Tesla Revista Científica, ISSN: 2796-9320
Vol. 4 Núm. 2 (2024), e464

Ingeniería
Artículo de Investigación Original

Optimización de la planificación del sílabo en la Universidad Técnica de Cotopaxi mediante Inteligencia Artificial Generativa: Un enfoque personalizado basado en LLAMA 2 (Large Language Model Meta AI)

Optimization of syllabus planning at the Technical University of Cotopaxi using Generative Artificial Intelligence: A personalized approach based on LLAMA 2 (Large Language Model Meta AI)

Juan Diego Chiluisa Gallardo[1][0009-0007-9797-3552], Gustavo Rodríguez Bárcenas[2][0000-0002-3669-5276]

[1,2] Universidad Técnica de Cotopaxi. Latacunga, Cotopaxi. Ecuador
[1]juan.chiluisa5@utc.edu.ec, [2]gustavo.rodriguez@utc.edu.ec

CITA EN APA:

Chiluisa Gallardo, J. D., & Rodríguez Bárcenas, G. (2024). Optimización de la planificación del sílabo en la Universidad Técnica de Cotopaxi mediante Inteligencia Artificial Generativa: Un enfoque personalizado basado en LLAMA 2 (Large Language Model Meta AI). *Tesla Revista Científica*, *4*(2).

Recibido: 2024-10-13
Revisado: 2024-10-13 al 2024-10-26
Corregido: 2024-11-04
Aceptado: 2024-11-14
TESLA
Revista Científica
ISSN: 2796-9320

Resumen: En la educación superior, una planificación curricular eficiente y personalizada es esencial para garantizar la calidad del proceso de enseñanza y aprendizaje. En la Universidad Técnica de Cotopaxi, la elaboración manual de sílabos consume tiempo y recursos significativos y no siempre se adapta a las necesidades específicas de los docentes, afectando negativamente la calidad educativa. Objetivo: Desarrollar e implementar un sistema automatizado basado en el modelo de lenguaje LLAMA 2 (Large Language Model Meta AI) para optimizar la planificación curricular mediante la generación de sílabos personalizados. Métodos: La metodología incluyó la recopilación y transformación de planificaciones curriculares previas para su procesamiento por LLAMA 2, ejecutando cinco experimentos con complejidad creciente y evaluando la precisión. Se utilizó la herramienta Pandas AI para el análisis de datos y la técnica de Recuperación Aumentada por Generación (RAG) para enriquecer la generación de sílabos. Resultados: Los resultados mostraron una precisión del 92,5% en las recomendaciones generadas, representando una mejora significativa respecto a los métodos tradicionales. Conclusiones: La implementación del sistema automatizado basado en LLAMA 2 demostró mejorar significativamente la eficiencia y precisión en la generación de sílabos personalizados, optimizando la planificación curricular y contribuyendo a mejorar la calidad educativa.

Palabras Clave: Planificación Silabo, LLAMA 2, Educación superior, RAG, Inteligencia artificial.

Abstract: Abstract: In higher education, efficient and personalized curricular planning is essential to ensure the quality of the teaching and learning process. At the Technical University of Cotopaxi, the manual creation of syllabi consumes significant time and resources, and does not always adapt to the specific needs of instructors, negatively affecting educational quality. Objective: To develop and implement an automated system based on the LLAMA 2 language model (Large Language Model Meta AI) to optimize curricular planning through the generation of personalized syllabi. Methods: The methodology included the collection and transformation of previous curricular plans for processing by LLAMA 2, executing five experiments with increasing complexity and evaluating accuracy. The Pandas AI tool was used for data analysis, and the Retrieval Augmented Generation (RAG) technique was employed to enhance the generation of syllabi. Results: The results showed an accuracy of 92.5% in the generated recommendations, representing a significant improvement over traditional methods.Conclusions: The implementation of the automated system based on LLAMA 2 significantly improved the efficiency and accuracy in the generating personalized syllabi, optimizing curricular planning and contributing to the enhancement of educational quality.

Keywords: Silabo Planning, LLAMA 2, Higher Education, RAG, Artificial Intelligence.

1. INTRODUCIÓN

La planificación curricular tiene un impacto significativo en la formación integral del ser humano. Por ello, es fundamental que se adapte a los cambios del entorno, tales como los avances científicos y tecnológicos, así como a las nuevas demandas de la sociedad (Navarro et al., 2010). Según (Jerez, 2015; Toapanta Pinta et al., 2018), "El sílabo o currículo es un instrumento que gestiona el proceso de aprendizaje conforme al perfil de egreso de cada carrera de titulación universitaria, constituye una herramienta al servicio del estudiante y representa el compromiso de transmisión de conocimientos y destrezas del docente y su unidad académica; tiene carácter público, es susceptible de análisis, revisión crítica y mejoramiento".

El diseño de sílabos que integren competencias genéricas, como la lectura crítica, el trabajo en equipo y la resolución de problemas, facilita el desarrollo de habilidades transversales que son esenciales para el desarrollo académico y profesional de los estudiantes (Salamanca Leguizamón et al., 2020). Estos sílabos no solo organizan el contenido, sino que también juegan un papel crucial en la planificación de actividades que desarrollen competencias genéricas.

Además, el sílabo es una herramienta esencial para la planificación micro curricular en la educación superior, ya que organiza los aspectos fundamentales de un curso académico, incluyendo los contenidos, la metodología, las actividades de aprendizaje y las formas de evaluación. Además, representa un compromiso entre el docente y el estudiante en el proceso educativo (Romero-Sacoto et al., 2021). Un currículo basado en competencias permite que los estudiantes desarrollen habilidades críticas y prácticas en diferentes áreas, fomentando un aprendizaje activo donde la evaluación se centra en la capacidad de los estudiantes para resolver problemas. Una planificación curricular bien estructurada no solo optimiza el proceso de enseñanza-aprendizaje, sino que también mejora la articulación entre los distintos niveles de planificación institucional y de aula, lo que contribuye al éxito académico de los estudiantes (Solari, 2018).

La tecnología en la educación no solo transforma los procesos de enseñanza y aprendizaje, sino que también plantea nuevos retos para las instituciones educativas, que deben adaptarse a esta realidad cambiante. Las aplicaciones de la inteligencia artificial, como los sistemas de tutoría inteligente, ya están desempeñando un papel clave en el aprendizaje individualizado, lo que permite una mayor personalización y dinamismo en el proceso educativo (Ocaña-Fernández et al., 2019). En este ámbito, la inteligencia artificial ha comenzado a desempeñar un papel crucial en la educación superior. Herramientas como Chat-GPT, desarrolladas por OpenAI y lanzadas en noviembre de 2022, se presentan como soluciones innovadoras que pueden transformar los procesos de enseñanza y aprendizaje.

La capacidad de Chat-GPT para sintetizar y presentar información de manera comprensible ha prometido ser una herramienta poderosa en la educación, facilitando el acceso a contenidos complejos (Ojeda et al., 2023). A nivel universitario, el uso de tecnologías de inteligencia artificial como Chat-GPT se percibe como un recurso viable para el desarrollo del pensamiento crítico en los estudiantes, siempre que se utilice de manera ética y responsable (Atencio-González et al., 2023).

En este contexto, la Universidad Técnica de Cotopaxi, realiza manualmente las planificaciones de los sílabos al momento de seleccionar las metodologías, actividades, recursos y formas de evaluar a utilizar, lo que implica un proceso laborioso y poco eficiente para los docentes. Esta situación ha generado preocupación en la institución debido al tiempo y esfuerzo dedicados a esta tarea, así como a la falta de consideración de las necesidades individuales de los docentes y las características específicas de las asignaturas.

La planificación manual del sílabo implica un consumo significativo de tiempo y esfuerzo por parte de los docentes, lo que puede afectar su carga laboral y su capacidad para dedicarse plenamente a otras tareas académicas. El objetivo principal de estudio es la optimización de estos procesos a través de la inteligencia artificial generativa LLAMA 2 (Large Language Model Meta AI), reduciendo drásticamente el tiempo necesario para elaborar planes de estudio y permitiendo que los docentes dediquen más tiempo a la enseñanza y la investigación.

El desafío de implementar un modelo de inteligencia artificial generativa como LLAMA 2 (Large Language Model Meta AI), para optimizar la planificación de los sílabos radico en la utilización de diversos estudios, métodos, herramientas y técnicas que permitieron obtener resultados con más presión y menos alucinaciones.

Llama 2 es una familia de modelos de lenguaje de gran escala (LLMs) pre entrenados y ajustados para tareas de diálogo. Estos modelos varían en tamaño desde 7 mil millones hasta 70 mil millones de parámetros. Los modelos ajustados, llamados Llama 2-Chat, están optimizados para casos de uso conversacional y, en comparación con otros modelos de código abierto, han demostrado un rendimiento superior en evaluaciones de utilidad y seguridad. El objetivo de Llama 2 es proporcionar una alternativa abierta a los modelos comerciales cerrados, permitiendo a la comunidad investigar y desarrollar tecnologías de lenguaje más seguras y eficaces (Touvron et al., 2023).

Los Large Language Models (LLMs) son modelos de IA altamente capaces que sobresalen en tareas de razonamiento complejo que requieren conocimientos expertos en una amplia variedad de campos. Estos modelos permiten interacciones con humanos a través de interfaces conversacionales, lo que ha llevado a una rápida adopción pública. Los LLMs se entrenan de manera auto regresiva sobre grandes volúmenes de datos autogenerados y se ajustan con técnicas como el aprendizaje por refuerzo con retroalimentación humana (RLHF) para alinearse mejor con las preferencias humanas (Touvron et al., 2023).

Los estudios de ablación en el campo de la inteligencia artificial, particularmente en las redes neuronales artificiales, constituyen una técnica analítica esencial para comprender cómo contribuyen componentes específicos al rendimiento global de un modelo. Estos estudios implican la eliminación o alteración de partes del sistema para evaluar el impacto en su desempeño. Esta práctica es análoga a la ablación en biología, donde se extraen secciones del cerebro para estudiar sus funciones (Meyes et al., 2019).

El objetivo de este enfoque es identificar qué elementos del sistema son indispensables y cuáles son redundantes o menos críticos. Se aplica a una amplia gama de modelos de IA, incluidas las redes neuronales, modelos generativos como LLAMA 2 y sistemas de visión por computadora. Un ejemplo típico es evaluar cómo la eliminación de capas o nodos afecta la capacidad de un modelo para realizar tareas como el reconocimiento de imágenes o el procesamiento del lenguaje natural.

El método de ensayo y error, consiste en realizar pruebas o experiencias al azar hasta encontrar la solución buscada. Este método puede ser muy fecundo cuando el interesado conoce el objeto en estudio, tiene una hipótesis, que puede ser vaga pero verosímil, y aplica el método de manera sistemática (Morles, 2002).

PandasAI es una biblioteca de Python que facilita la realización de preguntas a sus datos en lenguaje natural. Más allá de las consultas, PandasAI ofrece funcionalidades para visualizar datos a través de gráficos, limpiar conjuntos de datos abordando valores faltantes y mejorar la calidad de los datos a través de la generación de características, lo que lo convierte en una herramienta integral para científicos y analistas de datos (*Introduction to PandasAI*, s. f.).

Finalmente, el desarrollo de nuevas técnicas de procesamiento del lenguaje natural, como los sistemas de generación aumentada por recuperación (RAG), ha mejorado significativamente la capacidad de los modelos de inteligencia artificial para recuperar y generar información precisa en tareas complejas, lo que tiene un impacto positivo en la educación (Lewis et al., 2021). La integración de la recuperación de información en estos sistemas permite reducir errores fácticos y mejorar la fiabilidad de los contenidos generados, lo que representa una evolución importante en el uso de la inteligencia artificial en entornos educativos (Yu et al., 2024).

2. METODOLOGÍA O MATERIALES Y METODOS

Materiales

Para la realización de este estudio se emplearon diversos materiales que incluyen fuentes de producción científica relacionadas con las planificaciones de los sílabos en la educación superior. Estos documentos fueron convertidos de sus formatos originales (.pdf y .xlsx) a archivos de texto plano (.txt) con codificación UTF-8 utilizando las bibliotecas PyPDF2 v3.0 y Openpyxl v3.1.5 de Python (v. 3.11). Los

datos obtenidos de las planificaciones anteriores de sílabos, específicamente en lo relacionado con metodologías, recursos, actividades, y formas de evaluación, fueron fundamentales para los experimentos realizados. Además, se utilizó un marco personalizado de LLM (Large Language Model) basado en LLAMA 2 (lanzado por META en julio de 2023) (Touvron et al., 2023) y una serie de herramientas como Pandas AI para el análisis de datos.

Métodos

Selección de la producción científica

La producción científica fue seleccionada siguiendo criterios de relevancia y actualidad en relación con las planificaciones de sílabos en la educación superior. Para asegurar esta relevancia y actualidad, la búsqueda abarcó un lapso de los últimos 5 años, desde 2018 hasta 2022, y se tomaron en cuenta diversas formas de información, como artículos científicos y libros en formatos PDF, en los idiomas español e inglés. Se priorizaron estudios que incluyeran datos cuantitativos y cualitativos sobre metodologías, recursos, actividades y formas de evaluación aplicadas en contextos académicos similares. Las fuentes consultadas incluyeron plataformas como Scopus, Google Académico, Scielo, revistas en línea y sitios web de universidades, garantizando la incorporación de investigaciones recientes y relevantes. Estos documentos fueron posteriormente procesados y estructurados para ser integrados en el marco de análisis y experimentación del modelo LLAMA 2.

Para el procesamiento de las fuentes de producción científica, se realizaron los siguientes procesos “Conversión de PDF a Texto plano”, “Limpieza de datos”, “Estructuración y categorización”, optimizando así la entrada para el modelo LLAMA 2 y garantizando una mayor precisión en los resultados.

Conversión de PDF a Texto plano: Los documentos en formatos PDF de las fuentes de producción científica, se convirtieron a texto plano (formato .TXT). Se utilizó la biblioteca PyPDF2 para extraer el texto de los documentos PDF. El objetivo fue convertir el contenido de los PDFs en archivos de texto plano (.txt) para facilitar el procesamiento posterior.

Proceso de conversión de PDF a texto plano:

1. Importación de la biblioteca PyPDF2: Permite manipular archivos PDF en Python.

$$import\ PyPDF2$$

2. Apertura del archivo PDF en modo binario de lectura ('rb'): Garantiza una correcta lectura del contenido.

$$with\ open(pdf\_path, 'rb')\ as\ pdf\_file:$$

3. Creación de un objeto PdfReader: Utilizando PyPDF2 para acceder a las páginas del PDF.

$$pdf\_reader = PyPDF2.PdfReader(pdf\_file)$$

4. Extracción del texto de cada página: Iterando a través de cada página y almacenando el texto en una variable.

$$for\ page\ in\ pdf\_reader.pages:$$

$$text = page.extract\_text()$$

Limpieza de datos: Una vez en formato de texto plano, los documentos fueron sometidos a un proceso de limpieza. Este incluyó la eliminación de elementos no informativos como encabezados, números de página, y otros datos irrelevantes. También se removieron caracteres especiales o errores que podrían interferir con el análisis, asegurando que solo el contenido útil permaneciera en el conjunto de datos.

Proceso de limpieza de datos:

1. Importación de la biblioteca re: Para trabajar con expresiones regulares en Python.

$$import\ re$$

2. Eliminación de encabezados y pies de página comunes: Utilizando *re.sub()* para reemplazarlos por una cadena vacía.

$$text = re.sub(r'Encabezado\ común|Pie\ de\ página\ común', '', text)$$

3. Remoción de líneas que contienen solo dígitos: Reemplazándolas con una cadena vacía; la opción flags=re.MULTILINE permite que ^ y $ coincidan con el inicio y el final de cada línea.

$$text = re.sub(r'^\s*\d+\s*\$', '', text, flags = re.MULTILINE)$$

4. Eliminación de caracteres no deseados: Reemplazando todos los caracteres que no sean palabras o espacios.

$$text = re.sub(r'[^\w\s]', '', text)$$

5. Normalización de espacios en blanco: Sustituyendo secuencias de espacios o saltos de línea por un solo espacio.

$$text = re.sub(r'\s+', ' ', text)$$

Estructuración y categorización: Para facilitar el procesamiento por parte del modelo, los datos limpios fueron organizados en categorías temáticas. Se segmentaron en las siguientes secciones: "Metodologías", "Recursos", "Actividades" y "Formas de evaluación". Esta estructura permitió al modelo LLAMA 2 procesar el contenido de manera coherente y maximizar su precisión.

Proceso de estructuración y categorización:

1. Definición de las categorías temáticas:

$$categorie = ['Metodologías', 'Recursos', 'Actividades', 'Formas\ de\ evaluación']$$

2. Creación de un diccionario para almacenar los datos categorizados:

$$categorized\_data = \{category: '' \ for \ category \ in \ categories\}$$

3. Construcción del patrón de expresión regular: Para separar el texto en función de las categorías.

$$pattern = '|'.join([re.escape(category) \ for \ category \ in \ categories])$$

4. División del texto en secciones basadas en el patrón:

$$sections = re.split(r'({})'.format(pattern), text)$$

5. Inicialización de la variable de categoría actual:

$$current\_category = None$$

6. Iteración y asignación del texto a las categorías correspondientes:

```
for section in sections:
    if section in categories:
        current_category = section
    elif current_category:
        categorized_data[current_category] += section.strip() + ' '
```

Creación de prompts estandarizados

Los prompts se crearon con el objetivo de optimizar la precisión de las recomendaciones generadas por el modelo LLAMA 2. Se desarrollaron cinco configuraciones experimentales donde cada prompt fue diseñado mediante el método de ensayo y error, con el objetivo de guiar al modelo en la interpretación correcta de la información proporcionada. Esto incluyó la incorporación de información contextual sobre las metodologías, recursos, actividades, y formas de evaluación utilizadas en los periodos académicos anteriores, así como la implementación de estrategias de aprendizaje con pocos ejemplos (Kresevic et al., 2024).

Estudio de ablación: marco personalizado de LLM

Con la ayuda de una combinación de RAG (Retrieval-Augmented Generation) y con los datos de la planificaciones anteriores de los sílabos, específicamente en (Metodologías, Recursos, Actividades, Formas de Evaluar), se agregó información de fuentes de producción científica, en diferentes configuraciones experimentales con grados crecientes de complejidad en cuanto a reformateo de la información recolectada, arquitectura de prompts y aprendizaje con pocos ejemplos para crear un marco personalizado aplicado al modelo LLAMA 2 (lanzado por META en julio de 2023 con conocimientos actualizados hasta julio de 2023). Los experimentos con la Interfaz de Programación de Aplicaciones (API) de Ollama no pueden recuperar directamente información de archivos .pdf, .xlsx. Por lo tanto, los

documentos originales de la información en pdf y xlsx se convirtieron en archivos .txt con codificación UTF-8 utilizando las bibliotecas PyPDF2 v3.0, Openpyxl v3.1.5 de Python (v. 3.11).

Se realizó un estudio de ablación desde la base (Experimentos 1 al 5) para investigar cómo los diferentes ajustes en el reformateo de información, la arquitectura de prompts y el aprendizaje con pocos ejemplos impactan en la precisión y robustez de las salidas de los LLM (Figura 1).

Criterios experimentales

Base. Uso del LLAMA 2 básico sin ningún contexto. Para este experimento, solo se proporcionó el nombre de la materia o asignatura de la cual se requiera la recomendación sin ninguna instrucción adicional.

Experimento 1. Uso de LLAMA 2 básico con la información de producciones científicas sobre las planificaciones de los sílabos en la educación superior cargadas en contexto después de la conversión de pdf a texto en codificación UTF-8 sin ningún proceso de limpieza de texto adicional.

Experimento 2. Uso de LLAMA 2 con la información de producciones científicas sobre las planificaciones de los sílabos en la educación superior cargadas en contexto después de ser limpiadas manualmente con la eliminación de datos no informativos. Agregando información como: metodologías, recursos, actividades y formas de evaluar que se utilizaron en las planificaciones de los sílabos de los periodos académicos ya finalizados.

Experimento 3. Uso de LLAMA 2 básico con la información de producciones científicas sobre las planificaciones de los sílabos en la educación superior, agregando las metodologías, recursos, actividades y formas de evaluar utilizadas en los periodos académicos ya finalizados. Además, se empleó la biblioteca Pandas AI para el análisis de datos, lo cual nos permitió identificar los datos más significativos de las planificaciones de los sílabos con los periodos ya finalizados.

Experimento 4. Uso de LLAMA 2 básico con la información de producciones científicas sobre las planificaciones de los sílabos en la educación superior, incorporando metodologías, recursos, actividades y formas de evaluar, con su respectivo análisis de datos ejecutado mediante la biblioteca Pandas AI. Además, se proporcionó una serie de prompts (es decir, prompts estandarizados) que instruían al modelo sobre cómo interpretar la información enviada.

Experimento 5. Uso de LLAMA 2 básico con la información de producciones científicas sobre las planificaciones de los sílabos en la educación superior, incorporando metodologías, recursos, actividades y formas de evaluar, junto con su respectivo análisis de datos. Incluimos la serie de prompts (es decir, prompts estandarizados) y agregamos una colección de 50 planificaciones de sílabos de los periodos

académicos ya finalizados de diferentes materias o asignaturas (es decir, aprendizaje con pocos ejemplos).

Los experimentos, resumidos en la (Figura 1), se realizaron en un entorno local de Python con acceso a LLAMA 2. Se utilizó un modelo base de 7B de parámetros, seleccionando una temperatura de 0.9 y estableciendo un número máximo de 15,000 tokens en la salida.

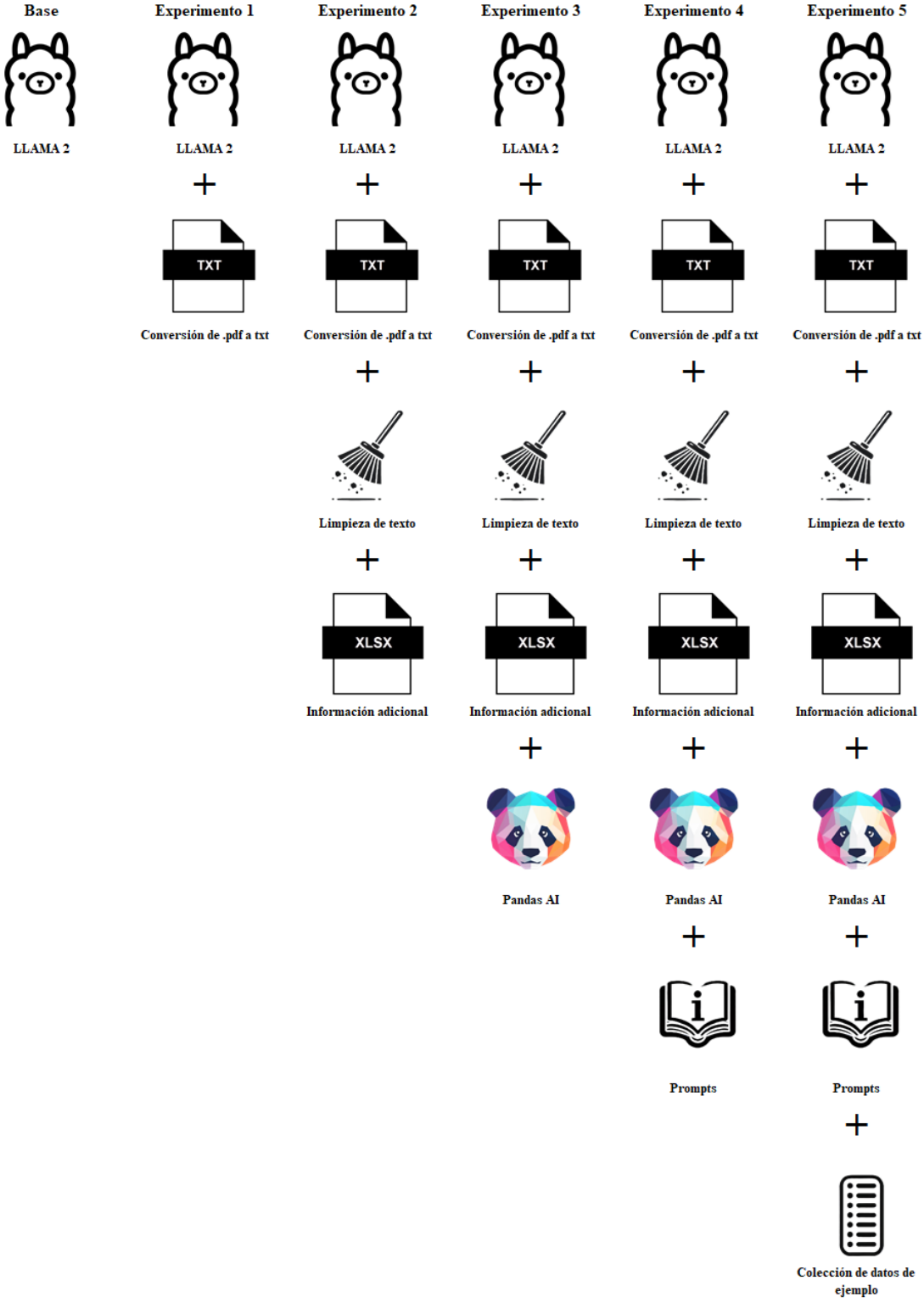
Criterios para la selección de las materias para los experimentos

La selección de las materias para los experimentos fue realizada bajo un enfoque de aleatoriedad, garantizando una muestra representativa y libre de sesgos que abarque diversas disciplinas académicas. Este criterio de selección aleatoria fue esencial para asegurar que el análisis y las recomendaciones del modelo LLAMA 2 fueran aplicables a un amplio espectro de contextos educativos, reflejando una diversidad de metodologías, recursos, actividades y formas de evaluación. Al no restringir la elección a un conjunto particular de materias, se evitó la introducción de sesgos y se promovió una evaluación más robusta de la capacidad del modelo para generar recomendaciones precisas y efectivas en distintos ámbitos del conocimiento.

Tabla 1. Materias seleccionadas para realizar los experimentos.

MATERIAS
Sistemas de Información Geográfica
Informática Aplicada
Formulación de Proyectos Agrícolas
Química Inorgánica y Orgánica
Física
Microbiología
Epistemología del Turismo
Comunicación y Lenguaje
Inventario de Recursos Turísticos
Marketing Estratégico
Fundamentos de Contabilidad
Matemática Financiera
Economía
Identidad Cultural
Desarrollo Organizacional
Filosofía de la Formación Profesional
Administración y Economía de Empresas
Campos Vectoriales
Metodología de la Investigación Científica
Calculo Diferencial e Integral

Figura 1. Evaluación cualitativa de la precisión entre todos los experimentos.



Procedimiento para la comparación de los resultados del modelo LLAMA 2 con las metodologías, recursos y formas de evaluación más utilizadas en la planificación de sílabos ya finalizados.

Para comparar los resultados, se realizaron cinco consultas de recomendaciones para cada una de las 20 materias seleccionadas (Tabla 1), en cada configuración experimental. Posteriormente, se evaluó la precisión de las recomendaciones generadas. Cada recomendación recibió una puntuación de 1 si el texto contenía información completamente precisa con la información obtenida “Metodologías, Recursos y

Formas de Evaluación" más utilizadas en la planificación de sílabos ya finalizados (Tabla 2), o 0 en caso contrario. Las recomendaciones consideradas inexactas se atribuyeron a "alucinaciones", es decir, la generación de información que, aunque plausible, podría ser incorrecta o no verificada.

Tabla 2. Metodologías, Recursos y Formas de Evaluación más utilizadas en la planificación de sílabos ya finalizados de las materias seleccionadas para los experimentos.

SISTEMAS DE INFORMACION GEOGRAFICA		
Actividad	Proponer Problemas reales	19
Forma de Evaluar	Proyecto	18
Metodología	Aprendizaje basado en proyectos	30
Recurso	Artículos	18
INFORMATICA APLICADA		
Actividad	Utilizar estas nuevas tecnologías se requiere cierto nivel de empatía con ellas	30
Forma de Evaluar	Reactivo	27
Metodología	Nuevas Tecnologías de la Información y de la Comunicación (TICs)	48
Recurso	Materiales audiovisuales: películas, vídeos, diapositivas, proyectores... Estos materiales	30
FORMULACION DE PROYECTOS AGRICOLAS		
Actividad	Investigar en territorio	18
Forma de Evaluar	Proyecto	27
Metodología	Aprendizaje basado en proyectos	27
Recurso	Artículos	27
QUIMICA INORGANICA Y ORGANICA		
Actividad	Analizar y comprender problemas, ejercicios o actividades	33
Forma de Evaluar	Ejercicios Prácticos	18
Metodología	Resolución de Ejercicios y Problemas	39
Recurso	Aula física o virtual	36
FISICA		
Actividad	Analizar y comprender problemas, ejercicios o actividades	117
Forma de Evaluar	Ejercicios teóricos y prácticos	93
Metodología	Resolución de Ejercicios y Problemas	156
Recurso	Aula física o virtual	132
MICROBIOLOGIA		
Actividad	Práctica de Laboratorio	54
Forma de Evaluar	Informe de la práctica	54
Metodología	Método basado en la demostración práctica	53
Recurso	Investigación en laboratorio	54
EPISTEMOLOGIA DEL TURISMO		
Actividad	Trabajar en grupos	9
Forma de Evaluar	Exposición grupal	9
Metodología	Metodología Aprendizaje cooperativo	11

Recurso	Recursos Virtuales	9
COMUNICACION Y LENGUAJE		
Actividad	Trabajar en grupos	21
Forma de Evaluar	Informe	24
Metodología	Aprendizaje cooperativo	42
Recurso	Materiales audiovisuales: películas, vídeos, diapositivas, proyectores... Estos materiales	21
INVENTARIO DE RECURSOS TURISTICOS		
Actividad	Exposiciones, Lecciones, talleres, ejercicios prácticos.	6
Forma de Evaluar	Reporte en equipo	6
Metodología	Aprendizaje cooperativo	9
Recurso	Conferencia	9
MARKETING ESTRATEGICO		
Actividad	Proponer conclusiones	15
Forma de Evaluar	Cuestionario	18
Metodología	Método expositivo	18
Recurso	Recursos Virtuales	12
FUNDAMENTOS DE CONTABILIDAD		
Actividad	Analizar y comprender problemas, ejercicios o actividades	21
Forma de Evaluar	Ejercicios Prácticos	21
Metodología	Resolución de Ejercicios y Problemas	39
Recurso	Aula física o virtual	18
MATEMATICA FINANCIERA		
Actividad	Repasar actividades realizadas	24
Forma de Evaluar	Reactivos	26
Metodología	Resolución de Ejercicios y Problemas	62
Recurso	Aula física o virtual	57
ECONOMIA		
Actividad	Trabajar en grupos	45
Forma de Evaluar	Ejercicios Prácticos	42
Metodología	Metodología Aprendizaje cooperativo	63
Recurso	Recursos Virtuales - Google	39
IDENTIDAD CULTURAL		
Actividad	Trabajar en grupos	87
Forma de Evaluar	Informe	76
Metodología	Aprendizaje cooperativo	158
Recurso	Materiales audiovisuales: películas, vídeos, diapositivas, proyectores... Estos materiales	71
DESARROLLO ORGANIZACIONAL		
Actividad	Asignar las funciones y roles a cada equipo	15
Forma de Evaluar	Ensayo	18
Metodología	Aprendizaje cooperativo	33
Recurso	Artículos científicos	15
FILOSOFIA DE LA FORMACION PROFESIONAL		
Actividad	Asignar las funciones y roles a cada equipo	12

Forma de Evaluar	Informe	12
Metodología	Aprendizaje basado en proyectos	13
Recurso	Diagnóstico	9
ADMINISTRACION Y ECONOMIA DE EMPRESAS		
Actividad	Trabajar en grupos	12
Forma de Evaluar	Reactivo	9
Metodología	Metodología Aprendizaje cooperativo	12
Recurso	Conferencia	9
CAMPOS VECTORIALES		
Actividad	Analizar y comprender problemas, ejercicios o actividades	15
Forma de Evaluar	Proyecto	9
Metodología	Resolución de Ejercicios y Problemas	18
Recurso	Aula física o virtual	18
METODOLOGIA DE LA INVESTIGACION CIENTIFICA		
Actividad	Elaboración de reporte escrito	9
Forma de Evaluar	Proyecto	9
Metodología	Aprendizaje basado en proyectos	15
Recurso	Artículos	13
CALCULO DIFERENCIAL E INTEGRAL		
Actividad	Analizar y comprender problemas, ejercicios o actividades	99
Forma de Evaluar	Ejercicios Prácticos	57
Metodología	Resolución de Ejercicios y Problemas	117
Recurso	Aula física o virtual	117

Pasos para calcular la presión y el valor de *p* a partir de los resultados obtenidos con el modelo LLAMA 2

Para calcular la precisión y el valor de *p* de los resultados de las cinco consultas realizadas al modelo LLAMA 2, para cada una de las veinte materias, se realizaron los siguientes cálculos en cada experimento.

1. **Cálculo de la Precisión (Porcentajes):**

La precisión en cada experimento se calcula utilizando la siguiente fórmula

$$\text{Precisión (\%)} = \left(\frac{Número\ de\ recomendaciones\ correctas}{Número\ total\ de\ consultas}\right) \text{x } 100$$

2. **Cálculo del valor de p (Prueba Chi-Cuadrado):**

El valor de p se calcula para determinar si la diferencia en la precisión entre dos grupos (por ejemplo, el experimento y la configuración base) es estadísticamente significativa. Esto se hace utilizando la prueba Chi-Cuadrado para la comparación de proporciones.

Paso 1: Construir una tabla de contingencia

Comparación de la precisión de un experimento con precisión de la configuración base. La tabla de contingencia tendría la siguiente forma:

Tabla 3. Contingencia para el cálculo de la precisión y la prueba Chi-Cuadrado

	Correcto (1)	**Incorrecto (0)**	**Total**
Experimento	A	B	A+B
Base	C	D	C+D
Total	A+C	B+D	N

Donde:

- A es el número de respuestas correctas en el experimento.
- B es el número de respuestas incorrectas en el experimento.
- C es el número de respuestas correctas en la configuración base.
- D es el número de respuestas incorrectas en la configuración base.
- N es el número total de respuestas.

Paso 2: Calcular el estadístico Chi-Cuadrado

El estadístico Chi-Cuadrado se calcula con la siguiente fórmula:

$$x^2 = \sum \frac{(O_i + E_i)^2}{E_i}$$

Donde:

O_i es el valor observado (A, B, C, D).

E_i es el valor esperado bajo la hipótesis nula, calculado como:

$$E_{ij} = \frac{Total\ Fila\ i\ x\ Total\ Columna\ j}{Total\ general}$$

Paso 3: Determinar el valor de p

El valor de p se obtiene comparando el valor de X^2 con una distribución Chi-Cuadrado con los grados de libertad adecuados, que es $(n-1)x(m-1)$, donde son el número de filas y columnas de la tabla de contingencia (en este caso 2-1 = 1).

Si el valor de p es menor que el nivel de significancia (típicamente 0.05), se rechaza la hipótesis nula, lo que indica que la diferencia en las proporciones es estadísticamente significativa.

Procedimiento para evaluar la similitud entre respuestas generadas por el LLM y la Información de planificaciones de sílabo finalizadas

Para evaluar la similitud entre las respuestas generadas por el LLM y la información recolectada de las planificaciones de los sílabos finalizados, se emplearon las siguientes métricas: Evaluación de Subgisting Orientada al Recuerdo (ROUGE) (Lin, 2004), Evaluación Bilingüe de Subgisting (BLEU) (Papineni et al., 2002), Métrica para la Evaluación de Traducción con Ordenación Explícita (METEOR) (Banerjee & Lavie, 2005) y una Puntuación Personalizada.

La puntuación personalizada de LLAMA 2 se basa en la similitud de coseno, mientras que las demás métricas se fundamentan en la superposición de palabras y la coherencia semántica entre las dos fuentes de texto. La similitud se evaluó comparando las respuestas generadas por el LLM con la información de las planificaciones de los sílabos. Todas las puntuaciones se expresan en una escala de 0 a 1, donde una puntuación de 1 indica una alineación perfecta entre las dos fuentes de texto comparadas. Se calcularon la media y la desviación estándar de las similitudes tras repetir la consulta 5 veces para cada una de las 20 materias (Tabla 1).

3. RESULTADOS Y DISCUSIÓN

Resultados:

Resultado del procedimiento para la comparación de los resultados del modelo LLAMA 2 con las metodologías, recursos y formas de evaluación más utilizadas en la planificación de sílabos ya finalizados.

- **Resultados obtenidos tras la aplicación de los experimentos en todas las materias seleccionadas.**

El marco personalizado de LLM alcanzó una precisión general del 92.5%, superando significativamente al modelo LLAMA 2 por sí solo (92.5% frente a 40.0%; $p < 0.001$). La incorporación de documentos relacionados con las planificaciones de los sílabos en el contexto mejoró la precisión (57.5% frente a 40.0%; $p = 0.023$). Cuando estos documentos fueron limpiados y se incluyeron las metodologías, recursos, actividades y formas de evaluación obtenidas de la base de datos del sistema de planificación de sílabos, la precisión aumentó al 67.5% (frente a 40.0%; $p < 0.001$). Posteriormente, al formatear los documentos con una estructura coherente junto con las metodologías, recursos, actividades y formas de evaluación, y aplicar la biblioteca Pandas AI para el análisis de datos, la precisión mejoró aún más, alcanzando un 77.5% (frente a 40.0%; $p < 0.001$). La adición de prompts estandarizados resultó en una mejora adicional, con una precisión del 87.5% (frente a 40.0%; $p < 0.001$). Finalmente, al incorporar el aprendizaje con pocos ejemplos, utilizando una colección de 50 planificaciones de sílabos de periodos

académicos ya finalizados, se logró alcanzar nuevamente una precisión del 92.5% (frente a 40.0%; p < 0.001) (Tabla 4) (Tabla 6) (Figura 2).

Figura 2. Estadísticas de los resultados de todos los experimentos aplicados en todas las materias seleccionadas.

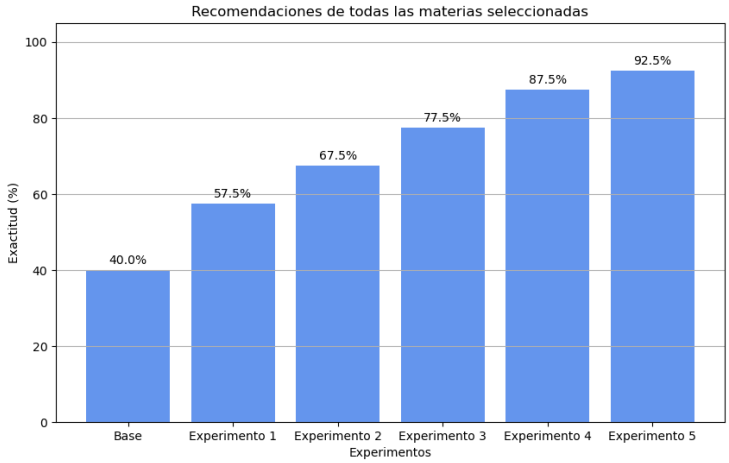


Tabla 4. Resultados del valor de ***p*** para cada uno de los experimentos, correspondiente a todas las materias seleccionadas.

Experimentos	**Valor de *p***
Base	1.0
Experimento 1	0.023588841712258
Experimento 2	0.000227808993789
Experimento 3	2.38683838e-07
Experimento 4	4.398e-12
Experimento 5	7e-15

Interpretación de los valores de p:

Base: El valor de p = 1.0 al comparar la configuración base consigo misma indica que no hay diferencia estadísticamente significativa en la precisión (p > 0.05). Esto es consistente con la hipótesis nula H_0:μ_{base} = μ_{base} donde μ_{base} es la precisión media de la configuración base. Sirve como control para validar el método estadístico empleado y establecer un punto de referencia para los experimentos subsecuentes.

Experimento 1: Un valor de p = 0.023588841712258 sugiere una diferencia estadísticamente significativa entre la precisión media de la configuración base (μ_{base}) y la del Experimento 1 (μ_1), con un nivel de significancia del 5% (p < 0.05). La hipótesis nula H_0: $\mu_{base} = \mu_1$ es rechazada en favor de la hipótesis alternativa H_1: $\mu_{base} \neq \mu_1$. Aunque el valor de p es cercano a 0.05, indicando un nivel moderado de significancia, se evidencia que agregar el contexto inicial mejora la precisión del modelo.

Experimento 2: Con un valor de p = 0.000227808993789, significativamente menor que 0.001 (p < 0.001), se observa una diferencia altamente significativa entre μ_{base} y μ_2 (precisión media del Experimento 2). La hipótesis nula es rechazada con mayor confianza, indicando que la limpieza de datos y la eliminación de información no informativa tienen un efecto positivo sustancial en la precisión.

Experimento 3: Un valor de p aproximado a cero (p ≈ 0) indica una diferencia extremadamente significativa entre μ_{base} y μ_3 . La probabilidad de que esta diferencia sea producto del azar es prácticamente nula. La incorporación de metodologías avanzadas, recursos adicionales y análisis de datos con Pandas AI mejoró notablemente la precisión, rechazando contundentemente la hipótesis nula en favor de H_1: $\mu_{base} \neq \mu_3$.

Experimento 4: El valor de p = 4.398e-12 refuerza la existencia de una diferencia estadísticamente significativa entre μ_{base} y μ_4.Este resultado, con un nivel de significancia mucho menor que 0.0001 (p < 0.0001), subraya la eficacia de los prompts estandarizados en la mejora de la interpretación del modelo. La hipótesis nula es rechazada con un alto grado de confianza estadística.

Experimento 5: Presenta el valor de p más bajo registrado, indicando la diferencia más significativa en precisión respecto a la base. La inclusión del aprendizaje con pocos ejemplos (few-shot learning) ha incrementado significativamente la precisión hasta μ_5.La hipótesis nula H_0: $\mu_{base} = \mu_5$ es rechazada a favor de la alternativa, con un nivel de significancia estadística extremadamente alto. Esto demuestra la relevancia de esta técnica para mejorar el rendimiento del modelo LLAMA 2.

- **Resultados obtenidos tras la aplicación de los experimentos en una materia en específico**

El marco personalizado de LLM alcanzó una precisión general del 95.4%, superando significativamente al modelo LLAMA 2 por sí solo (95.4% frente a 31.4%; p < 0.001). La incorporación de documentos sobre las planificaciones de los sílabos en el contexto mejoró la precisión (62.0% frente a 31.4%; p < 0.001). Después de limpiar el texto de los documentos y añadir las metodologías, recursos, actividades y formas de evaluación de las planificaciones de sílabos de periodos académicos finalizados, la precisión aumentó al 78.7% (frente a 31.4%; p < 0.001). Se logró una mejora sustancial al emplear la biblioteca Pandas AI para el análisis de datos, alcanzando un 86.8% de precisión (frente a 31.4%; p < 0.001). La adición de prompts estandarizados elevó la precisión al 93.6% (frente a 31.4%; p < 0.001). Finalmente, con el aprendizaje de pocos ejemplos, utilizando una colección de 50 planificaciones de sílabos de periodos académicos finalizados, la precisión alcanzó nuevamente el 95.4% (frente a 31.4%; p < 0.001) (Tabla 5) (Tabla 6) (Figura 3).

Figura 3. Estadísticas de los resultados de todos los experimentos aplicado a una materia en específico.

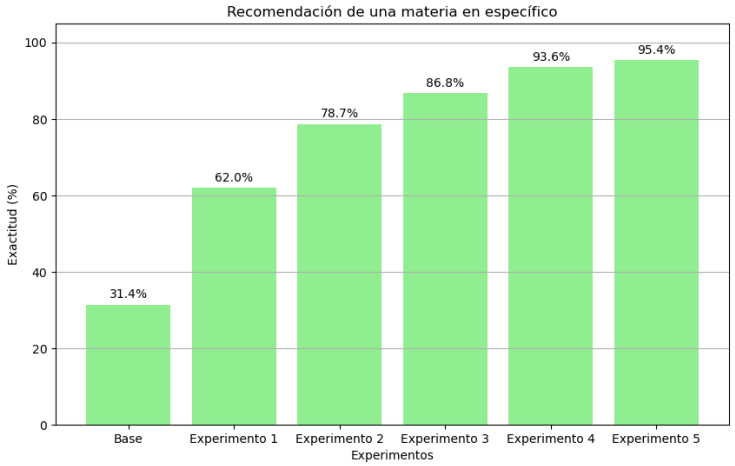


Tabla 5. Resultados del valor de ***p*** para cada uno de los experimentos, correspondiente a una materia en específico.

Experimentos	**Valor de *p***
Base	1.0
Experimento 1	2.726076965838658e-05
Experimento 2	4.658395728e-11
Experimento 3	5.11733e-15
Experimento 4	3.9e-19
Experimento 5	2e-20

Interpretación de los valores de p:

Base: El valor de p = 1.0 al comparar la configuración base consigo misma confirma que no hay diferencia estadísticamente significativa en la precisión ($p > 0.05$). Esto valida la hipótesis nula H_0:$\mu_{base} = \mu_{base}$ donde μ_{base} es la precisión media de la configuración base. Sirve como control para los experimentos posteriores y establece una línea de base para las comparaciones.

Experimento 1: Un valor de p = 2.726076965838658e-05 indica una diferencia estadísticamente significativa entre la precisión media de la configuración base μ_{base} y la del Experimento 1 (μ_1). Dado que $p < 0.05$, se rechaza la hipótesis nula $H_0: \mu_{base} = \mu_1$ en favor de la hipótesis alternativa $H_1: \mu_{base} \neq \mu_1$. La adición de contexto básico al modelo ha resultado en una mejora significativa en la precisión para esta materia específica.

Experimento 2: Con un valor de p = 4.658395728e-11, mucho menor que 0.001 ($p < 0.001$), se evidencia una diferencia altamente significativa entre μ_{base} y μ_2 (precisión media del Experimento 2). La hipótesis nula es rechazada con alta confianza estadística, indicando que la limpieza y estructuración de datos específicos han mejorado considerablemente la precisión en esta materia.

Experimento 3: El valor de p = 5.11733e-15 demuestra una diferencia extremadamente significativa entreμ_{base} y μ_3 . La probabilidad de que esta diferencia sea producto del azar es prácticamente nula. La introducción de análisis de datos mediante Pandas AI y la incorporación de recursos adicionales han resultado especialmente efectivas en mejorar el rendimiento del modelo. La hipótesis nula es rechazada contundentemente en favor de $H_1: \mu_{base} \neq \mu_3$.

Experimento 4: Con un valor de p = 3.9e-19, se refuerza la existencia de una diferencia estadísticamente significativa entre μ_{base} y μ_4. Este valor, extremadamente cercano a cero, subraya la eficacia de los prompts estandarizados en este contexto, mejorando sustancialmente la precisión en la generación de recomendaciones para esta materia. La hipótesis nula es rechazada con un nivel de significancia estadística muy alto.

Experimento 5: Presenta el valor de p más bajo de la tabla, p = 2e-20, indicando la mejora más significativa respecto a la configuración base. La incorporación de ejemplos específicos de la materia bajo un enfoque de aprendizaje con pocos ejemplos (few-shot learning) ha optimizado el rendimiento del modelo de manera notable. La hipótesis nula $H_0: \mu_{base} = \mu_5$ es rechazada a favor de la hipótesis alternativa con un nivel de significancia estadística extremadamente alto.

Figura 4. Comparación del valor de p en distintos experimentos: Todas las materias vs. Una materia específica.

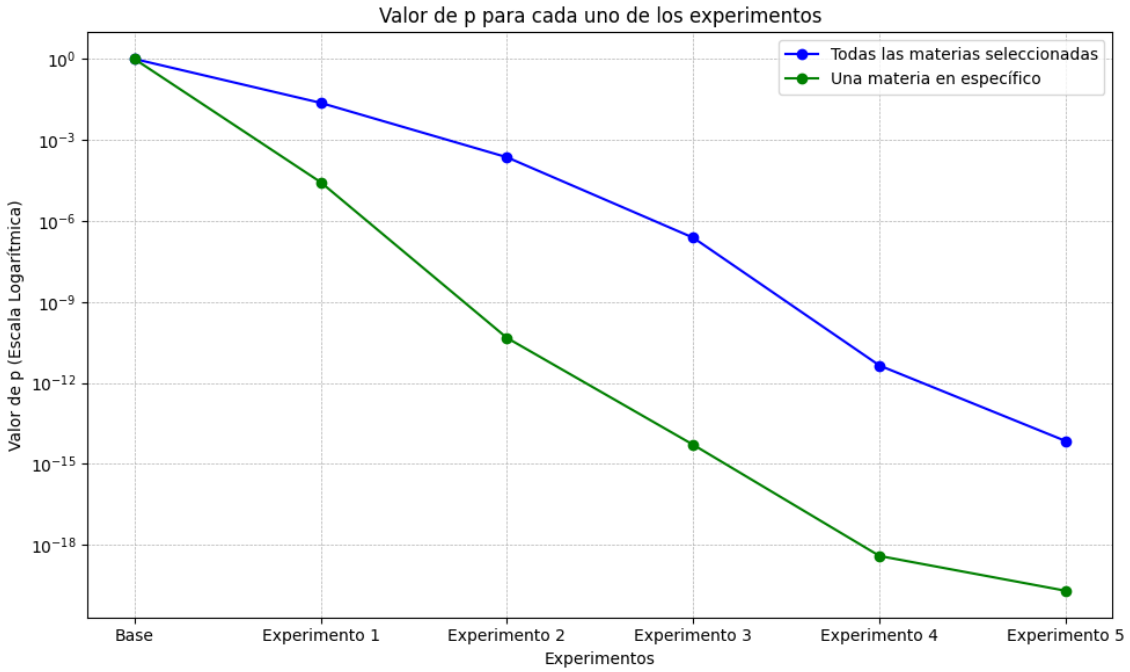


Tabla 6. Evaluación cualitativa de la precisión basada en la comparación de los resultados obtenidos de los experimentos con las metodologías, recursos, formas de evaluar más utilizadas.

Métrica	Base	Experimento 1	Experimento 2	Experimento 3	Experimento 4	Experimento 5
Todas las materias seleccionadas:						
Exactitud	40.0%	57.5%	67.5%	77.5%	87.5%	92.5%
Significancia estadística		$p = 0.023$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
Una materia en específico:						
Exactitud	31.4%	62.0%	78.7%	86.8%	93.6%	95.4%
Significancia estadística		$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Cuando se revisaron las salidas inexactas para detectar alucinaciones, encontramos 113 (89,68 %) alucinaciones en conflicto con hechos (FCH) y 13 (10,32%) alucinaciones en conflicto con entradas (ICH) en todos los experimentos. El tipo y la distribución de las alucinaciones en cada experimento se informan en la (Tabla 7). No se encontró alucinaciones en conflicto con el contexto (CCH) en ninguno de los experimentos.

Cuando los resultados se consideran inexactos, la inexactitud se debe a alucinaciones (es decir, la producción de sonidos plausibles pero potencialmente información no verificada o incorrecta) (Giuffrè et al., 2023; Giuffrè & Shung, 2024). Según las definiciones de (Zhang et al., 2023), definimos tres tipos de alucinaciones: FCH, ICH y CCH.

Tabla 7. Tipo de alucinaciones y distribución en todos los experimentos generados por LLM.

Alucinación	Total	Fact- conflicting	Input- conflicting	Contextual-conflicting
Base	59	50 (84.7%)	9 (15.3%)	-
Experimento 1	32	28 (87.5%)	4 (12.5%)	-
Experimento 2	20	20 (100%)	-	-
Experimento 3	12	12 (100%)	-	-
Experimento 4	2	2 (100%)	-	-
Experimento 5	1	1 (100%)	-	-

Resultado del procedimiento para evaluar la similitud de texto, de las respuestas generadas por el LLM (Large Language Model) con la información recolectada de las planificaciones de los sílabos.

Se encontró diferencias en el marco personalizado de LLM en comparación con el modelo básico en las puntuaciones de similitud (BLEU, ROUGE-LCS F1, METEOR F1 y la Puntuación Personalizada de LLAMA 2) para las asignaturas que fueron seleccionadas para realizar los experimentos (Tabla 8).

Tabla 8. Evaluación de la similitud de texto a texto entre los resultados generados por LLM.

Métrica	Base	Experimento 1	Experimento 2	Experimento 3	Experimento 4	Experimento 5
BLEU:						
Media (± SD)	0.028 (±0.024)	0.097 (±0.089)	0.110 (±0.142)	0.100 (±0.093)	0.139 (±0.117)	0.122 (±0.071)
Significancia	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
ROUGE-LCS F1:						
Media (± SD)	0.203 (±0.054)	0.335 (±0.122)	0.348 (±0.139)	0.337 (±0.115)	0.346 (±0.120)	0.360 (±0.096)
Significancia	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
METEOR F1:						
Media (± SD)	0.312 (±0.060)	0.420 (±0.103)	0.432 (±0.125)	0.411 (±0.100)	0.431 (±0.113)	0.424 (±0.079)
Significancia	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Puntaje personalizado para el modelo LLAMA 2:						
Media (± SD)	0.941 (±0.017)	0.955 (±0.018)	0.957 (±0.019)	0.957 (±0.017)	0.958 (±0.014)	0.959 (±0.018)
Significancia	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001

3.1 Discusión:

La incorporación de modelos de lenguaje de gran escala (LLMs) en la planificación de sílabos para la educación superior tiene el potencial de revolucionar la manera en que se diseñan y actualizan los planes de estudio. Al aprovechar las capacidades del procesamiento del lenguaje natural, es posible recomendar metodologías, actividades y recursos pedagógicos más efectivos. Estos modelos permiten a los docentes generar sílabos personalizados y alineados con las mejores prácticas educativas, lo que incrementa la calidad del proceso de enseñanza-aprendizaje. Además, la habilidad de los LLMs para analizar grandes volúmenes de datos educativos históricos y contextuales los convierte en una herramienta invaluable para adaptar la planificación a las necesidades y características específicas de los estudiantes y los cursos.

Los procedimientos y experimentos llevados a cabo confirman la efectividad del enfoque basado en LLAMA 2 para optimizar la planificación de sílabos en la Universidad Técnica de Cotopaxi. A lo largo de cinco experimentos, la precisión en la generación de recomendaciones mejoró sustancialmente: se inició con un 40% en el experimento base, donde solo se proporcionaba el nombre de la asignatura, y se alcanzó un 92.5% en el experimento final, que incorporaba aprendizaje con pocos ejemplos y 50 planificaciones previas de sílabos. Este incremento progresivo en la precisión resalta la importancia de proporcionar al modelo un contexto rico, datos depurados y una serie de prompts bien diseñados para guiar la interpretación de la información. Adicionalmente, el uso de herramientas como Pandas AI

permitió un análisis más detallado de los datos, mejorando la identificación de patrones relevantes en las metodologías, recursos y actividades de los sílabos. Estos hallazgos son consistentes con los observados en estudios previos con GPT-4, donde la incorporación de información contextual también resultó en mejoras significativas en la precisión de las respuestas generadas. La capacidad de LLAMA 2 para aprender y adaptarse a las necesidades específicas del contexto educativo, al igual que GPT-4 en el ámbito clínico, demuestra su versatilidad para diversas aplicaciones de generación de lenguaje natural.

Las puntuaciones de similitud, medidas a través de métricas como BLEU, ROUGE-L, METEOR y una puntuación personalizada, evidenciaron mejoras significativas entre la salida generada por LLAMA 2 y las recomendaciones educativas esperadas a lo largo de los experimentos. En particular, la precisión aumentó notablemente al agregar más contexto y aplicar ingeniería de prompts, alcanzando valores cercanos a 1, lo que indica una alta concordancia con los contenidos de las planificaciones de sílabos ya finalizadas. Esto demuestra que la capacidad de LLAMA 2 para generar recomendaciones altamente alineadas con las necesidades educativas se potencia cuando se combina con datos estructurados y ejemplos reales, consolidándolo como una herramienta poderosa para la optimización de sílabos.

En un estudio reciente, (Kresevic et al., 2024) desarrolló un marco basado en GPT-4 Turbo para interpretar guías clínicas relacionadas con la hepatitis C, empleando RAG y las recomendaciones de la Asociación Americana para el Estudio del Hígado (AASLD). Aunque el estudio mostró avances significativos en la generación de respuestas, también reveló limitaciones importantes en la capacidad del modelo para ofrecer respuestas completamente precisas, especialmente en escenarios clínicos complejos. Además, según la metodología descrita, no se especifica claramente cómo se convirtieron las guías en texto ni las estrategias empleadas para la fragmentación de la información, aspectos críticos que no se implementaron en el enfoque de la presente investigación. Tampoco se proporcionaron detalles exhaustivos sobre las tasas de precisión de las salidas. Por ende, aunque ambos estudios comparten objetivos similares, las diferencias metodológicas impiden una comparación directa de los hallazgos.

La optimización de la planificación del sílabo en la Universidad Técnica de Cotopaxi utilizando el modelo LLAMA 2 destacó por la precisión de las recomendaciones educativas, lograda mediante la integración de ingeniería de prompts y técnicas de aprendizaje con pocos ejemplos. Para contextualizar mejor estos resultados, es pertinente comparar este enfoque con un estudio previo que aplicó una metodología similar, pero con el modelo GPT-4 en un área distinta: la interpretación de guías clínicas para la hepatitis C. Aunque ambos estudios emplearon metodologías afines, sus diferencias radican en los dominios de aplicación y los modelos de lenguaje utilizados.

LLAMA 2, utilizado en este trabajo, ha demostrado eficacia en tareas de generación de contenido para la planificación educativa. Por otro lado, GPT-4, empleado en el estudio sobre guías clínicas, mostró una mayor robustez en contextos médicos complejos. Ambos modelos se beneficiaron de técnicas como

Retrieval-Augmented Generation (RAG) y la ingeniería de prompts, que mejoraron la precisión de las salidas en ambos casos, aunque la naturaleza de los dominios influyó en la magnitud de estos beneficios.

En el contexto educativo, LLAMA 2 alcanzó una precisión del 92.5% en la recomendación de sílabos, mejorando significativamente desde el rendimiento base del 40%. De manera similar, GPT-4 mejoró del 43% al 99% al ser empleado para interpretar guías clínicas, lo que demuestra que ambos modelos son sensibles al contexto cuando se les proporciona información relevante y estructurada. Sin embargo, mientras GPT-4 evidenció su eficacia en la interpretación médica, donde el análisis de texto técnico es esencial, LLAMA 2 fue más exitoso en un entorno educativo, donde la generación de recomendaciones personalizadas y precisas es crítica.

Un hallazgo clave en ambos estudios es la limitación de los modelos para procesar datos no textuales. En el caso de GPT-4, la conversión de tablas e imágenes a texto mejoró significativamente la precisión, pasando de un 28% en preguntas basadas en tablas a un 96% tras la conversión. De manera análoga, LLAMA 2 enfrentó desafíos al procesar las planificaciones de sílabos que contenían información tabular, los cuales se resolvieron mediante la integración de Pandas AI para el análisis de datos, conduciendo a un incremento notable en la precisión.

Ambos estudios exploraron el impacto del aprendizaje con pocos ejemplos. En el caso de GPT-4, se observó que esta técnica no aportó mejoras adicionales después de aplicar ingeniería de prompts. Por el contrario, en el presente estudio, la inclusión de ejemplos de 50 planificaciones previas de sílabos contribuyó a un aumento significativo en la precisión final de LLAMA 2, resaltando cómo la especificidad del dominio educativo favorece esta técnica.

En resumen, los dos estudios demuestran que el uso de modelos de lenguaje de gran escala en diferentes contextos puede mejorar significativamente los procesos de toma de decisiones, ya sea en la planificación de sílabos o en la interpretación de guías clínicas. La incorporación de contexto estructurado, la depuración de datos y el uso de herramientas de análisis resultaron ser factores decisivos en ambos enfoques. Aunque GPT-4 se destacó en el dominio médico, LLAMA 2 demostró ser una herramienta poderosa para la planificación educativa, abriendo oportunidades para futuras investigaciones en la personalización de planes de estudio.

La investigación de (Burgos et al., 2024) evidenció que la aplicación de ChatGPT en la planificación micro curricular puede mejorar significativamente el rendimiento académico. El estudio se desarrolló en dos grupos, experimental y control; el grupo experimental mostró un incremento en la media de calificaciones de **7.78 a 8.50**, en contraste con el grupo de control, que no presentó avances significativos. Además, el estudio destaca la percepción positiva de los docentes, quienes valoraron la precisión y la facilidad de uso de ChatGPT como elementos clave para adaptar eficazmente los planes de clase a las necesidades específicas del aula. Estos hallazgos son congruentes con los resultados obtenidos en el

presente artículo, donde la implementación de **LLAMA 2** en la planificación curricular también demostró su capacidad para mejorar la precisión y relevancia de las recomendaciones educativas, mediante un proceso de pre procesamiento y ajustes experimentales que optimizaron cada configuración del modelo.

Mientras que ChatGPT sobresalió por su aceptación y facilidad de integración en el entorno docente, el estudio sobre LLAMA 2 reveló que una metodología de estructuración de datos y el uso de *prompts* específicos permitieron alcanzar niveles de precisión estadísticamente significativos, con valores de **p** cercanos a cero en experimentos avanzados. Esto sugiere que ambos modelos de inteligencia artificial, aunque con enfoques metodológicos distintos, ofrecen ventajas complementarias en el contexto educativo. LLAMA 2 se posiciona como una herramienta eficaz para contextos donde la precisión de las recomendaciones es crucial, mientras que ChatGPT facilita un uso amplio y accesible en la planificación general del currículo. En conjunto, estos estudios resaltan el potencial de las herramientas de IA para mejorar las prácticas docentes, sugiriendo que una combinación estratégica de ambas tecnologías podría optimizar tanto la eficiencia en la planificación como la personalización educativa, adaptándose a diversos contextos y objetivos académicos.

4. CONCLUSIONES

La integración de modelos de lenguaje como LLAMA 2 ha demostrado ser altamente eficaz para optimizar la planificación de sílabos en la Universidad Técnica de Cotopaxi. Este estudio destaca cómo el uso de tecnologías avanzadas en inteligencia artificial puede mejorar significativamente las recomendaciones educativas, ofreciendo soluciones personalizadas que se adaptan a las necesidades tanto de estudiantes como de docentes.

Los experimentos realizados evidenciaron que la inclusión de documentos de planificación previos y los prompts estandarizados incrementó considerablemente la precisión de las recomendaciones generadas. Este hallazgo resalta la relevancia de proporcionar un contexto rico y bien estructurado para maximizar la efectividad de los modelos de IA en el ámbito educativo.

La aplicación de LLAMA 2 junto con técnicas avanzadas como la Recuperación Aumentada por Generación (RAG) ha permitido generar sílabos más alineados con las mejores prácticas pedagógicas. Este enfoque no solo mejora la calidad del proceso de enseñanza-aprendizaje, sino que también fortalece la personalización en la educación superior, enriqueciendo la experiencia educativa para estudiantes y docentes.

Eficacia de la ingeniería de prompts y el aprendizaje con pocos ejemplos: A través de cinco experimentos progresivos, el estudio demostró un aumento sustancial en la precisión de las recomendaciones, alcanzando hasta un 92.5%. Este progreso indica que la combinación de ingeniería de prompts, el análisis

de datos mediante Pandas AI y el aprendizaje con pocos ejemplos es esencial para mejorar el rendimiento de los modelos de IA en entornos educativos.

Este estudio abre nuevas perspectivas para investigaciones futuras sobre la implementación de IA en la educación, subrayando cómo los modelos de lenguaje pueden convertirse en herramientas poderosas para la planificación curricular. La capacidad de adaptarse a contextos específicos y la precisión alcanzada sugieren que la IA desempeñará un papel crucial en la evolución de los sistemas educativos en los próximos años.

AGRADECIMIENTOS

A la Dirección de TIC de la Universidad Técnica de Cotopaxi por todo el apoyo brindado para la presente investigación.

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación

CONFLICTO DE INTERESES

Los Autores declaran que no existe conflicto de intereses con su investigación

CONTRIBUCIÓN DE AUTORÍA

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (https://credit.niso.org/). Los autores declaran sus contribuciones en la siguiente matriz:

Participar activamente en:	*Chiluisa Diego*	*Rodríguez Gustavo*
Conceptualización	X	
Análisis formal	X	
Adquisición de fondos	X	
Investigación	X	X
Metodología	X	X
Administración del proyecto	X	X
Recursos	X	X
Redacción –borrador original	X	X
Redacción –revisión y edición	X	X
La discusión de los resultados	X	X
Revisión y aprobación de la versión final del trabajo.	X	X

REFERENCIAS BIBLIOGRÁFICAS:

Atencio-González, R. E., Bonilla-Ron, D. E., Miles-Flores, M. V., & López-Zavala, S. Á. (2023). Chat GPT como Recurso para el Aprendizaje del Pensamiento Crítico en Estudiantes Universitarios. *CIENCIAMATRIA*, *9*(17), 36-44. https://doi.org/10.35381/cm.v9i17.1121

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. En J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65-72). Association for Computational Linguistics. https://aclanthology.org/W05-0909

Burgos, J. J. P., Baque, C. J. S., Rosales, A. J. L., & Ramírez, C. N. J. (2024). Chatgpt como herramienta para la planificación microcurricular del currículo Ecuatoriano. *Conocimiento global*, *9*(3), Article 3.

Giuffrè, M., & Shung, D. L. (2024). Scrutinizing ChatGPT Applications in Gastroenterology: A Call for Methodological Rigor to Define Accuracy and Preserve Privacy. *Clinical Gastroenterology and Hepatology*, *22*(10), 2156-2157. https://doi.org/10.1016/j.cgh.2024.01.024

Giuffrè, M., You, K., & Shung, D. (2023). Evaluating ChatGPT in Medical Contexts: The Imperative to Guard Against Hallucinations and Partial Accuracies. *Clinical Gastroenterology and Hepatology*, *22*. https://doi.org/10.1016/j.cgh.2023.09.035

Introduction to PandasAI. (s. f.). PandasAI. Recuperado 13 de octubre de 2024, de https://docs.pandas-ai.com/intro

Jerez, O. (2015). *El diseño de Syllabus En La Educación Superior: Una Propuesta Metodológica*.

Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., Crocè, L. S., & Shung, D. L. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework. *Npj Digital Medicine*, *7*(1), 102. https://doi.org/10.1038/s41746-024-01091-y

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. http://arxiv.org/abs/2005.11401

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74-81. https://aclanthology.org/W04-1013

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation Studies in Artificial Neural Networks* (arXiv:1901.08644). arXiv. http://arxiv.org/abs/1901.08644

Morles, V. (2002). Sobre la metodología como ciencia y el método científico: Un espacio polémico. *Revista de Pedagogía*, *23*(66), 121-146.

Navarro, Y., Pereira, M., Pereira de Homes, L., & Fonseca Cascioli, N. (2010). Una mirada a la planificación estratégica curricular. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales*, *12*(2), 202-216.

Ocaña-Fernández, Y., Valenzuela-Fernández, L. A., & Garro-Aburto, L. L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones*, *7*(2). https://doi.org/10.20511/pyr2019.v7n2.274

Ojeda, A. D., Solano-Barliza, A. D., Alvarez, D. O., & Cárcamo, E. B. (2023). Análisis del impacto de la inteligencia artificial ChatGPT en los procesos de enseñanza y aprendizaje en la educación universitaria. *Formación universitaria*, *16*(6), 61-70. https://doi.org/10.4067/S0718-50062023000600061

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. En P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Romero-Sacoto, L. A., Yambay-Bautista, X. R., Ramírez-Coronel, A. A., Andrade-Molina, M. C., Cordero-Zumba, N. B., & Magdalena-Sarmiento, M. (2021). *Validation of the questionnaire of perception of the importance, usefulness and structure of the syllabus in microcurricular planning*. https://doi.org/10.5281/ZENODO.5557285

Salamanca Leguizamón, C., Neira Camacho, S. M., & Medina Díaz, E. (2020). Estrategia para el fortalecimiento y seguimiento de las competencias genéricas en el contexto universitario. *Revista Interamericana de Investigación, Educación y Pedagogía, RIIEP*, *13*(2), 283-307. https://doi.org/10.15332/25005421.5509

Solari, M. (2018). Tendiendo puentes para fortalecer la articulación entre la planificación institucional y la planificación de aula. *Cuadernos de Investigación Educativa*, *3*(18), 157-188. https://doi.org/10.18861/cied.2012.3.18.2713

Toapanta Pinta, P. C., Céspedes Granda, S. R., & Núñez Hurtado, P. J. (2018). Instrumento de evaluación docente en la Carrera de Obstetricia – Ecuador. *Investigación en Educación Médica*, *7*(28), 115-116. https://doi.org/10.22201/facmed.20075057e.2018.28.18131

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). *Evaluation of Retrieval-Augmented Generation: A Survey* (arXiv:2405.07437). arXiv. https://doi.org/10.48550/arXiv.2405.07437

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models* (arXiv:2309.01219). arXiv. http://arxiv.org/abs/2309.01219