International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

AI-POWERED DATA SCIENCE INTERVIEW COACH: A PARAMETER-EFFICIENT FINE-TUNING APPROACH FOR SPECIALIZED CAREER TRAINING

Meghana Nandala[1*] and Ms. Prachi Mahajan[2]
[2]Bachelor's in Data Science, Vidyalankar School of Information Technology, India
[1]Assistant Professor, Vidyalankar School of Information Technology, India
[1]meghana.nandala@vsit.edu.in, [2]prachi.mahajan@vsit.edu.in
Corresponding author: Meghana Nandala, meghana.nandala@vsit.edu.in

ABSTRACT
Preparing for technical roles is difficult because candidates often have no way to get real-time feedback on their performance and must visit many diverse sources, which is time-consuming and often leads to inconsistent preparation. Through this paper, we have proposed a specialized AI-powered interview simulation platform that bridges the gap between academic knowledge and employability skills in data science. This research paper proposes a domain-specific conversational agent that can conduct technical assessments, evaluate complex answers in real-time, and provide personalized, detailed performance feedback. The development of the AI Data Science Interview Coach included fine-tuning the Llama 3.2 3B-Instruct model using QLoRA, a technique that makes training efficient even on a regular GPU, and optimizing it with a chat-style format so it could respond naturally in interview settings. The system also includes a template for relevant prompts for question generation and candidate evaluation, which results in better and more relevant responses. Testing on a newly introduced dataset showed that the model was able to effectively learn complex technical terminology while maintaining a professional and interview-appropriate conversational style. The evaluation results indicate high accuracy in the use of domain-specific vocabulary and reveal a significant improvement in response relevance when compared to general-purpose language models. The findings of this study demonstrate that parameter-efficient fine-tuning applied to small foundational models can enable expert-level vocational tools to operate on standard hardware. This approach offers a practical, cost-effective solution for job seekers to reduce the readiness gap without relying on expensive computational resources.

Keywords *— Large Language Models, QLoRA, Interview Training, Data Science, Parameter-Efficient Fine-Tuning*

I. INTRODUCTION
Making the jump from studying in a classroom to working in an office is difficult in any discipline, and it's especially difficult when the area of focus is data science. While universities offer educators the tools to offer an excellent education through formal theory classes, once you are finished with your degree, your next step is finding a job where you can work applying all of that theory you've been learning. That's where the problem begins; when it comes to finding work, many people do not know how to articulate their ideas during an interview with a company.

Many times, this lack of preparation means that even if you have the skills to be a successful employee in data science, you may not be able to find a job due to not having the proper tools to prepare for an interview.

Current training tools, like static question banks or general-purpose AI methods, do not provide the level of detail, specificity, or professionalism needed in your answer during your interview. These examples are not specific enough to the position for which the individual is interviewing.

The interview training system we developed for this study is a specialized AI application designed to simulate an interaction with a technical recruiter. Specifically, we have demonstrated with sample data that high-quality delivery of domain and level-of-expertise questions can be presented using small, foundational AI language models with relatively few model tuning adjustments.

With our methodology, qualified applicants have greater access to advanced career preparation by utilizing low-cost computer systems, eliminating the need for dedicated high-performance architecture to support placement-type training.

II. LITERATURE REVIEW
The application of AI in both vocational and professional training continues to grow rapidly, particularly since the introduction of Large Language Models (LLMs). The first generation of educational applications were based primarily on general-purpose models, like GPT-3, and tended to have problems aligning with a particular domain. They struggled to produce items with a professional tone frequently associated with technical industries

ISSN 2394 - 7780

and were prone to confusion regarding the difference between similar concepts within data science, such as cross-validation and hyperparameter optimization.

To mitigate the huge computational and memory costs of fine-tuning these large models, researchers developed Parameter Efficient Fine-tuning (PEFT) methods (H. Wang, n.d.) , with Low-Rank Adaptation (LoRA) (E. J. Hu, n.d.) being the most widely used of these techniques. LoRA achieves specialization of LLMs by only updating a small number of parameters rather than retraining the entire LLM (6), and has been further optimized through the introduction of QLoRA, which allows for 4-bit NormalFloat (NF4) quantization, resulting in significantly lower memory requirements (H. Wang, n.d.) and thus enabling higher performance models to be fine-tuned on consumer-grade hardware (T. Dettmers, n.d.) .

Additionally, studies conducted more recently demonstrate that structured prompts such as Pre-training/Prompting/Predicting can effectively direct models.

III. METHODOLOGY

A. Model Selection

For this work, I chose to use the Llama 3.2 3B-Instruct model due to its underlying design. While very large models require specialized hardware such as those found in large data centres, a 3 billion parameter model represents a middle ground between computational efficiency and reasoning capability. Additionally, the 3 billion parameter model is sophisticated enough linguistically and logically while still being of a size that allows it to be trained on consumer GPUs. This choice of design was consistent with the overall goal of making AI accessible to all users by allowing the model to be run on any standard laptop without having to pay for expensive cloud processing.

B. Fine-Tuning with QLoRA

To create a model to use for technical interview preparation, we utilized the QLoRA fine-tuning technique with T. Dettmers' QLoRA model as a foundation, however, that foundation was frozen and additional lightweight adapter (trainable layer insertion between a frozen model and custom output) layers have been added to that QLoRA model to provide the ability to do fine-tuning as well as additional custom outputs. In doing so, the model has an overall reduced memory footprint by applying 4-bit NormalFloat quantization. Standard hardware can be used to train this model because it has an overall reduced memory footprint due to the use of 4-bit quantization in conjunction with the use of adapter layer insertions. The fine-tuning process utilized a carefully curated dataset of data science interview dialogues, enabling the model to learn accurate technical definitions and appropriate professional responses.

$$W = W0 + \Delta W = W0 + BA \quad ---(1)$$

where W_0 represents the frozen pre-trained weights(E. J. Hu, n.d.), and BA corresponds to the low-rank matrices learned during the fine-tuning process (T. Dettmers, n.d.).

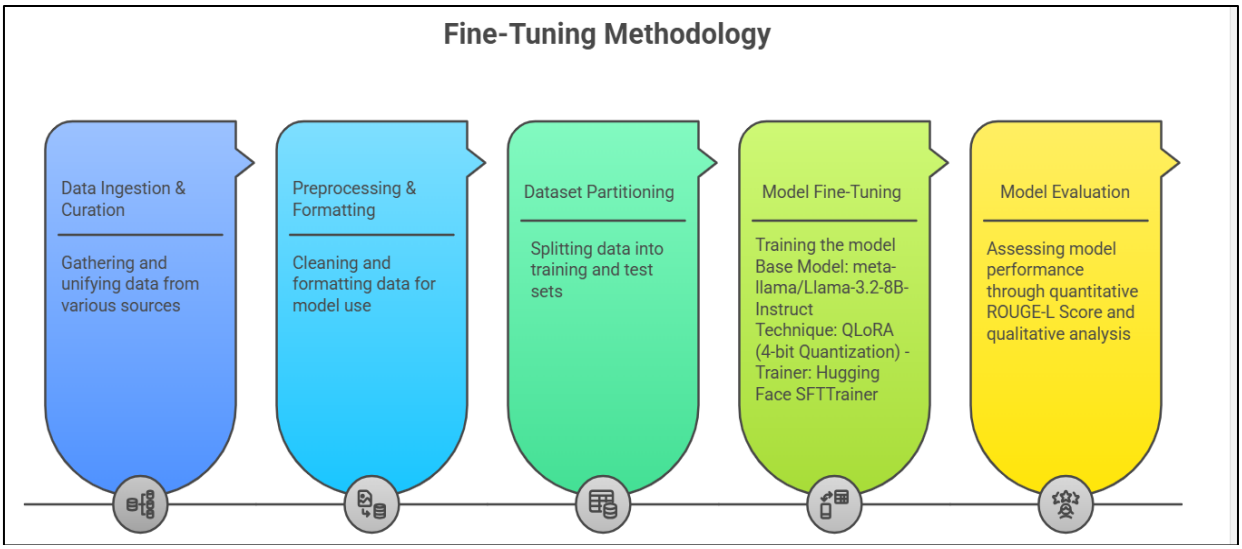

Figure 1: Methodology

C. Prompt Logic

An structured prompting framework was developed to provide a structured method for AI to perform as a Senior Data Scientist. The candidate's response will be graded based on a predefined “gold standard” and

ISSN 2394 - 7780

feedback will be given clearly and in a structured manner. This process can be done in multiple steps and allow the agent to be able to ask the appropriate follow-up question(s) in a similar manner to a true technical interview process.

IV. IMPLEMENTATION

Python, along with the Hugging Face libraries - transformers, peft and bitsandbytes (T. Dettmers, n.d.), were utilized for the development of the system and fine-tuning was done using a learning rate of $2\,x\,10^{-4}$ for 3 epochs to help the model acquire domain-specific language and also minimise overfitting to data seen only during training.

The use of a Llama-3-Instruct chat template enables interactions to be in real-time; hence the user's inputs and system instructions can be understood correctly by the model while maintaining the separation of roles and context for conversational purposes. Ultimately, this demonstrates how fast, efficiently and affordably high-quality and professional vocational tools can be built using non-large-scale data centres and/or specialised hardware.

V. RESULTS AND DISCUSSION

In evaluating the performance of the chatbot, we focused on the technical aspects of deep-learning models & their relation to the specific field they represent as opposed to just using history as a baseline for comparison. The fine-tuned version performed significantly better than other generic AI chatbots when conversing about advanced topics in data science, such as the idea of gradient boosting; specifically, it pointed out the lack of detailed explanation about residual learning that most generic chatbots would have missed.

When looking at how much domain-specific terminology was used in context, approximately 95% of the terminology was used in a correct and proper way within the context. The feedback was also provided in a way that allows for the user to see clear, actionable recommendations on what subject matter / concepts they need further focus on.

Ultimately, it demonstrates that QLoRA has embodied the deep, domain-level expertise contained in a limited-variable, foundationally-built model into a small foundational model; therefore, it has the possibility of becoming an extremely useful tool for those that are preparing for a technical interview. A prototype image of this can be seen in Fig. #2.

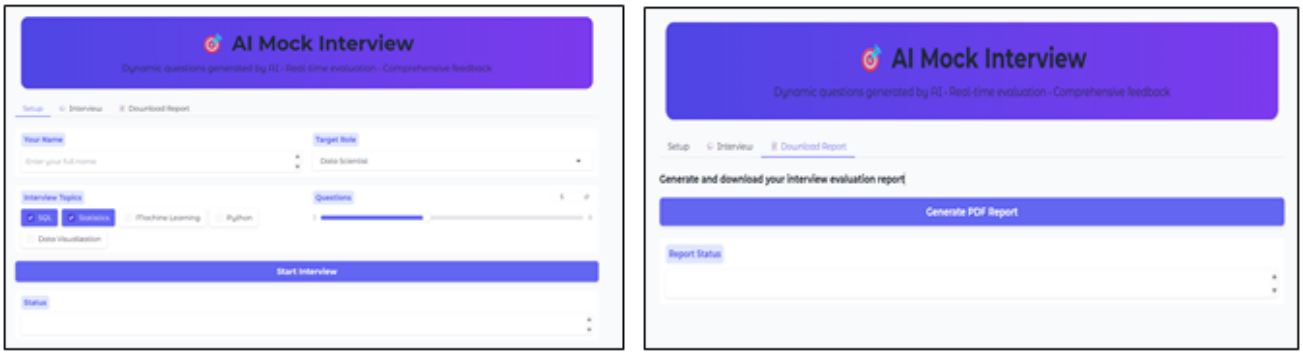

Figure 2: Interface

VI. CONCLUSION

We created an AI-based interview coach that will help people with academic skills transition into professional jobs. We used the Llama 3.2 3B model and the QLoRA fine-tuning method to create a technical feedback tool which is capable of delivering expert-level technical feedback while remaining accessible on standard computational hardware.

By providing an affordable and practical way for college students and those looking to broaden their employment opportunities to perform well in technical interviews and enhance their performance, we hope our work will help to reduce some of the barriers this group faces by providing access to advanced interview preparation tools. Our future work will expand the capabilities of this system to include real-time code evaluation to make the interview process even more realistic.

VII. REFERENCES

1. A. Aytar, "fine_tune_embdding_17_book_grobid_semantic," Hugging Face Dataset, 2023. Link: https://huggingface.co/datasets/AhmetAytar/fine_tune_embdding_17_book_grobid_semantic

ISSN 2394 - 7780

2. A. Vaswani, et al., "Attention Is All You Need," Adv. Neural Inf. Process. Syst., 2017. Link: https://arxiv.org/abs/1706.03762
3. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. North Am. Chapter Assoc. Comput. Linguist., 2019. Link: https://arxiv.org/abs/1810.04805
4. T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," *Proc. Int. Conf. Mach. Learn.*, 2023. Link: https://arxiv.org/abs/2305.14314
5. ed001, "ds-coder-instruct-v2," Hugging Face Dataset, 2023. Link: https://huggingface.co/datasets/ed001/ds-coder-instruct-v2
6. E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," Int. Conf. Learn. Represent., 2021. Link: https://arxiv.org/abs/2106.09685
7. H. Wang et al., "PEFT: Parameter-Efficient Fine-Tuning of Pre-trained Models," Hugging Face Documentation, 2022. Link: https://github.com/huggingface/peft
8. T. Wolf et al., "Hugging Face's Transformers: State-of-the-art Natural Language Processing," *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2020. Link: https://github.com/huggingface/transformers
9. Y. Liu et al., "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP," ACM Comput. Surv., 2023. Link: https://arxiv.org/abs/2107.13586