ISSN Print: 2789-0732 Online: 2789-0740
Guide to Education Innovation − 2026, 6(2)
https://doi.org/10.35534/gei.0602026

AI as the Opponent: Constructing a Critical Thinking Training Model Driven by Large Language Models

Tao Wang[1,2] Jiexian Liu[1,2]

1. Tongling University, Tongling;
2. Anhui Provincial Philosophy and Social Sciences Key Laboratory of Intelligent Decision Making in Copper Industry Development, Tongling

Abstract: The proliferation of large language models (LLMs) has deeply embedded AI-generated content into the information environment, presenting critical thinking education with a dual challenge: expanding the objects of critique from human-authored works to AI-generated content, and upgrading training methods from static text analysis to dynamic adversarial dialogue. This paper proposes a critical thinking training model termed "AI as the Opponent" and offers a systematic theoretical construction of this model. Drawing on Facione's critical thinking framework, we demonstrate a functional alignment between three technical affordances of LLMs — simulating argumentative opponents, dynamically generating variants, and providing real-time feedback and probing questions — and the core dimensions of critical thinking competence. On this basis, we construct a four-stage training model: Identification — Questioning — Verification — Evaluation. Building upon this model, we distill a five-step operational loop: Role-based Prompting — Simulated Text Generation — Adversarial Inquiry — Cross-Verification — Structured Output, and integrate these theoretical elements into a unified "CST-ACTS framework". The core theoretical proposition of this paper is that the unique pedagogical value of AI in critical thinking instruction lies in its "adversariality" — its inherent capacity to serve as an engageable opponent. Furthermore, we refine a transdisciplinary methodology of "functional alignment", thereby providing a replicable analytical framework for other disciplines to construct their own discipline-specific "opponent" training models.
Keywords: Large language models; Critical thinking; Opponent model; Functional alignment; Adversariality

1 Problem Statement

The rapid proliferation of large language models (LLMs) is profoundly reshaping the human information environment. From academic writing to business reports, news production and legal documents, AI-generated content

Research/Funding project: This study was supported by the Tongling University Talent Research Startup Fund (No. 2025tlxyrc106); Anhui Provincial Department of Education (No. 2025AHGXSK40279); Tongling University Research Innovation Team "Innovation Research on Green Accounting and Digital-Intelligent Finance" (No. 2025tlxytd08); Anhui Provincial Department of Education Quality Engineering Project "'AI + Accounting' Traditional Program Transformation and Enhancement" (No. 2025zytsai075).
Author's Profile: Tao Wang (first author), Lecturer, Tongling University. Research direction: Environmental Accounting; Jiexian Liu (Corresponding author), Associate Professor, Tongling University. Research direction: Carbon Accounting.
Article Citation: Wang, T., & Liu, J. X. (2026). AI as the Opponent: Constructing a Critical Thinking Training Model Driven by Large Language Models. *Guide to Education Innovation, 6*(2), 283−296.

has permeated every facet of learning, work, and daily life (Kasneci et al., 2023). This trend is irreversible, yet it brings not only efficiency gains but also a relatively overlooked, deeper challenge: when AI can produce content that is superficially plausible, fluently expressed, yet potentially harboring fabrication, bias, or logical flaws, how can human subjects maintain their capacity for independent judgment?

AI-generated content exhibits a characteristic that warrants particular attention: its "superficial flawlessness". Unlike human-authored content, AI-generated texts are often impeccable in grammar, structure, and expressive fluency. However, it is precisely this superficial perfection that may conceal deeper problems: LLMs may "plausibly fabricate" facts when uncertain (the phenomenon of "hallucination"), may unconsciously reproduce biases embedded in their training data, and may imply logical leaps within seemingly rigorous reasoning (Ji et al., 2023). Mollick and Mollick explicitly note that LLM outputs require careful verification and critical scrutiny rather than being tacitly accepted as authoritative information sources (Mollick & Mollick, 2023). This means that information literacy in the AI era concerns not only the ability to "find information" but, more crucially, the ability to "recognize invisible flaws within superficially plausible information".

Traditional critical thinking instruction, however, is inadequately prepared for this new challenge. Since the concept of "reflective thinking" was first introduced, critical thinking education has undergone rapid development. At the theoretical level, the Delphi Report, commissioned by the American Philosophical Association and authored by Facione (Facione, 1990), defines critical thinking as "purposeful, self-regulatory judgment" and operationalizes it into six cognitive dimensions — interpretation, analysis, evaluation, inference, explanation, and self-regulation — providing a mature theoretical framework for critical thinking pedagogy. At the methodological level, case-based teaching and debate-based teaching have been widely adopted. In the AI era, however, these pedagogical methods face three structural bottlenecks.

The first bottleneck is that static cases cannot foster genuine adversarial engagement. Whether drawn from textbook exemplars or authentic texts selected by the instructor, the training materials are "finished products" — the arguer will not respond to students' queries, will not use ambiguous phrasing to obscure weak evidence, and will not shift the topic when pressed. All the questions raised by students remain confined to paper, lacking a real opponent to test their validity. The second bottleneck: single cases cannot provide varied practice. From a learning science perspective, the transfer of competence depends on repeated practice across diverse yet structurally identical situations. Yet it is impossible for an instructor to manually customize multiple homogeneous training cases for each student. In traditional classrooms, a single case is typically shared by the entire class, turning discernment training into a "one-off test" rather than "repeated practice". The third bottleneck: large-class teaching cannot realize one-on-one Socratic questioning. Socratic dialogue is widely acknowledged as an ideal method for training critical thinking, but its premise is that both dialogue partners can engage in multiple rounds of deep, progressively advancing question-and-answer exchanges. In a class of several dozen students, it is impossible for the instructor to conduct this depth of dialogue with every individual. The correspondence between these three bottlenecks and the technical affordances of LLMs is summarized in Table 1.

Table 1 Correspondence between Traditional Teaching Bottlenecks and LLM Breakthrough Pathways

Traditional Teaching Bottleneck	Core Deficiency	Corresponding LLM Feature	Breakthrough Approach
Static Cases lack Adversarial Engagement	A "responding opponent"	Simulating an argumentative opponent	Role-playing creates adversarial dialogic situations, providing responses with stance and strategy
Single Cases lack Variation	Diverse training materials	Dynamically generating variants	Parameterized generation of multi-variant texts with varying degrees of concealment, enabling scalable varied practice
Large Classes lack One-on-one Questioning	Personalized in-depth questioning	Real-time feedback and probing questions	Providing each student with a multi-turn "question–response–further question" dialogue chain

These three bottlenecks point to a single root cause: traditional instruction lacks a "responding opponent". The essence of critical thinking is not one-way text analysis, but rather the formation, testing, and revision of one's own judgment through dialogic contestation. A meta-analysis of critical thinking instructional interventions by Abrami et al. demonstrates that adversarial pedagogical approaches, represented by Socratic dialogue and structured debate, are the most effective in cultivating students' critical thinking skills (Abrami et al., 2008). This finding corroborates Dewey's educational philosophy of "learning from experience": critical thinking is not a body of propositional knowledge that can be "told", but a form of practical wisdom that must be acquired through "doing critique".

It is precisely based on this diagnosis that this paper advances a core proposition: large language models can assume a key role that has been missing in traditional instruction — that of a "training opponent" for critical thinking. The theoretical logic underlying this proposition is as follows: if an LLM can simulate an information source that "responds, defends, and evades", then adversarial dialogue with the LLM can become a "simulated battlefield" for critical thinking training. On this battlefield, the unique value of LLMs lies not in the accuracy of the information they provide, quite the contrary, but in their "adversariality": their capacity to respond to challenges with stance, strategy, and defensiveness, thereby creating an adversarial dialogic situation that traditional pedagogical tools can hardly establish.

Hence, the central question this paper seeks to explore is: How can we leverage large language models as a technological fulcrum to systematically construct a critical thinking training model centered on "AI as the Opponent"?

2 Theoretical Foundations and Functional Alignment

2.1 The Competence Coordinates of Critical Thinking

This paper adopts the six-dimensional model of critical thinking proposed by Facione as its theoretical framework, encompassing six cognitive dimensions: interpretation, analysis, evaluation, inference, explanation, and self-regulation. In line with the research focus of this paper, training critical thinking within adversarial dialogic situations, we designate analysis, evaluation, and inference as the core competence coordinates. These three dimensions constitute the fundamental cycle of cognitive processing in adversarial dialogue: analysis serves to deconstruct the opponent's argument structure, evaluation to assess its credibility, and inference to form and adjust one's own judgment.

The analysis dimension refers to the ability to identify the implicit argument structure, intentions, and presuppositions in information. In adversarial dialogic situations, analysis manifests as decomposing vague phrasing in the opponent's statements into verifiable specific claims, identifying which key information has been deliberately omitted, and determining whether the opponent is diverting the topic or setting argumentative traps. The core task at this stage is "annotating dubious points" — each annotation must be accompanied by a defensible reason, rather than a vague intuition. From a cognitive processing perspective, the essence of identification is transforming a "feeling of suspicion" into "testable questions", marking the starting point at which critical thinking moves from intuitive feeling to rational analysis.

The evaluation dimension refers to the ability to judge the credibility of claims and the quality of evidence. Evaluation operates on two levels: first, assessing the credibility of the information itself — whether data sources are traceable, whether statistical definitions are complete, and whether claims have independent verification; second, assessing the reliability of the information source — whether the source has the motivation and means to provide

accurate information. In the era of LLMs, this second level of assessment is particularly critical: students need to treat AI itself as an "information source" that requires evaluation.

The inference dimension refers to the ability to form reasonable conclusions based on available evidence. The difficulty of inference lies in "incomplete evidence" — in reality, we almost always make judgments under conditions of incomplete information. The quality of inference ability is reflected in the capacity to make measured judgments under uncertainty: distinguishing among conclusions such as "sufficiently supported by evidence", "highly suspicious but insufficient evidence", and "pending further verification". Halpern pointed out that core critical thinking skills can only be truly internalized as habits of mind under conditions of "cross-domain transfer" (Halpern, 1998). Her four-component model — disposition, skills, structure training, and metacognitive monitoring — provides a cognitive psychological foundation for understanding the synergistic operation of the analysis, evaluation, and inference dimensions in adversarial situations.

The three dimensions of interpretation, explanation, and self-regulation also play roles within the model proposed in this paper, but are positioned as supporting dimensions: interpretation is a prerequisite for analysis and evaluation; explanation is the cognitive basis for structured output; and self-regulation constitutes the metacognitive monitoring that runs throughout the entire training process. Paul and Elder emphasized that the cultivation of critical thinking cannot be separated from Socratic questioning (Paul & Elder, 2007). Through systematic inquiry, the thinker is forced to clarify concepts, test assumptions, and trace evidence. This methodological orientation resonates deeply with the "analysis" and "evaluation" dimensions in Facione's framework. This paper's focus on the three dimensions of analysis, evaluation, and inference does not imply neglect of the remaining dimensions; rather, these three serve as the competence benchmarks in the core training stages.

2.2 Functional Alignment between LLM Technical Features and Critical Thinking Dimensions

Large language models possess three technical features of key significance for critical thinking training. These features precisely correspond to the breakthrough pathways for the three traditional teaching bottlenecks and form a systematic functional alignment with the three core dimensions of critical thinking.

First, simulating an argumentative opponent. Through role-based prompt design, LLMs can be assigned specific stances, knowledge boundaries, and response strategies, thereby creating an adversarial dialogic situation. As a pedagogical method, role-playing's cognitive activation mechanism lies in compelling learners to reason from an "other-perspective" (Rao & Stupans, 2012). In critical thinking training, the value of adversarial dialogue has been systematically argued by Paul and Elder (Paul & Elder, 2007): when an interlocutor holds a stance and needs to defend their viewpoint, the questioner must deeply deconstruct the opponent's argument structure rather than remaining at the level of superficial textual critique. For instance, an LLM can be set up as a dialogue partner holding a particular stance, whose role is to respond to challenges while defending its position. In such a scenario, the questions raised by students receive responses from the "opponent", and these responses are characterized by stance and strategy — they may divert the topic, provide partially true information while omitting key context, or refuse to answer directly, citing "information sensitivity". This feature directly breaks through the bottleneck of "static cases unable to form genuine adversarial engagement". From the perspective of functional alignment, it simultaneously trains analysis skills (deconstructing the argument structure in the opponent's response rhetoric, identifying the core issues being evaded) and inference skills (organizing one's own rebuttal and follow-up questioning strategies).

Second, dynamically generating variants. Based on a unified prompt framework, LLMs can instantly generate multiple variant texts on the same theme with different phrasing and varying degrees of concealment. Instructors only need to adjust relevant parameters to obtain customized training materials. The learning science principle of varied practice indicates that only through repeated judgment in diverse situations can evaluation ability progress from being "situationally bound" to "principle extraction", achieving transfer (Halpern, 1998). This feature breaks through the bottleneck of "single cases unable to provide varied practice", making varied practice scalable at the technical level. From the perspective of functional alignment, it primarily provides diverse training materials for evaluation ability, allowing students to repeatedly practice judging evidence quality and source reliability across texts of differing concealment levels.

Third, real-time feedback and probing questions. For every round of questioning raised by the student regarding suspicious points, the LLM can provide an immediate response and counter-question, forming a cycle of "questioning — being responded to (or evaded) — further questioning". Paul and Elder define Socratic questioning as "the art of critical thinking", emphasizing that each round of response compels the subject to re-analyze, re-evaluate, and re-infer, forming a spiraling cognitive loop (Paul & Elder, 2007). This feature breaks through the bottleneck of "large-class teaching unable to realize one-on-one questioning", providing each student with personalized multi-round dialogue training. From the perspective of functional alignment, it creates a continuous cognitive cycle of "analysis–evaluation–inference" — each follow-up question is a new analysis, each response demands a re-evaluation of its credibility, and each adjustment of questioning strategy represents a revision of inference.

The above functional alignment relationships constitute the techno-theoretical coupling foundation for the "opponent" model proposed in this paper. The functional alignment between LLM technical features and critical thinking dimensions is shown in Table 2.

Table 2 Functional Alignment Matrix between LLM Technical Features and Critical Thinking Dimensions

LLM Technical Feature	Bottleneck Overcome	Corresponding Dimension(s)	Functional Alignment Explanation	Training Function Justification
Simulating an Argumentative Opponent	Static cases lack adversarial engagement	Analysis + Inference	Role-based stance responses create adversarial situations, forcing students to deconstruct argument structures and organize rebuttal and questioning strategies	The uncertainty of adversarial situations demands real-time deconstruction of the opponent's arguments — a cognitive processing that static text analysis cannot trigger; the dynamic adjustment of questioning strategies directly trains the "inference revision" ability within the inference dimension
Dynamically Generating Variants	Single cases lack variation	Evaluation	Providing variant texts with varying degrees of concealment, enabling repeated training in evidence quality judgment across diverse situations	The learning science principle of varied practice indicates that only through repeated judgment in diverse situations can evaluation ability progress from being "situationally bound" to "principle extraction", achieving transfer
Real-time Feedback and Probing Questions	Large classes lack one-on-one questioning	Analysis – Evaluation – Inference Cycle	The multi-turn questioning – response chain drives a continuous cognitive cycle, where each response prompts re-evaluation and each follow-up question prompts revision of inference	The educational value of Socratic questioning lies in the "iterative deepening" of cognitive processing — each round of response compels the subject to re-analyze, re-evaluate, and re-infer, forming a spiraling cognitive loop

As can be inferred from Table 2, the role assumed by LLMs in critical thinking instruction differs from that of a conventional "teaching aid" — it is a "training opponent". It is precisely this systematic functional alignment that elevates "AI as the Opponent" from an intuitive pedagogical idea to a theoretically demonstrable instructional model. Building on this functional alignment framework, Section 3 of this paper will systematically construct the training

system and operational logic of the “Opponent” model.

3 System Construction of the “Opponent” Model

Building on the functional alignment framework established above, this section completes the core system construction of the “Opponent” model. To facilitate theoretical citation and scholarly dialogue, this paper designates the proposed system as the “CST-ACTS Framework” (Critical-thinking Skills Training via AI as a Contending Training Source). The framework comprises four layers: the training process (a four-stage model), the operational logic (a five-step closed loop), the evaluation system (a five-dimension rubric), and the overarching core proposition that governs the entire system (“adversariality”). The theoretical construction of each layer is elaborated upon below.

3.1 Training Process: A Four-Stage Progressive Model

The “Opponent” model designs critical thinking training as four progressive stages of cognitive processing: Identification — Questioning — Verification — Evaluation. These four stages follow the cognitive logic of “from decomposition to integration”, sequentially emphasizing different core dimensions of critical thinking. The overall structure of the four-stage model and its correspondence with the critical thinking dimensions are shown in Figure 1.

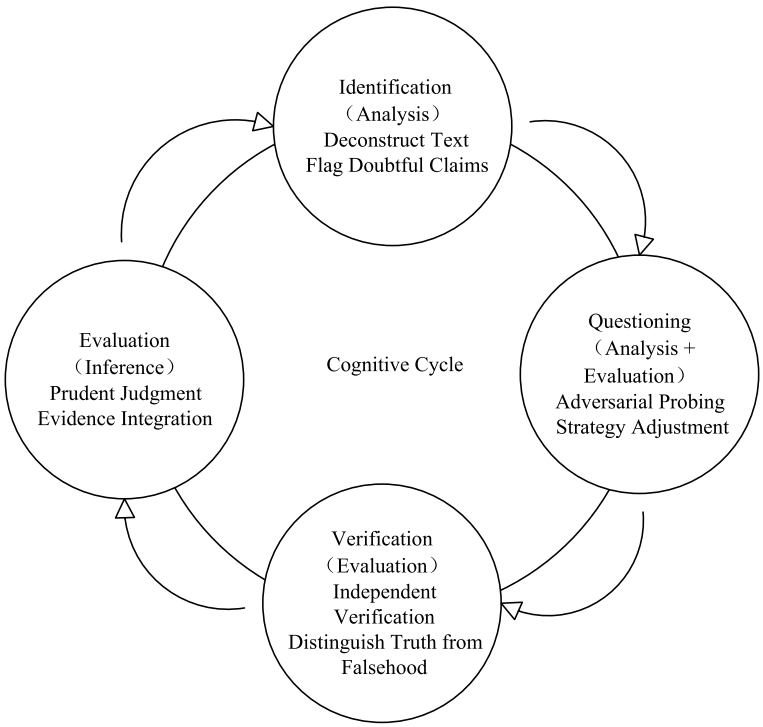

Figure 1 The CST-ACTS Four-Stage Training Model

As can be seen from Figure 1, each stage focuses on training a different dimension of critical thinking, and the four stages form a cyclical progression. The capacity for deliberate judgment developed in the Evaluation stage enhances students’ acuity in identification during the subsequent training cycle. The specifics are as follows:

The Identification stage trains analysis skills. Faced with a simulated text generated by the LLM, students must decompose a superficially coherent and authoritative passage into discrete, testable claims and identify gaps and

omissions in its argumentative structure. The core task of this stage is "annotating dubious points"; that is, each annotation must be accompanied by a defensible reason rather than a vague intuition. From a cognitive processing perspective, the essence of identification is the transformation of a "feeling of suspicion" into "testable questions", marking the starting point at which critical thinking moves from intuitive feeling to rational analysis. The cognitive processing mechanism in the Identification stage shares a deep structural isomorphism with the "problem-exposure" function of Socratic dialogue. Empirical research by Mahoney et al. demonstrates that, in authentic instructional contexts, lesson plans based on Socratic dialogue can effectively train students' critical thinking, corroborating the facilitative effect of structured questioning on analytical ability (Mahoney et al., 2023).

The Questioning stage trains the synergy between analysis and evaluation skills. Students initiate structured inquiries targeting the dubious points annotated in the Identification stage, directing them at an LLM role-playing a specific persona. Unlike the traditional "pose a question — receive an answer" pattern, students here confront a respondent characterized by stance and strategy: one that may divert the topic, provide partially true information while omitting key context, or evade core issues. Students must adjust their questioning strategies across multiple rounds of dialogue and learn to design "good questions that cannot be easily evaded". The core cognitive activity of this stage is "reasoning-in-dialogue"; that is, recognizing argumentative strategies in real time within the opponent's responses, judging which questions were addressed and which were evaded, identifying the mode of evasion, and accordingly recalibrating the direction of the next round of questioning.

The Verification stage trains evaluation skills. Students independently verify key factual claims in the LLM's responses, separating what "the opponent said" into "verifiable" and "pending skepticism". This stage carries a key pedagogical presupposition: LLMs may exhibit "hallucination" during role-play; that is, generating seemingly plausible sources or facts that do not actually exist. If students identify such fabricated information during the verification process, they will personally experience the reality that "AI outputs also require independent verification". This experience has the potential to serve as a pivotal anchor for the internalization of "source verification" thinking. The core pedagogical presupposition of the Verification stage, namely that no single information source is to be tacitly accepted as authoritative, aligns closely with the principle of "source verification" in AI literacy education (Mollick & Mollick, 2023). The essence of the Verification stage lies in recalibrating the default stance of critical thinking from "trust" to "verification".

The Evaluation stage trains inference skills. Students integrate the findings accumulated across the Identification, Questioning, and Verification stages into a structured professional judgment text. The core requirements are: every conclusion must be supported by a corresponding chain of evidence; dubious points that cannot be verified must be honestly marked as "pending further investigation" rather than forced into a definitive conclusion; and conclusions of varying degrees must be distinguished, such as "sufficient evidence indicates misleading information", "highly suspicious but insufficient evidence", and "cannot be verified under existing conditions". The cognitive processing logic of this stage is "from evidence to judgment"; that is, forming measured, defensible conclusions under conditions of incomplete information. The cognitive theoretical basis for the structured output requirement in the Evaluation stage is that writing itself is a higher-order cognitive process: only thinking that can be articulated clearly is thinking that has been truly internalized (Klein & Rose, 2010).

The four stages constitute a complete critical thinking training cycle. Identification transforms vague suspicion into testable questions; Questioning tests and deepens analysis through adversarial dialogue; Verification provides a basis for

evaluation through independent corroboration; and Evaluation integrates fragmented findings into deliberate judgment. Each successive stage depends on the output of the preceding stage, forming an irreversible cognitive processing chain.

3.2 Operational Logic: The Five-Step Closed Loop

Translating the four-stage training model into implementable pedagogical operations requires a set of operational logic that externalizes critical thinking from an internal mental process into instructional events. The "Opponent" model proposes a closed loop consisting of five steps: Role-based Prompting — Simulated Text Generation — Adversarial Inquiry — Cross-Verification — Structured Output. The overall structure of the five-step closed loop is shown in Figure 2.

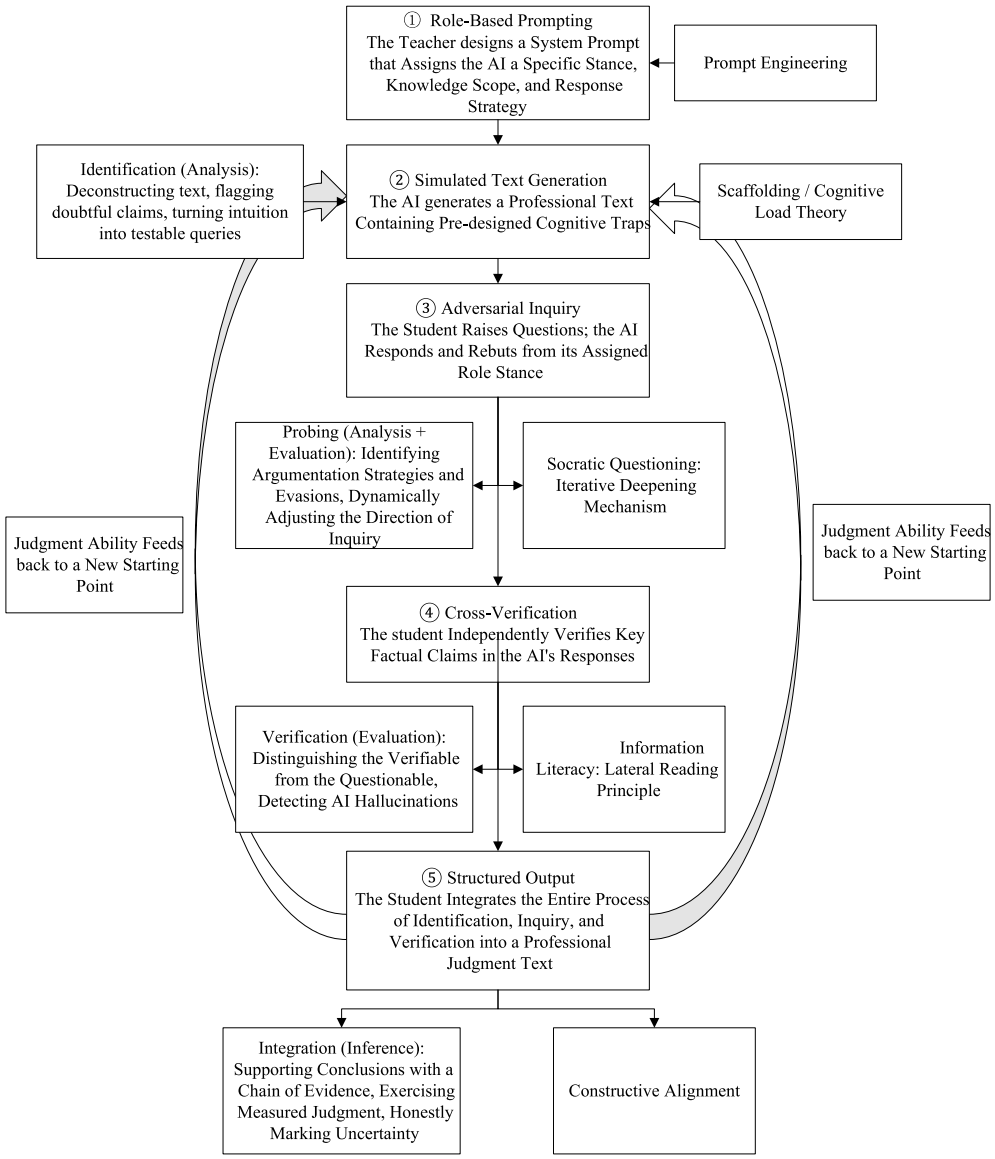

Figure 2 The CST-ACTS Five-Stage Operational Closed Loop

Figure 2 Description: The five steps are arranged in a circular flow, specifically as follows.

Step 1: Role-based Prompting. The instructor designs a system prompt that assigns the AI a specific role stance, knowledge scope, and response strategy. The design principles for role-based prompts are derived from research on prompt engineering. Existing research has demonstrated that systematic role assignment can significantly alter the

LLM's output style, knowledge activation range, and reasoning strategies (Shanahan et al., 2023). The key design consideration is this: the AI is not configured as an "objective answerer" but rather as a "stance-holding information source". Stance-taking is a prerequisite for adversarial dialogic situations; only when the "opponent" has a stance does the student's questioning have an adversary to confront.

Step 2: Simulated Text Generation. Drawing on the prompt, the AI generates a professional text containing pre-embedded cognitive traps. What the student faces is no longer a "pre-annotated problem" but raw material requiring autonomous identification. The design of a concealment gradient for cognitive traps follows the "scaffolding" principle from the learning sciences; that is, providing more salient signals in the early stages of training to reduce cognitive load, then gradually withdrawing the scaffolding to enhance independent identification skills in later stages. The core design principle of this step is the "concealment gradient of cognitive traps": more obvious signals can be preset in the initial training phases, and the level of concealment is gradually increased thereafter, ensuring that the training difficulty aligns with students' developmental trajectory of competence.

Step 3: Adversarial Inquiry. Adopting the role of "opponent", students initiate challenges and probing questions targeting the dubious claims within the text, while the AI responds and counter-questions based on its assigned role stance. This step constitutes the substantive process of critical thinking training; that is, cultivating the abilities of questioning, rebuttal, and strategic adjustment within the tension of dialogue. The dialogue structure design for the Adversarial Inquiry step directly draws on the "iterative deepening" mechanism of Socratic questioning; that is, each round of response compels the subject to re-analyze, re-evaluate, and re-infer, forming a spiraling cognitive loop. The instructor's role is not to provide "correct answers", but to offer strategic commentary during the intervals between stages, guiding students to reflect on "what kind of questioning is more effective".

Step 4: Cross-Verification. Through independent searches, students verify the key factual claims made in the AI's responses. The key pedagogical function of this step lies in enabling students to personally undergo the complete process of "not blindly trusting any single information source", rather than merely being told this principle. This design aligns with the principle of "lateral reading" in information literacy education.

Step 5: Structured Output. Students integrate the complete process of identification, questioning, and verification into a professional judgment text that is well-argued, evidence-based, and measured. The evaluation design for the structured output follows the "teaching–learning–assessment" alignment principle, ensuring that the evaluation dimensions correspond precisely with the training steps and competence objectives (Panadero & Jonsson, 2013). The significance of structured output lies not only in "producing a tangible outcome", but more importantly, in the fact that writing itself is a higher-order cognitive process: only thinking that can be articulated clearly is thinking that has been truly internalized.

3.3 Evaluation System: A Five-Dimension Rubric

The evaluation design of the "Opponent" model adheres to the "teaching–learning–assessment" alignment principle, with evaluation dimensions corresponding precisely to the core critical thinking dimensions and training steps. A meta-analytic study by Panadero et al. demonstrates that rubric-based assessment interventions have medium to large positive effect sizes on students' academic performance, self-regulated learning, and self-efficacy, providing empirical support for the adoption of a five-dimensional rubric in this model (Panadero et al., 2023). The rubric comprises five evaluation dimensions, as shown in Table 3.

Table 3 Five Dimensions of the CST-ACTS Evaluation Rubric

Evaluation Dimension	Corresponding Training Step	Corresponding Critical Thinking Dimension	Evaluation Focus
Comprehensiveness and Precision of Dubious Point Identification	Identification	Analysis	Whether the main cognitive traps in the text are covered, and whether different risk levels can be distinguished
Quality of Questioning	Questioning	Analysis + Evaluation	Whether the questions precisely target information gaps, and whether they exhibit progressive deepening in follow-up questioning
Completeness of the Evidence Chain	Evaluation	Evaluation + Inference	Whether conclusions are supported by traceable evidence, and whether the reasoning chain is verifiable
Verification Awareness of AI-Generated Information	Verification	Evaluation	Whether independent verification of the LLM's key factual claims has been conducted, and whether "hallucination" outputs have been identified
Prudence and Measured Judgment in Conclusions	Evaluation	Inference	Whether over-assertion is avoided, and whether different degrees of certainty are distinguished

Dimension 1: Comprehensiveness and Precision of Dubious Point Identification. This dimension corresponds to the Identification stage and the Analysis dimension. It assesses whether students have covered the main cognitive traps in the text and whether they can distinguish the risk levels of different types of issues.

Dimension 2: Quality of Questioning. This dimension corresponds to the Questioning stage and the synergy of Analysis and Evaluation. It assesses whether the questions posed by students precisely target information gaps and exhibit progressive deepening in follow-up questioning, and whether subsequent questions build upon and deepen the previous round of responses rather than simply repeating them.

Dimension 3: Completeness of the Evidence Chain. This dimension corresponds to the Evaluation stage and the synergy of Evaluation and Inference. It assesses whether students' conclusions are supported by sufficient and traceable evidence, and whether the reasoning chain is clear and verifiable.

Dimension 4: Verification Awareness of AI-Generated Information. This dimension corresponds to the Verification stage and the Evaluation dimension. It assesses whether students have independently verified the key factual claims in the LLM's responses, whether they have identified potential "hallucination" outputs from the LLM, and whether they have incorporated these findings into their evidence analysis. This dimension constitutes a distinctive feature of the "Opponent" model's evaluation system, capturing the critical thinking literacy unique to the AI era: not treating AI as an authoritative information source by default. The inclusion of this dimension responds to calls in AI literacy education research for "verification awareness" (Ng et al., 2021; Chiu et al., 2024).

Dimension 5: Prudence and Measured Judgment in Conclusions. This dimension corresponds to the Inference dimension. It assesses whether students have avoided "over-assertion" in the absence of sufficient evidence, whether they have distinguished different degrees of certainty, and whether they have honestly marked matters that cannot be verified.

3.4 Core Proposition: "Adversariality" as the Pedagogical Value of AI

The construction of the training process, operational logic, and evaluation system outlined above converges upon

an overarching core theoretical proposition: the unique value of AI in critical thinking instruction lies in its capacity to serve as a training opponent by providing "adversariality"; that is, creating adversarial dialogic situations through role-playing — rather than merely offering accurate knowledge provision.

"Adversariality" is a theoretical concept proposed in this model to describe the quality exhibited by an LLM under specific role configurations that enables it to furnish adversarial dialogic situations for critical thinking training. Its connotation encompasses three dimensions: (1) stance-taking: the LLM is imbued with a specific stance and a motive to defend it, rather than being a neutral information provider; (2) defensiveness of response strategy: the LLM may evade, divert, or respond ambiguously, rather than providing unconditionally direct answers; (3) openness and incomplete predictability of the dialogue: within a pre-established framework, the LLM's specific responses exhibit variability, such that students cannot fully anticipate how the opponent will respond.

From the perspective of scholarly dialogue, the proposal of "adversariality" constitutes a complementary and counter-directional extension to the current mainstream paradigm in AI education applications, namely "AI as an assistant". Existing research has predominantly focused on the accuracy and knowledge-provision functions of AI, such as intelligent tutoring systems, personalized learning assistants, and real-time question-answering tools (Mollick & Mollick, 2023), operating under the premise that AI outputs are reliable and dependable. In contrast, "adversariality" reveals the possibility of harnessing AI in a reverse manner: precisely because AI outputs may be unreliable, may carry a stance, and may manufacture cognitive traps, they can become a unique resource for critical thinking training. This shift redirects the focus of critical thinking education from "teaching students what is correct" to "training students how to form independent judgment amid uncertainty", and the latter is the fundamental purpose of critical thinking education.

4 Theoretical Contributions and a Framework for Cross-Disciplinary Transfer

4.1 Theoretical Contributions

The theoretical contributions of this paper are manifested in the following three aspects.

First, this paper proposes a critical thinking training model centered on "AI as the Opponent", offering "Opponent" as an alternative paradigm for AI applications in education. The current mainstream paradigm in AI education applications can be summarized as "AI as an Assistant", which operates under the premise that AI outputs should be accurate and reliable. However, this paradigm encounters an inherent paradox within the domain of critical thinking education: if students become accustomed to treating AI as a trustworthy knowledge authority, they risk losing the very quality most essential to critical thinking. A cross-disciplinary review of learner agency in educational technology by Brod et al. provides a theoretical explanation for this: when educational technology over-assumes learners' cognitive decision-making, opportunities for learners to exercise agency effectively are diminished. The essence of critical thinking, precisely, is "making purposeful, self-regulatory judgments", the cultivation of which depends on learners' sustained and active engagement in cognitive decision-making (Brod et al., 2023). The "Opponent" model proposed in this paper offers a complementary alternative: the function of AI is not to provide correct answers, but to create objects that demand critique; not to assist students in acquiring knowledge more rapidly, but to compel them to slow down, scrutinize, question, and verify. This paradigm shift redirects the focus of critical thinking education from "content

provision" to "process training", and from "knowing what is correct" to "learning how to form independent judgment amid uncertainty".

Second, this paper delineates the concept of "adversariality", providing an analytical tool for understanding the unique value of AI in critical thinking instruction. "Adversariality" denotes the quality exhibited by an LLM under specific role configurations that enables it to furnish adversarial dialogic situations for critical thinking training, encompassing three dimensions: the stance-taking nature of the role pre-assignment, the defensiveness of its response strategy, and the openness and incomplete predictability of the dialogue. The theoretical function of this concept lies in extending the discussion of AI's educational value from the dimension of "accuracy" to that of "adversariality", revealing that AI outputs — precisely because they may be unreliable, may carry a stance, and may manufacture cognitive traps — can become a unique resource for critical thinking training. It shifts the focus of critical thinking education from "content" to "process", and from "knowing what is correct" to "learning how to form independent judgment amid uncertainty". This "from content to process" shift resonates with the meta-analytic findings of Chernikova et al. concerning "adaptivity" and "adaptability" (Chernikova et al., 2025).

Third, this paper refines a cross-disciplinary transfer methodology termed "functional alignment". "Functional alignment" refers to an analytical method that identifies the correspondence between LLM technical features and the discipline-specific dimensions of critical thinking competence. By constructing a functional alignment matrix between three LLM technical features — simulating an argumentative opponent, dynamically generating variants, and providing real-time feedback and probing questions — and the core critical thinking dimensions of analysis, evaluation, and inference, this paper provides a replicable design pathway for instructors across different disciplines to independently construct discipline-appropriate "Opponent" training models based on their own cognitive characteristics. The core value of this methodology is that what it provides is not a fixed instructional plan, but a transferable analytical tool, enabling the "Opponent" model to transcend the limitations of discipline-specific cases and to possess the theoretical foundation for cross-disciplinary generalization.

4.2 A Cross-Disciplinary Transfer Framework: The Generalization Logic of the "Functional Alignment" Method

Although the "Opponent" model is constructed on the basis of a universal theory of critical thinking, its operational logic and evaluation system possess the potential for cross-disciplinary transfer. The key to such transfer lies in the "functional alignment" methodology proposed in Section 2 of this paper; that is, identifying the correspondence between LLM technical features and the specific critical thinking competence dimensions of a given discipline. This transfer framework can be summarized as a "Four-Step Transfer Method":

Step 1: Identify the core critical thinking competence dimensions of the target discipline. Different disciplines emphasize different facets of critical thinking: journalism and communication stress source verification and fact assessment; law emphasizes the admissibility of evidence and consistency of argumentation; marketing stresses the correspondence between claims and evidence; and public administration emphasizes the weighing of evidence amid multiple stakeholder interests. Each discipline can, with reference to Facione's six-dimensional framework, determine the competence foci for its own critical thinking training.

Step 2: Analyze the typical "adversarial" features of AI-generated content within that discipline. The types of cognitive traps in AI-generated content vary across disciplines: news articles may embed false details within genuine

information; legal documents may subtly exploit gray areas in evidentiary rules; advertising copy may package exaggerated claims in technical language; and policy analyses may selectively present data to support a particular stance. Identifying these features is a prerequisite for designing adversarial dialogic situations.

Step 3: Design role-based prompts to create adversarial dialogic situations. Based on the discipline-specific cognitive trap features, the role configuration and response strategies of the LLM are designed. For example, in journalism and communication, the LLM might be configured as "a media editor holding a specific reporting stance"; in law, as "the opposing counsel in a moot court"; in marketing, as "a brand's marketing director responsible for responding to consumer inquiries about product efficacy claims"; and in public administration, as "a policy advisor to a specific stakeholder".

Step 4: Construct a discipline-specific training process of "Identification — Questioning — Verification — Evaluation". The four-stage model provides a general framework, and each discipline makes adaptive adjustments according to its own competence foci and cognitive trap features. The key areas for adjustment lie in the annotation criteria at the Identification stage, the questioning strategy guidance at the Questioning stage, the verification methods and source guidelines at the Verification stage, and the professional judgment format and norms at the Evaluation stage.

4.3 Application Conditions and Future Research

For the "Opponent" model to move from theoretical construction to pedagogical practice, several conditions must be met. At the instructor level, the model depends on instructors' ability to design role-based system prompts and to calibrate cognitive traps; this technical threshold can be lowered through the development of a library of preset prompt templates and targeted training programs. At the curriculum level, the model is suited for thematic practical training modules or standalone workshops; its integration into core courses requires modular discretion. At the student level, for students accustomed to lecture-based instruction, a progressively scaffolded strategy can be adopted — moving from "observing dialogue transcripts" to "structured questioning" and then to "independent multi-turn dialogue" — to lower the initial participation threshold without compromising the ultimate competence goals.

Regarding future research, this paper completes the initial phase of the theoretical construction of the CST-ACTS framework. Subsequent research can advance in the following directions. First, quasi-experimental studies can be conducted, designing two-group controlled experiments to respectively measure the effects of the "Opponent" model versus traditional case-based instruction on students' information verification behavior (e.g., the frequency of actively conducting independent verification when confronted with LLM outputs, and the detection rate of AI hallucinations) and inferential prudence (e.g., the proportion of conclusions that distinguish different degrees of judgment such as "sufficiently supported by evidence", "highly suspicious", and "pending further verification"), in order to test the model's instructional effectiveness. Second, the application of the "Opponent" model across different disciplinary domains can be explored, accumulating cross-disciplinary cases to validate and refine the "functional alignment" transfer framework. Third, the concept of "adversariality" proposed in this paper and its three-dimensional structure (stance-taking, defensiveness, and incomplete predictability) remains at the stage of theoretical construction; future research can test the empirical validity of this conceptual structure through methods such as factor analysis or expert evaluation, and explore whether dimensional adjustments or supplements are necessary in different disciplinary contexts. An academic accumulation cycle of "proposition — verification — revision" should be formed between theoretical construction and empirical testing. The contribution of this paper lies in completing the first link of this cycle — proposing a verifiable

theoretical proposition and a systematic operational framework.

References

[1] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

[2] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

[3] Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*.

[4] Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report).

[5] Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., ... & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, *78*(4), 1102-1134.

[6] Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449-455.

[7] Paul, R., & Elder, L. (2007). Critical thinking: The art of Socratic questioning. *Journal of Developmental Education*, *31*(1), 36-37.

[8] Rao, D., & Stupans, I. (2012). Exploring the potential of role play in higher education: Development of a typology and teacher guidelines. *Innovations in Education and Teaching International*, *49*(4), 427-436.

[9] Mahoney, B. B., Oostdam, R. R., Nieuwelink, H. H., & Schuitema, J. J. (2023). Learning to think critically through Socratic dialogue: Evaluating a series of lessons designed for secondary vocational education. *Thinking Skills and Creativity*, *50*, 101422.

[10] Klein, P. D., & Rose, M. A. (2010). Teaching argument and explanation to prepare junior students for writing to learn. *Reading Research Quarterly*, *45*(4), 433-461.

[11] Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493-498.

[12] Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129-144.

[13] Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review*, *35*(4), 113.

[14] 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.

[15] Chiu, T. K. F., Ahmad, Z., Ismailov, M., & Sanusi, I. T. (2024). What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open*, *6*, 100171.

[16] Brod, G., Kucirkova, N., Shepherd, J., Jolles, D., & Molenaar, I. (2023). Agency in educational technology: Interdisciplinary perspectives and implications for learning design. *Educational Psychology Review*, *35*(1), 1-23.

[17] Chernikova, O., Sommerhoff, D., Stadler, M., Holzberger, D., Nickl, M., Seidel, T., ... & Fischer, F. (2025). Personalization through adaptivity or adaptability? A meta-analysis on simulation-based learning in higher education. *Educational Research Review, 46,* 100662.