An AI-Driven Lens on The Demand side of the Egyptian Labor Market (2021-To Date)

Part III: JobIt - A Conversational AI System for Interactive Labor Market Analysis

WP No. 243

July 2025

This working paper is the third in a series of papers by the Egyptian Center for Economic Studies (ECES), explaining the methodology for assessing skill demand in Egypt's labor market using AI. It is authored by Ahmed Habashy, AI Engineer. The core project team also includes Ahmed Dawoud, Head of the Data Analytics Unit; Sondos Samir, Research Analyst; Youssef Nasr, Research Analyst; Aya Saleh, Research Analyst; and Abdallah El-Lawah, External AI Consultant.

Abstract

While large-scale datasets offer immense potential, their complexity often creates a barrier to entry for non-technical users, limiting the scope of analysis to static reports and pre-defined dashboards. This paper introduces JobIt, the third and final component of the ECES labor market analysis system, designed to bridge this gap. JobIt is a proprietary conversational AI system that transforms the project's comprehensive labor market dataset into a dynamic, interactive resource. Built on an agentic Retrieval-Augmented Generation (RAG) framework, JobIt combines the reasoning of Large Language Models (LLMs) with specialized, high-fidelity tools, crucially, an SQL Agent that grounds all quantitative analysis in an ECES-verified dataset, ensuring accuracy and mitigating model hallucination. The system's dual-mode architecture enables users to generate complex, interactive visualizations and conduct deep, conversational inquiries using natural language. By translating user intent into executable code and structured queries, JobIt democratizes data exploration, empowering policymakers, researchers, and analysts to unlock nuanced insights from complex datasets without requiring specialized programming skills.

ملخص

على الرغم من القيمة الكبيرة لقواعد البيانات الضخمة، فإن تعقيدها غالباً ما يمنع غير المتخصصين من الاستفادة منها، مما يحصرهم في نطاق التقارير الجاهزة. لمعالجة هذه الفجوة، تقدم هذه الدراسة نظام "جوب إت" (JobIt) وهو نظام ذكاء اصطناعي حواري يتيح للمستخدمين أن يتحدثوا مع البيانات بشكل مباشر. ولضمان موثوقية النتائج، يستقي النظام إجاباته مباشرة من قاعدة بيانات المركز الموثقة، مما يمنع الذكاء الاصطناعي من اختلاق إجابات غير حقيقية. يمكن استخدام النظام بطريقتين رئيسيتين: الأولى هي إنشاء رسوم بيانية تفاعلية بطلبات بسيطة، والثانية هي الدخول في حوار تحليلي عميق وطرح أي سؤال والحصول على اجابات تفصيلية دقيقة. بهذه الطريقة، يزيل "جوب إت" الحاجز التقني، ويمكّن صانعي السياسات والباحثين من استكشاف البيانات بأنفسهم، واستخلاص رؤى قيمة دون الحاجة لأي مهارات برمجية.

Disclaimer

The findings and analysis presented in this study are based exclusively on online job postings sourced from trusted and credible platforms. While this approach does not capture the entirety of labor market demand in Egypt, it offers a valuable and timely perspective into employer needs. Online data provides real-time, continuously updated insights that reflect evolving market trends and recruitment practices. As digital platforms increasingly become the primary channel for job advertising, this method represents a forward-looking approach to understanding the demand side of the labor market with greater relevance, accuracy, and immediacy.

Contents

List of Figures

1. Introduction

The preceding parts of this series have detailed the systematic collection of labor market data (Part I) and the AI-powered classification of job postings according to the international ISCO-08 standard (Part II). The result of this multi-stage process is the Classified Clean Dataset (CCD)—a rich, high-fidelity repository of employer demand in Egypt. However, the existence of a quality dataset is only the first step; unlocking its full value requires powerful, intuitive, and reliable analytical tools that move beyond the confines of static quarterly reports.

This paper, Part III, presents the final component of our analytical pipeline: **JobIt**, a proprietary conversational AI system designed specifically for the deep exploration of the CCD. JobIt represents the culmination of our effort to not only gather and structure data but to make it fully accessible and queryable. It was conceived to address the "last-mile" challenge in data analytics: bridging the gap between a complex, structured database and the natural language questions of policymakers, researchers, and analysts.

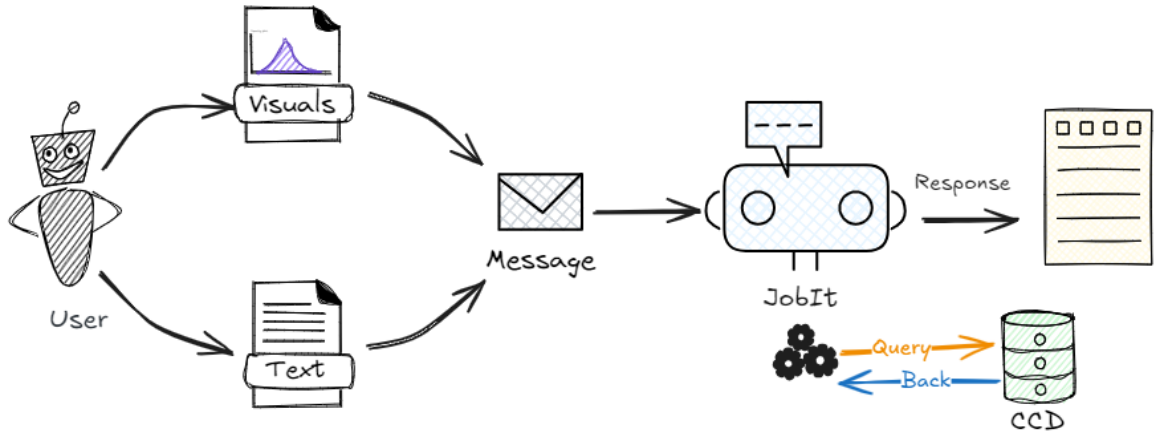
Unlike general-purpose data analysis tools, JobIt is a purpose-built analytical partner, exclusively tailored to the nuances of the ECES dataset. Its development was driven by the need for a system that could ensure data sovereignty, provide deep contextual relevance, and, most importantly, guarantee analytical reliability. This is achieved through an innovative agentic architecture that delegates quantitative tasks to a deterministic SQL Agent, effectively using the LLM for reasoning and translation while entrusting the structured data retrieval to a high-precision tool.

This paper details the rationale, architecture, and functionality of JobIt. We explain why a specialized solution was necessary and provide a comprehensive overview of its dual-mode operation, enabling both dynamic visualization and deep conversational inquiry. In doing so, we present a reliable framework for interactive engagement with complex economic data, demonstrating how conversational AI can serve as a robust and reliable interface for evidence-based policymaking.

2. JobIt - AI-Driven Visual and Textual Insights

JobIt, launched in Q4 2024, provides an AI-powered interface designed to facilitate deeper and more intuitive interaction with ECES's labor market data (Figure 2.1.).

Figure 2.1.: How JobIt Works?



Source: Author's own elaboration.

2.1. Why Jobit?

The development of JobIt was driven by several key factors:

- **Unlocking Textual Data:** The extensive accumulation of rich, unstructured text data within job descriptions and requirements presented a significant analytical opportunity. Traditional methods, primarily suited for tabular data, were insufficient to fully exploit the nuances embedded in this text. JobIt was conceived to analyze this textual corpus, providing a more granular understanding of employer demands.
- **Enhanced Contextual Relevance:** Generic Large Language Models (LLMs) lack the specific contextual understanding of the Egyptian labor market. JobIt, conversely, has exclusive access to ECES's comprehensive dataset. This grants it deep local context awareness – for example, recognizing "Shatoura" not just as a name, but specifically as a village within the Sohag governorate, enabling more accurate and relevant insights.
- **Data Sovereignty and Security:** A paramount design consideration was maintaining control and privacy over ECES's data. JobIt is architected to operate within a secure environment that explicitly prevents the sharing of underlying data with external LLM providers (e.g., OpenAI, Google).
- **Analytical Reliability and Accuracy:** A fundamental challenge of general-purpose LLMs is their tendency to "hallucinate"—inventing plausible but incorrect

information. To ensure trustworthy analysis, JobIt is architected from the ground up to mitigate this risk through a Retrieval-Augmented Generation (RAG) framework. At its core is a specialized SQL Agent. Instead of relying on the LLM's internal knowledge, JobIt tasks the LLM with translating user queries into precise SQL commands that are executed directly against the structured Classified Clean Dataset (CCD). The final answer is generated exclusively from this retrieved data, ensuring every quantitative insight is grounded in fact. Crucially, if information is not present in the dataset, JobIt is designed to state this limitation explicitly rather than invent a response. This design choice significantly enhances mathematical accuracy and makes JobIt a reliable tool for evidence-based analysis.

2.2. The Need for a Specialized Solution

While several general-purpose LLM-driven tools facilitate data exploration through natural language, their inherent design for broad applicability presents limitations when deep, dataset-specific insights are required. For instance, **LIDA** (Large Language Model-based Data Analytics Framework) excels in Exploratory Data Analysis (EDA), particularly when users approach new datasets without predefined analytical questions and seek suggestions for initial visualizations. Similarly, **RTutor** serves effectively as an interactive coding tutor, demonstrating simple visualizations in response to prompts, making it a valuable tool for learning basic visual tasks and R language basics.

However, the general-purpose nature of such tools means they are often less suited for complex querying or highly customized analytical workflows on a specific dataset. Their adaptability to any data source, while a strength, makes it challenging to deeply tailor their underlying code generation or analytical processes to the unique characteristics and semantic intricacies of a specialized domain, such as the Egyptian labor market.

Recognizing these constraints, **JobIt** was developed. The primary motivation was to create a solution explicitly customized for the ECES Egyptian job market database. This focused approach allows for the implementation of specific customizations and optimizations tailored to this dataset, enabling the extraction of more profound and contextually relevant insights with the assistance of LLMs. By concentrating on a single, well-understood data source, JobIt can leverage domain-specific knowledge and pre-processing to a degree not feasible with general-purpose tools, thereby maximizing the analytical potential.

To illustrate this critical distinction, consider a researcher posing the following, seemingly simple query: ***Show me the IT job count in Alexndria***

A general-purpose tool, designed for broad applicability, would likely fail on two fundamental levels. First, it would treat "Alexndria" as a literal string. Upon executing a

query like `WHERE city = 'Alexndria'`, it would find no matches and incorrectly report that there are zero jobs, unable to recognize the common misspelling of "Alexandria."

Second, and more subtly, it would struggle with the ambiguity of "IT job." Does this refer to any job at a company whose primary business is in the *'Information Technology' sector*, or does it refer to a specific *job function* like 'IT/Software Development', regardless of the employer's industry (e.g., an IT role at a bank)? The dataset contains both a `sector` column and a `job_category` column. Lacking specific domain knowledge about how labor markets are analyzed, the general tool would have to guess which column to query, or it might present the user with a confusing choice. This ambiguity leads to unreliable and inconsistent results.

In contrast, **JobIt** is engineered to resolve these issues by design. Its customized workflow handles the query as follows:

1. **Cardinality Resolution:** It immediately recognizes "Alexndria" through its pre-trained knowledge of common Egyptian place names and their variations, automatically mapping it to the correct, canonical value: **'Alexandria'**.
2. **Intent-to-Schema Mapping:** It understands that a query about "IT jobs" most commonly refers to the occupational role. Therefore, it correctly targets the `job_category` column and filters for the value **'IT/Software Development'**, rather than guessing between columns.

The resulting query is both technically correct and semantically aligned with the researcher's true intent. This ability to navigate linguistic and structural ambiguities is not a feature of a general-purpose model but a core strength derived from its specialized design, making it an indispensable tool for accurate labor market analysis.

2.3. How it Works?

JobIt operates as an **Agentic Retrieval-Augmented Generation (RAG)** system that combines the analytical reasoning capabilities of large language models (LLMs) with the power of code generation to deliver accurate insights and dynamic visualizations. The system supports two distinct operational modes:

2.3.1 Visualization Mode

In this mode, JobIt operates exclusively as a code-generation engine focused on delivering interactive visual representations of data. **Crucially, this mode does not generate any textual analysis or explanations; its sole output is chart code.** Rather than analyzing the full dataset on each request, JobIt utilizes a **pre-generated JSON summary** that describes the structure of the ECES labor market dataset. This summary

includes metadata such as column names, data types, unique values, and column summaries. It is provided to the LLM at runtime to help it understand the dataset schema and generate accurate visualization code. Alongside this JSON summary, a predefined visualization code template (a pre-written, parameterized code structure in Python) is supplied to guide the code generation process.

With these tools, the LLM produces fully functional, editable, and interactive graph code—typically written in Python using libraries such as Plotly. Each user interaction maintains a custom session history, allowing follow-up prompts to refine, reformat, or regenerate visuals with ease, all through natural language. This design ensures that users can quickly iterate on data visualizations without receiving extraneous or potentially distracting textual interpretations.

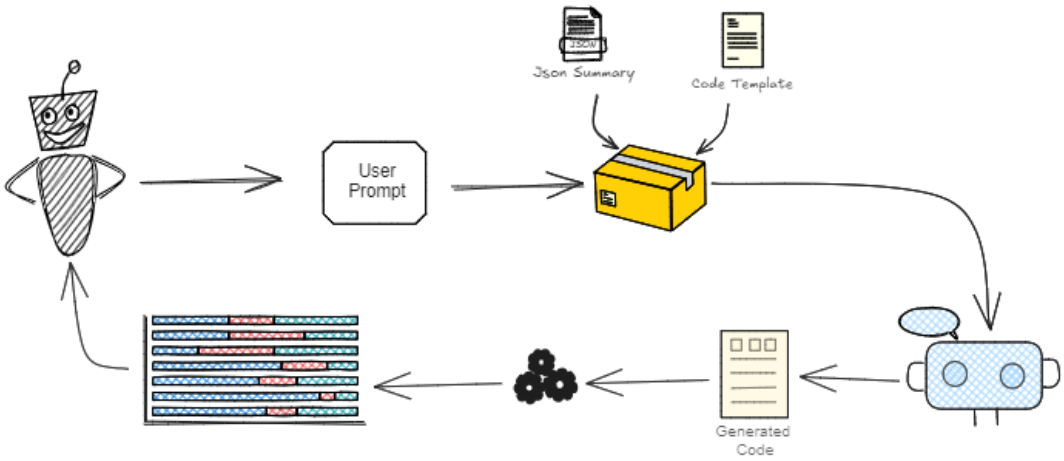
Here's an example workflow for the Visualization Mode:

1. The user enters a natural language prompt (e.g., "Show me the number of jobs in Alexndria").
2. This prompt is sent to the Large Language Model (LLM) along with the pre-generated JSON dataset summary and a visualization code template.
3. The LLM analyzes the prompt and edits the code template, determining the required graph type (e.g., bar chart).
4. Using the provided JSON summary, the LLM identifies that the user's query pertains to the "governorate" column and maps the input "Alexndria" to the correct, case-sensitive dataset value "Alexandria".
5. The LLM populates the code template with the necessary parameters to generate the final visualization code.
6. The generated code is sent to an execution environment, which runs the code and displays the interactive graph directly to the user.

2.3.2 Chat Mode

In this conversational mode, JobIt transforms from a code generator into a full-fledged analytical partner, leveraging the power of state-of-the-art LLMs (from providers like Google and OpenAI) to create a rich, interactive dialogue. While its core strength is accurate data retrieval, this mode also inherits the versatile capabilities of its underlying models. Users can engage in a natural, back-and-forth conversation, asking follow-up questions, requesting data to be summarized, translated into other languages, or reformatted into lists, tables, or paragraphs. This allows for a truly dynamic exploration where the initial query is just the starting point of a deeper analytical journey.

Figure 2.2.: Visualization Workflow



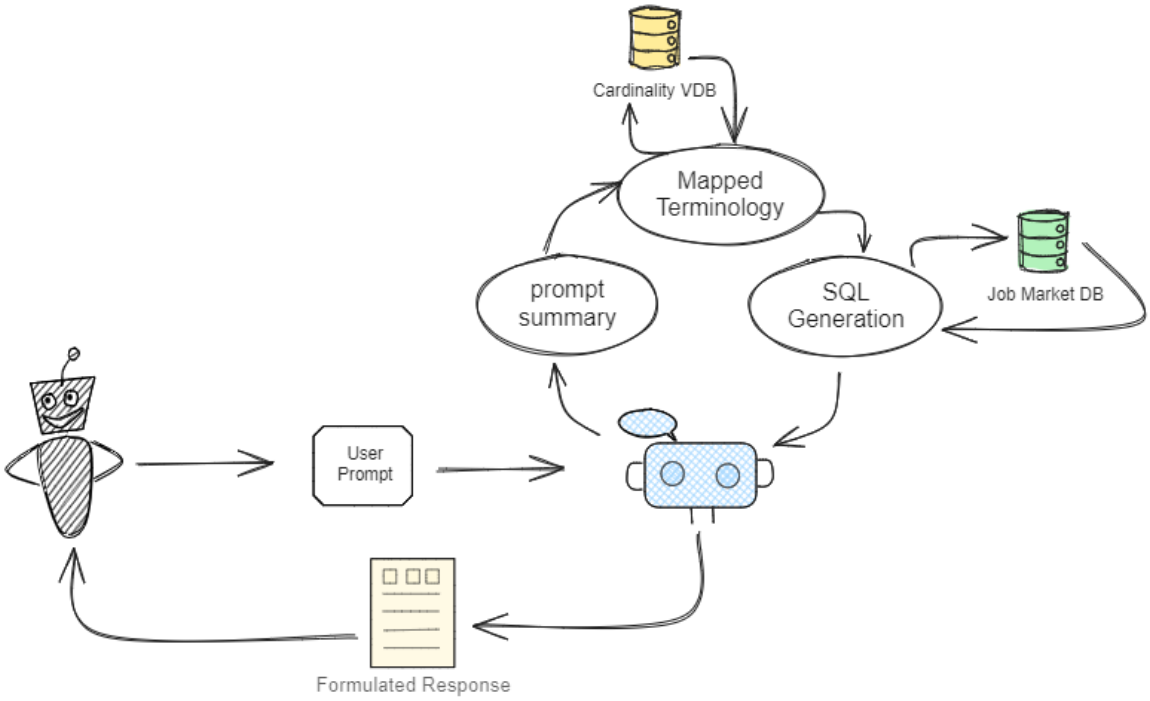
Source: Author's own elaboration.

To ensure this flexibility does not compromise accuracy, the system is designed to generate quantitative responses exclusively grounded in the ECES dataset, effectively minimizing hallucinations. Upon receiving a user query, the main LLM agent first evaluates if it can be answered directly without data access (e.g., general questions about JobIt's purpose or creators). If so, it responds immediately. For queries requiring data from the ECES dataset (which has been transformed into a structured SQL database), the following RAG workflow is initiated:

1. The user enters a natural language prompt (e.g., `"top 10 most asked for jobs in the IT sector in cairo"`).
2. The main LLM agent takes this prompt and simplifies it into a concise summary of the user's intent (e.g., `"10 frequent IT sector cairo"`).
3. The cardinality resolution tool then processes this simplified summary. This tool is crucial for handling variations in user terminology and maps them to precise dataset schema entities using a vector database. For example, it might rewrite the summary to `"10 frequent IT/Software Development job_category in Cairo"`.
4. An SQL generation tool takes this refined and unambiguous summary to formulate an appropriate SQL query.
5. This generated SQL query enters a feedback loop with an SQL query checker tool, which meticulously validates its syntax and logic.
6. Once confirmed to be error-free, the query is executed against the ECES structured SQL database.
7. The data retrieved from the database execution is passed back to the main LLM agent.

8. Finally, the agent synthesizes this retrieved data into a coherent, natural language answer, which is then presented to the user.

Figure 2.3.: Chat Workflow



Source: Author's own elaboration.

This structured process ensures that all quantitative claims are factually accurate. For instance, after retrieving the top 10 IT jobs, a user could ask a follow-up like, *"Can you count how many of those are related to cybersecurity?"* or *"Please present that list as a Markdown table and translate the job titles to French."* The LLM would perform these tasks using the data it has already retrieved, demonstrating a powerful fusion of high-fidelity data analysis with the flexible, multi-purpose capabilities of modern AI.

3. Conclusion

This paper has presented JobIt, the interactive analytical interface that completes the ECES AI-powered labor market intelligence system. Serving as the third pillar of our methodology, JobIt was conceived to address the "last-mile" challenge of data analysis: transforming a vast, clean, and classified dataset from a static asset into a dynamic and accessible resource for discovery. By enabling users to query complex data and generate bespoke visualizations through natural language, it significantly lowers the technical barrier to entry for deep labor market inquiry.

The system's true novelty lies not merely in the application of an LLM, but in its carefully designed agentic architecture. The integration of specialized, high-fidelity tools—most

notably the SQL Agent—is the cornerstone of its reliability. This framework transforms the LLM from a probabilistic text generator into a deterministic analytical engine for quantitative queries, grounding its responses in verifiable data and effectively mitigating the risk of factual inaccuracies or "hallucinations." This commitment to accuracy and data sovereignty ensures that JobIt serves as a trustworthy partner for high-stakes analysis.

Together, the three components detailed in this series—a robust data collection and cleaning pipeline (Part I), a high-accuracy ISCO-08 classification engine (Part II), and the JobIt conversational interface (Part III)—form a cohesive, end-to-end framework. This integrated system moves beyond traditional, static reporting, offering a new paradigm for real-time labor market intelligence. By empowering policymakers, researchers, and the public to engage directly and intuitively with data, this initiative aims to foster a more responsive and evidence-driven approach to workforce development and economic planning in Egypt.

Bibliography

Gupta, S., R. Ranjan, & S. N. Singh, 2024. "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions". *arXiv preprint arXiv:2410.12837*.

Liu, J., & S. Chu, 2025. "Text2SQL Business Intelligence System Based on Retrieval-Augmented Generation (RAG)". *Engineering Reports*, 7.

Ma, L., Zhang, R., Han, Y., Yu, S., Wang, Z., Ning, Z., Zhang, J., Xu, P., Li, P., Ju, W., Chen, C., Wang, D., Liu, K., Wang, P., Wang, P., Fu, Y., Liu, C., Zhou, Y., & Lu, C.-T. 2025. "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge". *arXiv preprint arXiv:2310.11703*.