Technological-Educational Journal DOCENTES 2.0
LA2017000128 | ISSN: 2665-0266 | ISNI: 0000 0005 0409 1664 | Ringgold ID 608948 | ROR: 037jqrk16
RTED
Vol. 19 (2), 2026
June – November 2026
233-250

ARTICLE

Evaluation of a Prompt for Academic Manuscript Review in Higher Education

Evaluación de un Prompt para Revisar Manuscritos Académicos en Educación Superior

César Augusto Reyes Alvarado[1], Guillermo Enrique Cevallos López[2], Jeessikha Ubillús Reyes[3], Marushka Chocobar Reyes[4], Jesús Enrique Reyes Acevedo[5], Remy Felipe Barreda Medina[6] and Emilio Chocobar-Reyes[7]

[1]Universidad Nacional de Piura, email: creyesa@unp.edu.pe
[2]Universidad Nacional de Piura, email: gcevallosl@unp.edu.pe
[3]Universidad Nacional de Piura, email: jubillusr@unp.edu.pe
[4]Propósito Perú SAC, Peru, email: mchocobar@propositoperu.pe
[5]Universidad Nacional Autónoma del Alto Amazonas, email: jreyes@unaaa.edu.pe
[6]OK Computer EIRL, email: cgerencia@okcomputer.com.pe
[7]Escuela de Negocios Zegel, email: echocobar@zegel.edu.pe

Received: May 13, 2026 | **Accepted:** September 14, 2026 | **Published:** November 29, 2026 | **DOI:** 10.37843/rted.v19i2.862

ABSTRACT

The critical review of manuscripts using generative artificial intelligence has emerged as a potential alternative for supporting editorial quality prior to submission to scientific journals. The objective was to design and preliminarily examine the applicability of a specialized prompt for reviewing academic manuscripts by comparing its performance on ChatGPT and Perplexity. The study was framed within the positivist paradigm, employing the hypothetico-deductive method, a quantitative approach, a non-experimental and instrumental-comparative design, an exploratory-applied scope, and a cross-sectional timeframe. The sample consisted of 46 open-access scientific articles published in 2025. Methods included document analysis, a systematized academic search, and the use of a prompt structured around ten criteria rated on a 0-to-3-point scale. A total of 92 evaluations were obtained. ChatGPT achieved a mean score of 24.85 ± 4.33, while Perplexity Pro with Deep Research obtained 21.11 ± 1.90. The Wilcoxon test indicated significant differences between the tools ($p < .001$; rrb = .80). Exact agreement between classifications was 26.1%, whereas Kappa and ICC indicators showed low agreement. Bland–Altman analysis showed a mean bias of 3.74 points and limits of agreement ranging from −3.96 to 11.44. It was concluded that the prompt enabled a structured, quantifiable critical evaluation, although the tools were not interchangeable. The findings did not constitute evidence of psychometric validity, and intra-model stability was not assessed.

KEYWORDS: Generative artificial intelligence, prompt, ChatGPT, Perplexity.

RESUMEN

La revisión crítica de manuscritos mediante inteligencia artificial generativa representó una alternativa emergente para apoyar la calidad editorial antes del envío a revistas científicas. El objetivo fue diseñar y examinar preliminarmente la aplicabilidad de un prompt especializado para revisar manuscritos académicos mediante su aplicación comparativa en ChatGPT y Perplexity. El estudio se enmarcó en el paradigma positivista, bajo el método hipotético-deductivo, con enfoque cuantitativo, diseño no experimental e instrumental-comparativo, alcance exploratorio-aplicado y corte transversal. La muestra estuvo constituida por 46 artículos científicos de acceso abierto publicados en 2025. Se emplearon el análisis documental, la búsqueda académica sistematizada y un prompt organizado en diez criterios valorados mediante una escala de 0 a 3 puntos. Se obtuvieron 92 evaluaciones. ChatGPT alcanzó una media de 24.85 ± 4.33 y Perplexity Pro con Investigación Profunda obtuvo 21.11 ± 1.90. La prueba de Wilcoxon mostró diferencias entre las herramientas ($p < .001$; rrb = .80). La coincidencia exacta de las clasificaciones fue de 26.1%, mientras que los indicadores kappa e ICC mostraron baja concordancia. Bland–Altman evidenció un sesgo medio de 3.74 puntos y límites de acuerdo entre −3.96 y 11.44. Se concluyó que el prompt permitió estructurar una evaluación crítica cuantificable, aunque las herramientas no resultaron intercambiables. Los hallazgos no constituyeron evidencia de validez psicométrica y la estabilidad intramodelo no fue evaluada.

PALABRAS CLAVE: Inteligencia artificial generativa, prompt, ChatGPT, Perplexity.

How to Cite

Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Introduction

The critical review of academic manuscripts using generative artificial intelligence represents an emerging alternative for strengthening editorial quality before submission to scientific journals. Writing, reviewing, and publishing academic manuscripts are central practices for ensuring the rigorous dissemination of knowledge and the formal communication of findings. These practices require authors to organize arguments, define research problems, structure sections, and substantiate contributions with up-to-date evidence, methodological soundness, and precise technical language (Kaklauskiene & Kaklauskas, 2024; Negi & Srivastava, 2021). Such requirements facilitate the identification of structural, argumentative, and formal weaknesses.

However, editorial processes face persistent barriers, including frequent rejections, extensive reviewer comments, lengthy publication timelines, academic fatigue, and increasing research waste (Menon et al., 2022; Ganga et al., 2022; Migheli et al., 2023). These difficulties are often associated with deficiencies in structure, writing, and formatting. They are also linked to a lack of novelty, misalignment with the journal, methodological weaknesses, unclear results, insufficient discussion, conclusions lacking contributions, and authorship problems (Paul, 2024; Kumar et al., 2023; Patel, 2022). Within this context, the research gap lies in the scarcity of evidence concerning structured instructions that have been preliminarily evaluated for systematic and quantifiable academic manuscript review.

In response to these limitations, generative artificial intelligence can support the detection of inconsistencies, argumentative gaps, methodological errors, and quality problems before editorial submission. Its use, however, requires specific, explicit, and methodologically organized instructions (Liu et al., 2023; Cevallos et al., 2025; Damián et al., 2025; Chocobar et al., 2025). Distributed cognition theory provides a basis for understanding scientific production as a sociotechnical process involving researchers, tools, practices, and digital environments (Gahrn-Andersen, 2025; Guile, 2023; Cavalcanti de Arruda et al., 2022). Nevertheless, ChatGPT and Perplexity do not replace researchers' critical judgment and require safeguards against hallucinations, nonexistent references, and superficial evaluations.

Accordingly, the objective of this research is to design and preliminarily examine the applicability of a specialized prompt for the critical review of academic manuscripts through its comparative application in GPT-4o and Perplexity Pro with Deep Research. Consistent with this purpose, the research question is: To what extent does a specialized prompt enable a structured, quantifiable critical evaluation of academic manuscripts, and what degree of agreement do ChatGPT and Perplexity show when applying it? The proposal contributes to strengthening editorial quality, academic innovation, and the responsible use of emerging technologies in higher education. It is also aligned with Sustainable Development Goals 4 and 9 (Chocobar & Barreda, 2025).

Methodology

The research was conducted within the positivist paradigm, an epistemological position in which phenomena are observed, measured, and analyzed through objective, verifiable, and replicable procedures. From this perspective, scientific knowledge is constructed through empirical evidence, observable indicators, and quantifiable relationships, while avoiding purely subjective interpretations of the phenomenon under study (Pretorius, 2024). Accordingly, the functionality of specialized instructions for academic manuscript review was addressed through explicit criteria. Structured measurements and quantifiable scores were employed to assess their ability to critically evaluate editorial quality before submission to scientific journals.

The hypothetico-deductive method was employed, starting from theoretical assumptions or initial hypotheses that are tested through observation, measurement, and empirical analysis. The study assumed that a set of specialized instructions, developed using explicit academic criteria, could function as a systematic tool. Its purpose was to identify errors, inconsistencies, gaps, and opportunities for improvement in academic manuscripts. Based on this premise, a matrix was

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

developed with ten critical evaluation criteria: Title; Abstract; Introduction; Methodology; Results; Discussion and Conclusions; Bibliographic References; Rigor, Consistency, and Depth; Originality and Contribution; and Relevance and Pertinence. This matrix was aligned with the logic of instrument development, which requires defining the construct, formulating criteria, establishing a scoring system, and interpreting results (Fei et al., 2025; Guinovart-Pedescoll & Palau, 2025).

The approach was quantitative, enabling observable phenomena to be transformed into numerical data for description, comparison, and interpretation through statistical procedures. Thus, manuscripts were critically evaluated using scores derived from a four-level Likert-type scale, with each criterion rated from 0 to 3: 0 = poor, 1 = basic, 2 = satisfactory, and 3 = excellent. Summing the ten criteria yielded a total score ranging from 0 to 30 points, which allowed the possible editorial decision suggested for each manuscript to be classified. Likert-type scales are appropriate when criteria, response levels, scoring, and interpretation are clearly defined (Ferrando et al., 2025; Koo & Yang, 2025).

The design was non-experimental, instrumental-comparative, exploratory-applied, cross-sectional, and focused on preliminary functional evaluation. Its non-experimental nature was justified because no independent variable was deliberately manipulated. The study observed and analyzed results generated by generative artificial intelligence tools using the same set of instructions. The cross-sectional timeframe corresponded to an evaluation conducted at a single point of application. This type of design examines data collected at a specific point in time without seeking to establish longitudinal causal relationships (Maier et al., 2023). The instrumental component focused on developing, organizing, and functionally examining a tool based on prompt engineering. The evaluation considered its structure, criteria, scoring system, and initial applicability (Costa et al., 2024; da Silva et al., 2024; de Figueirêdo et al., 2024). The comparative component enabled scores and decisions produced by GPT-4o and Perplexity Pro with Deep Research to be contrasted. Both tools applied the same evaluation procedure to the same documentary corpus.

The instrumental-comparative component was defined as a preliminary functional evaluation rather than a comprehensive psychometric validation. This methodological distinction was necessary because the study did not include expert judgment, Aiken's V, a content validity index, Cronbach's alpha, omega, test–retest assessment, factor analysis, or human interrater reliability. In studies involving instrument development and validation, validity and reliability provide necessary evidence of precision and consistency. Such evidence must be supported by specific measurement procedures that extend beyond the initial application of the tool (Ranganathan et al., 2024).

For this reason, the research preliminarily examined operational functioning, initial applicability, scoring capability, and comparative behavior between two generative artificial intelligence tools. This evidence describes the performance observed under the applied protocol and does not, in itself, constitute evidence of prompt validity.

The scope was exploratory-applied because the preliminary evaluation of specialized instructions represents a recent line of research in higher education and academic writing. The exploratory component enabled an initial examination of a phenomenon that remains methodologically underdeveloped: the use of structured instructions for quantifiable critical evaluations through ChatGPT and Perplexity. The applied component sought to offer a practical, realistic tool usable under an explicit protocol to detect weaknesses in writing, structure, methodology, results, discussion, references, originality, and relevance before submission. This approach is appropriate when the purpose is to assess the structure, content, scoring system, applicability, and capacity of a tool to produce organized and interpretable results (Costa et al., 2024; da Silva et al., 2024; de Figueirêdo et al., 2024).

The population consisted of scientific manuscripts available in academic databases, repositories, and platforms. Search Smart, a tool designed to guide the selection of appropriate academic search systems for scientific literature retrieval, was used to identify them (Gusenbauer, 2024). Initially, 100 academic databases,

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

repositories, or search systems were identified. Identification considered thematic coverage, availability of scientific records, and relevance to generative artificial intelligence, prompt engineering, ChatGPT, Perplexity, and higher education. Subsequently, 11 academic sources were selected through non-probability purposive sampling: SciELO, Redalyc, Dialnet, ScienceOpen, Google Scholar, Semantic Scholar, Wiley, Scopus, EBSCOhost, ScienceDirect, and Web of Science.

The final sample consisted of 46 open-access scientific articles published in 2025. They were written in English and related to generative artificial intelligence, prompt engineering, ChatGPT, Perplexity, and higher education. Selection of the 46 documents was based on availability, thematic relevance, and simultaneous compliance with the inclusion criteria. Full academic documents accessible online, with identifiable bibliographic data, a DOI or verifiable link, and a direct relationship with the object of study were included. Documents published outside the specified year, written in languages other than English, letters to the editor, book chapters, duplicates, and closed-access texts were excluded.

The corpus was organized in an Excel documentary matrix (Table 1), in which the fields needed to ensure academic and methodological traceability were recorded. The matrix included the database or academic source, the APA citation with authors, year, article title, academic journal, and DOI. When duplicate records were found across sources, only one entry was retained in the final matrix. This systematization made it possible to document the complete corpus transparently and present the 46 articles with verifiable bibliographic information.

Table 1

Matrix of 46 Articles Selected for the Sample.

#	Database	Citation	DOI
1	Web of Science (WoS)	Zhan and Yan (2025)	https://doi.org/10.1080/02602938.2025.2471821
2	Web of Science (WoS)	Kiyomiya et al. (2025)	https://doi.org/10.1177/20552076251323810
3	Web of Science (WoS)	Leypold et al. (2025)	https://doi.org/10.1016/j.wneu.2025.123753
4	Web of Science (WoS)	Peikos and Stavrou (2025)	https://doi.org/10.3390/educsci15030338
5	Web of Science (WoS)	Naranjo et al. (2025)	https://doi.org/10.3390/publications13020014
6	Scopus	Gonçalves et al. (2025)	https://doi.org/10.1002/ffo2.202
7	Scopus	Malek et al. (2025)	https://doi.org/10.1016/j.bja.2024.08.038
8	Scopus	Freire et al. (2025)	https://doi.org/10.1371/journal.pone.0323086
9	Scopus	Chan et al. (2025)	https://doi.org/10.1016/j.caeai.2024.100344
10	Scopus	Pristoupil et al. (2025)	https://doi.org/10.1186/s13244-025-01941-7
11	EBSCOhost	ElSayary et al. (2025)	https://doi.org/10.1155/hbe2/9986139
12	EBSCOhost	Tezbasaran and Ricci (2025)	https://doi.org/10.6339/25-JDS1184
13	EBSCOhost	Tupper et al. (2025)	https://doi.org/10.1080/14703297.2024.2316716
14	EBSCOhost	Ma et al. (2025)	https://doi.org/10.30191/ETS.202507_28(3).SP09
15	EBSCOhost	Lee and Palmer (2025)	https://doi.org/10.1186/s41239-025-00503-7
16	ScienceDirect	Keith et al. (2025)	https://doi.org/10.1016/j.ece.2025.01.002
17	ScienceDirect	Coen et al. (2025)	https://doi.org/10.2196/65848
18	ScienceDirect	Vissak and Torkkeli (2025)	https://doi.org/10.1108/RIBS-10-2024-0120
19	ScienceDirect	Kohnke and Moorhouse (2025)	https://doi.org/10.1016/j.chb.2025.108600
20	ScienceDirect	Ngo and Hastie (2025)	https://doi.org/10.1016/j.esp.2024.09.001
21	SciELO	Zhong (2025)	https://doi.org/10.5788/35-1-2037
22	SciELO	Şenyüz et al. (2025)	https://doi.org/10.5007/1518-2924.2025.e103465
23	Redalyc	Sánchez-Prieto et al. (2025)	https://doi.org/10.5944/ried.28.1.41548
24	Semantic Scholar	Leon (2025)	https://doi.org/10.25046/aj100201
25	Semantic Scholar	Reason et al. (2025)	https://doi.org/10.1007/s41669-025-00580-4
26	Semantic Scholar	Zhu et al. (2025)	https://doi.org/10.3390/electronics14112261
27	Semantic Scholar	Hsu (2025)	https://doi.org/10.1007/s11528-025-01052-6
28	Semantic Scholar	Schmitt and Brutzer (2025)	https://doi.org/10.54844/vte.2025.0940
29	Wiley	Walkington (2025)	https://doi.org/10.1111/ssm.18356
30	Wiley	Jooste et al. (2025)	https://doi.org/10.1002/cae.70064
31	Wiley	Gao et al. (2025a)	https://doi.org/10.1002/tea.70009

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

#	Database	Citation	DOI
32	Wiley	Topcu et al. (2025)	https://doi.org/10.1002/sys.21810
33	Wiley	Jones et al. (2025)	https://doi.org/10.1111/eje.13135
34	Dialnet	Pulari et al. (2025)	https://doi.org/10.9781/ijimai.2025.02.001
35	Dialnet	Kaya and Selcuk (2025)	https://doi.org/10.26668/businessreview/2025.v10i4.5426
36	Dialnet	Adomako et al. (2025)	https://doi.org/10.14201/adcaij.31857
37	Dialnet	PremaLatha et al. (2025)	https://doi.org/10.71459/edutech202524
38	ScienceOpen	Bystroński et al. (2025)	https://doi.org/10.48550/arXiv.2507.13874
39	ScienceOpen	Gao et al. (2025b)	https://doi.org/10.48550/arXiv.2502.02145
40	ScienceOpen	Wang et al. (2025)	https://doi.org/10.48550/arXiv.2507.12820
41	ScienceOpen	Singh and Singh (2025)	https://doi.org/10.48550/arXiv.2507.13113
42	ScienceOpen	Kusano et al. (2025)	https://doi.org/10.48550/arXiv.2507.13525
43	Google Scholar	Karakaya (2025)	https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/817
44	Google Scholar	Das et al. (2025)	https://doi.org/10.1016/j.omtn.2025.102546
45	Google Scholar	Morales et al. (2025)	https://doi.org/10.1007/s10270-024-01235-4
46	Google Scholar	Dertli and Yıldız (2025)	https://doi.org/10.29333/aje.2025.1015a

Note. The table lists the 46 articles from Web of Science (5), Scopus (5), EBSCOhost (5), ScienceDirect (5), SciELO (2), Redalyc (1), Semantic Scholar (5), Wiley (5), Dialnet (4), ScienceOpen (5), and Google Scholar (5). Authors' own work (2026).

The techniques used were document analysis, systematized academic searching, and controlled application of specialized instructions in generative artificial intelligence tools. The search strategy used the keywords and Boolean operators: "prompt engineering" AND "generative artificial intelligence" AND "ChatGPT" OR "Perplexity" AND "higher education". The main instrument was a prompt structured around ten critical evaluation criteria. The criteria addressed the title, abstract, introduction, methodology, results, discussion, conclusions, bibliographic references, rigor, consistency, depth, originality, contribution, relevance, and pertinence. Each criterion was scored from 0 to 3 points, yielding a cumulative total between 0 and 30 points. The structure of the specialized prompt is presented in Table 2.

Table 2

Structure of the Specialized Prompt for the Critical Review of Academic Manuscripts.

Act as an expert academic reviewer of scientific articles for multidisciplinary journals and analyze the attached file thoroughly and systematically, criterion by criterion, as a peer reviewer for high-impact academic journals would. Identify both strengths and weaknesses of the article, highlighting positive aspects before presenting constructive criticism. Read the entire article carefully (all sections and pages) to ensure a complete and consistent evaluation on each occasion, and assign a score for each criterion using the Likert scale (0, 1, 2, 3). Provide improvement suggestions for each criterion and write a report in Spanish exceeding 2,000 words in total, covering all ten evaluated criteria and strictly following these instructions. Your analysis must demonstrate consistency and rigor, reading every page of the file in depth without omitting any paragraph. The aim is to obtain the same result if you analyze the same article repeatedly at different times. Based on the foregoing, begin the analysis, scoring, and suggestions for the article according to the following ten criteria:

Evaluation Criterion	Aspect Evaluated	Scoring Scale	Interpretation
Title	Evaluate whether the title accurately reflects the manuscript's content, aligns with the research question and objectives, uses appropriate scientific terminology, is concise (around 20 words), avoids nonstandard abbreviations, includes the main variables, and identifies the type of research conducted.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the accuracy, clarity, relevance, and correspondence between the title and the actual manuscript content.
Abstract	Evaluate whether it contains no more than 250 words, follows a clear IMRaD structure (Introduction–Methods–Results–Discussion), includes keywords relevant to indexing, presents specific quantitative data when appropriate, avoids undefined abbreviations, bibliographic citations, and diagrams, is self-contained and understandable without reference to the full text, and uses the past tense to describe work performed.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the abstract's ability to independently, clearly, and completely summarize the essential manuscript content.

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Act as an expert academic reviewer of scientific articles for multidisciplinary journals and analyze the attached file thoroughly and systematically, criterion by criterion, as a peer reviewer for high-impact academic journals would. Identify both strengths and weaknesses of the article, highlighting positive aspects before presenting constructive criticism. Read the entire article carefully (all sections and pages) to ensure a complete and consistent evaluation on each occasion, and assign a score for each criterion using the Likert scale (0, 1, 2, 3). Provide improvement suggestions for each criterion and write a report in Spanish exceeding 2,000 words in total, covering all ten evaluated criteria and strictly following these instructions. Your analysis must demonstrate consistency and rigor, reading every page of the file in depth without omitting any paragraph. The aim is to obtain the same result if you analyze the same article repeatedly at different times. Based on the foregoing, begin the analysis, scoring, and suggestions for the article according to the following ten criteria:			
Introduction	Evaluate whether it predominantly uses the present tense for established facts and past tense for previous research describing the current problem, establishes the causes of the problem, identifies affected groups, clearly presents the scientific context through an up-to-date and relevant literature review, identifies research linked to theory, specifies existing knowledge gaps (epistemological gaps), states the consequences of not intervening, justifies the need and importance of the research, presents clearly formulated hypotheses where appropriate, explicitly connects with established theoretical frameworks, includes at least 70% of references from the last five years, and establishes clear links with the Sustainable Development Goals when applicable.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the stronger the contextual, theoretical, and argumentative foundation justifying the scientific need for the study.
Methodology	Evaluate whether the methods are appropriate for answering the research question, are described in sufficient detail to allow full replication, include explicit ethical considerations when appropriate, specify clear inclusion and exclusion criteria, describe sample-size calculations when applicable, include appropriate statistical power analyses, cite standard methodological sources for established techniques, fully describe new or modified methods, and transparently address methodological limitations.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the methodological rigor, transparency, coherence, and replicability.
Results	Evaluate whether findings are presented logically and sequentially without repeating methodological details, use appropriate visual elements (tables, figures, charts) that complement rather than duplicate the text, report all analyses conducted including nonsignificant findings, present confidence intervals alongside p-values when applicable, use an appropriate number of significant digits for measurement precision, include effect-size analyses when relevant, and avoid interpretations that belong in the Discussion section.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the clarity, statistical precision, and alignment of findings with the research question and objectives.
Discussion and Conclusions	Evaluate whether the discussion interprets results in the context of existing literature without extensively repeating findings, explicitly addresses study limitations and their implications for conclusions, discusses theoretical and practical implications, identifies specific directions for future research, establishes the study's unique contribution to existing knowledge, acknowledges unexpected or contradictory results with plausible explanations, and avoids causal claims unsupported by the study design.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the interpretive depth, argumentative coherence, and consistency between results, discussion, and conclusions.
Bibliographic References	Evaluate whether at least 30 references from indexed academic sources are included, with at least 70% published in the last five years; APA seventh-edition style is used consistently; all references have an accessible, verifiable DOI; at least 60% of references are in languages other than that of the manuscript (preferably English); excessive self-citation is avoided (maximum 10% of the total); there is a balanced distribution of original research, systematic reviews, and methodological sources; and all references are appropriately cited in the text without uncited entries.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the currency, traceability, academic quality, and formal consistency of the sources.
Rigor, Consistency, and Depth	Evaluate whether the manuscript strictly follows the IMRaD structure with logical transitions between sections, demonstrates internal coherence in argumentation and analysis, presents analytical depth appropriate for the journal, uses precise and consistent scientific terminology, shows rigor in data handling and statistical analysis, includes explicit ethical considerations where relevant, presents balanced arguments that consider alternative explanations, and demonstrates critical thinking in interpreting results.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the overall manuscript quality in terms of coherence, analytical depth, and scientific soundness.

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Act as an expert academic reviewer of scientific articles for multidisciplinary journals and analyze the attached file thoroughly and systematically, criterion by criterion, as a peer reviewer for high-impact academic journals would. Identify both strengths and weaknesses of the article, highlighting positive aspects before presenting constructive criticism. Read the entire article carefully (all sections and pages) to ensure a complete and consistent evaluation on each occasion, and assign a score for each criterion using the Likert scale (0, 1, 2, 3). Provide improvement suggestions for each criterion and write a report in Spanish exceeding 2,000 words in total, covering all ten evaluated criteria and strictly following these instructions. Your analysis must demonstrate consistency and rigor, reading every page of the file in depth without omitting any paragraph. The aim is to obtain the same result if you analyze the same article repeatedly at different times. Based on the foregoing, begin the analysis, scoring, and suggestions for the article according to the following ten criteria:			
Originality and Contribution	Evaluate whether the manuscript offers a clearly identifiable original contribution to scientific knowledge, addresses a specific literature gap, uses innovative methodological approaches or novel applications of established methods, generates new empirical data or reanalyzes existing data from original perspectives, proposes new theoretical frameworks or significant refinements of existing theories, offers practical implications or novel applications, and avoids duplication of previously published research.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the scientific contribution, novelty, academic value, and contribution to the field.
Relevance and Pertinence	Evaluate whether the manuscript clearly identifies and substantiates specific gaps in current scientific knowledge, demonstrates relevance to both the academic community and potential practical applications, aligns with research priorities in the discipline, contributes to sustainable development goals where applicable, has implications for public policy or professional practice when relevant, addresses contemporary issues of social or scientific importance, and convincingly justifies why the research was necessary.	0 = Poor; 1 = Basic; 2 = Satisfactory; 3 = Excellent.	The higher the score, the greater the importance, usefulness, applicability, and pertinence of the manuscript to the academic and professional community.

Note. The table presents the structure of the specialized prompt for critically reviewing academic manuscripts. Each criterion is scored using a four-level Likert-type scale: 0 = poor, 1 = basic, 2 = satisfactory, and 3 = excellent. Summing the ten criteria yields a total score from 0 to 30 points. The prompt used is available at https://doi.org/10.5281/zenodo.22111744. Authors' own work (2026).

The prompt was applied between May and December 2025 in GPT-4o and Perplexity Pro with Deep Research. The same documentary corpus, final version of the instructions, output language, scoring scale, and editorial decision categories were used. The definitive prompt version was fixed before formal application following pilot tests conducted solely to improve clarity, sequencing, and comprehension of the criteria. No further changes were made to the instructions during the formal phase to avoid retrospective optimization bias.

Each of the 46 manuscripts was evaluated once in GPT-4o and once in Perplexity Pro with Deep Research, producing 92 evaluations in total. Output was requested in Spanish, comprising a critical report, scores for each criterion, a total score, and a possible editorial decision. Parameters such as temperature, top-p, and penalties were not manually configured because the tools were accessed through user interfaces. To reduce variability, responses were not regenerated during formal evaluation, and the first complete response produced by each tool was retained.

To improve procedural traceability, an application log was incorporated. The log recorded the evaluation number, manuscript title, academic database or source, application date, approximate time, tool used, declared version or mode, account type, output language, run number, whether a response was regenerated, and control observations. Control observations were recorded narratively and were not coded a priori as independent variables for omitted criteria, arithmetic errors, inconsistency, or execution failure.

Human supervision first verified that each response included the ten required criteria, an individual score for each, the overall score, and the corresponding editorial decision. Next, arithmetic consistency between scores and categories was checked, together with the general correspondence between the observations provided and the content evaluated. This supervision did not replace the evaluations generated by the tools or introduce new human scores. Rather, it corrected transcription errors, identified obvious inconsistencies, and systematized data for statistical analysis. Accordingly, the results should be interpreted as GenAI-assisted evaluation under human oversight, not as independent expert peer review.

The total score for each manuscript ranged from 0 to 30 points, based on the sum of ten evaluated criteria. Four indicative classification

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

categories were established from these scores: “Publish as is” for scores of 27 or higher; “Publish after minor corrections and suggested improvements” for scores between 22 and 26; “Publish after major corrections and suggested improvements” for scores between 15 and 21; and “Not publishable in the journal” for scores of 14 or below. This classification allowed the evaluation severity of the two tools to be compared and their degree of agreement to be estimated.

These cutoffs were operational rules defined a priori by the researchers to summarize total scores. They were not externally validated and do not represent actual editorial decisions by a journal; they were used exclusively as ordinal categories for comparisons between tools.

Statistical analysis included descriptive statistics using frequencies, percentages, means, standard deviations, medians, interquartile ranges, minimums, and maximums. Before comparing GPT-4o and Perplexity Pro with Deep Research scores, the normality of paired differences was assessed using the Shapiro–Wilk test. Because the differences were not normally distributed, the Wilcoxon signed-rank test was applied. Effect size was also estimated using the rank-biserial correlation, and Bland–Altman analysis was performed to assess mean bias and limits of agreement between the tools. These procedures made it possible to establish whether GPT-4o and Perplexity Pro with Deep Research generated significantly different scores and whether their agreement was sufficient.

From an ethical perspective, only openly accessible academic manuscripts were used, without direct involvement of human participants, administration of instruments, or processing of sensitive personal data. Academic integrity was respected through the bibliographic identification of each article, and the authors interpreted and systematized the generated results. They considered transparency, human oversight, bias control, hallucination prevention, and research responsibility (Eaton, 2023; Lund et al., 2025; Wang et al., 2024).

Results

The results showed systematic differences in the evaluations generated by the two GenAI tools under the applied protocol. GPT-4o obtained higher overall scores (M = 24.85, SD = 4.33) than Perplexity Pro with Deep Research (M = 21.11, SD = 1.90), and the Wilcoxon signed-rank test confirmed a statistically significant difference between paired scores (W = 103.5, p <.001; rrb =.80). Exact agreement in the indicative classifications was 26.1% (12/46), while kappa and ICC coefficients indicated low agreement. Bland–Altman analysis yielded a mean bias of 3.74 points, with limits of agreement ranging from −3.96 to 11.44. Overall, the findings showed substantial differences in the scoring patterns produced by the two tools when the same specialized prompt was applied.

Results of Applying the Specialized Prompt in GPT-4o and Perplexity Pro with Deep Research

The main finding was the low agreement between the two tools: exact agreement for the indicative classifications was 26.1% (12/46), with unweighted kappa = −.036, linearly weighted kappa =.035, quadratically weighted kappa =.104, and ICC(2,1) =.192. Furthermore, Bland–Altman analysis showed a mean bias of 3.74 points and limits of agreement from −3.96 to 11.44, indicating that GPT-4o and Perplexity Pro with Deep Research were not interchangeable for assigning total scores under this protocol.

A total of 92 analyzable evaluations were obtained, 46 per tool. Table 3 preserves manuscript-level traceability by reporting total score, indicative classification, and the difference between the tools; the difference was calculated as the GPT-4o score minus the Perplexity Pro with Deep Research score.

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Table 3

Individual Scores and Indicative Classifications Generated by GPT-4o and Perplexity Pro with Deep Research.

#	Source	Article Evaluated	Perplexity Pro with Deep Research	Perplexity Classification	GPT-4o	ChatGPT Classification	Diff.
1	Web of Science	Zhan and Yan (2025)	24	Minor revisions	30	As is	6
2	Web of Science	Kiyomiya et al. (2025)	23	Minor revisions	30	As is	7
3	Web of Science	Leypold et al. (2025)	19	Major revisions	24	Minor revisions	5
4	Web of Science	Peikos and Stavrou (2025)	22	Minor revisions	30	As is	8
5	Web of Science	Naranjo et al. (2025)	20	Major revisions	22	Minor revisions	2
6	Scopus	Gonçalves et al. (2025)	18	Major revisions	19	Major revisions	1
7	Scopus	Malek et al. (2025)	22	Minor revisions	22	Minor revisions	0
8	Scopus	Freire et al. (2025)	21	Major revisions	28	As is	7
9	Scopus	Chan et al. (2025)	21	Major revisions	30	As is	9
10	Scopus	Pristoupil et al. (2025)	23	Minor revisions	21	Major revisions	−2
11	EBSCOhost	ElSayary et al. (2025)	22	Minor revisions	25	Minor revisions	3
12	EBSCOhost	Tezbasaran and Ricci (2025)	18	Major revisions	19	Major revisions	1
13	EBSCOhost	Tupper et al. (2025)	22	Minor revisions	28	As is	6
14	EBSCOhost	Ma et al. (2025)	23	Minor revisions	30	As is	7
15	EBSCOhost	Lee and Palmer (2025)	21	Major revisions	20	Major revisions	−1
16	ScienceDirect	Keith et al. (2025)	24	Minor revisions	21	Major revisions	−3
17	ScienceDirect	Coen et al. (2025)	20	Major revisions	27	As is	7
18	ScienceDirect	Vissak and Torkkeli (2025)	26	Minor revisions	27	As is	1
19	ScienceDirect	Kohnke and Moorhouse (2025)	21	Major revisions	30	As is	9
20	ScienceDirect	Ngo and Hastie (2025)	20	Major revisions	30	As is	10
21	SciELO	Zhong (2025)	19	Major revisions	20	Major revisions	1
22	SciELO	Şenyüz et al. (2025)	21	Major revisions	26	Minor revisions	5
23	Redalyc	Sánchez-Prieto et al. (2025)	21	Major revisions	30	As is	9
24	Semantic Scholar	Leon (2025)	19	Major revisions	22	Minor revisions	3
25	Semantic Scholar	Reason et al. (2025)	23	Minor revisions	20	Major revisions	−3
26	Semantic Scholar	Zhu et al. (2025)	22	Minor revisions	30	As is	8
27	Semantic Scholar	Hsu (2025)	22	Minor revisions	27	As is	5
28	Semantic Scholar	Schmitt and Brutzer (2025)	21	Major revisions	25	Minor revisions	4
29	Wiley	Walkington (2025)	20	Major revisions	18	Major revisions	−2
30	Wiley	Jooste et al. (2025)	19	Major revisions	20	Major revisions	1
31	Wiley	Gao et al. (2025a)	25	Minor revisions	28	As is	3
32	Wiley	Topcu et al. (2025)	21	Major revisions	26	Minor revisions	5
33	Wiley	Jones et al. (2025)	20	Major revisions	29	As is	9
34	Dialnet	Pulari et al. (2025)	19	Major revisions	22	Minor revisions	3
35	Dialnet	Kaya and Selcuk (2025)	21	Major revisions	28	As is	7
36	Dialnet	Adomako et al. (2025)	18	Major revisions	17	Major revisions	−1
37	Dialnet	PremaLatha et al. (2025)	18	Major revisions	17	Major revisions	−1
38	ScienceOpen	Bystroński et al. (2025)	19	Major revisions	20	Major revisions	1
39	ScienceOpen	Gao, et al. (2025b)	23	Minor revisions	29	As is	6
40	ScienceOpen	Wang et al. (2025)	20	Major revisions	25	Minor revisions	5
41	ScienceOpen	Singh and Singh (2025)	22	Minor revisions	30	As is	8
42	ScienceOpen	Kusano et al. (2025)	23	Minor revisions	21	Major revisions	−2
43	Google Scholar	Karakaya (2025)	20	Major revisions	26	Minor revisions	6
44	Google Scholar	Das et al. (2025)	22	Minor revisions	26	Minor revisions	4
45	Google Scholar	Morales et al. (2025)	23	Minor revisions	19	Major revisions	−4
46	Google Scholar	Dertli and Yıldız (2025)	20	Major revisions	29	As is	9

Note. Diff. = difference between the score assigned by GPT-4o and that assigned by Perplexity Pro with Deep Research. Authors' own work (2026).

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

The overall comparison showed a higher mean score for GPT-4o (24.85 ± 4.33) than for Perplexity Pro with Deep Research (21.11 ± 1.90). The respective medians were 26 and 21, and the observed ranges were 17–30 and 18–26 (Table 4).

Table 4

Descriptive Statistics of Scores Assigned by Each GenAI Tool.

GenAI Tool	n	Mean	SD	Median	IQR	Min.	Max.
Perplexity Pro with Deep Research	46	21.11	1.90	21	20–22	18	26
GPT-4o	46	24.85	4.33	26	21–29	17	30

Note. SD = standard deviation; IQR = interquartile range. Authors' own work (2026).

The indicative classifications derived from the cutoffs also differed between tools: Perplexity classified 58.7% as requiring major corrections and 41.3% as requiring minor corrections, whereas ChatGPT classified 43.5% as "publish as is", 26.1% as requiring minor corrections, and 30.4% as requiring major corrections (Table 5). These categories were operational rules of the study, not externally validated editorial decisions.

Table 5

Distribution of Indicative Editorial Classifications Generated by the Specialized Prompt.

Indicative classification	Perplexity Pro with Deep Research	%	GPT-4o	%
Publish as is	0	0.0	20	43.5
Publish with minor revisions	19	41.3	12	26.1
Publish with major revisions	27	58.7	14	30.4
Not publishable in the journal	0	0.0	0	0.0
Total	46	100.0	46	100.0

Note. The table shows the percentage distribution of indicative classifications generated by each GenAI tool. Authors' own work (2026).

Categorical agreement was limited: 12 of 46 evaluations (26.1%) matched exactly. In 20 cases ChatGPT assigned a classification one category higher than Perplexity, in 9 cases two categories higher, and in 5 cases one category lower (Table 6). These differences describe the relative behavior of the models under the operational cutoffs.

Table 6

Agreement Matrix for Indicative Editorial Classifications Generated by Both Tools.

Perplexity / ChatGPT classification	Not publishable	Major revisions	Minor revisions	As is	Total
Not publishable	0	0	0	0	0
Major revisions	0	9	9	9	27
Minor revisions	0	5	3	11	19
As is	0	0	0	0	0
Total	0	14	12	20	46

Note. The main diagonal represents exact agreement between the tools. Authors' own work (2026).

The Shapiro–Wilk test indicated non-normality of the differences (W =.948, p =.038), so the Wilcoxon test was applied. The score difference was significant (W = 103.5, p <.001), with a mean difference of 3.74 points, median of 4.5, and rank-biserial effect size rrb =.80 (Table 7). This comparison indicates a systematic difference between the tools, not prompt validity.

Table 7

Paired Comparison of Scores Between GenAI Tools.

Comparison	Mean difference	SD of differences	Median difference	95% CI	W	p	Effect size
ChatGPT – Perplexity	3.74	3.93	4.5	2.57 to 4.91	103.5	<.001	rrb =.80

Note. rrb = rank-biserial correlation. Authors' own work (2026).

Agreement indicators reinforced the main finding. Although Spearman's correlation showed a positive association (rho =.415, p =.004) and Kendall's tau-b =.329 (p =.003), agreement was low: unweighted kappa = −.036, linearly weighted kappa =.035, quadratically weighted kappa =.104, and ICC(2,1) =.192; even ICC(2,k) =.322 remained low (Table 8). Thus, association and agreement must not

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

be regarded as equivalent.

Table 8

Association and Agreement Indicators Between GPT-4o and Perplexity Pro with Deep Research.

Indicator	Value	p
Spearman correlation	.415	.004
Kendall's tau-b	.329	.003
Unweighted kappa	−.036	
Linearly weighted kappa	.035	
Quadratically weighted kappa	.104	
ICC(2,1), absolute agreement	.192	
ICC(2,k), tool average	.322	

Note. ICC = intraclass correlation coefficient. Kappa values were calculated using the ordinal editorial categories. Authors' own work (2026).

Bland–Altman analysis indicated low interchangeability: mean bias of 3.74 points, SD of the differences of 3.93, and limits of agreement from −3.96 to 11.44. The span between these limits was 15.40 points, indicating substantial discrepancies at the individual level under the applied protocol (Table 9 and Figure 1).

Table 9

Bland–Altman Analysis Between GPT-4o and Perplexity Pro with Deep Research.

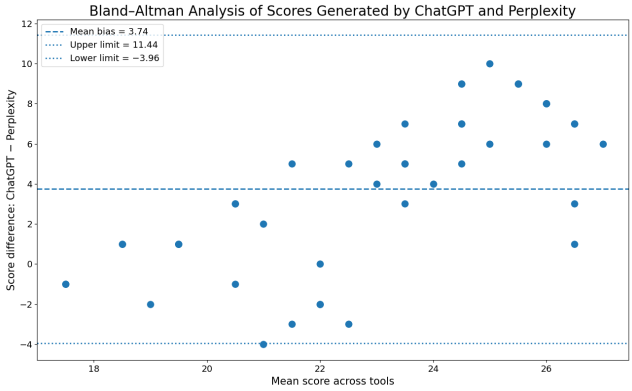
Indicator	Value
Mean bias	3.74
SD of differences	3.93
Lower limit of agreement	−3.96
Upper limit of agreement	11.44
Minimum observed difference	−4.00
Maximum observed difference	10.00

Note. The table presents the Bland–Altman analysis. Authors' own work (2026).

Figure 1 depicts the dispersion between the mean scores assigned by the two tools and the difference between GPT-4o and Perplexity Pro with Deep Research.

Figure 1

Bland–Altman Plot of Scores Generated by GPT-4o and Perplexity Pro with Deep Research.



Note. The figure shows the mean scores of both tools on the X-axis and the ChatGPT–Perplexity difference on the Y-axis. The central line represents the mean bias, and the outer lines represent the limits of agreement.

Discussion

The findings answer the research question by indicating that the prompt enabled a structured, quantifiable critical evaluation of academic manuscripts. It transformed qualitative review into an organized procedure comprising ten criteria, a 0-to-3-point scale, an overall score, and a preliminary editorial decision. In keeping with the research objective, the prompt was designed and its functionality preliminarily evaluated through comparative application in GPT-4o and Perplexity Pro with Deep Research. Nevertheless, this functionality must not be interpreted as definitive psychometric validation, but rather as preliminary evidence of operational applicability under the protocol used. The prompt made it possible to organize review using explicit criteria, scores, and indicative classifications.

The principal contribution of the study is to show that specialized instructions can function as a methodological support tool for reviewing academic manuscripts before editorial submission. Unlike general instructions, the designed prompt defined the reviewer's role, specified criteria, and established performance levels. It also required a score for each dimension, requested improvement suggestions, and derived a preliminary editorial decision. This structure allowed both GPT-4o and Perplexity Pro with Deep Research to process the

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

same manuscripts under a shared evaluation framework. The prompt can therefore be used in different generative artificial intelligence tools, provided that it is adapted to each system's technical conditions and accompanied by human oversight. Its applicability does not depend on identical results across models but on its capacity to structure a complex task and make it quantifiable.

The results show differences in evaluation severity between the tools examined. GPT-4o assigned higher overall scores, whereas Perplexity Pro with Deep Research assigned comparatively lower scores under the applied protocol. This difference does not invalidate the prompt; rather, it shows that GenAI-assisted evaluation depends on the interaction between the instructions, model, access, algorithmic architecture, and response criteria. Moreover, the quality of generated responses depends on the precision of instructions, contextual information, role formulation, and the use of each tool (Lee & Palmer, 2025; ElSayary et al., 2025; Tezbasaran & Ricci, 2025; Freire et al., 2025). Accordingly, the prompt should not be understood as a fixed universal formula, but as an adaptable matrix that requires calibration for the GenAI tool used.

The differences observed between GPT-4o and Perplexity Pro with Deep Research could be associated with the technical characteristics of each tool, including architecture, information retrieval mechanisms, configuration, updates, and response generation. However, these variables were neither manipulated nor directly examined in the study, so this explanation must be regarded as hypothetical rather than causal. This interpretation is consistent with research warning of hallucinations, biases, nonexistent references, variability, and consistency problems in the performance of generative artificial intelligence models in academic tasks (Topcu et al., 2025; Lund et al., 2025; Wang et al., 2024). Consequently, applying the prompt in different GenAI tools requires clearly defined conditions of use, identification of model version, output control, source verification, and human oversight of generated scores.

The lack of interchangeability between ChatGPT and Perplexity has important methodological implications. The mean difference of 3.74 points and Bland–Altman limits of agreement between −3.96 and 11.44 show that the tools do not directly produce equivalent results. On a 0-to-30-point scale, a difference of approximately four points can shift a manuscript from one editorial category to another. However, this finding does not indicate inefficiency in the prompt, but rather the need for intermodel calibration. Methodologically, the prompt successfully structured the evaluation, while the tools applied that structure with different levels of severity. Therefore, using it with multiple GenAI systems should involve adjustment rules, human review of discrepancies, and, wherever possible, comparison against expert reviewers.

The Wilcoxon test indicated systematic differences between scores, but it is not a validity test. By contrast, kappa, ICC, and Bland–Altman are central to assessing the degree of agreement and showed low agreement and wide limits. Accordingly, using the same prompt does not guarantee equivalence between models, and any comparative implementation requires external calibration and human oversight.

An evaluation instrument requires definition of the construct, dimensions, criteria, response levels, scoring system, and interpretation of results. When full validation is sought, it must incorporate evidence of content validity, reliability, temporal stability, and interrater agreement (Fei et al., 2025; Guinovart-Pedescoll & Palau, 2025; Costa et al., 2024; da Silva et al., 2024; de Figueirêdo et al., 2024). This research advanced the functional development of the instrument, its comparative application, and preliminary analysis of agreement between tools. The prompt therefore showed preliminary applicability for structuring and quantifying review under the evaluated conditions.

The applicability of the prompt should be understood as support for academic judgment. The tool does not replace human reviewers but can assist their work by organizing reading, making evaluation criteria explicit, generating preliminary feedback, and systematizing manuscript assessment. This usefulness is consistent with studies recognizing the potential of GenAI to support scientific writing, improve textual clarity, assist review processes, organize information, and detect inconsistencies, always subject to human verification. Thus, the prompt may be used in ChatGPT, Perplexity, or other GenAI systems as a pre-submission review

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

tool, provided the researcher adapts the instructions to the environment, reviews the outputs, contrasts the scores, and avoids automatically accepting decisions generated by the model.

Among the main limitations, the sample was selected through non-probability purposive sampling, which may introduce selection bias and limits the generalizability of the findings. Moreover, although open-access articles with verifiable data were used, a complete open corpus containing all evaluated files was not deposited. This restricts full external replication across the different GenAI versions. Their updates, internal configurations, internet access, and response policies may alter findings in future applications. Each manuscript was processed only once by each tool, so intramodel stability and reproducibility were not assessed.

An additional limitation was the absence of a disaggregated analytical matrix retaining the individual scores for the ten criteria for every manuscript and tool. Consequently, statistical analysis was conducted using total scores and indicative classifications, without estimating separate differences, associations, or agreement for each prompt dimension. The results therefore characterize the overall behavior of the tools under the applied protocol but cannot establish which specific criteria accounted for most of the observed discrepancies.

The findings are relevant because they show that the critical review of manuscripts may benefit from specialized, structured, and quantifiable instructions. The prompt should be presented as a preliminary methodological foundation for developing GenAI-assisted review processes. Its applicability is reflected in its ability to operationalize editorial criteria, generate comparable data, and guide preliminary classifications. The contribution of the study is therefore the proposal of a functional, adaptable, and supervisable tool capable of supporting improvements to academic manuscripts in higher education before submission to scientific journals.

Future research should expand this line of inquiry through more robust validation processes. It is recommended to apply the prompt to manuscripts undergoing actual preparation and to compare versions before and after suggested revisions to determine whether observable improvements occur in textual quality, methodological coherence, argumentative depth, and editorial suitability. Scores generated by GenAI should also be compared with evaluations by expert human reviewers. Researchers should obtain evidence of content validity using Aiken's V or CVI, estimate reliability by criterion, conduct test–retest assessments, analyze intra- and intermodel stability, and compare different tools, free and paid versions, languages, disciplines, and scientific article formats. The prompt should therefore be used as a preliminary, adaptable, and supervised tool.

Conclusions

The importance of this study lies in proposing a systematic structure to support the critical review of academic manuscripts through generative artificial intelligence. The prompt organizes assessment into ten criteria, transforming a broad review into a scored procedure that can be compared between tools. However, the results show that ChatGPT with GPT-4o and Perplexity Pro with Deep Research do not produce equivalent assessments under the applied protocol. The low agreement observed, together with the differences identified by kappa, ICC, and Bland–Altman analysis, indicates that the tools are not interchangeable when assigning overall scores or indicative classifications. The study's contribution therefore lies in preliminarily structuring an AI-assisted review procedure rather than demonstrating the psychometric validity of the prompt.

In relation to the research objective, the prompt enables the organization of a quantifiable critical evaluation of manuscripts through criteria concerning the title, abstract, introduction, methodology, results, discussion and conclusions, references, rigor, originality, and relevance. Nevertheless, the available results describe the tools' overall behavior through total scores and indicative classifications and do not establish which individual criteria account for most of the discrepancies between models. Similarly, conducting only one evaluation per manuscript and tool prevents determination of intramodel stability and score reproducibility. The prompt should therefore be regarded as a preliminary support tool whose use

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

requires human oversight, specification of the model employed, and caution when interpreting the results.

Future research should compare the scores generated by artificial intelligence tools with assessments by expert human reviewers, analyze intra- and intermodel stability, and estimate the reliability of the criteria comprising the prompt. It will also be necessary to retain disaggregated criterion-level scoring matrices to determine which dimensions show greater agreement or discrepancies between tools. Future studies may also incorporate content validity procedures, test–retest assessments, and comparisons between different models, languages, disciplines, and editorial contexts. These steps will allow the validity, stability, and usefulness of the prompt as a support instrument for academic review to be determined more precisely.

Acknowledgments

The authors express their gratitude to Guillermo Enrique Cevallos López, Jeessikha Ubillús Reyes, César Augusto Reyes Alvarado, Marushka Chocobar Reyes, Jesús Enrique Reyes Acevedo, Remy Felipe Barreda Medina, and Emilio José Chocobar Reyes for their valuable academic, methodological, technical, and editorial contributions to this study. The collaborative work undertaken to design the prompt and systematize the evaluation criteria is also acknowledged, together with the application of generative artificial intelligence tools and interpretation of the resulting findings.

Declaration of Conflicts of Interest

The authors declare that they have no conflicts of interest related to the conduct, analysis, interpretation, or publication of this study. They also declare that no personal, academic, professional, or financial relationships influenced the research process or the interpretation of the findings.

Funding Statement

The authors declare that this research received no specific funding from public, private, commercial, or nonprofit organizations. The study was conducted using the authors' own resources, and no external funding source influenced the design, execution, analysis, interpretation, or reporting of the results.

Ethics Statement

The study did not require approval from an institutional ethics committee because it exclusively analyzed openly accessible published scientific articles and did not involve human participants, interventions, informed consent procedures, or the processing of sensitive or personally identifiable data. Academic integrity, bibliographic traceability, responsible use of information, human oversight, and scientific rigor were maintained throughout the research process.

Statement on the Use of Artificial Intelligence

The authors declare that generative artificial intelligence tools were intentionally used as methodological components and objects of evaluation in this study. Specifically, GPT-4o and Perplexity Pro with Deep Research were applied to the same corpus of 46 scientific manuscripts using a predefined specialized prompt and standardized evaluation criteria. The outputs generated by both systems were subject to human supervision, verification, systematization, and statistical analysis. Generative artificial intelligence did not act as an autonomous author or final decision-maker, and the authors retained full responsibility for the methodological design, interpretation of the findings, conclusions, and final content of the manuscript.

References

Adomako, P., Khan, T., Ali, R. & Koutaly, R. (2025). Comparison of Leading AI Models an Analytical Study of ChatGPT Google Bard and Microsoft Bing. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 14, e31857. https://doi.org/10.14201/adcaij.31857

Bystroński, M., Hołysz, M., Piotrowski, G., Chawla, N. & Kajdanowicz, T. (2025). Large Language Models as Innovators: A Framework to Leverage Latent Space Exploration for Novelty Discovery. [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.13874

de Figueirêdo, R. C., de Siqueira Silva, Í., de Araújo, A. J., Silva, C. R. D. V., Martiniano, C. S., Brito, E. W. G., Xavier, P. B., &

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

da Costa Uchôa, S. A. (2024). Preparation and validation of the instrument "QualiAPS digital-Brazil" for assessing digital health care in primary health care: A required tool. *Frontiers in Public Health*, *12*, Article 1304148. https://doi.org/10.3389/fpubh.2024.1304148

Cavalcanti de Arruda, G., Medeiros de Freitas, D., Soares, C., Nawratek, K., & Miranda, B. (2022). The production of knowledge through religious and social media infrastructure: World making practices among Brazilian Pentecostals. *Popular Communication*, *20*(3), 208–221. https://doi.org/10.1080/15405702.2022.2074997

Cevallos, G., Ubillús, J., & Chocobar, E. (2025). Argumentos a favor de permitir o prohibir el uso de la inteligencia artificial generativa por estudiantes. Una revisión sistemática. *European Public & Social Innovation Review*, *11*, 1–30. https://doi.org/10.31637/epsir-2026-2155

Chan, K., Ali, F., Park, J., Sham, K., Thong, E., Chien, F., Qian, K., & Sze, G. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, *8*, Article 100344. https://doi.org/10.1016/j.caeai.2024.100344

Chocobar, E., & Barreda, R. (2025). Estructuras metodológicas PICO y PRISMA 2020 en la elaboración de artículos de revisión sistemática: Lo que todo investigador debe conocer y dominar. *Ciencia Latina Revista Científica Multidisciplinar*, *9*(1), 8525–8543. https://doi.org/10.37811/cl_rcm.v9i1.16491

Chocobar, E., Ubillús, J., & Cevallos, G. (2025). ¿Sólo el ChatGPT? Diferentes herramientas de inteligencia artificial generativa (GenAI) y sus usos: Una revisión de la literatura científica, 2021–2024. *Revista Científica de la UCSA*, *12*(1), 70–85. https://doi.org/10.18004/ucsa/2409-8752/2025.012.01.070

Coen, E., Del Fiol, G., Kaphingst, K., Borsato, E., Shannon, J., Smith, H., Masino, A., & Allen, C. (2025). Chatbot for the return of positive genetic screening results for hereditary cancer syndromes: Prompt engineering project. *JMIR Cancer*, *11*, Article e65848. https://doi.org/10.2196/65848

Costa, A., Rodrigues, F., Pitarma, R., & Ferreira, M. (2024). Design and validation of an instrument to evaluate the workgroup methodology in university students. *International Journal of Educational Research Open*, *7*, Article 100351. https://doi.org/10.1016/j.ijedro.2024.100351

Damián, M., Acosta, D., Delgado, I., & Rojas, G. (2025). Percepción de Estudiantes de Administración sobre la Utilidad de la Inteligencia Artificial en Artículos Científicos. *Revista Docentes 2.0*, 18(2), 107–126. https://doi.org/10.37843/rted.v18i2.678

Das, A., Bhattacharya, M., Abdelhameed, A., Lee, S., & Chakraborty, C. (2025). Generative artificial intelligence, integrative bioinformatics, and single-cell analysis reveal Alzheimer's genetic and immune landscape. *Molecular Therapy Nucleic Acids*, *36*(2), Article 102546. https://doi.org/10.1016/j.omtn.2025.102546

da Silva, L., Silva, R., Pereira, M., Herdeiro, M., & Baldoni, A. (2024). Construction and validation of an instrument to identify barriers to implementing pharmaceutical care. *Exploratory Research in Clinical and Social Pharmacy*, *16*, Article 100529. https://doi.org/10.1016/j.rcsop.2024.100529

Dertli, Z., & Yıldız, B. (2025). The use of prompt engineering in creating mathematical modelling activities with artificial intelligence tool ChatGPT. *Anatolian Journal of Education*, *10*(1), 59–80. https://doi.org/10.29333/aje.2025.1015a

Eaton, S. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity*, *19*(1), Article 23. https://doi.org/10.1007/s40979-023-00144-1

ElSayary, A., Kuhail, M., & Hojeij, Z. (2025). Examining the role of prompt engineering in utilizing generative AI tools for lesson planning: Insights from teachers' experiences and perceptions. *Human Behavior and Emerging Technologies*, *2025*(1), Article 9986139. https://doi.org/10.1155/hbe2/9986139

Fei, W., Jingqi, T., Xiaonan, S., Xinying, S., Jun, L., Xingxing, M., & Yibo, W. (2025). Design and development of scales in primary care: Practical steps and statistical methods. *Chinese General Practice Journal*, *2*(4), Article 100091. https://doi.org/10.1016/j.cgpj.2025.100091

Ferrando, P., Morales, F., Casas, J., & Muñiz, J. (2025). Likert scales: A practical guide to design, construction and use. *Psicothema*, *37*(4), 1–15. https://doi.org/10.70478/psicothema.2025.37.24

Freire, Y., Santamaría, A., Orejas, J., Ortiz, I., Gómez, M., Thuissard, I., Díaz-Flores, V., & Suárez, A. (2025). Evaluating the influence of prompt formulation on the reliability and repeatability of ChatGPT in implant-supported prostheses. *PLOS ONE*, *20*(5), Article e0323086. https://doi.org/10.1371/journal.pone.0323086

Gahrn-Andersen, R. (2025). Enlanguaged affordances in social practices: A critical rethinking of Gibson's approach to language. *Language & Communication*, *104*, 121–130. https://doi.org/10.1016/j.langcom.2025.07.004

Ganga, F., Alarcón, N., Suárez, W., Meleán, R., Ruiz, G., & Cueva, J. (2022). Causas que originan rechazo de artículos científicos en revistas científicas latinoamericanas. *Ingeniare. Revista Chilena de Ingeniería*, *30*(3), 602–618. https://doi.org/10.4067/S0718-33052022000300602

Gao, Y., Piccinini, M., Moller, K., Alanwar, A., & Betz, J. (2025a). From words to collisions: LLM-guided evaluation and adversarial generation of safety-critical driving scenarios. *arXiv*. https://doi.org/10.48550/arXiv.2502.02145

Gao, Y., Zhai, X., Li, M., Lee, G., & Liu, X. (2025b). A multimodal interactive framework for science assessment in the era of generative artificial intelligence. *Journal of Research in Science Teaching*, *62*(9), 2014–2028. https://doi.org/10.1002/tea.70009

Gonçalves, R., Spaniol, M., Rowland, N., & Malý, N. (2025). Reflections on building an artificial intelligence bot to prepare students to engage in strategic conversations during foresight fieldwork. *Futures & Foresight Science*, *7*(1), Article e202. https://doi.org/10.1002/ffo2.202

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Guile, D. (2023). Machine learning – a new kind of cultural tool? A "recontextualisation" perspective on machine learning + interprofessional learning. *Learning, Culture and Social Interaction, 42*, Article 100738. https://doi.org/10.1016/j.lcsi.2023.100738

Guinovart-Pedescoll, M., & Palau, R. (2025). Design and validation of a self-perception instrument on the use of scientific knowledge in teaching practice. *Journal of New Approaches in Educational Research, 14*(1), Article 3. https://doi.org/10.1007/s44322-024-00026-5

Gusenbauer, M. (2024). Searchsmart.org: Guiding researchers to the best databases and search systems for systematic reviews and beyond. *Research Synthesis Methods, 15*(6), 1200–1213. https://doi.org/10.1002/jrsm.1746

Hsu, H. (2025). From programming to prompting: Developing computational thinking through large language model-based generative artificial intelligence. *TechTrends, 69*(3), 485–506. https://doi.org/10.1007/s11528-025-01052-6

Jones, B., Desu, A., & Honig, C. (2025). Artificial intelligence chatbots as virtual patients in dental education: A constructivist approach to classroom implementation. *European Journal of Dental Education, 30*(2), 447–456. https://doi.org/10.1111/eje.13135

Jooste, J., Wolff, K., & Joubert, J. (2025). Engineering students' perceptions and use of generative artificial intelligence. *Computer Applications in Engineering Education, 33*(4), Article e70064. https://doi.org/10.1002/cae.70064

Kaklauskiene, L., & Kaklauskas, A. (2024). The mapping of research on workplace management and control using a SCALAR method and bibliometric DSS. *Studies in Informatics and Control, 33*(1), 37–48. https://doi.org/10.24846/v33i1y202404

Karakaya, K. (2025). Human-AI interaction with large language models in complex information tasks: Prompt engineering strategies. *Asian Journal of Distance Education, 20*(1). https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/817

Kaya, Y., & Selcuk, R. (2025). Assessing a fine-tuned scrum AI agent: Accuracy, utility, and expert validation. *International Journal of Professional Business Review, 10*(4), Article e05426. https://doi.org/10.26668/businessreview/2025.v10i4.5426

Keith, M., Keiller, E., Windows, C., Kings, I., & Robbins, P. (2025). Harnessing generative AI in chemical engineering education: Implementation and evaluation of the large language model ChatGPT v3.5. *Education for Chemical Engineers, 51*, 20–33. https://doi.org/10.1016/j.ece.2025.01.002

Kiyomiya, K., Aomori, T., & Ohtani, H. (2025). Medication counseling for OTC drugs using customized ChatGPT-4: Comparison with ChatGPT-3.5 and ChatGPT-4o. *Digital Health, 11*, Article 20552076251323810. https://doi.org/10.1177/20552076251323810

Kohnke, L., & Moorhouse, B. (2025). Enhancing the emotional aspects of language education through generative artificial intelligence (GenAI): A qualitative investigation. *Computers in Human Behavior, 167*, Article 108600. https://doi.org/10.1016/j.chb.2025.108600

Koo, M., & Yang, S. (2025). Likert-type scale. *Encyclopedia, 5*(1), Article 18. https://doi.org/10.3390/encyclopedia5010018

Kumar, N., Ali, Z., & Haldar, R. (2023). Novelty in research: A common reason for manuscript rejection! *Indian Journal of Anaesthesia, 67*(3), 245–247. https://doi.org/10.4103/ija.ija_143_23

Kusano, G., Akimoto, K., & Takeoka, K. (2025). Revisiting prompt engineering: A comprehensive evaluation for LLM-based personalized recommendation. *arXiv*. https://doi.org/10.48550/arXiv.2507.13525

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education, 22*(1), Article 7. https://doi.org/10.1186/s41239-025-00503-7

Leon, M. (2025). Generative artificial intelligence and prompt engineering: A comprehensive guide to models, methods, and best practices. *Advances in Science, Technology and Engineering Systems Journal, 10*(2), 1–11. https://doi.org/10.25046/aj100201

Leypold, T., Bahm, J., Beier, J., Guillaume, V., Ammo, T., Lauer, H., Kolbenschlag, J., & Schäfer, B. (2025). Evaluating ChatGPT o1's capabilities in peripheral nerve surgery: Advancing artificial intelligence in clinical practice. *World Neurosurgery, 196*, Article 123753. https://doi.org/10.1016/j.wneu.2025.123753

Liu, Y., Yang, Z., Yu, Z., Liu, Z., Liu, D., Lin, H., Li, M., Ma, S., Avdeev, M., & Shi, S. (2023). Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. *Journal of Materiomics, 9*(4), 798–816. https://doi.org/10.1016/j.jmat.2023.05.001

Lund, B., Mannuru, N., Teel, Z., Lee, T., Ortega, N., Simmons, S., & Ward, E. (2025). Student perceptions of AI-assisted writing and academic integrity: Ethical concerns, academic misconduct, and use of generative AI in higher education. *AI in Education, 1*(1), Article 2. https://doi.org/10.3390/aieduc1010002

Ma, Z., Cui, X., Liu, W., Tu, Y., & Hwang, G. (2025). ChatGPT-assisted collaborative argumentation: Effects of role-playing prompts on students' argumentation outcomes, processes, and perceptions. *Educational Technology & Society, 28*(3), 133–150. https://doi.org/10.30191/ETS.202507_28(3).SP09

Maier, C., Thatcher, J., Grover, V., & Dwivedi, Y. (2023). Cross-sectional research: A critical perspective, use cases, and recommendations for IS research. *International Journal of Information Management, 70*, Article 102625. https://doi.org/10.1016/j.ijinfomgt.2023.102625

Malek, M., van Velzen, M., Dahan, A., Martini, C., Sitsen, E., Sarton, E., & Boon, M. (2025). Generation of preoperative anaesthetic plans by ChatGPT-4.0: A mixed-method study. *British Journal of Anaesthesia, 134*(5), 1333–1340. https://doi.org/10.1016/j.bja.2024.08.038

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal, 19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Menon, V., Varadharajan, N., Praharaj, S., & Ameen, S. (2022). Why do manuscripts get rejected? A content analysis of rejection reports from the Indian Journal of Psychological Medicine. *Indian Journal of Psychological Medicine, 44*(1), 59–65. https://doi.org/10.1177/0253717620965845

Migheli, Q., Arthurs, S., & Wright, M. (2023). Regret to inform you: Editor's tips to avoid journal rejection. *Journal of Agricultural and Marine Sciences (JAMS), 28*(2), 124. https://go.docentes20.com/r5prg

Morales, S., Clarisó, R., & Cabot, J. (2025). Impromptu: A framework for model-driven prompt engineering. *Software and Systems Modeling, 24*(6), 1627–1645. https://doi.org/10.1007/s10270-024-01235-4

Naranjo, J., Llumiquinga, M., Vaca, W., & Espin, C. (2025). Generative AI vs. traditional databases: Insights from industrial engineering applications. *Publications, 13*(2), Article 14. https://doi.org/10.3390/publications13020014

Negi, P., & Srivastava, A. (2021). All-in-one guide for preparing and publishing a bioscience manuscript. *Current Science, 121*(9), 1162–1179. https://www.jstor.org/stable/27310779

Ngo, T., & Hastie, D. (2025). Artificial intelligence for academic purposes (AIAP): Integrating AI literacy into an EAP module. *English for Specific Purposes, 77*, 20–38. https://doi.org/10.1016/j.esp.2024.09.001

Patel, P. (2022). How to avoid manuscript rejection: Common reasons for rejection of manuscript. *National Journal of Community Medicine, 13*(6), 349–351. https://doi.org/10.55489/njcm.130620222179

Paul, J. (2024). Publishing in premier journals with high impact factor and Q1 journals: Dos and don'ts. *International Journal of Consumer Studies, 48*(3), Article e13049. https://doi.org/10.1111/ijcs.13049

Peikos, G., & Stavrou, D. (2025). ChatGPT for science lesson planning: An exploratory study based on pedagogical content knowledge. *Education Sciences, 15*(3), Article 338. https://doi.org/10.3390/educsci15030338

PremaLatha, V., Dinesh, A., & Parveen, N. (2025). Optimizing interactions: Strategies for prompt engineering in large language models. *Edu - Tech Enterprise, 3*, 24. https://doi.org/10.71459/edutech202524

Pretorius, L. (2024). Demystifying research paradigms: Navigating ontology, epistemology, and axiology in research. *The Qualitative Report, 29*(10), 2698–2715. https://doi.org/10.46743/2160-3715/2024.7632

Pristoupil, J., Oleaga, L., Junquero, V., Merino, C., Sureyya, O., Kyncl, M., Burgetova, A., & Lambert, L. (2025). Generative pre-trained transformer 4o (GPT-4o) in solving text-based multiple response questions for European Diploma in Radiology (EDiR): A comparative study with radiologists. *Insights into Imaging, 16*(1), Article 66. https://doi.org/10.1186/s13244-025-01941-7

Pulari, S., Umadevi, M., & Vasudevan, S. (2025). Improved fine-tuned reinforcement learning from human feedback using prompting methods for news summarization. *International Journal of Interactive Multimedia and Artificial Intelligence, 9*(2), 59–67. https://doi.org/10.9781/ijimai.2025.02.001

Ranganathan, P., Caduff, C., & Frampton, C. (2024). Designing and validating a research questionnaire – Part 2. *Perspectives in Clinical Research, 15*(1), 42–45. https://doi.org/10.4103/picr.picr_318_23

Reason, T., Klijn, S., Rawlinson, W., Benbow, E., Langham, J., Teitsson, S., Johannesen, K., & Malcolm, B. (2025). Using generative artificial intelligence in health economics and outcomes research: A primer on techniques and breakthroughs. *PharmacoEconomics Open, 9*(4), 501–517. https://doi.org/10.1007/s41669-025-00580-4

Sánchez-Prieto, J. C., Izquierdo-Álvarez, V., Del Moral-Marcos, M. T., & Martínez-Abad, F. (2025). Inteligencia artificial generativa para autoaprendizaje en educación superior: Diseño y validación de una máquina de ejemplos. *RIED-Revista Iberoamericana de Educación a Distancia, 28*(1). https://doi.org/10.5944/ried.28.1.41548

Schmitt, C., & Brutzer, A. (2025). Generative artificial intelligence in vocational education and training: A framework for sustainable teacher competence development. *Vocation, Technology & Education*. https://doi.org/10.54844/vte.2025.0940

Şenyüz, B., Özgen, E., & Oğuzcan, A. (2025). Integrating GenAI into communication education for 'generation prompt': An exploration of academics' perspectives on its benefits, challenges, and future prospects in Turkiye. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, 30*, 1–25. https://doi.org/10.5007/1518-2924.2025.e103465

Singh, P., & Singh, P. (2025). Leveraging language prior for infrared small target detection. *arXiv*. https://doi.org/10.48550/arXiv.2507.13113

Tezbasaran, A., & Ricci, S. (2025). Embracing the AI revolution: ChatGPT's role in advancing data science consultation services. *Journal of Data Science, 23*(2), 416–428. https://doi.org/10.6339/25-JDS1184

Topcu, T., Husain, M., Ofsa, M., & Wac, P. (2025). Trust at your own peril: A mixed methods exploration of the ability of large language models to generate expert-like systems engineering artifacts and a characterization of failure modes. *Systems Engineering, 28*(5), 583–604. https://doi.org/10.1002/sys.21810

Tupper, M., Hendy, I., & Shipway, R. (2025). Field courses for dummies: To what extent can ChatGPT design a higher education field course? *Innovations in Education and Teaching International, 62*(2), 512–526. https://doi.org/10.1080/14703297.2024.2316716

Vissak, T., & Torkkeli, L. (2025). Applying generative artificial intelligence applications for academic research on firms' nonlinear internationalization. *Review of International Business and Strategy, 35*(4), 436–484. https://doi.org/10.1108/RIBS-10-2024-0120

How to Cite

Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal, 19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND

RTED
Vol. 19 (2), 2026
June – November 2026
233-250

Walkington, C. (2025). The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*, *126*(4), 356–365. https://doi.org/10.1111/ssm.18356

Wang, H., Dang, A., Wu, Z., & Mac, S. (2024). Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines. *Computers and Education: Artificial Intelligence*, *7*, Article 100326. https://doi.org/10.1016/j.caeai.2024.100326

Wang, S., Fang, R., He, Z., Song, S., & Li, Y. (2025). Emotional support with LLM-based empathetic dialogue generation. *arXiv*. https://doi.org/10.48550/arXiv.2507.12820

Zhan, Y., & Yan, Z. (2025). Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *51*(5), 821–834. https://doi.org/10.1080/02602938.2025.2471821

Zhong, A. (2025). Prompts for language learners: A practical guide to using DeepSeek as a dictionary. *Lexikos*, *35*(1), 274–285. https://doi.org/10.5788/35-1-2037

Zhu, Z., Funabiki, N., Mentari, M., Aung, S., Kao, W., & Lee, Y. (2025). An automatic code generation tool using generative artificial intelligence for element fill-in-the-blank problems in a Java programming learning assistant system. *Electronics*, *14*(11), Article 2261. https://doi.org/10.3390/electronics14112261

How to Cite
Reyes Alvarado, C. A., Cevallos López, G. E., Ubillús Reyes, J., Chocobar Reyes, M., Reyes Acevedo, J. E., Barreda Medina, R. F., & Chocobar-Reyes, E. (2026). Evaluation of a Prompt for Academic Manuscript Review in Higher Education. *Docentes 2.0 Technological-Educational Journal*, *19*(2), 233-250. https://doi.org/10.37843/rted.v19i2.862

OPEN ACCESS
CC BY NC ND