Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327

Arab World English Journal (AWEJ) Special Issue on CALL Vol. 17, No.12. July 2026 Pp.418-434
DOI: https://dx.doi.org/10.24093/awej/call12.24

Artificial Intelligence in Medical Writing Education for Non-Native English Speakers: A Systematic Review of Applications, Benefits, and Risks

Wei Shen
Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Selangor, Malaysia;
Corresponding Author, P120860@siswa.ukm.edu.my

Afendi Hamat
Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Selangor, Malaysia
Email: fendi@ukm.edu.my

Anis Nadiah Che Abdul Rahman
Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Selangor, Malaysia
Email: **aniscar@ukm.edu.my**

Jun Wang
[4]School of Foreign Studies, Yanshan University, Qinhuangdao, China
Email: wywy@ysu.edu.cn

Received: 04/05/2026 Accepted:07/07/2026 Published:07/15/2026

Abstract
This review focuses on how artificial intelligence has been used to support English medical writing among non-native English speakers. The purpose was not to list every available tool, but to identify what the existing literature shows about their uses, benefits, limitations, and risks. Searches were conducted in Google Scholar, Scopus, and PubMed with terms related to artificial intelligence, medical or healthcare writing, and non-native English users. The search first produced 978 records. After duplicate removal, title and abstract screening, and full-text assessment, 24 sources were included in the review. These sources mainly discussed four forms of AI support: grammar and style checking, terminology and lexical assistance, summarisation or simplification, and large language model-based drafting or revision. Across the reviewed literature, AI tools were most often linked to grammar correction, vocabulary choice, readability, writer confidence, and the simplification of medical information for non-specialist readers. These benefits were useful but mostly short-term. Many studies used small samples, brief periods of tool use, or outcome measures that could not show whether learners became better independent writers. It is safer to view AI as a writing support tool rather than a replacement for teachers, supervisors, or writing instruction. Its use in English medical writing education should be guided by clear rules, disclosure of AI assistance, and careful checking factual accuracy in medical text.

Keywords: Artificial intelligence, English as a foreign language, English as a second language, large language models, medical education, medical writing, non-native English speakers, systematic review

Cite as: Shen, W., Hamat, A., Che Abdul Rahman, A. N., & Wang, J. (2026). Artificial Intelligence in Medical Writing Education for Non-Native English Speakers: A Systematic Review of Applications, Benefits, and Risks. *Arab World English Journal (AWEJ) Special Issue on CALLVol.17*. (12)418-434. https://dx.doi.org/10.24093/awej/call12.24

Introduction

Most medical and scientific journals publish in English. For researchers, clinicians, and students whose first language is not English, this creates a practical difficulty: they are often expected to report research, complete coursework, and communicate professional knowledge in a language that they did not grow up using. Medical students, researchers, and health professionals may need to write abstracts, reports, research articles, case-related texts, and other academic or professional materials in English. Writing these texts is not simply a matter of translating ideas from one language into another. Effective medical writing requires accurate terminology, grammatically controlled sentences, expected patterns of organisation, and an awareness of how claims are normally presented to academic, clinical, and public audiences.

For many non-native English speakers, English medical writing is a demanding task. Writers may understand the medical content but still find it difficult to express that content in a way that meets the expectations of journal editors, reviewers, teachers, or clinical readers. Amano et al. (2023) noted that language barriers can increase the time needed for writing and may also contribute to problems during revision and publication. In medical education, Kasai et al. (2024) also observed that students may be asked to complete report-writing tasks without sufficient prior training. This leaves many learners in a difficult position: they must communicate complex medical information accurately, but they may receive limited formal instruction in English medical writing. In some programmes, writing support is minimal; in others, feedback from supervisors may arrive too late or remain too general to guide revision effectively.

AI-assisted writing tools have become one possible form of support for these writing difficulties. Such tools may help writers identify language errors, compare alternative ways of expressing an idea, revise sentence structure, and simplify technical content for non-specialist readers. These forms of support are especially relevant for non-native English speakers who need to produce medical texts under academic or professional time pressure. However, the usefulness of AI depends heavily on how it is used. If students or researchers accept automated suggestions without checking them carefully, problems may arise in relation to factual accuracy, academic integrity, authorship, disclosure, and responsibility for the final text.

Recent studies have examined AI in academic writing and higher education from several perspectives, including learning support, assessment, and writing assistance (Aljuaid, 2024; Jen & Salam, 2024). These studies provide useful background for understanding AI-supported writing, but they do not focus specifically on how AI is being used to support English medical writing among non-native speakers. The relevant evidence is also scattered across different tools, learner groups, writing tasks, and research designs. A focused review is needed to clarify what has been studied, what benefits have been reported, and what limitations or risks remain.

This systematic review looks at AI-assisted English medical writing education for non-native English speakers. It reviews the types of AI tools discussed in the literature, the forms of writing support associated with these tools, and the reported benefits, risks, and methodological limitations. The review was guided by the following research questions:

1. What kinds of artificial intelligence applications have been reported in English medical

writing education for non-native English speakers?

2. What benefits of these applications have been reported?
3. What risks, limitations, and ethical concerns have been identified?

Literature Review

Conceptual Scope and Key Terms

In this review, artificial intelligence is used as a broad term for digital systems that support writing through automated feedback, text revision, text generation, or text simplification. This includes grammar-checking software that identifies spelling, grammar, and style problems, as well as large language models that can generate or revise extended passages from prompts. These tools do not work in the same way. Some are designed mainly to correct grammar, style, terminology, or vocabulary in an existing draft. Others can help writers reorganise ideas, rephrase sentences, prepare summaries, or simplify technical information for different audiences. For this reason, their educational value and risks need to be considered separately rather than treated as one single type of writing technology.

English medical writing education refers here to teaching, training, or support related to writing in medical, healthcare, and clinical contexts. The writing tasks considered in this review include research abstracts, reports, articles, assignments, case-related texts, and plain-language summaries. These tasks require more than grammatical accuracy. Writers also need to present specialised information clearly, accurately, and appropriately for academic, clinical, or public readers. Non-native English speakers refer to learners, trainees, researchers, and professionals who need to use English for these purposes but do not use English as their first language.

Pedagogical Framework: Artificial Intelligence as Scaffolding

This review understands AI-assisted writing support through the concept of pedagogical scaffolding. Scaffolding refers to temporary support that helps learners complete tasks they cannot yet manage fully on their own (Wood et al., 1976). In medical writing, AI may provide this kind of support by pointing out language problems, offering alternative expressions, helping writers revise unclear sentences, or adapting technical content for non-specialist audiences. Used in this way, AI can help learners notice features of medical English that they may otherwise miss. However, the support becomes less educational when the tool simply replaces the learner's own thinking or when students accept generated text without checking its accuracy, suitability, and rhetorical effect.

A scaffolding perspective is useful because research on AI in education has often been criticised for weak pedagogical grounding. Zawacki-Richter et al. (2019) argued that earlier studies of AI in higher education had a "weak connection to theoretical pedagogical perspectives" (p. 1). More recently, Bond et al. (2024) called for stronger attention to ethics, collaboration, and methodological rigour in AI-related higher education research. For the present review, scaffolding offers a practical way to distinguish between AI use that supports writing development and AI use that merely improves a single draft. The central question is not only whether AI makes a text cleaner, but whether it helps learners become more capable and responsible medical writers over time.

Applications and Reported Educational Value

Much of the research on automated writing support has examined tools that provide

feedback on grammar, punctuation, vocabulary, and sentence style. Grammarly is a common example. O'Neill and Russell (2019) found that students responded more positively to Grammarly when its feedback was supported by teacher guidance. They also cautioned that automated feedback should not replace writing instruction because the tool's suggestions were not always accurate. Yousofi (2022) similarly reported generally positive attitudes among teachers and students, although the study also pointed to limits in automated feedback. These findings suggest that grammar-oriented tools may support independent revision, but their value depends on whether learners know how to interpret, question, and apply the feedback.

Generative AI has widened the scope of writing assistance. Large language models can support idea generation, drafting, rewriting, summarising, analysing, and simplification. Reviews of AI in academic and essay writing have discussed uses such as language feedback, revision support, idea development, and general writing assistance (Aljuaid, 2024; Jen & Salam, 2024). Aljuaid (2024) concluded that AI tools are better understood as complements to academic writing instruction rather than replacements for traditional teaching. Broader reviews of AI in higher education have also covered adaptive systems, assessment, personalisation, and institutional use (Bond et al., 2024; Zawacki-Richter et al., 2019).

The literature points to two main kinds of AI support. One form focuses on local language correction, such as grammar, word choice, and sentence clarity. The other offers wider support for drafting, revising, paraphrasing, summarising, or simplifying medical information. Both forms may be useful, but neither automatically leads to writing development. A cleaner draft does not necessarily mean that the writer has gained stronger control of medical terminology, evidence, audience awareness, or genre expectations. Learners still need instruction in how medical claims are organised, how evidence is presented, and how language choices affect readers. AI support is best viewed as part of a wider system of writing guidance rather than as a substitute for teaching.

Risks, Governance, and Research Gap

Research on scientific communication has also raised concerns about generative AI. Gao et al. (2023) showed that ChatGPT could generate scientific abstracts that appeared realistic but contained fabricated information. Else (2023) likewise reported that AI-generated abstracts could be difficult to distinguish from human-written abstracts. These findings are important for medical writing education because factual accuracy is not optional in medical communication. A text may appear fluent and well organised while still being misleading or incorrect.

Authorship and responsibility are also major concerns. Stokel-Walker (2023) reported controversy over papers that listed ChatGPT as an author. The International Committee of Medical Journal Editors (ICMJE, 2025) states that human authors remain responsible for submitted material that includes AI-assisted content. This means that students, researchers, and supervisors must verify AI-assisted writing carefully, acknowledge or disclose AI use according to journal and institutional rules, and take responsibility for the final text.

Existing reviews have provided useful insights into AI in higher education, academic writing, and essay writing. Individual studies have also examined particular scientific or medical writing tasks. However, the literature has not yet offered a focused account of how AI supports English medical writing education for non-native English speakers. The evidence remains scattered across different technologies, learner groups, writing genres, outcome measures, and research designs. This review brings together the reported applications, educational benefits, risks, and methodological limitations of AI-assisted English medical writing support in this field

Methods

Protocol and Reporting

We structured this review according to the PRISMA 2020 guidelines (Page et al., 2021). The PRISMA checklist was used to guide the reporting of the search, screening, eligibility assessment, and final inclusion process. The flow diagram in Figure 1 shows how records moved from identification to screening, full-text assessment, and inclusion. No review protocol was registered before the study was conducted.

Information Sources and Search Strategy

Three electronic sources were searched: Google Scholar, Scopus, and PubMed. Google Scholar was included because it captures a wide range of materials, including journal articles, conference papers, and institutional sources. Scopus was used because of its coverage of peer-reviewed journals in education, linguistics, and related fields. PubMed was included because it indexes medical and health-science literature, where studies on medical writing and communication are likely to appear.

The search covered records available up to September 1, 2025, with no restriction on publication date. Search terms were organised around three concept blocks: artificial intelligence ("AI" OR "artificial intelligence"), medical and healthcare writing ("medical writing" OR "healthcare writing" OR "clinical writing"), and learner background ("ESL" OR "EFL" OR "non-native"). Because the three sources have different search functions, the search fields and Boolean operators were adjusted where necessary. When the database allowed it, searches were conducted in titles, abstracts, and keywords.

Eligibility Criteria

Sources were included if they examined the use of artificial intelligence to support medical, healthcare, or clinical writing in English and focused on non-native English-speaking learners, trainees, researchers, or professionals. Sources involving educators or supervisors were also included when their reported experiences or practices concerned writing support for these groups. Relevant writing tasks included research abstracts, reports, articles, assignments, case-related texts, and plain-language summaries.

Empirical studies using quantitative, qualitative, or mixed-methods designs were included. Non-empirical reports were also retained when they directly addressed the applications, educational implications, risks, or governance of AI in English medical writing. These reports were used to inform the narrative synthesis and discussion, but they were not included in the methodological quality appraisal.

Sources were excluded if they did not concern medical, healthcare, or clinical writing; did not involve an AI application; focused exclusively on native English speakers; or examined writing in languages other than English. Only sources published in English were included.

Screening and Selection

All records retrieved from the three electronic sources were exported to a spreadsheet for management and deduplication. After duplicate records were removed, titles and abstracts were screened against the eligibility criteria. Full texts were then assessed for records that appeared potentially relevant. During full-text assessment, reasons for exclusion were recorded using categories aligned with the eligibility criteria. These included lack of relevance to medical,

healthcare, or clinical writing; absence of an AI application; exclusive focus on native English speakers; target language other than English; and publication language other than English. The full selection process is presented in Figure 1.

Source Types and Outcomes of Interest

The relevant empirical studies were in the forms of quantitative, qualitative and mixed-methods research on AI in English medical writing education. The content of the review included indicators such as linguistic accuracy, readability, discourse quality, writer perception and confidence, and the quality or accessibility of task-specific outputs, as well as plain-language summaries. The studies that reported on this also examined the efficiency, revision methods and the experience of learners and teachers using AI-assisted writing tools.

Non-empirical reports were included only if they addressed the problems in the review questions directly. They are used to collect information on the application of AI, education, authorship and disclosure, ethics and governance. Since these sources did not empirically assess learners' results, they were considered to be in the context rather than evidence supporting writing enhancement.

Methodological Quality Appraisal

Methodological quality was assessed using appraisal tools appropriate to the study design. For quantitative studies, the relevant Joanna Briggs Institute critical appraisal checklists were used. For qualitative studies, the Critical Appraisal Skills Programme Qualitative Studies Checklist was applied. For mixed-methods studies, the Mixed Methods Appraisal Tool was used (Aromataris et al., 2024; Critical Appraisal Skills Programme, 2024; Hong et al., 2018). Appraisal was carried out at the study level, and the judgments were summarised descriptively. The assessment covered the 19 empirical studies only. The five non-empirical reports were not appraised because the selected tools were designed for empirical research.

Data Extraction

Data were extracted using a standardised form. The form recorded bibliographic information, country or setting, type of publication, study design, participant characteristics, AI application, writing task, outcome measures, main findings, and reported limitations. When available, participant information was also coded by field of study, level of education or training, and language background. Writing tasks were grouped into research abstracts, reports, articles, assignments, case-related texts, and plain-language summaries.

AI applications were coded according to their main writing-support function. Four functional groups were used: grammar and style support, terminology or lexical assistance, summarisation or simplification, and LLM-based drafting or revision. The extent of support was also noted. Surface-level support referred to grammar, style, and vocabulary. Discourse-level support referred to cohesion, organisation, and genre-related choices. Task-specific support referred mainly to summarisation and simplification.

Empirical and non-empirical sources were handled slightly differently during extraction. For empirical studies, data were extracted on participants, writing outcomes, learner or educator perceptions, and reported use of AI tools. For non-empirical reports, extraction focused on information relevant to the review questions, especially educational use, ethical concerns, authorship, disclosure, and governance. Contextual information, such as course integration,

supervision, and evaluation practices, was recorded when it was available.

Synthesis Approach

Meta-analysis was not feasible because the included sources differed in AI applications, participant groups, writing tasks, research designs, and outcome measures. A narrative synthesis was used. The synthesis followed the general approach described by Popay et al. (2006), including developing an initial account of the evidence, examining relationships across sources, and considering the strength and limitations of the evidence.

The synthesis focused on three areas: the forms of AI support reported in the literature, the benefits associated with these tools, and the risks, limitations, and ethical concerns identified across the sources. Results that appeared across multiple studies were noted, as were contradictions or gaps in the evidence. Quantitative results, including readability scores and other writing-related indicators, were reported descriptively because the measures were not sufficiently consistent for statistical pooling. Non-empirical reports were used to support the discussion of educational implications, authorship and disclosure, ethics, and governance.

Results

Study Selection

Figure 1 presents the study selection process. The searches identified 978 records across Google Scholar, Scopus, and PubMed. After 80 duplicate records were removed, 898 records were screened by title and abstract. Of these, 725 were excluded because they clearly did not meet the eligibility criteria. The remaining 173 full texts were assessed in more detail. A further 149 sources were excluded for the following reasons: no direct relevance to medical writing (n = 49), no AI application (n = 41), exclusive focus on native English speakers (n = 55), target language other than English (n = 1), or publication in a language other than English (n = 3). In total, 24 sources met the inclusion criteria and were included in the narrative synthesis.

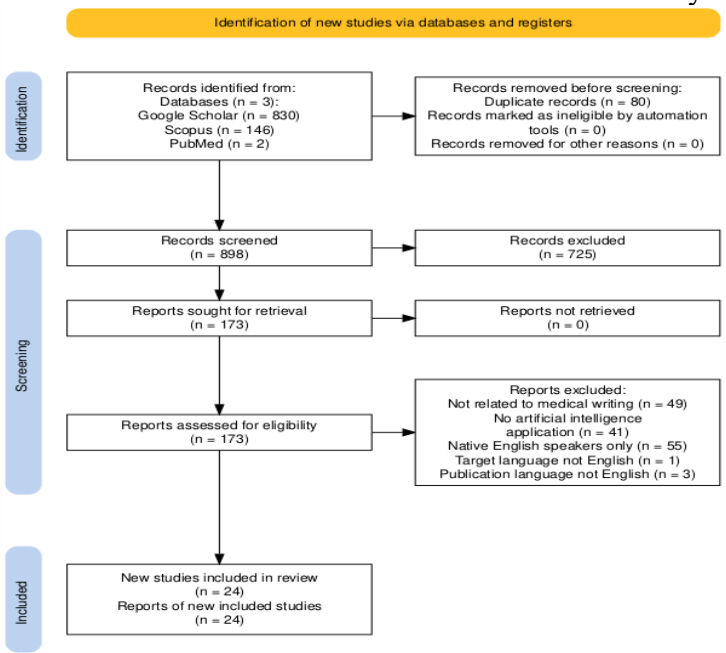


Figure 1. PRISMA 2020 flow diagram of study selection

Study Characteristics

The 24 included sources consisted of 19 empirical studies and five non-empirical reports published between 2019 and 2025. Among the empirical studies, six used quantitative designs, seven used qualitative designs, and six used mixed-methods designs. Sample sizes varied across the empirical studies, ranging from 10 participants in small qualitative studies to more than 300 participants in larger quantitative studies. Most participants were medical students at undergraduate or postgraduate level, although some studies also included teachers, faculty members, or supervisors. The studies were conducted in several settings, including China, Japan, Malaysia, Saudi Arabia, and European countries.

Table 1. *Distribution of included source types and sample size ranges (n = 24)*

Source type	Number of sources	Typical sample size range
Quantitative studies	6	n = 40 – 300+
Qualitative studies	7	n = 10 – 40
Mixed-methods studies	6	n = 15 – 100
Non-empirical reports	5	Not applicable

Note. Sample sizes refer to participant numbers reported in empirical studies. Non-empirical reports did not involve human participants.

Artificial Intelligence Applications

The included sources were classified according to their main form of AI-supported writing assistance. Nine sources focused mainly on grammar and style enhancement, eight examined LLM-assisted drafting or revision, four addressed summarisation or simplification, and three focused on terminology or lexical support (Table 2). Although some tools could serve more than one function, each source was assigned to the category that best represented its main use in the study or report.

Table 2. *Principal AI applications in English medical writing education (n = 24)*

AI application type	Number of sources	Representative tools
Grammar and style enhancement	9	Grammarly, Writefull
Terminology and lexical aids	3	Writefull, glossaries
Summarization/simplification	4	GPT-3.5/4, Claude, Gemini
Large Language Models (LLMs)	8	ChatGPT, Copilot, Perplexity, Meta AI

Note. The review assigned each source to the category that best represented its main AI application.

Methodological Quality Appraisal

Methodological quality was assessed for the 19 empirical studies. The five non-empirical reports were not appraised because they were used as contextual sources rather than empirical evidence of writing improvement. The appraisal showed uneven methodological strength across the empirical literature. Quantitative studies generally stated their aims clearly and used relevant outcome indicators, but several had small samples and did not use randomised designs. Qualitative studies often provided coherent thematic accounts, although reporting on reflexivity and ethics was inconsistent. Mixed-methods studies offered useful combined evidence, but some did not explain clearly how qualitative and quantitative findings were integrated.

To conclude, eight empirical studies were judged to have a low risk of bias, seven had a medium risk of bias, and four had a high risk of bias because of design limitations, incomplete reporting, or weak integration of findings.

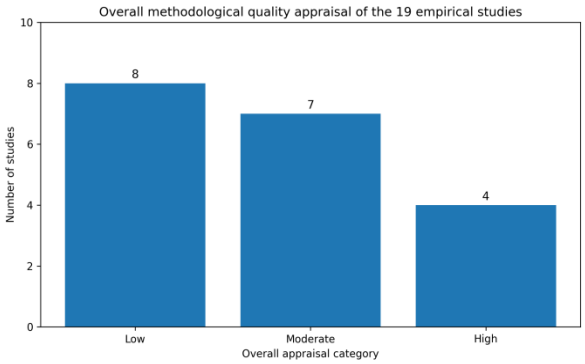


Figure 2. Overall methodological quality appraisal of the 19 empirical studies

Note. The appraisal included only the 19 empirical studies. The review did not appraise the five non-empirical reports because the selected critical appraisal tools were designed for empirical research.

Thematic Synthesis

The synthesis showed a mixed picture. On the positive side, the included sources most often associated AI-assisted writing with improvements in linguistic accuracy, readability, writer confidence, and the accessibility of medical information for non-specialist readers. These benefits were most clearly reported for grammar and style support, sentence-level revision, and plain-language summarisation.

The sources repeatedly raised concerns about factual accuracy, plagiarism, authorship, transparency, and dependence on automated revision. These concerns were especially relevant when AI tools generated new content, substantially rewrote existing drafts, or simplified technical medical information. The synthesis suggests that AI can support English medical writing for non-native speakers, but the evidence does not justify treating it as an independent substitute for writing instruction, expert feedback, or human responsibility.

Discussion

This review found that AI has been used in English medical writing education for non-native speakers in several ways. The reported applications included grammar and style support, terminology and lexical assistance, summarisation, simplification, and LLM-based support for brainstorming, drafting, and revision. Although the included studies differed in design and quality, they most often associated AI use with improved linguistic accuracy, better readability, and increased writer confidence. These findings are relevant to ESL/EFL medical writers because difficulties with grammar, article use, collocation, and sentence-level expression have long been reported in second-language writing (Ionin et al., 2004; Laufer & Waldman, 2011). Clear written communication is also important in medical education, academic publishing, and disciplinary participation (Hyland, 2003). Even so, the evidence does not support unrestricted use of AI. Given the concerns about factual reliability, academic integrity, and over-reliance on automated revision, AI is better understood as a teacher-supported writing aid than as an independent writing solution.

The reported benefits fell into two main areas. The first concerned visible improvements in the text itself, especially fewer surface-level errors and more readable sentences. Grammar and style tools, as well as LLM-assisted revision, were commonly reported to help writers correct local errors and improve sentence clarity. This is broadly consistent with findings from other higher-

education writing contexts, where automated feedback tools have been viewed as useful but incomplete forms of support (O'Neill & Russell, 2019; Yousofi, 2022). The second area concerned learner experience. Several studies reported that students felt more confident when they could use AI tools to check language, revise drafts, or simplify complex information. This confidence matters because non-native English-speaking medical students often need to write in high-pressure academic or professional contexts.

At the task level, plain-language summarisation appeared as a particularly relevant use of AI. Some studies suggested that AI tools can help convert technical medical information into language that is more accessible to non-specialist readers. This is useful because medical education increasingly expects students and researchers to communicate not only with specialists but also with patients, families, and the wider public. However, this benefit depends on careful verification. In medical communication, simplification cannot come at the cost of accuracy. An AI-generated summary may be easier to read, but it still needs to preserve the meaning, limits, and evidence of the original medical information.

The reviewed literature also raised serious concerns. The most prominent was factual accuracy. When LLMs generated new content or revised existing drafts, they could introduce false or unsupported information that appeared fluent and credible. This problem has also been reported in studies of AI-generated scientific abstracts, where both automated detectors and human reviewers had difficulty distinguishing generated abstracts from real ones (Else, 2023; Gao et al., 2023). A second concern was academic integrity. It is not always clear where acceptable language support ends and inappropriate text generation begins. Current publication guidance also makes clear that AI tools cannot be treated as authors and that human authors remain responsible for the final text (Committee on Publication Ethics [COPE], 2023; International Committee of Medical Journal Editors [ICMJE], 2025). A third concern was pedagogical. If students repeatedly allow AI to rewrite their work without reflection, they may become less able to draft, evaluate, and revise independently. Miao and Holmes (2023) also warned about this problem in their guidance on AI in education and research. These concerns suggest that AI use should be taught explicitly rather than left to informal trial and error.

The strongest evidence in the reviewed studies concerned surface-level writing support. AI tools were most useful for article use, tense consistency, subject-verb agreement, collocation, word choice, and sentence fluency. These are important areas for non-native English-speaking medical writers, and support at this level can make writing more readable and less frustrating. The evidence was weaker for higher-order writing. Fewer studies examined whether AI improved argument development, section coherence, genre awareness, or the rhetorical organisation of medical claims. Very few used validated instruments to assess genre conformity or discourse-level quality. For this reason, it remains uncertain whether AI support leads to deeper writing development rather than simply producing a cleaner version of a single draft. In medical programmes where formal writing instruction may already be limited (Hyland, 2016; Kasai et al., 2024; Wald et al., 2012), AI may be most defensible as a lower-level revision aid, provided that teachers continue to teach discourse, genre, and critical appraisal.

The findings should be interpreted cautiously. Many empirical studies used convenience samples from a single institution and lasted only a few weeks. Randomised trials were rare, and few studies followed learners beyond the immediate writing task. Several mixed-methods studies reported both qualitative and quantitative findings, but did not always explain clearly how the two forms of evidence were integrated. These design limitations make it difficult to compare studies

or determine whether reported improvements reflect lasting gains in independent writing ability. At present, the evidence mainly shows that AI-assisted writing can reduce surface errors, improve readability, and increase self-reported confidence in the short term.

Another limitation in the evidence base concerns learner differences. The empirical studies came from Asia, the Middle East, and Europe, which suggests that interest in AI-supported medical writing is not limited to one region. However, many studies treated non-native English speakers as a single group. Few compared beginners and advanced writers, or examined whether AI support worked differently across medical genres, such as abstracts, reports, case writing, or plain-language summaries. This may partly explain why the findings are uneven. Future studies need to compare proficiency levels, genres, and task conditions more directly so that educators can identify who benefits from AI assistance, for what kind of writing, and under what type of guidance.

Ethical and policy issues should also be considered alongside the reported benefits. AI-assisted plain-language writing may improve access to medical information, but trust can be damaged if users do not disclose AI assistance or if responsibility for the final text is unclear. These concerns are consistent with current guidance on AI use, authorship, and scholarly responsibility (COPE, 2023; ICMJE, 2025). For medical educators, the practical response is not simply to ban AI or allow it without limits. Students need to learn how to use AI-assisted writing tools responsibly, verify outputs, question revision suggestions, and disclose assistance when required. In this sense, writing instruction should treat AI use as part of professional communication training, not only as a language-polishing technique (Hyland, 2016; Norcini & Burch, 2007; ten Cate, 2005; Wald et al., 2012).

In summary, the safest reading of the evidence is a cautious one. AI tools can make English medical drafts clearer and less error-prone, especially for writers who struggle with grammar, vocabulary, or readability. Yet these tools also create risks when they generate claims, reshape medical meaning, or hide the writer's own contribution. The most reasonable use is supervised and limited: AI can assist with drafting and revision, but teachers, supervisors, and writers still need to check accuracy, genre appropriateness, disclosure, and final responsibility.

Educational Implications

Medical education programmes often have limited space for writing instruction because curricula are already crowded with clinical training, science courses, assessments, and professional requirements. In this context, the reviewed studies suggest that AI may be useful for routine language-polishing tasks, such as article use, tense consistency, subject-verb agreement, word choice, collocation, and sentence clarity. If these lower-level issues can be addressed more efficiently, teachers and supervisors may have more time to focus on argument quality, research accuracy, genre appropriateness, and disciplinary meaning. This distinction is important because writing problems in medical education are not only grammatical. Non-native English-speaking medical students often need support in turning medical knowledge into genre-appropriate academic or professional discourse (Kasai et al., 2024; Norcini & Burch, 2007; ten Cate, 2005).

AI-assisted writing tasks may also serve as early-stage support for learners who feel uncertain about writing in English. Several reviewed studies reported that students felt more confident when they used AI tools for abstract writing, case-related writing, or plain-language summaries. This kind of support may encourage participation, especially among learners who understand the medical content but hesitate to express it in English. However, support should be structured rather than passive. Students should be asked to compare AI suggestions with their own

drafts, decide whether to accept or reject the suggestions, and explain their revision choices. In this way, AI can be used to support learning without encouraging long-term dependence.

The findings also have implications for equity in medical writing education. Non-native English-speaking students may be judged unfavourably because of surface-level language errors, even when their disciplinary knowledge is strong. Carefully supervised AI use may reduce some of these barriers by helping students prepare clearer drafts for discussion with peers, teachers, or supervisors. This connects with broader concerns about fairness in medical education and academic publishing, where participation in research communication should not depend only on first-language English proficiency (Hyland, 2016; Wald et al., 2012). At the same time, equity should not be used to justify unrestricted AI use. Institutions need clear rules about what kinds of assistance are acceptable, how AI use should be disclosed, and how AI-assisted work should be evaluated (COPE, 2023; ICMJE, 2025).

Plain-language summary writing offers a practical setting for AI-assisted writing activities. As medical education pays more attention to patient involvement and public communication, students increasingly need to present research findings in ways that non-specialist readers can understand (Hyland, 2003; Schnell, 2024). AI tools can help students compare different versions of the same text, notice changes in readability, and examine whether simplification has affected scientific accuracy. Such activities allow students to learn with AI while also learning about the limits of AI. The tool may be used to draft or revise a summary, but students must still check whether the medical meaning has been preserved and whether any ethical or communicative risks have been introduced (ICMJE, 2025).

Teacher training is also necessary. Supervisors and writing instructors need practical strategies for responding to AI-assisted drafts, distinguishing acceptable language support from inappropriate text generation, and teaching students how to verify, revise, and disclose AI use. A guided approach is likely to be more sustainable than either banning AI completely or allowing unrestricted use. In English medical writing education, AI is best treated as a supervised resource: useful for early drafting, language revision, and audience adaptation, but still connected to explicit learning goals, human feedback, and the development of independent disciplinary writing ability.

Limitations and Future Research

The findings of this review should be interpreted with caution. AI-assisted English medical writing education is still a relatively new area, and much of the available research is small-scale and exploratory. Many studies used samples from a single institution, which limits the extent to which their findings can be generalised to other learner groups, medical programmes, or stages of training. Important learner-related factors were also not always reported or analysed systematically, including English proficiency, level of study, previous writing experience, and familiarity with medical genres.

The short duration of many studies is another limitation. Several studies examined AI use during a single assignment, workshop, or short teaching period. As a result, the current evidence tells us more about immediate writing support than about long-term writing development. Improvements in accuracy, readability, or confidence may reflect the support provided during a particular task rather than lasting gains in independent writing ability. Longer course-based studies and comparative designs are needed to examine whether AI functions mainly as short-term assistance or whether it can contribute to sustained writing development over time.

Outcome measurement was also inconsistent across the included studies. Some studies

counted language errors, others used readability measures, questionnaires, interviews, or teacher evaluations. These approaches are useful in different ways, but they make comparison across studies difficult. They also do not fully capture the complexity of medical writing. Few studies used multidimensional rubrics that assessed genre appropriateness, discourse organisation, rhetorical effectiveness, and the accuracy of medical content together. Future research would be stronger if it combined textual analysis, learner reflection, teacher assessment, and genre-based evaluation.

Transparency in reporting AI use was another concern. Some studies named the tool used but did not provide enough information about the version, prompts, task design, type of AI output, or amount of human revision after AI assistance. This matters because AI systems change quickly, and different prompts or settings can produce different results. Future studies should report AI use more clearly, including the tool and version, the writing task, prompt design, the type of feedback or output generated, and the role of learners, teachers, or researchers in checking and revising the text.

The geographical and linguistic coverage of the evidence was also uneven. The reviewed studies included settings in Asia, the Middle East, and Europe, but there was limited evidence from Africa and Latin America. The review also included only English-language publications, which means that relevant studies published in other languages may have been missed. Publication bias is possible as well, since studies reporting positive effects of AI-assisted writing may be more likely to appear in the published literature. Future research should include a wider range of educational contexts, learner populations, and medical writing genres, including case reports, reflective writing, clinical records, and patient-facing communication.

Finally, the methodological quality appraisal used established tools such as JBI, CASP, and MMAT, but these instruments were not designed specifically for AI-assisted writing research. They may be less sensitive to issues that are especially important in this field, such as tool transparency, prompt reporting, disclosure practices, authorship boundaries, and rapid changes in AI systems. Future research should develop or adapt evaluation criteria that better reflect the distinctive features of AI-supported writing education. The current evidence base is useful, but it remains limited. More attention is needed to discourse-level writing quality, critical AI literacy, institutional guidance, authorship norms, and the long-term development of independent medical writing ability.

Conclusion

This systematic review brought together 24 sources on the use of AI tools in English medical writing education for non-native English speakers. The reviewed literature suggests that AI can support grammar correction, sentence-level revision, readability improvement, terminology assistance, summarisation, simplification, and LLM-based drafting or revision. These forms of support may be useful for learners and professionals who need to communicate medical knowledge in English but have limited access to formal writing instruction. The most consistent reported benefits were improvements in local language accuracy, readability, confidence, and audience adaptation, especially in tasks such as abstract writing, case-related writing, and plain-language summaries.

However, the evidence remains limited and should not be interpreted as proof that AI produces long-term improvement in independent medical writing ability. Many studies were small, short-term, and methodologically uneven. Important questions also remain unresolved. AI-

generated or AI-revised text may contain inaccurate medical information, and the boundary between acceptable writing support and inappropriate text generation is not always clear. Issues of authorship, disclosure, responsibility, academic integrity, and learner dependence also require careful attention. AI tools are useful only when they remain part of supervised writing practice rather than becoming the main sources of the text.

In practice, AI should be used with clear limits in English medical writing education. It can help with early drafting, language revision, readability adjustment, and audience adaptation, but students should still learn how to judge the quality of suggestions, check medical accuracy, and explain their own revision choices. Future studies need to move beyond short writing tasks and examine whether AI-supported practice actually improves independent writing over time. Stronger evidence is also needed on discourse-level quality, genre awareness, prompt reporting, disclosure, and the responsibilities of human authors.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Acknowledgments: Not applicable.
Conflicts of Interest: The authors declare that there are no competing interests to disclose.
Authenticity: The authors confirm that this manuscript is original, has not been published previously, and is not currently under consideration by any other journal. All sources used in the review have been properly cited and acknowledged.
Artificial Intelligence Statement: During the preparation of this manuscript, the authors used ChatGPT only for language editing, readability improvement, and proofreading support after the review had been designed and conducted. The tool was not used to design the review, search or select studies, appraise methodological quality, extract data, analyse findings, or draw conclusions. All AI-assisted language suggestions were checked, revised, and approved by the authors. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

About the Authors:
Wei Shen is a PhD candidate in the English Language Studies Programme at the Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Malaysia. Her research areas are corpus linguistics, second language acquisition, and translation. Email: P120860@siswa.ukm.edu.my ORCID: https://orcid.org/0009-0005-3791-5719

Dr. Afendi Hamat is an Associate Professor at the Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Malaysia. His research areas are e-learning and web-based CALL. He has developed various e-learning platforms such as the Second Language Instruction and Management System (SLIM) and the intelligent English Language Literacy System (iELLS). Email: fendi@ukm.edu.my ORCID: https://orcid.org/0000-0003-2060-9203

Dr. Anis Nadiah Che Abdul Rahman is a senior lecturer at the Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Malaysia. Her area of expertise is in corpus linguistics and semantics. Her research interests are in

political and social media discourse. Email: aniscar@ukm.edu.my ORCID: https://orcid.org/0000-0001-7530-1113

Jun Wang is an Associate Professor in the School of Foreign Studies at Yanshan University, China. Her research area is TEFL (Teaching English as a Foreign Language). Email: wywy@ysu.edu.cn ORCID: https://orcid.org/0009-0005-6246-7030

References

Aljuaid, H. (2024). The impact of artificial intelligence tools on academic writing instruction in higher education: A systematic review. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*, 26–55. https://doi.org/10.24093/awej/ChatGPT.2

Amano, T., Ramírez-Castañeda, V., Berdejo-Espinola, V., Borokini, I., Chowdhury, S., Golivets, M., González-Trujillo, J. D., Montaño-Centellas, F., Paudel, K., & White, R. L. (2023). The manifold costs of being a non-native English speaker in science. *PLOS Biology, 21*(7), e3002184. https://doi.org/10.1371/journal.pbio.3002184

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI manual for evidence synthesis*. JBI. https://doi.org/10.46658/JBIMES-24-01

Bond, M., Khosravi, H., de Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education, 21*, Article 4. https://doi.org/10.1186/s41239-023-00436-z

Committee on Publication Ethics. (2023, February 13). *Authorship and AI tools*. https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools

Critical Appraisal Skills Programme. (2024). *CASP checklist: For qualitative research*. https://casp-uk.net/casp-checklists/CASP-checklist-qualitative-2024.pdf

Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature, 613*(7944), 423. https://doi.org/10.1038/d41586-023-00056-7

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine, 6*, Article 75. https://doi.org/10.1038/s41746-023-00819-6

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Hyland, K. (2003). *Second language writing*. Cambridge University Press.

Hyland, K. (2016). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing, 31*, 58–69. https://doi.org/10.1016/j.jslw.2016.01.005

International Committee of Medical Journal Editors. (2025). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals* (April 2025 update). https://icmje.org/news-and-editorials/icmje-recommendations_annotated_apr25.pdf

Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition, 12*(1), 3–69. https://doi.org/10.1207/s15327817la1201_2

Jen, S. L., & Salam, A. R. (2024). Using artificial intelligence for essay writing. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*, 90–99. https://doi.org/10.24093/awej/ChatGPT.5

Kasai, H., Asahina, M., Tajima, H., Shikino, K., Shimizu, I., Onodera, M., Kimura, Y., Araki, N., Tsukamoto, T., Yoshida, S., Yamauchi, K., & Ito, S. (2024). Effectiveness of a report writing training program using peer review: Evidence from first-year medical students. *BMC Medical Education, 24*, Article 1132. https://doi.org/10.1186/s12909-024-06041-6

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher, 29*(9–10), 855–871. https://doi.org/10.1080/01421590701775453

O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology, 35*(1), 42–56. https://doi.org/10.14742/ajet.3795

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme*. Lancaster University. https://doi.org/10.13140/2.1.1018.4643

Schnell, B. (2024). Multilingual scholarly publishing: Exploring the perceptions, attitudes, and experiences of plurilingual scholars in foreign language publication. *The Journal of Electronic Publishing, 27*(1), 165–193. https://doi.org/10.3998/jep.5416

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature, 613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z

ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education, 39*(12), 1176–1177. https://doi.org/10.1111/j.1365-2929.2005.02341.x

Wald, H. S., Borkan, J. M., Taylor, J. S., Anthony, D., & Reis, S. P. (2012). Fostering and evaluating reflective capacity in medical education: Developing the REFLECT rubric for assessing reflective writing. *Academic Medicine, 87*(1), 41–50. https://doi.org/10.1097/ACM.0b013e31823b55fa

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yousofi, R. (2022). Grammarly deployment (in)efficacy within EFL academic writing classrooms: An attitudinal report from Afghanistan. *Cogent Education, 9*(1), Article 2142446. https://doi.org/10.1080/2331186X.2022.2142446

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0