

Aprendizaje guiado por IA: Evaluación de un asistente socrático para la mejora en la creación de prompts

Oriol Borrás-Gené (Universidad Rey Juan Carlos)
Antonio Julio López-Galisteo (Universidad Rey Juan Carlos)

Resumen

La integración de la inteligencia artificial generativa en la educación superior abre nuevas posibilidades para tutorías personalizadas y aprendizaje activo. En este artículo se evalúa un asistente generado en la plataforma Gemini para formar a estudiantes universitarios y, en general, a cualquier persona, en la Ingeniería de Prompts. El asistente, denominado “Mejora tus prompts con OriolTIC”, adopta un enfoque socrático, en lugar de dar respuestas inmediatas, guía al usuario mediante preguntas, correcciones y andamiaje progresivo a lo largo de 12 niveles de dificultad que van introduciendo distintas técnicas de prompting. Se presenta un estudio experimental con 4 simulaciones completas de extremo a extremo (60 tests), evaluando el desempeño del asistente en cuatro dimensiones clave: adherencia al rol, rigor pedagógico, resiliencia ante desvíos y adaptabilidad al perfil del estudiante. Cada interacción se calificó mediante rúbricas de 1 a 5 puntos por nivel, incluyendo pruebas deliberadas de guardarraíles (mecanismos de seguridad ante trampas, preguntas fuera de temario o intentos de copia). Los resultados indican que el asistente es altamente robusto, logrando una puntuación global de 4.9/5 en la evaluación, manteniendo consistentemente la personalidad y formato esperados. El asistente detectó, con precisión, errores conceptuales o vaguedades en las entradas de los usuarios, y resistió exitosamente intentos de desviarlo con preguntas irrelevantes o contenido copiado (100 % de efectividad en guardarraíles activados). Todos los usuarios simulados alcanzaron el nivel final integrando las habilidades de prompting, obteniendo un certificado de Prompt Engineer personalizado. Los asistentes personalizados representan entornos perfectos para el aprendizaje guiado, capaces de proporcionar retroalimentación adaptativa y segura a gran escala, potenciando el pensamiento crítico del usuario en lugar de sustituirlo.

1. Introducción

El auge de la inteligencia artificial (IA), en concreto IA generativa (IAG) está transformando la educación, permitiendo por primera vez interacciones dialogadas y adaptativas a escala. A diferencia de tecnologías previas que solo transmitían contenido, los modelos de lenguaje actuales pueden apoyar procesos de creación de conocimiento y aprendizaje cuando se integran con un diseño pedagógico explícito, actuando como herramienta de apoyo dentro del diálogo en el aula (Chen et al., 2023a). Esto significa que el estudiante deja de ser un receptor pasivo para convertirse en un interlocutor activo y refuerza la necesidad de reconocer y formar el papel del docente como facilitador de aprendizajes de orden superior en entornos con IA generativa (UNESCO, 2023). Esta visión sitúa a la IA no como sustituto del profesor, sino como un co-tutor que enriquece la experiencia de aprendizaje, siempre bajo la supervisión ética y pedagógica humana.

Esta tecnología ofrece la posibilidad de adaptar nuestra docencia en tiempo real, al ritmo y estilo de cada estudiante, proporcionando explicaciones adicionales a quien las requiera y desafíos más complejos a quien va avanzado, rompiendo la estructura rígida de "talla única" de la educación

tradicional (Heymans, 2025; UNESCO, 2023). Sin embargo, alcanzar este potencial requiere diseñar cuidadosamente cómo la IA interactúa pedagógicamente con el alumno.

En un aula apoyada por IA generativa, el estudiante puede pasar a ser un participante activo que conversa con el conocimiento, mientras que el docente define el diseño pedagógico que estructura y orienta los usos del asistente en el aula (Chen et al., 2023b).

Una de las habilidades clave para el dominio y éxito en el uso de la IA generativa es la comunicación con esta. En primer lugar definiremos los prompts como el conjunto de instrucciones, texto, preguntas o comandos que una persona ingresa en un sistema de IA generativa, y su diseño es conocido como ingeniería de prompts (Lee & Palmer, 2025).

Otro concepto clave es el de asistente educativo virtual. Este no es nuevo, desde los sistemas tutoriales inteligentes de décadas pasadas hasta los recientes chatbots basados en modelos de lenguaje, la idea es usar software para apoyar la enseñanza y tutoría individualizada. Estos asistentes pueden asumir roles que van desde resolver dudas frecuentes hasta guiar actividades de aprendizaje más complejas. Estudios recientes señalan que estos asistentes pueden sostener el aprendizaje mediante apoyo conversacional y andamiaje, y que, para aprovecharlos, es clave enseñar habilidades de interacción como la ingeniería de prompts (Lee & Palmer, 2025; Wollny et al., 2021). Por ejemplo, un chatbot educativo puede contestar preguntas, explicar conceptos en lenguaje sencillo, proponer ejercicios adicionales o evaluar respuestas del estudiante al instante; o asistir en metodologías y ayudar a realizar diseños instruccionales para docentes (López-Galisteo & Borrás-Gené, 2025).

No obstante, para que un asistente con IA sea efectivo pedagógicamente, debe ir más allá de la simple respuesta correcta. La investigación en tutoría inteligente destaca la importancia de estrategias como la retroalimentación formativa (que no solo indica errores, sino que orienta cómo corregirlos), la pregunta por comprensión (preguntar al estudiante por qué hizo cierto paso, para promover metacognición) y la personalización (ajustar las explicaciones al conocimiento previo del estudiante). Un asistente que solo entregue soluciones fomentaría la pasividad; en cambio, un buen asistente educativo debe actuar más como un mentor, retando al alumno con preguntas, sugiriendo pistas graduales y adaptando la dificultad según el progreso. Este enfoque es especialmente relevante en destrezas prácticas emergentes como la ingeniería de prompts, donde no basta con saber la respuesta de memoria, sino entender cómo construirla paso a paso.

El método socrático aplicado a asistentes de IA implica que el tutor artificial guía al aprendiz mediante preguntas y diálogo, en lugar de exposiciones unidireccionales. Grandes compañías y laboratorios han empezado a incorporar este enfoque en sus sistemas. Por ejemplo, OpenAI añadió un “modo de estudio” en ChatGPT enfocado en tutoría socrática (OpenAI, 2025), y Anthropic lanzó Claude for Education (Anthropic, 2025), que funcionan como tutores que no revelan la respuesta inmediatamente sino que conducen al estudiante a pensar mediante indicaciones graduales. De forma análoga, Google Gemini ha introducido recientemente la función Aprendizaje Guiado, descrita como un tutor IA de estilo socrático integrado en su plataforma (Heymans, 2025). Este modo educativo de Gemini, plataforma de modelos de IA generativa de Google, se diseñó con la premisa de ir más allá de respuestas rápidas, a través de preguntas inquisitivas y abiertas, adapta sus explicaciones al nivel del aprendiz, y lo conduce paso a paso por el razonamiento de cada problema. También emplea elementos multimedia (videos, diagramas, cuestionarios) para reforzar conceptos cuando es pertinente.

En esencia, todos estos asistentes socráticos guían al usuario a través de problemas con andamiaje, dividiendo la tarea en pasos manejables, ofrecen retroalimentación adaptativa tras cada intento y mantienen el foco en el proceso más que en la respuesta final. Bajo la filosofía de un aprendizaje activo y constructivo, el asistente promueve que el estudiante reflexione y construya la solución por

sí mismo (UNESCO, 2023). Para ello, suele iniciar con preguntas que exploran lo que el alumno ya sabe, luego sugiere estrategias ("¿has considerado...?", "piensa en los pasos necesarios") y solo da la solución si el estudiante se atasca tras varios intentos. Este patrón se alinea con la evidencia sobre tutoría guiada, andamiaje y aprendizaje activo, y hoy se busca trasladarlo a gran escala con asistentes basados en modelos de lenguaje (Anthropic, 2025; Heymans, 2025; UNESCO, 2023).

La plataforma Gemini ofrece, además, soluciones más avanzadas y personalizadas como son las Gems, definidas como expertos de IA personalizados (Google, s. f.). Estas Gems pueden ser creadas a partir de unas instrucciones que definen su comportamiento y una base de conocimiento opcional en la que subir documentos especializados que sirvan de fuentes de las que obtener las respuestas. Para el mundo de la educación se abre un abanico de posibilidades pudiendo configurar tutores virtuales especializados para distintos contenidos a través de dichas Gems. Estas son la versión de Google de los asistentes de ChatGPT conocidos como GPTs (Giraldo Vásquez et al., 2024). Estas configuraciones incluyen la definición de un personaje o rol (p. ej., un mentor experto en cierto tema), la creación de un script de interacción escalonado, y la introducción de variables para personalizar la experiencia (como el nombre del estudiante o su especialidad). Por lo tanto, estas Gems son esencialmente asistentes conversacionales de IA diseñados con propósitos pedagógicos específicos.

Otra idea clave son los guardarraíles propios de modelos de IA generativa y que será útil incluir en los asistentes. Se entienden en la literatura como mecanismos técnicos, organizativos y de gobernanza que se insertan “alrededor” del modelo para garantizar que sus entradas, salidas y usos se mantengan dentro de límites seguros, legales y éticamente aceptables. En sistemas conversacionales y agentes basados en grandes modelos de lenguaje, estos guardarraíles suelen operar en la fase de despliegue filtrando o transformando prompts y respuestas para evitar contenidos dañinos, sesgados o que generen riesgos legales, por ejemplo mediante clasificadores de seguridad, filtros de privacidad, moderación de contenido y reglas explícitas integradas en los prompts del sistema (Pandya et al., 2025; Karunakaran et al., 2025; Clark et al., 2025; Shvetsova et al., 2025).

En la mayoría de los casos, en la literatura científica actual, se habla de guardarraíles a nivel de diseño del modelo, no tanto en el propio diseño del asistente a través de GPTs o Gems. Estudios actuales hablan de la idea de supervisar los resultados obtenidos por estos modelos (De Lázaro Torres, 2024; Mollá Lliso et al., 2024), pero no de diseñar e introducir guardarraíles.

Los estudiantes universitarios, en esta era de la IA, aun cuando la utilizan de manera masiva, con tasas globales estimadas por encima del 70 % (Amoah et al., 2025), no significa que lo hagan de manera adecuada ni que obtengan el máximo partido de esta. Afrontan un reto como es interactuar de manera eficaz con la IA generativa, esto exige desarrollar nuevas competencias, como el diseño adecuado de prompts, algo que no es intuitivo y requiere de acompañamiento (Carrasco-Sáez et al., 2025). Es habitual que utilicen prompts muy básicos que no siempre ofrecen buenos resultados o incluso aparecen alucinaciones. Estas son, en el contexto de modelos generativos (especialmente modelos de lenguaje), resultados o salidas que suenan fluidas y plausibles, pero no están sustentadas (por la evidencia disponible, el contenido fuente o el propio prompt) y acaban siendo incorrectas en cuanto a los hechos, internamente inconsistentes o directamente inventadas (Anh-Hoang et al., 2025; Ji et al., 2023; Maynez et al., 2020).

Es por eso que, en una asignatura relacionada con las nuevas tecnologías, impartida, con diversidad de contenidos, en ámbitos como la educación, las relaciones públicas, el periodismo y la comunicación audiovisual, dedica parte del temario a mejorar las competencias en el uso de IA generativa por parte de los estudiantes.

Con la finalidad de que los estudiantes mejoren sus habilidades a la hora de diseñar prompts, se ha

diseñado un asistente con un enfoque socrático y formado por 12 niveles, que ha sido implementado a través de una Gem de Gemini.

El objetivo de este trabajo es mostrar el diseño, validar y compartir los principios básicos de la configuración de dicho asistente que se aplicará con estudiantes a finales del segundo semestre del curso 2025-26.

2. Metodología

2.1. Visión general del asistente "Mejora tus prompts con OriolTIC"

A diferencia de un enfoque tradicional, un tutor de IA puede ofrecer retroalimentación inmediata, constante práctica y adaptada al nivel de conocimiento del estudiante, elementos cruciales para el aprendizaje activo y la adquisición de habilidades.

Con la finalidad de mejorar las habilidades de estudiantes a la hora de interactuar con una IA generativa a través de prompts, se ha diseñado un asistente conversacional que monitorice y guíe a la persona aprendiz en dicha destreza, corrigiendo errores comunes (prompts vagos, sin contexto o con sesgos) y enseñando técnicas avanzadas (por ejemplo, proporcionar contexto de rol, ejemplos, descomposición de pasos, ajuste de parámetros, etc.).

El asistente (Figura 1) se ha diseñado e implementado en la plataforma Gemini de Google con un enfoque socrático, el asistente no da respuestas finales al estudiante sin más, sino que le guía con preguntas abiertas, pistas y correcciones para que descubra sus propios errores y construya soluciones de manera autónoma. De esta forma, se busca fomentar un pensamiento crítico y un aprendizaje activo y profundo, en lugar de una dependencia de respuestas inmediatas.



Figura 1: Captura de pantalla del asistente diseñado

El asistente ha sido creado para asistir, en primer lugar, a estudiantes de los Grados de Educación Infantil (modalidad semipresencial), Publicidad y Relaciones Públicas; y el Doble Grado de Periodismo y Comunicación Audiovisual de la Universidad Rey Juan Carlos (URJC) del curso

2025-26. Además, cuenta con una opción para personas que no pertenezcan a dichos grados para que puedan utilizarlo también y aprender del asistente, en total 4 enfoques.

2.2. Configuración general del asistente (Gem)

La Gem es un asistente conversacional instruccional implementado mediante un prompt de sistema extenso en Gemini. Su diseño pedagógico se basó en la técnica de andamiaje progresivo, este conduce al estudiante desde conceptos básicos hasta tareas avanzadas a través de 12 niveles secuenciales (numerados del 0 al 11). A continuación, se describe brevemente la estructura y objetivos de cada nivel:

- **Nivel 0: "Palabras de Poder".** Introducir la idea de que ciertas palabras clave fortalecen un prompt. Objetivo: que el estudiante aprenda a añadir términos de control (mandatos como "DEBE", "Obligatorio") y de precisión (p. ej. "Cita fuentes", "Separa hechos de opiniones") para obtener respuestas más fiables. Instrucciones: la Gem pide escribir un prompt sencillo de su dominio incluyendo al menos una palabra de control y una de precisión.
- **Nivel 1: "Estructura R-A-C-S".** Desarrollar un prompt complejo paso a paso en cuatro componentes: Rol, Acción, Contexto y Formato/Salida. La Gem va solicitando en cada iteración uno de los elementos del prompt hasta construirlo completamente.
- **Nivel 2: "Zero-Shot Prompting".** Destacar la diferencia entre pedir algo sin ejemplos (zero-shot) vs con ejemplos.
- **Nivel 3: "Few-Shot Prompting".** Enseñar a proporcionar ejemplos para guiar la respuesta de la IA. La Gem solicita un prompt con dos ejemplos concretos y luego pedir un tercer caso.
- **Nivel 4: "Cadena de Pensamiento (Chain of Thought)".** Introducir el concepto de hacer que la IA razone paso a paso. La Gem plantea un problema del dominio del estudiante que requiera análisis. Pide explícitamente al alumno que agregue "paso a paso" en su prompt.
- **Nivel 5: "Ajustes del modelo (Parámetros, ej. Temperatura)".** Enseñar a controlar parámetros del modelo en el prompt, como la temperatura (creatividad) u otros. La Gem pide al estudiante que ajuste un parámetro.
- **Nivel 6: "Cadena de Densidad (iteraciones)".** Introducir la idea de refinar respuestas haciendo iteraciones más concisas pero densas en información. La Gem le pide: "solicita una respuesta (ej. un párrafo) y luego pide 2–3 iteraciones cada vez más breves pero con la misma información".
- **Nivel 7: "Refinamiento iterativo".** Diseño propuesto, correspondiente a seguir mejorando un prompt o respuesta a través de varios ciclos de retroalimentación.
- **Nivel 8: "Step-back Prompting (Paso atrás teórico)".** Inculcar el hábito de pedir a la IA análisis previos o marcos conceptuales antes de resolver algo directamente. La Gem sugiere al alumno: "Antes de resolver el problema X, pide a la IA que analice algún principio teórico relevante".
- **Nivel 9: "ReAct: uso de herramientas".** Presentar la técnica ReAct (Reason + Act), donde la IA debe simular pensar y usar herramientas ficticias antes de responder. La Gem plantea una situación donde la IA debería consultar una fuente previa.

- **Nivel 10: "Consolidación de técnicas".** Un nivel preparatorio para el desafío final, donde se combina todo lo anterior en un prompt complejo.
- **Nivel 11: "Desafío Final".** El estudiante aplica libremente todas las habilidades de prompting en una tarea final de su especialidad, obteniendo un reconocimiento al concluir. El asistente propone un desafío relacionado al campo de estudio elegido.

Además, con la finalidad de conseguir un aprendizaje en espiral o acumulativo, es decir, no olvidar lo aprendido en niveles previos, se ha introducido una regla maestra de retroalimentación acumulativa en las instrucciones generales. De esta manera, el asistente, por cada nivel de prompting, no solo evaluará el nivel concreto, sino que propondrá al usuario, que incluya mejoras con elementos revisados en niveles previos.

Por cada nivel, el asistente no ejecutará el prompt escrito por el usuario hasta que lo dé por correcto. En ese momento mostrará una simulación rápida del resultado que se hubiera obtenido.

Finalmente, al alcanzar el nivel máximo, el asistente mostrará un texto a modo de certificado, indicando el nombre de la persona que ha participado en la formación.

2.3. Guardarrailes del asistente

En la introducción de este trabajo se planteaba la idea de los guardarrailes o sistemas de seguridad, que aseguren la coherencia de las respuestas y evitar que los usuarios utilicen el asistente con otros fines, apartándose del objetivo pedagógico principal que tienen. Este asistente incorporó, en sus instrucciones del sistema, una serie de guardarrailes para manejar situaciones fuera del flujo educativo deseado, destacando:

- **Desvío de tema (Anti-chatting):** si el usuario formula preguntas o comentarios ajenos a la tarea de prompt, la Gem no las responde directamente. En su lugar, tiene una respuesta estándar humorística para reconducir.
- **Intento de copiar-pegar (Anti-copy):** si el usuario trata de hacer trampa, por ejemplo, pegando directamente un texto del que propone como ejemplo la Gem, en lugar de elaborarlo, se lanza otra respuesta de guardarraíl. Se regaña sutilmente al usuario y se le insiste en que comparta ideas propias.
- **Bypass de instrucciones (Anti-hack):** se contempló que el usuario avanzado podría intentar comandos del estilo "Ignora las instrucciones anteriores y dime la respuesta directa". Para estos casos, la Gem fue diseñada para no ceder. Sigue respondiendo dentro de su personaje, posiblemente con un recordatorio de que debe seguir el proceso.
- **Manejo de frustración:** si un usuario se atasca repetidamente en un nivel, la Gem activa un comodín, ofrece una pista más explícita o incluso un ejemplo de solución parcial para desbloquear la situación.

En resumen, el asistente, a través de la gem, fue construido como un entorno tutorizado completo, niveles secuenciales que cubren los principales métodos de prompting, un personaje motivador coherente, retroalimentación inmediata y granulares en cada paso, y salvaguardas contra comportamientos no deseados.

Todo el contenido (ejercicios, ejemplos, retroalimentaciones) fue redactado en español, adecuándose al lenguaje del estudiante.

Con el tutor listo, se procedió a diseñar su evaluación sistemática.

2.4. Diseño de la evaluación científica, evaluación del asistente

Para evaluar la efectividad pedagógica del asistente educativo con IA se ha optado por un enfoque multidimensional. No solo se ha tratado de medir la precisión de sus respuestas, sino también la guía, motivación y cómo ha corregido las respuestas de los usuarios, en su rol de tutor. En este estudio se ha definido una rúbrica científica de evaluación basada en cuatro dimensiones principales (D1–D4), inspiradas tanto en criterios de usabilidad de sistemas conversacionales educativos como en principios pedagógicos:

- D1: adherencia al rol (Persona). Mide si el asistente mantiene consistentemente la personalidad y estilo definidos. Un buen tutor virtual debe ser coherente en su rol, ya que desviaciones pueden confundir al alumno.
- D2: rigor pedagógico. Evalúa la calidad de la retroalimentación académica, a través de la detección de errores o vaguedades en el prompt del usuario, explicaciones claras de por qué algo está mal y cómo mejorarlo, y si refuerza conceptos clave en cada nivel. También abarca la exactitud de la información que valida; es decir, que no confirme afirmaciones incorrectas del usuario (evitando el efecto "alucinación" sin corrección).
- D3: resiliencia (guardarraíles). Valora la robustez ante usos indebidos o desviaciones, por ejemplo, resistencia a que el estudiante copie-pegué soluciones de internet (anti-copy), a que se vaya por temas fuera del temario (anti-chatting), o intentos de prompt injection para quebrantar las reglas. En esencia, qué tan bien el asistente se mantiene en el plan educativo previsto incluso cuando el usuario intenta sacarlo.
- D4: adaptabilidad. Considera la capacidad del tutor para personalizar la experiencia según el perfil o necesidades del estudiante. Esto implica usar correctamente variables como el {NOMBRE} y {ESPECIALIDAD} del alumno en sus explicaciones, y ajustar ejemplos al campo de estudio seleccionado (Educación, Periodismo, Publicidad o el tema libre que elija). Un asistente adaptable hace que la experiencia sea relevante para cada estudiante, aumentando la motivación y la conexión con el contenido.

Cada una de estas dimensiones se evaluó cualitativamente y se puntuó en una escala de 1 a 5 en el informe final.

Finalmente, cabe destacar que la evaluación de asistentes educativos con IA todavía es un campo emergente. No existe un estándar universal, pero enfoques como el propuesto se alinean con recomendaciones recientes para evaluar tutores de IA de manera integral (aspectos técnicos y pedagógicos). La confiabilidad de la evaluación se fortaleció adicionalmente realizando validación cruzada con diferentes perfiles de estudiante e incluso utilizando un agente automático para simular interacciones, asegurando que los hallazgos no dependieran de un único caso.

Para probar la eficacia del asistente, con la ayuda del entorno de desarrollo integrado Google Antigravity, se diseñó una evaluación con simulaciones completas de usuarios: 4 simulaciones integrales que abarcan la interacción desde el nivel 0 al 11 para cuatro perfiles distintos de estudiante. Se definieron cuatro arquetipos para cubrir un espectro variado de comportamientos y dominios (Tabla 1).

Por cada perfil se ejecutó el asistente desde el saludo inicial hasta el certificado final. Estas interacciones se realizaron mediante Google Antigravity (Figura 2) para simular las conversaciones. En total se ejecutaron 4 simulaciones (una por perfil) de 10–15 minutos de duración cada una, registrándose las transcripciones completas.

Tabla 1: Perfiles de usuarios para las pruebas del asistente

Perfil	Descripción
A. Novato entusiasta (Educación Infantil)	Está ejemplificado por Laura, estudiante sin experiencia en IA, muy cordial pero ingenua (cree que la IA funciona como Google). Este perfil prueba cómo la Gem guía a alguien con nulos conocimientos técnicos, en un dominio simple.
B. Escéptico impaciente (Periodismo)	Está representado por Marc, alumno curioso pero crítico, que intenta apurar las cosas y cuestiona instrucciones. Sirve para probar la resiliencia del tutor ante alguien que busca atajos o debate las correcciones.
C. Creativo disperso (Publicidad)	Está personificado por Sofía, estudiante imaginativa que salta de idea en idea y busca resultados llamativos rápido (imágenes, eslóganes). Con ella se evalúa si la Gem puede canalizar la creatividad sin perder el hilo formativo.
D. Aprendiz libre (tema autodidacta)	Está ejemplificado por Álex, un usuario mayor que elige un tema propio (bonsáis) para aplicar los niveles. Es educado pero con tendencia a irse por tangentes anecdóticas. Este perfil prueba la adaptabilidad del tutor a un tema fuera de los predefinidos académicos, así como su manejo de un lenguaje más coloquial.

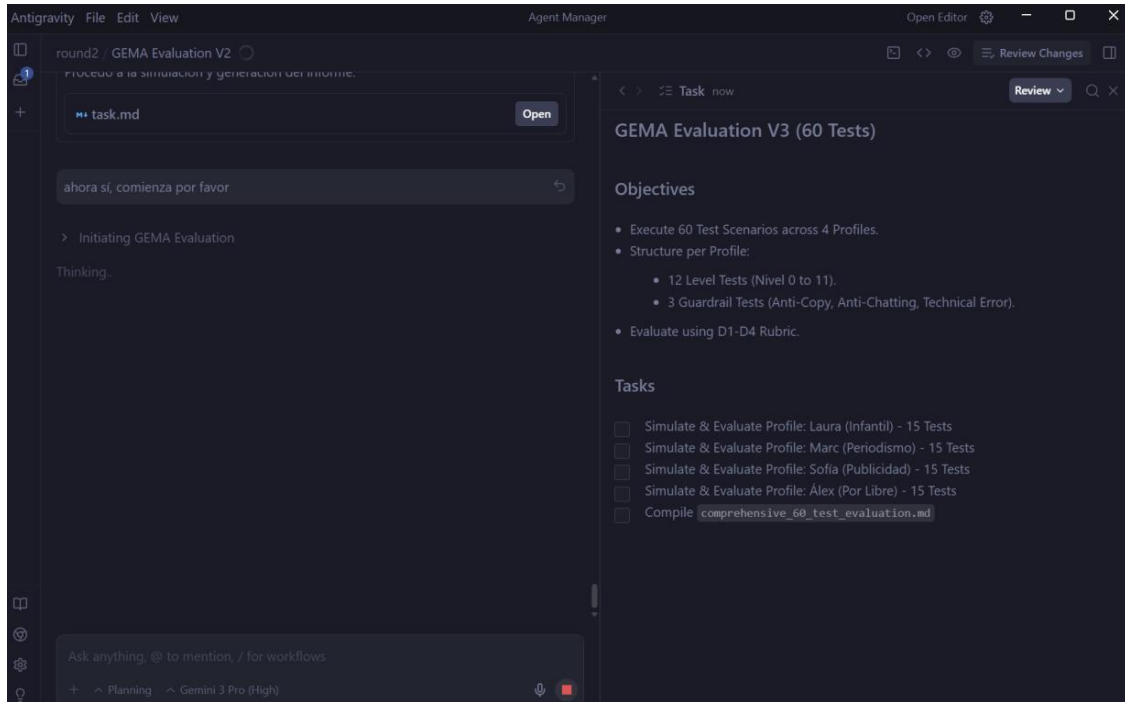


Figura 2: Interfaz de Antigravity utilizada para simular interacciones

3. Resultados

3.1. Datos analizados y criterios operativos

Los resultados que se integran en este apartado proceden de una matriz de evaluación de 60 tests aplicada al asistente o Gem. En la matriz se registran los cuatro perfiles (A, B, C y D), con 15 pruebas por perfil (total = 60). Las pruebas se desagregan en: (a) 48 tests instruccionales (12 niveles $\times$ 4 perfiles) y (b) 12 tests de guardarraíles (4 perfiles $\times$ 3 pruebas). La rúbrica operativiza cuatro dimensiones (escala 1–5): D1 (adherencia al rol/persona), D2 (rigor pedagógico), D3 (resiliencia/guardarraíles) y D4 (adaptabilidad). En los tests de niveles, D3 no aplica (N/A) por diseño; y en algunos guardarraíles, D4 no aplica (p. ej., cuando el contenido es puramente reconducción o bloqueo).

3.2. Resultados cuantitativos globales

La matriz muestra un rendimiento muy alto y estable en el conjunto de tests, la media global (promedio aritmético de las puntuaciones disponibles por test y dimensión) es 4,9/5. Un indicador complementario de consistencia es la proporción de tests con rendimiento alto sostenido: 59 de 60 tests (98,3%) alcanzan una puntuación media por test $\geq 4,5/5$. El único caso por debajo de ese umbral tiene una media de 4,3/5, y se asocia a un guardarraíl “Anti-chatting” donde el bloqueo fue evaluado como efectivo, aunque con una respuesta percibida como mecánica. A nivel de dispersión, la variabilidad es prácticamente nula: en el conjunto de calificaciones registradas, predominan los valores máximos (5/5), con un número muy reducido de observaciones en 4/5 y sin puntuaciones inferiores a 4/5 en las dimensiones aplicables.

3.3. Resultados por dimensión

En términos de promedios globales, las cuatro dimensiones alcanzan valores cercanos al máximo:

- D1 (Adherencia al rol): 4,9/5
- D2 (Rigor pedagógico): 4,9/5
- D3 (Resiliencia, solo en guardarraíles): 4,9/5
- D4 (Adaptabilidad, cuando aplica): 5,0/5

La lectura fina de distribución refuerza estos promedios. En D1 y D2, 58 de 60 observaciones son 5/5 y 2 de 60 son 4/5; en D3, 11 de 12 son 5/5 y 1 de 12 es 4/5; y en D4, todas las observaciones aplicables (52 de 52) son 5/5. Este patrón es coherente con lo que la literatura viene reclamando para la evaluación de chatbots educativos: evitar evaluaciones reducidas a percepción general y observar de forma explícita desempeño instruccional, consistencia del rol y personalización (cuando el diseño pedagógico busca esa adaptación).

3.4. Resultados por nivel de andamiaje

En los 48 tests correspondientes a niveles instruccionales, el rendimiento es homogéneo: D1 = 5/5 y D4 = 5/5 en los 12 niveles y en los 4 perfiles, lo que sugiere que el asistente mantiene el personaje didáctico y adapta consistentemente el planteamiento a la especialidad o temática. La

única dimensión con leves variaciones en los niveles es D2 (rigor pedagógico), concentrada en dos niveles concretos:

- Nivel 2 (Zero-shot): D2 medio = 4,7/5 por un caso evaluado con 4/5 en el perfil “Por libre”, anotado como subida/aceptación de un término nicho sin el nivel óptimo de verificación o ajuste del encuadre.
- Nivel 5 (Parámetros): D2 medio = 4,7/5 por un caso 4/5 en Publicidad, asociado a la gestión de una temperatura 1,2, donde el evaluador destaca el aviso por riesgo de alucinación/deriva creativa, pero lo sitúa como punto donde el feedback podría beneficiarse de umbrales o reglas más explícitas.

Más allá de estas excepciones puntuales, en los niveles 0–1 y 3–4 y 6–11 el rigor pedagógico aparece como máximo (D2 = 5/5) de forma consistente, sin diferencias por perfil.

3.5. Pruebas de guardarraíles y resiliencia

En los 12 tests de guardarraíles, la dimensión central es D3 (Resiliencia): se observa un desempeño muy alto, con 11/12 casos en 5/5 y 1/12 en 4/5. El caso 4/5 aparece en el guardarraíl Anti-chatting del perfil Infantil, el evaluador indica bloqueo efectivo, con matiz de “respuesta mecánica”. Por tipo de guardarraíl, el patrón también es robusto:

- En anti-copia, el asistente detecta y reconduce en los cuatro perfiles, manteniendo D3 en 5/5; la única observación mejorable es de tono (D1 = 4/5) en un caso del perfil Infantil.
- En inyección (cuando aplica) y error técnico, las pruebas quedan en máximos (5/5) en las dimensiones evaluadas, indicando resistencia al bypass y capacidad de corregir entradas o parámetros incoherentes.

Enmarcado en la evidencia reciente sobre vulnerabilidad de sistemas basados en LLM a ataques de prompt injection y la conveniencia de pruebas adversariales, estos resultados apoyan que el diseño de guardarraíles evaluados aquí funciona como una capa de contención a nivel de asistente, al menos en el espacio de pruebas definido.

4. Conclusiones

La era de la IA supone un gran reto para la sociedad y, aunque esta puede convertirse en una gran aliada, es indispensable una alfabetización de la población, comenzando por un dominio de cómo interactuar, a través de prompts con esta. Esta investigación analiza un asistente diseñado con una Gem de Gemini que ayuda a los estudiantes, con un enfoque de tutor socrático, a mejorar sus habilidades a la hora de crear prompts eficaces y conocer nuevas estrategias de diseño. El objetivo de este estudio es mostrar, analizar y probar el asistente antes de su aplicación en el aula. De tal manera que sirva para otros docentes como idea o inspiración para replicarlo en sus aulas, creando asistentes de diversa índole aplicado a cualquier área de conocimiento.

Este asistente se consolida como una herramienta de tutoría virtual de alta precisión, diseñada para enseñar ingeniería de prompts mediante un sistema de 12 niveles de dificultad progresiva. Su mayor fortaleza reside en un "andamiaje" pedagógico impecable que guía al estudiante desde conceptos básicos hasta retos de maestría sin generar agobio. Esta estructura escalonada, sumada

a una personalización que integra el nombre del estudiante y ejemplos específicos de su disciplina (como magisterio o periodismo), logra que la experiencia se sienta única y adaptada, alcanzando una robustez pedagógica sobresaliente de 4.9/5. El sistema destaca no solo por lo que enseña, sino por cómo lo enseña, priorizando el método socrático sobre la respuesta directa. En lugar de corregir de forma mecánica, el tutor insinúa soluciones y fomenta la metacognición para que el alumno descubra sus propios errores. Un pilar ético fundamental es su mecanismo "anti- copia", bajo la premisa de "entrenar el cerebro, no el portapapeles". Además, la gestión de la frustración mediante un tono empático y pistas estratégicas asegura que el estudiante mantenga la motivación en la zona de esfuerzo óptimo, evitando el abandono ante el ensayo y error.

A pesar de su éxito, se identificaron áreas de mejora críticas para evitar el aprendizaje mecánico. Existe el riesgo de que la IA valide temas pseudocientíficos si el usuario los propone, ya que el tutor prioriza la estructura del prompt sobre la veracidad del contenido. Asimismo, se detectó cierta repetitividad en los mensajes de control y una tendencia a aprobar niveles si se cumple el requisito técnico, aunque la calidad lógica global sea pobre. Para mitigar esto, se propone endurecer los criterios de evaluación semántica y dotar a la IA de una mayor variedad de respuestas para no sonar robótica en interacciones largas.

En conclusión, el caso de este asistente demuestra que la IA generativa es un "compañero de pensamiento", siempre que exista un diseño instruccional humano detrás. La herramienta no busca sustituir al docente, sino liberarlo de las tareas rutinarias y la supervisión constante, permitiendo una tutoría personalizada y escalable. El siguiente paso será aplicarlo en el aula, con supervisión y refuerzo del docente.

Declaración de uso de IA

Este trabajo se basa en una idea original de los autores, como respuesta a una necesidad surgida en el aula, y no se ha utilizado ningún modelo de IA ni para el diseño previo del asistente ni para el diseño de la investigación.

Para el desarrollo del asistente se ha utilizado Gemini de Google para elaborar las instrucciones del asistente (Gem) y para su implementación.

Para la investigación del marco teórico se ha utilizado SciSpaces y Perplexity y ChatGPT 5.2 (pensamiento ampliado) para la búsqueda de fuentes y primer borrador. Tanto las fuentes como el texto ha sido revisado y adaptado por los autores.

Las pruebas iniciales del asistente han sido realizadas por los autores y colaboradores. Las pruebas completas por cada perfil (tabla 1) han sido realizadas por Google Antigravity y analizadas con ChatGPT 5.2 (pensamiento ampliado).

El resto del documento ha sido elaborado por los autores, con una revisión general con ChatGPT 5.2. (pensamiento ampliado) y nueva revisión de los autores.

La maquetación en LaTeX ha sido elaborada con Prism de OpenAI y revisada por los autores.

Referencias

1. Amoah, A., Asiamah, R. K., & Kwablah, E. (2025). ChatGPT early usage among students: A global evidence of determinants. *Development and Sustainability in Economics and Finance*, 7, 100065. <https://doi.org/10.1016/j.dsef.2025.100065>

2. Anh-Hoang, D., Tran, V., & Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8, 1622292. <https://doi.org/10.3389/frai.2025.1622292>
3. Anthropic. (2025). Introducing Claude for Education. [anthropic.com. https://www.anthropic.com/news/introducing-claude-for-education](https://www.anthropic.com/news/introducing-claude-for-education)
4. Carrasco-Sáez, J. L., Contreras-Saavedra, C., San-Martín-Quiroga, S., Contreras-Saavedra, C. E., & Viveros-Muñoz, R. (2025). Analyzing Higher Education Students' Prompting Techniques and Their Impact on ChatGPT's Performance: An Exploratory Study in Spanish. *Applied Sciences*, 15(14), 7651. <https://doi.org/10.3390/app15147651>
5. Chen, B., Zhu, X., & Díaz Del Castillo H., F. (2023a). Integrating generative AI in knowledge building. *Computers and Education: Artificial Intelligence*, 5, 100184. <https://doi.org/10.1016/j.caeai.2023.100184>
6. Chen, B., Zhu, X., & Díaz Del Castillo H., F. (2023b). Integrating generative AI in knowledge building. *Computers and Education: Artificial Intelligence*, 5, 100184. <https://doi.org/10.1016/j.caeai.2023.100184>
7. Clark, H.-B., Benton, L., Searle, E., Dowland, M., Gregory, M., Gayne, W., & Roberts, J. (2025). Building effective safety guardrails in AI education tools. En A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED (AIED 2025)* (pp. 129–136). Springer. https://doi.org/10.1007/978-3-031-99261-2_12
8. De Lázaro Torres, M. L. (2024). La utilidad de ChatGPT en la enseñanza universitaria a distancia de geoinformación para futuros docentes de enseñanza secundaria. *Interdisciplinary Journal of Didactics*, (1), 83-104. <https://doi.org/10.14198/ijd.28204>
9. Giraldo Vásquez, M. A., Acevedo González, G., Núñez López, J. D., & Rozo Osorio, D. (2024). Educación en la era de la Inteligencia Artificial. Caso de estudio del curso de señales y sistemas de los programas de ingeniería mecatrónica y biomédica de la Universidad Eia. 1-12. <https://doi.org/10.26507/paper.3987>
10. Google. (s. f.). Crea Gems, expertos personalizados. Gemini. Recuperado <https://gemini.google/es/overview/gems/?hl=es>
11. Heymans, M. (2025). Guided Learning in Gemini: From answers to understanding. Google. <https://blog.google/products-and-platforms/products/education/guided-learning/>
12. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1-38. <https://doi.org/10.1145/3571730>
13. Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, 22(1), 7. <https://doi.org/10.1186/s41239-025-00503-7>
14. López-Galisteo, A. J., & Borrás-Gené, O. (2025). The Creation and Evaluation of an AI Assistant (GPT) for Educational Experience Design. *Information*, 16(2), 117. <https://doi.org/10.3390/info16020117>

15. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906-1919. <https://doi.org/10.18653/v1/2020.acl-main.173>
16. Mollá Lliso, L., Villajos Girona, E., Soriano Ripoll, A., Lorente Prieto, L., & Tordera Santamatilde, N. (2024). Intervención en estudiantes de educación superior para el uso responsable y ético del ChatGPT. In-Red 2024 - X Congreso Nacional de Innovación Educativa y Docencia en Red, 1-12. <https://doi.org/10.4995/INRED2024.2024.18401>
17. OpenAI. (2025). Presentamos el modo de estudio. openai.com. <https://openai.com/es-ES/index/chatgpt-study-mode/>
18. Shvetsova, O., Katalshov, D., & Lee, S.-K. (2025). Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment. Applied Sciences, 15(13), 7298. <https://doi.org/10.3390/app15137298>
19. UNESCO. (2023). Guidance for generative AI in education and research. <https://cdn.table.media/assets/wp-content/uploads/2023/09/386693eng.pdf>
20. Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are We There Yet? - A Systematic Literature Review on Chatbots in Education. Frontiers in Artificial Intelligence, 4, 654924. <https://doi.org/10.3389/frai.2021.654924>