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 (literally ”women's writing”) is a syllabic script created and used exclusively by women in the Jiangyong County of Hunan Province, China. Dating back to possibly the 15th century, Nüshu represents a rare example of a gender-specific writing system developed in response to women's exclusion from formal education in traditional Chinese society [1]. The script is characterized by its rhomboidal shapes and delicate, elongated strokes, with approximately 2,000 distinct characters identified to date. The transmission of Nüshu was primarily matrilineal, passed from mothers to daughters or among female friends, and was used for personal expression, correspondence, and recording folk songs. As China modernized and female literacy in standard Chinese became widespread, the practice of Nüshu declined dramatically, with the last proficient native writers passing away in the early 21st century. This makes the digitization and computational analysis of Nüshu particularly urgent for cultural preservation.

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 context. 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 particu-

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

larly 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 the Taiping Heavenly Kingdom period (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].

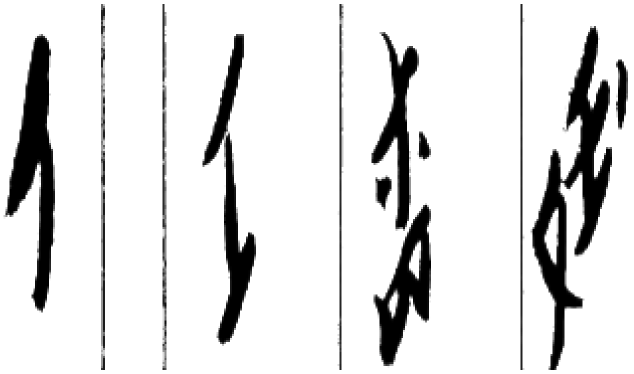

Fig. 1. The Nüshu writing of "Artificial Intelligence"

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 have emerged as powerful frameworks for representing complex domain-specific information through entity-relationship structures. Cultural heritage domains, with their rich interconnections between artifacts, historical contexts, and interpretations, are particularly well-suited for knowledge graph implementations. Carriero et al. demonstrated this effectively with ArCo, the Italian Cultural Heritage Knowledge Graph, which systematically organized cultural heritage data through seven networked modules capturing temporal, spatial, and contextual dimensions [10]. This approach provides valuable insights for our Nüshu knowledge representation, particularly in connecting characters with their cultural contexts and historical significance. For specialized domains like Nüshu, the implementation of domain ontologies within knowledge graphs significantly enhances semantic representation capabilities. Dou et al. illustrated this approach in their work on Chinese intangible cultural heritage, developing a robust knowledge graph that integrated domain-specific terminology with natural language processing techniques [11]. Their framework incorporated hierarchical taxonomies of cultural elements alongside entity extraction methods, creating a comprehensive representation system that overcame the limitations of traditional database approaches. Similarly, our Nüshu knowledge graph implements a domain-specific ontology that captures the unique characteristics of this writing system, including its visual features, phonetic properties, and semantic relationships.

Recent advancements in retrieval-augmented generation (RAG) have transformed how language models interact with specialized knowledge. Gao et al. systematically categorize RAG architectures into retriever-reader frameworks, demonstrating how they significantly mitigate hallucination issues in large language models by grounding responses in verified external knowledge [3]. This approach is particularly valuable for low-resource domains like Nüshu, where general language models lack adequate training data. Our implementation extends these principles by integrating a specialized knowledge graph as the retrieval source, providing more nuanced contextual information than traditional document-based retrieval systems.

The evolution of RAG systems has moved beyond simple document retrieval to incorporate structured knowledge representations. Contemporary approaches demonstrate superior performance in handling complex queries about specialized domains by combining dense vector retrieval with graph traversal methods. This hybrid approach allows for both semantic similarity matching and relationship-based information retrieval, addressing the multifaceted nature of queries about cultural writing systems like Nüshu. Our system builds upon these advances, leveraging both embedding-based similarity search and graph relationship patterns to provide comprehensive information about Nüshu characters and their cultural context.

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 N"ushu, 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 research has shown that combining retrieval-augmented generation (RAG) with domain-adapted LLMs further enhances performance in knowledge-intensive tasks. By integrating external knowledge sources—such as knowledge graphs or curated document collections—RAG frameworks help mitigate hallucination and improve factual accuracy, especially in domains where training data is scarce. The synergy between RAG and LoRA-based adaptation allows for both robust generalization and precise domain alignment, making it well-suited for applications in digital humanities and cultural heritage preservation.

In summary, the intersection of advanced LLM architectures, retrieval-augmented methods, and efficient adaptation techniques like LoRA provides a promising foundation for building intelligent agents capable of handling complex, domain-specific queries. Our system leverages these advances to deliver accurate and context-aware responses for N"ushu character analysis and retrieval.

3. DESCRIPTION OF MODELS

In this section, you should briefly describe the model you are using for each task, along with the rationale. You may opt to use a single learning algorithm to solve the problem or multiple ones, but bear in mind there are page limitations and that you should explain your rationale behind your choices. That is, the algorithmic description must detail your reasons for selecting a particular model.

You can clarify them with flow charts, figures or equations. An example of how to draw an image is demonstrated in Fig. ??.

3.1. System Overview and Pipeline

Our N”ushu character retrieval-augmented agent is designed as a modular pipeline that integrates knowledge representation, retrieval, and generation components. The system workflow begins with user queries, which are first processed and analyzed for intent. Relevant entities and relationships are then retrieved from the Neo4j-based knowledge graph using both semantic similarity and graph traversal techniques. The retrieved knowledge is passed to a retrieval-augmented generation (RAG) module, which combines this structured information with the generative capabilities of a fine-tuned DeepSeek-R1-Distill model. This pipeline ensures that responses are both contextually accurate and grounded in domain-specific knowledge, effectively bridging the gap between symbolic reasoning and neural language generation.

3.2. Knowledge Graph Structure

The core of our system is a Neo4j-based knowledge graph that encodes the multifaceted relationships among N”ushu characters. Each node in the graph represents a character and is annotated with attributes such as pronunciation, meaning, visual features, and historical context. Edges capture relationships including phonetic similarity, semantic equivalence, variant forms, and cultural associations. The schema is designed to support flexible queries, enabling both direct lookups and complex traversals (e.g., finding all characters with similar meanings or tracing the evolution of a character). This structured representation not only facilitates efficient retrieval but also supports downstream tasks such as visualization and statistical analysis of the script’s internal structure.

3.3. DeepSeek-R1 Model

For the generative component, we employ the DeepSeek-R1-Distill-Qwen-1.5B model, a compact yet powerful large language model optimized for reasoning and factual accuracy. DeepSeek-R1 is pre-trained on a diverse multilingual corpus and further distilled for efficiency, making it suitable for integration with retrieval-augmented frameworks. In our system, the model is further adapted to the N”ushu domain using LoRA fine-tuning, allowing it to generate context-aware and culturally informed responses. The combination of DeepSeek-R1’s strong generalization ability and domain-specific adaptation ensures high-quality outputs even for complex or low-resource queries.

Task	Model	Train Acc	Val Acc	Test Acc
A				
B				
...				
...				

4. IMPLEMENTATION

This section must provide the detailed implementation of your models. In particular, you must provide the name and use of external libraries, explain hyperparameter selection, training pipeline (if any) and key modules/classes/functions/algorithms.

You also must provide a detailed description of the dataset (content, size, format, etc.), any data preprocessing that was applied and how you separate your dataset into training, validation and test sets.

The execution of your models also should be reported here. In particular, this section should include a thorough discussion on the training convergence and stopping criterion (it is recommended that learning curves graphs be used to this effect).

4.1. Task A: the task name

4.1.1. module name

Hello world!

4.2. Task B: the task name

Hello world! ...

5. EXPERIMENTAL RESULTS AND ANALYSIS

This section describes and discusses your results. Additionally, this section should include accuracy prediction scores on a separate test dataset, provided by the module organizers, but not used during your training and validation process.

We recommend you use a table to list the tasks, models and results before analysis.

6. CONCLUSION

This last section summarizes the findings and suggests directions for future improvements.

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.