UNIVERSIDAD EAFIT

# INTELIGENCIA DEL MERCADO LABORAL COLOMBIANO: DETECCIÓN AUTOMATIZADA DE HABILIDADES MEDIANTE MODELOS GRANDES DE LENGUAJE (LLM) Y RECUPERACIÓN AUMENTADA (RAG)

Colombian Labor Market Intelligence: automated skills detection using Large Language Models (LLM) and Retrieved Augmented Generation (RAG)

JORGE MARIO ZAPATA POSADA

Tesis

Asesores

Claudia Patricia Alvarez Barrera

Jorge Ivan Padilla Buritica

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2025

UNIVERSIDAD EAFIT

## CONTENIDO

UNIVERSIDAD EAFIT

# LISTA DE FIGURAS

UNIVERSIDAD EAFIT

## RESUMEN

La demanda de habilidades en el mercado laboral ha evolucionado significativamente en las últimas décadas, impulsada por cambios en el entorno económico y los constantes avances tecnológicos. En este contexto, la descripción detallada de cada oferta laboral, disponible en portales web de empleo, proporciona información precisa sobre las habilidades específicas que requiere el mercado en tiempo real. La investigación en inteligencia del mercado laboral *(Labour Market Intelligence, LMI)* utiliza estos datos en conjunto con algoritmos de aprendizaje automático para anticipar tendencias y comprender la evolución de la demanda de talento.

A pesar de los avances en inteligencia artificial y la disponibilidad de grandes volúmenes de datos, sigue existiendo una brecha en la adaptación de estas tecnologías al contexto local. Los mercados regionales, como el colombiano, requieren enfoques personalizados para garantizar que las soluciones tecnológicas respondan a las necesidades específicas del mercado laboral, alineando de manera efectiva la oferta y la demanda de talento.

Este estudio se enfoca en analizar los datos del portal web de empleo *Talent.com* para Colombia, mediante un enfoque de última generación basado en Modelos Grandes de Lenguaje *(LLM)* combinados con Recuperación Aumentada *(RAG)*, para identificar habilidades emergentes, tradicionales, técnicas y blandas. En una primera etapa, un *LLM* multilingüe extrae menciones de habilidades a partir de descripciones de ofertas de trabajo. En una segunda etapa, un módulo de recuperación semántica consulta la taxonomía abierta de clasificación de habilidades de la Comisión Europea *(ESCO)* para proponer candidatos normalizados, el *LLM* selecciona la etiqueta más adecuada y devuelve resultados estructurados en formato *JSON* validado.

Los resultados preliminares apuntan a mejoras en precisión, cobertura y auditabilidad frente a enfoques puramente supervisados, al reducir alucinaciones mediante la selección sobre candidatos recuperados y al estandarizar las categorías mediante la clasificación de habilidades *ESCO*.

Este marco proporciona información valiosa que en futuros trabajos permita a las instituciones universitarias desarrollar programas académicos alineados con las necesidades del mercado, facilitando así la toma de decisiones estratégicas por parte de empleadores, formuladores de políticas y educadores, contribuyendo al desarrollo del talento y a la reducción del desempleo en Colombia.

**Palabras clave:** *Inteligencia del mercado laboral (LMI), Procesamiento de Lenguaje Natural (NLP), Talent.com, Mercado laboral colombiano, Modelos Grandes de Lenguaje (LLM), Recuperación Aumentada (RAG)*

UNIVERSIDAD EAFIT


## ABSTRACT

The demand for skills in the labor market has evolved significantly in recent decades, driven by changes in the economic environment and constant technological advances. In this context, the detailed description of each job offer, available on employment web portals, provides accurate information on the specific skills required by the market in real time. Labor Market Intelligence (LMI) research uses this data along with machine learning algorithms to anticipate trends and understand the evolution of talent demand.

Despite advances in artificial intelligence and the availability of large data volumes, there remains a gap in adapting these technologies to local contexts. Regional markets, such as Colombia, require customized approaches to ensure that technological solutions respond to the specific needs of the labor market, effectively aligning talent supply and demand.

This study analyzes data from the Talent.com employment platform for Colombia using a state-of-the-art approach based on Large Language Models (LLM) combined with Retrieval Augmented Generation (RAG) to identify emerging, traditional, technical, and soft skills. In the first stage, a multilingual LLM extracts skill mentions from job descriptions. In the second stage, a semantic retrieval module queries the European Commission's open ESCO skills taxonomy to propose standardized candidate labels, the LLM then selects the most appropriate label and delivers validated, structured JSON outputs.

Preliminary results show improvements in precision, coverage, and auditability compared to purely supervised approaches, reducing hallucinations through candidate-constrained selection and standardizing categories using ESCO skill classification.

This framework provides valuable insights that, in future work, may support universities in designing academic programs aligned with labor market needs, thus facilitating strategic decision-making for employers, policymakers, and educators, and contributing to talent development and the reduction of unemployment in Colombia.

**Keywords:** *Labour Market Intelligence (LMI), Natural Language Processing (NLP), Talent.com, Colombian labor market, Large Language Models (LLM), Retrieval Augmented Generation (RAG)*

UNIVERSIDAD EAFIT

# INTRODUCCIÓN

El mercado laboral se encuentra en constante transformación debido a múltiples factores, entre los que destacan los avances tecnológicos, la digitalización y los cambios en las dinámicas económicas a nivel global. Estas transformaciones han generado una evolución en las habilidades demandadas por las empresas, creando desafíos tanto para los trabajadores como para los empleadores. La diferencia entre las habilidades disponibles y las requeridas en el mercado, conocida como brecha de habilidades, es una preocupación creciente que impacta la productividad, la empleabilidad y el desarrollo económico. En este contexto, la Inteligencia del Mercado Laboral *(Labour Market Intelligence, LMI)* ha surgido como una disciplina clave para analizar en tiempo real los cambios en las demandas laborales.

Las ofertas de empleo publicadas en plataformas digitales representan una fuente valiosa para este análisis, al reflejar de manera precisa las habilidades específicas requeridas en cada sector y región. Estas fuentes proporcionan una alternativa más dinámica y accesible que las encuestas tradicionales, permitiendo una mejor comprensión de las tendencias emergentes y las habilidades obsoletas que requieren ser reemplazadas. Sin embargo, el manejo eficiente de este volumen de datos requiere la implementación de algoritmos avanzados de Procesamiento de Lenguaje Natural *(NLP)* y técnicas de aprendizaje automático.

En este trabajo se propone un pipeline moderno de extracción y normalización de habilidades basado en Modelos Grandes de Lenguaje *(LLM)* + Recuperación Aumentada *(RAG)*. El *LLM* se emplea para identificar menciones de habilidades en descripciones de empleo. Posteriormente, un recuperador semántico consulta la taxonomía abierta de clasificación de habilidades de la Comisión Europea *(ESCO)* para ofrecer candidatos normalizados y el modelo elige la mejor correspondencia, devolviendo el resultado estructurado en formato *JSON* listo para análisis. Este

diseño reduce la dependencia de grandes corpus anotados, disminuye errores por sinónimos o ambigüedad y mejora la trazabilidad.

El objetivo principal es demostrar que la combinación *LLM + RAG* permite identificar y normalizar de manera confiable las habilidades demandadas en el mercado laboral colombiano a partir de datos digitales, habilitando indicadores comparables por sector, región y familia ocupacional. Este análisis permitirá categorizar las competencias en emergentes, tradicionales, técnicas y blandas, proporcionando información clave para alinear los programas educativos con las demandas del mercado. Aunque el diseño de recomendaciones específicas para instituciones educativas y empleadores será abordado en trabajos futuros, este proyecto sienta una base sólida para reducir la brecha de talento y mejorar la comprensión de las dinámicas del mercado laboral colombiano.

UNIVERSIDAD EAFIT

## PLANTEAMIENTO DEL PROBLEMA

La brecha de habilidades en el mercado laboral representa un desafío para el bienestar social y económico. Según Abadía et al. (2023) “tener altos desajustes de habilidades genera ineficiencias en el mercado laboral y frena el progreso tanto social como económico del país” (p. 7). En Colombia, esta brecha es especialmente preocupante debido a la constante transformación del mercado laboral, que genera un desajuste entre las competencias que poseen los trabajadores y las que demandan las empresas.

Estudios recientes han destacado la gravedad de este problema en la región. Según Díaz & Salas (2020), una encuesta de *Manpower Group* revela que al menos la mitad de los empleadores formales en América Latina “no encuentran trabajadores con las habilidades requeridas para los cargos, especialmente en sectores de ciencia y tecnología” (p. 1). En el caso colombiano, un informe de la Corporación Andina de Fomento (CAF) señala que los jóvenes presentan los niveles más altos de desajuste en las habilidades específicas requeridas por las empresas (Díaz & Salas, 2020, p. 3). Este desajuste limita la empleabilidad, reduce las oportunidades de los trabajadores y afecta directamente la competitividad del país.

A pesar de que plataformas de empleo como *Talent.com* ofrecen datos en tiempo real sobre las habilidades demandadas, esta valiosa fuente de información no ha sido aprovechada de manera óptima. La ausencia de herramientas avanzadas que procesen y analicen estos datos dificulta la rápida adaptación de las instituciones educativas y los empleadores a las tendencias del mercado laboral. Esto resalta la necesidad de un análisis detallado y eficiente que permita comprender mejor las habilidades en demanda.

En este contexto, esta investigación se centra en la identificación y clasificación de habilidades demandadas en el mercado laboral colombiano, agrupándolas en

categorías como emergentes, tradicionales, técnicas y blandas. Este análisis inicial permitirá establecer una base para alinear programas educativos con las necesidades del mercado, aunque el diseño de recomendaciones específicas será abordado en trabajos futuros. La clasificación precisa de habilidades es un primer paso esencial para reducir el desajuste entre oferta y demanda laboral, promoviendo el desarrollo del talento y el crecimiento económico del país.

UNIVERSIDAD EAFIT

## JUSTIFICACIÓN

Esta investigación busca optimizar el uso de datos digitales provenientes de plataformas de empleo mediante la aplicación de herramientas avanzadas de inteligencia artificial, como los Modelos Grandes de Lenguaje *(LLM)* combinados con Recuperación Aumentada *(RAG).* Este enfoque moderno permitirá extraer información precisa sobre las diferentes habilidades demandadas, facilitando su clasificación en categorías como emergentes, tradicionales, técnicas y blandas. Este enfoque es crucial para reducir la brecha de habilidades, ya que proporciona un análisis detallado de las competencias requeridas en el mercado laboral y su evolución.

Desde una perspectiva científica, esta investigación contribuye al campo de la Inteligencia del Mercado Laboral *(LMI)* al demostrar cómo el análisis de grandes volúmenes de datos textuales, mediante inteligencia artificial, puede proporcionar soluciones prácticas a problemas relacionados con la clasificación de competencias. Además, ofrece un modelo replicable para otros contextos que busquen entender mejor las tendencias de habilidades y su relevancia en sectores estratégicos.

El aporte de este estudio es especialmente relevante en el contexto colombiano, donde los desajustes entre oferta y demanda laboral afectan la empleabilidad y limitan el crecimiento económico. La identificación y clasificación de habilidades no solo proporcionará insumos fundamentales para alinear los programas educativos con las necesidades del mercado, sino que también sentará las bases para futuros estudios enfocados en el diseño de políticas públicas y estrategias educativas. Este análisis inicial es esencial para fomentar la sincronización entre la oferta educativa y la demanda laboral, promoviendo un desarrollo económico sostenible y una mayor competitividad en sectores clave.

Después de esta introducción, se presentará el objetivo general junto con los objetivos específicos de este trabajo. Posteriormente, se describirá el marco conceptual que sustenta el estudio, seguido por la metodología diseñada para alcanzar los objetivos planteados. Por último, se presentarán los resultados del estudio, desde el análisis exploratorio de datos hasta la extracción, normalización y clasificación de habilidades, seguidos de la discusión de los principales hallazgos, las conclusiones y las posibles líneas de trabajo futuro.

UNIVERSIDAD EAFIT

# OBJETIVOS

## GENERAL

Identificar las habilidades emergentes, tradicionales, técnicas y blandas demandadas en el mercado laboral colombiano mediante el análisis automatizado de ofertas de empleo recopiladas de *Talent.com*, utilizando técnicas avanzadas de procesamiento de lenguaje natural basadas en Modelos Grandes de Lenguaje *(LLM)* y Recuperación Aumentada *(RAG).*

## ESPECÍFICOS

1. Recolectar y preprocesar datos relevantes de ofertas de empleo publicadas en *Talent.com* el en período comprendido entre Junio - Noviembre de 2025 para obtener información precisa sobre las habilidades demandadas en diferentes sectores y regiones.
2. Aplicar algoritmos de procesamiento de lenguaje natural, como Modelos Grandes de Lenguaje *(LLM)* en combinación con Recuperación Aumentada *(RAG)*, para analizar las descripciones textuales de las ofertas de empleo, identificando patrones y tendencias.
3. Evaluar y clasificar las habilidades identificadas en categorías relevantes (emergentes, tradicionales, técnicas, blandas) para diferentes sectores y regiones.

UNIVERSIDAD EAFIT

# MARCO CONCEPTUAL

## 1. Introducción al Mercado Laboral y la Brecha de Habilidades

El mercado laboral puede definirse como “el mecanismo que facilita la satisfacción de una demanda de servicios laborales por parte de quienes desean o pueden suministrar servicios laborales” (World Health Organization, 2021, p. 10). Este mercado está en constante transformación debido a la digitalización, la automatización de procesos productivos y los avances tecnológicos, los cuales se desarrollan a un ritmo cada vez más acelerado. Dichas transformaciones han modificado significativamente las habilidades requeridas por los empleadores, planteando interrogantes sobre qué ocupaciones crecerán en el futuro y en qué ubicaciones específicas, así como cuáles serán las habilidades más demandadas en los próximos años (Colombo et al., 2019, p. 27).

La brecha de habilidades, que se refiere al desajuste entre las habilidades que poseen los trabajadores y las requeridas por los empleadores, representa un obstáculo significativo para el crecimiento y desarrollo económico (Rahhal et al., 2024, p. 18). A nivel empresarial, esta brecha reduce la innovación y competitividad, ya que las empresas tienen dificultades para determinar las habilidades necesarias en determinados puestos (Papoutsoglou et al., 2022, p. 1). En el caso de los trabajadores, la falta de competencias relevantes reduce sus oportunidades laborales y aumenta el riesgo de desempleo o subempleo; afectando tanto a los poco cualificados, que tienden a bajar la productividad y calidad, como a los sobrecalificados, quienes pueden desmotivarse y perder habilidades o buscar mejores oportunidades fuera del mercado laboral local (World Health Organization, 2021, p. 207).

Estudios recientes sugieren que esta brecha afecta especialmente a los mercados emergentes, como podría ser el caso del mercado colombiano, donde los cambios

tecnológicos requieren una adaptación más rápida de los programas educativos (Díaz & Salas, 2020, p. 3). En este contexto, la Inteligencia del Mercado Laboral *(Labour Market Intelligence, LMI)* se ha convertido en una disciplina clave para monitorear las demandas laborales y proporcionar datos que orienten las políticas públicas y las decisiones de los empleadores.

## 2. Inteligencia del Mercado Laboral *(LMI):* Concepto y Aplicaciones

La Inteligencia del Mercado Laboral *(Labor Market Intelligence, LMI)* es un campo interdisciplinario que combina la ciencia de datos, el análisis económico laboral y las tecnologías de la información para examinar en tiempo real las tendencias, habilidades y demandas del mercado (Mezzanzanica & Mercorio, 2019, p. 8). Surgió como respuesta a las limitaciones de los métodos tradicionales de recopilación de datos, como encuestas y censos, los cuales no lograban capturar con precisión las fluctuaciones dinámicas del mercado laboral (Boselli et al., 2018, p. 320). En este contexto, *LMI* aprovecha fuentes digitales masivas, como las ofertas de empleo publicadas en plataformas web como *LinkedIn*, *Xing*, *Talent.com*, entre otras, que proporcionan información actualizada y granular sobre las competencias requeridas en cada sector y región (Papoutsoglou et al., 2019, p. 157599), con el objetivo de orientar tanto a las universidades como a los empleadores en la formación y reclutamiento de talento especializado.

El desarrollo de *LMI* ha sido impulsado por el crecimiento exponencial del uso de tecnologías digitales y la expansión de la automatización en los procesos productivos (Colombo et al., 2019, p. 28). La posibilidad de acceder a datos en tiempo real ha transformado la manera en que se diseñan políticas educativas y estrategias empresariales, permitiendo una mayor alineación entre la oferta de competencias y las demandas del mercado laboral (Rahhal et al., 2024, p. 18).

El análisis basado en *LMI* es esencial para identificar habilidades emergentes, en riesgo de obsolescencia o subvaloradas, contribuyendo a una adaptación ágil de los programas formativos y las estrategias de reclutamiento (Khaouja et al., 2021, p. 118143). Por ejemplo, en la industria de vehículos eléctricos, donde las competencias tecnológicas evolucionan rápidamente, la integración de datos digitales ha demostrado ser crucial para ajustar los programas educativos relacionados con ingeniería y tecnología de acuerdo con las demandas emergentes del sector (Papoutsoglou et al., 2022, p. 10). Un segundo ejemplo de la aplicación de *LMI* se encuentra en la industria de tecnologías de la información. En este sector, donde las competencias en programación, ciberseguridad y análisis de datos están en constante cambio, los empleadores enfrentan el desafío de cubrir vacantes con habilidades específicas que cambian rápidamente (Rahhal et al., 2019, p. 535).

En resumen, *LMI* ha evolucionado como una herramienta clave para el desarrollo de políticas públicas y empresariales, al proporcionar información basada en datos que permite una alineación efectiva entre las competencias de los trabajadores y las necesidades del mercado. La capacidad de detectar en tiempo real los cambios en las demandas laborales no solo permitiría optimizar los programas educativos, sino también mejorar la empleabilidad y la competitividad en sectores estratégicos de la economía (Boselli et al., 2018, p. 320).

## 3. Procesamiento del Lenguaje Natural (NLP) y Aprendizaje Automático

El manejo eficiente del volumen de datos textuales disponibles en plataformas digitales, como las ofertas de empleo en *Talent.com*, requiere el uso de algoritmos avanzados de procesamiento de lenguaje natural *(Natural Language Processing, NLP)* y técnicas de aprendizaje automático. Estas herramientas permiten analizar datos no estructurados y extraer información significativa sobre las competencias demandadas en tiempo real (Khaouja et al., 2021, p. 118136).

En los últimos años la literatura ha evolucionado desde enfoques basados en modelos estáticos y algoritmos clásicos, hacia arquitecturas de gran escala impulsadas por los Modelos Grandes de Lenguaje *(Large Language Models, LLM)*, los cuales muestran un rendimiento superior en extracción de información, razonamiento semántico y alineación con taxonomías de clasificación laboral formales como *SOC (Standard Occupational Classification) y ESCO (European Skills, Competences, Qualifications and Occupations)* (Nguyen et al., 2024, p. 27).

A continuación, se presenta una síntesis de estas metodologías, iniciando con los enfoques previos y finalizando con la arquitectura más reciente basada en *LLM + RAG (Retrieval Augmented Generation)*, que constituye la principal aplicación de esta investigación.

### 3.1. Métodos previos de extracción y análisis basados en *Transformers* y modelado temático

Los primeros trabajos en *LMI* se apoyaron en una combinación de métodos de modelado temático, representaciones distribuidas y algoritmos supervisados. Entre las técnicas más utilizadas en estudios se encuentran:

- ***LDA (Latent Dirichlet Allocation):*** Esta técnica de modelado temático no supervisado permite identificar temas en grandes conjuntos de datos textuales, facilitando la clasificación de habilidades en categorías relevantes (Mezzanzanica & Mercorio, 2019, p. 38). *LDA* es útil para detectar patrones subyacentes en las ofertas de empleo, aunque presenta limitaciones al capturar matices semánticos complejos y al interpretar tópicos poco coherentes en comparación con modelos más recientes (Ao et al., 2023, p. 12).

- ***Word Embeddings* estáticos:** Métodos como *Word2Vec* y *GloVe* representan palabras en espacios vectoriales de baja dimensión, lo que permite capturar relaciones semánticas basadas en la proximidad entre términos. Estos modelos permiten medir similitud entre habilidades, agrupar competencias relacionadas y construir índices de búsqueda semántica (Rahhal et al., 2024, p. 9). Sin embargo, su incapacidad para modelar adecuadamente el contexto (una misma palabra tiene un único vector) limita su desempeño frente a modelos contextualizados basados en *Transformers* (Aleisa et al., 2023, p. 23).
- ***Random Forest:*** En aplicaciones de aprendizaje automático para *LMI*, este algoritmo supervisado ha sido empleado para predecir la demanda futura de habilidades, y clasificar características de las vacantes a partir de atributos estructurados derivados de las ofertas. Si bien ha demostrado ser efectivo en modelos de recomendación y predicción de demanda (Parida et al., 2022, p. 89), requiere conjuntos de datos etiquetados y no operan de forma directa sobre texto sin estructurar.
- **Redes neuronales recurrentes *LSTM (Long Short-Term Memory):*** Este tipo de red neuronal recurrente ha sido utilizada para analizar secuencias temporales, predecir tendencias de demanda de habilidades a lo largo del tiempo y anticipar necesidades futuras (Senthurvelautham & Senanayake, 2023, p. 6). La principal desventaja de este enfoque es que “el modelo está limitado a las soluciones tecnológicas y habilidades incluidas en el conjunto de datos de entrenamiento” (Alharbi & Al-Alawi, 2024, p. 479), lo que reduce su capacidad de generalizar a nuevas competencias emergentes.

Posteriormente, con la adopción de la arquitectura *Transformer*, se popularizaron modelos como *BERT (Bidirectional Encoder Representations from Transformers)* y sus variantes. En el contexto de la extracción y agrupación de habilidades destacan:

- ***RoBERTa (Robustly Optimized BERT Pretraining approach):*** Es un modelo basado en *Transformers* que supera a *BERT* al entrenarse con un corpus más extenso y sin la tarea de predicción de la siguiente oración *(Next Sentence Prediction, NSP)*, lo cual permite una mejor convergencia y mayor rendimiento (Aleisa et al., 2023, p. 23). Su arquitectura genera representaciones contextuales que han demostrado ser superiores a *Word2Vec*, *TF-IDF* y *BoW* en tareas como clasificación, medición de similitud y clusterización de textos (Aleisa et al., 2023, p. 28).
- ***BERTopic:*** Modelo que combina representaciones contextuales derivadas de *BERT* con técnicas de reducción de dimensionalidad y agrupamiento para obtener tópicos interpretables. Ha demostrado generar tópicos más coherentes y mejor diferenciados que *LDA* y *PLSA*, lo que facilita la identificación de grupos de habilidades emergentes en grandes corpus de ofertas laborales (Ao et al., 2023, p. 12).

En conjunto, estos enfoques han permitido avances importantes en la extracción y análisis de habilidades, especialmente en la detección de tendencias y la agrupación temática. Sin embargo, presentan limitaciones como el manejo de menciones complejas o implícitas de habilidades, y la alineación de habilidades extraídas con taxonomías estandarizadas de forma precisa y auditable (Nguyen et al., 2024, p. 27).

### 3.2. Modelos Grandes de Lenguaje *(LLM)* para extracción de habilidades

Los Modelos Grandes de Lenguaje *(Large Language Models, LLM)* como las familias LLaMA o Gemma, han introducido un cambio de paradigma en tareas de extracción de información. En lugar de entrenar modelos específicos para cada dominio, los *LLMs* pueden resolver nuevas tareas mediante instrucciones en lenguaje natural y ejemplos en contexto (*in-context learning*), incluso en escenarios de *zero-shot* o *few-shot* (Clavié & Soulié, 2023, p. 2).

En el dominio del mercado laboral, estudios recientes han analizado el uso de *LLMs* para la extracción de habilidades desde ofertas de empleo. Por ejemplo, se ha mostrado que, aunque los *LLMs* no siempre superan a los modelos supervisados entrenados con grandes corpus anotados, sí son especialmente competentes en la detección de menciones compuestas y habilidades que aparecen de forma implícita en las responsabilidades, reduciendo la necesidad de esquemas de etiquetado complejos y la dependencia de anotaciones manuales extensivas (Nguyen et al., 2024, p. 27). Este replanteamiento de la tarea de *skill extraction* sugiere que los *LLMs* pueden funcionar como extractores generales, siempre que se combinen con mecanismos adecuados de control, normalización y evaluación.

Otra línea de trabajo propone aprovechar los *LLMs* para generar datos sintéticos de alta calidad. En lugar de anotar manualmente miles de ejemplos, se diseña un flujo en dos etapas: Primero el *LLM* genera descripciones de puestos y listas de habilidades siguiendo una taxonomía normalizada como *ESCO.* A continuación, un modelo más ligero aprende de este conjunto sintético para realizar extracción y mapeo a gran escala (Magron et al., 2024, p. 2). De esta manera se generan ofertas realistas combinando múltiples habilidades, las cuales se vinculan a la taxonomía normalizada, obteniendo mejoras frente a modelos supervisados.

Sin embargo, tanto los enfoques basados en ejemplos en contexto, como los esquemas apoyados en datos sintéticos, siguen dependiendo del conocimiento

interno del modelo y de la calidad de los ejemplos utilizados. Cuando el objetivo es alinear las habilidades detectadas con taxonomías normalizadas, se hace necesario complementar la capacidad generativa de los *LLM* con mecanismos explícitos de acceso y anclaje a fuentes externas de conocimiento (Zhao et al., 2024, p. 3).

### 3.3. Recuperación Aumentada *(RAG)*

En este contexto, una de las principales innovaciones recientes es el uso de Recuperación Aumentada *(Retrieval Augmented Generation, RAG)*, donde el *LLM* no responde únicamente a partir de sus pesos, sino que también consulta una base externa de conocimiento. En un sistema *RAG* típico, la mención de una habilidad extraída (por ejemplo, “atención a clientes”) se transforma en un vector mediante un modelo de *embeddings,* y se utiliza para recuperar las habilidades más similares de la taxonomía normalizada. El *LLM* recibe entonces tanto el texto original de la oferta como los candidatos recuperados, y se le instruye para que seleccione la opción más adecuada (Clavié & Soulié, 2023, p. 3).

La literatura reciente sobre RAG muestra estudios que analizan diferentes decisiones de diseño (número de documentos recuperados, estrategias de *re-ranking*, elección del modelo de *embeddings)* y muestran que no existe una configuración universalmente óptima, sino buenas prácticas que deben adaptarse a cada dominio concreto (Zhao et al., 2024, p. 1). En este sentido, el enfoque *OG-RAG (Ontology-Grounded RAG)* propone anclar explícitamente la recuperación a una ontología formalizada, de forma que el contexto que se pasa al *LLM* no sea un conjunto arbitrario de fragmentos de texto, sino subconjuntos consistentes de una taxonomía normalizada (por ejemplo, grupos de habilidades relacionadas dentro de *ESCO*). Esto facilita la trazabilidad, mejora la precisión y reduce las alucinaciones, al limitar el espacio de posibles salidas del modelo a la ontología de referencia (Sharma et al., 2025, p. 32952).

En el ámbito aplicado, herramientas como *ESCOX (ESCO Skill Extractor)* implementan precisamente este tipo de enfoque, al combinar *LLMs* con modelos de *embeddings* para detectar y normalizar tanto habilidades como ocupaciones a partir de texto libre, alineando sistemáticamente las salidas con *ESCO*, y proporcionando una interfaz *(API)* que permite explorar estos resultados en análisis de brechas, vigilancia de tendencias y apoyo a la formulación de políticas (Kavargyris et al., 2025, p. 3).

### 3.4. Síntesis y justificación del enfoque propuesto

En síntesis, los algoritmos de NLP y aprendizaje automático clásicos *(LDA, word embeddings, Random Forest, LSTM, RoBERTa, BERTopic)* han demostrado ser útiles para explorar grandes corpus de ofertas laborales, identificar temas, agrupar habilidades y construir modelos predictivos a partir de datos etiquetados. No obstante, los avances recientes en *LLMs* y *RAG* permiten abordar de forma más flexible y precisa el problema de extracción y normalización de habilidades desde ofertas de trabajo. La implementación de estas herramientas contribuye no solo a una mejor comprensión de las tendencias del mercado, sino también a una alineación más efectiva entre la oferta educativa y la demanda laboral en sectores clave de la economía (Rahhal et al., 2024).

La integración de datos provenientes de plataformas de empleo como *Talent.com,* combinada con estas técnicas de inteligencia artificial, facilita la identificación y clasificación de habilidades en categorías como emergentes, tradicionales, técnicas y blandas. Este enfoque proporciona una base sólida para alinear los programas educativos con las demandas del mercado, sentando las bases para futuras recomendaciones dirigidas a empleadores e instituciones educativas.

UNIVERSIDAD EAFIT

## DISEÑO METODOLÓGICO

El diseño metodológico de esta investigación se centra en el análisis de datos textuales obtenidos de 44,077 ofertas de empleo publicadas en la plataforma *Talent.com* para Colombia en el período comprendido entre junio y noviembre de 2025. Este enfoque busca identificar, analizar y clasificar las habilidades demandadas en diferentes sectores y regiones del mercado laboral colombiano, en concordancia con los objetivos específicos planteados.

### 1. Recolección y preprocesamiento de datos

La primera etapa del estudio consiste en la recolección de datos relevantes provenientes de ofertas de empleo disponibles en *Talent.com* para el dominio colombiano. Los datos se obtienen a partir de un extracto en formato *JSON* procedente del índice interno de la compañía, el cual se transforma a un formato tabular *(CSV)* para su posterior procesamiento en *Python*. Las variables principales incluidas se detallan en la Tabla 1.

*Tabla 1. Variables relevantes obtenidas de Talent.com*

| **Variable** | **Descripción** |
|---|---|
| *source_title* | Título de la oferta de trabajo |
| *job_description* | Descripción de la oferta de trabajo |
| *enrich_geo_region1* | Departamento |
| *enrich_geo_region2* | Ciudad |
| *enrich_soc_major_group* | Clasificación ocupacional estándar: Grupo mayor |
| *enrich_soc_minor_group* | Clasificación ocupacional estándar: Grupo menor |
| *enrich_soc_broad_group* | Clasificación ocupacional estándar: Ocupación amplia |
| *enrich_soc_detailed_group* | Clasificación ocupacional estándar: Ocupación detallada |

A partir de estas variables, se llevará a cabo un proceso de preprocesamiento de datos que garantiza consistencia y calidad para el análisis posterior.

### 1.1. Normalización y limpieza de texto

La columna principal para el análisis es *job_description*, la cual puede contener *HTML*, etiquetas especiales y texto en distintos idiomas. El preprocesamiento incluye:

- Conversión del *HTML* a texto plano mediante el uso de la librería *BeautifulSoup*, eliminando etiquetas <*script*>, <*style*> y otros elementos no relevantes.
- Normalización de espacios y eliminación de caracteres de control, manteniendo la estructura básica de frases.
- Unificación de codificación en *UTF-8* para preservar caracteres propios del español.

A diferencia de enfoques basados en modelos clásicos, no se realiza tokenización explícita ni lematización para alimentar el modelo, ya que los *LLM* modernos son capaces de trabajar directamente con texto libre en lenguaje natural.

En el caso de las columnas geográficas *(enrich_geo_region1, enrich_geo_region2),* se aplica estandarización de nombres y se contrasta el listado oficial de departamentos y municipios de Colombia, disponible en www.datos.gov.co.

### 1.2. Manejo de valores faltantes

Los registros con descripciones de empleo ausentes (*job_description* vacío o nulo) se excluyen del análisis desde el comienzo, dado que no aportan información para la extracción de habilidades.

En el caso de columnas auxiliares (departamento, ciudad, y variables *SOC*), se aplica imputación simple cuando es posible reconstruir el valor a partir de otros campos o metadatos.

### 1.3. Exploración de datos preprocesados

Sobre el conjunto de datos depurado se realiza un análisis exploratorio de datos *(EDA)* que incluye:

- Distribución de registros por departamento *(enrich_geo_region1)* y ciudad *(enrich_geo_region2).*
- Distribución de registros por categorías de clasificación ocupacional estándar *(SOC):*
    - *enrich_soc_major_group*
    - *enrich_soc_minor_group*
    - *enrich_soc_broad_group*
    - *enrich_soc_detailed_group.*
- Frecuencia de títulos de ofertas de empleo *(source_title)* para identificar roles dominantes.

Esta etapa permite detectar tendencias geográficas y ocupacionales, así como validar qué títulos de oferta son más comunes.

## 2. Extracción y normalización de habilidades mediante *LLM + RAG*

La segunda etapa de la investigación consiste en aplicar un pipeline moderno basado en Modelos Grandes de Lenguaje *(LLM)* y Recuperación Aumentada *(RAG)* para extraer y normalizar habilidades a partir de las descripciones textuales de las ofertas de empleo. El objetivo de este módulo es transformar cada descripción de trabajo en un conjunto de habilidades explícitas, escritas en español, y posteriormente mapear dichas habilidades a etiquetas oficiales de la taxonomía *ESCO (European Skills, Competences, Qualifications and Occupations)* de la Comisión Europea.

### 2.1. Extracción de habilidades con *LLM*

Para la extracción inicial se emplea un modelo grande de lenguaje de código abierto de la familia *Gemma (Google),* en su variante *instruct-tuned*, entrenada para seguir instrucciones y devolver salidas estructuradas. El modelo se invoca mediante la librería *Transformers* de *Hugging Face* en un entorno con aceleración por *GPU*.

La interacción con el modelo se organiza en forma de diálogo entre el sistema y el usuario. En el mensaje de sistema se define al modelo como un “extractor de habilidades” y se le indica que debe responder exclusivamente con un *JSON* que contenga una lista de cadenas en español, sin duplicados. El mensaje de usuario especifica qué debe incluir, qué debe excluir y cómo comportarse si la descripción está en inglés u otro idioma.

A partir de este *prompt*, para cada descripción de empleo se construye la secuencia de entrada al modelo, se genera la respuesta y se extrae el bloque *JSON* que contiene la lista de habilidades en español. Para reducir el ruido en el conjunto de datos, se implementa además un mecanismo de caché que reutiliza las habilidades ya extraídas en caso de descripciones duplicadas.

### 2.2. Construcción de la ontología de habilidades a partir de *ESCO*

La normalización de habilidades se realiza utilizando la taxonomía abierta *ESCO*, a partir del archivo oficial de habilidades en español *(skills_es.csv),* para el cual se seleccionan las columnas relevantes descritas en la Tabla 2.

*Tabla 2. Columnas relevantes taxonomía ESCO*

| **Variable** | **Descripción** |
|---|---|
| *preferredLabel* | Nombre canónico de la habilidad |
| *altLabels* | Sinónimos o formas alternativas de la misma habilidad |
| *description* | Definición descriptiva de la habilidad |

Con esta información se construye una tabla de alias en la que, para cada habilidad *ESCO*, se generan múltiples variantes textuales a partir de *preferredLabel* y *altLabels*. Todas las variantes se normalizan (minúsculas, eliminación de tildes y caracteres especiales) con el fin de unificar expresiones equivalentes (por ejemplo, “Atención al cliente” y “Servicio al cliente”), y permitir coincidencias exactas entre las habilidades extraídas y los alias normalizados.

Esta tabla de alias constituye la base ontológica sobre la cual se llevará a cabo la recuperación semántica y el posterior anclaje de las habilidades detectadas a etiquetas *ESCO* oficiales.

### 2.3. Indexación semántica con *EmbeddingGemma* y *FAISS*

Para poder recuperar de forma eficiente, y por similitud semántica, los candidatos *ESCO* más cercanos a una habilidad extraída, se construye un índice vectorial utilizando un modelo de *embeddings* de la familia Gemma: *embeddinggemma-300m* (también de código abierto).

Primero, se define una función de *embedding* que:

- Recibe una lista de cadenas (por ejemplo, cada alias *ESCO* concatenado con su descripción).
- Tokeniza el texto y lo procesa con el modelo *embeddinggemma-300m.*
- Obtiene la última capa oculta y aplica *mean pooling* enmascarado (promedio solo sobre los tokens válidos).
- Normaliza los vectores resultantes en norma *L2*, de forma que el producto punto corresponda a una similitud coseno.

Los vectores generados para todos los alias se almacenan en una matriz densa y se indexan con *FAISS,* lo que permite realizar consultas de k-vecinos más cercanos de manera eficiente. De esta forma, dada una habilidad extraída en texto libre, el sistema puede recuperar los alias *ESCO* más similares y sus etiquetas canónicas (*preferredLabel*).

## 2.4. Estrategia robusta de normalización de habilidades a ESCO a través de RAG

Una vez recuperados los candidatos más similares desde la ontología *ESCO* para cada habilidad superficial extraída por el *LLM*, se implementa una estrategia de normalización robusta basada en un esquema de Recuperación Aumentada *(RAG).* En primer lugar, el modelo de lenguaje recibe la habilidad original junto con una lista numerada de candidatos *ESCO* y se le instruye para escoger únicamente uno de ellos (o declarar que ninguno aplica), restringiendo su salida a identificadores numéricos. De esta forma, el modelo no puede inventar nuevas etiquetas, sino que debe decidir siempre dentro del conjunto propuesto por el recuperador semántico.

Si esta primera decisión no resulta válida o consistente, se recurre a un segundo esquema en el que el modelo devuelve su elección en un pequeño objeto estructurado, manteniendo el mismo conjunto de candidatos. Cuando, aun así, no se obtiene una selección fiable, la decisión final se toma a partir del candidato con mayor similitud en el espacio de *embeddings.* Este procedimiento se aplica a cada habilidad detectada en las descripciones de empleo y genera, para cada oferta, un conjunto de etiquetas *ESCO* canónicas, junto con información detallada sobre cómo se llegó a cada correspondencia. El *dataset* resultante constituye la base para los análisis posteriores de distribución de habilidades técnicas, blandas, emergentes y tradicionales por región, ciudad y familia ocupacional.

## 3. Evaluación y clasificación de habilidades

La última etapa del estudio tiene dos componentes. Primeramente, la evaluación del rendimiento del pipeline *LLM + RAG.* Segundo la clasificación de las habilidades en las cuatro categorías de interés (emergentes, tradicionales, técnicas, blandas).

### 3.1. Métricas de evaluación de la normalización a *ESCO*

Para evaluar el desempeño del pipeline de extracción y normalización de habilidades se construye un conjunto de referencia a partir de una muestra estratificada equivalente al 1% de las ofertas de empleo con extracción de habilidades válida. La estratificación se realiza sobre las columnas de región geográfica *(enrich_geo_region1, enrich_geo_region2),* y de clasificación por grupo ocupacional estándar *(enrich_soc_major_group, enrich_soc_minor_group, enrich_soc_broad_group y enrich_soc_detailed_group),* de modo que la muestra mantenga la diversidad geográfica y ocupacional de la base completa. Para cada registro de esta muestra se revisa manualmente la lista de habilidades normalizadas

por el sistema *(esco_preferred_labels)* y se corrigen aquellas que no corresponan a la habilidad original según la taxonomía *ESCO*, generando así el conjunto de referencia *esco_gold.*

Sobre este conjunto se evalúa el sistema de forma extremo a extremo, comparando directamente la lista de habilidades *ESCO* producida por el pipeline *LLM + RAG* con la lista de habilidades *ESCO* considerada correcta por evaluación humana. Para cada oferta $i$ se definen:

- $P_i$: conjunto de habilidades predichas por el sistema *(esco_preferred_labels)*
- $G_i$: conjunto de habilidades de referencia *(esco_gold)*

A partir de estos conjuntos se calculan, agregando sobre todas las ofertas de la muestra, las siguientes métricas clásicas de clasificación multietiqueta:

- **Precisión micro *(micro-precision)***: proporción de habilidades propuestas por el sistema que son efectivamente correctas

$$\text{Precision} = \frac{\sum \text{i} \mid P_i \cap G_i \mid}{\sum \text{i} \mid P_i \mid}$$

- **Cobertura micro *(micro-recall)***: proporción de las habilidades correctas presentes en el conjunto de referencia que son recuperadas por el sistema

$$\text{Recall} = \frac{\sum \text{i} \mid P_i \cap G_i \mid}{\sum_i \mid G_i \mid}$$

- ***F1* micro**: media armónica entre *precision* y *recall*, que resume el equilibrio entre ambas medidas.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot Recall}{\text{Precision} + Recall}$$

Adicionalmente se calculan métricas a nivel de oferta:

- **Índice de *Jaccard* por oferta**: cuyo promedio resume el grado de solapamiento entre las listas de habilidades predichas y las de referencia.

$$J_i = \frac{| P_i \cap G_i |}{| P_i \cup G_i |}$$

- ***Exact match****:* proporción de ofertas en las que el conjunto predicho coincide exactamente con el conjunto de referencia

$$(P_i = G_i).$$

Estas métricas se reportan tanto de forma global como desagregadas por grandes grupos ocupacionales *(enrich_soc_major_group)* y por región *(enrich_geo_region1),* lo que permite analizar si el desempeño del sistema es homogéneo o si existen segmentos donde la normalización resulta más problemática. Es importante señalar que, por restricciones de tiempo, la evaluación se centra en la normalización a *ESCO* a partir de las habilidades extraídas por el modelo. Es decir, el conjunto *gold* no intenta recuperar todas las habilidades implícitas en la descripción, sino corregir las etiquetas *ESCO* de las habilidades que el sistema decidió extraer. En consecuencia, las métricas obtenidas reflejan la calidad del pipeline de extracción y normalización tal y como está implementado, pero no permiten estimar el *recall* absoluto respecto a todas las habilidades posibles presentes en el texto.

### 3.2. Clasificación de habilidades en categorías: emergentes, tradicionales, técnicas y blandas

En la etapa final del pipeline se clasifican las habilidades normalizadas en cuatro categorías analíticas: habilidades técnicas tradicionales, habilidades técnicas emergentes, habilidades blandas tradicionales y habilidades blandas emergentes. El punto de partida son las habilidades *ESCO* ya asignadas a cada oferta

*(esco_preferred_labels).* A partir de este campo se construye un listado de habilidades únicas, enriquecido con su descripción oficial en *ESCO* y con información básica de frecuencia (número de ofertas en las que aparece cada habilidad).

La dimensión emergente/tradicional busca capturar el grado de novedad relativa de la habilidad en el contexto del mercado laboral actual. Se etiquetan como emergentes aquellas competencias asociadas a tecnologías recientes (inteligencia artificial, ciencia de datos, ciberseguridad, computación en la nube, automatización avanzada), nuevos modelos de trabajo (teletrabajo, plataformas digitales, economía verde) o campos que han ganado importancia en la última década. Se consideran tradicionales las habilidades que han estado presentes de forma estable en ocupaciones consolidadas (tareas administrativas básicas, oficios manuales clásicos, atención al cliente presencial).

La dimensión técnica/blanda se define de la siguiente manera: se consideran habilidades técnicas aquellas vinculadas a conocimientos específicos de una profesión, manejo de herramientas, tecnologías, procedimientos o normativas (por ejemplo, “programar en *Python*”, “realizar conciliaciones contables”, “operar sistemas de posicionamiento global”). Se consideran habilidades blandas las competencias transversales de tipo interpersonal, comunicativo o actitudinal, como “trabajar en equipos”, “gestionar conflictos” o “pensar de forma proactiva”.

La clasificación se realiza mediante un modelo grande de lenguaje de la familia *Gemma (Google),* en su variante *instruct-tuned,* al igual que se hizo en la fase de extracción de habilidades. Para cada habilidad *ESCO* se proporcionarán al modelo: el nombre canónico *(preferredLabel),* su descripción oficial *(description)* y un indicador de frecuencia relativa en la muestra (muy alta, alta, media o baja). A partir de este contexto, el modelo devolverá, en formato *JSON* estructurado, la categoría asignada en las dos dimensiones anteriores.

Finalmente, esta tabla de clasificación se integra con el *dataset* principal mediante una unión por *preferredLabel*. Esto permitirá analizar la distribución de la demanda por tipo de habilidad y por carácter temporal, tanto a nivel agregado como desagregado por región y grupo ocupacional *SOC*.

En conclusión, el diseño metodológico planteado permitirá alcanzar los objetivos específicos del estudio de manera sistemática y estructurada. Al integrar herramientas avanzadas de inteligencia artificial con una base de datos robusta de ofertas de empleo, se espera generar información valiosa sobre las dinámicas del mercado laboral colombiano, estableciendo una base sólida para futuras investigaciones y aplicaciones en el diseño de políticas públicas y estrategias educativas.

UNIVERSIDAD EAFIT

# RESULTADOS

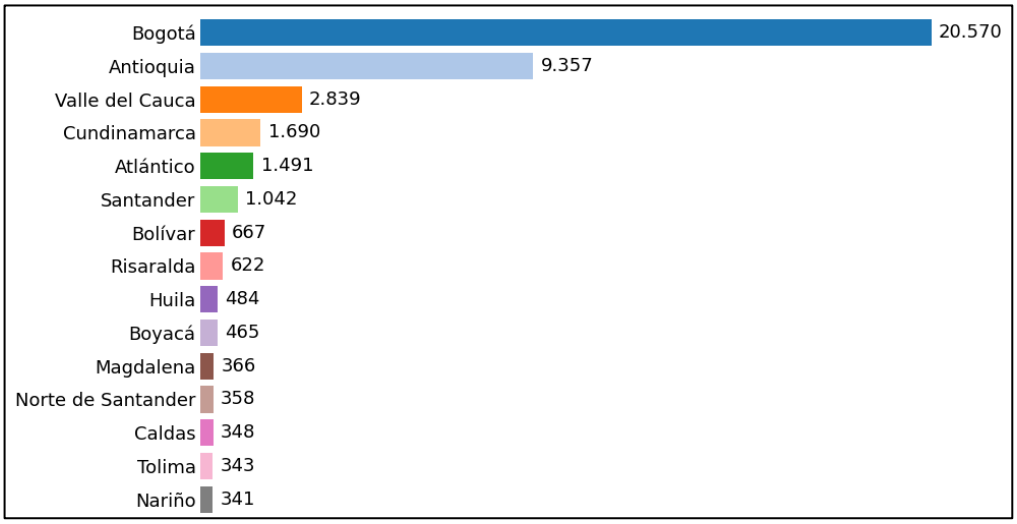
## 1. Análisis Exploratorio de Datos *(EDA)*

Primeramente, se realizó un análisis exploratorio de datos sobre las 44,077 ofertas de empleo recopiladas. Este análisis tuvo como propósito caracterizar la muestra desde el punto de vista geográfico y ocupacional, identificar tendencias de concentración (por departamento, ciudad y grupos ocupacionales *SOC*) y verificar la calidad de las variables disponibles. A partir de esta exploración se establecieron los principales ejes de análisis que guían el resto de la investigación, así como los segmentos prioritarios para el estudio detallado de las habilidades demandadas.

### 1.1. Análisis por Región

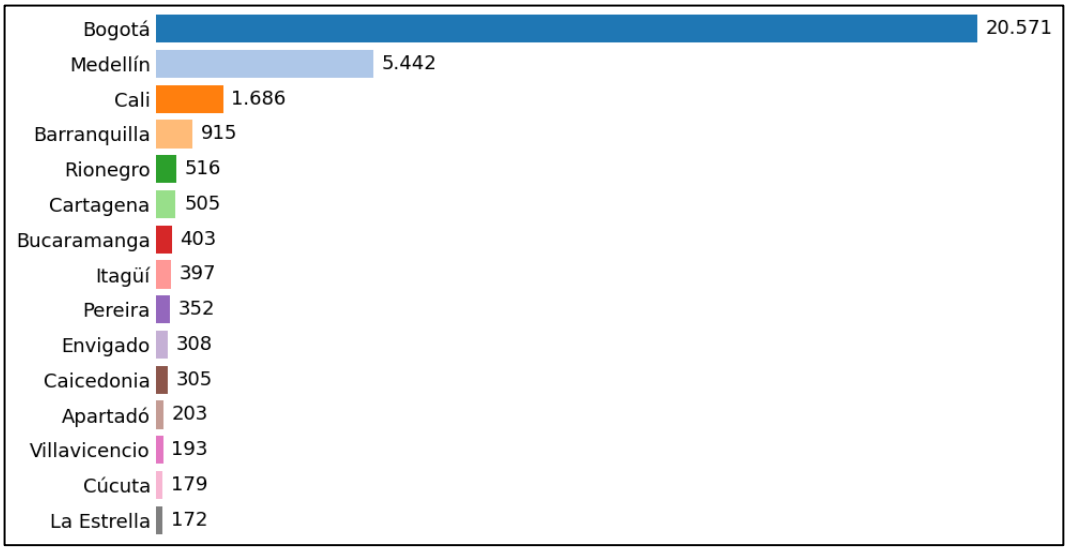
El análisis preliminar de los datos muestra tendencias relevantes de ofertas laborales concentradas en regiones específicas. A nivel departamental (Figura 1), encontramos que el 46.6% del total de vacantes de empleo son requeridas en Bogotá, la capital colombiana. Seguido de Antioquia con el 21.2%, Valle del Cauca con 6.4% y Cundinamarca 3.8%. Estos tres departamentos, junto con el distrito capital, concentran cerca del 80% de las ofertas laborales en la muestra, lo cual evidencia una marcada centralización de las oportunidades de empleo en las principales regiones del país. Este panorama resalta la importancia de analizar las habilidades demandadas en estas regiones estratégicas para entender mejor las dinámicas del mercado laboral colombiano.

*Figura 1. Top 15 ofertas laborales por Departamento*




A nivel de ciudad (Figura 2), se observa una tendencia similar a la departamental, con Bogotá liderando con el 46.6 % de las ofertas laborales, seguida de Medellín con el 12.3%, Cali con el 3.8% y Barranquilla con el 2%. Este patrón confirma la concentración de oportunidades laborales en las principales ciudades del país, que representan los principales centros de actividad económica y laboral.

*Figura 2. Top 15 ofertas laborales por Ciudad*

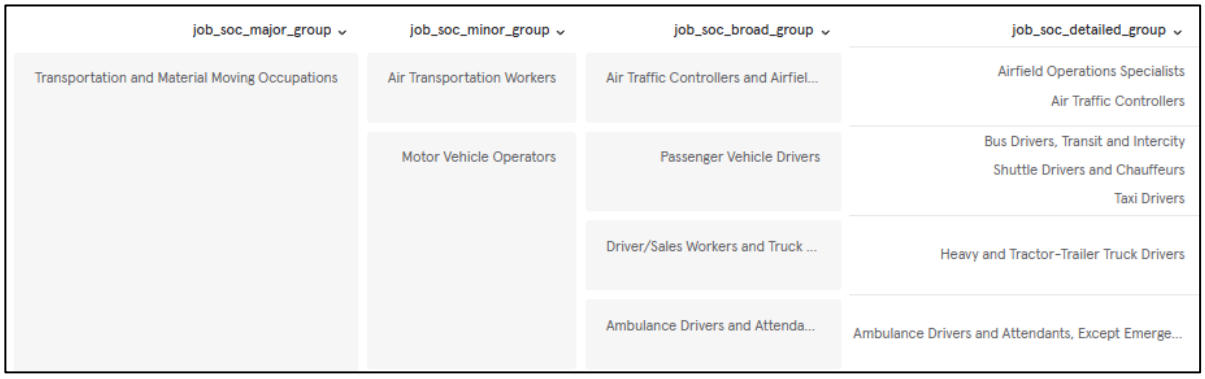


## 1.2. Análisis por Clasificación Ocupacional Estándar *(SOC)*

El Sistema de Clasificación Ocupacional Estándar (Standard Occupational Classification, SOC) es un marco ampliamente utilizado para organizar y categorizar ocupaciones laborales según sus funciones y competencias principales. Está estructurado en varios niveles que permiten clasificar desde amplias áreas profesionales hasta roles específicos (Figura 3). Las divisiones del SOC se estructuran de la siguiente manera:

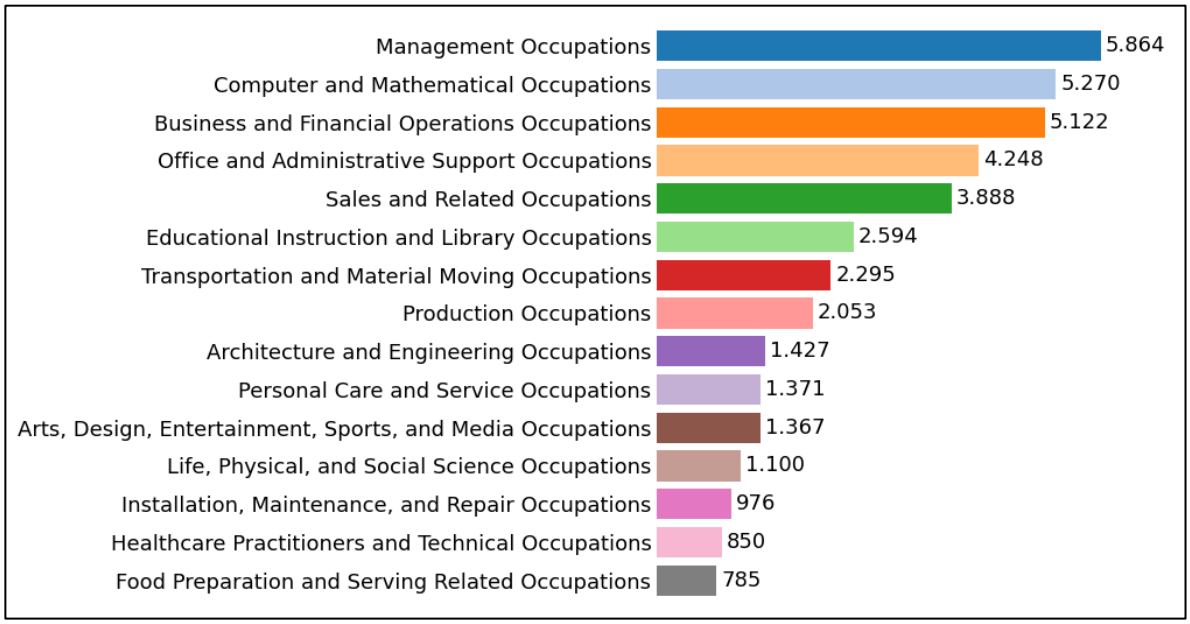
- *SOC major group:* Es el nivel más general en la clasificación. Agrupa amplias categorías profesionales basadas en áreas generales de actividad laboral, como "Gerencia", "Ventas", o "Transporte".
- *SOC minor group:* En este nivel se subdividen los grupos mayores en categorías más específicas. Agrupa ocupaciones que comparten características funcionales o competencias similares, como "Transportadores aéreos" o “Transportadores en vehículo” dentro del grupo mayor de "Transporte".
- *SOC broad ocupation:* Proporciona un desglose intermedio dentro de los grupos menores, clasificando ocupaciones relacionadas en conjuntos más definidos. Por ejemplo “Conductores de ambulancia” o “Conductores de vehículo de pasajeros” dentro del grupo menor “Transportadores de vehículo”.
- *SOC detailed ocupation:* Es el nivel más detallado del sistema SOC, que clasifica ocupaciones específicas con roles concretos y claramente definidos. Por ejemplo “Conductores de bus” o “Conductores de taxi” dentro de la ocupación amplia “Conductores de vehículo de pasajeros”.

*Figura 3. Estructura del sistema de Clasificación Ocupacional Estándar (SOC)*



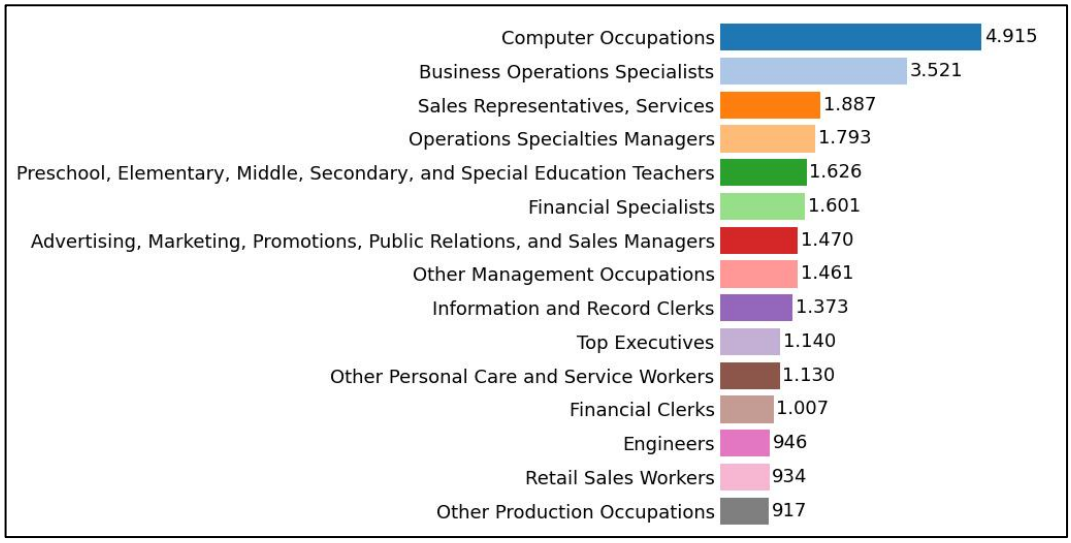

En cuanto a la clasificación por grupos de estándar ocupacional, en el grupo mayor (Figura 4) destacan las ocupaciones gerenciales *(management occupations),* las cuales representan el 13.3% del total de ofertas laborales en el período evaluado. Otros grupos que destacan son las ocupaciones de computación y matemáticas (11.9%), operaciones comerciales y financieras (11.6%), y de apoyo administrativo y de oficina (9.6%).

*Figura 4. Top 15 ofertas laborales por grupo mayor SOC*



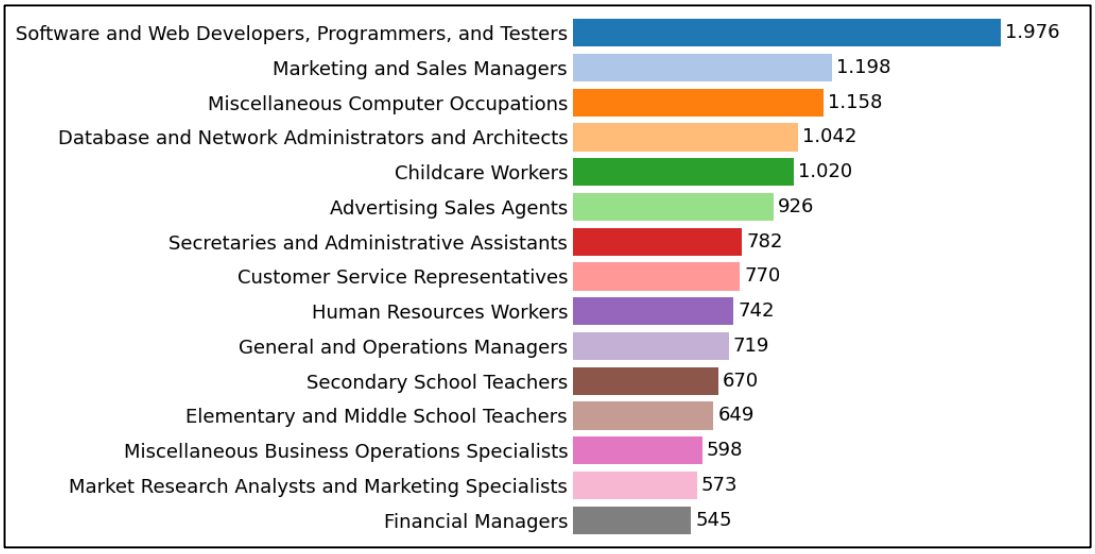
En el caso del grupo menor de la clasificación estándar (Figura 5), las ocupaciones informáticas *(computer ocupations)* encabezan la lista representando el 11,1% de las ofertas laborales. Le siguen los especialistas de operaciones de negocio (8%), representantes de ventas (4.3%) y gerencia de especialidades operativas (4%).

*Figura 5. Top 15 ofertas laborales por grupo menor SOC*



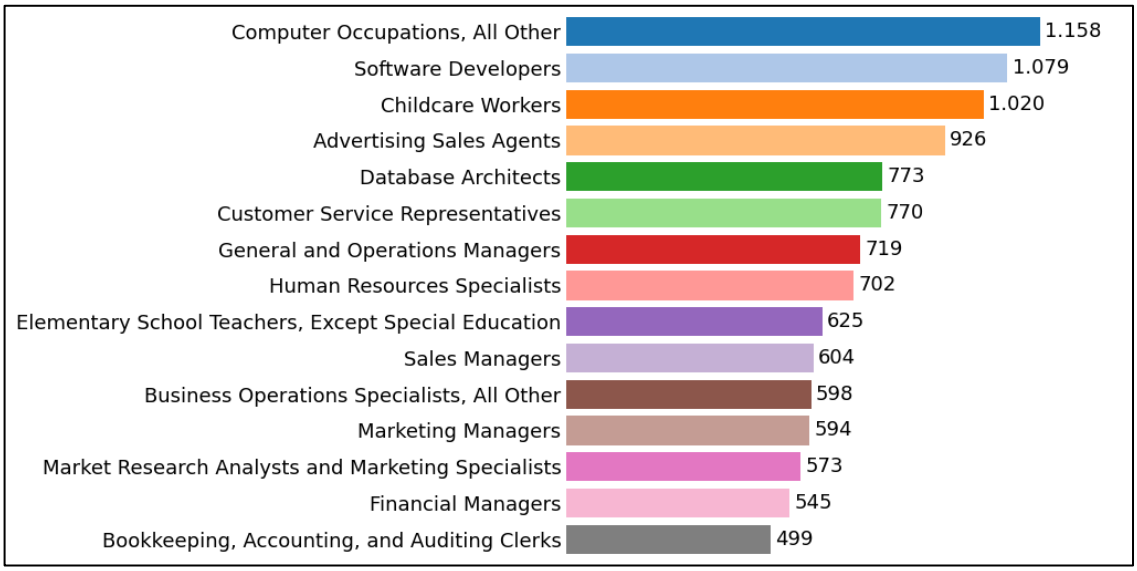

Por otra parte, en la clasificación de ocupación amplia (Figura 6) dominan las ofertas de desarrolladores, programadores y testers de software y web *(software and web developers, programmers and testers)* las cuales representan el 4.5% de la muestra. Le siguen la gerencia de mercadeo y ventas (2.7%), ocupaciones informáticas diversas (2.6%), y administradores y arquitectos de bases de datos y redes (2.3%).

*Figura 6. Top 15 ofertas laborales por ocupación amplia SOC*




Finalmente, para el caso de las ocupaciones detalladas (Figura 7), encabezan de nuevo las ocupaciones de computación *(computer ocupations, all other)* con un 2.6% de las ofertas. A continuación, destacan los desarrolladores de software (2.4%), trabajadores de cuidado infantil (2.3%) y agentes de ventas publicitarios (2.1%).

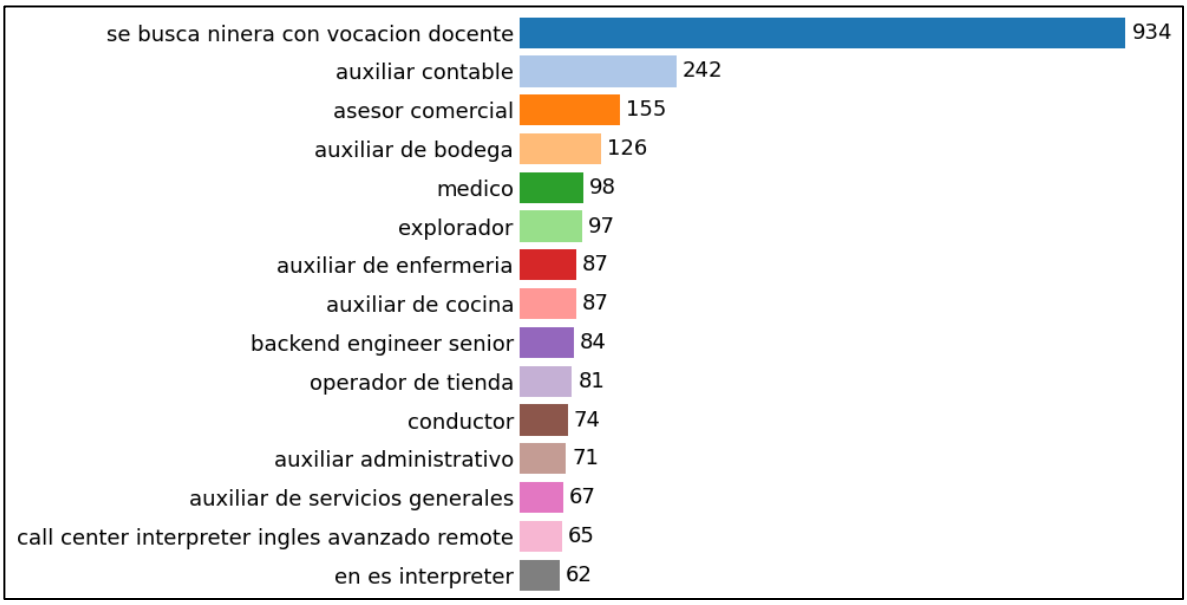
*Figura 7. Top 15 ofertas laborales por ocupación detallada SOC*



## 1.3. Análisis por Título de Oferta Laboral

Las ofertas laborales a nivel país presentan una alta concentración en un conjunto de títulos recurrentes. Como se observa en la Figura 8, el más frecuente es el de “niñera con vocación docente”, el cual representa el 2.1% de toda la muestra y corresponde a procesos masivos de contratación en el sector de cuidado infantil. Le siguen “auxiliar contable” y “asesor comercial”, que reflejan la fuerte demanda de perfiles de apoyo administrativo y de fuerza comercial. Otros títulos relevantes, aunque con menor frecuencia, son “auxiliar de bodega”, “médico”, “explorador”, “auxiliar de enfermería”, entre otros. En conjunto, estos quince títulos concentran el 5.2% de vacantes en la muestra, indicando que el mercado laboral colombiano se estructura alrededor de un núcleo de ocupaciones de soporte operativo, servicios personales y actividades comerciales.

*Figura 8. Top 15 títulos de ofertas laborales en todo el país*

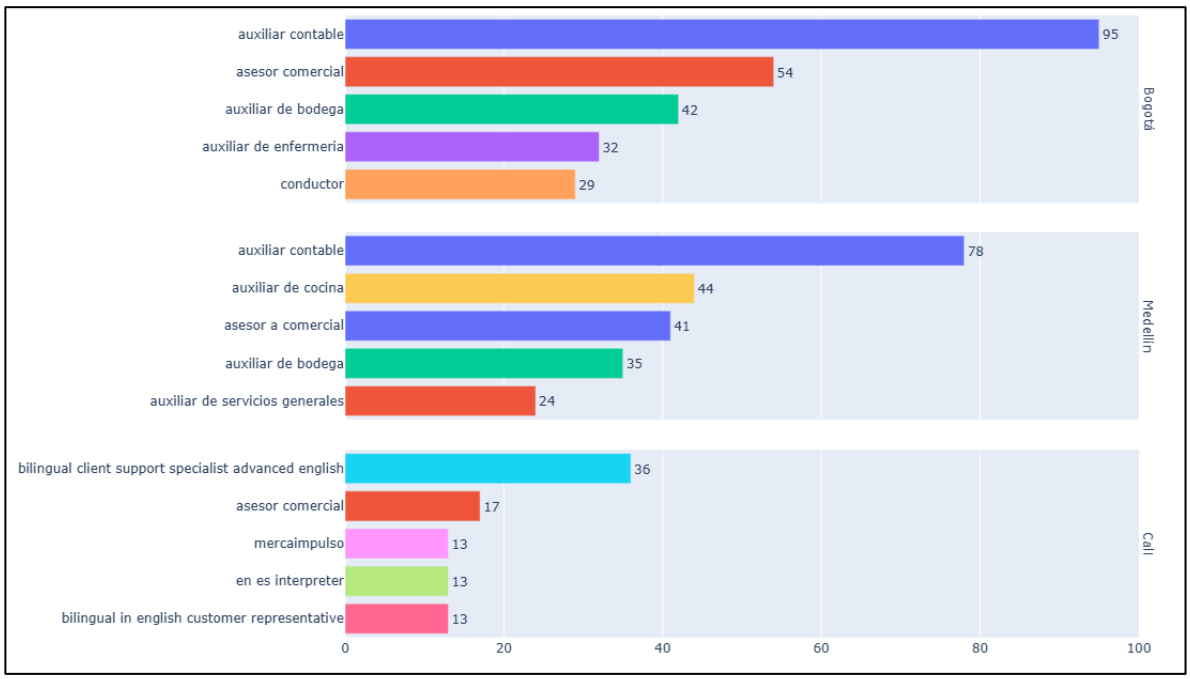


En cuanto a las principales ciudades del país (Bogotá, Medellín y Cali) representan más del 60% del total de vacantes de la muestra. La posición más demandada a nivel general es la de auxiliar contable (Figura 9), especialmente en Bogotá donde representa el 1% del total de vacantes en la ciudad. Otras posiciones demandadas en la capital son asesor comercial (0.5%) y auxiliar de bodega (0.4%).

En el caso de Medellín, también destaca la alta demanda de auxiliares contables, representando el 1.7% de las ofertas publicadas en la ciudad. Le siguen las ofertas de auxiliar de cocina (1%) y asesor comercial (0.9%).

Por último, en Cali la posición más demandada es la de soporte al cliente bilingüe, la cual representa el 2.2% de las vacantes en la ciudad. Le siguen posiciones como asesor comercial (1.1%) y mercaimpulso (1%).

*Figura 9. Top 5 ofertas laborales en ciudades principales (Bogotá, Medellín, Cali)*



## 2. Resultados de extracción de habilidades con *LLM*

Tal como se describió en el diseño metodológico, la extracción de habilidades se implementó mediante un modelo grande de lenguaje de la familia *Gemma*: g*emma-3-4b-it (instruct-tuned),* configurado para devolver listas de habilidades en español en formato JSON. El modelo se ejecutó en un entorno con *GPU Nvidia H100* (Figura 10), lo que permitió procesar las 44,077 descripciones de empleo en fragmentos *(shards)* de 2,000 registros.

*Figura 10. Especificaciones técnicas GPU Nvidia H100*

| | H100 |
|---|---|
| Form Factor | SXM5 |
| Max Power | 700W |
| FP64 TC \| FP32 TFLOPS[2] | 67 \| 67 |
| TF32 TC \| FP16 TC TFLOPS[2] | 989 \| 1979 |
| FP8 TC \| INT8 TC TFLOPS/TOPS[2] | 3958 \| 3958 |
| GPU Memory / Speed | 80GB HBM3 |
| Multi-Instance GPU (MIG) | Up to 7 |
| NVLink Connectivity | Up to 256 |

Se diseñó un *prompt* en español (Figura 11) compuesto primeramente por un mensaje de sistema que define al modelo como “extractor de habilidades”, y le da instrucción de responder únicamente con un *JSON* que contenga una lista de cadenas en español (sin duplicados). A continuación, se incluyó un mensaje de usuario que especifica:

- Qué se debe incluir (habilidades técnicas, blandas, lenguajes de programación, herramientas, metodologías).

- Qué se debe excluir (estudios básicos, criterios geográficos, beneficios, cultura organizacional, cargos, nombres de empresa).
- Cómo comportarse si la descripción está en inglés u otro idioma (interpretar el texto en el idioma original, pero devolver las habilidades en español, manteniendo solo acrónimos y nombres técnicos como *Python, SQL, Excel*).

*Figura 11. Prompt del extractor de habilidades*

```
# Prompts
SYSTEM = """
Eres un extractor de habilidades. Tu única tarea es leer descripciones de empleo
y devolver una lista de habilidades en español en formato JSON (lista de cadenas únicas).

Si la descripción de empleo está en inglés u otro idioma, primero interpreta el texto
en ese idioma y luego devuelve las habilidades traducidas al español.

Reglas:
- Devuelves SIEMPRE las habilidades en español.
- Conservas en su forma original los nombres propios técnicos (p.ej. "Python", "SQL", "Excel", "JavaScript").
- No añades explicaciones, texto extra ni comentarios: solo el JSON final.
""".strip()

USER_TMPL = """Extrae una lista de *habilidades* de la siguiente descripción de empleo.

Incluye:
- Habilidades técnicas (p.ej., contabilidad, carpintería, operación de maquinaria)
- Habilidades blandas (p.ej., comunicación, trabajo en equipo, gestión del tiempo)
- Lenguajes de programación (p.ej., Python, JavaScript)
- Herramientas y software (p.ej., Excel, Salesforce, AutoCAD)
- Metodologías (p.ej., Agile, Scrum)

Excluye:
- Términos generales (p.ej., capacitación, desarrollo profesional)
- Estudios básicos (p.ej., educación primaria, educación secundaria, bachiller, formación técnica)
- Criterios de zona de residencia (p.ej., residir en zonas específicas)
- Beneficios, cultura organizacional, nombres de empresa, cargos, frases genéricas.

Si la descripción está en inglés u otro idioma, interpreta el texto en ese idioma
pero DEVUELVE siempre las habilidades en español (salvo nombres propios como "Python", "SQL", etc.).

Devuelve **solo** un JSON con una lista de cadenas en español (sin duplicados), nada más.

Descripción de empleo:
\"\"\"{jd}\"\"\"
```

Para cada oferta de empleo, la función *extract_skills_gemma* (Figura 12):

- Construye la estructura de mensajes *(system + user)* con la descripción *(job_description).*

- Usa *apply_chat_template* y el tokenizador de *Gemma* para generar el *prompt* interno del modelo.
- Ejecuta *generate* con decodificación determinista (sin muestreo, temperatura 0) y recupera la respuesta.
- Extrae el bloque *JSON* del texto generado, lo interpreta con *json.loads* y obtiene la lista *skills* en español.

*Figura 12.Función extract_skills_gemma*

```python
MODEL_ID = "google/gemma-3-4b-it"

model = Gemma3ForConditionalGeneration.from_pretrained(
    MODEL_ID,
    device_map="auto",
    dtype=torch.bfloat16 if torch.cuda.is_available() else torch.float32,
).eval()

# ---- FUNCTION ----
def extract_skills_gemma(job_description: str, max_new_tokens: int = 200) -> list[str]:
    """Extracts skills in Spanish as JSON from a job description using Gemma-3-1B-IT."""

    # Build message list
    messages = [
        {"role": "system", "content": [{"type": "text", "text": SYSTEM}]},
        {"role": "user", "content": [{"type": "text", "text": USER_TMPL.format(jd=job_description)}]},
    ]

    # Build prompt string
    prompt_str = tokenizer.apply_chat_template(messages, add_generation_prompt=True, tokenize=False)

    # Tokenize and send to device
    inputs = tokenizer(prompt_str, return_tensors="pt").to(model.device)

    # Generate
    with torch.inference_mode():
        output = model.generate(**inputs, max_new_tokens=max_new_tokens, do_sample=False, temperature=0.0)

    decoded = tokenizer.decode(output[0], skip_special_tokens=True)

    # Try to extract valid JSON
    match = re.search(r"\[.*\]", decoded, flags=re.S)
    json_str = match.group(0) if match else "[]"

    try:
        skills = json.loads(json_str)
        skills = sorted({s.strip() for s in skills if isinstance(s, str) and s.strip()})
    except Exception:
        skills = []
```

Se implementó también un caché a través de la función *extract_with_cache* (Figura 13), con la finalidad de evitar recalcular habilidades para descripciones repetidas, y se creó la columna *skills_list* para cada registro. Esta función fue desplegada en fragmentos *(shards)* de 2,000 registros cada uno, para aprovechar el procesamiento en paralelo en varias sesiones.

*Figura 13. Función extract_with_cache*

```
START_ROW = 28000      # INCLUSIVE
END_ROW   = 30000      # EXCLUSIVE

SHARD_OUT = f"jobs_with_skills_gemma_shard_{START_ROW}_{END_ROW}.csv"

mask = (df["row_id"] >= START_ROW) & (df["row_id"] < END_ROW)
df_shard = df.loc[mask].copy().reset_index(drop=True)

# Run extract_skills_gemma
# Tip: small memoization to speed up duplicates
_cache = {}
def extract_with_cache(txt: str):
    key = txt.strip()
    if key in _cache:
        return _cache[key]
    res = extract_skills_gemma(key)
    _cache[key] = res
    return res

df_shard["skills_list"] = df_shard["job_text"].progress_apply(extract_with_cache)
df_shard.to_csv(SHARD_OUT, index=False, encoding="utf-8")
```

## 2.1. Cobertura y volumen de habilidades extraídas

De los 44,077 registros iniciales, la función de extracción obtuvo habilidades para 38,959, los cuales representan un 88.4% de la muestra. Mientras que los registros que arrojaron listas vacías fueron 5,118, un total del 11.6% de la muestra (Figura 14).

*Figura 14. Estadísticas descriptivas de habilidades extraídas*

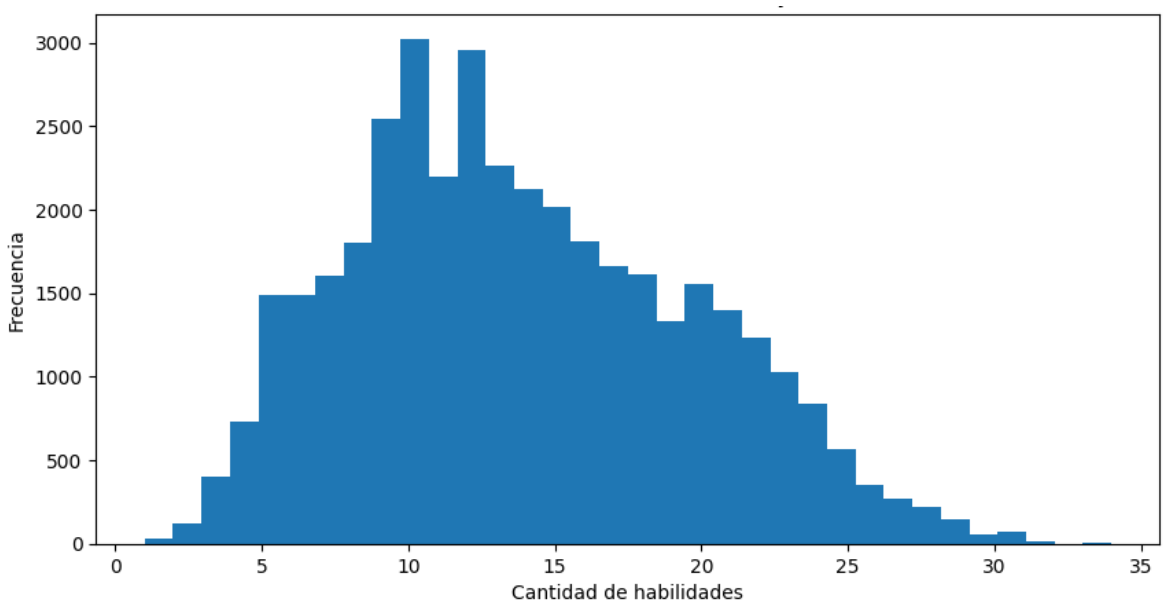
| | skills_count |
|---|---|
| count | 44077.000000 |
| mean | 12.182204 |
| std | 7.132960 |
| min | 0.000000 |
| 25% | 8.000000 |
| 50% | 12.000000 |
| 75% | 17.000000 |
| max | 34.000000 |
| non_zero_count | 38959.000000 |
| zero_count | 5118.000000 |

## 2.2. Distribución de habilidades por oferta

En cuanto a la distribución del número de habilidades extraídas por registro, se concentra principalmente entre 10 y 12 habilidades, que constituyen el pico del histograma, considerando únicamente aquellos casos en los que la lista no estaba vacía (Figura 15). La mayoría de los registros se encuentran dentro del rango de 5 a 20 habilidades, lo cual indica un comportamiento relativamente consistente en el proceso de extracción, con dispersión moderada y pocos casos excepcionales.

Existen algunos casos extremos, aunque poco frecuentes, con listas que alcanzan hasta 34 habilidades; pero en promedio, los registros contienen alrededor de 12 habilidades, lo que coincide con la zona de mayor densidad observada en la gráfica.

*Figura 15. Distribución del número de habilidades (excluyendo listas vacías)*




## 3. Resultados de normalización de habilidades a *ESCO* mediante *RAG*

En la tercera etapa del análisis se implementó el módulo de normalización basado en *RAG* descrito en el diseño metodológico. A partir del archivo oficial de *ESCO* en español, se construyó el índice semántico con *embeddinggemma-300m* y se implementaron las estrategias de selección de etiqueta canónica mediante el *LLM Gemma-3-4b-it.*

### 3.1. Construcción del índice de habilidades a partir de *ESCO*

En primer lugar, se construyó un espacio semántico de habilidades *ESCO* utilizando un modelo de *embeddings* de la familia *Gemma (Google): embeddinggemma-300m,* también de código abierto. La función *embed_texts* recibe una lista de cadenas y las convierte en vectores densos normalizados *L2* (Figura 16).

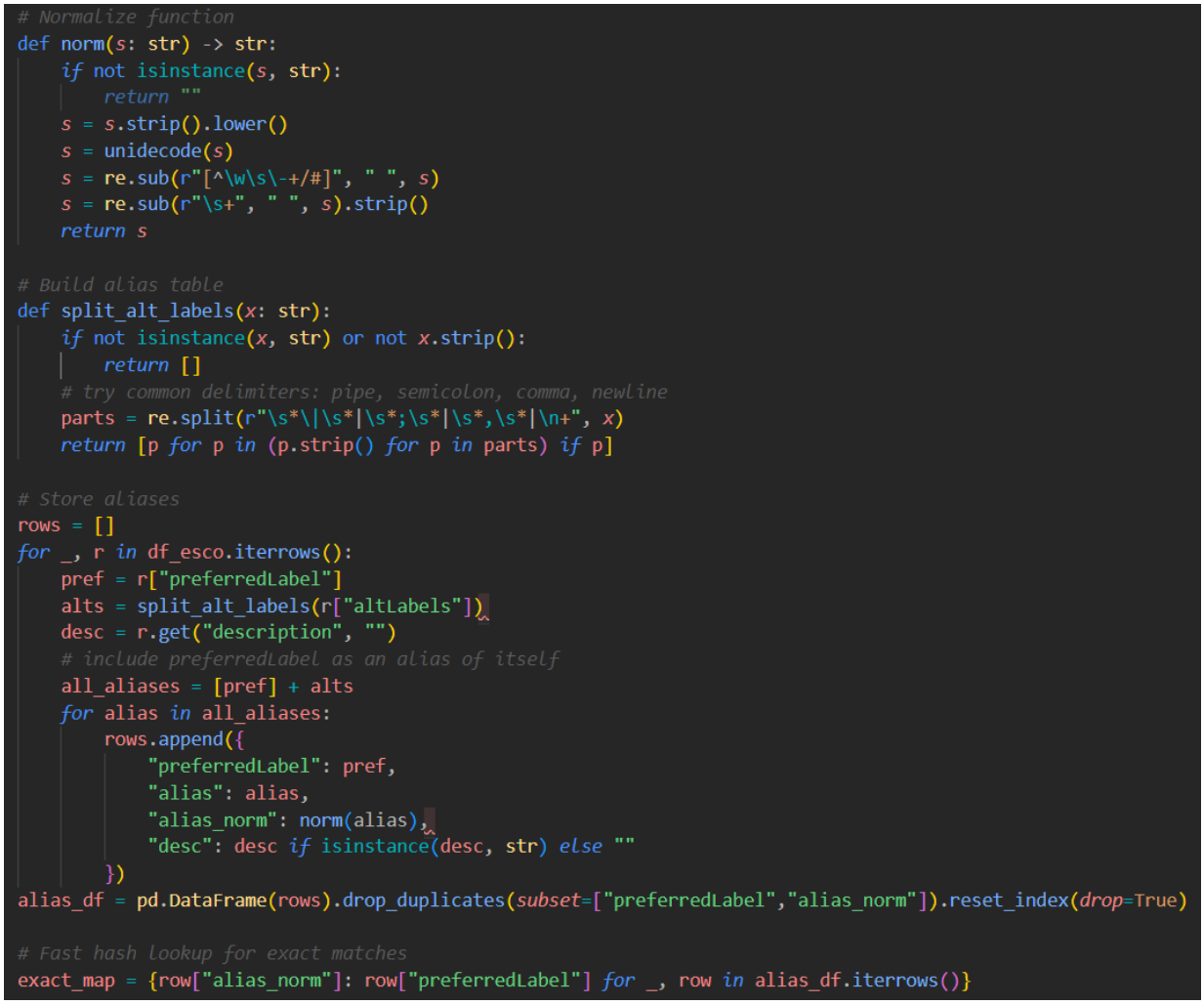
*Figura 16. Función de embedding*

```python
EMB_MODEL_ID = "google/embeddinggemma-300m"

emb_tokenizer = AutoTokenizer.from_pretrained(EMB_MODEL_ID)
emb_model = AutoModel.from_pretrained(
    EMB_MODEL_ID,
    device_map="auto",
    torch_dtype=torch.bfloat16 if torch.cuda.is_available() else torch.float32,
).eval()

# Embed function
def embed_texts(texts, batch_size: int = 64) -> np.ndarray:
    """
    Convierte una lista de strings en embeddings normalizados (L2) usando embeddinggemma-300m.
    Usa mean pooling sobre la última capa oculta con máscara de atención.
    """
    all_embs = []

    for i in range(0, len(texts), batch_size):
        batch = texts[i:i+batch_size]

        enc = emb_tokenizer(
            batch,
            padding=True,
            truncation=True,
            return_tensors="pt"
        ).to(emb_model.device)

        with torch.inference_mode():
            outputs = emb_model(**enc)
            # [B, T, H]
            last_hidden = outputs.last_hidden_state
            # Máscara de atención para hacer mean pooling adecuado
            attn_mask = enc["attention_mask"].unsqueeze(-1)  # [B, T, 1]
            masked = last_hidden * attn_mask
            summed = masked.sum(dim=1)                       # [B, H]
            lengths = attn_mask.sum(dim=1).clamp(min=1)      # [B, 1]
            embs = summed / lengths                          # mean pooling

        # Normalizamos para similitud coseno (FAISS con producto punto)
        embs = torch.nn.functional.normalize(embs, p=2, dim=1)
        all_embs.append(embs.detach().cpu().to(torch.float32).numpy())

    return np.vstack(all_embs)
```

Para ello se realizó tokenización del texto, se obtuvo la última capa oculta del modelo y se aplicó un *mean pooling* enmascarado (promedio solo sobre los tokens válidos), de modo que cada habilidad quede representada por un único vector. Posteriormente, estos vectores fueron normalizados para que el producto punto corresponda a una similitud coseno, lo que permite utilizarlos de forma eficiente en

un índice vectorial. Seguidamente, se preparó una tabla de alias de *ESCO* a partir del archivo *skills_es.csv,* como se especifica en la Figura 17.

*Figura 17. Preparación de tabla de alias de ESCO*



```python
# Normalize function
def norm(s: str) -> str:
    if not isinstance(s, str):
        return ""
    s = s.strip().lower()
    s = unidecode(s)
    s = re.sub(r"[^\w\s\-+/#]", " ", s)
    s = re.sub(r"\s+", " ", s).strip()
    return s

# Build alias table
def split_alt_labels(x: str):
    if not isinstance(x, str) or not x.strip():
        return []
    # try common delimiters: pipe, semicolon, comma, newline
    parts = re.split(r"\s*\|\s*|\s*;\s*|\s*,\s*|\n+", x)
    return [p for p in (p.strip() for p in parts) if p]

# Store aliases
rows = []
for _, r in df_esco.iterrows():
    pref = r["preferredLabel"]
    alts = split_alt_labels(r["altLabels"])
    desc = r.get("description", "")
    # include preferredLabel as an alias of itself
    all_aliases = [pref] + alts
    for alias in all_aliases:
        rows.append({
            "preferredLabel": pref,
            "alias": alias,
            "alias_norm": norm(alias),
            "desc": desc if isinstance(desc, str) else ""
        })
alias_df = pd.DataFrame(rows).drop_duplicates(subset=["preferredLabel","alias_norm"]).reset_index(drop=True)

# Fast hash lookup for exact matches
exact_map = {row["alias_norm"]: row["preferredLabel"] for _, row in alias_df.iterrows()}
```

Para cada fila se toma *preferredLabel*, sus *altLabels* y la descripción, y se generan múltiples alias normalizados. De este modo, expresiones como “Atención al cliente”, “Servicio al cliente” o variantes con mayúsculas y acentos quedan unificadas. El resultado fue almacenado en *alias_df*, junto con un diccionario *exact_map* para mapear coincidencias exactas en el espacio normalizado.

### 3.2. Indexación semántica con *EmbeddingGemma* y selección de la mejor etiqueta *ESCO*

Sobre la tabla de alias fue construido el índice de recuperación semántica. Para cada alias se realizó la concatenación del nombre con su descripción, para generar un *embedding.* Todos los vectores fueron agrupados en la matriz *emb_matrix* e indexados mediante *FAISS,* para buscar rápidamente los alias más similares en una consulta. La función *retrieve_esco_candidates* (Figura 18) utiliza este índice para recuperar los *k* alias más similares de *ESCO* junto con su *preferredLabel,* descripción y puntuación de similitud dado un texto de habilidad original.

*Figura 18. Función retrieve_esco_candidates*

```python
# -------- Build embeddings + FAISS index ---------
alias_text_for_embed = (alias_df["alias"].fillna("") + " — " + alias_df["desc"].fillna("")).tolist()
emb_matrix = embed_texts(alias_text_for_embed).astype("float32")

index = faiss.IndexFlatIP(emb_matrix.shape[1])  # cosine (with normalized vectors → dot product)
index.add(emb_matrix)

# Map from row id to (preferredLabel, alias)
id2canon = list(zip(alias_df["preferredLabel"].tolist(), alias_df["alias"].tolist(), alias_df["alias_norm"].tolist()))

# Candidate retrieval
_cand_cache = {}
def retrieve_esco_candidates(query: str, k: int = 8):
    key = ("cand", query, k)
    if key in _cand_cache:
        return _cand_cache[key]
    q = embed_texts([query]).astype("float32")
    sims, idxs = index.search(q, k)
    out = []
    for sim, idx in zip(sims[0], idxs[0]):
        pref, alias, alias_norm = id2canon[idx]
        desc = alias_df.iloc[idx]["desc"]
        out.append({"id": int(idx), "preferredLabel": pref, "alias": alias, "desc": desc, "sim": float(sim)})
    _cand_cache[key] = out
    return out
```

Sobre estos candidatos recuperados se hizo un *re-ranking,* en el cual el *LLM* debe escoger siempre entre los candidatos *ESCO* obtenidos en la fase previa. Este procedimiento sigue un esquema en tres niveles:

- *constrained_id:* Selección de *ID* con decodificación restringida.
- *json_select:* Selección de *JSON*, cuando la primera opción no produce una salida válida.

- *embed_fallback:* Selección del candidato con mayor similitud de *embeddings*, cuando el *LLM* no logra tomar una decisión.

Finalmente, a través de la función *normalize_one_skill_robust* (Figura 19) se aplicó este procedimiento a todas las habilidades extraídas de cada oferta de trabajo, para obtener un diccionario que incluye: *original, preferredLadel, method, sim, y raw* (Figura 20).

*Figura 19. Función normalize_one_skill_robust*

```python
_norm_cache = {}
def normalize_one_skill_robust(skill_surface: str, k: int = 8):
    """
    Try constrained-decoding → JSON-select → embed fallback.
    Returns dict: {'original':..., 'preferredLabel':..., 'method':..., 'sim':..., 'raw':...}
    """
    key = ("norm", skill_surface, k)
    if key in _norm_cache:
        return _norm_cache[key]

    # Tier 1
    pref, method, meta = select_esco_id_constrained(skill_surface, k=k)
    if pref is not None:
        rec = {"original": skill_surface, "preferredLabel": pref, "method": method, "sim": float(meta.get("sim", 0.0)), "raw": meta.get("raw","")}
        _norm_cache[key] = rec
        return rec

    # Tier 2
    pref, method, meta = choose_esco_with_llm_json(skill_surface, k=k)
    if pref is not None:
        rec = {"original": skill_surface, "preferredLabel": pref, "method": method, "sim": float(meta.get("sim", 0.0)), "raw": meta.get("raw","")}
        _norm_cache[key] = rec
        return rec

    # Tier 3 (embed top)
    cands = retrieve_esco_candidates(skill_surface, k=k)
    fb = cands[0]
    rec = {"original": skill_surface, "preferredLabel": fb["preferredLabel"], "method": "embed_fallback", "sim": float(fb["sim"]), "raw": ""}
    _norm_cache[key] = rec
    return rec
```

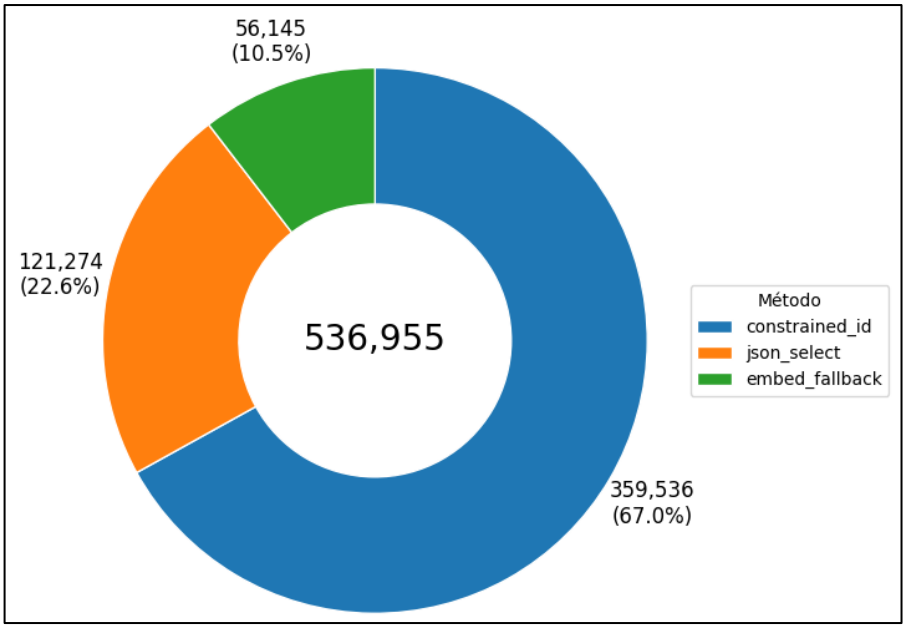
*Figura 20. Formato de salida de la función de normalización con RAG*

```
{
  "original": <texto original de la habilidad>,
  "preferredLabel": <habilidad ESCO seleccionada>,
  "method": <constrained_id | json_select | embed_fallback>,
  "sim": <similitud de embeddings>,
  "raw": <respuesta cruda dada por el LLM al seleccionar candidato>
}
```

### 3.3. Desempeño del módulo de normalización

La Figura 21 resume el comportamiento del módulo de normalización a ESCO según el método de decisión utilizado. En total se normalizaron 536.955 menciones de habilidades. De ellas, el 67,0% se resolvió mediante el primer nivel *(constrained_id)*, en el que el *LLM* selecciona un único candidato de la lista a partir de una decodificación restringida a dígitos. Un 22,6% adicional se normalizó mediante el segundo nivel *(json_select),* donde el modelo devuelve un pequeño objeto *JSON* indicando el número de candidato elegido, sin restricciones duras sobre los *logits.* Finalmente, solo el 10,5% de los casos requirió recurrir al tercer nivel *(embed_fallback),* en el que se toma directamente el alias *ESCO* con mayor similitud en el espacio de *embeddings.* En conjunto, estos resultados muestran que en casi nueve de cada diez menciones la decisión final proviene del *LLM*, pero siempre acotada al conjunto de candidatos recuperados, mientras que el *fallback* basado únicamente en similitud semántica funciona como un mecanismo de cobertura para menciones ruidosas o poco frecuentes.

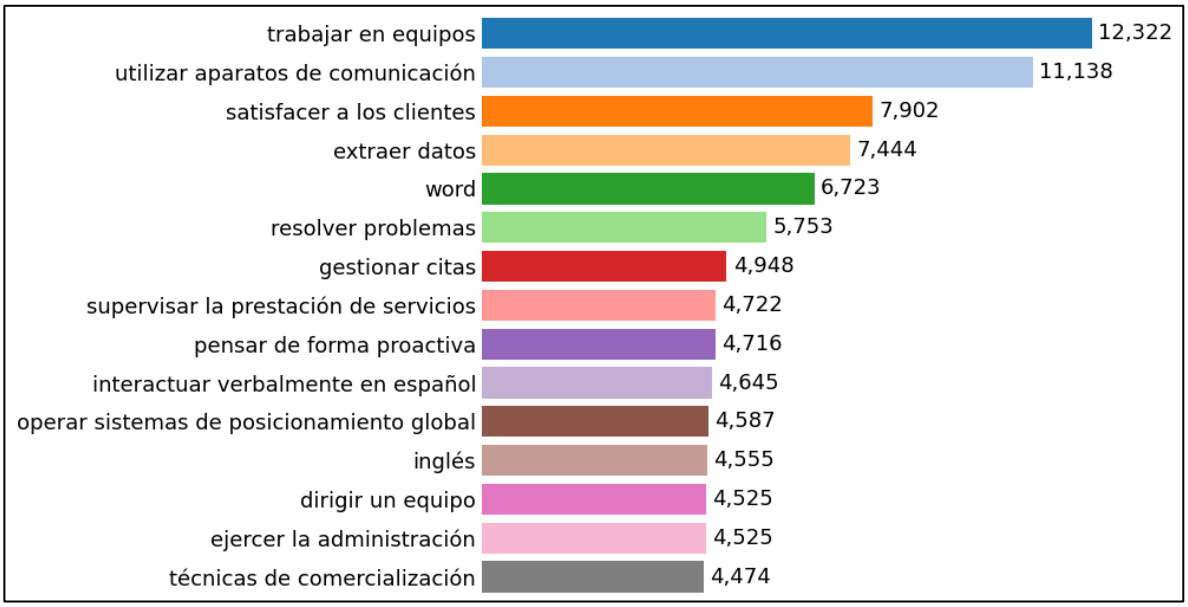
*Figura 21. Proporción de habilidades normalizadas por método*



### 3.4. Distribución de habilidades *ESCO*

La Figura 22 presenta las quince habilidades *ESCO* más frecuentes una vez completado el proceso de normalización. La habilidad con mayor número de menciones es “trabajar en equipos”, seguida de “utilizar aparatos de comunicación”, “satisfacer a los clientes” y “extraer datos”. En los siguientes lugares aparecen tanto habilidades digitales básicas (como “Word” u “operar sistemas de posicionamiento global”); como competencias transversales de carácter interpersonal (“resolver problemas”, “pensar de forma proactiva”, “interactuar verbalmente en español”, “dirigir un equipo”) y conocimientos lingüísticos (“inglés”). Este patrón sugiere que el mercado laboral colombiano se articula principalmente alrededor de una combinación de habilidades blandas orientadas a la colaboración y al servicio al cliente, junto con competencias digitales y administrativas generales, más que alrededor de habilidades técnicas altamente especializadas.

*Figura 22. Top 15 habilidades ESCO en la muestra*



## 4. Evaluación cuantitativa del enfoque *LLM + RAG*

Tal como se indica en el diseño metodológico, para valorar el desempeño del enfoque propuesto se optó por construir un conjunto de referencia representativo, con una muestra aproximada del 1% de los registros con extracción de habilidades válida (390 de 38,959), estratificada según las variables de región geográfica y de clasificación por grupo ocupacional estándar.

### 4.1. Resultados globales de evaluación

Sobre esta muestra estratificada del 1% de ofertas, el sistema alcanzó una precisión global $P = 0.8668$, una cobertura $R = 0.8735$ y un $F1\,micro = 0.8701$ (Figura 23). Esto significa que, en promedio, alrededor de 9 de cada 10 habilidades propuestas por el modelo son correctas según el conjunto de referencia, y que el sistema recupera cerca del 87% de las habilidades *ESCO* consideradas válidas por el evaluador humano. El equilibrio entre *precision* y *recall* indica que el pipeline no solo evita en buena medida asignaciones erróneas, sino que también logra capturar una fracción elevada de las habilidades que deberían estar presentes.

*Figura 23. Métricas de evaluación a nivel global*

| Métrica | Valor |
|---|---|
| Precisión (micro) | 0.8668 |
| Recall (micro) | 0.8735 |
| F1 (micro) | 0.8701 |
| Jaccard medio | 0.7852 |
| Exact match | 0.1949 |

A nivel de oferta, el índice de *Jaccard* promedio es $J_i = 0.7852$, lo que implica que, para una vacante típica, casi el 80% del conjunto de etiquetas *ESCO* coincide entre la salida del sistema y el conjunto de referencia. Sin embargo, la métrica de *exact match* alcanza únicamente el 19.49%, es decir, en aproximadamente una de cada cinco ofertas la lista de habilidades predicha coincide exactamente con la lista de referencia. Este resultado es coherente con la naturaleza estricta de la métrica: basta con que falte una habilidad correcta o se incluya una etiqueta adicional para que la coincidencia exacta deje de cumplirse. En conjunto, estos indicadores sugieren que el sistema produce listas de habilidades muy cercanas a las de referencia, aunque todavía existe un margen para refinar la cobertura y reducir pequeñas discrepancias en cada vacante.

### 4.2. Resultados desagregados por grupo ocupacional

Al desagregar las métricas por grandes grupos ocupacionales *SOC* (Figura 24) se observa que el rendimiento del sistema es consistentemente alto en todos los segmentos analizados, con valores de *F1* que se sitúan entre 0.86 y 0.97. Los mejores resultados se obtienen en dominios como *Healthcare Support Occupations,* donde el *precision* y el *recall* se sitúan alrededor del 97%, y en *Community and Social Service Occupations* o *Educational Instruction and Library Occupations,* que muestran F1 superiores al 90%. Estos grupos se caracterizan por una terminología relativamente estandarizada, lo que facilita tanto la recuperación de candidatos *ESCO* como la elección de la etiqueta correcta por parte del *LLM.*

En el extremo inferior se encuentran grupos como *Management Occupations* o algunas categorías de producción y servicios, donde el *F1* se sitúa alrededor del 86%. En estos dominios las descripciones de habilidades son más genéricas (“liderazgo”, “orientación al resultado”, “capacidad de negociación”), lo que aumenta la ambigüedad semántica y hace más difícil distinguir entre etiquetas *ESCO*

próximas. Aun así, los valores siguen siendo elevados, lo que indica que incluso en estos casos el sistema suele asignar habilidades razonables, aunque no siempre coincidan exactamente con la selección del evaluador humano.

Estas diferencias entre grupos ocupacionales sugieren que el pipeline *LLM + RAG* es especialmente robusto en segmentos técnicos o con vocabulario especializado, mientras que en ocupaciones de gestión o de perfil más generalista podría beneficiarse de ajustes adicionales, por ejemplo afinando el índice de recuperación, enriqueciendo el contexto pasado inicialmente al *LLM* o introduciendo reglas específicas para habilidades muy genéricas.

*Figura 24. Métricas de evaluación según grupo mayor SOC*

| enrich_soc_major_group | tp | fp | fn | precision | recall | f1 |
|---|---|---|---|---|---|---|
| Healthcare Support Occupations | 33.0 | 1.0 | 1.0 | 0.970588 | 0.970588 | 0.970588 |
| Community and Social Service Occupations | 103.0 | 10.0 | 8.0 | 0.911504 | 0.927928 | 0.919643 |
| Educational Instruction and Library Occupations | 171.0 | 17.0 | 17.0 | 0.909574 | 0.909574 | 0.909574 |
| Farming, Fishing, and Forestry Occupations | 10.0 | 1.0 | 1.0 | 0.909091 | 0.909091 | 0.909091 |
| Healthcare Practitioners and Technical Occupat... | 85.0 | 9.0 | 9.0 | 0.904255 | 0.904255 | 0.904255 |
| Arts, Design, Entertainment, Sports, and Media... | 235.0 | 28.0 | 24.0 | 0.893536 | 0.907336 | 0.900383 |
| Personal Care and Service Occupations | 36.0 | 4.0 | 4.0 | 0.900000 | 0.900000 | 0.900000 |
| Office and Administrative Support Occupations | 572.0 | 73.0 | 72.0 | 0.886822 | 0.888199 | 0.887510 |
| Business and Financial Operations Occupations | 632.0 | 83.0 | 78.0 | 0.883916 | 0.890141 | 0.887018 |
| Protective Service Occupations | 29.0 | 4.0 | 4.0 | 0.878788 | 0.878788 | 0.878788 |
| Sales and Related Occupations | 471.0 | 67.0 | 64.0 | 0.875465 | 0.880374 | 0.877912 |
| Installation, Maintenance, and Repair Occupations | 61.0 | 9.0 | 9.0 | 0.871429 | 0.871429 | 0.871429 |
| Food Preparation and Serving Related Occupations | 112.0 | 19.0 | 15.0 | 0.854962 | 0.881890 | 0.868217 |
| Production Occupations | 165.0 | 28.0 | 24.0 | 0.854922 | 0.873016 | 0.863874 |
| Management Occupations | 630.0 | 108.0 | 102.0 | 0.853659 | 0.860656 | 0.857143 |

A pesar de las limitaciones del tamaño muestral y de contar con un único anotador humano, los resultados apoyan la viabilidad del enfoque *LLM + RAG* para mapear de manera sistemática las habilidades extraídas a etiquetas *ESCO*, ofreciendo una base cuantitativa para futuros trabajos de mejora y ampliación del sistema.

## 5. Resultados de la clasificación de habilidades técnicas, blandas, emergentes y tradicionales

A partir de las habilidades *ESCO* normalizadas se construyó un listado de las 7,429 habilidades presentes en toda la muestra, las cuales fueron clasificadas automáticamente en categorías por tipo (técnicas/blandas) y carácter (emergentes/tradicionales), a través de un modelo grande de lenguaje de la familia *Gemma*: g*emma-3-4b-it (instruct-tuned),* usando el *prompt* descrito en la Figura 25.

*Figura 25. Prompt del clasificador de habilidades*

```
SYSTEM_PROMPT = """Eres un experto en análisis de habilidades laborales.
Vas a clasificar habilidades ESCO según estas dimensiones:

1) TIPO:
- "tecnica": conocimiento específico de una ocupación, herramienta, tecnología, procedimiento o normativa.
- "blanda": competencia interpersonal, comunicativa, de gestión personal o actitudinal (trabajo en equipo, liderazgo, etc.).

2) CARÁCTER TEMPORAL:
- "emergente":   habilidad asociada a tecnologías recientes, digitalización avanzada, datos, IA, ciberseguridad, sostenibilidad,
                 nuevos modelos de trabajo, etc.
- "tradicional": habilidad presente desde hace décadas en ocupaciones consolidadas (oficios clásicos, tareas administrativas
                 básicas, atención al cliente presencial, etc.).

Debes elegir SIEMPRE una combinación consistente y devolver un JSON con los campos:
- "tipo": "tecnica" | "blanda"
- "caracter": "emergente" | "tradicional"

No escribas nada más fuera del JSON.
"""

USER_TMPL = """Clasifica la siguiente habilidad ESCO.

Nombre de la habilidad: "{skill}"
Descripción oficial ESCO: "{desc}"

Frecuencia relativa en la muestra: {freq_band} (aparece en {n_offers} ofertas).

Devuelve SOLO un JSON con la forma:
{{
  "tipo": "...",
  "caracter": "..."
}}"""
```

A nivel de tipo de habilidad, el análisis muestra que prácticamente todas las ofertas de la muestra requieren al menos una habilidad técnica, con el 99,5% de las vacantes incluyendo este tipo de competencias. Las habilidades blandas también tienen una presencia muy extendida, ya que el 93,4% de las ofertas demandan alguna competencia interpersonal, comunicativa o de gestión personal (Figura 26). Esto sugiere que, en el mercado laboral colombiano, los perfiles buscados combinan de forma sistemática requisitos técnicos específicos con capacidades blandas transversales.

*Figura 26. Ofertas por tipo de habilidad (técnica / blanda)*

| tipo | cantidad_ofertas | porcentaje |
|---|---|---|
| técnica | 38778 | 0.995354 |
| blanda | 36414 | 0.934675 |

En cuanto al carácter temporal de las habilidades, casi la totalidad de las ofertas (99,9%) mencionan al menos una habilidad clasificada como tradicional, es decir, asociada a ocupaciones y tareas consolidadas. En contraste, sólo el 32,7% de las vacantes incluyen habilidades emergentes vinculadas a tecnologías recientes, digitalización avanzada o nuevos modelos de trabajo (Figura 27). Esto indica que, aunque las competencias emergentes empiezan a tener un peso relevante, el núcleo de la demanda laboral sigue anclado en habilidades de corte tradicional.

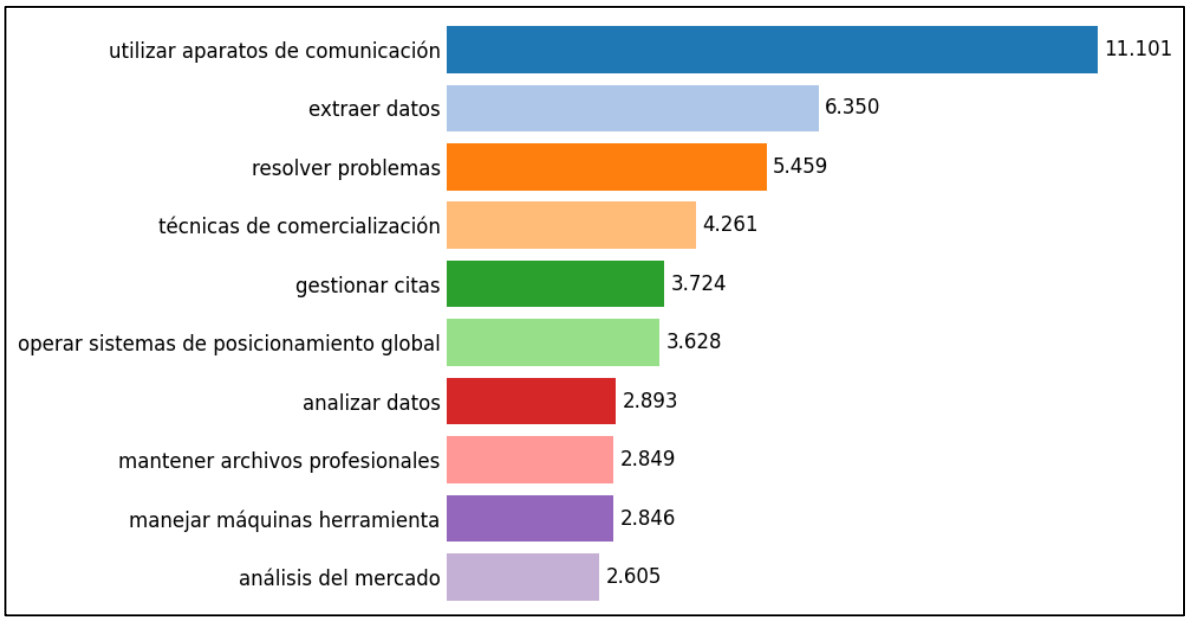
*Figura 27. Ofertas por carácter de habilidad (tradicional / emergente)*

| caracter | cantidad_ofertas | porcentaje |
|---|---|---|
| tradicional | 38957 | 0.999949 |
| emergente | 12763 | 0.327601 |

## 5.1. Habilidades técnicas / blandas

En la Figura 28 se presentan las diez habilidades técnicas *ESCO* más demandadas. Destaca en primer lugar “utilizar aparatos de comunicación”, que supera las 11,000 apariciones en la muestra y casi duplica la frecuencia de la siguiente habilidad, “extraer datos”. Le siguen “resolver problemas” y “técnicas de comercialización”, asociadas al análisis cuantitativo y las actividades comerciales. Completan el ranking habilidades como “gestionar citas”, “operar sistemas de posicionamiento global”, “analizar datos” o “mantener archivos profesionales”, que remiten a tareas de soporte administrativo y uso intensivo de herramientas digitales. En conjunto, este patrón sugiere un mercado laboral fuertemente orientado a la operación de tecnologías de la comunicación, la gestión de datos y los procesos comerciales.

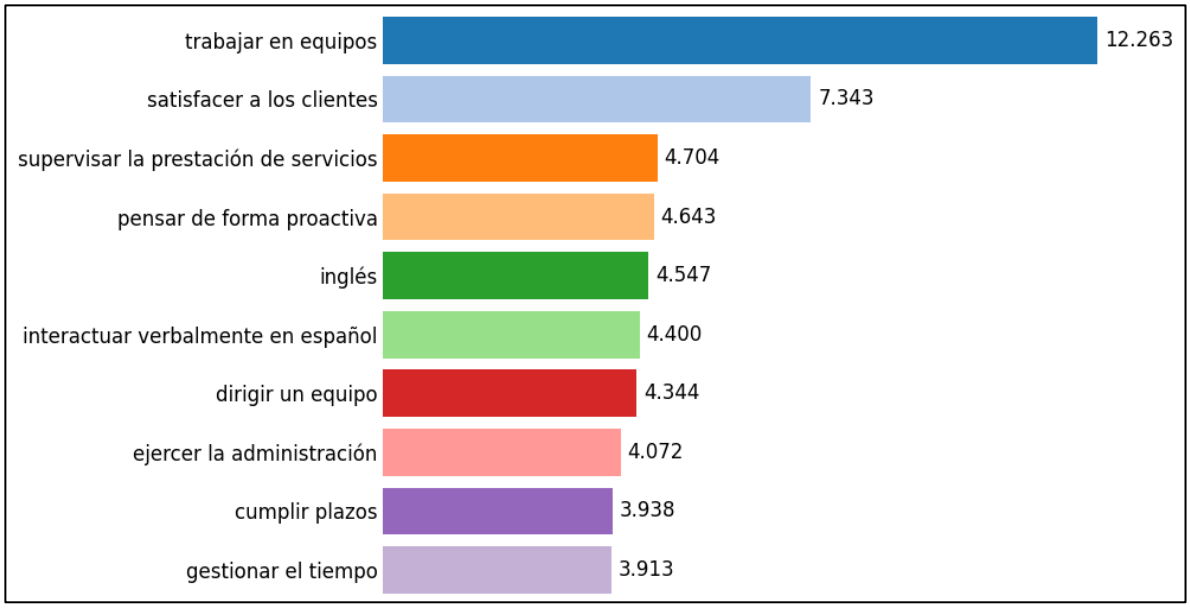
*Figura 28. Top 10 habilidades técnicas*




En cuanto a habilidades blandas, la más frecuente es “trabajar en equipos”, con más de 12,000 registros, seguida por “satisfacer a los clientes” y “supervisar la prestación

de servicios", lo que evidencia el peso de la atención al cliente y del trabajo colaborativo (Figura 29). Otras competencias destacadas son "pensar de forma proactiva", "interactuar verbalmente en español", "dirigir un equipo", "ejercer la administración", "cumplir plazos" y "gestionar el tiempo". Este conjunto refleja que, además del dominio técnico, las empresas valoran fuertemente las capacidades de organización, liderazgo, comunicación y orientación al servicio, coherentes con un mercado de servicios donde la interacción con clientes y equipos de trabajo es central.

*Figura 29. Top 10 habilidades blandas*

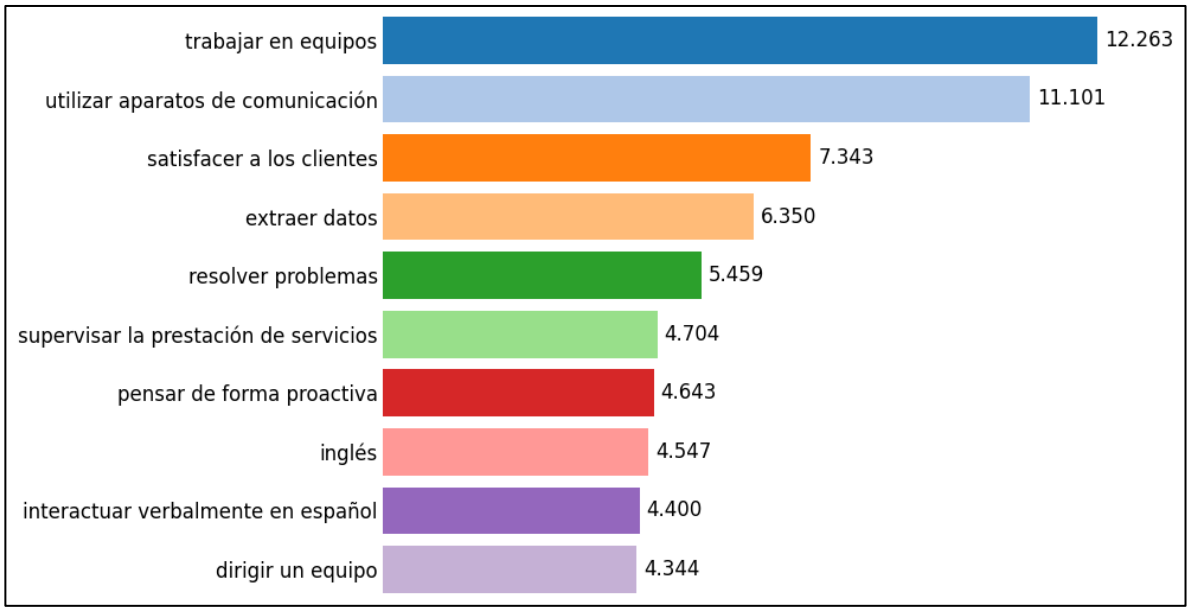



## 5.2. Habilidades tradicionales / emergentes

En el caso de las habilidades tradicionales (Figura 30), el ranking está liderado por "trabajar en equipos" y "utilizar aparatos de comunicación", ambas presentes en más de 11,000 ofertas. Esto confirma el peso de las competencias transversales ligadas al trabajo colaborativo y al uso de herramientas de comunicación en contextos

laborales consolidados. Les siguen “satisfacer a los clientes”, “resolver problemas” y “supervisar la prestación de servicios”, lo que refuerza la centralidad de la atención al cliente y la resolución de incidencias en el mercado laboral colombiano. Llama la atención la presencia de “inglés” e “interactuar verbalmente en español” entre las diez primeras, lo que sugiere que, incluso en roles tradicionales, las habilidades comunicativas y lingüísticas siguen siendo un requisito clave. En conjunto, el top tradicional dibuja un núcleo de habilidades asociadas a servicios, relaciones interpersonales y coordinación de equipos.

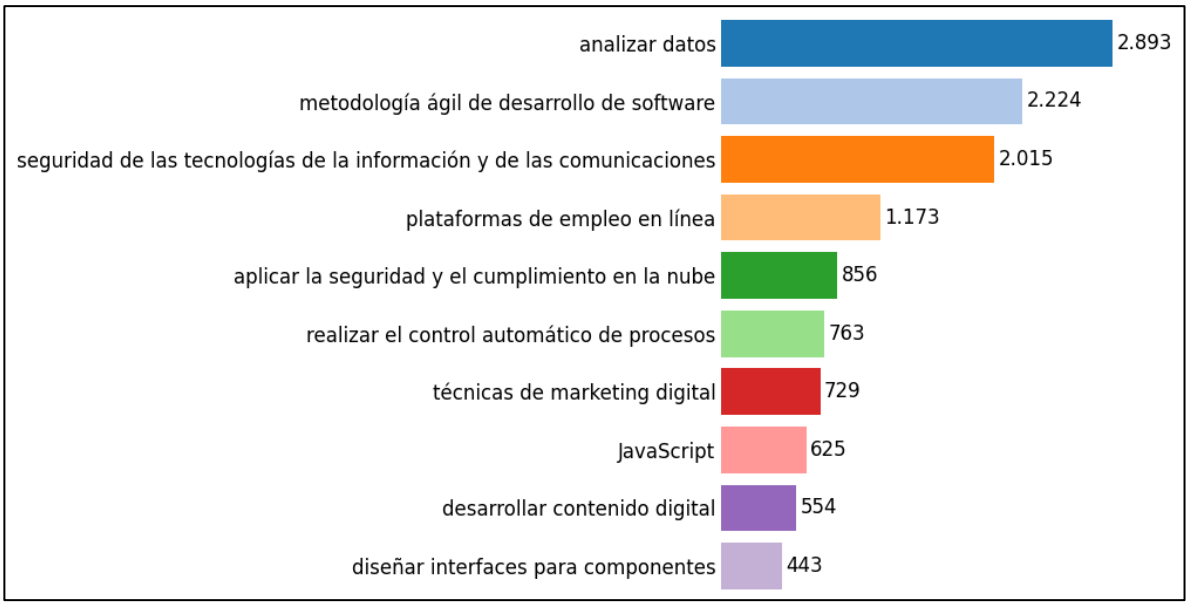
*Figura 30. Top 10 habilidades tradicionales*




En contraste, el ranking de habilidades emergentes (Figura 31) está claramente dominado por competencias vinculadas a la transformación digital. La habilidad más frecuente es “analizar datos”, seguida de “metodología ágil de desarrollo de software” y “seguridad de las tecnologías de la información y de las comunicaciones”, lo que refleja la creciente demanda de perfiles capaces de trabajar

con datos, desarrollo ágil y ciberseguridad. También destacan “plataformas de empleo en línea”, “aplicar la seguridad y el cumplimiento en la nube” y “realizar el control automático de procesos”, que apuntan hacia la expansión de servicios en la nube, automatización e industria 4.0. Finalmente, habilidades como “técnicas de marketing digital”, *“JavaScript”* y “desarrollar contenido digital” muestran cómo la economía digital exige capacidades en desarrollo web, experiencia de usuario y creación de contenido. Aunque sus frecuencias absolutas son menores que las tradicionales, este conjunto emergente marca con claridad los frentes tecnológicos donde se está concentrando la nueva demanda de talento.

*Figura 31. Top 10 habilidades emergentes*




Aunque la clasificación se apoya en un *LLM* y, por tanto, no equivale a un etiquetado humano exhaustivo, los resultados ofrecen una tipología consistente y suficientemente interpretable para el análisis agregado. Este esquema de cuatro categorías proporciona una base cuantitativa para discutir qué tipo de habilidades están impulsando actualmente el mercado laboral colombiano y en qué medida las

instituciones educativas deberían reforzar competencias técnicas emergentes frente a habilidades tradicionales ya consolidadas.

## 6. Líneas de trabajo futuro

Los resultados obtenidos muestran que el enfoque *LLM + RAG* es viable para extraer y normalizar habilidades a gran escala, pero también abren varias oportunidades de profundización. En primer lugar, sería recomendable ampliar la base de datos incorporando ofertas de otros portales, periodos de tiempo adicionales y actualizaciones periódicas del índice. Esto permitiría realizar análisis temporales de la demanda de habilidades, identificar tendencias de aparición y declive de competencias, y contrastar si los patrones observados en *Talent.com* se replican en otras fuentes de información laboral.

En segundo lugar, el pipeline técnico podría perfeccionarse en dos frentes. Por un lado, evaluando modelos de lenguaje de mayor tamaño, así como variantes multilingües que capturen mejor las ofertas redactadas en inglés o en registros mixtos. Por otro lado, refinando el módulo de normalización a *ESCO* mediante ajustes en el índice vectorial, técnicas de *re-ranking* más avanzadas y experimentos con *prompts* especializados para dominios ocupacionales concretos. Una línea complementaria sería separar explícitamente la evaluación de la fase de extracción y de la fase de normalización, construyendo conjuntos de referencia más amplios y con varios anotadores humanos para estimar la variabilidad entre evaluadores.

Finalmente, la clasificación en habilidades técnicas, blandas, emergentes y tradicionales puede aprovecharse para desarrollar herramientas aplicadas. Entre ellas se encuentran *dashboards* interactivos de inteligencia de mercado laboral para entidades públicas, observatorios sectoriales de habilidades emergentes y estudios de brechas entre la demanda identificada en las ofertas y la oferta formativa de

programas educativos específicos. Sería deseable integrar información adicional sobre el lado de la oferta, (como hojas de vida, egresados, formación) para avanzar hacia modelos de emparejamiento entre perfiles y vacantes, así como simulaciones de impacto de políticas de capacitación sobre la adecuación de habilidades en el mercado laboral colombiano.

UNIVERSIDAD EAFIT

## CONCLUSIONES

1. El proceso de recolección y limpieza de ofertas de empleo de *Talent.com* permitió obtener un corpus coherente y consistente, sobre el cual fue posible caracterizar con detalle la demanda laboral en Colombia por departamento, ciudad y grupos *SOC*. Existe una concentración espacial de oportunidades en Bogotá y Antioquia, y una estructura ocupacional dominada por grupos de gerencia, computación, operaciones financieras y apoyo administrativo, lo que confirma que el conjunto de datos captura adecuadamente las principales dinámicas del mercado laboral colombiano en el período analizado.

2. El pipeline propuesto, basado en un *LLM* de la familia *Gemma* combinado con un módulo *RAG* sobre la taxonomía *ESCO*, demostró ser una herramienta eficaz para extraer y normalizar habilidades a partir de texto libre. Los resultados indican que este enfoque permite mapear de manera sistemática las descripciones de ofertas a habilidades estandarizadas, reduciendo las alucinaciones y capturando patrones claros.

3. La clasificación de habilidades en las cuatro categorías analíticas permitió cuantificar con precisión la estructura de la demanda de competencias, mostrando que prácticamente todas las vacantes requieren habilidades técnicas y tradicionales, una proporción muy elevada exige también habilidades blandas, y un tercio incorpora habilidades emergentes. En conjunto, estos hallazgos muestran que el mercado laboral colombiano sigue anclado en un núcleo de competencias tradicionales, pero ya presenta rutas claras de demanda en áreas emergentes, ofreciendo insumos concretos para la planificación educativa y de políticas de formación.

UNIVERSIDAD EAFIT

## REFERENCIAS


Abadía, L. K., Bernal, G. L., Lizarazo, L., Ramos, J., Fernandez, Y., & Garzón, O. (2023). *Brecha de habilidades digitales, técnicas y blandas: Colombia antes y durante la pandemia*.

Aleisa, M. A., Beloff, N., & White, M. (2023). Implementing AIRM: a new AI recruiting model for the Saudi Arabia labour market. *Journal of Innovation and Entrepreneurship*, *12*(1), 1–41. https://doi.org/10.1186/s13731-023-00324-w

Alharbi, F., & Al-Alawi, A. I. (2024). Labor Market Prediction Using Machine Learning Methods: A Systematic Literature Review. *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, 478–482. https://doi.org/10.1109/ICETSIS61505.2024.10459632

Ao, Z., Horváth, G., Sheng, C., Song, Y., & Sun, Y. (2023). Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions. *Information Processing and Management*, *60*(2), 1–16. https://doi.org/10.1016/j.ipm.2022.103185

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Classifying online Job Advertisements through Machine Learning. *Future Generation Computer Systems*, *86*, 319–328. https://doi.org/10.1016/j.future.2018.03.035

Clavié, B., & Soulié, G. (2023). Large Language Models as Batteries-Included Zero-Shot ESCO Skills Matchers. *ArXiv Preprint*, 1–9. http://arxiv.org/abs/2307.03539

Colombo, E., Mercorio, F., & Mezzanzanica, M. (2019). AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, *47*, 27–37. https://doi.org/10.1016/j.infoecopol.2019.05.003

Díaz, A. M., & Salas, L. M. (2020). *Brecha de habilidades de los jóvenes en el mercado laboral colombiano*. www.onetline.org

Kavargyris, D. C., Georgiou, K., Papaioannou, E., Petrakis, K., Mittas, N., & Angelis, L. (2025). ESCOX: A tool for skill and occupation extraction using LLMs from unstructured text. *Software Impacts*, *25*, 1–9. https://doi.org/10.1016/j.simpa.2025.100772

Khaouja, I., Kassou, I., & Ghogho, M. (2021). A Survey on Skill Identification from Online Job Ads. *IEEE Access*, *9*, 118134–118153. https://doi.org/10.1109/ACCESS.2021.3106120

Magron, A., Dai, A., Zhang, M., Montariol, S., & Bosselut, A. (2024). JOBSKAPE: A Framework for Generating Synthetic Job Postings to Enhance Skill Matching. *ArXiv Preprint*, 1–16. http://arxiv.org/abs/2402.03242

Mezzanzanica, M., & Mercorio, F. (2019). *Big Data for Labour Market Intelligence*.

Nguyen, K. C., Zhang, M., Montariol, S., & Bosselut, A. (2024). Rethinking Skill Extraction in the Job Market Domain using Large Language Models. *Proceedings of the First Workshop on Natural Language Processing for Human Resources*, 27–42. https://huggingface.co/datasets/jjzha/green

Papoutsoglou, M., Ampatzoglou, A., Mittas, N., & Angelis, L. (2019). Extracting Knowledge from On-Line Sources for Software Engineering Labor Market:

A Mapping Study. *IEEE Access*, *7*, 157595–157613. https://doi.org/10.1109/ACCESS.2019.2949905

Papoutsoglou, M., Rigas, E. S., Kapitsaki, G. M., Angelis, L., & Wachs, J. (2022). Online labour market analytics for the green economy: The case of electric vehicles. *Technological Forecasting and Social Change*, *177 (121517)*, 1–14. https://doi.org/10.1016/j.techfore.2022.121517

Parida, B., KumarPatra, P., & Mohanty, S. (2022). Prediction of recommendations for employment utilizing machine learning procedures and geo-area based recommender framework. *Sustainable Operations and Computers*, *3*, 83–92. https://doi.org/10.1016/j.susoc.2021.11.001

Rahhal, I., Kassou, I., & Ghogho, M. (2024). Data science for job market analysis: A survey on applications and techniques. In *Expert Systems with Applications: Vol. 251 (124101)* (pp. 1–25). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2024.124101

Rahhal, I., Makdoun, I., Mezzour, G., Khaouja, I., Carley, K., & Kassou, I. (2019). Analyzing Cybersecurity Job Market Needs in Morocco by Mining Job Ads. *2019 IEEE Global Engineering Education Conference (EDUCON)*, 535–543.

Senthurvelautham, S., & Senanayake, N. (2023). A Machine Learning-Based Job Forecasting And Trend Analysis System To Predict Future Job Markets Using Historical Data. *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1–7.

Sharma, K., Kumar, P., & Li, Y. (2025). OG-RAG: Ontology-grounded retrieval-augmented generation for large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32951–32970. https://tinyurl.com/3e8pc2xr

World Health Organization. (2021). *Health labour market analysis guidebook*.

Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024). Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. *ArXiv Preprint*, 1–27. http://arxiv.org/abs/2409.14924

UNIVERSIDAD EAFIT

# ANEXOS

## Repositorio del proyecto

https://github.com/jorgeposadaz/LMI_Colombia