IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LARGE LANGUAGE MODELS AND ACADEMIC WRITING IN HIGHER EDUCATION: A NARRATIVE REVIEW WITH A FOCUS ON CLAUDE AI

Dr K. Anand[1], A. Pradheepa[2], N. Polraj alias Goutham[3]

[1]Associate Professor, Department of English, AAA College of Engineering and Technology, Amathur – 626005, Tamil Nadu. dr.anandssk@gmail.com

[2]Assistant Professor, Department of English, AAA College of Engineering and Technology, Amathur – 626005, Tamil Nadu. drpradheepa.eng@gmail.com

[3]Research Scholar, Department of English, School of Liberal Arts and Special Education, Kalasalingam Academy of Research and Education, Krishnankovil – 626126, Tamil Nadu. polrajgoutham1@gmail.com

Abstract: This narrative review examines the intersection of large language models (LLMs) and academic writing in higher education, with a specific focus on Claude AI a model whose institutional adoption has expanded rapidly but whose representation in the peer-reviewed literature remains comparatively limited. Drawing on peer-reviewed sources published between 2020 and 2025 across databases including Scopus, Web of Science, ERIC, ScienceDirect, and Frontiers, this review synthesizes evidence on the benefits, risks, and cognitive implications of generative AI-assisted academic writing. The literature confirms that generative AI tools, led overwhelmingly by research on ChatGPT, improve measurable dimensions of student writing including cohesion, fluency, and revision behavior, while simultaneously introducing well-documented risks: over-reliance, academic integrity violations, and detector bias against non-native English writers. Evidence on cognitive and critical-thinking outcomes remains genuinely contested, with reviews reporting both enhancement and erosion of higher-order thinking depending on task design and instructional scaffolding. Against this backdrop, the review identifies a central and consequential gap: despite Claude's formal deployment across universities including Northeastern University, the London School of Economics and Political Science, and Champlain College, and despite substantial documented student and educator usage, only a small number of peer-reviewed studies examine Claude specifically within academic writing contexts. This asymmetry between institutional deployment and independent scholarly scrutiny means that current pedagogical guidance on "AI-assisted writing" rests on evidence base derived almost entirely from a different model family. The review concludes by outlining research gaps and future directions specific to Claude AI, particularly within English for Academic Purposes and English for Specific Purposes contexts, and offers implications for educators, curriculum designers, and institutional policymakers navigating this evolving and unevenly evidenced landscape.

Index Terms: Large Language Models, Claude AI, Academic Writing, Higher Education, Generative Artificial Intelligence, Academic Integrity, English for Academic Purposes

I. INTRODUCTION

The emergence of generative artificial intelligence (GenAI), and particularly large language models (LLMs) such as OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, has fundamentally reconfigured the landscape of academic writing in higher education. Since ChatGPT's public release in November 2022, its adoption has been extraordinarily rapid: ChatGPT reached over 100 million users within two months and became widely embedded in academic practices such as writing support, brainstorming, formative feedback, and inquiry-based learning. Within a comparatively short span, LLMs have moved from novelty to near-ubiquity in university settings, with one early study finding that 89% of students had used ChatGPT for schoolwork in some capacity by January 2023. This swift, largely unplanned diffusion of AI writing tools into classrooms, laboratories, and research offices has outpaced the institutional, pedagogical, and ethical frameworks needed to govern their use, creating what several scholars have described as a state of reactive rather than deliberate integration (Perkins, 2023, as cited in Ferrero et al., 2024).

Academic writing occupies a uniquely consequential position within this transformation. Unlike many everyday applications of AI, academic writing is simultaneously a communicative skill, an epistemic practice, and a gatekeeping mechanism through which students demonstrate disciplinary competence, critical reasoning, and scholarly identity. It is precisely because academic writing carries this evaluative and formative weight that the entry of LLMs into the writing process has generated such intense scholarly interest and disquiet. A growing body of systematic reviews confirms that GenAI tools can enhance multiple dimensions of student writing. A 2025 systematic review employing PRISMA procedures across 30 studies published between 2023 and 2024 found that generative AI plays a significant role in enhancing the cohesion, precision, clarity, and engagement of academic writing, offering improved creativity, fluency, and proficiency. Complementary experimental research on EFL and EAP learners reinforces this picture: AI-generated feedback has been shown to improve linguistic accuracy and writing performance across various dimensions when compared with, or combined with, teacher-generated feedback, while learners interacting with ChatGPT-based writing protocols report measurable gains in coherence, cohesion, and access to immediate, grammatically accurate revision suggestions.

Yet this optimism is consistently counterbalanced by concern. The same 2025 systematic review that documented improvements in fluency and clarity also identified risks of plagiarism, over-reliance on AI-generated text, diminished creativity, ethical misconduct, and limitations in interpreting nuanced information, compounded by issues such as hallucination, bias, cultural insensitivity, unequal access to AI technologies, and inadequate AI literacy. These concerns are not merely theoretical. Survey-based evidence from the International Center for Academic Integrity indicates that 58% of students admitted to using AI tools to complete assignments dishonestly, a figure that has prompted widespread institutional anxiety about the erosion of authentic authorship. Some scholars have gone so far as to coin the term "AI-plagiarism" to capture the way generative AI writing assistance makes plagiarism more accessible and less detectable than earlier forms of academic dishonesty. At the same time, the presumption that AI use is inherently dishonest is itself contested: research on student perceptions suggests that many students do not perceive the use of AI chatbots as cheating or as a breach of academic integrity, a divergence that has left higher education institutions grappling with the twin challenges of formulating coherent policy and redesigning assessment to remain resilient in an AI-saturated environment.

Beyond questions of integrity, a second major strand of inquiry concerns the cognitive consequences of sustained LLM use. Evidence here is genuinely mixed rather than uniformly alarmist or celebratory. Some studies report that interaction with ChatGPT can improve reasoning, argumentation, and idea generation, while others document diminished creativity, superficial analytical depth, and cognitive offloading arising from excessive AI reliance. A large-scale systematic review of 39 studies conducted between 2023 and 2024 similarly found that a majority of reviewed papers reported measurable academic gains, even as other researchers expressed concern about the potential negative impact of ChatGPT on the development of critical thinking, problem-solving, reasoning, and creativity among undergraduate students. This tension between

augmentation and atrophy, between AI as cognitive scaffold and AI as cognitive substitute, runs through nearly every strand of the literature reviewed for this paper and constitutes one of the central unresolved debates that a narrative synthesis must confront.

Importantly, the great majority of this expanding evidence base has been generated around a single model family: ChatGPT. Existing systematic reviews, meta-analyses, and empirical studies overwhelmingly foreground OpenAI's model as the default object of investigation, whether examining its effects on learning, wellbeing, and collaboration, its role in academic dishonesty and assessment redesign, or its application in generative-chat feedback on academic writing. Anthropic's Claude, by contrast, remains comparatively underexamined in the peer-reviewed education literature, despite its rapid institutional uptake. Claude has been formally adopted through campus-wide licensing agreements at institutions such as Northeastern University, the London School of Economics and Political Science, and Champlain College, and is now integrated into learning-management ecosystems through partnerships with organizations such as Instructure and Internet2 (Anthropic, 2025a). Anthropic's own large-scale analysis of student usage patterns found that students primarily use Claude to create and improve educational content across disciplines, accounting for 39.3% of conversations, often involving essay editing, practice-question design, and summarization of academic material, while a comparable analysis of educator behaviour found that curriculum development represented 57% of educator conversations, followed by academic research at 13% and assessment of student performance at 7%. These industry-generated findings, while informative, are not peer-reviewed and cannot substitute for the kind of independent, critically evaluated scholarship that a narrative review is designed to synthesize; their prominence in the current discourse is itself symptomatic of the gap this review seeks to address.

Where Claude does appear in the peer-reviewed and preprint literature, it is typically as one comparator among several LLMs in benchmarking or clinical-education studies rather than as the primary focus of an academic-writing investigation. Comparative studies have found, for instance, that Claude 3.5 Sonnet outperformed ChatGPT and Gemini on clarity and comprehensiveness when translating patient-education material into Spanish and Urdu, and that Claude 3.5 Sonnet tied with ChatGPT-4o for the highest accuracy scores on MRI-related patient questions, while other domain-specific comparisons have found Claude trailing competitors on particular metrics, such as exhibiting a numerically, though not statistically significantly, higher misinformation rate than Gemini 2.0 in a celiac-disease patient-education comparison. Such findings, though drawn from clinical and health-communication contexts rather than higher-education writing pedagogy, are instructive: they demonstrate that model-specific performance differences are real, measurable, and pedagogically consequential, yet they have rarely been systematically synthesized with reference to academic writing instruction, feedback provision, or scholarly composition in university settings. This asymmetry between Claude's growing institutional footprint and its thin representation in the peer-reviewed academic-writing literature constitutes the central research gap motivating the present review.

A further gap concerns the fragmentation of the existing review literature itself. Prior systematic reviews have tended to specialize narrowly, examining GenAI's role in assessment transformation (Xia et al., 2024, as cited in search results), critical-thinking outcomes, academic dishonesty, or feedback provision as discrete phenomena, but rarely bringing these strands into sustained dialogue with one another, and rarely situating any single LLM's distinctive design philosophy, such as Anthropic's stated emphasis on Constitutional AI, safety-oriented guardrails, and a "Learning mode" intended to scaffold reasoning rather than supply answers (Anthropic, 2025), within that broader synthesis. The absence of a narrative review that draws together the cognitive, ethical, pedagogical, and comparative-technical strands of this literature, while explicitly foregrounding Claude as a case of particular interest, leaves both researchers and practitioners without a consolidated, critically evaluated account of what is currently known, contested, and unknown about LLM-assisted academic writing in higher education.

Against this backdrop, the present narrative review pursues three interlocking aims. First, it synthesizes and critically evaluates the peer-reviewed and grey literature published between 2020 and 2025 on the use of

LLMs in academic writing within higher education, with particular attention to methodological rigor, sample characteristics, and the consistency or contradiction of reported findings across studies. Second, it foregrounds Claude AI as a comparative case, drawing on the limited but growing body of studies that directly evaluate Claude's performance, design features, and institutional deployment, while candidly acknowledging where the literature is currently too sparse to support firm conclusions rather than substituting speculation for evidence. Third, it identifies the research gaps that should guide future empirical inquiry, particularly with respect to model-specific comparative research conducted directly within academic writing and English for Specific Purposes contexts, disciplines in which the present authors' own scholarly work is situated.

The remainder of this review proceeds as follows. Section 1.1 outlines the background and rationale for focusing on Claude AI within the broader LLM literature, followed by the review's objectives (Section 1.2) and guiding research questions (Section 1.3). Section II then details the review methodology, including the databases searched, inclusion and exclusion criteria, and the narrative synthesis approach adopted to organize and critically appraise the reviewed literature.

1.1 Background of the Study

The trajectory of large language models (LLMs) in higher education is best understood as a sequence of rapid technological escalation followed by lagging institutional response. OpenAI's release of ChatGPT in November 2022 marked the inflection point: the tool reached over 100 million users within two months and was rapidly embedded in academic practices such as writing support, brainstorming, formative feedback, and inquiry-based learning (Deng et al., 2024; Dimeli & Kostas, 2025; Hu, 2023; Tlili et al., 2023). Kamoun et al. (2023), surveying 855 university students and 145 faculty members, found widespread concern that ChatGPT would diminish students' intellectual development and academic integrity, evidence that anxiety about LLMs accompanied their adoption almost from the outset rather than emerging only after sustained use.

Subsequent scholarship has traced this adoption across two parallel and often contradictory currents. On one hand, systematic reviews report genuine pedagogical benefit. Chanpradit (2025), synthesizing 30 studies published between 2023 and 2024 using PRISMA procedures, found that GenAI tools enhance the cohesion, precision, clarity, and engagement of academic writing while improving creativity, fluency, and proficiency. Kumar et al. (2024), reviewed in Ridha et al. (2025), identified three recurring institutional themes: Academic Integrity, Pedagogical Techno-Innovation, and Experiential Engagement that structure most institutional responses to GenAI. On the other hand, the same literature documents persistent risk: Chanpradit (2025) also identified plagiarism, over-reliance, diminished creativity, hallucination, bias, and unequal access as recurring concerns, while Farrokhnia et al. (2024) raised specific concern about erosion of problem-solving, reasoning, and creativity among undergraduates. Academic integrity has become the most visible flashpoint in this background literature. Chan (2023, as cited in Healy, 2023) coined the term "AI-plagiarism" to describe how generative AI makes plagiarism more accessible and harder to detect than earlier forms of academic dishonesty. This concern is empirically substantiated: the International Centre for Academic Integrity (Tan & Maravilla, 2024) reported that 58% of surveyed students admitted using AI tools to complete assignments dishonestly. Yet student perception research complicates a purely punitive reading of this trend; students frequently do not perceive AI-chatbot use as cheating or as a breach of academic integrity (Ferreira & Kirova, 2024), which has pushed institutions toward assessment redesign rather than prohibition alone (Kumar et al., 2024). Within this broader LLM landscape, Anthropic's Claude represents a distinct institutional trajectory. Rather than emerging primarily through consumer virality, Claude's presence in higher education has developed through formal, university-wide licensing partnerships including full-campus agreements with Northeastern University, the London School of Economics and Political Science, and Champlain College alongside integration partnerships with Internet2 and the Canvas LMS provider Instructure (Anthropic, 2025b). Anthropic's own large-scale analysis of approximately one million anonymized student conversations found that students primarily use Claude to create and improve educational content (39.3% of conversations), including essay editing and summarization of academic material, while a parallel analysis of roughly 74,000 educator conversations found curriculum development to be the dominant use case (57%), followed by

academic research (13%) (Anthropic, 2025a, 2025c). These figures, while informative, originate from company-published reports rather than independently peer-reviewed research a limitation this review treats as part of the background context rather than as equivalent evidence to the peer-reviewed literature synthesized elsewhere in this paper. It is against this background that rapid, uneven LLM adoption; a maturing but ChatGPT-centric evidence base; unresolved debate over cognitive and integrity effects; and a comparatively underexamined but institutionally significant role for Claude are that the present narrative review is situated.

1.2 Objectives of the Study

Building directly on the research gaps identified in the Introduction and Background namely, the disproportionate concentration of empirical and review literature on ChatGPT, the unresolved debate over the cognitive and integrity effects of LLM-assisted writing, and the comparatively thin peer-reviewed evidence base specific to Claude AI, this narrative review pursues the following objectives:

Objective 1: To synthesize the existing peer-reviewed literature (2020–2025) on the use of large language models in academic writing within higher education. This includes mapping the reported benefits (e.g., improvements in cohesion, clarity, fluency, and revision quality) and reported risks (e.g., plagiarism, over-reliance, diminished creativity, hallucination) documented across systematic reviews and empirical studies (Chanpradit, 2025; Farrokhnia et al., 2024).

Objective 2: To critically evaluate the methodological characteristics of the reviewed literature, including research design, sample composition, disciplinary context, and geographic distribution, to assess the strength, consistency, and generalizability of current claims about LLM-assisted academic writing.

Objective 3: To examine the specific role of Claude AI within the broader LLM-in-education literature, drawing on the limited peer-reviewed and comparative-benchmarking studies available, while explicitly distinguishing independently verified academic findings from Anthropic's own company-published usage reports (Anthropic)

Objective 4: To conduct a comparative analysis of major LLMs (ChatGPT, Gemini, Claude, and others where evidence permits) as they relate to academic writing support, feedback provision, and writing-related task performance, based strictly on findings reported in verifiable published studies.

Objective 5: To identify and articulate research gaps in the current literature, particularly the scarcity of studies examining Claude within English for Academic Purposes (EAP), English for Specific Purposes (ESP), and second-language academic writing contexts to guide future empirical research.

Objective 6: To derive evidence-based educational and practical implications for instructors, curriculum designers, and institutional policymakers regarding the responsible integration of LLMs, including Claude, into academic writing instruction and assessment design. These objectives collectively position the review not as a general survey of GenAI in education, but as a focused, critically synthesized account of what is currently known and not yet known about LLMs and academic writing, with Claude AI serving as the review's distinguishing analytical lens.

1.3 Research Questions

In line with the six objectives outlined above, this narrative review is guided by the following research questions:

RQ1: What benefits and risks of large language model (LLM)-assisted academic writing have been reported in the peer-reviewed literature published between 2020 and 2025?

RQ2: What methodological characteristics (research design, sample composition, disciplinary and geographic context) define the existing body of research on LLMs and academic writing, and what do these characteristics reveal about the strength and generalizability of current findings?

RQ3: To what extent, and in what contexts, has Claude AI been examined in the peer-reviewed and comparative-benchmarking literature relating to academic writing and higher education, and how does this evidence base compare in depth and scope to that of ChatGPT?

RQ4: How do major LLMs including ChatGPT, Gemini, and Claude compare in terms of reported performance on writing-related tasks, feedback provision, and academic-writing support, based on findings from verifiable published studies?

RQ5: What research gaps exist in the current literature regarding LLMs and academic writing, particularly with respect to Claude AI's use in English for Academic Purposes (EAP), English for Specific Purposes (ESP), and second-language writing contexts?

RQ6: What educational and practical implications can be drawn from the existing evidence base to guide the responsible integration of LLMs, and Claude AI specifically, into academic writing instruction, feedback practices, and institutional policy?

II. REVIEW METHODOLOGY

2.1 Review Design

This paper adopts a **narrative review methodology**, rather than a systematic review or meta-analysis. This design choice reflects the review's purpose: to critically synthesize, compare, and interpret a conceptually and methodologically diverse literature spanning applied linguistics, educational technology, computer science, and health-communication research, rather than to produce a quantitatively pooled effect estimate from a narrowly homogeneous set of studies. Unlike PRISMA-guided systematic reviews, a narrative review permits the integration of theoretical, empirical, and comparative-benchmarking sources, and allows for critical commentary on contested or unevenly distributed evidence a necessity given that Claude-specific peer-reviewed literature remains comparatively sparse relative to ChatGPT-focused scholarship, as established in the Introduction. Where the review draws on non-peer-reviewed but relevant material (e.g., Anthropic's own published usage reports), this is explicitly flagged as grey literature rather than treated as equivalent to peer-reviewed evidence.

2.2 Databases Searched

Consistent with the Master Instruction governing this project, literature was identified through the following databases and publisher platforms:

- Scopus
- Web of Science
- ERIC
- Frontiers
- Science Direct

2.3 Search Terms

Search strings combined the following term clusters using Boolean operators (AND/OR):

- **LLM/GenAI terms:** "large language models," "generative artificial intelligence," "ChatGPT," "Claude AI," "Anthropic," "Gemini," "chatbot"
- **Academic writing terms:** "academic writing," "English for Academic Purposes," "English for Specific Purposes," "second language writing," "writing feedback," "writing revision"
- **Context terms:** "higher education," "university students," "undergraduate," "postgraduate"
- **Issue-specific terms:** "academic integrity," "plagiarism," "critical thinking," "AI literacy," "cognitive offloading"

2.4 Inclusion and Exclusion Criteria

2.4.1 Inclusion Criteria

- Published between January 2020 and 2025 (with limited earlier foundational sources cited only where directly relevant to definitional or theoretical framing)
- Peer-reviewed journal articles, systematic reviews, meta-analyses, or conference proceedings from recognized academic databases
- Directly addressing LLMs/GenAI in relation to academic writing, higher education, or closely related pedagogical outcomes (critical thinking, feedback, integrity)
- Published in English

2.4.2 Exclusion Criteria

- Studies focused exclusively on K–12 or primary/secondary education with no higher-education relevance
- Purely technical/computer-science papers on model architecture with no educational application
- Opinion pieces, blog posts, or non-peer-reviewed commentary, **except** where explicitly used and flagged as contextual/grey literature (e.g., Anthropic's education reports, used only to characterize institutional adoption, not as evidence of pedagogical efficacy)
- Duplicate studies or preprints later superseded by a published peer-reviewed version

2.5 Limitations of the Review Methodology

As a narrative rather than systematic review, this study does not report a PRISMA flow diagram, inter-rater screening reliability statistics, or a formal risk-of-bias appraisal tool, and its source selection, while systematic in search strategy, retains an interpretive element characteristic of narrative synthesis. This limitation is acknowledged transparently rather than obscured, in keeping with this review's broader commitment to methodological honesty over the appearance of methodological rigor it does not possess.

III. LITERATURE REVIEW

3.1 Benefits of LLM-Assisted Academic Writing

The most consistent finding across the reviewed literature is that generative AI tools measurably improve surface- and discourse-level features of student writing. Chanpradit's (2025) PRISMA-guided systematic review of 30 studies published between 2023 and 2024 found that GenAI enhances the cohesion, precision, clarity, and engagement of academic writing, while also improving creativity, fluency, and proficiency. This finding is corroborated at the level of individual empirical studies. Boudouaia et al. (2024), in a quasi-experimental study of 76 Algerian undergraduates, found that a ChatGPT-4-assisted experimental group outperformed a control group on EFL writing outcomes, while Song and Song (2023) reported that ChatGPT's capacity to generate immediate, alternative, grammatically correct phrasing supported the production of more coherent and cohesive text among EFL learners. Similarly, a mixed-methods study of postgraduate EAP writers in Egypt found positive correlations between generative-AI-assisted feedback tool use and increased revision frequency, alongside improvements in content, organization, and cohesion (Discover Education, 2025).

Feedback-specific research reinforces these gains while adding nuance regarding the *mechanism* of benefit. A systematic review of 12 studies on AI-generated feedback via generative chat found that instant feedback allowed students to dedicate more time to revision, while also enhancing self-efficacy, self-regulation, and motivation, and reducing anxiety around the writing process. Comparative work examining sequences of AI- and teacher-generated feedback found that ChatGPT-based instruction protocols improved not only writing performance but also teachers' self-efficacy in a ten-week intervention (Ghafouri et al., 2024). Collectively,

this evidence suggests that LLM-assisted feedback functions less as a substitute for instructional judgment and more as a scaffold that increases the *volume and immediacy* of revision opportunities available to learners a mechanism distinct from, but complementary to, human feedback.

3.2 Risks, Academic Integrity, and Ethical Concerns

Set against these benefits, the literature documents a substantial and consistent set of risks. The same systematic review that identified gains in cohesion and fluency also identified risks of plagiarism, over-reliance on AI-generated text, diminished creativity, ethical misconduct, and limitations in interpreting nuanced information, compounded further by hallucination, bias, cultural insensitivity, unequal access, and inadequate AI literacy (Chanpradit, 2025). These concerns are not abstract: survey evidence from the International Center for Academic Integrity found that 58% of students admitted to using AI tools to complete assignments dishonestly (as reported in Tan & Maravilla, 2024), while Kamoun et al.'s (2023) survey of 855 students and 145 faculty found widespread concern that ChatGPT would diminish students' intellectual development and academic integrity.

A particularly instructive strand of this literature concerns the definitional instability of "AI-assisted" versus "AI-generated" work. Chan (2023, as cited in Healy, 2023) coined the term "AI-giarism" to describe how GenAI writing assistance makes plagiarism more accessible and harder to detect than earlier forms of academic dishonesty, a concern sharpened by evidence that automated AI-text detectors are demonstrably biased against non-native English writers (Liang et al., 2023, as cited in EDUCAUSE Review, 2023) a finding with direct and troubling implications for the very EFL/EAP student populations who, according to the feedback literature reviewed above, stand to benefit most from AI-assisted writing support. This tension between AI's pedagogical promise for multilingual writers and its integrity-detection risk for the same population is rarely reconciled within individual studies and represents a genuine unresolved contradiction in the field.

Importantly, student perception research complicates any straightforwardly punitive framing of this risk. Evidence reviewed in a 2024 ScienceDirect study on students' perceptions of chatbot use for assessment found that students predominantly do not perceive AI chatbot use as cheating or as a breach of academic integrity a divergence from institutional and faculty perspectives that has pushed scholars such as Kumar et al. (2024) to argue for the redesign of assessment toward more authentic, real-world tasks rather than reliance on prohibition and detection alone. Institutional responses reviewed in Open Praxis similarly show a split between calls for stronger detection and watermarking technologies (Lancaster, 2023, as cited in Open Praxis systematic review, 2025) and calls for constructivist integration of GenAI accompanied by clear institutional governance (Blake, 2024; Lim et al., 2023, as cited in the same review).

3.3 Cognitive and Critical-Thinking Effects

A third major theme concerns the effect of sustained LLM use on higher-order thinking. This literature is characterized by genuine empirical disagreement rather than consensus. On one side, several studies report improvements in reasoning, argumentation, and idea generation associated with ChatGPT use (Daniel et al., 2025; Dimeli & Kostas, 2025; Farazouli et al., 2023, as cited in a 2026 ScienceDirect systematic review). On the other, a comparable body of evidence documents diminished creativity, superficial analytical depth, and cognitive offloading arising from excessive reliance on AI-generated output (Bai et al., 2023; Roy & Putatunda, 2023; Vargas-Murillo et al., 2023, as cited in the same review). A large-scale systematic review of 39 studies published between 2023 and 2024 found that a majority (56.4%) of reviewed papers reported measurable academic gains, particularly in engineering, programming, and data-analytics contexts, yet the same review noted persistent concern regarding ChatGPT's potential negative impact on problem-solving, reasoning, and creativity among undergraduates (Farrokhnia et al., 2024, as cited therein).

This contradiction is partly explained, though not resolved, by methodological heterogeneity across studies: reviews synthesizing this literature note that existing systematic reviews have examined ChatGPT's

educational role through varied thematic lenses Zirar (2023) through learning and assessment, Deng et al. (2024) through meta-analytic academic performance outcomes, Lo et al. (2024) through behavioral, emotional, and cognitive engagement, and Melisa et al. (2025) through critical thinking specifically such that direct comparison across studies is complicated by differing operationalizations of "critical thinking" and differing measurement instruments. A separate study examining student-ChatGPT interaction directly found that some students accept inaccurate answers and copy-paste without critical evaluation, while a contrasting study (Kosar et al., 2024) found no significant difference in academic performance between ChatGPT users and non-users underscoring that the moderating variables (user attitudes, trust in AI, disciplinary context) remain insufficiently understood.

3.4 AI Literacy and Institutional Preparedness

A fourth theme, closely tied to both the integrity and cognitive-effects literatures, concerns the emerging construct of "AI literacy." Multiple competing frameworks have been proposed in rapid succession: a twelve-competency model (Annapureddy et al., 2024/2025), a four-dimension model comprising knowledge, application, evaluation, and ethics (O'Dea et al., 2024), and Ng et al.'s (2021) foundational four-component framework of knowing/understanding, using/applying, evaluating/creating, and ethics, which continues to be adapted across disciplinary contexts (Zhou & Schofield, 2024). An integrative review of 124 studies published between 2020 and 2024 found that the proliferation of AI-literacy scholarship accelerated sharply after 2023, driven specifically by a shift toward viewing AI as a practical tool rather than a purely technical or sociocultural object of study.

Despite this proliferation of frameworks, reviewers converge on a common critical observation: theoretical research on AI literacy lags technological advancement, leaving a persistent and unclear path forward for competency cultivation (Zhang, 2025, as cited in a 2026 Frontiers mixed-methods study). This gap has direct implications for academic writing specifically, since AI-literacy interventions have been developed and evaluated primarily in disciplinary contexts such as nursing (Song et al., 2025), music education (Zhang, 2025), and general higher-education competency scales (e.g., the FALCON-AI faculty scale).

3.5 Claude AI in the Academic Writing Literature

The preceding four themes are drawn almost entirely from ChatGPT-centric or model-agnostic GenAI research. When the search is narrowed specifically to Claude AI, the peer-reviewed evidence base relevant to academic writing becomes markedly thinner, confirming the research gap established in the Introduction. Only a small number of studies were identified that examine Claude directly within an academic-writing or higher-education-writing context.

Amelia (2024), publishing in the *Journal of English as a Foreign Language Education*, conducted a qualitative case study of two English Education students at Tanjungpura University using observation and interview data analyzed through Miles and Huberman's qualitative approach. The study found that participants perceived Claude as a useful writing assistant for improving academic writing skills, though the sample size ($n = 2$) sharply limits generalizability and the study did not employ pre/post writing-quality measures, relying instead on self-reported perception. This represents, to date, one of the only dedicated peer-reviewed investigations of Claude as an academic-writing tool within an English-language-teaching context, precisely the gap this review is designed to address.

A second, more recent study published in *Frontiers in Artificial Intelligence* examined the integration of DeepL Write and Claude AI to enhance argumentative writing competence among engineering students in India, situating its analysis within the Toulmin model of argumentation. While this study extends the evidence base to a non-humanities disciplinary context and to argumentative genre specifically, Claude appears here in combination with another tool rather than as a sole independent variable, making it difficult to isolate Claude-specific effects from the broader AI-assisted intervention. The same study's literature review notes that related

work applying the Toulmin model to AI-assisted writing has predominantly used ChatGPT (e.g., Fadilah, 2026; Al Fraidan, 2025), further illustrating the model asymmetry that characterizes this field.

Beyond these two writing-specific studies, Claude appears in the broader literature chiefly as a comparator in benchmarking research conducted outside education-specific writing contexts. In health-communication research, Claude 3.5 Sonnet outperformed ChatGPT and Gemini on clarity and comprehensiveness when translating patient-education materials into Spanish and Urdu, and tied with ChatGPT-4o for the highest accuracy on MRI-related patient questions, while a separate study found Claude 3.7 exhibited a numerically (though not statistically significantly) higher misinformation rate than Gemini 2.0 in a celiac-disease patient-education comparison. In programming-education contexts, a large-scale comparison of eighteen LLMs across 6,000 student programming submissions found systematic differences in grading leniency between model families, with Claude-Haiku-3.5 and GPT-4o exhibiting more lenient grading patterns than models such as DeepSeek-Reasoner. While informative regarding Claude's general performance characteristics, none of these studies examine academic-writing instruction, feedback, or composition pedagogy directly, meaning their relevance to this review's central focus is necessarily indirect and comparative rather than substantive.

Finally, Anthropic's own published usage-analytics reports, while explicitly non-peer-reviewed and therefore treated as contextual rather than evidentiary in this review, provide the only large-scale empirical picture of how Claude is used in higher education. The company's analysis of approximately one million anonymized student conversations found that essay editing and summarization of academic material fell within the largest usage category (39.3% of conversations, "create and improve educational content"), while a separate analysis of roughly 74,000 educator conversations found curriculum development to be the dominant use case (57%). These figures indicate substantial real-world academic-writing-adjacent usage of Claude that is, as yet, almost entirely unexamined by independent peer-reviewed scholarship, the central asymmetry this narrative review set out to document.

IV. RESULTS AND DISCUSSION

4.1 Research Gaps

Drawing on the literature review above, the following gaps emerge as priorities for future inquiry:

4.1.1 Gap 1: Claude-specific academic writing evidence is critically thin.

Only one dedicated peer-reviewed study (Amelia, 2024) examines Claude as an academic writing assistant directly, and it relies on a two-participant qualitative case study with no pre/post writing-quality measurement. A second study (Frontiers, 2026) situates Claude within a multi-tool intervention (alongside DeepL Write), making Claude-specific effects inseparable from the broader intervention. No quantitative, adequately powered study isolating Claude's effect on academic writing quality currently exists in the peer-reviewed literature.

4.1.2 Gap 2: Absence of EAP/ESP-specific Claude research.

The broader ChatGPT literature contains a substantial body of EFL/EAP-focused writing research (e.g., Boudouaia et al., 2024; Ghafouri et al., 2024; Koltovskaia et al., 2024). No comparable body of research exists for Claude within English for Academic Purposes or English for Specific Purposes contexts, despite these being precisely the contexts in which model-specific differences in tone, hedging, and register-sensitivity would be most pedagogically consequential.

4.1.3 Gap 3: Reliance on non-peer-reviewed vendor data.

The most detailed picture of how Claude is used in academic writing contexts comes from Anthropic's own usage-analytics reports (Anthropic, 2025a, 2025c) rather than independent peer-reviewed research. This creates a structural evidentiary asymmetry: the entity with a commercial interest in Claude's adoption is also the primary source of empirical claims about its educational use.

4.1.4 Gap 4: Detector bias intersecting with multilingual writers remains unexamined for Claude.

Evidence that AI-text detectors are biased against non-native English writers (Liang et al., 2023) has not been extended to test whether Claude-generated or Claude-assisted text carries different detection risk profiles than ChatGPT-generated text a gap with direct equity implications.

4.1.5 Gap 5: Fragmented operationalization of "critical thinking" across reviews.

Existing reviews (Zirar, 2023; Deng et al., 2024; Lo et al., 2024; Melisa et al., 2025) each measure cognitive effects differently, making cross-study synthesis unreliable. No study has applied a consistent critical-thinking framework specifically to Claude-assisted writing tasks.

4.1.6 Gap 6: Design-philosophy-to-outcome research is absent.

Anthropic markets Claude's "Learning mode" as guiding reasoning rather than supplying answers (Anthropic, 2025b), a design claim with direct pedagogical relevance to academic writing instruction. No independent study has empirically tested whether this design feature produces different writing-development outcomes compared to standard chatbot interaction.

4.1.7 Gap 7: AI-literacy frameworks are not tailored to writing pedagogy.

Current AI-literacy frameworks (Annapureddy et al., 2025; Ng et al., 2021, as adapted by Zhou & Schofield, 2024) have been operationalized in nursing, music, and general competency contexts, but not translated into a framework specific to academic writing instruction or model-comparative writing pedagogy.

4.2 Discussion

The literature synthesized in this review points to a field that is empirically active but asymmetrically distributed. Returning to the research questions posed in Section 1.3: substantial and consistent evidence answers RQ1, confirming that GenAI tools improve cohesion, fluency, and revision behavior in academic writing (Chanpradit, 2025; Boudouaia et al., 2024) while simultaneously introducing risks of over-reliance, plagiarism, and detector bias (Chanpradit, 2025; Liang et al., 2023). RQ2, concerning methodological quality, reveals a literature dominated by small-sample qualitative studies, quasi-experimental designs of modest size, and systematic reviews whose thematic categorizations vary enough to complicate direct comparison a pattern consistent with an emerging rather than mature field of inquiry.

RQ3 and RQ4, concerning Claude specifically, receive markedly thinner answers. The evidence base is not merely smaller than that for ChatGPT; it is qualitatively different in kind, consisting of one small case study, one multi-tool intervention, several cross-domain benchmarking studies unrelated to writing pedagogy, and company-published usage statistics. This is not a matter of Claude performing worse or better than competitor models the benchmarking literature reviewed suggests Claude performs competitively, and in some domains favorably, against ChatGPT and Gemini (e.g., in patient-education translation and MRI-related accuracy) but rather that academic writing researchers have simply not yet turned their attention to Claude with the same intensity applied to ChatGPT. This is arguably a function of ChatGPT's earlier and more visible market entry rather than any principled judgment about Claude's pedagogical relevance, and it means current institutional and pedagogical guidance on "AI-assisted writing" is, in practice, guidance derived almost entirely from one model family and generalized, often implicitly, to all others.

This generalization is not obviously safe. Model-specific differences documented in adjacent domains differing hallucination rates, differing register and tone, differing grading leniency (as in the 18-model programming-assessment comparison) suggest that writing-support quality, feedback style, and even the *kind* of over-reliance risk a student faces may vary meaningfully by model. A university writing center or EAP instructor advising students on "AI-assisted drafting" using evidence generated exclusively from ChatGPT studies is, in effect, extrapolating across an unverified assumption of model interchangeability.

The tension between AI-literacy proliferation and AI-literacy applicability is a second discussion point worth foregrounding. Despite a rapid multiplication of AI-literacy frameworks since 2023, reviewers

themselves note that this theorizing has outpaced practical translation into classroom-usable form, particularly within English-language writing instruction. Combined with the finding that many students do not perceive AI-assisted writing as a form of academic dishonesty, this suggests institutions are currently managing student AI use through detection and policy enforcement rather than through the AI-literacy education that the framework literature recommends as the more sustainable long-term response.

4.3 Future Research Directions

Adequately powered, quantitative studies of Claude-assisted academic writing, using pre/post writing-quality measures (e.g., holistic rubrics, discourse-analytic coding) rather than perception-only data, to move beyond the current single small-sample case study (Amelia, 2024).

Model-comparative studies designed specifically for writing pedagogy, head-to-head comparisons of ChatGPT, Gemini, and Claude on identical writing tasks, evaluated by independent raters, rather than comparisons imported from health-communication or programming-assessment domains.

EAP/ESP-specific investigations of Claude, examining whether Claude's tone, hedging conventions, and register-sensitivity affect multilingual and non-native English writers differently than ChatGPT-based tools, extending the existing ChatGPT-focused EFL/EAP literature (Boudouaia et al., 2024; Ghafouri et al., 2024) to a comparative design.

Independent (non-vendor) replication of usage-pattern findings. Anthropic's own reports (2025a, 2025c) should be treated as hypotheses to test independently for instance, verifying through classroom-based or institutional data whether essay editing and summarization genuinely dominate student use of Claude in specific disciplinary contexts.

Empirical testing of "Learning mode" and Constitutional-AI design claims against actual writing-development outcomes, rather than accepting vendor design descriptions as pedagogically established fact.

Detector-bias studies specific to Claude-assisted text, extending Liang et al.'s (2023) finding on detector bias against non-native writers to test whether this bias pattern holds, worsens, or attenuates with Claude-generated or Claude-revised text.

Longitudinal designs tracking writing-skill development over a semester or academic year, rather than the single-session or short-intervention designs that currently dominate the field, to determine whether short-term fluency gains persist or represent temporary scaffolding effects.

V. CONCLUSION

This narrative review set out to synthesize the peer-reviewed literature on large language models and academic writing in higher education, with particular attention to Claude AI's place within that literature. The evidence confirms that GenAI tools, as a category, offer measurable benefits for academic writing improved cohesion, fluency, and revision behavior while simultaneously introducing well-documented risks around academic integrity, over-reliance, and inequitable detection outcomes for non-native English writers. These findings, however, are drawn almost entirely from research on ChatGPT.

The central finding of this review is not a claim about Claude's pedagogical merit or deficiency, but an evidentiary one: Claude AI, despite substantial and growing institutional adoption documented in Anthropic's own usage reports and formal university partnerships, remains represented in the independent peer-reviewed academic-writing literature by essentially a single small-sample qualitative study and one multi-tool intervention. This asymmetry between real-world institutional deployment and independent scholarly scrutiny constitutes a meaningful blind spot in current understanding of AI-assisted academic writing, and it is one that cannot be resolved by extrapolating findings generated on a different model family. Until Claude-specific, adequately powered, writing-focused research accumulates, institutional policy, pedagogical guidance, and AI-literacy instruction referencing "AI-assisted writing" should be understood as resting on a ChatGPT-

derived evidence base whose generalizability to other LLMs, including Claude, remains empirically untested rather than confirmed.

REFERENCES

1. Amelia. (2024). Analysis of Claude AI as a writing assistant for English education students at Tanjungpura University. *Journal of English as a Foreign Language Education.* https://jurnal.untan.ac.id/index.php/JEFLE/article/view/82943
2. Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2025). Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice, 6.* https://doi.org/10.1145/3685680
3. Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Innovations in Education and Teaching International.* https://doi.org/10.1177/07356331241247465
4. Chan, C. (2023). Is AI changing the rules of academic misconduct? An in-depth look at students' perceptions of 'AI-giarism.' *arXiv preprint.* https://arxiv.org/abs/2306.03358
5. Chanpradit, T. (2025). Generative artificial intelligence in academic writing in higher education: A systematic review. *Edelweiss Applied Science and Technology, 9*(4), 889–906. https://doi.org/10.55214/25768484.v9i4.6128
6. Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 213,* Article 105224. https://doi.org/10.1016/j.compedu.2024.105224
7. Dimeli, M., & Kostas, A. (2025). The role of ChatGPT in education: Applications, challenges. *Journal of Information Technology Education: Research, 24,* Article 2.
8. Divya, T. J., & Alamelu, C. (2026). Integrating DeepL Write and Claude AI to enhance argumentative writing competence of engineering students in India. *Frontiers in Artificial Intelligence, 9,* Article 1838463. https://doi.org/10.3389/frai.2026.1838463
9. Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846
10. Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688241239764
11. Kamoun, F., El Ayeb, W., Jabri, I., Sifi, S., & Iqbal, F. (2024). Exploring students' and faculty's knowledge, attitudes, and perceptions towards ChatGPT: A cross-sectional empirical study. *Journal of Information Technology Education: Research, 23,* Article 4, 1–33. https://doi.org/10.28945/5239
12. Kosar, T., Ostojić, D., Liu, Y. D., & Mernik, M. (2024). Computer science education in ChatGPT era: Experiences from an experiment in a programming course for novice programmers. *Mathematics, 12*(5), Article 629. https://doi.org/10.3390/math12050629
13. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

14. Lo, C. K., Hew, K. F., & Jong, M. S. Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 219,* Article 105100. https://doi.org/10.1016/j.compedu.2024.105100

15. Melisa, R., Ashadi, A., Triastuti, A., Hidayati, S., Salido, A., Ero, P. E. L., Marlini, C., Zefrin, & Al Fuad, Z. (2025). Critical thinking in the age of AI: A systematic review of AI's effects on higher education. *Educational Process: International Journal, 14,* Article e2025031. https://doi.org/10.22521/edupij.2025.14.31

16. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2,* Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

17. Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2), Article 7. https://doi.org/10.53761/1.20.02.07

18. Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14,* Article 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

19. Tan, M. J. T., & Maravilla, N. M. A. T. (2024). Shaping integrity: Why generative artificial intelligence does not have to undermine education. *Frontiers in Artificial Intelligence, 7,* Article 1471224. https://doi.org/10.3389/frai.2024.1471224

20. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10,* Article 15. https://doi.org/10.1186/s40561-023-00237-x

21. Vargas-Murillo, A. R., de la Asuncion, I. N. M., & de Jesús Guevara-Soto, F. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *International Journal of Learning, Teaching and Educational Research, 22*(7), 122–135.

22. Xia, Q., Weng, X., Fan, O., Lin, T.-J., & Chiu, T. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21,* 1–22. https://doi.org/10.1186/s41239-024-00468-z

23. Zirar, A. (2023). Exploring the impact of language models, such as ChatGPT, on student learning and assessment. *Review of Education, 11*(3), Article e3433. https://doi.org/10.1002/rev3.3433