A MEDICAL-RESEARCH SPECIFIC RESEARCH METHOD QUERYING SYSTEM WITH LORA

Sinjoy Saha
Penn State University
sks7620@psu.edu

Xin Dong
Penn State University
xjd5036@psu.edu

1 Research Objective

This project introduces a Medical-research-specific Research Method Querying System (M-RMQ), which generates structured methodological content—including background, objective, and methods—based on biomedical research questions. We fine-tune a pre-trained LLaMA 3.2 model with 1 billion parameters ("Llama-3.2-1B") using Low-Rank Adaptation (LoRA), enabling efficient task adaptation with minimal parameter updates. Training data is derived from a curated corpus of medical literature, reformatted into instruction-style input-output pairs. The resulting M-RMQ model is optimized to reflect the structure and language of real-world biomedical methods. We evaluate model performance using BLEU and ROUGE for surface-level accuracy, and BERTScore for semantic alignment. The overall pipeline is illustrated in Figure 1.

To ensure reproducibility and transparency, we have made the code files for this project publicly available on our GitHub repository, accompanied by detailed documentation to facilitate the replication of our results [1] [2].

2 Dataset Construction

For this study, we use the PubMedQA dataset [1], a biomedical question-answering corpus constructed from PubMed abstracts. It comprises three subsets: PQA-L (expert-labeled), PQA-A (automatically generated), and PQA-U (unlabeled). We utilize the PQA-L subset, which includes 1,000 high-quality expert-annotated samples. Each sample contains a research question derived from a biomedical article title, along with a long answer corresponding to the article's conclusion. Although originally intended for QA classification, the long answer often includes methodological insights and outcome summaries. This makes the dataset well-suited for instruction tuning of large language models (LLMs) to generate research methodologies conditioned on biomedical research questions.

To prepare the dataset for our task, we apply the following pre-processing steps:

- **Filtering**: Instances with missing/empty long answers are excluded to ensure complete outputs. Samples lacking details like either "objective" and "background context" or "Methodology" content are removed.
- **Instruction Formatting**: Each input is reformatted into an instruction-style prompt containing: 1) "Research Question"—the central inquiry driving the study; 2)"Introduction" that outlines the study's "objective" and "background context" details; and 3) "Methodology"—the primary methodology or approach used. Additional details such as "participants", "settings", or "datasets" are intentionally omitted to maintain consistency, as these elements are not uniformly available across all papers. This approach ensures standardized prompts that focus exclusively on the core components present in every study.
- **Tokenizing**: Prompts and responses are tokenized using the tokenizer corresponding to the base language model, with a maximum sequence length of 512 tokens. Padding and truncation are applied as necessary.

[1] https://github.com/sinjoysaha/CSE587-midterm-project.git
[2] https://github.com/XDNG2024/CSE587-midterm-project.git

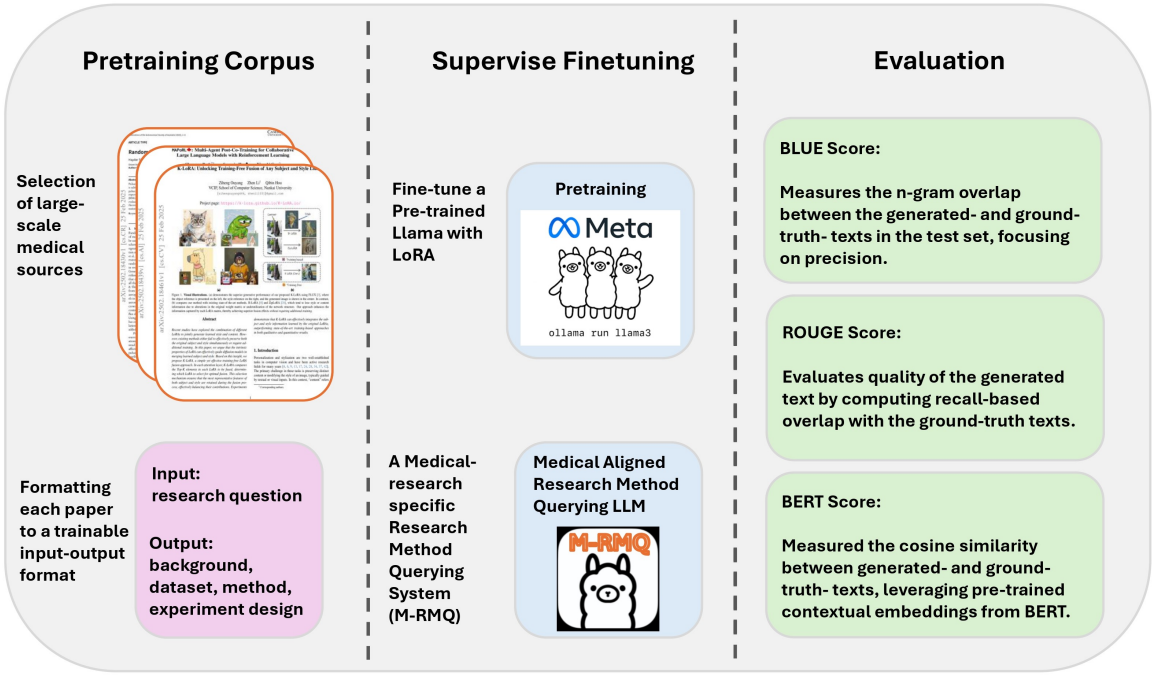

Figure 1: Framework of A Medical-research specific Research Method Querying System (M-RMQ)

- **Train-Test Splitting**: We reserve 20 percent of the data for testing and 80 percent of the data for training. Each example is formatted as an instruction–response pair, where the input is a research question, and the output is the corresponding methodological description.

To evaluate whether the 512-token limit adequately accommodates the data, we analyze the character length distribution of the concatenated "Introduction" and "Methodology" features. As shown in Figure 2, most samples fall between 500 and 1000 characters, corresponding to approximately 125–250 tokens. A few longer samples exceed 1500 characters (about 375 tokens). The figure confirms that the outputs are sufficiently detailed yet compatible with the model's input constraints for instruction fine-tuning.

3 LLM Selection and Training Details

We used a base model as the Llama 3.2 model with 1 billion parameters, which is "meta-llama/Llama-3.2-1B", and we adopted the Low-Rank Adaptation (LoRA) technique to fine-tune the pre-trained LLM.

3.1 LoRA Fine-Tuning Methodology

Large language models often consist of billions of parameters, making full-parameter fine-tuning computationally expensive and memory-intensive. Traditional fine-tuning updates all weights of the model, requiring significant GPU memory and storage. To address this, we adopt LoRA, a parameter-efficient fine-tuning technique that injects trainable low-rank matrices into the attention layers of pre-trained transformer models while freezing the original model weights. In a typical transformer model, attention layers contain weight matrices such as $W_q \in \mathbb{R}^{d\times k}$ and $W_v \in \mathbb{R}^{d\times k}$ for the query (q) and value (v) projections, respectively. LoRA replaces these with the following modified formulation during training:

$$W_q^{\text{LoRA}} = W_q + \Delta W_q = W_q + BA$$

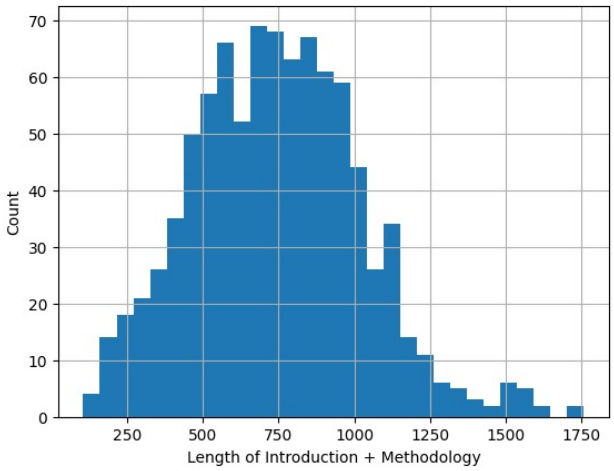

Figure 2: Characteristics of Introduction and Methodology

$$W_v^{\text{LoRA}} = W_v + \Delta W_v = W_v + B'A'$$

where:

- $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{k \times r}$ are low-rank matrices ($r \ll d, k$),
- $\Delta W_q = BA$ and $\Delta W_v = B'A'$ are the trainable low-rank updates,
- W_q and W_v are kept frozen during training.

These low-rank updates are learned while keeping the original weights unchanged, allowing efficient storage and inference without degrading model performance. We use a causal language modeling objective to maximize the log-likelihood of the output tokens given the input prompt.

$$\mathcal{L}_{\text{LoRA}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x; \theta, \phi)$$

where:

- x is the input prompt (research question),
- $y_{1:T}$ is the target response (methodological explanation),
- θ are the frozen base model parameters,
- ϕ are the trainable LoRA parameters (A, B).

3.2 Training Details

Table 1 summarizes the key training configurations and hyperparameters used during the fine-tuning of the language model using LoRA. The setup includes core training parameters such as batch size, learning rate, and number of epochs, as well as LoRA-specific settings like the rank (r), scaling factor (α), and targeted modules for adaptation. We set the rank to 8 and applied LoRA to the query and value projection layers with a dropout rate of 0.05. The base model was loaded in 4-bit precision using FP16 data types to reduce memory consumption. This configuration balances computational efficiency with adaptation quality, enabling effective instruction tuning on limited resources.

4 Evaluation Metrics

To evaluate the effectiveness of our fine-tuned model in generating high-quality research methodologies, we conduct a series of experiments using a held-out test set. The model outputs are assessed using a suite of standard natural language generation metrics, including BLEU, ROUGE-L, and BERTScore, which capture different aspects of textual overlap and semantic similarity with human-authored reference answers:

Table 1: LoRA Configuration

Category	Parameter	Value
Training Setup	Per device train batch size	4
	Gradient accumulation steps	8
	Number of training epochs	10
	Learning rate	2e-4
	Logging steps	10
	Save steps	200
	Save total limit	3
	FP16	True
	BF16	False
LoRA Hyperparameters	Rank (r)	8
	LoRA alpha	16
	LoRA dropout	0.05
	Bias	none
	Task type	CAUSAL_LM
	Target modules	["q_proj", "v_proj"]
Model Configuration	Load in 4bit	True
	Torch dtype	torch.float16

- **BLEU (Bilingual Evaluation Understudy) Score**: Measures the n-gram overlap between the generated method (including methodology and others) and the ground-truth texts in the test set, focusing on precision.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score**: Evaluates the quality of the generated text by computing recall-based overlap with the ground-truth texts.
- **BERT Score**: Measured the cosine similarity between generated- and ground-truth- texts, leveraging pre-trained contextual embeddings from BERT.

5 Experiments and Analysis

In this section, we include the training curves and report evaluation metrics used to assess model performance, including BLEU, ROUGE-L, precision, recall, and BERTScore. We also present an example used for instruction, along with the output from the pre-trained language model and our fine-tuned model.

Table 2 presents a comparison of evaluation metrics between the pre-trained and fine-tuned models across different content sections. While both models achieve full section coverage, the fine-tuned model demonstrates improved semantic alignment and more balanced precision-recall tradeoffs. For the Introduction section, it outperforms the pre-trained model in ROUGE-L (0.0759 vs. 0.0641), and achieves more balanced precision (0.6113 vs. 0.6939) and recall (0.6047 vs. 0.5682), resulting in a higher BERT Score (0.6075 vs. 0.6209). In the Methodology section, although BLEU drops to 0, the fine-tuned model slightly improves on BERT Score (0.5427 vs. 0.5281), indicating better semantic alignment despite lexical sparsity. Overall, the fine-tuned model provides a competitive or improved BERT Score (0.6069 vs. 0.5959) while maintaining similar recall, suggesting enhanced contextual understanding. These results support the effectiveness of instruction fine-tuning with LoRA for improving content quality in a structured research generation task.

Table 2: Evaluation Results Across Models and Sections

Model Type	Section	BLEU	ROUGE-L	Precision	Recall	BERT Score
Pre-trained Model	Overall	0.0703	0.1961	0.6199	0.5772	0.5959
	Introduction	0.0641	0.2370	0.6939	0.5682	0.6209
	Methodology	0.0145	0.1483	0.5297	0.5298	0.5281
Fine-tuned Model	Overall	0.0577	0.1805	0.6118	0.6033	0.6069
	Introduction	0.0759	0.2057	0.6113	0.6047	0.6075
	Methodology	0.0000	0.1280	0.5334	0.5566	0.5427

Figure 3 illustrates the training dynamics of our LoRA fine-tuned language model over 10 epochs. Figure 3 (a) shows a rapid decrease in training loss during the first two epochs, followed by a stable convergence around 1.3, indicating effective learning. Figure 3 (b) presents the average token accuracy, which improves significantly in the early stages and converges near 74%, reflecting consistent token-level predictions. Figure 3 (c) depicts the gradient norm, where an initial spike is followed by stabilization, suggesting improved gradient flow and convergence. Finally, Figure 3 (d) shows the linear decay of the learning rate throughout training, following a scheduled strategy to enhance generalization and prevent overfitting. These trends collectively confirm that the model successfully learns from the instruction-tuning data with stable optimization behavior.

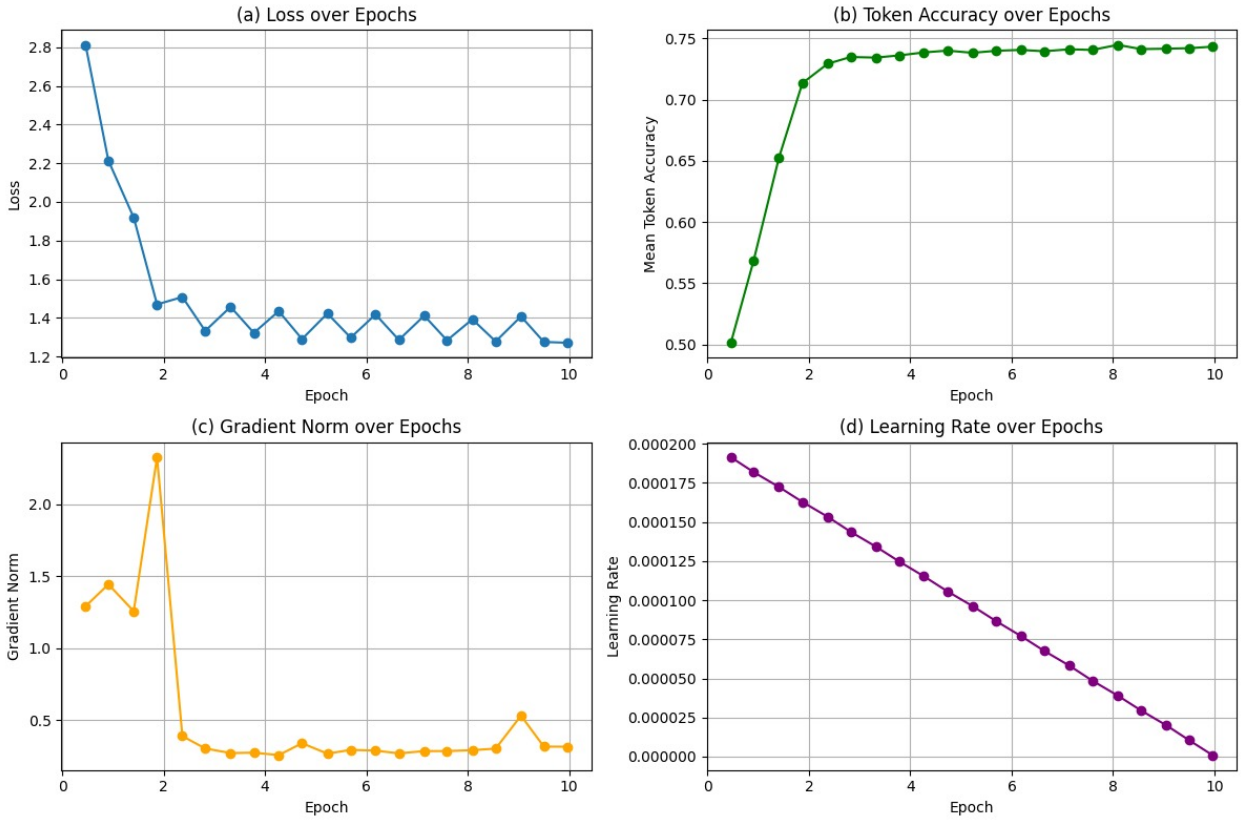

Figure 3: Training Curve

We have a total of 190 test samples (all results are attached in GitHub), and here we present one of them randomly:

An Example of Instructions

$< bos >< start\_of\_turn >$ user

You are a helpful assistant. Given the following research question, please provide a methodology to conduct research on it which would enable the researcher to answer the question.

Research Question: Preoperative locoregional staging of gastric cancer: is there a place for magnetic resonance imaging?

Please be concise within 2-4 sentences for Introduction and Methodology. Do not use markdown format. Format the output as:
Introduction:
Methodology:

$< end\_of\_turn >$

Expected Output of the Example

Introduction: The aim of this study was to prospectively compare the diagnostic performance of magnetic resonance imaging (MRI), multidetector computed tomography (MDCT) and endoscopic ultrasonography (EUS) in the preoperative locoregional staging of gastric cancer.

Methodology: This study had Institutional Review Board approval, and informed consent was obtained from all patients. Fifty-two patients with biopsy-proven gastric cancer underwent preoperative 1.5-T MRI, 64-channel MDCT and EUS. All images were analysed blind, and the results were compared with histopathological findings according to the seventh edition of the TNM classification. After the population had been divided on the basis of the local invasion (T1-3 vs T4a-b) and nodal involvement (N0 vs N+), sensitivity, specificity, positive and negative predictive value, and accuracy were calculated and diagnostic performance measures were assessed using the McNemar test.

Pre-trained Model Output of the Example

Introduction: Gastric cancer surgery often necessitates precise staging to determine the extent of the disease. Magnetic resonance imaging (MRI) offers superior visualization of tumors compared to conventional imaging, potentially improving surgical planning and treatment decisions. This research will investigate the feasibility and efficacy of MRI staging for preoperative locoregional gastric cancer.

Methodology: This study will utilize a retrospective review of patient data from a database of gastric cancer surgeries, focusing on cases where MRI staging was performed prior to surgery. Data will be analyzed to assess the correlation between MRI findings and surgical margins, tumor size, and overall disease extent. Statistical analysis will be employed to determine if MRI staging significantly improves surgical planning and treatment outcomes.

Fine-tuned Model Output of the Example

Introduction: The preoperative locoregional staging of gastric cancer is a critical step in determining the appropriate treatment strategy for patients with this disease. Magnetic Resonance Imaging (MRI) has been increasingly recognized as a valuable tool in this process due to its high sensitivity and specificity for detecting tumor recurrence and metastasis. This study aims to evaluate the role of MRI in preoperative locoregional staging of gastric cancer.

Methodology: This study will involve a retrospective review of 100 consecutive patients with gastric cancer who underwent preoperative MRI before undergoing surgery. The MRI scans will be analyzed using standard imaging protocols, and the location of the tumor, lymph nodes, and metastasis will be assessed. The accuracy of the MRI-based staging will be compared to the traditional preoperative staging methods, including CT scans and clinical examination. The results will be used to determine whether MRI plays a place in the preoperative locoregional staging of gastric cancer.

The comparison between the outputs of the pre-trained and fine-tuned models reveals notable differences in both specificity and coherence. The pre-trained model produces a structurally valid response, but its content remains relatively generic. While it correctly frames the clinical relevance of MRI in gastric cancer staging (as shown by the blue-highlighted text describing MRI's potential to improve surgical planning), it lacks concrete details regarding patient population, imaging protocol, and comparative evaluation. In contrast, the fine-tuned model provides a more targeted and context-aware output. The blue-highlighted segments illustrate that it specifies a concrete study design involving 100 consecutive patients, defines standard imaging protocols for tumor, lymph node, and metastasis assessment, and clearly compares MRI and conventional staging methods. Furthermore, the introduction from the fine-tuned model articulates the clinical importance of MRI with greater precision and justification.

These differences suggest that instruction fine-tuning significantly improves the model's ability to generate detailed, relevant, and medically grounded research methodologies.

6 Our Thoughts

This project provided valuable insights into the process of instruction fine-tuning using parameter-efficient adaptation methods, specifically LoRA. By applying LoRA to a pre-trained language model, we were able to achieve notable improvements in semantic alignment and generation quality across different research sections, such as the Introduction and Methodology, while keeping the computational overhead low. One key takeaway is the effectiveness of LoRA in adapting large language models to specialized tasks without requiring full model retraining. The significantly smaller number of trainable parameters enabled faster experimentation and resource-efficient fine-tuning, making it a practical choice for academic and real-world applications alike.

Through detailed comparison between pre-trained and fine-tuned outputs, we observed that fine-tuning substantially enhanced the model's ability to produce more detailed, contextually relevant, and medically grounded research methodologies. Specifically, the fine-tuned model demonstrated greater specificity by incorporating concrete study designs, patient cohort information, comparative evaluation strategies, and domain-specific justifications—capabilities that were largely absent in the pre-trained baseline. Highlighted text segments in our examples clearly illustrate this improvement, confirming that instruction fine-tuning not only improves surface-level similarity metrics but also meaningfully enriches content quality and scientific soundness.

Throughout the project, we encountered several challenges, including tuning the right balance between precision and recall, and ensuring that improvements in metrics like BLEU or ROUGE corresponded to actual improvements in content quality. Additionally, we noticed that lexical metrics sometimes failed to capture the full semantic richness of generated text, emphasizing the importance of using BERT-based evaluations to assess deeper semantic alignment.

In future work, we plan to extend the system by introducing multiple specialized agents to collaboratively generate research instructions. Each agent will focus on a specific aspect of the research process—for example, one agent may specialize in identifying relevant datasets, another in outlining experimental design, and another in suggesting evaluation protocols. These agents will operate either sequentially or in parallel and communicate through a shared memory or coordination mechanism. By incorporating multiple agents with domain-specific expertise, we aim to produce more comprehensive, modular, and context-aware research methodologies tailored to diverse scientific questions.

References

[1] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.