https://doi.org/10.69639/arandu.v13i3.2598

Confianza y juicio técnico frente a la inteligencia artificial en la formación en Ingeniería en Alimentos

Trust and Technical Judgment Regarding Artificial Intelligence in Food Engineering Education

Jonathan Alexis Tapia Chamba
jtapiac3@uteq.edu.ec
https://orcid.org/0009-0001-0121-9923
Universidad Técnica Estatal de Quevedo
Quevedo – Ecuador

Odalis Kassandra Gonzalez Duarte
ogonzalezd@uteq.edu.ec
https://orcid.org/0009-0002-6990-2689
Universidad Técnica Estatal de Quevedo
Quevedo – Ecuador

Jessica Lisbeth Parrales Cedeño
jparralesc4@uteq.edu.ec
https://orcid.org/0009-0004-9333-3687
Universidad Técnica Estatal de Quevedo
Quevedo – Ecuador

Bella Elena Chamba Palacios
bella.chamba@educacion.gob.ec
https://orcid.org/0009-0009-7623-3787
Escuela de Educación Básica "Destacamento Banderas"
Orellana – Ecuador

Artículo recibido: 10 agosto 2026- Aceptado para publicación:16 septiembre 2026
Conflictos de intereses: Ninguno que declarar.

RESUMEN

La posibilidad de evaluar las respuestas que la inteligencia artificial generativa entrega ante una solicitud exige un examen de sus fundamentos y suficiencia mediante conocimientos disciplinares que los estudiantes todavía están consolidando. En torno a lo anterior, el objetivo del presente estudio fue analizar la asociación entre la capacidad autopercibida para evaluar estas respuestas y el acierto al clasificarlas, en estudiantes de Ingeniería en Alimentos de la Universidad Técnica Estatal de Quevedo, Ecuador. Se realizó un estudio cuantitativo, transversal y descriptivo-correlacional, con muestreo no probabilístico por accesibilidad. Participaron 127 estudiantes de cuarto a décimo semestre; el análisis principal incluyó 125 registros completos. El cuestionario, administrado presencialmente mediante Google Forms, reunió cinco ítems de autopercepción, cinco casos alimentarios y la primera acción declarada ante cada respuesta. El desempeño se puntuó por coincidencia exacta con una clave técnica. La autopercepción media fue de 3,46 en una escala de 1 a 5, y el desempeño medio fue de 1,07 aciertos de cinco. La correlación de

Spearman fue positiva y débil (ρ = 0,138; IC del 95 %: −0,037 a 0,309), estimada con 5.000 remuestreos por estudiante. Consultar una fuente confiable o a un especialista fue la primera acción más elegida en todos los casos. La asociación fue débil e incierta, por lo que la autopercepción requiere examinarse junto con tareas de desempeño. Los resultados fundamentan el interés de actividades que hagan explícitas las razones de la clasificación y permitan observar la comprobación, sin establecer sobreconfianza ni conductas efectivas de verificación.

Palabras clave: inteligencia artificial generativa, autopercepción, juicio técnico, ingeniería en alimentos, educación superior

ABSTRACT

The ability to evaluate the responses generated by generative artificial intelligence in response to a prompt requires examining their underlying rationale and adequacy using disciplinary knowledge that students are still in the process of consolidating. Against this background, the objective of the present study was to analyze the association between students' self-perceived ability to evaluate these responses and their accuracy in classifying them among Food Engineering students at the Universidad Técnica Estatal de Quevedo, Ecuador. A quantitative, cross-sectional, descriptive-correlational study was conducted using non-probability convenience sampling. Participants comprised 127 students in semesters four through ten; the primary analysis included 125 complete records. The questionnaire, administered in person through Google Forms, included five self-perception items, five food-related cases, and the first action participants reported they would take regarding each response. Performance was scored by exact agreement with a technical answer key. Mean self-perception was 3.46 on a scale from 1 to 5, and mean performance was 1.07 correct classifications out of five. Spearman's correlation was positive and weak (ρ = 0.138; 95% CI: −0.037 to 0.309), estimated using 5,000 student-level bootstrap resamples. Consulting a reliable source or a specialist was the most frequently selected first action in every case. The association was weak and uncertain, indicating that self-perception needs to be examined alongside performance tasks. The findings support considering learning activities that elicit reasons for classifications and allow verification processes to be observed, without establishing overconfidence or actual verification behavior.

Keywords: generative artificial intelligence, self-perception, technical judgment, food engineering, higher education

Todo el contenido de la Revista Científica Internacional Arandu UTIC publicado en este sitio está disponible bajo licencia Creative Commons Atribution 4.0 International.

INTRODUCCIÓN

Cuando un estudiante recibe una respuesta de inteligencia artificial generativa se encuentra ante una solución que no ha construido cognitivamente. Esa circunstancia modifica el punto desde el cual se aproxima al problema: dispone de una explicación en la que un sistema ya ha seleccionado la información, relacionado los conceptos y establecido una conclusión. El estudiante puede leer la respuesta y comprender lo que afirma e incluso reconocer en ella contenidos estudiados durante su formación. Sin embargo, para decidir si esa respuesta puede utilizarse, se requiere volver sobre un razonamiento que se le presenta al estudiante como dado y externo a él, al mismo tiempo que necesita examinar las fundamentaciones que lo sostienen.

Por tanto, para decidir si la respuesta que da la inteligencia artificial generativa es correcta y utilizable, el estudiante debe reconocer si los datos de la consulta fueron considerados, si los conceptos se aplicaron bajo las condiciones pertinentes y si la conclusión conserva aquello que el problema exigía resolver. La disponibilidad de la respuesta sitúa así al estudiante ante una exigencia de evaluación cuyo cumplimiento depende, en parte, de los conocimientos que todavía está aprendiendo a utilizar.

En Ingeniería en Alimentos, esta dificultad adquiere un contenido preciso. Juzgar una explicación sobre un proceso, una propiedad de un alimento o un resultado microbiológico exige comprender qué puede afirmarse a partir de la información disponible y bajo qué condiciones esa afirmación resulta válida. El reconocimiento del vocabulario técnico ofrece un punto de entrada, aunque puede dejar intacto el problema de su aplicación.

Un estudiante puede comprender las afirmaciones de una respuesta sin advertir que la conclusión depende de una condición omitida; puede también recordar una regla y carecer todavía de los recursos necesarios para reconocer sus límites. En ambos casos, la dificultad se encuentra en la relación entre lo aprendido y lo que la situación demanda hacer con ese conocimiento. Evaluar la respuesta obliga a establecer esa relación, a decidir qué aspecto merece revisión y a reconocer cuándo la información resulta insuficiente para sostener una conclusión.

La necesidad de este examen puede apreciarse en la diversidad de aplicaciones de inteligencia artificial (IA) en el campo alimentario. Pennells et al. (2025) identifican usos en seguridad y calidad alimentaria, control y optimización de procesos, trazabilidad y evaluación sensorial, entre otros ámbitos, y señalan que las respuestas de herramientas basadas en modelos de lenguaje requieren evaluaciones estructuradas y específicas del dominio antes de apoyar decisiones prácticas.

Esta exigencia permite precisar el problema educativo: si la validez de una respuesta debe establecerse atendiendo al conocimiento del campo y a las condiciones de la tarea, aprender a utilizar estas herramientas implica también aprender a reconocer qué tendría que comprobarse en cada situación. La recomendación general de verificar una respuesta deja pendiente ese

aprendizaje. Para llevarla a cabo, el estudiante necesita identificar dónde podría encontrarse una insuficiencia y disponer de razones para decidir si afecta la solución.

Es en esta relación entre conocimiento disponible y exigencia de evaluación donde situamos el *juicio técnico en formación*. Con esta expresión nos referimos a la valoración que el estudiante realiza al confrontar una respuesta con los datos del caso y con los recursos disciplinares que consigue movilizar. Nos interesa esa valoración porque permite interrogar cómo participa el conocimiento en una decisión concreta: qué alcanza a distinguir quien evalúa y hasta dónde puede sostener su elección.

El carácter formativo del juicio remite a esta condición de aprendizaje. El estudiante debe pronunciarse sobre una explicación cuando su capacidad para reconocer errores, omisiones o límites de aplicación puede encontrarse todavía en desarrollo. Por ello, el resultado de su valoración requiere interpretarse en relación con las exigencias del caso y con el alcance de los conocimientos que este demanda.

Evaluación crítica y conocimiento disciplinar

La alfabetización en IA ofrece un marco para examinar la capacidad de los estudiantes para evaluar las respuestas entregadas por la IA generativa. Long y Magerko (2020) incluyen la evaluación crítica entre las competencias necesarias para relacionarse con estas tecnologías, junto con su utilización y la comunicación y colaboración con ellas. En el problema que nos ocupa, esa evaluación crítica requiere adquirir una forma disciplinar: debe permitir examinar una afirmación concreta, reconocer qué razones la sostienen y decidir si responde a la consulta.

Saber que un sistema puede equivocarse puede despertar cautela, pero la cautela necesita encontrar un objeto y un procedimiento para convertirse en un examen crítico. Quien sospecha de una respuesta todavía tiene que determinar qué revisará y con qué criterio lo hará. La formación en IA y la formación disciplinar se encuentran, por tanto, en una tarea cuya resolución depende de ambas, aunque el acierto obtenido no permita distinguir por separado cuánto ha contribuido cada una.

Esta dificultad se hace visible cuando se considera qué tendría que hacer el estudiante para comprobar una explicación. Por ejemplo, ante un cálculo, necesitaría examinar la relación entre los datos, las unidades y el procedimiento empleado; ante una afirmación sobre microorganismos o una inferencia acerca de un lote, tendría que reconocer qué permite concluir la información presentada. En cada caso, la revisión supone identificar una exigencia particular del problema.

El grado de manejo que el estudiante tenga sobre la herramienta de IA puede ayudarle a formular consultas o a solicitar explicaciones adicionales, pero a pesar de esto, aún debe juzgar si lo que recibe resuelve la cuestión que motivó esa consulta. Por eso, al estudiar como los estudiantes clasifican sus propias respuestas, también observamos una actuación que demanda

conocimiento disciplinar. Un desacierto admite explicaciones relacionadas con ese conocimiento, cuya contribución no puede aislarse mediante la clasificación por sí sola.

La amplitud del concepto de alfabetización en IA obliga a precisar dónde se sitúa esta actuación. Ng et al. (2021) organizan el campo en torno al conocimiento y la comprensión, el uso, la evaluación y las cuestiones éticas. Desde esa organización, examinar una respuesta técnica corresponde a una dimensión evaluativa que tiene sentido dentro de una formación más amplia, pero que no permite dar cuenta de todos sus componentes. La comprensión del funcionamiento de los sistemas, la capacidad de crearlos y el análisis de sus implicaciones éticas plantean otro tipo de exigencias cognitivas y formativas. Esta delimitación resulta necesaria para interpretar una prueba: sus resultados adquieren significado por aquello que se pidió hacer y por los conocimientos que esa tarea requirió, de modo que una puntuación obtenida al clasificar respuestas no puede extenderse al conjunto de la alfabetización en IA.

El trabajo de Hornberger et al. (2023) permite apreciar la importancia de esas condiciones, toda vez que en el desarrollo de una prueba objetiva de alfabetización en IA aplicada a 1.286 estudiantes de seis universidades, estos autores observaron mejores resultados entre quienes tenían formación técnica o experiencia previa con IA. Su instrumento examinó conocimientos generales sobre estas tecnologías, por lo que el hallazgo no establece cómo resolverían los participantes una tarea alimentaria. No obstante, lo que sí muestra es que los antecedentes formativos merecen atención cuando se interpreta el desempeño. En nuestra investigación, esa consideración conduce a especificar los contenidos de los casos y la trayectoria académica de la población, sin asumir que el semestre cursado acredita el dominio individual de los temas.

Los cinco casos seleccionados en esta investigación abordan humedad y actividad de agua, congelación, pH y acidez titulable, análisis microbiológico de un lote y preparación de una salmuera. Su diversidad permite plantear exigencias técnicas diferentes, a la vez que circunscribe el alcance de lo observado: el desempeño corresponde al acierto en esas tareas.

Cada clasificación se contrasta con una clave, pero el formulario no recoge una justificación abierta que permita reconstruir las razones de la elección. Podemos establecer si el estudiante coincide con la clasificación técnica esperada; para comprender cómo llegó a ella necesitaríamos examinar su razonamiento. Esta distancia entre el resultado de la valoración y sus fundamentos introduce la pregunta por las razones que permiten conceder crédito a una respuesta.

La autoridad de la respuesta y la responsabilidad de juzgar

Cuando una respuesta se acepta como base para resolver un problema, sus afirmaciones, procedimientos o información no se quedan pasivamente en la mente del estudiante, sino que pasan a orientar una decisión. Y en ese movimiento, aparece la cuestión de la autoridad: interesa comprender qué crédito se concede al contenido y sobre qué razones podría sostenerse. En este artículo empleamos la expresión *autoridad de la respuesta* para examinar esa relación. Su alcance

es interpretativo; no designa una variable medida directamente por el formulario ni presupone que todos los estudiantes otorguen autoridad a la IA. Permite, más bien, formular una pregunta sobre la aceptación de una explicación

Comprender un texto constituye una condición para examinarlo, aunque entre seguir una explicación y establecer su validez todavía hay un trabajo por realizar. El estudiante puede reconocer los conceptos y encontrar plausible el desarrollo de una conclusión sin haber identificado los supuestos de los que depende. En esas condiciones, la respuesta podría recibir crédito antes de que su adecuación al problema haya sido comprobada. También es posible que su aceptación esté respaldada por conocimientos que el estudiante moviliza correctamente. El interés epistemológico, por tanto, reside en distinguir estas posibilidades y en preguntar por las razones de la decisión; una elección final, observada de manera aislada, deja abierta esa diferencia. De ahí que el juicio no pueda reducirse a la impresión que produce el texto ni reconstruirse por completo a partir de su aceptación o rechazo.

Para nuestro problema formativo, evaluar supone conservar la posibilidad de examinar una respuesta, señalar una insuficiencia o suspender su aceptación cuando faltan razones para utilizarla. Su desarrollo en el campo alimentario encuentra un criterio en la validación específica del dominio señalada por Pennells et al. (2025): las afirmaciones deben confrontarse con las condiciones del problema y con el conocimiento pertinente. La corrección se establece mediante ese examen, cuya exigencia permanece incluso cuando la exposición resulta clara y ordenada.

Llamamos *responsabilidad de juzgar* a la exigencia de atender a las razones que permitirían utilizar una respuesta o mantenerla pendiente de comprobación. El sentido de esta responsabilidad es epistemológico y debe entenderse dentro de la situación de aprendizaje. El estudiante puede realizar un examen y equivocarse porque desconoce un aspecto del contenido o porque aplica de manera inadecuada un criterio que conoce. Atribuir el error a una renuncia al juicio eliminaría precisamente el problema que interesa investigar: la dificultad de evaluar cuando los recursos para hacerlo todavía se están formando. Por ello, la responsabilidad se refiere al trabajo de valoración que la tarea exige, sin convertir cada desacierto en una atribución de culpa.

La dependencia apropiada estudiada por Schemmer et al. (2023) aporta un punto de comparación para pensar esta relación. Los autores la conceptualizan mediante la aceptación de consejos correctos y el mantenimiento del juicio propio ante consejos incorrectos, atendiendo a decisiones anteriores y posteriores a recibir la recomendación. Este planteamiento permite advertir por qué la corrección del consejo debe formar parte del examen: aceptar una recomendación correcta y aceptar una incorrecta pueden compartir una conducta aparente, pero tienen consecuencias distintas para la decisión. Nuestro estudio no registra aquella secuencia ni el uso efectivo de las respuestas. La referencia ayuda a precisar el problema, mientras que el instrumento ofrece un acceso más acotado a la clasificación y a la primera acción declarada.

Ese acceso obliga a mantener separadas las observaciones y sus posibles explicaciones. Declarar que se contrastaría una respuesta indica una vía de acción, pero deja pendiente conocer si la comprobación podría ejecutarse y qué criterio se emplearía para valorar su resultado. Del mismo modo, utilizar directamente una respuesta correcta puede apoyarse en razones que el formulario no registra. La confianza general tampoco permite reconstruir la autoridad concedida a cada texto, pues se refiere a una expectativa sobre las herramientas.

Estas diferencias son necesarias para que la discusión sobre el juicio conserve su contenido: si una clasificación acertada se interpretara de inmediato como autonomía intelectual, o una desacertada como delegación en la IA, desaparecería la pregunta por los fundamentos de la decisión. La relación entre lo que el estudiante cree poder evaluar y lo que efectivamente clasifica ofrece una aproximación empírica a ese problema, siempre que se conserve el referente de cada medida.

Autopercepción y desempeño

Considerarse capaz de evaluar una respuesta implica una valoración de los propios recursos técnicos, teóricos y metodológicos de los que se dispone. El estudiante se pronuncia sobre su posibilidad de detectar errores, advertir insuficiencias, revisar cálculos o reconocer cuándo necesita una comprobación adicional. Esta valoración interesa porque permite conocer cómo se sitúa ante la tarea, aunque no establece por sí misma qué logrará distinguir cuando deba examinar un contenido concreto.

Para estudiar esa relación es necesario recoger ambos registros y preguntar si quienes se atribuyen mayor capacidad también resuelven con mayor acierto las tareas propuestas. La comparación adquiere sentido precisamente porque cada registro aporta información sobre un aspecto diferente del problema.

El contenido de la autopercepción merece, por ello, una delimitación cuidadosa. Carolus et al. (2023), al desarrollar la escala MAILS con datos de 300 adultos germanoparlantes, distinguieron facetas de alfabetización percibida y componentes relacionados con autoeficacia y autogestión. Esta diferenciación permite reconocer que una valoración favorable de la propia capacidad de uso puede coexistir con apreciaciones distintas sobre otras capacidades. El antecede de MAILS permite precisar el campo del autoinforme, mientras que el contenido concreto de nuestros ítems delimita aquello sobre lo que se pide al estudiante que se valore.

La necesidad de comparar esa valoración con una tarea de desempeño encuentra un antecedente en GLAT. Jin et al. (2025) desarrollaron una prueba de 20 ítems de alfabetización en IA generativa con 355 participantes y examinaron su validez externa en otros 83 estudiantes. En esta última evaluación, la prueba aportó capacidad predictiva sobre el rendimiento en tareas apoyadas por IA generativa por encima de una medida de competencia autopercibida en ChatGPT. El resultado muestra la utilidad de disponer de información sobre lo que las personas resuelven bajo condiciones definidas, además de lo que afirman acerca de su capacidad.

La comparación en nuestro estudio requiere atender a una diferencia de referencia. Los ítems de autopercepción preguntan por una capacidad general para examinar respuestas, mientras que el desempeño se obtiene mediante cinco casos seleccionados. Una persona puede valorarse a partir de experiencias más amplias que las contenidas en esos casos, y cada caso exige conocimientos particulares. Por esta razón, expresar ambas puntuaciones en un mismo rango numérico no haría equivalentes sus contenidos. Su diferencia tampoco constituiría una medida validada de sobreconfianza. Lo que puede examinarse es la asociación entre ambas: si las posiciones relativas de los estudiantes en la capacidad que se atribuyen guardan relación con sus posiciones en el acierto obtenido.

El interés pedagógico de esta asociación depende de conservar esas condiciones de interpretación. Una correspondencia débil plantearía preguntas sobre la relación entre la valoración de los propios recursos y las exigencias de las tareas, pero no identificaría por sí sola el origen de esa relación. Habría que considerar los conocimientos movilizados, el contenido de los ítems y las características de los estímulos. Una correspondencia positiva tampoco resolvería la cuestión de los fundamentos del juicio, porque el acierto seguiría sin revelar las razones de cada elección. Ambas posibilidades permiten discutir la formación del criterio desde un registro empírico delimitado. Para situarlo mejor, conviene examinar también qué esperan los estudiantes de las herramientas y cómo declaran que procederían ante las respuestas.

Confianza y primera acción declarada

La expectativa acerca de la capacidad de la herramienta de IA utilizada tiene un referente distinto de la valoración de los propios recursos. Cuando un estudiante indica cuánto confía en la corrección de las respuestas de IA, se pronuncia sobre la información que espera recibir; cuando valora su capacidad para evaluarlas, se refiere a lo que cree poder hacer frente a esa información. Distinguir ambos referentes permite estudiar su relación sin asumir que una confianza elevada implica una capacidad autopercibida igualmente elevada, o que una baja confianza demuestra habilidad para reconocer errores. Glikson y Woolley (2020) muestran que la confianza en IA depende de aspectos como su representación, capacidad, transparencia y fiabilidad, y distinguen componentes cognitivos y emocionales. El ítem general sobre corrección técnica utilizado aquí recoge una parte de ese campo y debe interpretarse dentro de ese alcance.

También es necesario examinar qué relación guarda la confianza expresada con una decisión concreta. Mehrotra et al. (2024) diferencian las creencias declaradas sobre la IA de la confianza inferida a partir de la actuación. Esta distinción tiene consecuencias para la lectura de un cuestionario: afirmar que se confía en una herramienta no permite establecer si se adoptará una recomendación determinada.

Para conocer esa actuación sería necesario observarla en la situación correspondiente. Algo semejante ocurre con la verificación. Un estudiante puede declarar que comprobaría una respuesta, pero el formulario no muestra cómo desarrollaría el procedimiento, qué información

encontraría ni por qué la consideraría suficiente. La declaración permite estudiar la vía elegida sin atribuirle el alcance de una comprobación realizada.

En este marco, la primera acción declarada aproxima la pregunta al contenido de cada caso. El estudiante debe indicar si utilizaría la respuesta tal como está, revisaría el razonamiento con conocimientos o cálculos propios, consultaría a una IA, contrastaría el aspecto decisivo con una fuente confiable o una persona especializada, o resolvería por otra vía. La elección informa cómo afirma que comenzaría a proceder ante esa explicación.

Su pertinencia depende del problema y de los recursos disponibles: revisar un cálculo propio puede ser adecuado para una formulación sencilla, mientras que una afirmación sobre inocuidad puede requerir un respaldo específico. Incluso una misma opción admite procedimientos de calidad distinta, que el cuestionario no permite reconstruir. Por ello, las acciones se conservan como categorías nominales, sin ordenarlas en una escala universal de calidad.

Al reunir la clasificación y la primera acción puede describirse qué vías declaran quienes aciertan y quienes eligen otras categorías, manteniendo abierta la explicación de esas relaciones. Se trata de información distinta de la secuencia de decisiones requerida para estudiar la dependencia apropiada en los términos de Schemmer et al. (2023). Además, como no se solicita una probabilidad de acierto para cada clasificación, tampoco es posible examinar calibración por caso. La confianza general, la autopercepción y la acción declarada contribuyen así a contextualizar el desempeño desde referentes diferentes. La literatura sobre educación en ingeniería permite apreciar por qué esas diferencias resultan necesarias al interpretar la relación de los estudiantes con la IA.

Antecedentes en educación en ingeniería

La incorporación de IA generativa a la formación en ingeniería ha dado lugar a investigaciones sobre experiencias de uso, valoraciones de utilidad y prácticas de interacción. La revisión de Álvarez Ariza et al. (2025), que reúne 146 estudios en educación en ingeniería y computación, plantea la necesidad de articular la incorporación de estas herramientas con prácticas reflexivas y con el currículo. Para el problema que aquí se examina, esa articulación exige precisar qué se aprende cuando se utiliza una herramienta y qué evidencia permite reconocerlo. El acceso a una respuesta puede formar parte de una actividad académica, pero la frecuencia de uso, la utilidad que se le atribuye y el acierto al evaluarla informan sobre aspectos distintos de esa experiencia.

La heterogeneidad del uso aparece en el estudio de Dai (2025) sobre un chatbot empleado en proyectos de diseño con 403 estudiantes. El 59,8 % registró un uso sustancial y el 40,2 % mostró un uso limitado o no documentado; las decisiones de utilización se relacionaron con características de la tarea, del estudiante y del contexto. Estas diferencias muestran por qué la disponibilidad de una herramienta no basta para describir la experiencia formativa que se

construye con ella. Alkabaa y Alamri (2025), mediante una encuesta a 403 estudiantes de siete ramas de ingeniería, también identificaron diferencias entre especialidades en experiencias y percepciones sobre IA generativa. Aunque sus medidas autoinformadas no permiten determinar qué grupo reconoce objetivamente más errores, el estudio invita a precisar la población y los contenidos cuando se interpretan resultados bajo la denominación general de estudiantes de ingeniería.

La utilidad percibida añade una dificultad: una misma herramienta puede valorarse de manera diferente según la función que se le atribuya. Mohammed et al. (2026) examinaron una actividad de Ingeniería Mecánica mediante una encuesta a 56 estudiantes, de los cuales 47 declararon haber utilizado IA. La valoración de su ayuda para resolver problemas y aclarar conceptos fue diferente de la valoración de su uso para validar cálculos y conceptos. La distinción importa para comprender la evaluación de respuestas, porque encontrar apoyo para seguir una explicación no resuelve necesariamente el examen de su corrección. El estudio recoge percepciones en una asignatura y, desde ese registro, permite preguntar qué espera el estudiante cuando considera útil una respuesta; la evidencia directa de aprendizaje requiere otras observaciones.

La coexistencia de utilidad y limitaciones técnicas se aprecia en Shallari et al. (2026), quienes compararon valoraciones estudiantiles con una evaluación docente de respuestas en dos cursos de sistemas digitales, con 32 participantes de pregrado y 20 de posgrado. Los estudiantes podían atribuir utilidad a contenidos en los que los docentes identificaban errores o la necesidad de apoyo adicional.

Este resultado permite continuar la pregunta anterior: si una explicación puede ser valorada favorablemente pese a contener dificultades técnicas, conviene observar qué distingue el estudiante cuando se le pide juzgarla. El antecedente no aplicó a cada participante una prueba equivalente a nuestros casos; su contribución consiste en mostrar por qué la utilidad percibida deja abierta una pregunta sobre la valoración técnica que nuestro estudio aborda mediante estímulos comunes.

Las investigaciones sobre formas de interacción ofrecen otra aproximación a esa pregunta. Hou et al. (2025), con 808 respuestas posteriores a una actividad con IA generativa, encontraron asociaciones entre pensamiento crítico y formas reflexivas, cautelosas y colaborativas de uso; la confianza mostró su asociación más fuerte con el uso irreflexivo. Amoozadeh et al. (2024), en una encuesta a 253 estudiantes de informática, identificaron también heterogeneidad en la confianza en IA generativa. Al proceder de autoinformes, estas observaciones permiten estudiar cómo se relacionan las valoraciones y las prácticas declaradas, sin establecer una dirección causal ni demostrar conductas observadas. Para nuestra investigación, sostienen el interés de registrar confianza y experiencia de uso conservando la diferencia entre esos datos y el acierto en una tarea técnica.

El examen de la interacción requiere, además, atender a lo que se hace con las respuestas durante la actividad. López-Vargas et al. (2026) estudiaron una tarea de depuración con 340 estudiantes y analizaron datos completos de interacción en 301 casos. El éxito declarado se relacionó con la eficiencia de las consultas y con las prácticas de verificación; una mayor cantidad de consultas no aseguró mejores resultados dentro del tiempo disponible. La ausencia de grupo control y el carácter autodeclarado del éxito limitan la interpretación causal, pero la relación observada permite precisar el objeto del examen: contar interacciones deja pendiente conocer si contribuyeron a reconocer y resolver el problema.

La verificación plantea a su vez preguntas sobre aquello que se comprueba. Borowczak y Borowczak (2026) describen, entre 147 estudiantes, frecuencias declaradas de verificación diferentes para código y explicaciones técnicas. La distinción entre ambos contenidos permite advertir que preguntar con qué frecuencia se verifica no agota el estudio de esa práctica. Para comprenderla habría que conocer qué aspecto se examina y cómo se decide que la evidencia obtenida es suficiente. Nuestro formulario recoge la primera vía de acción, de manera que conserva una referencia al caso sin reconstruir el desarrollo de la estrategia. Esta delimitación conduce de nuevo al problema disciplinar: el examen que una respuesta requiere depende de lo que afirma y de las condiciones bajo las cuales se pretende utilizarla.

Antecedentes alimentarios y objetivo del estudio

En Ciencia de los Alimentos, Sabbione y Patrignani (2024) presentan una propuesta didáctica que incorpora ChatGPT y solicita contrastar sus respuestas con libros y trabajos científicos. La comprobación documental forma parte de la actividad, de modo que la información recibida debe someterse a examen durante el aprendizaje. El trabajo describe una encuesta docente y el desarrollo de la propuesta, sin presentar resultados posteriores de desempeño estudiantil. Su aportación permite situar la revisión como una exigencia de enseñanza en la disciplina y plantea una pregunta que requiere observación empírica: qué logran reconocer los estudiantes cuando deben evaluar una explicación técnica.

La validación específica del dominio señalada por Pennells et al. (2025) ayuda a comprender el alcance de esa pregunta. Una respuesta destinada a apoyar decisiones alimentarias debe examinarse mediante conocimientos pertinentes; una actividad que incorpora ese examen debe considerar cómo aprenden los estudiantes a utilizarlos. Entre ambas exigencias se encuentra el juicio técnico en formación. El problema consiste en conocer qué valoración puede formular quien debe decidir sobre una respuesta mientras todavía está consolidando los criterios para hacerlo. Estudiar esa actuación requiere definir los contenidos, las condiciones de la tarea y aquello que se considerará acierto, pues la pertinencia del problema educativo no garantiza por sí misma la calidad de los estímulos seleccionados.

Los antecedentes revisados aportan pruebas de alfabetización, escalas de autopercepción y estudios sobre uso, confianza y verificación. La diversidad de sus medidas permite reconocer

varias dimensiones de la relación con la IA, aunque también exige precisar cómo se conectan cuando el objeto es una respuesta alimentaria. En el corpus temático de los trabajos utilizado para fundamentar esta investigación no se identificó la misma combinación de estudiantes de Ingeniería en Alimentos, autopercepción evaluativa, clasificación individual de cinco respuestas técnicas adaptadas y primera acción declarada.

Esta observación se limita a los trabajos examinados: su selección no constituye una revisión sistemática ni permite afirmar prioridad universal. El interés del estudio reside en observar cómo se relaciona la capacidad que el estudiante se atribuye con su acierto ante tareas que demandan conocimientos del campo, y en describir cómo afirma que procedería frente a esas mismas respuestas.

La pregunta principal es qué asociación existe entre la capacidad autopercibida para evaluar respuestas de IA y el acierto al clasificar cinco respuestas técnicas adaptadas en estudiantes de Ingeniería en Alimentos de la Universidad Técnica Estatal de Quevedo. El objetivo general es analizar esa asociación en estudiantes de cuarto a décimo semestre. No se establece una hipótesis direccional, porque la relación se plantea sin presuponer concordancia o discrepancia entre las medidas. De manera complementaria, se busca describir la experiencia de uso, la formación previa, la verificación habitual y la confianza general; caracterizar las clasificaciones y las primeras acciones declaradas ante cada caso; y explorar la relación del desempeño con las variables formativas y de uso.

Esta aproximación permite discutir necesidades de enseñanza a partir de resultados de valoración y de la capacidad que los estudiantes se atribuyen, conservando abierta la pregunta por los fundamentos de sus elecciones. La discusión sobre autoridad y responsabilidad orienta esa lectura sin convertir una asociación en prueba de autonomía intelectual o de cesión del juicio a la IA. La interpretación depende de mantener vinculadas las medidas con las tareas de las que proceden. Por ello, el diseño que se describe a continuación distingue la clasificación técnica, la autopercepción y las acciones declaradas, y establece qué información aporta cada una al problema de aprender a juzgar una respuesta disponible.

MATERIALES Y MÉTODOS

Diseño y participantes

Se realizó un estudio cuantitativo, no experimental y transversal, de alcance descriptivo-correlacional, en estudiantes de Ingeniería en Alimentos de la Universidad Técnica Estatal de Quevedo, Ecuador, durante el PPA 2026-2027. La convocatoria comprendió todos los paralelos de cuarto a décimo semestre, mediante muestreo no probabilístico por accesibilidad y participación voluntaria. Estos niveles se seleccionaron porque habían cursado contenidos considerados necesarios para abordar los casos, sin asumir que el semestre acreditara su dominio individual.

Se estableció previamente una semana de vigencia de la encuesta. La exportación registra envíos entre el 6 y el 11 de septiembre de 2026; el cierre se comunicó antes de examinar las asociaciones. De 130 envíos, tres correspondieron a personas que no otorgaron consentimiento y dejaron vacío el instrumento. Se incluyeron 127 estudiantes, todos con respuestas a los cinco casos. Dos omitieron un ítem de autopercepción, por lo que el análisis principal comprendió 125 registros completos. Los tres participantes del piloto quedaron fuera de esta muestra. La unidad de análisis fue el estudiante, conservando vinculadas sus respuestas a los cinco casos.

Instrumento y revisión

El formulario de Google incluyó consentimiento y tres bloques. El primero registró semestre, frecuencias de uso académico y disciplinar de IA generativa, herramienta más utilizada, formación previa para evaluar respuestas, verificación habitual y confianza general en su corrección técnica. Las frecuencias y la confianza se trataron como ordinales; la herramienta y la procedencia de la formación, como nominales.

El segundo bloque reunió cinco ítems propios del proyecto sobre capacidad autopercibida: identificar errores técnicos, reconocer respuestas correctas pero insuficientes, detectar errores en cálculos o explicaciones, distinguir una respuesta convincente de otra fundamentada y reconocer cuándo se necesita comprobación adicional. Las respuestas variaron de 1, «totalmente en desacuerdo», a 5, «totalmente de acuerdo», sin ítems inversos. El índice fue la media de los cinco ítems.

El tercer bloque presentó cinco respuestas adaptadas de un banco de 20 consultas a ChatGPT (versión gratuita), sobre humedad y actividad de agua, congelación y microorganismos, pH y acidez titulable, análisis microbiológico de un lote y preparación de una salmuera. Fueron estímulos seleccionados para evaluar clasificaciones técnicas, no una muestra aleatoria de respuestas del sistema. Cada respuesta se clasificó como correcta y suficiente; correcta, pero incompleta; contiene algún error técnico; incorrecta en lo esencial; o no puedo determinarlo. Una guía explicó las diferencias entre categorías.

Después de cada clasificación se preguntó qué haría primero el estudiante: utilizar la respuesta directamente, revisarla con conocimientos o cálculos propios, consultar a una IA, contrastarla con una fuente confiable o especialista, o resolver por otra vía. Las acciones se conservaron como categorías nominales, sin una jerarquía universal de calidad. No se recogieron justificaciones abiertas, confianza por caso ni comprobaciones ejecutadas.

Dos revisores con doctorado en Ciencias de Alimentos examinaron los casos y sus claves. Solicitaron precisiones o adiciones de términos, incorporadas antes de la encuesta principal, y mantuvieron las clasificaciones previstas. Un piloto con tres estudiantes no registró dificultades reportadas de comprensión o funcionamiento. Estas actividades correspondieron a revisión técnica y comprobación de la aplicación, sin constituir una validación psicométrica integral.

Aplicación y resguardo

La aplicación fue presencial, con autorización verbal de la facultad y la carrera y asistencia de un colaborador que atendió dudas operativas sin interpretar los casos. Los participantes recibieron información sobre voluntariedad, uso académico de los datos y ausencia de consecuencias sobre sus calificaciones. El formulario no recogió información personal. Se solicitó responder con conocimientos propios, sin consultar IA, internet, materiales u otras personas. Los casos mantuvieron el mismo orden y no se reportaron incidencias. Se conservaron la exportación original, una base de trabajo y una base analítica con identificadores secuenciales, junto con el registro de depuración y el código. El investigador tiene acceso a la base, cuya conservación será indefinida. No se eliminaron participaciones por baja puntuación o semejanza de respuestas.

Puntuación y análisis

La clave para los casos 1 a 5 fue, respectivamente, correcta y suficiente; contiene algún error técnico; correcta, pero incompleta; incorrecta en lo esencial; y correcta y suficiente. Cada coincidencia exacta recibió un punto; las demás categorías, incluida la abstención, recibieron cero. El rango posible del total fue de 0 a 5. La abstención se conservó además descriptivamente y las omisiones se trataron como faltantes. Ambos índices requirieron sus cinco componentes, sin imputación ni prorrateo; los registros incompletos contribuyeron a las variables disponibles.

Se calcularon frecuencias, porcentajes, mediana, rango intercuartílico, rango observado, media y desviación estándar. Por caso se presentó la distribución de clasificaciones y el acierto con intervalo de Wilson del 95 %. La consistencia interna de la autopercepción se examinó mediante correlaciones entre ítems, ítem-total corregidas y alfa de Cronbach con intervalo bootstrap, utilizando las puntuaciones originales y correlaciones de Pearson.

La asociación principal se estimó mediante Spearman, con intervalo del 95 % por percentiles de 5.000 remuestreos de estudiantes completos. Se conservaron los empates mediante rangos promedio y se utilizó la semilla 20260905. Exploratoriamente se relacionó el desempeño con confianza, semestre, uso y verificación; la formación se comparó mediante Kruskal–Wallis y tamaño del efecto. Los cruces entre acierto y acción se describieron con porcentajes y V de Cramér. Los contrastes utilizaron 9.999 permutaciones; Benjamini–Hochberg se aplicó separadamente a las cinco correlaciones ordinales y los cinco cruces. Kruskal–Wallis se informó individualmente. La sensibilidad omitió sucesivamente cada caso, manteniendo como principal la puntuación completa. El análisis utilizó Python 3.12.14, NumPy 2.3.5 y pandas 3.0.1. Las inferencias se circunscriben a las asociaciones de esta muestra y estas tareas.

RESULTADOS

Participantes y características de la muestra

Se recibieron 130 envíos, de los cuales tres correspondieron a personas que no otorgaron su consentimiento y no respondieron el resto del instrumento. La muestra estuvo conformada por

127 estudiantes de cuarto a décimo semestre. Los tres participantes del piloto no formaron parte de esta base y no se reportaron incidencias de uso de apoyos durante la aplicación. Todos los participantes respondieron las cinco clasificaciones y las cinco preguntas de acción. Dos omitieron el ítem 10 de autopercepción, por lo que el índice de cinco ítems y su asociación con el desempeño se calcularon con 125 respuestas completas, sin imputación.

Tabla 1

Distribución de participantes por semestre

Semestre	n	%
4	9	7,1
5	10	7,9
6	4	3,1
7	40	31,5
8	16	12,6
9	26	20,5
10	22	17,3
Total	127	100,0

Séptimo semestre concentró la mayor proporción de participantes (31,5 %; tabla 1). El uso académico de IA al menos una vez por semana fue declarado por 111 estudiantes (87,4 %), y el uso con esa frecuencia para problemas de la carrera por 93 (73,2 %). Las herramientas más utilizadas fueron ChatGPT (66; 52,0 %), Claude (31; 24,4 %) y Gemini (20; 15,7 %).

La formación para evaluar respuestas de IA fue personal en 49 estudiantes (38,6 %), institucional en nueve (7,1 %) y ambas en 13 (10,2 %); 56 (44,1 %) no reportaron formación. La verificación frecuente o constante fue declarada por 51 (40,2 %) y la confianza moderada por 98 (77,2 %).

Autopercepción y distribución del desempeño

El índice de autopercepción presentó una media de 3,46 (DE = 0,74), mediana de 3,40 y cuartiles de 3,00 y 4,00, con un rango observado de 1 a 5 (n = 125). Las medias por ítem se presentan en la tabla 2. El alfa de Cronbach fue 0,851 (IC del 95 %: 0,776–0,899), calculado con los 125 registros completos. Las correlaciones ítem–total corregidas estuvieron entre 0,603 y 0,736.

Tabla 2

Autopercepción de la capacidad de evaluación técnica

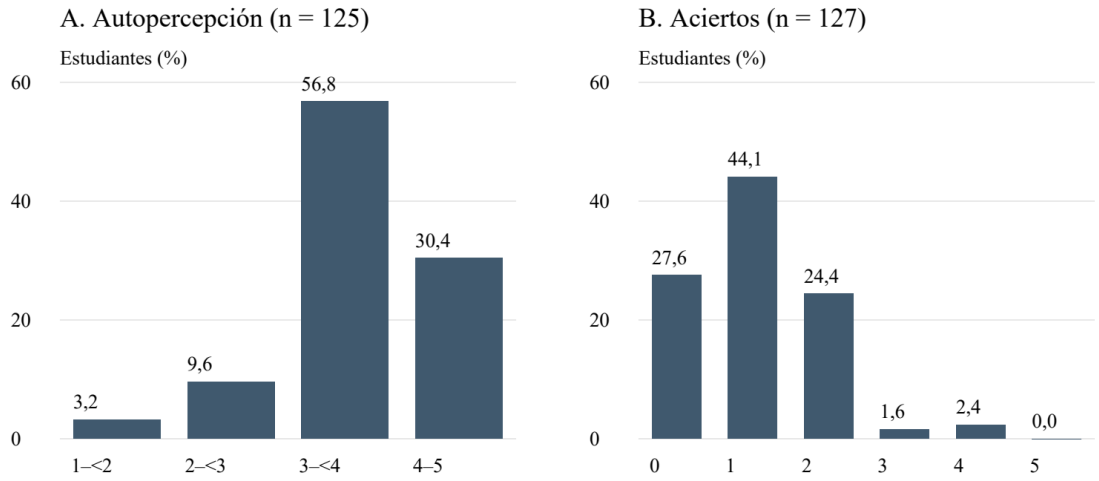
Ítem abreviado	n	Media	DE	r ítem–total
8. Identificar errores técnicos	127	3,46	1,00	0,656
9. Reconocer información insuficiente	127	3,58	0,80	0,677
10. Detectar errores de cálculo o explicación	125	3,48	1,04	0,603
11. Distinguir persuasión y fundamento técnico	127	3,39	0,96	0,657
12. Determinar necesidad de comprobación	127	3,43	0,93	0,736

Nota. Respuestas de 1 a 5. Los descriptivos utilizan las respuestas disponibles por ítem; las correlaciones ítem–total corregidas utilizan n = 125. IC del alfa: bootstrap percentil de 5.000 muestras de estudiantes.

El desempeño medio fue de 1,07 aciertos de cinco (DE = 0,89), con mediana de 1 y cuartiles de 0 y 2 (n = 127). Obtuvieron cero aciertos 35 estudiantes (27,6 %); uno, 56 (44,1 %); dos, 31 (24,4 %); tres, dos (1,6 %); y cuatro, tres (2,4 %). Ninguno obtuvo cinco aciertos (figura 1).

Figura 1

Distribuciones de autopercepción y total de aciertos



Nota. Los porcentajes se calculan dentro de cada panel. La autopercepción se agrupa solo para visualizarla: [1,2), [2,3), [3,4) y [4,5]. Los cálculos utilizan su valor original. Elaboración propia con la base del estudio.

Clasificación técnica de las respuestas

La proporción de coincidencias con la clave fue de 18,1 % en el caso 1, 29,9 % en el caso 2, 30,7 % en el caso 3, 6,3 % en el caso 4 y 22,0 % en el caso 5. La tabla 3 conserva todas las categorías elegidas, incluida la respuesta «No puedo determinarlo».

Tabla 3

Clasificaciones elegidas y acierto exacto por caso

Caso y clave	A	B	C	D	E	Acierto % [IC 95 %]
1 (A)	23	59	26	7	12	18,1 [12,4; 25,7]
2 (C)	28	40	38	14	7	29,9 [22,6; 38,4]
3 (B)	23	39	38	16	11	30,7 [23,4; 39,2]
4 (D)	18	39	46	8	16	6,3 [3,2; 11,9]
5 (A)	28	37	27	19	16	22,0 [15,7; 30,0]

Nota. n = 127 por caso. Las columnas A–E expresan frecuencias. A: correcta y suficiente; B: correcta, pero incompleta; C: contiene algún error técnico; D: incorrecta en lo esencial; E: no puedo determinarlo. Clave preespecificada y confirmada por el investigador: A–C–B–D–A. IC de Wilson del 95 %.

En el caso 1, sobre humedad y actividad de agua, 59 estudiantes (46,5 %) eligieron «correcta, pero incompleta», frente a 23 (18,1 %) que coincidieron con la clave «correcta y suficiente». En el caso 2, sobre congelación y microorganismos, la categoría más frecuente fue «correcta, pero incompleta» (40; 31,5 %), seguida de «contiene algún error técnico» (38; 29,9 %), que correspondía a la clave.

En el caso 3, sobre pH y acidez titulable, las elecciones se concentraron en «correcta, pero incompleta» (39; 30,7 %) y «contiene algún error técnico» (38; 29,9 %). En el caso 4, sobre el análisis microbiológico de un lote, ocho estudiantes (6,3 %) eligieron «incorrecta en lo esencial» y 46 (36,2 %) señalaron «contiene algún error técnico». En el caso 5, sobre preparación de salmuera, 37 estudiantes (29,1 %) eligieron «correcta, pero incompleta» y 28 (22,0 %) coincidieron con «correcta y suficiente».

La opción «No puedo determinarlo» fue seleccionada por 12, siete, 11, 16 y 16 estudiantes en los casos 1 a 5, respectivamente. Estas respuestas se conservaron como categoría descriptiva y recibieron cero puntos conforme al criterio de acierto exacto.

Primera acción declarada ante las respuestas

La comprobación mediante una fuente confiable o consulta especializada fue la acción más frecuente en los cinco casos (tabla 4). Su elección osciló entre 49 estudiantes (38,6 %) en el caso 5 y 57 (44,9 %) en el caso 4. El uso directo sin otra comprobación fue elegido por seis a diez estudiantes por caso.

Tabla 4

Primera acción declarada en cada caso

Caso	Uso directo	Revisión propia	Contraste con IA	Fuente o especialista	Otra vía
1	7 (5,5)	27 (21,3)	32 (25,2)	56 (44,1)	5 (3,9)
2	9 (7,1)	24 (18,9)	32 (25,2)	55 (43,3)	7 (5,5)
3	10 (7,9)	22 (17,3)	31 (24,4)	51 (40,2)	13 (10,2)
4	6 (4,7)	21 (16,5)	29 (22,8)	57 (44,9)	14 (11,0)
5	10 (7,9)	31 (24,4)	27 (21,3)	49 (38,6)	10 (7,9)

Nota. Frecuencias y porcentajes entre paréntesis; n = 127 por caso. Las acciones son categorías nominales y expresan la primera acción declarada, no una conducta observada.

Los cruces entre acierto exacto y acción declarada se presentan en la tabla 5. En el caso 2, 19 de los 38 estudiantes con acierto (50,0 %) eligieron contrastar con IA, frente a 13 de los 89

sin acierto exacto (14,6 %). En el caso 5, siete de los 28 estudiantes con acierto (25,0 %) eligieron usar directamente la respuesta, frente a tres de los 99 sin acierto exacto (3,0 %).

Tabla 5

Asociación entre acierto exacto y primera acción declarada

Caso	n	V de Cramér	p Monte Carlo	p ajustado
1	127	0,114	0,8526	0,8526
2	127	0,419	0,0002	0,0010
3	127	0,194	0,3196	0,5327
4	127	0,139	0,6791	0,8489
5	127	0,354	0,0034	0,0085

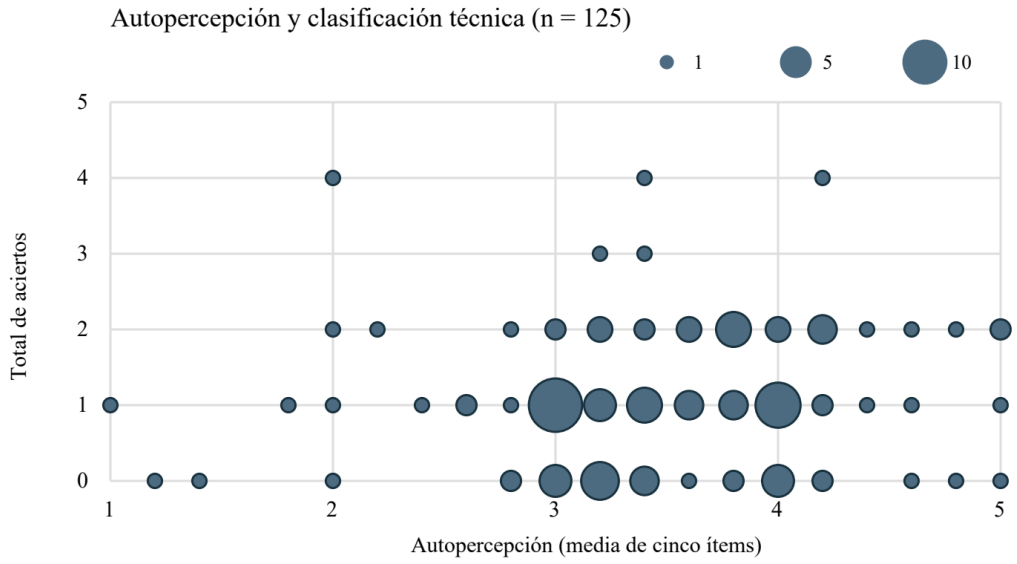
Nota. Tablas de 2 × 5: acierto/no acierto exacto por acción. Se incluyó la abstención en no acierto. Pruebas de independencia con estadístico chi-cuadrado de Pearson y 9.999 permutaciones de etiquetas, conservando los márgenes. Ajuste Benjamini–Hochberg para la familia de cinco casos.

Asociación entre autopercepción y desempeño

La correlación de Spearman entre el índice de autopercepción y el total de aciertos fue de 0,138 (IC del 95 %: −0,037 a 0,309; n = 125). La estimación puntual fue positiva y de baja magnitud; el intervalo incluyó el valor cero. La figura 2 muestra la frecuencia de las combinaciones observadas de ambas puntuaciones.

Figura 2

Relación entre autopercepción y total de aciertos



Nota. Cada círculo representa una combinación de puntuaciones y su área es proporcional al número de estudiantes. La leyenda indica frecuencias de referencia. Se conservaron los 125 registros con los cinco ítems de autopercepción completos. Elaboración propia con la base del estudio.

El intervalo se estimó mediante 5.000 remuestreos de estudiantes completos y los percentiles 2,5 y 97,5 de los coeficientes obtenidos. Se utilizó la semilla 20260905 y no se registraron remuestreos con correlación no estimable.

Al omitir sucesivamente cada caso, la correlación fue de 0,020 sin el caso 1; 0,291 sin el caso 2; 0,103 sin el caso 3; 0,107 sin el caso 4; y 0,014 sin el caso 5 (n = 125 en todas las estimaciones). Se mantuvo como resultado principal el coeficiente calculado con los cinco casos.

Asociaciones exploratorias

Las asociaciones del desempeño con confianza general, semestre, frecuencia de uso académico, frecuencia de uso disciplinar y verificación habitual se presentan en la tabla 6. La mayor estimación positiva correspondió a la verificación habitual (rho = 0,206; p sin ajuste = 0,0197; p ajustado = 0,0985). Ninguna de las cinco asociaciones obtuvo un valor p ajustado inferior a 0,05.

Tabla 6

Asociaciones exploratorias del desempeño con variables ordinales

Variable	n	rho	p	p ajustado
Confianza general	127	0,076	0,3932	0,4915
Semestre	127	-0,138	0,1199	0,2998
Uso académico de IA	127	-0,092	0,3089	0,4915
Uso disciplinar de IA	127	0,000	0,9987	0,9987
Verificación habitual	127	0,206	0,0197	0,0985

Nota. Correlaciones de Spearman. Valores p bilaterales calculados con 9.999 permutaciones y ajustados mediante Benjamini–Hochberg en esta familia de cinco asociaciones. Los índices ordinales mantienen el orden de las opciones del cuestionario.

La comparación del desempeño según formación previa produjo H de Kruskal–Wallis = 1,941 (p por permutación = 0,5987). El tamaño del efecto calculado como (H − k + 1)/(n − k), truncado en cero, fue 0. Las medias de aciertos fueron 1,04 (DE = 0,87; n = 56) entre quienes no reportaron formación; 1,00 (DE = 0,87; n = 49) entre quienes reportaron formación personal; 1,56 (DE = 1,24; n = 9) entre quienes reportaron formación institucional; y 1,15 (DE = 0,80; n = 13) entre quienes reportaron ambas. La mediana fue de un acierto en los cuatro grupos. No se realizaron comparaciones por pares.

DISCUSIÓN

Autopercepción y juicio técnico ante respuestas de IA

La relación entre la capacidad que los estudiantes se atribuyeron y su acierto al clasificar respuestas técnicas fue positiva y débil: rho = 0,138, con un intervalo de confianza del 95 % de −0,037 a 0,309 y 125 participantes en el análisis principal. Este resultado sitúa una dificultad en la evaluación de la formación: conocer cómo se valora un estudiante ofrece información limitada sobre la posición que ocupa frente a otros cuando debe examinar los casos propuestos. El intervalo, sin embargo, admite desde una asociación ligeramente negativa hasta una positiva de mayor magnitud que la estimación puntual. Por ello, la relación observada debe discutirse atendiendo tanto a su debilidad en la muestra como a la incertidumbre que permanece acerca de

su magnitud. Los datos no establecen independencia entre las medidas y tampoco respaldan una correspondencia estrecha entre ellas.

Para comprender qué significa esta relación conviene volver sobre aquello que se pidió a los particiántes en cada momento. Al responder los ítems de autopercepción, el estudiante valoró su capacidad general para reconocer errores e insuficiencias; al clasificar las respuestas, tuvo que aplicar conocimientos a consultas específicas y elegir entre categorías próximas. El primer registro permite conocer cómo se reconoce ante la exigencia de evaluar, mientras que el segundo muestra el resultado de una tarea definida. Esta diferencia da sentido a la comparación con Jin et al. (2025), quienes encontraron que una prueba de alfabetización en IA generativa aportaba capacidad predictiva sobre el rendimiento en tareas asistidas por IA por encima de una medida de competencia autopercibida en ChatGPT. Aunque los contenidos y las poblaciones impiden comparar directamente las puntuaciones, ambos trabajos sostienen la necesidad de observar lo que se resuelve además de preguntar por la capacidad que se cree poseer.

La media de autopercepción de 3,46 sobre 5 y el promedio de 1,07 aciertos sobre cinco podrían invitar a describir el resultado como una sobrevaloración de la propia capacidad. Esa lectura supondría, sin embargo, que las dos puntuaciones expresan una misma cantidad y pueden restarse para establecer cuánto se equivocó el estudiante al valorarse. El acuerdo con una afirmación general sobre la capacidad no expresa una probabilidad de acierto en cada caso, y el total de clasificaciones correctas depende de los contenidos seleccionados. La diferencia numérica entre ambas medias carece, por tanto, del significado que tendría una comparación entre confianza y acierto referidos a una misma decisión. En consecuencia, nuestro hallazgo se encuentra en la asociación entre las posiciones relativas de los participantes, sin que esa asociación permita diagnosticar sobreconfianza.

La consistencia interna de los ítems, con un alfa de 0,851, tampoco elimina esta diferencia. Una persona puede responder de forma coherente acerca de distintas capacidades que se atribuye y encontrar dificultades cuando debe utilizarlas frente a un contenido determinado. El coeficiente describe la relación entre los ítems de autopercepción; la pertinencia de esa medida para comprender el juicio técnico requiere examinar, además, cómo se relaciona con la tarea. Del lado del desempeño, la concentración de puntuaciones bajas y el reducido número de casos limitan la diferenciación entre estudiantes. La correlación varió de 0,014 a 0,291 al omitir sucesivamente cada caso, lo que muestra cuánto puede cambiar su estimación al modificar la composición de la prueba. Interpretar la asociación exige entonces examinar qué reúne el total de aciertos y qué diferencias quedan ocultas cuando se utiliza como única descripción del desempeño.

Qué expresa el desacierto en la clasificación

En el caso de análisis microbiológico, ocho estudiantes eligieron la categoría prevista, «incorrecta en lo esencial», mientras que otros 46 indicaron que la respuesta contenía algún error técnico. Las dos elecciones expresan valoraciones distintas del alcance del problema, aunque solo

la primera recibe un punto. Los 46 estudiantes no clasificaron el texto como plenamente correcto; su elección reconoce la presencia de un error sin coincidir con la categoría que la clave asigna a su gravedad. Si se leyera únicamente la puntuación, esta diferencia desaparecería y podría atribuirse a todos los desaciertos una aceptación del contenido que las respuestas registradas no sostienen. El examen por categorías permite precisar la dificultad: además de advertir un error, la tarea exige determinar cuánto compromete el razonamiento o la conclusión.

La suficiencia plantea un problema relacionado desde otra dirección. En humedad y actividad de agua y en preparación de una salmuera, ambas con clave «correcta y suficiente», la opción más frecuente fue «correcta, pero incompleta». Estas elecciones sitúan la pregunta en los límites de lo que una consulta requiere: qué información basta para responder y qué omisión impide hacerlo. Es posible que algunos estudiantes hayan esperado una explicación más extensa, considerado necesaria una condición adicional o interpretado de otra manera las categorías. El instrumento no permite elegir entre esas explicaciones, porque no recogió justificaciones. Lo observado es la frecuencia con la que se atribuyó insuficiencia a respuestas consideradas suficientes en la clave; comprender las razones de esa valoración requiere escuchar o registrar el razonamiento que la sostuvo.

La puntuación por coincidencia exacta conserva un criterio común y mantiene la clave fijada antes del análisis. Al mismo tiempo, reúne bajo cero puntos elecciones cuyo contenido puede ser diferente. Reconocer esta propiedad del instrumento permite utilizar el total para la comparación prevista y acompañarlo de una lectura que preserve los matices de cada caso. La distribución de categorías no autoriza a corregir retrospectivamente la puntuación para premiar respuestas alternativas; permite identificar qué distinciones deberían examinarse con mayor detalle en una nueva aplicación. La frontera entre información insuficiente, error localizado e incorrección esencial forma parte de la tarea evaluativa y también exige comprobar cómo la entienden los participantes.

Esta frontera adquiere sentido en relación con el problema alimentario que debe resolverse. Pennells et al. (2025) señalan la necesidad de una evaluación estructurada y específica del dominio de las respuestas de sistemas de lenguaje especializados en alimentos antes de utilizarlas en decisiones prácticas. Desde esa exigencia, los resultados permiten formular una necesidad de enseñanza: trabajar cómo se determina el alcance de una afirmación y su importancia para la solución. Reconocer una expresión cuestionable constituye un paso; establecer si puede corregirse sin alterar la conclusión, si deja información necesaria sin resolver o si compromete la respuesta en lo esencial exige relacionarla con el conjunto del razonamiento. Los cinco casos permiten observar resultados de esa valoración, cuya justificación continúa siendo el problema central para comprender el juicio técnico en formación.

Autoridad de la respuesta y responsabilidad de juzgar

La pregunta por la autoridad de la respuesta surge al considerar qué razones permitirían utilizarla después de evaluarla. La revisión de las categorías muestra que el examen puede exigir tanto reconocer una explicación suficiente como precisar por qué otra necesita corrección. Desconfiar de un contenido atribuido a IA no garantiza, por sí mismo, una valoración fundada: señalar insuficiencia donde la clave establece suficiencia también requiere una justificación. La evaluación crítica incluida por Long y Magerko (2020) entre las competencias de alfabetización en IA adquiere aquí una exigencia concreta: disponer de criterios para decidir qué afirmaciones merecen crédito y con qué límites pueden orientar una decisión.

Esta exigencia permite desarrollar la responsabilidad de juzgar como un problema de formación. El estudiante debe aprender a reconocer qué relación necesita examinar, qué conocimiento resulta pertinente y hasta dónde puede sostener su conclusión. Desde esta perspectiva, el error interesa porque abre una pregunta por el modo en que se realizó la valoración. Atribuirlo directamente a pasividad intelectual o a una cesión del juicio a la IA impediría examinar otras posibilidades, como dificultades de conocimiento disciplinar, comprensión de la consulta o interpretación de las opciones. El cuestionario no distingue esas explicaciones y tampoco compara la actuación ante los mismos textos atribuidos a una fuente humana. La relación débil con la autopercepción no aporta la evidencia que falta para identificar el origen de cada elección.

La contribución del marco epistemológico consiste en orientar la interpretación hacia las razones que hacen defendible una decisión, conservando la diferencia entre esas razones y el resultado registrado. Acertar en la clasificación informa una coincidencia con el criterio técnico; conocer si el estudiante puede justificarla exige observar algo más. Por ello, la formación del juicio plantea una tarea que continúa después de elegir una categoría: explicar qué permite aceptar el contenido, qué debe corregirse o qué información falta para decidir. La primera acción declarada ofrece una aproximación a cómo el estudiante afirma que proseguiría, aunque para comprender ese paso es necesario atender al contenido de cada caso y al alcance de lo que efectivamente se preguntó.

Comprobación declarada y condiciones de uso

La consulta a una fuente confiable o a una persona especializada fue la primera acción más elegida en los cinco casos. Esta preferencia muestra que buscar respaldo externo ocupa un lugar reconocible entre las vías que los estudiantes declaran ante una respuesta técnica. Su elección deja abierto el trabajo necesario para convertir esa intención en una comprobación: identificar la afirmación que debe revisarse, seleccionar una fuente pertinente y establecer qué hacer si la información encontrada discrepa de la respuesta. El formulario registra la vía elegida, pero no permite conocer cómo resolverían los participantes esas exigencias. Por tanto, la frecuencia de esa opción debe leerse como una disposición declarada cuyo desarrollo requiere observación adicional.

Los cruces entre clasificación y acción permiten apreciar por qué las opciones deben interpretarse dentro de cada tarea. En congelación, quienes coincidieron con la clave eligieron con mayor frecuencia consultar a una IA; en la salmuera, el uso directo fue más frecuente entre quienes clasificaron correctamente la respuesta. Los valores p ajustados de los cruces completos fueron 0,0010 y 0,0085, respectivamente. Estas asociaciones describen cómo se distribuyeron las intenciones después de clasificar y no establecen que las acciones hayan producido el acierto. Sí muestran la dificultad de asignar una calidad universal a cada opción: utilizar una respuesta correcta puede ser compatible con una valoración adecuada, mientras que solicitar otra respuesta a una IA deja pendiente el examen de ese nuevo contenido. La categoría elegida adquiere significado en relación con aquello que se pretende utilizar y con las razones, no recogidas aquí, de quien la selecciona.

Schemmer et al. (2023) permiten precisar qué observaciones serían necesarias para estudiar esa relación como dependencia apropiada. Su planteamiento considera el juicio inicial, la corrección del consejo recibido y la decisión posterior. Así puede examinarse si una persona adopta consejos correctos y mantiene su juicio frente a consejos incorrectos. Nuestro diseño recoge una clasificación y una intención posterior, sin observar el uso efectivo ni una decisión anterior al consejo. Las combinaciones encontradas no miden aquella dependencia, pero ofrecen situaciones concretas para una investigación posterior: observar qué comprueba quien decide consultar, qué razones sostiene quien utiliza directamente una respuesta y cómo se modifica la decisión cuando aparece información adicional.

Las asociaciones con los antecedentes de uso también deben interpretarse según la información que aportan esas variables. La verificación habitual presentó una asociación positiva con el desempeño, aunque su valor p ajustado fue 0,0985; las frecuencias de uso académico y disciplinar, la confianza general y el semestre tampoco mostraron evidencia estadística de asociación tras el ajuste. Estos resultados no establecen que la experiencia o la formación carezcan de importancia. La frecuencia indica cuánto se declara utilizar o comprobar, pero informa poco sobre qué se hizo, con qué exigencia y mediante qué procedimientos. El semestre sitúa una trayectoria curricular sin medir el dominio de los contenidos evaluados. La incertidumbre de las asociaciones debe mantenerse junto con esta limitada descripción de la experiencia, evitando convertirla en una explicación concluyente de los aciertos.

Implicaciones formativas y alcance del estudio

Una propuesta educativa derivada de estos resultados consistiría en hacer visible el trabajo que la clasificación deja sin registrar. Ante respuestas correctas, insuficientes y erróneas, el estudiante podría identificar la afirmación decisiva y explicar cómo se relaciona con la consulta. En una respuesta correcta tendría que justificar por qué la información basta para resolverla; ante un error, precisar qué parte del razonamiento compromete y qué cambiaría al corregirlo. La utilidad de esta actividad estaría en someter a examen las razones de la elección, especialmente

cuando dos estudiantes llegan a categorías distintas o coinciden en una categoría por motivos diferentes. El estudio permite fundamentar la pertinencia de esa propuesta a partir de las distinciones observadas, pero su eficacia requiere una intervención que la evalúe.

La comprobación podría incorporarse a la misma actividad como un procedimiento ejecutado. Seleccionar una referencia, rehacer un cálculo o contrastar una afirmación permitiría registrar qué evidencia utiliza el estudiante y si esta modifica su juicio inicial. Sabbione y Patrignani (2024) ofrecen un antecedente disciplinar al proponer actividades de Ciencia de los Alimentos en las que las respuestas de ChatGPT se comparan con libros y trabajos científicos. Ese planteamiento permite situar la verificación dentro del aprendizaje y vincularla con la justificación de una decisión. Su pertinencia didáctica no demuestra que una actividad equivalente mejoraría las puntuaciones de esta muestra; ofrece una vía para estudiar cómo los conocimientos intervienen cuando la revisión debe llevarse a cabo y explicarse.

CONCLUSIONES

La autopercepción de capacidad para evaluar respuestas de IA mostró una asociación positiva débil con el acierto en cinco casos de Ingeniería en Alimentos. La incertidumbre de la estimación impide establecer una correspondencia estrecha entre ambas medidas y tampoco permite concluir que sean independientes. En esta muestra, la valoración personal aporta información que necesita examinarse junto con tareas de desempeño, conservando el alcance específico de cada registro.

La distribución de clasificaciones permitió reconocer diferencias que el total de aciertos no expresa por sí solo. Algunos estudiantes señalaron errores sin coincidir con el alcance previsto en la clave, mientras que otros consideraron incompletas respuestas clasificadas como suficientes. Estas elecciones sitúan el interés formativo en cómo se distinguen corrección, suficiencia y gravedad del error. La preferencia por consultar una fuente confiable o a un especialista añade una intención de comprobación, cuyo desarrollo efectivo permanece desconocido.

En consecuencia, la formación del juicio técnico requiere hacer visibles las razones con las que se acepta una respuesta, se propone corregirla o se mantiene pendiente de verificación. Los resultados permiten fundamentar el interés de actividades que soliciten esas razones y observen la comprobación, sin demostrar todavía su eficacia. Avanzar en esa dirección permitiría estudiar cómo el conocimiento del estudiante sostiene sus decisiones, más allá de la clasificación que finalmente elige.

REFERENCIAS

Alkabaa, A. S., & Alamri, N. M. (2025). Exploring the effectiveness of generative AI as a learning tool in engineering education: An analysis of student experiences and perceptions. *Computer Applications in Engineering Education, 34*, e70110. https://doi.org/10.1002/cae.70110

Álvarez Ariza, J., Benítez Restrepo, M., & Hernández Hernández, C. (2025). Generative AI in engineering and computing education: A scoping review of empirical studies and educational practices. *IEEE Access, 13*, 30789–30810. https://doi.org/10.1109/ACCESS.2025.3541424

Amoozadeh, M., Daniels, D., Nam, D., Kumar, A., Chen, S., Hilton, M., Ragavan, S. S., & Alipour, M. A. (2024). Trust in generative AI among students: An exploratory study. En *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. Association for Computing Machinery. https://doi.org/10.1145/3626252.3630842

Borowczak, M., & Borowczak, A. C.-B. (2026). Trust but verify: Engineering students' adoption of generative AI and pedagogical implications across disciplines. *Proceedings of the American Society for Engineering Education Annual Conference & Exposition* (Paper ID 49945).

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans, 1*, 100014. https://doi.org/10.1016/j.chbah.2023.100014

Dai, Y. (2025). Why students use or not use generative AI: Student conceptions, concerns, and implications for engineering education. *Digital Engineering, 4*, 100019. https://doi.org/10.1016/j.dte.2024.100019

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence, 5*, 100165. https://doi.org/10.1016/j.caeai.2023.100165

Hou, C., Zhu, G., & Sudarshan, V. (2025). The role of critical thinking on undergraduates' reliance behaviours on generative AI in problem-solving. *British Journal of Educational Technology, 56*, 1919-1941. https://doi.org/10.1111/bjet.13613

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence, 9*, 100436. https://doi.org/10.1016/j.caeai.2025.100436

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. En *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

López-Vargas, A., Rodríguez-Vidal, J., García-Beltrán, A., Vigara, P., & de León-Torres, L. (2026). Performance, interaction, and ethical evaluation in the use of generative artificial intelligence in engineering education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-026-14112-y

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing, 1*(4), Article 26. https://doi.org/10.1145/3696449

Mohammed, Y., Qandil, H., & Lund, B. (2026). Engineering student perceptions of generative AI use in learning, trust, judgment and information seeking. *Discover Artificial Intelligence, 6*, 1004. https://doi.org/10.1007/s44163-026-02086-2

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology, 58*, 504-509. https://doi.org/10.1002/pra2.487

Pennells, J., Watkins, P., Bowler, A. L., Watson, N. J., & Knoerzer, K. (2025). Mapping the AI landscape in food science and engineering: A bibliometric analysis enhanced with interactive digital tools and company case studies. *Food Engineering Reviews, 17*, 465-489. https://doi.org/10.1007/s12393-025-09413-w

Sabbione, A. C., & Patrignani, M. (2024). Usos de la IA en la enseñanza de la ciencia de los alimentos: Potencialidades, limitaciones y su aplicación en el desarrollo de propuestas didácticas. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, número especial 37, 110-116. https://doi.org/10.24215/18509959.37.e10

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. En *Proceedings of the 28th International Conference on Intelligent User Interfaces*. Association for Computing Machinery. https://doi.org/10.1145/3581641.3584066

Shallari, I., Gallo, V., Carratù, M., Hussain, M., Krapohl, D., & Mousavirad, S. J. (2026). Perception-performance gap in generative AI: An exploratory study across two engineering education contexts. *Education Sciences, 16*, 803. https://doi.org/10.3390/educsci16050803