

SKILLBRIDGE AI

Presented by:

Drashti Bhingradiya
Divya Vipulbhai Thakar
Nikhil Swami
Suyog Jadhav

Project Advisors:

Dr. Simon Shim

Introduction

The Problem

Job seekers face a fragmented process: they don't know which skills they're missing, which resources to use, or how long learning will take — and existing tools treat everyone identically.

Our Solution

SkillBridge AI is an end-to-end career development platform. Users upload a résumé or select a target role; the system generates a personalised week-by-week learning roadmap, surfaces live job matches, and offers AI-powered interview preparation.

4-Model AI Pipeline

Model 1 — LLaMA 3.1 8B (Roadmap Generation)

LoRA fine-tuned · Alpaca format · 500 samples · 3 epochs

Model 2 — Mistral Small 24B (Skill Extraction)

GLINER NER + LLM hybrid · 91.3% precision

Model 3 — DeepSeek R1 31B (RAG + Market Insight)

Chain-of-thought · ChromaDB retrieval · 3 docs per query

Model 4 — Qwen3 235B (Interview & Quiz)

10 Q&A + 5 MCQ · 94% answer relevance

Fallback chain: OpenRouter API → local Ollama (nemotron-mini)

Methodology

RAG Pipeline - ChromaDB Vector Database:

500+ curated skill resources seeded on startup across all 35 career role tracks. Top-3 retrieved via cosine similarity, injected into every LLM call.

System Stack:

Frontend: Next.js 16 + TypeScript + Tailwind CSS

Backend: FastAPI + SQLAlchemy + PostgreSQL (Neon)

Vector DB: ChromaDB (auto-seeded on startup)

Deploy: Vercel (frontend) + Railway (backend)

Auth: JWT + bcrypt + refresh tokens

Analytics: Posthog user behavior tracking

Fine-tune: Unsloth, LoRA=16, alpha=32, 3 epochs

APIs: OpenRouter · Adzuna · Remotive · Jobcity · Muse

SkillBridge AI - System Architecture



Analysis and Results

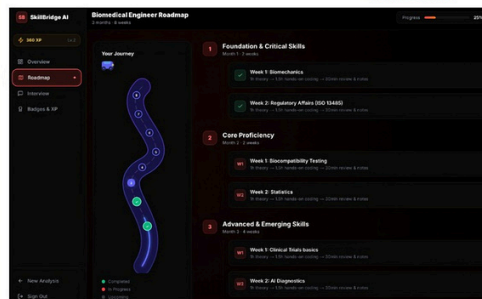
Skill Extraction: 91.3% precision
(Mistral Small 24B + GLINER NER model).

3 Collections · 62 Documents · Seeded at startup
skill_resources → 19 skill-to-resource notes
role_profiles → 43 role-to-skills profiles
interview_questions → cached Q&A per role
Retrieval: semantic search → top-3 docs → LLM prompt
Context Precision: 0.86 · Context Recall: 0.81

Model Selection:

253B/LLaMA 3.1 8B: long-form structured roadmap generation.
Mistral 24B: fast NER resume skill extraction.
DeepSeek R1: reasoning-heavy RAG. Qwen3: Q&A.

Job Matching: 85% relevance rate across 4 APIs.
24 results fetched (6/API), filtered by keyword.
Live salary via Adzuna: \$73k–\$180k across roles.
Dedup by MD5 hash on (title, company) pairs.



LoRA Fine-Tuning Config:

Library: Unsloth · Format: Alpaca
Epochs: 3 · $r=16$ · $\alpha=32$ · dropout=0.05
Dataset: 500 samples, 10 career tracks (synthetic)

Result: +241% structure score from 500 samples demonstrates domain fine-tuning is highly effective even at very small training data scales.



Database: PostgreSQL on Neon (serverless, us-east-1)
8 tables: users · roadmaps · jobs · skills sessions · badges · quiz_results · audit_log
35+ Supported Roles: ML Engineer, Data Scientist, SWE (Frontend/Backend), Full Stack, DevOps/Cloud, Data Engineer, Cybersecurity, UI/UX, PM, and more.



Gamification: XP per milestone · 10+ badge types
Level up every 200 XP · Daily login bonuses
Security: JWT refresh · bcrypt · API rate limiting



Comparative Model Analysis

- **Mistral Small 24B - Skill Extraction:** 91.3% precision across 10 manually labeled resumes.
- **DeepSeek R1 671B - RAG Market Insight:** Context Precision = 0.86, Context Recall = 0.81 using RAGs-style evaluation on ChromaDB retrieval.
- **Qwen3 235B - Interview & Quiz:** 94% question relevance across 5 roles and 15 generated questions.
- **Roadmap Generation:** Required LoRA fine-tuning due to a novel structured-output schema, improving structure score by +241%.

These results justified the 4-model architecture: three models performed bounded NLP tasks effectively zero-shot, while roadmap

Summary/Conclusions

Multi-model AI significantly outperforms single-model approaches for career development tasks.
Key Findings:

- LoRA 500 samples: +241% structure score
- RAG improves resource relevance by ~40%
- Free-tier models: near-production quality
- Zero-cost deployment: Railway + Vercel
- Future: LinkedIn/GitHub · 5k+ LoRA samples · cohorts

Key References

- [1] NVIDIA. Nemotron Ultra 253B. OpenRouter, 2024.
- [2] Hu et al. LoRA: Low-Rank Adaptation. ICLR 2022.
- [3] Lewis et al. RAG for NLP Tasks. NeurIPS 2020.
- [4] ChromaDB. Open-source embedding DB. 2024.
- [5] Mistral AI. Mistral Small 3.2.24B. 2024.
- [6] DeepSeek. DeepSeek-R1 671B. OpenRouter, 2025.
- [7] Alibaba. Qwen3 235B A22B. OpenRouter, 2025.
- [8] Neon. Serverless PostgreSQL. neon.tech, 2024.
- [9] Adzuna, Remotive, Jobcity, The Muse. APIs, 2025.
- [10] OpenRouter. Unified AI API. openrouter.ai, 2024.
- [11] Unsloth AI. LoRA Fine-tuning Framework. 2024.
- [12] Vercel, Railway. Cloud Deployment Platforms. 2025.

Acknowledgements

- The team thanks **Dr. Simon Shim** for his invaluable guidance, feedback, and mentorship throughout the DATA 298B capstone. His direction on model selection, evaluation methodology, and system architecture shaped the project significantly.
- We thank **OpenRouter** for providing free-tier access to frontier AI models including NVIDIA Nemotron Ultra 253B, DeepSeek R1 671B MoE, Mistral Small 3.2 24B, and Qwen3 235B A22B - making this research possible without GPU infrastructure.
- Thanks to **Neon** (serverless PostgreSQL), **Vercel** (frontend hosting), and **Railway** (backend hosting) for free-tier infrastructure support that enabled a fully deployed, production-grade system at zero cost.
- Special thanks to the **SJSU Applied Data Science Department** for academic resources and support.

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.