Nicholas Michaels

11/10/2025

CUE Project Updates

This week, I've decided to check out an old document that was sent to me relating to LLMs, and I would like to inform you of one particular section from the document. The paper, *"AI Hiring with LLMs: A Context-Aware and Explainable Multi-Agent Framework for Resume Screening"* by Jianing Qiu, Zeyu Wang, et al., introduces a novel multi-agent LLM framework integrated with Retrieval-Augmented Generation (RAG) to provide a more objective, context-aware, and explainable solution for automating the time-intensive process of resume screening.

The Related Work section outlines the evolution of AI-driven hiring technologies and establishes the need for the proposed multi-agent, RAG-enhanced Large Language Model (LLM) framework. Early AI systems for resume screening relied on traditional machine learning (ML), such as Bag-of-Words and Support Vector Machines, which treated resumes as structured data and lacked semantic understanding, often failing due to their reliance on exact keyword matching. The field advanced to deep learning (DL), incorporating models like RNNs, LSTMs, and later, context-aware transformer-based models (e.g., BERT). These improvements allowed AI to capture semantic and sequential information, recognizing terms like "software engineer" and "software developer" as contextually related, but they still required extensive labeled data for training, limiting their adaptability across diverse hiring contexts.

The shift to modern LLMs enabled zero-shot and few-shot learning, allowing models to leverage large-scale pre-training for deeper contextual reasoning, such as inferring implicit or transferable skills, which is a significant advance over earlier models. However, the paper notes that existing LLM-driven screening systems typically operate as monolithic models. This single-LLM approach creates a lack of modularity and requires resource-intensive retraining or fine-tuning (e.g., using LoRA) every time the scoring logic or hiring criteria change. Furthermore, LLMs rely on static pretraining data, hindering their ability to adapt to the constantly changing requirements of the job market. The section, therefore, highlights Retrieval-Augmented Generation (RAG) as a key solution. While RAG is already widely explored in domains like legal research and finance to integrate real-time, domain-specific information, its application in resume screening is still limited, providing a clear gap for the paper's novel multi-agent, RAG-LLM approach to address.

One document that I found on this subject is *"Application of LLM Agents in Recruitment: A Novel Framework for Automated Resume Screening"* by Chengguang Gan, Qinghao Zhang, Tatsunori Mori. This presents an innovative, efficiency-driven system that uses a multi-agent

framework built on Large Language Models (LLMs) to automate the initial phases of talent acquisition. The core objective of this framework is to significantly reduce the time and effort recruiters spend on manual screening by transforming the process into an intelligent, automated pipeline. The system first converts diverse resume formats into a structured format. It then utilizes a fine-tuned, open-source LLM (such as LLaMA2) to perform sentence classification, tagging every line of the resume with categories like "experience," "skills," and "education." This step is crucial for preparing the data, as it is also used to remove personal identifying information (PII), thereby enhancing fairness and reducing the risk of algorithmic bias.

Following the data preparation, the specialized LLM agents take over the core evaluation tasks. These agents are designed to simulate the decision-making process of human HR professionals by summarizing each resume and assigning a final grade or rating. By using a simulated dataset of real resumes, the authors demonstrated the framework's superior efficiency and high accuracy. Specifically, the system was able to process resumes *11 times faster* than traditional manual methods, freeing up human recruiters for more strategic tasks. Furthermore, the framework's fine-tuned model achieved a robust F1 score of *87.73%* for resume sentence classification, confirming its reliability in understanding and structuring the complex, unstructured data found in job applications.

This article is relevant to my project because it is incredibly valuable to my resume analysis, which is aiming for a Groq-based format. What I've learned is that the key to building a robust, high-performance system isn't the model itself, but the architecture. I need to shift my thinking from "How can a single LLM analyze a resume?" to "How can I orchestrate multiple fast agents to collaborate on the analysis?" Groq's low latency is essential here; it allows me to execute the sequential or parallel steps of the proposed framework—where one agent extracts data, a second evaluates it against the job description using RAG (Retrieval-Augmented Generation), and a third formats the score—with near-instantaneous speed. This confirms that the right path for my project is to implement a modular, multi-agent system that leverages RAG for context, which will, in turn, fully capitalize on Groq's unique advantage in fast, repetitive inference.

Speaking of my project, I have tested out a dozen more resumes that were generated by ChatGPT, hoping that they would work well with the Ollama aspect of the interface despite the seeming differences in logic, bringing strong, useful, reasonable, convincing feedback to the user upon their resume. Based on the two articles, I am on the stance that *the reliability and quality of my analysis hinges less on the source of the resume (be it human or ChatGPT) and more on the rigorous implementation of a multi-agent architecture and Retrieval-Augmented Generation (RAG).* This means I must focus my development efforts on designing specialized agents—an Extractor for transforming unstructured resume text, an Evaluator to ground the assessment in job criteria retrieved via RAG, and a Summarizer to generate the final, convincing feedback—because this modular approach, accelerated by Groq's low-latency inference, is the

only way to guarantee the analysis is objective, context-aware, and fast enough for a real-world application, regardless of the input data's origin.

Throughout my experimentation, I found that most of the results of the resumes prompted out suggestions for priorities like “Update resume with quantified achievements, or Build portfolio of 3-5 relevant projects, or Seek mentorship from senior professionals, or Build professional network of 500+ connections”, which reveals to me that while the resumes weren’t fully authentic, they were able to result in ideal feedback for any user to go about their next steps. When it came to the Market Intelligence and Risk Assessment, though, it appears that a majority of the applications need “Implementation of keyword optimization, use of standard formatting, avoidance of tables and complex layouts, or Revision resume to include numbers: percentages, dollar amounts, team sizes, project scopes or **Job Market:** High demand in Arts, **Growth Rate:** 15-20% annually, **Remote Opportunities:** 60% of positions”. I take this to mean that this is the most common denominator for a person to be within the right range of employment despite what their current beliefs are. Also, a lot of times “Lack of metrics may weaken your impact stories and Several in-demand skills are missing from your profile” were warnings for the dummy resumes due to their seemingly small amount of material for each resume. What was worst of all though, was that when the jobs listed, especially for places that were outside of the US like Pune, Maharashtra or Gurugram, Haryana, no matches were found after clicking the links in question. It went to show that most of those kinds of jobs were prevalent in those areas, so ChatGPT was a big hindrance in that situation as it was making randomizations without any second thought or logic.

This finding matters to my resume analysis project because it exposes critical limitations in the system's geographic and contextual accuracy. While the AI successfully identified universal areas for professional improvement—such as quantifying achievements, optimizing keywords, and strengthening impact stories—the recommendations were undermined by significant data mismatches. The system's US-centric career database (featuring companies like Google, Microsoft, and Amazon with salary ranges in USD) was fundamentally incompatible with India-based test resumes representing professionals in Kochi, Mumbai, Pune, and Bengaluru. This resulted in LinkedIn job searches that returned few or no relevant results, as the system paired Indian locations with Western companies that may not have substantial operations in those regions.

The geographic disconnect revealed deeper structural issues**:** the hardcoded `CAREER_FIELDS` data lacks region-specific employers (such as TCS, Infosys, Wipro for IT in India), uses inappropriate salary formats (dollars instead of lakhs), and omits locally significant sectors like **BPO/KPO**, textiles, and agriculture technology. Additionally, the "geographic hotspots" recommendations (San Francisco, New York, Seattle, Austin) were irrelevant to users seeking opportunities in Indian markets.

Despite these limitations, the recurring themes in feedback around metrics, formatting, and keyword usage do indicate foundational factors influencing resume effectiveness across industries. However, for the system to provide truly actionable, realistic suggestions, it requires location-aware intelligence that dynamically adapts company recommendations, salary benchmarks, certification requirements, and market trends to the user's actual geographic and professional context. Without this adaptation, the market intelligence insights—while conceptually valuable—fail to meaningfully improve a candidate's competitiveness in their target job market.

Following the initial testing phase, I recognized the need to evaluate the system with authentic resumes. After analyzing three real resumes from friends—one in Psychology, one in Teaching, and another in Accounting—I observed a dramatic shift in the quality and relevance of the AI-generated feedback. The recommendations became notably more personalized and contextually grounded. For instance, the psychology graduate received tailored advice like *"Connect with licensed clinicians in your area" and "Consider pursuing certifications in CBT or trauma-informed care,*" while the teaching candidate saw specific guidance such as *"Prepare for common Education interview questions. Practice STAR method responses. Research target companies thoroughly. Prepare questions to ask interviewers."* The accounting professional received field-specific suggestions including *"Study the history of your company and how accounting standards evolved" and "Pursue CPA licensure or specialized tax certifications."*

This marked improvement revealed a critical insight about AI resume analysis: **the system's effectiveness is directly proportional to the authenticity and completeness of the input data.** My friends had invested considerable effort into crafting resumes that genuinely reflected their professional identities—highlighting specific coursework, internships, volunteer experiences, technical proficiencies, and career aspirations. These resumes weren't generic templates; they contained the narrative threads, quantifiable achievements, and domain-specific terminology that allowed the AI to perform nuanced pattern recognition and generate actionable, career-stage-appropriate recommendations.

The contrast between dummy and authentic resumes illuminated several factors:

1. **Narrative Coherence** - Real resumes tell a career story with logical progressions, whereas dummy resumes often present disconnected job titles without the contextual tissue that connects them.
2. **Specificity of Language** - Authentic resumes use precise terminology (e.g., "implemented trauma-informed practices with at-risk youth" vs. generic "worked with students"), giving the AI richer semantic material to analyze.
3. **Emotional Investment** - My friends' resumes reflected genuine professional goals and personal stakes, which translated into more detailed skill descriptions and carefully curated experiences that the AI could leverage for meaningful guidance.

4. **Industry-Specific Details** - Real professionals naturally include certifications, software proficiencies, and methodologies relevant to their fields (Tally for accounting, STAR method awareness for teaching interviews), creating hooks for the AI to provide specialized advice.

Perhaps most significantly, the tone of the **AI feedback shifted.** With authentic resumes, the system generated encouragement that felt genuinely supportive rather than formulaic—phrases like *"Your background in educational psychology positions you well for student counseling roles"* or *"Your experience with auditing and GST compliance makes you competitive for senior accounting positions."* This wasn't just technical analysis; it was career counseling that acknowledged the candidate's journey and validated their preparation efforts.

This experience underscored a fundamental principle: AI-powered career tools require a "full, personalized backstory" to function optimally. The machinery doesn't just parse keywords—it identifies patterns, recognizes professional trajectories, and contextualizes individual experiences within industry standards. When fed shallow or fabricated data, it produces shallow outputs. But when given the rich, authentic material that real job seekers provide, the system demonstrates its true potential to deliver insights that are not only technically accurate but also emotionally resonant and strategically valuable. The authentic backgrounds of my friends allowed the AI to do what it was designed for: provide guidance that felt less like algorithmic output and more like advice from a knowledgeable mentor.

This report ultimately concludes that the next generation of AI-driven talent acquisition hinges on a paradigm shift from **monolithic models to orchestrated**, multi-agent frameworks that prioritize modularity, context, and speed. By successfully defining a specialized architecture—featuring dedicated Extractor, RAG-enhanced Evaluator, and Summarizer agents—we have established a robust solution that overcomes the key limitations of conventional systems: the high cost of retraining and the lack of real-time market awareness. Crucially, this modular design is uniquely positioned to capitalize on hardware innovations like Groq's low-latency inference, proving that high-velocity, objective, and explainable resume analysis is not just possible, but the immediate next step for high-stakes HR applications. This focus on architectural rigor over sheer model size provides a scalable and adaptable blueprint for delivering personalized, data-driven career feedback regardless of the input data’s origin.