

# Contenido

Introducción .....	1
LoRA:.....	1
Hiperparámetros      2	
Resultados: .....	3
Análisis: .....	3
RAG .....	4
Funcionamiento (arquitectura general) .....	4
Modelos.....	5
Resultados .....	5
Consideraciones .....	7
ANEXO I: .....	7
LoRA.....	7
RAG .....	10

## Introducción

En esta práctica hemos intentado mejorar la gestión de cartas de presentación (cover letters) para ambos el solicitante y la empresa.

Para ello hemos aplicado la técnica LoRA (Low Rank Adaptation), para mejorar el rendimiento de 2 modelos en la generación de cartas de presentación.

Para la empresa hemos implementado un RAG (Retrieval Augmented Generation), para que introduciendo los requisitos que se buscan, te devuelva los top-K candidatos inscritos que más se adaptan a lo que se busca.

Para entrenar ambos modelos, se han utilizado las columnas **input** y **cover letter** del dataset de cartas de presentación en inglés de HuggingFace; <https://huggingface.co/datasets/dhruvvaidh/cover-letter-dataset-llama3>

## LoRA:

LoRA es una técnica de PEFT (parameter efficient fine tuning) muy eficiente, permitiendo mejorar el rendimiento de un modelo, modificando menos del 1% de los parámetros.

Se han elegido 2 modelos para comparar en la generación de cartas de presentación:

- Qwen2.5-3B cuantizado a 4bit
- IBM Watson 1B

Se han elegido modelos de este número de parámetros, porque es el tamaño más grande que cabían en la P100 de Google Colab o Kaggle.

Siguiendo la estructura del dataset, se ha decidido que la entrada tenga los campos de título del puesto, cualificaciones buscadas en el puesto, empresa, nombre del candidato, experiencia laboral, cualificaciones del candidato y competencias del candidato.

El prompt que se envía al modelo es el siguiente:

*SYSTEM PROMPT:*

*You are an assistant that writes professional cover letters.*

*Given the job description and candidate information,  
generate a tailored cover letter that highlights relevant experience,  
skills, and enthusiasm for the company.*

*INSTRUCTION:*

*Job Title: Senior Java Developer*

*Preferred Qualifications: 5+ years of experience in Java Development*

*Hiring Company: Google Applicant*

*Name: John Doe*

*Working Experience: Java Developer at XYZ for 3 years Senior Java Developer at ABC for 2 years*

*Qualifications: BSc in Computer Science Skillsets: Java, Spring Boot, Hibernate, SQL*

*RESPONSE:*

## Hiperparámetros

Los hiperparámetros se escogieron con los siguientes criterios:

El hiperparámetro **task\_type** se fijó en CAUSAL\_LM porque el objetivo era generar texto de manera secuencial, específicamente cartas de recomendación.

El valor de **r** se estableció en 16, un rango intermedio buscando el equilibrio entre la capacidad de adaptación y el coste computacional. Con valores más bajos se perdería flexibilidad para capturar patrones, mientras que con valores más altos el entrenamiento se volvería más pesado sin necesidad real.

El parámetro **lora\_alpha** se configuró en 32, actuando como factor de escalado moderado para dar estabilidad al entrenamiento y evitar que las actualizaciones fueran demasiado agresivas, reduciendo así el riesgo de sobreajuste.

El **lora\_dropout** se fijó en 0.05, un porcentaje pequeño para mejorar la generalización y prevenir que el modelo memorice ejemplos concretos del dataset, pero sin que se pierda demasiada información durante el ajuste.

En **target\_modules** se incluyeron los principales componentes de las capas de atención y feed-forward del Transformer, como q\_proj, k\_proj, v\_proj, o\_proj, gate\_proj, up\_proj y down\_proj.

Esta configuración nos ha permitido realizar un fine-tuning ligero pero efectivo, logrando que los modelos se adapten a la tarea sin necesidad de entrenamientos largos ni un consumo excesivo de recursos.

## Resultados:

Ambos modelos mejoraron respecto a la base, lo cual significa que el LoRA ha tenido efecto.

Modelo	BLEU	ROUGE
Qwen2.5-3B Base	46,16	0,62
Qwen2.5-3B LoRA	<b>47,01</b>	0,63
Watson 1B Base	46,81	0,62
Watson 1B LoRA	46,9	<b>0,64</b>

Los resultados muestran que aplicar LoRA produce mejoras pequeñas pero consistentes en ambos modelos: en **Qwen2.5-3B** se observa un incremento moderado en BLEU (+0,85), lo que indica una mayor precisión en la generación de texto, mientras que el aumento en ROUGE (+0,01) es casi marginal; en cambio, en **Watson 1B** la mejora en BLEU es mínima (+0,09), pero el ROUGE sube un poco más (+0,02), sugiriendo que el ajuste fino le ayuda a cubrir mejor el contenido de referencia.

En conjunto, los incrementos son modestos, típicos de LoRA, y reflejan que este método funciona como un refinamiento, que ajusta el modelo en la dirección correcta sin transformar radicalmente su rendimiento.

## Análisis:

Se ha pedido a los 4 modelos que generen una carta de recomendación con el mismo input. Se pueden ver las salidas en el Anexo I.

Las mejoras del LoRA se pueden apreciar en cómo las predicciones incluyen con mayor precisión los requisitos del puesto y el stack técnico esperado, reduciendo el relleno genérico de los modelos base y aumentando alineación con la “salida esperada”.

En Qwen2.5-3B, la versión base tiende a generalidades (“modelos y algoritmos”, “estrategias de datos”), mientras que la LoRA introduce menciones concretas a Azure Databricks, Data Factory, Synapse, Data Lake y la construcción de pipelines, además de dashboards en Power BI y SQL, acercándose más al énfasis del texto objetivo en visualización, automatización y herramientas de Azure. Por ejemplo, Qwen LoRA pasa de afirmar liderazgo y comunicación de forma abstracta a describir tareas específicas (mantener pipelines y crear reportes), lo que mejora la relevancia y el “recall” de habilidades clave.

En Watson, la base ya nombra con detalle el ecosistema de Azure (Databricks, Data Factory, SSAS, Synapse, Function) y la gestión de calidad y productos; sin embargo, la versión LoRA refina el enfoque hacia colaboración con equipos de producto, visualización orientada a decisiones y comunicación a públicos técnicos y no técnicos, además de corregir redundancias y enfatizar experiencia práctica (liderar analistas, estrategias de gestión de datos). Esto refuerza la adecuación al rol.

Comparando ambos modelos: Qwen gana más en especificidad técnica y mapeo de herramientas, mientras que Watson mejora en contextualización del impacto (cómo se usa la analítica para decisiones y coordinación entre equipos).

En ambos casos, LoRA reduce vaguedades, añade evidencias de tareas concretas y ajusta el tono hacia logros y responsabilidades que el reclutador espera leer, lo que se traduce en textos más útiles y persuasivos.

Aun así se puede ver que, sobretodo Watson, ha hecho un poco de overfitting, aunque el entrenamiento no fue muy largo, se observa cierto sobreajuste, repitiendo “Azure Data Factory” dos veces seguidas (“including Azure Data Factory, SSAS, Data Factory, Databricks...”), quitando claridad al texto. Esto se puede deber a que ha aprendido demasiado a adaptarse a la entrada. Esto explica que en Watson el ROUGE aumente más que el BLEU, ya que el modelo repite términos del dataset en vez de mejorar más la precisión de la redacción.

## **RAG**

El objetivo de este proyecto es implementar un sistema de Recuperación Aumentada por Generación (RAG) para automatizar la preselección de candidatos a partir de una colección de cartas de presentación.

El sistema RAG se utiliza para buscar documentos relevantes basándose en una descripción de trabajo y, posteriormente, utiliza un LLM para analizar, rankear y justificar la selección del candidato más adecuado, proporcionando un informe.

El sistema recibe como entrada una descripción del puesto (job description) y devuelve los candidatos más adecuados entre los *top-K* perfiles existentes, ordenados por similitud semántica. Esto permite a la empresa automatizar una parte del proceso de selección al identificar, de forma eficiente, los perfiles más alineados con las necesidades del rol.

## **Funcionamiento (arquitectura general)**

### **1. Embeddings**

La primera parte del proceso consiste en cargar un modelo de embeddings preentrenado especializado en representar texto en forma de embeddings.

Este modelo procesa cada cover letter generando su embedding correspondiente.

Cada carta de presentación queda así convertida en un vector que captura su significado semántico, esencial porque toda búsqueda se realiza en el espacio vectorial.

### **2. Creación de índices vectoriales (FAISS)**

Una vez generados los embeddings de todas las cover letters, se insertan en una base vectorial FAISS preparada para realizar consultas de tipo nearest neighbors (por proximidad).

### **3. Recuperación de candidatos**

Cuando un reclutador introduce una job description, el sistema genera con el mismo modelo de embeddings un vector de la oferta (asegurando que esté en el mismo espacio vectorial que las cover letters), realiza una búsqueda de proximidad en el índice FAISS y obtiene los *k* candidatos cuya carta de presentación es más cercana semánticamente a la oferta.

Estos top-K candidatos representan los perfiles más alineados con la descripción del puesto según el embedder.

#### 4. Generación / Razonamiento con LLM

La última fase consiste en analizar a los candidatos recuperados. Aquí entra en juego el LLM que recibe un prompt estructurado que incluye la descripción del puesto, la información de los candidatos recuperados y una plantilla de instrucciones que pide identificar al mejor candidato, justificar la decisión, y asignar una puntuación del 1 al 10 que refleje el nivel de adecuación al puesto.

Hemos determinado que se va a priorizar seleccionar solo un ganador en vez de un ranking completo de los candidatos por motivos que explicaremos más adelante, priorizando razonar sobre la selección y justificarla.

### Modelos

#### Modelo A

- **Retriever (embedder):** BGE (BAAI/bge-large-en-v1.5)
  - Tiene 335 millones de parámetros.
  - Es un modelo para transformar texto en vectores densos de alta calidad (dimensión 1024) centrado en tareas de búsqueda semántica y recuperación. Ha demostrado un buen desempeño en benchmarks de recuperación por su distribución de similitud que mejora la discriminación entre textos similares.
- **Generador (LLM):** Qwen2.5-3B (Qwen/Qwen2.5-3B)
  - Cuenta con 3 mil millones de parámetros.
  - Tiene un largo contexto (hasta 32 768 tokens) y soporte multilingüe, pero nosotros solo tenemos cartas en inglés.

#### Modelo B

- **Retriever (embedder):** E5 (intfloat/e5-large-v2)
  - Tiene 300 millones de parámetros.
  - Modelo preentrenado con un enfoque contrastivo para producir embeddings con dimensión de 1024. Su tamaño y rendimiento lo hacen ideal para tareas de recuperación semántica.
- **Generador (LLM):** Phi-3-mini (microsoft/Phi-3-mini-4k-instruct)
  - Tiene 3.8 mil millones de parámetros.
  - Destaca por su ventana de contexto de 128k tokens, que le permite analizar instrucciones largas. Aun siendo compacto, mantiene buena capacidad de razonamiento.

### Resultados

Aunque el pipeline es el mismo, en la práctica cada modelo exigió un prompt distinto para obtener resultados usables ya que la misma instrucción producía salidas muy diferentes en distintos y determinamos que la forma de expresar la petición condiciona fuertemente la respuesta del LLM. En nuestros modelos tuvimos que simplificar la instrucción del phi3 para que lograra entender lo que se le pedía.

Se realizaron ejecuciones comparativas utilizando la *Job Description*:

"We need a Project Manager with AWS certification, strong leadership, experience in other projects related to finances" para evaluar el rendimiento de los dos pipelines RAG.

#### Diferencias en candidatos recuperados

los candidatos propuestos por cada retriever difieren entre sí por que la diferente forma de calcular los embeddings supone vectores muy diferentes entre una misma carta. Cada embedder mapea texto a un espacio vectorial con sus propias prioridades semánticas causando un top-K diferente para cada modelo.

#### Resultados de Retrieval:

Los modelos Retriever produjeron conjuntos de candidatos recuperados y métricas de similitud de manera diferente:

Retriever	Similitud	Candidatos (IDs)	Observaciones Clave
BGE-large-en-v1.5	-0.0083, - 0.0114, -0.0116	810, 778, 117	Los valores de similitud son bajos y negativos, lo cual es inusual para <i>embeddings</i> normalizados y puede indicar un fallo en la alineación del espacio vectorial entre la <i>query</i> y los documentos.
E5-large-v2	0.7958, 0.7947, 0.7940	409, 193, 171	Los valores de similitud son altos con una búsqueda exitosa de documentos relevantes.

**Resultados de Generación**

La fase de Generación es la encargada de analizar y justificar la selección del mejor candidato de entre los top-K propuestos por el Retriever.

Generador (LLM)	Ganador (ID)	Racional de Decisión	Resultado
Qwen2.5-3B (3B Parámetros)	778	<i>The candidate with ID 778 has 8 years of experience in mobile app development, which is more than the required 5 years...</i>	Incoherente. El LLM ignora la Job Description (AWS/Finanzas) y alucina habilidades irrelevantes (mobile apps), indicando un fallo en el ranking a partir de un contexto de baja calidad.
Phi-3-mini-4k-instruct (3.8B Parámetros)	171	<i>Candidate 171 is the best fit for the Project Manager position with AWS certification and finance-related project experience. The cover letter explicitly mentions proficiency in SQL, Python, Scala, and AWS services...</i>	El LLM realiza un reranking efectivo y su justificación se alinea directamente con los requisitos de la Job Description.

**Análisis de resultados finales**

A pesar de que los candidatos propuestos por el *Retriever* son distintos, al pasar la información al LLM, la arquitectura RAG logró su objetivo en el Pipeline B: elegir un candidato bueno y lograr una buena explicación de su decisión.

- **Pipeline A (Qwen)** fracasó en la coherencia debido al mal rendimiento del retrieval BGE y a una propensión del LLM a la alucinación.
- **Pipeline B (Phi-3-mini)** seleccionó al candidato ID 171 (que no era el más similar según el E5), pero demostró una efectiva capacidad de razonamiento ya que su justificación se centró en las palabras clave específicas de la Job Description. Este resultado valida la hipótesis de que la fase de Generación puede compensar variaciones en la fase de Recuperación.

El Pipeline B demostró una mayor capacidad para seguir instrucciones complejas y realizar un reranking coherente que el A, a pesar de tener una diferencia pequeña en el número de parámetros la calidad de la respuesta del fue radicalmente mejor.

## Consideraciones

En un inicio, el objetivo del sistema era generar un ranking completo de candidatos en lugar de seleccionar únicamente a uno. Sin embargo, tras probar varios modelos distintos a los incluidos en el código entregado, observamos que solo Qwen fue capaz de producir un ranking, mientras que el resto generaba respuestas aberrantes. Esto parece deberse a la reducida cantidad de parámetros de muchos modelos, que dificulta que comprendan instrucciones complejas con la precisión de modelos mucho más grandes como ChatGPT o Gemini. Para reforzar el razonamiento de la generación, solicitamos al modelo una puntuación del 1 al 10 que a asignar al candidato final para valorar como de correcta era su selección para la oferta requerida.

Tuvimos de limitación la infraestructura disponible: al trabajar con GPUs P100 del entorno Kaggle como nuestra máxima GPU, la memoria resultaba insuficiente (16GB) para cargar modelos completos. Por ello, nos vimos obligados a cuantizar todos los modelos, ya que en muchos casos era muy costoso, e incluso imposible, ejecutarlos en su formato original.

Por último, observamos que la forma en la que se redacta el prompt influye enormemente en el resultado. Aunque las instrucciones transmitían las mismas ideas, pequeños cambios en la formulación alteraban significativamente la calidad y la estabilidad de la generación. Esto confirma la sensibilidad de los modelos pequeños a la ingeniería de prompts y que instrucciones muy complejas en longitud le suponen un problema.

## ANEXO I:

### LoRA

#### ***Salida esperada:***

Dear Hiring Manager,

I am writing to express my interest in the Data Scientist position at XYZ Data Solutions. With my strong background in data analysis and visualization, I believe I am well-suited for this role.

In my current position as a Data Scientist at XYZ Data Solutions, I have successfully led a project team in creating cutting-edge visualizations to meet our clients' needs. I have also automated build and deployment processes, resulting in more efficient data management. Additionally, I have been responsible for investigating new product features and identifying new opportunities for our clients.

I have a deep knowledge and experience within Azure, including Databricks, Data Factory, SSAS, Synapse, and Function. I am proficient in Power BI and SQL, and I have a strong understanding of data visualization and management. My strong communication skills and ability to work well within a team make me an ideal candidate for this position.

I hold a Bachelor's degree in Computer Science and have over 3 years of experience in data and analytics. I am highly motivated and continuously strive for improvement. I am confident that my skills and qualifications make me a valuable asset to XYZ Data Solutions.

Thank you for considering my application. I look forward to the opportunity to discuss how my skills and experience align with the needs of XYZ Data Solutions.

Sincerely,

John Smith

### ***Predicción de Qwen2.5-3B Base:***

Dear Hiring Manager,

I am writing to express my interest in the Data Scientist position at XYZ Data Solutions. With my strong background in data analysis and visualization, I believe I am well-suited for this role.

In my current position as a Data Scientist at XYZ Data Solutions, I have successfully developed and implemented data models and algorithms to extract valuable insights from complex datasets. I have also collaborated with cross-functional teams to design and execute data strategies that drive business value. My proficiency in Azure, Power BI, and SQL has allowed me to effectively analyze and visualize data, providing actionable insights to stakeholders.

I am a strong communicator with excellent leadership skills, and I am able to effectively lead and influence teams. My ability to tackle ambiguous problems and my strong problem-solving skills have enabled me to find innovative solutions to complex data challenges. Additionally, my knowledge of Microsoft Office, including Excel and PowerPoint, has allowed me to effectively prepare and present data to stakeholders.

I am excited about the opportunity to join XYZ Data Solutions and contribute to the development of data strategies that drive business value. I am confident that my skills and qualifications make me a strong fit for this role. Thank you for considering my application.

Sincerely,

John Smith

### ***Predicción de Qwen2.5-3B LoRA:***

Dear Hiring Manager,

I am writing to express my interest in the Data Scientist position at XYZ Data Solutions. With my strong background in data analysis and experience in utilizing Azure tools such as Databricks, Data Factory, and Synapse, I believe I would be a valuable asset to your team.

Throughout my career as a Data Analyst at ABC Company, I have gained extensive experience in data management and analysis. I have successfully developed and maintained data pipelines using Azure data factory and Azure Data Lake, ensuring efficient data processing and storage. Additionally, my experience with Power BI and SQL has allowed me to create insightful dashboards and reports for decision making.

I am highly proficient in Python and have a solid understanding of data visualization techniques. My strong communication skills enable me to effectively communicate complex data concepts to both technical and nontechnical stakeholders. Furthermore, my leadership and problem-solving abilities make me a valuable team member.



I am excited about the opportunity to contribute to XYZ Data Solutions and further enhance my skills in data science. Thank you for considering my application. I look forward to the possibility of discussing how my qualifications align with your company's needs.

Sincerely,

John Smith

### ***Predicción de Watson 3B Base:***

Dear Hiring Manager,

I am writing to express my interest in the Data Scientist position at XYZ Data Solutions. With my strong background in data analysis and expertise in Azure, Power BI, SQL, and Python, I believe I am wellsuited for this role.

In my current position as a Data Scientist at XYZ Data Solutions, I have successfully built scalable data pipelines and implemented data management and quality initiatives. I have also led the development of data products and solutions, ensuring alignment with business objectives. My strong communication skills have allowed me to effectively collaborate with various stakeholders, including business leaders and technical teams.

I have a proven track record of delivering quality results and have a solid understanding of Microsoft Azure data services, including Databricks, Data Factory, SSAS, Synapse, and Function. My proficiency in Power BI, SQL, and Python enables me to effectively analyze and visualize complex data sets. Additionally, my strong leadership and problemsolving skills have allowed me to successfully lead crossfunctional teams and deliver results on time and within budget.

I am excited about the opportunity to contribute to the growth and success of XYZ Data Solutions. Thank you for considering my application. I look forward to the possibility of discussing my qualifications further.

Sincerely,

John Smith

### ***Predicción de Watson 3B LoRA:***

Dear Hiring Manager,

I am writing to express my interest in the Data Scientist position at XYZ Data Solutions. With my strong background in data analysis and proficiency in Azure and Power BI, I believe I am wellsuited for this role.

In my current position as a Data Scientist at XYZ Data Solutions, I have successfully led a team of analysts and developed data management strategies to support business decisions. I have also collaborated with product teams to understand their needs and deliver valuable insights through data visualization. My strong communication skills have allowed me to effectively communicate complex data concepts to both technical and non-technical stakeholders.

I have a deep knowledge and experience within Azure, including Azure Data Factory, SSAS, Data Factory, Databricks, and Power BI. I am proficient in Python and have a strong understanding of data visualization and SQL. Additionally, my leadership skills and problemsolving abilities have enabled me to effectively manage data projects and drive results.

I hold a Bachelor's degree in Computer Science and have over 3 years of experience in data and analytics. I am confident that my skills and qualifications make me a strong candidate for this position.

Thank you for considering my application. I look forward to the opportunity to contribute to the success of XYZ Data Solutions.

Sincerely,

John Smith

## **RAG**

### ***Job description***

We need a Project Manager with AWS certification, strong leadership, experience in other projects related to finances

### ***BGE + QWEN***

Assistant: \*\*WINNER ID: 778

\*ATCH SCORE: 9

DECISION RATIONALE: The candidate with ID 778 has 8 years of experience in mobile app development, which is more than the required 5 years mentioned in the job description. Additionally, they have a Master's degree in Computer Science, which is a strong requirement for the position. They are proficient in iOS, Android, Swift, Java, and Kotlin, which are all relevant technologies for mobile app development. Furthermore, they possess excellent communication skills and the ability to work independently, which are essential qualities for a project manager. Overall, their extensive experience and qualifications make them the best fit for this role.

### ***E5 + phi3***

WINNER ID: 171

MATCH SCORE: 9

DECISION RATIONALE: Candidate 171 is the best fit for the Project Manager position with AWS certification and finance-related project experience. The cover letter explicitly mentions proficiency in SQL, Python, Scala, and AWS services, aligning closely with the job requirements.