Original Article

Journal of Advances in Medical Education & Professionalism

Benchmarking Large Language Models for Scientific Writing: A Mixed-methods Evaluation of ChatGPT, DeepSeek, and Human Authors

KULYASH R. ZHILISBAYEVA[1], MSc Student; NASRIN SHOKRPOUR[2,3*], PhD; MOHAMMAD MEHDI PARVIZI[4,5], MD

[1]Department of Languages, West Kazakhstan Marat Ospanov Medical University, Aktobe, Kazakhstan; [2]English Department, Faculty of Paramedical Sciences, Shiraz University of Medical Sciences, Shiraz, Iran; [3]Clinical Education Research Center, Shiraz University of Medical Sciences, Shiraz, Iran; [4]Molecular Dermatology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran; [5]Department of Medical Journalism, School of Paramedical Sciences, Shiraz University of Medical Sciences, Shiraz, Iran

Abstract

Introduction: Modern large language models (LLMs) like ChatGPT (based on the GPT4 architecture) and DeepSeek offer unprecedented capabilities for generating scientific text. However, their performance in replicating structured, high-quality scientific writing, especially compared to human-authored abstracts, remains insufficiently evaluated. To compare the abstract quality produced by human authors, ChatGPT/GPT4, and DeepSeek across six evaluation criteria: Clarity, Coherence, Conciseness, Accuracy, IMRaD Structure, and Language Quality, using blinded expert ratings and non-parametric statistical methods, specifically the Kruskal–Wallis test followed by pairwise Wilcoxon rank-sum tests with false discovery rate correction.

Methods: We selected 23 medical and healthrelated research topics, each yielding three abstracts (human, ChatGPT, DeepSeek), for a total of 69 abstracts. Three raters scored each abstract. Kruskal–Wallis tests assessed group differences; Cliff's Delta (δ) was calculated as a nonparametric effect size for each comparison, suitable for ordinal data.

Results: Across criteria, ChatGPT and DeepSeek significantly outperformed human authors in Clarity, Coherence, IMRaD Structure, and Language Quality. In contrast, Conciseness and Accuracy showed negligible effect sizes ($|\delta| < 0.10$), suggesting parity across all three sources.

Conclusions: ChatGPT and DeepSeek achieved significantly higher scores in clarity, coherence, structure, and language quality, while showing comparable performance in conciseness and accuracy. These findings complement recent evaluations showing competitive medical and reasoning performance of DeepSeek models compared to proprietary LLMs. While shortform abstracts, expert oversight, and domain expertise remain critical, the results suggest that LLMs—particularly GPT4 and DeepSeek—can serve as effective tools in drafting scientific abstracts.

Keywords: Artificial intelligence; Natural language processing; Deep learning; Abstracting and indexing; Medical writing

**Corresponding author:*
Nasrin Shokrpour, PhD;
English Department, Faculty of Paramedical Sciences,
Shiraz University of Medical Sciences,
Shiraz, Iran
Tel: +98-7132270328
Email: shokrpourn@gmail.com

Please cite this paper as:
Zhilisbayeva KR, Shokrpour N, Parvizi MM. Benchmarking Large Language Models for Scientific Writing: A Mixed-methods Evaluation of ChatGPT, DeepSeek, and Human Authors. J Adv Med Educ Prof. 2026;14(3):281-290.
DOI: 10.30476/jamp.2026.109761.2366.

Received: 23 January 2026
Accepted: 21 April 2026

Introduction

The rapid advancement of large language models (LLMs) such as OpenAI's ChatGPT and DeepSeek has transformed the landscape of scientific communication (1). These models, trained on vast datasets encompassing a wide array of knowledge domains, are increasingly being used to support tasks such as literature review, text summarization, abstract generation, and even full manuscript drafting (2). Their capacity to produce fluent, coherent, and grammatically sound texts has sparked both enthusiasm and concern within the academic community (3).

Scientific abstracts serve as the critical gateway to a research article, offering readers a concise summary of the study objectives, methodology, results, and conclusions (4). Given their prominence in indexing, retrieval, and citation contexts, abstract quality has direct implications for the visibility and impact of a study (5). Traditionally, abstract writing has been the domain of human authors, often requiring multiple iterations to balance clarity, conciseness, and completeness (6). However, the emergence of LLMs challenges this paradigm by offering instant generation of abstracts with stylistic and structural sophistication comparable to—or even exceeding—that of human experts (7).

While recent studies have evaluated the readability and coherence of AI-generated content, comprehensive, blinded, and comparative assessments of abstracts authored by LLMs versus humans remain limited (8). Furthermore, head-to-head comparisons between leading models like ChatGPT (GPT-4 pro) and newer entrants such as DeepSeek-V2 are still in their infancy (9). In line with this topic, recent JAMP publications have discussed the opportunities, challenges, and educational implications of AI/ChatGPT in medical scholarship, providing relevant context for evaluating AI-generated scientific writing and structured reporting (10-12). Such evaluations are critical not only for understanding the strengths and limitations of these tools but also for guiding ethical and effective integration into academic workflows (13).

To address this gap, the present study aimed to systematically evaluate whether large language models (ChatGPT and DeepSeek) produce scientific abstracts of comparable or superior quality to human-written abstracts across six predefined dimensions: clarity, coherence, conciseness, accuracy, IMRaD structure, and language quality. Therefore, the following research question and hypothesis were posed:

Research question: Do AI-generated abstracts differ significantly from human-written abstracts across these six dimensions when rated by blinded expert reviewers?

Hypothesis: We hypothesized that AI-generated abstracts would differ significantly from human-written abstracts across these six dimensions.

Methods

Study design

This study employed a within-subjects comparative experimental design, in which each research topic served as a matched unit across three writing conditions: Human, ChatGPT, and DeepSeek. For each of the 23 selected research topics, three abstracts were generated—one human-written and two AI-generated—allowing direct comparison within the same topical context. This design minimizes between-topic variability and enables controlled comparison of writing performance across sources. For each topic, three versions of the abstract were created—one written by a human, and the other two generated using ChatGPT and DeepSeek—resulting in a total of 69 abstracts.

Abstract generation

The 23 research topics were selected using purposive sampling from peer-reviewed studies in clinical medicine, epidemiology, and public health journals with an impact factor greater than 5. Articles were screened against predefined eligibility criteria to ensure structured abstract format and applied medical relevance. Basic laboratory sciences and preclinical experimental studies were not included. Topics were chosen to reflect applied medical research contexts in which structured abstracts are commonly required. Human-written abstracts were extracted from peer-reviewed original research articles published between 2020 and 2023 in medical, epidemiological, and public health journals with an impact factor greater than 5. Only structured abstracts (IMRaD or equivalent format) were included. No human abstract was rewritten or modified for this study. Additional details regarding abstract selection and eligibility criteria are provided in Supplementary Material 2. All selected abstracts were independently verified by a senior author to confirm adherence to standard abstract structure and absence of major factual inconsistencies. AI-generated abstracts were produced in May 2024 using ChatGPT (GPT-4 model, OpenAI; version available May 2024) and DeepSeek-V2 (DeepSeek AI; version available May 2024). Both models were accessed through their publicly available web-based interfaces

under default settings. The same standardized prompt was used for both models to ensure consistency. The identical standardized prompt used for both ChatGPT (GPT-4) and DeepSeek-V2 is provided in Supplementary Material 1 to ensure methodological transparency and reproducibility. The prompt was applied without modification across all 23 research topics to ensure consistency and reproducibility.

Evaluation criteria and scoring

The six evaluation criteria were selected based on established principles of scientific writing and prior literature on abstract quality assessment (4-7). Clarity and coherence reflect logical organization and readability, which are central to effective scientific communication. Conciseness corresponds to standard abstract length constraints and the requirement for efficient information delivery. Accuracy assesses factual consistency and alignment with the described research topic. IMRaD structure assesses adherence to the conventional Introduction–Methods–Results–Discussion format recommended in biomedical reporting standards. Language quality captures grammatical correctness and stylistic fluency. Collectively, these dimensions represent widely recognized formal and structural characteristics of high-quality scientific abstracts.

For human-written abstracts, accuracy was defined as correspondence with the original article's main findings and absence of factual inconsistencies relative to the published study. For AI-generated abstracts, which were produced based solely on article titles without access to original datasets or full manuscripts, accuracy was assessed based on the absence of gross inconsistencies with general medical knowledge and the internal coherence of the inferences presented. Therefore, the accuracy metric in this study does not represent verification against actual study data for AI-generated texts but rather an assessment of logical plausibility and absence of obvious factual errors.

Each abstract was independently evaluated according to six pre-defined criteria: Clarity; Coherence; Conciseness; Accuracy; Structure (IMRaD compliance); and Language Quality. Scores for each criterion ranged from 0 (poor) to 2 (excellent). In addition, evaluators were asked to select the "Best Overall" abstract among the three for each topic, resulting in a binary outcome for that measure (1 = best; 0 = not selected).

A detailed scoring rubric was developed before evaluation to standardize rating procedures and enhance inter-rater consistency. Each criterion was scored on a three-point ordinal scale:
0 = Poor: Major deficiencies; criterion not adequately met.
1 = Acceptable: Minor deficiencies; criterion generally met but with noticeable limitations.
2 = Excellent: Criterion fully met; clear, complete, and consistent with high-quality scientific standards.
Specific definitions for each evaluation dimension were as follows:
Clarity:
2 = Ideas expressed clearly and unambiguously;
1 = Generally clear but minor ambiguity present;
0 = Difficult to understand or unclear statements.
Coherence:
2 = Logical flow and smooth transitions between components;
1 = Minor disruptions in logical progression;
0 = Disorganized or fragmented structure.
Conciseness:
2 = Efficient and focused writing within word limits;
1 = Minor redundancy or verbosity;
0 = Excessively verbose or unfocused.
Accuracy:
2 = No factual inconsistencies; conclusions align with described methods and results;
1 = Minor imprecision not affecting interpretation;
0 = Major factual inconsistencies or unsupported conclusions.
Structure (IMRaD compliance):
2 = All IMRaD components clearly identifiable;
1 = Minor structural omissions or unclear separation;
0 = Major structural deficiencies.
Language Quality:
2 = Grammatically correct with professional scientific tone;
1 = Minor grammatical or stylistic issues;
0 = Frequent grammatical or stylistic errors.

Prior to formal evaluation, raters participated in a calibration session using sample abstracts to ensure consistent interpretation of the scoring criteria.

Raters and blinding

Three experienced reviewers with academic backgrounds in scientific writing and publication independently scored the abstracts. To minimize bias, we anonymized and randomly shuffled the abstracts. Raters were blinded to the source (Human, ChatGPT, or DeepSeek) during the evaluation process. All raters used the same scoring rubric and participated in a calibration

session before evaluation.

Statistical analysis for comparison of scores

The evaluation dataset consisted of 23 abstracts, each independently written by a human author, ChatGPT, or DeepSeek. Each abstract was assessed across six criteria: Clarity, Coherence, Conciseness, Accuracy, Structure (IMRaD), and Language Quality, using an ordinal scoring system. A total of 69 abstracts (23 per group) were analyzed. For statistical analysis, each individual rating was treated as an independent ordinal observation, resulting in 69 scores per group (23 abstracts × 3 raters).

To compare the distributions of ordinal scores across the three models, we used non-parametric statistical methods appropriate for ordinal data. The data were first reshaped into long format, and the model type was categorized into three groups: Human, ChatGPT, and DeepSeek.

For each evaluation criterion, a Kruskal–Wallis rank sum test was performed to assess overall differences in distribution of scores among the three models. If the Kruskal–Wallis test indicated a statistically significant difference ($p<0.05$), we conducted pairwise Wilcoxon rank-sum tests (also known as Mann–Whitney U tests) between each pair of models. To control multiple comparisons, we applied a false discovery rate (FDR) correction.

Because evaluation scores were measured on an ordinal 0–2 scale, non-parametric statistical methods were selected. The Kruskal–Wallis test was used to assess overall group differences, as it does not assume normal distribution of data. When significant differences were detected, pairwise Wilcoxon rank-sum tests were performed with false discovery rate (FDR) correction for multiple comparisons.

Independence of observations was maintained because each abstract was treated as an independent analytical unit within each writing condition. Although the study employed a within-topic design, statistical comparisons were conducted at the abstract level with no repeated scoring of the same text within groups.

Cliff's Delta (δ) was selected as the effect size measure because it is appropriate for ordinal data and does not rely on distributional assumptions. Unlike parametric effect size measures such as eta-squared, which are derived from ANOVA and assume interval-level data and normality, Cliff's Delta provides a robust measure of dominance between groups for non-parametric comparisons.

The results were visualized using box-and-whisker plots, faceted by evaluation criterion. Significant pairwise differences were annotated on the plots using asterisk markers (< 0.05, $p < 0.01$, $p < 0.001$). Comparisons that were not statistically significant were omitted from the visual output to improve clarity. All analyses were conducted in R version 4.4.1 using the tidyverse, readxl, and ggpubr packages. A detailed summary of statistical assumptions and analytical rationale is provided in Supplementary Material 3.

Effect size analysis

To evaluate the magnitude of differences in the abstract quality among the three writing methods—ChatGPT, DeepSeek, and Human authorship—we calculated Cliff's Delta (d) for each pairwise comparison across the six evaluation criteria: Clarity, Coherence, Conciseness, Accuracy, Structure (IMRaD), and Language Quality.

Cliff's Delta is a non-parametric effect size measure suitable for ordinal data, indicating the probability that a randomly selected observation from one group will have a higher score than that from another group. It ranges from -1 to +1, where 0 represents no effect.

For each criterion, comparisons were made between ChatGPT vs. Human, DeepSeek vs. Human, and ChatGPT vs. DeepSeek. Effect sizes were interpreted using established thresholds: negligible ($|d| <0.147$), small ($0.147 \leq |d| <0.33$), medium ($0.33 \leq |d| <0.474$), and large ($|d| \geq 0.474$). All computations were performed using the effsize package in R (version 4.4.1).

Radar chart visualization

To visually compare the overall evaluation profiles of abstracts produced by ChatGPT, DeepSeek, and human authors, we constructed a radar chart. Six evaluation criteria were included: Clarity, Coherence, Conciseness, Accuracy, Structure (IMRaD), and Language Quality.

Mean scores for each criterion were computed separately for the three groups. To generate the radar chart, we standardized the data to a common scale with a minimum of 0 and a maximum of 2, corresponding to the scoring rubric used in the study. The visualization was created using the fmsb package in R version 4.4.1. This method enables simultaneous comparison of all six qualitative dimensions across the three writing methods.

Ethical approval

This study was approved by the research ethics committee of Shiraz University of Medical Sciences with the code of IR.SUMS.REC.1404.482.

Results

Comparative evaluation of abstract quality across writing methods

A total of 69 abstracts (23 per group) generated by Human authors, ChatGPT, and DeepSeek were evaluated across six ordinal criteria: Clarity, Coherence, Conciseness, Accuracy, Structure (IMRaD), and Language Quality. Figure 1 displays box-and-whisker plots of the score distributions for each criterion across the three models, with statistically significant differences annotated.

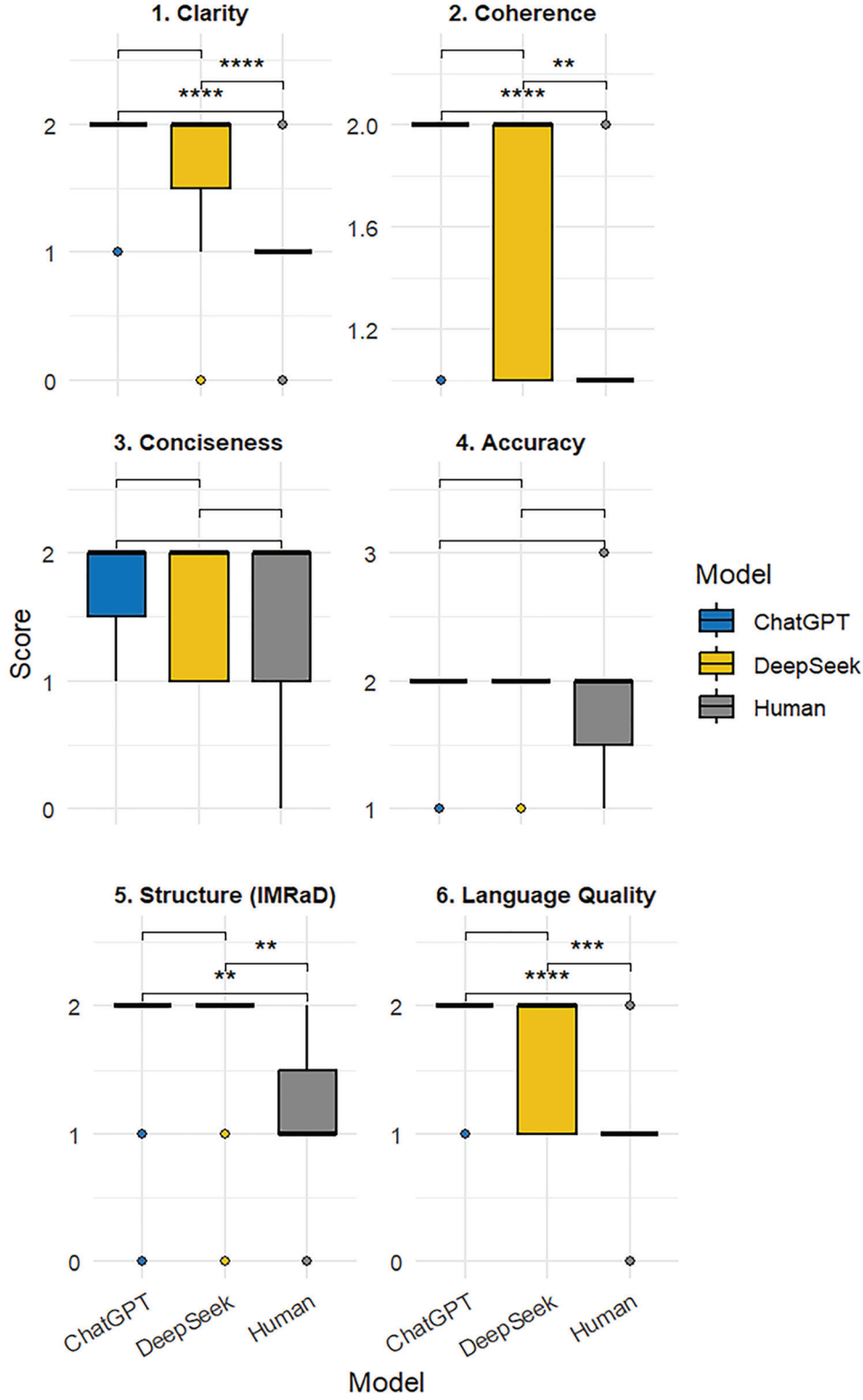

Figure 1. Boxplots of evaluation criteria scores for Human, ChatGPT, and DeepSeek-generated abstracts
Figure 1 illustrates the distribution of scores across evaluation criteria. AI-generated abstracts demonstrate higher median values and tighter distributions, particularly for clarity and coherence. Human-authored abstracts exhibit greater variability and lower central tendencies.

Distribution of scores for six evaluation criteria—Clarity, Coherence, Conciseness, Accuracy, Structure (IMRaD), and Language Quality— across abstracts written by Human authors, ChatGPT, and DeepSeek (n=23 per group) is shown in the Figure 1. Boxplots display median (horizontal line), interquartile range (boxes), and outliers (points). Pairwise statistical differences were assessed using Wilcoxon rank-sum tests and are shown only when significant ($p < 0.05$). Kruskal–Wallis tests were used to assess the overall differences among the three groups for each criterion.

The Kruskal–Wallis test revealed significant differences among the three writing conditions in Clarity ($H(2) = 107.58$, $p < 0.001$). Pairwise Wilcoxon comparisons showed that both ChatGPT and DeepSeek achieved significantly higher clarity scores than Human-authored abstracts ($p < 0.01$ for both comparisons), while no significant difference was observed between the two AI models.

A similar pattern was observed for Coherence ($H(2) = 89.39$, $p < 0.001$), with both ChatGPT and DeepSeek scoring significantly higher than Human authors ($p < 0.01$), and no significant difference between the AI systems.

For Conciseness, a significant overall difference was detected ($H(2) = 44.98$, $p < 0.001$). Pairwise comparisons indicated that DeepSeek scored significantly higher than Human authors ($p < 0.05$). In contrast, ChatGPT did not differ significantly from Human-authored abstracts. No significant difference was observed between ChatGPT and DeepSeek.

In contrast, Accuracy did not differ significantly among the three groups ($H(2) = 1.87$, $p = 0.392$), indicating comparable performance across Human, ChatGPT, and DeepSeek-generated abstracts.

Significant differences were observed in Structure (IMRaD compliance) ($H(2) = 83.09$, $p < 0.001$), with both AI models producing more structurally complete abstracts than Human authors ($p < 0.05$), and no significant difference between ChatGPT and DeepSeek.

Finally, Language Quality demonstrated marked differences among groups ($H(2) = 93.75$, $p < 0.001$). Both ChatGPT and DeepSeek significantly outperformed Human-authored abstracts ($p < 0.001$), while the difference between the two AI models was not statistically significant.

Descriptive statistics for each evaluation criterion are presented in Table 1. Median scores were generally higher for ChatGPT and DeepSeek in Clarity, Coherence, IMRaD Structure, and Language Quality, while median values for Conciseness and Accuracy were similar across groups.

Table 1 demonstrates that both ChatGPT and DeepSeek consistently achieved higher median scores compared to human-authored abstracts across most criteria. The most pronounced differences were observed in clarity, coherence, structure, and language quality, where human-authored abstracts showed lower median values. In contrast, conciseness and accuracy displayed comparable median scores across all groups, indicating similar performance in these domains.

Effect size results

As shown in Table 2, Cliff's Delta analysis revealed substantial differences in abstract quality between AI-generated texts and human-written abstracts across several evaluation criteria (Table 1). Large effect sizes were observed for clarity, with ChatGPT demonstrating a delta of 0.830 and DeepSeek 0.624 compared to human authors. Similarly, for coherence, ChatGPT achieved a delta of 0.652 and DeepSeek 0.478, while for structure (IMRaD format), both AI models yielded a delta of 0.499. In the domain of language quality, ChatGPT again outperformed with a delta of 0.743, followed by DeepSeek at 0.535. These results indicate that AI-generated abstracts—particularly those produced by ChatGPT—were consistently rated higher than those authored by humans in terms of clarity, coherence, structural organization, and linguistic quality. In contrast, negligible effect sizes were found for conciseness ($|d| < 0.07$) and accuracy ($|d| < 0.10$) across all pairwise comparisons, suggesting comparable performance between AI and human authors in these domains.

Table 1. Median (IQR) scores for each evaluation criterion by writing source

Criterion	ChatGPT Median (IQR)	DeepSeek Median (IQR)	Human Median (IQR)
Clarity	2 (2–2)	2 (2–2)	1 (1–1)
Coherence	2 (2–2)	2 (2–2)	1 (1–1)
Conciseness	2 (2–2)	2 (2–2)	1 (1–2)
Accuracy	2 (2–2)	2 (2–2)	2 (2–2)
Structure (IMRaD)	2 (2–2)	2 (2–2)	1 (1–1)
Language Quality	2 (2–2)	2 (2–2)	1 (1–1)

Each group included 69 individual ratings (23 abstracts evaluated by 3 raters).

Table 2. Cliff's Delta effect sizes comparing writing methods across evaluation criteria

Criterion	Comparison	Cliff's Delta (d)	Magnitude
Clarity	ChatGPT vs Human	0.830	Large
	DeepSeek vs Human	0.624	Large
	ChatGPT vs DeepSeek	0.178	Small
Coherence	ChatGPT vs Human	0.652	Large
	DeepSeek vs Human	0.478	Large
	ChatGPT vs DeepSeek	0.174	Small
Conciseness	ChatGPT vs Human	0.066	Negligible
	DeepSeek vs Human	0.026	Negligible
	ChatGPT vs DeepSeek	0.043	Negligible
Accuracy	ChatGPT vs Human	0.093	Negligible
	DeepSeek vs Human	0.051	Negligible
	ChatGPT vs DeepSeek	0.043	Negligible
Structure (IMRaD)	ChatGPT vs Human	0.499	Large
	DeepSeek vs Human	0.499	Large
	ChatGPT vs DeepSeek	0.000	Negligible
Language quality	ChatGPT vs Human	0.743	Large
	DeepSeek vs Human	0.535	Large
	ChatGPT vs DeepSeek	0.217	Small

Comparisons between ChatGPT and DeepSeek generally resulted in small effect sizes, including clarity (d = 0.178), coherence (d = 0.174), and language quality (d = 0.217), indicating relatively minor differences between the two AI models. Overall, these findings demonstrate that AI-generated abstracts achieved significantly higher scores in four of the six evaluated dimensions—clarity, coherence, IMRaD structure, and language quality—while showing comparable performance to human-authored abstracts in conciseness and accuracy.

Table 2 highlights the magnitude of differences between writing methods using Cliff's Delta. Large effect sizes were observed for clarity, coherence, structure, and language quality, confirming the superior performance of AI-generated abstracts in these domains. Conversely, negligible effect sizes for conciseness and accuracy indicate no meaningful differences across groups. Color coding: negligible, small, medium, large effect size.

Radar chart profile of writing quality

The radar chart (Figure 2) provides a visual summary of the average performance of abstracts generated by ChatGPT, DeepSeek, and Human authors across the six evaluated criteria.

ChatGPT displayed consistently high scores across all dimensions, forming a nearly

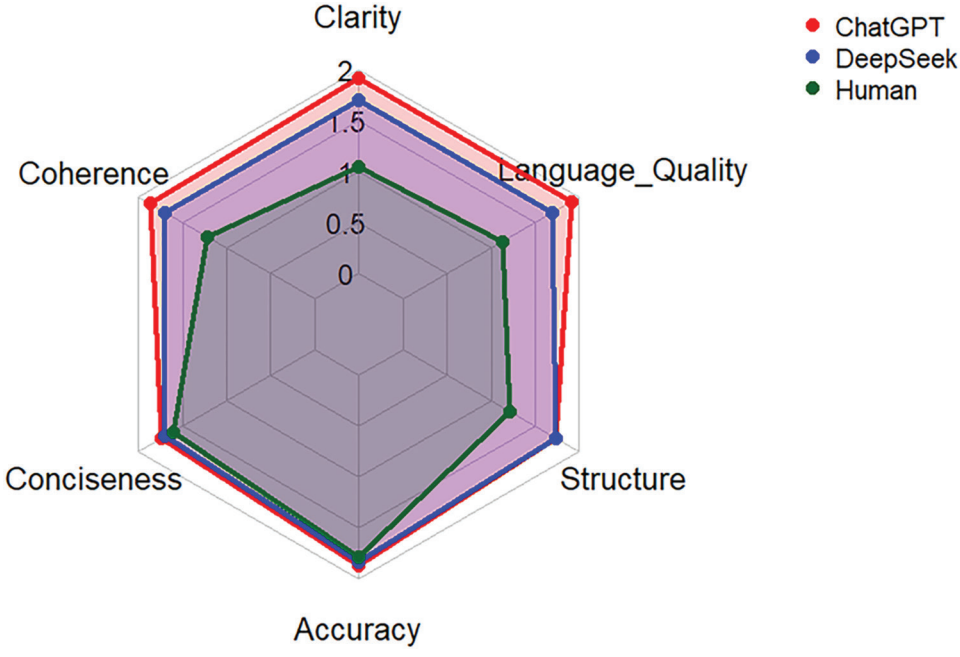

Figure 2. Radar chart comparing mean evaluation scores across writing methods.

symmetrical and expansive radar shape, with a particularly strong performance in Clarity, Coherence, Structure, and Language Quality. DeepSeek also demonstrated high performance in these dimensions, with a slightly reduced profile compared to ChatGPT. Human-authored abstracts showed noticeably lower mean scores, especially in Clarity and Language Quality, resulting in a more compressed radar area.

The radar chart provides a visual summary of the mean performance profiles across the six evaluated criteria. However, precise between-group comparisons should rely on the corresponding statistical analyses rather than visual interpretation alone.

Figure 2 visually summarizes the overall performance profiles across the six evaluation criteria. ChatGPT exhibits the most balanced and highest performance across dimensions, followed closely by DeepSeek. Human-authored abstracts demonstrate comparatively lower scores, particularly in clarity and language quality.

Discussion

This research provides a comparative evaluation of the quality of scientific abstracts produced by human authors, ChatGPT, and DeepSeek, focusing on six core dimensions of writing quality. Our findings extend the emerging JAMP literature highlighting both the strengths of generative AI in producing fluent, well-structured text and the need for careful interpretation and expert oversight when AI is used in medical academic work (14, 15). The results indicated that AI-generated abstracts—particularly those produced by ChatGPT—achieved significantly higher scores on clarity, coherence, structural organization, and language quality compared to human-authored texts. However, no meaningful differences were observed in conciseness and accuracy. These findings align with recent evidence suggesting that LLMs can produce scientific writing that rivals or even surpasses human output on several quality metrics (16, 17).

While our findings indicate that AI-generated abstracts performed strongly in stylistic and structural domains, several studies have raised concerns regarding the broader use of large language models in scientific writing. Reports have highlighted risks such as superficial reasoning, overgeneralized conclusions, fabricated citations, and reduced interpretative depth (18-20). In contrast to studies evaluating conceptual reasoning or methodological rigor, the present investigation focused primarily on formal and structural dimensions of abstract quality. This distinction may explain why our findings differ from more cautionary reports. Abstract-level writing emphasizes clarity, structure, and linguistic fluency—areas in which LLMs are particularly optimized—whereas deeper scientific reasoning and originality require domain expertise and critical interpretation. Therefore, differences across studies may reflect variation in evaluation criteria rather than fundamental contradictions in performance.

The superior performance of ChatGPT and DeepSeek in clarity and language quality may reflect their extensive pretraining on well-edited corpora and reinforcement learning protocols that optimize for fluency and user satisfaction (1). Notably, human-written abstracts lagged in linguistic polish and structural consistency, potentially due to variability in writing skill or time constraints during initial drafting. These observations are in the same line with the findings of Salvagno, et al. (21), who reported that LLM-generated medical abstracts scored higher in grammar and readability than those written by physicians.

However, the accuracy and conciseness dimensions revealed negligible differences across all models. This is particularly important as it suggests that while LLMs excel in stylistic and structural domains, they do not demonstrate significant advantages in factual correctness. This reinforces concerns noted in prior studies about the potential for hallucination or imprecision in AI-generated scientific content (19). Despite prompts designed to maintain factual accuracy, both ChatGPT and DeepSeek may still generate plausible-sounding but unverifiable statements, highlighting the importance of domain expert oversight when using LLMs in scholarly writing (18). Importantly, the operational definition of accuracy differed between human and AI-generated abstracts. Human abstracts were evaluated for consistency with their original published findings, whereas AI-generated abstracts were assessed only for absence of obvious inconsistencies with general medical knowledge and internal logical coherence. Because AI models did not have access to the original study data, direct verification of factual correspondence was not possible. Despite this definitional difference, the negligible effect size observed for accuracy likely reflects the models' ability to generate coherent and medically plausible text without gross errors. However, this finding should not be interpreted as confirmation of scientific validity or equivalence to data-grounded reporting.

The radar chart analysis further supports the robustness of AI-generated abstracts in terms

of uniform performance across qualitative dimensions. The radar visualization reflects the overall distribution of mean scores across criteria. Nevertheless, interpretation of relative performance is grounded primarily in the statistical test results rather than graphical shape comparison. This visualization resonates with a recent comparative evaluation by Yan, et al. (22), which emphasized GPT-4's balance across multiple human language understanding benchmarks. DeepSeek, while similarly effective, trailed slightly in coherence and clarity, likely due to its more recent development and smaller user calibration base. The concentration of median scores at the upper boundary of the 0–2 scale suggests potential ceiling effects and limited discriminatory resolution of the three-point ordinal scale. Future studies may benefit from a more granular scoring framework to increase sensitivity.

Blinded evaluations by experienced raters strengthen the validity of these findings, minimizing potential biases related to familiarity with AI writing styles. Additionally, the use of Cliff's Delta provided a nuanced quantification of effect sizes, revealing large differences in key areas and small differences between ChatGPT and DeepSeek—suggesting that both tools are viable alternatives to manual abstract drafting under expert supervision.

Despite these strengths, the study has limitations. First, while the evaluation focused on writing quality, it did not assess scientific depth or novelty—key attributes in manuscript evaluation. Second, the use of short-form abstracts may not generalize to longer or more complex scientific texts. Future studies should explore whether similar advantages persist across full-length manuscripts and domain-specific writing tasks such as methods sections or data interpretation.

Furthermore, ethical and academic implications warrant scrutiny. While AI can augment scientific productivity and reduce language barriers (23), overreliance without critical oversight could erode research integrity or perpetuate misinformation (20). Transparent disclosure and human validation remain essential when integrating AI tools into scholarly workflows. A major limitation concerns the asymmetry in source material. Human abstracts were derived from complete peer-reviewed studies, whereas AI-generated abstracts were produced solely from article titles without access to original datasets or manuscripts. This structural difference limits the comparability of the "accuracy" metric across groups and should be considered when interpreting findings related to factual correctness.

A key methodological limitation concerns the asymmetry in information access between groups. Human abstracts were derived from complete published studies, whereas AI-generated abstracts were produced solely from article titles without access to original datasets, methods, or full manuscripts. Consequently, the evaluation of "accuracy" for AI-generated abstracts was limited to assessment of logical plausibility and absence of obvious inconsistencies with general medical knowledge rather than verification against real study findings. This difference constrains direct comparability of factual accuracy across groups and should be considered when interpreting these results.

It is important to emphasize that the present study evaluated primarily formal and structural aspects of abstract quality, including clarity, coherence, structural completeness, and language fluency. These dimensions reflect communicative effectiveness rather than scientific originality or interpretative depth. The intrinsic value of a scientific abstract extends beyond eloquence and structural correctness to include conceptual insight, methodological rigor, and meaningful interpretation of results. An abstract may be linguistically polished yet conceptually superficial. Therefore, higher scores in stylistic domains should not be equated with superior scientific contribution. Future research should incorporate measures of conceptual depth and analytical reasoning to provide a more comprehensive assessment of AI-assisted scientific writing.

From a practical standpoint, scientific abstracts—whether written by humans or generated by AI—are rarely published without expert revision. The present findings suggest that AI-generated abstracts may provide a structurally coherent and linguistically polished starting point, potentially reducing the time required for structural editing. In contrast, human-authored abstracts may offer greater contextual nuance, interpretative flexibility, or originality of framing, even if they require refinement in clarity or structure. Rather than positioning AI and human writing as competing approaches, these results support a collaborative model in which AI serves as an assistive drafting tool, with domain experts providing critical oversight, conceptual depth, and scientific validation.

Conclusion

Our findings indicate that ChatGPT and DeepSeek generate scientific abstracts with significantly higher scores in clarity, coherence,

structure, and language quality compared to human-authored abstracts. Importantly, no significant differences were observed in conciseness or factual accuracy. These results suggest that large language models may offer advantages in stylistic and structural aspects of abstract drafting, while maintaining comparable performance in content-sensitive dimensions. Nevertheless, expert oversight remains essential to ensure scientific rigor and interpretative depth.

Authors' Contributions

KR.Z Conceptualization, methodology, formal analysis, data curation, writing the first draft, visualization, N.Sh Conceptualization, validation, review, and editing of the first draft, supervision.

Conflicts of interest

The authors declare no conflicts of interest.

Declaration of AI Use

The authors have not used AI in preparation of this article.

References

1. Jin I, Tangsrivimol JA, Darzi E, Hassan Virk HU, Wang Z, Egger J, et al. DeepSeek vs. ChatGPT: prospects and challenges. Front Artif Intell. 2025;8:1576992.
2. Azam S, Ali ME, Ahmad J, Mim MMJ, Sakib S, Fahad NM, et al. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access. 2024;12:26839-74.
3. Phogat R, Arora D, Mehra PS, Sharma J, Chawla D, editors. A comparative study of large language models: ChatGPT, DeepSeek, Claude and Qwen. 2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT); 2025.
4. Khatri BB. Writing an effective abstract for a scientific paper. Nep J Dev Rural Stud. 2022;19(01):1-7.
5. Majhi S, Sahu L, Behera K. Practices for enhancing research visibility, citations and impact: review of literature. Aslib J Inf Manag. 2023;75(6):1280-305.
6. Luby S, Southern DL. Achieving clarity and conciseness. The Pathway to Publishing: A Guide to Quantitative Writing in the Health Sciences. Singapor: Springer; 2022. pp. 73-86.
7. Sollini M, Pini C, Lazar A, Gelardi F, Ninatti G, Bauckneht M, et al. Human researchers are superior to large language models in writing a medical systematic review in a comparative multitask assessment. Sci Rep. 2025;16(1):173.
8. Ripoll Y Schmitz LM, Sonnleitner P. Evaluating AI-generated vs. human-written reading comprehension passages: an expert SWOT analysis and comparative study for an educational large-scale assessment. Large-Scale Assess Educ. 2025;13(1):20.
9. Madupati B, Jonnalagadda AK, Madupathi S, Kassetty N, Jakku PC, Raghavan P, et al. A technical comparison of ChatGPT and DeepSeek: Architecture, efficiency, and performance. Int J Glob Innov Solut. 2025.
10. Keykha A, Behravesh S, Ghaemi F. ChatGPT and medical research: a meta-synthesis of opportunities and challenges. J Adv Med Educ Prof. 2024;12(3):135-47.
11. Keykha A, Fazlali B, Behravesh S, Farahmandpour Z. Integrating artificial intelligence in medical education: a meta-synthesis of potentials and pitfalls of ChatGPT. J Adv Med Educ Prof. 2025;13(3):155-72.
12. Jafari F, Keykha A, Taheriankalati A, Taghavi Monfared A. The role of AI in shaping medical education: insights from an umbrella review of review studies. J Adv Med Educ Prof. 2025;13(4):270-93.
13. Khalifa M, Albadawy M. Using artificial intelligence in academic writing and research: An essential productivity tool. Comput Methods Programs Biomed Update. 2024;5:100145.
14. Alizadeh M, Jafar Sameri M. Intelligent assessment systems in medical education: A systematic review. J Adv Med Educ Prof. 2025;13(3):173-90.
15. Mir MM, Mir GM, Raina NT, Mir SM, Mir SM, Miskeen E, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. J Adv Med Educ Prof. 2023;11(3):133-40.
16. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15(4):616-26.
17. Thelwall M. Research quality evaluation by AI in the era of large language models: advantages, disadvantages, and systemic effects – An opinion paper. Scientometrics. 2025;130(10):5309-21.
18. Gao R, Yu D, Gao B, Hua H, Hui Z, Gao J, et al. Legal regulation of AI-assisted academic writing: challenges, frameworks, and pathways. Front Artif Intell. 2025;8:1546064.
19. Lee C, Kim J, Lim JS, Shin D. Generative AI risks and resilience: How users adapt to hallucination and privacy challenges. Telemat Inform Rep. 2025;19:100221.
20. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ. 2024;11(1):28.
21. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care. 2023;27(1):75.
22. Yan J, Yan P, Chen Y, Li J, Zhu X, Zhang Y. Benchmarking LLMs against human translators: A comprehensive evaluation across languages, domains, and expertise levels. IEEE Trans Big Data. 2025;99:1-16.
23. Yener H, Yener S. Overcoming Language Barriers With Artificial Intelligence. Harnessing AI for multigenerational English language learning. USA: IGI Global Scientific Publishing; 2026. pp. 97-122.