CareerCraft AI

An Intelligent Resume Optimization System Using Generative AI

Submitted by:
Aditi Mishra
Lasit Vyas
Likhit Kumar CR

Instructor:
Prof. Onur Barut

1. Executive Summary

CareerCraft AI is an intelligent, end-to-end resume optimization system designed to assist students and job seekers in generating professional, ATS-optimized resumes tailored to specific job descriptions. In today’s competitive hiring landscape, resumes are frequently screened by automated Applicant Tracking Systems before reaching human recruiters. As a result, many qualified candidates are rejected due to poor formatting, weak phrasing, or missing keywords rather than a lack of skills or experience.

The objective of CareerCraft AI is to automate the resume enhancement process using Generative AI while maintaining accuracy, clarity, and professional formatting. The system allows users to either upload an existing resume in PDF or DOCX format or manually enter their resume details through a structured interface. Users also provide a job description for the role they are applying for. The platform then analyzes both the resume content and job description to generate a refined, role-specific resume in PDF format.

From a technical standpoint, CareerCraft AI integrates multiple components into a unified workflow. A Streamlit-based frontend provides a clean and intuitive user interface. An n8n workflow automation engine orchestrates backend logic, routing inputs through extraction, prompt construction, AI inference, and document generation stages. A Flask-based microservice handles resume text extraction, while a fine-tuned Phi-3 Mini language model hosted on Hugging Face performs the resume rewriting and optimization.

This project demonstrates how lightweight large language models, when combined with workflow orchestration and structured prompting, can deliver production quality AI applications. CareerCraft AI highlights practical applications of Generative AI in career development and showcases scalable, cost-efficient system design.

2. Problem Statement & Solution Overview

2.1 Problem Statement

Students and job seekers often face several recurring challenges during the job application process:

- Resumes are poorly structured or inconsistently formatted
- Resume bullet points lack measurable impact or clarity
- Critical keywords required by Applicant Tracking Systems (ATS) are missing
- One resume is reused for multiple job roles without personalization
- Manually rewriting resumes for different job descriptions is time-consuming

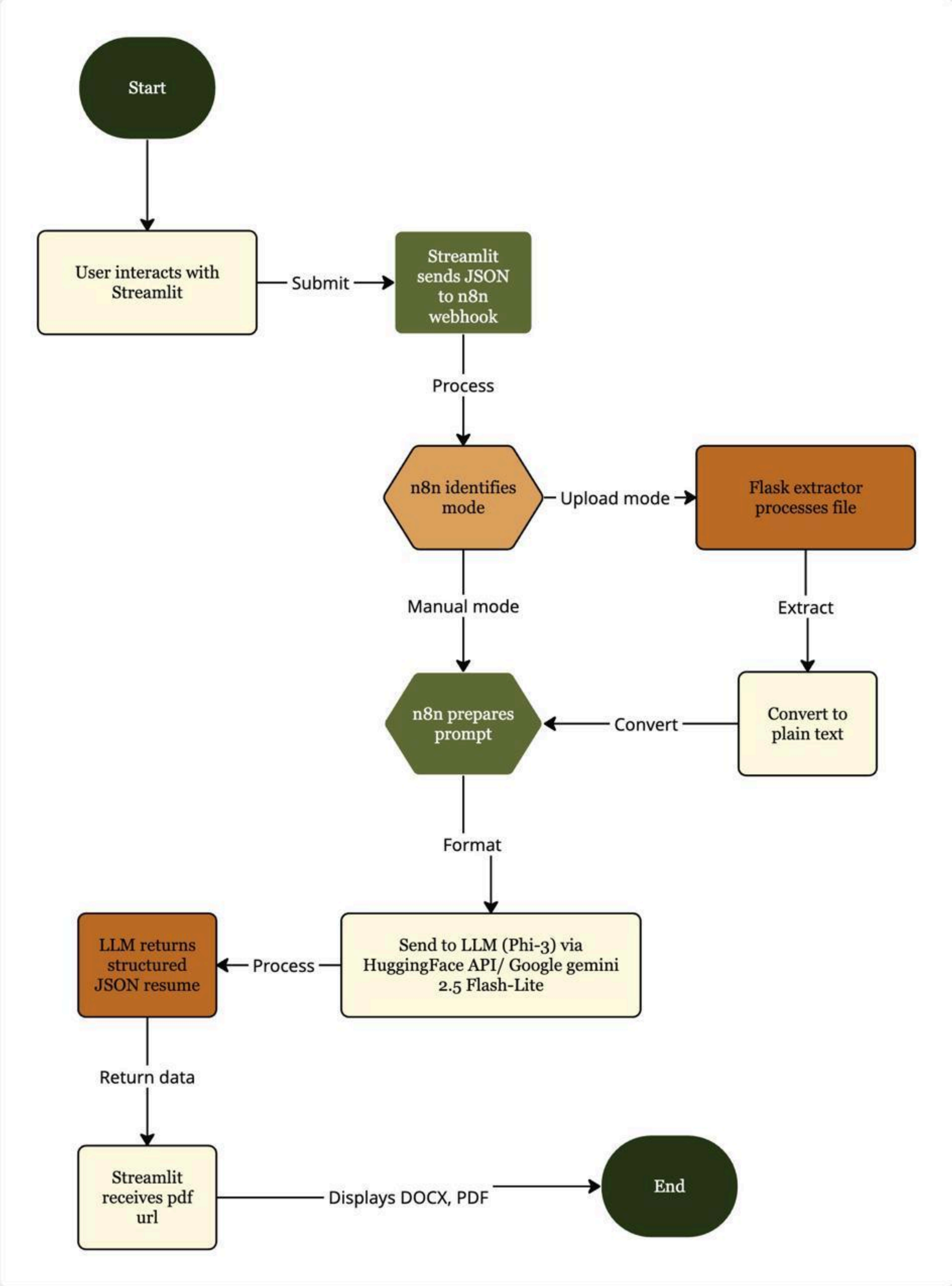
These challenges result in lower interview callback rates, even for candidates who are technically qualified. Additionally, many resume optimization tools either require manual effort or lack true contextual understanding of job descriptions.

2.2 Solution Overview

CareerCraft AI addresses these challenges by introducing an automated AI-driven resume optimization agent. The system performs the following tasks automatically:

1. Extracts resume text from uploaded PDF or DOCX files
2. Accepts structured resume data when users choose manual entry
3. Analyzes job descriptions to identify relevant skills and keywords
4. Rewrites resume content using professional, ATS-friendly language
5. Enhances weak bullet points with quantified achievements
6. Generates a clean, professionally formatted PDF resume

By automating these steps, CareerCraft AI significantly reduces the effort required to tailor resumes while improving their effectiveness in automated screening systems.


Start
User interacts with Streamlit
Submit
Streamlit sends JSON to n8n webhook
Process
n8n identifies mode
Upload mode
Flask extractor processes file
Manual mode
Extract
n8n prepares prompt
Convert
Convert to plain text
Format
LLM returns structured JSON resume
Process
Send to LLM (Phi-3) via HuggingFace API/ Google gemini 2.5 Flash-Lite
Return data
Streamlit receives pdf url
Displays DOCX, PDF
End

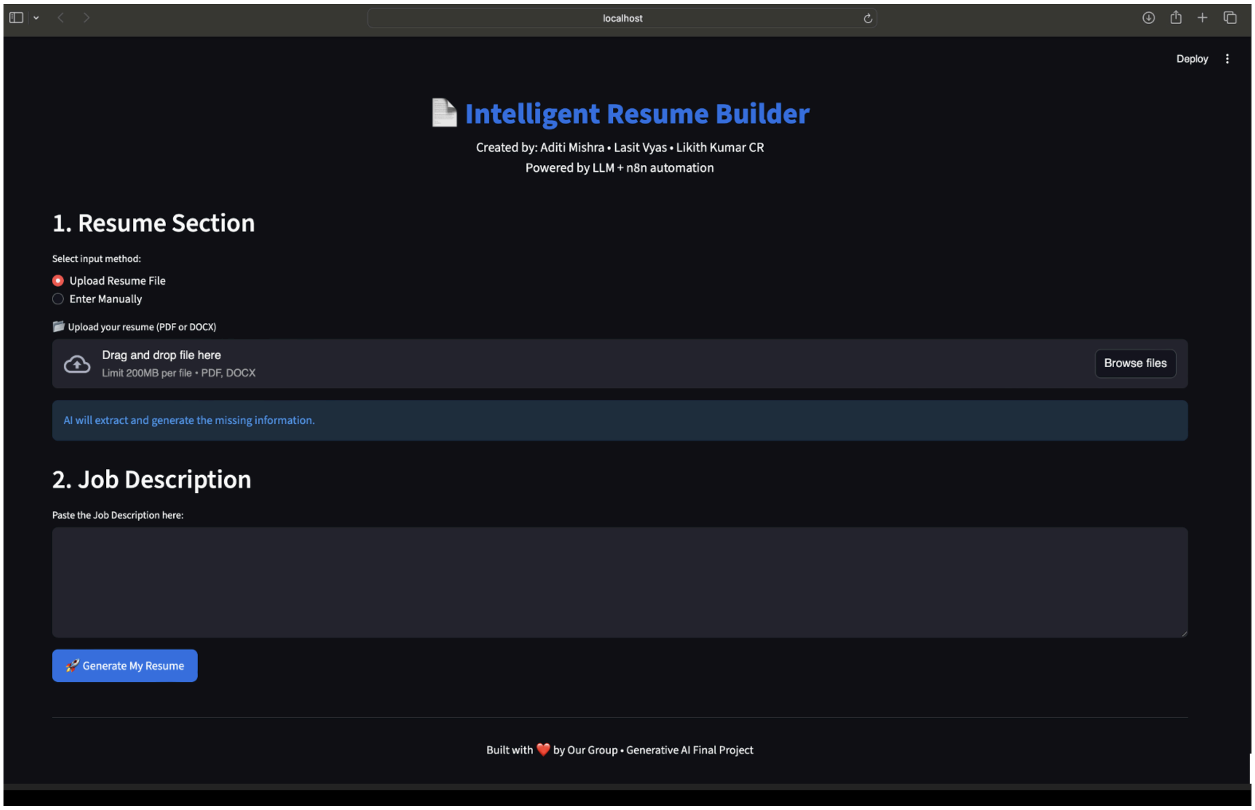
3. Technical Architecture & Design Specifications

3.1 High-Level Architecture

CareerCraft AI follows a modular, service-oriented architecture that separates concerns across frontend, automation, AI inference, and document generation layers. This design improves scalability, maintainability, and extensibility.

The major architectural components include:

- **Frontend Layer (Streamlit):** Handles user interaction and input validation
- **Automation Layer (n8n):** Orchestrates the workflow logic and service communication
- **Extraction Layer (Flask Microservice):** Extracts text from uploaded resumes
- **AI Inference Layer (LLM):** Enhances resume content using Generative AI
- **Document Generation Layer:** Converts structured output into PDF format



📄 Intelligent Resume Builder

Created by: Aditi Mishra • Lasit Vyas • Likith Kumar CR

Powered by LLM + n8n automation

1. Resume Section

Select input method:

● Upload Resume File

○ Enter Manually

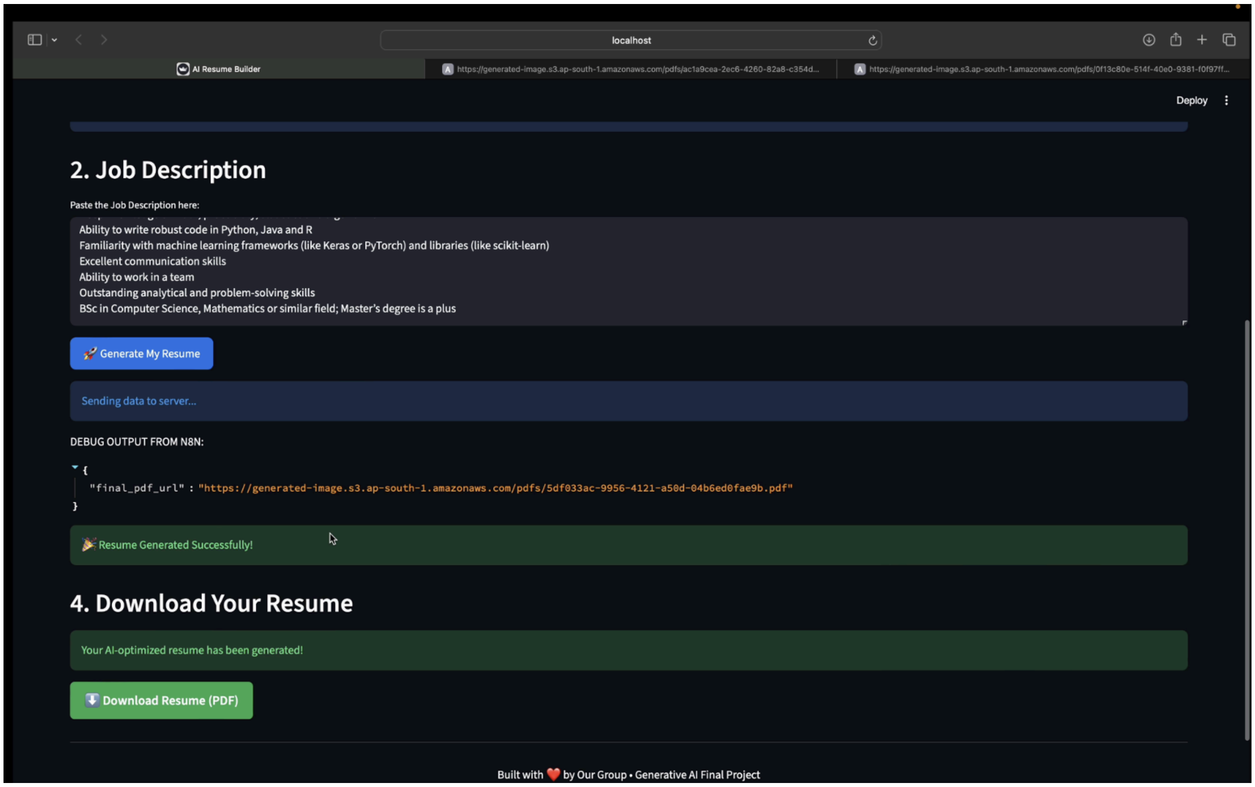
📁 Upload your resume (PDF or DOCX)

AI will extract and generate the missing information.

2. Job Description

Paste the Job Description here:

Built with ❤️ by Our Group • Generative AI Final Project


localhost
AI Resume Builder
https://generated-image.s3.ap-south-1.amazonaws.com/pdfs/ac1a9cea-2ec6-4260-82a8-c354d...
https://generated-image.s3.ap-south-1.amazonaws.com/pdfs/0f13c80e-514f-40e0-9381-f0f97ff...
Deploy

2. Job Description

Paste the Job Description here:

Ability to write robust code in Python, Java and R
Familiarity with machine learning frameworks (like Keras or PyTorch) and libraries (like scikit-learn)
Excellent communication skills
Ability to work in a team
Outstanding analytical and problem-solving skills
BSc in Computer Science, Mathematics or similar field; Master's degree is a plus

Sending data to server...

DEBUG OUTPUT FROM N8N:

```
{
  "final_pdf_url" : "https://generated-image.s3.ap-south-1.amazonaws.com/pdfs/5df033ac-9956-4121-a50d-04b6ed0fae9b.pdf"
}
```

🎉 Resume Generated Successfully!

4. Download Your Resume

Your AI-optimized resume has been generated!

Built with ❤️ by Our Group • Generative AI Final Project

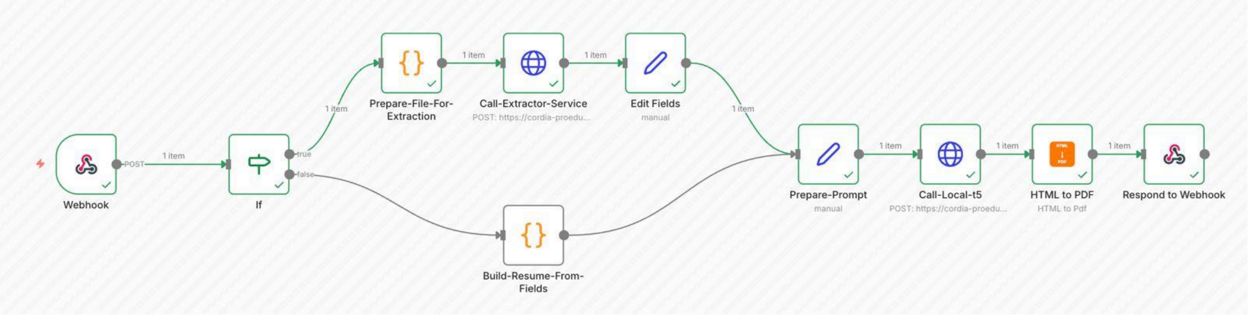
3.2 Workflow Orchestration Using n8n

The n8n workflow serves as the central control mechanism for the system. It begins with a webhook trigger that receives input from the Streamlit frontend. An IF node determines whether the user uploaded a resume or entered data manually.

- In **upload mode**, the workflow sends the file to the Flask extraction microservice

- In **manual mode**, the workflow directly uses structured user input

The workflow then constructs a combined prompt containing resume content and job description and sends it to the language model. The AI-generated structured response is parsed and forwarded to the document generation module before being returned to the frontend.



4. Methodologies and Technologies Used

4.1 Frontend Development

The frontend is built using **Streamlit**, chosen for its rapid development capabilities and seamless Python integration. Streamlit allows for dynamic form creation, file uploads, and real-time interaction without requiring extensive frontend frameworks.

4.2 Resume Extraction Methodology

A Flask-based microservice is used to extract text from uploaded resumes. The service utilizes:

- **PyPDF2** for PDF parsing
- **python-docx** for DOCX file extraction

The extracted content is cleaned and normalized before being passed into the AI pipeline.

4.3 Language Model Selection

The project uses **Phi-3 Mini 4K Instruct**, a lightweight yet highly capable language model. Phi-3 was selected due to:

- Fast inference speed
- Strong instruction-following ability
- Compatibility with LoRA fine-tuning
- Cost efficiency compared to larger LLMs

4.4 Fine-Tuning Methodology

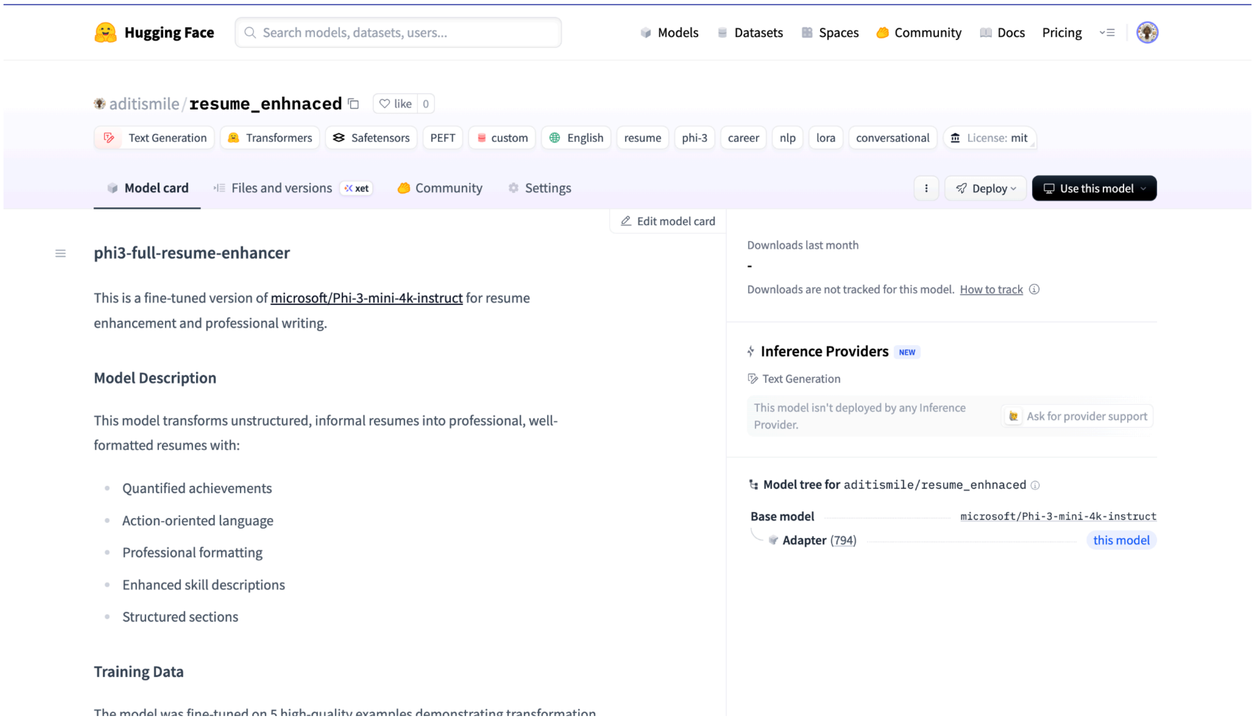
The model was fine-tuned to improve resume-specific rewriting tasks. Key techniques include:

- **LoRA (Low-Rank Adaptation):** Training only a small subset of parameters
- **4-bit NF4 Quantization:** Reducing memory usage by approximately 75%
- **Targeted Layer Training:** q_proj, k_proj, v_proj, o_proj layers

Training configuration included:

- Epochs: 3
- Learning rate: 2e-4
- Optimizer: AdamW (8-bit)

This approach enabled efficient training on limited hardware while preserving output quality.



phi3-full-resume-enhancer

This is a fine-tuned version of microsoft/Phi-3-mini-4k-instruct for resume enhancement and professional writing.

Model Description

This model transforms unstructured, informal resumes into professional, well-formatted resumes with:

- Quantified achievements
- Action-oriented language
- Professional formatting
- Enhanced skill descriptions
- Structured sections

Training Data

The model was fine-tuned on 5 high-quality examples demonstrating transformation

Downloads last month

-

Downloads are not tracked for this model. How to track

Inference Providers NEW

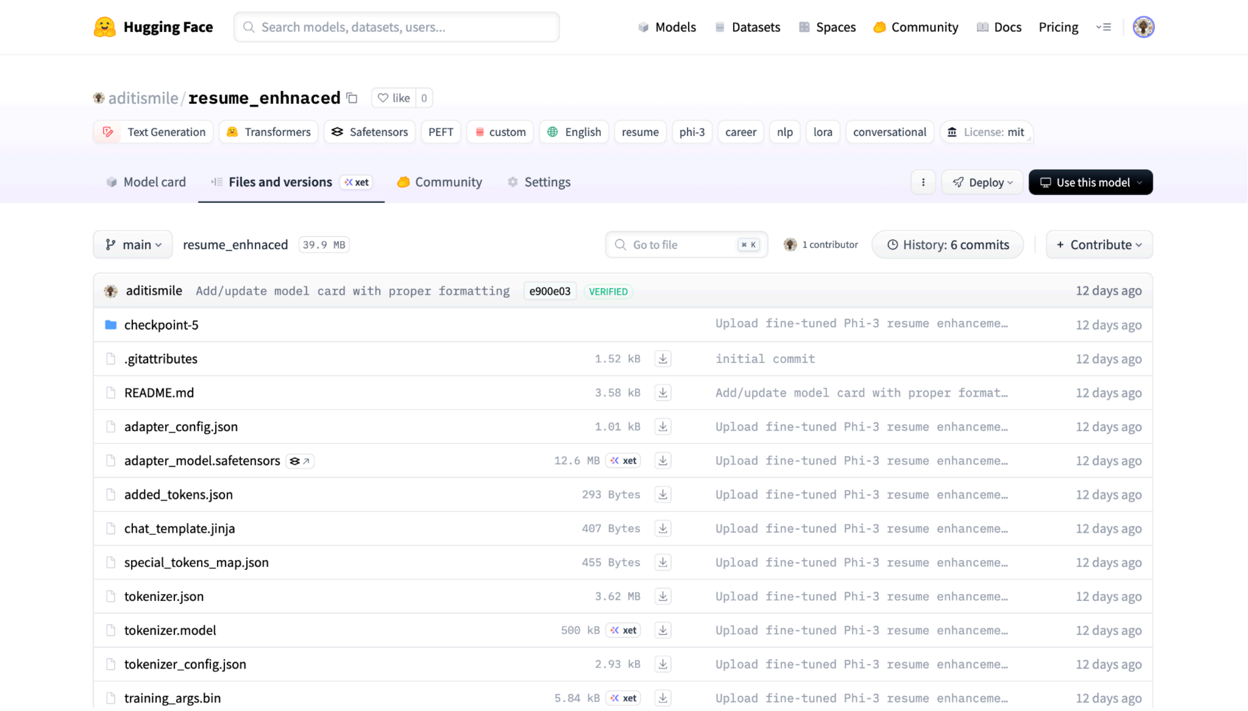
Text Generation

This model isn't deployed by any Inference Provider.

Ask for provider support

Model tree for aditismile/resume_enhnaced

Base model	microsoft/Phi-3-mini-4k-instruct
Adapter (794)	this model



resume_enhnaced 39.9 MB

aditismile Add/update model card with proper formatting e900e03 VERIFIED 12 days ago

checkpoint-5		Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
.gitattributes	1.52 kB	initial commit	12 days ago
README.md	3.58 kB	Add/update model card with proper format…	12 days ago
adapter_config.json	1.01 kB	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
adapter_model.safetensors	12.6 MB xet	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
added_tokens.json	293 Bytes	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
chat_template.jinja	407 Bytes	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
special_tokens_map.json	455 Bytes	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
tokenizer.json	3.62 MB	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
tokenizer.model	500 kB xet	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
tokenizer_config.json	2.93 kB	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago
training_args.bin	5.84 kB xet	Upload fine-tuned Phi-3 resume enhanceme…	12 days ago

5. Deployment Details

The deployment strategy for CareerCraft AI was designed with modularity, scalability, and cloud readiness in mind. Rather than building a monolithic application, the system is composed of loosely coupled services that communicate through HTTP requests and webhooks. This architectural decision allows individual components to be developed, tested, deployed, and scaled independently.

During development and testing, the application operates in a local environment. The Streamlit frontend runs as a local service, providing users with a browser-based interface for resume upload, manual data entry, and job description submission. To enable communication between the frontend and backend automation pipeline, a secure tunnel is established using ngrok. This exposes the n8n webhook endpoint to the internet without requiring immediate cloud deployment, enabling rapid iteration and debugging.

The automation layer, powered by n8n, is responsible for orchestrating all backend logic. n8n is deployed either locally or on a lightweight virtual machine and configured with a webhook trigger node that receives structured JSON data from the frontend. This webhook acts as the system's entry point, ensuring consistent input handling regardless of user workflow. Conditional logic within the workflow determines whether the user has uploaded a resume file or opted for manual entry. This branching enables flexible processing without duplicating logic.

For resume extraction, a Flask-based microservice is deployed as an independent service. This microservice accepts resume files in PDF or DOCX format and extracts raw text using PyPDF2 and python-docx. By isolating extraction logic into a microservice, the system avoids tight coupling between document parsing and AI inference, allowing future upgrades such as OCR support or multilingual extraction without redesigning the core workflow.

The AI inference layer is deployed using the Hugging Face Inference API, which serves the fine-tuned Phi-3 Mini model. This approach eliminates the need to manage GPU infrastructure directly while providing reliable scaling and availability. Requests sent from n8n include structured prompts containing resume content and job descriptions, and responses are returned in structured JSON format. This design ensures deterministic parsing and reduces the risk of malformed outputs.

The final stage of deployment involves document generation. The system converts structured AI output into a clean HTML template, which is then rendered into a PDF using document generation libraries. The resulting resume PDF is sent back to the Streamlit frontend through the webhook response, allowing users to download the file immediately.

The system is designed for full production deployment on cloud platforms such as Microsoft Azure. Each component—frontend, n8n workflow engine, extraction microservice, and AI inference—can be containerized and deployed using cloud-native services. This architecture supports horizontal scaling, fault isolation, and future enterprise-grade deployment requirements.

6. UI/UX Design Process and Outcomes

6.1 Design Principles

The UI design emphasizes:

- Simplicity and clarity
- Minimal cognitive load
- Logical grouping of inputs
- Step-by-step user guidance

6.2 User Input Modes

The application supports two user workflows:

Resume Upload Mode:
Users upload an existing resume, which is automatically extracted and optimized.

Manual Entry Mode:
Users enter structured information including:

- Name and job title
- Work experience and achievements
- Skills and education
- Contact information

This ensures accessibility for users without an existing resume.

7. Commercial Viability & Market Strategy

CareerCraft AI addresses a well-defined and growing market: individuals actively seeking employment in increasingly automated hiring ecosystems. The global rise of Applicant Tracking Systems has shifted the hiring process toward keyword-driven and format-sensitive screening, creating a strong demand for tools that optimize resumes for automated evaluation while maintaining human readability.

The primary target market includes university students, recent graduates, and early-career professionals who lack access to professional resume-writing services. A secondary market consists of international job seekers who must adapt resumes to regional hiring standards, particularly in markets such as the United States and Europe. Additionally, career services departments at universities represent an institutional market segment where CareerCraft AI could be deployed at scale.

CareerCraft AI offers several competitive advantages over traditional resume tools. Unlike static resume templates or keyword scanners, the system performs contextual analysis of both resumes and job descriptions. By leveraging a fine-tuned language model, it generates role-specific improvements rather than generic suggestions. Furthermore, the fully automated pipeline reduces user effort, making the tool accessible to non-technical users.

From a cost perspective, the use of a lightweight language model and workflow automation significantly reduces operational expenses compared to systems built on large proprietary models. This enables flexible pricing strategies, including a freemium model where basic resume generation is free and advanced features such as multiple versions, cover letters, or premium templates are offered via subscription.

Potential monetization strategies include:

- Subscription-based access for individual users
- Institutional licensing for universities and career centers
- API-based offerings for recruitment platforms and job boards
- White-label solutions for HR consultancies

By focusing on automation, personalization, and affordability, CareerCraft AI demonstrates strong commercial potential in a competitive but underserved market.

8. Results and Impact Assessment

The effectiveness of CareerCraft AI was evaluated through qualitative analysis of generated resumes and comparison of input versus output quality. Results consistently show a significant improvement in resume clarity, professionalism, and alignment with job descriptions.

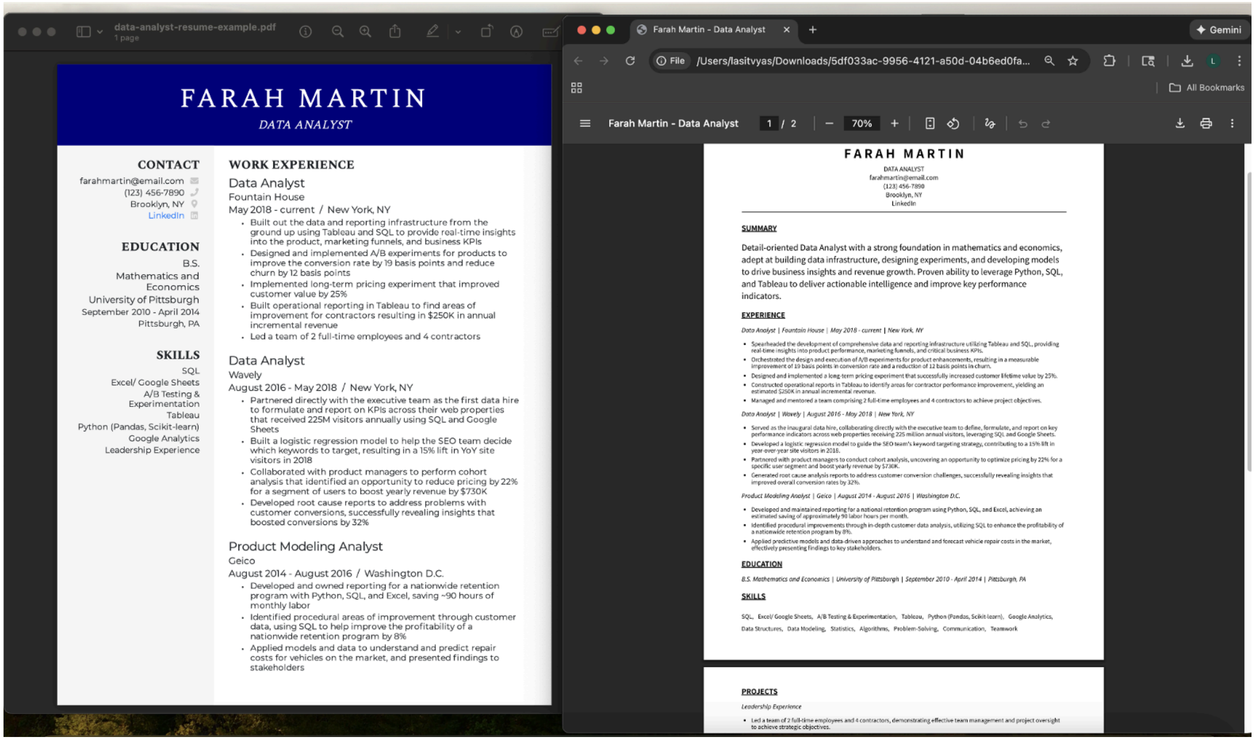
One of the most notable impacts is the transformation of weak or generic resume bullet points into strong, quantified statements. The fine-tuned language model consistently enhances descriptions by adding measurable outcomes, action-oriented language, and industry-relevant terminology. This aligns resumes more closely with ATS parsing algorithms while also improving readability for human recruiters.

Another major impact is efficiency. Tasks that traditionally require hours of manual rewriting and formatting are completed within minutes. This reduction in time and effort lowers the barrier for users to tailor resumes for multiple job applications, increasing their chances of success in competitive hiring pipelines.

The system also improves consistency. Because resume generation follows a structured pipeline, formatting remains professional and uniform across different resumes and job roles. This eliminates common issues such as inconsistent spacing, poorly organized sections, or uneven bullet formatting.

From a technical standpoint, the project validates the effectiveness of combining LoRA fine-tuning with 4-bit quantization. The model delivers high-quality outputs while remaining lightweight and cost-efficient, demonstrating that high performance does not require large-scale models when tasks are well-scoped and domain-specific.

Overall, CareerCraft AI demonstrates measurable impact in resume quality, user efficiency, and system scalability, confirming the feasibility of deploying Generative AI for real-world career development applications.


data-analyst-resume-example.pdf
1 page
Farah Martin - Data Analyst
Gemini
File
/Users/lasitvyas/Downloads/5df033ac-9956-4121-a50d-04b6ed0fa...
All Bookmarks
Farah Martin - Data Analyst
1 / 2
70%

FARAH MARTIN

DATA ANALYST

CONTACT

farahmartin@email.com
(123) 456-7890
Brooklyn, NY
LinkedIn

EDUCATION

B.S.
Mathematics and Economics
University of Pittsburgh
September 2010 - April 2014
Pittsburgh, PA

SKILLS

SQL
Excel/ Google Sheets
A/B Testing & Experimentation
Tableau
Python (Pandas, Scikit-learn)
Google Analytics
Leadership Experience

WORK EXPERIENCE

Data Analyst
Fountain House
May 2018 - current / New York, NY

- Built out the data and reporting infrastructure from the ground up using Tableau and SQL to provide real-time insights into the product, marketing funnels, and business KPIs
- Designed and implemented A/B experiments for products to improve the conversion rate by 19 basis points and reduce churn by 12 basis points
- Implemented long-term pricing experiment that improved customer value by 25%
- Built operational reporting in Tableau to find areas of improvement for contractors resulting in $250K in annual incremental revenue
- Led a team of 2 full-time employees and 4 contractors

Data Analyst
Wavely
August 2016 - May 2018 / New York, NY

- Partnered directly with the executive team as the first data hire to formulate and report on KPIs across their web properties that received 225M visitors annually using SQL and Google Sheets
- Built a logistic regression model to help the SEO team decide which keywords to target, resulting in a 15% lift in YoY site visitors in 2018
- Collaborated with product managers to perform cohort analysis that identified an opportunity to reduce pricing by 22% for a segment of users to boost yearly revenue by $730K
- Developed root cause reports to address problems with customer conversions, successfully revealing insights that boosted conversions by 32%

Product Modeling Analyst
Geico
August 2014 - August 2016 / Washington D.C.

- Developed and owned reporting for a nationwide retention program with Python, SQL, and Excel, saving ~90 hours of monthly labor
- Identified procedural areas of improvement through customer data, using SQL to help improve the profitability of a nationwide retention program by 8%
- Applied models and data to understand and predict repair costs for vehicles on the market, and presented findings to stakeholders

FARAH MARTIN

DATA ANALYST
farahmartin@email.com
(123) 456-7890
Brooklyn, NY
LinkedIn

SUMMARY

Detail-oriented Data Analyst with a strong foundation in mathematics and economics, adept at building data infrastructure, designing experiments, and developing models to drive business insights and revenue growth. Proven ability to leverage Python, SQL, and Tableau to deliver actionable intelligence and improve key performance indicators.

EXPERIENCE

Data Analyst | Fountain House | May 2018 - current | New York, NY

- Spearheaded the development of comprehensive data and reporting infrastructure utilizing Tableau and SQL, providing real-time insights into product performance, marketing funnels, and critical business KPIs.
- Orchestrated the design and execution of A/B experiments for product enhancements, resulting in a measurable improvement of 19 basis points in conversion rate and a reduction of 12 basis points in churn.
- Designed and implemented a long-term pricing experiment that successfully increased customer lifetime value by 25%.
- Constructed operational reports in Tableau to identify areas for contractor performance improvement, yielding an estimated $250K in annual incremental revenue.
- Managed and mentored a team comprising 2 full-time employees and 4 contractors to achieve project objectives.

Data Analyst | Wavely | August 2016 - May 2018 | New York, NY

- Served as the inaugural data hire, collaborating directly with the executive team to define, formulate, and report on key performance indicators across web properties receiving 225 million annual visitors, leveraging SQL and Google Sheets.
- Developed a logistic regression model to guide the SEO team's keyword targeting strategy, contributing to a 15% lift in year-over-year site visitors in 2018.
- Partnered with product managers to conduct cohort analysis, uncovering an opportunity to optimize pricing by 22% for a specific user segment and boost yearly revenue by $730K.
- Generated root cause analysis reports to address customer conversion challenges, successfully revealing insights that improved overall conversion rates by 32%.

Product Modeling Analyst | Geico | August 2014 - August 2016 | Washington D.C.

- Developed and maintained reporting for a national retention program using Python, SQL, and Excel, achieving an estimated saving of approximately 90 labor hours per month.
- Identified procedural improvements through in-depth customer data analysis, utilizing SQL to enhance the profitability of a nationwide retention program by 8%.
- Applied predictive models and data-driven approaches to understand and forecast vehicle repair costs in the market, effectively presenting findings to key stakeholders.

EDUCATION

B.S. Mathematics and Economics | University of Pittsburgh | September 2010 - April 2014 | Pittsburgh, PA

SKILLS

SQL, Excel/ Google Sheets, A/B Testing & Experimentation, Tableau, Python (Pandas, Scikit-learn), Google Analytics, Data Structures, Data Modeling, Statistics, Algorithms, Problem-Solving, Communication, Teamwork

PROJECTS

Leadership Experience

- Led a team of 2 full-time employees and 4 contractors, demonstrating effective team management and project oversight to achieve strategic objectives.

9. Conclusion & Future Roadmaps

CareerCraft AI successfully demonstrates how Generative AI can be applied to a practical, real-world problem using a thoughtfully engineered system. By integrating frontend design, workflow automation, microservices, and a fine-tuned language model, the project delivers a complete and scalable solution rather than a standalone AI demo.

The project highlights several key lessons. First, domain-specific fine-tuning significantly improves output quality compared to generic language models. Second, workflow orchestration tools such as n8n enable reliable and maintainable AI pipelines. Third, user-centered UI design is essential for making advanced AI systems accessible to non-technical users.

Looking forward, CareerCraft AI has a clear and extensible roadmap. Planned enhancements include LinkedIn profile ingestion to automatically extract professional data, generation of cover letters and portfolios using the same AI pipeline, and chat-based resume editing that allows users to iteratively refine content. Additional resume templates will support different design preferences and industry standards.

From a deployment perspective, future work includes full migration to Azure cloud infrastructure, improved logging and monitoring, and enhanced security controls for handling sensitive user data. Advanced features such as sentiment-aware rewriting and multilingual support would further broaden the system's applicability.

In conclusion, CareerCraft AI serves as a strong example of how Generative AI can be responsibly and effectively deployed to solve meaningful problems. The project not only meets academic objectives but also demonstrates real-world viability, making it a strong foundation for future research, commercialization, or product development.

THANK YOU

GitHub:
https://github.com/aditimishra28/airesumeportfoliobuilder.git

HuggingFace:
https://huggingface.co/aditismile/resume_enhnaced