Conversational AI for Job Search

using Phi-3.5-mini-instruct and LoRA

By: Karl Cao, Sarthak Bisht, Ahmad Furqan

ERTIONS
LOCATIONS
LOB STIRG
BIBCAMTIONS
JOB SURC
LOZATIONS
JOB
Jobe Model
FINE-TUNED

Agenda

- Project Overview
- Dataset Exploration
- Data Cleaning & Preprocessing
- Model Selection
- Fine-Tuning the Model
- Evaluation Metrics
- Conclusion & Future Work

Project Overview

- **Objective**: Develop a **conversational AI** model for querying jobs by skills, location, salary, experience.
- **Dataset**: Job descriptions, resumes, location, etc.
- **Techniques**: Fine-tuning with **LoRA** for efficient adaptation of the **Phi3 3B** model,
- **Evaluation**: BERTScore, ROUGE, ARES metrics.
- **Deliverable**: Model that provides accurate job-related recommendations

Data Exploration

- **Columns**: Resume, Job Title, Location, Salary, and Description.
- **Data Shape**: Train (596, 5) & Test (149, 5)
- **Duplicated Rows**: Train (289) & Test (27)
- **Missing Values in Salary**: Train (220) & Test (85)

Data Cleaning and Preprocessing

- **Lowercasing Text:** to reduce variability
- **Handling Duplicated Rows:**

  Train (48,5%) and Test (18%) was removed
- **Handling Missing Values in Salary column**

  1) Define Salary Range Buckets

  2) Imputation Based on Job Title and Location
- **Cleaning Pretext Columns:** Punctuation, URLs, Newline, Emojis, and Symbols
- **Summarizing Resume dan Description Columns:**

  Transformer library and BART

Key Findings

Summarizing Resume and Description Fields Using Transformers Library

- **To address the issue of overly long text in the resume and description columns, we employed the Transformers Library and used the BART model for text summarization**.
- This step was necessary because long text entries significantly increase the computational resources and time required for fine-tuning.
- By summarizing the text in these fields, we ensured that the content remains concise and relevant without losing important information, allowing for more efficient fine-tuning of the model while maintaining the quality of the dataset.
- This approach balances the need for comprehensive data with the constraints of our limited resources and time.

Tokenization for Fine-Tuning

Job Query & Description Generation

- Queries created from predefined templates across skills, location, salary, and experience categories.
- Each query type generated with 3 variations for better model generalization.

Tokenization Process:

- GPT-2 tokenizer used to convert Job Queries and Descriptions into numerical tokens for model training. Example Tokens: "What jobs are available for Python developer?" → ["What", "jobs", "are", ...]

Training, Validation, and Test Splits

- Training and Validation Split: We split the training dataset into an 80% training set and a 20% validation set to monitor model performance and optimize hyperparameters.

Key Findings

"students are encouraged to explore multiple types of job-related queries and analyze system performance in each case."

We used predefined templates to create a variety of meaningful questions across four categories: skills (from resume and description columns), location, salary, and experience.

For each query, we generated 10 (decreased to 3) variations of questions to create a comprehensive dataset for training

"Through this approach, we succesfully augmented our data from 596 to 3,684 query-description pairs for the training set and 149 to 1.464 for the test set, ensuring that the model has a wide range of examples to learn from"

Model Selection

Chosen Model: Phi-3.5-mini-instruct

Reasons for Selection:

- **Size & Efficiency**: 3B parameters—balances performance and resource requirements.
- **Instruction Following**: Pre-trained for instruction-based tasks.
- **Community Support**: Active development and documentation.

Library: Utilized Unsloth (built on HuggingFace) for efficient model adaptation in resource-limited environments

Fine-Tuning Technique

Technique Used: Low-Rank Adaptation (LoRA)

Why LoRA?

- **Efficiency**: Reduces computational resources and memory footprint.
- **Flexibility**: Allows quick adaptation without full model retraining.
- **Performance**: Maintains model quality while optimizing for our specific task.

Fine-Tuning Technique

Fine-Tuning Technique: LoRA (Low-Rank Adaptation) applied to specific layers of the model (q_proj, v_proj, k_proj, o_proj, gate_proj).

Parameters:

- r = 16: Low-rank adaptation dimension.
- per_device_train_batch_size = 2.
- gradient_accumulation_steps = 4.
- max_steps = 60 (2 mins) | max_steps = -1 (3 hours)
- Efficiency: Enabled efficient fine-tuning without the computational overhead of full model retraining.
- Max Sequence Length: Set to 2048 tokens to handle long job queries and descriptions.

Training Process

Fine-Tuning Steps:

- Loaded preprocessed data into the model.
- Configured LoRA parameters for optimal training.

Types of User Queries Handled:

- **Skills-Based**: "What jobs are available for Python developers?"
- **Location-Based**: "Are there any jobs in New York?"
- **Salary-Based**: "What jobs offer over $100k?"
- **Experience-Based**: "Jobs suitable for 5 years of experience?"

Sample Response - Baseline

<User> What jobs are open in New York?

<AI> As of my knowledge cutoff in early 2023, I cannot provide real-time job listings. However, New York City is a bustling hub for a wide range of industries, and there are always numerous job openings across various sectors. Here are some common categories where you might find job listings:

1. *Finance*: New York is home to Wall Street and many major financial institutions, so there are often openings for roles such as financial analysts, accountants, investment bankers, and financial advisors.

2. *Technology*: With the rise of tech companies in the city, there are opportunities in software development, IT support, cybersecurity, and data analysis.

3. *Healthcare*: The city's large population demands a robust healthcare system, so there are jobs available in hospitals, clinics, and healthcare administration.

4. *Education*: New York has a vast array of educational institutions, so there are teaching positions, educational consultants, and administrative roles.

Sample Response - Finetuned

<User> What jobs are open in New York?

<AI> Job available in New York

Job-Title: senior frontend developereloper

Salary: Over $150K

Description: senior frontend developer will be responsible for working with the product team to understand the vision for the companys online presence and collaborate on the design and development of the frontend

Evaluation Metrics

Automated Metrics Used:

- **BERTScore**: Evaluates semantic similarity between generated responses and references.
- **RoGUE**: Measures overlap, considering various aspects such as n-gram overlap, word sequences, and word pairs.

Evaluation Metrics - Baseline

Metrics Used: Precision, Recall, F1-Score, BERTScore, ROUGE.

Baseline Results:

- BERTScore: 0.792
- ROUGEScore:
  - Rouge1: 0.386
  - Rouge2: 0.073
  - RougeL: 0.277
- Purpose: Ensures the model generates relevant, coherent job-related responses.

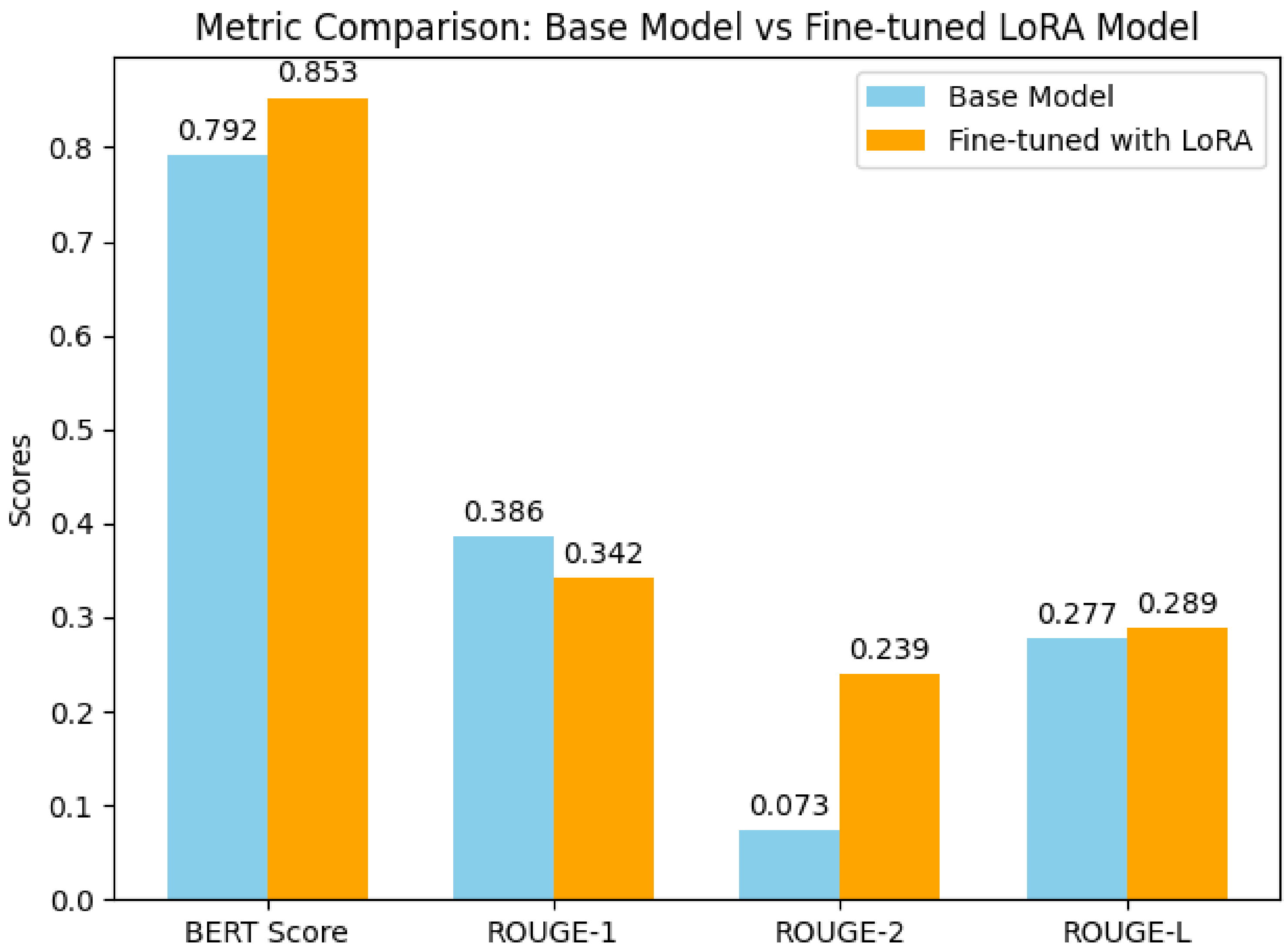
Evaluation Metrics - Finetuned

Metrics Used: Precision, Recall, F1-Score, BERTScore, ROUGE.

Finetuned Results:

- BERTScore: 0.853
- ROUGEScore:
  - Rouge1: 0.342
  - Rouge2: 0.239
  - RougeL: 0.289
- Purpose: Ensures the model generates relevant, coherent job-related responses.

Evaluation Metrics - Finetuned



Key Findings

- BERTScore shows clear improvement, indicating the finetuned model generates semantically closer predictions.
- ROUGE1 drops slightly, meaning the finetuned model captures fewer exact word matches.
- ROUGE2 shows a major improvement, meaning the finetuned model captures better local phrase structure.
- ROUGEL improves modestly, indicating better alignment in longer segments

Thank You

Q & A