Responsible Use of LLM Models for Labor Market Foresight: The Experience of a Vocational Education and Training Institution

Cicero Braga[1], Yuri Lima[1,2], Inês Pereira[1]

[1] Gerência de Prospecção e Avaliação Educacional - Serviço Nacional de Aprendizagem Comercial (SENAC) – Departamento Nacional.
Rio de Janeiro – RJ – Brasil

[2] Laboratório do Futuro –Programa de Engenharia de Sistemas e Computação – COPPE/UFRJ

cicero.braga@senac.br; yuri.lima@senac.br; ines.pereira@senac.br

Abstract. *This article examines the experience of a Vocational Education and Training institution in utilizing Artificial Intelligence to analyze labor market dynamics and guide the provision of VET programs. It presents the structuring of methodologies capable of delivering efficient and robust results, derived from large datasets, in an ethical and responsible manner.*

1. Introduction

Vocational Education and Training (VET) institutions must continuously adapt to ensure that their course offerings keep pace with social and technological changes while addressing the demands of the labor market. Mapping these shifts can be challenging, given Brazil's diverse economic, social, and demographic landscapes, which are all shaped by hard-to-measure technological advancements. Moreover, labor market studies, especially when future-oriented, face significant limitations due to the large volume of data and its inherent constraints – such as periodicity, limited access, lack of longitudinal databases, and delays in data release – or due to methodological costs, including the complexity of econometric models and the trade-offs between external and internal validity. Additionally, there is a wide range of topics that must be addressed when studying labor markets, which limits the possibility of observing it holistically (Eloundou *et al.*, 2024).

In this context, Artificial Intelligence (AI) emerges as a powerful tool for rethinking the intersection between the labor market and VET by expanding, enhancing, and expediting research. However, it's essential to ensure that AI is applied thoughtfully, with clear standards that validate the reliability and robustness of the results. Put simply, AI must uphold methodological rigor to be used ethically and responsibly, as several agencies have demonstrated (UNESCO, 2023; European Comission, 2024).

From this perspective, the goal of this paper is to document two projects that use LLM models as a core part of their methodology, emphasizing ethical and responsible practices. Though differing in scope, both projects were developed within a framework aimed at providing data-driven insights for a VET institution in Brazil. In addition to leveraging AI, these initiatives are linked by their common focus on analyzing the evolving dynamics of the Brazilian labor market.

The institution, National Commercial Apprenticeship Service (SENAC), is the main provider of professional education focused on the trade of goods, services and tourism, and one of Brazil's largest VET providers, with more than 1.6 million enrollments each year across 600+ schools distributed in all 27 states. Given the size of the institution and its impact on people's lives (mostly young and from low-income families), the use of AI for future-oriented labor market studies in the institution must be done responsibly and ethically, from goal definition to how the results are used.

In this sense, we deal with ethical challenges from two perspectives (Jonas, 2006). The first one regarding the use of new technology (LLM, as available via ChatGPT and similar) that gives us considerable powers and has little to no previous use cases. The second one related to visualizing the long-term impact of AI and other technologies and socioeconomic trends on the labor market.

2. Projects

Although the two projects presented have different goals and approaches, both rely on AI tools, LLMs more specifically, as central components of their methodologies. Each study began with a literature review to define typologies and establish other parameters that informed the prompts[1]. These prompts were embedded in Python scripts, which sent the inputs to the GPT-4 Turbo, GPT-4o, or Claude 2 APIs and processed the model responses. While the prompts were designed to meet the specific goals of each project, they shared structural elements and underwent human evaluation to ensure robust and reliable results. Structurally, the prompts provided clear instructions on the task at hand, including the goal of the analysis, the premises that should be taken into account (e.g., the pedagogical model of the institution), the expected response format, a definition of some important elements (e.g., definition of economies of the future, automation technologies), examples of correct and wrong answers, and some prompt "hacks" (e.g. asking the model to think step by step and tipping it) (Bsharat and Myrzakhan, 2023). For each response, the model was asked to justify its choices to help with explainability and quality assurance. In the following subsections, we outline the goals and data for each project, focusing on the human evaluation experience of the AI-generated outputs.

2.1 Project 1: Automation Impact (*Impacto da Automação*)

The "Automation Impact" project analyzed how automation technologies impact technical courses, with a focus on Senac's 35 courses in this modality (Lima and Pereira, 2024). The technologies studied were divided into two groups: Enabling Technologies (biotechnology, blockchain, cloud and edge computing, advanced connectivity, nanotechnology, and Web 3) and Automation Technologies (data analytics, applied AI, 3D/4D printing and modeling, IoT and connected devices, digital platforms and apps, extended reality, and robotics). AI was used here to evaluate the impact of these technologies on the 2,100 Courses Curricular Indicators (CCIs) which are like work tasks.

The writing of the prompt was an iterative process, where each new version was tested in ChatGPT and refined based on the accuracy and quality of the responses received. This iterative process was repeated over 30 times during the writing stage alone. The prompt was followed by a human evaluator that selected six courses from different

[1] Prompts utilizados disponíveis em: https://bit.ly/prompts_laai

economic sectors to both test if the instructions were clear as well as creating examples of correct evaluations for the prompt.

To further refine the prompt, it was implemented in a Python code where the full list of CCIs was evaluated in two equal rounds. After each round, 101 randomly selected CCIs were manually reviewed, with each response classified as: correct (no improvement needed), partially correct (acceptable divergence between AI and human evaluation), or incorrect (unsatisfactory responses). In cases of incorrect evaluations, the column with the issue and a description of the problem were recorded for correction.

Between the first and second evaluation rounds, the accuracy rate increased from 71% to 81%. As we were unable to eliminate the incorrect results (10% in both rounds), partly due to GPT-4's limitations, the responses were verified by another AI model, Claude 2 from Anthropic. The prompt for Claude 2 presented the task originally given to GPT-4, attaching the previous prompt and GPT-4’s responses, and asked for an evaluation of each. When GPT-4's answers were deemed unsatisfactory, Claude 2 was prompted to provide a correct answer and an explanation for its decision. This process resulted in a review of 522 out of the 2,100 evaluated CCIs. These cases were manually reviewed and consolidated by the researchers.

As a result, for each CCI (*e.g.* course: Gastronomy Technician – CCI: organizes work schedules for the kitchen team) we had a technology (scheduling systems, in this case), an automation level (Medium 40-60%) and a rationale for the answer (scheduling systems can generate optimized work schedules but manual adaptations based on unforeseen circumstances are still frequently necessary, requiring human supervision). These results were sent to the consideration of experts to update national curricular plans.

2.2 Project 2: Future Economies (*Economias do Futuro*)

This study identified five economic clusters that highlight specific labor market dynamics: Green; Creative; Digital; Care; Tourism. The goal is to identify and characterize the occupations and economic sectors associated with these "Future Economies" to ultimately align the institution's educational offerings with these clusters. Once each Future Economy was defined, we used GPT-4o to determine whether each of the over 143 thousand activities performed by each occupation, based on the Brazilian Classification of Occupations (CBO), belonged to one or more economic cluster. The prompt incorporated definitions for each economy, including evaluation examples and criteria. After the typologies were set, two researchers and the AI model engaged in an iterative three stages process to ensure the consistency of the results.

In the first stage, 150 random activities were selected, regardless of occupation, and the researchers independently assessed the classification of each. The responses from the researchers were compared both with each other and with the GPT model. Interestingly, the level of agreement between the AI and one of the researchers (61%) was higher than between the researchers themselves (57%), and only 39% of the cases showed full agreement across all parties. Based on these results, and drawing on GPT's justifications for its classifications, the researchers discussed the inconsistencies to reduce subjective interpretations and refine the prompt instructions to align more precisely with the economic clusters definitions.

For the second round, the same process was repeated with another set of 150 random activities, using updated instructions based on the discussions from the first round. While there were still some divergent evaluations, the levels of agreement between the three parts improved to 79%. Once again, GPT's justifications were used to align the human evaluations, ensuring a uniform classification across all activities. As a result, additional classification examples were incorporated into the prompt to further clarify what should be considered correct or incorrect.

Following these adjustments, a final robustness check was conducted with a new set of 150 random activities, in which 91% of cases were classified in agreement between the human evaluators and the AI model. Acknowledging the natural margin of error inherent in inference studies, this version of the prompt was adopted for the evaluation of all 143 thousand activities targeted for assessment. Additionally, the process helped establish an initial evaluation workflow and set parameters for expected error rates.

3. Conclusion and Next Steps

The use of AI in research undoubtedly represents significant methodological progress by expediting analysis and enabling new studies. However, its application requires careful consideration to ensure that the inferences are robust, and the results are presented responsibly. This challenge is especially important when the decision-making processes being supported are sensitive such as in the case of Senac where providing the right professional education can make a considerable difference in the lives of over a million of people each year. Therefore, the execution of our studies using AI go through extensive human evaluation and the results are presented with an emphasis on the methodology used (and its limitations), so other units in the institution can apply them in a critical manner, understanding that AIs can, and will, make mistakes.

In this paper, we highlighted two distinct projects aimed at establishing parameters that ensure the robustness of AI use in future-oriented labor market research. Our experience shows that prompts must undergo several refinements to reach a satisfactory level of response accuracy ensuring alignment between commands and results (reducing hallucination effects) and given the societal relevance of these topics, help mitigate subjective influences (minimizing endogeneity).

Additional projects are currently in development, with an ongoing commitment to data robustness. Ultimately, through these experiences, the goal is to create best practices for internal use of AI in the institution. More importantly, the aim is to expand the scope of research on educational offerings and vocational demand by leveraging innovative yet responsible and ethical AI-driven results.

References

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171.*

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*(6702), 1306-1308.

European Comission (2024). Living guidelines on the responsible use of generative ai in research. ERA Forum Stakeholders' document. *Research and Innovation.*

Jonas, H. (2006). O Princípio da Responsabilidade. Rio de Janeiro: Contraponto

Lima, Y.; Pereira, I. (2024) Estimando o impacto da automação sobre a Educação Profissionalizante: o caso dos cursos técnicos. Scielo Preprints.

UNESCO (2023). *Guidance for generative AI in education and research*.