International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

DIANGPT: A Review on a Domain-Specific Question Answering System Using LORA and RAG

Pradeep Nayak, Nischitha, Omkar KS, Omkar JK, Pallavi SK
Department of Information Science and Engineering,
Alva's Institute of Engineering and Technology, Mijar, Karnataka, India.

***Abstract*—DianGPT is a domain-specific question-answering (QA) framework designed to transform institutional archival content into an accurate, context-aware conversational knowledge system. Integrating Low-Rank Adaptation (LoRA), Retrieval-Augmented Generation (RAG), FAISS-based retrieval, and structured dataset engineering, the system optimizes both precision and computational efficiency. This review evaluates DianGPT's architectural pipeline, including dataset processing, supervised fine-tuning, retrieval workflows, and automated evaluation strategies. Further, the paper critically analyzes the reported experimental outcomes, highlights methodological strengths, and identifies gaps and potential improvements. The work serves as a comprehensive reference for researchers developing lightweight, domain-optimized large language models.**

***Index Terms*—Domain-Specific QA, LoRA, Retrieval-Augmented Generation, Dataset Engineering, Large Language Models.**

I. INTRODUCTION

Large Language Models (LLMs) have transformed natural language processing by enabling machines to understand and generate human-like text across diverse tasks. Despite this progress, general-purpose LLMs face limitations when applied to highly specialized domains. These environments contain unique terminology, historical context, and structural dependencies that are absent from mainstream datasets. As a result, generic LLMs often hallucinate facts, misinterpret domain-specific terminology, or provide incomplete responses when tasked with answering queries derived from institutional archives.

Organizations such as academic institutions, enterprises, and government bodies maintain extensive document collections including reports, minutes, guidelines, yearbooks, and policy documents. These archives represent valuable knowledge but are difficult to access through traditional keyword search mechanisms. The DianGPT model addresses these challenges through a hybrid approach, combining LoRA-based fine-tuning with Retrieval-Augmented Generation (RAG) and structured dataset engineering. This framework enables efficient adaptation of large models to domain-specific environments without requiring full retraining.

This review provides a detailed analysis of the DianGPT system's pipeline and highlights the significance of integrating retrieval mechanisms with parameter-efficient fine-tuning to build reliable QA systems.

II. RELATED WORK

Domain-specific QA research spans multiple directions including model adaptation, retrieval-augmented systems, and evaluation frameworks.

A. Domain Adaptation

Early models such as SciBERT demonstrated that training on domain-specific corpora significantly improves performance. Gururangan et al. further introduced domain-adaptive pretraining (DAPT), which enhanced models' alignment with specialized text by exposing them to domain-relevant documents prior to fine-tuning.

B. Parameter-Efficient Fine-Tuning

Full LLM fine-tuning is computationally expensive. LoRA (Low-Rank Adaptation) revolutionized fine-tuning by updating only low-rank weight matrices while keeping the pretrained weights frozen. This reduces memory usage, improves training speed, and enables deployment on mid-range GPUs.

C. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances factual precision by retrieving relevant context before generation. Dense Passage Retrieval (DPR) and FAISS-based similarity search have become standard for embedding-based document retrieval due to their scalability and high recall.

D. Evaluation Strategies

Lexical metrics such as BLEU and ROUGE provide coarse-grained similarity scores but struggle to evaluate correctness in domain-specific QA. Recent work has shifted toward LLM-based judge scoring that correlates strongly with human evaluations. DianGPT adopts a similar approach by correlating judge-model scores with expert annotations.

III. PROBLEM AND OBJECTIVES

A. Problem Statement

Institutional archives contain a wide variety of document types, formats, and domain-specific content. Conventional LLMs, trained on broad and generic text, are ill-equipped to respond accurately to queries referencing such information. Furthermore, the computational requirements of full-scale fine-tuning inhibit adoption in resource-limited environments. Additionally, archival data often contains redundancy, inconsistency, and noise that necessitate extensive preprocessing.

B. Objectives

DianGPT establishes the following objectives:

- Enhance the accuracy and specificity of answers for domain-related queries.
- Utilize LoRA to enable parameter-efficient fine-tuning suitable for moderate hardware.
- Integrate retrieval mechanisms to reduce hallucination and improve grounding.
- Implement a scalable automated evaluation system aligned with human judgment.

IV. DIANGPT SYSTEM OVERVIEW

A. Dataset Engineering

Dataset preparation involves several key steps:

- **Segmentation**: Archival documents are divided into meaningful units such as sections or paragraphs.
- **Noise Filtering**: Non-text content and inconsistencies are removed to maintain high-quality inputs.
- **Entity Extraction**: LLM-assisted extraction identifies roles, events, organizations, and relationships.
- **MinHash Deduplication**: Removes redundant entries to prevent bias and overfitting.
- **Synthetic Q&A Generation**: Multi-turn question-answer pairs are generated to enhance supervised fine-tuning quality.

The result is a clean, structured dataset optimized for downstream model training and retrieval indexing.

B. Supervised Fine-Tuning (LoRA)

LoRA enables fine-tuning large models by injecting trainable low-rank matrices into the attention layers. This adaptation significantly reduces GPU memory requirements and training time. Unlike full fine-tuning, LoRA preserves the general-purpose knowledge of the pretrained model while adding domain-specific patterns learned during supervised training.

C. Retrieval-Augmented Generation (RAG)

RAG enhances QA accuracy by retrieving relevant passages before generating answers. DianGPT uses BCEmbedding for encoding and FAISS for efficient search. The top retrieved passages are re-ranked and appended to the model input, enabling grounded and context-consistent responses.

D. Evaluation Framework

DianGPT employs an LLM-based judge scoring system to evaluate output quality across criteria such as accuracy, completeness, and coherence. The reported 0.728 correlation between judge-model scores and human ratings indicates strong evaluation reliability. Rouge-L metrics provide additional structure-based comparison.

V. EXPERIMENTAL FINDINGS

TABLE I
REPORTED DIANGPT PERFORMANCE METRICS

Strategy	Score	Rouge-L	Latency(s)
Base Model	3.087	0.137	3.132
SFT (LoRA)	3.273	0.164	2.856
Base + RAG	3.587	0.242	6.779
SFT + RAG	3.329	0.217	6.123

The findings indicate:

- RAG provides the highest improvement in accuracy and Rouge-L metrics.
- LoRA reduces inference latency and enhances task-specific reasoning.
- Combined SFT + RAG configuration shows interference effects requiring further study.
- Retrieval increases computational cost but significantly enhances factual grounding.

VI. CRITICAL ANALYSIS

A. Strengths

- Modular pipeline ensures reproducibility and ease of deployment.
- LoRA fine-tuning enables resource-efficient domain adaptation.
- RAG significantly reduces hallucination by grounding responses.
- Automated judge scoring enhances scalability of evaluation.
- The framework suits real institutional environments with limited computational resources.

B. Limitations

- Evaluation dataset size is limited, affecting generalizability.
- Entity extraction accuracy depends on upstream LLM quality.
- Retrieval latency increases with larger archival datasets.
- SFT–RAG interaction is poorly understood and needs ablation studies.

VII. FUTURE WORK

Future improvements include:

- Expanding evaluation datasets to cover broader query diversity.
- Incorporating multimodal retrieval, including images and scanned documents.
- Enhancing multi-turn reasoning using memory-augmented networks.
- Exploring hybrid sparse-dense retrieval architectures.
- Integrating user feedback loops for real-time system refinement.

VIII. CONCLUSION

DianGPT demonstrates the effectiveness of combining parameter-efficient fine-tuning and retrieval augmentation to construct a robust domain-specific QA system. With LoRA reducing training overhead and RAG improving factual accuracy, the system provides a viable blueprint for deploying efficient and reliable domain-aligned LLMs. Continued advancements in retrieval, dataset expansion, and evaluation methodologies will enhance its applicability across diverse institutional do- mains.

REFERENCES

1) Chen, X., Jiang, Y., Pei, H., Hei, X., "DianGPT: Bridging Precision and Efficiency for a Domain-Specific Question-Answering System," 2025.
2) Beltagy, I., Lo, K., Cohan, A., "SciBERT: A Pretrained Language Model for Scientific Text," 2019.
3) Gururangan, S. et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," ACL, 2020.
4) Hu, J. E., et al., "LoRA: Low-Rank Adaptation of Large Language Models," 2021.
5) Lewis, P., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020.
6) Karpukhin, V., et al., "Dense Passage Retrieval for Open-Domain QA," 2020.
7) Papineni, K., et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," 2002.
8) Lin, C.-Y., "ROUGE: A Package for Automatic Evaluation of Summaries," 2004.
9) Kamalloo, E., et al., "Evaluating Open-Domain Question Answering in the Era of Large Language Models," 2023.