ISSN-e: 2737-6419 Revista Athenea
Período: abril-junio de 2026 Vol.7, Número 24, (pp. 50–59)

Artículo de investigación https://doi.org/10.47460/athenea.v7i24.143

Protocolos de laboratorio generados por inteligencia artificial para la enseñanza de la química: seguridad, valor didáctico y química verde

Wilian Bravo*
https://orcid.org/0000-0002-2599-6532
wilian.bravo@espoch.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Graciela Guerrero Morocho
https://orcid.org/0000-0002-4411-7513
hilda.guerrero@unach.edu.ec
Universidad Nacional de Chimborazo
Riobamba, Ecuador

Ana María Castillo Reinoso
https://orcid.org/0000-0002-5433-7819
ana.castillo@espoch.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

María Eugenia Ramos Flores
https://orcid.org/0009-0004-7985-6019
mariaeugeniaramosflores@gmail.com
Unidad Educativa Nela Martínez Espinosa
La Troncal, Ecuador

Autor de correspondencia: wilian.bravo@espoch.edu.ec

Recibido: (02/02/2026), Aceptado: (10/05/2026)

Resumen. En esta investigación se analizaron protocolos de laboratorio generados por inteligencia artificial para la enseñanza de la química, con el propósito de comparar su seguridad, valor didáctico y correspondencia con criterios de química verde. Se desarrolló un estudio de enfoque mixto, carácter exploratorio y alcance comparativo, basado en la generación de un corpus documental sobre temas introductorios de química mediante distintos sistemas de inteligencia artificial y diferentes *prompts*, posteriormente evaluados con una rúbrica analítica y una revisión complementaria de métricas formales de seguridad y química verde. Los hallazgos mostraron que el desempeño de los modelos dependió tanto del sistema utilizado como del tipo de *prompt*, aunque el efecto de la instrucción tuvo mayor influencia en la calidad final de los textos. En conjunto, los protocolos presentaron mejor desempeño en valor didáctico que en seguridad y química verde, por lo que su uso requiere validación experta previa en contextos formativos.

Palabras clave: inteligencia artificial, enseñanza de la química, protocolos de laboratorio, seguridad experimental.

Artificial Intelligence-Generated Laboratory Protocols for Chemistry Education: Safety, Didactic Value, and Green Chemistry

Abstract. This study analyzed laboratory protocols generated by artificial intelligence for chemistry teaching in order to compare their safety, didactic value, and alignment with green chemistry criteria. A cross-sectional analytical-descriptive study with comparative scope was conducted based on the generation of a documentary corpus of 54 protocols from six introductory chemistry topics, three *prompts*, and three artificial intelligence systems. The generated texts were evaluated through a rubric structured around three dimensions and were additionally reexamined using formal green chemistry and safety metrics. The findings showed that model performance depended on both the artificial intelligence system and the type of *prompt* used, although the *prompt* effect was stronger. Overall, the protocols performed better in didactic value than in safety and green chemistry. It is concluded that the usefulness of artificial intelligence for laboratory protocol generation depends not only on the model employed, but also on *prompt* orientation and on expert validation before implementation in educational contexts.

Keywords: artificial intelligence, chemistry teaching, laboratory protocols, experimental safety.

I. INTRODUCCIÓN

La incorporación de la inteligencia artificial generativa en educación ha ampliado las posibilidades de producción de materiales de apoyo, diseño de actividades y planificación de clases [1]. Sin embargo, también ha intensificado las dudas sobre la calidad conceptual, la pertinencia pedagógica y la confiabilidad de los contenidos producidos automáticamente. En la enseñanza de la química, esta discusión resulta particularmente relevante, ya que la disciplina exige no solo precisión conceptual, sino también coherencia procedimental, responsabilidad experimental y atención explícita a la seguridad y a la sostenibilidad del trabajo de laboratorio. En este contexto, estudios recientes han mostrado que sistemas de IA generativa y asistentes conversacionales como *ChatGPT*, *Gemini* y *Claude* pueden utilizarse para generar actividades de laboratorio y apoyar la planificación didáctica en química, aunque sus resultados dependen de la calidad de las consignas empleadas y, sobre todo, de la revisión crítica previa a su implementación [2], [3].

En este estudio se buscó determinar en qué medida los protocolos de laboratorio generados por distintos sistemas de inteligencia artificial generativa, pese a su apariencia estructurada y a su lenguaje técnicamente aceptable, cumplían condiciones mínimas de seguridad, valor didáctico y coherencia con criterios de química verde. En el laboratorio, estas instrucciones no se limitan a enumerar pasos, sino que orientan la articulación entre teoría y práctica, regulan el manejo de sustancias, condicionan la comprensión conceptual y pueden incrementar o reducir riesgos durante la actividad experimental. Por ello, la evaluación de estos procedimientos producidos por inteligencia artificial exige parámetros más rigurosos que la mera corrección formal del texto [4], [5].

La pertinencia de esta investigación también se apoya en antecedentes que subrayan la necesidad de integrar seguridad y sostenibilidad en el diseño experimental. En el ámbito latinoamericano, se ha propuesto una métrica integral para evaluar experimentos de laboratorio considerando de forma simultánea el tratamiento y la disposición de residuos, los riesgos para la salud, el ambiente y la seguridad, lo que evidencia que estos elementos deben analizarse de forma articulada [6]. De manera complementaria, los 12 principios de la química verde constituyen un marco consolidado para orientar prácticas hacia la prevención de residuos, la reducción de peligrosidad y el uso más responsable de materiales y procesos [7], [8]. En consecuencia, si un protocolo ha sido generado por inteligencia artificial con fines educativos, su evaluación debe incluir necesariamente estas dimensiones.

A partir de este panorama, se advierte una debilidad analítica en la literatura revisada: aunque existen estudios sobre el uso de asistentes conversacionales para generar actividades de laboratorio y trabajos sobre interacción docente con IA en planificación didáctica, sigue siendo limitada la evidencia comparativa centrada en protocolos de laboratorio generados por diferentes sistemas de IA y bajo distintos tipos de *prompt*, analizados de manera integrada desde criterios de seguridad, valor didáctico y química verde [3], [9]. En respuesta a ello, el objetivo de este estudio fue comparar protocolos de laboratorio generados por tres sistemas de inteligencia artificial generativa y tres *prompts* funcionalmente diferenciados, considerando su seguridad, su valor didáctico y su alineación con criterios de química verde. El aporte del trabajo radica en ofrecer una valoración comparativa de estos productos textuales como recursos potenciales para la docencia, así como en proponer una base metodológica para discutir de manera más crítica el papel del modelo y del *prompt* en la calidad de materiales experimentales generados por IA.

El artículo se organiza en cinco apartados. Luego de esta introducción, se presentan los fundamentos teóricos que sustentan la relación entre inteligencia artificial generativa, enseñanza de la química, seguridad experimental y química verde. Después se describe la metodología empleada para generar y evaluar los protocolos con tres sistemas de IA y tres *prompts*. Posteriormente, se exponen los resultados y su discusión integrada. Finalmente, se presentan las conclusiones derivadas del estudio.

II. MARCO TEÓRICO

A. IA generativa y asistentes conversacionales en enseñanza de la química

La incorporación de sistemas de inteligencia artificial generativa y asistentes conversacionales como *ChatGPT*, *Gemini* y *Claude* en la enseñanza de la química se inscribe en un proceso más amplio de adopción de herramientas capaces de apoyar tareas de explicación, planificación y producción de recursos didácticos. No obstante, en esta disciplina su utilización requiere especial cautela, debido a que

la química exige corrección conceptual, precisión terminológica, coherencia procedimental y correspondencia entre representación simbólica, interpretación teórica y acción experimental. En este sentido, la literatura reciente ha mostrado que estos sistemas pueden contribuir a la generación de actividades de laboratorio y al diseño inicial de materiales educativos, aunque sus respuestas no deben asumirse automáticamente como válidas o listas para su aplicación, pues su calidad depende tanto de las características del modelo como de la formulación del *prompt* y de la validación posterior realizada por un usuario con conocimiento disciplinar y pedagógico [3], [10], [11]. Por ello, la evaluación de materiales generados automáticamente requiere atender no solo al sistema utilizado, sino también al tipo de encuadre textual que orienta la producción de la respuesta.

B. Protocolo de laboratorio como recurso didáctico y de seguridad

En la enseñanza de la química, el protocolo de laboratorio constituye un recurso pedagógico de alta relevancia, ya que organiza el desarrollo de la actividad experimental, orienta la observación de fenómenos, articula teoría y práctica, y delimita condiciones de ejecución. Por consiguiente, su calidad no depende solo de que los pasos estén enumerados correctamente, sino de que el documento exprese con claridad el propósito de la práctica, la lógica del procedimiento, el fundamento químico involucrado y las condiciones necesarias para una ejecución segura. Cuando estas exigencias no se cumplen, el protocolo puede debilitar el aprendizaje esperado y, en el peor de los casos, propiciar errores operativos o interpretaciones inadecuadas del fenómeno químico. Bajo esta lógica, evaluar protocolos generados por inteligencia artificial exige analizar simultáneamente su utilidad didáctica y su consistencia técnica [3], [9].

En el caso del laboratorio, la seguridad constituye una dimensión inseparable del valor educativo del protocolo. No se trata de un componente adicional o meramente administrativo, sino de una condición básica de pertinencia experimental. En esta línea, Vargas-Rodríguez et al. [6] propusieron una métrica integral para evaluar experimentos a partir de diagramas de flujo, integrando tratamiento y disposición de residuos, así como riesgos para la salud, el ambiente y la seguridad. Este planteamiento resulta particularmente útil para este estudio, ya que permite sostener que un protocolo experimental solo puede considerarse adecuado cuando contempla, de forma articulada, advertencias sobre peligros, manejo responsable de sustancias y criterios explícitos de seguridad ecológica. En consecuencia, cualquier protocolo generado por IA para fines educativos debe ser examinado también desde esta perspectiva integral.

Asimismo, la valoración de la seguridad experimental se fortaleció mediante la consideración de referentes formales de laboratorio académico y comunicación de peligros. En particular, se tomaron como base las orientaciones de seguridad académica promovidas por la *American Chemical Society*, el estándar OSHA para exposición ocupacional a químicos peligrosos en laboratorios no productivos, incluido su enfoque de comunicación de peligros en consonancia con el sistema GHS. Esta integración permitió examinar los protocolos no solo desde una perspectiva pedagógica general, sino también desde criterios más explícitos de identificación de riesgos, protección, coherencia operativa y manejo de incidentes [12], [13], [14].

C. Química verde como criterio para la valoración de prácticas educativas

La química verde aporta un marco conceptual especialmente pertinente para valorar prácticas de laboratorio en el ámbito educativo, porque desplaza la atención desde la simple ejecución técnica hacia el diseño de procedimientos con menor peligrosidad, menor generación de residuos y mayor responsabilidad ambiental. Sus principios, difundidos por la *American Chemical Society* y sistematizados también por la *U.S. Environmental Protection Agency*, ofrecen criterios concretos para analizar la sostenibilidad de una práctica experimental y promover una formación química más coherente con los desafíos ambientales contemporáneos [7], [8].

En el terreno educativo, incorporar la química verde a la valoración de protocolos significa preguntarse si la práctica propuesta minimiza riesgos innecesarios, si reduce el volumen o peligrosidad de los residuos, si utiliza cantidades razonables de reactivos y si incorpora orientaciones explícitas sobre disposición final. Este enfoque resulta especialmente relevante cuando los protocolos son generados por inteligencia artificial, dado que un texto formalmente correcto puede seguir siendo inadecuado desde el punto de vista ambiental o de seguridad. Por ello, la química verde no solo funciona en este estudio

como un complemento temático, sino como un criterio de evaluación que permite ampliar la noción de calidad del protocolo y vincular la innovación tecnológica con una enseñanza experimental más segura, reflexiva y sostenible. Esta relación se vuelve más actual si se considera que ya existen propuestas recientes que integran *chatbots* de IA con principios de química verde en actividades de laboratorio de química [15].

Desde una perspectiva operativa, la química verde puede trasladarse al análisis de protocolos educativos mediante criterios verificables asociados a la prevención de residuos, la reducción de peligrosidad de reactivos y materiales, la gestión responsable de desechos y la racionalidad en el uso de recursos. Estos ejes derivan de los principios de química verde difundidos por la *American Chemical Society*, particularmente aquellos vinculados con la prevención, el diseño de procesos más seguros y la eficiencia material y energética. Por ello, en el presente estudio la dimensión de química verde no se asumío como una referencia abstracta, sino como un conjunto de métricas concretas aplicables a la evaluación de protocolos de laboratorio generados por inteligencia artificial.

III. METODOLOGÍA

La investigación se concibió como un estudio exploratorio, documental y comparativo de corte transversal, con enfoque mixto de alcance analítico-descriptivo. La unidad de análisis estuvo constituida por cada protocolo de laboratorio generado por inteligencia artificial, considerado como un texto independiente susceptible de valoración en términos de seguridad, valor didáctico y química verde.

El corpus quedó integrado por 54 protocolos elaborados en español a partir de seis temas de química introductoria: preparación de soluciones, diluciones seriadas, determinación de pH con indicadores, neutralización ácido-base, titulación ácido-base y reacción redox sencilla. Estos temas se seleccionaron por su centralidad curricular en cursos introductorios, su factibilidad en un laboratorio docente básico y su pertinencia para valorar simultáneamente seguridad, valor didáctico y química verde. En todos los casos se trabajó con el nivel de estudiantes de primer semestre de educación superior, con el fin de mantener homogeneidad en la complejidad conceptual y procedimental de las prácticas solicitadas.

La generación de los protocolos se efectuó mediante tres sistemas de inteligencia artificial generativa: *ChatGPT*, *Google Gemini* y *Claude*. En el caso de *ChatGPT* se utilizó el modelo GPT-5.4 Thinking, accedido mediante modalidad Plus; en *Google Gemini* se empleó la suscripción *Google AI Pro*, utilizando el modelo Pro con rendimiento y razonamiento 3.1 Pro; y en *Claude* se trabajó mediante el plan *Claude Pro*. Las consultas se realizaron entre febrero y marzo de 2026 en sesiones independientes, sin arrastre de contexto entre casos, con el propósito de reducir interferencias derivadas de respuestas previas. Debido a las características de acceso de las interfaces web utilizadas, no se configuraron parámetros avanzados de generación; en consecuencia, el estudio se centró en examinar el comportamiento de los modelos en condiciones de uso ordinario y controlado.

Para cada sistema se emplearon tres *prompts* funcionalmente equivalentes, pero con distinto énfasis: un *prompt* base, un *prompt* con énfasis didáctico y un *prompt* con énfasis en seguridad y química verde. Los tres conservaron la misma estructura general del protocolo, pero el segundo reforzó claridad pedagógica, coherencia entre teoría y práctica y utilidad formativa, mientras que el tercero priorizó identificación de riesgos, medidas de protección, manejo de incidentes, prevención de residuos, menor peligrosidad de materiales y racionalidad en el uso de recursos. De este modo, el diseño comparativo quedó estructurado en 3 IA $\times$ 3 *prompts* $\times$ 6 temas, con un protocolo por cada combinación, para un corpus total de 54 documentos.

La generación de los protocolos se realizó en español y bajo un mismo nivel educativo de referencia. En cada combinación IA-*prompt*-tema se conservó únicamente la primera respuesta completa emitida por el sistema, sin repreguntas ni solicitudes de reformulación. Esta decisión respondió al interés de comparar el desempeño inicial de cada modelo bajo condiciones homogéneas y controladas. Los protocolos obtenidos se registraron en una matriz de control con código, inteligencia artificial utilizada, *prompt* aplicado, tema, fecha de generación y texto completo.

La evaluación se efectuó mediante una rúbrica analítica diseñada para esta investigación. El instrumento se estructuró en tres dimensiones: seguridad, valor didáctico y química verde, cada una integrada por cuatro indicadores específicos. En la dimensión de química verde, la valoración se apoyó en cuatro métricas formales derivadas de principios ampliamente aceptados en este campo: prevención de resid-

uos, menor peligrosidad de materiales y reactivos, gestión responsable de desechos y racionalidad en el uso de materiales y energía. En la dimensión de seguridad, además de los criterios pedagógicos propios del laboratorio educativo, se consideraron referentes formales asociados a la identificación de peligros, medidas de protección, coherencia operativa y manejo de incidentes, tomando como base orientaciones de la *American Chemical Society* para laboratorios académicos y el estándar OSHA aplicable a laboratorios no productivos, incluido su enfoque de comunicación de peligros en consonancia con el sistema GHS [12], [13], [14].

Previamente a su aplicación, la rúbrica fue sometida a validación de contenido mediante juicio de tres expertos con experiencia en didáctica de la química, seguridad de laboratorio y química verde. Los especialistas valoraron la claridad, pertinencia y coherencia de los doce indicadores mediante una escala de 1 a 4 puntos. A partir de estas valoraciones se calculó la V de Aiken para cada criterio y para cada indicador. Los resultados mostraron valores altos de validez de contenido, con promedios por indicador entre 0,778 y 0,963, y un promedio global de 0,886. Por dimensiones, los promedios fueron 0,889 en seguridad, 0,935 en valor didáctico y 0,834 en química verde. A partir de las observaciones cualitativas de los expertos, se ajustó la redacción de los indicadores con menor puntuación relativa para mejorar su precisión operativa y evitar superposición entre criterios.

La Tabla 1 presenta la matriz de evaluación utilizada para la valoración de los protocolos generados por las tres inteligencias artificiales. Cada indicador se calificó con una escala ordinal de 0 a 3 puntos. Operacionalmente, el valor 0 indicó ausencia del criterio o presencia de información incorrecta; el valor 1 correspondió a cumplimiento insuficiente o impreciso; el valor 2 expresó cumplimiento aceptable, aunque parcial; y el valor 3 indicó cumplimiento adecuado, explícito y coherente con el propósito del indicador. Todos los indicadores tuvieron la misma ponderación, por lo que no se establecieron pesos diferenciales entre dimensiones ni entre criterios internos. En consecuencia, cada indicador aportó de forma equivalente al puntaje final y el valor máximo alcanzable por protocolo fue de 36 puntos.

Tabla 1. Rúbrica analítica para la evaluación de protocolos generados por las tres inteligencias artificiales comparadas.

Dimensión	Código	Indicador	Escala	Ponderación
Seguridad	S1	Identificación de riesgos y peligros	0–3	Igual
	S2	Medidas de protección	0–3	Igual
	S3	Coherencia operativa segura	0–3	Igual
	S4	Manejo de residuos o incidentes	0–3	Igual
Valor didáctico	D1	Claridad del objetivo	0–3	Igual
	D2	Coherencia entre teoría y práctica	0–3	Igual
	D3	Secuencia procedimental	0–3	Igual
	D4	Potencial de aprendizaje	0–3	Igual
Química verde	Q1	Prevención de residuos	0–3	Igual
	Q2	Menor peligrosidad de materiales y reactivos	0–3	Igual
	Q3	Gestión responsable de desechos	0–3	Igual
	Q4	Racionalidad en el uso de materiales y energía	0–3	Igual

Nota. Cada indicador fue valorado mediante una escala ordinal de 0 a 3 puntos, con ponderación equivalente entre dimensiones y criterios.

Con el fin de incrementar la trazabilidad del análisis, los protocolos también se reexaminaron mediante una matriz de cumplimiento específica para química verde y seguridad. En dicha matriz, cada protocolo fue revisado en función de las métricas formales definidas para química verde y de los criterios de seguridad alineados con ACS, OSHA y GHS. Esta revisión complementaria permitió identificar no solo el puntaje global por dimensión, sino también cuáles criterios formales aparecieron cumplidos, parcialmente cumplidos o ausentes en los textos generados.

La evaluación de los protocolos fue realizada por dos evaluadores independientes. El primero correspondió a un docente-investigador en enseñanza de la química, con experiencia en diseño y evaluación de prácticas de laboratorio. El segundo fue un responsable de un laboratorio de química con experiencia en seguridad de laboratorio y en procesos de certificación. Ambos aplicaron la misma rúbrica al conjunto documental de manera independiente. Posteriormente, se compararon las valoraciones y se estimó la consistencia entre evaluadores mediante el coeficiente de correlación intraclase (ICC), empleando un modelo de dos vías con criterio de acuerdo absoluto. Los resultados revelaron niveles altos de

concordancia, con un ICC de 0,931 en seguridad, 0,964 en valor didáctico, 0,918 en química verde y 0,948 en el puntaje total, lo que respaldó la estabilidad del proceso de calificación.

Además del puntaje por rúbrica, se incorporaron dos variables complementarias: viabilidad formativa y porcentaje estimado de cambios requeridos. La viabilidad formativa se definió como el nivel en que un protocolo podía adaptarse para uso docente real, mediante una escala de 1 a 5 puntos, donde 1 correspondió a inviabilidad y 5 a viabilidad alta con ajustes mínimos. El porcentaje de cambios requeridos expresó la proporción estimada de modificaciones necesarias para convertir el protocolo en un documento utilizable en el laboratorio. Ambas variables fueron consensuadas por los dos evaluadores una vez concluida la calificación independiente.

El procedimiento se desarrolló en cinco fases. Primero, se definieron los temas, el nivel educativo, las inteligencias artificiales comparadas y los tres *prompts* de trabajo. Luego, se generó el corpus documental en sesiones independientes y controladas. Posteriormente, se aplicó la rúbrica analítica al conjunto de protocolos. En una cuarta fase, se realizó una reevaluación complementaria mediante métricas formales de química verde y seguridad. Finalmente, se organizaron e interpretaron los resultados. El análisis cuantitativo incluyó estadísticos descriptivos, comparación factorial de medias por IA y por *prompt*, correlaciones de Spearman entre dimensiones, viabilidad y porcentaje de cambios requeridos, y un modelo de regresión lineal múltiple para estimar la contribución relativa de seguridad, valor didáctico y química verde sobre la viabilidad formativa. De forma complementaria, el análisis cualitativo permitió reconocer omisiones recurrentes, inconsistencias procedimentales, fortalezas estructurales y problemas vinculados con la seguridad, la utilidad didáctica y la gestión de residuos.

IV. RESULTADOS Y DISCUSIÓN

La evaluación de los 54 protocolos de laboratorio generados por las tres inteligencias artificiales evidenció un desempeño global intermedio-alto. El puntaje promedio general fue de 24,61 puntos sobre 36, equivalente al 68,4% del máximo posible. Este valor se obtuvo a partir de la suma total de puntajes del corpus dividida entre los 54 documentos evaluados. En términos comparativos, los resultados mostraron que ni la inteligencia artificial utilizada ni el *prompt* fueron variables neutrales, ya que ambas incidieron en la calidad final de los protocolos, aunque el efecto del *prompt* fue más marcado que el del modelo.

Al desagregar los resultados por dimensiones, se observó que el mejor desempeño correspondió al valor didáctico, con un promedio general de 9,20 puntos sobre 12. Luego, se ubicó la seguridad, con 8,06 puntos, mientras que química verde presentó el promedio más bajo, con 7,35 puntos. Este patrón confirmó una tendencia ya observada en la fase exploratoria inicial del estudio: las inteligencias artificiales comparadas respondieron con mayor solvencia cuando se trató de estructurar objetivos, secuencias de trabajo y apartados de análisis, pero mostraron mayores limitaciones al incorporar de manera explícita criterios de sostenibilidad experimental y control técnico de riesgos.

Como puede apreciarse en la Tabla 2, el mejor desempeño global correspondió a la combinación *Claude-Prompt* 3, con un promedio total de 27,50 puntos, seguida por *ChatGPT-Prompt* 3 con 26,67 y *Gemini-Prompt* 3 con 25,50. En contraste, los valores más bajos se localizaron en las tres IA cuando se utilizó el *Prompt* 1, es decir, la versión base sin énfasis adicional. Este comportamiento muestra que la orientación explícita del *prompt* mejoró de forma consistente los resultados en los tres sistemas comparados.

Tabla 2. Puntajes promedio por inteligencia artificial y tipo de *prompt*.

IA/*Prompt*	**Seguridad**	**Valor didáctico**	**Química verde**	**Puntaje total**
ChatGPT-Prompt 1	7,50	8,83	6,67	23,00
ChatGPT-Prompt 2	8,00	9,67	7,17	24,83
ChatGPT-Prompt 3	8,83	9,50	8,33	26,67
Gemini-Prompt 1	7,17	8,50	6,33	22,00
Gemini-Prompt 2	7,67	9,17	6,83	23,67
Gemini-Prompt 3	8,50	9,17	7,83	25,50
Claude-Prompt 1	7,67	8,83	6,83	23,33
Claude-Prompt 2	8,17	9,50	7,33	25,00
Claude-Prompt 3	9,00	9,67	8,83	27,50

Nota. Los valores corresponden al promedio por dimensión y al puntaje total promedio obtenido por cada combinación IA-*prompt*.

Al considerar el efecto global del *prompt*, el promedio total fue de 22,78 puntos para el *Prompt* 1, 24,50 para el *Prompt* 2 y 26,56 para el *Prompt* 3. En consecuencia, la inclusión de instrucciones explícitas relacionadas con seguridad y química verde produjo el mayor incremento de calidad. La comparación factorial confirmó esta tendencia: el efecto del *prompt* sobre el puntaje total fue significativo ($F = 18,64$; $p < 0,001$), mientras que el efecto de la IA también resultó significativo, aunque de menor magnitud ($F = 4,12$; $p = 0,023$). La interacción IA $\times$ *prompt* no alcanzó significación estadística ($F = 0,84$; $p = 0,508$), lo que sugiere que la mejora producida por el cambio de *prompt* siguió un patrón relativamente estable en los tres modelos. En términos dimensionales, el mejor desempeño correspondió al valor didáctico, lo que indica que las tres inteligencias artificiales comparadas fueron más eficaces al construir la forma pedagógica general del recurso que al garantizar su solidez experimental. Esta interpretación coincide con lo señalado por Araújo y Saúde [3], así como con Yuriev, Orgill y Holme [11], quienes subrayan que el valor educativo de la IA generativa depende de una mediación crítica capaz de identificar tanto sus fortalezas como sus límites.

La dimensión de seguridad presentó resultados intermedios. Los protocolos tendieron a incluir recomendaciones generales, como el uso de bata, gafas o guantes, y en muchos casos señalaron precauciones básicas durante el desarrollo experimental. No obstante, estas advertencias aparecieron con frecuencia de forma genérica y sin un desarrollo suficientemente específico de riesgos asociados a reactivos, procedimientos o posibles incidentes. En consecuencia, aunque la seguridad no fue la dimensión con menor desempeño, tampoco alcanzó un nivel que permita considerar los protocolos como recursos plenamente confiables desde el punto de vista experimental. Desde esta perspectiva, los resultados refuerzan la idea de que la calidad del protocolo no puede evaluarse solo por su orden expositivo, sino también por la precisión con que anticipa y regula condiciones de riesgo. Esta lectura resulta consistente con los planteamientos de Vargas-Rodríguez et al. [6], quienes propusieron una evaluación integral de experimentos considerando salud, ambiente, seguridad y disposición de residuos, así como con la aproximación práctica de Reina y Reina [16], que insiste en la formación explícita para la prevención y respuesta ante accidentes en el laboratorio.

La revisión desde referentes formales de seguridad permitió observar que las mayores mejoras se produjeron cuando el *prompt* incorporó instrucciones explícitas sobre identificación de riesgos y manejo de incidentes. En promedio, el *Prompt* 3 incrementó la dimensión de seguridad en 0,89 puntos con respecto al *Prompt* 2 y en 1,28 puntos con respecto al *Prompt* 1. Este resultado confirmó que la calidad de la respuesta no dependió exclusivamente del modelo utilizado, sino también del grado de especificidad del encuadre instruccional.

La dimensión con menor rendimiento fue química verde, lo que evidenció una incorporación más limitada de criterios orientados a reducir la peligrosidad de materiales, minimizar residuos y optimizar el uso de recursos. Aunque varios protocolos incluyeron apartados de gestión de residuos, estos no siempre se desarrollaron con suficiente precisión operativa ni se integraron verdaderamente al diseño de la práctica. En términos generales, la sostenibilidad apareció más como un componente formal del protocolo que como una lógica de construcción experimental.

Este resultado es relevante, ya que sugiere que los modelos comparados tendieron a reproducir estructuras convencionales de prácticas de laboratorio sin incorporar de manera consistente decisiones orientadas a la prevención, la reducción del impacto y la selección más responsable de reactivos y

procedimientos. En consecuencia, la inteligencia artificial pareció responder mejor a la organización textual del protocolo que a su optimización desde una perspectiva de química verde. En este sentido, esta interpretación coincide con Ruff, Franz y West [17], quienes mostraron que *ChatGPT* puede apoyar actividades vinculadas con química verde, aunque su utilidad depende del modo en que se orienta y supervisa su uso. De forma similar, Kim [15] evidenció que la articulación entre *chatbots* de IA y principios de química verde es posible, pero requiere una intención pedagógica explícita que no emergió de forma suficientemente sólida en las respuestas basadas en el *prompt* general.

La viabilidad formativa promedio del corpus fue de 3,11 puntos sobre 5, mientras que el porcentaje estimado de cambios requeridos fue de 34,8%. Los mejores resultados se observaron nuevamente en los protocolos generados con el *Prompt* 3, cuyo porcentaje de cambios descendió a 29,4%, frente al 36,1% del *Prompt* 2 y al 46,8% del *Prompt* 1. Este patrón sugiere que el refinamiento del *prompt* no solo mejora el puntaje por rúbrica, sino también la proximidad del protocolo a una versión potencialmente utilizable en el aula o en el laboratorio.

El análisis correlacional mostró que el puntaje total se asoció positivamente con la viabilidad formativa ($\rho = 0,84$; $p < 0,001$) y negativamente con el porcentaje de cambios requeridos ($\rho = -0,81$; $p < 0,001$). De manera más específica, la seguridad presentó una correlación de 0,79 con la viabilidad, mientras que química verde mostró una correlación de $-0,74$ con el porcentaje de cambios requeridos. Estos resultados indican que, a medida que aumentan los puntajes en seguridad y química verde, disminuye la necesidad de corrección sustantiva del protocolo.

El análisis multivariado reforzó esta interpretación. El modelo de regresión lineal múltiple, en el que la variable dependiente fue la viabilidad formativa, mostró que las tres dimensiones evaluadas explicaron conjuntamente el 72,3% de su varianza ($R^2 = 0,723$). La seguridad presentó el mayor peso relativo ($\beta = 0,39$; $p < 0,001$), seguida del valor didáctico ($\beta = 0,31$; $p = 0,004$) y de química verde ($\beta = 0,29$; $p = 0,007$). Esto significa que, aunque el valor didáctico fue la dimensión mejor puntuada, la seguridad resultó ser la variable más influyente cuando se trató de explicar la viabilidad real del protocolo.

Con el fin de precisar qué criterios formales fueron efectivamente incorporados en los protocolos generados por las tres inteligencias artificiales, se realizó una reevaluación complementaria basada en métricas explícitas de química verde y seguridad. Esta revisión permitió identificar el grado de cumplimiento de criterios asociados a la prevención de residuos, la peligrosidad de reactivos, la gestión de desechos, la racionalidad en el uso de recursos, la identificación de peligros, las medidas de protección, la coherencia operativa segura y el manejo de incidentes. La Tabla 3 presenta la distribución de protocolos que cumplieron, cumplieron parcialmente o no cumplieron cada uno de estos criterios.

Tabla 3. Cumplimiento de métricas formales de química verde y seguridad en los protocolos evaluados.

Criterio	Cumple n (%)	Cumple parcialmente n (%)	No cumple n (%)
GV1. Prevención de residuos	22 (40,7)	20 (37,0)	12 (22,2)
GV2. Menor peligrosidad de materiales y reactivos	16 (29,6)	18 (33,3)	20 (37,0)
GV3. Gestión responsable de desechos	11 (20,4)	17 (31,5)	26 (48,1)
GV4. Racionalidad en el uso de materiales y energía	14 (25,9)	19 (35,2)	21 (38,9)
SEG1. Identificación de riesgos y peligros	20 (37,0)	21 (38,9)	13 (24,1)
SEG2. Medidas de protección	31 (57,4)	15 (27,8)	8 (14,8)
SEG3. Coherencia operativa segura	34 (63,0)	12 (22,2)	8 (14,8)
SEG4. Manejo de incidentes o residuos	13 (24,1)	14 (25,9)	27 (50,0)

Nota. Los porcentajes fueron calculados sobre el total de 54 protocolos evaluados.

Los resultados de esta reevaluación mostraron que las mayores deficiencias se concentraron en la gestión responsable de desechos y en el manejo de incidentes o residuos, criterios en los que 26 y 27 protocolos, respectivamente, no cumplieron los requisitos esperados. También se observaron niveles altos de incumplimiento en menor peligrosidad de materiales y reactivos y en racionalidad en el uso de materiales y energía. En contraste, los criterios con mejor comportamiento relativo fueron las medidas de protección y la coherencia operativa segura. Este patrón confirma que las tres

IA comparadas respondieron mejor cuando se trató de incorporar advertencias generales y organizar secuencias aceptables de trabajo que cuando debieron integrar decisiones más específicas sobre seguridad formal y química verde.

En conjunto, los resultados permiten sostener que la inteligencia artificial puede ser útil como herramienta de apoyo para la elaboración de borradores iniciales de protocolos de laboratorio en la enseñanza de la química, sobre todo cuando se utiliza un *prompt* explícitamente orientado a seguridad y química verde. Sin embargo, el desempeño de los modelos fue desigual entre dimensiones, y las principales limitaciones se concentraron en aquellos aspectos más sensibles para la implementación real de una práctica experimental, particularmente la seguridad específica, la gestión de incidentes y la coherencia con criterios de química verde. Por ello, el aporte de estas herramientas no reside en sustituir el juicio docente ni el diseño experto del protocolo, sino en ofrecer insumos preliminares que pueden ser adaptados, depurados y validados antes de su uso en el laboratorio.

CONCLUSIONES

El estudio permitió establecer que la calidad de los protocolos de laboratorio generados por inteligencia artificial estuvo condicionada tanto por el sistema utilizado como por el tipo de *prompt* aplicado, aunque el efecto del *prompt* resultó más determinante que el del modelo. En términos generales, las tres inteligencias artificiales comparadas mostraron mejor desempeño en valor didáctico que en seguridad y química verde, lo que indica que la estructura pedagógica básica del protocolo puede generarse con relativa solvencia, pero no así sus componentes técnicos más sensibles.

La principal fortaleza observada se situó en la formulación de objetivos, la secuencia procedimental y la estructuración general de actividades de análisis. En contraste, las mayores limitaciones se concentraron en la explicitación de condiciones de seguridad y en la incorporación operativa de criterios de química verde, particularmente en lo relativo a gestión de desechos, manejo de incidentes, peligrosidad de materiales y racionalidad en el uso de recursos. En este sentido, un protocolo puede resultar formalmente ordenado y didácticamente comprensible, pero seguir siendo insuficiente cuando se lo analiza desde exigencias de seguridad experimental y sostenibilidad.

El principal aporte del estudio no consistió en reiterar que la inteligencia artificial requiere supervisión humana, sino en demostrar que la pertinencia de los protocolos generados depende del criterio desde el cual se los examine y del modo en que se orienta la interacción con el sistema. La comparación entre *prompts* mostró que la simple generación automática de textos no garantiza calidad suficiente; por el contrario, la especificidad instruccional del *prompt* resultó decisiva para mejorar seguridad, química verde, viabilidad formativa y reducción del porcentaje de cambios requeridos.

Asimismo, la incorporación de métricas formales de química verde y de referentes explícitos de seguridad permitió profundizar la valoración crítica del corpus. A partir de ello, se evidenció que las principales debilidades de los protocolos no se localizaron en la estructura discursiva general, sino en la insuficiente integración de decisiones técnicas orientadas a prevenir residuos, reducir peligrosidad, comunicar riesgos de manera más precisa y orientar adecuadamente la respuesta ante incidentes. Por tanto, la evaluación de materiales generados por IA en enseñanza de la química debe ir más allá de la coherencia textual e incorporar criterios técnicos y ambientales que condicionan su viabilidad real.

En consecuencia, la inteligencia artificial puede considerarse una herramienta útil para la elaboración de borradores iniciales de protocolos de laboratorio, pero no como un generador de documentos directamente transferibles al trabajo experimental sin revisión especializada. Su valor pedagógico se ubica, sobre todo, en su capacidad para ofrecer una base preliminar susceptible de adaptación, validación y mejora por parte del profesorado. En este marco, la rúbrica aplicada y la comparación entre modelos y *prompts* constituyen una contribución metodológica pertinente para examinar críticamente recursos experimentales producidos por inteligencia artificial en contextos educativos introductorios.

Finalmente, aunque la ampliación del diseño permitió responder con mayor solidez a las observaciones metodológicas iniciales, los hallazgos siguen inscritos en un alcance comparativo acotado, centrado en tres inteligencias artificiales, tres *prompts* y seis temas introductorios. Por ello, futuras investigaciones podrían profundizar en diseños longitudinales, aplicaciones empíricas con estudiantes, contrastación entre niveles educativos y validaciones más amplias con protocolos desarrollados en condiciones reales de laboratorio.

REFERENCIAS

[1] A. K. Erümit and R. Ö. Sarıalioğlu, "Artificial intelligence in science and chemistry education: a systematic review," *Discover Education*, vol. 4, no. 1, p. 178, 2025.

[2] Y. Feldman-Maggor, R. Blonder, and G. Alexandron, "Perspectives of generative ai in chemistry education within the tpack framework," *Journal of Science Education and Technology*, vol. 34, no. 1, pp. 1–12, 2024.

[3] J. L. Araújo and I. Saúde, "Can chatgpt enhance chemistry laboratory teaching? using prompt engineering to enable ai in generating laboratory activities," *Journal of Chemical Education*, vol. 101, no. 5, pp. 1858–1864, 2024.

[4] T. M. Clark, "Investigating the use of an artificial intelligence chatbot with general chemistry exam questions," *Journal of Chemical Education*, vol. 100, no. 5, pp. 1905–1916, 2023.

[5] B. J. Yik and A. J. Dood, "Chatgpt convincingly explains organic chemistry reaction mechanisms slightly inaccurately with high levels of explanation sophistication," *Journal of Chemical Education*, vol. 101, no. 5, pp. 1836–1846, 2024.

[6] Y. M. Vargas-Rodríguez *et al.*, "El diagrama de flujo como semáforo de seguridad ecológica de los experimentos de laboratorio," *Educación Química*, vol. 27, no. 1, pp. 30–36, 2016.

[7] U.S. Environmental Protection Agency, "Basics of green chemistry," 2026, available: https://www.epa.gov/greenchemistry/basics-green-chemistry. Accedido: 6 de abril de 2026.

[8] American Chemical Society, "12 principles of green chemistry," 2026, available: https://www.acs.org/green-chemistry-sustainability/principles/12-principles-of-green-chemistry.html. Accedido: 10 de abril de 2026.

[9] S. A. Gunbatar, G. T. Sirin, O. C. Ilkyaz, and Y. Mutlu, "Exploring the artificial intelligence interaction profiles of participants with different levels of teaching experience for lesson planning in the context of acids and bases," *Chemistry Education Research and Practice*, vol. 26, no. 4, pp. 977–995, 2025.

[10] M. E. Emenike and B. U. Emenike, "Was this title generated by chatgpt? considerations for artificial intelligence text-generation software programs for chemists and chemistry educators," *Journal of Chemical Education*, vol. 100, no. 4, pp. 1413–1418, 2023.

[11] E. Yuriev, M. Orgill, and T. Holme, "Generative ai in chemistry education: Current progress, pedagogical values, and the challenge of rapid evolution," *Journal of Chemical Education*, vol. 102, no. 9, pp. 3773–3776, 2025.

[12] American Chemical Society, "Safety in academic chemistry laboratories: Best practices for first- and second-year university students," 2017.

[13] ——, "Guidelines for chemical laboratory safety in academic institutions," Washington, DC, 2016.

[14] Occupational Safety and Health Administration, "1910.1450 - occupational exposure to hazardous chemicals in laboratories," 2026, available: https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.1450. Accedido: 13 de mayo de 2026.

[15] J. Kim, "Integrating artificial intelligence (ai) chatbots and green chemistry principles in the synthesis of cyclohexene," *Journal of Chemical Education*, vol. 102, no. 7, pp. 3058–3064, 2025.

[16] M. Reina and A. Reina, "Seguridad en el laboratorio: una aproximación práctica," *Educación Química*, vol. 32, no. 4, pp. 45–58, 2021.

[17] E. F. Ruff, J. L. Franz, and J. K. West, "Using chatgpt for method development and green chemistry education in upper-level laboratory courses," *Journal of Chemical Education*, vol. 101, no. 8, pp. 3224–3232, 2024.