A Multi-model Deliberative Generative AI Framework: Enhancing Critical Thinking and Content Veracity in Higher Education

Shyh-Jian Tan, WuFeng University, Taiwan
Y. C. K. Chen, WuFeng University, Taiwan

The European Conference on Education 2026
Official Conference Proceedings

Abstract

Generative Artificial Intelligence (GAI), powered by Large Language Models (LLMs), has rapidly entered higher education, offered instant content generation but raising concerns about student overreliance and cognitive slacking. While GAI outputs often appear credible, they remain vulnerable to bias and hallucinations, creating risks of misinformation. Current teaching practices focus mainly on tool operation, lacking systematic approaches to guide students in verifying authenticity or engaging in critical evaluation. This study proposes the Multi-model Deliberative Generative AI Framework (MDGAIF), which employs cross-model verification and structured guidance to transform students from passive recipients into active evaluators and ethical stewards of AI. A quasi-experimental design was conducted with students in an "Introduction to Machine Learning" course, comparing an experimental group (MDGAIF-integrated teaching) with a control group (traditional tool-based teaching). Mixed-methods analysis revealed that MDGAIF significantly improved students' habits of verifying information and enhanced their critical thinking and problem-solving skills. Findings demonstrate that MDGAIF not only achieves its intended goals but also provides a valuable reference for universities seeking to design integrated AI teaching strategies and curricula.

Keywords: generative artificial intelligence, critical thinking, higher education, MDGAIF, content veracity

iafor
The International Academic Forum
www.iafor.org

Introduction

The landscape of higher education has undergone a profound transformation with the advent of Generative Artificial Intelligence (GAI) technologies powered by Large Language Models (LLMs). Since the public release of ChatGPT in late 2022, followed by rapid advancements in competing platforms such as Google's Gemini, Anthropic's Claude, and Meta's Llama series, students have gained unprecedented access to AI systems capable of generating sophisticated text, code, images, and even video content with remarkable fluency and apparent expertise. This technological revolution has fundamentally altered the dynamics of academic work, enabling students to produce complex assignments, solve technical problems, and generate research content at speeds previously unimaginable.

Research Context and Motivation

However, the convenience of Generative Artificial Intelligence (GAI) technologies have precipitated a critical pedagogical crisis that we term the "Cognitive Slacking" phenomenon. Students increasingly exhibit a pattern of over-reliance on AI-generated outputs, submitting content without rigorous verification, critical evaluation, or genuine intellectual engagement with the material. This behavioral pattern represents more than mere academic dishonesty; it constitutes a fundamental abdication of cognitive responsibility that threatens the core educational mission of developing independent critical thinking and problem-solving capabilities. The ease with which students can obtain plausible-sounding answers has created what Mollick and Mollick (2023) describe as a "delegation of thinking," where the cognitive labor of analysis, synthesis, and evaluation is outsourced entirely to algorithmic systems.

The risks associated with this uncritical adoption of GAI outputs extend beyond pedagogical concerns to encompass serious issues of content veracity and academic integrity. Despite their impressive linguistic capabilities, LLMs are fundamentally probabilistic systems that generate text based on statistical patterns in training data rather than genuine understanding or factual knowledge. This architectural limitation manifests in several critical failure modes: (1) *hallucinations*, where models confidently generate factually incorrect information that appears logically coherent; (2) *outdated information*, as models are trained on historical data and lack awareness of recent developments; (3) *biases*, reflecting problematic patterns in training data; and (4) *logical inconsistencies*, where reasoning chains contain subtle but significant errors (Zhang et al., 2025).

The emergence of so-called "reasoning models" in 2025—exemplified by OpenAI's o1 and o3 series (OpenAI, 2024) and Anthropic's Claude 3.7 Sonnet (Anthropic, 2025)—has paradoxically intensified these concerns. While these advanced systems demonstrate enhanced Chain-of-Thought (CoT) capabilities and can engage in more sophisticated multi-step reasoning, research indicates they remain vulnerable to producing "highly plausible hallucinations" (Zhang et al., 2025). The very sophistication of their outputs makes errors more difficult to detect, as they are embedded within otherwise coherent and logically structured responses. This creates what we term the "hallucination paradox": as models become more capable, their errors become more dangerous precisely because they are more convincing.

Furthermore, the recent shift toward "Agentic Workflows"—where AI systems can autonomously execute complex multi-step tasks with minimal human intervention—has amplified the cognitive slacking problem. When students deploy autonomous agents capable of independently researching, coding, writing, and debugging, the degree of automation often

triggers “deep cognitive delegation.” Students become mere initiators and acceptors of AI-generated work, abandoning the essential quality audit processes that should accompany any intellectual production.

Research Objectives

This study addresses the implementation gap by proposing and empirically validating the Multi-Model Deliberative Generative AI Framework (MDGAIF), a comprehensive pedagogical intervention designed to operationalize UNESCO’s 2025 principles within the context of technical higher education. The framework transforms the student’s role from that of a “prompt sender” who passively receives AI outputs to that of an “auditor and quality controller” who actively deliberates on, verifies, and takes responsibility for AI-assisted work.

The research pursues three primary objectives.

Objective 1: Framework Development

To design and articulate a systematic, replicable pedagogical framework that embeds deliberative verification procedures directly into AI-assisted learning workflows, specifically tailored for technical education contexts where content veracity is paramount.

Objective 2: Empirical Validation

To rigorously assess the framework’s effectiveness through a quasi-experimental study measuring its impact on two critical dimensions: (a) critical thinking capabilities, as measured by standardized instruments adapted from Facione (1990) and Ng et al. (2021), and (b) complex problem-solving behaviors, as assessed through systematic analysis of student-AI interaction patterns based on OECD (2014) and Wing (2006) frameworks.

Objective 3: Pedagogical Translation

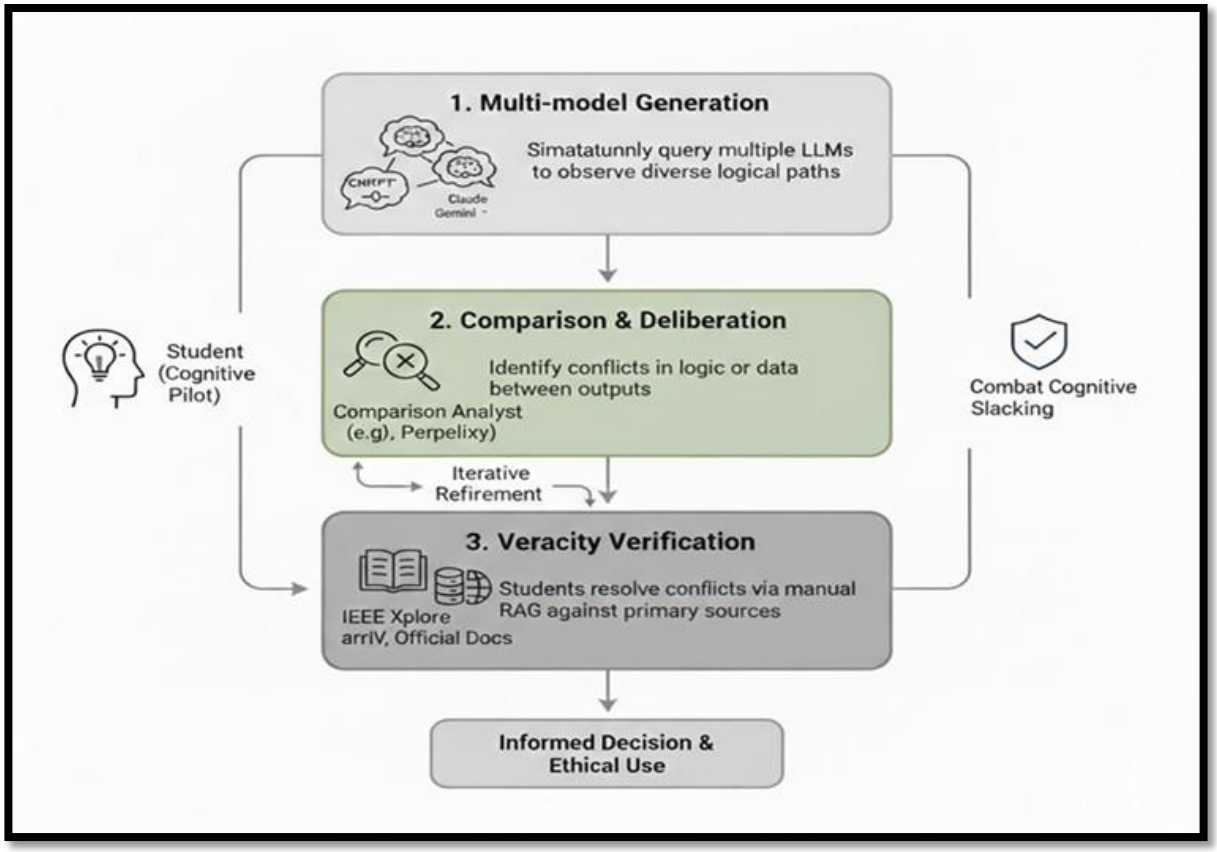
To provide concrete implementation guidance, including specific instructional protocols, assessment rubrics, and operational procedures that enable other educators to adopt and adapt the framework within their own institutional contexts.

MDGAIF Architecture and Process

The Multi-Model Deliberative Generative AI Framework (MDGAIF) transforms students from passive consumers into active deliberative evaluators. The MDGAIF architecture and processes are shown as Figure 1 which consists of three interconnected modules.

Figure 1

Framework Architecture and Processes



Module 1: Multi-model Generation

Students formulate a structured prompt and submit it to three distinct LLMs: ChatGPT, Claude, and Gemini. This induces "cognitive diversity," exposing students to multiple approaches and preventing premature acceptance of a single AI answer.

Module 2: Comparison and Deliberative Conflict Identification

Students use a fourth model, such as Perplexity, as a "Comparison Analyst." This model identifies consensus, conflicts in logic, and specific technical differences. This meta-level critical thinking teaches students to evaluate the AI's evaluative process.

Module 3: Veracity Verification

This represents the core deliberative component. Students resolve conflicts through "Manual RAG," consulting authoritative sources like IEEE Xplore, official library documentation, or textbooks. Students are prohibited from resolving conflicts by asking another AI model for the "correct" answer.

Research Methodology

The study employed a mixed-methods approach combining quantitative psychometric instruments with qualitative behavioral analysis. This triangulation strategy enhances validity by capturing both self-reported cognitive capabilities and actual behavioral performance.

Quantitative Analysis by Psychometric Instruments

Quantitative Instrument 1: Critical Thinking Scale (CTS)

The Critical Thinking Scale (CTS) served as the primary quantitative measure of cognitive development. This instrument was adapted from two established sources: Facione's (1990) California Critical Thinking Disposition Inventory (CCTDI) and Ng et al.'s (2021) AI Literacy framework.

Quantitative Instrument 2: Complex Problem-Solving Assessment (CPS)

The Complex Problem-Solving (CPS) Assessment represents a behavioral coding and scoring tool rather than a self-report instrument. This assessment was designed to capture actual performance in AI-assisted problem-solving contexts, addressing the limitation that self-report measures may not accurately reflect actual behavior. The CPS Assessment draws on two frameworks: (1) the OECD's (2014) PISA framework for assessing complex problem-solving, which emphasizes exploration, understanding, and planning in dynamic environments; and (2) Wing's (2006) computational thinking framework, which identifies key cognitive processes in technical problem-solving including decomposition, pattern recognition, abstraction, and algorithm design.

Qualitative Behavioral Analysis

In addition to quantitative instruments, the study collected rich qualitative data.

Interaction Logs

As described above, complete logs of student-AI interactions provided detailed behavioral data. Beyond quantitative coding, these logs were analyzed qualitatively to identify patterns, strategies, and changes in interaction styles.

Deliberation Logs

Students submitted Deliberation Logs with each assignment, documenting their multi-model comparison, conflict identification, verification process, and reflections. These logs provided insight into students' deliberative reasoning and metacognitive awareness.

Reflection Journals

Students maintained weekly reflection journals throughout the study, responding to prompts such as "What did you learn about AI capabilities and limitations this week?" and "How has your approach to using AI changed?" These journals captured students' evolving understanding and attitudes.

Focus Group Interviews

At the conclusion of the study, three focus group interviews were conducted (10 students per group) to gather in-depth qualitative feedback on the MDGAIF experience, challenges encountered, and perceived benefits.

Research Results

Quantitative Findings

- CTS scores increased from 24.35 to 33.95 ($p < .001$, $d = 2.65$).
- Information filtering improved by 75%.
- Logical inference improved by 44%.
- CPS showed 237% growth, with review and identification improving 300%.

Qualitative Findings

- Students shifted from vague, single prompts to multi-stage deliberative workflows.
- Error detection rates rose to 85% (logic) and 92% (factual).
- Reflection journals revealed increased awareness of AI limitations and ethical responsibility.

Conclusion

The introduction of generative AI into higher education marks a significant milestone in educational technology evolution. Previously, most tools focused on making learning more efficient or convenient, but generative AI has the potential to fundamentally transform what "learning" means. Consider what if students could instantly come up with flawless answers without thinking. Then, what would become education's fundamental goal: to "develop critical thinking skills"? However, this study indicates that with the right teaching strategies, AI can support and enhance human thinking rather than replace it. Positioning AI as a tool for reflective inquiry, rather than a substitute for thought, facilitates the cultivation of critical thinking. This is achieved through a systematic process of student-led content verification and the establishment of clear accountability for the resulting contents. In response to this paradigm shift, MDGAIF offers a possible solution. It is grounded in solid theory, provides clear instructional steps, and has been thoroughly validated through research. While it is not a universal remedy, it offers a practical framework for educators aiming to integrate AI into teaching. This is without compromising academic quality and integrity. As AI capabilities advance unpredictably, MDGAIF principles-critical assessment, cross-verification, self-awareness, and ethical responsibility—will become increasingly vital. Regardless of whether teaching materials originated from humans or AI, teaching students to quickly evaluate information authenticity will inevitably become a key component of higher education. This study demonstrates that education, by keeping pace with AI and harnessing its power for positive development, is not only a contemporary trend but also an attainable goal. It gives us hope—higher education can undoubtedly adapt to the AI era and continue to fulfill its irreplaceable role of "inspiring the human intelligence."

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that AI technologies were used in the preparation of this manuscript. Specifically, Gemini Nanbanana was used to visually render Figure 1 based on the author's original conceptual design. Additionally, Gemini AI Tools was used for language refinement and proofreading to ensure clarity and grammatical accuracy. The ideas, research procedures, data analysis, and the core deliberative logic of the MDGAIF framework remain the original work of the author.

References

Anthropic. (2025). *Claude 3.7 Sonnet: The first hybrid reasoning model*. Anthropic News. https://www.anthropic.com/news/claude-3-7-sonnet

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. California Academic Press. https://eric.ed.gov/?id=ED315423

Mollick, E. R., & Mollick, L. (2023). *Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts*. Wharton School Working Paper. SSRN. https://ssrn.com/abstract=4391243

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

OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. OECD Publishing. https://doi.org/10.1787/9789264208070-en

OpenAI. (2024). *Technical report: Enhancing reasoning and verification in o1 models*. OpenAI Blog. https://openai.com/index/learning-to-reason-with-llms/

UNESCO. (2025). *AI and the future of education: Disruptions, dilemmas and directions*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000395236

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Zhang, Y., Wang, L., Chen, C., & Liu, J. (2025). *Reasoning large language model errors arise from hallucinating critical problem features*. arXiv. https://doi.org/10.48550/arXiv.2505.12151