Fine-Tuning Setup Report

1. Data

- **Sources:**
  - RACE (Reading Comprehension Dataset, MCQ format)
  - SQuAD (Stanford Question Answering Dataset, short-answer format)
- **Preparation:**
  - Sampled 25,000 examples each from RACE and SQuAD for diversity and balance.
  - Converted all samples to a unified JSONL format with fields: `context`, `type`, `instruction`, `question`, `options` (for MCQ), and `answer`.
  - Combined, shuffled, and saved as `train_50k.jsonl` for training.

2. Method

- **Base Model:**
  - Used `distilgpt2` for its lightweight architecture and efficiency.
- **Parameter-Efficient Fine-Tuning:**
  - Applied LoRA (Low-Rank Adaptation) via the PEFT library for memory- and storage-efficient adaptation.
  - Targeted GPT-2's attention and projection layers (`c_attn`, `c_proj`).
- **Prompt Construction:**
  - For each example, constructed a prompt including instruction, context, question, options (if MCQ), and answer.
  - Example prompt:

    For MCQ:

    ```
    [Instruction]
    Context: [context]
    Question: [question]
    Options: [option1 | option2 | ...]
    Answer: [answer]
    ```

    For Short Answer:

```
[Instruction]
Context: [context]
Question: [question]
Answer: [answer]
```

- **Tokenization:**
  - Used the base model's tokenizer with truncation and padding to a max length of 128 tokens.
- **Training Details:**
  - 3.5 epochs, batch size 2, AdamW optimizer, learning rate 2e-4, weight decay 0.001.
  - Saved checkpoints every 50 steps, logged every 10 steps.

3. Results

- **Model Output:**
  - The fine-tuned model can generate both MCQ and short-answer questions from a given context.
  - Output format is consistent with the training prompt, supporting robust downstream parsing.
- **Qualitative Observations:**
  - The model produces mostly relevant and fluent questions.
  - MCQs generally have plausible options and correct answers, though some repetition and generic outputs may occur.
  - Short answers are usually correct but sometimes lack detail or diversity.
- **Evaluation:**
  - The detailed evaluation report, including human ratings and analysis, is provided separately.

Summary

This fine-tuning setup enabled efficient adaptation of a lightweight language model for the dual task of MCQ and short-answer question generation, using a balanced and diverse dataset and parameter-efficient training.