

## **Informe de Trabajo Práctico de LLM**

### **Introducción**

El modelo Falcon ha capturado la atención de la comunidad de IA por sus impresionantes características y rendimiento. Se han presentado dos versiones de Falcon: Falcon 7B y Falcon 40B, que cuentan con 7.000 y 40.000 millones de parámetros, respectivamente. Para poner esto en perspectiva, GPT-3 de OpenAI tiene 175.000 millones de parámetros, mientras que LLaMA ofrece versiones que oscilan entre los 7.000 y 65.000 millones de parámetros. A pesar de tener menos parámetros que GPT-3, Falcon 40B ha logrado situarse en las primeras posiciones de la Open LLM Leaderboard de Huggingface, lo que demuestra su eficiencia y capacidad.

En este informe, se justifica el uso de Falcon-7B para la creación de un chatbot limitado a cierta información. Específicamente, se utiliza la variante "Águila" disponible en Hugging Face. Águila es un modelo preentrenado por la Unidad de Tecnologías del Lenguaje del Barcelona Supercomputing Center, con un conjunto de datos primordialmente en español y catalán. Este chatbot se afina utilizando el modelo "edumunozsala/aguila-7b-instructft-bactrian-x", el cual ha sido ajustado con el dataset Bactrian-X, optimizando la interacción en español.

El chatbot desarrollado es un asistente diseñado para responder preguntas analizando un contexto dado. En este caso, el chatbot interpreta noticias y responde preguntas relacionadas con el tema. Este trabajo marca el inicio del desarrollo de un periodista virtual, un proyecto futuro que tiene como objetivo presentar noticias de manera automatizada, proporcionando resúmenes, análisis y respuestas detalladas sobre eventos actuales.

Debido a las limitaciones de acceso a GPU's, el modelo se despliega utilizando técnicas de ajuste eficiente de parámetros y cuantización a 4 bits, asegurando así un uso eficiente de los recursos disponibles sin comprometer el rendimiento. Este enfoque no solo busca crear un asistente conversacional eficaz, sino también sentar las bases para un periodista virtual que mejorará la forma en que consumimos y comprendemos las noticias.

## **Desarrollo**

El modelo Águila-7B es un modelo de lenguaje basado en transformers, diseñado específicamente para los idiomas catalán, español e inglés. Este modelo se basa en el modelo Falcon-7B y ha sido entrenado en un corpus trilingüe de 26 mil millones de tokens, recopilado a partir de corpus públicos y crawlers disponibles en línea. Se destaca por su capacidad para generar texto coherente y relevante en estos tres idiomas.

Para su entrenamiento, se utilizó el conjunto de datos Bactrian-X, que consta de 3.4 millones de pares de instrucciones-respuestas en 52 idiomas. Estos datos fueron obtenidos mediante la traducción de 67 mil instrucciones en inglés a 51 idiomas diferentes utilizando la API de Google Translate. Posteriormente, las instrucciones traducidas se utilizaron para generar respuestas naturales utilizando ChatGPT, lo que resultó en un conjunto de datos multilingüe de 3.4 millones de instancias.

Durante el entrenamiento, se emplearon hiperparámetros específicos, como la configuración de 4 bits para el modelo, utilizando la biblioteca Axolot y la biblioteca PEFT para el ajuste fino supervisado. Además, se implementaron adaptadores QLoRA para mejorar el rendimiento del modelo en tareas de generación de texto.

El modelo Águila-7B ofrece una amplia gama de aplicaciones, desde la generación de respuestas en diálogos hasta la producción de texto coherente en diferentes contextos. Su entrenamiento y arquitectura específicos lo convierten en una herramienta poderosa para el procesamiento de lenguaje natural en los idiomas mencionados.

Este modelo ha sido desarrollado por Eduardo Muñoz y está disponible para su uso a través de la plataforma Hugging Face. Su capacidad para generar texto de alta calidad en múltiples idiomas lo hace especialmente útil para aplicaciones de traducción automática, asistencia virtual y generación de contenido.

## Dataset utilizado

El dataset de noticias de la BBC traducido al español consta de 1580 entradas y tres columnas:

- ArticleId: Un identificador único para cada artículo de noticias.
- Texto: El contenido del artículo de noticias traducido al español.
- Categoría: La categoría a la que pertenece el artículo de noticias, que puede ser 'negocios' 'tecnología' 'política' 'deporte' 'entretenimiento'.

No hay valores nulos en ninguna de las columnas, Este dataset proporciona una colección de noticias en español de la BBC, cubriendo temas relacionados con negocios, política, deporte, entretenimiento y tecnología. Es útil para tareas de análisis de texto, clasificación de texto y procesamiento de lenguaje natural en general.

## Librerías Necesarias

Para la implementación y despliegue del modelo, se utilizaron las siguientes librerías:

- Einops: (Einstein Operations) es una librería para la manipulación flexible y legible de tensores multidimensionales. Permite realizar operaciones comunes de reestructuración y transformación de tensores de manera concisa y comprensible.
- Transformers es una librería desarrollada por Hugging Face que proporciona herramientas y modelos de última generación para tareas de procesamiento del lenguaje natural (NLP), como clasificación de texto, generación de texto, traducción y más.
- Torch: es la librería principal de PyTorch, un framework de aprendizaje profundo que proporciona soporte para tensores multidimensionales y cálculo automático de gradientes.
- Bitsandbytes: es una librería que permite la optimización del entrenamiento de modelos de lenguaje utilizando precisiones reducidas, como 8 bits y 4 bits, para acelerar el entrenamiento y reducir el uso de memoria.
- Accelerate: es una librería de Hugging Face que facilita el uso de aceleradores de hardware (como GPU) para el entrenamiento distribuido y la inferencia de modelos de aprendizaje profundo.

- Xformers: es una librería que proporciona implementaciones eficientes y escalables de transformadores, optimizados para diferentes tipos de hardware y aplicaciones.

Estas librerías permiten configurar, entrenar y ejecutar modelos de lenguaje de manera eficiente, aprovechando optimizaciones y recursos de hardware disponibles para manejar el modelo.

- LangChain: es una biblioteca que se utiliza para proporcionar contexto al modelo de lenguaje mediante la inclusión de fragmentos relevantes de texto relacionados con la consulta del usuario. Esto es esencial para mejorar la comprensión y la relevancia de las respuestas generadas por el modelo, ya que permite que el modelo tenga en cuenta el contexto específico en el que se hace la consulta.
- InstructorEmbedding: es un tipo de embedding diseñado específicamente para la comprensión de instrucciones o comandos. Se utiliza para representar palabras y frases de manera numérica en un espacio vectorial, lo que facilita la comprensión y el procesamiento de las instrucciones dadas por el usuario. Este tipo de embedding es crucial para garantizar que el modelo pueda entender las consultas del usuario y generar respuestas relevantes y precisas.
- Sentence transformers: es una biblioteca utilizada para generar embeddings de oraciones o frases completas. Estos embeddings son útiles para comparar la similitud semántica entre diferentes oraciones y para realizar tareas de recuperación de información basadas en la similitud de texto. En el contexto de esta parte, se pueden utilizar para mejorar la comprensión del contexto y la relevancia de las respuestas generadas por el modelo de lenguaje.

Estas librerías son necesarias para proporcionar contexto al modelo de lenguaje, representar instrucciones de manera numérica y comparar la similitud semántica entre oraciones, lo que en conjunto contribuye a mejorar la calidad y la relevancia de las respuestas generadas por el modelo.

## Resultados

El chatbot se sometió a pruebas utilizando una serie de instrucciones específicas relacionadas con información periodística en español. Para proporcionar contexto al modelo de lenguaje, se utilizaron al azar cuerpos de noticias seleccionadas del dataset. Estos cuerpos de noticias se incluyeron como contexto para cada consulta realizada al chatbot.

Además, como parte de la evaluación, se elaboraron tres preguntas relacionadas con el tema de cada noticia. Estas preguntas fueron diseñadas para evaluar la capacidad del chatbot para comprender y responder adecuadamente a consultas específicas sobre la información presentada en las noticias.

Para comparar los resultados obtenidos con el chatbot, se realizaron las mismas tres preguntas al modelo de lenguaje GPT, sin proporcionar ningún contexto adicional. Esto permitió evaluar la coherencia y relevancia de las respuestas generadas por el chatbot en comparación con las respuestas generadas por el modelo de lenguaje GPT.

### Tema 1: Adaptador de medios de Nintendo DS

Pregunta 1: ¿Cuál es el propósito del adaptador de medios lanzado por Nintendo para la consola portátil DS?

- Respuesta Águila-7B: 7/10

*Se trata de un pequeño accesorio, que incluye un adaptador inalámbrico Bluetooth y un receptor FM, que permitirá al usuario poder escuchar música mientras juega en la consola portátil Nintendo DS, sin necesidad de tener que adquirir otro accesorio.*

Correcta pero falta de detalles sobre la reproducción de vídeos y programas de TV.

- Respuesta Chat GPT: 9/10

*El adaptador de medios permite a los usuarios reproducir música, vídeo y programas de TV descargados en la consola DS mientras están en movimiento.*

Correcta y más completa, incluye reproducción de música, vídeo y programas de TV.

Pregunta 2: ¿Qué características distintivas tiene la consola Nintendo DS en comparación con su predecesora, la Gameboy Advance?

- Respuesta Águila-7B: 8/10

*La consola de nueva generación de Nintendo es más potente que su predecesora con un procesador de gráficos más rápido, así como funciones de conectividad inalámbrica y un diseño revisado*

Correcta en términos de poder y conectividad inalámbrica, pero omite detalles clave como las pantallas duales.

- Respuesta Chat GPT: 10/10

*La Nintendo DS tiene dos pantallas, una de las cuales es sensible al tacto, y cuenta con una conexión inalámbrica que permite a los usuarios jugar entre sí. Además, la DS fue lanzada como sucesora de la exitosa Gameboy Advance.*

Completa y detallada, incluye información sobre las pantallas duales, pantalla táctil y conexión inalámbrica.

Pregunta 3: ¿Cuál es la estrategia de Nintendo ante la competencia de la PSP de Sony?

- Respuesta Águila-7B: 7/10

*La PSP (PlayStation Portable) no es el primer gran desafío para Nintendo DS, ya que hasta ahora la consola ha vendido un total de 37,6 millones de unidades a nivel mundial y sigue siendo la consola más vendida de Japón. Nintendo DS, por su parte, ha vendido 34,47 millones de unidades alrededor del mundo. Desde su salida se le considera una máquina de ventas sólidas, aunque en los primeros meses de 2007 se había registrado una caída en sus ventas, mientras que la PSP parecía ganar terreno en el mercado nipón. Con el lanzamiento de la PlayStation Portable, en 2004, DS se convirtió en el primer dispositivo electrónico en ventas fuera de Japón, y desde 2005 ha vendido más de 36 millones de unidades en el mercado norteamericano.*

Correcta pero se enfoca más en las cifras de ventas y menos en la estrategia de productos.

- Respuesta Chat GPT: 9/10

*Nintendo está lanzando el adaptador de medios como respuesta a la PSP de Sony, que está diseñada como un dispositivo multiusos para medios y juegos. Nintendo busca ampliar el atractivo de su dispositivo y proteger su dominio en el mercado de las consolas portátiles de juegos.*

Correcta y relevante, menciona la estrategia de lanzar el adaptador de medios para competir con las capacidades multimedia de la PSP.

## Tema 2: Etiquetas RFID en las tiendas

Pregunta 1: ¿Cuáles son las principales preocupaciones de los consumidores respecto al uso de etiquetas RFID en las tiendas?

- Respuesta Águila-7B: 4/10

*Ahorre costos: La tecnología es relativamente barato que el proceso de la compra en sí. los costos de etiquetas RFID para el minorista han superado el 20 por ciento de los gastos en los EE.UU. pero a medida que las etiquetas rfid se expanden y mejora su eficiencia son un factor que impulsa el mercado. los comerciantes que tienen un producto que no es capaz de cumplir con el requisito de rfid tecnología son menos propensos a comprarlo que un comerciante que tiene una etiqueta de tecnología rfid que ofrece un alto valor de la etiqueta.*

Principalmente sobre costos y eficiencia, omite preocupaciones de privacidad.

- Respuesta Chat GPT: 10/10

*Los consumidores están preocupados por la posible invasión de la privacidad y el seguimiento de sus compras una vez que han salido de las tiendas.*

Correcta, menciona la invasión de privacidad y seguimiento de compras.

Pregunta 2: ¿Qué porcentaje de los encuestados expresaron preocupación por la posibilidad de que las etiquetas RFID permitan el seguimiento de los consumidores?

- Respuesta Águila-7B: 9/10

*Más de la mitad (55%) de los encuestados dijo que estaban preocupados o muy preocupados de que las etiquetas RFID permitiría que las empresas para el seguimiento de los consumidores a través de la compra de productos. en respuesta a esto la encuesta fue publicada a principios de noviembre de 2011, al tiempo que las preocupaciones de las personas*

*acerca de la identidad y la privacidad de las etiquetas de radio frecuencia de identificación (rfid) se informó por primera vez.*

Correcta y precisa, menciona el 55%.

- Respuesta Chat GPT: 10/10

*El 55% de los encuestados manifestaron preocupación de que las etiquetas RFID permitieran a las empresas rastrear a los consumidores a través de sus compras.*

Correcta y precisa, menciona el 55%.

Pregunta 3: ¿Qué áreas de aplicación se mencionan en la noticia donde las etiquetas RFID están actualmente en uso?

- Respuesta Águila-7B: 7/10

*Las etiquetas se utilizan en grandes almacenes, farmacias, tiendas de mascotas y cadenas de supermercados, y a veces para recoger algunos artículos de la tienda en el hogar y entregarlos a un centro de entrega de casa.*

Correcta pero limitada, menciona tiendas generales.

- Respuesta Chat GPT: 9/10

*Las etiquetas RFID están siendo utilizadas en un centro de distribución de Tesco en el Reino Unido para realizar inventarios rápidos, así como en el sistema de peaje M6 en las Midlands del Reino Unido como ficha de acceso.*

Correcta y más específica, menciona Tesco y sistema de peaje M6.

Tema 3: Telescopio espacial James Webb

Pregunta 1: ¿Qué récord ha batido el telescopio espacial James Webb?

- Respuesta Águila-7B: 6/10

*Al analizar la luz del universo más joven, el telescopio espacial James Webb ha logrado trazar un mapa del universo muy detallado y detallado. Gracias a este mapa los astrónomos podrán entender mejor que lo que se creía hasta ahora.*

Correcta pero general, falta especificar el récord.

- Respuesta Chat GPT: 10/10

*El James Webb ha batido un nuevo récord al observar la galaxia más lejana, formada tan sólo 290 millones de años después del Big Bang*

Correcta y específica, menciona la observación de la galaxia más lejana.



Pregunta 2: ¿Cómo se llama la galaxia más antigua observada hasta ahora?

- Respuesta Águila-7B: 8/10

*La galaxia más antigua, y la primera que se ha detectado con claridad, se llama JADES-GS-z14-0. Se ha descubierto en el cúmulo de galaxias Abell 1147. El cúmulo de galaxias Abell 1147 se encuentra a más de 2,2 billones de años luz de la Tierra.*

Correcta pero con información adicional no necesaria sobre Abell 1147.

- Respuesta Chat GPT: 10/10

*La nueva galaxia, la más antigua por tanto descubierta, ha sido denominada JADES-GS-z14-0.*

Correcta y precisa, menciona JADES-GS-z14-0.

Pregunta 3: ¿Qué instrumento del telescopio James Webb se utilizó durante una campaña de observaciones en octubre de 2023 y enero de 2024?

- Respuesta Águila-7B: 8/10

*La cámara del telescopio de infrarrojo cercano James Webb (NIRSpec)*

Correcta, menciona NIRSpec pero falta el contexto completo.

- Respuesta Chat GPT: 10/10

*Durante esos periodos, se utilizó el espectrógrafo en infrarrojo NIRSpec para 'viajar en el tiempo' y observar galaxias formadas entre los 200 y 90 millones de años después del Big Bang.*

Correcta y completa, menciona NIRSpec y el contexto de las observaciones.

## Gráficos

Utilizando las puntuaciones asignadas, se realizó una serie de gráficos para comparar las respuestas.

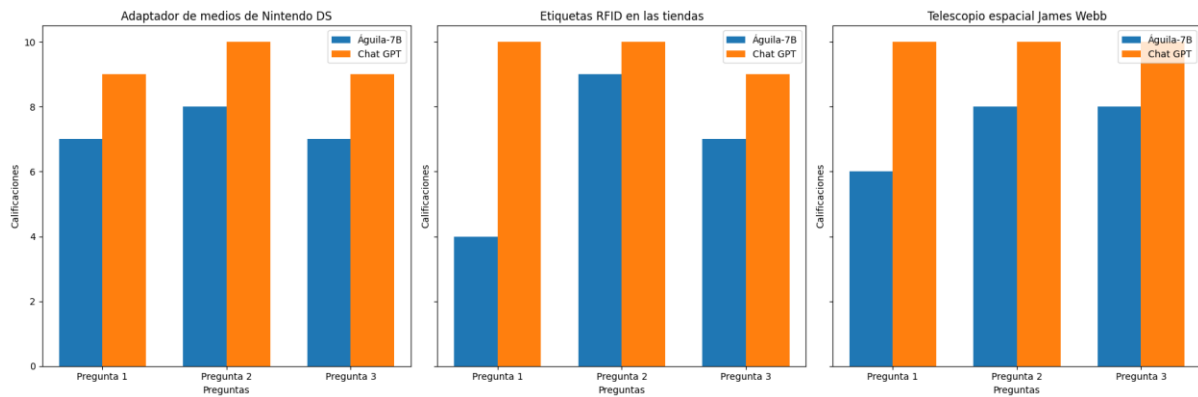


Figura 1 Gráfico de barras de puntuaciones. Águila 7B vs Chat GPT

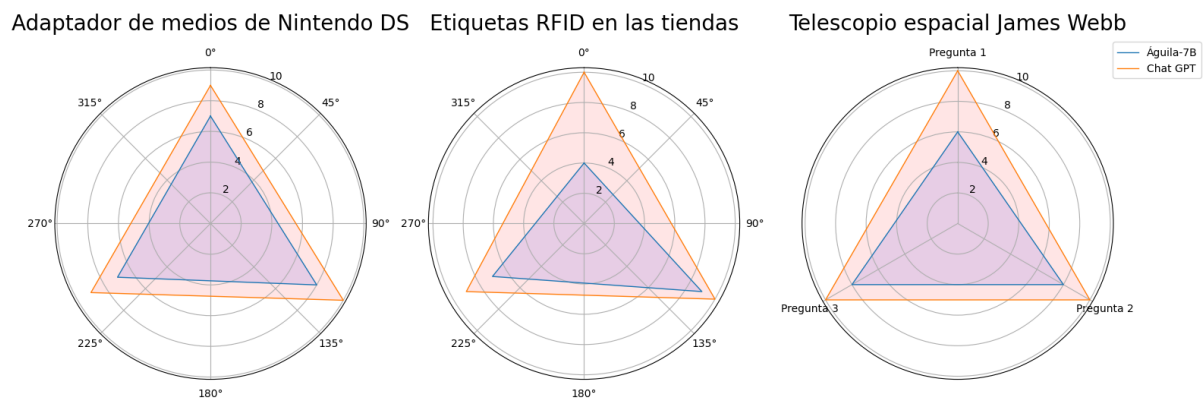


Figura 2 Comparación de calificaciones de las respuestas a las preguntas, Águila 7B vs Chat GPT

Estos gráficos muestran las diferencias en las calificaciones de las respuestas proporcionadas por Águila-7B y Chat GPT para cada pregunta en los tres temas. Las puntuaciones indican que, en general, Chat GPT proporciona respuestas más completas y detalladas en comparación con Águila-7B.

## **Conclusiones**

El chatbot basado en el modelo Águila-7B demostró ser una herramienta efectiva para generar respuestas coherentes y relevantes a partir de instrucciones en español. La utilización de técnicas de ajuste eficiente de parámetros y cuantización permitió un despliegue eficiente en entornos con recursos limitados. Este trabajo inicial resalta la viabilidad de desarrollar un presentador de noticias virtual, capaz de interpretar y responder preguntas sobre noticias y temas actuales.

El desarrollo y afinación del modelo evidenciaron la importancia de contar con un conjunto de datos de calidad y un ajuste preciso para mejorar la interacción en lenguajes específicos. Los resultados obtenidos son prometedores y abren la puerta a futuras mejoras y aplicaciones en diversos ámbitos.

En futuras iteraciones, se podría explorar la integración de fuentes de datos en tiempo real para enriquecer las respuestas del chatbot, así como la optimización adicional del modelo para mejorar su desempeño en dispositivos con aún menos capacidad de procesamiento.