

# Cultivating Media Critical Thinking with a Socratic AI Coach: A Methodology

Panayiotis Kapetanakis\*, Ioannis Elissaios Paparigopoulos, Dora Katsamori, Eirini Papachristou,  
Georgios Petasis

Institute of Informatics and Telecommunications, National Centre for Scientific Research (N.C.S.R.)  
“Demokritos” GR-153 10, P.O.BOX 60228, Aghia Paraskevi, Athens, Greece

\*Corresponding Author

DOI: <https://dx.doi.org/10.47772/IJRISS.2025.910000678>

Received: 27 October 2025; Accepted: 01 November 2025; Published: 20 November 2025

## ABSTRACT

The increasing sophistication of disinformation necessitates a fundamental shift from reactive fact-checking to proactive cognitive and affective resilience. In today's hybrid digital ecosystem, manipulative content exploits cognitive shortcuts like motivated reasoning and emotional vulnerabilities to bypass analytical scrutiny. Existing interventions often fail due to insufficient scope, slow speed, and the inert skill problem—where individuals possess critical thinking skills but fail to apply them in emotionally charged, real-world contexts. This concept paper introduces Media Critical Thinking (MCT), a unified pedagogical framework that integrates Media Information Literacy (MIL), critical thinking and manipulation-discernment skills, and critical-thinking dispositions into a single practice. MCT is operationalized through the Socratic AI Coach, a chatbot developed in the EU TITAN project, designed not as a truth arbiter, but as a ‘media thinking coach’ for users. The coach trains users to recognize the mechanics of manipulation tactics—such as conspiracism, polarization, and discrediting—rather than focusing solely on verifying facts. It utilizes Retrieval-Augmented Generation (RAG) architecture and a ‘Tactic Profiling’ methodology to implement dialogical inoculation. Through structured, Socratic questioning, the system guides users to evaluate media sources (“WHO”) and recognize specific manipulative tactics (“WHAT”). A formative mixed-method pilot study (N = 12) evaluated usability, engagement, and educational impact through focus groups, questionnaires, and reflective discussion. Findings provided initial empirical validation of the Socratic approach, confirming its high relevance for stimulating critical analysis while revealing a key design challenge: balancing user expectations for definitive answers with the tool’s role as a facilitator of reflection. Overall, the study outlines a scalable model for media education that cultivates durable meta-literacy—the reflective habits essential for autonomous reasoning in complex information environments.

**Keywords:** Critical Thinking, Media Literacy, Disinformation, AI Chatbot, Socratic Method, Psychological Inoculation .

## INTRODUCTION

In today’s algorithmically mediated post-truth era, disinformation extends beyond simple factual inaccuracy [32][47]. Unlike unintentional misinformation, disinformation is strategically engineered to manipulate cognition, emotion, and social perception, polarizing audiences and undermining trust in democratic and scientific institutions [14][21]. Digital platforms amplify this threat through algorithmic curation, repetition, and clickbait that exploit cognitive shortcuts, bypass reflective reasoning, and accelerate virality [8].

System-level interventions such as legislation, fact-checking, and algorithmic moderation remain fragmented and insufficient [42]. They face challenges of scale, speed, and declining public trust, often provoking skepticism or reactance [17][45]. Alone, these strategies cannot counter the ‘illusory truth’ effect, the ‘continued influence’ effect, habitual superficial scrolling, or affective manipulation [8][22].

Individual-level interventions, including critical-thinking training, gamified inoculation, verification heuristics, and media literacy curricula, offer some benefits in developing reasoning skills, awareness of misinformation, and cognitive resistance but they face important limitations [26][42]. Media literacy interventions are criticized for heavy geographic and demographic bias, their impact is inconsistent, long-term evaluations are scarce, and they may foster a false sense of confidence [42]. Critical thinking (CT) interventions rely on weak or small-scale evidence, with inconclusive long-term benefits [42]. Also, generic CT is limited, as its transferability is debated and it often requires domain-specific knowledge [1][16]. The effectiveness of critical thinking is limited by identity-protective cognition and motivated reasoning, particularly when essential dispositions such as openmindedness are lacking [27][27][48]. Moreover, CT's rhetoric (e.g., “do your own research”) is often co-opted by conspiracy theorists [42]. Inoculation (prebunking) depends on voluntary uptake, decays over time, can induce dysfunctional cynicism that reduces trust even in reliable news, and may fail if the source is seen as untrustworthy or politicized [26][42].

These challenges highlight the need for a hybrid, AI-supported pedagogy that integrates the contextual heuristics of Media and Information Literacy (MIL) with CT's generic skills and dispositions [18][29][31]. To be effective, critical thinking requires the domain-specific knowledge that comes from MIL. Media literacy provides knowledge to evaluate sources, detect bias, and verify claims, but its effectiveness depends on generic critical thinking to analyze, interpret, and reason rigorously, and on CT dispositions—such as open-mindedness and intellectual humility—to sustain reflective reasoning under cognitive load [4][37]. Without this integration, even highly analytical individuals may apply reasoning selectively to protect identity rather than pursue accuracy [27].

This synthesis defines Media Critical Thinking (MCT): the reflective, situated capacity to discern manipulative from neutral media content through analytical skill, contextual awareness, and epistemic virtue [39][37][43]. MCT reframes reasoning as a media-situated practice, interrogating how rhetoric, emotion, and technology shape interpretation, while cultivating affective regulation, intellectual humility, open-mindedness, and reflective selfmonitoring [11][13].

We operationalize MCT through the Socratic Coach, an AI-supported ‘boosting’ intervention developed by NCSR Demokritos within the EU Horizon Europe project TITAN. Unlike automated fact-checkers or belieftargeted bots [7][33], the Coach fosters guided, reflective dialogue, strengthening reasoning and the capacity to detect manipulateness rather than debating content. By targeting media manipulation rather than ideology, it reduces psychological reactance [35]. Iterative Socratic questioning cultivates analytical reasoning and dispositional resilience—including self-awareness and open-mindedness [5][9][36]—embodying MCT by using AI not to correct beliefs, but to reinforce reflective cognition in an attention-driven media ecosystem [28].

## METHODOLOGY

### Dialogical Design and Pedagogical Framework

The Socratic AI Coach was developed as part of the EU Horizon Europe TITAN project to operationalize the Media Critical Thinking (MCT) framework in an interactive, scalable form. The methodology was designed to cultivate manipulation discernment and reflective reasoning through guided dialogue. It integrates principles from cognitive psychology, media literacy, and constructivist learning to enable users to recognize manipulative tactics and strengthen their epistemic resilience [9][11] [23][48].

The design process followed three core pedagogical pillars:

1. **Nudging Analytical Thinking** – slowing down intuitive, “System 1” responses to promote deliberate, reflective reasoning [20].
2. **Dialogical Inoculation** – combining prebunking (exposure to weakened forms of misinformation) with active Socratic questioning to teach recognition of manipulation tactics [6][40].
3. **Fostering Self-Monitoring** – encouraging metacognitive reflection on personal biases, reasoning, and emotional responses to media messages [6].

Together, these pillars form a cognitive architecture for manipulation discernment, supporting both skill acquisition and the dispositions required for critical engagement with digital content [30][37].

## Theoretical Foundations

The Socratic method provides the conceptual foundation for the Coach. In educational psychology and coaching, Socratic dialogue promotes critical thinking by prompting learners to examine their assumptions, justify conclusions through guided questioning and consider alternatives [34][36]. This approach shifts users from passive reception to active co-construction of knowledge, aligning with constructivist and experiential learning theories [3][23]. To complement this dialogical foundation, Media and Information Literacy (MIL) frameworks—such as SIFT and CRAAP—were embedded to guide users in evaluating sources, authorship, and credibility [4]. The Coach thus merges analytical questioning with practical heuristics, reinforcing both conceptual understanding and procedural knowledge.

## Knowledge Base and Tactic Profiling

A structured Knowledge Base supports the chatbot's reasoning process through a *Tactic Profiling* methodology. This framework categorizes manipulation tactics based on psychological, rhetorical, and emotional markers, drawing from recent misinformation taxonomies [15][21][41]. Each profile defines the tactic's narrative structure, logical fallacies, rhetorical ploys, cognitive bias and emotional triggers, and counterarguments.

The current system emphasizes six recurring tactics: Discrediting, Polarization, Conspiracism, Trolling, Pseudoscience, and Science Denialism [41]. Associated cues—such as ad hominem attacks, sensationalist framing, and cherry-picked evidence—help the chatbot support user learning [25]. This knowledge base is further enriched with MIL heuristics for source verification and contextual evaluation [4][31].

## Socratic Inoculation Engine

At the technical and pedagogical core lies the 'Socratic Inoculation Engine'—a dialogue engine that transforms tactic profiles into dynamic learning interactions. The system engages users in questioning sequences that prompt analysis of 'WHO' (source credibility) and 'WHAT' (content manipulation) dimensions of a media item. Through iterative questioning, the Coach encourages users to interpret messages and identify core claims, detect reasoning fallacies and emotional triggers, generate independent counterarguments, and reflect on their own interpretive biases [9]. In contrast to belief-based correction systems, this design minimizes psychological reactance by focusing on *technique recognition* rather than ideological persuasion [17][24][41].

## 2.4 Dialogue Structure and Adaptive Support

The chatbot's dialogue follows a **two-block structure**:

1. **Block 1: Passive Prebunking** – Users are shown short infographics illustrating manipulation tactics with real-world examples, fostering awareness before active engagement.
2. **Block 2: Active Coaching** – Guided Socratic questions lead users through source ('WHO') and content ('WHAT') evaluation.

Example 'WHO' prompts include questions about Author Identity, Author Expertise & Authenticity, Source Presentation, etc., such as "Can you tell me if the author uses his/her real name or if he/she uses a pseudonym? Are his/her credentials verifiable?" Example 'WHAT' prompts include questions about signs of manipulation tactics, such as questions about conspiracism "*Does the evidence come from 'silenced' sources or 'insiders'?*" To accommodate users of varying literacy levels, the dialogue employs adaptive, support mechanisms. Novice users receive explanatory hints, definitions, and step-by-step guidance; experienced users encounter more openended, reflective questions [13][18].

**Table 1.** Example WHO Question Analysis: “Do you think that an expert’s opinion is needed to address this particular topic? If so, can you tell me if the author is himself/herself an expert on the subject or if he/she relies on other experts’ opinions?”

Mechanism Name	Example / Content
<b>Description</b>	This question checks whether the author has the necessary qualifications (education, experience, credentials) to speak with authority on the topic, or if they cite experts who do.
<b>Rationale</b>	For complex topics (like health, science, or economics), a simple opinion is not enough; specialized knowledge is required. This helps you distinguish a random person's opinion from an evidence-based conclusion.
<b>Example</b>	An article about a new medical treatment should be written by a doctor (MD) or a medical journalist referencing peer-reviewed studies, and not by a wellness blogger citing “a feeling.”
<b>Tip</b>	<i>Lateral Reading:</i> Don’t just trust the “About Us” page. Google the author’s name along with their credentials (e.g., “Dr. John Doe cardiology”) to see what other verifiable, independent sources say about them.
<b>Micro-lesson</b>	“What is ‘Expertise’? (And how to spot a ‘Fake Expert’)”
<b>Warning</b>	Be careful: Sometimes the rhetoric of expertise (e.g., “do your own research,” “I am a critical thinker”) is used to imply authority, even without actual credentials.
<b>Pre-emptive Refutation</b>	<b>Claim:</b> “This ‘brave doctor’ is being silenced by the establishment.” <b>Refutation:</b> Extraordinary claims require extraordinary evidence. A “lone wolf” expert claiming to overturn decades of science is less credible than the consensus of thousands of experts in that field. Check if this “silenced” expert is recognized by their peers.

**Table 2.** Example WHAT Question Analysis: “Does the evidence come from ‘silenced’ sources or ‘insiders’?”

Mechanism Name	Example / Content
<b>Description</b>	The theory relies on a “whistleblower,” a “persecuted doctor,” or a “brave researcher.”
<b>Rationale</b>	This tactic tries to make a questionable source seem credible by portraying them as a “victim” telling the “truth they don’t want you to hear.”
<b>Definition</b>	<i>Appeal to Questionable/Maverick Authority</i> — using an authority not accepted by the legitimate field.
<b>Tip</b>	<i>Lateral Reading:</i> Search the “expert’s” name on other, reliable websites. Are they actually an expert in this specific field?
<b>Pre-emptive Refutation</b>	Extraordinary claims require extraordinary evidence. A theory that overturns decades of science needs more than one “lone wolf” expert; it needs solid, verifiable evidence that other experts in the field can review and confirm.
<b>Mechanism Name</b>	<b>Example / Content</b>
<b>Micro-lesson</b>	“How to Use ‘Lateral Reading’.”

## Technical Implementation

To ensure reliability and pedagogical consistency, the Coach operates on a Retrieval-Augmented Generation (RAG) architecture [12]. Verified content and tactic definitions are stored as vector embeddings, which the large language model retrieves to construct contextually relevant, evidence-based dialogue turns. The initial prototype was implemented using Meta's LLaMA 3.1 8B Instruct, later upgraded to Google's Gemma 3 12B Instruct for improved multilingual support and context length. The hybrid RAG design ensures both semantic relevance and transparency—reducing hallucination risk while allowing pedagogical control over dialogue flow. This semiscripted model aligns with ethical and reproducibility principles for educational AI: it maintains instructor-level control while adapting to each learner's input [10].

## User Engagement and Conversational Design

The dialogue is designed to be dynamic, conversational, and non-judgmental, facilitating a supportive environment where users feel comfortable exploring ideas. Concise, focused interactions prevent cognitive overload, while interactive elements such as definitions, examples, and micro-lessons maintain engagement [44]. This careful design ensures users can actively participate in the learning process and apply insights immediately.

## Research Questions

The methodology includes ongoing assessment of the chatbot's effectiveness via user feedback, interaction analytics, and pilot testing to inform iterative refinement of question phrasing, scaffolding, and dialogue structure. Metrics focus on improving critical thinking, media literacy, and counterargument generation. The evaluation of the Socratic Coach was guided by four research questions:

1. **Expectations:** What are users' expectations when using an interactive AI-driven chatbot for evaluating news content?
2. **Relevance:** How relevant do users find a Socratic chatbot's questions in helping them detect manipulation and critically analyze media items?
3. **Experience and Usability:** How do users evaluate the experience, usability, and value of interacting with an AI-driven Socratic chatbot?
4. **Pedagogical Alignment:** Does the chatbot's dialogue design align with users' needs for refuting manipulative arguments and enhancing their critical thinking skills?

## Evaluation Method

### Design

The study employed a convergent mixed-methods design integrating quantitative and qualitative data. Quantitative components included baseline test scores and Likert-scale questionnaire responses. Qualitative components comprised focus-group transcripts, observational notes, and chatbot–user dialogue logs. This approach allowed triangulation across (a) behavioral evidence of interaction, (b) self-reported perceptions, and (c) objective baseline ability. The design emphasized formative validation rather than hypothesis testing.

### Participants

The empirical study was conducted as a pilot workshop at the National Centre for Scientific Research 'Demokritos' (NCSR-D) in Athens, Greece, on 15 October 2024. Twelve adults (8 female, 4 male) participated voluntarily. All were mid-career professionals and researchers from NCSR-D departments, representing an analytically skilled but usability-sensitive group. Participant demographics were as follows:

**Table 3.** Participant Demographics (N=12)

Characteristic	Category	Number of Participants	Percentage of Total
<b>Gender</b>	Female	8	66.7%
	Male / Other	4	33.3%
<b>Age Group</b>	41–50 years	6	50%
	Other Age Groups	6	50%
<b>Education Level</b>	Graduate or Professional Degree	9	75%
	Undergraduate Degree or Lower	3	25%

Inclusion criteria were (a) familiarity with digital media use and (b) ability to engage in reflective discussion in English. No participants reported previous experience with AI-based educational chatbots. Sampling followed a convenience design, appropriate for early-stage, formative evaluation. Although non-representative, this sample was considered optimal for identifying conceptual and functional issues prior to large-scale field deployment.

## Materials and Instruments

### a. TITAN Socratic Chatbot Prototypes

The TITAN Socratic Coach was implemented using a Retrieval-Augmented Generation (RAG) model designed to operationalize the Media Critical Thinking (MCT) framework through guided Socratic dialogue. The system retrieves verified tactic profiles and pedagogical resources to generate contextually relevant, evidence-based questions and support prompts. Each exchange encourages users to analyze source credibility (“WHO”) and manipulative tactics (“WHAT”). The RAG design ensures transparency, semantic accuracy, and a consistent educational flow, serving as a formative prototype for reflective media critical thinking training.

### b. Manipulative Online Content Recognition Inventory (MOCRI)

The MOCRI, a 24-item instrument measuring manipulateness recognition, was administered as a pre-test and served as a baseline diagnostic of participants’ Media Critical Thinking [30]. Its focus on *discernment* rather than factual recall aligns with epistemic vigilance [37][38]. Descriptive statistics from this pre-test (see Findings 4.6) were used to establish this baseline and contextualize subsequent self-reported perceptions.

### c. Post-Interaction Questionnaire

A 28-item Likert-type instrument was developed to assess usability, engagement, attitudes, and perceived learning. In particular, items Q1 through Q23 collectively addressed the experiential and usability domains of the Socratic Coach, covering user perceptions of tasks, interaction quality, language clarity, and overall satisfaction. Items Q24 through Q28 specifically targeted critical-thinking development, awareness of manipulation, and metacognitive reflection. Two additional open-ended items (Q29–Q30) invited participants to elaborate on perceived learning, engagement, and dialogue quality. Questionnaire items were adapted from established conversational-agent evaluation scales and pilot-tested for clarity.

### d. Qualitative Materials

Focus-group discussions and transcribed dialogue logs were used for thematic analysis [2]. Field notes from facilitators documented behavioral reactions and usability incidents.

## Procedure

The study was conducted in two 2-hour sessions, each facilitated by three researchers. The procedure comprised four sequential stages:

1. **Pre-test phase.** Participants completed a demographic and media-use questionnaire, followed by the MOCRI test to establish baseline manipulation-recognition ability.
2. **Brainstorming and expectation mapping.** A guided focus-group discussion explored participants' experiences with online misinformation, trust criteria, and expectations of an AI-based fact-checking or coaching tool.
3. **Hands-on interaction.** Each participant interacted individually with the Socratic Coach prototype (RAG) using authentic media articles. Dialogues followed a Socratic structure encouraging the identification and reflection on disinformation tactics.
4. **Post-test and reflection.** After the interaction, participants completed the usability and learning questionnaire, and the open-ended reflection items. Facilitators debriefed participants on the system's aims and collected verbal feedback on interaction flow and educational relevance.

All sessions were recorded, anonymized, and transcribed. Total participation time per person was approximately two hours.

## Data Analysis

Quantitative data were analyzed descriptively (means, frequencies) to identify usability patterns. Qualitative data from focus groups and open-ended responses were coded inductively [2], with emergent categories grouped under usability, pedagogical engagement, and reflective reasoning. Integration followed a triangulation logic, comparing convergence among behavioral indicators, self-reports, and MOCRI outcomes—consistent with formative AI-education research [7][33][42]. The final report focuses on key items selected for their relevance to the research questions (RQs) and the primary design constraints (e.g. Q1, Q13, Q24, Q25, Q27), which serve as the primary indicators of system performance.

## FINDINGS

The findings provide empirical validation of the TITAN Socratic Coach as an educational innovation that cultivates critical awareness and reflective reasoning in digital media contexts. Across all data sources, participants confirmed the system's relevance to media-literacy development, while also identifying opportunities for refinement that can strengthen engagement and adaptability.

### Research Question 1: Expectations

#### User Perceptions and Expectations (Pre-Interaction)

Before the interaction, participants described digital-first news habits (e.g., social media, Google feeds, online newspapers) and identified reliability cues such as author credibility, verified sources, and linguistic accuracy. Disinformation was typically defined as intentional manipulation aimed at shaping opinion or behavior through multimedia content.

Expectations for TITAN involved transparency, credibility, and educational purpose. Users valued clear explanations, accessible yet formal language, and evidence-based reasoning. Many envisioned the chatbot as a *learning companion* offering structured, time-efficient sessions (~15 minutes) with optional micro-lessons for deeper exploration.

**Table 4.** User Expectations for the TITAN Socratic Coach (Pre-Interaction)

Key Expectation	Typical User Expression
Transparency of source and data	“I want to know who built it and where the data comes from.”
Reliability and clarity	“It should explain clearly and cite evidence.”
Educational support	“I want to feel I’m being trained, not just corrected.”
Efficient sessions	Preference for concise, guided dialogues with optional depth

### Research Question 3: Experience and Usability

#### Usability and Overall Experience (Post-Interaction)

Questionnaire results reflected generally positive usability perceptions, complemented by constructive feedback on pacing and flexibility. Two-thirds of participants found tasks comprehensible, and half agreed that the chatbot was easy to use. Users praised linguistic clarity and logical task flow, noting that smoother response timing and navigational options would further enhance experience.

**Table 5.** Usability Evaluation (Post-Interaction, N = 12)

Aspect	Mean (M/5)	Summary Insight
Tasks were comprehensible	3.83	Clear task framing and coherent flow
Assistance was adequate	2.91	Highlights need for more dynamic support
Easy to use overall	3.50	Moderate usability; improvements in pacing would help

#### Interaction Quality and Engagement (Post-Interaction)

Engagement data indicated strong educational relevance, paired with valuable insights for enhancing conversational variety. Participants appreciated the content’s reflective focus but suggested more adaptive phrasing to sustain interest.

**Table 6.** Interaction Quality and Engagement Metrics (N = 12)

Item	Mean (M/5)	Interpretation
Responses non-repetitive	—	Identified as key area for improvement
Responses relevant to input	2.75	Content informative but could better adapt to user input
Discussion engaging	3.25	Moderate engagement; cognitive rather than entertainment-oriented
Chatbot kept user engaged	2.50	Indicates need for more variety and navigational freedom
Qualitative insights	—	Users valued reflection but wanted clearer Socratic structure

Such feedback is typical for early-stage Socratic prototypes, where depth of reasoning is prioritized over



conversational fluency—offering a clear path for iterative enhancement.

### Dialogue-Level Analysis (RAG Chatbot)

Analysis of fourteen logged RAG dialogues reinforced the need for flexibility and personalization. Sessions averaged 12 exchanges (maximum 26), confirming user preference for concise 15-minute interactions. About half of the responses were five words or fewer (e.g., “Yes,” “I don’t know”), showing the importance of more open-ended prompts and guided navigation.

**Table 7.** Dialogue Analysis of RAG Prototype (N = 14)

Observed Aspect	Empirical Pattern	Interpretation
Dialogue length	12 exchanges (max 26)	Aligns with preferred session duration (~15 min)
Short replies ( $\leq 5$ words)	50% of responses	Indicates need for elaboration prompts
Requests for guidance/help	64% of users	Shows active engagement and need for structured support

### Language and Accessibility

Participants evaluated TITAN’s language as clear, accurate, and approachable. Ratings for tone and phrasing were high ( $M = 4.16/5$ ), with only isolated mentions of abrupt expressions.

**Table 8.** Language and Accessibility Ratings

Aspect	Mean (M/5)	Interpretation
Language easy to understand	4.16	Strong linguistic accessibility
Questions clearly phrased	3.91	High clarity and logical sequencing
Information aided responses	3.41	Slightly lower score indicates opportunity for more guidance cues

These consistently high clarity scores confirm that linguistic accessibility forms a strong foundation for future improvements in interactivity.

### Research Questions 2 & 4: Educational Impact and Validation

#### Baseline Manipulation Recognition (MOCRI Pre-Test)

Baseline MOCRI results showed high but non-uniform manipulation-recognition ability, even among a highly educated sample. This variability validates the need for reflective, skill-based interventions that strengthen epistemic vigilance—confirming TITAN’s pedagogical relevance.

#### Educational Value and Critical Thinking Outcomes (Post-Interaction)

Post-interaction findings confirmed that TITAN enhanced reflective awareness and media-critical reasoning. Two-thirds of users reported learning new insights, and three-quarters recognized its potential for long-term application.

**Table 9.** Educational and Cognitive Impact (N = 12)

Evaluation Dimension	Mean (M/5)	Summary Interpretation
Learned new things	3.67 (67% agree/strongly agree — 8/12)	TITAN promoted new insights and reflective learning
Helps resist manipulation	3.25 (58% agree — 7/12)	Moderate perceived efficacy; potential to grow with extended use
Questions relevant for detecting and analyzing manipulation	—	75% rated “Quite” or “Very relevant”
Evaluation Dimension	Mean (M/5)	Summary Interpretation
Clear guidance and explanations for refuting manipulative arguments	—	75% agreement confirms effective pedagogical structure
Interpretation	—	TITAN promoted metacognitive reflection and provided useful refutation guidance; mean scores indicate moderate immediate efficacy with room to strengthen perceived future resilience.

### Design Lessons and Implications

Synthesizing evidence from all datasets, several empirically grounded design lessons emerged, confirming TITAN’s pedagogical robustness while guiding its next iteration toward adaptive, user-centered learning. **Table 10.** Empirically Derived Design Lessons

Observed Aspect	Evidence	Recommended Enhancement
Conversational variety	Users noted repetitive phrasing	Implement adaptive rephrasing logic
Fatigue after ~15 min	Consistent across logs	Shorter sessions + summary checkpoints
User agency	64% asked for guidance	Add “guided” and “open” session modes
Educational demand	Requests for summaries and micro-lessons	Include “Tell me more” features
Conceptual clarity	Occasional confusion (e.g., pseudoscience vs. science denial)	Refine definitions and examples
Accessibility	Requests for native languages	Develop multilingual and tone-adaptive profiles

### Summary of Empirical Impact

Overall, findings from the pilot study indicate that the Socratic approach is pedagogically valuable and highly relevant for enhancing critical analysis. Concurrently, user feedback provided a clear and practical roadmap for

future improvements, particularly concerning dialogue flow and interactivity. These results validate the tool's conceptual soundness and set a strong empirical foundation for its further evolution.

## DISCUSSION

### Overview and Interpretation of Findings

This exploratory pilot provides the first empirical validation of the Media Critical Thinking (MCT) framework operationalized through the TITAN Socratic AI Coach, illustrating how AI-mediated dialogue can serve as a vehicle for cultivating reflective reasoning in media contexts. These results build upon the qualitative findings of the initial case study, confirming earlier observations on user expectations, trust, and the pedagogical potential of Socratic dialogue while providing new quantitative evidence of educational impact. Qualitative and observational evidence suggest that the chatbot effectively fosters reflective engagement and critical awareness—core components of epistemic resilience. The results highlight a constructive alignment between pedagogical design and technical implementation: the Coach's cognitive framework successfully supports users' reflective reasoning, while its current dialogue flow and pacing indicate areas for future enhancement. Baseline findings from MOCRI test further demonstrated that even a highly educated sample remained vulnerable to manipulation, validating the project's rationale that manipulation literacy cannot be assumed, even among experts. This finding supports prior research showing that analytic ability alone does not ensure epistemic vigilance [38] and underscores the need for interventions targeting cognitive style, emotion regulation, and reflective reasoning rather than factual recall alone [16][48].

### User Experience and Interactional Dynamics

Users described the TITAN Coach as clear, structured, and intellectually stimulating. The system's methodical questioning style sometimes challenged users' expectations of fast, automated responses. Yet, this very structure created the kind of cognitive friction that supports reflection and deeper reasoning. Rather than being experienced as a limitation, this deliberate pacing encouraged users to slow down, reconsider assumptions, and recognize manipulative intent within messages. While some participants initially anticipated more direct feedback, many later acknowledged that guided inquiry helped them articulate their reasoning and engage more critically with content. Future iterations could enhance engagement by offering adaptive pacing, optional summaries, and interactive control, while preserving the Socratic rigor that defines Media Critical Thinking (MCT) [36].

### Educational Value and Cognitive Outcomes

Within the current design parameters, the TITAN Coach demonstrated meaningful educational value by transforming abstract literacy concepts into active reasoning practice. Through guided questioning, users not only identified manipulative tactics but also reflected on their own interpretive processes, an early sign of metacognitive development. This shift from fact verification to reasoning about influence exemplifies the pedagogical goal of Media Critical Thinking (MCT): fostering awareness of how emotion, framing, and intent shape perception. Even brief interactions encouraged users to articulate judgments, internalize reflective habits, and approach media with greater epistemic vigilance. These learning outcomes are consistent with prior literature on metacognitive transfer suggesting that Socratic dialogue can strengthen cognitive resilience when supported by adaptive design and iterative engagement.

### Design Implications and Future Refinements

Synthesizing the evidence identifies clear priorities for development:

1. **Adaptive Interaction Flow:** add dynamic rephrasing, skip-question logic, and user-initiated navigation to enhance conversational diversity.
2. **Hybrid Guidance Model:** offer dual modes (“guided” vs. “open”) to vary support.

3. **Micro-Learning Integration:** embed short tutorials via “Tell me more” prompts to balance brevity with depth. • **Feedback and Summarization:** provide end-session summaries outlining reasoning steps.
4. • **Conceptual Clarity:** refine tactic definitions (e.g., pseudoscience vs. science denial).
5. **Multilingual & Personalized Adaptation:** adjust tone, pacing, and language complexity to user profiles.

These refinements will transform the prototype from static dialogue tool to adaptive learning environment capable of sustaining engagement while preserving Socratic rigor.

### Theoretical and Practical Contributions

The findings contribute to three domains. First, they empirically support Socratic AI coaching as digital metacognition training—linking analytical reasoning with emotional regulation in misinformation contexts. Second, they extend MIL pedagogy by showing that question-driven interfaces can elicit reflection without triggering psychological reactance. Third, they provide methodological insight for AI-education evaluation, emphasizing combined behavioral logs, self-reports, and standardized measures (e.g. MOCRI) to assess both subjective and objective outcomes. Overall, the study highlights transparency-by-design as both ethical and pedagogical—modeling epistemic accountability through explainable system behavior.

### Limitations and Future Research

As an exploratory pilot, the study is constrained by its small, homogeneous sample. High baseline literacy may have limited observable gains, while short exposure precluded long-term assessment. The RAG model’s semiscripted design also ensured pedagogical consistency but limited spontaneity. Future research should employ a pre/post-test assessment of manipulateness discernment with larger and diverse populations, longitudinal follow-ups, and multimodal analytics (e.g., latency, sentiment trajectories) to capture engagement in real time. Integrating affective-computing components could further enable empathetic and context-aware feedback.

## CONCLUDING REMARKS

Overall, this pilot confirms that Socratic AI can meaningfully engage users in reflective reasoning about media manipulation. By transforming misinformation from passive exposure to active analysis, the TITAN Coach demonstrates how AI can serve as an educational partner cultivating manipulateness awareness, openmindedness, cognitive reflection and epistemic control. Continued refinement toward adaptive, multimodal, and multilingual interaction will allow the system to progress from conceptual prototype to scalable public intervention—empowering individuals to cultivate critical reflection *with* AI rather than *against* it.

## REFERENCES

1. Bailin, S., Case, R., Coombs, J.R., & Daniels, L.B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies*, 31(3), 285–302. <https://doi.org/10.1080/002202799183133>
2. Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA Handbook of Research Methods in Psychology: Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57-71). American Psychological Association.
3. Bruner, J. S. (1966). *Toward a theory of instruction*. Belknap Press of Harvard University Press.
4. Caulfield, M., & Wineburg, S. (2023). *Verified: How to think straight, get duped less, and make better decisions about what to believe online*. University of Chicago Press.
5. Cohen, J. S., Edmunds, J. M., Brodman, D. M., Benjamin, C. L., & Kendall, P. C. (2013). Using selfmonitoring: Implementation of collaborative empiricism in cognitive-behavioral therapy. *Cognitive and Behavioral Practice*, 20(4), 419–428. <https://doi.org/10.1016/j.cbpra.2012.07.001>
6. Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE*, 12(5), e0175799. <https://doi.org/10.1371/journal.pone.0175799>

7. Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714). <https://doi.org/10.1126/science.adq1814>
8. Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., & Fazio, L. K. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1, 13–29. <https://doi.org/10.1038/s44159-021-00006-y>
9. Facione, P.A. (2015). *Critical Thinking: What It Is and Why It Counts*. Millbrae, CA: Measured Reasons LLC.
10. Floridi, L. and Cowls, J. (2022). A Unified Framework of Five Principles for AI in Society . In *Machine Learning and the City*, S. Carta (Ed.). <https://doi.org/10.1002/9781119815075.ch45>
11. Gabaree, S. (2022). A metacognitive approach to reduce the spread of online misinformation. *The Journal of Media Literacy*. International Council for Media Literacy. <https://ic4ml.org/journal-article/ametacognitive-approach-to-reduce-the-spread-of-online-misinformation/>
12. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2312.10997>
13. Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America*, 117(27), 15536–15545. <https://doi.org/10.1073/pnas>.
14. Hameleers, M., & van der Meer, T. G. L. A. (2019). Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research*, 47(2), 227-250. <https://doi.org/10.1177/0093650218819671> (Original work published 2020)
17. Harjani, T., Roozenbeek, J., Biddlestone, M., van der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., & Graham, M. (2022). *A Practical Guide to Prebunking Misinformation*.
18. Hitchcock, D. (2024). Critical thinking. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024 Edition). Stanford University. <https://plato.stanford.edu/archives/sum2024/entries/critical-thinking/>
19. Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 8, 1545 - 1553.
20. Jones-Jang, S.M., Mortensen, T., & Liu, J. (2021). Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, 65(2), 371-388. <https://doi.org/10.1177/0002764219869406>
21. Kahan, D.M. (2017). Misinformation and Identity-Protective Cognition. *Political Communication eJournal*.
22. Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
23. Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5), 1301-1326. <https://doi.org/10.1177/1461444820959296>
24. Kaye, B. K., & Johnson, T. J. (2024). I can't stop myself! Doomscrolling, conspiracy theories, and trust in social media. *Atlantic Journal of Communication*, 32(3), 471–483. <https://doi.org/10.1080/15456870.2024.2316844>
25. Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.
26. Koppel, L., Robertson, C. E., Doell, K. C., Javeed, A. M., Rasmussen, J., Rathje, S., Vlasceanu, M., & Van Bavel, J. J. (2023). Individual-level solutions may support system-level change—if they are internalized as part of one's social identity. *Behavioral and Brain Sciences*, 46, e165. <https://doi.org/10.1017/S0140525X2300105X>
27. Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21(3), 103– 156. <https://doi.org/10.1177/1529100620946707>
28. Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., et al. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 8, 1044–1052. <https://doi.org/10.1038/s41562024-01881-0>

29. Kwek, A., Peh, L., Tan, J., & Lee, J. X. (2023). Distractions, analytical thinking and falling for fake news: A survey of psychological factors. *Humanities & social sciences communications*, 10(1), 319. <https://doi.org/10.1057/s41599-023-01813-9>
30. Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, 58, 101964. <https://doi.org/10.1016/j.gloenvcha.2019.101964>
31. Machete, P., & Turpin, M. (2020). The use of critical thinking to identify fake news: A systematic literature review. In M. Hatching, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible design, implementation and use of information and communication technology* (Vol. 12067, pp. 257–267). Springer. [https://doi.org/10.1007/978-3-030-45002-1\\_20](https://doi.org/10.1007/978-3-030-45002-1_20)
32. Maertens, R., Roozenbeek, J., & van der Linden, S. (2023). The Manipulative Online Content Recognition Inventory (MOCRI): A scale to measure the ability to spot manipulation techniques in online content. *PsyArXiv*. <https://doi.org/10.31234/osf.io/g68mc>
33. McGrew, S., & Breakstone, J. (2023). Civic online reasoning across the curriculum: Developing and testing the efficacy of digital literacy lessons. *AERA Open*, 9, 1–16. <https://doi.org/10.1177/23328584231176451>
34. McIntyre, L. (2018). *Post-truth*. MIT Press.
35. Meyer, M., Enders, A., Klofstad, C., Stoler, J., & Uscinski, J. (2024). Using an AI-powered “street epistemologist” chatbot and reflection tasks to diminish conspiracy theory beliefs. *Harvard Kennedy School (HKS) Misinformation Review*, 5(6). <https://doi.org/10.37016/mr-2024-070>
36. Neenan, M. (2008). Using Socratic questioning in coaching. *Journal of Rational-Emotive & Cognitive-*
37. *Behavior Therapy*, 27(4), 249–264. <https://doi.org/10.1007/s10942-007-0076-z>
38. Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* (32), 303–330. [10.1007/s11109-010-9112-2](https://doi.org/10.1007/s11109-010-9112-2).
39. Paul, R., & Elder, L. (2016). *The thinker’s guide to the art of Socratic questioning*. Foundation for Critical Thinking.
40. Paul, R., & Elder, L. (2020). *Fact over fake: A critical thinker’s guide to media bias and political propaganda*. Rowman & Littlefield.
41. Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
42. Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
43. Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(65). <https://doi.org/10.1057/s41599-019-0279-9>.
44. Roozenbeek, J., Traberg, C., & van der Linden, S. (2022). Correction: Technique-based inoculation against real-world misinformation (print 2023), by Roozenbeek et al. *Royal Society Open Science*, 10, Article 231235. <https://doi.org/10.1098/rsos.231235>
45. Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. *European Psychologist*, 28(3), 189–205. <https://doi.org/10.1027/1016-9040/a000492>
46. Sartori, R., Tommasi, F., Ceschi, A., Falser, M., Genero, S., & Belotto, S. (2022). Enhancing critical thinking skills and media literacy in initial vocational education and training via self-nudging: The contribution of NERD VET project. *Frontiers in Psychology*, 13, 935673. <https://doi.org/10.3389/fpsyg.2022.935673>
47. Singh, A., Guan, Z., & Rieh, S. Y. (2025). Enhancing critical thinking in generative AI search with metacognitive prompts (arXiv:2505.24014). *arXiv*. <https://doi.org/10.48550/arXiv.2505.24014>
48. Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, 359, 11461151. [10.1126/science.aap9559](https://doi.org/10.1126/science.aap9559).
49. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.



50. Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe Report. <https://rm.coe.int/information-disorder-towardan-interdisciplinary-framework-for-research/1680765e2b>
51. Ziemer, C.-T., & Rothmund, T. (2024). Psychological underpinnings of misinformation countermeasures: A systematic scoping review. *Journal of Media Psychology: Theories, Methods, and Applications*, 36(6), 397–409. <https://doi.org/10.1027/1864-1105/a000407>

## Declarations for Manuscript Submission Ethical Considerations

The study received ethical approval from the Institutional Review Board (IRB) at NCSR ‘Demokritos’. All participants were fully informed about the study's purpose, its voluntary nature, and confidentiality safeguards, and provided written informed consent prior to participation. Data anonymization was ensured through the use of self-generated four-digit personal codes to link the pre- and post-test questionnaires without collecting personally identifiable information.

## Conflict of Interest

The author(s) declare(s) no competing interests.

## Funding

This paper was funded by the TITAN and AI4TRUST projects. TITAN has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No.101070658, and by UK Research and innovation under the UK government's Horizon funding guarantee grant numbers 10040483 and 10055990. AI4TRUST project has received funding from the Horizon Europe (AI to fight disinformation) under grant agreement No.101070190.

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.”

## Data Availability Statement

The raw data (qualitative transcripts and quantitative scores) generated during the pilot study workshop are not publicly available due to privacy and confidentiality protocols agreed upon with the participants during the consent process. Data may be made available upon reasonable request to the corresponding author, subject to ethical approval and anonymity preservation.

## Copyright and Licensing

Upon acceptance, all articles are published under the Creative Commons Attribution License (CC BY 4.0), permitting unrestricted use, sharing, and reproduction of the work with proper attribution.

## Appendix 1: The Questioning Process Educational Example Dialogue

This dialogue process dynamically adapts to users’ media literacy levels, ensuring an engaging and educational experience.

### User Input:

Users upload or reference a media item (e.g., article or social media post) for evaluation. Users specify whether they want to analyze the **source** (‘WHO’), **content** (‘WHAT’), or both.

### Introduction:

The chatbot introduces itself as a coach in media critical thinking, explaining its key features. Users can choose to start with either the ‘WHO’ or the ‘WHAT’ section.

### Evaluating the ‘WHO’ (Medium/Author):

#### • “WHO” Questions (Tailored to Media Type):

For websites: e.g., *“Is the website a trustworthy, established source?”*, *“Is the author a real person with a real name?”*, *“Does the author have any actual expertise in this topic?”*

For social media: e.g., *“Does the account have a verification badge?”*

### Support Mechanisms (Activated Based on User Input):

**Rationales:** When users ask why a question is relevant, the chatbot provides context.

**Chatbot:** “Is the author a real person, or are they using a fake name?”

**User:** “Why does that matter?”

**Chatbot:** “Well, if the author is hiding behind a fake name or doesn’t have verifiable credentials, it could be a sign of something shady—like trolls, bots, or even organized groups trying to spread propaganda or misinformation. It’s a tactic often used in influence campaigns to mislead or manipulate people.”

**Media Literacy Tips:** If users find it difficult to spot signs of deception, the chatbot can offer step-by-step tips, clear definitions, or practical examples to help them navigate the information more effectively.

○ **User:** “I’m not sure how to figure out if the author is a real person.” ○ **Chatbot:** “No problem! Here are a few steps to help you verify:

1. Check their bio for detailed background information—real authors usually have clear credentials.
2. Look for a professional social media presence or contact details, like an email.
3. Confirm their institutional affiliation through official websites.
4. Search for their name in reputable publications or third-party mentions to see if they’re recognized elsewhere.”

**User:** “Hmm, I didn’t really check the website. What should I be looking for?”

**Chatbot:** “Great question! Trustworthy sites tend to have clear editorial standards and a history of reliable reporting. They often have well-established reputations and citations. If the site is obscure or seems designed to push a specific agenda, it’s worth looking into more carefully. Always check if they’ve been recognized by other reputable sources.

**Warnings & Preemptive Refutations:** If users tend to distrust mainstream verification methods and dismiss traditional markers of credibility, the chatbot provides nuanced insights to address their concerns.

○ **User:** “I trust alternative or independent voices and media.”

○ **Chatbot:** “Independent perspectives can be valuable, but expertise is key. Watch out for vague titles like ‘independent researcher,’ ‘health guru,’ or ‘truth warrior’—these often lack credible evidence or professional backing.”

○

### Trust Reflection:

At the end of the ‘WHO’ section, the chatbot summarizes findings and asks, “Based on this evaluation, do you find the source trustworthy?” ○ Depending on user feedback, they can proceed to the ‘WHAT’ section or conclude the evaluation.



## Evaluating the WHAT (Content)

### ‘WHAT’ Questions (Based on chatbot’s detected signs):

- Prompts user to identify the central claim or message of the media item and its purpose.
  1. **Chatbot:** “Let’s begin by figuring out the main point this article is trying to make. What’s the core claim here?”
  2. **Chatbot:** “Does this article seem to be pushing a particular agenda, narrative or ideology? Is there any indication that the author might have a bias, or are they simply presenting the facts? Who might benefit from this message being spread?”

Uses follow-up questions to help user spot the detected signs of disinformation in the headlines and content of the article/post.

3. **Chatbot:** “Now, let’s take a closer look at the headline. Does it seem overly dramatic or like clickbait?”
4. **Chatbot:** “One common sign of disinformation is the mention of secret plots or hidden agendas by powerful people/groups.”

### Support Mechanisms:

**Rationales:** These clarify why specific tactics are problematic.

- **Chatbot:** “Overly Dramatic or Clickbait Headlines play on emotions like fear or outrage, which can cloud judgment and make readers less likely to critically analyze the content.”
- **Chatbot:** “Claims of secret plans or agendas often lack concrete evidence and exploit the human tendency to find patterns or conspiracies where none exist.”

**Tips, Examples, Definitions:** These assist users when they are unsure about certain signs.

- Chatbot: “This article claims that vaccines are part of a secret population control plan.”
- “Sensationalist language can involve exaggerated claims, such as ‘shocking’, or ‘exposes hidden truths’, ‘The Secret Population Control Agenda No One Wants You to See!’, ‘What They Don’t Want You to Know: The Dark Side of Global Health Policies!’, Here’s the Hidden Truth Big Pharma Doesn’t Want You to Know!’”

**Preemptive Refutations:** The chatbot provides pre-emptive refutations or counterarguments. For example, if the article claims that powerful groups suppress the truth (despite the lack of evidence):

- User: “They’ve been hiding the truth for years. With the right resources, they can keep this under wraps forever.”
- Chatbot: “Big conspiracies are unlikely to stay secret; eventually they leak; the more people involved, the higher the chance of leaks.”

### Trust Reflection:

- At the end of the ‘WHAT’ section, the chatbot summarizes the findings and asks the user, “So, based on everything we’ve reviewed—suspicious headlines, emotional language, lack of credible sources, and unverified claims—do you think the content is trustworthy?”
- This summary prompts users to reflect on the information they’ve processed and decide whether the content remains credible. Depending on their feedback, the chatbot then encourages re-thinking about the context behind the content and provides a holistic view of the media item, moving beyond just factual accuracy and examining the broader intention or biases.

1. **Chatbot:** “Now that we’ve looked at the content and medium, let’s think about the bigger picture. Who might benefit from this message being spread?”
2. **Chatbot:** “We’ve talked about the content’s credibility, but let’s re-think about its purpose. Does the article seem to align with a particular group’s interests or agenda?”

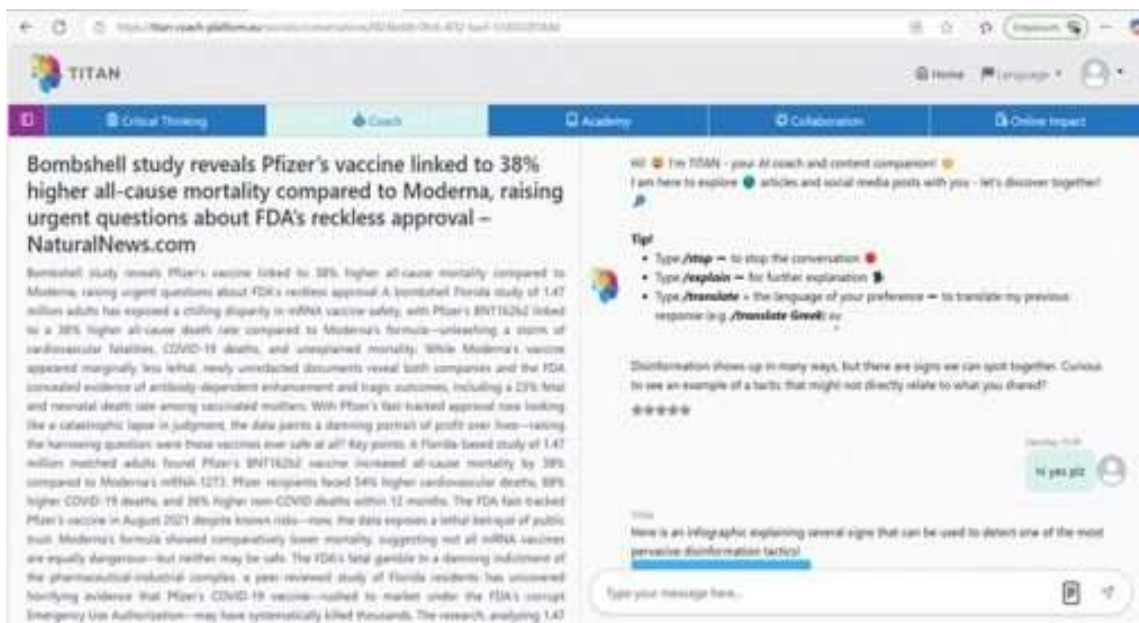
## APPENDIX 2: DIALOGUE SCREENSHOTS

This appendix presents screenshots from an interactive dialogue facilitated by the Socratic chatbot. The images demonstrate the Socratic questioning approach used to evaluate media content and enhance critical thinking. Each screenshot corresponds to a specific stage of the questioning process, as outlined in Appendix 2.

In this example, the user submitted an article for evaluation, titled *"Bombshell study reveals Pfizer's vaccine linked to 38% higher all-cause mortality compared to Moderna, raising urgent questions about FDA's reckless approval"* (<https://www.naturalnews.com/2025-05-01-pfizers-vaccine-linked-to-38higher-all-cause-mortality.html>)

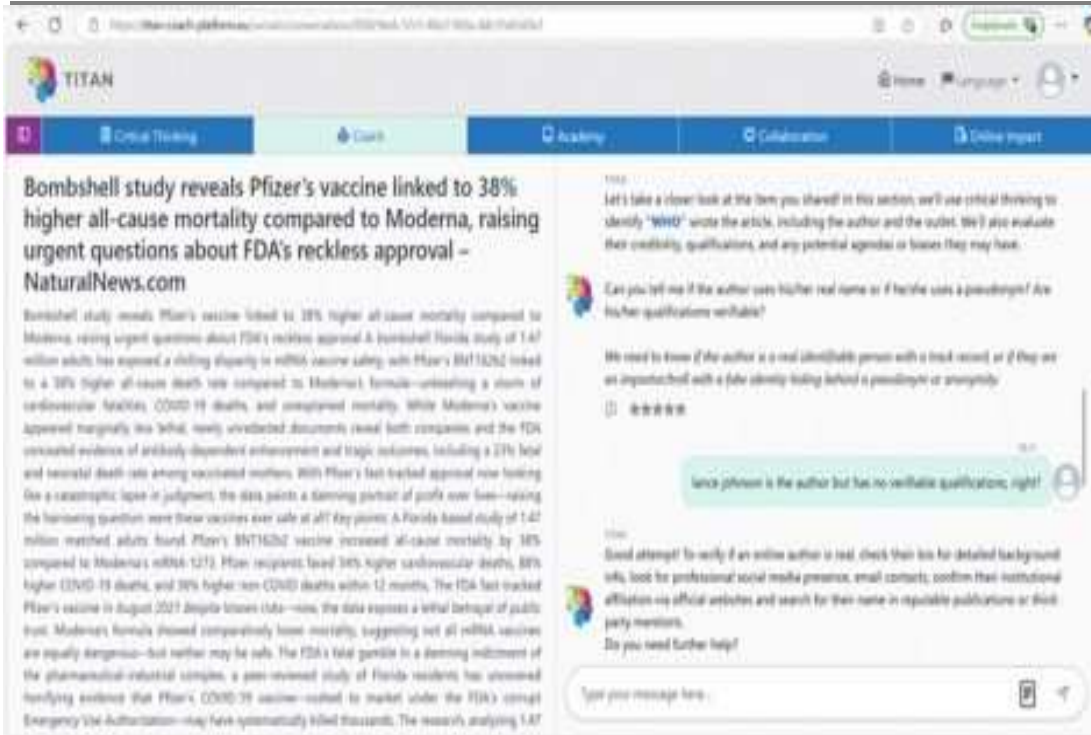
### SCREENSHOT 1

TITAN introduces itself with friendly language and emojis: "Hi! 🙋 I'm TITAN - your AI coach and content companion! 😊 I am here to explore 🔍 articles and social media posts with you - let's discover together! 🤝" Below this, TITAN provides a "Tip!" box explaining how to interact with the chatbot using commands like /stop, /explain, and /translate. This greeting establishes a helpful and collaborative tone for the user's interaction with TITAN, framing the process as a joint exploration of media content. After the friendly greeting, TITAN offers to show the user an example of a disinformation tactic directly related to the article they've shared. The prompt reads: "Disinformation shows up in many ways, but there are signs we can spot together. Curious to see an example of a tactic that might not directly relate to what you shared?" This clearly asks the user if they are interested in viewing an infographic illustrating a disinformation tactic as a next step in their learning process. The user's positive response ("hi yes plz") confirms their interest.



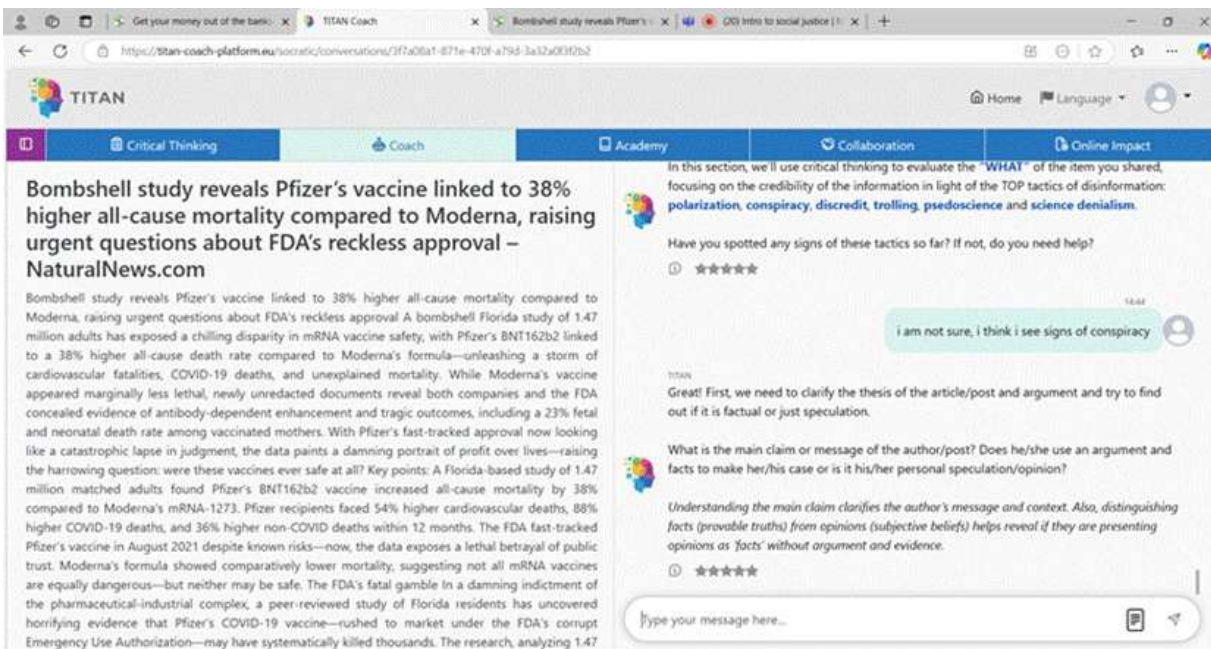
### SCREENSHOT 2:

This screenshot demonstrates the chatbot focusing on author verification within the 'WHO' analysis. The chatbot asks the user to check if the author uses a real name and has verifiable qualifications, explaining the rationale behind checking for anonymity. After the user identifies the author but states they lack verifiable qualifications, the chatbot provides positive feedback ('Good attempt!') and offers concrete steps ("check their bio," "look for professional social media presence," etc.) to guide the user on how to verify an author's credentials independently.



### SCREENSHOT 3:

This screenshot marks the transition to the 'WHAT' (content evaluation) phase. The chatbot introduces this stage by listing relevant disinformation tactics (polarization, conspiracy, etc.) and asks the user for initial observations. After the user mentions possibly seeing 'signs of conspiracy,' the chatbot guides the dialogue towards a foundational analysis step: prompting the user to identify the article's main claim or message and to consider whether it's presented as factual or speculative, explaining the rationale for this inquiry.



### SCREENSHOT 4:

This screenshot illustrates a critical thinking lesson within the chatbot interaction. The chatbot prompts the user to identify the central claim of the article ('WHAT'), a foundational step in critical analysis and signal detection. The user points to the title. The chatbot affirms this and immediately pivots to the next layer of critical thinking, guiding the user to examine the reasoning behind the claim by looking for logical fallacies and emotional appeals. This demonstrates the chatbot's method of progressively building critical thinking skills.



## SCREENSHOT 5:

This screenshot showcases the possibility that the user asks the chatbot to point out detected signs. The chatbot responds by listing specific disinformation signs present in the NaturalNews.com article: CONSPIRACY THEORY (suggesting a conspiracy involving the FDA and Pfizer), ATTACK CREDIBILITY (discrediting the FDA), and SECRET PLOT (a "bioweapon" narrative). Then it prompts further critical evaluation by asking whether the user detects the main signs of conspiracism, starting with the following sign: "Does it mention that powerful groups are secretly planning something behind our backs?" This illustrates the chatbot's method of teaching users to recognize common misinformation strategies.

