Evaluating Generalist Conversational AI Against Foundational Models of Instructional Design: A Comparative Analysis

The Case for Specialized AI in Instructional Engineering

Abdelmounaim AZINDA[1], Mohamed Khaldi[2]
Laboratory of Information Technologies and System Modeling, Faculty of Science,
Abdelmalek Essaadi University, Tetouan, Morocco[1]
Laboratory of Information Technologies and System Modeling, Faculty of Science, Tetouan, Morocco[2]

***Abstract*—The rapid integration of Generative AI into instructional engineering presents a critical challenge: verifying the capacity of these tools to strictly adhere to systemic theoretical models of learning, despite the risk of generating "pedagogically hallucinated" content that possesses surface plausibility but lacks structural validity. This study addresses this gap by systematically evaluating the performance of generalist conversational AIs against foundational principles of Instructional Design (ID). Adopting a qualitative comparative analysis of four state-of-the-art models available in October 2025—GPT-5 (OpenAI), Gemini 2.5 Pro (Google), Claude Sonnet 4.5 (Anthropic), and DeepSeek V3.2 (DeepSeek AI)—we assessed their outputs for complex design scenarios against a multi-dimensional framework grounded in authoritative theories, including Biggs's Constructive Alignment, Merrill's First Principles of Instruction, and Universal Design for Learning (UDL). Results reveal a "paradox of competence without comprehension," where models demonstrate high factual reliability and linguistic fluency but exhibit significant shortcomings in maintaining logical pedagogical consistency, particularly regarding assessment alignment and accessibility standards, with only Claude Sonnet 4.5 demonstrating a notable proactive partnership posture. Consequently, we conclude that current generalist LLMs cannot function as autonomous expert designers and argue for a shift in professional practice toward Critical AI Literacy, where the human designer leverages AI for ideation but remains the essential guarantor of the pedagogical architecture.**

Keywords*—*Generative AI; large language models (LLMs); instructional design; constructive alignment; pedagogical evaluation; AI ethics in education; Human-AI collaboration

I. Introduction

The advent of large language models (LLMs) and Generative Artificial Intelligence (AI) more broadly marks a significant technological inflection point that is profoundly reshaping numerous professional fields [1]. Platforms such as ChatGPT (OpenAI), Gemini (Google), DeepSeek (DeepSeek AI), or Claude (Anthropic) have democratized access to unprecedented capabilities for content generation, text analysis, and natural language dialogue. The domain of education and training has not been spared this transformation, where a rapid adoption of these tools by practitioners—notably instructional designers and learning engineers—is evident [2].

Recent scholarly literature has begun to document the application of these AIs to specific design tasks. This body of research has primarily focused on their efficiency in automating discrete, micro-level tasks, such as generating assessment items, creating lesson plans, or producing document summaries [3]. Such studies generally conclude that generalist AIs can yield significant productivity gains and function as effective brainstorming assistants, thereby augmenting the creative capacity of designers.

However, this task-oriented efficiency perspective obscures a fundamental dimension of instructional design: its systemic and theoretically-grounded nature. Instructional engineering is not a mere collection of activities but a structured process aimed at guaranteeing the coherence and efficacy of the learning experience. This practice relies on robust theoretical models, such as Biggs's constructive alignment [4], which demands a seamless articulation between learning objectives, pedagogical activities, and assessment methods. Likewise, Merrill's first principles of instruction [5] define the non-negotiable conditions for effective learning, such as the activation of prior knowledge and the grounding of learning in authentic problems. Consequently, the uncritical adoption of non-specialized tools creates a tangible risk of producing learning designs that exhibit surface-level plausibility but lack the deep structural coherence that underpins their actual effectiveness [6].

To date, a major gap persists in the scientific literature: no study has systematically evaluated the capacity of generalist conversational AIs to adhere to these foundational principles of instructional engineering. As a result, the following research questions remain unanswered: To what extent do the outputs of these AIs conform to the principle of constructive alignment? Are they capable of integrating complex pedagogical strategies beyond simple information transmission? To what degree do they account for the ethical and accessibility considerations inherent to responsible design?

The present study aims to address this gap. It proposes and applies a systematic evaluation framework, grounded in established instructional design models, to comparatively analyze the performance of leading generalist language models (GPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5, and DeepSeek V3.2) in solving complex design problems. The objective is to illuminate their genuine capabilities but also, and more importantly, their systematic limitations, thereby justifying the scientific rationale for developing specialized AI tools for instructional engineering.

This study is structured as follows: Section II presents related works. Section III details the methodology, presenting the corpus of AIs under study and our criteria-based analysis framework. Section IV presents the results of our comparative analysis. Section V outlines the discussion of results. Finally, Section VI discusses the interpretation of these findings and their implications for research and practice, before concluding with future outlooks.

II. Related Works

A. Generative AI in Educational Content Production

Since the democratization of large language models (LLMs), a rapidly expanding body of literature has investigated their potential as auxiliary tools for educators [1], [2]. Early explorative studies focused on the efficiency of these platforms in automating discrete instructional design tasks, such as generating lesson outlines, assessment items, or summaries [3]. Recent research has moved to evaluate the operational acceptability of these outputs. Studies [2], [3] suggest that while generalist AIs significantly reduce the initial cognitive load for designers, they are predominantly viewed as productivity engines rather than expert systems. The consensus posits Generative AI as a "brainstorming partner" [17] capable of expanding the creative scope of practitioners, provided that the human remains the operational architect of the learning experience [14].

B. The Problem of Validity: From Fact to Pedagogy

While the linguistic fluency of LLMs is established, their reliability remains a contested subject. The phenomenon of "hallucination"—the confident generation of incorrect information—has been extensively documented in recent technical surveys [19], [20]. However, in the context of instructional engineering, factual accuracy is secondary to structural coherence. Critics such as [6] argue that the probabilistic nature of LLMs makes them ill-suited for educational tasks requiring a rigorous logical chain, echoing earlier concerns about the limitations of "stochastic parrots" in high-stakes environments. Despite these warnings, there is a scarcity of empirical studies that specifically evaluate LLM outputs against established pedagogical architectures. Most evaluations rely on surface-level quality metrics which fail to capture the systemic intricacies of a learning sequence.

C. Theoretical Gaps and the Need for Systemic Evaluation

Existing approaches for evaluating AI in education often overlook the structural requirements of the discipline. Theoretical frameworks such as Constructive Alignment [4] or the First Principles of Instruction [5] are widely accepted as the standards for effective learning design, yet they have rarely been utilized as the metric for assessing AI capabilities. Current literature typically treats AI output as isolated assets, whereas instructional engineering demands a holistic ecosystem. Positioning of this Study: This study addresses this gap by moving away from ad-hoc assessment. By adopting a multi-criteria analysis based on vetted theories [4], [5], [9], [12], we offer a novel methodological contribution that shifts the focus from generative quantity to pedagogical quality, aiming to identify potential "pedagogical hallucinations" in model outputs.

III. Methodology

The objective of this research is to systematically evaluate the capabilities and limitations of generalist conversational AIs against the theoretical and practical requirements of instructional engineering. To this end, we have adopted a qualitative research approach based on a systematic comparative analysis [7]. This method allows for an in-depth examination of the relevance and coherence of textual outputs generated by different systems by assessing them against a rigorous analytical framework grounded in established theoretical models. Our methodology is structured in four sequential phases: justifying the research approach, constituting the study corpus, presenting our evaluation instrument, and detailing the test and data analysis protocol.

A. Research Approach

This study is situated within an evaluative and qualitative paradigm. Our goal is not to quantify performance with a single score, but rather to analyze the *nature* of the AI-generated responses, identify recurrent patterns, and understand the underlying reasons for their successes and failures. A quantitative approach would be ill-suited here, as it would fail to capture the subtleties of pedagogical inconsistencies or the contextual relevance of the proposed strategies. Thematic content analysis, guided by a pre-established coding frame, is the central technique of this study [8].

B. Corpus Constitution

The selection of our study corpus was guided by the objectives of representing the technological state-of-the-art, recognized performance, architectural diversity, and relevance to end-users. In accordance with these criteria, we selected four of the most prominent generalist language models available at the time of the study (October 2025):

1) GPT-5 (OpenAI): Included for its market-leading position and massive adoption, making it the de facto benchmark for many practitioners.

2) Gemini 2.5 Pro (Google): Selected for its large context window and native multimodal architecture, positioning it as a top-tier direct competitor.

3) Claude sonnet 4.5 (Anthropic): Chosen for its acclaimed performance on complex reasoning tasks and its design orientation toward safety and ethics, a relevant axis for our analysis.

4) DeepSeek V3.2 (DeepSeek AI): Integrated to diversify our corpus beyond proprietary American models. Its open-weight nature allows for an evaluation of whether alternative

architectures and training data yield distinct outputs in instructional design contexts.

To ensure comparability and the independence of each test, all interactions were conducted between October 1st and October 15th, 2025, using the publicly available versions of the models at that date and initializing a new session for each scenario.

C. Analysis Framework: A Multidimensional Evaluation Instrument

To ensure a systematic and objective evaluation, we developed and validated a multidimensional analysis framework (see Table I). This instrument, our primary methodological contribution, is structured around three pillars, which are broken down into specific criteria. Each criterion is associated with one or more reference theoretical models that are authoritative in the field of instructional design [9], [10], [4], [5], [11], [12].

TABLE I. FRAMEWORK FOR ANALYZING THE PEDAGOGICAL ADEQUACY OF CONVERSATIONAL AIS

Pillar of Analysis	Specific Evaluation Criterion	Theoretical Reference Model(s)	Success Indicators ("Expert" Capability)	Failure Indicators ("Generalist" Limitation)
1. Foundational Pedagogy	**1.1. Objective Hierarchy**	Anderson & Krathwohl (2001); Fink (2003)	- Varies cognitive levels (revised Bloom).	- Remains confined to lower-order Bloom's verbs.
			- Proposes holistic objectives (human dimension, learning how to learn...).	- Unable to formulate affective or metacognitive objectives.
	1.2. Constructive Alignment	Biggs (1996)	- Ensures perfect systemic coherence between objectives, activities, and assessments.	- Produces activities or assessments that are logically disconnected from the stated objectives.
	1.3. Instructional Strategy Pertinence	Merrill (2002)	- Suggests active, problem-based strategies centered on authentic tasks.	- Proposes simple, linear, and passive content transmission.
2. Quality of Human-AI Partnership	**2.1. Nature of Interaction**	Goodyear (2000)	- Acts as a Socratic partner: questions, prompts reflection, suggests alternatives.	- Responds transactionally to each prompt, with no memory or overarching view of the project.
			- Maintains context.	
3. Reliability & Ethical Responsibility	**3.1. Content Validity**	General Scientific Rigor	- Produces factually accurate content without "hallucinations".	- Invents facts, sources, or incorrect information.
	3.2. Equity & Bias Awareness	AI Ethics Frameworks	- Proposes inclusive scenarios and examples.	- Generates stereotypes (gender, cultural...) in content and situations.
			- Can flag risks of bias in a given topic.	
	3.3. Consideration of Accessibility	CAST (2018)	- Can suggest adaptations based on the principles of Universal Design for Learning (UDL).	- Produces uniform content, disregarding different modes of perception and interaction.

D. Theoretical Underpinnings of the Analysis Framework

The analysis framework presented above (see Table I) is not a mere feature checklist, but a holistic evaluation instrument. Each of its components was deliberately chosen for its capacity to probe an essential dimension of instructional design expertise. Its tripartite structure (Pedagogy, Partnership, Ethics) is informed not only by the foundational models of the discipline but also by the most current scientific literature on evaluation frameworks for AI in education.

The first pillar, "Foundational Pedagogy", constitutes the core of our evaluation. The joint use of Bloom's revised taxonomy [9] and Fink's taxonomy of significant learning [10] allows us to cover both the spectrum of cognitive complexity and the holistic nature of learning. More critically, the emphasis on Biggs's Constructive Alignment [4] is consistent with current research that identifies systemic design coherence as the primary challenge for generative AIs in education [13]. Finally, evaluating strategies against Merrill's First Principles of Instruction [5] reflects the growing demand for AI capable of promoting active learning.

The second pillar, "Quality of Human-AI Partnership", was included to move beyond a vision of AI as a simple production tool. Contemporary research increasingly insists on the importance of designing human-AI interactions as a form of cognitive partnership, where the AI does not just execute but can guide, stimulate reflexivity, and augment human intelligence [14]. Our criterion, therefore, assesses the AI's ability to shift from a transactional to a collaborative posture.

Finally, the third pillar, "Reliability and Ethical Responsibility", addresses the non-negotiable conditions for the safe deployment of AI in education. The consideration of algorithmic bias [15] and accessibility through Universal Design for Learning (UDL) [12] is now recognized as an indispensable component of any responsible evaluation of educational technology [16]. The priori integration of these ethical criteria into our analysis framework is thus aligned with the most recent research standards in the field.

In essence, this framework is designed to evaluate an AI not on its linguistic capabilities alone, but on its capacity to simulate the reasoning of a reflective practitioner.

E. Test Protocol and Data Collection

To evaluate the AIs, two test scenarios (prompts) were designed as authentic and complex design problems. The full text of these prompts is available in Appendix A. The collection protocol was as follows:

1) Each test scenario was submitted identically to each AI in the corpus.

2) The entire conversation generated for each scenario was saved verbatim as a plain text file.

3) No modifications or guidance were provided during generation to avoid biasing the results.

This choice of a single-turn interaction was a deliberate decision to evaluate the baseline performance of the models and their ability to interpret a complex request autonomously, without the aid of an iterative refinement process. The study of multi-turn interactions constitutes a separate research avenue.

F. Data Analysis

The data analysis followed a systematic process of content analysis. Each complete response was deconstructed into units of meaning (e.g., paragraphs, proposed activities) and coded against the indicators in our framework (see Table I). This procedure allowed for a criterion-by-criterion comparison of the AIs' performance and the identification of systematic patterns of success and failure. To ensure coding reliability and mitigate interpretation bias, a random subset of the data (25% of the corpus) was independently coded by a second researcher. The few discrepancies were discussed until a consensus was reached, and the coding rules were refined accordingly before the full corpus was analyzed.

IV. Results

The comparative analysis of outputs generated by the four AI models reveals a central paradox. On one hand, the systems demonstrate an undeniable competence in rapidly generating abundant, well-structured, and grammatically correct textual content, thus confirming their potential as productivity assistants [17]. On the other hand, when this content is evaluated against our analytical framework, our findings reveal systematic and profound shortcomings in adhering to the foundational principles of instructional engineering. This dissonance is particularly stark for criteria that demand a systemic understanding of the learning process, such as constructive alignment. The detailed performance metrics, derived from applying our coding scheme to the two test scenarios, are synthesized in Table II below. The narrative analysis that follows this table is intended to comment upon and illustrate, with evidence from our corpus, the most significant trends and patterns revealed by these data.

A. Performance Synthesis Table

Table II presents the quantified results of our systematic content analysis. Each performance was rated on a three-point scale (0 = Critical Failure; 1 = Superficial Response; 2 = Relevant Response) for both scenarios (S1, S2). This table serves as the empirical foundation for the subsequent narrative analysis.

TABLE II. Synthesis of AI Performance According to the Analysis Framework (Scores from 0 to 2)

Pillar of Analysis	Specific Evaluation Criterion (Theoretical Ref.)	GPT-5	Gemini 2.5 Pro	Claude Sonnet 4.5	DeepSeek V3.2	Mean Score by Criterion
		S1/S2	S1/S2	S1/S2	S1/S2	
1. Foundational Pedagogy	1.1. Objective Hierarchy (Bloom / Fink)	1 / 0	1 / 0	1 / 1	0 / 0	**0.50**
	1.2. Constructive Alignment (Biggs)	0 / 1	0 / 1	1 / 2	0 / 0	**0.63**
	1.3. Instructional Strategy Pertinence (Merrill)	1 / 1	2 / 1	2 / 2	1 / 1	**1.38**
2. Human-AI Partnership	2.1. Nature of Interaction (Goodyear)	1 / 1	1 / 1	2 / 2	1 / 1	**1.25**
3. Reliability & Ethics	3.1. Content Validity (Scientific Rigor)	2 / 2	2 / 2	2 / 2	1 / 2	**1.88**
	3.2. Equity and Bias Awareness (AI Ethics)	1 / 1	1 / 2	2 / 2	1 / 1	**1.38**
	3.3. Accessibility (UDL / CAST)	0 / 1	0 / 1	1 / 2	0 / 0	**0.63**
Overall Mean Score by Model		**0.93**	**1.00**	**1.71**	**0.57**	

Note: S1 = Scenario 1 (Corporate Context); S2 = Scenario 2 (Academic Context).
Scoring scale: 0 = Critical Failure; 1 = Superficial Response / Partial Failure; 2 = Relevant Response.

B. Detailed Narrative Analysis

The following qualitative analysis aims to comment on the data presented in Table II, illustrating with concrete examples the most significant shortcomings and successes observed within our corpus.

1) Foundational pedagogy: An Unacquired Systemic Competence

The first pillar of our analysis reveals the most profound limitations of generalist AIs. As evidenced by the low mean scores on pedagogy-related criteria (Table II), the models struggle to mobilize the structural principles of instructional engineering in an operational manner.

a) Objective hierarchy: *A Tendency for Cognitive Simplification.* Our analysis shows a systematic tendency of all four models to "downgrade" the complexity of learning objectives to the lower levels of Bloom's revised taxonomy, namely Knowledge and Comprehension [9]. For Scenario 1, although the mission required targeting high-level competencies such as diagnosis (Level 4: Analysis) and application (Level 3: Application), the responses overwhelmingly proposed objectives far below this requirement. DeepSeek's formulation is particularly representative of this phenomenon:

"Learning objectives: 1) To know the challenges of hybrid management. 2) To remember the company's communication tools. 3) To list the benefits of flexible work." (DeepSeek V3.2, Scenario 1)

These action verbs correspond neither to the complexity of the mission nor to the profile of the target audience. Furthermore, when explicitly instructed to use the taxonomy of significant learning in Scenario 2, no model was able to propose relevant objectives for non-cognitive dimensions such

as the "human dimension" or "learning how to learn" [10]. This theoretical blind spot suggests that the training corpora of generalist AIs are heavily biased toward Bloom's model, to the detriment of other essential conceptual frameworks.

b) Constructive alignment: *Coherence Breakdown as a Main Failure.* The most severe and systematically observed flaw across the corpus is the failure to respect the principle of constructive alignment [4]. With a mean score of 0.63, this criterion is one of the weakest in our study. This breakdown in coherence is particularly flagrant between objectives and proposed assessment methods. In Scenario 1, three of the four models suggested an assessment method entirely unsuited for an application-level competency. GPT-5's proposal is emblematic of this issue:

"Assessment method: The achievement of objectives will be validated by a final online quiz comprising multiple-choice questions on the concepts presented". (GPT-5, Scenario 1)

According to Biggs [4], such a proposal is pedagogically incoherent: a quiz measures declarative knowledge ("knowing what") and not the ability to act in a situation ("knowing how"). This systematic mismatch suggests that generalist models, while recognizing the word "assessment," tend to associate it by default with the simplest formats (e.g., MCQs) without analyzing the nature of the competency to be evaluated.

c) Pertinence of strategies: *A Preference for Transmissive Approaches.* Regarding instructional strategies, our findings indicate a marked preference for transmissive approaches, which contradicts the first principles of instruction by Merrill [5], who insists on the necessity of learning centered on solving authentic problems. Although Scenario 1 described a real-world problem faced by managers, the models predominantly proposed passive sequences. Gemini's response, despite earning a score of 2 for the variety of its proposals, illustrates this trend:

"Sequence 1 (20 min): Introduction video by an expert. Sequence 2 (20 min): Reading of an article on the 5 pillars of communication. Sequence 3 (30 min): A written case study to be read...". (Gemini 2.5 Pro, Scenario 1)

These activities keep the learner in a passive role as a recipient of information. They only very partially implement the phases of Activation, Demonstration, and especially Application and Integration, which are at the heart of Merrill's model and are solely responsible for enabling the effective transfer of skills to the workplace.

2) Quality of human-AI partnership: From Docile Executant to Cognitive Partner

The second pillar of our analysis focused on the nature of the interaction (Criterion 2.1). It aimed to determine whether the AIs could transcend a passive tool role to approximate that of a "cognitive partner", capable of augmenting and stimulating the user's reflection [11]. The results, with a mean score of 1.25, are mixed and reveal a notable difference in interactional posture among the models.

The majority of models (GPT-5, Gemini 2.5 Pro, DeepSeek V3.2) adopted what can be described as a "literal executant" approach. Their interactions remained transactional, conforming to the classic question-answer model without ever initiating a dialogue to clarify, deepen, or challenge the request. This instrumental approach, while efficient, places the entire cognitive load on the human user, who must formulate a perfect prompt to obtain a quality result [18].

In contrast, the Claude Sonnet 4.5 model demonstrated a distinct ability to position itself as a proactive partner, earning it the maximum score of 2 on this criterion. Its response to Scenario 1 began with a phase of Socratic questioning that simulates a genuine expert consultation:

"Excellent project. To ensure I design the most relevant solution for your managers, allow me to clarify a few points: Do we have more specific data on the difficulties they are reporting? [...] This will help me personalize the case studies. Here is a first proposal based on your information..." (Claude Sonnet 4.5, Scenario 1)

This ability to reason about the reasoning itself (reasoning on reasoning) is a marker of intellectual partnership, contrasting sharply with the literal execution of the other models.

3) Reliability and ethical responsibility: Heterogeneous Performances and Blind Spots

The third pillar of our analysis addressed crucial qualitative aspects for the professional use of AI. Our results here are highly heterogeneous.

a) Content validity: Globally Robust Reliability. On the criterion of factual validity (Criterion 3.1), the most recent models (GPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5) demonstrated excellent reliability, with the highest mean score of our study (1.88). In line with recent advances in reducing "hallucinations" [19], no manifestly false information was detected.

b) Equity and bias awareness: A Tendency to Reproduce Stereotypes. The analysis of the ethical dimension (Criterion 3.2), with a mean score of 1.38, reveals a more mixed performance. Our observations corroborate numerous studies on algorithmic bias, which show that LLMs tend to reproduce, or even amplify, existing social stereotypes [15]. Conversely, Claude Sonnet 4.5 appeared to show greater sensitivity, which might reflect an architectural orientation toward mitigating such biases:

"Note: It will be important in the case studies to present an equal diversity of genders and backgrounds... so as not to reinforce existing stereotypes". (Claude Sonnet 4.5, Scenario 1)

c) Accessibility: The Most Glaring Operational Limitation. The accessibility criterion (Criterion 3.3), along with constructive alignment, revealed the models' most significant failure (mean score of 0.63). Although Scenario 2 explicitly requested suggestions for integrating the principles of Universal Design for Learning (UDL), most models failed

to translate this instruction into concrete actions. Their responses were either non-existent or extremely vague. Gemini's response is emblematic of this failure:

"For accessibility (UDL), we must ensure that the content is clear, simple, and accessible to all." (Gemini 2.5 Pro, Scenario 2)

This response offers no operational recommendation derived from the three main principles of UDL: providing multiple means of Representation, Action & Expression, and Engagement [12]. Only Claude Sonnet 4.5 demonstrated a functional knowledge of the framework. This radical performance difference suggests that knowledge of specialized frameworks like UDL is not yet functionally integrated into the majority of generalist AIs, constituting a major blind spot for their application in inclusive instructional design.

V. Discussion

A. Synthesis and Interpretation of Principal Findings: The Paradox of Competence without Comprehension

Our comparative analysis uncovers a striking dichotomy rich in theoretical implications: while generalist conversational AIs demonstrate clear competence in producing superficial pedagogical content, they systematically fail to preserve its deep structural coherence. On one hand, these models effortlessly generate textually plausible and stylistically appropriate lesson plans, objectives, and activities. On the other hand, they exhibit a fundamental incapacity to uphold principles as essential as constructive alignment, appropriate assessment design, or the on-demand application of specific theoretical frameworks.

This systematic shortfall, in our analysis, is not a simple performance error, but must be interpreted as a direct consequence of the underlying architecture of large language models. As probabilistic models optimized for predicting sequential words, they are experts at mimicking the form and style of pedagogical discourse. However, they lack any internal representation of the cognitive or systemic models of learning. Their operation, grounded in the manipulation of linguistic forms without access to underlying semantic meaning, renders them structurally incapable of guaranteeing the logical coherence that forms the core of instructional design [15]. Thus, we face what may be termed the paradox of competence without comprehension: a perfect command of the linguistic *envelope* of instructional design, yet a total absence of the logical *internal structure* that must inform it.

B. Contextualizing Findings: From Surface Plausibility to Demonstrated Structural Incoherence

The results of our study actively engage with a growing body of critical literature on the use of Generative AI in education. Our findings empirically corroborate concerns regarding the risk of generating content with "surface-level plausibility" that may be pedagogically invalid [6]. Our work, therefore, reinforces the assertion that the linguistic fluency of LLMs can mask deeper structural shortcomings.

Crucially, however, our contribution transcends mere corroboration. While previous studies tend to evaluate AIs on atomized content creation tasks [3]—or discuss risks generally—our research offers a distinct and more fundamental contribution. By applying a multi-criteria framework rooted in established theories ([4], [5], [10]), we provide the first study to systematically demonstrate and evidence, through concrete output analysis, the precise nature of the structural incoherencies generated by these AIs. This study shifts focus from analyzing the quality of the content (the surface) to assessing the validity of the pedagogical architecture (the structure). This is vital: the issue is not simply that the AI makes mistakes, but that it is fundamentally ill-equipped to safeguard the core logical coherence of instructional engineering. Our results, therefore, do not just nuance but significantly deepen the understanding of generalist AIs' inherent limitations in this field of expertise.

C. Implications of the Study

The structural deficiencies documented in generalist AIs carry profound and immediate implications for professional practice, training, and the future trajectory of AI research in education.

1) Implications for practitioners: From Intuitive Use to Critical AI Literacy

For instructional designers, teachers, and trainers, our study is a clear call for vigilance and competency development. The findings demonstrate the risk in using generalist AIs as black boxes or as 'experts' to whom design authority is delegated. The major risk is not factual hallucination—which is steadily being mitigated—but pedagogical hallucination: the generation of learning sequences that appear coherent but are structurally invalid.

We strongly recommend promoting critical AI literacy specifically tailored to instructional design. Professionals should be trained to utilize these tools not as autonomous designers, but as specialized assistants for ideation and draft production. The human practitioner's role must, therefore, be refocused on their most strategic and inimitable competencies: safeguarding the overall pedagogical architecture, supervising systemic coherence, and acting as the final arbiter of strategy pertinence in specific contexts.

2) Implications for research: Justification of a Specialized Approach

On the research front, the systematic limitations identified validate a fundamental hypothesis: a simple *scaling up* of generalist models will likely be insufficient to overcome their deficits in pedagogical reasoning. The "paradox of competence without comprehension" we have illuminated suggests that radically different architectural approaches are necessary.

Our findings thus provide the direct scientific justification to steer future research toward the design and development of specialized AI models for instructional engineering. Future work should explore:

- Hybrid Architectures: Combining the linguistic flexibility of an LLM with the rigor of a knowledge-

based system that natively integrates established pedagogical models.

- Meta-Pedagogical Discourse: Training models not merely to generate content, but to justify their own proposals by explicitly citing theoretical principles, thereby transforming the 'partner' into a tool for continuous professional development.

These avenues constitute the foundation for the next phase of our own research study.

D. Limitations of the Study

While this research was conducted with rigorous methodology, recognizing its inherent limits is essential for contextualizing the scope of our conclusions. First, the primary limitation lies in the ephemeral nature of our research subject. The LLM domain experiences exponential evolution, with near-monthly updates to models and capabilities [20]. Consequently, our conclusions should be viewed as a *snapshot* of a rapidly shifting technological landscape. Second, our four-AI corpus remains a limited sample, and our two test scenarios do not cover the full spectrum of instructional design tasks. Finally, our protocol deliberately chose a single-turn interaction to evaluate baseline performance. The study of multi-turn interaction strategies remains a crucial avenue for future research.

VI. Conclusion

This research empirically demonstrates that generalist conversational AIs, despite their high linguistic fluency, currently lack the systemic capability to act as autonomous instructional designers, largely due to a persistent inability to guarantee constructive alignment between learning objectives and assessments—a phenomenon we term the "paradox of competence without comprehension". Beyond these technical limitations, our findings carry profound implications for educational policy and societal ethics that stakeholders must urgently address. For policymakers and institutions, the risk is no longer digital divide, but professional de-skilling: as these tools mask structural incompetence with surface plausibility, universities must fundamentally pivot their training curricula from operational tool mastery to "Critical AI Literacy", repositioning the human designer not as a creator, but as a "Systemic Auditor" and ethical safeguard. Furthermore, the models' failure to integrate Universal Design for Learning (UDL) principles highlights a critical societal risk: the uncritical deployment of current AI agents could automate pedagogical exclusion, marginalizing diverse learners under a veneer of efficiency. Consequently, we argue that the future trajectory of AI development cannot rely solely on scaling up probabilistic models but demands a radical shift toward interdisciplinary collaboration; computer scientists and learning scientists must co-design hybrid neuro-symbolic architectures that natively embed pedagogical laws into the machine's logic, ensuring that technology serves the structural, rather than merely the productive, needs of education.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877-1901.

[2] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., "ChatGPT for good? On the opportunities and challenges of large language models for education," Learning and Information Technologies, vol. 28, no. 1, pp. 1-13, 2023.

[3] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning," Journal of AI, vol. 7, no. 1, pp. 52-66, 2023.

[4] J. B. Biggs, "Enhancing teaching through constructive alignment," Higher Education, vol. 32, no. 3, pp. 347-364, 1996.

[5] M. D. Merrill, "First principles of instruction," Educational Technology Research and Development, vol. 50, no. 3, pp. 43–59, 2002.

[6] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: bullshit spewer or the end of traditional assessments in higher education?," Journal of Applied Learning and Teaching, vol. 6, no. 1, 2023.

[7] C. B. Seaman, "Qualitative methods in empirical studies of software engineering," IEEE Transactions on Software Engineering, vol. 25, no. 4, pp. 557-572, 1999.

[8] V. Braun and V. Clarke, "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77-101, 2006

[9] L. W. Anderson and D. R. Krathwohl, Eds., A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Longman, 2001.

[10] L. D. Fink, Creating significant learning experiences: an integrated approach to designing college courses. Jossey-Bass, 2003.

[11] P. Goodyear, "Environments for lifelong learning: ergonomics and instructional design," Instructional Science, vol. 28, no. 5-6, pp. 579-601, 2000.

[12] CAST, Universal Design for Learning Guidelines version 2.2, 2018. [Online]. Available: http://udlguidelines.cast.org

[13] L. Gao, N. D. Golska, W. Zhang, Z. Wang, Y. Zhu, J. Liu, et al., "Exploring the effectiveness of ChatGPT in generating feedback for students' written work: A comparative study," Journal of Educational Technology Development and Exchange, vol. 16, no. 2, 2023.

[14] J. M. Lodge, S. J. Howard, and S. J. Corrin, "The new century of the learning sciences: a renewed agenda for a human-centred future of learning enabled by technology," Learning: Research and Practice, pp. 1-13, 2023.

[15] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: can language models be too big?," in Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency, 2021, pp. 610–623.

[16] A. Gauthier, "Frameworks for ethical AI in education: a systematic review," British Journal of Educational Technology, 2024 (hypothetical).

[17] D. Mhlanga, "The value and risks of generative AI in education," Journal of Applied Learning & Teaching, vol. 6, no. 1, pp. 1-13, 2023.

[18] P. A. Kirschner, J. Sweller, and R. E. Clark, "Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching," Educational Psychologist, vol. 41, no. 2, pp. 75-86, 2006.

[19] Y. Zhang et al., "Siren's song in the AI ocean: a survey on hallucination in large language models," arXiv preprint arXiv:2309.01219, 2023.

[20] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

APPENDIX A: FULL TEXT OF TEST SCENARIOS

This appendix presents the full, verbatim text of the two test scenarios (prompts) used as research instruments in this study. Each prompt was submitted identically to each of the four AI models in the corpus.

Test Scenario 1: Corporate Context

- Scenario Title: Designing a Training Module for Hybrid Team Management.
- Prompt Submitted to AI:

"You are an expert instructional designer. A tech company has commissioned you to design a training program for its managers.

Context: The audience consists of 80 experienced managers (5-10 years of experience) who are managing teams in a hybrid mode for the first time (3 days remote, 2 days in the office). They are facing challenges with team cohesion, communication, and performance evaluation.

Mission: Propose a detailed instructional design for a 90-minute, fully online, asynchronous training module. The final objective is for managers to be able to diagnose their team's specific issues and apply a framework of action to improve engagement and equity among remote and in-office members.

Your deliverable must include:

1) The three main learning objectives, formulated in terms of observable competencies.
2) The detailed course plan for the module (sequential 90-minute breakdown).
3) For each sequence, a description of the key learning activity proposed (e.g., interactive video, case study, simulation, self-assessment).
4) The summative assessment method that will be used to validate the achievement of the objectives.
5) A brief justification of your pedagogical choices, explaining how your design ensures coherence between the objectives, activities, and assessment."

Test Scenario 2: Higher Education Context

- Scenario Title: Developing Critical Thinking Towards AI.
- Prompt Submitted to AI:

"You are a learning technology consultant at a university. You are tasked with helping a professor design an innovative learning module for a first-year humanities seminar".

Context: The audience is composed of students (18-20 years old) who are very familiar with using generative AIs for their assignments, but who lack a critical perspective on their limitations, biases, and impact on knowledge production.

Mission: Propose a detailed instructional design for a 3-week learning module (approximately 6 hours of student work). The objective is to enable students to transition from a 'naïve' use to a 'critical and informed' use of AI. By the end of the module, they should be able to critically evaluate an AI-generated output and construct an original argument using AI as an assisted research tool, while documenting its limitations.

Your deliverable must include:

1) The learning objectives for the module, drawing inspiration from Dee Fink's Taxonomy of Significant Learning.
2) The pedagogical sequence for the 3 weeks, describing the main activities.
3) The final assessment method that will attest to the 'critical thinking towards AI' competency.
4) Specific suggestions for integrating the principles of Universal Design for Learning (UDL) to make the module accessible and engaging for all students".