Toward Trustworthy AI in Higher Education: An Empirical Evaluation of GenAI in Management Education

Yicheng Sun[1], Jacky Keung[1], Yihan Liao[1], Hi Kuen Yu[1], Xiaoxue Ma[2*]
[1] City University of Hong Kong, Hong Kong, China
[2] Hong Kong Metropolitan University, Hong Kong, China
[*] kxma@hkmu.edu.hk

Abstract: *Recent advances in generative artificial intelligence (GenAI) have significantly expanded the role of large language models (LLMs) in higher education, offering new opportunities for automated learning support and knowledge dissemination. However, their accuracy in discipline-specific fields such as management remains underexplored. To address this gap, we adopt a mixed evaluation strategy that combines objective metrics (accuracy and semantic similarity scores) with expert-based assessments of correctness, logical coherence, and disciplinary relevance. Our findings show that GPT-4 achieves the highest overall accuracy and consistency, particularly excelling in case-based reasoning. LLaMA performs competitively on definitional and factual questions but struggles with complex, multi-step reasoning, while DeepSeek demonstrates efficiency and contextual relevance, though with lower accuracy in advanced analytical scenarios. These results highlight both the potential and limitations of current LLMs in management education and provide actionable insights for educators and curriculum designers on integrating GenAI tools into higher education.*
Keywords: GenAI, Management Education, Educational Technology Integration

1. Introduction

The mainstreaming of GenAI has prompted universities to incorporate these technologies into curricula and policy frameworks. Several higher education institutions now provide orientation materials, training programs, or even GenAI-enabled classroom tools that promote ethical, transparent, and pedagogically aligned use (Ning et al., 2024). These initiatives reflect growing awareness that GenAI is not merely a novelty, but a lasting component of the digital education landscape (Poth et al., 2024). Nevertheless, the majority of these efforts have focused on technically oriented disciplines where the tasks GenAI performs are more rule-based or fact-centered. In contrast, the role of GenAI in management education remains under-investigated. Management is a complex, interdisciplinary field that encompasses economics, strategy, operations, marketing, and organizational behavior (Pantazatos et al., 2024). Unlike STEM domains, management education emphasizes analytical judgment, qualitative synthesis, contextual reasoning, and problem solving under ambiguity—all of which require nuanced understanding of human and institutional dynamics (Meli et al., 2024). These demands may pose significant challenges for current LLMs, which, despite their fluency, may struggle with tasks that require multi-step logic, trade-off analysis, or contextual alignment with business theories.

To address these gaps, our study conducts a fine-grained, discipline-specific evaluation of three leading LLMs—GPT-4, LLaMA, and DeepSeek—on tasks relevant to management education. We focus on three representative question types: definitional, case-based, and problem-solving tasks. Each reflects a distinct cognitive demand that students regularly encounter in business school coursework and assessments. Using a mixed-methods approach that combines objective accuracy and semantic similarity scoring with expert-based evaluations of logical coherence and disciplinary relevance, we aim to provide a comprehensive assessment of GenAI's performance in management contexts. Our findings contribute to the emerging discourse on trustworthy GenAI in education, offering actionable insights for educators, developers, and policymakers on how to responsibly integrate these tools into management teaching and learning

environments. Based on the background and identified research gaps, the following research questions are proposed to guide this study:

RQ1: *How accurately do different GenAI models (GPT-4, LLaMA, and DeepSeek) respond to discipline-specific questions in management education across three major cognitive task types: definitional, case analysis, and problem-solving?*

RQ2: *To what extent do GenAI-generated responses demonstrate logical coherence and disciplinary relevance when evaluated by domain experts in management education?*

2. Methodology

2.1. Participants

To evaluate the quality and appropriateness of GenAI-generated responses within a management education context, we recruited seven domain experts as evaluators. All participants held senior academic positions in the field of management and had more than ten years of experience in teaching, curriculum development, and student assessment at the undergraduate or postgraduate level. Their areas of specialization included strategic management, organizational behavior, marketing, operations management, and business ethics.

2.2. Data Collection

The dataset used in this study was compiled from course materials and exam archives provided by two major research universities in China. In total, 258 distinct questions were collected, each originally developed and used in actual teaching or assessment contexts. The dataset includes questions representative of core areas within the management curriculum and spans multiple cognitive levels to reflect realistic student learning outcomes.

The distribution of question types is summarized in Table 1. As shown, the dataset provides a balanced yet realistic representation of the types of questions commonly encountered in management courses, allowing for comprehensive evaluation across task types. Each question was input into three different GenAI models (GPT-4, LLaMA, and DeepSeek) using consistent prompts to ensure comparability. Responses were then anonymized and randomized for expert evaluation to minimize bias.

Table 1.

Distribution of Question Types in the Dataset.

Question Type	Description	Count
Definitional	Conceptual definitions and theory articulation	92
Case Analysis	Scenario-based evaluation and reflection	84
Problem Solving	Applied decision-making and strategic planning	82

3. Results

3.1 Objective Performance Across Question Types (RQ1)

In this section, we conducted a quantitative analysis based on objective evaluation metrics. This included binary accuracy for definitional questions and semantic similarity scores for case analysis and problem-solving questions. These metrics provide an initial, model-agnostic assessment of how well each system reproduces expected or reference-aligned outputs. For definitional questions, each model-generated response was compared to a gold-standard answer manually written by domain experts. Responses that closely matched the key conceptual content of the gold-standard were labeled as accurate (score = 1); others were scored as inaccurate (score = 0). A summary of the objective evaluation results is presented in Table 2 below. The results confirm that model performance varies significantly across question types, underscoring the need to match GenAI tools to specific pedagogical contexts. In detail, GPT-4 achieved the highest accuracy rate of 93.5%, followed by LLaMA at 88.5%, and DeepSeek at 85.0%. These results suggest that all three models are relatively proficient at factual recall and concise articulation of core management concepts, though GPT-4 demonstrates a measurable edge in both precision and completeness.

Table 2.

Objective Performance of GenAI Models Across Question Types.

Question Type	Metric	GPT-4	LLaMA	DeepSeek
Definitional	Accuracy (Binary)	93.5%	88.5%	85.0%
Case Analysis	Cosine Similarity (SBERT)	0.89	0.71	0.77
Problem Solving	Cosine Similarity (SBERT)	0.87	0.62	0.74

For the more open-ended case analysis tasks, we employed Sentence-BERT to compute cosine similarity scores between model responses and instructor-provided reference answers. This approach captures the semantic closeness of model-generated insights to pedagogically appropriate responses. GPT-4 again outperformed the other models with a mean similarity score of 0.89, followed by DeepSeek (0.77) and LLaMA (0.71). The higher performance of GPT-4 is likely attributable to its stronger contextual understanding and ability to generate structured reasoning aligned with management theory.

A similar pattern was observed for problem-solving questions, which required models to provide strategic recommendations, justify decisions, and apply management frameworks. These tasks represent the highest level of cognitive complexity among the three categories. GPT-4 maintained strong performance with a mean similarity score of 0.87, while DeepSeek and LLaMA scored 0.74 and 0.62, respectively. Notably, LLaMA's responses to problem-solving questions were often generic or incomplete, lacking specific application of management tools such as SWOT, Porter's Five Forces, or the BCG matrix.

3.2 Expert Evaluation of Reasoning and Relevance (RQ2)

In this section, we analyzed the results from the structured rubric-based evaluation conducted by seven senior management faculty members. Each model-generated response was assessed along three key dimensions: Correctness, Logical Coherence, and Disciplinary Relevance, using a 3-point Likert scale (0 = poor, 1 = acceptable, 2 = strong). The maximum possible score per response was 6. This evaluation allowed us to move beyond surface-level metrics and probe the quality of reasoning, clarity of structure, and alignment with domain knowledge. A summary of expert-based evaluation scores is presented in Table 3 below. The results reaffirm GPT-4's superiority in generating responses that are not only accurate but also pedagogically meaningful and contextually aligned with the management discipline. GPT-4 consistently achieved the highest average scores across all three dimensions, with particularly strong ratings for logical coherence (1.82) and disciplinary relevance (1.89), suggesting that its responses were not only factually accurate but also well-structured and educationally appropriate. LLaMA performed moderately well on correctness (1.32) but received lower scores for coherence (1.01) and relevance (1.18), reflecting challenges in maintaining structured argumentation and theoretical precision. DeepSeek, while efficient in producing concise outputs, scored lower overall, with more variability in response depth and alignment with management frameworks.

Table 3.

Average Expert Evaluation Scores by Dimension and Model.

Model	Correctness	Logical Coherence	Disciplinary Relevance
GPT-4	1.85	1.82	1.89
LLaMA	1.32	1.01	1.18
DeepSeek	1.27	0.93	1.06

For definitional questions, all three models received relatively high correctness scores, though GPT-4's explanations were more likely to include relevant sub-concepts or theoretical qualifiers. In contrast, case analysis and problem-solving responses revealed greater divergence in expert assessments. In these categories, GPT-4 maintained consistent performance, often applying appropriate models (e.g., SWOT, 5Cs, stakeholder analysis), while LLaMA and DeepSeek produced more generalized or partially structured answers that lacked depth or contextual tailoring.

The dimension of disciplinary relevance proved especially discriminative. Evaluators noted that GPT-4's responses often reflected the tone and depth expected in academic or applied business settings, while LLaMA and DeepSeek occasionally introduced vague, overly generic statements or omitted key management constructs. For instance, in a prompt requiring strategic planning, GPT-4 referenced the BCG matrix and market entry strategies, while DeepSeek merely suggested "improving operations" without theoretical justification.

4. Discussion

4.1. Findings of the Results

The results of this study reveal clear patterns in how different GenAI models respond to the cognitive and pedagogical demands of management education. Rather than offering uniform performance across task types, each model demonstrated distinct strengths and limitations, shaped by its underlying architecture, scale, and training data. These findings emphasize that GenAI tools must be selected and deployed with pedagogical intentionality. No single model meets all instructional demands, and their effectiveness depends on alignment with task type, learning objectives, and instructional context. Educators and curriculum designers should consider both technical performance and cognitive fit when integrating GenAI into management education.

4.2. Educational Implications

The results of this study underscore the importance of task-aware GenAI integration in management education. Given the varying strengths of each model across cognitive task types, educators should avoid treating GenAI as a monolithic solution. From a curriculum design perspective, GenAI can be embedded as both a scaffold and a reflective prompt. For instance, students could be asked to critique AI-generated strategies, identify missing assumptions, or compare AI outputs to textbook models. This approach not only demystifies GenAI but also positions it as a tool for inquiry, not authority. Importantly, educators must provide clear guidelines on acceptable GenAI use, differentiating between supportive usage (e.g., drafting, brainstorming) and prohibited actions (e.g., submitting AI-generated assignments without review).

References

Ning, Y., Teixayavong, S., Shang, Y., Savulescu, J., Nagaraj, V., Miao, D., ... & Liu, N. (2024). Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. The Lancet Digital Health, 6(11), e848-e856.

Poth, A., Wildegger, A., & Levien, D. A. (2024, September). Considerations about integration of GenAI into products and services from an ethical and legal perspective. In European conference on software process improvement (pp. 155-171). Cham: Springer Nature Switzerland.

Pantazatos, D., Grammatikou, M., & Maglaris, V. (2024, May). Enhancing soft skills in network management education: A study on the impact of GenAI-based virtual assistants. In 2024 IEEE Global Engineering Education Conference (EDUCON) (pp. 1-5). IEEE.

Meli, K., Taouki, J., & Pantazatos, D. (2024). Empowering educators with generative ai: The genai education frontier initiative. In EDULEARN24 Proceedings (pp. 4289-4299). IATED.