Evaluación de la calidad de historias de usuario usando modelos de lenguaje de gran tamaño: un estudio en la industria

Erika Hernández-Agüero[1,2], Christian Quesada-López[1,2], José P. Chaves-Sánchez[2]

[1]Posgrado en Computación e Informática, Escuela de Ciencias de la Computación e Informática – Universidad de Costa Rica (UCR) –San José –Costa Rica

[2]Dirección de Tecnologías de Información y Comunicaciones – Universidad Estatal a Distancia (UNED) – San José – Costa Rica

{erika.hernandezaguero,cristian.quesadalopez}@ ucr.ac.cr,
{ehernandez,cquesadal,jpchaves}@uned.ac.cr

Abstract. *The specification and maintenance of high-quality user stories are critical and challenging activities in agile software development, due to the dynamic nature of projects, the ambiguity of natural language, and the effort required for manual evaluations. This study investigates the use of large language models (LLMs) to assess the quality of user stories in an industrial software project, using the criteria defined by the INVEST framework. The performance of three LLM tools is compared with two evaluations conducted by requirements engineering experts. The results indicate that LLMs have the potential to support the automated assessment of user stories based on INVEST.*

Keywords: software requirements, quality, NLP4RE, automation, LLM, user stories, INVEST

Resumo. *La especificación y mantenimiento de historias de usuario de alta calidad son actividades críticas y desafiantes en el desarrollo de software ágil, debido a la naturaleza dinámica de los proyectos, la ambigüedad del lenguaje natural y el esfuerzo requerido para las evaluaciones manuales. Este estudio investiga el uso de modelos de lenguaje de gran tamaño (LLM) para evaluar la calidad de historias de usuario en un proyecto de software de industria, utilizando los criterios del marco INVEST. Se compara el desempeño de tres herramientas LLM con dos evaluaciones de personas expertas en ingeniería de requisitos. Los resultados indican que los LLM tienen potencial para apoyar la evaluación automatizada de historias basado en INVEST.*

Palabras clave. requerimientos de software, calidad, NLP4RE, automatización, LLM, historias de usuario, INVEST

1. Introducción

La Ingeniería de Software Continua proporciona una perspectiva integral del flujo de trabajo que vincula la estrategia de negocio, el desarrollo de software y las operaciones,

un enfoque conocido como BizDevOps [Fitzgerald and Stol 2017]. Esta disciplina abarca elementos clave como la mejora continua, la innovación, la adopción de metodologías ágiles en el desarrollo de software y la automatización de procesos [Fitzgerald and Stol 2017; Bosch 2014]. La integración fluida entre el negocio y el desarrollo de software es crucial, ya que una gestión eficaz de los requisitos permite la entrega de productos alineados con las necesidades de las partes interesadas y los objetivos estratégicos de la organización [Bourque and Fairley 2014]. Los enfoques de automatización se han enfocado en las áreas de DevOps, pero se requieren esfuerzos en la automatización de procesos que involucren el área Biz [Fitzgerald and Stol 2017]. En este contexto, el Procesamiento de Lenguaje Natural (NLP) para la Ingeniería de Requisitos (NLP4RE) ha surgido como un área de investigación que aplica sus técnicas NLP para abordar desafíos en la ingeniería de requisitos, tales como la generación, captura, análisis, validación y gestión de requisitos redactados en lenguaje natural [Marques, Silva, and Bernardino 2024; Belzner, Gabor and Wirsing 2024]. Es necesario valorar estos enfoques en proyectos reales, para determinar su potencial con el fin de apoyar este tipo de tareas en la industria.

En enfoques ágiles como Scrum, los requisitos suelen expresarse mediante historias de usuario organizadas en la pila de producto, las cuales se refinan y priorizan para facilitar eventos clave, como la planificación, la revisión y retrospectiva en cada iteración [Schwaber and Sutherland 2020]. Al redactarse en lenguaje natural, las historias de usuario incorporan una de las formas más utilizadas para representar requisitos, y capturar de forma concisa las necesidades del usuario final [Zhang et al. 2024]. Este enfoque busca responder a la evolución continua del negocio y acelerar la entrega de valor. La evaluación de la calidad de estas historias representa un desafío complejo, pues requiere adaptarse a cambios frecuentes en las especificaciones, además de analizar atributos clave como la completitud, la verificabilidad y la ambigüedad inherente al lenguaje natural, además de la variabilidad en los requisitos [Parra et al. 2015; Bosch 2014; Subedi et al. 2021]. La calidad de las historias de usuario puede evaluarse mediante marcos como INVEST, el cual define criterios específicos que deben cumplir que sean: independientes, negociables, valiosas, estimables, pequeñas y testeables [Subedi et al. 2021; Ronanki, Berger and Horkoff 2023; Zhang et al. 2024]. Se ha propuesto la adopción de técnicas de aprendizaje automático, principalmente, modelos de lenguaje de gran tamaño (LLM), como una herramienta de apoyo para las personas expertas en ingeniería de requisitos [Parra et al. 2015; Subedi et al. 2021; Ronanki et al. 2023; Krishna et al. 2024; Bosch 2014; Zhang et al. 2024] con el potencial de apoyar los procesos de evaluación y mejora de las historias de usuario, utilizando INVEST [Subedi et al. 2021; Zhang et al. 2024].

Este estudio explora el potencial uso de los LLMs en el proceso de automatización de la evaluación de la calidad de historias de usuario, integrando el marco INVEST. Se analizaron los resultados obtenidos mediante ChatGPT, Gemini y Copilot contrastándolos con la revisión de dos personas expertas en el contexto de un proyecto de desarrollo de un departamento de TI de una institución pública. El enfoque de estudio proporciona herramientas de apoyo para personas expertas, con el fin de facilitar la detección de inconsistencias, ambigüedades y deficiencias en los requisitos. Para analistas de nivel inicial, el uso de LLMs podría proporcionar una guía estructurada basada en INVEST, facilitando la adopción de buenas prácticas y la redacción de

historias de usuario alineadas con principios ágiles, basado en el contexto de los proyectos de la organización. Para personas expertas en ingeniería de requisitos, estas herramientas podrían proporcionar apoyo en la revisión de especificaciones complejas y el alto volumen de requisitos. El artículo reporta la experiencia en el uso de las herramientas y los resultados obtenidos de la comparación. La unidad de desarrollo de la organización ha implementado procesos de mejora en las actividades de requisitos [Hernandez, Quesada and Chaves 2024] y busca automatizar las evaluaciones de calidad de requisitos para apoyar a las personas expertas. Se analiza la consistencia, diferencias y coincidencias entre las evaluaciones realizadas por las personas expertas y las generadas por los LLMs. Asimismo, se estudian variaciones en la evaluación de cada criterio del marco INVEST, determinando hasta qué punto los LLMs pueden complementar el juicio experto humano en la validación de historias de usuario. Además, se espera que este estudio aporte en la discusión sobre el uso de estas herramientas en dichas tareas de la ingeniería de requisitos y el alcance de los LLMs para valorar la calidad de las historias de usuario utilizando INVEST, en el contexto de un proyecto de industria.

2. Trabajo Relacionado

Distintos estudios han explorado la automatización de tareas de la ingeniería de requisitos. Los estudios recientes se centran en el uso de modelos de lenguaje de gran tamaño (LLMs) para mejorar la eficiencia y calidad en diversas fases del ciclo de vida del desarrollo de software. Estos trabajos abordan la generación y validación de documentos de especificaciones de requisitos de software, la asistencia en la elicitación de requisitos, la mejora de historias de usuario en entornos ágiles y el apoyo en diseño de sistemas, generación de código y pruebas.

[Krishna et al. 2024] analizó cómo los LLMs, específicamente GPT-4 y CodeLlama, pueden contribuir a la generación y validación de Especificaciones de Requisitos de Software (SRS). Para esto, compararon documentos generados por estos modelos con los elaborados por ingenieros de software de nivel inicial, evaluando atributos de calidad como claridad, consistencia y completitud. Los resultados mostraron que los LLMs pueden generar documentos de calidad comparable, con una reducción significativa del tiempo de elaboración. [Ronanki et al. 2023] exploró el potencial de ChatGPT para asistir en la elicitación de requisitos. Se formularon preguntas específicas para obtener requisitos tanto de ChatGPT como de expertos humanos, y se evaluaron en atributos de calidad como abstracción, atomicidad, consistencia, corrección, claridad (no ambigüedad), comprensibilidad y factibilidad. Los hallazgos indicaron que ChatGPT destaca en abstracción, atomicidad y comprensibilidad, aunque presenta deficiencias en claridad y factibilidad.

[Zhang et al. 2024] analizó el uso del sistema ALAS, que utiliza agentes basados en LLM (GPT-3.5 y GPT-4) para mejorar historias de usuario en proyectos ágiles. La metodología consistió en la implementación de agentes autónomos que colaboran para refinar historias de usuario, evaluadas posteriormente mediante el marco INVEST. Los resultados mostraron mejoras en claridad y alineación con objetivos de negocio, aunque persisten desafíos relacionados con la complejidad de las historias generadas. Por su parte, [Belzner et al. 2024] revisó los beneficios y desafíos de integrar LLMs en la ingeniería de software. Se evaluó el apoyo de modelos como ChatGPT y Bard en la

ingeniería de requisitos, diseño, pruebas y generación de código. La metodología incluyó interacciones iterativas con los LLMs para desarrollar requisitos y proponer soluciones, identificando tanto el potencial para mejorar la productividad como las limitaciones en la integración con procesos existentes. Finalmente, estudios previos han aplicado técnicas de aprendizaje automático para evaluar automáticamente la calidad de requisitos, emulando el juicio de expertos con base en atributos IEEE 830, alcanzando hasta un 86,1 % de precisión [Parra et al. 2015]. En el caso de historias de usuario, se ha evaluado la calidad mediante los atributos Testeable y Valiosa del marco INVEST, destacando el clasificador de árboles de decisión y el uso de SMOTE para mejorar el recall sin comprometer la precisión [Subedi et al. 2021]. El presente estudio integra la evaluación de historias de usuario basándose en los seis atributos del marco INVEST, y busca aportar evidencia empírica sobre la capacidad de los LLMs para evaluar la calidad de los requisitos en entornos ágiles de un proyecto en la industria.

3. Metodología

El estudio investiga el uso de herramientas LLM para evaluar las historias de usuario en un proyecto de software de industria. Se evalúa su efectividad para la automatización de la evaluación en el contexto de desarrollo ágil de software. Las evaluaciones obtenidas a partir del uso de las herramientas LLMs son comparadas con las realizadas por dos personas expertas en ingeniería de requisitos del proyecto seleccionado. Para evaluar la calidad de los requisitos se utiliza el marco INVEST. El trabajo plantea la siguiente pregunta de investigación: *RQ1: ¿Cómo se comparan los resultados de la evaluación de calidad de las historias de usuario realizadas por las personas ingenieras de requisitos en un proyecto ágil con las obtenidos con el apoyo de las herramientas LLM?* En el análisis se busca identificar las diferencias y similitudes que existen entre las evaluaciones obtenidas de las herramientas LLM y de las personas expertas de acuerdo al marco INVEST.

3.1. Proyecto, Conjunto de Datos, Participantes y Herramientas

El proyecto analizado corresponde al desarrollo de una aplicación para la gestión de notas de cursos universitarios, en la cual interactúan los roles de administración, tutoría y coordinación de cátedra. El desarrollo se realiza en un entorno ágil, utilizando Azure DevOps Server para gestionar los requisitos mediante historias de usuario organizadas en la pila de producto. El equipo de trabajo sigue formalmente el marco Scrum, y la persona administradora de la aplicación cumple también el rol de persona dueña del producto. La aplicación es utilizada por un total de 1,820 personas usuarias. En este estudio, se seleccionó un conjunto de 60 historias de usuario de una pila total de 112 registros de diferentes módulos funcionales, como: mantenimiento de usuarios, mantenimiento de instrumentos de evaluación, asignación de modelos de evaluación, asignación de tutores, ingreso de notas e historial de notas, entre otras funcionalidades. Las evaluaciones de calidad se realizaron tanto por personas expertas en ingeniería de requisitos como mediante herramientas basadas en LLM. Las personas expertas cuentan con más de 15 años de experiencia en desarrollo de software, dominio de metodologías ágiles y gestión de historias de usuario bajo el marco INVEST. Las herramientas LLM utilizadas fueron ChatGPT Pro 4, Gemini Advanced 1.5 y Microsoft Copilot, todas en sus versiones de pago. La selección de estas herramientas se basa en su uso extendido en

la industria y su aplicación en investigaciones previas relacionadas con la evaluación de requisitos.

Disponibilidad de artefactos: por tratarse de un estudio realizado en la industria no es posible compartir los datos.

3.2. Marco de Evaluación INVEST

La Tabla 1 muestra el marco de evaluación INVEST, basado en [Zhang et al. 2024; CertiProf 2022]. Para evaluar la calidad de las historias de usuario, se aplica este marco asignando a cada historia una puntuación en una escala de Likert del 1 al 5, donde 1 indica un fuerte desacuerdo con el cumplimiento del criterio, y 5 representa un fuerte acuerdo.

Table 1. Variables to be considered on the evaluation of interaction techniques

ID	Declaraciones y las características correspondientes al marco INVEST para las historias
I	Independiente: es autónoma, comprensible por sí sola y no depende de otras historias.
N	Negociable: tiene un nivel de detalle adecuado que permite ajustes y priorización.
V	Valiosa: proporciona un beneficio claro y medible para el sistema o usuario final.
E	Estimable: es técnicamente alcanzable y se puede estimar en términos de tiempo o esfuerzo.
S	Pequeña: presenta un alcance acotado para completarse dentro de un sprint (máximo 4 semanas).
T	Testeable: incluye criterios de aceptación que permiten validar su implementación con pruebas.

Durante el proceso de evaluación de calidad, cada persona experta evalúa la calidad de cada historia de usuario de manera independiente utilizando un instrumento estructurado, que contiene el conjunto de historias por evaluar con los espacios para la calificación de cada criterio INVEST con la escala de Likert y un espacio para sus observaciones. Cada una de las personas ingenieras de requisitos asignó la calificación de 1 a 5 siguiendo la escala por criterio. Por separado, la persona investigadora ejecuta en cada herramienta LLM el *prompt* de evaluación de calidad de las historias de usuario, obteniendo las calificaciones correspondientes para el mismo conjunto de requisitos. El conjunto de datos fue segmentado en 12 subconjuntos por las limitaciones de cantidad de tokens de las herramientas. El proceso de evaluación se basa en lo recomendado por [Zhang et al. 2024].

3.3. Implementación del Prompt

La implementación del *prompt* se basa en las recomendaciones de [Zhang et al. 2024; Ronanki et al. 2023; Krishna et al. 2024]. Se construye a partir de dos etapas de trabajo: etapa de preparación y etapa de ejecución. En la etapa de preparación se crean los *prompts* y se considera la siguiente estructura para el *prompt* inicial y de seguimiento:

$$\textit{Initial Prompt}_i = \textit{Profile}_i + \textit{Task} + \textit{Context} + \textit{Subtask}_i,\ (1 \le i \le k)$$

$$\textit{Follow-up Prompt}_i = \textit{Subtask}_i + \textit{Response}_{i-1},\ (i > k)$$

Donde para el *prompt* inicial:

Profile$_i$: Describe el rol, habilidades o responsabilidades específicas.
Task: Es una descripción amplia del objetivo que se busca alcanzar.
Context: Proporciona detalles sobre el entorno, limitaciones o condiciones específicas.
Subtask$_i$: Es un paso específico dentro de la tarea general.

Para el *prompt* de seguimiento:

Subtask$_i$: Es un paso específico dentro de la tarea general.
Response$_{i-1}$: La respuesta generada tras completar la subtarea anterior, permite generar las instrucciones que simulan la interacción de diferentes actividades que lleva a cabo este evaluador.

En la etapa de ejecución se realizan las actividades relacionadas con la aplicación y calibración del *prompt* generado en el diseño en cada una de las herramientas seleccionadas ChatGPT, Gemini y Copilot. Esto permite la recolección de las evaluaciones generadas por cada LLM de manera automatizada [Zhang et al. 2024].

Evaluación de los Requisitos de Software

Primero, se definió un flujo de trabajo para la evaluación de la calidad de las historias de usuario utilizando el marco INVEST, como se muestra en la Figura 1. Este proceso puede ser ejecutado tanto por una persona analista, a través del perfil de una persona ingeniera de requisitos, quien, según el contexto del proyecto, realiza la evaluación utilizando el *prompt* diseñado.

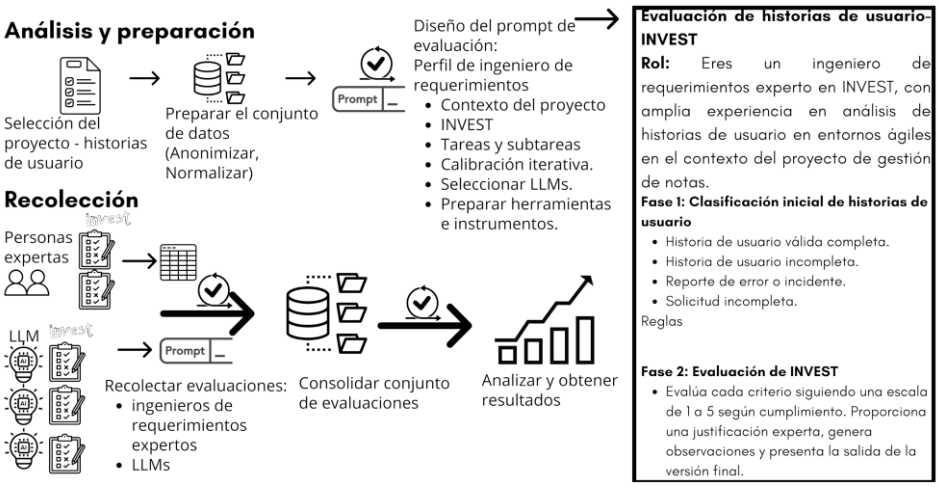


Figura 1. Evaluación de calidad de requisitos

Análisis y Preparación de los Datos

En la etapa de preparación para la recolección de datos, se llevaron a cabo actividades de preprocesamiento y normalización de las historias de usuario. Este proceso incluyó la eliminación de registros incompletos o sin información relevante, como aquellos creados en Azure DevOps que solo contenían un título sin descripción, casos generados como recordatorios de situaciones particulares que no representan funcionalidades reales, o registros correspondientes a ceremonias Scrum (planificación, refinamiento, revisión y retrospectiva). Asimismo, registros que representan funcionalidades específicas, y no cumplían con la estructura estándar de redacción de historias de usuario, fueron incluidos en el conjunto de datos con el objetivo de prever casos similares en otros proyectos y permitir que parte del proceso de validación considere la verificación de este tipo de historias. La etapa de anonimización de datos, implicó la detección y eliminación de información confidencial de la institución. Por ejemplo, se sustituyeron nombres de sistemas, identificadores personales, referencias a bases de datos y otros elementos vinculados a la infraestructura técnica del proyecto. Esto es fundamental para mitigar los riesgos asociados al uso de modelos externos o en la nube, y para garantizar el cumplimiento de los principios de seguridad y privacidad desde las

fases iniciales del análisis. La estructura estándar de la historia de usuario definida por la organización es la siguiente: *Como [rol] quiero [funcionalidad/requerimiento] para [beneficio/objetivo]*. Además, cada historia de usuario debe incluir criterios de aceptación, definidos bajo la estructura: Dado *[contexto inicial], cuando [acción], entonces [resultado esperado]*. Los registros que no contaban con criterios de aceptación definidos fueron etiquetados con la leyenda "[No se han definido AC]", con el propósito de que los LLMs evaluaran las historias tal como fueron registradas originalmente, sin inferir ni completar la información ausente durante el proceso de evaluación. El conjunto de datos final se conformó por tres tipos de registros: (1) historias de usuario con criterios de aceptación completos; (2) historias de usuario sin criterios de aceptación explícitos; y (3) funcionalidades del sistema que no siguen la estructura estándar de redacción.

Recolección de las Evaluaciones

Previo al proceso de recolección de datos, se llevaron a cabo pruebas para calibrar el *prompt,* ajustando las instrucciones de la tarea con el fin de asegurar que los resultados fueran comparables con las evaluaciones de las personas expertas. Como parte del proceso de calibración, se implementaron mejoras para la evaluación: garantizar la lectura y calificación individual de cada historia de usuario, asegurar que cada historia de usuario reciba su propia calificación incluyendo observaciones, estandarizar la estructura de salida de las evaluaciones para que cada historia de usuario ocupe una única fila en la tabla de evaluación, evitando duplicaciones y mejorando la claridad de los resultados. Asimismo, se definen instrucciones detalladas para el procesamiento de datos, incluyendo la lectura de archivos, el formato de despliegue de los resultados y la descarga de la evaluación.

Primero se define el perfil del evaluador y su rol dentro del contexto de evaluación de la calidad de requisitos, como se ilustra en la Figura 2. Se trata de un ingeniero de requisitos que desempeña funciones específicas y realiza tareas concretas en un entorno ágil de desarrollo de software. La fase 1 del proceso abarca las actividades relacionadas con los refuerzos en la evaluación que permite establecer las reglas adicionales. Por ejemplo: si una historia de usuario no cuenta con criterios de aceptación definidos debe recibir la calificación mínima en el criterio Testeable, verificar el cumplimiento del formato estándar de la historia de usuario válidas que deben cumplir con la descripción y criterios de aceptación, historias de usuario incompletas que carecen de una descripción estructurada o criterios de aceptación definidos, registros que no corresponden a requisitos (reportes de errores o incidentes) que deben recibir la calificación más baja e incluir la observación que corresponde a otro tipo gestión para esos elementos de trabajo. Asimismo, se incorporaron ejemplos historias de usuario habilitadoras, las cuales también son relevantes en el contexto de evaluación como se ilustra en la Figura 3. La fase 2 del proceso permite realizar las evaluaciones de manera que se asigne la calificación por historia de usuario. Se realizó una segmentación en subgrupos de 5 historias de usuario por ejecución, para evitar exceder los límites de procesamiento de los modelos, facilitar la revisión y validación progresiva de las respuestas generadas y aplicar calibraciones en tiempo real para mejorar la precisión del análisis. Una vez obtenidas las evaluaciones de los LLM se integran en el consolidado junto con las evaluaciones proporcionados por los expertos para proceder con el análisis de resultados como se muestra en la Figura 3.

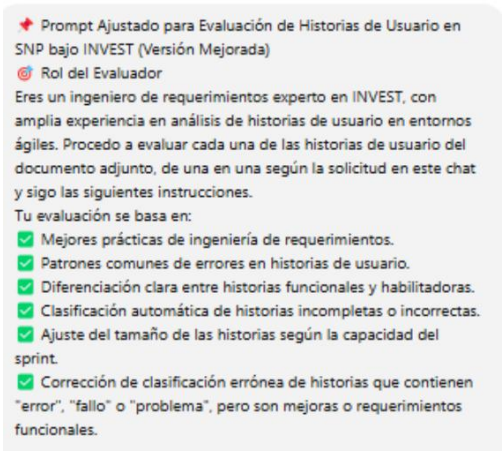
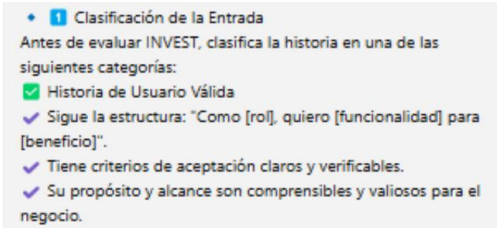


Figura 2. Prompt- Rol evaluador



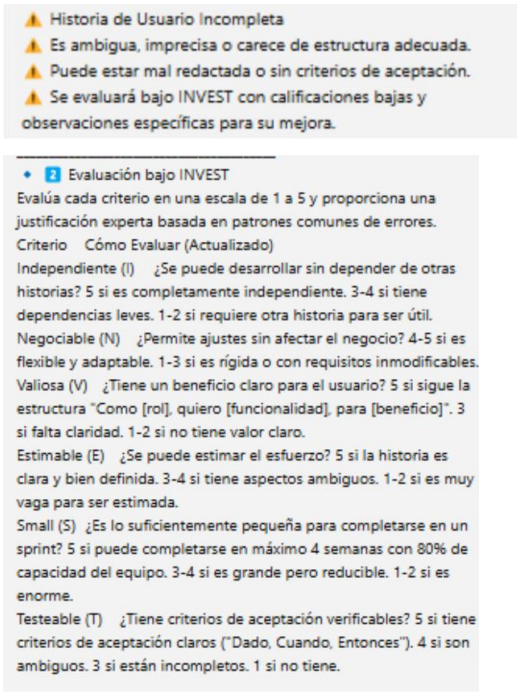


Figura 3. Prompt-Fases

4. Análisis de Resultados

El estudio compara las evaluaciones de historias de usuario bajo el marco INVEST, analizando las diferencias entre las evaluaciones realizadas por las personas expertas (Experto1, Experto2) y las herramientas LLM (ChatGPT, Gemini, Copilot). Se procesaron 60 historias de usuario, asignando calificaciones en seis criterios: I: Independiente, N: Negociable, V: Valioso, E: Estimable, S: Pequeña y T: Testeable.

La Figura 4 presenta las diferencias en las evaluaciones de cada una de las 60 historias de usuario, comparando los resultados de los dos expertos humanos con los obtenidos por las herramientas LLM. Se observan agrupamientos que reflejan tanto similitudes como variabilidad en las calificaciones asignadas en la escala de 1 a 5. En cada caso, se identifican concordancias y discrepancias respecto al juicio experto, lo que permite analizar la capacidad de los modelos para replicar evaluaciones humanas. Los resultados indican que los LLMs, en particular ChatGPT y Copilot, presentan un desempeño favorable bajo el marco INVEST, aunque tienden a sobrestimar atributos como Negociable y Valiosa. Estas diferencias resaltan la necesidad de calibrar y ajustar los modelos para mejorar su alineación con la perspectiva de expertos, especialmente en contextos ágiles donde la evaluación precisa de historias de usuario es clave para el éxito de los proyectos. El análisis de las 60 historias evidencia variabilidad y similitudes en los distintos criterios INVEST. La Figura 5 amplía este análisis mediante el cálculo de diferencias absolutas [Roumeliotis, Tselikas y Nasiopoulos, 2024], que permiten cuantificar la magnitud de las discrepancias en la calificación. Estas se obtienen al calcular el valor absoluto de la diferencia entre la puntuación asignada por una persona experta y la otorgada por un LLM, para cada historia y para cada criterio del marco INVEST.

Se expresa como:

$$Diferencia\ absoluta = |Evaluación_{Experto} - Evaluación_{LLM}|$$

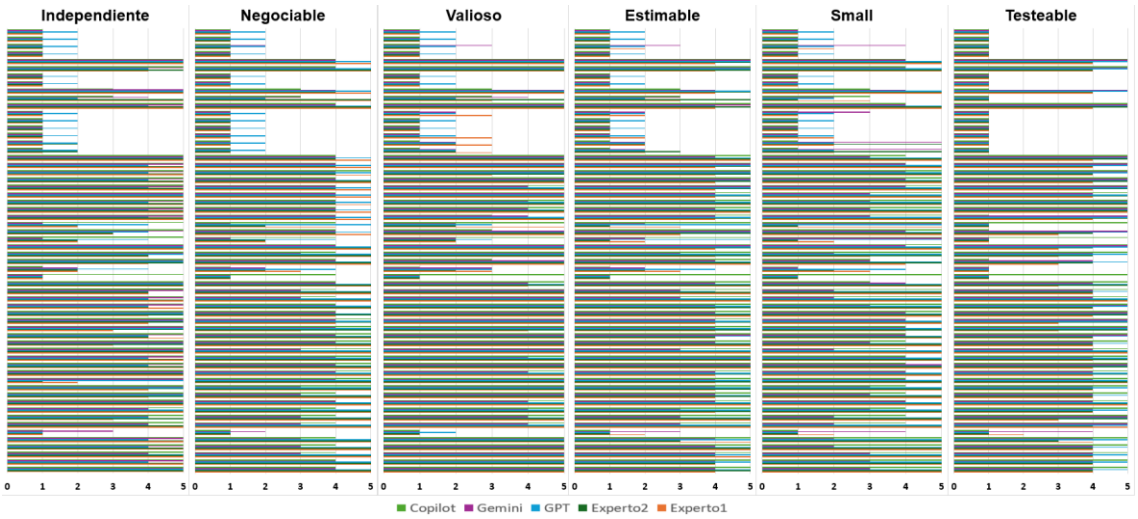


Figura 4. Evaluación de historias de usuario

El uso de diferencias absolutas ofrece una forma objetiva de medir la discrepancia, ya que elimina la influencia de la dirección del error, es decir, no importa si el LLM sobreestima o subestima en comparación con el experto, lo relevante es la magnitud de la diferencia. Esto facilita la comparación entre pares de evaluadores, permitiendo identificar tanto diferencias sistemáticas como casos atípicos que podrían reflejar inconsistencias significativas en la interpretación de los criterios INVEST.

La Figura 5 presenta la distribución de las diferencias absolutas entre las personas expertas (Experto 1 y Experto 2) y los modelos LLM (ChatGPT, Gemini y Copilot), revelando patrones relevantes en las evaluaciones de los criterios del marco INVEST. En general, la mayoría de las diferencias se concentra en el rango de 0 a 2, lo que sugiere una alta consistencia entre las evaluaciones humanas y automatizadas. Esta alineación es particularmente evidente en los criterios Independiente y Valioso, donde se observa una menor variabilidad y escasa presencia de valores atípicos. En estos casos, la objetividad de los criterios parece facilitar su interpretación por parte de los modelos. En contraste, los criterios Pequeña (S) y Testeable (T) presentan las mayores diferencias absolutas, lo que sugiere que los LLM encuentran más dificultades para evaluarlos adecuadamente. En el criterio Pequeña, la complejidad radica en estimar correctamente el tamaño o granularidad de las historias, mientras que en Testeable, los desafíos están vinculados a la ausencia o ambigüedad de los criterios de aceptación. Las comparaciones entre Experto 2 y Gemini, y entre Experto 1 y Copilot, evidencian las mayores discrepancias en estos dos atributos, reflejando limitaciones específicas en su interpretación. Al comparar el rendimiento general, ChatGPT mostró mayor consistencia con las evaluaciones humanas, mientras que Gemini y Copilot presentaron mayor dispersión, sobre todo en los criterios más complejos. Estos resultados indican que, si bien los LLM pueden aproximarse al juicio experto, es necesario fortalecer la calibración en criterios como Pequeña, Negociable y Testeable, a través de instrucciones más específicas y refuerzos dirigidos

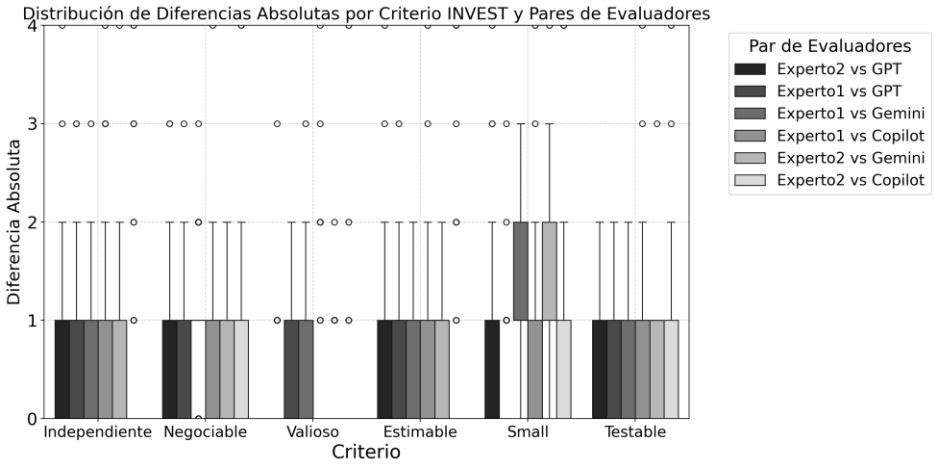


Figura 5. Distribución de diferencias absolutas

Desde la perspectiva del proyecto, los resultados indican que estas herramientas pueden apoyar los procesos de evaluación de requisitos, con supervisión de las personas expertas, aunque se requiere mejorar su precisión y consistencia, enfocándose en su capacidad para interpretar de forma más alineada con el juicio experto en algunos criterios. Se espera que, al ofrecer un contexto más amplio, junto con ejemplificaciones de evaluaciones previas realizadas por personas expertas, los LLM mejoren su rendimiento, considerando además el balance con la capacidad de procesamiento de tokens de cada herramienta. Para analizar la consistencia, la Figura 6 presenta la matriz de correlación de Spearman, que muestra asociaciones positivas entre las evaluaciones de las personas expertas y cada LLM. La correlación más alta se observa entre Experto1 y Experto2 ($\rho = 0.80$), lo que indica una fuerte correlación en sus evaluaciones. ChatGPT presenta correlaciones elevadas con ambas personas expertas ($\rho = 0.75$ con Experto1 y $\rho = 0.77$ con Experto2), lo que sugiere una mayor consistencia que permite utilizar estas medidas como base para la comparación. En contraste, Copilot y Gemini muestran correlaciones moderadas, siendo más bajas con Experto1 ($\rho = 0.56$ y $\rho = 0.59$, respectivamente) y ligeramente superiores con Experto2. Estos resultados indican que, aunque ambos LLM logran un alineamiento con los expertos, ChatGPT muestra mayor consistencia. Por otro lado, se analizó la fiabilidad de la consistencia de las escalas de medición mediante el Alfa de Cronbach [Essel et al. 2024] para cada uno de los pares evaluadores:

$$\alpha = k/(k-1)\,(1 - \Sigma/(\sigma\_t\text{^}2)\,\sigma\_i\text{^}2)$$

k = Número de elementos evaluados (criterios I N V E S T).

σ_t^2 = Varianza de cada criterio evaluado (I N V E S T).

σ_i^2 = Varianza total del conjunto de evaluaciones.

En este caso, un valor alto en la consistencia de las evaluaciones sugiere que la herramienta LLM pueden apoyar la evaluación automatizada de la calidad de los requisitos, al mostrar concordancia con las evaluaciones expertas. Los valores bajos indican diferencias en la interpretación de los criterios y la necesidad de mejorar la calibración. Se establece que un valor de $\alpha \geq 0.9$ representa una alta consistencia entre las evaluaciones expertas y las generadas por los LLM, lo que sugiere que los LLM

pueden replicar el juicio humano con un alto grado de confianza. Los valores por debajo de 0.9 reflejan diferencias en la forma en que las herramientas LLM evalúan los criterios, evidenciando la necesidad de ajustes en su calibración. La Figura 7 presenta los valores de consistencia calculados mediante el Alfa de Cronbach. Los resultados muestran una consistencia aceptable en la mayoría de los criterios INVEST, aunque se identifican áreas específicas de mejora. En particular, la herramienta Gemini presenta los valores más bajos de consistencia en el criterio Pequeña (S), con $\alpha = 0.61$ y $\alpha = 0.63$ frente a las evaluaciones de los expertos, lo que indica una mayor variabilidad en sus juicios. En particular, ChatGPT y Copilot presentan la mayor consistencia con las personas expertas, lo que sugiere que sus evaluaciones son más estables y confiables. En contraste, Gemini evidencia mayores diferencias, lo que indica la necesidad de ajustes en su calibración.

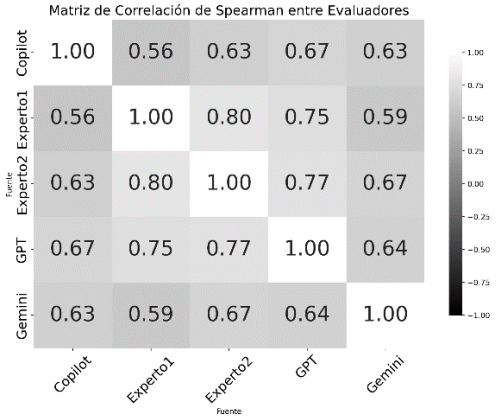


Figura 6. Correlación entre evaluadores

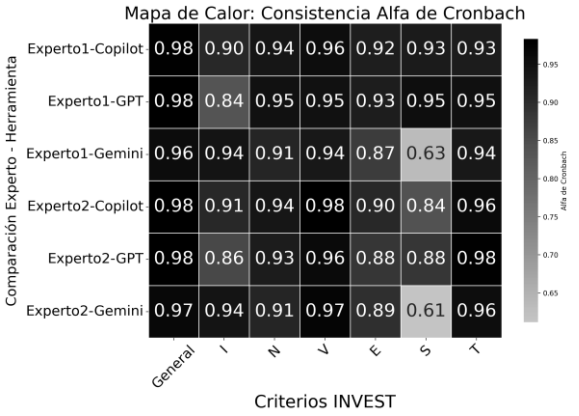


Figure 7. Consistencia de las evaluaciones

La prueba Kruskal-Wallis, permitió evaluar las diferencias globales, obteniendo un estadístico de 63.8148 y un p-valor de 0.0000, lo que confirma diferencias significativas entre los evaluadores. Previamente, se verificaron los supuestos de normalidad mediante las pruebas de Kolmogorov-Smirnov y Shapiro-Wilk, obteniendo p-valores de 0.0000 en todos los casos, lo cual indica que las distribuciones no siguen una normalidad estadística. La prueba de Homogeneidad de Levene (estadístico = 2.9207, $p = 0.0201$) mostró diferencias en la varianza entre los grupos, lo que justifica el uso de pruebas no paramétricas. Como análisis post hoc, se aplicó la prueba de Dunn con corrección de Bonferroni la cual reveló diferencias significativas entre las evaluaciones expertas y los LLM. Sin embargo, esto se debe a las diferencias que se han presentado en las evaluaciones de Gemini con todos los evaluadores ($p < 0.01$), lo que indica que su enfoque de evaluación difiere del criterio experto y de otros LLM. Copilot no presentó diferencias significativas con las personas expertas, lo que sugiere una mayor alineación con la evaluación experta. ChatGPT, mostró diferencias con una evaluación experta ($p = 0.000573$), pero no con la otra persona experta, ni con Copilot, lo que indica una alineación consistente con la evaluación humana. Los resultados sugieren que, para una automatización alineada con el juicio de las personas expertas, ChatGPT y Copilot son más consistentes. Mientras que Gemini requiere mejoras para alinearse a los criterios humanos.

Las diferencias significativas ($p < 0.0001$) en varias comparaciones refuerzan la necesidad de un análisis cualitativo para identificar los factores que influyen en la divergencia de Gemini. Estos hallazgos subrayan la importancia de calibrar los LLM para lograr evaluaciones más precisas y alineadas con los estándares utilizados en la ingeniería de requisitos. El coeficiente de concordancia de Kendall (W) obtenido fue de 0.7048, con un p-valor de 0.0000, que representa un acuerdo significativo entre los evaluadores, aunque no perfecto. Estas diferencias explican la variabilidad en la concordancia total y sugieren la necesidad de ajustes en la calibración Gemini para mejorar su alineación con el criterio experto. El análisis comparativo entre personas expertas y herramientas LLM mediante la prueba de Mann-Whitney revela que no se encontraron diferencias estadísticamente significativas en los criterios Independiente ($p = 0.4992$), Valioso ($p = 0.5693$) y Estimable ($p = 0.3365$), lo que sugiere que los LLM presentan un desempeño alineado con las personas expertas en estos criterios. Se detectaron diferencias significativas en los criterios Negociable ($p = 0.0016$), Pequeña ($p = 0.0009$) y Testeable ($p = 0.000056$), lo que indica que los LLM podrían evaluar estos criterios de manera diferente. Los hallazgos del estudio indican que los LLM pueden apoyar el criterio experto en el proceso de evaluaciones de calidad realizadas por personas expertas, para algunos de los criterios de evaluación.

Para analizar el grado de alineación entre las evaluaciones humanas y automáticas de la calidad de historias de usuario, se aplicó una clasificación semántica a las puntuaciones asignadas en cada criterio del marco INVEST. Dado que las calificaciones se otorgaron en una escala de 1 a 5, estas se agruparon en tres categorías: Positivo (4–5), Neutral (3) y Negativo (1–2). A partir de esta categorización, se definieron tres niveles de acuerdo semántico entre pares de evaluadores (persona experta y LLM): (1) Concordancia semántica, cuando ambos asignaron la misma categoría; (2) Desacuerdo leve, cuando uno evaluó como Neutral y el otro como Positivo o Negativo; y (3) Desacuerdo opuesto, cuando uno calificó como Positivo y el otro como Negativo. Esta metodología permitió cuantificar de manera cualitativa y cuantitativa la coincidencia conceptual entre evaluaciones, así como identificar discrepancias interpretativas significativas. Para garantizar la validez del análisis, se aplicó esta estrategia a las 60 historias de usuario, evaluadas por las dos personas expertas y los tres modelos LLM, de forma consistente para los seis criterios del marco INVEST.

Los resultados del análisis evidencian una alta concordancia semántica (superior al 80 %) en la mayoría de los pares de comparación, particularmente en los criterios Independiente, Pequeña y Testable, donde las definiciones más objetivas favorecen la consistencia en las evaluaciones. En contraste, se identificaron mayores niveles de desacuerdo leve y opuesto en los criterios Valioso, Estimable y Negociable, los cuales tienden a involucrar un mayor grado de subjetividad. El criterio Pequeña destaca por un desacuerdo opuesto más pronunciado en la comparación entre Experto 2 y Gemini, lo que sugiere diferencias en la interpretación del tamaño o granularidad adecuada de las historias de usuario. Estos resultados permiten identificar los criterios en los que los LLM se alejan del juicio experto, lo que constituye un insumo valioso para orientar el ajuste de los modelos y priorizar el análisis cualitativo de los casos más críticos.

Por ejemplo, al analizar los niveles de desacuerdo semántico entre las personas expertas y los LLM, el análisis detallado por historia de usuario revela que los casos con ID 4643 y 4821 presentan las mayores frecuencias de desacuerdo opuesto, con 18 y 13

casos, respectivamente. A estos les sigue el caso ID 4645, con un total de 8 desacuerdos (4 leves y 4 opuestos). Estos hallazgos sugieren que dichas historias podrían contener ambigüedades o múltiples interpretaciones asociadas a su claridad, valor o granularidad, por lo que fueron seleccionadas para un análisis cualitativo en profundidad.

La historia con ID 4643 corresponde a una solicitud de reporte de actividades evaluativas que no sigue la estructura estándar definida por la organización ni incluye criterios de aceptación. Esta falta de estructura influye a que los modelos LLM asignaran calificaciones más optimistas en varios atributos INVEST como Independiente, Negociable, Valiosa y Estimable en comparación con las personas expertas. No obstante, en el atributo Testeable, sí se observó consenso entre expertos y modelos, asignando una calificación mínima (valor 1), lo que indica el reconocimiento compartido de la ausencia de condiciones verificables. Un patrón similar se identificó en la historia ID 4821, también correspondiente a una solicitud de reporte, formulada sin estructura definida ni criterios de aceptación. Este comportamiento recurrente sugiere que los requisitos mal estructurados, especialmente aquellos relacionados con reportes sin elementos verificables, tienden a generar mayores divergencias en las evaluaciones automáticas, lo cual resalta la necesidad de establecer reglas claras de redacción para mejorar la consistencia y coherencia en los procesos de evaluación de calidad asistidos por LLM. El análisis agrupado por criterios INVEST revela que los criterios con mayor frecuencia de desacuerdo opuesto son Pequeña (11 casos), Negociable (10) y Estimable (8), lo que sugiere que los modelos LLM presentan mayores dificultades para alinearse con las valoraciones humanas en aspectos que requieren juicios más subjetivos, como la granularidad del requerimiento, su negociabilidad o estimabilidad. En contraste, el atributo Valioso registra la mayor cantidad de desacuerdos leves (17 casos), lo que indica que, aunque existen diferencias, estas tienden a ser más matizadas y posiblemente vinculadas a interpretaciones del valor percibido. En conjunto, estos hallazgos refuerzan la hipótesis de que ciertos atributos INVEST son más propensos a generar discrepancias semánticas, particularmente cuando son evaluados de manera automática, lo que subraya la importancia de complementar las técnicas automáticas con revisión cualitativa experta en contextos reales.

6. Conclusiones

Los resultados muestran el potencial de las herramientas LLM para el apoyo en la evaluación de la calidad de requisitos de software, específicamente en la valoración de historias de usuario siguiendo los criterios INVEST. Los LLMs presentan un rendimiento aceptable y que, en conjunto con las personas expertas en ingeniería de requisitos, pueden realizar evaluaciones de las historias de usuario. A pesar de la necesidad de supervisión humana, se observó que mediante procesos de calibración y mejora de *prompts* diseñados siguiendo los ejemplos prácticos indicados en otros estudios como [Krishna et al. 2024; Zhang et al. 2024; Ronanki et.al 2023], es posible ajustar el comportamiento de los modelos para que simulen de forma más precisa el criterio experto, integrando reglas específicas y ejemplos de evaluaciones previas.

Los modelos de lenguaje muestran un desempeño prometedor al alcanzar altos niveles de concordancia semántica con evaluadores humanos en la mayoría de los criterios del marco INVEST, lo que evidencia su potencial para apoyar procesos de evaluación automatizada de requisitos. Sin embargo, los resultados también revelan

oportunidades de mejora, especialmente en criterios como Pequeña, Negociable y Estimable, donde los modelos presentan mayor variabilidad frente a las evaluaciones humanas. Estas diferencias son más notorias en historias de usuario sin estructura estandarizada ni criterios de aceptación, lo que resalta la importancia de seguir buenas prácticas de redacción. En este contexto, el análisis cualitativo experto resulta fundamental para complementar las evaluaciones automáticas, aportando una visión contextual y especializada que fortalece la precisión, la coherencia y la confiabilidad del proceso de validación en entornos reales de ingeniería de requisitos.

En términos de alineación con el juicio experto, Copilot y ChatGPT mostraron un desempeño más cercano al de las personas evaluadores expertas. En contraste, Gemini presentó oportunidades de ajuste. Para mejorar los resultados se requieren análisis cualitativos para identificar los aspectos que requieren calibración. Por lo que, lo anterior implica el análisis de las diferencias de interpretación de los criterios INVEST entre los LLM y expertos, ya que en algunos criterios los modelos presentan similitudes, pero en otras diferencias, y hasta qué punto pueden ser ajustados para interpretar de manera más precisa. La necesidad de calibración y validación de personas expertos en la evaluación automatizada aún parece necesaria, identificando estrategias de ajuste y calibración continua para mejorar su alineación al juicio experto.

En cuanto a generalización de los resultados, las evaluaciones se realizaron en el contexto de un proyecto real con historias de usuario, y se reconoce que su generalización puede estar limitada a el dominio y estilo de documentación o estructuras organizativas. Sin embargo, el proceso parece ser aplicable en organizaciones con procesos similares. Como trabajo futuro se recomienda investigar estrategias híbridas combinando evaluaciones de LLM con validación posterior por parte de personas ingenieras de requisitos, según criterios de calidad de la organización. Por un lado, ampliar el conjunto de datos de evaluación, incluyendo historias de usuario, con el contexto particular de cada proyecto de la organización, con el fin de reducir el sesgo en las evaluaciones y por otro lado la posibilidad de ofrecer contexto más detallado durante las evaluaciones. Desde la perspectiva práctica, se debe evaluar la posibilidad de desarrollar herramientas de apoyo integradas en plataformas de gestión de requisitos, permitiendo una automatización controlada y supervisada. Como siguientes pasos se proyecta llevar a cabo un proceso iterativo de mejora de la evaluación de las historias y de mejora de la especificación de las historias de usuario evaluadas en tres fases y utilizando dos perfiles clave: la persona ingeniera de requisitos y la dueña del producto.

7. Agradecimientos

Esta investigación fue parcialmente apoyada por el proyecto No. 0013-2024 UNED, Nodos-Investiga DTIC-UNED, No. 834-C1-011 UCR, el Posgrado en Computación e Informática y el Centro de Investigaciones en Tecnologías de la Información y Comunicación de la UCR. Se agradecen los invaluables aportes de las personas colaboradoras de los equipos de la DTIC-UNED.

8. Referencias

Belzner, L., Gabor, T. and Wirsing, M. (2023) "Large language model assisted software engineering: prospects, challenges, and a case study", In: International Conference on

Bridging the Gap between AI and Reality (pp. 355-374). Springer Nature Switzerland.

Bosch, J. (2014) "Continuous software engineering: An introduction", In: Bosch, J. (eds) Continuous Software Engineering (pp. 3-13). Springer , Cham.

Bourque, P., and Fairley, R. (2014). Guide to the Software Engineering Body of Knowledge (Swebok). 335.

CertiProf. (2022), Scrum Master Professional Certificate.

Essel, H., Vlachopoulos, D., Essuman, A. and Amankwa, J. (2024) "ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs)", In: Computers and Education: Artificial Intelligence, 6, 100198.

Fitzgerald, B. and Stol, K. (2017) "Continuous software engineering: A roadmap and agenda", In Journal of Systems and Software, 123, 176-189.

Hernandez-Agüero, E., Quesada-López, C., & Chaves-Sánchez, J. P. (in press). "Integración de enfoques ágiles para el mejoramiento continuo de procesos de software", In: Proceedings of the 13th International Conference on Software Process Improvement (CIMPS 2024), Mérida, Yucatán, México. IEEE Xplore.

Krishna, M., Gaur, B., Verma, A. and Jalote, P. (2024). "Using LLMs in software requirements specifications: an empirical evaluation", In: 2024 IEEE 32nd International Requirements Engineering Conference (RE) (pp. 475-483). IEEE.

Marques, N., Silva, R. and Bernardino, J. (2024). Using chatgpt in software requirements engineering: A comprehensive review. Future Internet, 16(6), 180.

Parra, E., Dimou, C., Llorens, J., Moreno, V. and Fraga, A. (2015) "A methodology for the classification of quality of requirements using machine learning techniques", In: Information and Software Technology, 67, 180-195.

Ronanki, K., Berger, C. and Horkoff, J. (2023) "Investigating ChatGPT's potential to assist in requirements elicitation processes", In: 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (pp. 354-361). IEEE.

Roumeliotis, K., Tselikas, N. and Nasiopoulos, D. (2024). "LLMs in e-commerce: a comparative analysis of GPT and LLaMA models in product review evaluation", In: Natural Language Processing Journal, 6, 100056.

Schwaber, K. and Sutherland, J. (2020), La Guía de Scrum.

Subedi, I., Singh, M., Ramasamy, V. and Walia, G. (2021), "Classification of testable and valuable user stories by using supervised machine learning classifiers", In: 2021 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW) (pp. 409-414). IEEE.

Zhang, Z., Rayhan, M., Herda, T., Goisauf, M., and Abrahamsson, P. (2024). "Llm-based agents for automating the enhancement of user story quality: An early report", In: International Conference on Agile Software Development (pp. 117-126). Springer Nature Switzerland.