PathFinder: Intelligent Career Development with Skill Gap Analysis and Personalized Upskilling Pathways

Kamarul Adha
Knowledge Innovation Centre

Anthony F. Camilleri
Knowledge Innovation Centre

ABSTRACT

This paper presents PathFinder, an intelligent career development system that leverages large language models (LLMs) to bridge the gap between individuals' current skills and their career aspirations. The system creates sentence embeddings from user resumes and career goals, then employs vector-based matching against the European Skills, Competences, and Occupations (ESCO) database to identify skill gaps. PathFinder then generates personalised learning pathways by recommending relevant courses from educational platforms like EdX. By combining natural language processing with structured occupational data, PathFinder transforms traditional career guidance into a precise, data-driven process that aligns personal development with specific career objectives. This approach offers a scalable solution for career advancement in an increasingly dynamic job market. PathFinder was developed as part of the LCAMP Project (European Commission [EC], 2024), funded by the European Union through the ERASMUS+ Programme.

KEYWORDS

Skill Embeddings, Vector Search, Career Pathways

1. INTRODUCTION

The rapid evolution of the modern job market has made continuous professional development an imperative rather than an option. While traditional frameworks offer structured approaches to skill acquisition, they often lack the flexibility to address individual career trajectories. The emergence of artificial intelligence (AI), particularly large language models (LLMs), presents an opportunity to revolutionise career development through data-driven approaches to learning (Li, 2022). LLMs have demonstrated remarkable capabilities in understanding complex language patterns and retrieving contextual knowledge across various domains. In education and career development, these technologies excel at processing vast amounts of professional information, making them ideal tools for addressing upskilling and reskilling challenges (Proceedings of the 16th International Conference on Information Systems, 2023).

However, a significant gap exists between the theoretical potential of these technologies and their practical application in delivering actionable career guidance. This paper presents **PathFinder**, an AI-driven system that bridges this gap by leveraging LLM-derived sentence embeddings for skill gap analysis and targeted recommendations. The system's strength lies in its ability to synthesise multiple data sources - including resumes, career goals, and industry-standard frameworks - into concrete development pathways. PathFinder demonstrates how AI can transform traditional career guidance into a precise, data-driven process for navigating modern career landscapes. This development emerged from the objectives of the LCAMP Project Deliverable 5.4 which emphasises human-centric learning pathways tailored to the evolving needs of the advanced manufacturing sector. By employing a systematic approach that integrates ESCO's occupational framework and personalised upskilling pathways, PathFinder exemplifies how AI-driven tools can enhance human-centric career development processes. This initiative underscores the broader objective of LCAMP to promote innovation and adaptability in an evolving industrial landscape.

2. LITERATURE REVIEW

This literature review outlines the core technologies that form the foundation of PathFinder, focusing on how embedding models, sentence embeddings, and retrieval-augmented generation (RAG) enhance personalised career recommendations. By understanding these components, it becomes evident how AI-driven systems can transform traditional career guidance into a dynamic, data-driven process.

2.1 The Role of Embedding Models

Textual data, being unstructured, requires transformation into numerical representations for computational tasks such as similarity measurement, classification, and clustering. Traditional embedding models like Word2Vec, GloVe, and FastText addressed this need but struggled with capturing context in complex or lengthy documents. Large language models (LLMs) have revolutionised embeddings through attention mechanisms that relate words to their context across entire documents and corpora, enabling richer semantic representations (Tao et al., 2024). For this study, the BGE-M3 model, developed by Chen et al. (2024), was selected due to its ability to process multilingual data using the Multi-Lingual Document Retrieval (MLDR) dataset. This model not only understands multiple languages but also possesses extensive knowledge across diverse topics. Running locally on an RTX 4090 GPU, it generates 1,024-dimensional sentence embeddings, effectively capturing long contexts and nuanced semantics. Its open-source availability and performance make it an ideal choice for generating embeddings foundational to the analyses in this paper.

2.2 Sentence Embeddings & Vector Searching Techniques

In modern NLP applications, sentence embeddings offer a more effective way to represent textual data compared to traditional word embeddings. While word embeddings focus on individual tokens, they often fail to capture the broader semantic relationships within a document, leading to a loss of context (Rygl et al., 2017). Sentence embeddings address this issue by converting entire sentences or documents into high-dimensional vectors that preserve semantic meaning across a longer context. This transformation enables more efficient computations for similarity and relevance by mapping documents into a shared vector space, where similar texts are positioned closer together and dissimilar texts are further apart (Sawarkar, Mangal, & Solanki, 2024). For the PathFinder system, the cosine similarity metric was chosen due to its efficiency and ability to calculate the angle between vectors rather than the direct distance, making it particularly suitable for determining semantic proximity between embeddings. This approach ensures that course recommendations align with users' identified skill gaps based on nuanced contextual matching.

2.3 Retrieval-Augmented Generation (RAG) in Information Retrieval

Retrieval-Augmented Generation (RAG) has emerged as a powerful approach to mitigate the limitations of large language models (LLMs) in knowledge-intensive tasks. LLMs are prone to hallucination and have finite sequence lengths, restricting their ability to provide accurate and contextually rich responses without extensive parameter updates. RAG addresses this by integrating a retrieval mechanism that supplements LLMs with external, non-parametric memory, such as dense vector indices, which can be updated without retraining the model (Lewis et al., 2020). This approach enables RAG systems to deliver more accurate and relevant outputs by dynamically retrieving contextual information from large document collections. Furthermore, RAG enhances search and recommendation processes by refining results through semantic similarity, ensuring that retrieved documents are closely aligned with users' queries in meaning rather than merely matching keywords. As noted by Salemi and Zamani (2024), the ability of RAG systems to evaluate and rank retrieved documents based on their relevance to the task significantly improves the quality and accuracy of generated outputs, making RAG particularly valuable for applications requiring up-to-date and precise information.

3. METHODOLOGY

The PathFinder system integrates multiple data sources to provide personalised career development recommendations. These sources include EdX for course data (EdX, n.d.) and ESCO for skill and occupation data (European Commission, n.d.). Each of these sources plays a critical role in matching users with suitable learning pathways based on their career goals and skill gaps.

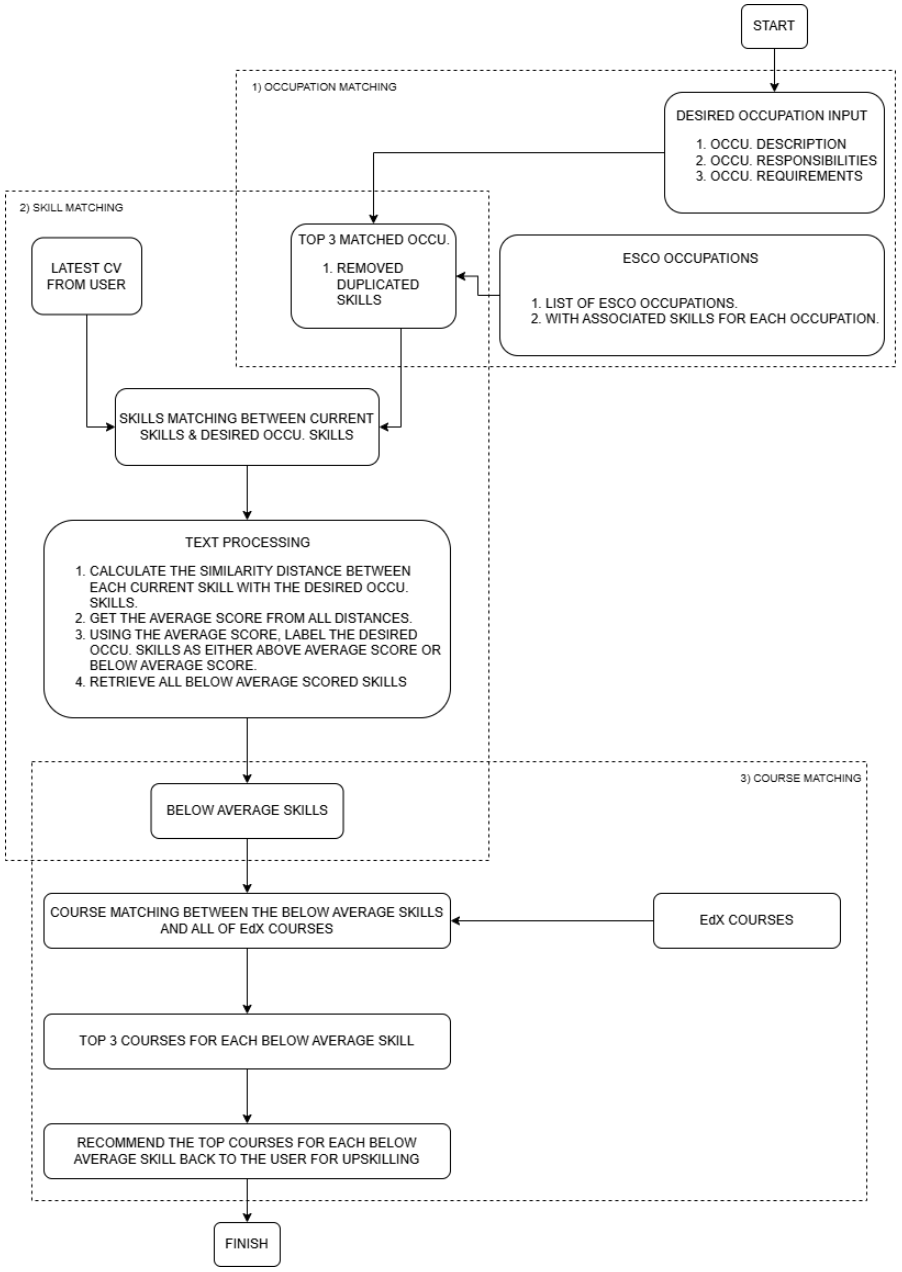


Figure 1. Complete system architecture and data flow from input processing to final output

3.1 Prerequisites

The data required for PathFinder consists of five key components: the user's desired occupation, ESCO occupations, ESCO skills, the user's CV or credentials to display their current skill set, and course data from EdX.

The user inputs, including their desired occupation and CV, are transformed into sentence embeddings to facilitate semantic matching. The ESCO occupations are similarly converted into embeddings, with the associated skills appended to provide additional context. Each ESCO skill is also embedded individually to ensure precise skill matching. Finally, course data from EdX, gathered manually using Python Selenium, is processed into sentence embeddings by combining the title, description, learning outcomes, and other relevant details. These embeddings form the foundation for PathFinder's occupation, skill, and course matching processes.

3.2 Occupation Matching

The first input used in PathFinder is the user's desired occupation, including relevant details such as responsibilities, job descriptions, required skills, and experience. This input is converted into a sentence embedding and mapped against pre-embedded ESCO occupations to identify the closest matches.

The system calculates semantic similarity between the user's input and the ESCO database, selecting the top three occupations with the highest similarity scores. Each selected occupation comes with a list of associated skills, which will be used in the next step to identify skill gaps. Since many skills are shared across different occupations, a filtering process is applied to remove duplicate entries. This ensures that the system processes only unique skills, reducing both computation time and resource usage during the next phase of analysis.

3.3 Skill Matching

The next step involves matching the skills gathered from the selected occupations with the user's current skill set. The user's CV or credentials, uploaded as a PDF file, are parsed and converted into sentence embeddings to extract their existing skills. These embeddings are then compared against the required skills identified from the desired occupations.

The system calculates a similarity score for each required skill by comparing it with the user's current skills. Using these scores, an average similarity value is determined to benchmark the user's skill level. Skills that meet or exceed the average score are classified as above average, while those with lower similarity scores are marked as below average. The below-average skills are flagged as gaps that require improvement and are used in the next process to recommend relevant courses for upskilling.

3.4 Course Matching

Once the below-average skills have been identified, each is compared against the course database from EdX to find relevant learning resources. The system calculates the semantic similarity between the skill embeddings and the course embeddings, selecting the top three courses that best match the context of each skill gap.

The selected courses are then compiled and mapped to the corresponding skill gaps, creating a personalised learning pathway tailored to the user’s specific needs. By recommending targeted courses for each deficient skill, the system provides users with actionable upskilling opportunities that align with the requirements of their desired occupation. This ensures that users can address their skill gaps effectively and progress toward their career goals with relevant educational resources.

4. BENCHMARKING

Benchmarking was conducted to evaluate the retrieval performance of PathFinder's embedding spaces for occupations, courses, and skills. Each embedding space was created using extensive attribute data to capture the context of each record. To test retrieval accuracy, a small segment from each record—such as a title or label—was extracted, converted into embeddings, and compared against the full embedding space using cosine similarity. The goal was to assess how well the system retrieves relevant results with limited input.

Table 1. Performance benchmarking results across different parameters

Embedding Space	Top-3	Top-5	Top-10	Top-25	Top-50	Sum	Total Records	Top-3 (%)
Occupation	2919	60	36	15	5	3035	3039	**96.05%**
Skill	13882	31	15	9	1	13938	13939	**99.59%**
Course	3454	25	21	48	1	3549	3549	**97.32%**

The results show that 96.05% of occupations, 97.32% of courses, and 99.59% of skills were found in the Top-3 most relevant records, demonstrating the effectiveness of the embedding spaces in retrieving contextually similar matches. These high percentages highlight the system's ability to provide accurate recommendations even with minimal input.

However, the sum of matched records is not equal to the total number of records. This indicates that some segmented records did not appear in the Top-50, likely due to insufficient contextual information. Despite these cases, the overall performance shows that the embedding model effectively captures semantic meaning, ensuring reliable retrieval and accurate recommendations in practical use cases.

5. CONCLUSION

PathFinder revolutionizes career development by leveraging LLM-based embeddings and ESCO's framework to deliver precise occupation (96.05%), skill (99.59%), and course (97.32%) matches. Its three-stage process—occupation matching, skill gap analysis, and course recommendations—bridges career goals with market demands. The system's effectiveness in addressing evolving job markets hinges on maintaining up-to-date datasets, ensuring its data-driven insights remain relevant and impactful for personalized professional growth.

ACKNOWLEDGEMENT

This research was conducted under the framework of the LCAMP Project, whose support and resources made this innovative development possible. The LCAMP Project is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

I extend my heartfelt gratitude to the Knowledge Innovation Centre (KIC) for fostering an environment of technological innovation and supporting the development of *PathFinder*. Special thanks to Prof. Fauziah Ahmad for her invaluable guidance, and to my family for their unwavering support throughout this journey. Their collective encouragement and wisdom were instrumental in the realization of this project.

REFERENCES

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv preprint*, arXiv:2402.03216. https://arxiv.org/abs/2402.03216

EdX, 2024. *EdX Learning Platform*. Available at: https://www.edx.org/ [Accessed 10 November 2024]

ESCO, 2024. *European Skills, Competences, Qualifications and Occupations*. Available at: https://esco.ec.europa.eu/en [Accessed 10 November 2024]

European Commission. (2024). Learner Centric Advanced Manufacturing Platform for Centres of Vocational Excellence (LCAMP). https://lcamp.eu/

IADIS, 2023. *Proceedings of the 16th International Conference on Information Systems*. IADIS Press, Lisbon, Portugal.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Li, L. (2022). Reskilling and upskilling the future-ready workforce for industry 4.0 and beyond. Information Systems Frontiers, 1-16.

Rygl, J., Pomikálek, J., Řehůřek, R., Růžička, M., Novotný, V., & Sojka, P. (2017). Semantic vector encoding and similarity search using fulltext search engines. arXiv preprint arXiv:1706.00957.

Salemi, A., & Zamani, H. (2024, July). Evaluating retrieval quality in retrieval-augmented generation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2395-2400).

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. arXiv preprint arXiv:2404.07220.

Tao, C., Shen, T., Gao, S., Zhang, J., Li, Z., Tao, Z., & Ma, S. (2024). LLMs are Also Effective Embedding Models: An In-depth Overview. arXiv preprint arXiv:2412.12591.