AI Presentation Coach: Data Science Report

Fine-Tuning and Evaluation of the Synthesis Agent

Author:

MD Misbah Ur Rahman

B.Tech (Hons.), Chemical Engineering
Indian Institute of Technology Kharagpur

Abstract

This report details the data science methodologies employed in the development of the AI Presentation Coach. It covers the creation of a specialized fine-tuning dataset, the process of parameter-efficient fine-tuning (PEFT) using the QLoRA technique on a state-of-the-art language model, and the design of a multi-faceted evaluation protocol to measure the agent's performance. The objective of this work was to transform a general-purpose language model into an expert communication coach, and this document presents the setup, methods, and outcomes of that process.

Contents

September 13, 2025

1 Fine-Tuning Setup: Forging a Specialist

The core of the AI Presentation Coach's intelligence lies in its **Synthesis-Worker**, the component responsible for generating the final feedback report. A generic, pre-trained language model, while capable, lacks the specific expertise and stylistic nuance of a professional communication coach. To fulfill the mandatory project requirement and, more importantly, to create a truly effective agent, I undertook a fine-tuning process to specialize a model for this exact task.

1.1 Dataset Curation: The Textbook for Our Agent

No public dataset exists for the task of "generate expert feedback from multi-modal presentation metrics." Therefore, I constructed a custom, high-quality dataset from scratch.

- **Hybrid Strategy:** The dataset was built using a hybrid strategy to ensure a balance of real-world data and controlled diversity. It includes one "ground-truth" sample from an actual video analysis, combined with several meticulously crafted synthetic samples.
- **Diverse Personas:** The synthetic samples were designed to cover a range of speaker archetypes, including a "Nervous Novice," a "Confident Expert," and a "Rushed Presenter," as well as different formats like technical presentations and behavioral interview answers. This diversity is crucial for teaching the model to provide relevant feedback across various scenarios.
- **Data Structure:** Each entry in the dataset is a JSON object with three keys: 'instruction', 'input', and 'output'. The 'input' contains the complete, structured data packet from our analysis workers, and the 'output' contains the corresponding "gold-standard" expert report that I authored for the model to learn from.

1.2 Methodology: Parameter-Efficient Fine-Tuning (PEFT)

To specialize the model, I employed a state-of-the-art technique called **QLoRA (Quantized Low-Rank Adaptation)**.

- **Base Model:** The chosen base model was `google/gemma-2b-it`. This model was selected after a rigorous process of elimination revealed that larger models were physically incompatible with the available free-tier cloud GPU resources. Gemma's 2B instruction-tuned variant provided the optimal balance of performance, a small memory footprint, and a permissive license.
- **QLoRA Technique:** This method is exceptionally powerful for this task. It involves loading the base model with its weights quantized to 4-bit precision, which dramatically reduces memory usage. Then, a small number of "adapter" layers are added to the model. During training, the billions of parameters in the base model remain frozen, and only these tiny adapter layers are trained.
- **Justification:** This approach directly fulfills the assignment's requirement for a **parameter-efficient tuned model**. More importantly, it is the most effective way to teach a massive model a new task with a very small, high-quality dataset, making it the perfect technical choice for this project.

2 Evaluation Methodology & Outcomes

Building and fine-tuning an agent is only half the challenge, proving its effectiveness is the other, equally important half. To ensure the AI Presentation Coach is a genuinely useful tool, I designed a multi-faceted evaluation strategy to measure the performance of its components and the quality of its final output. This was not just about getting a single score, but about deeply understanding the system's strengths and weaknesses.

2.1 Quantitative Evaluation: Worker-Level Metrics

For the data-gathering workers that produce objective outputs, we can use standard quantitative metrics to measure their accuracy.

- **Transcription-Worker Evaluation:** To measure the accuracy of our Whisper model, I calculated the **Word Error Rate (WER)**. This is the industry standard for speech-to-text systems. I created a small, manually-verified ground-truth transcript and compared the model's output to it. The formula is:

$$WER = \frac{S + D + I}{N}$$

  Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference. Our final 'small.en' model achieved a WER that was a significant improvement over the initial baseline, confirming the success of the model upgrade.

- **Visual-Worker Evaluation (Future Work):** A quantitative evaluation of the visual metrics would involve manually annotating a test video frame-by-frame for ground-truth data on gaze, smiles, and gestures. This is a time-intensive process that I have outlined as a next step for future development to further validate and refine these complex heuristics.

2.2 Qualitative Evaluation: Synthesis Agent Rubric

The final, synthesized report is too complex and nuanced for a simple numerical score. Its quality is not just about accuracy, but about its usefulness as a coaching tool. Therefore, I designed a qualitative rubric to score the agent's output based on three key criteria:

Actionability (Score: 1-5): Does the report provide specific, concrete advice that the user can actually implement? Or is it vague and generic? A high score here was my primary goal.

Data-Driven Reasoning (Score: 1-5): Does the report's advice logically follow from the quantitative metrics gathered by the workers? Or does it contradict the data or hallucinate facts (a key failure mode of early prototypes)?

Tone & Empathy (Score: 1-5): Does the report sound like an encouraging, supportive coach, in line with the "I'm beside you" ethos? Or is it overly critical, robotic, or generic?

2.3 Outcomes and Results

After fine-tuning the 'gemma-2b-it' model on our curated dataset, I performed a final evaluation run. The results were a dramatic improvement. The synthesized reports are now not only stylistically correct but also demonstrate strong **data-driven reasoning**, correctly correlating the input metrics with relevant feedback. While there is always room for improvement, the agent consistently scores highly on the qualitative rubric, particularly in its ability to provide **actionable** advice. The fine-tuning process was a clear success, validating the entire data science approach and confirming that we have created a genuinely intelligent and specialized agent.