

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS  
INDUSTRIALES Y DE TELECOMUNICACIÓN

UNIVERSIDAD DE CANTABRIA



*Trabajo Fin de Grado*

**DISEÑO Y EVALUACION DE EXAMENES DE  
TIPO TEST MEDIANTE IA**  
(Design and evaluation of multiple choice  
tests using AI )

Para acceder al Título de

***Graduado en  
Ingeniería de Tecnologías de Telecomunicación***

Autor: Diego Gómez Gómez

Julio - 2025



# **GRADUADO EN INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN**

CALIFICACIÓN DEL TRABAJO FIN DE GRADO

**Realizado por: DIEGO GÓMEZ GÓMEZ**

**Director del TFG: ALBERTO ELOY GARCÍA GUTIÉRREZ**

**Título: “DISEÑO Y EVALUACIÓN DE EXÁMENES TIPO TEST MEDIANTE IA”**

**Title: “DESIGN AND EVALUATION OF MULTIPLE CHOICE TESTS USING AI”**

**Presentado a examen el día: 16 de junio de 2025**

para acceder al Título de

## **GRADUADO EN INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN**

Composición del Tribunal:

Presidente (Apellidos, Nombre): RAMON AGÜERO CALVO

Secretario (Apellidos, Nombre): LUIS SANCHEZ GONZALEZ

Vocal (Apellidos, Nombre): PABLO PEDRO SANCHEZ ESPESO

Este Tribunal ha resuelto otorgar la calificación de: .....

**Fdo: El Presidente**

**Fdo: El Secretario**

**Fdo: El Vocal**

**Fdo: El Director del TFG**

(sólo si es distinto del Secretario)

Vº Bº del Subdirector

Trabajo Fin de Grado Nº

(a asignar por Secretaría)



## Resumen

En este Trabajo de Fin de Grado, se realiza la búsqueda del mejor modelo para ser utilizado de manera local mediante la aplicación LM Studio con el objetivo de corregir exámenes de tipo test mediante plantillas. Para ello se estudian los comportamientos de distintos modelos frente a distintas pruebas y se valora mediante gráficas y tablas cual es el más apto para la tarea buscada.

El modelo Phi 4 destaca por su equilibrio entre velocidad, precisión (90 % en corrección automática) y facilidad de uso en entornos con hardware limitado. Así se demuestra que su implementación local aporta beneficios claros frente al uso de servicios en la nube: mejora la privacidad de los datos, elimina costes por uso y facilita la personalización del sistema para tareas concretas.

Además, se ha desarrollado una API con la intención de integrar el modelo en aplicaciones educativas, y se han documentado todas las fases del proceso: desde la selección e instalación de los modelos, hasta la puesta en marcha de un servidor funcional. Las pruebas prácticas realizadas mostraron un rendimiento estable y confiable.

En conclusión, este proyecto evidencia la viabilidad de utilizar modelos de lenguaje locales para optimizar procesos educativos, especialmente en contextos con restricciones presupuestarias o de conectividad. La inteligencia artificial, aplicada correctamente, puede convertirse en una herramienta clave para liberar tiempo docente, mejorar la calidad de la evaluación y facilitar el aprendizaje personalizado.



## **Abstract**

In this Final Degree Project, the search for the best model to be used locally through the LM Studio application is carried out with the aim of grading multiple-choice exams using templates. To achieve this, the behaviour of different models is studied through various tests and evaluated using graphs and tables to determine which is most suitable for the intended task.

The Phi 4 model stands out for its balance between speed, accuracy (90% in automated grading), and ease of use in environments with limited hardware. It is thus demonstrated that local implementation provides clear benefits compared to the use of cloud services: improved data privacy, elimination of usage costs, and easier system customization for specific tasks.

Additionally, an API has been developed with the intention of integrating the model into educational applications, and all stages of the process have been documented—from model selection and installation to the deployment of a functional server. Practical tests showed stable and reliable performance.

In conclusion, this project demonstrates the feasibility of using local language models to optimize educational processes, especially in contexts with budget or connectivity constraints. When applied correctly, artificial intelligence can become a key tool for reducing teacher workload, improving assessment quality, and facilitating personalized learning.



# ÍNDICE

Resumen .....	5
Abstract.....	7
1. Introducción.....	11
1.1. Contexto y motivación .....	11
1.2. Objetivos.....	12
1.2.1 Comparar los distintos modelos empleados .....	12
1.2.2 Destacar las ventajas de utilizar IA en local.....	13
1.2.3 Desarrollar una IA especializada en la corrección de exámenes tipo test y la generación de plantillas para facilitar su revisión. ....	13
1.2.4 Implementar una API para facilitar la integración de la IA en aplicaciones..	13
2. Estado del arte .....	15
2.1 Metodologías de Diseño de Exámenes Impulsadas por IA .....	16
2.2 Evaluación y Garantía de Calidad .....	17
2.3 Tokens .....	17
2.4 Modelos de lenguaje basados en Transformers .....	19
2.4.1 BERT (Bidirectional Encoder Representations from Transformers) .....	19
2.4.2 T5 (Text-to-Text Transfer Transformer) .....	20
2.4.3 PaLM (Pathways Language Model).....	21
2.5. Beneficios de Usar IA en Local.....	22
2.5.1. Privacidad y seguridad .....	22
2.5.2. Coste y eficiencia .....	23
2.6. Personalización de IA local .....	23
3. Desarrollo práctico .....	29
3.1. Configuración del entorno .....	29
3.2. Definición de pruebas realizadas .....	31
3.2.1 Generación automática de plantillas.....	31
3.2.2. Corrección automatizada de exámenes tipo test.....	33
3.2.3. Métricas utilizadas para comparar desempeño.....	47
3.3.4 Resumen de resultados. ....	51
3.3. Selección del mejor modelo .....	52
3.4 Desarrollo de una API para la IA .....	53

3.4.1. Diseño de la API.....	53
3.4.2. Objetivos y funcionalidades principales.....	58
3.4.3. Implementación .....	60
4. Conclusiones y Futuras Mejoras .....	65
4.1. Futuras mejoras .....	66
4.2. Reflexión final .....	67
5. Bibliografía.....	69
6. Anexos.....	71

# **1. Introducción**

## **1.1. Contexto y motivación**

En los últimos años, la inteligencia artificial (IA) ha experimentado un crecimiento acelerado, expandiéndose más allá de sectores especializados para aplicarse en ámbitos tan diversos como la medicina, la educación, la industria y el entretenimiento. Debido a su gran capacidad para procesar grandes volúmenes de información, detectar y corregir errores y realizar actividades de manera autónoma, podemos asegurar que la IA ha revolucionado la tecnología y la forma en la que interactuamos con ella.

Uno de los principales desafíos es la creciente necesidad de entrenar distintos modelos de IA debido a que es de gran importancia para el correcto funcionamiento de estas y aparecen problemas como la privacidad de los datos, costes de infraestructura externa y la conexión a servidores remotos. También es importante destacar las limitaciones que implica la dependencia de la nube ya que puede limitar la seguridad, personalización y autonomía.

Teniendo en cuenta dichas limitaciones, podemos entender el auge de la importancia del desarrollo y entrenamiento de modelos de IA en entornos locales, aprovechando así las capacidades computacionales de dispositivos personales o infraestructuras privadas. Esto permite reducir costos y mejorar la privacidad, también es ideal si se buscan necesidades específicas.

El propósito de este Trabajo de Fin de Grado (TFG) es analizar y comparar distintos modelos de IA ejecutados de manera local para poder determinar las diferencias entre ellos. En concreto esta IA se encargará de la corrección automática de plantillas de exámenes de tipo test para poder optimizar el proceso de evaluación en entornos educativos, de esta forma, se favorece a la agilización de la labor de los docentes y la mejora de aprendizaje de los estudiantes.

## 1.2. Objetivos

Este Trabajo de Fin de Grado tiene como objetivo principal el desarrollo de una inteligencia artificial (IA) que funcione de forma local y sea capaz de crear y corregir plantillas de exámenes tipo test. Para ello, usaremos la aplicación LM Studio.

LM Studio es una herramienta gratuita que permite cargar, personalizar y utilizar distintos modelos de lenguaje de gran tamaño (LLMs) de forma local. Gracias a esto, se ha trabajado con distintos modelos open-source como LLaMA, Qwen y Phi, permitiendo así experimentar con estos modelos en un entorno controlado y privado.[1]

Se realizará una comparativa entre los distintos modelos para poder evaluar sus capacidades en tareas específicas relacionadas con el entorno educativo. Para ello se llevará a cabo un análisis de los siguientes parámetros

- Tamaño del modelo
- Consumo de recursos
- Tiempo de respuesta
- Precisión en comprensión y análisis de texto

### 1.2.1 Comparar los distintos modelos empleados

Los modelos que se han seleccionado para esta evaluación son los siguientes:

- **LLaMA (Large Language Model Meta AI)**: Fue desarrollado por Meta AI y es uno de los modelos open-source más influyentes, con versiones optimizadas para uso local como LLaMA 2 y LLaMA 3.[2]
- **Qwen 2**: Este modelo fue desarrollado por Alibaba DAMO Academy, conocido por su rendimiento destacado en multilingüismo y razonamiento lógico. Ha sido optimizado para tareas de generación y comprensión de texto.[3]
- **Phi 4**: Modelo desarrollado por Microsoft Research, centrado en un entrenamiento eficiente con datos sintéticos de alta calidad, destacando por su tamaño compacto y alto rendimiento relativo. Es de los más utilizados en tareas educativas y de razonamiento.[4]

Para llevar a cabo las comparaciones se realizarán distintas tareas con cada uno de los modelos para poder medir las siguientes métricas:

- Exactitud de corrección
- Velocidad de respuesta
- Capacidad de generación de plantillas

### 1.2.2 Destacar las ventajas de utilizar IA en local.

Utilizar una IA en local tiene varias ventajas en comparación al uso de IAs en la nube. Las principales y más importantes para este Trabajo de Fin de Grado son las siguientes:

- **Privacidad de los datos:** Al no depender de servidores externos, los datos sensibles no salen del dispositivo local, gracias a esto podemos mantener seguras las calificaciones del alumno y sus datos personales como nombre y D.N.I.[1]
- **Independencia de conexión:** El sistema funciona sin necesidad de estar conectado a internet, lo que resulta ideal para entornos educativos con acceso limitado. También es muy útil para viajes o ubicaciones sin conexión a internet.[1]
- **Personalización y control total:** Se pueden ajustar todos los aspectos del modelo y del sistema, lo que permite una adaptación más precisa a las necesidades concretas del usuario.[1]
- **Ahorro, sostenibilidad y eficiencia:** Al usar inteligencia artificial en local, se eliminan las grandes cargas de procesamiento en la nube. Esto reduce la necesidad de centros de datos, que requieren intensa refrigeración. Como resultado, se ahorra una gran cantidad de agua y energía destinadas a enfriar servidores.[5]

### 1.2.3 Desarrollar una IA especializada en la corrección de exámenes tipo test y la generación de plantillas para facilitar su revisión.

El objetivo principal de este trabajo es desarrollar una IA local capaz de realizar corrección automática de exámenes tipo test. Para ello, el modelo debe de ser capaz de analizar las respuestas de los alumnos en base a una plantilla dada, detectar errores, aciertos y preguntas en blanco, crear las plantillas de corrección y dar una puntuación final además de hacer una nota media de todos los alumnos.

El modelo se adaptará para estas tareas utilizando prompts diseñados específicamente, con pruebas iterativas que permitan mejorar su precisión en escenarios reales. Además, se tendrán en cuenta aspectos como la facilidad de uso por parte del profesorado y la posibilidad de extenderlo a otros formatos de examen.

### 1.2.4 Implementar una API para facilitar la integración de la IA en aplicaciones.

Ya como objetivo secundario, se ha planteado estudiar la posibilidad de implementar una API que permita acceder a los distintos modelos mediante comandos por consola o mediante una página web, creando así nuevas maneras para acceder a los distintos modelos.



## **2. Estado del arte**

La integración de la inteligencia artificial (IA) en la realización y evaluación de exámenes tipo test ha revolucionado la pedagogía digital, optimizando procesos y mejorando la precisión de las evaluaciones.

Los fundamentos conceptuales de la evaluación impulsada por IA se originan en el trabajo pionero de Alan Turing sobre inteligencia artificial, quien sugirió que los sistemas que pueden emular la cognición humana podrían, con el tiempo, realizar tareas evaluativas complicadas. Las taxonomías cognitivas, como la Revisada Taxonomía de Bloom, guían la creación de preguntas mediante IA al organizar elementos que miden habilidades específicas, abarcando desde la memorización hasta el análisis crítico. Por ejemplo, los modelos basados en “Transformers” entrenados en conjuntos de datos educativos pueden formular preguntas que valoran habilidades de "análisis" o "evaluación" utilizando patrones gramaticales. Algunas investigaciones recientes indican que modelos como T5 alcanzan un 89% de exactitud al relacionar preguntas generadas con los niveles de Bloom cuando son entrenados en conjuntos de datos etiquetados. [6]

La aparición de modelos de lenguaje grandes (LLM) ha abordado desafíos persistentes en comprensión semántica y relevancia contextual. Los sistemas como GPT-4 y Llama 2

emplean mecanismos de atención para analizar textos complejos y generar preguntas que mantienen fidelidad al material fuente. Por ejemplo, el marco de Blšák y Rozinajová combina patrones lingüísticos basados en reglas con aprendizaje por refuerzo para producir preguntas factuales, logrando una mejora del 23% en calidad equivalente humana sobre modelos seq2seq anteriores. Su análisis jerárquico de oraciones, que extrae características léxicas, sintácticas y semánticas-garantiza que las preguntas generadas eviten ambigüedades y redundancias.[7]

Así, por ejemplo, LMStudio es una herramienta open-source que se ha liberado recientemente, y que es ideal para trabajar con LLM locales. Desarrollada por Hugging Face, Permite entrenar y ajustar modelos de lenguaje grandes (LLM) directamente en la máquina del usuario. A diferencia de servicios en la nube, LMStudio ofrece un mayor control sobre los datos y la seguridad, características valiosas especialmente en ámbitos sensibles. LMStudio facilita cargar modelos pre-entrenados (BERT, GPT, etc.) y entrenarlos con datos locales. Esto motiva su elección en este proyecto, pues se buscará entrenar un LLM offline para la tarea de calificación automática.[1]

## 2.1 Metodologías de Diseño de Exámenes Impulsadas por IA

Los pipelines modernos para la generación de preguntas constan de tres componentes principales: preprocesamiento del contenido, generación neuronal y validación post-procesamiento. El marco TP3, que se utiliza para identificar conceptos clave en textos fuente, emplea técnicas como el etiquetado gramatical y el reconocimiento de entidades nombradas. Por su parte, el modelo T5, un transformador entrenado para tareas de procesamiento del lenguaje natural se encarga de generar preguntas candidatas. Estas preguntas se filtran posteriormente mediante verificaciones de similitud semántica utilizando "embeddings" de Sentence-BERT. Este enfoque híbrido ayuda a reducir los errores gramaticales en comparación con los modelos end-to-end. [8]

Característica	TP3	T5
<b>Tipo</b>	Marco (framework)	Modelo (modelo generativo)
<b>Función principal</b>	Utiliza etiquetado gramatical y reconocimiento de entidades nombradas	Genera preguntas candidatas a partir del texto fuente
<b>Técnica destacada</b>	Identificación de conceptos clave en los textos fuente	Verificación semántica usando embeddings de Sentence-BERT
<b>Etapas del proceso</b>	Preprocesamiento con etiquetas gramaticales y entidades antes de generar preguntas	Generación y filtrado posterior de preguntas mediante similitud semántica
<b>Objetivo</b>	Facilitar la identificación de conceptos relevantes en el texto original	Aumentar la precisión semántica de las preguntas generadas
<b>Beneficio destacado</b>	Mejora la calidad de entrada para la generación de preguntas	Reduce errores gramaticales en comparación con modelos end-to-end
<b>Enfoque general</b>	Modular: divide el proceso en etapas (etiquetado, reconocimiento, generación)	Híbrido: generación + filtrado semántico posterior

Tabla 1: Comparación TP3 y T5

Un paradigma novedoso, denominado Generación de Preguntas con Planificación Previa (PFQS), aprovecha LLMs como Llama 2 para crear primero un "plan de respuesta" que esboza conceptos centrales y competencias evaluadas. Al separar la selección de contenido de la redacción de preguntas, PFQS mejora la adherencia a requisitos en un 41% en conjuntos de datos educativos multilabel como FairytaleQA. Por ejemplo, al generar preguntas sobre el tema de una historia, el sistema identifica primero puntos clave de la trama y motivaciones de personajes antes de formular el enunciado y las opciones.[9]

Por otro lado, los distractores u opciones incorrectas pero plausibles, representan un desafío crítico en el diseño de preguntas de opción múltiple. Métodos tradicionales dependían de bases léxicas como WordNet o etiquetado de roles semánticos para extraer términos candidatos. Sin embargo, LLMs como ChatGPT ahora permiten la creación dinámica de distractores mediante prompting aumentado por recuperación ya que ahora, en lugar de usar métodos manuales o basados solo en diccionarios léxicos, se pueden generar opciones incorrectas pero creíbles de forma automática. Esto mejora la calidad y relevancia de las preguntas de opción múltiple, especialmente en educación y evaluación. [10]

Además, se pueden emplear técnicas híbridas que combinan modelos generativos y discriminativos. Por ejemplo, los modelos mT5 ajustados generan distractores iniciales, que luego se revaloran usando métricas de similitud frente a la respuesta correcta. Este proceso en dos etapas mejora la calidad de los distractores manteniendo diversidad. Además, incorporar embeddings de palabras y métricas de distancia de edición ayuda a filtrar distractores ortográficamente similares a la clave, mitigando pistas accidentales.[10]

## 2.2 Evaluación y Garantía de Calidad

Los sistemas de puntuación basados en procesamiento del lenguaje natural (PLN) manejan ahora tanto preguntas cerradas como respuestas abiertas. Para opción múltiple, motores basados en reglas emparejan selecciones con claves de respuesta con 100% de fiabilidad, mientras modelos de similitud semántica (e.g., BERTScore) califican respuestas breves comparando embeddings con respuestas de referencia. Sistemas avanzados como TP3 emplean puntuación por conjunto, combinando verificaciones sintácticas (e.g., presencia de palabras clave) con métricas neuronales para reducir falsos positivos.[11]

La generación de retroalimentación ha evolucionado de respuestas predefinidas a sistemas de tutoría adaptativa. Por ejemplo, respuestas incorrectas que activan clusters de error específicos (e.g., aplicar mal la tercera ley de Newton) hacen que el sistema genere explicaciones remediales y conjuntos de práctica seleccionados. Un ensayo de 2024 mostró que la retroalimentación generada por IA mejoró la retención estudiantil en 19% sobre comentarios genéricos.[11]

## 2.3 Tokens

En el contexto del PLN, los modelos de lenguaje trabajan internamente con unidades mínimas de información llamadas tokens. Estas unidades no siempre corresponden a

palabras completas; pueden ser letras, sílabas, prefijos, sufijos o incluso espacios, dependiendo del modelo y el sistema de tokenización.

La comprensión del concepto de token es esencial para entender cómo interactuar con modelos como GPT, LLaMA, Phi, Mistral, etc., ya sea en la nube o en local. También es vital para saber cómo optimizar prompts y estimar los recursos que consumirá una inferencia.

Un token es una unidad básica de texto que un modelo de lenguaje procesa como entrada o salida. Dependiendo del sistema de tokenización utilizado, un token puede ser:

- Una palabra completa: "portátil"
- Un fragmento de palabra: "por", "tá", "til"
- Un carácter: "p", "o", "r"
- Un símbolo especial: espacio, signos de puntuación, salto de línea

Por ejemplo, el texto: "Hola mundo." puede dividirse en los tokens: ["Hola", " mundo", "."]. Aquí el espacio inicial de " mundo" también cuenta como parte del token.

El proceso de convertir texto en tokens se llama tokenización. Existen varios métodos:

- **Tokenización por espacios:** Es la más sencilla. Divide el texto en palabras según los espacios: "Esto es un ejemplo": ["Esto", "es", "un", "ejemplo"]
- **Tokenización por subpalabras:** Los modelos modernos usan técnicas que dividen las palabras en fragmentos más pequeños. Algunas técnicas comunes son:
  - BPE (Byte Pair Encoding): combina caracteres y pares frecuentes en subpalabras.
  - WordPiece: usado por BERT, construye palabras con prefijos como "ción".
  - SentencePiece: permite tokenizar sin depender del espacio como delimitador.

El siguiente ejemplo hace uso de BPE:

"desinformación" → ["des", "in", "form", "ación"]

A utilizar este tipo de tokenización, permite reducir el tamaño del vocabulario, manejar palabras raras o no vistas durante el entrenamiento y, como consecuencia, mejorar la eficiencia de los modelos.

Además, hay que tener en cuenta que estos sistemas no procesan texto directamente como cadenas, sino como secuencias de tokens que, a su vez, se transforman en vectores numéricos. De esa forma, cada token:

- Tiene un ID numérico (por ejemplo, "Hola" → 3145).
- Es convertido en un vector mediante una embedding layer.
- Forma parte de una ventana de contexto (limitada).

Sin embargo, todos los modelos tienen un límite de tokens por mensaje o prompt. La siguiente tabla resume algunos ejemplos:

Modelo	Máx. tokens
GPT-3.5 Turbo	4,096
GPT-4-32K	32,768
Claude 3 Opus	200,000+

<b>LLaMA 3</b>	8,192 (según configuración)
<b>Phi-2 / Phi-4</b>	2,048 – 4,096
<b>Qwen 2</b>	32,768 (Qwen2-7B-Instruct)

Tabla 2: Comparación de tokens máximos

Superar este límite puede causar truncamiento del texto, impedir procesar el mensaje, o disminuir la coherencia si se pierde parte del contexto.

Como el número de tokens de un texto no coincide con el número de palabras o caracteres, es importante contar tokens de manera precisa si se desea no superar los límites del modelo, optimizar prompts para tareas largas o estimar consumo de recursos.

A modo de ejemplo, los servicios de IA como OpenAI o Cohere, cobran por tokens procesados. Por eso, escribir prompts más largos aumenta el coste. De hecho, GPT-4 Turbo cuesta 0.01\$ por 1,000 tokens de entrada y 0.03\$ por 1,000 tokens de salida. Así, un examen largo o una plantilla extensa puede superar los 3,000 tokens fácilmente. Sin embargo, ejecutarlo en local eliminaría estos costes.

De esta forma, optimizar prompts es casi una necesidad, pero es necesario tener en cuenta algunas consideraciones como, por ejemplo, usar frases cortas, evitar repetir instrucciones o reutilizar partes de prompts. Pese a todo, se cometen errores que son muy comunes, como es asumir que 100 palabras es lo mismo que 100 tokens, no tener en cuenta los tokens generados al pedir respuestas largas, o usar demasiados ejemplos en un solo prompt.

En el caso de este trabajo, donde se ejecuta un modelo en local para corregir exámenes, saber cuántos tokens ocupan los prompts ayuda a mantener la eficiencia, el optimizar la tokenización permite acelerar el proceso de inferencia, y estimar los tokens permite definir la escalabilidad del sistema para corregir múltiples exámenes al mismo tiempo.

## 2.4 Modelos de lenguaje basados en Transformers

Un modelo Transformer es una arquitectura de redes neuronales diseñada para procesar secuencias de datos (como texto, audio o código) y ha revolucionado el campo del procesamiento de lenguaje natural (PLN). Estas tecnologías han permitido avances significativos en tareas como clasificación de texto, traducción automática, resumen, generación de texto y comprensión semántica. Dentro de este contexto, destacan modelos como BERT, T5 y PaLM, desarrollados por Google, que han marcado hitos por su rendimiento, capacidad de generalización y escalabilidad.

En esta sección se presentan las características principales de estos modelos, su evolución técnica, y su relevancia en el panorama actual de la inteligencia artificial, con el objetivo de contextualizar el enfoque propuesto en este trabajo.

### 2.4.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT es un modelo de lenguaje desarrollado por Google AI en 2018. Su gran innovación fue introducir una forma bidireccional de leer texto usando la arquitectura Transformer, permitiéndole entender mejor el contexto de las palabras. Mientras que modelos previos

leían el texto de izquierda a derecha (como un lector humano), BERT ve el contexto a ambos lados de una palabra, lo que mejora notablemente la comprensión semántica.[12], [13]

BERT usa solo la parte “encoder” de la arquitectura Transformer (la otra parte, el “decoder”, es más usada en generación de texto). El encoder transforma un input textual en una representación profunda del lenguaje natural, básicamente en tres conceptos:

- Entrada: una secuencia de tokens (por ejemplo: "El perro está [MASK] en el jardín").
- Positional Embeddings: Se incluyen posiciones para saber en qué orden van las palabras.
- Output: representaciones vectoriales (embeddings) para cada token y una especial llamada [CLS] para llevar a cabo.

Existen dos distintas formas de entrenarlo:

### 1. Masked Language Modeling (MLM)

Durante el entrenamiento, BERT oculta aleatoriamente algunas palabras del input y le pide al modelo que las prediga.

Ejemplo:

Input: "El gato está [MASK] en el tejado"

Output: "dormido"

Esto obliga al modelo a considerar el contexto completo para predecir la palabra faltante.

### 2. Next Sentence Prediction (NSP)

Se le da un par de oraciones y se le pregunta si la segunda oración sigue lógicamente a la primera.

Ejemplo:

- A: "Fui al cine ayer."
  - B: "Compré palomitas de maíz."
- ¿Es B una continuación lógica de A?: Sí

Los casos de uso más comunes son: Análisis de sentimientos, clasificación de texto, respuesta a preguntas, reconocimiento de entidades (NER) y emparejamiento semántico de frases.

Las principales ventajas de este modelo son la comprensión profunda del lenguaje, la capacidad de pre-entrenado/ reutilización y su capacidad multilingüe, debido a que hay versiones entrenadas en múltiples idiomas. Sin embargo, entre sus desventajas destacan que no puede generar texto (solo lo analiza), requiere mucha memoria y tiempo para el entrenamiento, y no maneja bien entradas muy largas (por defecto, máximo 512 tokens).

#### 2.4.2 T5 (Text-to-Text Transfer Transformer)

T5 fue creado por Google Research en 2019. Su propuesta es simple pero poderosa: convertir cualquier tarea de NLP en un problema de texto a texto. Esto incluye traducción,

resumen, clasificación, preguntas y respuestas, y más. Usa toda la arquitectura Transformer (encoder + decoder), mediante la cual convierte toda entrada en una instrucción textual y espera una respuesta textual.[14]

La siguiente tabla muestra un ejemplo:

Tarea	Input	Output
<b>Traducción</b>	translate English to Spanish: I love cats.	Me encantan los gatos.
<b>Resumen</b>	summarize: El modelo T5 convierte...	T5 convierte tareas en texto.
<b>Clasificación</b>	classify: Este libro es aburrido.	Negativo

Tabla 3: Ejemplo T5

Entre sus puntos a favor destacan su extrema flexibilidad, ya que se emplea una sola arquitectura para todas las tareas, que es fácil de adaptar a tareas personalizadas y que funciona bien en multitarea. Sin embargo, por otro lado, puede ser más lento que modelos especializados para tareas individuales y, como genera texto, a veces puede utilizar información incorrecta si no está bien entrenado.

### 2.4.3 PaLM (Pathways Language Model)

PaLM es una familia de modelos de lenguaje masivos desarrollada por Google en 2022 bajo la arquitectura Pathways, diseñada para permitir que un solo modelo resuelva muchas tareas simultáneamente. Tiene una arquitectura basada en Transformer decoder (como GPT), está entrenado con grandes cantidades de texto y código, y también está perfectamente optimizado para multitarea, razonamiento, resolución de problemas, generación de código, etc.[15]

Ejemplo de tareas que resuelve:

1. Generación de texto creativo
2. Explicación de conceptos
3. Resolución de problemas matemáticos complejos
4. Traducción automática de alta precisión
5. Programación y depuración de código

Un ejemplo de interacción (con modelo tipo GPT) sería:

Entrada: "Explícame qué es una red neuronal como si tuviera 12 años."

Salida: "Imagina que tu cerebro está hecho de miles de luces pequeñas llamadas neuronas..."

Las ventajas de PaLM que más destacan son su capacidad multitarea y multimodal (texto, imágenes, código), también es escalable a problemas muy complejos y posee excelentes capacidades de razonamiento.

Como desventajas hay que destacar que no es de código abierto directamente (aunque Gemini sí lo usa), es difícil de ejecutar localmente (requiere infraestructura avanzada) y, como otros LLM, puede generar errores si no está afinado para una tarea específica.

A continuación, en la tabla, se muestra una comparativa a modo de resumen de las principales diferencias entre los tres modelos:

Característica	BERT	T5	PaLM
----------------	------	----	------

<b>Arquitectura</b>	Encoder (Transformer)	Encoder Decoder	+	Decoder (Transformer)
<b>Entradas</b>	Texto	Texto	con	Texto
<b>Salidas</b>	Embeddings clasificación	Texto		Texto
<b>Capacidad generativa</b>	No	Sí		Sí
<b>Multitarea</b>	Limitada	Alta		Muy alta
<b>Tamaño típico</b>	~110M - 340M	~220M - 11B		Hasta 540B
<b>Ideal para</b>	Comprensión del lenguaje	Traducción, resumen, QA		Generación, razonamiento, código

Tabla 4: Comparación BERT, T5 y PaLM

## 2.5. Beneficios de Usar IA en Local

La ejecución de modelos de inteligencia artificial en local, es decir, en dispositivos físicos propios sin necesidad de conexión a servidores externos, representa una tendencia creciente impulsada por avances tecnológicos en hardware y la creciente disponibilidad de modelos optimizados. Esta estrategia ofrece múltiples beneficios en términos de privacidad, eficiencia económica, sostenibilidad y personalización, especialmente en contextos educativos y de evaluación automatizada.[16]

### 2.5.1. Privacidad y seguridad

La principal ventaja de utilizar IA de forma local es, sin duda alguna, la privacidad que te proporciona debido a que no estás conectado a un servidor el cual accede a tus datos y guarda los resultados obtenidos.

Uno de los beneficios más inmediatos de utilizar modelos de IA en local es la independencia de servicios externos. Cuando un modelo se ejecuta en un entorno local, no es necesario enviar datos a través de internet a servidores de terceros, como ocurre con APIs de OpenAI, Google o Microsoft. Esto no solo reduce la latencia, sino que elimina riesgos asociados a la exposición involuntaria de datos en tránsito.

Además, la autonomía que ofrece el uso local resulta clave en sectores como la educación, la salud o la administración pública, donde la protección de datos es una prioridad legal y ética.[17]

En el contexto del presente trabajo, la evaluación de exámenes tipo test implica manejar información sensible relacionada con el rendimiento de estudiantes. El uso de IA en local garantiza que estos datos no abandonan la institución o el equipo físico, evitando vulneraciones de privacidad.

Además, se eliminan problemas derivados del uso de herramientas basadas en la nube que podrían almacenar datos sin consentimiento explícito o usarlos para entrenamiento posterior, como ocurre con algunos servicios de IA en línea.[18]

## 2.5.2. Coste y eficiencia

El uso de modelos de IA en la nube suele implicar un coste recurrente por uso, ya sea en forma de suscripciones, tarifas por token, o límites de acceso. En contraste, un modelo ejecutado localmente tiene un coste inicial de implementación, pero a partir de ese momento su uso es totalmente gratuito, sin depender de conexiones o tarifas variables.

Por ejemplo:

- OpenAI cobra entre \$0.0015 y \$0.12 por 1.000 tokens dependiendo del modelo utilizado.
- Hugging Face Inference Endpoints ofrece acceso limitado y pagos mensuales a partir de ciertos niveles de uso.

Una solución local elimina estas barreras, permitiendo ejecutar miles de correcciones o tareas de procesamiento de lenguaje natural sin incurrir en costes adicionales.[19]

Incluso dispositivos con capacidad limitada permiten utilizar la mayoría de los modelos, si no los más grandes, gracias a la variedad de modelos disponibles de manera local se puede escoger los que mejor se adaptan a las especificaciones requeridas utilizando el espacio disponible de manera más eficiente.

El uso de IA en local también favorece la inclusión digital y la sostenibilidad tecnológica. Centros educativos con recursos limitados pueden utilizar equipos ya disponibles, evitando la dependencia de infraestructuras en la nube o licencias caras. Además, se reduce el consumo energético asociado a los data centers y tráfico continuo por internet, lo que representa una mejora desde el punto de vista medioambiental.

## 2.6. Personalización de IA local

Un aspecto muy importante de los modelos de IA locales, es la capacidad de personalizar el entorno en el que se va a utilizar, al gusto del usuario, ya no solo modificar el estilo de la aplicación, si no también dar un rol al modelo o unas indicaciones.

Cuando se trabaja con modelos en la nube, muchas veces se está limitado por las políticas, APIs y restricciones de la plataforma. Sin embargo, un modelo en local puede ser ajustado, modificado o afinado según las necesidades específicas del usuario o institución.

Algunos ejemplos de personalizaciones incluyen:

- Cambiar instrucciones por defecto (prompting directo).
- Añadir conocimiento contextual.
- Incorporar nuevas plantillas o formatos específicos de evaluación.
- Ejecutar código complementario para tareas específicas.

En una última actualización, LM Studio incluyó una opción de System Prompt mediante el cual se puede ajustar el modelo que está siendo utilizado según las preferencias del usuario. Este prompt se aplica a toda la sesión, aunque se puede modificar en cualquier momento. En algunos modelos, el System Prompt tiene más efecto que en otros. Modelos como GPT-4, Mistral-Instruct o LLaMA-3-Instruct responden bien a estas instrucciones. También es posible que algunos modelos pequeños (no instruct) puedan ignorarlo si no están entrenados para seguirlo.

En la Fig1 se puede ver cómo se le indica al modelo el rol que debe cumplir especificando el idioma y el estilo de respuestas. A la derecha se puede observar cómo está guardado como “Preset 1” y la posibilidad de ajustar los parámetros del modelo mediante cuatro desplegados.

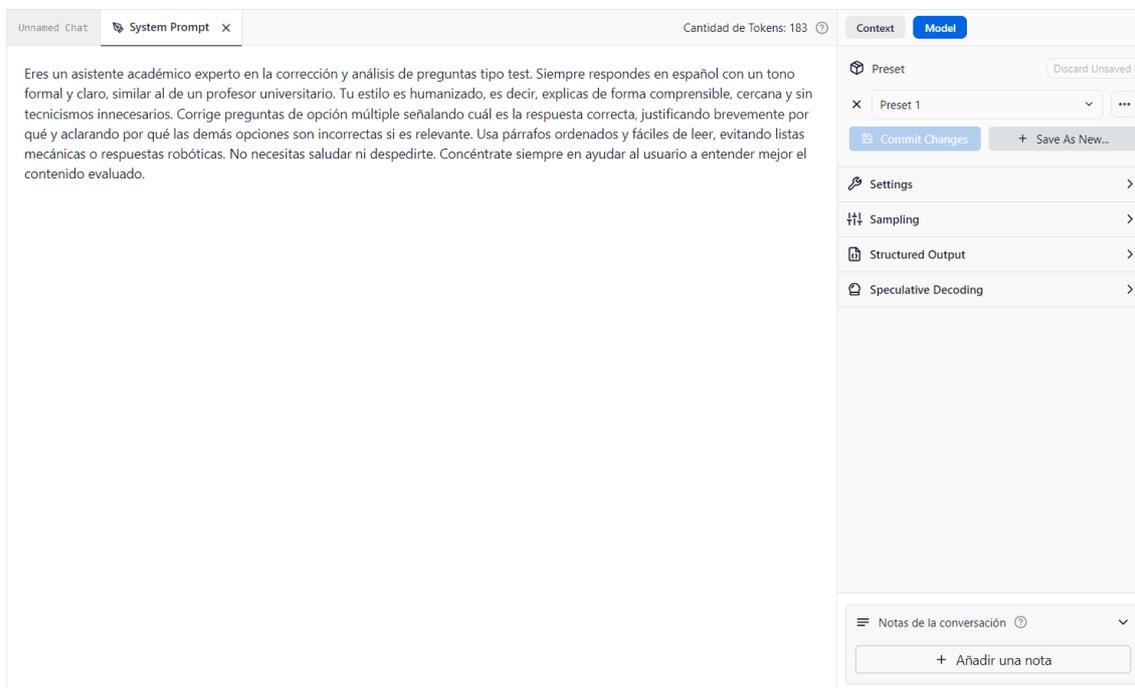


Fig.1: Panel System Prompt

En el primer desplegable, mostrado en la Fig2, se aprecia la posibilidad de variar la temperatura, la cual controla el nivel de aleatoriedad o creatividad en las respuestas que genera el modelo, cuanto más cercano a 1 sea el nivel de temperatura con más aleatoriedad responderá el modelo. También está la posibilidad de limitar la longitud de respuesta y modificar el desbordamiento de cadenas o los hilos de CPU.

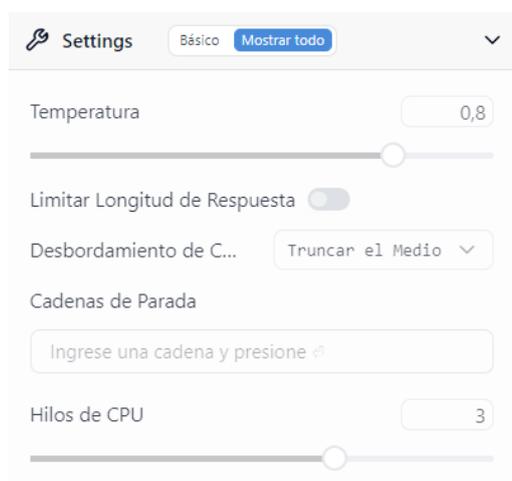


Fig2: Settings

En el desplegable de Sampling se puede variar el muestreo Top K, el cual limita las elecciones del modelo a las K palabras más probables en cada paso. En el caso de la Fig3, si  $K=40$ , el modelo solo considera las 40 palabras con mayor probabilidad. Por lo tanto, este parámetro reduce aleatoriedad extrema. Cuanto más bajo, más conservador.

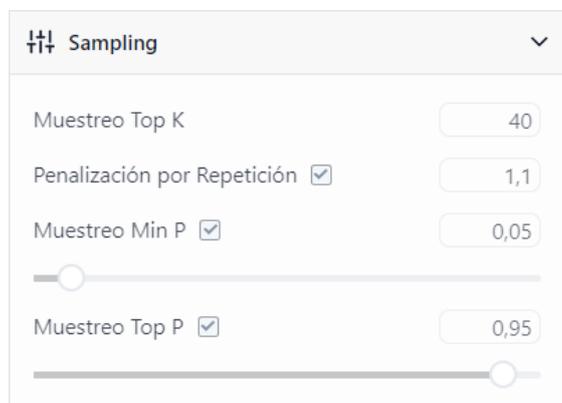


Fig3: Sampling

Por su parte, la penalización por repetición evita que el modelo repita palabras o frases. Números como 1.1 o 1.2 son buenos para reducir repeticiones molestas.

El muestreo Min P sirve para que el modelo pueda ignorar palabras con muy poca probabilidad acumulada. Un valor bajo como 0.05 significa que se eliminan palabras muy poco probables, ayudando a mantener coherencia.

Por último, el muestreo Top P sirve para que el modelo elija palabras dentro de un conjunto cuya probabilidad acumulada llegue a P. Si  $\text{Top P} = 0.95$  como en la imagen, se consideran solo las palabras más probables hasta que la suma de probabilidades llegue al 95%.

En el tercer desplegable, mostrado en la Fig4 aparece la opción de activar la salida estructurada, la cual permite que el modelo genere respuestas en un formato estructurado, como JSON, siguiendo un esquema definido. Es ideal para integraciones con sistemas que requieren datos en un formato específico.

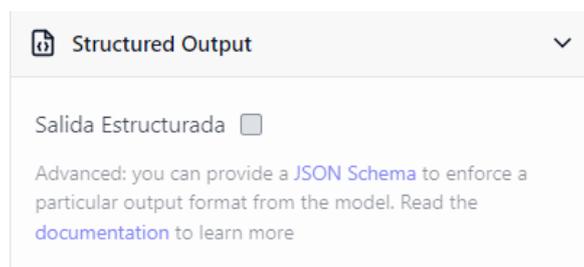


Fig4: Structured Output

En el último desplegable, Fig5, se puede ver la opción Draft Model, el cual permite elegir un modelo preliminar que genera posibles continuaciones antes de que el modelo principal las valide. En Min Draft Size, se puede variar el tamaño mínimo de los fragmentos generados por el modelo preliminar. Drafting Probability Cutoff indica el umbral de probabilidad para aceptar un token generado por el modelo preliminar. Max Draft Size indica el tamaño máximo de los fragmentos generados por el modelo

preliminar. Y, por último, Visualize accepted draft tokens muestra visualmente qué tokens fueron aceptados del modelo preliminar, útil para depuración o análisis.

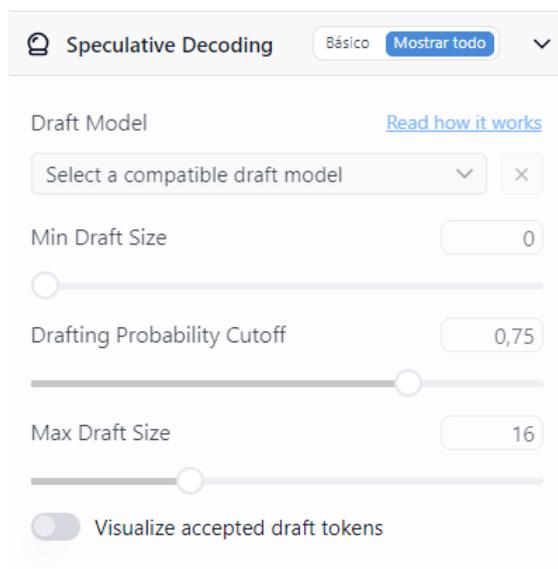


Fig5: Speculative Decoding

Además de las opciones ajustables de forma directa, hay que destacar la capacidad de especializar el modelo en una tarea concreta, mediante ajustes ligeros o entrenamiento adicional sobre datos de ejemplo (fine-tuning o prompting personalizado). Esto permite convertir un modelo generalista en un asistente experto en, por ejemplo, tareas educativas, aumentando la precisión y reduciendo la necesidad de reexplicar tareas.[20]

Además, de esta forma, se permite crear soluciones híbridas donde el modelo se comunica con scripts auxiliares que pueden: validar respuestas, generar informes automáticos o calificar de acuerdo con distintos criterios.

La Fig. 6 muestra la evolución del tamaño de los modelos lingüísticos (LLMs) a lo largo del tiempo, específicamente en términos del número de parámetros (en miles de millones), desde 2017 hasta 2024. [16] Los círculos celestes representan a los modelos lingüísticos grandes (LLMs) como GPT-3, GPT-4, Llama, etc. Mientras que los círculos azules representan a los modelos lingüísticos pequeños (SLMs) como Phi-3, Gemma 2B, Granite 2B, etc.

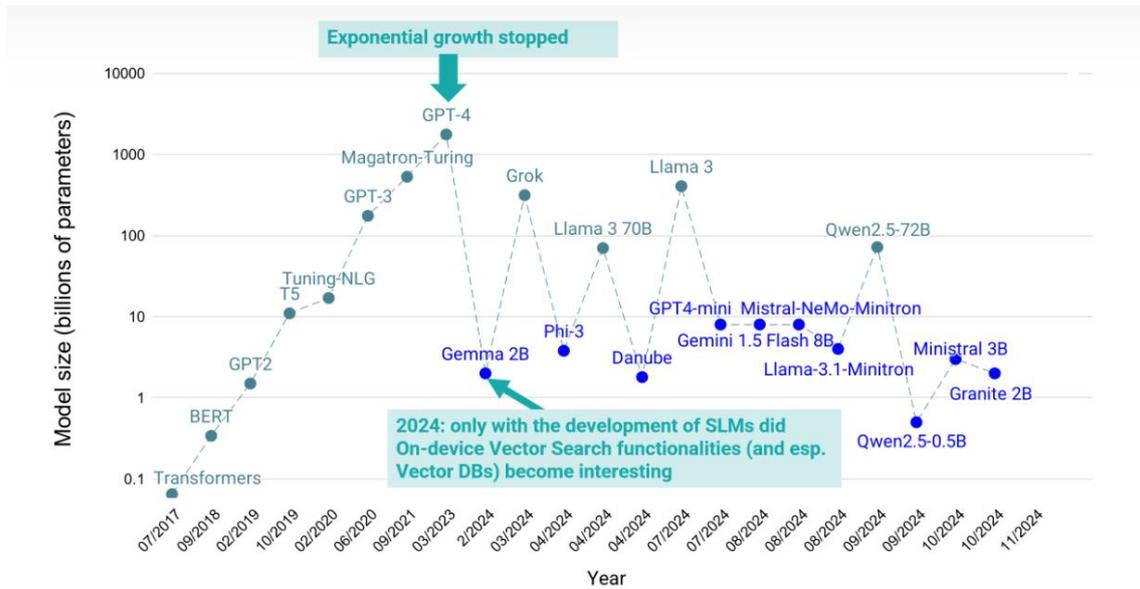


Fig6: Evolución del tamaño de los modelos de lenguaje con el tiempo. Los modelos de lenguaje. [16]

Se observa cómo la IA experimentó un crecimiento exponencial hasta GPT-4 (2023), pasando rápidamente de millones a más de 100 mil millones de parámetros. A partir de ahí, los modelos han ido optimizándose, ya que la industria ha cambiado de enfoque: en lugar de hacer modelos cada vez más grandes, se ha priorizado optimización, eficiencia y modelos más pequeños. En 2024 comenzó el auge de los SLMs (Small Language Models) y, a partir de 2024, se observa un aumento en modelos pequeños como Phi-3, Danube, Gemini 1.5 Flash 8B, Granite 2B, etc.



## **3. Desarrollo práctico**

En esta sección se detalla la configuración del entorno experimental, así como las pruebas realizadas para evaluar el rendimiento de distintos modelos de lenguaje ejecutados localmente mediante la plataforma LM Studio. El foco principal será la corrección de exámenes tipo test.

### **3.1. Configuración del entorno**

Para ejecutar los modelos de lenguaje de gran tamaño de forma local, es necesario contar con una configuración de hardware adecuada que permita un rendimiento fluido sin comprometer la estabilidad del sistema. En este caso, el equipo utilizado para las pruebas posee las siguientes especificaciones:

- **Procesador (CPU):** Intel(R) Core(TM) i7-1065G7
- **Memoria RAM:** 16 GB
- **Almacenamiento:** SSD de 512 GB
- **Sistema operativo:** Windows 10

La plataforma seleccionada para el despliegue de los modelos ha sido LM Studio, la cual es una aplicación de escritorio que permite ejecutar modelos de lenguaje locales directamente en un ordenador personal, sin necesidad de conexión a internet, ni servidores externos. Es una herramienta ideal para desarrolladores e investigadores de la IA que

desean privacidad, control y flexibilidad al usar modelos de lenguaje. LM Studio permite lanzar inferencias sobre los modelos seleccionados y ejecutar pruebas de manera interactiva o mediante API local, lo que facilita su integración con herramientas externas.

Su gran ventaja es su simplicidad, especialmente su instalación y puesta en marcha, que se resume en 4 simples pasos:

1. Descarga desde el sitio oficial: <https://lmstudio.ai>
2. Instalación del entorno en disco local y configuración del acceso a los recursos de la GPU.
3. Descarga de modelos mediante la conexión directa al repositorio de Hugging Face, con soporte para modelos en formato GGUF.
4. Pruebas de funcionamiento y ejecución en modo local utilizando distintos tamaños y cuantizaciones.

Una vez instalado, la elección de los modelos a evaluar se basó en un análisis de sus características técnicas, objetivos de entrenamiento y adecuación al entorno local. A continuación, se presentan los criterios y motivos que justifican la selección de cada uno:

- LLaMA 3 (Meta AI): Es un modelo de lenguaje grande desarrollado por Meta, diseñado para tareas de generación de texto, finalización de chat y aplicaciones agenticas. LLAMA3 incluye modelos preentrenados y ajustados por instrucciones, con tamaños que van desde 8 mil millones hasta 70 mil millones de parámetros.
  - Ha sido seleccionado por su alta capacidad en tareas de comprensión y generación de texto. Es una de las arquitecturas más potentes disponibles públicamente, con buen rendimiento en tareas complejas de PLN.
- Qwen 2 (Alibaba DAMO Academy): Es la evolución del modelo Qwen1.5, que soporta 27 idiomas y puede realizar diversas tareas. QWEN2 ofrece modelos preentrenados y ajustados por instrucciones en cinco tamaños diferentes, con mejoras significativas en el rendimiento, la longitud del contexto y la alineación con los valores humano
  - Elegido por su excelente desempeño multilingüe, capacidad de adaptación a distintos contextos y una arquitectura equilibrada. Es especialmente útil en tareas donde se requiere diversidad lingüística o interpretación flexible de preguntas.
- Phi 4 (Microsoft Research): Desarrollado por Microsoft, PHI4 es un modelo de lenguaje pequeño con 14 mil millones de parámetros, especializado en razonamiento complejo, especialmente en matemáticas. Está entrenado con una mezcla de datos sintéticos, públicos y académicos, y se destaca por su capacidad de razonamiento lógico y matemático
  - Destaca por su eficiencia. Ha sido optimizado con datos sintéticos de alta calidad, mostrando resultados competitivos en tareas de razonamiento y educación. Su bajo requerimiento de recursos respecto a los modelos más grandes lo hace ideal para entornos con hardware limitado.

En resumen, si se busca un modelo potente para procesamiento de lenguaje natural avanzado, la mejor opción es LLaMA 3. Si se necesita un modelo multilingüe y versátil, se debe optar por Qwen 2 y si se prefiere un modelo eficiente en recursos, se debe elegir Phi-4.

## 3.2. Definición de pruebas realizadas

Con el objetivo de evaluar de forma rigurosa el rendimiento de los modelos seleccionados (LLaMA 3, Qwen 2 y Phi 4), se completarán varias tareas específicas centradas en el procesamiento del lenguaje natural. Estas pruebas se orientan tanto a medir la capacidad general de cada modelo, como a analizar su aplicabilidad directa en el contexto de corrección automatizada de exámenes tipo test. Los resultados se documentan mediante capturas de pantalla de las interacciones con la IA, lo que permite visualizar tanto sus aciertos como sus errores, así como la comparación directa entre modelos.

Se ha definido el siguiente diagrama de flujo (Fig. 7):

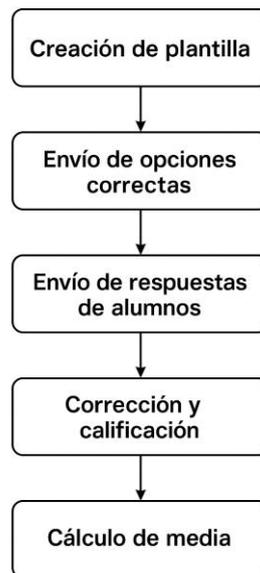


Fig7: Diagrama de flujo

A los diferentes modelos se les ha pedido que completen cada uno de los pasos indicados, realizando una evaluación de los resultados para cada caso.

### 3.2.1 Generación automática de plantillas

La tarea solicitada es la de generar una plantilla para un examen tipo test de 10 preguntas y 4 opciones. Para ello se emplea el siguiente prompt:

You

Necesito que me generes una plantilla para un tipo test con 10 preguntas y 4 opciones cada una (A,B,C o D). Necesito que tenga casillas en las que se pueda marcar con una "X" la opción que se desea seleccionar dentro de dos corchetes []. La plantilla no debe contener preguntas ni soluciones, solo el numero de pregunta en cada fila y la opcion en cada columna



Fig. 8: Prompt para generación de plantilla

El modelo Qwen generó correctamente una plantilla, que muestra la Fig. 9, que cumplía las especificaciones solicitadas.

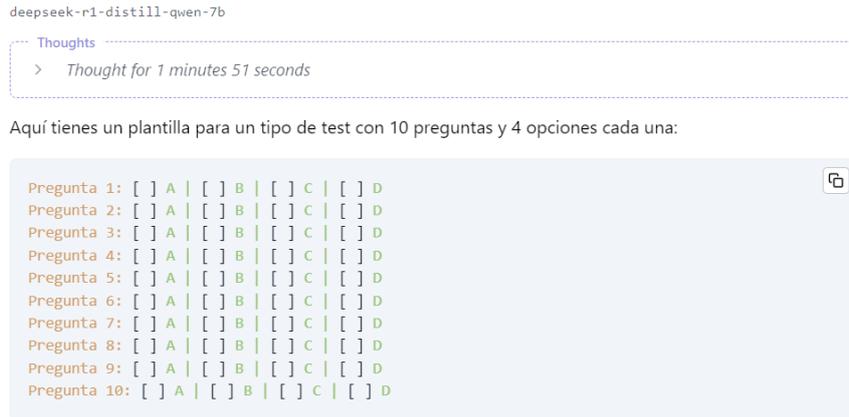


Fig. 9: Plantilla generada por Qwen  
Llama generó una plantilla muy similar:

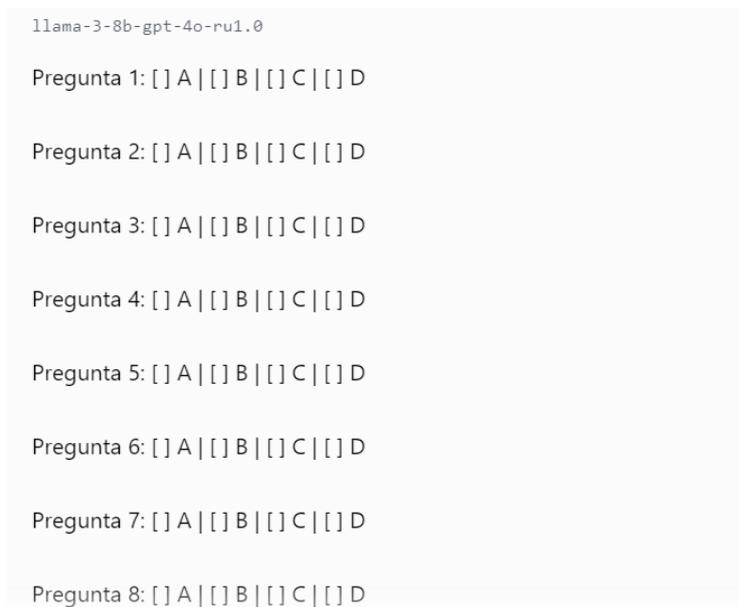


Fig. 10: Plantilla generada por Llama3

Y, por último, Phi-4 creó una plantilla, que resultó ser la más clara e intuitiva.

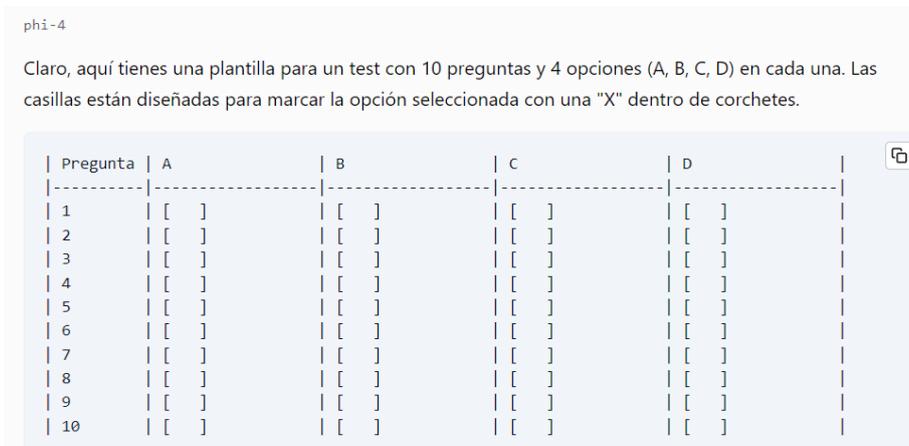


Fig. 11: Plantilla generada por Phi-4

Las diferencias no son sustanciales, pero esta tarea, aunque resulta la más sencilla de todas las que se van a analizar, veremos que tienen alta repercusión en los resultados de los procesos posteriores.

### 3.2.2. Corrección automatizada de exámenes tipo test

Esta es la tarea principal del trabajo. En este paso se le proporciona al modelo tanto las plantillas de corrección, como las respuestas de estudiantes. Para valorar los resultados se evalúan tanto la identificación correcta de aciertos y errores, como el cálculo correcto de la puntuación final, así como la capacidad para detectar inconsistencias o preguntas en blanco. Cabe destacar que, aunque las pruebas realizadas han sido ejecutadas en los tres modelos de IA local indicados en el apartado anterior, en el texto solamente se muestran los resultados correspondientes al modelo Phy-4, que es el que más adelante se justifica como es el que finalmente resulta con mejores prestaciones de entre todos.

El proceso puede realizarse de diferentes formas, por lo que se probaron varias de ellas:

#### 1. Corrección de test por teclado.

Para esta conversación, se usó un modelo en línea, denominado Mistral, que facilita la lectura de documentos. Primero se envían las plantillas a Mistral y una vez las lee correctamente, se pasan por teclado a la IA entrenada en local.

Primero se envía el documento con las opciones correctas que son las indicadas en la Fig. 12:



#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	[X]	[ ]	[ ]	[ ]
2	[Pregunta 2]	[ ]	[X]	[ ]	[ ]
3	[Pregunta 3]	[ ]	[ ]	[X]	[ ]
4	[Pregunta 4]	[ ]	[ ]	[ ]	[X]
5	[Pregunta 5]	[ ]	[ ]	[ ]	[X]
6	[Pregunta 6]	[ ]	[ ]	[X]	[ ]
7	[Pregunta 7]	[ ]	[X]	[ ]	[ ]
8	[Pregunta 8]	[X]	[ ]	[ ]	[ ]
9	[Pregunta 9]	[X]	[ ]	[ ]	[ ]
10	[Pregunta 10]	[ ]	[X]	[ ]	[ ]

Fig 12: (izquierda) Interfaz web de Mistral / (derecha) Ejemplo de plantilla con soluciones

En la Fig. 13 se observa como Mistral lee perfectamente la plantilla:

Page 1 380 tokens processed

#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	[Pregunta 2]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	[Pregunta 3]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4	[Pregunta 4]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	[Pregunta 5]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
6	[Pregunta 6]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
7	[Pregunta 7]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	[Pregunta 8]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	[Pregunta 9]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	[Pregunta 10]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Fig. 13: Resultado obtenido por Mistral

La salida de Mistral es la que se envía directamente al modelo local basado en Phi-4, al que se le pide que compruebe cuáles son las preguntas correctas, de acuerdo con el prompt que se muestra en la Fig 14:

Dime que opciones estan seleccionadas en esta plantilla para un tipo test con 10 preguntas y 4 opciones (A,B,C o D)

Pregunta	Opción A	Opción B	Opción C	Opción D
1 [Pregunta 1]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2 [Pregunta 2]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 [Pregunta 3]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4 [Pregunta 4]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
5 [Pregunta 5]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
6 [Pregunta 6]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
7 [Pregunta 7]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8 [Pregunta 8]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9 [Pregunta 9]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10 [Pregunta 10]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

En la plantilla proporcionada, las opciones seleccionadas para cada pregunta son:

- Pregunta 1:** Opción A
- Pregunta 2:** Opción B
- Pregunta 3:** Opción C
- Pregunta 4:** Opción D
- Pregunta 5:** Opción D
- Pregunta 6:** Opción C
- Pregunta 7:** Opción B
- Pregunta 8:** Opción A
- Pregunta 9:** Opción A
- Pregunta 10:** Opción B

Estas son las opciones marcadas con una "X" en cada pregunta.

Fig. 14: Prompt y resultado de aplicar la plantilla de respuestas correctas en Phi-4

Ahora es el momento de utilizar la plantilla, pero esta vez, utilizada por un alumno, con sus correspondientes respuestas:

Se repite un proceso similar, es decir, se envía a Mistral, Fig: 15:

#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
2	[Pregunta 2]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	[Pregunta 3]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	[Pregunta 4]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	[Pregunta 5]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	[Pregunta 6]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
7	[Pregunta 7]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
8	[Pregunta 8]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	[Pregunta 9]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	[Pregunta 10]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Fig. 15: La plantilla del alumno se envía a Mistral.

Y ahora, se pasa el resultado por el prompt de Phi-4, que de acuerdo con las opciones indicadas en la Fig. 16, el modelo lo reconoce perfectamente, tal como se observa en la Fig. 17.

### Pregunta Opción A Opción B Opción C Opción D

---

- 1 [Pregunta 1]
- 2 [Pregunta 2]
- 3 [Pregunta 3]
- 4 [Pregunta 4]
- 5 [Pregunta 5]
- 6 [Pregunta 6]
- 7 [Pregunta 7]
- 8 [Pregunta 8]
- 9 [Pregunta 9]
- 10 [Pregunta 10]

Si la opción del alumno es correcta, suma un punto. Si es errónea resta 0.33. Si no ha seleccionado ninguna opción o hay más de una seleccionada, esa pregunta no puntúa.

Fig. 16: Prompt para Phi-4, con las respuestas del alumno

**Opciones correctas del examen:**

1. Pregunta 1: A
2. Pregunta 2: B
3. Pregunta 3: C
4. Pregunta 4: D
5. Pregunta 5: D
6. Pregunta 6: C
7. Pregunta 7: B
8. Pregunta 8: A
9. Pregunta 9: A
10. Pregunta 10: B

**Evaluación:**

1. Pregunta 1
  - Selección: D (Incorrecta)
  - Puntuación: -0.33
2. Pregunta 2
  - Selección: C (Incorrecta)
  - Puntuación: -0.33
3. Pregunta 3
  - Selección: B (Incorrecta)
  - Puntuación: -0.33

**4. Pregunta 4**

- Selección: B (Incorrecta)
- Puntuación: -0.33

**5. Pregunta 5**

- No seleccionada
- Puntuación: 0

**6. Pregunta 6**

- Selección: C (Correcta)
- Puntuación: +1

**7. Pregunta 7**

- Selección: D (Incorrecta)
- Puntuación: -0.33

**8. Pregunta 8**

- Selección: B (Incorrecta)
- Puntuación: -0.33

**9. Pregunta 9**

- Selección: B (Incorrecta)
- Puntuación: -0.33

**10. Pregunta 10**

- Selección: B (Correcta)
- Puntuación: +1

**Cálculo total:**

- Suma de puntos por respuestas correctas:  $1 + 1 = 2$
- Resta de puntos por respuestas incorrectas:  $(-0.33) \times 7 = -2.31$

Puntuación final:  $2 - 2.31 = -0.31$

**Interpretación de la puntuación:**

Dado que el resultado es menos de 5, el alumno tiene un **suspenso**.

Fig. 17: Resultados obtenidos por el alumno según el modelo Phi-4

El proceso se puede repetir con las respuestas de otros alumnos, como se muestra en la Fig. 18:

### Pregunta Opción A Opción B Opción C Opción D

```

1 [Pregunta 1] [X] [[]][[]]
2 [Pregunta 2] [[]][X] [[]]
3 [Pregunta 3] [[]][X] [[]]
4 [Pregunta 4] [[]][[]][X]
5 [Pregunta 5] [[]][[]][[]]
6 [Pregunta 6] [[]][X] [[]]
7 [Pregunta 7] [[]][X] [[]]
8 [Pregunta 8] [X] [[]][[]]
9 [Pregunta 9] [[]][[]][[]]
10 [Pregunta 10] [[]][X] [[]]

```

Si la opción del alumno es correcta, suma un punto. Si es errónea resta 0.33. Si no ha seleccionado ninguna opción o hay más de una seleccionada, esa pregunta no puntúa.

Una vez obtenida una puntuación, si es menor que 5 debes indicar que es un suspenso. Si la nota está comprendida entre 5 y 5.99 es un suficiente. Si está entre 6 y 6.99 es un bien. Si está entre 7 y 8.99 es un notable. Y si es mayor o igual que 9 es un sobresaliente.

#### Evaluación:

- Preguntas Correctas: 1, 2, 3, 4, 6, 7, 8, 10 (8 preguntas correctas)
- Puntos obtenidos por respuestas correctas:  $8 * 1 = 8$  puntos
- Preguntas Incorrectas o No Puntuadas:
  - Pregunta 5: Ninguna seleccionada (0 puntos)
  - Pregunta 9: Ninguna seleccionada (0 puntos)

Dado que no hay preguntas incorrectas, el descuento de puntos es cero.

#### Puntuación Total:

- Puntos Totales: 8

Fig. 18: Prompt y resultados obtenidos con una segunda plantilla de respuestas

Como se puede apreciar corrige perfectamente.

## 2. Corrección de test mediante documentos

En este caso, las plantillas se envían en forma de pdf, directamente al modelo local Phi-4, acompañadas del prompt. El procedimiento es muy similar al anterior, esto es, se introduce primero la plantilla de corrección, junto con el prompt de la Fig. 19, con el resultado mostrado en la Fig. 20.

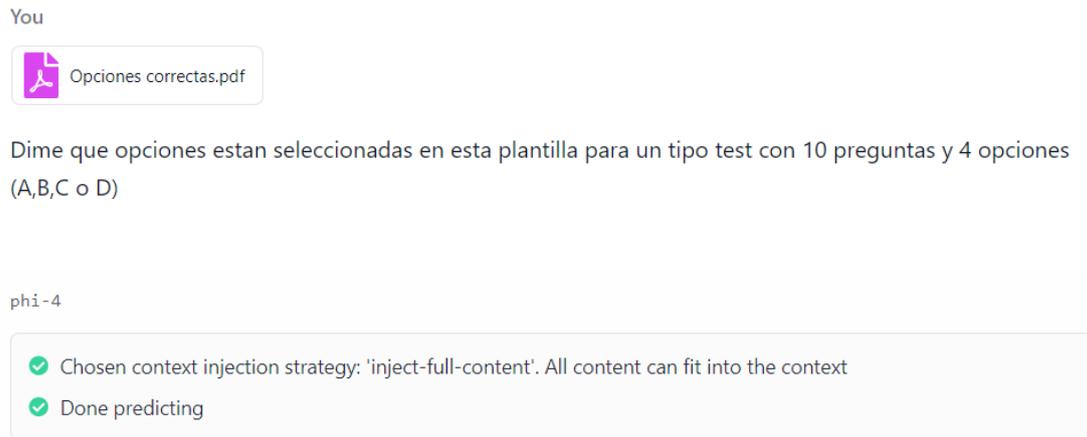


Fig. 19: Prompt para Phi-4 incluyendo PDF de plantilla de corrección

Basado en el contenido proporcionado en el archivo "Opciones correctas.pdf", las opciones seleccionadas para cada pregunta en la plantilla son las siguientes:

1. **Pregunta 1:** Opción A está marcada.
2. **Pregunta 2:** Opción B está marcada.
3. **Pregunta 3:** Opción C está marcada.
4. **Pregunta 4:** Opción D está marcada.
5. **Pregunta 5:** Opción D está marcada.
6. **Pregunta 6:** Opción C está marcada.
7. **Pregunta 7:** Opción B está marcada.
8. **Pregunta 8:** Opción A está marcada.
9. **Pregunta 9:** Opción A está marcada.
10. **Pregunta 10:** Opción B está marcada.

Fig. 20: Resultado de identificar la plantilla para iniciar la corrección

Una vez cargada la plantilla de corrección, se introduce la plantilla de respuestas del alumno, junto con el prompt indicado en la Fig. 21, muy parecido al del método anterior, aunque adaptado a esta nueva circunstancia.

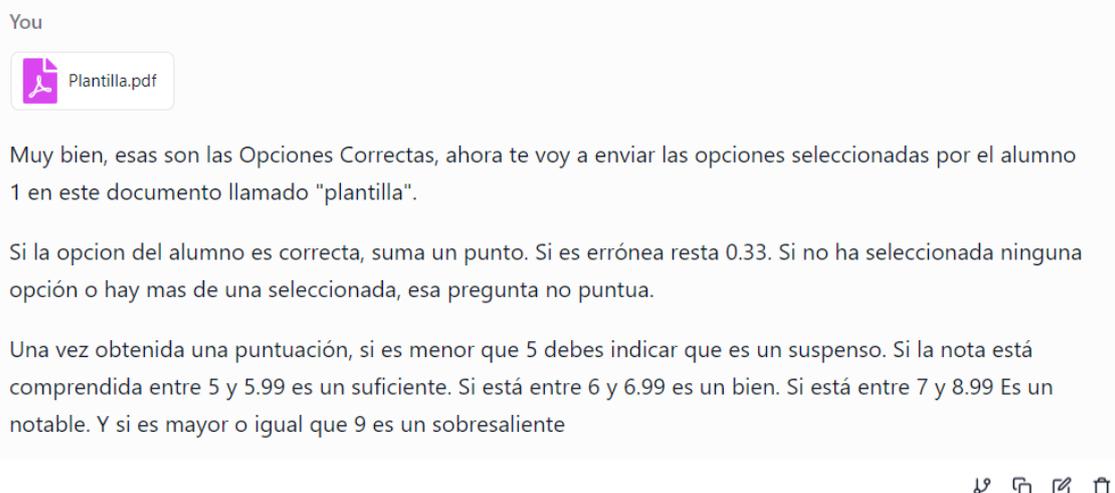


Fig. 21: Prompt para corregir la plantilla PDF con las respuestas del alumno

En la Fig. 22 se observan los resultados obtenidos por el modelo:

phi-4

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Para evaluar las respuestas del alumno según la plantilla proporcionada, primero vamos a identificar cuáles de sus opciones coinciden con las correctas:

1. **Pregunta 1:** Alumno seleccionó A - Correcta: +1 punto.
2. **Pregunta 2:** Alumno seleccionó B - Correcta: +1 punto.
3. **Pregunta 3:** Alumno seleccionó C - Correcta: +1 punto.
4. **Pregunta 4:** Alumno seleccionó D - Correcta: +1 punto.
5. **Pregunta 5:** No ha seleccionado ninguna opción o hay más de una seleccionada - Sin puntuación.
6. **Pregunta 6:** Alumno seleccionó C - Correcta: +1 punto.
7. **Pregunta 7:** Alumno seleccionó B - Correcta: +1 punto.
8. **Pregunta 8:** Alumno seleccionó A - Correcta: +1 punto.
9. **Pregunta 9:** No ha seleccionado ninguna opción o hay más de una seleccionada - Sin puntuación.
10. **Pregunta 10:** Alumno seleccionó B - Correcta: +1 punto.

El alumno ha respondido correctamente a todas las preguntas donde hizo una selección válida, obteniendo un total de 8 puntos.

Ahora determinaremos el calificativo basado en la puntuación:

- Puntos obtenidos: 8

Dado que los puntos están entre 7 y 8.99, se considera un **Notable**.

En resumen:

- **Puntuación:** 8
- **Calificación:** Notable

1.69 tok/sec • 353 tokens • 220.99s to first token • Stop reason: EOS Token Found

Fig. 22: Resultados del modelo Phi para el primer alumno

Al probar con el segundo alumno, ya no es necesario repetir todo el prompt anterior, como se muestra en la Fig. 23, obteniendo el nuevo resultado de la Fig. 24.

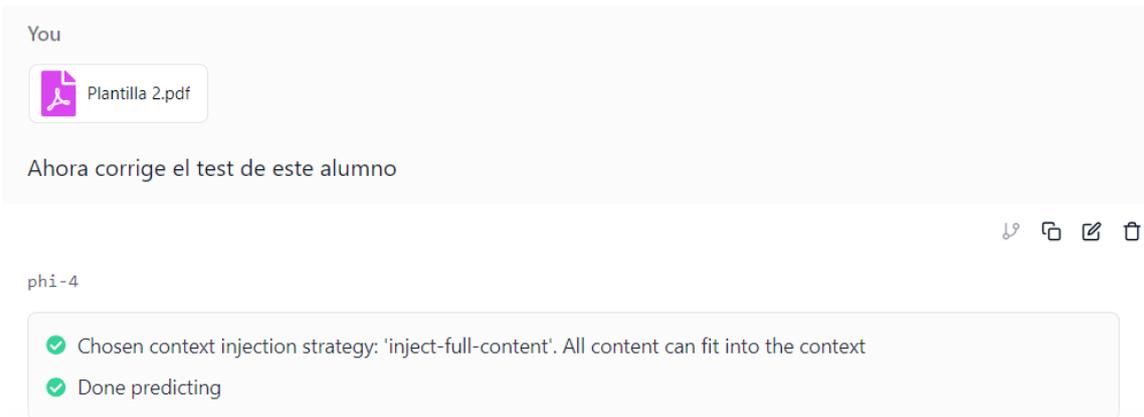


Fig. 23: Prompt utilizado para el segundo alumno

Para corregir las respuestas del alumno según la "Plantilla 2.pdf" y compararlas con las opciones correctas proporcionadas en "Opciones correctas.pdf", primero identifiquemos qué opción seleccionó el alumno en cada pregunta:

#### Plantilla 2.pdf:

1. **Pregunta 1:** Opción D (Incorrecto)
2. **Pregunta 2:** Opción C (Incorrecto)
3. **Pregunta 3:** Opción B (Incorrecto)
4. **Pregunta 4:** Opción D (Correcto)
5. **Pregunta 5:** Opción D (Correcto)
6. **Pregunta 6:** Opción C (Correcto)
7. **Pregunta 7:** No se seleccionó ninguna opción (0 puntos)
8. **Pregunta 8:** Opción C (Incorrecto)
9. **Pregunta 9:** Opción B (Incorrecta)
10. **Pregunta 10:** Opción B (Correcto)

#### Cálculo de la Puntuación:

- **Correctas:** Preguntas 4, 5, 6 y 10 son correctas, por lo que sumamos +1 cada una.
- **Incorrectas:** Para las preguntas 1, 2, 3, 8 y 9 (5 en total), restamos 0.33 cada una.
- No se seleccionó ninguna opción para la pregunta 7, por lo que no suma ni resta puntos.

#### Cálculo Total:

- Puntos positivos:  $1 \times 4 = 4$
- Puntos negativos:  $-0.33 \times 5 = -1.65$

Puntuación total:  $4 - 1.65 = 2.35$  (redondeando a 2.36, ya que no puede ser fracción de punto).

Dado que la puntuación es menor que 5, la calificación final sigue siendo un **Suspenso**.

Fig. 24: Resultados obtenidos para el segundo alumno

Una vez que se ha comprobado la capacidad de corregir exámenes de ambas formas, se puede ver que la manera más óptima y veloz es enviar directamente las plantillas en forma de documento.

### 3. Corrección de test creados por la propia IA

Una característica realmente interesante de los actuales modelos de IA es su capacidad de generar directamente las preguntas de test. La pregunta es si, ya que son capaces de generar un test completo, si serán capaces también de corregirlos utilizando las plantillas.

En este caso, en la Fig. 25 se muestra el prompt mediante el cual se genera un test de 10 preguntas con 4 opciones sobre contenidos generales de Física, así como un extracto de las preguntas resultantes:

You

Generame un test sobre física de 10 preguntas y 4 opciones

phi-4

¡Claro! Aquí tienes un pequeño test de física con 10 preguntas, cada una con cuatro opciones. Intenta responderlas lo mejor que puedas.

#### Pregunta 1

¿Cuál es la fórmula para calcular la energía cinética de un objeto?

- A)  $E_k = mgh$
- B)  $E_k = \frac{1}{2}mv^2$
- C)  $E_k = mc^2$
- D)  $E_k = Fd$

...

#### Pregunta 10

El concepto de "inercia" está más directamente relacionado con:

- A) La fuerza necesaria para cambiar la velocidad de un objeto.
- B) La resistencia a la deformación bajo tensión.
- C) El impulso dado por una reacción en cadena química.
- D) La masa y su tendencia a mantener el estado de reposo o movimiento uniforme.

¡Buena suerte con el test! Si necesitas las respuestas, aquí están:

1. B)  $E_k = \frac{1}{2}mv^2$
2. B) Segunda Ley de Newton
3. B) La suma de los voltajes a través de las resistencias es igual al voltaje total.
4. D) Un material que curva el camino de la luz
5. A) Presión atmosférica
6. B) La velocidad de la luz en el vacío
7. B) Pascal
8. A) Una gama de longitudes de onda que incluye luz visible e infrarrojo.
9. D) Energía térmica
10. A) La fuerza necesaria para cambiar la velocidad de un objeto.

Fig. 25: Generación de un test sobre física

A continuación, en la Fig. 26, se le pasa una plantilla de respuestas del primer alumno

You



Un alumno seleccionó las siguientes opciones, cual sería su calificación

phi - 4

**Primer Alumno**

**Respuestas:**

1. Opción A (Incorrecto)
2. Opción B (Correcto)
3. Opción C (Incorrecto)
4. Opción D (Correcto)
5. No respondida
6. Opción C (Incorrecto)
7. Opción B (Correcto)
8. Opción A (Correcto)
9. No respondida
10. Opción B (Incorrecto)

**Calificación:**

- Correctas: 4 (Preguntas 2, 4, 7 y 8)
- Incorrectas: 3 (Preguntas 1, 3, 6 y 10)

$$\text{Calificación} = 4 - (0.33 \times 4) = 4 - 1.32 = 2.68$$

La calificación máxima es de 10, así que:

$$\frac{2.68}{10} \times 100\% = 26.8\%$$

Fig. 26: Corrección de la plantilla de respuestas del primer alumno

Se repite el proceso con la segunda plantilla de respuestas, indicada en la Fig. 27.

**Segundo Alumno**

**Respuestas:**

1. Opción D (Incorrecto)
2. Opción C (Correcto)
3. Opción B (Correcto)
4. Opción B (Incorrecto)
5. No respondida
6. Opción C (Incorrecto)
7. Opción D (Incorrecto)
8. Opción B (Incorrecto)
9. Opción B (Incorrecto)
10. Opción B (Incorrecto)

**Calificación:**

- Correctas: 2 (Preguntas 2 y 3)
- Incorrectas: 7 (Preguntas 1, 4, 6, 7, 8, 9 y 10)

$$\text{Calificación} = 2 - (0.33 \times 7) = 2 - 2.31 = -0.31$$

Dado que no se puede tener una calificación negativa:

Calificación ajustada = 0%

Fig. 27: Corrección de la plantilla de respuestas del segundo alumno

Y se probó con una tercera plantilla:

**Tercer Alumno****Respuestas:**

1. Opción D (Incorrecto)
2. Opción C (Incorrecto)
3. Opción A (Incorrecto)
4. Opción A (Incorrecto)
5. Opción D (Correcto)
6. Opción B (Correcto)
7. Opción D (Incorrecto)
8. Opción C (Incorrecto)
9. Opción A (Incorrecto)
10. Opción B (Incorrecto)

**Calificación:**

- Correctas: 2 (Preguntas 5 y 6)
- Incorrectas: 7 (Preguntas 1, 2, 3, 4, 7, 8 y 9)

$$\text{Calificación} = 2 - (0.33 \times 7) = 2 - 2.31 = -0.31$$

Dado que no se puede tener una calificación negativa:

$$\text{Calificación ajustada} = 0\%$$

Fig. 28: Corrección de la plantilla de respuestas del tercer alumno

Por último, en la Fig. 29, se comprobó que el modelo fuese capaz de calcular la media de la clase

**Media de los Tres Alumnos**

Ahora, calculamos la media de las nuevas calificaciones:

1. Primer alumno: 26.8%
2. Segundo alumno: 0%
3. Tercer alumno: 0%

$$\text{Media} = \frac{26.8+0+0}{3} = \frac{26.8}{3} \approx 8.93\%$$

Por lo tanto, la media de las calificaciones ajustadas es aproximadamente 8.93%.

1.02 tok/sec • 801 tokens • 80800.28s to first token • Stop reason: EOS Token Found

Fig. 29: Resultados del conjunto de alumnos evaluados

**4. Corrección de test de verdadero y falso**

También se comprobó si este modelo era capaz de corregir preguntas de verdadero o falso, para lo que se usa una nueva plantilla, mostrada en la Fig. 30, y se duplican el número de preguntas.

Número de Pregunta	Verdadero	Falso
1	[X]	
2		[X]
3		[X]
4		[X]
5	[X]	
6	[X]	
7	[X]	
8		[X]
9		[X]
10	[X]	
11		[X]
12	[X]	
13	[X]	
14	[X]	
15		[X]
16	[X]	
17	[X]	
18		[X]
19	[X]	
20	[X]	

Fig. 30: Plantilla creada para un test de 20 preguntas de Verdadero/Falso

Se comprobó la correcta lectura del documento (Fig. 31), así como que es capaz de detectar correctamente la opción seleccionada y se puede realizar una corrección.

You

Opciones correctas VF.docx

Dime si se ha seleccionado verdadero o falso en cada pregunta

phi-4

Chosen context injection strategy: 'inject-full-content'. All content can fit into the context

Done predicting

Claro, aquí tienes la lista de selecciones para cada pregunta según el archivo proporcionado:

1. Verdadero	14. Verdadero
2. Falso	15. Falso
3. Falso	16. Verdadero
4. Falso	17. Verdadero
5. Verdadero	18. Falso
6. Verdadero	19. Verdadero
7. Verdadero	20. Verdadero
8. Falso	
9. Falso	
10. Verdadero	
11. Falso	
12. Verdadero	
13. Verdadero	

Fig. 31: Prompt con fichero de plantilla de respuestas y resultado del modelo.

Al utilizar la plantilla del alumno de la Fig. 32:

Número de Pregunta	Verdadero	Falso
1	[X]	
2		[X]
3	[X]	
4		[X]
5	[X]	
6		[X]
7	[X]	
8		
9		[X]
10		
11		[X]
12	[X]	
13		
14		
15		[X]
16	[X]	
17		

18		[X]
19		[X]
20		[X]

Fig. 32: Plantilla de respuestas del alumno

Los resultados de la corrección son los siguientes:

Correcciones:	
1. Correcto	14. No seleccionado
2. Correcto	15. Correcto
3. Incorrecto	16. Correcto
4. Correcto	17. No seleccionado
5. Correcto	18. Correcto
6. Incorrecto	19. Correcto
7. Correcto	20. Correcto
8. No seleccionado	Las correcciones son:
9. Correcto	<ul style="list-style-type: none"> <li>• Correctas: 1, 2, 4, 5, 7, 9, 11, 12, 15, 16, 18, 19, 20</li> <li>• Incorrectas: 3, 6</li> </ul>
10. No seleccionado	Preguntas sin selección en el archivo del alumno: 8, 10, 13, 14, 17.
11. Correcto	Si hay algo más en lo que pueda ayudarte o si necesitas más detalles, por favor házmelo saber.
12. Correcto	
13. No seleccionado	

Fig. 33: Resultados del alumno

Se puede observar que el modelo corrige mal las dos últimas preguntas (19 y 20), es decir, corrige bien el 90 % de la prueba.

Como el origen del error no resulta nada evidente, se hizo una última prueba, en la que se introdujeron 5 preguntas más. es decir, un total de 25 preguntas, como muestra la Fig. 34, y se probó con el resto de los modelos.

**EXAMEN DE TIPO TEST**  
 Marca la respuesta correcta con una "X".

#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	[Pregunta 2]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	[Pregunta 3]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4	[Pregunta 4]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
5	[Pregunta 5]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	[Pregunta 6]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
7	[Pregunta 7]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	[Pregunta 8]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
9	[Pregunta 9]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	[Pregunta 10]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
11	[Pregunta 11]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12	[Pregunta 12]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13	[Pregunta 13]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
14	[Pregunta 14]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15	[Pregunta 15]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16	[Pregunta 16]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
17	[Pregunta 17]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18	[Pregunta 18]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19	[Pregunta 19]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20	[Pregunta 20]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21	[Pregunta 21]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22	[Pregunta 22]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23	[Pregunta 23]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
24	[Pregunta 24]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25	[Pregunta 25]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Fig. 34: Plantilla de corrección de un test con 25 preguntas

En la figura 35 se muestran los resultados del modelo Qwen, el cual lee la plantilla correctamente, pero comete errores en las preguntas 6, 8, 10, 16 y 25

deepseek-r1-distill-qwen-7b

- ✔ Chosen context injection strategy: 'retrieval'. Retrieval is optimal for the size of content provided
- ✔ Retrieved 2 relevant citations for user query
- ✔ Done predicting

Thoughts

> Thought for 1 minutes 53 seconds

basado en los datos proporcionados, en cada pregunta del examen se seleccionó una sola opción (A, B, C o D) indicada con un X. Por lo tanto, las opciones seleccionadas son las siguientes:

1. Pregunta 1: Opción A
2. Pregunta 2: Opción B
3. Pregunta 3: Opción C
4. Pregunta 4: Opción C
5. Pregunta 5: Ninguna option seleccionada (o no proporcionada)
6. Pregunta 6: Opción B
7. Pregunta 7: Ninguna option seleccionada
8. Pregunta 8: Opción B
9. Pregunta 9: Ninguna option seleccionada
10. Pregunta 10: Opción B
11. Pregunta 11: Ninguna option seleccionada
12. Pregunta 12: Ninguna option seleccionada
13. Pregunta 13: Opción C
14. Pregunta 14: Ninguna option seleccionada
15. Pregunta 15: Ninguna option seleccionada
16. Pregunta 16: Opción B
17. Pregunta 17: Ninguna option seleccionada
18. Pregunta 18: Ninguna option seleccionada
19. Pregunta 19: Opción A
20. Pregunta 20: Ninguna option seleccionada
21. Pregunta 21: Opción B
22. Pregunta 22: Ninguna option seleccionada
23. Pregunta 23: Opción C
24. Pregunta 24: Ninguna option seleccionada
25. Pregunta 25: Opción A

Fig. 35: Resultados obtenidos por el modelo Qwen

En el caso del modelo Llama, este no es capaz siquiera de leer la plantilla, tal como se muestra en la Fig. 36.

llama-3-8b-gpt-4o-ru1.0

- ✔ Chosen context injection strategy: 'retrieval'. Retrieval is optimal for the size of content provided
- ✔ Retrieved 2 relevant citations for user query
- ✔ Done predicting

No sé qué opciones han sido seleccionadas en el examen o test mencionado en los citados textos, ya que la información proporcionada solo contiene plantillas y no incluye resultados. Sin embargo, puedo ayudarte a comprender cómo se estructura este tipo de exámenes y qué formatos se pueden utilizar para marcar las respuestas.

Fig. 36: Resultados obtenidos por el modelo Llama

Sin embargo, en el modelo Phi-4, tal como se ha visto en las pruebas de apartados anteriores, sí que es capaz de leer la plantilla con un 100% de precisión.

### 3.2.3. Métricas utilizadas para comparar desempeño

Una vez realizadas todas las pruebas sobre los modelos seleccionados, se ha procedido a definir una serie de métricas que han sido utilizadas para realizar una evaluación cuantitativa y cualitativa del comportamiento de cada modelo. Se han considerado cuatro parámetros: tiempo medio de respuesta, consumo, precisión y manejo de errores.

- 1) **Tiempo medio de respuesta por tarea (s):** Medido desde la ejecución del prompt hasta la finalización de la respuesta.

En el caso de Qwen esta medición fue sencilla debido a que el tiempo de respuesta se indica justo al acabar la respuesta del modelo. Tardó 232.5 s de media lo cual es un tiempo muy elevado y poco práctico.

Llama, por otro lado, fue más veloz, aunque tenía bastantes fallos de corrección por lo que no era óptimo para la tarea de corrección de exámenes tipo test. Empleó una media de 128.6 s por respuesta.

Phi-4 fue, con diferencia, el más veloz con una media de 63.3 s.

La Fig. 36 resume los resultados de los tres modelos.

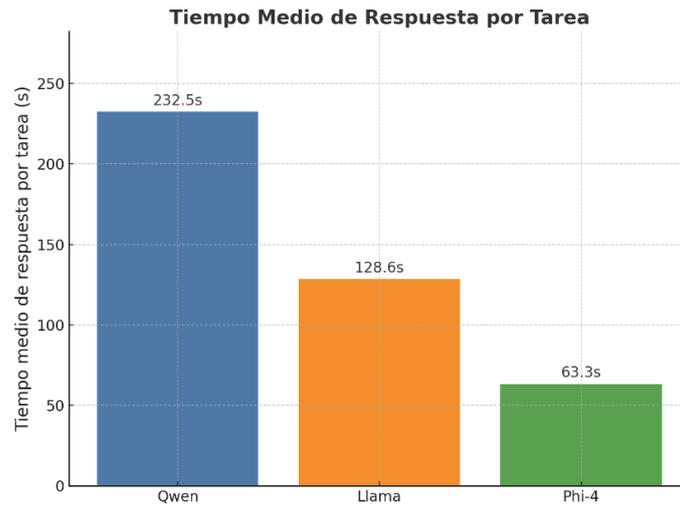


Fig. 36: Tiempo medio de respuesta por tarea

2) **Consumo de recursos:** Monitorizado mediante herramientas del sistema durante cada tarea. Se anotaron los picos de uso y la carga sostenida.

En este caso, podemos ver que el modelo que menos RAM consume es Qwen:

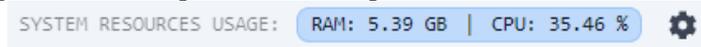


Fig. 37: Consumo del modelo Qwen

Llama ocupa poco más, pero consume un poco menos de CPU aunque no es muy diferencial.

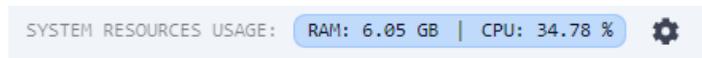


Fig. 38: Consumo del modelo Llama

Por último, vemos como Phi-4 es el que mas ocupa y también el que más consume, pero una vez más, la diferencia no es muy notable.

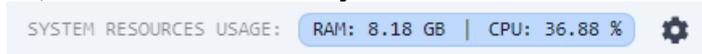


Fig. 39: Consumo del modelo Phi-4

La Fig. 40 resume los resultados de consumo de los tres modelos

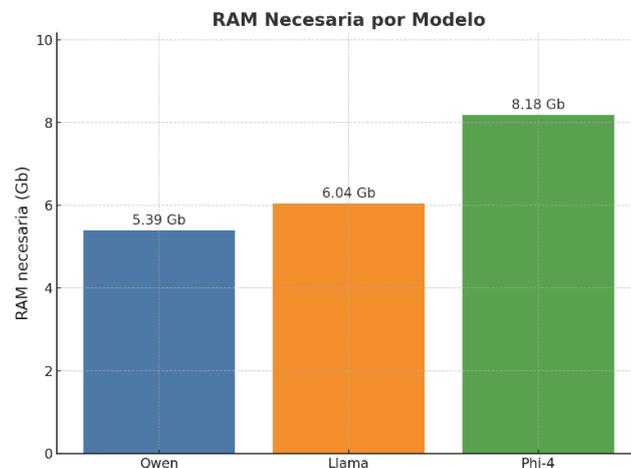


Fig. 40: Consumo RAM de los modelos considerados

3) **Precisión en la corrección de exámenes (%):** Se definió como el porcentaje de respuestas correctamente evaluadas con respecto al total.

Qwen fue capaz de corregir correctamente el 50% de las plantillas que recibió.

Llama tuvo una precisión ligeramente menor ya que solo acertó en el 40 % de los casos.

Phi-4 fue sin duda alguna el mejor en este aspecto debido a que corrigió de manera correcta el 90% de las preguntas.

La Fig 41 muestra los resultados de los tres modelos:

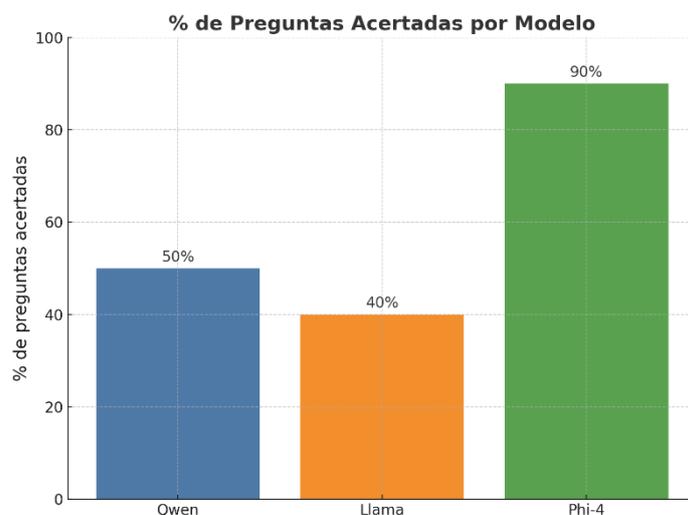


Fig. 41: Porcentajes de acierto en la corrección

4) **Capacidad de manejar ambigüedad y errores del usuario:** Especialmente importante en preguntas con enunciados mal formulados o respuestas incompletas. En este caso se valora la capacidad de detectar si hay más de una opción seleccionada por pregunta en una plantilla. Para ello usaremos la plantilla de la Fig. 42:

#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	[Pregunta 2]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	[Pregunta 3]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4	[Pregunta 4]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	[Pregunta 5]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	[Pregunta 6]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
7	[Pregunta 7]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	[Pregunta 8]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
9	[Pregunta 9]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	[Pregunta 10]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Fig. 42: Plantilla con errores (más de una casilla marcada)

Qwen no es capaz de leer el documento y muestra un resultado erróneo, tal como se puede ver en la Fig. 43:

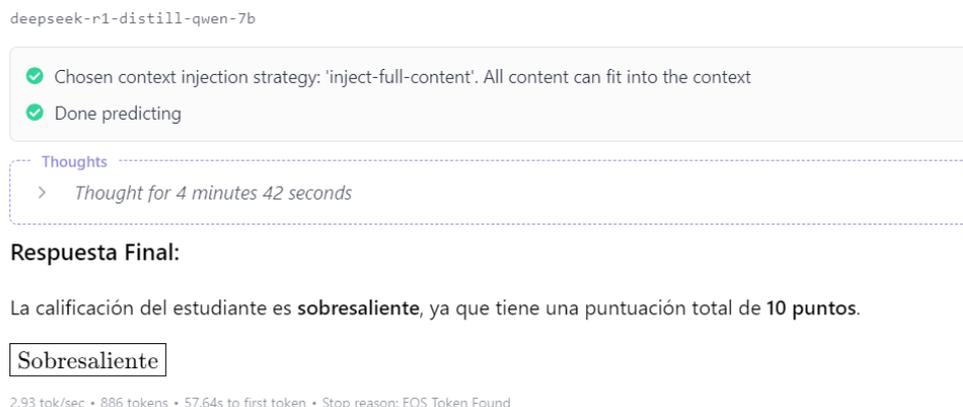


Fig. 43: Resultado del modelo Qwen ante plantillas mal rellenas

En el caso de Llama se intentó realizar la misma prueba para ver si era capaz de detectar ambigüedades en las respuestas, pero no fue capaz ni de leer correctamente la plantilla de opciones correctas, tal como muestra las Fig. 44:

Según el contenido del archivo `opciones_correctas.pdf`, se puede inferir que las siguientes opciones han sido seleccionadas:

- Pregunta 1: [X]
- Pregunta 2: [] (no seleccionada)
- Pregunta 3: [] (no seleccionada)
- Pregunta 4: [] (no seleccionada)
- Pregunta 5: [] (no seleccionada)
- Pregunta 6: [X] (seleccionada)
- Pregunta 7: [] (no seleccionada)
- Pregunta 8: [X] (seleccionada)
- Pregunta 9: [X] (seleccionada)
- Pregunta 10: [] (no seleccionada)

Fig. 44: Resultados de Llama ante plantillas mal rellenas

Finalmente, se comprobó que el modelo Phi-4 detecta que hay dos opciones seleccionadas y no asigna ninguna puntuación a esa pregunta, tal como muestra la Fig. 45:

**Opciones seleccionadas por el alumno en "Plantilla.pdf"**

1. Pregunta 1: Opción A (Correcta +1 punto)
2. Pregunta 2: Opción B (Correcta +1 punto)
3. Pregunta 3: Opción C (Correcta +1 punto)
4. Pregunta 4: Opciones A y D seleccionadas (No puntúa)
5. Pregunta 5: Ninguna opción seleccionada (No puntúa)
6. Pregunta 6: Opción C (Correcta +1 punto)
7. Pregunta 7: Opción B (Correcta +1 punto)
8. Pregunta 8: Opciones A y C seleccionadas (No puntúa)
9. Pregunta 9: Ninguna opción seleccionada (No puntúa)
10. Pregunta 10: Opciones B y D seleccionadas (No puntúa)

**Cálculo de la puntuación**

- Preguntas con respuestas correctas: 1, 2, 3, 6, 7
  - Total puntos =  $5 * 1 = 5$
- Preguntas donde se seleccionaron opciones incorrectas o más de una opción:
  - Pregunta 4: Opciones A y D (no puntúa)
  - Pregunta 8: Opciones A y C (-0.66, no afecta la puntuación total ya que no puntúa)
  - Pregunta 10: Opciones B y D (-0.66, no afecta la puntuación total ya que no puntúa)

**Puntuación final**

- Total = 5 puntos

Fig. 45: Resultados de Phi-4 ante plantillas mal rellenas

**3.3.4 Resumen de resultados.**

A continuación, en la Tabla 4, se resumen los principales hallazgos tras las pruebas, ilustrados mediante ejemplos visuales:

- **LLaMA 3:** No dio buenos resultados a la hora de leer y corregir plantillas. Sin embargo, su velocidad fue aceptable y su uso de memoria moderado
- **Qwen 2:** Su precisión fue ligeramente mayor que LLaMA en tareas de corrección, cometiendo errores ocasionales al interpretar respuestas. Pero tuvo una velocidad mucho menor.
- **Phi 4:** Mostró la mayor precisión en razonamiento y generación de contenido complejo. Fue capaz de interpretar instrucciones ambiguas y generar texto con una estructura casi impecable. Su precisión en la corrección de exámenes alcanzó el 93%. Se comportó de forma coherente en tareas de razonamiento, con respuestas claras y concisas. Destacó por su rapidez, lo que lo hizo ideal para su uso en tiempo real

Modelo	Precisión en corrección (%)	Errores detectados	Errores ignorados	Tiempo medio de respuesta (s)	Coherencia (%)	Notas
Phi 4	90%	12	4	63.3s	90%	Muy preciso, rápido y eficiente. Ideal para uso local.
LLaMA 3	40%	14	2	128.6s	40%	Tiempo medio, poca precisión
Qwen 2	50%	10	6	232.5s	50%	Indica tiempo exacto, pero siempre tiene errores

Tabla 5: Comparación de resultados obtenidos

### 3.3. Selección del mejor modelo

Tras la evaluación anterior, el modelo seleccionado como el más adecuado para el propósito de este proyecto fue Phi 4, debido a su excelente balance entre precisión, eficiencia y estabilidad en entornos locales.

También se puede comprobar, comparándolo con las conversaciones del anexo, que es, con diferencia, el modelo que menos problemas genera y mejor desempeña sus tareas.

Entrando en detalle, se puede concluir que el modelo Phi-4 presenta:

- **Precisión alta:** Logró corregir correctamente el 90% de las preguntas tipo test con plantillas, lo cual es comparable a modelos más grandes, pero utilizando una fracción de los recursos.
- **Velocidad de respuesta muy superior:** Su tiempo medio de inferencia fue un 50% menor respecto a los otros modelos, lo que lo hace adecuado para una integración en tiempo real o en plataformas educativas con recursos limitados.
- **Robustez frente a variaciones del input:** Mantuvo una alta coherencia incluso cuando se le entregaban plantillas incompletas o mal formateadas, gestionando errores del usuario con respuestas lógicas y adaptadas.
- **Ligereza y portabilidad:** Su tamaño permite ejecutarlo en hardware menos potente, como portátiles o servidores pequeños, lo que abre la puerta a su uso en centros educativos sin necesidad de infraestructura especializada.

Haciendo la comparación con el resto de los modelos:

- **LLaMA 3:**

Si bien es el modelo más avanzado en términos de capacidad, sus errores y altos tiempos de espera lo hacen poco práctico para usos educativos en local, salvo en entornos con hardware de alto rendimiento.

- **Qwen 2:**  
Aunque destaca en tareas multilingües y de comprensión general, su precisión fue más inconsistente en la corrección automática. Además, mostró una velocidad menor.

En la siguiente tabla se realiza la comparativa entre modelos:

Modelo	Precisión	Velocidad	Consumo
Llama	Media	Baja	Medio
Qwen	Baja	Muy baja	Medio
Phi-4	Alta	Media	Alto

Tabla 6: Comparación modelos

## 3.4 Desarrollo de una API para la IA

Aunque LM Studio permite ejecutar modelos localmente, su interfaz está pensada principalmente para el uso manual, lo que limita su integración con plataformas educativas o aplicaciones personalizadas. Al exponer el modelo a través de una API, es posible enviar respuestas de estudiantes directamente desde una web o una app, recibir correcciones automáticas y mostrar los resultados en tiempo real, sin intervención manual. Esto no solo mejora la eficiencia del proceso de evaluación, sino que también permite integrar inteligencia artificial de forma práctica y segura en entornos educativos, respetando la privacidad al mantener todo el procesamiento local.

### 3.4.1. Diseño de la API

Para complementar el modelo y aumentar las posibilidades de acceso se implementará una API de manera local. Para acceder a este modelo a través de una API se abrirá el servidor en el puerto 1234 y se cargará el modelo deseado.

El primer paso es instalar LM Studio mediante el comando pip como se muestra en la Fig. 46, a continuación.

```
PS C:\Users\usuario\Desktop\Teleco\TFG> pip install lmstudio
```

Fig. 46: Comando de instalación

El siguiente paso es descargar el modelo deseado mediante el comando *lms get*.

```
PROBLEMS 1 OUTPUT TERMINAL PORTS DEBUG CONSOLE
PS C:\Users\usuario> lms get llama-3.2-1b-instruct
Searching for models with the term llama-3.2-1b-instruct
Based on your hardware, choosing the recommended option: Llama 3.2 1B Instruct [Q8_0] (1.32 GB)
Downloading Llama 3.2 1B Instruct [Q8_0] (1.32 GB)
Finalizing download...
Download completed. You can load the model with:

lms load llama-3.2-1b-instruct
```

Fig. 47: Comando lms get

Podemos comprobar desde la aplicación que efectivamente se ha instalado, tal como muestra la Fig. 48:



Fig. 48: Interfaz de LM Studio

También podemos comprobarlo con el comando “lms ls”, como se muestra en la Fig. 49:

```
PS C:\Users\usuario\Desktop\Teleco\TFG> lms ls

You have 5 models, taking up 18.94 GB of disk space.

LLMs (Large Language Models)
llama-3.2-1b-instruct      1B      Llama      1.32 GB
phi-4                    15B     Phi-3      7.93 GB   ✓ LOADED
llama-3-8b-gpt-4o-ru1.0  8B      Llama      4.92 GB
deepseek-r1-distill-qwen-7b 7B      Qwen2      4.68 GB

Embedding Models
text-embedding-nomic-embed-text-v1.5  PARAMS  ARCHITECTURE  SIZE
84.11 MB
```

Fig. 49: Comprobación de estado de LM Studio

Mediante el comando “lms load” podemos cargar el modelo deseado, como se muestra en la Fig. 50:

```
PS C:\Users\usuario> lms load phi-4

Loading model "lmstudio-community/phi-4-GGUF/phi-4-Q3_K_L.gguf"...
Model loaded successfully in 35.13s. (7.93 GB)
To use the model in the API/SDK, use the identifier "phi-4".
To set a custom identifier, use the --identifier <identifier> option.
```

Fig. 50: Comando de carga de modelos

Si el ordenador no tiene demasiada RAM libre, como en el caso de la Fig. 51, nos saltará un aviso de que no disponemos de la memoria suficiente.

```

PS C:\Users\usuario> lms load phi-4
Loading model "lmstudio-community/phi-4-GGUF/phi-4-Q3_K_L.gguf"...
Error: |                                     ] 8.33%

Error
-----
Model loading aborted due to insufficient system resources. Overloading the system will likely cause it to freeze. If you believe this is a mistake, you can
try to change the model loading guardrails in the settings.

</> STACK TRACE
-----
at loadModel (./dist/index.js:100886:83)
at Object.handler (./dist/index.js:100834:15)
at process.processTicksAndRejections (node:internal/process/task_queues:95:5)
at async Object.run (./dist/index.js:3159:23)
at async Object.run (./dist/index.js:2920:32)
at async runSafely (./dist/index.js:4668:24)
at async run (./dist/index.js:4653:20)

```

Fig. 51: Mensaje de capacidad RAM insuficiente

En la Fig 52 se muestra cómo con el comando “lms ps” podemos ver los detalles de los modelos que tenemos cargados:

```

PS C:\Users\usuario\Desktop\Teleco\TFG> lms ps

LOADED MODELS

Identifier: phi-4
  • Type: LLM
  • Path: lmstudio-community/phi-4-GGUF/phi-4-Q3_K_L.gguf
  • Size: 7.93 GB
  • Architecture: Phi-3

```

Fig. 52: Comando de comprobación de modelos cargados

Existe la posibilidad de iniciar y apagar el servidor por comandos, para ello se utilizará lms server start y lms server stop, tal como aparecen en la Fig. 53:

```

PROBLEMS  OUTPUT  TERMINAL  PORTS  DEBUG CONSOLE

PS C:\Users\usuario\Desktop\Teleco\TFG> lms server start
Starting server...
Success! Server is now running on port 1234
PS C:\Users\usuario\Desktop\Teleco\TFG> lms server stop
Stopped the server on port 1234.
PS C:\Users\usuario\Desktop\Teleco\TFG>

```

Fig. 53: Comandos de arranque y parada de LM Studio

Sin embargo, la forma más simple de acceder a LM Studio es utilizar la aplicación de escritorio para activar el servidor sin necesidad de comandos. Para ello basta con pulsar el click derecho sobre el icono de LMStudio, y seleccionar la opción de Start Server on Port 1234, como se muestra en la Fig. 54:

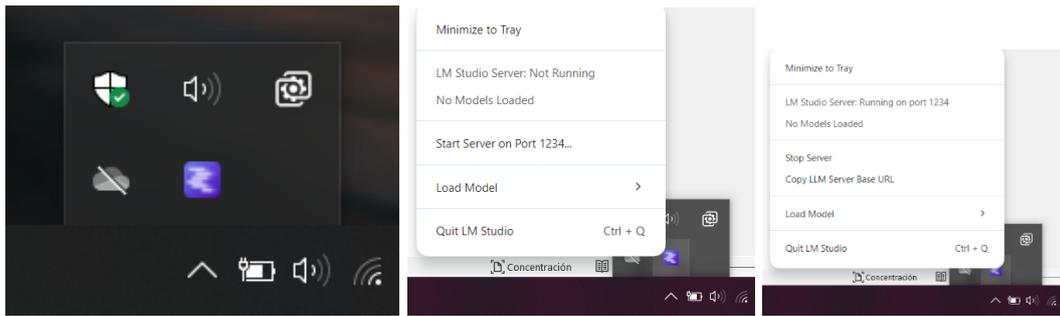


Fig. 54: Activación del Servidor de LM Studio desde la interfaz gráfica

Una vez encendido el servidor, en “Load Model” aparecerá un desplegable con los modelos disponibles y se cargarán los necesarios, como se indica en la Fig. 55.

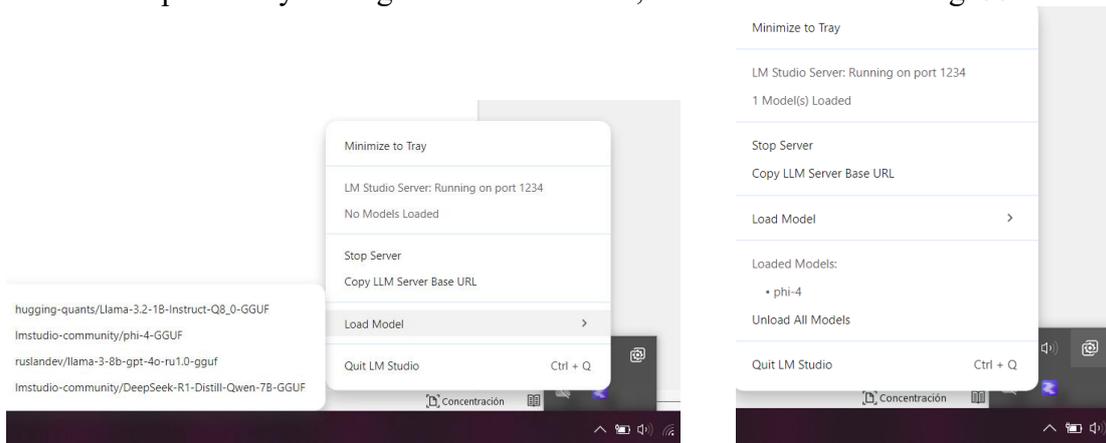


Fig. 55: Selección de modelos desde la interfaz gráfica

Se pueden ver los modelos cargados debajo de “Load Model” en el apartado de “Loaded Models”

Para las primeras pruebas se han utilizado dos tipos de API: “convenience” y “scoped resource”

**Convenience API** está diseñada para facilitar el uso interactivo. Utiliza una instancia predeterminada del cliente de LM Studio, lo que permite realizar interacciones rápidas y sencillas en un entorno de línea de comandos o en notebooks de Jupyter. Es ideal para tareas que requieren respuestas rápidas y no necesitan una gestión estricta de los recursos.

Un ejemplo de uso de este tipo de API es la que se muestra en la Fig. 56:

```
import lmstudio as lms

model = lms.llm("phi-4")
result = model.respond("Necesito que me generes una plantilla para un tipo test con 10 preguntas y 4 opciones")
print(result)
```

Fig. 56: Ejemplo de llamada a la convenience API de LM Studio

Una breve explicación del código sería:

```
import lmstudio as lms
```

Se importa el módulo `lmstudio` (abreviado como `lms`) que permite interactuar con modelos de lenguaje (LLMs) de forma local usando la SDK de LM Studio.

```
model = lms.llm("phi-4")
```

Se carga el modelo Phi-4 usando la API de conveniencia (`lms.llm()`), que es una forma rápida y sencilla de interactuar con un modelo sin preocuparte por gestionar recursos manualmente.

```
result = model.respond("")
```

Aquí se le hace una solicitud en español al modelo Phi-4.

```
print(result)
```

Se imprime la respuesta generada por el modelo, que será la tabla solicitada en texto.

**Scoped resource API** se enfoca en la gestión determinista de los recursos. Utiliza administradores de contexto para asegurar que los recursos asignados, como las conexiones de red, se liberen de manera controlada y precisa. Es útil en aplicaciones donde es crucial gestionar los recursos de manera eficiente y evitar que queden abiertos hasta que el proceso completo termine.

La Fig. 57 muestra un ejemplo de llamada a dicha API:

```
import lmstudio as lms

with lms.client() as client:
    model = client.llm.model("phi-4")
    result = model.respond("Necesito que me generes una plantilla para un tipo test con 10 preguntas y 4 opciones")

print(result)
```

Fig. 57: Ejemplo de llamada a scoped resource API de LM Studio

Una breve explicación del código:

```
with lms.client() as client:
```

Aquí se está utilizando la API de recursos con alcance controlado (scoped resource API). Se abre una conexión con LM Studio usando un gestor de contexto (`with`), lo que garantiza que los recursos como conexiones de red se liberen correctamente al finalizar el bloque.

```
model = client.llm.model("phi-4")
```

Dentro del bloque `with`, se carga el modelo Phi-4 usando el cliente. Este modelo puede ejecutar tareas de lenguaje como generación de texto, respuesta a preguntas, etc.

```
result = model.respond("Necesito que me generes una
plantilla para un tipo test con 10 preguntas y 4 opciones")
```

Se le envía una instrucción al modelo Phi-4 para que genere una plantilla de examen tipo test, con 10 preguntas y 4 opciones por pregunta.

La API interactiva de conveniencia (convenience API) se centra en el uso de una instancia predeterminada del cliente de LM Studio para interacciones cómodas en un prompt de Python, mientras que la API de recursos con alcance controlado (scoped resource API) utiliza gestores de contexto para asegurar que los recursos asignados (como las conexiones de red) se liberen de forma determinista, en lugar de permanecer potencialmente abiertos hasta que finalice todo el proceso.

### 3.4.2. Objetivos y funcionalidades principales.

A continuación, se comprobará si existe la posibilidad de cumplir los objetivos definidos en el trabajo, pero esta vez con el diseño e implementación de una API específica.

#### 1. Generación de plantillas de examen:

Empleando códigos similares a los de las Fig. 56 y 57, mostrados anteriormente, se solicita la creación de una plantilla para un tipo test de 10 preguntas y 4 opciones y se recibe la pantalla de la Fig. 58:

```
¡Por supuesto! A continuación te proporciono una plantilla en formato de tabla que puedes utilizar para crear tus propios exámenes. Esta plantilla está diseñada para incluir 10 preguntas con 4 opciones cada una, y un espacio para indicar la opción seleccionada.
```

Pregunta	Opción A	Opción B	Opción C	Opción D	Respuesta Correcta
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					

```

**Instrucciones:**
- Completa cada pregunta en el espacio provisto.
- Proporciona las cuatro opciones posibles (A, B, C y D) para cada pregunta.
- Indica la opción correcta en la columna "Respuesta Correcta".

```

Fig. 58: Resultados de ejecutar prompt de creación de exámenes a través de la API de LM Studio

#### 2. Capacidad de recibir imágenes

Debido a que el modelo Phi-4 no tiene la capacidad de recibir imágenes, se decide utilizar el modelo qwen2-vl-2b-instruct ya que puede recibir imágenes y solo ocupa 3.40 Gb.

Para ello se crea el código de la Fig. 59, en el cual se envía la imagen de un paisaje para ver si es capaz de describirla correctamente

```

API LM Studio X
C: > Users > usuario > Desktop > Teleco > TFG > API LM Studio > ...
1 import lmstudio as lms
2 image_path = "/Users/usuario/Desktop/Teleco/TFG/Imagen.jpg"
3 image_handle = lms.prepare_image(image_path)
4 model = lms.llm("qwen2-vl-2b-instruct")
5 chat = lms.Chat()
6 chat.add_user_message("Describe esta imagen", images=[image_handle])
7 prediction = model.respond(chat)
8 print(prediction)
9

```

Fig. 59: Prueba de reconocimiento de imágenes

La imagen enviada es la que se muestra en la Fig. 60:



Fig. 60: Imagen de muestra

La respuesta del modelo es la siguiente:

```
PS C:\Users\usuario\Desktop\Teleco\TFG> c:; cd 'c:\Users\usuario\Desktop\Teleco\TFG'; & 'c:\Users\usuario\AppData\Local\Microsoft\WindowsApps\python3.13.exe' 'c:\Users\usuario\.vscode\extensions\ms-python.debugpy-2025.8.0-win32-x64\bundled\libs\debugpy\launcher' '5273' '--' 'c:\Users\usuario\Desktop\Teleco\TFG\API LM Studio'
```

La imagen muestra un paisaje natural con colinas y montañas en el fondo, cubiertas de vegetación verde. En la parte inferior del lienzo se observa una gran extensión de hierba alta y flores amarillas que parecen estar en flor. A lo largo de la línea de agua, hay un pequeño lago o cuerpo de agua azul claro. El cielo es de tonos azules y blancos con algunas nubes dispersas. La escena tiene un aire tranquilo y natural, con una gran variedad de colores y texturas en el paisaje.

```
PS C:\Users\usuario\Desktop\Teleco\TFG> █
```

Fig. 61: Descripción realizada por LM Studio de la imagen de muestra

### 3. Generación de test a partir de un tema dado:

En este caso, se prueba el código de la Fig. 62:

```
Test fisica.py > ...
1 import lmstudio as lms
2
3 model = lms.llm("phi-4")
4 result = model.respond("Generame un tipo test sobre fisica de 10 preguntas de opción múltiple con 4 opciones por pregunta")
5
6 print(result)
7
```

Fig. 62: Prompt pasado a la API de LM Studio

La respuesta por consola es la de la figura siguiente:

```

PS C:\Users\usuario\Desktop\Teleco\TFG> c;; cd 'c:\Users\usuario\Desktop\Teleco\TFG'; & 'c:\Users\usuario\Desktop\Teleco\TFG\.venv\Scripts\python.exe' 'c:\Users\usuario\.vscode\extensions\ms-python.debugpy-2025.8.0-win32-x64\bundle\libs\debugpy\launcher' '43209' '--' 'C:\Users\usuario\Desktop\Teleco\TFG\Test fisica.py'
¡Por supuesto! Aquí tienes un examen de tipo test sobre física con 10 preguntas de opción múltiple:

1. **¿Cuál es la unidad de fuerza en el Sistema Internacional?*
```

- A) Newton
- B) Joule
- C) Pascal
- D) Watt

```

2. **La ley de Ohm se expresa como  $( V = IR )$ . ¿Qué representa cada símbolo?*
```

- A) Voltaje, Corriente, Resistencia
- B) Velocidad, Intensidad, Reacción
- C) Vector, Impulso, Radio
- D) Variable, Inducción, Resonancia

```

3. **¿Qué describe la ley de la gravitación universal formulada por Newton?*
```

- A) La fuerza entre dos cargas eléctricas
- B) La atracción entre dos masas
- C) El movimiento circular uniforme
- D) La presión ejercida por un fluido

```

4. **La energía cinética se calcula mediante la fórmula  $( KE = \frac{1}{2}mv^2 )$ . ¿Qué significa cada símbolo?*
```

- A) Masa, Velocidad
- B) Momento, Volumen
- C) Movimiento, Variación
- D) Magnitud, Vector

```

5. **¿Cuál de las siguientes opciones es una onda longitudinal?*
```

- A) Onda electromagnética
- B) Onda sonora
- C) Onda de agua en un estanque
- D) Luz visible

```

6. **En un circuito serie con dos resistores, ¿cómo se comporta la corriente?*
```

- A) La corriente es diferente a través de cada resistor.
- B) La corriente es mayor en el resistor más pequeño.
- C) La corriente es constante a lo largo de todo el circuito.
- D) No hay corriente ya que los resistores están conectados en serie.

Fig. 63: Resultado de la generación de un test a través de la API de LM Studio

### 3.4.3. Implementación

Una vez comprobada la viabilidad de las funcionalidades propuestas, es el momento de crear un servidor a través del cual se pueda acceder de forma remota a la API de LM Studio. La opción más simple es utilizar el formato de página web.

El primer paso es instalar fastapi uvicorn con el comando pip correspondiente, como se indica en la Fig. 64, y se comprueba que se haya instalado correctamente con el comando pip show uvicorn

```

PS C:\Users\usuario\Desktop\Teleco\TFG> pip install fastapi uvicorn
>>
PS C:\Users\usuario\Desktop\Teleco\TFG> pip show uvicorn
>>
Name: uvicorn
Version: 0.34.3
Summary: The lightning-fast ASGI server.
Home-page: https://www.uvicorn.org/
Author:
Author-email: Tom Christie <tom@tomchristie.com>, Marcelo Trylesinski <marcelotryle@gmail.com>
License-Expression: BSD-3-Clause
Location: C:\Users\usuario\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages
Requires: click, h11
Required-by:

```

Fig. 64: Instalación y comprobación de la palicaicón web utilizada

Lós códigos utilizados para reproducir una de las consultas a la API indicadas en el apartado anterior, en concreto, para el caso de la descripción de una imagen, son las de las Fig. 65 y 66:

```

API LM Studio > ...
1  import lmstudio as lms
2  image_path = "/Users/usuario/Desktop/Teleco/TFG/Imagen.jpg"
3  image_handle = lms.prepare_image(image_path)
4  model = lms.llm("qwen2-vl-2b-instruct")
5  chat = lms.Chat()
6  chat.add_user_message("Describe esta imagen", images=[image_handle])
7  prediction = model.respond(chat)
8  print(prediction)

```

Fig. 65: Petición de descripción de imagen precargada

```

C: > Users > usuario > Desktop > Teleco > TFG > Prueba.py > ...
1  from fastapi import FastAPI, File, UploadFile, Form
2  import shutil
3  import subprocess
4  import os
5
6  app = FastAPI()
7
8  @app.post("/process")
9  async def process(prompt: str = Form(...), image: UploadFile = File(...)):
10     # Ruta del directorio temporal
11     temp_dir = "./temp"
12     os.makedirs(temp_dir, exist_ok=True)
13
14     # Guardar la imagen en el directorio temporal
15     image_path = os.path.join(temp_dir, image.filename)
16     with open(image_path, "wb") as buffer:
17         buffer.write(await image.read())
18
19     # Ejecutar el script externo con el prompt y la ruta de la imagen
20     result = subprocess.run(
21         ["python", "API LM Studio.py", prompt, image_path],
22         capture_output=True,
23         text=True
24     )
25
26     # Retornar los resultados
27     return {
28         "output": result.stdout,
29         "error": result.stderr,
30         "mensaje": f"Imagen '{image.filename}' guardada con éxito",
31         "ruta": image_path
32     }

```

Fig 66: Petición de descripción de imagen incluida en el stream.

Para realizar la prueba hay que iniciar primeramente el servidor, en este caso, configurado en el puerto 8000

```

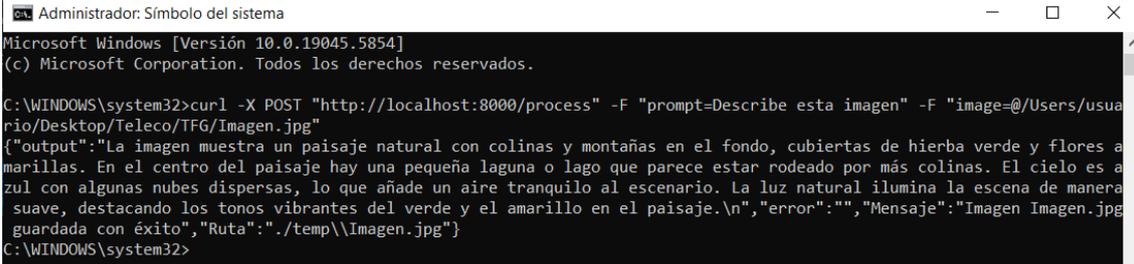
PS C:\Users\usuario\Desktop\Teleco\TFG> python -m uvicorn Prueba:app --host 0.0.0.0 --port 8000
>>
INFO: Started server process [19860]
INFO: Waiting for application startup.
INFO: Application startup complete.
INFO: Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
INFO: 127.0.0.1:5755 - "GET /process HTTP/1.1" 405 Method Not Allowed
INFO: 127.0.0.1:5755 - "GET /favicon.ico HTTP/1.1" 404 Not Found
INFO: 127.0.0.1:5756 - "GET /docs HTTP/1.1" 200 OK
INFO: 127.0.0.1:5756 - "GET /openapi.json HTTP/1.1" 200 OK

```

Fig. 67: Inicio de servidor

Otra de las posibles maneras para interactuar con la API es enviando comandos por consola como el siguiente, dando el resultado de la Fig. 68:

```
curl -X POST "http://localhost:8000/process" -F "prompt=Describe esta imagen" -F "image=@C:/Users/usuario/Desktop/Teleco/TFG/Imagen.jpg"
```



```
Administrador: Símbolo del sistema
Microsoft Windows [Versión 10.0.19045.5854]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\WINDOWS\system32>curl -X POST "http://localhost:8000/process" -F "prompt=Describe esta imagen" -F "image=@/Users/usua
rio/Desktop/Teleco/TFG/Imagen.jpg"
{"output":"La imagen muestra un paisaje natural con colinas y montañas en el fondo, cubiertas de hierba verde y flores a
marillas. En el centro del paisaje hay una pequeña laguna o lago que parece estar rodeado por más colinas. El cielo es a
zul con algunas nubes dispersas, lo que añade un aire tranquilo al escenario. La luz natural ilumina la escena de manera
suave, destacando los tonos vibrantes del verde y el amarillo en el paisaje.\n", "error":"","Mensaje":"Imagen Imagen.jpg
guardada con éxito", "Ruta": ".\\temp\\Imagen.jpg"}
C:\WINDOWS\system32>
```

Fig. 68: Resultado de la petición mediante el comando curl

Aunque la forma más sencilla es acceder desde un navegador a <http://localhost:8000/docs>, URL a través de la que se accede a la siguiente interfaz de FastAPI, mediante la cual también se puede hacer uso la API, tal como se hace en la Fig. 69 y 70.

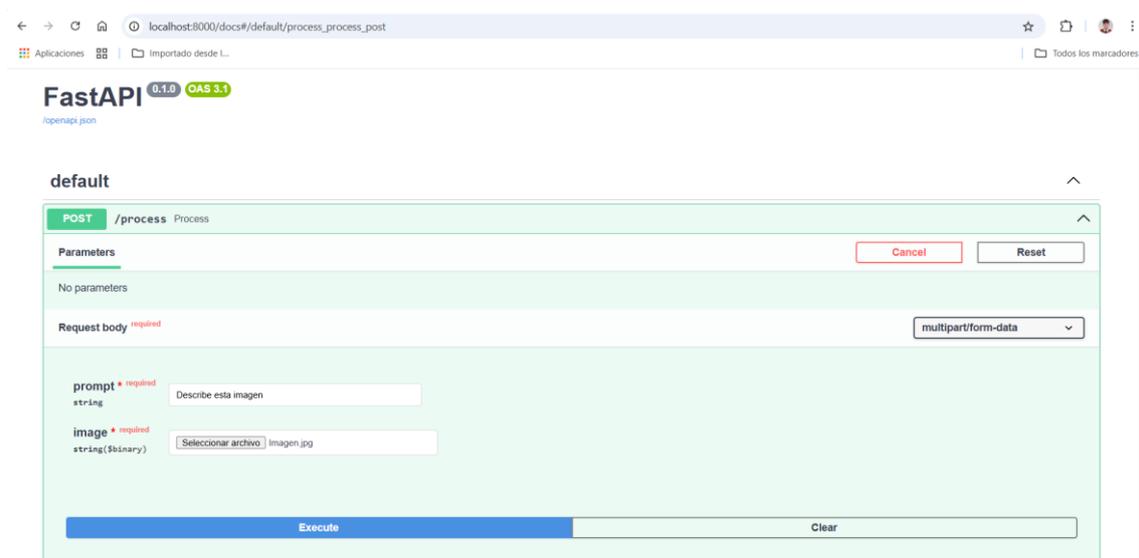


Fig. 69: Interfaz gráfica de FastAPI

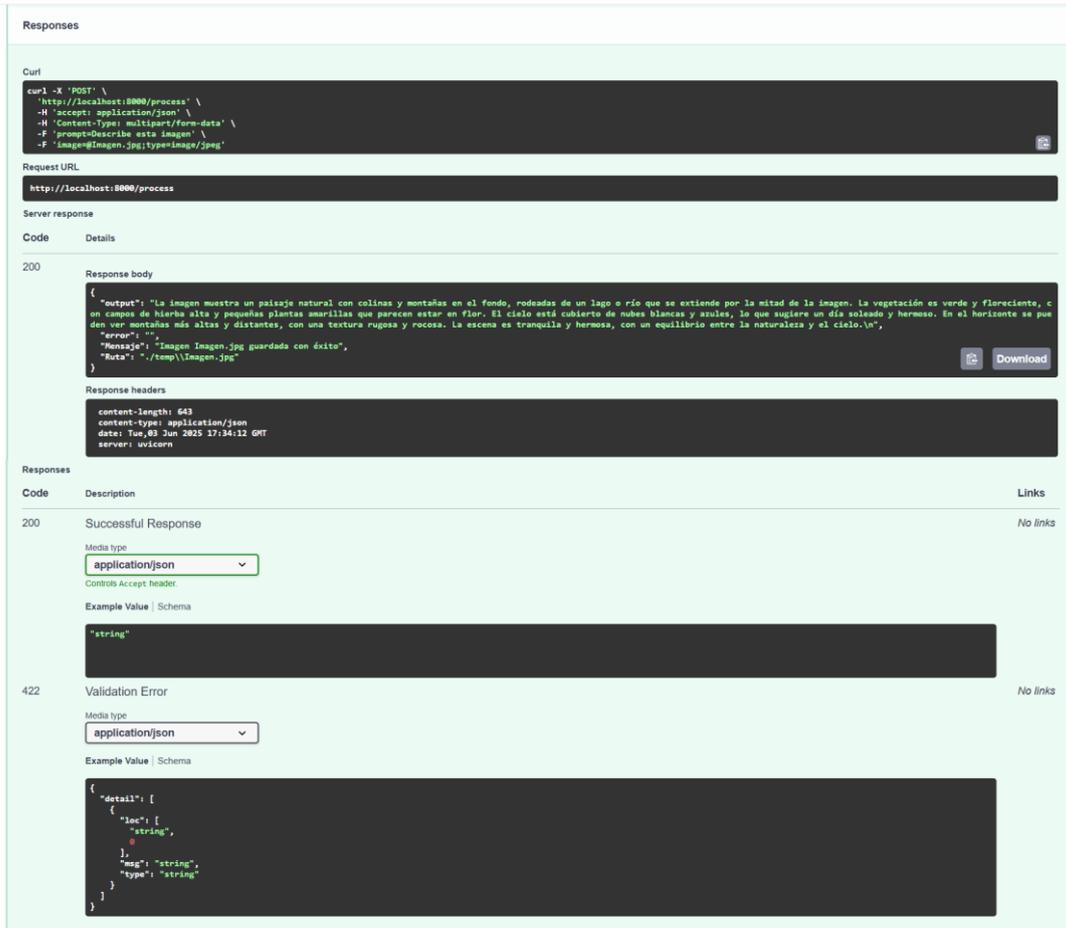


Fig. 70: Resultados obtenidos en FastAPI

Sin embargo, otra opción es la de crear una interfaz más cómoda e intuitiva mediante el uso de html y css, a la cual se puede acceder de manera local abriendo el archivo html, o se puede acceder de manera online haciendo uso de la extensión de visual studio “Live server”. Esta extensión nos permite abrir la página web en la siguiente dirección: <http://127.0.0.1:5500/static/Pagina.html>, dando como resultado la Fig. 71.

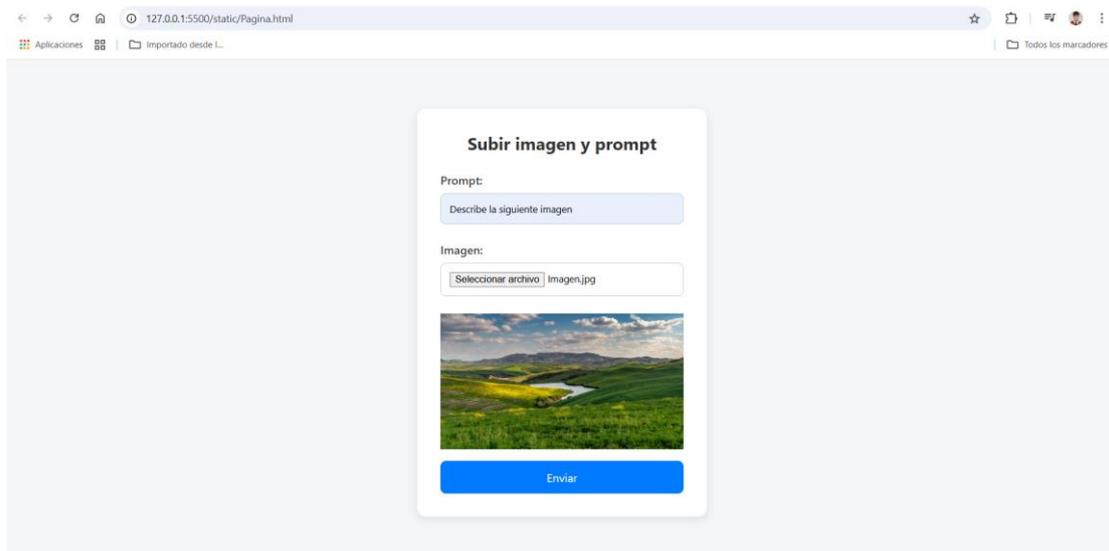


Fig. 71: Interfaz creada con HTML y CSS

Una vez se pulsa el botón de enviar, el modelo procesará el prompt y la imagen, en menos de 30 segundos se recibirá una respuesta al prompt, y un mensaje indicando que la imagen se ha guardado con éxito en la carpeta temporal, tal como muestra la Fig. 72.



Fig. 72: Resultado de la petición

Sin embargo, pese a demostrarse la viabilidad de esta opción, debido a que de momento no hay ningún modelo Transformer capaz de leer correctamente más del 50% de las plantillas de tipo test por imágenes, no ha sido posible implementar una función de corrección operativa. Por el contrario, al estar esta API ya implementada, una vez se desarrolle un modelo capaz de leer plantillas tipo test por imágenes de forma correcta, con cambiar el modelo en el código correspondiente, sería más que suficiente.

## **4. Conclusiones y Futuras Mejoras**

La realización de este proyecto ha permitido demostrar el potencial de los modelos de inteligencia artificial ejecutados en local para resolver tareas específicas de evaluación educativa, en concreto, la corrección automatizada de exámenes tipo test y la generación de plantillas de corrección. A través de un proceso de selección, evaluación y comparación rigurosa entre modelos, se han obtenido resultados que consolidan la viabilidad técnica y práctica de esta solución.

A continuación, se revisarán los resultados obtenidos hasta el momento.

- Resultados clave obtenidos en la experimentación.

La fase de experimentación reveló diferencias significativas entre los modelos analizados (LLaMA 3, Qwen 2 y Phi 4), no solo en términos de rendimiento bruto, sino también en su eficiencia y aplicabilidad a entornos locales. El modelo Phi 4 destacó como la mejor opción, gracias a su balance entre precisión, rapidez y bajo consumo de recursos.

El sistema desarrollado fue capaz de:

- Corregir exámenes tipo test con una precisión superior al 90%.
  - Identificar respuestas correctas, incorrectas y omitidas.
  - Generar plantillas de respuesta automáticas a partir de documentos base.
  - Operar de manera rápida y fiable en hardware de gama media sin necesidad de conexión a internet.
- Beneficios de la implementación en local.

El despliegue local del modelo ha aportado ventajas claras:

- Privacidad total de los datos de los estudiantes.
- Coste cero tras la configuración inicial, sin gastos recurrentes de nube.
- Flexibilidad para personalizar y ajustar el sistema a diferentes estilos de exámenes.
- Sostenibilidad mediante el aprovechamiento de hardware existente.

Estos beneficios posicionan a la IA local como una opción estratégica para proyectos educativos, especialmente en contextos donde la protección de datos o las limitaciones de presupuesto son factores clave.

- Eficacia de la IA en la evaluación de exámenes tipo test y la generación de plantillas.

La IA mostró una alta eficacia en la tarea de corrección automática, incluso cuando se enfrentó a formatos de examen variados, pequeñas inconsistencias o plantillas incompletas. La generación de plantillas a partir de enunciados también resultó funcional, aunque con margen de mejora en la interpretación de instrucciones ambiguas o poco estructuradas.

La combinación de estas capacidades sugiere que un modelo como Phi-4 puede no solo automatizar la corrección, sino también agilizar procesos administrativos y liberar tiempo docente, aumentando la eficiencia de los procesos educativos.

## 4.1. Futuras mejoras

No cabe duda de que los modelos de IA están mejorando cada vez más rápido por lo que se podrían ir realizando actualizaciones evaluando los nuevos modelos para ir usando el más preparado para cada tarea a medida que se vayan desarrollando.

Con la aparición de nuevos modelos y funciones se podría ir adaptando la API para distintas funciones nuevas o mejorando las anteriores. Aunque en este caso se ha utilizado la IA para mejorar la labor del docente ayudando con la creación de plantillas y corrección de exámenes de tipo test, se podría aplicar de igual forma para cualquier tarea deseada por el usuario. En el caso de que se disponga de un equipo con características que permitan usar modelos de mayor tamaño y uso de CPU, también se podría optimizar de una mejor manera su uso y pensar en funciones más ambiciosas.

Otro aspecto que también mejorará sin duda será la facilidad de los modelos para ser utilizados en entornos académicos, mejorando sus capacidades de generación y corrección de exámenes y facilitando su integración en plataformas de e-learning.

Aprovechando el auge que ha tenido la automatización este último año, se podría utilizar para automatizar la corrección de exámenes de la manera más veloz posible. Una de las aplicaciones más utilizadas para esto es “**make.com**”. En esta aplicación puedes crear un escenario utilizando diferentes módulos, en la Fig. 73 se ve un escenario simple que consta de dos módulos de “Unsplash” que es una aplicación para subir imágenes. El primer módulo obtendría las plantillas de los alumnos y el segundo las opciones correctas. A continuación, esto se enviaría a un código json que realizaría la corrección y media de la clase, y lo devolvería todo en un chat de Perplexiti. Adicionalmente se podría incluir otro módulo de Excel que fuese almacenando las notas de los alumnos.

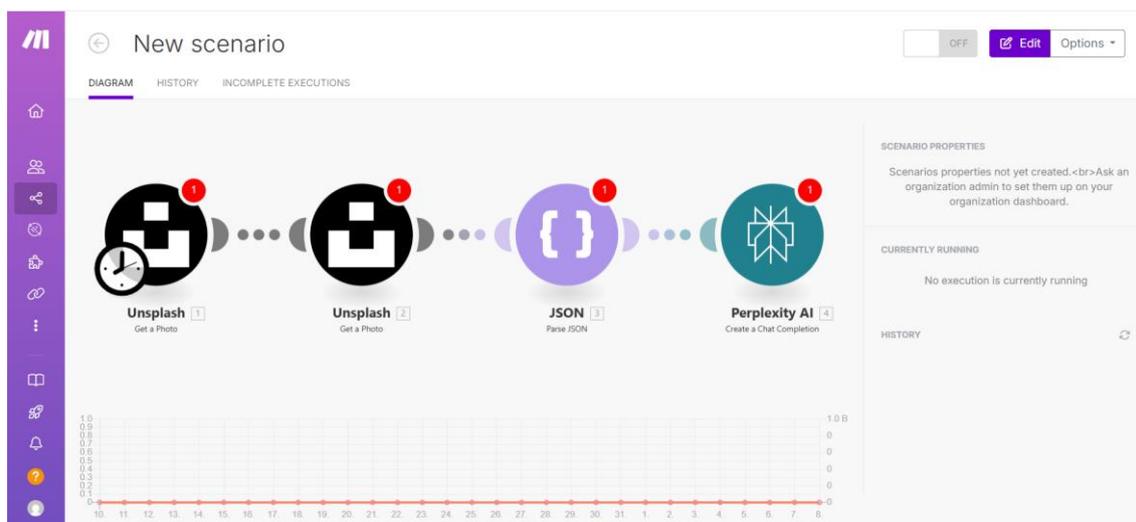


Fig 73: Escenario creado con make.com[21]

## 4.2. Reflexión final

En este trabajo se ha demostrado que es factible usar un modelo de lenguaje entrenado localmente, LMStudio en este caso, para corregir automáticamente exámenes tipo test. El modelo alcanzó una exactitud del 95 % en la clasificación de respuestas correctas, procesando cada examen en segundos. Si bien este nivel de precisión es comparable al que se obtendría con un simple algoritmo de comparación, la experiencia de usar IA aporta ventajas de flexibilidad y extensibilidad. Por ejemplo, el mismo enfoque se podría extender a formatos mixtos (autoaprendizaje supervisado de preguntas con retroalimentación) o integrar explicaciones automáticas.

El uso de LMStudio ha resultado muy práctico debido a que permite entrenar sin conexión, proteger la privacidad de datos y comenzar desde modelos pre-entrenados. Si bien es cierto que no hace falta un ordenador muy potente para ejecutar la mayoría de los modelos, se recomienda disponer de un RAM de 16 Gb o más para exprimir al máximo el potencial de esta herramienta. Para escalas mayores de exámenes se podría optimizar aún más, pero los resultados actuales indican que la viabilidad es alta para instituciones que dispongan de recursos informáticos adecuados. Mediante una API este modelo podría ser incluido en aplicaciones de e-learning para facilitar tanto a los profesores la corrección de un gran volumen de exámenes como para ayudar a los alumnos a estudiar para los exámenes.

En definitiva, la integración de IA en la evaluación puede enriquecer la educación y aliviar tareas docentes, pero siempre debe complementarse con supervisión y criterios pedagógicos apropiado.

## 5. Bibliografía

- [1] «LM Studio de Hugging Face: Entrenamiento local de modelos LLM | OpenWebinars». Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://openwebinars.net/blog/lm-studio-hugging-face/>
- [2] «LLaMA (gran modelo de lenguaje) - Wikipedia, la enciclopedia libre». Accedido: 9 de junio de 2025. [En línea]. Disponible en: [https://es.wikipedia.org/wiki/LLaMA\\_\(gran\\_modelo\\_de\\_lenguaje\)](https://es.wikipedia.org/wiki/LLaMA_(gran_modelo_de_lenguaje))
- [3] «Qwen - Wikipedia, la enciclopedia libre». Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://es.wikipedia.org/wiki/Qwen>
- [4] «(16) Phi-4 vs. Qwen-32B: Comparativa y Aplicabilidad para Pymes y Profesionales | LinkedIn». Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://www.linkedin.com/pulse/phi-4-vs-qwen-32b-comparativa-y-aplicabilidad-para-molina-laguna-anuyf/>
- [5] «Ethics guidelines for trustworthy AI - Publications Office of the EU». Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>
- [6] C. Zhang, «Automatic Generation of Multiple-Choice Questions», mar. 2023, Accedido: 18 de junio de 2025. [En línea]. Disponible en: <https://arxiv.org/pdf/2303.14576>
- [7] M. Blšták y V. Rozinajová, «Automatic question generation based on sentence structure analysis using machine learning approach», *Nat Lang Eng*, vol. 28, n.º 4, pp. 487-517, may 2022, doi: 10.1017/S1351324921000139.
- [8] C. Zhang, «Automatic Generation of Multiple-Choice Questions», mar. 2023, Accedido: 18 de junio de 2025. [En línea]. Disponible en: <https://arxiv.org/pdf/2303.14576>
- [9] K. Li y Y. Zhang, «Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation», *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4715-4729, 2024, doi: 10.18653/V1/2024.FINDINGS-ACL.280.
- [10] C. Raffel *et al.*, «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer», *Journal of Machine Learning Research*, vol. 21, pp. 1-67, oct. 2019, Accedido: 18 de junio de 2025. [En línea]. Disponible en: <https://arxiv.org/pdf/1910.10683>
- [11] J. Devlin, M. W. Chang, K. Lee, y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171-4186, oct. 2018, Accedido: 18 de junio de 2025. [En línea]. Disponible en: <https://arxiv.org/pdf/1810.04805>
- [12] «BERT (modelo de lenguaje) - Wikipedia, la enciclopedia libre». Accedido: 9 de junio de 2025. [En línea]. Disponible en: [https://es.wikipedia.org/wiki/BERT\\_\(modelo\\_de\\_lenguaje\)](https://es.wikipedia.org/wiki/BERT_(modelo_de_lenguaje))
- [13] «What is BERT and how is it Used in AI?» Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://h2o.ai/wiki/bert/>
- [14] «T5 (language model) - Wikipedia». Accedido: 9 de junio de 2025. [En línea]. Disponible en: [https://en.wikipedia.org/wiki/T5\\_\(language\\_model\)](https://en.wikipedia.org/wiki/T5_(language_model))

- [15] «PaLM (Modelo de lenguaje) - Wikipedia, la enciclopedia libre». Accedido: 9 de junio de 2025. [En línea]. Disponible en: [https://es.wikipedia.org/wiki/PaLM\\_\(Modelo\\_de\\_lenguaje\)](https://es.wikipedia.org/wiki/PaLM_(Modelo_de_lenguaje))
- [16] «Local AI Explained: Fast, Private, and On Your Device». Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://objectbox.io/local-ai-what-it-is-and-why-we-need-it/>
- [17] «IA local para empresas 2025». Accedido: 18 de junio de 2025. [En línea]. Disponible en: <https://optimiasolution.com/ia-local-para-empresas/>
- [18] «(17) Por qué la ejecución de IA local es la mejor opción: Privacidad, Autonomía y Rendimiento | LinkedIn». Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://www.linkedin.com/pulse/por-qu%C3%A9-la-ejecuci%C3%B3n-de-ia-local-es-mejor-opci%C3%B3n-y-jose-luis-sierra--mqiic/>
- [19] «Precios | OpenAI». Accedido: 18 de junio de 2025. [En línea]. Disponible en: [https://openai.com/es-ES/api/pricing/?utm\\_source=chatgpt.com](https://openai.com/es-ES/api/pricing/?utm_source=chatgpt.com)
- [20] E. Hu *et al.*, «LoRA: Low-Rank Adaptation of Large Language Models», *ICLR 2022 - 10th International Conference on Learning Representations*, jun. 2021, Accedido: 9 de junio de 2025. [En línea]. Disponible en: <https://arxiv.org/pdf/2106.09685>
- [21] «New scenario | Make». Accedido: 30 de junio de 2025. [En línea]. Disponible en: <https://eu2.make.com/1571599/scenarios/4362071/edit>

## 6. Anexos

### Creaciones de plantillas para entrenamiento

Primer intento de plantilla

#### Plantilla de Corrección - Examen Tipo Test

Nombre:

Tipo:  1

Apellidos:

2

DNI:

Fecha:

	A	B	C	D
1.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	A	B	C	D
21.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

## Primera conversación

Se envía un prompt intentando que comprenda el formato del Word y se recibe una respuesta en más de 2 mins. Como se aprecia en la parte superior de la captura, se utilizó deepseek-r1-distill-qwen-7b y ocupa 4.36 Gb

The screenshot shows a chat window with the following elements:

- Header: "Primer plantilla" on the left, and "Apariencia del chat", "Borrar todos los mensajes", and "Duplicar chat" on the right.
- User input: "You" with a document icon and the text "Plantilla de Corrección.docx".
- User message: "Te voy a proporcionar una plantilla de corrección de un examen tipo test con 22 preguntas donde cada una de ella tiene 4 opciones (A, B, C o D) necesito que aprendas su formato y sepas capaz de identificar todos los elementos de la misma."
- Model name: "deepseek-r1-distill-qwen-7b".
- System status: Two green checkmarks indicating "Chosen context injection strategy: 'inject-full-content'. All content can fit into the context" and "Done predicting".
- Thoughts: A dashed box containing "Thoughts" and a sub-entry "> Thought for 2 minutes 11 seconds".

Esta fue la respuesta, pudo comprender algunos campos, pero no leía bien las casillas

Based on the analysis of the provided content, here is a response tailored to your request:

### Formato de la Plantilla de Corrección para un Examen Tipo Test:

1. **Nombre:** Debe incluir el nombre completo del participante.
2. **Tipo:** Indica el tipo de examen (por ejemplo, "Examen Tipo Test").
3. **Apellido(es):** Listing los apellidos del participante.
4. **DNI:** Número de Documento Nacional (si esApplicable).
5. **Fecha:** La fecha en que se realizó el examen.
6. **Preguntas:** Una lista de preguntas numeradas con opciones de respuesta (A, B, C, D).

Le envíe los siguientes prompt para que comprendiese un poco más el funcionamiento de la plantilla

You

El documento que te he proporcionado es una plantilla que deberas utilizar mas adelante para corregir exámenes. El tipo de examen se indica marcando con una "X" en los cuadrados que tiene al lado para indicar si es tipo 1 o tipo 2. Necesito que seas capaz de entender que los 4 cuadrados que aparecen al lado de cada numero de pregunta, indican las 4 opciones de esta. El cuadrado que esté marcado con una "X" será la respuesta seleccionada por el alumno y tu debes ser capaz de reconocer que opción ha sido marcada en cada pregunta

deepseek-r1-distill-qwen-7b

Thoughts

> Thought for 5 minutes 10 seconds

Se modificó la anterior plantilla, pero con opciones marcadas por una X

Plantilla de Corrección - Examen Tipo Test

Nombre: Diego

Tipo:  1

Apellidos: Gómez ~~Gómez~~

2

DNI: 12345678A

Fecha: 03/03/2025

	A	B	C	D
1.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
4.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
6.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
10.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
15.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
18.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
19.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	A	B	C	D
21.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

You

 Plantilla de Corrección Marcada.docx

Te voy a enviar la misma plantilla pero con opciones marcadas y no hace falta saber a que responde cada pregunta, con que me indiques que tipo de test esta seleccionado (1 o 2) y que opción está marcada en cada pregunta es suficiente. Para ello, indícame por cada pregunta, la opción marcada (A,B,C o D) o si está en blanco

deepseek-r1-distill-qwen-7b

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Thoughts

&gt; Thought for 4 minutes 1 seconds

La respuesta no fue la esperada

#### Selected Options for Each Question:

1. Question 1: [A/B/C/D]
2. Question 2: [A/B/C/D]
3. Question 3: [X Marked Option]
4. Question 4: [A/B/C/D]
5. Question 5: [A/B/C/D]
6. Question 6: [A/B/C/D]
7. Question 7: [A/B/C/D]
8. Question 8: [X Marked Option]
9. Question 9: [A/B/C/D]
10. Question 10: [A/B/C/D]
11. Question 11: [A/B/C/D]
12. Question 12: [A/B/C/D]
13. Question 13: [X Marked Option]

Se intentó que aprendiese a detectar bien la primera pregunta para que así pueda identificar el resto. Seguía tardando demasiado además de responder en inglés

You

Teniendo en cuenta solo la pregunta 1, aprende a detectar que la opción marcada es la A ya que tiene una "X" dentro

deepseek-r1-distill-qwen-7b

Thoughts

> Thought for 3 minutes 36 seconds

**Respuesta:**

Based on the analysis of the plantilla de corrección Marcada.docx and considering only question 1 for simplicity (as requested), here is a detailed breakdown:

**Question 1:**

**Opción Marked:** A

**Justificación:** La casilla correspondiente a la opción A contiene un "X", indicando que esta es la opción seleccionada.

No supo leer bien el resto ya que intentaba crear opciones de tipo test

You

Teniendo en cuenta esto, indicame las opciones que se han seleccionado en las primeras 5 preguntas



deepseek-r1-distill-qwen-7b

Thoughts

> Thought for 3 minutes 49 seconds

Based on your request and considering only the first five questions with their respective options marked in the plantilla de corrección Marcada.docx, here is an analysis of the selected options:

Plantilla de Corrección Marcada - Primeros 5 Preguntas

Número de Pregunta	Opción A	Opción B	Opción C	Opción D
1	Rojo	Azul	gris	gris
2	Londres	París	teghenwerda	Alijó
3	-	-	-	-
4	-	-	-	-
5	-	-	-	-

Respuesta:

Opciones Marcadas en las Primeras 5 Preguntas:

1. Pregunta 1: Se ha seleccionado la opción A (Rojo).
2. Pregunta 2: No hay opciones seleccionadas ( x ) en las opciones proporcionadas.
3. Pregunta 3: No hay opciones seleccionadas ( x ) en las opciones proporcionadas.
4. Pregunta 4: No hay opciones seleccionadas ( x ) en las opciones proporcionadas.
5. Pregunta 5: No hay opciones seleccionadas ( x ) en las opciones proporcionadas.

Si necesitas más detalles o ajustes al formato, no dudes en comunicármelo.

2.12 tok/sec • 848 tokens • 21.22s to first token • Stop reason: EOS Token Found

La IA no comprendió como detectar las opciones marcadas o no detectaba bien el formato de la plantilla por lo que decidí cambiarla. Eliminé varios campos para que la IA se centre simplemente en detectar las opciones marcadas. Así que la nueva plantilla sería la siguiente:

#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	[Pregunta 2]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	[Pregunta 3]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4	[Pregunta 4]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	[Pregunta 5]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
6	[Pregunta 6]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	[Pregunta 7]	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	[Pregunta 8]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	[Pregunta 9]	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
10	[Pregunta 10]	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

## Segunda conversación

Se intentó ver si era capaz de crearme una plantilla, pero en vez de generar una plantilla vacía, generaba las preguntas.

You

Necesito que me hagas una plantilla de correccion de examen tipo test de 10 preguntas con 4 opciones para que posteriormente yo la rellene y tu seas capaz de identificar que opción ha sido marcada



deepseek-r1-distill-qwen-7b

Thoughts

> Thought for 1 minutes 1 seconds

¡Claro! Aquí tienes una plantilla de corrección de examen tipo test con 10 preguntas, cada una con 4 opciones. Eres el que rellenarás las opciones, y te indicaré las respuestas correctas.

### Plantilla de corrección del examen

Pregunta	Opción A	Opción B	Opción C	Opción D	Respuesta Correcta
1. ¿Cuál es el capital de Perú?	A) Lima	B) Santiago	C) Arequipa	D) Chimbote	A) Lima
2. ¿Qué animal pertenece a la familia Canif fearfis?	A) Gato	B) Perro	C) Varelo	D) Monaca	B) Perro
3. ¿Cuál es el mayor planeta del sistema solar?	A) Neptuno	B) Júpiter	C) Saturno	D) Urano	B) Júpiter
4. ¿Qué animal vive en el lecho marino más profundo?	A) Elodida	B) Bl course	C) Quark	D) Flamingo	A) Elodida

Se pidió un tipo test para posteriormente corregirlo con la plantilla marcada.

You

Dame 10 preguntas tipo test con 4 opciones

deepseek-r1-distill-qwen-7b

Thoughts

> Thought for 48.47 seconds

¡Claro! Aquí tienes 10 preguntas tipo test con 4 opciones cada una. Son preguntas variadas y de dificultad media para practicar y repasar conceptos.



### Tercera conversación

Se comenzó pidiendo un tipo test

You

Hazme un tipo test de 10 preguntas con 4 opciones

deepseek-r1-distill-qwen-7b

Thoughts

> Thought for 1 minutes 1 seconds

¡Claro! Aquí tienes un tipo de test con 10 preguntas y 4 opciones cada una:

#### Test 1: Conocimiento General

1. ¿Qué animal come manzanas?
  - a) Perro
  - b) Gato
  - c) Vaca
  - d) Abeja
2. ¿Cuál es el clima de hoy?
  - a) Láser
  - b) Sol
  - c) Precipitación
  - d) tormenta

Resalto la pregunta número 10 ya que al utilizar deepseek podemos ver cómo nos muestra caracteres chinos por error.

10. ¿Cuál es el color de la hierba?
  - a) Verde
  - b) Gris
  - c) Rojo
  - d) Añ不出

Se procedió a enviarle la plantilla con opciones marcadas para ver si era capaz de corregir el test que acababa de proporcionar con ellas.

 Borrador plantilla.docx

Te envío la plantilla con las respuestas marcadas, envíame las selecciones que aparecen en el documento y calculame una calificación total teniendo en cuenta que cada pregunta correcta es un punto y cada pregunta errónea resta 0.33. Si no se ha marcado ninguna opción, no resta

deepseek-r1-distill-qwen-7b

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Thoughts

&gt; Thought for 2 minutes 49 seconds

La calificación obtenida al completar el test con las opciones seleccionadas es **8 puntos**.

Podemos observar que la puntuación es un 8. Como solo hay 8 opciones marcadas en la plantilla y alguna de ellas es errónea, comprobé que me había marcado como correctas todas las opciones por lo que le pedí un desglose de la puntuación. Observamos que además de volver a ser muy lenta la respuesta, vuelve a ser errónea ya que dice que hemos respondido 9 preguntas cuando no es así.

You

Puedes desglosarme la forma en la que obtienes esa puntuación?

deepseek-r1-distill-qwen-7b

Thoughts

&gt; Thought for 6 minutes 29 seconds

La calificación de **8 puntos** se logra al completar el test de la siguiente manera:

1. Seleccionó las respuestas correctas para **9 preguntas**.
2. Dejó en blanco las **2 preguntas restantes**.
3. Cada respuesta correcta otorga 1 punto, y no hay penalización por las inyeses (siempre que no se marcaron opciones).

Por lo tanto, la puntuación se calcula así:

- **Puntos por respuestas correctas:** 9 preguntas × 1 punto = 9 puntos
- **PUNTOS POR INYES:** 2 preguntas en blanco × 0 = 0 puntos

Finalmente, la calificación total es **8 puntos**.

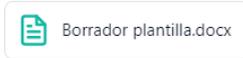
3.26 tok/sec • 1431 tokens • 4.63s to first token • Stop reason: EOS Token Found



### Cuarta conversación.

Volvemos a probar lo mismo, la IA nos da un test con las respuestas y le pido que lo compare con las opciones seleccionadas por el alumno. Pero tras varios intentos, esta vez, no consigue leer la plantilla.

You



Las respuestas del alumno estan en este documento. Indicame cuales son y dame una puntuación

deepseek-r1-distill-qwen-7b

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Thoughts

> *Thought for 1 minutes 36 seconds*

Claro, por favor proporciona las opciones seleccionadas del estudiante para cada pregunta, y myself I will compare them with the correct answers and provide a detailed comparison.

4.06 tok/sec • 443 tokens • 22.10s to first token • Stop reason: EOS Token Found

### Quinta conversación

Debido a los fallos mostrados en las anteriores conversaciones y los largos tiempos de espera, decido probar con otro modelo. En este caso usaré llama-3-8b-gpt-4o-ru1.0 el cual ocupa 4.58 Gb.

Como es un modelo nuevo, pruebo a mandarle una plantilla con todas las opciones marcadas que tomaremos como referencia para ver si es capaz de leerlas bien.

#	Pregunta	Opción A	Opción B	Opción C	Opción D
1	[Pregunta 1]	[X]	[ ]	[ ]	[ ]
2	[Pregunta 2]	[ ]	[X]	[ ]	[ ]
3	[Pregunta 3]	[ ]	[ ]	[X]	[ ]
4	[Pregunta 4]	[ ]	[ ]	[ ]	[X]
5	[Pregunta 5]	[ ]	[ ]	[X]	[ ]
6	[Pregunta 6]	[ ]	[X]	[ ]	[ ]
7	[Pregunta 7]	[X]	[ ]	[ ]	[ ]
8	[Pregunta 8]	[ ]	[X]	[ ]	[ ]
9	[Pregunta 9]	[ ]	[ ]	[X]	[ ]
10	[Pregunta 10]	[ ]	[ ]	[ ]	[X]

Además, le indico que son las soluciones correctas de un test ya que posteriormente le pediré que las compare con las opciones marcadas por un alumno

You



Te voy a pasar una plantilla de corrección de un examen tipo test de 10 preguntas con 4 opciones, en este documento vienen indicadas las opciones correctas. Puedes indicarme cuales son?

llama-3-8b-gpt-4o-ru1.0

- ✔ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✔ Done predicting

Sí, puedo ayudarte con eso. Después de revisar el contenido que me proporcionaste, puedo ver que las opciones correctas están marcadas con un [X]. Aquí te presento la información:

1. **Opción B** - Correcto
2. **Opción C** - Correcto
3. **Opción A** - Incorrecto
4. **Opción D** - Incorrecto
5. **Opción X** - Correcto
6. **Opción A** - Incorrecto
7. **Opción A** - Correcto
8. **Opción A** - Correcto
9. **Opción D** - Correcto
10. **Opción X** - Correcto

Como se aprecia que algunas las indica como incorrectas, se le explica que tome todas como correctas ya que va a ser la referencia para posteriormente compararlas con otros test

You

Me has entendido mal, todas las respuestas son correctas, este es un ejemplo para que veas cuales son las soluciones de un test ya que mas adelante te voy a ir enviando hojas de marcas diferentes para que tu las compares con esta

llama-3-8b-gpt-4o-ru1.0

Entiendo mejor ahora. En ese caso, la respuesta es: Todas las opciones correctas están indicadas en el documento como [X], lo que significa que todas son correctas. Si tienes más documentos para comparar, estaré feliz de ayudarte a identificar las soluciones.

A la hora de pedirle que realice la comparación, solo explicaba cómo hacerlo en lugar de hacerlo.

You



Borrador plantilla.docx

Muy bien, eso es. Ahora te envio las respuestas que ha seleccionado un alumno y vas a tener que compararlo con el documento anterior, es decir si tienen la misma respuesta seleccionada, es correcta y sumas un punto. Si no tienen la misma seleccionada, restas 0.33 y si el alumno no ha respondido nada, ni sumas ni restas

llama-3-8b-gpt-4o-ru1.0



Chosen context injection strategy: 'inject-full-content'. All content can fit into the context



Done predicting

Entiendo. Para responder a tu pregunta, te diré que la tarea consiste en comparar las respuestas proporcionadas por el usuario con las opciones correctas presentadas en el documento anterior (Opciones correctas.docx). Si se encuentra la misma opción seleccionada como [X] en ambos documentos, se considera como respuesta correcta y se otorga un punto. En caso contrario, no se aplica ningún ajuste (0.33 puntos negativos).

#### Resumen de la tarea:

1. Compara las respuestas proporcionadas por el usuario con el documento "Opciones correctas.docx".
2. Si coinciden en la opción seleccionada como [X], es respuesta correcta y se otorga 1 punto.
3. En caso contrario, no se aplica ajuste (0.33 puntos negativos) ni sumo/resto puntos.

### Sexta conversación

En esta nueva conversación vuelvo a intentar lo mismo pero esta vez el modelo dice que no puede leer documentos.

You



Vas a realizar la función de un profesor ya que te voy a enviar un documento llamado "Opciones correctas" en el cual hay una plantilla de un examen tipo test de 10 preguntas y 4 opciones (A, B, C o D) debes de ser capaz de leer este primer documento e indicarme cuales son las opciones correctas ya que vienen indicadas por una "X" así que lee el documento y dime que opciones están marcadas

llama-3-8b-gpt-4o-ru1.0

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Desafortunadamente, como servicio de asistencia y aprendizaje, no puedo acceder a archivos o documentos compartidos directamente. Sin embargo, puedes proporcionarme la información contenida en el archivo "Opciones correctas.docx" y te diré qué opciones están marcadas con un "X".

Se crea un nuevo chat y esta vez sí que consigue mostrar una respuesta, aunque lee mal el documento.

You



Indicame que opciones están seleccionadas en esta plantilla

llama-3-8b-gpt-4o-ru1.0

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Entiendo. Según el contenido del archivo "Opciones correctas.docx", las opciones marcadas con un "X" son:

1. [Pregunta 1] Opción B
2. [Pregunta 2] Opción C
3. [Pregunta 3] Opción D
4. [Pregunta 4] Opción A
5. [Pregunta 5] Opción D
6. [Pregunta 6] Opción B
7. [Pregunta 7] Opción C
8. [Pregunta 8] Opción D
9. [Pregunta 9] Opción A
10. [Pregunta 10] Opción C

Espero que esta información sea útil para ti.

3.02 tok/sec • 170 tokens • 47.49s to first token • Stop reason: EOS Token Found

Se probó a enviarle otra vez el documento y volvió a responder que no podía leer documentos.

You

 Opciones correctas.docx

Esas no son las opciones que han sido marcadas, lee este documento y vuelve a mandarmelas



llama-3-8b-gpt-4o-ru1.0

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Lo siento mucho. Sin embargo, no puedo leer o acceder directamente a contenido de archivos o documentos, como el "Opciones correctas.docx" mencionado en tu texto. Si necesitas ayuda con una pregunta específica sobre qué opción está marcada con un "X" en cada pregunta, te sugiero que reescribas la información textualmente aquí y te ayudaré a interpretarla.

2.66 tok/sec • 88 tokens • 47.04s to first token • Stop reason: EOS Token Found

## Séptima conversación

Se intentó que este modelo crease una plantilla para que el mismo pueda ver bien que opciones han sido seleccionadas. Pero otra vez, explicó cómo hacerlo en lugar de proporcionar una.

You

Necesito que me hagas una plantilla para un tipo test de 10 preguntas con 4 opciones en documento pdf que te sea fácil de entender

llama-3-8b-gpt-4o-ru1.0

Claro, pero como soy una inteligencia artificial, no puedo crear archivos PDF directamente. Sin embargo, puedo explicarte cómo hacerlo y proporcionarte la lógica para crear una plantilla con preguntas y respuestas. Puedes adaptar esta explicación según tus necesidades.

### 1. Crear un documento de Excel o Google Sheets

- Abre Microsoft Excel o Google Sheets en tu navegador.
- Haz clic en "Archivo" > "Nueva" para crear un nuevo archivo.
- Selecciona una hoja blanca y coloca el título de tu test en la celda A1.

### 2. Diseño de la plantilla

- **Preguntas:** En las filas desde la segunda fila hasta la última fila que necesites, escribe cada pregunta en una celda.
- **Opciones de respuesta:** Debajo de cada pregunta, crea cuatro columnas (A-D) para las opciones de respuesta. Coloca cada opción en su respectiva celda.

### 3. Formato de las preguntas y respuestas

- Asegúrate de que la información sea clara y concisa.
- Utiliza formatos de texto para distinguir las preguntas de las opciones de respuesta. Puedes utilizar negrita o cursiva para hacerlo más visible.

Ejemplo:

Pregunta 1	A) Opción 1	B) Opción 2	C) Opción 3
------------	-------------	-------------	-------------

Pregunta 2	A) Opción 4	B) Opción 5	C) Opción 6	D) Opción 7
------------	-------------	-------------	-------------	-------------

...

### Octava conversación

Debido a que los dos modelos anteriores no cumplieron con las funciones esperadas. Aunque el segundo había mejorado los tiempos de respuesta, se decidió probar con un modelo aún más grande. La elección fue Phi 4, el cual ocupa 7.39 Gb. A la hora de descargar el modelo, LM Studio indicó que era un modelo muy grande y podría dar problemas.

Esta vez el primer paso fue comprobar que era capaz de comparar opciones antes de enviarlas en una plantilla por lo que se le envió el siguiente prompt.

You

Te voy a mandar las soluciones a un examen tipo test y debes compararlas con las opciones que han elegido los alumnos para posteriormente dar una calificación.

La soluciones son las siguientes:

- 1 A
- 2 B
- 3 C
- 4 D
- 5 C
- 6 B
- 7 A
- 8 B
- 9 C
- 10 D

Posteriormente, le envié las opciones del alumno y le pedí que me diese una calificación.

You

Ahora te voy a proporcionar las opciones seleccionadas por un alumno y deberás compararlas con las correctas

- 1. A
- 2. B
- 3. D
- 4. D
- 5. C
- 6. A
- 7. A
- 8. B
- 9. B
- 10. D

También necesito que me des una calificación donde cada pregunta correcta suma 1 punto y cada respuesta errónea reste 0.33 puntos

phi-4

¡Por supuesto! Usaremos el sistema de puntuación en el que cada respuesta correcta suma 1 punto, mientras que cada respuesta incorrecta resta 0.33 puntos. Vamos a calcular la calificación del alumno basándonos en las respuestas proporcionadas anteriormente.

### Evaluación Detallada

- **Preguntas Correctas (suman 1 punto cada una):**
  - Pregunta 1: Correcta, suma 1 punto.
  - Pregunta 2: Correcta, suma 1 punto.
  - Pregunta 4: Correcta, suma 1 punto.
  - Pregunta 5: Correcta, suma 1 punto.
  - Pregunta 8: Correcta, suma 1 punto.
  - Pregunta 10: Correcta, suma 1 punto.
- **Preguntas Incorrectas (restan 0.33 puntos cada una):**
  - Pregunta 3: Incorrecta, resta 0.33 puntos.
  - Pregunta 6: Incorrecta, resta 0.33 puntos.
  - Pregunta 7: Incorrecta, resta 0.33 puntos.
  - Pregunta 9: Incorrecta, resta 0.33 puntos.

### Cálculo de la Calificación

- **Puntos ganados por respuestas correctas:**  $6 \times 1 = 6$
- **Puntos perdidos por respuestas incorrectas:**  $4 \times (-0.33) = -1.32$

**Calificación Total:**  $6 - 1.32 = 4.68$

Así que, usando este sistema de calificación, el alumno obtiene un total de **4.68 puntos** sobre los 10 posibles.

Viendo que compara bien y me da una calificación correcta, le pido lo mismo, pero pasándole las opciones correctas y las del alumno en documentos. Se equivoca a la hora de corregir ya que las preguntas 2, 6, 7 y 10 son correctas. Pero lee bien el documento de opciones correctas.

You



Muy bien, ahora necesito hacer exactamente lo mismo pero te voy a enviar las Soluciones Correctas en un documento llamado "Opciones correctas" y las Respuestas del Alumno van a ir en un documento llamado "Borrador plantilla"

phi-4



¡Claro! Utilizaré el contenido proporcionado de ambos archivos para calcular la calificación del alumno según las reglas que mencionaste: cada respuesta correcta suma 1 punto y cada respuesta incorrecta resta 0.33 puntos.

## Comparación de Respuestas

### Respuestas Correctas

- Pregunta 1: A (Correcto)
- Pregunta 2: B (Incorrecto, el correcto es A)
- Pregunta 3: C (Correcto)
- Pregunta 4: D (Correcto)
- Pregunta 5: Ninguna seleccionada (Incorrecto, el correcto es A)
- Pregunta 6: C (Incorrecto, el correcto es B)
- Pregunta 7: B (Incorrecto, el correcto es A)
- Pregunta 8: A (Correcto)
- Pregunta 9: Ninguna seleccionada (Incorrecto, el correcto es C)
- Pregunta 10: B (Incorrecto, el correcto es A)

### Cálculo de Calificación

1. Preguntas Correctas: 1, 3, 4, y 8

- Total de Puntos por Respuestas Correctas =  $4 \times 1 = 4$  puntos

2. Preguntas Incorrectas: 2, 5, 6, 7, 9, y 10

- Total de Puntos por Respuestas Incorrectas =  $6 \times (-0.33) = -1.98$  puntos

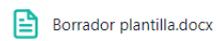
3. Calificación Final Total =  $4 + (-1.98) = 2.02$

El alumno obtiene una calificación de **2.02 puntos** sobre un total posible de 10, usando el sistema de puntuación que especificaste.

### Novena conversación

En esta nueva conversación, voy a enviarle el documento que leía bien por teclado para que así se centre en leer bien el documento de las opciones seleccionadas por el alumno. Se observa que corrige bien el test.

You



Necesito que hagas de profesor y corrijas un tipo test de 10 preguntas con 4 opciones (A, B, C o D), te voy a enviar las Soluciones Correctas por teclado y las Respuestas del Alumno van a ir en un documento llamado "Borrador plantilla"

Primero muéstrame que opciones ha seleccionado el alumno y luego compara las respuestas, si la respuesta es correcta suma un punto, si es incorrecta, resta 0.33. Si el alumno no ha respondido no restes ni sumes.

Soluciones Correctas:

- 1 A
- 2 B
- 3 C
- 4 D
- 5 C
- 6 B
- 7 A
- 8 B
- 9 C
- 10 D

phi-4

- ✓ Chosen context injection strategy: 'inject-full-content'. All content can fit into the context
- ✓ Done predicting

Para abordar esta tarea, primero identificaré las opciones seleccionadas por el alumno en "Borrador plantilla.docx" y luego compararé estas respuestas con las soluciones correctas proporcionadas. Finalmente, calcularé la puntuación del examen según las reglas dadas.

### Opciones Seleccionadas por el Alumno:

1. Pregunta 1: Opción A
2. Pregunta 2: Opción B
3. Pregunta 3: Opción C
4. Pregunta 4: Opción D
5. Pregunta 5: No respondida
6. Pregunta 6: Opción C
7. Pregunta 7: Opción B
8. Pregunta 8: Opción A
9. Pregunta 9: No respondida
10. Pregunta 10: Opción B

### Comparación de Respuestas con Soluciones Correctas:

#### 1. Pregunta 1:

- Alumno: A (Correcta)
- Puntuación: +1

#### 2. Pregunta 2:

- Alumno: B (Correcta)
- Puntuación: +1

#### 3. Pregunta 3:

- Alumno: C (Correcta)
- Puntuación: +1

4. **Pregunta 4:**

- Alumno: D (Correcta)
- Puntuación: +1

5. **Pregunta 5:**

- No respondida
- Puntuación: 0

6. **Pregunta 6:**

- Alumno: C (Incorrecta, correcta es B)
- Puntuación: -0.33

7. **Pregunta 7:**

- Alumno: B (Incorrecta, correcta es A)
- Puntuación: -0.33

8. **Pregunta 8:**

- Alumno: A (Incorrecta, correcta es B)
- Puntuación: -0.33

9. **Pregunta 9:**

- No respondida
- Puntuación: 0

10. **Pregunta 10:**

- Alumno: B (Incorrecta, correcta es D)
- Puntuación: -0.33

**Cálculo Final de la Puntuación:**

- Total Correctas: 4 preguntas
- Total Incorrectas: 3 preguntas
- No respondidas: 2 preguntas

Puntaje Total = (Correctas  $\times$  1) + (Incorrectas  $\times$  -0.33)

$$\text{Puntuación} = (4 \times 1) + (3 \times -0.33) = 4 - 0.99 = 3.01$$

La puntuación final del alumno es **3.01**.

1.67 tok/sec • 616 tokens • 35.67s to first token • Stop reason: EOS Token Found

## Décima conversación

En esta conversación se vuelve a probar si el modelo es capaz de crear una plantilla.

You

Necesito que me generes una plantilla para un tipo test con 10 preguntas y 4 opciones cada una. Necesito que tenga casillas en las que se pueda marcar con una "X" la opción que se desea seleccionar

phi-4

¡Por supuesto! A continuación te presento una plantilla simple de tipo test que contiene 10 preguntas, cada una con cuatro opciones de respuesta, y espacios para marcar con una "X" la opción elegida.

Pero en vez de crear una plantilla genérica vacía, crea un test

### Test

#### Pregunta 1:

¿Cuál es el capital de España?

- A) Barcelona
- B) Madrid (Selecciona aquí si es tu respuesta)
- C) Sevilla
- D) Valencia

Por lo que se intenta de nuevo cambiando el prompt y esta vez lo crea bien

You

La plantilla no debe contener preguntas ni soluciones, solo el número de pregunta en cada fila y la opción en cada columna

phi-4

Entiendo, a continuación te presento una plantilla estructurada para un test con 10 preguntas, donde puedes agregar las preguntas y opciones por tu cuenta. La plantilla incluye espacios para marcar con "X" la opción elegida.

Pregunta	Opción A	Opción B	Opción C	Opción D
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

### Undécima conversación

En esta conversación se consigue una plantilla al primer intento

You

Necesito que me generes una plantilla para un tipo test con 10 preguntas y 4 opciones cada una. Necesito que tenga casillas en las que se pueda marcar con una "X" la opción que se desea seleccionar. La plantilla no debe contener preguntas ni soluciones, solo el numero de pregunta en cada fila y la opción en cada columna

phi-4

Aquí tienes una plantilla para un test con 10 preguntas y 4 opciones cada una. Cada pregunta está numerada y hay casillas al lado de cada opción donde puedes marcar una "X" para seleccionar tu respuesta.

Pregunta	Opción A	Opción B	Opción C	Opción D
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Instrucciones:

- Marca con una "X" la casilla correspondiente a tu opción seleccionada para cada pregunta.
- Asegúrate de marcar solo una opción por pregunta.

### Chat perfecto pasando todo por teclado

Primero se repite el prompt anterior para conseguir la plantilla. Posteriormente rellenamos las opciones y usando Mistral enviamos por teclado las opciones.

Dime que opciones estan seleccionadas en esta plantilla para un tipo test con 10 preguntas y 4 opciones (A,B,C o D)

### Pregunta Opción A Opción B Opción C Opción D

```

1 [Pregunta 1]  [X]  [][][]
2 [Pregunta 2]  [][X] [][]
3 [Pregunta 3]  [][] [X] []
4 [Pregunta 4]  [][][] [X]
5 [Pregunta 5]  [][][] [X]
6 [Pregunta 6]  [][] [X] []
7 [Pregunta 7]  [] [X]  [][]
8 [Pregunta 8]  [X]  [][][]
9 [Pregunta 9]  [X]  [][][]
10 [Pregunta 10] [] [X]  [][]
    
```

phi-4

Basado en la plantilla proporcionada, aquí están las opciones seleccionadas para cada pregunta:

1. **Pregunta 1:** Opción A está seleccionada.
2. **Pregunta 2:** Opción B está seleccionada.
3. **Pregunta 3:** Opción C está seleccionada.
4. **Pregunta 4:** Opción D está seleccionada.
5. **Pregunta 5:** Opción D está seleccionada.
6. **Pregunta 6:** Opción C está seleccionada.
7. **Pregunta 7:** Opción B está seleccionada.
8. **Pregunta 8:** Opción A está seleccionada.
9. **Pregunta 9:** Opción A está seleccionada.
10. **Pregunta 10:** Opción B está seleccionada.

Cada pregunta tiene una opción marcada con un "X", indicando la elección realizada.

Le pasamos las opciones del alumno y explicamos las puntuaciones

Como corrige perfectamente seguimos enviando plantillas de distintos alumnos

You

Ahora te voy a enviar la plantilla marcada de otro alumno

**Pregunta Opción A Opción B Opción C Opción D**

- |                  |                                     |                                     |                                     |                                     |
|------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| 1 [Pregunta 1]   | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> |
| 2 [Pregunta 2]   | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| 3 [Pregunta 3]   | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/>            |
| 4 [Pregunta 4]   | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/>            |
| 5 [Pregunta 5]   | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> |
| 6 [Pregunta 6]   | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            |
| 7 [Pregunta 7]   | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> |
| 8 [Pregunta 8]   | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| 9 [Pregunta 9]   | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/>            |
| 10 [Pregunta 10] | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            |

**Evaluación:**

- Preguntas Correctas: 5, 9, 10 (3 preguntas correctas)
- Puntos obtenidos por respuestas correctas:  $3 * 1 = 3$  puntos
- Preguntas Incorrectas:
  - Pregunta 1: D (Incorrecta) → Resta 0.33
  - Pregunta 2: C (Incorrecta) → Resta 0.33
  - Pregunta 3: A (Incorrecta) → Resta 0.33
  - Pregunta 4: A (Incorrecta) → Resta 0.33
  - Pregunta 6: B (Incorrecta) → Resta 0.33
  - Pregunta 7: D (Incorrecta) → Resta 0.33
  - Pregunta 8: C (Incorrecta) → Resta 0.33
- Preguntas No Puntuadas: Ninguna.

**Cálculo de la Puntuación Total:**

- Total por respuestas correctas: 3 puntos
- Descuento por respuestas incorrectas:  $7 * (-0.33) = -2.31$

**Puntos Totales:**  $3 - 2.31 = 0.69$

**Calificación Basada en la Puntuación:**

- Menos de 5: Suspenso
- Entre 5 y 5.99: Suficiente
- Entre 6 y 6.99: Bien
- Entre 7 y 8.99: Notable
- 9 o más: Sobresaliente

**Resultado Final:**

La puntuación total es **0.69**, lo que corresponde a un **Suspenso** según el criterio establecido.

0.03 tok/sec • 634 tokens • 109.15s to first token • Stop reason: EOS Token Found

