

AI-DRIVEN AUTONOMOUS CAREER MENTORSHIP & MARKET ANALYZER

Presented by:

Edara Bindu Madhavi
Eadara Harika
Katakam Hemanvitha
Madapati Sai Swetha
Gode Sri Gopi Sarath

Project Advisors:

Dr. Simon Shim
Dr. Lee C. Chang

Introduction

Job seekers often receive career advice that is generic, outdated, or not aligned with real job-market expectations. CareerMentor AI analyzes resumes, identifies skill gaps, recommends learning paths, and provides personalized AI mentorship.

The system guides users through RESUME_UPLOAD, SKILL_ANALYSIS, LEARNING_PATH, PROGRESS_TRACKING, and AI_MENTOR_SUPPORT.



Problem Statement

- Job seekers need a clear way to measure their readiness for target roles. We target the gap between **resume skills** and **real job-market expectations**.
- Current platforms give job listings or generic advice, but do not show precise missing skills.
- Learners often lack a structured path for what to learn, practice, and track next.
- CareerMentor AI aims to provide personalized career guidance through resume analysis, skill-gap detection, learning recommendations, and AI mentor support.

Methodology



- Collected real-world data from job boards, course platforms, and resume datasets
- Cleaned and structured data into a skill taxonomy and career tracks
- Generated career Q&A pairs to teach the AI domain knowledge
- Fine-tuned four language models from Alibaba, Meta, Microsoft, and Hugging Face
- Evaluated models on response accuracy, quality, and speed
- Built a full-stack web platform with eight career mentorship features
- Deployed a live AI career mentor freely accessible to job seekers

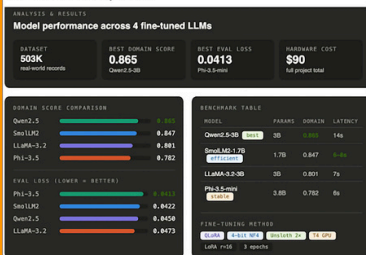
Analysis and Results

1. System Architecture



Tier 1 - User Layer: Job seekers access the platform through a web browser on desktop or mobile.
Tier 2 - Frontend: Netflix hosts the React + TypeScript interface for resume upload, dashboards, learning paths, quizzes, mentor chat, and model comparison.
Tier 3 - Backend: Render hosts the FastAPI service that manages resume parsing, skill-gap analysis, learning resources, progress tracking, quizzes, and REST APIs.
Tier 4 - AI Inference: Google Colab T4 GPU runs four fine-tuned LLMs for mentor chat and live model comparison.
Data Flow: User requests move from browser → frontend → backend → LLM inference through HTTPS/REST and local tunnel.
Deployment: GitHub CI/CD automatically deploys frontend updates to Netflix and backend updates to Render.

2. Models and Components



- Qwen2.5-3B achieved the highest domain score (0.865) - Alibaba's multilingual pre-training improved career-specific understanding above all 3B competitors.
- Phi-3.5-mini had the lowest eval loss (0.0413) but the lowest domain score (0.782) - proving eval loss alone is an insufficient predictor of real-world advice quality.

- SmolLM2-1.7B is the best parameter-efficiency choice - a 1.7B model matched 3B model quality (0.847) at 6-8s inference, the fastest response time in the study.
- Pre-training architecture fundamentally shapes fine-tuning outcomes - all 4 models trained on identical data yet produced distinctly different response styles.
- 503K real-world records ground the advice- training on LinkedIn, Indeed, Glassdoor postings and Udemy/Coursera courses eliminates the generic advice problem of base LLMs.

3. Visualizations

Domain accuracy vs response speed

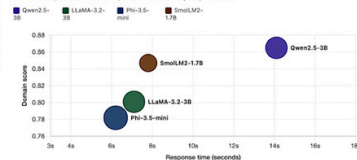


Figure 3.1-Domain Accuracy vs Response Speed

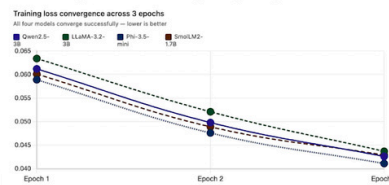


Figure 3.2- Training Loss Convergence Across 3 Epochs

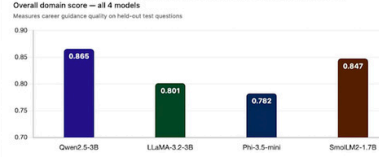


Figure 3.3-Overall Domain Score Comparison

Figure 3.1 shows the tradeoff between domain accuracy and response speed across all four models, where Qwen2.5-3B achieves the highest accuracy but at the slowest response time, while SmolLM2-1.7B delivers competitive accuracy at nearly half the latency making it the optimal choice for real-time career mentorship. Figure 3.2 demonstrates that all four models converged successfully across three training epochs with loss consistently decreasing below 0.045, confirming that the 72,741 career domain training examples were sufficient to adapt each architecture effectively on a single T4 GPU, with Phi-3.5-mini achieving the most stable fine-tuning.

Figure 3.3 compares the overall domain scores evaluated on a held-out test set, where all four models exceed the 0.78 target threshold. Qwen2.5-3B leading at 0.865, SmolLM2-1.7B at 0.847, LLaMA-3.2-3B at 0.801, and Phi-3.5-mini at 0.782 with the variation in scores despite identical training data confirming that pre-training methodology independently shapes fine-tuned model quality, a key research contribution of this project.

Summary/Conclusions

Our project bridges the gap between job seekers and industry by fine-tuning four large language models on real-world hiring data delivering personalized skill gap analysis, curated learning paths, and AI-powered career guidance in one free platform.

Key outcomes:

- Four LLMs fine-tuned on real job market data using QLoRA
- Skill gap analysis grounded in real hiring patterns
- Eight career features deployed as a live public platform
- Proven that small fine-tuned models rival large commercial AI at near-zero cost

Impact: Democratizing career mentorship data-driven, personalized, and accessible to all. By eliminating the cost barrier of human career coaching, CareerMentor AI empowers every job seeker regardless of background, location, or financial means to compete confidently in today's job market.

Key References

- Hu et al. (2022). LoRA: Low-rank adaptation of large language models. *ICLR* 2022.
- Dettmers et al. (2023). QLoRA: Efficient finetuning of quantized LLMs. *NeurIPS* 2023.
- Bai et al. (2025). Qwen2.5 technical report. *arXiv:2501.12599*.
- Abdin et al. (2024). Phi-3 technical report. *arXiv:2404.14219*.
- Grattafiori et al. (2024). The LLaMA 3 herd of models. *arXiv:2407.21783*.
- Allal et al. (2024). SmolLM2: Best language models to the edge. *HuggingFace* 2024.
- Li et al. (2025). Skill graph construction from job postings using LLMs. *EMNLP* 2025.

Acknowledgements

We thank our project advisors and the Applied Data Science Department at San José State University for their invaluable guidance, constructive feedback, and continuous support throughout this project. We also thank the open-source ecosystem including Hugging Face, Unsloth, FastAPI, and React for the tools and frameworks that made this project possible.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.