



INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

A Generative AI-Integrated Framework for Teaching Operations Management: Case-Based Projects

Sajad Ebrahimi, Raj Bridgelall

To cite this article:

Sajad Ebrahimi, Raj Bridgelall (2026) A Generative AI-Integrated Framework for Teaching Operations Management: Case-Based Projects. INFORMS Transactions on Education

Published online in Articles in Advance 30 Oct 2026

. <https://doi.org/10.1287/ited.2026.0224>

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “*INFORMS Transactions on Education*.” Copyright © 2026 The Author(s). <https://doi.org/10.1287/ited.2026.0224>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.”

Copyright © 2026 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes. For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

A Generative AI-Integrated Framework for Teaching Operations Management: Case-Based Projects

Sajad Ebrahimi,^{a,*}  Raj Bridgelall^b 

^aManagement Department, Seidman College of Business, Grand Valley State University, Allendale, Michigan 49401; ^bDepartment of Transportation and Supply Chain Management, College of Business, North Dakota State University, Fargo, North Dakota 58108

*Corresponding author

✉ ebrahims@gvsu.edu (S. Ebrahimi), raj.bridgelall@ndsu.edu (R. Bridgelall)

 <https://orcid.org/0000-0001-8655-6330> (S. Ebrahimi), <https://orcid.org/0000-0003-3743-6652> (R. Bridgelall)

Received: May 12, 2026

Revised: July 7, 2026; August 17, 2026

Accepted: September 18, 2026

Published Online in Articles in Advance:
October 30, 2026

<https://doi.org/10.1287/ited.2026.0224>

Copyright: © 2026 The Author(s)

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "INFORMS Transactions on Education. Copyright © 2026 The Author(s). <https://doi.org/10.1287/ited.2026.0224>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>."

Abstract. As generative artificial intelligence (GAI) reshapes business education, instructors must simultaneously safeguard academic integrity and develop students' AI literacy for an increasingly AI-driven workplace. This study developed the GAI Learning Framework (GLF), a replicable, step-by-step experiential learning approach for case-based projects in undergraduate business courses, illustrated through an operations management (OM) application. Rather than enabling passive cognitive offloading, the GLF promotes active learning by requiring students to craft context-rich prompts grounded in course materials, systematically compare outputs across multiple GAI platforms (e.g., ChatGPT, Claude, and Gemini), and critically evaluate discrepancies against independently verified sources. The framework was implemented across two sections of an undergraduate OM course and evaluated using a student survey that combined Likert-scale items with open-ended reflections ($n = 50$). All three constructs exceeded the scale midpoint, with pooled means ranging from 3.93 (Constructive Alignment) to 4.08 (Cognitive Load/Scaffolding). Qualitative student reflections further indicated that students perceived the framework as fostering active learning, AI platform literacy, and professional communication skills. This study offers educators a practical, classroom-tested framework for integrating GAI tools into case-based projects in business courses while preserving analytical rigor and promoting experiential learning.

Supplemental Material: The online appendices are available at <https://doi.org/10.1287/ited.2026.0224>.

Keywords: generative artificial intelligence • large language models • operations management learning • instructional framework

1. Introduction

Operations management is a required course in most undergraduate business programs, yet the limited duration of a semester typically restricts instruction to foundational concepts, leaving students underprepared for real-world application (Drake 2019, Petersen et al. 2020). Recognizing this gap, educators have long employed active learning strategies, such as simulations, case studies, and project-based assignments, to help students connect theory with practice (Bromley 2013, Freeman et al. 2014).

Business education has long relied on experiential instructional approaches to bridge the gap between conceptual knowledge and real-world decision making, particularly in applied fields such as operations management, supply chain management, and business analytics (Bonwell and Eison 1991, Prince 2004). Among these approaches, case-based projects occupy a distinctive position, combining the analytical depth of the case method with the sustained engagement characteristic of project-based learning (Thomas 2000,

Drake 2019). In a case-based project, students examine a real company and apply concepts to a series of analytical questions, connecting theoretical frameworks to real or simulated challenges (Dean and Fornaciari 2002, Prince 2004). Within the business education literature, closely related activities have been described using a variety of terms, including case articles, real-world project-based cases, experiential case exercises, and consulting-style projects, reflecting both disciplinary differences and the evolving nature of experiential pedagogy (Heriot et al. 2008, Korfhage Smith 2010). Despite these terminological differences, such approaches share core design principles: sustained engagement with a real organization, scaffolded deliverables, and assessment mechanisms that emphasize synthesis and analytical rigor (Dean and Fornaciari 2002, Hillon et al. 2012).

At the same time, the rapid rise of generative artificial intelligence (GAI) tools creates both new opportunities and challenges for case-based projects, particularly regarding how students produce analyses,

verify insights, maintain academic integrity, and demonstrate individual understanding (Gruenhagen et al. 2024, Larson et al. 2024). As GAI continues to reshape professional business environments, business schools play a crucial role in equipping students to harness these technologies responsibly and in providing proactive learning opportunities that prepare them for a technology-driven professional future (Greenland et al. 2022, Gupta et al. 2024). A compelling need, therefore, exists to revisit and refine case-based project design, specifically regarding how to structure assignments to support learning, critical thinking, and academic integrity in AI-augmented environments.

The integration of GAI into academic coursework introduces several well-documented concerns. First, students risk relying too much on GAI outputs as a substitute for their own reasoning; a form of cognitive offloading that may reduce active processing unless assignments explicitly require evaluation, justification, and source verification (Jose et al. 2025, Wang 2026). Fan et al. (2025) report that learners using a generative assistant improved their essay scores more than comparison groups while showing no corresponding gains in knowledge acquisition or transfer, a pattern they describe as metacognitive laziness. Second, when case-based projects demand higher-order thinking, the cognitive load imposed by both the subject matter and the tool-use expectations can substantially influence how effectively students engage with GAI. Li et al. (2026) found that differing degrees of cognitive load within an assigned project significantly affected the quality of GAI integration. This finding highlights the need for structured frameworks that scaffold task decomposition and guide the use of AI tools (Patac and Patac 2025). Third, the use of platforms such as ChatGPT has raised persistent questions about assessment integrity in higher education, prompting discussions about rethinking evaluation design to include in-class assessments and oral presentations (Kirwan 2024, Lee et al. 2024). Educators must therefore develop proficiency in using GAI effectively and establish clear guidelines for students' responsible use of the tools.

Prior studies proposed ways to incorporate GAI into coursework, but relatively few provided an auditable, replicable, step-by-step framework for using multiple GAI tools to support case-based learning in undergraduate business courses, partly due to the limited educator time and ongoing uncertainty about effective prompt design (Carrasco-Sáez et al. 2025). To address this gap, this study developed a GAI Learning Framework (GLF) for undergraduate case-based projects in business courses, demonstrated through an application to operations management. This research makes two primary contributions:

(1) **A replicable GLF** outlining the process for assigning and completing GAI-assisted case-based

projects in undergraduate business courses, providing educators with guidance to implement the approach. The framework is supported by a prompt-engineering guide and an assessment rubric designed to promote analytical rigor, active engagement, and responsible AI use.

(2) **A perception-based evaluation of the GLF** through an anonymous student survey administered across two course sections ($n = 50$), providing preliminary evidence of the framework's perceived pedagogical value in supporting active learning, AI literacy, and cognitive load management.

The remainder of this study is organized as follows. Section 2 provides background on GAI and its applications in higher education. Section 3 presents the proposed GLF. Section 4 illustrates the framework through a Walmart example case. Section 5 describes the assessment design and instructor guidance. Section 6 presents a survey-based evaluation of the GLF across two sections of an undergraduate OM course. Section 7 presents the study's conclusions, and Section 8 identifies limitations and directions for future research.

2. GAI and Its Applications in Higher Education Pedagogy

GAI refers to a class of AI systems capable of producing new content, including text, images, audio, and code, in response to user-provided prompts (Dwivedi et al. 2023). Unlike earlier AI systems designed primarily for prediction, classification, or optimization, GAI models generate novel outputs based on training across large and diverse data sets. Because OpenAI publicly released ChatGPT in late 2022, GAI tools have rapidly entered higher education because of their ability to support a wide range of academic activities through natural-language interaction (Grant and Metz 2022, Lee et al. 2024).

Despite these promising capabilities, the growing use of GAI in education raises significant concerns. Batista et al. (2024) grouped these concerns into three categories, particularly relevant to case-based project design. The first concern pertains to academic integrity, particularly uncertainty regarding authorship and the potential risk of students submitting AI-generated content as their own work. The second concern is imbalanced assistance: students may rely on GAI outputs rather than engaging in independent analytical reasoning, a pattern of cognitive offloading that suppresses active processing unless assignments explicitly require justification and source verification (Vendrell and Johnston 2026). The third concern is assessment design: GAI systems can produce plausible responses to conventional assignments, and traditional evaluation methods may no longer accurately measure student understanding (Batista et al. 2024).

These concerns are particularly acute in business education, where students must develop critical thinking, analytical reasoning, and decision-making skills for complex organizational environments. Business schools have therefore begun exploring ways to integrate GAI into their curricula while ensuring students retain the ability to evaluate information critically and apply theoretical frameworks appropriately (Greenland et al. 2022, Mehmet et al. 2025). In this context, GAI tools can serve as supportive resources that help students locate relevant information, synthesize insights, and examine alternative interpretations of business problems, provided assignments are designed to require genuine analytical engagement rather than passive acceptance of GAI outputs.

Although the volume of empirical research on GAI in business education is growing rapidly, most studies focus on student attitudes or general AI use patterns rather than on structured frameworks that guide multi-step, platform-specific engagement with AI tools. Li and Ho (2024) evaluated ChatGPT's performance on OM problems and found that the system performed adequately on qualitative and deterministic tasks but struggled with stochastic problems requiring higher-order cognitive reasoning. They further found that some AI-generated business analytics responses evaded conventional detection, highlighting the need for assessment designs that evaluate the reasoning process rather than the output alone.

Taken together, this body of work identifies three conditions a pedagogically sound GAI integration framework must satisfy: structuring task complexity to manage cognitive load (Sweller 1994, Li et al. 2026), requiring active evaluation rather than passive acceptance of outputs (Freeman et al. 2014, Vendrell and Johnston 2026), and aligning assessment with the reasoning process rather than the output alone (Batista et al. 2024, Li and Ho 2024). However, no published framework operationalizes all three conditions simultaneously within a semester-long, multiplatform, case-based OM project. The GLF, presented in Section 3, is designed to satisfy all three conditions within the context of a semester-long, case-based OM project.

3. GLF

This section presents the eight-step GLF, designed to guide undergraduate students through the process of producing, evaluating, and refining GAI-assisted case-based projects within an OM course. As illustrated in Figure 1, the GLF proceeded sequentially: students selected a focal company (Step 3.1), completed the relevant module concepts (Step 3.2), designed structured prompts (Step 3.3), generated and collected outputs across multiple GAI platforms, with each session documented via a conversation record (Step 3.4), evaluated

and refined those outputs (Step 3.5), developed evidence-based recommendations (Step 3.6), wrote the conclusion (Step 3.7), and compiled and presented findings (Step 3.8).

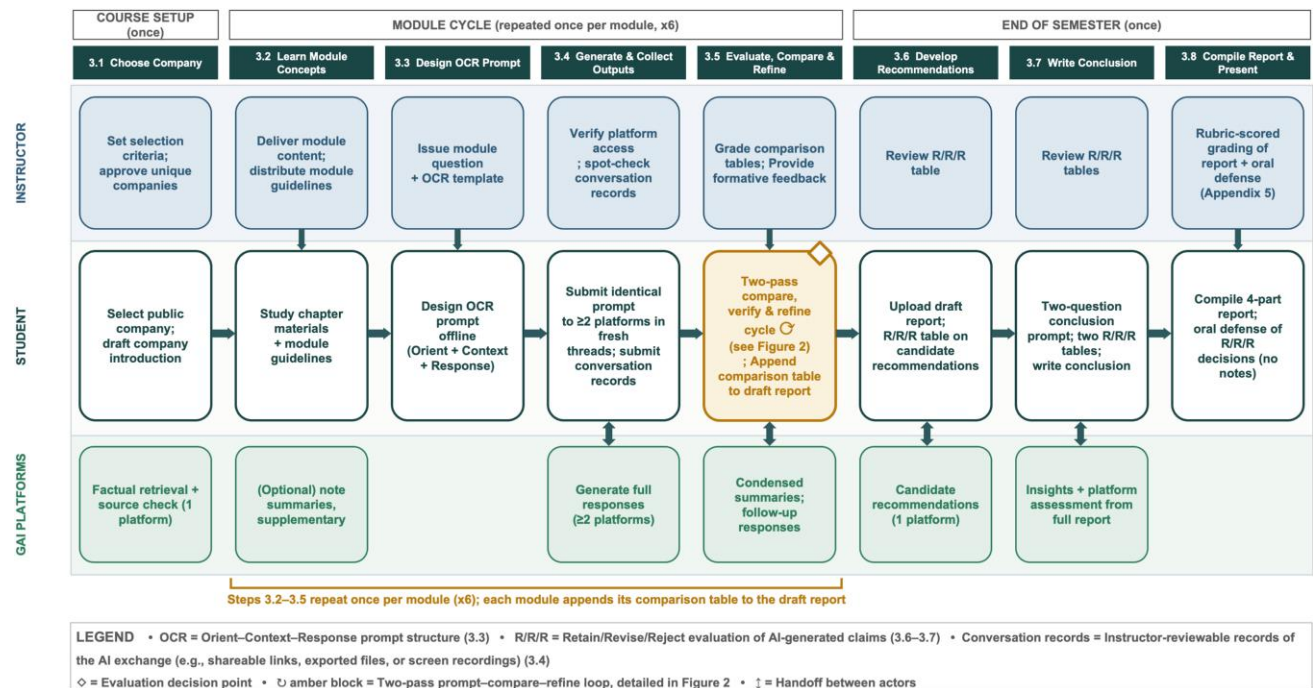
Across these eight steps, students produce two primary semester-long deliverables: a written case report and an oral presentation, both organized around the same four-part structure, including Company Introduction, Module Analyses, Recommendations, and Conclusion. Throughout this paper, a module refers to one instructional unit of the course, a bundle of textbook chapters, lecture sessions, slide decks, and practice exercises covering one OM topic area (e.g., operations strategy, demand management), and a module-level analysis refers to the student's written response to the project question assigned for that module, produced through Steps 3.3–3.5 and appended to the draft report before the next module begins. The written report is not submitted as a single end-of-semester document; rather, it is built incrementally as a living draft, with each module analysis completed and added to the report, and submitted to the instructor at the close of the corresponding module before the next module begins. The complete written report is submitted as a single compiled document at the end of the semester. This staged submission structure creates natural accountability checkpoints throughout the semester and prevents end-of-term compression of the analytical work.

The GLF was grounded in three complementary pedagogical principles. First, constructive alignment (Biggs et al. 2022) ensured that each step directly supported the intended learning outcomes: analytical reasoning, AI literacy, and communication of findings. Second, cognitive load theory (Dunlosky et al. 2013) informed the scaffolded, module-by-module structure that broke complex analytical tasks into manageable stages. Third, active learning principles (Freeman et al. 2014) underpinned the emphasis on cross-platform comparison and source verification, requiring students to critically evaluate GAI outputs rather than passively accept them. Together, these principles distinguish the GLF from generic AI-use guides by embedding responsible, learning-centered engagement into each step of the workflow.

3.1. Choosing the Company

Students select a real organization as the focal case for the semester-long project. To improve response quality and reduce the risk of AI hallucination, students are encouraged to select a company with a substantial and verifiable public presence, such as a large retailer, manufacturer, or logistics firm, for which credible information is available across company filings, industry reports, and news sources (Huang et al. 2025). Although this approach may not eliminate hallucinations entirely, it

Figure 1. GLF Roadmap



Notes. Three columns show instructor, student, and GAI-platform activities across Steps 3.1–3.8. Diamond nodes are evaluation decisions; the amber block at Step 3.5 denotes the two-pass prompt-compare-refine cycle detailed in Figure 2.

increases the likelihood that GAI models can draw on well-documented, retrievable information rather than generating plausible but unverifiable responses.

After selecting a focal company, students complete a structured company introduction that will serve as the opening section of the written report and the first segment of the oral presentation. Using a GAI tool with appropriate prompting, they compile a concise overview of the company's history, mission and vision, main products and services, industry segment and key competitors, market position, and strategic priorities, and then verify AI-generated claims against independent sources before incorporating the introduction into their report. Section 4.1 illustrates this step with an example prompt and company introduction for Walmart.

3.2. Learning Module Concepts

Before engaging with GAI tools for the project, students are expected to complete all assigned learning materials for the relevant course module, including readings, lectures, and case exercises, as a foundational context. Familiarity with course concept terminology and theories is necessary for students to evaluate whether GAI outputs are accurate and appropriately scoped (Biggs et al. 2022).

At the start of each course module, the instructor provides students with a set of module-specific project completion guidelines: a document that defines the

module's project question, specifies the scope of OM concepts to be addressed, and identifies the types of sources and evidence acceptable for verification. These guidelines serve as the anchor for all subsequent steps in the GLF and should be reviewed carefully before prompt design begins.

Students may use GAI tools in a supplementary capacity during this step, for example, to organize lecture notes into structured summaries or to identify gaps in their understanding. Comparing an AI-generated concept summary against one's own notes can help reinforce understanding and surface misalignments before the related module's work on the project begins (Dunlosky et al. 2013). Instructors are encouraged to incorporate case discussions and worked examples into the module to deepen students' familiarity with how the given OM concepts manifest in practice, which, in turn, supports the quality of prompts students can design or customize.

3.3. Designing Module-Level Analytical Prompts

This step applies specifically to the module-level project questions that form the core analytical work of the GLF. The company introduction prompt, described in Section 3.1, follows a different design logic and verification approach, as it targets factual retrieval rather than concept-driven analysis. However, this step guides students in constructing module-level prompts that elicit accurate, course-aligned, and concept-grounded

analytical responses from GAI tools, a practice known as prompt engineering (Al-kfairy et al. 2024). Prompt engineering has emerged as a critical skill for producing reliable outputs from large language models (LLMs), and there are currently no established prompt engineering standards tailored specifically to higher education learning activities (Carrasco-Sáez et al. 2025, Lee and Palmer 2025). The GLF provides students with a structured prompt template to reduce this uncertainty.

Each prompt should be composed of three elements (Orient, Context, and Response), forming the OCR framework:

(1) **Orient:** A role instruction that positions the GAI as a subject-matter expert (e.g., “You are an expert in operations management”), establishing the analytical lens and level of expertise the response should reflect.

(2) **Context:** The project question drawn directly from the module-specific report guidelines, together with the uploaded course materials relevant to the module, which prime the platform with domain-specific knowledge through a process known as retrieval-augmented generation (RAG).

(3) **Response:** A specification of the desired output format, including a structural example of the expected answer, a technique known as one-shot prompting. Providing this structural example prevents the GAI from generating overly verbose or mismatched responses, directly reducing the iterative burden of reprompting that students often experience (Farrokhnia et al. 2026). The Response element should specify the desired output structure, such as headers and labeled categories, but should not impose sentence counts or word limits on the first pass. (Length constraints are applied only when requesting a condensed follow-up summary in a subsequent prompt, a process described in Step 3.5.)

Section 4.2 presents two fully developed module-level prompts that demonstrate the three-element OCR structure applied to course-specific analytical questions, in contrast to the introductory prompt illustrated in Section 4.1.

The initial prompt for each module is designed offline, before any platform interaction, for a methodological reason: Submitting an identical, OCR-structured opening prompt to every platform in a fresh thread ensures that the cross-platform differences students analyze in Step 3.5 reflect platform behavior rather than differences in what each platform was requested. Constructing the prompt against the module guidelines and course materials, rather than through trial-and-error in the chat window, is itself the intended learning activity, requiring students to translate course concepts into a precise analytical question. This does not mean interaction is discouraged: After the initial submission, Step 3.5 requires interactive engagement, including the second-pass condensation prompt, follow-up prompts probing discrepancies between

platforms, and elaboration requests. All of these exchanges occur within the same recorded thread and are visible to the instructor in the submitted conversation record, so the full interaction history, not only the opening prompt, is part of the assessed record. Students who discover through interaction that an element of their initial prompt was ambiguous document this in the process summary and may carry the improvement into the next module’s prompt, making prompt design an iteratively improving skill across the semester rather than a one-shot task.

3.4. Generating and Collecting GAI Outputs

Students must begin a new conversation thread for each module’s project question to prevent the GAI from conflating topics across modules and to ensure focused, contextually clean responses. Once the prompt is finalized, students submit it, along with the uploaded course materials, to at least two GAI platforms (e.g., ChatGPT, Claude, Gemini). GAI platform features and access tiers change frequently, so instructors should confirm which retrieval and citation features are available to their students and provide updated information about current model capabilities and limitations before the project begins (Francis et al. 2025, Sun et al. 2026). Section 5.1 provides detailed guidance on managing unequal access tiers and midsemester platform changes. If organized in teams, each student group should consistently utilize the same set of platforms across all project modules throughout the semester to ensure comparability of outputs over time.

To document compliance with the prompt construction requirements outlined in Step 3.3, students must submit a conversation record for each platform used in each module. A conversation record is any platform-supported or instructor-approved format that preserves the relevant AI exchange for review, such as a shareable conversation link, exported conversation file, PDF, or screen recording. These records give the instructor access to the full exchange for auditing, including the opening prompt, follow-up interactions, and generated responses, and serve as the primary compliance record for this step. Students should save the conversation record immediately after completing each platform session and include the record, or a reference to it, in the corresponding row of the comparison table. In the reported implementation, shareable links were used where available; when a platform did not support an accessible shareable link, students submitted an exported PDF or a screen recording of the conversation. Two constraints made this necessary: on some platforms a conversation containing an uploaded file cannot be shared by public link at all, and on others link sharing is limited to members of the same institutional workspace.

Instructors are encouraged to conduct a midsemester check-in, ideally after the second module's comparison table is submitted, by reviewing two or three students' conversation records per section to verify the prompt structure and the quality of follow-up exchanges. On the platforms used here, shared links do not display uploaded files as of this writing; to verify that course materials have been uploaded, instructors can periodically and randomly request students to submit a screenshot of the platform interface showing the attached files at the time of submission. This targeted review provides an early-warning signal before the final report is assembled, without requiring instructors to review all submissions.

3.5. Evaluating, Comparing, and Refining GAI Outputs

After collecting responses from all selected platforms, students systematically evaluate and compare the outputs across three dimensions, including accuracy, depth and relevance, and consistency. Students must record their findings in a structured comparison table (rows listing the questions and columns showing the responses received using the GAI tools) and include the table in their case report for each question in each course module. The three dimensions are as follows.

- **Accuracy:** Students verify whether each platform's response aligns with the concepts and frameworks covered in the module. GAI models can confidently produce inaccurate information, a phenomenon known as hallucination (Ji et al. 2023, Bridgelall 2024), so students should request supporting references from each platform and cross-check them against course materials and independent sources. When a platform cannot provide verifiable references, students must consult their instructor, refer to company filings or industry reports, or seek input from a subject-matter practitioner (Bernik et al. 2025, Liu et al. 2025). When activating deep search or Internet retrieval features, students should explicitly instruct the platform to consult only scholarly, peer-reviewed sources and official company filings or government reports, and to exclude blogs, opinion pieces, and marketing materials such as promotional brochures or investor-targeted reports, which may present biased or unverifiable claims. For student-conducted verification, the module guidelines specify a three-tier source hierarchy. Tier 1 (authoritative): audited company filings (10-K/annual reports), regulatory documents, and peer-reviewed academic sources; Tier 2 (credible secondary): established industry reports, government statistics, and major business press with named authors and editorial standards; Tier 3 (excluded): blogs, opinion pieces, promotional or investor-marketing materials, user-generated content, and AI-generated summaries cited as sources. A claim is verified when supported by at least one Tier 1 source or two independent Tier 2

sources. When acceptable sources conflict, students apply a precedence rule: Tier 1 outranks Tier 2; among equal-tier sources, the more recent and more specific to the claim prevails; and if the conflict survives these tests, the student reports both positions in the process summary, states which they adopted and why, and flags the claim as contested. Conflicts that cannot be adjudicated from available sources are legitimate findings: the GLF treats an honestly documented unresolved conflict as stronger work than a forced resolution, and the rubric's critical-evaluation dimension rewards the documentation, not the tidiness of the outcome.

- **Depth and relevance:** Students assess whether each response substantively addresses the project question or provides only a surface-level answer. Responses should be evaluated for alignment with the prompt's scope and the level of analytical detail specified in the module guidelines (Kasneji et al. 2023, Bai and Wang 2025).

- **Consistency across platforms:** Students identify agreements and discrepancies among responses generated by different platforms. Discrepancies are not errors to be disregarded; they serve as verification tasks and learning opportunities. When platforms present conflicting information, students must identify the disputed claims and resolve them against independently verified sources before finalizing their comparison table.

Before importing findings into the comparison table, students should use a two-pass prompting approach for each platform. In the first pass, students submit the prompt without word or length restrictions, allowing the platform to generate its most complete and detailed response and ensuring that no analytical depth is lost to premature output constraints. In the second pass, students submit a follow-up prompt instructing the platform to summarize its previous response, retaining only the most important and relevant takeaways in approximately 50–100 words. This condensed summary is then entered into the comparison table. The two-pass approach ensures that students engage with the full analytical content of each platform's response before deciding what to document, preventing superficial comparison based on truncated outputs.

The evaluated implementation included cross-platform comparison, independent source verification, and documented follow-up prompting. The three safeguards discussed next, including repeated sampling, alternative prompt framings, and comparison against a course-concept checklist, are proposed extensions for instructors and were not evaluated as separate components of the present study.

Recommended safeguard 1: Run-to-run variation. Run-to-run stochasticity remains a limitation because LLM outputs can vary even when the same prompt is

submitted repeatedly. A single regeneration or resubmission in a fresh thread can help students detect an obvious instability: A changed response signals that the original difference may have reflected sampling variation. A repeated response, however, does not establish a stable platform characteristic, particularly when users cannot observe or control temperature and related sampling parameters, and pockets of output variability persist even in newer and more capable models (Zhou et al. 2024). The repeated element remains only a candidate platform difference that must be adjudicated against course materials and independent sources. Instructors seeking a stronger diagnostic may use repeated sampling and modal aggregation when feasible, following the self-consistency approach developed for reasoning tasks (Wang et al. 2023), but this refinement was not part of the evaluated implementation and is not required by the GLF.

Recommended safeguard 2: Sycophancy and framing sensitivity. Neutral and two-sided framing can reduce overtly leading language but cannot eliminate sycophancy, the tendency of models to align with a user's stated or implied position, which has been documented in educational settings where leading assistants affirmed misconceptions embedded in practice-oriented prompts (Richter et al. 2025). Determining whether a prompt is genuinely neutral is itself difficult. Recommended safeguards include eliciting an initial assessment before revealing the student's position, posing opposing framings in separate fresh threads, and comparing responses under alternative prompt wordings. These procedures can reveal sensitivity to framing, but they do not prove that a model is unbiased or free of agreement-seeking behavior. They are therefore presented as recommendations for instructors and as priorities for future evaluation rather than as components tested in the present study.

Recommended safeguard 3: Common omissions across platforms. Cross-platform agreement bounds what the platforms produced, not what is true. Therefore, a category omitted by every platform remains invisible to a purely cross-platform check. As an AI-independent safeguard, instructors should ask students to compare the union of platform-identified categories against the concept checklist in the module guidelines and course materials. When a course-covered concept is absent from all responses, students should investigate it independently, document whether it applies, and note that the addition came from the course-material check rather than from a platform.

After completing the evaluation, students synthesize the most accurate, well-supported elements from each platform's response into a coherent answer. Follow-up prompts, such as asking a platform to expand on a point, reconsider a conclusion, or explain a discrepancy, are encouraged as part of this iterative refinement

process. For example, if GAI platforms provide inconsistent responses to a question, students should follow up by asking the platform that provided the conflicting answer about the discrepancy exposed by another platform, and request elaboration and reasoning for it. Students are required to document the process they followed to reach a final response for each question, including a conversation record for each AI platform used, in an instructor-approved format, and providing the condensed second-pass response in the comparison table.

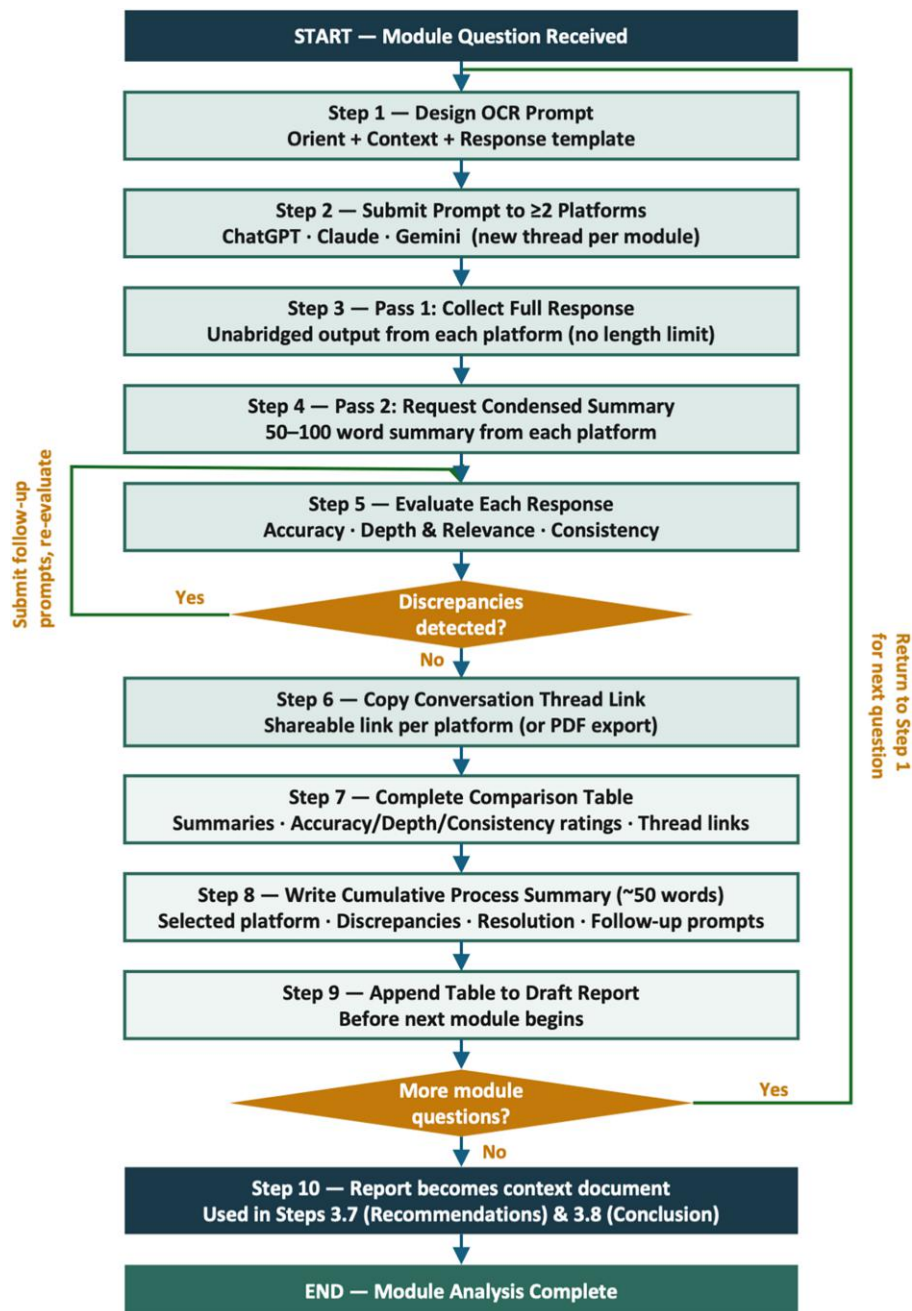
A process summary of approximately 50 words at the bottom of the table should identify which platform's response was selected as the basis for the final answer, what discrepancies were encountered, how they were resolved, and what follow-up prompts were submitted. When shareable conversation links are used as the conversation record, the final report becomes a hyperlinked reference document, with each synthesized paragraph traceable to the AI exchange that informed it, enabling both retrospective review by the student and targeted verification by the instructor. Figure 2 outlines the iterative workflow students follow during Step 3.5, illustrating how each platform's output is collected, condensed, and systematically evaluated across three dimensions before the completed comparison table is appended to the draft report.

Once finalized, the completed comparison table for each module is appended directly to the relevant section of the draft report before the next module begins, ensuring the written report grows as a continuous, cumulative record of the student's analytical engagement throughout the semester. This incrementally built report later serves as the uploaded context document for Steps 3.6 and 3.7, enabling a GAI tool to generate recommendations and a draft conclusion grounded in the student's own verified findings.

3.6. Developing Recommendations

Before initiating this step, students must consolidate all completed module analyses and comparison tables into a single draft report. Students then upload their draft report to a new thread in a GAI tool and submit a structured recommendation prompt (illustrated in Section 4.4). They develop a Recommendations section that proposes four to six actionable strategies to improve the focal company's operations, drawing on insights from all modules. The GAI tool serves as a strategic brainstorming partner, generating candidate recommendations that students must then critically evaluate through either retaining, revising, or rejecting each suggestion with written justification.

A written retain/revise/reject (R/R/R) table must be added to the draft report, accompanied by the conversation record for that session. The table should list all AI claims within a single column, accompanied by

Figure 2. Step 3.5 Student Workflow: Evaluating, Comparing, and Refining GAI Outputs

Note. Completed once per module question, and the loop repeats for each additional question within the module.

the decision (retain, revise, reject), justification, and module source in the remaining columns. Students are also required to defend their recommendation decisions analytically during the oral presentation.

3.7. Writing the Conclusion

The conclusion is the component of the written report and oral presentation in which students demonstrate their capacity to synthesize, evaluate, and move

beyond the GAI outputs generated across the semester (Dunlosky et al. 2013, Eaton 2023).

Students upload their compiled draft report to one GAI tool and submit a structured conclusion prompt asking the platform to (a) identify the most significant operational insights about the focal company that emerged across all modules; (b) assess the comparative capabilities and limitations of the GAI platforms used throughout the project; and (c) identify any remaining gaps or unresolved questions in the analysis.

Then, students critically review the AI-generated response and complete two R/R/R tables, one for the operational insights (prompt a) and one for the platforms assessment (prompt b), following the same format used in Step 3.6. Each table must include a minimum of three rows. For any gaps or unresolved questions identified under prompt (c), students must either locate additional verified sources to address them or explicitly acknowledge the limitation in the written conclusion rather than accepting unverifiable AI claims.

Students submit (1) their conclusion with the two completed R/R/R tables appended and (2) the conversation record from the GAI session. During the oral presentation, students must explain and defend their R/R/R decisions, with specific reference to module-level findings.

3.8. Compiling and Presenting the Final Report

Having completed the module analyses, recommendations, and conclusion, students compile these components into the final written report and oral presentation. This structure mirrors the cognitive progression of the framework, from descriptive context building, through analytical engagement, to strategic synthesis and individual reflection.

The written report serves as a process documentation record, integrating comparison tables, conversation records, source verification records, two-pass prompting logs, and R/R/R decisions from the Recommendation and Conclusion sections into a structured, semester-long audit trail of the student's analytical engagement. The report provides the evidentiary foundation for the oral presentation, allowing students to review their documented process and findings as they prepare their slides and talking points.

The oral presentation follows the same four-part structure, ensuring alignment between written and oral deliverables. The oral presentation serves two purposes: it discourages passive reliance on AI-generated text by requiring students to explain and defend their analyses in their own words (Eaton 2023, Sullivan et al. 2023), and it develops professional communication competencies, including structured argumentation and evidence-based presentation skills that are increasingly expected in AI-augmented workplaces (Mehmet et al. 2025). In an oral presentation, although the module analyses may be presented at a summary level, students are expected to discuss the Recommendations and Conclusion sections in analytical depth, explaining what the AI initially proposed, what they independently added or modified, and the reasoning behind each decision. Discussing the mechanics of the comparison table is optional and at the student's discretion, as including it may exceed the presentation time specified by the instructor. Students are encouraged to use GAI

tools to create visual aids for their presentations, making them more engaging for the audience.

4. Framework Application: A Walmart Case Example

This section illustrates how students should engage with each step of the GLF through a concrete OM topic. All prompts, outputs, and evaluation decisions presented in this section are author-generated illustrations of how the GLF steps operate in practice; they are not drawn from student submissions and are intended solely to demonstrate the framework's workflow and expected level of analytical engagement. The illustrative outputs in Online Appendices C and D were generated on September 1, 2025, using GPT-5 (OpenAI), Claude Sonnet 4 (Anthropic), and Gemini 2.5 (Google), with deep search enabled on each platform; these represent the model versions available at the time of course implementation. Walmart serves as the focal company throughout this example. The implementation proceeds through five subsections, each corresponding to one or more GLF steps discussed in Section 3: Section 4.1 demonstrates the development of the company introduction (Step 3.1); Section 4.2 presents two fully developed module-level prompts (Step 3.3), Step 3.2 is preparatory and not separately illustrated; Section 4.3 evaluates and compares GAI outputs across three platforms (Step 3.5); Section 4.4 demonstrates the development of recommendations (Step 3.6); and Section 4.5 demonstrates the two-stage R/R/R evaluation for the conclusion and the optional gap-check prompt (Step 3.7). Two project questions drawn from distinct OM modules are used throughout: (1) competitive priorities (order qualifiers and order winners) and (2) demand management strategies (promotional pricing).

4.1. Developing the Company Introduction (GLF Step 3.1)

The following prompt was used to generate a draft introduction for Walmart.

Company Introduction Prompt: "You are an expert in business and industry analysis. Please draft a concise company overview for an undergraduate Operations Management case report on Walmart. Include the following components: (1) a brief history of the company; (2) its mission and vision; (3) its primary products and services; (4) its industry segment and key competitors; (5) its approximate market position, such as size, market share, or role in the industry; and (6) any notable strategic priorities relevant to operations management. Use clear headings and short sentences suitable for an undergraduate business student. After each factual claim, do an online search for a publicly accessible source (please verify it), such as an annual report, the

company website, or an industry report, that can be used to verify the information.”

In this illustrative example, the AI-generated draft was verified against Walmart’s 2023 Annual Report, the company’s official website, and publicly available industry reports, confirming the accuracy of key claims regarding revenue, competitive positioning, and strategic priorities. Claims that could not be verified through independent sources were revised or removed. The final company introduction was then rewritten in plain, student-appropriate language before being incorporated into the sample report. This verification process mirrors the source-checking practices the GLF requires of students in all subsequent modules and establishes information literacy as a foundation before analytical work begins (Kasneji et al. 2023, Thüs et al. 2024)

4.2. Designing Module-Level Prompts for GAI-Assisted Analysis (GLF Step 3.3)

Unlike the company introduction prompt in Section 4.1, which targets factual retrieval and verification, the module-level prompts developed here are designed to elicit course concept-aligned analytical responses grounded in uploaded instructional materials. Each prompt addresses a specific OM question drawn from the module-specific project completion guidelines described in Section 3.2. The prompts require the GAI platform to reason within the scope of course terminology, analytical frameworks, and the depth expected for the assignment. This distinction reflects the GLF’s scaffolded design: factual context is established first (Step 3.1), and analytical engagement follows module by module (Steps 3.3–3.5). The following two prompts illustrate the three-element OCR structure described in Section 3.3.

Prompt 1:

[Orient]: You are an expert in operations management.

[Context]: Please answer the following question using the concepts covered in the uploaded course materials and the information available about Walmart.

Question: Identify Walmart’s order qualifiers and order winners from the end-users’ perspective.

Please ensure your response aligns with the educational materials uploaded for this chapter (e.g., lecture notes, readings, or summaries) and incorporates references where applicable. When using your deep search or Internet retrieval features, consult only scholarly, peer-reviewed journal articles, academic textbooks, and official company documents such as annual reports or regulatory filings. Do not use blogs, opinion pieces, news editorials, or marketing and promotional materials as sources.

[Response]: Use short sentences that are easy for a freshman undergraduate student to understand. Do not impose any word or sentence limits. Provide a thorough and fully supported response for each entry. Use the following exact structure for your output:

[Example Output Structure]

Order Qualifier: [Name of Qualifier]

- **Definition:** [Define the order qualifier based on uploaded course notes]

- **Applicable Product/Service Types:** [Explain why customers expect this criterion as a baseline threshold, and how Walmart’s performance on it compares to key competitors]

- **Walmart Application:** [Explain how Walmart achieves this with supporting evidence]

- **Source:** [Reference].

Prompt 2:

[Orient]: You are an expert in operations management.

[Context]: Please answer the following question using the concepts covered in the uploaded course materials, as well as the information available about Walmart.

Question: Evaluate the application of promotional pricing as one of the demand management strategies at Walmart. Demand for which types of products can be managed by this option?

Please ensure your response aligns with the educational materials uploaded for this chapter (e.g., lecture notes, readings, or summaries) and incorporates references where applicable. When using your deep search or Internet retrieval features, consult only scholarly, peer-reviewed journal articles, academic textbooks, and official company documents such as annual reports or regulatory filings. Do not use blogs, opinion pieces, news editorials, or marketing and promotional materials as sources.

[Response]: Use short sentences that are easy for a freshman undergraduate student to understand. Use the following exact structure for your output. Do not impose any word or sentence limits. Provide a thorough and fully supported response for each entry:

[Example Output Structure]

Demand Management Strategy: [Name of Strategy]

- **Definition:** [define the demand management strategy based on uploaded notes]

- **Applicable Product/Service Types:** [identify which types of products at Walmart this strategy applies to]

- **Walmart Application:** [explain how Walmart implements this strategy in practice]

- **Source:** [Reference]

Both prompts, along with the uploaded course materials, were submitted to ChatGPT, Claude, and Gemini with deep search enabled and a new conversation thread for each; condensed platform responses are provided in Online Appendix C (Prompt 1: order qualifiers and order winners) and Online Appendix D (Prompt 2: promotional pricing). The course-material summaries uploaded as context for the two prompts are reproduced in Online Appendices A and B, respectively.

4.3. Evaluating and Comparing GAI Outputs (GLF Steps 3.4 and 3.5)

With outputs collected from all three platforms as described in Section 4.2, the evaluation phase proceeds

across the three dimensions specified in Step 3.5: accuracy, depth and relevance, and consistency across platforms. For each prompt, platform responses are documented side by side in a structured comparison table (Step 3.5), and any discrepancies are treated as verification tasks rather than errors to be dismissed. The following subsections illustrate this process for both Prompt 1 (order qualifiers and order winners) and Prompt 2 (promotional pricing as a demand management strategy). Summarized platforms responses are provided in Online Appendix C for Prompt 1 and Online Appendix D for Prompt 2.

4.3.1. Evaluation of Responses to Prompt 1. Table 1 displays the order qualifiers identified by each GAI platform in response to Prompt 1. The detailed responses from all three platforms are provided in Online Appendix C.

Table 1 shows that all three platforms identified consistent product quality and product variety/availability as order qualifiers. This agreement is an initial comparison result, not evidence that the labels are necessarily accurate or stable, so students must still verify them against course definitions and independent sources. The remaining qualifiers illustrate platform-specific differences that require investigation. Claude did not identify on-time delivery, whereas ChatGPT and Gemini did. Rather than accepting the majority view, students should treat the difference as a verification task and ask Claude to reason through whether on-time delivery qualifies in Walmart’s operating context. GPT-5’s inclusion of brand reputation illustrates another possible error: conflating a marketing-level competitive advantage with the operational threshold criteria used to define order qualifiers in the course materials. Students should identify this conceptual mismatch and verify the distinction before incorporating the finding into the final report. This evaluation process, rather than platform agreement itself, is one of the GLF’s intended learning activities. Table 2 displays the order winners identified by each platform.

All three platforms identified low-cost operations and volume flexibility as order winners. Only two platforms identified convenience/multichannel access, variety across categories, and delivery speed. Students

Table 1. Walmart Order Qualifiers Reported by Each GAI Platform

Category	ChatGPT	Claude	Gemini
Consistent product quality	*	*	*
Product variety/availability	*	*	*
On-time delivery	*		*
Customer service standards		*	
Store accessibility		*	
Brand reputation	*		

Table 2. Walmart Order Winners Reported by Each GAI Platform

Category	ChatGPT	Claude	Gemini
Low-cost operations	*	*	*
Convenience/multichannel access	*	*	
Volume flexibility/handling peaks	*	*	*
Variety across categories		*	*
Delivery speed	*	*	

should treat these differences as verification tasks: identify the contested claim, substantiate it with independent sources such as company filings or reputable industry reports, and document how the evidence supports or refutes it. A student might, for example, question whether product variety remains an order winner for Walmart when major competitors offer comparable variety, suggesting that it may instead function as an order qualifier. Differences can also reflect prompt interpretation, retrieval coverage, model training, or sampling variation.

Illustrative Extension. These checks were not evaluated in the two course sections reported in this study. An instructor could apply the recommended safeguards in Step 3.5 by asking a student to resubmit the same opening prompt to Gemini in a fresh thread. A changed response would reveal obvious run-to-run instability, whereas a repeated omission would remain only a candidate discrepancy requiring external verification and would not constitute evidence of a stable platform characteristic. For the Table 1 discrepancy involving Claude’s omission of on-time delivery, the instructor could elicit an initial assessment in a fresh thread before revealing the other platforms’ positions or compare responses generated under opposing prompt framings. These checks supplement rather than replace independent source verification. Completeness should also be assessed against course materials rather than platform consensus. Online Appendix A lists nine competitive priorities, so a student noticing that no platform assessed development speed should evaluate that omission independently and document why it was or was not applicable.

4.3.2. Evaluation of Responses to Prompt 2. Table 3 compares the key points identified by each platform in response to Prompt 2 on promotional pricing. A summary of the full platforms responses is included in Online Appendix D.

All three platforms identified promotional pricing as a temporary strategy to stimulate demand during slack periods and for specific product categories. Seasonal and impulse products were unanimously highlighted as primary targets. However, notable differences emerged across platforms: ChatGPT uniquely noted that constant-demand necessities such as milk and

Table 3. Key Points on Promotional Pricing Identified by ChatGPT, Claude, and Gemini

Category (key points)	ChatGPT	Claude	Gemini
Definition of promotional pricing (temporary price reductions to shift demand)	*	*	*
Works for seasonal products (e.g., clothing, holiday goods)	*	*	*
Works for durable/stock-up goods (nonperishables, household essentials)	*	*	
Works for discretionary/impulse goods (electronics, toys, décor)	*	*	*
Less effective for necessities (milk, bread, items with constant demand)	*		
Inventory management benefit (avoids overstock, stockouts, waste)	*	*	*
Customer attraction and loyalty (reinforces Walmart's "everyday low price" or competitive image)	*	*	*
Example tactics (rollbacks, BOGO, ^a clearance, seasonal markdowns)	*	*	*
Strategic benefits (stable sales, cash flow, smoother customer traffic)		*	*
Link with other demand strategies (e.g., complementary products, prescheduled pickup)		*	

^aBuy One, Get One, a promotional pricing tactic where a customer purchases one item at full price and receives a second item either free or at a reduced price (e.g., Buy One, Get One 50% off).

bread are less suited to promotional pricing at Walmart, an important boundary condition worth including in the final report. Claude and Gemini offered the most comprehensive coverage of tactical examples (rollbacks, BOGO, clearance, seasonal markdowns) and strategic benefits such as stable cash flow and smoother customer traffic. Only Claude recognized the effectiveness of combining promotional pricing with complementary product strategies, a point present in the uploaded course materials but missed by the other two platforms, illustrating how cross-platform comparison can surface coverage gaps that a single-platform approach would miss. Students should document these differences in their comparison table and use them to construct a more complete, synthesized response than any single platform provided. Additionally, follow-up inquiries should be made regarding why one of the three platforms did not recognize a category that the other two did. For example, "Works for Durable/Stock-up Goods" was overlooked by Gemini, and "Strategic Benefits" was overlooked by ChatGPT.

4.4. Developing Recommendations (GLF Step 3.6)

As described in Section 3.6, after completing all module-level analyses, students upload their full draft report to a GAI tool and submit a structured prompt to generate candidate recommendations. The following prompt can be used at this stage.

Prompt 3:

"You are an expert in operations and supply chain strategy. Based on the case report uploaded here, propose 4–6 concrete recommendations to improve Walmart's operations and supply chain performance. For each recommendation: (1) link it explicitly to a specific finding from the uploaded report; (2) identify the OM concept it draws on, such as inventory management, demand management, or capacity planning; and (3) briefly explain the expected benefit and any implementation challenge. Do not restate the analysis; build on it. Use concise language and short paragraphs suitable for an undergraduate business audience."

In response to Prompt 3, the GAI tool generated six candidate recommendations: (1) expanding the use of promotional pricing for seasonal and impulse categories to smooth demand during peak periods; (2) implementing more dynamic safety stock policies for high-velocity nonperishable goods; (3) leveraging multichannel fulfillment capabilities to strengthen convenience as an order winner; (4) investing in supplier collaboration programs to reduce lead time variability; (5) adopting complementary product bundling strategies in conjunction with promotional pricing; and (6) increasing the frequency of real-time inventory audits to reduce overstock in discretionary categories.

To illustrate the evaluation process, each AI-generated recommendation was assessed against the findings in this example report. Recommendations (1), (3), and (5) were retained because they were directly supported by findings from the order qualifier, order winner, and promotional pricing analyses in Section 4.3, and could be substantiated with independent sources. Recommendation (2) was revised: Its scope was narrowed upon determining that Walmart's existing cross-docking infrastructure already addresses safety stock for most high-velocity categories, making the original suggestion overly broad. Recommendation (4) was rejected because the report contained no module-level evidence related to supplier lead times, and the GAI tool did not provide a verifiable source to support the claim. Recommendation (6) was retained with modification, linking it specifically to the inventory management benefit identified in Table 3. The summary of the evaluation process of the responses is provided in Table 4.

4.5. Writing the Conclusion (GLF Step 3.7)

As described in Section 3.7, the Conclusion begins with students uploading their full compiled draft report to one GAI tool and submitting a structured two-question conclusion prompt. To illustrate this step, the following prompt was submitted using the complete Walmart case report as context.

Table 4. Illustrative Evaluation of GAI-Generated Recommendations for the Walmart Case (GLF Step 3.6)

No.	GAI-generated Recommendation	Decision	Justification
1	Expand use of promotional pricing for seasonal and impulse categories to smooth demand during peak periods	Retain	Directly supported by the promotional pricing analysis (Table 3) and Walmart’s publicly documented rollback strategy; aligns with demand management module findings
2	Implement more dynamic safety stock policies for high-velocity nonperishable goods	Revise	Overly broad; Walmart’s existing cross-docking infrastructure already addresses safety stock for most high-velocity categories, narrowed to a targeted application in understocked regional distribution centers
3	Leverage multichannel fulfillment capabilities to strengthen convenience as an order winner	Retain	Grounded in the order winner analysis (Table 2); supported by independent data on Walmart’s curbside pickup and delivery expansion
4	Invest in supplier collaboration programs to reduce lead time variability	Reject	No module-level evidence related to supplier lead times was generated; the GAI tool did not provide a verifiable source to substantiate the claim
5	Adopt complementary product bundling strategies in conjunction with promotional pricing	Retain	Supported by the cross-platform finding in Table 3 that only Claude identified this connection; verified against Walmart’s seasonal promotional practices
6	Increase frequency of real-time inventory audits to reduce overstock in discretionary categories	Retain (revised)	Retained with modification; linked specifically to the inventory management benefit identified in Table 3, with scope narrowed to discretionary and seasonal product lines

Prompt 4:
“You are an expert in operations management and business strategy. Based on the full case report uploaded here, please: (a) identify the most significant operational insights about Walmart that emerged across all modules of this report; and (b) assess the competitive capabilities and limitations of the GAI platforms used throughout this project in supporting the analysis. For each part, use concise language and short paragraphs suitable for an undergraduate business audience.”

The GAI tool responded with two sections corresponding to parts (a) and (b) from prompt 4. Students critically reviewed each section and completed two R/R/R tables (one per prompt) before writing their conclusion. For exposition, the Walmart illustration develops two of the course’s six modules in full, the Ch. 2 module (operations strategy) and the Ch. 9 module (forecasting and demand management); In deployment, the same cycle repeats for each remaining module (Figure 2), so the compiled report evaluated at this step contains all six module analyses, and the Module Source column in Tables 5 and 6 accordingly references the two modules developed in this example. The illustrative evaluation process is shown in Tables 5 and 6.

Once both tables are completed, students synthesize the retained and revised insights into their final written conclusion in their own words. The completed tables are appended to the written report, along with the conversation record from this GAI session. After completing the R/R/R tables, students may optionally submit a follow-up prompt asking the platform to identify any operational insights, platform-specific observations, or

analytical limitations that the conclusion may have overlooked. Unlike the required R/R/R tables for parts (a) and (b) in prompt 4, this gap-check does not produce a required deliverable as it serves as a self-auditing tool students may use at their discretion before finalizing the written conclusion.

5. Assessment Design and Instructor Guidance

This section first reports the implementation and assessment procedures used in the two course sections and then presents optional adaptations for other instructional settings. Practices identified as recommended, optional, or GLF-Lite configurations were not evaluated as separate components of the present study.

The implementation context was a 300-level undergraduate OM course required of all business majors, delivered face-to-face in two sections during a 15-week semester, taught by the same instructor with identical syllabi. The course covered six OM modules, each corresponding to one textbook chapter: Operations and Value Creation (Ch. 1), Operations Strategy and Competitive Priorities (Ch. 2), Process Strategy (Ch. 3), Quality Management (Ch. 4), Forecasting and Demand Management (Ch. 9), and Inventory Management (Ch. 10). Each module consisted of an assigned textbook chapter (Krajewski and Malhotra 2024), recorded lecture-video content with embedded quizzes, in-class discussion, and, for most chapters, a short, applied video case. The course grade comprised 15% homework (lecture-video quizzes and video cases), 15% quantitative practice tests, 10% simulations, 45% distributed across three noncumulative exams, and 15%

Table 5. Illustrative R/R/R Evaluation of AI-Generated Operational Insights for the Conclusion Section of the Walmart Case (GLF Step 3.7, Prompt 4, Part a)

AI-generated insight	Decision	Justification	Module source
Walmart's operational advantage rests on a mutually reinforcing system of low-cost operations, volume flexibility, and demand smoothing	Retain	Directly supported by order winner analysis (Table 2) and promotional pricing findings (Table 3); system-level framing adds synthesis depth	Ch. 2 and Ch. 9
Any operational improvement must be evaluated for whether it preserves or disrupts this system-level balance	Retain	Consistent with the recommendation to narrow safety stock policy (Table 4, Rec. 2); adds strategic coherence to the conclusion	Ch. 2
Walmart's digital transformation strategy connects multichannel access to its subscription service model	Revise	Partially supported, multichannel access was confirmed as an order winner (Table 2), but the subscription service link was not addressed in any module analysis; narrowed to confirmed multichannel findings only	Ch. 2
Walmart faces no significant operational gaps based on the analysis	Reject	Contradicts the rejected recommendation on supplier lead times (Table 4, Rec. 4); the report identified at least one unresolved gap that the AI understated	Ch. 9

for the GLF case project. Rubric percentages reported in Online Appendix E are weights within this 15% project component and not shares of the overall course grade.

In the OM course, students received a comprehensive rubric outlining expectations for both the written report and the oral presentation. The rubric emphasized not only the thoroughness and accuracy of student engagement with each GLF step but also the quality of students' evaluation, verification, and synthesis of GAI-generated outputs. The full rubric, including performance-level descriptors for each component, is provided in Online Appendix E. Online Appendix G provides annotated exemplars of strong and weak submissions for three key deliverables, illustrating the rubric's performance-level distinctions; instructors may adapt these directly into module

guidelines or use them in an in-class calibration exercise. The rubric assessed eight components.

(1) *OM concept understanding.* The accuracy with which students identify, explain, and apply the operations management concepts relevant to each module question, including appropriate use of course terminology, frameworks, and relationships among concepts.

(2) *Prompt engineering and documentation.* The quality of the prompts submitted for each module, including adherence to the OCR structure described in Section 3.3, alignment with the module question and course materials, and clear documentation of the prompting process.

(3) *Report generation and multiplatform analysis.* The completeness and quality of each module-level analysis, including the use of at least two GAI platforms,

Table 6. Illustrative R/R/R Evaluation of AI-Generated Platform Assessment: Walmart Conclusion (GLF Step 3.7, Prompt 4, Part b)

AI-generated platform observation	Decision	Justification	Module source
ChatGPT provided the broadest coverage but included unverifiable claims (e.g., brand reputation as order qualifier)	Retain	Consistent with comparison table finding that GPT-5 conflated marketing advantages with operational threshold criteria	Ch. 2
Claude offered the most conceptually precise responses and identified complementary product bundling that other platforms missed	Retain	Supported by Table 3 finding that only Claude identified this connection, which is present in the uploaded course materials	Ch. 9
Gemini performed comparably to Claude across both illustrated modules	Revise	Partially accurate; Gemini missed "Works for Durable/Stock-up Goods" in Table 3; revised to reflect this specific coverage gap	Ch. 9
All three platforms are equally reliable for source verification	Reject	Not supported; follow-up prompting revealed platform-specific gaps in retrieval coverage; equal reliability cannot be concluded from this analysis	Ch. 2 and Ch. 9

organization of the resulting responses in the required comparison table, and development of a clear synthesized response to the module question.

(4) *Critical evaluation and refinement.* The rigor with which students assess GAI outputs for accuracy, depth and relevance, and consistency across platforms; identify and investigate discrepancies; verify substantive claims against course materials and independent sources; use follow-up prompting where appropriate; and justify how competing or incomplete outputs were resolved in the final synthesized response.

(5) *Written report synthesis.* The coherence and integration of the semester-long written report, including the extent to which module-level analyses are connected into a logically organized account of the focal company rather than presented as isolated responses to individual questions.

(6) *Recommendations, conclusion, and R/R/R tables.* The quality of the higher-level synthesis conducted in the Recommendations and Conclusion stages, including the R/R/R evaluation of AI-generated claims, the extent to which decisions are justified with specific module-level evidence, and the quality of the final written conclusion in integrating operational findings and reflecting critically on the GAI platforms used.

(7) *Conversation-record documentation.* The completeness of the instructor-reviewable records required in Step 3.4, including evidence of the submitted opening prompts, generated responses, required second-pass condensation prompts, substantive follow-up exchanges, and other documented interactions needed to verify compliance with the GLF process.

(8) *Oral presentation.* The clarity and professionalism of the presentation and, particularly, the student's ability to discuss module findings at an appropriate summary level while analytically defending the Recommendations and Conclusion in real time, including explaining what the AI proposed, what the student retained, revised, rejected, or independently contributed, and the reasoning supporting those decisions without relying on a prepared script.

These eight dimensions were assigned the following approximate weights: OM concept understanding (20%), prompt engineering and documentation (15%), report generation and multiplatform analysis (15%), critical evaluation and refinement (20%), written report synthesis (5%), recommendations, conclusion, and R/R/R tables (5%), conversation-record documentation (5%), and oral presentation (15%). These weights reflected the GLF's pedagogical emphasis on critical evaluation and independent synthesis as the primary evidence of student analytical engagement. Instructors may adjust these weights to reflect local course policies or course-level learning objectives, provided the weighting scheme continues to incentivize process documentation rather than output quality alone.

Regarding academic integrity, the rubric shifts the evidentiary emphasis from detecting AI use to verifying documented process and student understanding. Comparison tables and conversation records provide an auditable basis for checking prompts, follow-up exchanges, discrepancies, and verification activity; the R/R/R tables require students to justify decisions against findings in their own reports; and the oral presentation requires students to explain and defend those decisions in real time. These mechanisms do not make the written artifacts immune to AI assistance, but they create linked evidence that can be checked against the student's demonstrated understanding. Section 5.3 examines the verification strength and limitations of these mechanisms in greater detail.

5.1. Platform Access, Tier Equity, and Midsemester Changes

The GLF requires students to use at least two GAI platforms, which raises a practical equity concern because students may have different subscription tiers or no paid access. This section first reports the access conditions observed in the evaluated implementation and then presents three access-equity measures (verifying baseline access, anchoring assessment in process, and documenting tier heterogeneity), together with a continuity rule for midsemester platform changes.

During the reported implementation, the instructor did not identify any student who was unable to complete the required activities because of platform access. However, this observation was not based on systematic access logs or a dedicated access survey and should not be interpreted as evidence that free-tier availability ensured equitable access. Free-tier use may still be constrained by token caps, file-upload limits, temporary throttling, device or connectivity constraints, and changes in platform policy.

First, instructors adopting the GLF may verify baseline access before the semester through a short trial task and provide more than one approved route. Options include institutionally licensed accounts, student discounts or course-funded subscriptions, a required subscription treated transparently as course material when institutional policy and affordability permit, and no-cost open-access or locally hosted open-weight models that satisfy the learning objectives. Any required paid option should be accompanied by an equivalent no-cost route. If file upload is unavailable, students may paste the relevant course-material excerpt directly into the prompt. The GLF requires functional access to at least two comparison tools and not a particular vendor or paid tier.

Second, the GLF anchors assessment in process rather than output quality. The rubric in Online Appendix E emphasizes prompt structure, discrepancy documentation, verification records, and R/R/R justifications. This

design is intended to reduce, rather than eliminate, grading advantages associated with access to a more capable model: A student using constrained access can still demonstrate rigorous evaluation, whereas a student using a paid tier receives no credit merely for producing a polished answer. Instructors should nevertheless review tier declarations and intervene when usage limits systematically restrict a student's ability to complete the expected processes.

Third, tier heterogeneity should be made visible rather than assumed away. Instructors may require each comparison table to identify the platform, displayed model version, and access tier used for each column. This documentation helps instructors interpret differences with tier information in view, teaches students that model version and tier are part of the provenance of AI-assisted analysis, and creates a record of midsemester changes. It does not make access conditions equivalent, but it provides information that can help instructors identify and mitigate inequities.

A related continuity rule may be applied when a platform changes substantially or becomes unavailable during the semester. Under this rule, the student would document the change in the affected module's process summary, substitute an approved alternative platform for the remainder of the term, and not be penalized for cross-module inconsistency attributable to the change. Because each module begins in a new thread and each comparison table is a self-contained record, the substitution would not affect earlier modules; the audit trail would simply record the dated change in platform configuration.

5.2. Instructor Workload and Scalable Implementation Options

The GLF's scaffolded submission structure raises a practical feasibility question: How much additional instructor workload does a distributed, process-based project require relative to a conventional end-of-semester case project? In the implementation evaluated here, the GLF required more grading time than the conventional single-submission case project used in prior semesters. The additional effort arose primarily from the six module-level checkpoints, although distributing the work across the semester allowed corrective feedback to be provided while students could still apply it to subsequent modules. This section first reports the review procedure, implementation timeline, and observed grading time, and then presents optional grading and implementation adaptations for instructors for whom the full implementation may not be feasible.

Conversation-record review. Conversation records serve as compliance and process documentation rather than a second essay to grade. Review focuses on four elements: (1) whether the opening prompt follows the

OCR structure and matches the prompt reported in the comparison table; (2) whether course materials were provided to the platform (verified via spot-requested interface screenshots, since shared links do not display uploads); (3) whether the second-pass condensation prompt was actually issued rather than the student condensing the output themselves post hoc; and (4) whether at least one substantive follow-up exchange probes a discrepancy or requests elaboration, as claimed in the process summary. Applied as a checklist, this review was estimated to require approximately two minutes per conversation record. The rubric's conversation-record dimension assesses these items; conversation-record review is not a holistic quality judgment, which is instead made on the comparison table and synthesized answer.

Sampling protocol. In the reported implementation, conversation records were reviewed by sampling rather than exhaustively. The instructor reviewed all conversation records for Module 1, when compliance problems were most likely and corrective feedback had the greatest downstream value, and then sampled the conversation records of two to three students per section per module thereafter, together with any record whose comparison table raised a concern, such as a missing record, inconsistency between the process summary and reported outputs, or an implausibly complete resolution of a discrepancy. Since students did not know which records would be reviewed, the sampling approach retained a compliance check while reducing the number of conversation records requiring review.

Implementation timeline. In the implementation evaluated here, the project ran Weeks 2 through 14 of a 15-week semester: company selection and introduction (Step 3.1) in Weeks 2 and 3, then six module cycles (Steps 3.2–3.5) across the weeks matching each chapter's coverage, Recommendations and Conclusion (Steps 3.6–3.7) in Weeks 13 and 14, and oral presentations in Week 15. Module tables were submitted incrementally, one per module, so the instructor returned six tables during the term rather than grading one large end-of-term deliverable.

Timely formative feedback. Because module submissions were distributed across the semester, the instructor reviewed and returned each completed module before students proceeded to subsequent modules. This sequencing provided timely formative feedback on areas such as prompt quality, discrepancy evaluation, and process documentation, allowing students to identify weaknesses and apply the feedback in later module cycles.

Observed grading time. For a section of approximately 35 students, Module 1 required an estimated 4.5 hours of grading, including review of all conversation records. Once conversation-record sampling was implemented,

each of the remaining five modules required approximately 2.7 hours on average. This included grading all students' comparison tables and process summaries, together with roughly 10 minutes of conversation-record review for the two to three students sampled, with somewhat more time when flagged submissions were also reviewed. Across all six modules, module-level grading therefore totaled approximately 18 hours. The compiled final report, with module analyses already reviewed, added approximately 5 hours of grading focused on the Recommendations and Conclusion R/R/R tables and overall synthesis. Oral presentations required class time plus approximately 4.5 hours of rubric scoring. The full implementation therefore required approximately 27.5 hours of grading and scoring per section, distributed across the semester rather than concentrated at the end of the term. In the instructor's experience this was somewhat more than a conventional single-submission case project of comparable scope requires. These figures are retrospective instructor estimates rather than logged time records and are reported to characterize the workload tradeoff rather than as evidence of time neutrality.

Instructors can draw on established assessment strategies that predate widespread GenAI adoption and on more recent AI-assisted grading approaches. The former do not depend on LLMs, but they generalize to GAI-integrated settings because they reduce the volume of work requiring exhaustive review while preserving targeted evidence of student engagement. The following grading and implementation strategies are proposed adaptations for instructors who may find the full implementation infeasible. Except where noted otherwise, these alternatives were not evaluated as separate components of the present study.

Grading-model flexibility. Instructors may reduce the grading load by shifting when and by whom summative judgment is applied, without abandoning the framework's process anchoring. Under formative module grading, each comparison table is checked only for completion and protocol compliance during the term, with summative quality assessment concentrated on the final report and oral defense; because every module table reappears in the final report, quality review is relocated rather than duplicated, consistent with evidence that frequent low-stakes feedback is among the most effective assessment interventions (Black and Wiliam 1998). Under deep-sampled grading, the sampling logic applied above to conversation records is extended to the graded submissions themselves: Module 1 is fully scored for every student, and a random subset of the remaining modules (say, one or two of five) receives full scoring while the rest receive completion checks. Because students cannot predict which of the later modules will be scored, the incentive to document every module is preserved with substantially

lower grading effort, following the limited-ground-truth logic shown to sustain evaluation effort in peer assessment (Gao et al. 2019). Scoring Module 1 for all students guarantees the early corrective checkpoint that purely random selection would leave to chance, and drawing the remaining sample as submissions arrive keeps later feedback usable. Under weight redistribution, where modules are graded formatively or lightly, a larger rubric weight shifts to the oral presentation, concentrating summative assessment on the least delegable component (Section 5.3), consistent with evidence that interactive oral examinations deter misconduct while building communication skills (Sotiriadou et al. 2020). Finally, under AI-assisted first-pass grading, an approach examined in recent higher-education research, the instructor supplies an LLM with the rubric, guidelines, and anonymized submissions to draft provisional scores and feedback for human review and approval. Rubric-anchored LLM grading can approach human reliability on structured work (Flodén 2025, Yavuz et al. 2025), but documented risks of inconsistency, fabricated feedback, and unfairness make unsupervised use inappropriate; the instructor remains responsible for every grade, and institutional data-processing policies must be observed (Yan et al. 2024). Used this way, AI-assisted grading models the discipline the GLF expects from students: AI output is a first draft to be verified, not an authority to accept. All four grading approaches may be combined with the reduced configurations below, provided the scheme still rewards documented process over polished output.

Minimum viable configurations. For larger sections or courses with limited grading support, we propose two reduced configurations that preserve the framework's core mechanism while reducing workload.

GLF-Lite A (team-based): Students work in teams of three to four, producing one comparison table and one report per team, with individually assessed oral defenses; this divides written grading volume by team size while keeping the individual accountability instrument intact. **GLF-Lite B (reduced cadence):** The number of assessed module cycles is reduced from six to two (one early module graded with full feedback and one late module), with intermediate modules completed as ungraded practice or peer-reviewed drafts, or, alternatively, with the project scoped from the outset to a selected subset of course topics rather than every chapter; this can reduce full table-grading volume by approximately two-thirds while retaining the early-feedback checkpoint that supports improvement. In both configurations, the oral presentation and the R/R/R tables for Recommendations and Conclusion are retained unreduced, since these carry the primary integrity and synthesis functions of the framework. Instructors with teaching assistants can additionally

delegate the four-point conversation-record checklist, which requires no OM subject-matter judgment.

Recommended feedback carry-forward practice. Instructors may ask students to open each module submission after the first with a two-to-four-sentence note identifying the main improvement requested in the previous module and stating how the current submission implements it (for example, “Module 2 feedback noted my follow-up prompts accepted Claude’s revised answer without a source; here every discrepancy resolution cites the verifying source”). This creates a visible feedback loop across module cycles and a quick way to confirm that prior feedback was considered. This practice was not part of the evaluated implementation.

5.3. What the GLF Verifies and What It Cannot: AI-Completeness and Binding Points

A fair challenge to any GAI-integrated assignment is that the assignment’s own deliverables can be produced by the tools it teaches. We address this directly by auditing each GLF deliverable for AI-completeness and identifying where verification binds.

Steps 3.1–3.5 produce written artifacts (prompts, comparison tables, process summaries, synthesized answers), and we acknowledge that a sufficiently motivated student could generate plausible versions of each using GAI itself, including fabricated discrepancy resolutions. The framework’s integrity design therefore does not rest on the unverifiability of these artifacts. It rests on three mechanisms that are progressively harder to delegate.

First, the conversation records (Step 3.4) document the exchange as it occurred: A student claiming a follow-up exchange took place must produce the record of it, and the four-point review (Section 5.2) checks the claimed process against that record. The evidentiary strength of this mechanism varies by record type. A platform-hosted link provides stronger provenance than a standalone PDF export or screen recording because it points to a platform-hosted exchange; however, no record type alone establishes that the interaction occurred exactly as part of the student’s genuine analytical process, and an exchange could in principle be staged or recreated. The framework therefore relies on conversation records in combination with the two mechanisms that follow. Second, the R/R/R justifications (Steps 3.6–3.7) must trace to specific findings in the student’s own report; a justification generated without the report in view fails the traceability check the rubric applies. Third, and decisively, the oral defense requires students to explain in real time and in their own words what the AI proposed, what they changed, and why. Accordingly, the oral defense functions as the framework’s primary verification point, and we designed it as such: It is weighted at 15% of the project grade, it specifically targets the R/R/R

decisions (the least delegable artifacts), and a student who delegated Steps 3.1–3.5 wholesale is less likely to provide a coherent, specific defense tied to the documented analytical process.

The GLF’s claim is therefore not that each step resists delegation individually, but that the chain terminates in a real-time assessment the student must perform personally, whatever preparation preceded it, and that the intermediate documentation makes a coherent defense easier for students who engaged with the work and harder for those who did not.

What this design cannot guarantee is that engagement with the process produces measurable learning gains; the evaluation reported in Section 6 measures perceived learning only. Establishing objective gains (through rubric-score comparisons against non-GLF cohorts, pre/post assessment, or analysis of oral-defense performance relative to written-component performance) is the central priority for future work (Section 8).

6. Student Perceptions of the GLF: Survey-Based Evaluation

To assess student perception of the GLF and gather preliminary evidence of its perceived pedagogical value, a postproject survey was administered at the end of the Fall 2025 semester across two sections of an undergraduate OM course. The OM course is a 300-level core business course mainly taken by juniors. The survey protocol was reviewed and approved by the institutional review board (IRB) of the authors’ institution for both course sections prior to data collection. Participation was voluntary and anonymous, had no impact on course grades, and no identifiable personal information was gathered.

The evaluation reported here is exploratory, descriptive, and nonexperimental in intent. The design includes two intact, nonrandomly assigned sections without a comparison group, a modest sample ($n = 50$), and a newly developed instrument, and supports characterization of student perceptions, but not causal inference or inferential generalization. Accordingly, this section foregrounds descriptive results (Section 6.3), reports section-level comparisons as exploratory observations with effect sizes rather than confirmatory tests (Section 6.4) and defers assumption diagnostics to Online Appendix H. Statistical significance markers, where reported, should be read in this exploratory spirit.

6.1. Survey Design

The survey instrument consisted of 21 items organized across four components: one demographic item (academic major), one background item (prior familiarity with AI tools), 14 Likert-scale items rated on a five-point scale ranging from one (strongly disagree) to five

(strongly agree), and five open-ended reflection questions. The full list of survey items is provided in Online Appendix F. The survey was developed specifically for this study, with no prior validation and served as an exploratory instrument. Responses from open-ended questions were analyzed inductively using open coding, in which themes were identified through iterative review until saturation was reached.

Items were assigned to constructs based on conceptual alignment with the corresponding pedagogical principle; the full item-to-construct mapping and representative items are detailed in Table 7. These groupings are conceptually assigned composites rather than empirically derived factors: No factor analysis was conducted, and the instrument has no validation history. Throughout Section 6, we therefore use the grouping labels as organizational shorthand for sets of face-valid items and not as claims about validated latent constructs.

6.2. Participants

The survey was distributed to students enrolled in two sections of the same undergraduate OM course, taught by the same instructor with identical syllabi, assignments, rubrics, and grading criteria. Section 1 yielded 26 responses from 36 enrolled students (response rate: 72%), and Section 2 yielded 24 responses from 35 enrolled students (response rate: 69%), for a combined sample of $n=50$. All responses were completed anonymously.

The sample was demographically diverse across business majors. Nearly 88% of respondents reported majors outside of operations management or supply chain management, with approximately half from accounting and finance programs. This demographic profile is representative of the broader business student population that typically enrolls in undergraduate OM courses as a required core course, suggesting that the GLF’s perceived usefulness extends beyond students

with a direct interest in operations or supply chain careers.

Students in both sections reported comparable levels of prior familiarity with generative AI tools. Across the full sample ($n=50$), the most common response was moderately familiar (50.0%), followed by slightly familiar (24.0%), very familiar (22.0%), and not familiar at all (4.0%). Section-level distributions were nearly identical: 73.1% of Section 1 ($n=19$ of 26) and 70.8% of Section 2 ($n=17$ of 24) reported moderately or very familiar, whereas 26.9% and 29.2%, respectively, reported slightly or not familiar at all. Both sections showed a similar median familiarity level (moderately familiar). Statistical tests confirmed no significant between-section differences in prior AI familiarity ($\chi^2(3) = 0.372$, $p = 0.946$) or major field distribution ($\chi^2(6) = 4.327$, $p = 0.633$), supporting the decision to pool responses for the primary analysis and to treat the section-level difference in active learning and AI use as attributable to unmeasured cohort factors rather than to systematic pre-existing differences.

6.3. Pooled Descriptive Results

Table 8 presents the pooled construct means weighted by section sample size, along with the corresponding GLF steps and theoretical grounding for each survey construct. All three construct means exceeded the scale midpoint (3.0), ranging from 3.93 to 4.08. Cognitive load/scaffolding received the highest rating ($M = 4.08$), indicating that students found the module-by-module structure manageable and both the rubric and framework guidance clear. The result provides preliminary support for the GLF’s grounding in cognitive load theory (Sweller 1994): Decomposing the semester-long project into bounded, modular tasks reduced the perceived burden of a cognitively complex assignment. This construct comprised only three items focused on perceived manageability and rubric clarity; the narrow scope of measurement may partly explain its higher pooled

Table 7. Conceptual Grouping of Survey Items into Composites and Measurement Categories

Construct	Measurement dimensions	Representative survey items
Cognitive load/scaffolding	Workload manageability; rubric clarity; framework guidance	“The module-by-module structure made the project workload manageable”; “The rubric clearly explained what was expected at each step”; “The framework guidance helped me understand how to proceed”
Active Learning and AI use	Prompt engineering skills; cross-platform comparison; source verification; cross-output evaluation	“Designing the prompts helped me engage more deeply with the course content”; “Comparing outputs across platforms improved my ability to evaluate AI-generated information”; “Verifying AI outputs against independent sources strengthened my analytical skills”
Constructive alignment	OM concept understanding; theory-practice linkage; oral communication; applied OM strategy knowledge	“The project helped me better understand how OM concepts apply to real companies”; “Presenting my findings orally helped me consolidate what I learned”; “The project connected course theory to real-world business practice”

Table 8. Pooled Descriptive Statistics Across All Respondents ($n = 50$)

Item grouping	Items (N)	Pooled mean	Pooled standard deviation	Item range	GLF steps
Cognitive load/scaffolding	3	4.08	0.85	4.02–4.12	3.2 Guidelines, Rubric (Supplemental Appendix E), Workload
Active learning and AI use	7	3.94	0.93	3.76–4.12	3.3–3.5 Prompting, Comparison, Verification, AI Literacy
Constructive alignment	4	3.93	0.92	3.90–3.94	3.2 OM Learning, 3.8 Oral Presentation

Note. Pooled means are weighted by section sample sizes ($n_1 = 26$, $n_2 = 24$).

rating relative to the broader active learning and AI use construct.

Active learning and AI use received a pooled mean of $M = 3.94$ and the largest between-section effect size in the survey ($g = 0.642$; Table 9). This construct captures students' perceived development of platform comparison, prompt engineering, source verification, and cross-output evaluation skills, which together form the behavioral core of the GLF's active learning design (Freeman et al. 2014, Farrokhnia et al. 2026).

Constructive alignment received a pooled mean of $M = 3.93$, reflecting students' perceived gains in OM concept understanding, theory-practice linkage, applied OM strategy knowledge, and oral communication skills. The moderate cross-sectional difference is consistent with the argument of Biggs et al. (2022) that constructive alignment depends on alignment between task design, learning activities, and assessment, and suggests that minor variations in how instructors frame the connection between the case project and course concepts can meaningfully influence perceived learning outcomes. The qualitative themes reported below further reinforce and extend each of these findings.

6.4. Exploratory Section-Level Comparisons

Section-level differences were examined with Welch t -tests, with nonparametric checks and assumption diagnostics reported in Online Appendix H; conclusions were unchanged across methods. Given nonrandom assignment and the exploratory instrument, these comparisons are reported to characterize implementation variation between sections, not to test hypotheses. Table 9 presents the results.

Section 2 scored higher than Section 1 across all three constructs; $|t|$ values were reported as absolute values accordingly. Welch t -tests revealed a statistically significant section-level difference for active learning and AI use ($|t|(42.4) = 2.345$, $p = 0.024$, $g = 0.642$), and Section 2

scored higher ($M = 4.24$) than Section 1 ($M = 3.66$). Differences for cognitive load/scaffolding ($|t|(47.7) = 1.851$, $p = 0.070$, $g = 0.512$) and constructive alignment ($|t|(46.0) = 1.689$, $p = 0.098$, $g = 0.465$) did not reach statistical significance, although both effect sizes fall in the medium range, suggesting potentially meaningful difference in implementation experience between sections.

Despite this variation, all three construct means exceeded the scale midpoint in both sections, supporting the decision to pool responses for the primary analysis. Internal consistency was high for all three item groupings ($\alpha = 0.824 - 0.955$), supporting their use as composite scores for descriptive summary; we note that alpha indexes interitem consistency and not construct validity, and the high alpha for the seven-item active learning and AI use grouping may partly reflect item redundancy. Interpretation of the composites as measures of the named theoretical constructs awaits formal validation (Section 8).

6.5. Qualitative Reflections

The five open-ended survey questions invited students to reflect on their experiences with the GLF in their own words. Thematic review of responses provided deeper context for the quantitative findings, revealing four recurring themes in students' GLF engagement.

6.5.1. Theme 1: Active Processing and Translation. Students reported that designing prompts and summarizing AI outputs forced them to actively process the course material rather than passively read it. The process required them to tailor their questions to the specific company and deeply engage with the content. As one student emphasized, the cognitive act of putting AI outputs into their own words proved particularly effective for retention:

Table 9. Section-Level Comparison of Construct Means

Construct (theory anchor)	Items (N)	α	Section 1: Mean (standard deviation)	Section 2: Mean (standard deviation)	$ t (df)$	p	g
Cognitive load/scaffolding	3	0.824	3.87 (0.89)	4.31 (0.76)	1.851 (47.7)	0.070	0.512
Active learning and AI use	7	0.955	3.66 (1.05)	4.24 (0.66)	2.345 (42.4)	0.024	0.642
Constructive alignment	4	0.932	3.72 (1.01)	4.14 (0.76)	1.689 (46.0)	0.098	0.465

“Summarizing the AI prompts in my own words really helped what I was reporting stick to my brain ... I find myself comfortable in explaining what I learned to family when they ask.”

6.5.2. Theme 2: Value and Limits of Cross-Platform Comparison. The requirement to use multiple GAI platforms (GLF Step 3.5) was frequently cited as the most valuable aspect of the project, although it also surfaced practical limitations. Students appreciated gaining exposure to “different perspectives” and understanding the strengths and weaknesses of how GAI platforms research and compile information. One student noted a lasting impact on how they envision using these tools professionally:

“I thought the comparison of the two LLM models was the best part of the project. It made me consider what each model excelled at, and it made me warm up to the idea of using LLMs for practical research in my day-to-day life.”

However, several students noted that leading platforms such as ChatGPT and Gemini often produced highly similar responses, leading some to feel that the comparison occasionally felt redundant:

“Most of the time it gave me the same answer, but in different ways of saying or explaining it ... so sometimes I felt like it was a slight waste of time using both.”

This concern about platform convergence is addressed as a design limitation in Section 8.

6.5.3. Theme 3: Oral Presentation as a Mechanism for Internalization. Although the oral presentation requirement was designed to assess communication skills and deter passive reliance on AI-generated text, students explicitly recognized it as a tool for deeper internalization. Verbalizing AI-generated findings required a higher degree of comprehension than merely submitting a written report. One student reflected as follows:

“It was useful for making me look over all the information on the project and extract the main takeaways from comparing the two AI models. It helped me find the reliable information from the responses.”

6.5.4. Theme 4: Structured Prompt Guidance as a Cognitive Scaffold. Although the iterative nature of prompt engineering is a well-documented challenge in educational AI contexts, arising in part from the absence of standardized prompting frameworks tailored to higher education (Carrasco-Sáez et al. 2025, Lee and Palmer 2025), students in this implementation recognized the value of the GLF’s structured three-element OCR prompt template in navigating that complexity. Rather

than approaching each module as an open-ended interaction with an unfamiliar tool, students reported that the prompt structure helped them focus their questions, connect AI outputs to the specific company context, and engage more meaningfully with the course material. As one student noted:

“Prompting the generative AI tools helped me the most because it made me tailor the question to the specific company and understand the information about it better.”

6.6. Summary and Implications for GLF Design

Together, the section-level comparisons and pooled results converge on three implications for GLF design, indicating practically meaningful variation in implementation experience between sections. A post hoc comparison of background characteristics confirmed that the two sections did not differ significantly in prior AI familiarity ($\chi^2(3) = 0.372$, $p = 0.946$) or academic major distribution ($\chi^2(6) = 4.327$, $p = 0.633$), thus ruling out pre-existing demographic differences as a confound. The section-level difference in active learning and AI use may therefore reflect unmeasured cohort-level factors, including section scheduling, peer group dynamics, or naturally occurring engagement variation, which is consistent with the known limitations of small, intact classroom samples (Shadish et al. 2002). This finding reinforces the importance of replication across multiple cohorts in future work.

7. Conclusion

As GAI reshapes both the modern workplace and higher education, business schools face a dual mandate: safeguarding the integrity of the learning process while equipping students with the AI literacy required for future careers. This study addressed that challenge by developing the GLF, a structured, step-by-step pedagogical approach that uses GAI platforms to augment case-based projects in undergraduate business courses.

Rather than treating GAI as a substitute for independent research and reasoning, the GLF is designed to inhibit cognitive offloading and promote active, critical engagement with course material. By requiring students to engineer effective prompts from course materials, systematically compare outputs across multiple AI platforms (e.g., ChatGPT, Claude, and Gemini), and reconcile discrepancies against independent sources, the framework is designed to support the development of evaluative habits of mind.

Preliminary survey evidence supported the framework’s perceived pedagogical value: Students reported that the scaffolded structure was manageable, cross-platform comparison deepened evaluative skills, and the R/R/R tables and oral presentation requirements

supported genuine synthesis rather than passive acceptance of AI outputs.

The structural features of the GLF most readily transferable across disciplines include the OCR prompt template, the two-pass prompting protocol, the cross-platform comparison table, and the R/R/R evaluation structure; the module-by-module scaffolding tied to sequential OM concept coverage and the oral defense of R/R/R decisions are more tightly coupled to the case-based project format typical of OM and other applied business courses.

8. Limitations and Future Research

Although the GLF provides a viable approach for integrating AI into OM coursework, several limitations warrant acknowledgment.

First, the framework's reliance on cross-platform comparisons is subject to the rapid, continuous evolution of GAI tools. As underlying models improve or converge in data training, discrepancies between platforms may narrow, potentially reducing the cross-platform variation the framework relies on to stimulate critical comparison. Instructors must remain adaptable, for example, by shifting comparison targets from competing platforms to AI outputs versus student-generated analyses or by introducing domain-specific or locally verified data sets as reference points. Platform access is subject to the same volatility: Free-tier availability does not ensure sufficient token allowances, file-upload capability, or uninterrupted access, and the present study did not systematically log access failures (see Section 5.1). Future research should measure whether institutional, discounted, open-access, or locally hosted options provide equitable participation as models, interfaces, and access policies evolve.

Second, the evaluation in this study was based on student-perceived learning outcomes rather than objective performance metrics. Although perceived effectiveness is a critical indicator of student engagement and instructional design clarity, future empirical studies should employ quasi-experimental designs to measure objective knowledge retention by comparing exam scores or rubric performance between cohorts using the GLF and those using traditional case-based instruction. Consequently, claims throughout this paper about critical thinking, AI literacy, and related learning outcomes should be read as statements of instructional design intent and student-perceived experience rather than demonstrated learning gains; Section 5.3 discusses which framework mechanisms can and cannot be verified from the artifacts students submit.

Third, this framework was evaluated in a single undergraduate OM course, although the participants' diverse demographic profile (88% non-OM majors) suggests potential for broader applicability. Future

research should explore the implementation of the GLF in other business disciplines, such as finance, marketing, or business analytics, to evaluate how different subject matters and quantitative requirements impact the effectiveness of multiplatform AI evaluation.

Fourth, the small per-section sample sizes ($n_1 = 26$, $n_2 = 24$) limited the statistical power of between-section comparisons and the sensitivity of normality tests; Shapiro-Wilk flagged nonnormality in five of six distributions, although the Welch *t*-test is robust to moderate nonnormality at these sample sizes (Lumley et al. 2002), and nonparametric Mann-Whitney tests confirmed identical significance conclusions across all three constructs. The two intact classroom sections also preclude random assignment, meaning the study is best characterized as an exploratory, nonexperimental evaluation conducted in two intact course sections rather than a controlled experiment (Shadish et al. 2002). Future implementations should replicate the GLF across multiple instructors, institutions, and course contexts with larger randomly assigned cohorts to establish more generalizable evidence of effectiveness.

Fifth, the survey instrument was developed specifically for this implementation and has not undergone formal psychometric validation. This is a substantive limitation: The construct structure, item wording, and factor assignments reflect the authors' theoretical judgments rather than an independently validated scale. Future research should subject the instrument to confirmatory factor analysis across diverse course contexts before drawing generalizable conclusions about construct validity. For this reason, the composite labels used in Section 6 should be read as organizational shorthand rather than validated construct measurements.

Sixth, the run-to-run stochasticity and sycophancy discussed in Section 3.5 bear directly on what the cross-platform comparison at the center of the GLF can support, because the framework's process records cannot distinguish either effect from substantive analytical difference. The safeguards identified there were offered as recommendations and were not tested as separate components of the evaluated implementation. Future research should evaluate them directly, including repeated-sampling protocols, alternative prompt wordings, fresh threads with opposing framings, and initial assessments obtained before a student's position is disclosed.

Seventh, the framework carries a workload cost that faculty should plan for. As reported in Section 5.2, the evaluated implementation required approximately 27.5 hours of grading and scoring per section, which in the instructor's experience was somewhat more than a conventional single-submission case project of comparable scope requires. That total already reflects the sampled conversation-record review described in Section 5.2; it does not reflect any reduction in summative scoring,

because every module's comparison table received full rubric assessment. The grading and implementation strategies described there are proposed adaptations that were not separately evaluated, so instructors should budget for the additional time rather than expect its distribution across the term alone to absorb it.

The GLF represents a foundational step toward a broader pedagogical shift in AI-integrated business education. By providing educators with a replicable, adaptable structure, this framework demonstrates how business programs can move beyond simply reacting to the threat of AI and instead harness it to cultivate the critical thinking and technological adaptability that will define the next generation of management professionals.

Acknowledgments

The authors thank the students who generously shared reflections on their experiences with the proposed framework and the associate editor and anonymous reviewers for constructive feedback and valuable suggestions that significantly enhanced the quality and clarity of this study.

Generative AI tools were used in one capacity in this manuscript: the illustrative prompts in Section 4 and the platform outputs in Online Appendices C and D were generated on September 1, 2025, using GPT-5 (OpenAI), Claude Sonnet 4 (Anthropic), and Gemini 2.5 (Google). These outputs serve as author-constructed illustrations of the GLF workflow and are not drawn from student submissions.

References

- Al-Kfairy M, Mustafa D, Kshetri N, Insiew M, Alfandi O (2024) Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics* 11(3):58.
- Bai Y, Wang S (2025) Impact of generative AI interaction and output quality on university students' learning outcomes: A technology-mediated and motivation-driven approach. *Sci. Rep.* 15(1):24054.
- Batista J, Mesquita A, Carnaz G (2024) Generative AI and higher education: Trends, challenges, and future directions from a systematic literature review. *Information* 15(11):676.
- Bernik A, Radošević D, Čep A (2025) A comparative study of large language models in programming education: Accuracy, efficiency, and feedback in student assignment grading. *Appl. Sci.* 15(18):10055.
- Biggs, JB, Tang CS, Kennedy G (2011). *Teaching for Quality Learning at University*. (Open University Press, Maidenhead).
- Black P, Wiliam D (1998) Assessment and classroom learning. *Assessment Ed.: Principles, Policy Practice* 5(1):7–74.
- Bonwell CC, Eison JA (1991) *Active Learning: Creating Excitement in the Classroom*, ASHE-ERIC Higher Education Report No. 1 (School of Education and Human Development, George Washington University, Washington, DC).
- Bridgelall R (2024) Unraveling the mysteries of AI chatbots. *Artificial Intelligence Rev.* 57(4):89.
- Bromley P (2013) Active learning strategies for diverse learning styles: Simulations are only one method. *Political Sci. Politics* 46(4):818–822.
- Carrasco-Sáez JL, Contreras-Saavedra C, San-Martín-Quiroga S, Contreras-Saavedra CE, Viveros-Muñoz R (2025) Analyzing higher education students' prompting techniques and their impact on ChatGPT's performance: An exploratory study in Spanish. *Appl. Sci.* 15(14):7651.
- Dean KL, Fornaciari CJ (2002) How to create and use experiential case-based exercises in a management classroom. *J. Management Ed.* 26(5):586–603.
- Drake MJ (2019) Teaching OR/MS with cases: A review and new suggestions. *INFORMS Trans. Ed.* 19(2):57–66.
- Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT (2013) Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psych. Sci. Public Interest* 14(1):4–58.
- Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, Baabdullah AM, et al. (2023) "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Internat. J. Inform. Management* 71(March): 102642.
- Eaton SE (2023) Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *Internat. J. Ed. Integration* 19(1):23.
- Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, Shen Y, et al. (2025) Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British J. Ed. Tech.* 56(2):489–530.
- Farrokhnia M, Latifi S, Papadopoulos PM, Hogenkamp L, Gijlers H, Khosravi H, Noroozi O (2026) Generative AI offers more, but students revise less: Comparing the effects of teacher and AI feedback on student essay revisions. *Internat. J. Ed. Tech. High Ed.* 23(1):6.
- Flodén J (2025) Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Ed. Res. J.* 51(1):201–224.
- Francis NJ, Jones S, Smith DP (2025) Generative AI in higher education: Balancing innovation and integrity. *British J. Biomed. Sci.* 81:14048.
- Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP (2014) Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. USA* 111(23):8410–8415.
- Gao XA, Wright JR, Leyton-Brown K (2019) Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence* 275:618–638.
- Grant N, Metz C (2022) A new chat bot is a "code red" for Google's search business. *International New York Times* (December 21), <https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html>.
- Greenland S, Saleem M, Misra R, Mason J (2022) Sustainable management education and an empirical five-pillar model of sustainability. *Internat. J. Management Ed.* 20(3):100658.
- Gruenhagen JH, Sinclair PM, Carroll JA, Baker PRA, Wilson A, Demant D (2024) The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers Ed.: Artificial Intelligence* 7:100273.
- Gupta P, Mahajan R, Badhera U, Kushwaha PS (2024) Integrating generative AI in management education: A mixed-methods study using social construction of technology theory. *Internat. J. Management Ed.* 22(3):101017.
- Heriot KC, Cook R, Jones RC, Simpson L (2008) The use of student consulting projects as an active learning pedagogy: A case study in a production/operations management course. *Decision Sci. J. Innovative Ed.* 6(2):463–481.
- Hillon ME, Cai-Hillon Y, Brammer D (2012) A brief guide to student projects with industry. *INFORMS Trans. Ed.* 13(1):10–16.
- Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, et al. (2025) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inform. Systems* 43(2):42:1–42:55.

- Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, et al. (2023) Survey of hallucination in natural language generation. *ACM Comput. Survey* 55(12):1–38.
- Jose B, Cherian J, Verghis AM, Varghise SM, S M, Joseph S (2025) The cognitive paradox of AI in education: Between enhancement and erosion. *Frontiers Psych.* 16.
- Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, et al. (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individual Differences* 103:102274.
- Kirwan A (2024) ChatGPT and university teaching, learning and assessment: Some initial reflections on teaching academic integrity in the age of large language models. *Irish Ed. Stud.*
- Korfage Smith R (2010) A case study in project-based learning: An international partnership. *J. Teaching Internat. Bus.* 21(3):178–188.
- Krajewski LJ, Malhotra MK (2024) *Operations Management: Processes and Supply Chains*, 14th ed (Pearson, Harlow, UK).
- Larson BZ, Moser C, Caza A, Muehlfeld K, Colombo LA (2024) Critical thinking in the age of generative AI. *Acad. Management Learn. Ed.* 23(3):373–378.
- Lee D, Palmer E (2025) Prompt engineering in higher education: A systematic review to help inform curricula. *Internat. J. Ed. Tech. High Ed.* 22(1):7.
- Lee D, Arnold M, Srivastava A, Plastow K, Strelan P, Ploekf L, Lekkas D, et al. (2024) The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Comput. Ed.: Artificial Intelligence* 6:100221.
- Li Y, Ho JC (2024) Generative AI for operations management education and learning: A critical assessment and analysis. *Internat. J. Services Standards* 14(4):381–396.
- Li H, Yang F, Liu J (2026) Task complexity and AI dependency among college students: The mediating roles of cognitive load, future anxiety, and task motivation. *BMC Psych.*
- Liu Y, Li H, Ouyang J, Xue Z, Wang M, He H, Song B, et al. (2025) Evaluating large language models for preoperative patient education in superior capsular reconstruction: Comparative study of Claude, GPT, and Gemini. *JMIR Perioperative Medicine* 8(1):e70047.
- Lumley T, Diehr P, Emerson S, Chen L (2002) The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* 23:151–169.
- Mehmet M, Papakosmas M, Kyriazis E, Nikidehaghani M (2025) Re-imagining marketing education for career readiness in the GenAI Era. *J. Marketing Ed.* 47(2):170–190.
- Patac LP, Patac AV (2025) Using ChatGPT for academic support: Managing cognitive load and enhancing learning efficiency: A phenomenological approach. *Soc. Sci. Humanities Open* 11:101301.
- Petersen CI, Baeppler P, Beitz A, Ching P, Gorman KS, Neudauer CL, Rozaitis W, et al. (2020) The tyranny of content: "Content coverage" as a barrier to evidence-based teaching approaches and ways to overcome it. *CBE Life Sci. Ed.* 19(2):ar17.
- Prince M (2004) Does active learning work? A review of the research. *J. English Ed.* 93(3):223–231.
- Richter E, Spitzer MWH, Morgan A, Frede L, Weidlich J, Moeller K (2025) Large language models outperform humans in identifying neuromyths but show sycophantic behavior in applied contexts. *Trends Neurosci. Ed.* 39:100255.
- Shadish WR, Cook TD, Campbell DT (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston).
- Sotiriadou P, Logan D, Daly A, Guest R (2020) The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Stud. Higher Ed.* 45(11): 2132–2148.
- Sullivan M, Kelly A, McLaughlan P (2023) ChatGPT in higher education: Considerations for academic integrity and student learning. *J. Appl. Learn. Teaching* 6(1):31–40.
- Sun D, Ba S, Cha Y, Yu J, Chiang F, Dai HM, Lim CP (2026) Empowering university teachers in higher education: A generative AI-responsive competency framework. *Comput. Ed.: Artificial Intelligence* 10:100542.
- Sweller J (1994) Cognitive load theory, learning difficulty, and instructional design. *Learn. Instruction* 4(4):295–312.
- Thomas JW (2000) A review of research on project-based learning. Research report, The Autodesk Foundation, San Rafael, CA, 1–45.
- Thüs D, Malone S, Brünken R (2024) Exploring generative AI in higher education: A RAG system to enhance student engagement with scientific literature. *Frontiers Psych.* 15.
- Vendrell M, Johnston SK (2026) Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. *Comput. Ed.: Artificial Intelligence*: 100572.
- Wang J (2026) Cognitive offloading through digital tools and its relationship with critical thinking, task persistence, and learning depth. *Frontiers Psych.* 17.
- Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, Chowdhery A, Zhou D (2023) Self-consistency improves chain of thought reasoning in language models. *11th Internat. Conf. Learn. Representations* (Appleton, WI).
- Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, Li X, et al. (2024) Practical and ethical challenges of large language models in education: A systematic scoping review. *British J. Ed. Tech.* 55(1):90–112.
- Yavuz F, Çelik Ö, Yavaş Çelik G (2025) Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British J. Ed. Tech.* 56(1):150–166.
- Zhou L, Schellaert W, Martínez-Plumed F, Moros-Daval Y, Ferri C, Hernández-Orallo J (2024) Larger and more instructable language models become less reliable. *Nature* 634(8032):61–68.