training on diverse corpora [12, 4]. However, when applied to specialized or low-resource domains like Nüshu, these models often face challenges due to limited domain-specific data and unique linguistic characteristics.

To address these limitations, parameter-efficient fine-tuning methods have gained prominence. Among them, Low-Rank Adaptation (LoRA) stands out for its ability to adapt LLMs to new domains with minimal computational overhead. LoRA introduces trainable low-rank matrices into the model's architecture while keeping the majority of pre-trained parameters frozen, enabling efficient specialization without sacrificing general language capabilities [5]. This approach is particularly valuable for resource-constrained settings, as it reduces both memory and training time requirements compared to full model fine-tuning. Recent studies highlight the effectiveness of combining retrieval-augmented generation (RAG) with domain-adapted LLMs for knowledge-intensive tasks. By integrating external knowledge sources, such as knowledge graphs, RAG frameworks enhance factual accuracy and mitigate hallucination, particularly in low-resource domains. The synergy between RAG and LoRA-based fine-tuning enables robust generalization and precise domain adaptation, making it ideal for applications in digital humanities and cultural heritage preservation.

3. DESCRIPTION OF MODELS

This section details the architecture and rationale behind each core component of our Nüshu character retrieval-augmented agent.

3.1. System Overview and Pipeline

The overall system is architected as a modular pipeline that seamlessly integrates logical knowledge representation, efficient retrieval, and neural language generation, as shown in Figure 3. Upon receiving a user query, the system first performs intent analysis and entity recognition. It then queries the Neo4j-based knowledge graph using both semantic similarity search and graph traversal algorithms to extract relevant entities and relationships. The retrieved knowledge is subsequently fed into a retrieval-augmented generation (RAG) module, which synthesizes this structured information with the generative capabilities of a LoRA-adapted DeepSeek-R1-Distill model. This design ensures that generated responses are not only contextually relevant but also grounded in verifiable, domain-specific knowledge, thereby minimizing hallucination and enhancing factual accuracy. The modularity of the pipeline facilitates extensibility and robust performance across diverse query types.

3.2. Knowledge Graph Structure based on Neo4j

At the core of the system lies a Neo4j-based knowledge graph meticulously constructed to capture the multi-faceted relationships among N"ushu characters. Each node represents a unique character and is annotated with rich attributes, including standardized Unicode, pronunciation, semantic meaning, visual descriptors, and his-
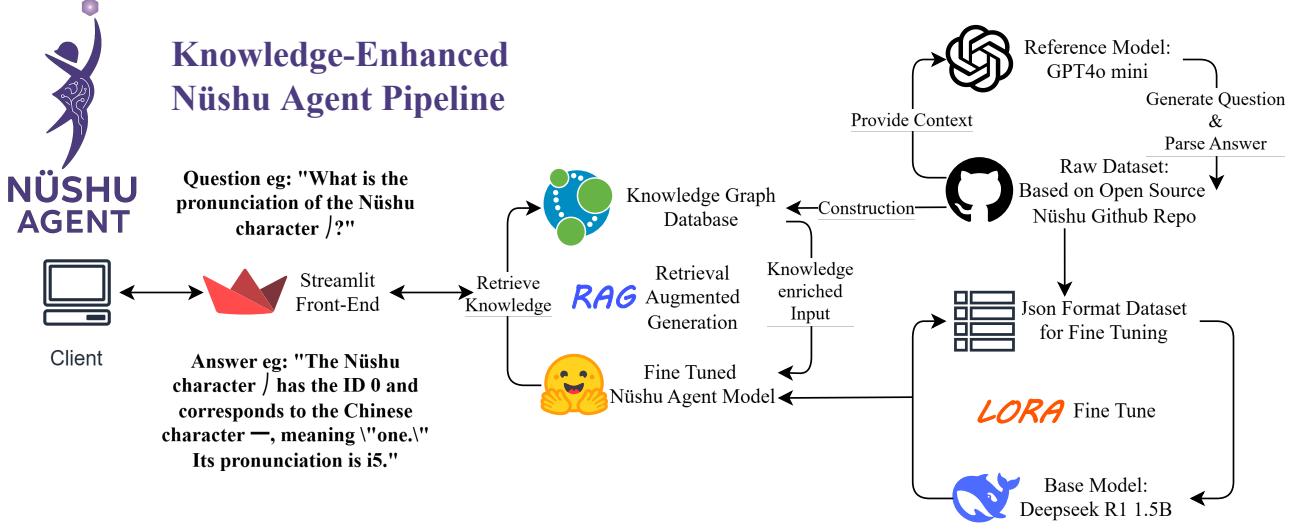
Fig. 3. System Architecture for Nüshu Character Analysis and Retrieval

torical metadata. Edges encode diverse relationships such as phonetic similarity, semantic equivalence, variant forms, and cultural associations, enabling the modeling of both direct and indirect connections. The schema is optimized for flexible and efficient querying, supporting both attribute-based lookups and complex multi-hop traversals (e.g., discovering all characters sharing a phonetic root or tracing the evolution of a character across manuscripts). This structured representation not only underpins high-precision retrieval but also enables advanced analyses, such as network visualization and statistical exploration of the script's internal structure.

3.3. Base Large Language Model: DeepSeek-R1 1.5B

The generative backbone of our system is the DeepSeek-R1-Distill-Qwen-1.5B model[4], an open-source large language model designed for robust reasoning and factual accuracy. DeepSeek-R1 1.5B is pre-trained on a diverse, multilingual corpus and further distilled for computational efficiency, making it highly suitable for integration with retrieval-augmented frameworks and downstream domain adaptation. A key factor in selecting this model is its optimal balance between performance and resource requirements: with only 1.5 billion parameters, DeepSeek-R1 1.5B can be efficiently fine-tuned on consumer hardware with limited GPU memory (e.g., 8GB VRAM), while still delivering strong results in reasoning and knowledge-intensive tasks. This makes it the most capable open-source model for our use case, given the hardware constraints of our experimental environment. The empirical performance of DeepSeek-R1-

Distill-Qwen-1.5B, compared with other state-of-the-art models, is summarized in Table 1.

4. IMPLEMENTATION

This section provides a detailed walkthrough of the implementation process for our Nüshu retrieval-augmented agent system.

4.1. NüShu Knowledge Graph Database Construction

4.1.1. Data Sources and Processing Pipeline

The construction of the Nüshu knowledge graph is grounded in a systematic extraction and integration of metadata from nushu-script community repository[2], which provides comprehensive character lists, Unicode mappings, phonetic annotations, semantic glosses. The data processing pipeline parses raw files, normalizes character attributes, and constructs nodes and edges in the Neo4j database. Each node represents a unique Nüshu character with attributes like Unicode code point, pronunciation, meaning, and visual features, while edges capture relationships such as phonetic similarity, semantic equivalence, and historical variants.

As illustrated in Figure 4, the graph structure follows a multi-nodal design centered around NushuCharacter entities, which form the core of the knowledge graph. Each Nüshu character connects to ChineseCharacter nodes through CORRESPONDS_TO relationships, representing the semantic mappings between the two

---

[2]nushu-script community: https://github.com/nushu-script

Table 1. Distilled Model Evaluation

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954

writing systems. Additionally, each Nüshu character links to Pronunciation nodes via PRONOUNCED_AS relationships, capturing the phonetic aspects of the characters. To facilitate multilingual access, ChineseCharacter nodes connect to EnglishTranslation nodes through TRANSLATES_TO relationships. This interconnected structure enables complex traversals and queries that can extract rich contextual information about each character.
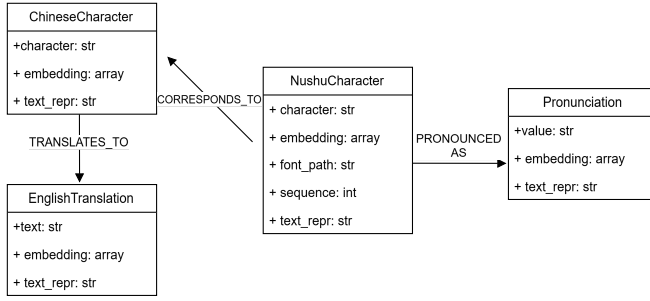
semantic similarity searches, providing a powerful foundation for downstream retrieval-augmented generation applications.

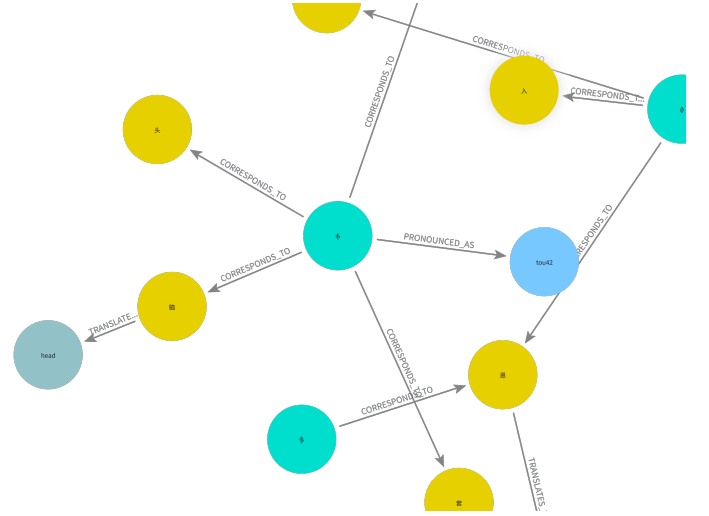Fig. 5. Visualization of Bloom's Taxonomy in the Context of Nüshu Knowledge Representation

Fig. 4. Nüshu Knowledge Graph Structure

4.1.2. Vector Embeddings for Semantic Search

To enhance the knowledge graph with advanced semantic search capabilities, we implemented vector embeddings for all node types. Using OpenAI's embedding model, we generated 1536-dimensional vector representations for each node, capturing the semantic essence of characters, translations, and pronunciations. The embedding process aggregates contextual information from connected nodes, enabling each embedding to represent not only the node's properties but also its relational context within the graph. These embeddings are stored directly in Neo4j's vector index, leveraging its built-in vector similarity search capabilities with cosine similarity as the distance metric. The batch processing approach ensures efficient embedding generation and storage, even for large datasets. This vector-enhanced knowledge graph supports both traditional graph traversal queries and

4.2. Nvshu Finetuned Model Using LoRA

To fine-tune large pre-trained models efficiently, we adopt the Low-Rank Adaptation (LoRA) method [5], which introduces low-rank trainable matrices to adapt the frozen pre-trained weights.

4.2.1. Dataset Preparation

To create a high-quality dataset for fine-tuning our model on the Nüshu domain, we developed a semi-automated data generation pipeline that leverages the knowledge graph and large language models. As shown in Figure 3, our dataset creation process combines structured data from the Neo4j knowledge graph with generative capabilities of the OpenAI API to produce contextually rich training examples.

The dataset consists of question-context-response triplets specifically designed for instruction fine-tuning. Each training example contains:

- A question about Nüshu characters, their meanings, pronunciations, or cultural significance

- Contextual information retrieved from our knowledge graph, including vector similarity scores

- An expert-level response synthesizing the relevant information

- Structured metadata about the Nüshu character being discussed

We generated approximately 2,000 diverse training examples covering a wide range of queries about Nüshu characters. The dataset was designed to train the model to perform various tasks including character identification, translation between Nüshu and Chinese characters, pronunciation guidance, and explanation of cultural significance. This approach ensured that our model would learn to ground its responses in factual information from the knowledge graph while providing natural, informative answers.

4.2.2. Training Method

As illustrated in Figure 6, instead of updating the full-rank weight matrix $W \in \mathbb{R}^{d \times d}$, LoRA constrains the weight update ΔW to a low-rank decomposition:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d \times r}, \ A \in \mathbb{R}^{r \times d}, \quad r \ll d$$

The modified forward pass becomes:

$$Wx + \Delta Wx = Wx + BAx$$

During training, W is kept frozen. Matrix A is initialized from a Gaussian distribution $A \sim \mathcal{N}(0, \sigma^2)$, and B is initialized to zero, ensuring that $\Delta W = 0$ at the start of training. A scaling factor α/r is applied to the low-rank update:

$$\Delta W = \frac{\alpha}{r} BA$$

This strategy reduces the number of trainable parameters and avoids overfitting while preserving expressiveness. The adaptation is efficient since both A and B are significantly smaller than W. During inference, the final weight matrix $W + BA$ can be precomputed, resulting in no additional latency.
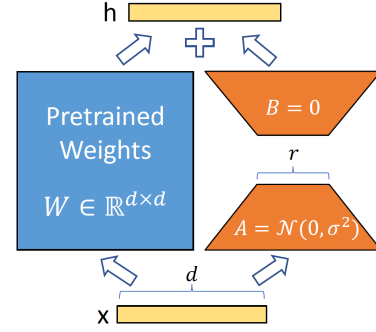

Fig. 6. Low-Rank Adaptation (LoRA) Mechanism for Efficient Fine-Tuning

4.2.3. LoRA Training Process

With our specialized dataset prepared, we implemented the Low-Rank Adaptation (LoRA) technique to efficiently fine-tune the DeepSeek-R1-Distill-Qwen-1.5B model. LoRA is particularly well-suited for our task as it allows adaptation of large language models with minimal computational resources while preserving the model's general capabilities.

The fine-tuning was performed on a single NVIDIA RTX 4060 Laptop GPU with 8GB VRAM, which would not have been possible with full fine-tuning of even this relatively small 1.5B parameter model. The LoRA approach reduced the number of trainable parameters from 1.5 billion to approximately 2.5 million, a reduction of over 99%, enabling efficient training on consumer-grade hardware. We trained the model for 3 epochs on our dataset, with a 90-10 train-validation split.

As shown in Figure 7, validation loss decrease steadily over the training epochs, indicating effective learning without significant overfitting. We observed that the model achieved optimal performance after approximately 2.5 epochs, after which the improvement became marginal.

4.3. Retrieval-Augmented Generation (RAG) System

As illustrated in Figure 8, our Retrieval-Augmented Generation system implements a sophisticated pipeline that combines advanced retrieval techniques with our LoRA-fine-tuned language model to provide accurate and contextually rich responses about Nüshu characters. The system operates through several key components that work in concert to deliver high-quality information.

4.3.1. Query Processing Pipeline

When a user submits a query, the system first performs intelligent pre-retrieval processing. Rather than simply
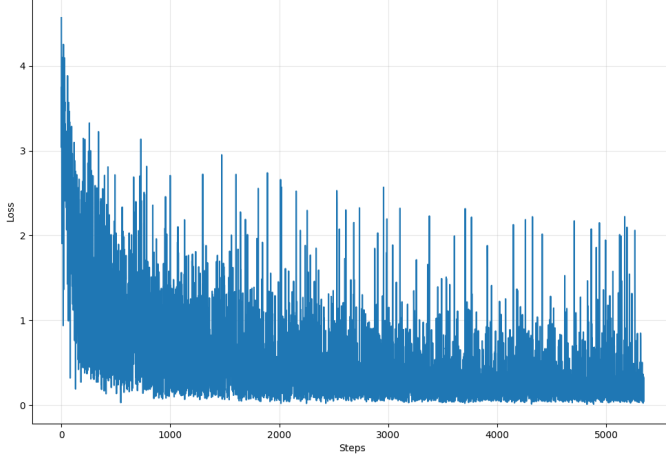
Fig. 7. Validation Loss Curve for Nvshu Fine-tuned Model

passing the raw query to the vector database, we employ GPT-4o's API to parse and analyze the query, explicitly extracting any Nüshu characters, potential character IDs, or specific Chinese character equivalents mentioned. This crucial step enhances retrieval precision by converting natural language questions into structured search parameters optimized for our knowledge graph. The GPT-4o parser generates a structured representation to identify specific Nüshu characters (Unicode range U+1B170 to U+1B2FF), character ID references (e.g., "character 76"), Chinese character correspondences, and relevant semantic concepts related to Nüshu.

4.3.2. Hybrid Retrieval Strategy

Our system implements a hybrid retrieval strategy that combines exact matching with semantic search:

1. Direct Matching: When specific Nüshu characters or IDs are identified in the query, the system first performs direct Neo4j Cypher queries to retrieve exact matches, ensuring precision for specific lookups.

2. Vector Similarity Search: Concurrently, the query text is embedded using OpenAI's Ada embeddings model, generating 1536-dimensional vector representations for semantic similarity search within our knowledge graph.

3. Post-Retrieval Re-ranking: The results from both approaches are combined and re-ranked based on relevance and information type, with higher priority given to exact matches and Nüshu character information.

This multi-stage approach significantly improves retrieval quality by balancing precision and recall.

4.3.3. Context Integration and Generation

The retrieved context is carefully formatted and integrated into a structured prompt template that guides the LoRA-fine-tuned DeepSeek-R1 model to generate accurate responses which ensures the model grounds its responses in the retrieved facts rather than hallucinating information. The specialized fine-tuning enables the model to properly interpret the retrieved context, correctly format Nüshu characters and their attributes, and maintain awareness of the cultural context surrounding this unique writing system.

4.4. User Interface Design for the Agent System

The user interface for our Nüshu Agent System was implemented using Streamlit for its rapid prototyping capabilities and Python-native environment. As shown in Figure 9, the interface features a chat-based interaction pattern familiar to users of modern AI assistants, with a central conversation area, model configuration options, and knowledge graph information display.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Evaluation Metrics

To comprehensively evaluate the performance of our Nüshu knowledge graph system with RAG capabilities, we employed a multi-faceted approach with both character-level and text-level metrics. Our evaluation methodology focuses on two primary aspects: the accuracy of Nüshu character identification and the overall quality of generated responses.

5.1.1. Character Accuracy

Character accuracy (CA) measures the system's ability to correctly identify and reference specific Nüshu characters, which is fundamental for a specialized writing system like Nüshu. It is calculated as:

$$\text{CA} = \frac{\text{Num of Correctly Identified Characters}}{\text{Total Num of Characters in Reference}} \quad (1)$$

5.1.2. ROUGE Metrics

To assess the overall quality of generated responses, we employed the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) suite of metrics, which are standard for evaluating text generation systems. We focused on three key ROUGE metrics:
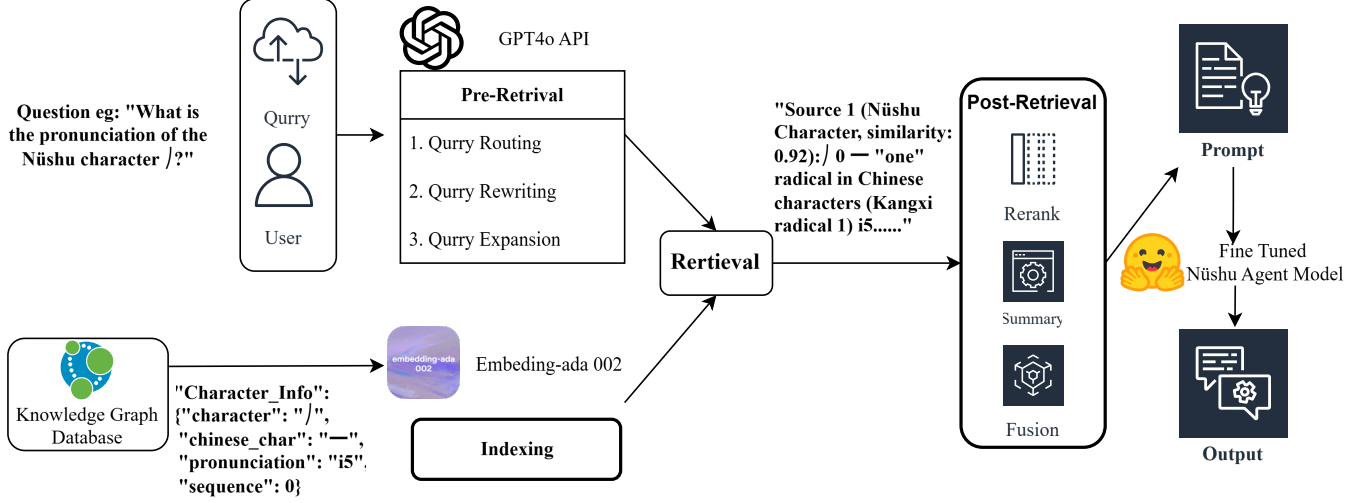
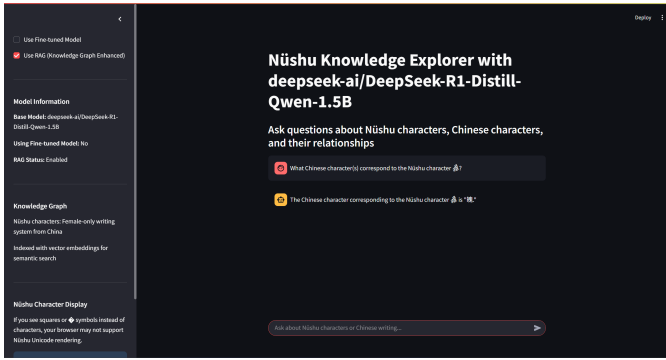Fig. 8. Retrieval-Augmented Generation System for Nüshu Analysis

Fig. 9. Streamlit-based User Interface for Nüshu Agent System showing the chat interface, model configuration options, and information sidebar

- ROUGE-1 F1: Measures the overlap of unigrams (single words) between the generated response and reference.

$$\text{ROUGE-1 F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

where precision is the ratio of matched unigrams to total unigrams in the generated text, and recall is the ratio of matched unigrams to total unigrams in the reference text.

- ROUGE-2 F1: Measures the overlap of bigrams (word pairs) between the generated response and reference, providing insight into the fluency and syntactic structure of the generated text.

- ROUGE-L F1: Uses the Longest Common Subse-

quence (LCS) approach to evaluate text similarity, capturing both lexical and partial structural information.

5.2. Model Performance Analysis

We evaluated three different models: the base DeepSeek-R1-Distill-Qwen-1.5B model without fine-tuning, the same model with LoRA fine-tuning for the Nüshu domain, and GPT-4o Mini as a commercial comparison point. Figure 10 presents a comparative analysis of these models across our evaluation metrics.

As shown in Figure 10, all three models achieved perfect character accuracy (1.0), indicating that they were all able to correctly identify the specific Nüshu characters in the test set. This demonstrates that both the base and fine-tuned models have sufficient capability to handle the specialized character identification tasks for Nüshu, which is crucial for a domain-specific knowledge system.

While character accuracy was consistent across models, significant differences emerged in the quality of the generated text. The LoRA fine-tuned model consistently outperformed both the base model and GPT-4o Mini across all ROUGE metrics, with improvements of 6.7% in ROUGE-1 F1 (0.521 vs. 0.488), 7.4% in ROUGE-2 F1 (0.373 vs. 0.347), and 6.8% in ROUGE-L F1 (0.510 vs. 0.478) compared to the base model.

GPT-4o Mini showed the lowest performance across all ROUGE metrics (0.333 for ROUGE-1 F1, 0.163 for ROUGE-2 F1, and 0.315 for ROUGE-L F1), highlighting the challenges general-purpose models face with specialized domains like Nüshu without proper context and fine-tuning.
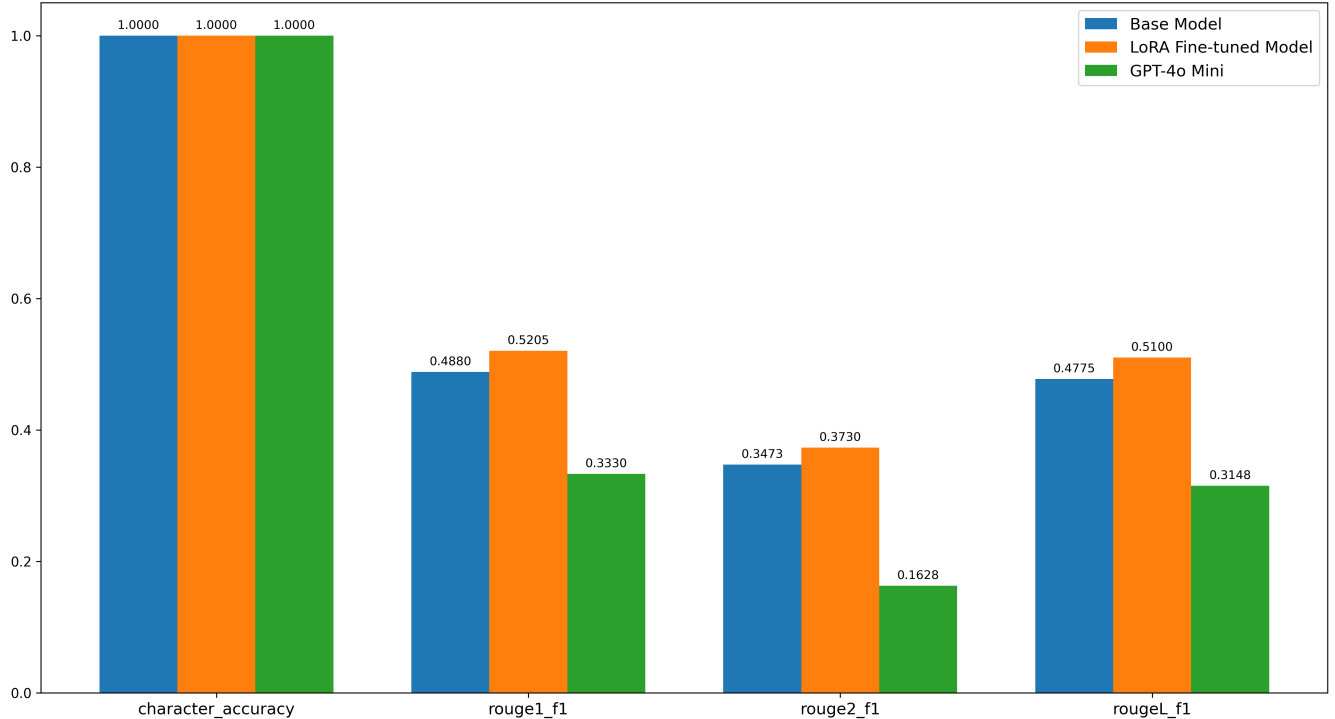
Fig. 10. Comparison of Metrics Across Different Models

These results demonstrate the effectiveness of our LoRA fine-tuning approach for the Nüshu domain. The fine-tuned model's superior performance in generating accurate, relevant, and well-structured responses validates our approach of combining knowledge graph construction with targeted model adaptation. The significant performance gap between our models and GPT-4o Mini underscores the importance of domain-specific adaptation for specialized language tasks, even when compared to commercial models.

6. CONCLUSION

My system integrates: (1) a Neo4j-based knowledge graph capturing Nüshu character relationships; (2) a retrieval-augmented generation framework combining structured knowledge with DeepSeek-R1-Distill models; and (3) efficient domain-specific adaptation using LoRA with minimal computational overhead.

Experimental results demonstrated the efficacy of our approach, with the LoRA-tuned model outperforming both the base model and GPT-4o Mini across all metrics (improvements of 6.7%, 7.4%, and 6.8% in ROUGE-1, ROUGE-2, and ROUGE-L F1 scores respectively). While all models achieved perfect character accuracy, significant differences in text quality metrics validate the importance of domain-specific adaptation for specialized

writing systems like Nüshu, even when using parameter-efficient methods like LoRA. This work contributes to digital humanities by applying modern NLP techniques to preserve endangered writing systems of cultural and historical value. Future directions include extending the knowledge graph with temporal relationships, incorporating multi-modal capabilities for processing physical manuscripts, developing cross-lingual translation beyond Chinese, and creating interactive educational tools to promote awareness of this unique cultural heritage.

By making our system publicly available, we contribute to ongoing efforts to preserve Nüshu, ensuring this remarkable writing system remains accessible to future generations. Moreover, the study of Nüshu represents more than just linguistic preservation—it highlights the extraordinary wisdom that emerged under conditions of cultural deprivation, systematic gender oppression, and limited educational opportunity. These insights from Nüshu offer valuable lessons for contemporary society's development of equality and diversity, while demonstrating the potential of combining structured knowledge representation with neural language models for specialized domains.

7. REFERENCES

[1] Qi Guang Chen, [Nu Han Zi Dian: Nüshu Chinese Character Dictionary], [Central University for Nationalities Press], 2006, This dictionary compiles over 3400 characters from original manuscripts, copies, and rubbings of Nüshu texts. It annotates the form, pronunciation, and meaning of each character, and documents their evolution into standard Chinese characters.

[2] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann, "Knowledge Graphs," ACM Computing Surveys, vol. 54, no. 4, pp. 1–37, May 2022.

[3] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 2024.

[4] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," 2025.

[5] J. E. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models," ArXiv, June 2021.

[6] Liming Zhao, [Collected Works of Chinese Nüshu (5 volumes)], 5 volumes. [Zhonghua Book Company], 2005, This is the most comprehensive compilation of Nüshu texts, the world's only known women-only script, including over 652 manuscripts with scanned facsimiles, translations, and annotations. The collection covers more than 90% of all extant Nüshu materials and provides valuable resources for linguistics, philology, literature, anthropology, folklore, and sociology.

[7] Xiurong Zhang, Kun Guo, and Jingquan He, [A Comparative Study of Nüshu Characters], [Intellectual Property Publishing House], 2006, This book provides a comparative study of Nüshu characters, including analyses of the social and cultural foundations of Chinese Marxism, its theoretical achievements, and innovations.

[8] Jovana Mitric, Igor Radulovic, Tomo Popovic, Zoja Scekic, and Sandra Tinaj, "AI and Computer Vision in Cultural Heritage Preservation," in 2024 28th International Conference on Information Technology (IT), Feb. 2024, pp. 1–4.

[9] Dessy Harisanty, Kathleen Lourdes Ballesteros Obille, Nove E. Variant Anna, Endah Purwanti, and Fitri Retrialisca, "Cultural heritage preservation in the digital age, harnessing artificial intelligence for the future: A bibliometric analysis," Digital Library Perspectives, vol. 40, no. 4, pp. 609–630, Sept. 2024.

[10] Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata, "ArCo: The Italian Cultural Heritage Knowledge Graph," in The Semantic Web – ISWC 2019, Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, Eds., Cham, 2019, pp. 36–52, Springer International Publishing.

[11] Jinhua Dou, Jingyan Qin, Zanxia Jin, and Zhuang Li, "Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage," Journal of Visual Languages & Computing, vol. 48, pp. 19–28, Oct. 2018.

[12] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu, "Qwen Technical Report," 2023.