CJBEHM
CANADIAN JOURNAL OF BUSINESS, ECONOMICS AND HEALTH MANAGEMENT

ORIGINAL ARTICLE
CJBEHM. 2026 August; 8(2): 08-20
Published by www.cjbehm.ca
DOI: https://doi.org/10.62510/cjbehm.2026.v2i2.p2

Impact of Generative Artificial Intelligence Use on the Academic Performance of Undergraduate Students in Canada

Pragalvha Sharma[1]

1 Ivey Business School, University of Western Ontario, Canada

Corresponding author*

Pragalvha Sharma
Ivey Business School,
University of Western Ontario

Email:
pragalvhasharma@gmail.com

Article info
Received: 19 May 2026
Accepted: 29 July 2026
Published: 31 August 2026

Keywords

generative AI, ChatGPT, academic performance, undergraduate students, higher education, Canada

Abstract

Generative artificial intelligence (AI) has entered everyday undergraduate study faster than the evidence base has kept up with it, and the evidence that does exist points in opposite directions. This paper synthesizes peer-reviewed and carefully delimited contextual research on how generative AI use relates to undergraduate academic performance, with attention to what Canadian universities can already act on. Searches of Google Scholar, Scopus, Web of Science, ERIC, and ScienceDirect covering 2022 to 2026 produced a two-tier evidence base: Tier 1 peer-reviewed empirical and synthetic studies of student learning outcomes, and Tier 2 supplementary sources admitted under stated justifications, including contextual Canadian and international surveys, one secondary-school field experiment, and one preprint mechanistic study. Experimental syntheses report medium to large short-term gains when ChatGPT is built into instruction. Survey work points the other way: frequent unstructured use tracks with procrastination, self-reported memory problems, and slightly lower grades, and unrestricted access during practice has been shown to depress later unaided performance. Purpose of use reconciles most of that disagreement, because a tool that scaffolds thinking behaves very differently from one that replaces it. Canadian peer-reviewed studies document heavy campus adoption and considerable student ambivalence about integrity and learning, yet almost none link purpose-differentiated use to measured performance. Closing that gap matters for assessment redesign, AI-literacy programming, and the credibility of the credentials Canadian universities issue.

Introduction

Generative artificial intelligence (AI) tools such as ChatGPT, Microsoft Copilot, and Google Gemini have become part of ordinary academic life for university students. A global survey of 3,839 students across 16 countries found that 86% already used AI in their studies and 54% used it at least weekly (Digital Education Council, 2024). Students ask these tools to explain difficult concepts, summarize readings, tighten their writing, generate practice questions, and, in many cases, produce assignments outright. Used well, generative AI works something like a tutor who never sleeps. Used carelessly, it can weaken memory, erode independent problem-solving, and hollow out the skills a degree is meant to certify (Abbas et al., 2024; Bastani et al., 2025). Canadian adoption has climbed quickly. In a non-probability online panel of Canadian students aged 18 and over, drawn from university, college, high school, vocational, and postgraduate settings, reported reliance on generative AI for schoolwork rose from 52% in 2023 to 59% in 2024 and 73% in 2025 (KPMG in Canada, 2024, 2025). Institutional responses have not kept pace. Some universities restrict the tools, others encourage them, and many leave the decision to individual instructors, so students improvise. A large share concede they are learning less when they use AI (KPMG in Canada, 2024). Nearly half believe their critical-thinking skills have slipped, and a majority say that using the tools feels like cheating (KPMG in Canada, 2025). The research Canadian institutions would need in order to respond well is thin and scattered. The underlying problem is that one technology plausibly helps and harms at the same time. Asked for explanations, feedback, and practice, it may deepen understanding. Asked for finished essays and worked solutions, it may lift grades in the short run while quietly removing the effort that produces skill. Published work already reflects the split. A systematic review of 69 experimental studies, 62 of them meta-analysed, found a large positive pooled effect of ChatGPT interventions on academic performance, strongest in classrooms rather than laboratories (Deng et al., 2025). A separate meta-analysis of college students reported a medium positive effect on achievement (Sun & Zhou, 2024). A large field experiment found the opposite pattern: students who practised with an unrestricted AI tutor scored worse than peers once the tool was withdrawn, which suggests the tool had stood in for learning rather than supported it (Bastani et al., 2025). Survey research links heavy ChatGPT use to procrastination, memory complaints, and lower grades (Abbas et al., 2024). Early neuroscience work finds that writers who lean on AI assistants show weaker cognitive engagement and remember less of what they wrote (Kosmyna et al., 2025). Read together, these results suggest that outcomes depend less on whether students use generative AI than on how, why, and when they use it. Most studies still count frequency. Few separate learning-oriented use, in which the student seeks explanation, feedback, or practice, from outsourcing-oriented use, in which the tool produces the work. Much of the available evidence also comes from short experiments or from samples outside Canada. Canadian peer-reviewed studies document adoption, purposes, and student concerns well enough (Miller et al., 2025; Mohammad et al., 2025), but they stop short of connecting purpose-differentiated use to measured performance. Domestic and international students may also differ, since AI doubles as language support for many international students, and disciplinary variation remains largely unexamined.

This review treats the Canadian gap as part of its contribution rather than a caveat. It synthesizes international peer-reviewed evidence in order to clarify what Canadian universities can already infer and what they still need to establish. Academic performance here means grades, examination or practice scores, and the closely related outcomes (study habits, critical thinking, and cognitive engagement) that help explain them. Two learning frameworks guide the analysis. Cognitive load theory separates extraneous load, which consumes working memory without building anything, from germane processing, which constructs schemas (Sweller, 1988; Sweller et al., 1998). On that account, generative AI helps when it strips away unnecessary difficulty and harms when it removes the mental work through which understanding forms. Research on desirable difficulties points the same way: conditions that feel harder during study often produce better long-term retention (Bjork, 1994; Bjork & Bjork, 2020). The test for any given use of AI, then, is whether it preserves productive struggle or replaces it. This study examines the relationship between generative AI use and academic performance among undergraduate students, with particular attention to the Canadian university context. The aim is to establish what is currently known about how different purposes and patterns of use relate to academic outcomes. Four objectives follow from that aim:

1. To map the patterns and purposes of generative AI use reported among undergraduate students.
2. To evaluate how generative AI use relates to academic performance and to the study habits, critical thinking, and cognitive engagement that support it.
3. To test whether the distinction between learning-oriented and outsourcing-oriented use accounts for the divergence in reported effects.
4. To appraise the strength of Canadian evidence and identify the gaps most consequential for university policy and practice.

Organizing the literature around purpose rather than frequency is what allows this review to explain why findings diverge and to specify the evidence Canadian universities most need as they rewrite assessment and academic-integrity policy. The contribution is a clearer map of where the literature agrees, where it contradicts itself, and where it is silent, together with practical implications for AI-literacy programming and student support.

Methodology

This study followed a narrative literature review design to examine the relationship between generative artificial intelligence (AI) use and academic performance among undergraduate students, with particular attention to the Canadian university context. A narrative review was chosen because the relevant evidence is methodologically uneven: experimental, survey, qualitative, and mixed-methods studies address the same question with incompatible designs and outcome measures, and a narrative synthesis can compare them interpretively where a statistical pooling could not. The review is purposive and critical rather than exhaustive, prioritizing depth of thematic comparison across the sources most directly relevant to the four objectives. Relevant studies were searched through Google Scholar, Scopus, Web of Science, ERIC, and ScienceDirect. These databases were selected because they index peer-reviewed research on educational technology, higher education, and student learning. Searches were conducted in July 2026 and covered studies published between 2022 and 2026, the period in which generative AI tools such as ChatGPT became widely available to students. Keywords included "generative AI," "ChatGPT," "large language model," "AI chatbot," "academic performance," "learning performance," "learning outcomes," "undergraduate students," "university students," and "higher education," combined with the Boolean operators AND and OR. A sample search string was ("generative AI" OR ChatGPT OR "large language model" OR "AI chatbot") AND ("academic performance" OR "learning performance" OR grades OR "learning outcomes") AND ("undergraduate students" OR "university students" OR "higher education"). Selection used a two-tier evidence framework. Tier 1 sources were peer-reviewed empirical studies or structured syntheses that examined generative AI use in relation to academic performance or closely related learning outcomes among undergraduate or university students, were written in English, were published between 2022 and 2026, and had accessible full texts. Tier 2 sources were admitted only under pre-specified justifications and carried an explicit evidentiary discount: (a) national or international student surveys, used solely for adoption and attitude context; (b) one high-quality secondary-school field experiment (Bastani et al., 2025), included because it supplies the only large randomized comparison of unrestricted versus guardrailed generative AI access identified for this review; and (c) one preprint mechanistic study (Kosmyna et al., 2025), included for neurocognitive evidence unavailable in peer-reviewed form at the time of writing. Records identified from the selected databases were screened for duplicates, assessed by title and abstract, and then assessed by full text against the eligibility criteria. Because the review is purposive rather than a full systematic review, it does not present a PRISMA-style inventory of every retrieved record. The Results instead characterize the included evidence base and synthesize its findings thematically. Data extraction was structured. For each included source the review recorded author and year, country or study location, research objective, design, sample size, population, data-collection method, main variables or outcomes, key findings, and limitations. Two further notes were recorded for every source: whether it distinguished learning-oriented from outsourcing-oriented use, and whether it examined domestic and international students or disciplinary differences. No formal appraisal instrument such as AMSTAR-2 or MMAT was applied. Methodological quality was judged narratively through clarity of objectives, appropriateness of design, sampling, measurement validity, analytical transparency, and reporting of limitations, weighting purpose-of-use measurement and the strength of academic-outcome indicators most heavily. Each core source was then assigned a confidence rating of high, moderate, or low for the specific claims this review draws from it. Findings were analysed through thematic narrative synthesis and grouped into four themes corresponding to the four objectives: patterns and purposes of generative AI use; academic performance and related learning outcomes; learning-oriented versus outsourcing-oriented use; and the strength of Canadian evidence across student groups and disciplines. Because the review draws entirely on previously published studies and involved no direct participation of human subjects, formal ethical approval was not required. All sources are cited, and findings are reported with attention to study design and evidentiary weight. Several limits apply. The review covers only English-language sources indexed in the selected databases, so relevant work in other languages, other databases, or unavailable in full text may have been missed. Generative AI research is expanding quickly, and newer studies will appear after the search window closed. Reliance on Tier 2 surveys for Canadian adoption patterns, and on a secondary-school experiment and a preprint for key mechanistic claims, further limits the strength of inference for Canadian undergraduates.

Results

Overview of the included evidence

The synthesis draws on 10 empirical sources: 7 core empirical or synthetic sources summarized in Tables 1 and 2 (5 Tier 1 and 2 Tier 2), plus 3 contextual Tier 2 surveys (Digital Education Council, 2024; KPMG in Canada, 2024, 2025). Two theoretical strands, cognitive load theory and desirable difficulties, are used for interpretation rather than as evidence. Core Tier 1 sources comprise two meta-analyses of experimental or intervention evidence (Deng et al., 2025; Sun & Zhou, 2024), a multi-wave survey of ChatGPT use and academic outcomes among university students (Abbas et al., 2024), and two Canadian campus surveys of student AI use and perceptions (Miller et al., 2025; Mohammad et al., 2025). The two Tier 2 mechanistic sources are Bastani et al. (2025), from secondary-school mathematics in Turkey, and Kosmyna et al. (2025), which offers preliminary neurocognitive evidence.

Tables 1 and 2 summarize the characteristics and findings of the core empirical and synthetic sources.

Table 1 Study Characteristics of Core Sources

Source	Setting	Design and sample	Confidence
Deng et al. (2025)	International university pool; no Canadian primary studies identified	Meta-analysis of 69 articles (62 meta-analysed); mostly university students	Moderate: publication bias and high heterogeneity
Sun and Zhou (2024)	International college/universi-ty pool; primary studies main-ly Asian	Meta-analysis of 28 articles (65 effect sizes; 1,909 participants)	Moderate: college focus, limit-ed Canadian coverage
Abbas et al. (2024)	Pakistan	Three-wave university survey (Study 2 $N = 494$)	Moderate: validated scale, self-reported GPA

Miller et al. (2025)	Canada	Undergraduate survey (N = 125)	Moderate for perceptions; low for performance effects
Mohammad et al. (2025)	Canada (University of Guelph)	Campus survey (N = 1,265)	Moderate for adoption; low for performance effects
Bastani et al. (2025)*	Turkey	Field experiment with nearly 1,000 high-school mathemat-ics students	High internal validity; low transfer to Canadian undergraduates
Kosmyna et al. (2025)*	United States (lab)	EEG experiment (N = 54; crossover n = 18); preprint	Low: preprint status and small crossover sample

*Tier 2 source included under pre-specified justification.

Table 2 ***Outcomes and Key Findings of Core Sources***

Source	**Main outcomes**	**Purpose of use distinguished?**	**Key finding**
Deng et al. (2025)	Academic performance, moti-vation, higher-order thinking, mental effort	Indirectly, via instructional design	Large positive pooled effect on academic performance in ChatGPT interventions
Sun and Zhou (2024)	Academic achievement	Indirectly, via activity type	Medium positive effect of generative AI on achievement (g = 0.533)
Abbas et al. (2024)	ChatGPT use, procrastination, memory loss, self-reported CGPA	No (overall usage only)	Heavier use linked to more procrastination and memory loss and lower CGPA
Miller et al. (2025)	Perceptions, uses, and concerns	Partially (task purposes)	Students see GenAI as useful for learning but worry about integrity and personal learning
Mohammad et al. (2025)	Usage patterns and perceived learning impact	Partially (task types)	86.4% used AI tools; mixed perceived effects on learning; integrity concerns common
Bastani et al. (2025)*	Practice and exam performance	Yes (via tool design)	Unrestricted GPT access boosted practice but later un-aided exam scores fell by about 17%
Kosmyna et al. (2025)*	Neural connectivity, owner-ship, recall	Yes (via tool condition)	LLM assistance linked to weaker connectivity, lower ownership, and poorer recall

*Tier 2 source included under pre-specified justification.

Outcome measures were heterogeneous. They ranged from objectively scored practice and examination performance to self-reported cumulative GPA, self-reported learning and retention, and related constructs such as procrastination, memory, higher-order thinking, and neural engagement. That heterogeneity limits direct cross-study comparison, so the review weights supervised performance measures and meta-analytic estimates more heavily than perceptual survey items whenever it makes claims about academic performance.

Patterns and purposes of generative AI use

University students report using generative AI across a wide range of academic tasks. In the Canadian campus survey by Mohammad et al. (2025), 86.4% of 1,265 University of Guelph respondents reported using AI tools, mainly for academic support such as explaining concepts, solving homework problems, and clarifying what an assignment actually required. Miller et al. (2025), surveying 125 Canadian undergraduates, found much the same breadth of use, concentrated at moments of cognitive or time-management pressure. Those students generally judged the tools effective for learning while remaining uneasy about integrity, accuracy, and the effect on their own learning. National Canadian survey evidence shows similar task profiles at larger scale. Students commonly report using generative AI for research, idea generation, editing or reviewing assignments, and summarizing information, with a smaller but substantial share using it to write essays or reports (KPMG in Canada, 2024, 2025). Globally, Digital Education Council (2024) documents near-ubiquitous student AI use. Across all of these sources, reported purposes fall along a continuum running from scaffolding, such as explanation, feedback, and brainstorming, to substitution, such as full-draft generation or answer provision. Very few studies measure that continuum with validated purpose-of-use scales. Most capture overall frequency, or task categories that quietly mix learning-oriented and outsourcing-oriented behaviour together.

Academic performance and related learning outcomes

Evidence on academic performance is mixed and highly sensitive to study design and outcome type. Deng et al. (2025) reviewed 69 experimental articles published between 2022 and 2024 and meta-analysed 62 of them. For academic performance the pooled effect was large and positive ($g+ = 0.712$, 95% CI [0.497, 0.926], based on 51 effect sizes). ChatGPT interventions also improved affective-motivational states and higher-order thinking propensities, reduced mental effort, and showed no significant effect on self-efficacy. Most interventions ran at the university level and involved direct student use of ChatGPT. Sun and Zhou (2024) similarly found that generative AI improved college students' academic achievement with a medium effect size (Hedges's $g = 0.533$, 95% CI [0.408, 0.659]). Deng et al. nevertheless cautioned against reading these numbers too confidently: limited power analysis, concerns about post-intervention assessment design, evidence of publication bias for academic performance, and very high between-study heterogeneity all argue for restraint. Gains on short, well-defined tasks may reflect the quality of AI-generated answers as much as durable skill acquisition. The mental-effort finding rests on a small number of effect sizes and is best treated as provisional.

Survey-based evidence points the other way once use becomes frequent and largely unstructured. In a three-wave study of university students in Pakistan ($N = 494$), Abbas et al. (2024) found that heavier ChatGPT use was associated with greater procrastination ($\beta = 0.309$), greater self-reported memory loss ($\beta = 0.274$), and a small negative association with self-reported cumulative GPA ($\beta = -0.104$). Students facing higher academic workload and time pressure were more likely to reach for ChatGPT. Indirect associations from workload and time pressure to GPA through ChatGPT use were negative but only marginally significant, so stronger causal claims are not warranted. Reverse causality also remains plausible, since students already struggling academically may adopt ChatGPT more heavily in the first place. Canadian industry surveys sharpen the tension between grades and learning as perceived outcomes. In the 2025 KPMG panel, 71% of student users said their grades improved after using generative AI, while 66% said they did not think they were learning or retaining as much (KPMG in Canada, 2025). An earlier wave found that 67% of student users doubted they were learning or retaining as much even though 75% reported improved assignment quality (KPMG in Canada, 2024). These are perceptions from a mixed student panel rather than measured undergraduate GPA effects. Set beside the experimental and survey literature, though, they suggest that generative AI can raise short-term performance indicators while threatening the learning those indicators are supposed to represent. The literature also links generative AI use to changes in study habits and cognitive engagement. Abbas et al. (2024) found that ChatGPT use predicted higher procrastination and greater self-reported memory impairment. Canadian survey evidence echoes those concerns at scale: 48% of student users say their critical-thinking skills have deteriorated since they began using generative AI, 65% say peers use AI to avoid critical thinking, and 45% report that their first instinct on receiving an assignment is to turn to AI rather than draft anything themselves (KPMG in Canada, 2025). The same students also report using AI for research, idea generation, editing, and summarization, all of which could support learning if paired with genuine evaluation and revision. Canadian peer-reviewed work shows the same duality: confidence in the tools' usefulness alongside worry about academic integrity and personal learning (Miller et al., 2025; Mohammad et al., 2025). Emerging neurocognitive evidence strengthens the concern that convenience carries a cognitive cost. Kosmyna et al. (2025) compared essay writing with an LLM assistant, with a search engine, and with no tools, using EEG alongside behavioural and linguistic measures. Across 54 participants, brain connectivity was strongest in the brain-only condition, moderate with search-engine support, and weakest with LLM assistance. In a crossover fourth session (n = 18), participants reassigned from an LLM to unaided writing showed reduced alpha and beta connectivity, a pattern suggestive of under-engagement. LLM users also reported lower ownership of their essays and struggled to quote their own writing. Given the small crossover subsample and the preprint status of the work, these results should be read as suggestive rather than established. Deng et al. (2025) complicate any purely negative reading by reporting that experimental ChatGPT interventions can improve higher-order thinking propensities and affective-motivational states while reducing mental effort. Interpreted through cognitive load theory and desirable difficulties (Bjork, 1994; Sweller et al., 1998), reduced mental effort looks beneficial when it removes extraneous load and frees capacity for germane processing, and harmful when it replaces the generative struggle through which understanding is built. The same tool can scaffold thinking or substitute for it, and those two paths lead to different learning consequences.

Learning-oriented versus outsourcing-oriented use

Purpose and design of use offer the strongest available explanation for the contradictory performance findings. Bastani et al. (2025) provide direct experimental evidence for the distinction, though in a secondary-school mathematics setting. In a field experiment with nearly 1,000 students in Turkey, access to an unrestricted GPT-4 interface improved practice performance substantially, by roughly 48% relative to control, while a tutor version built with learning safeguards improved practice performance even more, by roughly 127%. Once AI access was removed, students who had used the unrestricted interface scored about 17% worse than students who never had AI access at all. The guarded tutor condition largely eliminated that negative exam effect, although it still did not produce a clear positive learning gain relative to control. Interaction analyses showed why: students often used the unrestricted tool as a crutch, requesting and copying solutions, whereas the guarded tutor drew out more substantive help-seeking and more independent attempts at answers. Students also appeared largely unaware that copying solutions had damaged their later performance. These findings map directly onto the distinction that structures this review. Learning-oriented use, in which the tool explains, hints, scaffolds, or gives feedback while the student remains responsible for producing the answer, is associated with better immediate performance and less damage to later unaided performance. Outsourcing-oriented use, in which the tool does the intellectual work, can inflate practice or assignment scores while weakening the skill that gets measured once the tool is gone. Deng et al. (2025) reinforce that reading by arguing that future assessments should move beyond well-defined problems AI can solve directly and toward tasks requiring demonstrated skill, originality, or proctored performance. Their subject-area pattern is suggestive as well: science showed a smaller positive effect than arts and humanities, which is consistent with the possibility that domains with clear right answers are especially vulnerable to solution-copying. Abbas et al. (2024), by measuring overall ChatGPT usage rather than purpose, almost certainly capture a mixture of both patterns, which helps explain why frequency alone predicts poorer academic outcomes in survey settings where outsourcing is easy and common.

Canadian evidence, student groups, and disciplines

Direct Canadian peer-reviewed evidence linking generative AI usage patterns to measured academic performance remains scarce, and that scarcity is itself a central finding. What Canadian peer-reviewed studies show most clearly is high campus adoption paired with ambivalent student experience. Mohammad et al. (2025) document very high AI use at one Ontario university and mixed perceptions of learning impact. Miller et al. (2025) show that Canadian undergraduates often judge GenAI effective for learning while remaining concerned about integrity, accuracy, and effects on their own learning. Neither study estimates GPA or examination effects by purpose of use. National industry surveys extend the picture, with clear limits. Reported student use of generative AI for schoolwork rose from 52% in 2023 to 59% in 2024 and 73% in 2025, alongside rising daily use and persistent integrity anxiety: a majority of users still feel as though they are cheating, and many fear being caught (KPMG in Canada, 2024, 2025). In 2025, 77% wanted their educational institution to offer courses on ethical and safe AI use (KPMG in Canada, 2025). These are non-probability panels spanning several education levels, so they indicate national trends rather than establish undergraduate-specific causal effects. Digital Education Council (2024) reports similarly high adoption across countries, which places Canada inside a broader shift rather than outside it. Disciplinary and student-group differences are under-reported relative to their practical importance. Deng et al. (2025) note that experimental work clusters in language education and university settings, but fine-grained comparisons by discipline, by domestic versus international status, or by language background remain limited. Bastani et al. (2025) show that tool design matters enormously within mathematics practice, which implies that STEM contexts with clear right answers may be especially exposed to solution-copying. Writing-intensive contexts may instead be exposed to the ownership and recall deficits suggested by Kosmyna et al. (2025). International students, for whom generative AI may function as language support as well as content support, are largely absent as an analytically distinct group in the current peer-reviewed outcome literature. The Canadian undergraduate setting, with its large international student population and uneven institutional AI policies, remains an important empirical gap. Three patterns recur across the evidence base. Generative AI can improve short-term academic performance, particularly within structured instructional interventions. Unrestricted or outsourcing-oriented use can damage later unaided performance, study habits, and cognitive engagement. And Canadian evidence documents high, rising adoption and student unease about a grades-versus-learning trade-off while still lacking peer-reviewed, purpose-differentiated estimates of effects on undergraduate academic performance.

Discussion

The central finding of this review is that generative AI's relationship to academic performance is conditional. When studies embed ChatGPT in instructional designs that keep students cognitively active, performance and related learning indicators often improve (Deng et al., 2025; Sun & Zhou, 2024). When students use unrestricted tools to obtain answers or outsource production, short-term scores may rise while later unaided performance, memory, and sense of ownership fall (Abbas et al., 2024; Bastani et al., 2025; Kosmyna et al., 2025). That conditionality explains why the literature can look optimistic and alarming at the same time. The studies are frequently measuring different uses of the same technology. The learning-oriented versus outsourcing-oriented distinction is a more useful organizing principle than frequency of use. Frequency still matters, because heavier use under workload and time pressure tracks with worse academic outcomes in survey research (Abbas et al., 2024). But Bastani et al. (2025) show that two students can both use AI intensively and experience opposite learning consequences, depending only on whether the tool withholds answers and demands thinking or simply supplies solutions. For Canadian universities, this means that policies framed purely as banning or allowing AI are incomplete. The more consequential question is what kinds of use their assessments, course designs, and AI interfaces make possible in the first place. Canadian evidence raises the policy stakes while exposing the research gap that motivated this review. Peer-reviewed campus studies confirm that generative AI is already ordinary student infrastructure in Canadian universities (Miller et al., 2025; Mohammad et al., 2025). National surveys report that many students believe their grades have improved while also believing they are learning less (KPMG in Canada, 2024, 2025). That combination is exactly what widespread outsourcing-oriented use would produce: grades become noisier signals of learning. Student reports of deteriorated critical thinking and first-instinct reliance on AI suggest that habitual outsourcing may already be reshaping study practices. Because Canadian peer-reviewed studies still emphasize perceptions and usage patterns rather than purpose-linked performance outcomes, and because the national surveys are industry panels rather than academic outcome studies, Canada currently has a strong adoption story attached to a weak causal evidence base. Peer-reviewed Canadian studies connecting purpose-differentiated AI use to grades, supervised assessments, and skill measures are the priority.

The findings carry implications for assessment and AI literacy. If AI can complete conventional take-home tasks, then leaving assessment regimes unchanged risks rewarding outsourcing. Deng et al. (2025) recommend more complex, project-based, originality-sensitive, or proctored assessments. Canadian student demand for institutional AI-literacy courses (KPMG in Canada, 2025) indicates receptivity to guidance rather than prohibition alone. Effective institutional responses will likely combine clear norms about acceptable use, assessment redesign that preserves unaided demonstration of skill, and instruction that teaches students to use AI for explanation, critique, and revision rather than substitution. Several limitations of the reviewed evidence constrain certainty. Experimental syntheses concentrate on language-education and classroom contexts, and their post-intervention assessments often use well-defined problems that AI can solve directly (Deng et al., 2025). Survey studies rely on self-reported use and grades, rarely separate learning- oriented from outsourcing-oriented behaviour, and cannot rule out reverse causality or confounding by prior achievement (Abbas et al., 2024). The most direct causal evidence on guardrails comes from secondary-school mathematics rather than Canadian undergraduates (Bastani et al., 2025). The neurocognitive evidence is preliminary, not peer-reviewed, and based on a small crossover subsample (Kosmyna et al., 2025). Canadian adoption statistics, while timely, are not academic outcome studies (KPMG in Canada, 2024, 2025). The review itself is purposive, limited to English-language sources in selected databases, and cannot claim exhaustive coverage of a literature expanding this quickly. Future research in Canada should measure purpose of AI use directly, combine self-reports with behavioural traces or supervised performance tasks, and compare domestic and international students as well as disciplines. Longitudinal designs are needed to test whether early grade gains persist or reverse as dependency deepens. Research should also evaluate institutional interventions themselves: AI-literacy courses, assessment redesign, and guardrailed tutoring tools. Until such evidence exists, universities should read high adoption not as proof of educational benefit but as a reason to align tool design, pedagogy, and assessment with learning rather than output production.

Conclusion and Policy Implications

Generative AI neither helps nor harms undergraduate learning on its own. Its effect depends on purpose: structured, learning-oriented use tends to improve performance, while unrestricted, outsourcing-oriented use can inflate short-term scores and weaken later unaided achievement, study habits, and cognitive engagement. Canada has abundant evidence of rapid adoption and student ambivalence, and almost none linking purpose-differentiated use to measured undergraduate performance. That gap, not adoption itself, is the problem worth solving.

Impact of the research and its policy relevance

The practical contribution of this review is a shift in the unit of measurement. Policy conversations in Canada currently run on adoption statistics, which say how many students use AI and how often, and adoption statistics cannot distinguish a student who asks for a worked explanation from one who submits a generated essay. By showing that those two behaviours produce opposite learning outcomes, the review reframes the policy question from how much AI students use to what kind of use institutions permit, design for, and assess. That reframing is directly usable: it converts an unresolved debate about banning or allowing AI into a set of testable design choices about assessment format, tool configuration, and instruction. The review also carries a warning that reaches beyond education policy. If outsourcing-oriented use is common and grades still rise, then grades decouple from skill, and every downstream system that trusts a transcript, including graduate admissions, professional licensing, and employer hiring, inherits a measurement problem it did not create and cannot see. Treating that as a credential-integrity issue rather than a classroom-discipline issue changes who has standing to act on it.

How the findings help government

Education falls under provincial and territorial jurisdiction in Canada, so the most actionable audience is provincial ministries responsible for colleges and universities. Three uses follow from this review. First, ministries can require or fund purpose-differentiated measurement rather than frequency counts as a condition of AI-related programme funding, since frequency data cannot support the distinction the evidence turns on. Second, existing national instruments, including Statistics Canada's postsecondary and graduate outcome surveys and the annual Canadian Undergraduate Survey Consortium survey, could carry purpose-of-use items at low marginal cost, which would give governments comparable national data instead of one-campus studies. Third, quality-assurance and academic-integrity expectations can be revised around demonstrated skill rather than submitted output, which is a lever ministries already hold. Federal departments have a narrower but real interest. Skills and labour-market policy depends on credentials meaning what they claim to mean, so evidence that AI use inflates grades without building competence is labour-market information, not merely an academic concern. Canada's federal AI strategy emphasizes adoption and literacy; this review supplies the specific finding that AI literacy must teach the difference between scaffolding and substitution, because students in the reviewed studies could not reliably detect when their own AI use was hurting them. Given that international students form a large share of Canadian undergraduate enrolment and may use AI as language support, immigration and postsecondary policy will also need evidence disaggregated by student status before drawing conclusions about integrity.

How the findings help universities

For universities the implications are immediate and mostly within existing authority. Assessment redesign is the highest-leverage response: where an assignment can be completed end-to-end by a chatbot, the assessment measures tool access rather than learning, and moving toward supervised, project-based, originality-sensitive, or oral components restores the signal. Second, AI-literacy programming should be built around purpose rather than prohibition, teaching students to use AI for explanation, critique, and revision while keeping them responsible for producing the work. Student demand for exactly this kind of instruction is already documented. Third, institutional policy should specify acceptable use at the level of course and assessment design instead of issuing campus-wide bans, since the evidence shows tool configuration, not tool availability, determines the outcome. Finally, teaching and learning centres and institutional research offices are well placed to run the missing Canadian studies, since they can link purpose-of-use measures to actual grade and assessment records under existing ethics frameworks. If Canadian institutions leave purpose of use unexamined, they risk mistaking higher grades for deeper learning at precisely the moment generative AI makes that distinction hardest to see.

Author note

Generative AI tools were used to assist with literature organization, drafting support, and editorial revision of this manuscript. All source selection decisions, interpretations, and final wording were reviewed and approved by the author. No competing interests are declared. This review received no external funding.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education, 21(1), Article 10. https://doi.org/10.1186/s41239-024-00444-7

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). MIT Press.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. Journal of Applied Research in Memory and Cognition, 9(4), 475–479. http s://doi.org/10.1016/j .j armac.2020.09.003

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 227, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Digital Education Council. (2024). Digital Education Council global AI student survey 2024. http s://www.digitaleducationcouncil.com/p ost/digital-education-council-global-ai-student-survey -2024

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task [Preprint]. arXiv. http s://doi.org/10.48550/arXiv.2506.08872

KPMG in Canada. (2024, October). With most Canadian students now using generative AI, they confess they're not learning as much http s://kpmg.com/ca/en/home/media/p ress-releases/2024/10/students-using-gen-ai-say -they - are-not-learning-as-much.html

KPMG in Canada. (2025, October). Generative AI boom among Canadian students raises dilemmas https://kpmg.com/ca/en/home/media/press-releases/2025/10/generative-ai-boom-among-canadian-students-raises-dilemmas.html

Miller, M., Muñoz Bocanegra, B., & Choi, Y. (2025). A precarious partnership: Student perceptions of generative AI in post-secondary learning. The Open/Technology in Education, Society, and Scholarship Association Journal, 4(3), 1–18. https://doi.org/10.18357/otessaj.2024.4.3.74

Mohammad, N. M., Demers, M., McCubbin, E., Mitchell, J., & Fulmer, S. M. (2025). How college students use ChatGPT. Pedagogical Research, 10(4), Article em0250. https://doi.org/10.29333/pr/17428

Sun, L., & Zhou, L. (2024). Does generative artificial intelligence improve the academic achievement of college students? A meta-analysis. Journal of Educational Computing Research, 62(7), 1896–1933. http s://doi.org/10.1177/07356331241277937

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251–296. https://doi.org/10.1023/A:1022193728205