Finetune a un LLM para la creación de chats de apoyo al aprendizaje. Caso de estudio para la Ingeniería de Software

Juan José López Gómez, Francisco José García Peñalvo, and Alicia García Holgado

Universidad de Salamanca
Departamento de Informática y Automática
Facultad de Ciencias.
Plaza de los Caídos s/n
37008 Salamanca, España
{juanjoselopez,fgarcia,aliciagh}@usal.es

Resumen El avance continuo en la inteligencia artificial ha provocado un aumento de la popularidad y usos en distintas disciplinas de los modelos de lenguaje a gran escala, ofreciendo oportunidades sin precedentes en la educación mediante desarrollo de Chatbots debido a la flexibilidad y facilidad de interacción mediante lenguaje natural. En la actualidad el desarrollo de modelos de lenguaje a gran escala está liderado por la empresa OpenAI con sus modelos ChatGPT 3.5, ChatGPT 4 y el reciente ChatGPT-4o; otras empresas están tratando de desarrollar sus modelos de código libre para poder competir con los modelos privados de OpenAI. Haciendo uso de estos modelos de código libre esta investigación pretende abordar el desarrollo, y la comparación con los modelos privados mencionados, de un modelo de lenguaje a gran escala de código libre que será especializado mediante un proceso de finetuning realizado con un conjunto de datos, o corpus, desarrollado tanto en inglés como en español para poder observar cómo varía el comportamiento enfocado en la disciplina de la Ingeniería de Software. Los modelos de lenguaje a gran escala desarrollados, y el corpus creado para la especialización, como resultado de la investigación cumplen la función de poderse utilizar como herramientas de apoyo al aprendizaje para los estudiantes, pero la conclusión que se obtiene es que el proceso de especialización que se lleve a cabo no es viable si no se tiene un conjunto de datos con la calidad suficiente y el hardware necesario para ello no se van a poder obtener resultados como los que ofrecen los modelos de OpenAI tanto en su versión gratuita como en su versión de pago para ser utilizados como herramientas de apoyo al estudiante en la disciplina de la Ingeniería de Software.

Keywords: Modelos de lenguaje a gran escala, especialización, educación, Ingeniería de Software, código libre, corpus.

1. Introducción

En los recientes años la inteligencia artificial (IA) ha cobrado un papel muy importante en la sociedad. En especial desde 2022 un enfoque en específico ha tenido un mayor desarrollo, la inteligencia artificial generativa que se define como la producción de contenidos sintéticos desconocidos, en cualquier formato y para cualquier tarea, mediante modelos generativos preentrenados [1].

Esta tecnología tiene la capacidad de generar distintos tipos de información a partir de una entrada dada dependiendo del modelo que se utilice, el mayor uso que se les ha dado es como asistentes conversacionales (ChatGPT), pero también se pueden generar imágenes a partir de texto (DALL-E 2), generar código fuente (Copilot) o ser modelos multimodales que permiten recibir distintos tipos de entradas para generar salidas de distintos tipos (ChatGPT 4, ChatGPT 4o, Kosmos-1) [2]. Esta flexibilidad ha conseguido que puedan ser utilizados en múltiples disciplinas, especialmente en medicina o ciberseguridad pero también para tareas de *Natural Language Processing* o generación de arte [1].

El uso de la IA en la educación es un recurso que ha sido ampliamente estudiado como presenta Wang et al., 2024 [3] en su revisión sistemática en al que muestra como en la última década se han presentado artículos para utilizar distintos tipos de IA (*Machine Learning*, *Deep Learning* o Chatbots) en distintos niveles educativos y asignaturas relacionadas con STEM (*Science, Technology, Engineering and Mathematics*). A pesar de los beneficios que se puedan observar la implementación de estas tecnologías es un proceso lento debido a que se enfrenta a distintas barreras como se definen en el artículo de Wang y Cheng, 2021 [4], una primera barrera centrada en factores estructurales, de recursos o políticos que no están bajo el control del personal (incertidumbre en la compra de material o de integración) docente mientras que las barreras de segundo orden se refieren a los obstáculos que aportan los docentes a estas tecnologías (inseguridad ante el valor pedagógico de las herramientas, falta de interés por parte de los docentes o rechazo por desconocimiento de las tecnologías).

La IA generativa ha provocado una serie de problemas a nivel ético y moral sobre la procedencia y los métodos de obtención de los datos utilizados para realizar el preentrenamiento a los modelos, esta controversia afecto al asistente de programación desarrollado por GitHub llamado Copilot, cuando se comprobó que generaba código almacenado en repositorios privados o bajo licencia de autor aunque la empresa negaba que pudiese hacerlo [5], del mismo modo Meta ha cambiado su política de empresa informando de que los datos públicos de los usuarios van a poder ser utilizados para entrenar inteligencias artificiales [6]. Junto a la problemática presentada surge la posibilidad de filtrar información debido a que la inferencia que realicen los usuarios con los modelos se puede utilizar como método de entrenamiento automático [7] además la información generada es mediante procesos matemáticos sin que los modelos entiendan el significado de lo generado cabiendo la posibilidad de generar información falso o errónea, por ello se ha regularizado que la última decisión que se tome ante cualquier información generada mediante inteligencia artificial sea tomada por una persona [8].

La implementación más conocida de la inteligencia artifical generativa son los *Large Language Models* (LLM) que trabajan sobre el paradigma del NLP (*Natural Language Processing*) junto a redes neuronales permitiendo que puedan interpretar, manipular, comprender y generar conocimiento a través del lenguaje natural. Está tecnología tiene un gran potencial en el entorno educativo [9] ofreciendo nuevas formas de enseñanzas y aprendizaje, permitiendo personalizar el ritmpo de aprendizaje, generando nuevos contenidos educativos [10], mejorando las metodologías educativas centradas en el

estudiante [11].

Por los beneficios que se han presentado, múltiples empresas han desarrollado su implementación de los LLM en especial la empresa OpenAI, con sus modelos ChatGPT 3.5 [12], ChatGPT 4 [13] o el reciente ChatGPT 4o, debido al ecosistema que han construido alrededor y la facilidad de uso aunque alguna de esas funcionalidades y modelos sean de pago. El movimiento del *open source* ha tratado de desarrollar modelos de código libre que consigan hacerle frente a los desarrollados por OpenAI, como las distintas versiones de Llama desarrollados por Meta o Mistral desarrollado por Mistral.AI; por este hecho el objetivo principal que pretende abordar la investigación es el desarrollo de un LLM *open source* especializado en la disciplina de la Ingeniería de Software mediante un proceso denominado *finetuning*. Este proceso necesita un conjunto de datos para poder ser realizado por lo que se propone como objetivo específico el desarrollo de un corpus multilenguaje (castellano e inglés) con el que se pueda realizar este proceso, además se propone un estudio comparativo entre los modelos *open source* desarrollados frente a los modelos de Open AI con el fin de obtener conclusiones de las diferencias entre ambos tipos de modelos.

El presente trabajo se organiza en seis apartados. El siguiente apartado (Sec. 9) presenta los conceptos teóricos que es necesario entender antes de realizar la investigación. El segundo apartado (Sec. 9) presenta la revisión del estado de la cuestión que se ha desarrollado para la investigación, el siguiente apartado (Sec. 9) presenta la metodología de estudio seguida en la investigación. La Sec. 9 presenta los resultados de la investigación mientras que la Sec. 9 presenta la discusión sobre estos. Finalmente, el último apartado resume las principales conclusiones y líneas de trabajo futuras.

2. Conceptos y Técnicas

2.1 Large Language Models

Los *Large Language Models* son la implementación de las redes neuronales artificiales para tareas complejas de NLP (*Natural Language Processing*) que han sido entrenados con una gran cantidad de datos que les permite entender y generar lenguaje natural mediante una serie de técnicas que se aplican de forma secuencial [14], [15].

En la Fig. 1 se muestra un esquema del funcionamiento de un LLM, que al tratarse de una red neuronal están basados en una sucesión de capas con distintos números de parámetros y funciones de activación implementando la arquitectura *Transformers* presentada por Google en 2017 [17].

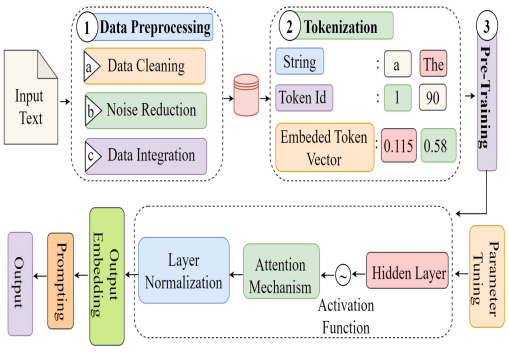


Figura 1: Esquema de un LLM [16]

Esta arquitectura permite tomar entradas de forma textual que sufrirán un proceso de tokenización, división de la entrada en las unidades más simples posibles denominadas tokens [18] dichos tokens pueden ser caracteres, sílabas, palabras en función del tamaño y el tipo del modelo.

Permitiendo que los modelos tengan una gran capacidad de adaptación. Junto a la arquitectura también se presentó el mecanismo de atención que permite realizar una representación de las secuencias de las entradas relacionando las diferentes posiciones que pueden ocupar los tokens de la entrada al modelo. La arquitectura de un *Transformer* se presenta en la Fig. 2, cabe destacar los siguientes componentes:

- *Positional Encodings*: Permiten a la arquitectura tener consciencia sobre la posición de los tokens una vez se han codificado. Este proceso permite entender y generar una respuesta que sea semánticamente correcta.
- *Inputs*: Tras la tokenización de la entrada al modelo, se asigna un identificador único a cada token, añadiéndose a un vector que servirá para entender el significado de las palabras.
- *Outputs*: Mediante un mecanismo de aprendizaje autorregresivo, se aprende a a precedir la siguiente palabra en al secuencia al analizar las que le preceden, permitiendo la generación de las secuencias de forma coherente y siguiendo el contexto de la entrada al modelo.

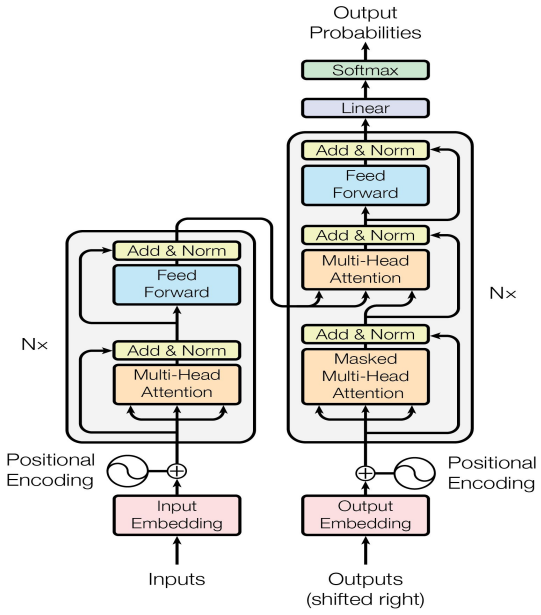


Figura 2: Arquitectura de un *Transformer* [17]

A raíz de que Google presentará esta arquitectura se han desarrollado muchos LLM tanto privados como los ofrecidos por OpenAI (ChatGPT) o por Anthropic (Claude3), como *open source* como son los desarrollados por Meta como Llama2 o Llama3 o los desarrollados por Mistral.ai como es Mistral para poder hacerle frente a los modelos privados.

2.1.1 Llama2 Desarrollado por Meta y lanzado en julio de 2023 [19] cuenta con versiones de diferentes número de parámetros desde 7B hasta 70B (miles de millones). Estos modelos fueron desarrollados para ser un competidor principal de los modelos privados más populares. En la Fig. 3 se muestra la comparación que realizó la empresa desarrollador frente a otros LLM y se observa como se comporta con bastante similitud a los modelos privados mientras que obtenía un mejor resultado para los modelos *open source*.

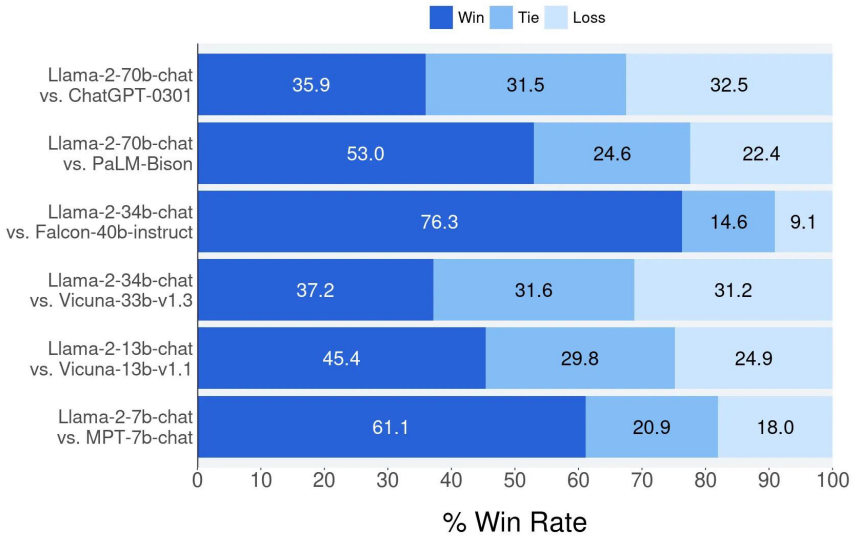


Figura 3: Evaluación de la utilidad del Llama2 [19]

2.1.2 Mistral En septiembre de 2023, Mistral.AI lanzaba su modelo llamado Mistral [20] que fue comparado con el modelo desarrollado por Meta, Llama2 como se puede observar en la Fig. 4.

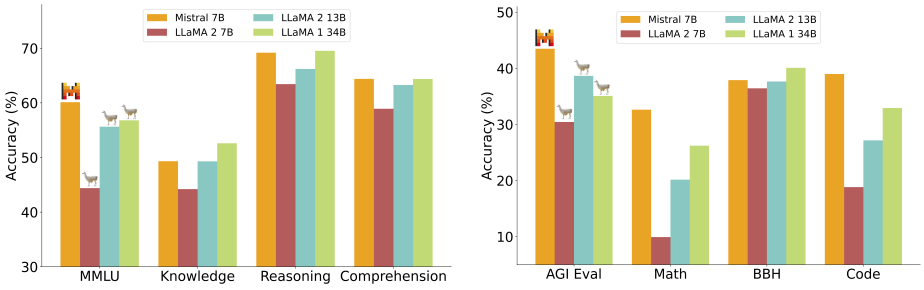


Figura 4: Comparaciones entre Mistral y Llama [21]

Obteniendo mejores resultados en las pruebas que los modelos de Llama que se presentan en la imagen.

2.2 Finetune

Esta técnica es muy importante en el campo de la inteligencia artificial debido a la capacidad de adaptación que ofrece a los LLM.

Como ya se ha presentado los LLM se basan en la arquitectura *Transformers* basados en un mecanismo de atención que permite modelar de forma conceptual las relaciones entre las palabras de un texto.

Para ello se tienen que modificar los pesos de las capas que están inicialmente adaptados para capturar patrones y terminologías generales, mediante este proceso de especialización se refinan las capas para que se pueden detectar patrones más específicos, minimizando errores en el texto generado.

3. Antecedentes

Debido al aumento de la popularidad de la inteligencia artificial y los LLM en los últimos años, es conveniente tener una visión de los artículos publicados que hacen referencia al uso de estas tecnologías en la educación. Para ello se ha realizado una revisión de la literatura en la que se propusieron las siguientes preguntas de investigación:

- **RQ1.** ¿Cuál es la viabilidad de utilizar los LLM para el aprendizaje autónomo de un ámbito en concreto?
- **RQ2.** ¿Cuáles son los beneficios de utilizar los LLM en la educación?

En los artículos se ha encontrado un mayoritario uso de los LLM en la educación, pero el uso de métodos basados en *Machine Learning* o *Deep Learning* como se menciona en el artículo de Wang et al., 2024 [3] entre otros.

Los estudios abarcan una gran variedad de beneficios y usos en la educación, algunos de los estudios como el de Jeon y Lee, 2023 [22] está centrado en como pueden ser utilizado para mejorar las labores pedagógicas de los docentes, otros se centran en la capacidad de automatizar procesos y mejorar la calidad del material que se utilice como indican Huang et al., 2023 [10], Pradhan et al., 2024 [23] o Li et al., 2024 [24].

Otros estudios se centran en como afectan estas tecnologías a los estudiantes, por ejemplo Bernabei et al., 2023 [25] o Huang et al., 2023 [10] indican en sus estudios que el uso de los LLM beneficia en la capacidad de aprendizaje y de retención de los conocimientos por parte de los estudiantes, otros inciden en la mejora en el rendimiento escolar como Chang et al., [26] 2023 o Al Shloul et al., 2024 [27]. Sin embargo Crawford et al., 2024 [28] en su artículo obtiene la conclusión de que tener una dependencia hacia la inteligencia artificial produce una disminución del rendimiento escolar y una mayor tendencia al abandono escolar en los universtarios, también Cross et al., 2023 [29] menciona lo ya comentado de la posibilidad de la generación de información errónea y los problemas que puede acarrear si no se le aplica un proceso crítico sobre la salida de los modelos.

Debido al cambio de paradigma en la educación, implicando más al alumno en las actividades y en el proceso de aprendizaje respecto a la forma tradicional, se ha observado que los artículos como el de Ng et al. [11] mencionan como el uso de los Chatbots basados en LLM pueden tener una mejora en las metodologías como el *Self-Regulated Learning* [30], debido a que ofrecen una gran flexibilidad y adaptabilidad respecto al estudiante, mientras que por ejemplo Lin y Chang, 2023 [31] se centran en como pueden mejorar el *Active Learning* [32] favoreciendo la participación mediante la exploración y la colaboración mediante los LLM.

4. Metodología

4.1 Elección de los Modelos

Los LLM que se utilicen tienen que cumplir los siguiente requisitos:

- *Open source*
- Versión preentrenada para actuar como un Chatbot
- Versión computacionalmente viable para la investigación en un entorno doméstico.

La búsqueda fue realizada en la plataforma de Chatbot Arena [33], debido a que tiene un ranking de evaluación de los LLM mediante un sistema de votación a ciegas entre una gran variedad de modelos tanto privados como de código libre.

Tras el análisis y las pruebas correspondientes los modelos seleccionados fueron Mistral 7B [34] y Llama2 7B [35] ambos en sus versiones optimizadas para actuar como un Chatbot.

4.2 Creación del Corpus

El proceso de *finetuning* necesita una gran cantidad de datos con una sintaxis específica, la escogida es la definida Huggingface para ambos modelos en la documentación de sus librerías [36]:

```
<s>[INST]<<SYS>>System Prompt<SYS>> User Question [/INST] Assistant Response </s>
```

Las fuentes de las que se han extraído la información son repositorios de información en español públicos que tratan sobre la Ingenería del Software [37], [38] a los que se les ha realizado un procesamiento de los datos mediante la aplicación de herramientas de minería de datos y un posterior adecuación a la sintaxis que se ha específicados haciendo uso de herramientas de inteligencia artifical generativas como ChatGPT 4. Para poder realizar pruebas con distintos contextos se ha creado el corpus tanto en castellano como en inglés.

4.3 Proceso de Especialización

La especialización o también denominada como el proceso de *finetuning* se realiza mediante el corpus construido utilizando las abstracciones que provee HuggingFace de los *Transformers* ajustando una serie de parámetros como el *Batch Size* o el *Learning rate* entre otros.

Debido a que los LLM son redes neuronales con una gran cantidad de parámetros que no pueden ser entrenados de forma completa en ordenadores domésticos se necesitan utilizar técnicas de optimización computacional, conocidas como LoRA (*Low Rank Adaptation*) cuyo objetivo es modificar los parámetros del modelo sin necesidad de ajustarlos todos a la vez, entrenando unos pocos mientras que se congelan el resto de parámetros obteniendo una menor carga computacional sin afectar a las métricas [39] o QLoRA (*Quantum Low Rank Adaptation* que tiene el mismo funcionamiento y objetivo pero aplica una cuantización a los parámetros (reducir su precisión) mejorando el rendimiento a costa de un impacto negativo en las métricas.

En la investigación se utilizará LoRA ya que se ha decidido primar la obtención de mejores métricas a la rápidez en el entrenamiento.

4.4 Proceso de Evaluación

En toda investigación es necesario evaluar los resultados para ello se proponen dos métodos de evaluación:

- **Evaluación cuantitativa**: clasificación basada en métodos más científicos dejando de lado el análisis de la calidad de las respuestas.
  - Análisis de coherencia: Utilizando BERT (*Bidirectional Encoder Representations from Transformers*), debido a que tiene la capacidad de entender el contexto de una palabra tanto por las palabras que la predecen como por las que le siguen mediante *Masked language Model* que reemplaza algunas palabras por un token con el objetivo de predecirlas en base al contexto, se puede llegar a analizar la coherencia de un texto mediante la función de pérdida de este modelo [21].
- **Evaluación humana**: clasificación numérica de una serie de métricas por parte de un ser humano. Las métricas que se van a evaluar son las siguientes:
  - Veracidad: La información exacta, verdadera y contenidos de calidad.
  - Relevancia: La respuesta debe abordar correctamente la pregunta en el contexto propuesto.
  - Claridad: La redacción debe ser clara y comprensible.
  - Fluidez: El texto debe leerse naturalmente, como si fuese escrito por un ser humano.

Estos métodos de evaluación serán utilizados a las respuestas generadas en las siguientes pruebas de evaluación:

- ***Zero Shot***: Realizar la inferencia a los modelos base sin haberles realizado ningún tipo de entrenamiento previo para poder ver el conocimiento base sobre la Ingeniería del Software con el que han sido preentrenados [41].
- ***Finetune***: Realizar la inferencia a los modelos después de realizarles la especialización sobre la disciplina elegida para poder comparar resultados.

5. Resultados

5.1 Entrenamiento de los Modelos

Esta especialización se le va a realizar a los modelos Llama2 7B Chat [35] y Mistral 7B instruct [34], en el mismo entorno y los mismos parámetros del entrenamiento. Para evaluar la calidad de cada uno los modelos se hará uso de dos métricas:

- **Pérdida de entrenamiento**: Diferencia entre las predicciones del modelo y los valores verdaderos, utilizada para ajustar los parámetros mediante un proceso de optimización.
- **Pérdida de evaluación**: Obtenida al aplicar una función de pérdida sobre los datos de evaluación que tiene como propósito proporcionar una estimación imparcial. Si aumente o se estanca puede deberse a un sobreajuste.

Tras realizar pruebas con valores en los parámetros para poder obtener una visión del comportamiento de los modelos, se observó que las pruebas que fueron realizadas con la distribución de los datos 90 % para entrenamiento y 10 % para evaluación y en castellano obtuvieron mejores valores en las métricas que se han utilizado para evaluar. Acorde con esto en la Tab. 1 se presentan los resultados que han obtenido los modelos en

Modelo	**Épocas**	***Batch Size***	**Pérdida de entrenamiento**	**Pérdida de evaluación**	**Tiempo de entrenamiento**
Llama2	50	3	0.0751	1.812	10h 16m 1s
Mistral	50	3	0.0747	3.737	10h 41m 3s

Tabla 1: Resultados de los modelos finales

las métricas cuando se han sometido al proceso de *finetuning* con unos parámetros reales.

Ambos modelos entrenados han obtenido unos valores bajos en la pérdida de entrenamiento ya que es muy cercana a cero, mientras que en la pérdida de evaluación siguen el mismo comportamiento que cuando se realizaron las pruebas piloto, el aumento exponencial de esta métrica y dispar entre cada uno de los modelos.

5.2 Evaluación de los Modelos

Para la evaluación de los modelos se necesitan crear unos datos de prueba, en este caso se ha construido una batería de preguntas en inglés y castellano.

```
{
 "test1":
  [
   {
     "role": "system", "content": "system prompt"
   },
   {
     "role": "user", "content": "user question"
   }
  ],
}
```

Figura 5: Ejemplo de estructura de los datos de prueba (1)

En la Fig. 5 se representa una conversación nueva con el asistente sin haberle dado ningún tipo de contexto previo, mientras que en la Fig. 6 se muestra la estructura cuando ya se le ha realizado una pregunta generándole un contexto previo para responder la pregunta del usuario de forma más acertada.

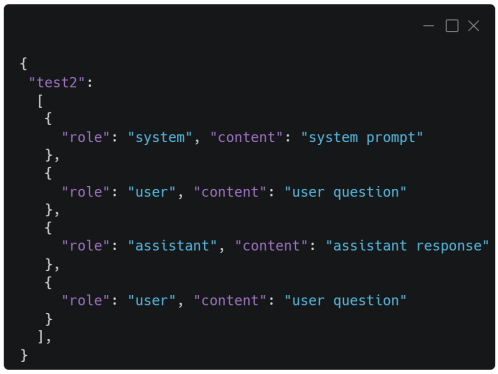


Figura 6: Ejemplo de estructura de los datos de prueba (2)

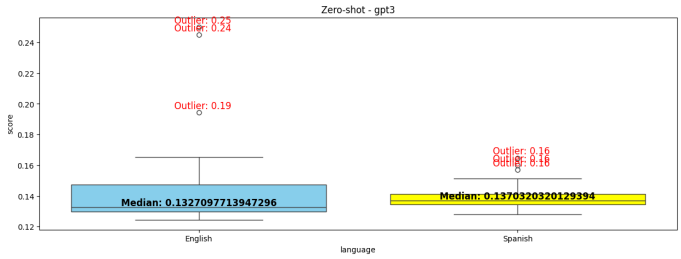
5.2.1 Prueba de Evaluación: Zero Shot Esta prueba se va a realizar sobre los modelos base de Mistral 7B y Llama2 7B junto a ChatGPT 3.5 y ChatGPT 4. La finalidad de esta prueba es comparar el comportamiento de los modelos cuando trabajan en el mismo ambiente sin haber sido alterados ni especializados de forma alguna.

Evaluación cuantitativa En la Tab. 2 se presenta el número medio de caracteres que han generado los modelos en base a la batería de preguntas.

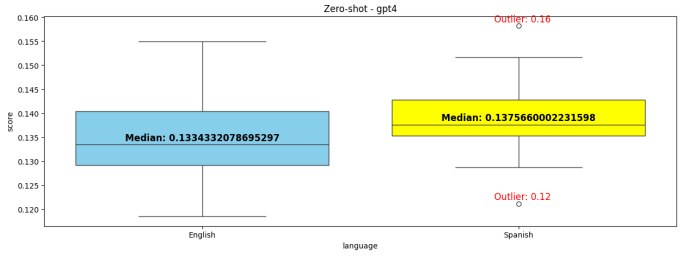
Modelo	Castellano	Inglés
ChatGPT 3.5	2045.8	2338.72
ChatGPT 4	2143.12	2410.96
Llama2 7B	2863.56	3075.8
Mistral 7B	1847.76	2043.44

Tabla 2: Número medio de caracteres generados en la prueba de Zero Shot

Como parte de la evaluación cuantitativa se ha realizado un análisis de coherencia mediante BERT [21] que se puede ver en las Figs. 7 y 8.

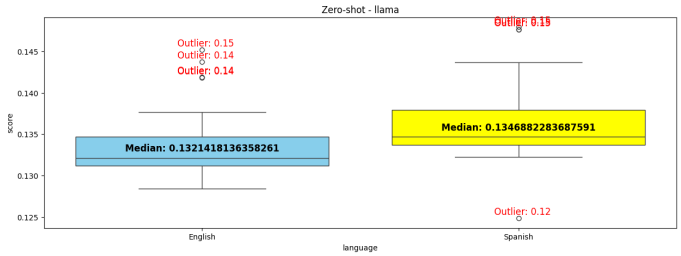


(a) ChatGTP 3.5

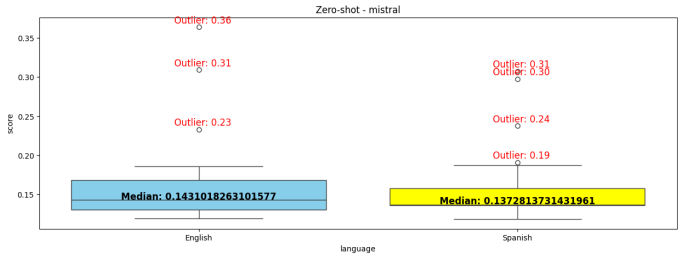


(b) ChatGTP 4

Figura 7: Coherencia obtenida en la prueba de evaluación: Zero Shot (1)



(a) Llama2



(b) Mistral

Figura 8: Coherencia obtenida en la prueba de evaluación: Zero Shot (2)

Evaluación humana En la Tab. 3 se muestran las métricas que ha obtenido en la evaluación humana el modelo de Llama2 en la prueba de evaluación de *Zero Shot*.

Métrica	**Castellano**	**Inglés**	**Total**
Veracidad	7.36	9.00	8.18
Relevancia	8.12	9.64	8.87
Claridad	7.48	9.00	8.24
Fluidez	7.28	9.36	8.32

Tabla 3: Evaluación humana - Llama2 (Zero Shot)

En la Tab. 4 se muestran las métricas que ha obtenido en la evaluación humana el modelo de Mistral en la prueba de evaluación de *Zero Shot*.

Métrica	Castellano	Inglés	Total
Veracidad	4.60	8.04	6.32
Relevancia	5.56	9.72	7.64
Claridad	5.04	9.08	7.06
Fluidez	4.96	8.84	6.90

Tabla 4: Evaluación humana - Mistral (Zero Shot)

En la Tab. 5 se muestran las métricas que ha obtenido en la evaluación humana el modelo de ChatGPT 3.5 en la prueba de evaluación de *Zero Shot*.

Métrica	Castellano	Inglés	Total
Veracidad	8.40	8.40	8.40
Relevancia	9.12	9.12	9.12
Claridad	8.80	8.80	8.80
Fluidez	8.80	8.80	8.80

Tabla 5: Evaluación humana - GPT 3.5 (Zero Shot)

5.2.2 Prueba de Evaluación: Finetune Esta prueba ha realizado sobre los modelos de Llama2 7B, Mistral 7b y ChatGPT 3.5 Turbo tras realizarles el *finetuning* con el corpus creado para la investigación.

Evaluación cuantitativa En la Tab. 6 se muestra el número medio de caracteres que han generado los modelos en base a la batería de preguntas.

Modelo	Castellano	Inglés
ChatGPT 3.5 Finetune	438.6	387.76
Mistral 7B Finetune	374.48	421.48
Llama2 7B Finetune	171.08	161.12

Tabla 6: Número medio de caracteres generados en la prueba de Finetune

Mientras que en la Fig. 9, la Fig. 10 y la Figure 11 se muestran los resultados del mismo análisis de coherencia realizado para la otra prueba de evaluación.

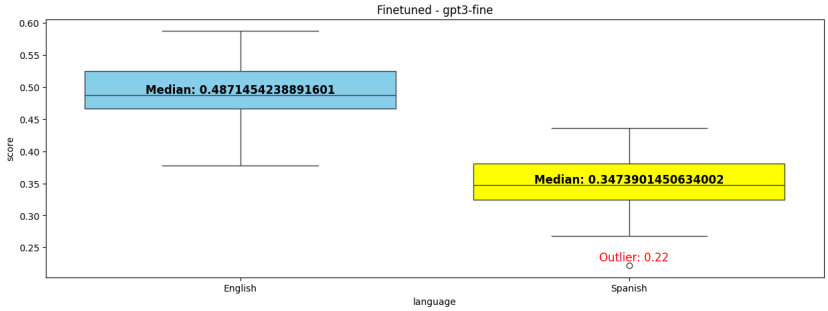


Figura 9: Coherencia ChatGTP 3.5

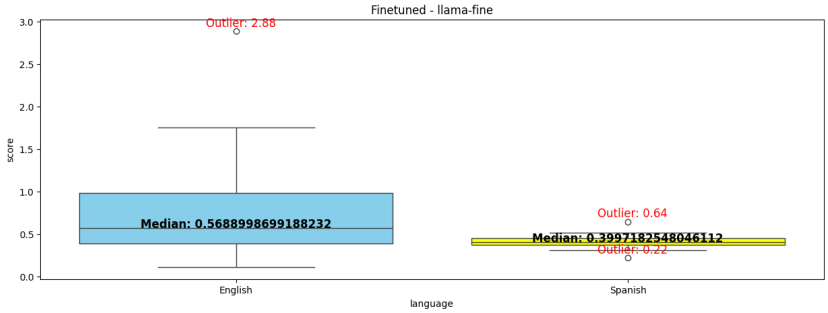


Figura 10: Coherencia Llama2

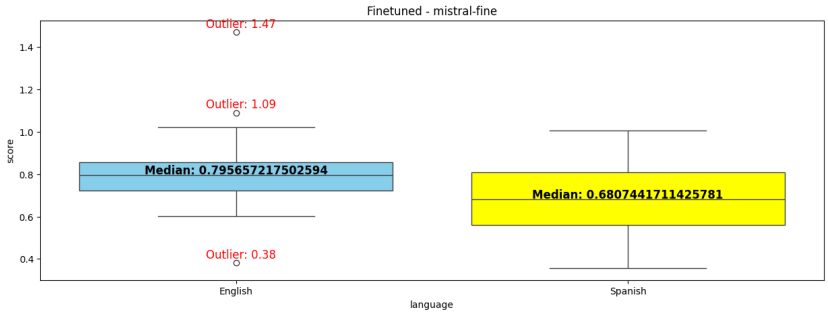


Figura 11: Coherencia Mistral

Evaluación humana En la Tab. 7 se muestran las métricas que ha obtenido en la evaluación humana el modelo de Llama2 en la prueba de evaluación de *finetuning*.

Métrica	Castellano	Inglés	Total
Veracidad	4.60	0.00	2.30
Relevancia	7.72	0.00	3.86
Claridad	8.84	0.00	4.42
Fluidez	8.64	0.00	4.32

Tabla 7: Evaluación humana - Llama2 (Finetuning)

En la Tab. 8 se muestran las métricas que ha obtenido en la evaluación humana el modelo de Mistral en la prueba de evaluación de *finetuning*.

Métrica	Castellano	Inglés	Total
Veracidad	1.20	0.00	0.60
Relevancia	4.28	0.00	2.14
Claridad	6.64	0.00	3.32
Fluidez	6.52	0.00	3.26

Tabla 8: Evaluación humana - Mistral (Finetuning)

En la Tab. 9 se muestran las métricas que ha obtenido en la evaluación humana el modelo de ChatGPT 3.5 en la prueba de evalución de *finetuning*.

Métrica	Castellano	Inglés	Total
Veracidad	6.04	6.04	6.04
Relevancia	9.44	9.44	9.44
Claridad	9.12	9.12	9.12
Fluidez	9.12	9.12	9.12

Tabla 9: Evaluación humana - GPT 3.5 (Finetuning)

6. Discusión

En la Fig. 12 se muestra gráficamente los valores que han obtenido en las pruebas de *finetuning* para observar el comportamiento a pequeña escala y poder determinar el mejor comportamiento que replicar a una mayor escala.

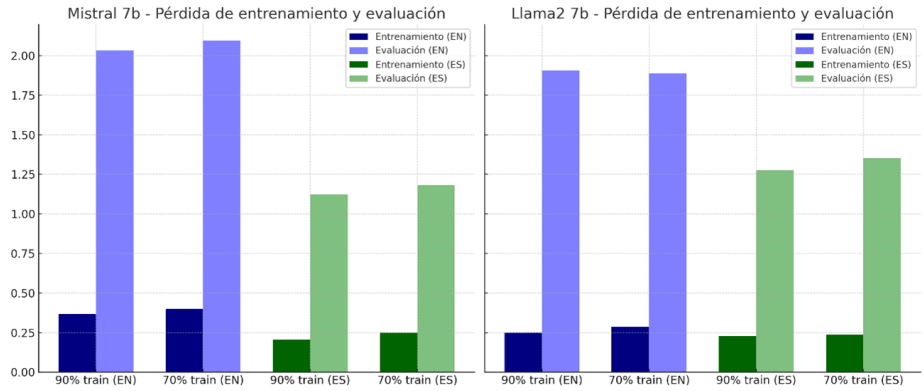


Figura 12: Distribución de las pérdidas

Ambos modelos obtuvieron un mejor comportamiento cuando el corpus estaba construido en español, siendo esto un hecho anómalo a lo que se podría pensar debido a que generalizando la mayoría de los datos con los que estén entrenados estos modelos sean en inglés.

Las valores que se obtuvieron para las métricas en ambos modelos seguían el mismo comportamiento, en español tanto la pérdida de entrenamiento como la de evaluación son mejores, pero a pesar de que la pérdida de entrenamiento obtiene valores buenos (cercanos al cero) la pérdida de evaluación tiende a aumentar indicando que puede haber un sobreajuste en el entrenamiento debido a la poca calidad del corpus creado.

Los modelos finales desarollados fueron con el corpus en español y una distribución del 90 % de los datos para entrenamiento y 10 % para evaluación ya que fue con la que mejores valores se obtuvieron en las pruebas piloto, siendo entrenados con valores más reales se obtuvieron unos valores de pérdida de entrenamiento prácticamente cero mientras que la pérdida de evaluación ha tenido un aumento exponencial y diferenciado entre los modelos ya que Mistral tiene un 106 % más de pérdida que Llama2, es decir, de forma teórica ha tenido un mayor sobreajuste.

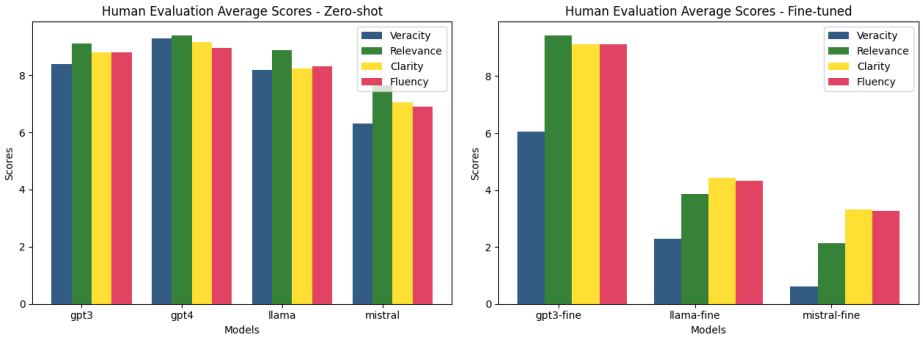
Se ha observado cómo el proceso de especialización ha influido negativamente. La longitud de las cadenas que generaban los modelos base respecto a los modelos que han sido especializados son entorno un 80 % y un 95 % mayor. Este hecho no representa la calidad de las respuestas pero hace ver la importancia de crear un corpus de calidad ya que es un componente decisivo a la hora de realizar el *finetuning* por la capacidad de aprendizaje de patrones y sintaxis que tienen los LLM.

La coherencia obtenida también se ha visto afectada negativamente según lo observado en las Figs. 7, 8 y las figuras referentes a la prueba de *finetuning* (Figs. 9, 10, 11). Los modelos base han obtenido valores medios cercanos al cero, siendo lo óptimo, en ambos idiomas mientras que al realizar la especialización, ha habido un 357 % de aumento en la métrica respecto a la coherencia en inglés y un 247 % respecto al español.

Cómo la finalidad de estos modelos es la de generar texto no hay que tener solo en cuenta los análisis cuantitativos que se han presentado si no que también hay que analizar la calidad de las preguntas desde una perspectiva de contenidos y de forma lingüística mediante una evalución por parte de personas.

Para ello se ha realizado una clasificación de la veracidad, la relevancia, la claridad

y la fluencia otorgándole una puntación de entre cero y diez a las respuestas generadas por cada uno de los modelos a cada una de las preguntas de la batería. En las Figs. 13a y 13b se presenta la distribución por cada modelo y cada prueba que se ha llevado a cabo.

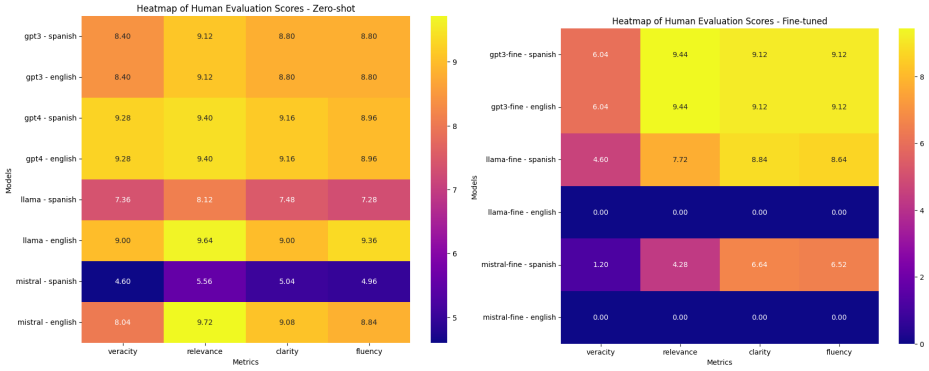


(a) Zero Shot (b) Finetune

Figura 13: Representación de la evaluación humana

Analizando los resultados se observa como los modelos base obtiene mejores valores en las métricas como se ha ido advirtiendo con el análisis cuantitativo aunque estando todos muy a la par ChatGPT 4 es el que mejores métricas ha obtenido. Viendo los resultados de la prueba de *finetuning* se observa como ha afectado a la calidad de las respuestas la especialización, ChatGTP 3.5 ha sido el único de los modelos que ha sido capaz de responder correctamente en ambos idiomas, mientras que Llama2 y Mistral han perdido la capacidad de responder en otro idioma que no sea con el que han sido especializados.

En la Fig. 14b se observa como en los modelos especializados cuando se les realiza la inferencia con la batería de preguntas en inglés obtienen una puntación de cero, excepto con ChatGPT 3.5, en cualquiera de las métricas ya que responden en español sin llegar a ser coherentes ni obteniendo respuestas completas, aún así en el resto de condiciones en comparación a los que se obtienen con los modelos base (Fig. 14a) se observa como se ha producido una reducción significativa de la calidad en las respuestas debido al proceso de especialización.



(a) Zero Shot (b) Finetune

Figura 14: Heatmaps de la evaluación humana

7. Conclusiones

El trabajo realizado se resume en la investigación de los diferentes LLM open source para ser utilizados como Chatbots siendo escogidos los modelos de Llama2 y Mistral, la creación de un conjunto de datos con el formato necesario multilenguaje (español e inglés) enfocado en la disciplina de la Ingeniería de Software; aplicar un proceso de finetuning con los modelos elegidos y el corpus desarrollado para desarrollar asistentes tipo Chatbot que estén especializados en la Ingeniería de Software y proceder a realizar una comparación con los modelos privados más populares como son ChatGPT 3.5 y ChatGPT 4.

Se ha conseguido desarrollar dos LLM que sirven como herramienta de apoyo al aprendizaje autónomo de la disciplina de Ingeniería de Software junto a dos dataset enfocados en esta disciplina en diferentes idiomas con el trabajo que conlleva todo el preprocesado y traducción de los datos extraídos, aunque el proceso de especialización que se ha llevado a cabo no ha resultado en ninguna mejora e incluso empeora los resultados que fueron obtenidos al utilizar los modelos base; de la investigación realizada se pueden extraer las siguientes conclusiones:

- El corpus es un factor importante a la hora de realizar el *finetuning*, debido a que la calidad que ofrezca junto a la variedad de formato y longitudes entre preguntas y respuestas va a ser un factor determinante sobre el comportamiento final de los modelos especializados. La construcción del *dataset* tanto en inglés, debido a que la mayoría de software y recursos están enfocados para angloparlantes, como en español ya que, el número de hispanohablantes es muy elevado, es relevante para esta y futuras investigaciones. La dificultad de crear estos *dataset* es el preprocesamiento de los datos en crudo ya que resulta en una tarea compleja y en muchas ocasiones tiene dificultades de ser automatizadas acaban generando conjuntos de datos repetitivos y sin la calidad suficiente para poder realizar un proceso de *finetuning* que aporte buenos resultados.
- El proceso de especialización es viable si se cuenta con un corpus de muy buena calidad y extenso junto a los recursos computacionales necesarios aunque se utilicen técnicas de optimización y reducción de carga.

- En las pruebas de evaluación tanto Llama2 como Mistral han obtenido valores similares a ChatGPT 3.5
- Los modelos base *open source* han obtenido buenos valores en las evaluaciones, pero el que mejores resultados ha obtenido es ChatGPT 4
- Las opciones más viables para funcionar como herramientas de apoyo son las desarrolladas por OpenAI tanto la versión gratuita ChatGPT 3.5 como la de pago ChatGTP 4, debido al ecosistema que ofrece la empresa, la facilidad de uso y la gran base de conocimiento con la que han sido entrenados los modelos.

- Extender y mejorar el corpus, aumentando la variedad de contenidos y de formatos de forma coherente y veraz, además de hacerlo multilenguaje para evitar los sesgos cuando se realiza la inferencia en distintos idiomas.
- Mejorar el hardware para repetir el proceso con modelos de mayor tamaño para poder extraer más conclusiones.

Referencias

1. F. García-Peñalvo and A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, p. 7, 2023.
2. R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of Generative AI Applications," June 2023. arXiv:2306.02781 [cs].
3. D. Wang, Y. Tao, and G. Chen, "Artificial intelligence in classroom discourse: A systematic review of the past decade," *International Journal of Educational Research*, vol. 123, p. 102275, Jan. 2024.
4. T. Wang and E. C. K. Cheng, "An investigation of barriers to Hong Kong K-12 schools incorporating Artificial Intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100031, 2021.
5. S. Bennet, "GitHub lanza el plan 'Copilot for Business' en medio de problemas de infracción de derechos de autor," Dec. 2022.
6. P. D. E. Ruiz, "Meta usará publicaciones de Facebook e Instagram para entrenar su IA: ¿está en riesgo tu privacidad?," June 2024. Section: Tecno.
7. Y. Li, Z. Tan, and Y. Liu, "Privacy-Preserving Prompt Tuning for Large Language Model Services," May 2023. arXiv:2305.06212 [cs].
8. Jefatura del Estado, "Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación," July 2022.
9. F. J. García Peñalvo, F. Llorens-Largo, and J. Vidal, "La nueva realidad de la educación ante los avances de la inteligencia artificial generativa," *RIED-Revista Iberoamericana de Educación a Distancia*, vol. 27, pp. 9–39, July 2023.
10. R. S. T. Huang, K. J. Q. Lu, C. Meaney, J. Kemppainen, A. Punnett, and F.-H. Leung, "Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study," *JMIR MEDICAL EDUCATION*, vol. 9, p. e50514, 2023. Num Pages: 9 Place: Toronto Publisher: Jmir Publications, Inc Web of Science ID: WOS:001085615600001.
11. D. T. K. Ng, C. W. Tan, and J. K. L. Leung, "Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study," *British Journal of Educational Technology*, vol. n/a, no. n/a. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13454.

12. "Introducing ChatGPT."
13. "GPT-4."
14. "What Are Large Language Models (LLMs)? | IBM," Nov. 2023.
15. "What is an LLM? A Guide on Large Language Models and How They Work."
16. M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024. Conference Name: IEEE Access.
17. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023. arXiv:1706.03762 [cs].
18. H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A Comprehensive Overview of Large Language Models," Apr. 2024. arXiv:2307.06435 [cs] version: 7.
19. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," July 2023. arXiv:2307.09288 [cs].
20. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," Oct. 2023. arXiv:2310.06825 [cs].
21. M. AI, "Mistral 7B-announcment," Sept. 2023. Section: news.
22. J. Jeon and S. Lee, "Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT," *Education and Information Technologies*, vol. 28, pp. 15873–15892, Dec. 2023.
23. F. Pradhan, A. Fiedler, K. Samson, M. Olivera-Martinez, W. Manatsathit, and T. Peeraphatdit, "Artificial intelligence compared with human-derived patient educational materials on cirrhosis," *Hepatology Communications*, vol. 8, p. e0367, Mar. 2024.
24. Y. Li, C. Zeng, J. Zhong, R. Zhang, M. Zhang, and L. Zou, "Leveraging Large Language Model as Simulated Patients for Clinical Education," Apr. 2024. arXiv:2404.13066 [cs].
25. M. Bernabei, S. Colabianchi, A. Falegnami, and F. Costantino, "Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100172, Jan. 2023.
26. D. H. Chang, M. P.-C. Lin, S. Hajian, and Q. Q. Wang, "Educational Design Principles of Using AI Chatbot That Supports Self-Regulated Learning in Education: Goal Setting, Feedback, and Personalization," *Sustainability*, vol. 15, p. 12921, Jan. 2023. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.
27. T. Al Shloul, T. Mazhar, Q. Abbas, M. Iqbal, Y. Y. Ghadi, T. Shahzad, F. Mallek, and H. Hamam, "Role of activity-based learning and ChatGPT on students' performance in education," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100219, June 2024.
28. J. Crawford, K.-A. Allen, B. Pani, and M. Cowling, "When artificial intelligence substitutes humans in higher education: the cost of loneliness, student success, and retention," *Studies in Higher Education*, vol. 0, no. 0, pp. 1–15, 2024. Publisher: Routledge _eprint: https://doi.org/10.1080/03075079.2024.2326956.
29. J. Cross, R. Robinson, S. Devaraju, A. Vaughans, R. Hood, T. Kayalackakom, P. Honnavar, S. Naik, and R. Sebastian, "Transforming Medical Education: Assessing the Integration of ChatGPT Into Faculty Workflows at a Caribbean Medical

School," *CUREUS JOURNAL OF MEDICAL SCIENCE*, vol. 15, p. e41399, July 2023. Num Pages: 11 Place: London Publisher: Springernature Web of Science ID: WOS:001055896400011.
30. "Self-regulated learning," Apr. 2024. Page Version ID: 1221497274.
31. M. P.-C. Lin and D. Chang, "CHAT-ACTS: A pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100167, Jan. 2023.
32. "Active learning," Mar. 2024. Page Version ID: 1214258589.
33. W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," Mar. 2024. arXiv:2403.04132 [cs].
34. "mistralai/Mistral-7B-Instruct-v0.2 · Model."
35. "meta-llama/Llama-2-7b-chat-hf · Model," Apr. 2024.
36. C. Zheng, "chujiezheng/chat_templates," May 2024. original-date: 2023-11-23T08:29:54Z.
37. "Curso: Ingeniería del Software I (2011)."
38. F. J. García-Peñalvo, A. García-Holgado, and A. Vázquez-Ingelmo, "Recursos docentes de la asignatura Ingeniería de Software I. Grado en Ingeniería Informática. Curso 2023-2024," Apr. 2024. Version Number: 1.0.
39. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021. arXiv:2106.09685 [cs].
40. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019. arXiv:1810.04805 [cs].
41. M. Oleszak, "Zero-Shot and Few-Shot Learning with LLMs," Mar. 2024.