Comunicar, n. 82, v. XXXIII, 2025 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

Evaluación de documental educativo con IAG utilizando prompt engineering

Evaluation of Educational Documentary using GAI and Prompt Engineering

José Miguel Suárez-Martínez*, Consellería Educación. Generalitat Valenciana (Spain) (jm.suarezmartinez@edu.gva.es) (https://orcid.org/0000-0002-4817-7783)
Roberto Arnau Roselló, Universitat Jaume I (Spain) (rarnau@uji.es) (https://orcid.org/0000-0003-2484-7399)
Rubén Nieto-González, Universidad Jaume I (Spain) (ruben.nieto@uji.es) (https://orcid.org/0009-0006-5195-3081)

RESUMEN

El estudio aborda el uso de la inteligencia artificial generativa (IAG) en la evaluación de material audiovisual educativo, la unidad de análisis es el documental NOMADS producido en el Instituto de Educación Secundaria Cabo de la Huerta de Alicante, España, dentro de un proyecto Erasmus+ sobre derechos humanos. Se diseñó un sistema multiagente utilizando técnicas de prompt engineering (PE) en ChatGPT 4.0, con el objetivo de evaluar multidimensionalmente el documental mediante métricas basadas en indicadores clave. La base metodológica está fundamentada en la ingeniería semiótica, disciplina que estudia la interacción hombre-máquina, consta de cuatro fases en las que integra una variante del método Delphi, una técnica estructurada de obtención de consenso en un grupo de agentes expertos evaluadores. La primera fase del PE genera métricas de evaluación, después se definen perfiles de agentes expertos, en tercer lugar se modela la iteración de prompts para recolectar evaluaciones y en cuarto lugar la validación estadística de los resultados generados. Se despliegan unos instrumentos denominados factores de interacción adaptativa para trazar la evaluación multidimensional con técnicas de PE, integrando perspectivas de cinco expertos del ámbito educativo y audiovisual. Las métricas generadas lograron mapear aspectos como impacto educativo, narrativa transmedia y calidad audiovisual. El resultado aporta una propuesta metodológica con resultados validados desde el propio modelo con diferentes coeficientes, lo que requiere de posteriores validaciones tanto con herramientas cuantitativas como con expertos reales.

ABSTRACT

The study explores the use of generative artificial intelligence (GAI) in evaluating educational audiovisual material, focusing on the documentary *NOMADS*, produced at the Cabo de la Huerta Secondary School in Alicante, Spain, as part of an Erasmus+ project on human rights. A multi-agent system was developed using prompt engineering (PE) techniques in ChatGPT 4.0, aiming for a multidimensional evaluation of the documentary based on key performance indicators. The methodology is grounded in semiotic engineering, a discipline that analyzes human-computer interaction. The process comprises four phases and incorporates a variant of the Delphi method, a structured technique for achieving expert consensus. In the first phase, PE generates evaluation metrics. Then, expert agent profiles are defined. The third phase models the iteration of prompts to gather evaluations, and the final phase involves the statistical validation of results. Adaptive interaction factors are introduced as tools to map multidimensional evaluations using PE, integrating insights from five experts in education and audiovisual fields. The resulting metrics successfully captured elements such as educational impact, transmedia storytelling, and audiovisual quality. The outcome presents a validated methodological framework, with internally confirmed results through different coefficients, highlighting the need for further validation through both quantitative tools and real expert panels

PALABRAS CLAVE / KEYWORDS

Inteligencia artificial, evaluación educativa, ChatGPT, narrativa transmedia, indicadores clave, método delphi.
Artificial Intelligence, Educational Evaluation, ChatGPT, Transmedia Storytelling, Key Indicators, Delphi Method.

Received: 2025-03-10 | Reviewed: 2025-03-28 | Accepted: 2025-03-30 | Online First: 2025-07-21 | Published: 2025-07-24
DOI: https://doi.org/10.5281/zenodo.16121533 | Pages: 90-102

1. Introducción

1.1. Contextualización y justificación

La creciente integración de la Inteligencia Artificial Generativa (IAG) en educación plantea retos y oportunidades significativas, particularmente en la evaluación tanto de aprendizajes como de la implementación de proyectos de innovación. Un buen número trabajos de bibliografía reciente al respecto aborda temáticas sobre el balance riesgo oportunidad en la educación superior (Cordón García, 2023), sobre la problemática del plagio (Díaz Arce, 2024), subraya las contradicciones del cambio de paradigma emergente en el ámbito educativo (Prendes-Espinosa, 2023), la conciliación del potencial de las herramientas y las resistencias al cambio (Diaz Vera et al., 2024) y un variado abanico de aspectos tangenciales al uso operativo de estas herramientas para la generación de conocimiento o la sistematización de la resolución de problemas en el campo educativo e investigador. Sin embargo es más escasa la bibliografía en español que aborde la sistematización de las herramientas IAG para su adaptación a contextos educativos mediante un enfoque instrumental que permita explorar y explotar todo el potencial de generación de conocimiento e innovación. Este trabajo se circunscribe en un proyecto de investigación principal (PIP) enmarcado en la metodología de investigación basada en diseño, IBD o DBR por sus siglas en inglés, Reinking (2021) señala las ventajas de las investigaciones basadas en diseño al promover una relación más estrecha entre teoría e implementación práctica, respondiendo así a necesidades reales de estudiantes y educadores, sin olvidar el fomento de la innovación y la adaptabilidad en entornos educativos dinámicos; aspectos todos ellos que se integran perfectamente con el propósito de conciliar teoría y práctica como una de las motivaciones de este estudio.

Nos fundamentamos en dos pilares teóricos que abordan la interacción humano-computadora. El primero de ellos es la ingeniería semiótica, desde donde se abordará la analítica de metacomunicación (De Souza, 2005). Otro pilar mucho más reciente, el Prompt Engineering (PE), lo podemos conceptualizar como el proceso de diseñar entradas textuales que actúan como instrucciones para un modelo de IAG, buscando maximizar su desempeño en una tarea específica. La emergencia en los últimos años del enfoque operativo del PE busca adaptar modelos de lenguaje como ChatGPT a necesidades específicas de usuarios de diversos contextos. En un campo en plena ebullición la revisión de la literatura de Roumeliotis y Tselikas (2023) destaca el papel de los prompts para optimizar la generación de respuestas y mitigar sesgos, otros estudios recientes introducen estrategias que integran generación enriquecida por recuperación para aumentar la relevancia y coherencia de las salidas del modelo (Zhou et al., 2023). Estos enfoques permiten realizar aproximaciones recursivas para realizar ajustes por refinamiento de los prompts (Chaubey, Tripathi y Ranjan, 2024) y contribuyen a enriquecer las innovaciones que llegan de la mano de los large language models (LLMs) como ChatGPT, son modelos de inteligencia artificial entrenados con grandes conjuntos de datos de texto para predecir, generar y comprender lenguaje humano. Buscando la simplificación conceptual los denominamos en este trabajo IAG de forma genérica.

Este trabajo se enmarca como el sexto resultado de investigación dentro del proyecto principal donde se realiza un estudio de caso ex post facto, desde el mencionado paraguas de la IBD, de dos proyectos Erasmus+ desarrollados en el Instituto de Educación Secundaria Cabo de la Huerta de Alicante, España. Entre los cursos 2015-16 y 2017-18 se desarrolló Human Rights in a European Community of Values, en adelante HUMREV y un segundo proyecto entre los cursos 2018-19 y 2021-22, Bread Way From Hands to Heart, en adelante BWH2H. Entre los múltiples outputs audiovisuales de HUMREV (2016) destaca el documental NOMADS (HUMREV, 2017), orientado a concienciar sobre las migraciones y los derechos humanos y producido colaborativamente en el seno de dicho centro educativo (Suárez-Martínez, 2024, p. 1). En distintos outputs previos del PIP se desarrolló un modelo de evaluación de evidencias en soporte digital mediante indicadores clave que forman la columna vertebral del PIP intersectando todo el material de entrenamiento de un panel multiagente IA. Se trata en este contexto de aplicar las técnicas de PE, usando ChatGPT 4.0, para la evaluación de material audiovisual, el documental educativo NOMADS es tomado como caso de estudio de investigación y los materiales generados en la investigación realizada en Suárez-Martínez (2024) como instrumentos de entrenamiento combinando desde el paraguas de la IBD el enfoque descriptivo y el experimental.

Se establece, por consiguiente, el reto de simular un contexto de cinco expertos en el terreno audiovisual utilizando ChatGPT 4.0 y el plugin Video-Summarizer como herramientas de análisis de contenidos en Youtube con el propósito de evaluar tanto la calidad audiovisual como la educativa de la pieza atendiendo a

un conjunto de métricas creadas por el propio modelo, todo ello realizando una metaevaluación posterior en el seno del mismo en términos de coeficientes cuantitativos que permitan complementar la evaluación cualitativa previa. Se entiende metaevaluación en este contexto como un proceso de validación cuantitativa sobre un conjunto de evaluaciones generadas por el sistema de IAG, que reutiliza sus propios outputs como insumos.

1.2. Marco teórico y estado de la cuestión

1.2.1. Ingeniería semiótica

El primer pilar teórico para fundamentar este estudio se asienta en la obra The semiotic engineering of human-computer interaction De Souza (2005), presenta una teoría en la que la interacción humano-computadora (HCI) se conceptualiza como un acto de comunicación entre el diseñador del sistema y el usuario. Su objetivo principal es diseñar, evaluar y analizar sistemas interactivos desde una perspectiva comunicativa y semiótica; aporta el marco conceptual para entender la interacción con ChatGPT como un proceso comunicativo, donde el diseño de prompts es un acto de ingeniería semiótica. Esta visión sostiene metodológicamente la evaluación de NOMADS, al permitir estructurar las interacciones como procesos de metacomunicación evaluable, ajustable y replicable. Nos fundamentamos para mapear las estrategias de PE en esta teoría previa con vocación de contextualizarla en PE en varios de sus elementos fundamentales: el interfaz, como eje de intercomunicación HCI, desde el modelo de los LLM el prompt pasa a ocupar este rol; la metacomunicación, es decir, cómo debe interpretarse y producirse la propia comunicación entre usuario y sistema, para lo que se despliega un mapa de factores de interacción adaptativa posteriormente; la sistematización, al ser sistematizado el proceso metacomunicativo de PE ayuda a entender qué hacer, cómo y por qué, con estrategias de interacción comprensibles; la conversación, como flujo de intercambio que promueva la replicabilidad del modelo entre contextos; la contextualización, atendiendo a un intercambio de carácter interpretativo donde aparecen cuestiones como dónde y para qué; por último para armonizar lo anterior la evaluación de la comunicación, adaptamos el método denominado CEM (Communicability Evaluation Method).

En la tabla 1 abordamos los elementos clave del método CEM armonizado con las intenciones del enfoque PE, sacrificamos en cierta medida la impronta nominativa del original para adaptarlo al propósito que nos ocupa:

Tabla 1: Elementos clave del Communicability Evaluation Method (De Souza, 2005).

CEM	Denominación	Descripción
CEM1	Preparación y definición del contexto	Definir objetivos, roles protagonistas, escenarios, respondemos cuestiones como dónde, para qué, quién
CEM2	Definición de tareas representativas	Seleccionar tareas específicas y representativas de la experiencia especificando aspectos de la metacomunicación
CEM3	Implementación de la interacción	Recolectar respuestas del sistema con la puesta en práctica del modelo
CEM4	Validación de la respuesta del sistema	Validar si está de acuerdo al propósito inicial denotando errores, sesgos y desviaciones
CEM5	Análisis metacomunicativo	Interpretar los patrones de desviación y sesgos identificando puntos de mejora
CEM6	Heurísticas de rediseño metacomunicativo	Implementar propuestas de mejora iterativamente
CEM7	Documentación para retroalimentación	Documentar los objetivos con implementación exitosa para servir de retroalimentación en el siguiente ciclo

1.2.2. Prompt engineering y los factores de interacción adaptativa

Definimos los factores de interacción adaptativa (FIAs) en prompt engineering como un conjunto de indicadores clave que contribuyen a guiar y modelar el diseño del ajuste dinámico de los *prompts*, traducidos como instrucciones o indicaciones para sistemas de IAG, con el objetivo de mejorar la calidad de las respuestas dinámicamente en función de las necesidades del usuario y del contexto específico de forma sistematizada. Distintos autores como Singh, Samborowski y Mentzer (2023) destacan el uso de tipologías de prompts de configuración de contexto o context-setting y de roles o role-plays para guiar la producción de respuestas como un proceso co-creativo humano-IA. Moreno et al. (2024) proporcionan un patrón para elaborar un «prompt perfecto» en su revisión de la literatura sobre PE: [Contexto] + [Información específica] + [Propósito] + [Formato de la respuesta] = Prompt perfecto. Autores como Joshi et al. (2024) afirman la sistematización iterativa mejora el diseño de prompts mediante un enfoque estructurado y dinámico. El proceso iterativo permite a los usuarios ajustar y refinar criterios y prompts según los resultados obtenidos, asegurando una mejora continua. Todos los autores enfatizan la necesidad

de estrategias de validación y evaluación en sistemas de IAG buscando la consistencia en los resultados obtenidos. En aras de una posterior sistematización para su explotación metodológica recopilamos en acrónimos numerados algunos de los FIAs más significativos encontrados en la literatura especializada con objeto de ser instrumentalizados en la discusión de resultados.

Zhou et al. (2023) proponen una metodología innovadora llamada DYNAICL para optimizar la eficiencia de los modelos generalistas de lenguaje mediante la asignación dinámica de ejemplos en contexto dentro del prompt a modo de feed-forward. Esta dinámica de alimentación contextual de los prompts es similar a técnicas como Retrieval-Augmented Generation (RAG) o generación enriquecida por recuperación (Li et al., 2024) que es una combinan los modelos de lenguaje generativo con sistemas de recuperación de datos relevantes provenientes de salidas previas del prompting para enriquecer el contexto y la precisión de las salidas Esta metodología se enfoca en superar las limitaciones de los modelos generativos puros, que a menudo producen información genérica o incorrecta cuando no tienen suficiente conocimiento contextual. El modelo RAG que conforma una base importante en este trabajo parte de dos conceptos fundamentales: recuperador y generador. El recuperador, RAG1, es un dataset obtenido fuera del modelo de lenguaje (Suárez-Martínez, 2024) y que sirve para alimentar al mismo, el generador, RAG2 (ChatGPT, 2025a, 2025b, 2025c), es la respuesta de precisión aumentada propiciada por el recuperador.

Las investigaciones de White et al. (2023) muestran cómo el ajuste dinámico de los prompts permite a la IA adaptarse mejor a las necesidades contextuales de los usuarios, describen un catálogo de patrones para optimizar la interacción y personalización de respuestas de modelos de lenguaje como ChatGPT. Estos patrones ofrecen soluciones reutilizables y están organizados en cinco categorías de patrones clave que referimos como CPs. El artículo *Prompt Engineering a Prompt Engineer* de Ye et al. (2023) explora técnicas avanzadas para optimizar la ingeniería de prompts centrándose en un método denominado PE2. Recopilamos tres de los principios esenciales para mejorar la calidad de los prompts y potenciar la capacidad del modelo para adaptar prompts a situaciones complejas y para elaborar soluciones específicas para tareas inéditas que referimos como PEPs. En la tabla 2 seleccionamos algunos de los FIAs empleados en nuestra investigación.

Tabla 2: FIAs con patrones clave y optimización de prompts.

Siglas	Denominación	Descripción
CP1	Semántica de entrada	Usar patrones que ayudan a definir cómo el modelo debe interpretar la entrada y su contextualización
CP2	Personalización de salida	Ajustar el tipo y formato de salida en forma de patrones de rol
CP3	Identificación de errores	Utilizar listas de verificación de hechos o condiciones que permiten que el modelo genere una lista de validación
CP4	Optimización de prompts	Incluir patrones de *refinamiento de preguntas*, que ayuda a mejorar las preguntas iniciales del usuario descomponiéndose en subpreguntas más sencillas
CP5	Interacción y control de contexto	Formular prompts con *interacción invertida*, donde el modelo hace preguntas al usuario para guiar la conversación hacia un objetivo específico
PEP1	Contextualización	Aclarar cómo el prompt y el texto de entrada deben interactuar, ajustándose a la estructura de la tarea
PEP2	Secuenciación evaluable	Desglosar el proceso de generación de resultados en fases, se evalúa el prompt actual y luego sugiere mejoras mediante plantilla de razonamiento paso a paso
PEP3	Ajustar iterativamente los errores	Proporcionar al modelo una guía estructurada para analizar ejemplos de errores, refinando los prompts consecutivamente a la aparición de errores

Tabla 3: FIAs para entrenamiento y mejora de resultados de modelos IAG.

Siglas	Descripción	Vínculos a otros FIAs
XAI1	Entrada precisa y estructurada	CP1, PEP1, FT2
XAI2	Filtros epistémicos incorporando mecanismos de selección o descarte de resultados en función del determinados sesgos de los usuarios	CP3, CP4, PEP3
XAI3	Factor de explicabilidad demandando al modelo explicaciones claras y rastreables del origen de los resultados o los mecanismos deductivos conducentes al mismo	CP4, CP5, FT3
FT1	Definir los objetivos y el dominio del modelo	PEP1, XAI1, CP1
FT2	Estructurar los datos de entrada para aumentar la precisión	XAI1, PEP1, FT4
FT3	Crear muestras preformateadas que reduzcan sesgos	CP1, PEP1, XAI1
FT4	Refinamiento iterativo	PEP2, PEP3, CP5
FT5	Documentación y evaluación final	PEP3, CP3, FT3

Moruzzi, Ferrari y Riscica (2024) abordan una metodología integrada de PE con filtros epistémicos y explicabilidad, sus FIAs los referimos como XAI, de las siglas Explainable Artificial Intelligence (XAI) la metodología combina el diseño de entradas textuales estratégicas para modelos de lenguaje con herramientas de explicabilidad, considerando los sesgos cognitivos y contextuales de los usuarios y sistemas. Su propósito principal es maximizar la relevancia, transparencia y efectividad de las interacciones humano-máquina mediante la comprensión y mitigación de sesgos. Guo et al. (2024) cuando determinan los principios del diseño de muestras de entrenamiento mediante un proceso al que denominaremos fine-tuning de los modelos destacan varios factores que definimos como FTs. Se muestran ambos enfoques en la tabla 3 armonizados con los FIAs de la tabla 2.

1.2.3. Método Delphi para creación de un panel de expertos

El Método Delphi es una técnica estructurada que se utiliza para obtener consenso entre expertos en temas complejos, inciertos o polémicos. Este método es ampliamente reconocido en áreas como la planificación estratégica, la prospectiva y la evaluación de políticas públicas. Según Okoli y Pawlowski (2004), el método puede ser reformulado como un proceso sistemático de construcción de conocimiento colectivo que no depende únicamente de la interacción directa entre participantes, sino de la iteración controlada de juicios expertos. Esta perspectiva habilita su uso en entornos mediados por agentes computacionales, siempre que se mantenga la lógica estructural del método: anonimato, retroalimentación iterativa y síntesis progresiva de respuestas.

Las etapas del Método Delphi incluyen la selección de expertos (EMD1): los participantes son seleccionados con base en su conocimiento y experiencia en el área de estudio. Su anonimato garantiza la neutralidad de las respuestas (Okoli y Pawlowski, 2004). Formulación de preguntas (EMD2): se diseñan cuestionarios estructurados o semiestructurados, enfocados en los objetivos del estudio. Estas preguntas suelen ser abiertas en la primera ronda y luego se refinan (Rowe y Wright, 1999). Rondas iterativas (EMD3): en cada ronda, los expertos responden al cuestionario y proporcionan su razonamiento. El facilitador analiza las respuestas, identificando patrones, consensos y divergencias, y reenvía un resumen para una nueva evaluación (Hsu y Sandford, 2007). Análisis y consenso (EMD4): las iteraciones continúan hasta alcanzar un nivel satisfactorio de acuerdo o identificar claramente las diferencias irreconciliables. Esto se mide mediante métricas como la desviación estándar de las respuestas o el porcentaje de consenso (Turoff y Linstone, 2002).

Procedemos, por tanto, a justificar la pertinencia de combinar la triangulación de técnicas de prompt engineering e ingeniería semiótica con métodos de evaluación por consenso experto como Delphi. Las dos últimas comparten una fundamentación sistematizada y validada metodológicamente que aportan un marco estructurado que impacta en el modelado y la replicabilidad de este estudio orientado a generar una propuesta metodológica que llamamos Prompt Engineering Evaluation Method.

1.3. Objetivos

El objetivo principal de este trabajo es modelar las estrategias de prompt engineering para la evaluación del documental educativo NOMADS creando un sistema multiagente de IAG. Entendemos aquí el concepto de sistema multiagente como la simulación de un grupo de expertos en el ámbito educativo, transmedia y audiovisual modelados mediante prompting en ChatGPT utilizando el plugin de análisis de vídeos en Youtube Video Summarizer.

Atendemos a varios objetivos secundarios vinculados a sendas fases del proceso.

- OS1, obtener métricas de evaluación del documental formuladas desde el esquema Goal-Question-Metric, en adelante GQM, estableciendo con ello un context-setting.
- OS2, configurar un panel de 5 expertos del modelo multiagente IA que permite modelar un role-play setting.
- OS3, desarrollar y documentar las diferentes iteraciones del PE que permiten obtener las evaluaciones de los distintos expertos para mapear una adaptación del método Delphi generando un dataset de resultados de evaluación.
- OS4, realizar una metaevaluación del dataset de resultados para complementar cuantitativamente la evaluación realizada por el panel de expertos.
- OS5, fundamentar todo el proceso para crear un marco sistematizado, documentado y replicable en contextos diferentes.

2. Materiales y método

2.1. Materiales de entrenamiento del modelo

Distinguimos en este punto varios tipos de materiales previos que se deben contextualizar en el marco del PIP. Si tomamos el modelo RAG (Li et al., 2024) podremos conceptualizar estos materiales como el recuperador, RAG1, es decir, un conjunto de datos obtenido fuera del modelo de lenguaje, dentro del proyecto de investigación, en forma de papers de investigación (Suárez-Martínez, 2024) y que sirven para entrenar al mismo.A continuación, se describen y categorizan los recuperadores obtenidos en fases previas del PIP.
RAG1a: la descripción del documental NOMADS en (Suárez-Martínez, 2024, p. 1), ofrece al modelo la posibilidad de análisis del material residente en Youtube contenido en el documental educativo (HUMREV, 2016) y toda la descripción del desarrollo de dicha producción colaborativa.
RAG1b: los indicadores clave de diseño, o KDIs, de la innovación docente obtenidos del análisis del contexto de HUMREV en Suárez-Martínez (2024, p. 2), ofrecen un paper que mapea el conocimiento preciso del contexto del centro al inicio del proyecto Erasmus+ y los principales desafíos abordados.
RAG1c: se utilizan los indicadores clave de aprendizaje derivados de los objetivos del proyecto que sustentan los aprendizajes competenciales y metodológicos más significativos y su impacto en la dinámica del centro en Suárez-Martínez (2024, p. 3) como este tercer recuperador.
RAG1d: el modelo de evaluación de las evidencias multimedia de HUMREV mediante las métricas obtenidas en los indicadores clave precedentes se realizan en Suárez-Martínez (2024, p. 4). Este recuperador está basado en ofrecer métricas de evaluación sistematizadas mediante la técnica Goal Question Metric que inicialmente nació en entornos de ingeniería del software (Basili, 1992).
RAG1e: al ser los indicadores clave la columna vertebral de la investigación es crucial ver de qué manera apoyan la labor de redefinición a otros contextos. Esto se obtiene en el recuperador propuesto en Suárez-Martínez (2024, p. 5) donde se rediseñan indicadores clave para evaluar las evidencias del segundo proyecto Erasmus+, BWH2H desarrollado entre los cursos 2017-18 y 2021-22.
RAG1f: el requerimiento de realizar una analítica con indicadores de la dimensión transmedia del documental es propiciado por este recuperador basado en Suárez-Martínez (2023) donde se formula la implementación de estos instrumentos sobre material educativo.

2.2. Materiales de entrada/salida recursiva o regenerada

Describimos la salida del generador, RAG2 (ChatGPT, 2025a, 2025b, 2025c)[1], ya definido como la respuesta de precisión aumentada propiciada por el entrenamiento del modelo con los recuperadores anteriores (Li et al., 2024). Con este material de entrenamiento y diferentes técnicas de PE los expertos modelados mediante el panel deben ser capaces de generar unos indicadores rediseñados para evaluar el documental: indicadores clave de diseño, KDIs; indicadores de clave de aprendizaje, KLIs; indicadores clave de calidad audiovisual o KAVIs e indicadores clave de producción transmedia o KTIs, generados desde el propio modelo para implementar las métricas de evaluación (ChatGPT, 2025b), le denominamos RAG2a a este generador. Con estos indicadores clave se logrará extraer un libro de datos de métricas cualitativas según el enfoque GQM (ChatGPT, 2025a) al que denominamos RAG2b.

2.3. Método

Se describen las diferentes fases del método representadas en la figura 1. La primera fase, de generación de métricas de evaluación, se trata de alimentar al modelo para el proceso RAG1 (Li et al., 2024), usando como entrada los cinco recuperadores de RAG1a hasta RAG1e obtenidos de la investigación previa del PIP. El generador o dataset de salida de esta fase, RAG2a, son las tablas de métricas de evaluación con varios tipos de indicadores (ChatGPT, 2025b, pp. 23-25). El objetivo de esta fase es la primera parte del context-setting que ofrecerá estructuras de datos mediante el enfoque Goal Question Metric para que sirvan de material de entrenamiento del panel de expertos desde las salidas anteriores (Zhou et al., 2023).

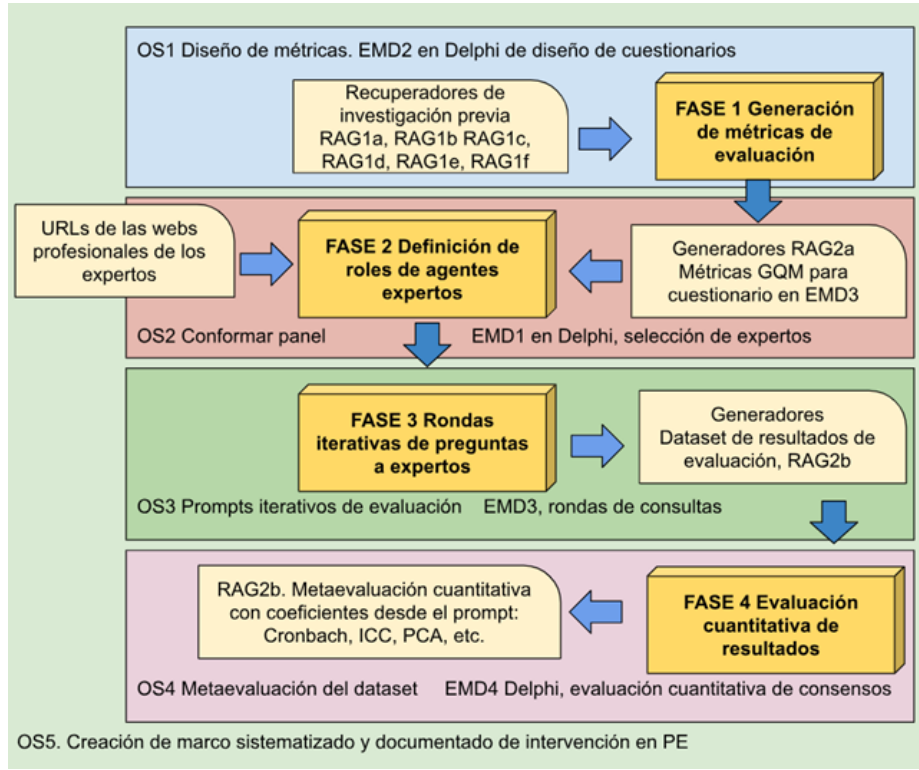
Segunda fase, definimos los roles actanciales de los expertos modelados desde sus webs académicas trazando cinco tipos de perfiles, el uso de prompts de configuración de contexto o context-setting y de roles o role-plays para guiar la producción de respuestas específicas de IA (Singh et al., 2023). Los datos de entrada son cinco URLs obtenidas de los perfiles profesionales de sendos expertos reales en materia audiovisual, educativa y transmedia. Se seleccionan dichos expertos reales por estar en el ámbito del

proyecto de investigación y poder contar eventualmente con su colaboración para una validación real en el futuro. La conformación del panel resumida por perfiles se explicita en la tabla resumen del final del generador (ChatGPT, 2025b, pp. 35-37).

La tercera fase, que denominamos de rondas iterativas de consulta al panel de expertos, simula la etapa tres de rondas de preguntas del método Delphi o EMD3. Más que una fase secuencial en sí es un proceso iterativo de recopilación de las respuestas más adecuadas por cada uno de los agentes con PE y la generación de dos tipos de outputs, un dataset para su retroalimentación posterior, o feedforward en la fase de análisis y validación cuantitativa; RAG2b como dataset de evaluación (ChatGPT, 2025a) y un registro documental del hilo de ChatGPT para mapear con los FIAs instrumentalizados (ChatGPT, 2025c) y así se genera un documento histórico de todo el proceso de cara a analizar, refinar o replicar el experimento que contiene toda la lógica experimental desplegada y mapeable con los FIAs,

En la cuarta fase, de análisis cuantitativo, usamos como input el generador obtenido de RAG2b (ChatGPT, 2025a) donde reside el dataset de métricas y criterios generados, se convierte ahora en recuperador y con él realizamos desde el propio modelo una evaluación aplicando diferentes coeficientes de validación estadística como Alpha de Cronbach, Split-half, Pearson, Spearman, Análisis de Concordancia Interjueces (ICC), Análisis de Componentes Principales (PCA), etc. Se trata de ofrecer un correlato cuantitativo para una validación de los datasets obtenidos e instrumentalizados como recuperadores del proceso de PE. Los coeficientes cuantitativos y sus respectivos marcos teóricos asociados no los detallamos aquí por exceder este propósito la extensión de este trabajo pero obedecen a una validación previa en el ámbito de la literatura de investigación y son propuestos por el propio modelo desde un esquema de interacción invertida, con filtros epistémicos y explicabilidad (Moruzzi et al., 2024; White et al., 2023). La fiabilidad de la propuesta metodológica está condicionada a una validación posterior tanto de tipo cuantitativo para convalidar las métricas ofrecidas por los indicadores del modelo como de tipo cualitativo contrastando la evaluación con expertos reales.

Figura 1: Mapa de flujo de recuperadores y generadores RAG vinculados a objetivos, fases y etapas del método Delphi. Fuente: elaboración propia.



3. Resultados

3.1. Fase de generación de métricas y FIAs aplicados

El propósito es cubrir OS1, la entrada son los recuperadores de RAG1a a RAG1f, se persigue crear los cuestionarios de la EMD2 en Delphi, la salida es RAG2a como generador de esta fase que se sintetiza en las métricas obtenidas (ChatGPT, 2025b, pp. 26-28).

Hemos puesto varios tipos de FIAs evidenciados en ChatGPT (2025b). Desde la tabla 2 recopilamos CP1 semántica de entrada (p. 2), CP2 con los patrones de rol o role-play settings, actúa como, CP4 se realizan a lo largo de RAG2a distintos tipos de refinamiento de preguntas hasta obtener las métricas orientando al modelo con diversos materiales, CP5 el modelo de interacción invertida permite plantear un prompt del tipo «Qué preguntas puedo hacerte para elaborar un prompt que te ayude a ofrecer un marco teórico donde basarnos para obtener los 6 indicadores mencionados»(p. 6). Hemos usado la contextualización, PEP1, la secuenciación evaluable por el usuario, PEP2 y hemos conseguido ajustar iterativamente los errores, PEP3. Desde la tabla 3 tenemos: entrada precisa y estructurada, XAI1; factores de explicabilidad, XAI3 (p.8), FT1, definir los objetivos y el dominio del modelo (pp. 2-6); FT2, estructurar los datos de entrada para aumentar la precisión (p. 2); FT3, crear muestras preformateadas que reduzcan sesgos; FT4, refinamiento iterativo durante todas las aproximaciones realizadas y FT5, documentación y evaluación final.

3.2. Fase de definición de roles del panel de expertos

Atendiendo al OS2, con el propósito de configurar un panel de cinco expertos del modelo multiagente pasamos a identificar los resultados del role-play setting coincidente con la primera de las fases de la adaptación del método Delphi que establece la selección de expertos, estamos simulando en Delphi la EMD1 usando varios FIAs como CP1 y PEP1. En ella los expertos participantes son seleccionados con base en su conocimiento y experiencia en el área de estudio y están vinculados a expertos reales del grupo de investigación ITACA[2] de la Universidad Jaume I de Castellón, España. Lo cual partiendo de que la información del perfil de los diferentes expertos se obtiene de sus websites profesionales, es de dominio público y se proporciona a ChatGPT (2025b, pp. 28-35) mediante las URLs de cada uno de ellos integrando CP1, CP2, PEP1, FT4. Finalmente integrando FT5 se solicita una tabla resumen (p.35) que incluiremos en nuestro RAG2a. Esto proporciona una posibilidad de validación de los resultados en el futuro recurriendo a estos expertos y justifica operativamente la muestra implementada. El propio ChatGPT (2025b) aporta respuestas que facilitan la explicabilidad como «Si deseas que elabore un perfil detallado o adapte sus áreas de expertise específicas a las métricas propuestas, házmelo saber» (p. 29) vinculadas a los XAIs. La formulación de preguntas de la adaptación fundamentada en el método Delphi desde EMD2, en la que se diseñan cuestionarios estructurados o semiestructurados, enfocados en los objetivos del estudio que se han establecido específicamente mediante los indicadores clave y las métricas de la fase anterior. El enfoque GQM permite una definición consistente de los criterios de evaluación desde los indicadores clave.

3.3. Fase de rondas iterativas del panel de expertos y FIAs asociados

Atendemos en esta tercera fase a OS3, con el propósito de desarrollar y documentar los diferentes procesos del PE que permiten obtener las evaluaciones de los distintos expertos apoyándonos en los FIAs para mapear una adaptación del método Delphi, generando con ello un dataset de feed-forwarding para la cuarta y última fase de metavalidación cuantitativa. Desde nuestra implementación de la variante del método Delphi estas rondas iterativas o EMD3, se caracterizan porque en cada ronda los expertos responden las métricas GQM y proporcionan su razonamiento en una columna llamada criterio. Se formulan preguntas en búsqueda de análisis y consenso de manera colegiada. Las iteraciones continúan hasta alcanzar un nivel satisfactorio de acuerdo o identificar claramente las diferencias irreconciliables. Esto se mide posteriormente también de forma cuantitativa en la fase siguiente.

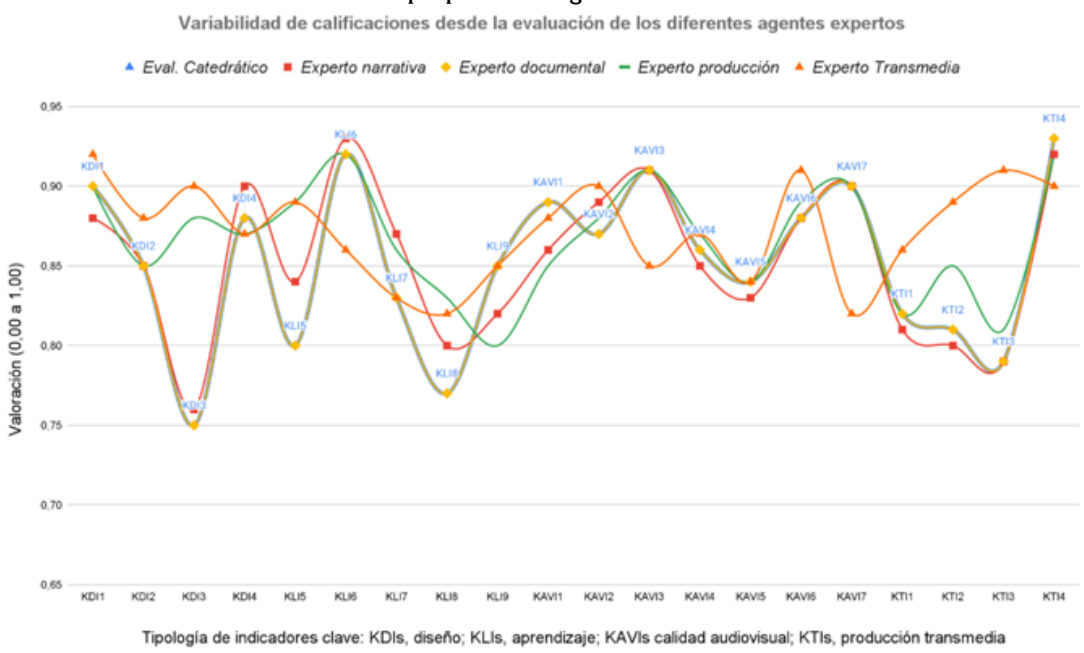
Los datos de entrada son RAG1a y la tabla de métricas GQM de evaluación de ChatGPT (2025b, pp. 26-28) que denominamos previamente como RAG2a, ya que pasa de ser un generador a realizar el rol de recuperador según el modelo de RAG propuesto por Li et al. (2024). Se destacan así las conceptualizaciones realizadas (Joshi et al., 2024; Singh et al., 2023) al formular la potencia de las estrategias de sistematización iterativa y adaptativa para reducir sesgos del modelo, generar mayor precisión o ajustar y mejorar el desempeño de los modelos de IAG. La salida o generador de esta fase es por tanto RAG2b disponible en ChatGPT (2025c) y resumida en el dataset de métricas de en el libro de Google Sheets de ChatGPT (2025a).

Analizando los FIAs de la tabla 2, en ChatGPT (2025c) se integran: CP1 semántica de entrada (p. 2); CP2 con los patrones de rol o role-play settings para las evaluaciones de todos los expertos; CP4 se realizan a lo largo del proceso distintos tipos de refinamiento de preguntas hasta obtener las métricas orientando al modelo con diversos materiales como «RESULTADO PARCIALMENTE CORRECTO, REFINAMIENTO de presentación, la tabla debe aparecer en pantalla…» (p. 10); CP5 el modelo de interacción invertida propone debates entre los expertos con explicaciones tabuladas e intercambio de preguntas, «los tres agentes utilizados: [...] deben establecer un consenso» (p. 41). Integramos PEP1, la secuenciación evaluable por el usuario, PEP2 y al ajustar iterativamente los errores, PEP3. De la tabla 3 se implementan: entrada precisa y estructurada, XAI1; filtros epistémicos, XA2 y la explicabilidad, XAI3, hay que reseñar la importancia de sustentarse en la tabla de métricas GQM para definir los objetivos, FT1, al igual que FT2, al estructurar los datos de entrada para aumentar la precisión. Debemos poner de manifiesto todo el potencial de debate de expertos desplegado en ChatGPT (2025c, pp. 41-69) donde se evidencia la solvencia y consistencia análisis cualitativo realizado por cada uno de los expertos, las preguntas cruzadas en formato focus-group, el enfoque de debate y consenso desde una precisión ciertamente congruente con los perfiles demandados a cada uno de ellos.

3.4. Fase de análisis cuantitativo y metavalidación desde el modelo

Atendemos aquí al OS4, con el propósito de realizar una metaevaluación con distintos coeficientes del dataset de resultados generado (RAG2b) que ahora se convierte en recuperador para esta fase, tras tabular todos los datos en ChatGPT (2025a) en un libro de Google Sheets con objeto de complementar cuantitativamente la evaluación realizada por el panel de expertos. Entendemos aquí el concepto de metavalidación cuantitativa como la evaluación con distintos índices cuantitativos de las evaluaciones generadas en la tabla de métricas de RAG2b.

Figura 2: Gráfico de la congruencia de la evaluación entre distintos expertos. Fuente: elaboración propia en Google Sheets.



El abordaje de puesta a prueba de la metaevaluación cualitativa comienza en ChatGPT (2025c, p. 69) pidiendo al modelo que sugiera algunas fórmulas de validación estadística de los datos. Observamos una discordancia significativa antes y después de introducir el recuperador RAG2b en ChatGPT (2025c) lo cual valida las referencias del marco teórico. El coeficiente de Alpha de Cronbach muestra antes resultado negativo (-0.29) y tras adjuntar RAG2b afirma «El Alpha de Cronbach calculado para las valoraciones de los expertos

utilizando los datos del PDF es 0.84, lo que indica una alta consistencia interna entre las valoraciones» (pp. 75-76) . Se reproducen aquí otras salidas de esta fase: «a partir del Análisis de Componentes Principales (PCA) El PCA revela que, en general, los expertos coinciden en sus valoraciones sobre la narrativa, la accesibilidad transmedia y el impacto emocional del documental» (pp. 71-74). «El coeficiente de Spearman-Brown de 0.91 indica una alta consistencia interna entre las dos mitades de las evaluaciones. Este resultado respalda la fiabilidad de las valoraciones» (p. 79). En definitiva, a pesar de no ser validada cuantitativamente de forma rigurosa esta metaevaluación se puede afirmar que el modelo es capaz de ofrecer una orientación para profundizar en una segunda confirmación de los datos haciendo uso del generador RAG2b y las indicaciones realizadas para fórmulas en hojas de cálculo. En la figura 2 se obtiene un gráfico generado en Google Sheets desde ChatGPT (2025a) que aproxima una cierta congruencia interjueces desde un análisis exógeno al modelo.

4. Discusión y conclusiones

4.1. Discusión de resultados

El método CEM, mencionado como *Communicability Evaluation Method* (De Souza, 2005), armonizado con el resto del marco metodológico presentado, podemos redefinirlo como *Prompt Engineering Evaluation Method* (PEEM) y ver las diferentes dimensiones que cubre desde el marco ofrecido por su precedente, atendiendo a nuestros objetivos.

CEM1, preparación y definición del contexto, se cubre en los OS1 y OS2 de las fases correspondientes a través de los FIAs como CP1 y PEP1, que permiten configurar entradas precisas y contextualizadas (White et al., 2023). CEM2, con la definición de tareas representativas, se operacionaliza mediante los FIAs que seleccionan interacciones prototipadas y contextualizadas, en especial desde la metacomunicación adaptada de De Souza (2005) y por medio de estrategias de retroalimentación estructurada (Joshi et al., 2024). CEM3, implementación de la interacción, se sustenta en la recolección iterativa de respuestas del sistema, utilizando ajustes progresivos de prompts desde una lógica de in-context learning dinámica o DYNAICL (Zhou et al., 2023). CEM4, validación de la respuesta del sistema de acuerdo al propósito inicial, se realiza de forma recursiva e iterativa tal como formulan las fases de la IBD, convalidando con el marco teórico de diseño iterativo (Easterday, Rees Lewis y Gerber, 2018) y de aprendizaje adaptativo estructurado por prompts (Li et al., 2024). CEM5 corresponde al análisis metacomunicativo, se sufraga así en el OS5 al fundamentar todo el proceso del PEEM como un proceso sistemático de interacción humano-máquina donde el diseño del mensaje es también un acto evaluable (De Souza, 2005). CEM6, que muestra las heurísticas de diseño de la interacción metacomunicativa, se traduce en la aplicación de FIAs que permiten mapear y reajustar los errores, como reclaman Guo et al. (2024) para entrenar modelos con muestras más eficientes y menos sesgadas. Finalmente, CEM7, documentación para retroalimentación, se materializa en los ciclos RAG propuestos como estructura para capturar, ajustar y regenerar salidas, lo cual concuerda con la lógica de recuperación aumentada de contexto (Li et al., 2024) y con la documentación de patrones reutilizables de prompts (White et al., 2023). Todo ello contribuye a establecer un marco de sistematización replicable en contextos diversos, que es, en última instancia, el objetivo del Prompt Engineering Evaluation Method (PEEM).

Los resultados de la fase del OS1, muestran en ChatGPT (2025b) las métricas de evaluación del documental formuladas desde el enfoque Goal-Question-Metric (Basili, 1992) estableciendo con ello un context-setting de evaluación efectivo y versátil, con posibilidad de ser adaptado a múltiples contextos. Los resultados de la fase del OS2 permiten configurar un panel de cinco expertos del modelo implementando un role-play setting sencillo, económico y con resultados consistentes ahorrando los importantes costes asociados al método Delphi tradicional (Turoff y Linstone, 2002) minimizando muchas de sus desventajas. El OS3 es transversal a todo el estudio pues persigue desarrollar y documentar las diferentes fases del prompting que permiten obtener las evaluaciones de los distintos expertos apoyándonos en los FIAs y sus resultados se evidencian en los RAG1s y los RAG2s. OS4, realizar una evaluación del dataset de resultados para complementar cuantitativamente la evaluación realizada por el modelo utilizando diferentes instrumentos. Medir el alcance de OS4 requiere una aproximación sistematizada y rigurosa que excede el propósito de este trabajo y al mismo tiempo lo complementa. La figura 2 refiere una aproximación visual, aunque superficial, que permite observar un boceto de la coherencia de las valoraciones con algunas disparidades en ciertos indicadores clave (ChatGPT, 2025c). Respecto a OS5, este trabajo persigue transversalmente fundamentar el proceso para crear un marco sistematizado, documentado y replicable en contextos diferentes, al que nos referimos como PEEM.

Tabla 4: Matriz de corresponencias de FIAs y claves del método CEM con objetivos.		
Factores CEM	Correspondencia con FIAs	Objetivos de investigación vinculados a etapas del método Delphi
CEM1	Definición precisa contexto (CP1, PEP1, XAI1, FT1), control sesgos (XAI2, FT3)	OS1 y OS2, configurar panel de expertos con métricas de evaluación EMD1 y EMD2
CEM2	Personalización tareas (CP2), secuenciación (PEP2), explicabilidad (XAI3)	OS3, documentar todas las fases del PE con los FIAs asociados
CEM3	Optimización iterativa (CP4, FT4), filtros epistémicos (XAI2), interacción clara (PEP1)	OS1, OS2 y OS3 precisan de recopilar las respuestas del modelo, la EMD3 que aborda las rondas de consultas
CEM4	Validación explícita, detección errores/sesgos (CP3, XAI2), refinamiento iterativo (PEP3, FT4)	OS4 se realiza metavalidación de la evaluación aunque el proceso de validación iterativa es transversal en el proceso
CEM5	Análisis explicabilidad (XAI3), interpretación errores (CP3), ajustes iterativos (PEP3, FT4)	OS5 se documenta todo el proceso metacomunicativo para crear un marco metodológico sistematizado llamado PEEM
CEM6	Rediseño adaptativo (todos los FIAs mapeados contribuyen al proceso de rediseño y replicabilidad de los experimentos)	OS5, los FIAs son los instrumentos que mapean la heurística metacomunicativa del PEEM
CEM7	Documentación evaluación (FT5), retroalimentación iterativa (PEP3), explicabilidad (XAI3)	OS5, documentación del proceso mediante los hilos de prompt de recuperadores RAG1s y generadores RAG2s

4.2. Conclusiones

Se puede concluir que las técnicas de PE pueden ser eficientes y flexibles en contextos educativos, sin requerir entrenamientos costosos si se utilizan combinadas como refieren Chaubey et al. (2024), evidenciando cómo los indicadores clave formulados en papers de las etapas previas de la investigación (Suárez-Martínez, 2024) permiten diseñar muestras de entrenamiento muy eficientes para modelos LLM reduciendo muchos sesgos gracias a las entradas semiestructuradas de entrenamiento con métricas en la línea de lo que reclaman Guo et al. (2024). Se confirma que el uso estratégico de ChatGPT en contextos aplicados debe apoyarse en metodologías robustas en hibridación con control humano mediante PE con refinamientos iterativos como reclaman Joshi et al. (2024) y un abanico de técnicas de mapeo de la metacomunicación con el modelo sostenidos en los FIAs como instrumentos de apoyo en la trazabilidad del proceso del PE. Aunque Roumeliotis y Tselikas (2023) advierten que modelos como ChatGPT pueden tener problemas con la objetividad y precisión semántica si no están correctamente dirigidos, los indicadores clave provistos en los papers previos (Suárez-Martínez, 2024) instrumentalizados como recuperadores RAG1s en combinación con el enfoque de Ye et al. (2023) sugieren que el diseño iterativo de prompts con retroalimentación puede ayudar a reducir errores semánticos de precisión y objetividad. Como conclusiones colaterales, se ha enfocado este trabajo tanto a nivel teórico como operativo desde la aproximación epistémica del paraguas de la IBD con: el diagnóstico del problema, para identificar necesidades y oportunidades; el prototipado, que persigue crear soluciones iniciales para probar ideas; la evaluación, que busca analizar cómo las soluciones funcionan en contextos y con casos reales y el refinamiento iterativo, para extraer implicaciones para el conocimiento científico (Easterday et al., 2018; Reinking, 2021) lo que se postula como una aproximación complementaria para sostener el enfoque metodológico PEEM.

Es preciso insistir y enfatizar nuevamente en que las limitaciones más evidentes a nivel metodológico están en lo concerniente al contraste de los resultados del modelo mediante validación cualitativa con agentes humanos reales junto con la validación cuantitativa mediante herramientas de análisis estadístico de los resultados del libro de datos generado (ChatGPT, 2025a). Todo ello abre la posibilidad a nuevas investigaciones que repercutan en apuntalar al PE como un modelo emergente de investigación experimental, evidenciable, mapeable y transferible a múltiples contextos educativos (Moreno et al., 2024) y a los LLMs, como ChatGPT, como un elemento más del engranaje sistémico de los proyectos de investigación en innovación educativa, triangulando colaborativamente con docentes e investigadores de forma multidimensional. Concluimos por tanto que la ingeniería de prompts no debe entenderse únicamente como una técnica operativa, sino como un componente estratégico del diseño pedagógico contemporáneo aunque requieran como señalan Moreno et al. (2024) un enfoque cuidadoso para abordar desafíos éticos, técnicos y pedagógicos.

Apoyos

Proyecto subvencionado: Alfabetización mediática en los medios de comunicación públicos. Análisis de estrategias y procesos de colaboración entre medios e instituciones educativas en Europa y España (AMI-EDUCOM). PROYECTO PID2022-138840NB-I00. Código: 231305.01/1.

Notas

[1] Nota metodológica al respecto de las citas: las referencias citadas como *ChatGPT (2025a, 2025b, 2025c)* corresponden a documentos generados por el estudio mediante técnicas de PE con el modelo ChatGPT-4.0 de OpenAI. Estas salidas fueron obtenidas a través de interacciones controladas, documentadas y replicables, se considera fundamental dejarlas disponibles públicamente mediante URLs específicas incluidas en la bibliografía, lo que permite su verificación y análisis posterior.

Dado que dichos documentos se han utilizado como parte integral del diseño metodológico, en calidad de *generadores* y, posteriormente, como entradas o *recuperadores* dentro del modelo RAG, se consideran productos *recuperables*, *citables y parte esencial de la validación experimental*, por ello se ha optado por incluirlos en la sección de referencias bajo el formato estándar para recursos digitales. Se ha opta por indizarlos e incorporar epígrafes al material obtenido del hilo de prompt para facilitar su consulta sin alterar la lógica del proceso que conforma parte esencial del experimento. Aunque estas prácticas en la citación están en proceso de consolidación en APA7, se persigue contribuir a conformar criterios emergentes en investigación con IAG, haciendo énfasis en garantizar la trazabilidad del proceso experimental, alineándose con los principios de transparencia, reproducibilidad y aplicabilidad científica.

[2] https://www.culturavisual.uji.es/category/investigadores/

Referencias

Basili, V. R. (1992). *Software Modeling and Measurement: The Goal/Question/Metric Paradigm*. University of Maryland at College Park. https://hdl.handle.net/1903/7538

ChatGPT. (2025a). *Libro de datos de evaluación documental NOMADS en Google Sheets* [Data set obtenido de ChatGPT (2025c)]. https://bit.ly/libro-datos-evaluación-NOMADS

ChatGPT. (2025b). *Respuesta generada con ChatGPT mediante ingeniería de prompts. Elaboración de Métricas de material audiovisual educativo* [Documento PDF generado con modelo de lenguaje GPT-4, indizado y parcialmente editado para facilitar su análisis operativo]. OpenAI. https://bit.ly/hilo-chat-gpt-elaboracion-metricas-NOMADS

ChatGPT. (2025c). *Respuesta generada mediante prompting con ChatGPT, diseñada con técnicas específicas de ingeniería de prompts para evaluación de documental educativo transmedia* [Documento PDF generado con modelo de lenguaje GPT-4, indizado y parcialmente editado para facilitar su análisis operativo]. OpenAI. https://bit.ly/evaluacion-agentesIA-NOMADS-hilo-ChatGPT

Chaubey, H. K., Tripathi, G. y Ranjan, R. (2024). Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development. En *2024 International Conference on Future Technologies for Smart Society (ICFTSS)* (pp. 169-172). IEEE. https://doi.org/10.1109/ICFTSS61109.2024.10691338

Cordón García, O. (2023). Inteligencia Artificial en Educación Superior: Oportunidades y Riesgos. *RiiTE Revista interuniversitaria de investigación en Tecnología Educativa*, (15), 16-27. https://doi.org/10.6018/riite.591581

De Souza, C. S. (2005). *The Semiotic Engineering of Human-Computer Interaction*. MIT Press. https://doi.org/10.7551/mitpress/6175.001.0001

Díaz Arce, D. (2024). Herramientas para detectar el Plagio a la Inteligencia Artificial: ¿cuán útiles son? Tools to detect Plagiarism in Artificial Intelligence: how useful are they? *Revista Cognosis*, *9*(2), 144-150. https://doi.org/10.33936/cognosis.v9i2.6195

Diaz Vera, J. P., Molina Izurieta, R., Bayas Jaramillo, C. M. y Ruiz Ramírez, A. K. (2024). Asistencia de la inteligencia artificial generativa como herramienta pedagógica en la educación superior. *Revista de Investigación en Tecnologías de la Información*, *12*(26), 61-76. https://doi.org/10.36825/RITI.12.26.006

Easterday, M. W., Rees Lewis, D. G. y Gerber, E. M. (2018). The logic of design research. *Learning: Research and Practice*, *4*(2), 131-160. https://doi.org/10.1080/23735082.2017.1286367

Guo, B., Wang, H., Xiao, W., Chen, H., Lee, Z., Han, S., et al. (2024). Sample Design Engineering: An Empirical Study on Designing Better Fine-Tuning Samples for Information Extraction with LLMs. En *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 573-594). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.43

Hsu, C.-C. y Sandford, B. A. (2007). The Delphi Technique: Making Sense of Consensus. *Practical Assessment, Research, and Evaluation*, *12*(1), 10. https://doi.org/10.7275/pdz9-th90

HUMREV. (2016). *Canal de Youtube del proyecto HUMREV* [Video]. YouTube. https://bit.ly/canal-youtube-HUMREV

HUMREV. (2017). *NOMADS, Migrations and Human Rights* [Video]. YouTube. https://bit.ly/NOMADS-documentary-film

Joshi, I., Shahid, S., Venneti, S., Vasu, M., Zheng, Y., Li, Y., et al. (2024). CoPrompter: User-Centric Evaluation of LLM Instruction Alignment for Improved Prompt Engineering. *arXiv preprint arXiv:2411.06099*. https://doi.org/10.48550/arXiv.2411.06099

Li, D., Zhao, Y., Wang, Z., Jung, C. y Zhang, Z. (2024). Large Language Model-Driven Structured Output: A Comprehensive Benchmark and Spatial Data Generation Framework. *ISPRS International Journal of Geo-Information*, *13*(11), 405. https://doi.org/10.3390/ijgi13110405

Moreno, Y., Ortega, L., Reyes, J. y Saldana-Barrios, J. J. (2024). Revisión Sistemática de la Literatura Acerca de Prompt Engineering Enfocado en la Educación. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E74), 328-345. https://dialnet.unirioja.es/servlet/articulo?codigo=9929876

Moruzzi, S., Ferrari, F. y Riscica, F. (2024). Biases, Epistemic Filters, and Explainable Artificial Intelligence. En *Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI)*. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3825/short1-3.pdf

Okoli, C. y Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), 15-29. https://doi.org/10.1016/j.im.2003.11.002

Prendes-Espinosa, M. P. (2023). La revolución de la Inteligencia Artificial en tiempos de negacionismo tecnológico. *RiiTE Revista interuniversitaria de investigación en Tecnología Educativa*, (15), 1-15. https://doi.org/10.6018/riite.594461

Reinking, D. (2021). *Design-Based Research in Education: Theory and Applications*. Guilford Publications.

Roumeliotis, K. I. y Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*, *15*(6), 192. https://doi.org/10.3390/fi15060192

Rowe, G. y Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, *15*(4), 353-375. https://doi.org/10.1016/S0169-2070(99)00018-7

Singh, J., Samborowski, L. y Mentzer, K. (2023). A Human Collaboration with ChatGPT: Developing Case Studies with Generative AI. *Proceedings of the ISCAP Conference*, *9*, n6039. https://bit.ly/3C64Nfu

Suárez-Martínez, J. M. (2023). *Propuesta de Indicadores de logro para evaluación de estrategia transmedia en educación secundaria* [Trabajo Fin de Máster]. Universidad Jaume I. https://doi.org/10.35542/osf.io/epyb7_v1

Suárez-Martínez, J. M. (2024). *Evaluación de proyectos Erasmus + con Indicadores clave e inteligencia artificial generativa*. Anexo Publicaciones. https://bit.ly/anexo-publicaciones

Turoff, M. y Linstone, H. A. (2002). *The Delphi Method-Techniques and Applications*. http://is.njit.edu/pubs/delphibook/

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382

Ye, Q., Axmed, M., Pryzant, R. y Khani, F. (2023). Prompt Engineering a Prompt Engineer. *arXiv preprint arXiv:2311.05661*. https://doi.org/10.48550/arXiv.2311.05661

Zhou, W., Jiang, Y. E., Cotterell, R. y Sachan, M. (2023). Efficient Prompting via Dynamic In-Context Learning. *arXiv preprint arXiv:2305.11170*. https://doi.org/10.48550/arXiv.2305.11170