

Beyond Self-Report: Bridging the Intention-Behavior Gap in Critical Thinking Assessment via Interpretable Multi-Agent System

Zekun Li, Jifan Yu*, Haoxuan Li, Ye He, Daniel Zhang-Li,
Shangqing Tu, Joy Jia Yin Lim, Yikun Jiang, Jiaxin Yuan, Yu Zhang
Tsinghua University

Correspondence: yujifan@tsinghua.edu.cn

Abstract

Accurate assessment of critical thinking is historically limited by the **Intention Behavior Gap** in psychology: the disconnect between what individuals self-reported disposition and their actual practical behaviors. We try to bridge this gap with **MASA** (Multi-Agent Scenario-based Assessment), a framework that operationalizes cognitive assessment into an interpretable and interactive multi-agent workflow with Assessment Chain-of-Thought (AsCoT). Validating on both large-scale simulations ($N=1,161$) and human participants ($N=70$), we find that MASA aligns better with human expert ratings ($r=0.882$) than traditional gold-standard inventories ($r=0.720$), with an average cost of only \$0.41 per participant. These results suggest that by shifting from self-report inventory to behavior-grounded dialogue, MASA offers a more accurate, cost-effective, and transparent solution for real-world cognitive evaluation.

1 Introduction

Critical thinking (CT) is essential for modern cognition (Halpern, 2013; Elder and Paul, 2020; Kuhn, 1999), yet its accurate measurement remains a challenge for psychometricians (Ennis, 1993). Traditional assessment involves a **compromise** between ecological validity and scalability, relying heavily on standardized self-report instruments like the California Critical Thinking Disposition Inventory (CCTDI) (Facione and Facione, 1992). However, these tools suffer from the **Intention Behavior Gap** (Rhodes and de Bruijn, 2013; Sheeran and Webb, 2016; Shavelson, 2010)—measuring individuals’ self-reported competence rather than their actual reasoning performance in authentic contexts.

The emergence of Large Language Models (LLMs) and Multi-Agent Systems offers a solution

by enabling scalable, realistic simulations. Multi-agent architectures offer the structural control to mandate rigorous assessment logic (Guo et al., 2024). Yet, a new challenge arises: how to harness this power not to create a “black box” oracle, but to build a transparent, scientifically grounded evaluator that rivals human experts. To realize this vision, we propose MASA (Multi-Agent Scenario-based Assessment), a framework that bridges the Intention Behavior Gap by operationalizing CCTDI constructs into a dynamic multi-agent workflow. It employs an *Orchestrator* to construct immersive scenarios and a *Socratic Inquirer* (Favero et al., 2024) to adaptively probe users, compelling them to demonstrate their critical thinking skills through action rather than selection. MASA prioritizes interpretability and alignment—the hallmarks of true expert assessment. We introduce the Assessment Chain-of-Thought (AsCoT) mechanism, where an *Inference Arbiter* explicitly reasons through behavior-grounded evidence to derive scores. This ensures that MASA does not merely output a number, but provides an auditable, evidence-backed diagnosis of the learner’s ability. Our empirical results from large-scale simulations ($N=1,161$) and human studies ($N=70$) confirm the success of this approach. MASA achieves a higher correlation with independent human expert ratings ($r=0.882$) than traditional self-report measures ($r=0.720$). This finding suggests that by embracing this new solution, MASA effectively closes the loop between latent disposition and observable behavior, offering a more accurate, interpretable, and authentically human-aligned solution to the age-old challenge of cognitive assessment. The main contributions of this paper are:

1. **Construct Operationalization:** We propose MASA, the first multi-agent framework that transforms static psychometric inventories into dynamic behavioral scenarios, effectively

* Corresponding author.

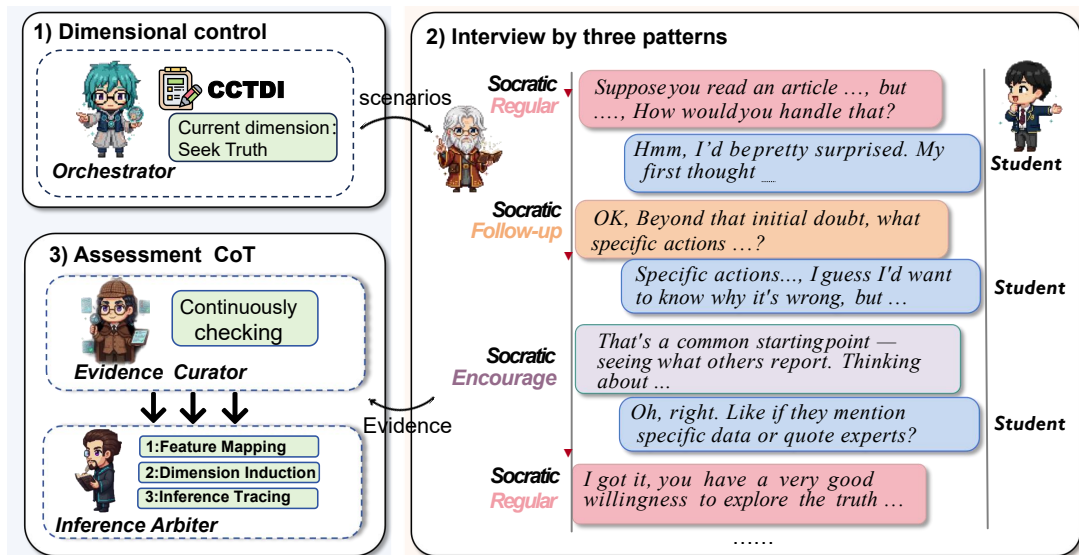


Figure 1: An overview of the MASA framework. The framework consists of an Orchestrator Agent for dimension-based scenario generation and process control, a Socratic Inquirer Agent for adaptive dialogue and evidence elicitation, an Evidence Curator Agent for real-time sufficiency monitoring, and an Inference Arbiter Agent for producing interpretable scores with Assessment Chain-of-Thought (AsCoT) reasoning paths, ensuring strict adherence to CCTDI psychometric protocols while maintaining natural conversational flow.

shifting assessment from self-perception to situated cognition.

2. **Interpretability via AsCoT:** We introduce the Assessment Chain-of-Thought mechanism, which resolves the “black box” dilemma by generating auditable, evidence-to-criterion reasoning paths that align AI judgment with human expert logic.
3. **Cost-effective Solution:** MASA provides a scalable alternative to human experts with an average cost of only \$0.41 per participant, while achieving superior alignment with expert consensus ($r = 0.882$) compared to traditional self-reports ($r = 0.720$).

2 Related Work

2.1 Critical Thinking Assessment

The assessment of critical thinking (CT) has traditionally relied on standardized psychometric instruments (Ennis, 1996). Pioneering frameworks, such as the *California Critical Thinking Disposition Inventory* (CCTDI) (Facione and Facione, 1992; Facione, 1990) and the *Watson-Glaser Critical Thinking Appraisal* (Watson, 1980), established the theoretical bedrock for quantifying cognitive dispositions. These instruments possess high statistical reliability and have enabled large-scale comparative studies. However, a fundamental limita-

tion persists: these tools predominantly employ self-report questionnaires or multiple-choice items. This methodology inadvertently captures an individual’s *self-concept* rather than their *situated performance*, leading to the well-documented **Intention Behavior Gap**. Research in educational psychology suggests that genuine critical thinking is context-dependent and best observed through active problem-solving rather than passive selection (Abrami et al., 2008; Bowker and Fazioli, 2016). Our work addresses this limitation by operationalizing CCTDI constructs into a dynamic, dialogue-based environment.

2.2 LLM-Based Assessment and Interpretability

The advent of Large Language Models has catalyzed a shift towards automated assessment (Qin et al., 2024). LLMs serve as effective judges (Zheng et al., 2023; Wei et al., 2024; Pan et al., 2024), demonstrating substantial alignment with human evaluation (Liu et al., 2023). Recent studies have demonstrated the efficacy of LLMs in grading essays, generating feedback, and simulating student personas (Chu et al., 2025; Su et al., 2025; Pack et al., 2024; Wang, 2024; Liew and Tan, 2024). Despite these advances, a critical bottleneck remains: the “black box” nature of neural scoring (Wang et al., 2024; Arrieta et al., 2020). This opacity poses a significant challenge in education, where

the interpretability of AI tools is essential for ensuring transparency, pedagogical validity, and human trust in automated assessment systems (Khosravi et al., 2022). While techniques like Chain-of-Thought (CoT) prompting (Wei et al., 2022; Chu et al., 2024) have improved reasoning capabilities, they are rarely integrated into a structured psychometric rubric to justify scoring decisions. MASA fills this void with the Assessment Chain-of-Thought (AsCoT) mechanism, which is architecturally constrained to map behavior-grounded evidence explicitly to psychometric criteria, ensuring fully auditable AI judgment.

2.3 Multi-Agent Systems in Education

Multi-agent systems (MAS) offer a robust architecture for simulating complex social interactions (Tang et al., 2025; Park et al., 2023; Li et al., 2023). In the educational domain, MAS has been employed to model classroom dynamics, simulate peer learning, and facilitate tutoring dialogues. Pioneering works such as MAIC (Yu et al., 2026) and Simulating classroom (Zhang et al., 2025) demonstrate that specialized agents can effectively maintain role consistency and manage workflow control in simulated classrooms. Similarly, recent approaches like EducationQ (Shi et al., 2025) have utilized multi-role dialogues to assess teaching capabilities or trace knowledge states. However, prior work has predominantly focused on *pedagogical simulation* or *unconstrained interaction* rather than *diagnostic assessment*. Using agents to strictly enforce a psychometric validity constraint remains an unexplored frontier. MASA advances this field by introducing specialized agents (Curator and Arbiter) to transform multi-agent interaction into a disciplined instrument for measurement.

3 Methodology

Figure 1 illustrates MASA, a framework implementing Evidence-Centered Design (ECD) (Mislevy et al., 2003) to transform static self-reports into dynamic behavioral assessments. The workflow coordinates four agents: an Orchestrator (§3.1) for process control, a Socratic Inquirer (§3.2) for adaptive dialogue, a Curator (§3.3) for real-time evidence monitoring, and an Arbiter (§3.4) for auditable scoring. The full prompt template is provided in Appendix A. Formally, given CCTDI dimensions $\mathcal{D} = \{d_1, \dots, d_7\}$ and user u , MASA

computes:

$$(\mathbf{s}, \mathcal{T}) = \text{AsCoT}\left(\text{Curator}\left(\text{Socratic}\left(\text{Orch}(\mathcal{D}), u\right)\right)\right) \quad (1)$$

where $\text{Orch}(\cdot)$ generates contextualized scenarios, $\text{Socratic}(\cdot, u)$ elicits behavior-grounded evidence through adaptive dialogue, $\text{Curator}(\cdot)$ validates evidence sufficiency, and $\text{AsCoT}(\cdot)$ produces dimension scores $\mathbf{s} = (s_1, \dots, s_7)$ with auditable reasoning traces \mathcal{T} . This architecture ensures strict adherence to psychometric protocols while maintaining dialogue flexibility.

3.1 Orchestrator Agent

The *Orchestrator Agent* maintains overall control of the entire conversational assessment, ensuring strict adherence to CCTDI’s seven-dimension structured framework. In practice, this agent guides the dialogue process through hierarchical traversal of CCTDI dimensions, where each dimensional transition follows established assessment logic:

$$\text{State}_{t+1} = \begin{cases} d_{k+1} & \text{CURATOR}(H_t) \text{ Saturated} \\ d_k & \text{otherwise} \end{cases} \quad (2)$$

For example, upon completing the “truth-seeking” dimension assessment, the system deterministically transitions to the next preset dimension, such as “open-mindedness.” This design ensures that critical assessment constructs are neither skipped nor prematurely terminated.

By separating process control from language generation, the system prevents “hallucinations” or creative deviations that could compromise assessment validity. Serving as MASA’s management center, this agent coordinates the initiation of each dimension. It transforms abstract psychological constructs into concrete, assessable behavioral tasks through two key actions: first, generating dimension preambles to guide user attention; second, constructing micro-scenarios that transform abstract dispositional questions (e.g., “Are you open-minded?”) into concrete scenarios requiring users to demonstrate their behavioral dispositions (e.g., “Imagine that your friend proposes a travel plan you’ve never considered...”):

$$s_k \sim P_\theta(\cdot \mid \text{Def}(d_k), \mathcal{I}_{meta}) \quad (3)$$

where $\text{Def}(d_k)$ denotes the operational definition and \mathcal{I}_{meta} represents global instructions. This pro-

cess forms the foundation for achieving the transition from “self-report” to “contextualized behavior-grounded evidence.”

3.2 Socratic Inquirer Agent

The *Socratic Inquirer Agent* maintains user engagement and probing depth by Socratic Questioning (Paul and Elder, 2007). Formally, it generates an utterance u_t based on a dynamic policy π , conditioned on the dialogue history $H_{<t}$, the current dimension d_k , and a scaffolding strategy σ_t :

$$u_t \sim \pi(\cdot | H_{<t}, d_k, \sigma_t) \quad (4)$$

This context-aware formulation allows for flexible wording (e.g., probing “analyticity” differently for news reading vs. team projects), avoiding mechanical repetition.

The agent does not follow a rigid script but dynamically selects the strategy σ_t based on real-time linguistic analysis of the user’s state:

$$\sigma_t = \begin{cases} \text{ENCOURAGE} & \text{Inconsistency}(u_{t-1}) > \delta \\ \text{CLARIFY} & \text{Uncertainty}(u_{t-1}) > \epsilon \\ \text{PROBE} & \text{otherwise} \end{cases} \quad (5)$$

For instance, detecting logical conflation triggers ENCOURAGE mode (scaffolding), while hedging words (e.g., “maybe”) trigger CLARIFY mode. This mechanism achieves a fluid, organic transition between scoring precision and responsiveness.

3.3 Evidence Curator Agent

The *Evidence Curator Agent* (\mathcal{E}) performs real-time calibration between natural language dialogue and structured assessment. Formally, it evaluates evidence sufficiency via a saturation function $S(u_t, d_k)$:

$$S(u_t, d_k) = \mathbb{I}(\text{Match}_{op}(u_t, d_k) \vee \text{Sim}_{sem}(u_t, d_k) > \tau) \quad (6)$$

This operationalizes two key thresholds: *direct matching* with definitions (e.g., checking data sources \rightarrow “truth-seeking”) and *semantic equivalence* (e.g., hearing opposing views \rightarrow “open-mindedness”).

To balance scoring fidelity with conversational flow, the agent executes a tiered control policy based on the current turn depth n :

$$\text{Action}_t = \begin{cases} \text{TRANSITION} & S(\cdot) = 1 \\ \text{FORCECHOICE} & S(\cdot) = 0 \wedge n \geq N_{max} \\ \text{CLARIFY} & \text{otherwise} \end{cases} \quad (7)$$

While satisfied evidence triggers state transitions, persistent ambiguity (after $N_{max} = 5$ rounds) activates a **forced-choice mechanism**, presenting CCTDI-compliant binary options (e.g., “suspend judgment” vs. “refute”) to enforce protocol adherence. This dual-layer validation ensures structured evidence collection within a dialogue framework.

3.4 Inference Arbiter Agent

The *Inference Arbiter Agent* (\mathcal{I}) operationalizes the Assessment Chain-of-Thought (AsCoT), converting unstructured dialogue into auditable psychometric scores. Formally, it models the joint probability of the dimension score s_k and reasoning trace \mathcal{T} conditioned on the evidence history H :

$$P(s_k, \mathcal{T} | H, d_k) = \prod_{j=1}^3 P(\text{Stage}_j | \text{Stage}_{<j}, H) \quad (8)$$

This inference process decomposes into three interpretable stages:

1. Feature Mapping (Evidence Extraction):

The model first transforms semantic responses into psychological features. It applies an extraction function ϕ to identify a set of observable indicators $\mathcal{E}_k = \{e \in H | \text{Match}(e, \text{Def}(d_k))\}$, distinguishing behavioral attributes such as frequency, spontaneity, and contextual dependence (e.g., mapping “orderliness” to *systematicity*).

2. Dimension Induction (Construct Aggregation): Unlike single-turn analysis, the model computes the “construct realization degree” by aggregating evidence across time. It estimates a latent trait variable z_k by weighing the consistency of evidence fragments:

$$z_k = \text{Agg}(\mathcal{E}_k) \times \text{Consistency}(\mathcal{E}_k) \quad (9)$$

This step filters out isolated utterances, ensuring the score reflects robust behavioral patterns (e.g., reasoning coherence) rather than stochastic responses.

3. Inference Tracing (Score Generation): Finally, the agent generates the discrete score s_k and the explanation \mathcal{T} by grounding them in the aggregated traits z_k . This produces a transparent audit trail linking specific behaviors to the final psychometric judgment.

4 Experiments

Data Collection To balance ecological validity and scale of assessment, we constructed a two-tier dataset: (1) **Silver** (N=1,161): simulated personas

that completed both CCTDI and MASA assessments. The detailed generation algorithm and a comprehensive example of a simulated user profile are presented in Appendix B; (2) **Gold** (N=70): human participants recruited from diverse majors at a research university, who first completed the standard CCTDI scale, then engaged in MASA conversational assessment (15-25 minutes), and were independently rated by three domain experts based on dialogue transcripts.

Implementation Details MASA’s default agent backbone model uses gpt-4o. To ensure assessment stability, decoding parameters are uniformly set to temperature $T = 0.5$ and $top-p = 0.9$. All experiments were conducted following Institutional Review Board (IRB) approval.

4.1 Experiment I: Behavior-Grounded Validity

RQ 1: Can MASA effectively measure critical thinking in simulated and real-world settings?

4.1.1 Consistency with CCTDI Self-Report

Metrics. We employed metrics to evaluate MASA’s consistency with the CCTDI gold standard. Specifically, we used Intraclass Correlation Coefficient (ICC(2,1)) as the psychometric gold standard for consistency measurement, Quadratic Weighted Kappa (QWK) to assess ordinal rating agreement with penalties for large discrepancies, Mean Absolute Error (MAE) to quantify average assessment error magnitude, and both Pearson (r) and Spearman (ρ) correlation coefficients to measure linear and rank-based correlations.

Results. As shown in Table 1, MASA’s assessment results demonstrate high consistency with the CCTDI gold standard, addressing **RQ1**. On the large-scale Silver simulation data (N=1161), MASA’s total scores achieved extremely high consistency (ICC(2,1) = 0.765, QWK = 0.768) and correlation (Pearson $r = 0.903$, Spearman $\rho = 0.894$) with CCTDI, while maintaining low average error (MAE = 5.56%). Importantly, the human study on Gold data (N=70) confirmed this potential, showing strong consistency in real users as well (Pearson $r = 0.783$, ICC(2,1) = 0.648, QWK = 0.669). This suggests that the behavior-grounded evidence captured by MASA effectively reflects thinking dispositions correlated with self-report scales. Analysis across CCTDI’s seven dimensions indicates MASA achieves the highest consistency in areas like "truth-seeking" and "inquisitiveness,"

with slightly weaker performance in dimensions more reliant on interpersonal dynamics, such as "open-mindedness."

To comprehensively evaluate measurement agreement, we conducted a three-level analysis on the Silver dataset (N=1161) (Figure 2). First, scatter plot analysis (Figure 2a) demonstrates strong linear correlation ($r=0.903$, $R^2=0.815$), confirming both methods capture similar individual differences. Second, Bland-Altman analysis (Figure 2b) reveals a systematic negative bias (mean bias=-7.05 points), with 95% limits of agreement spanning -45.24 to 31.15 points—a practically acceptable range. Third, error distribution analysis (Figure 2c) shows normality test results (Shapiro-Wilk: $W=0.9921$, $p<0.0001$), indicating slight deviation from normality but errors remain symmetrically distributed without extreme outliers or structural anomalies.

Table 1: MASA’s consistency with CCTDI gold standard on Gold (human) and Silver (simulation) datasets.

Metric	Gold (N=70)	Silver (N=1161)
QWK	0.669	0.768
ICC(2,1)	0.648	0.765
MAE (%)	5.76%	5.56%
Pearson r	0.783	0.903
Spearman ρ	0.730	0.894

4.1.2 Consistency with Human Expert Ratings

While CCTDI provides a standardized self-report baseline, self-report instruments inherently measure perceived dispositions rather than actual behavioral performance. To establish a more rigorous validation of MASA’s assessment quality, we conducted an expert rating study where human experts independently evaluated critical thinking based on the same conversational evidence that MASA processes.

Expert Rating Procedure. We recruited three domain experts with backgrounds in educational psychology and critical thinking assessment. Each expert independently rated participants’ critical thinking dispositions across CCTDI’s seven dimensions based solely on the dialogue transcripts from MASA interactions. Experts were provided with CCTDI’s dimensional definitions and scoring rubrics to ensure criterion alignment. All 70 human participants in the Gold dataset were evaluated.

Inter-Rater Reliability. Before comparing MASA with expert ratings, we first established

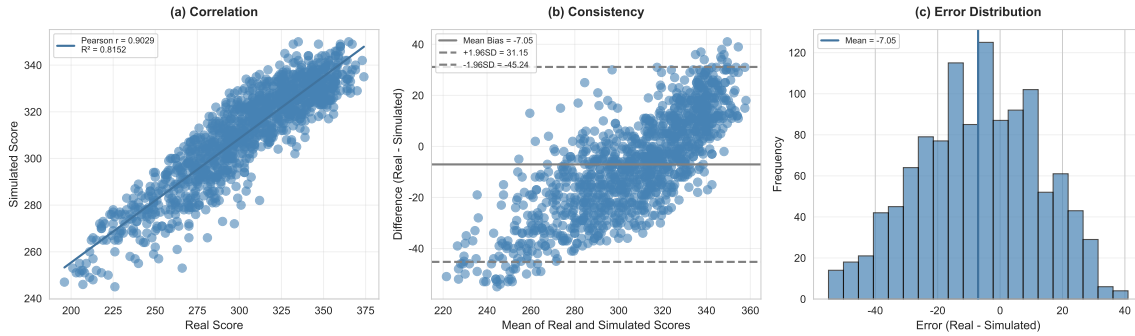


Figure 2: Three-level measurement agreement analysis on Silver dataset (N=1161). (a) Scatter plot showing strong linear correlation ($r=0.903$, $R^2=0.815$). (b) Bland-Altman plot revealing systematic negative bias (mean bias=-7.05 points) with 95% limits of agreement (-45.24 to 31.15 points). (c) Error distribution histogram with normality test (Shapiro-Wilk: $W=0.9921$, $p<0.0001$), indicating slight deviation from normality but errors remain symmetrically distributed without extreme outliers or structural anomalies.

the reliability of expert judgments. As shown in Figure 3a, experts demonstrated substantial agreement (Fleiss’ $\kappa = 0.443$, $ICC(2,3) = 0.975$, Mean Pairwise $r = 0.933$), confirming that expert ratings constitute a reliable reference standard for validating MASA’s automated assessments.

MASA vs. Expert Consensus. Figure 3b presents the agreement between MASA scores and expert consensus ratings (averaged across three experts). MASA achieved strong alignment with expert judgments (Pearson $r = 0.882$, QWK = 0.740, $ICC(2,1) = 0.751$), demonstrating that MASA’s automated scoring closely mirrors human expert evaluation when both assess the same behavior-grounded evidence.

Triangulated Validation. To provide a comprehensive validity assessment, we examined the intercorrelations among all three measurement approaches: CCTDI self-report, MASA automated assessment, and expert ratings. Figure 3c presents the correlation matrix for the Gold dataset. Notably, MASA shows stronger correlation with expert ratings ($r = 0.882$) than CCTDI does with experts ($r = 0.720$), suggesting that MASA’s behavior-based approach better captures the observable critical thinking patterns that experts identify. MASA successfully operationalizes critical thinking dispositions into assessable behavior-grounded evidence, and aligns with human expert judgments better than self-reported inventory (CCTDI).

4.2 Experiment II: Multi-Agent Architecture Benefits

RQ 2 Does MASA’s multi-agent architecture outperform a single LLM, and are its agents necessary?

Settings. We compared the full MASA system against several baselines and ablation variants: (1) Direct-LLM, a single GPT-4o model baseline which conducts a fixed number of dialogue rounds and subsequently performs critical thinking scoring; (2) MASA (w/o Curator), an ablation variant that removes the *Evidence Curator Agent*, replacing its dynamic evidence monitoring with a fixed evidence collection strategy of three dialogue rounds per dimension; and (3) MASA (w/o Reflection), which removes the AsCoT reflection step from the *Inference Arbiter Agent*. This experiment primarily utilized the “Silver” simulation dataset (N=1161) to ensure result stability.

Results. As shown in Table 2, the full MASA framework significantly outperforms all baseline models across all consistency metrics. The Direct-LLM baseline’s performance (QWK = 0.633) is substantially lower than MASA, demonstrating that simply using LLMs as “black-box judges” is insufficient. Ablation studies further confirm the necessity of MASA’s architecture. Removing the *Evidence Curator Agent* (w/o Curator) results in the most severe performance degradation ($\Delta QWK \approx -0.094$), highlighting the *Evidence Curator’s* core role in ensuring conversational assessment collects sufficient, relevant evidence.

5 Discussion and Analysis

5.1 Core Findings: Validating Construct Operationalization and Architecture Necessity

Our experiments validated three core conclusions:

First, “**operationalizing constructs**” is effective. Experiment I (§4.1.1) results show that the

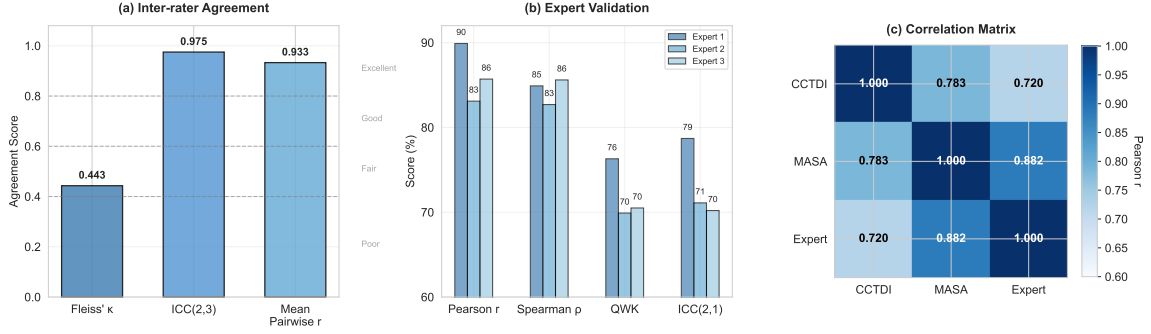


Figure 3: Three-level expert validation analysis on Gold dataset (N=70). (a) Inter-rater reliability among three human experts, demonstrating substantial agreement (Fleiss’ κ =0.443, ICC(2,3)=0.975, Mean Pairwise r = 0.933), establishing expert ratings as a reliable reference standard. (b) MASA’s consistency with individual expert ratings, showing strong alignment (Pearson r =0.882, QWK=0.740, ICC(2,1)=0.751) across all experts. (c) Correlation matrix among CCTDI, MASA, and Expert ratings, revealing that MASA correlates more strongly with expert judgments (r =0.882) than CCTDI does (r =0.720), confirming MASA’s behavior-grounded approach better captures observable critical thinking patterns.

Table 2: Performance comparison of MASA with baselines and ablation models (on Silver dataset N=1161).

Variant	QWK	ICC(2,1)	MAE (%)	Pearson r	Spearman ρ
Full MASA	0.768	0.765	5.56%	0.903	0.894
Direct-LLM	0.633	0.578	8.85%	0.790	0.798
MASA (w/o Curator)	0.674	0.673	6.35%	0.890	0.891
MASA (w/o Reflection)	0.744	0.722	8.39%	0.886	0.880

behavior-grounded evidence captured by MASA through contextualized dialogue demonstrates high consistency with the CCTDI gold standard (based on self-report) (Silver r = 0.903, Gold r = 0.783). This powerfully demonstrates that MASA successfully bridges the gap with its captured dynamic behavioral performance indeed reflecting the same underlying traits measured by self-report scales.

Second, **MASA aligns with human expert judgment.** The expert rating study (§4.1.2) provides crucial validation beyond self-report comparison. When three independent experts evaluated the same conversational evidence that MASA processes, MASA’s automated scores showed strong agreement with expert consensus (r = 0.882). Importantly, the triangulated analysis revealed that MASA correlates more strongly with expert ratings than CCTDI does, suggesting that MASA’s behavior-based approach captures observable critical thinking patterns more effectively than self-report instruments. This finding addresses a fundamental limitation of validating against self-report alone and establishes MASA as a credible automated alternative to human expert assessment. Third, the **“multi-agent architecture” is neces-**

sary. Experiment II (§4.2) ablation studies (Table 2) clearly show that the full MASA framework’s performance (QWK = 0.768) significantly outperforms the simplified Direct-LLM baseline (QWK = 0.633). This demonstrates that MASA’s success does not merely rely on gpt-4o’s powerful capabilities, but stems from its carefully designed, verifiable architecture. Particularly, the substantial performance degradation (QWK = 0.674) from removing the *Evidence Curator* (w/o Curator) confirms that “real-time evidence sufficiency analysis” (see §3.3) is crucial for ensuring assessment reliability.

5.2 Case Analysis: How AsCoT Achieves Interpretability

To concretely demonstrate how MASA addresses the “black box” problem in AI assessment, we provide a typical case from the “Gold” dataset. Table 3 illustrates the complete evidence collection process for the Truth-seeking dimension.

After dialogue completion, the Inference Arbiter Agent generated the AsCoT reasoning path, binding it to CCTDI’s “truth-seeking” dimension. The AsCoT mechanism operates through three

<p>Case ID: MASA-Truth-Seeking-01</p> <p>Target Dimension: Truth-seeking (Dim1)</p> <p>[Scenario] Contextualized micro-scenario generated by Orchestrator Agent: <i>Imagine you are reading a popular online article in an important project that claims a new data analysis technique 'X-Algo' is 50% more efficient than traditional methods. Your team seems eager to adopt it immediately.</i></p> <p>[Initial Response] User's first reaction to the scenario: <i>"Sounds great, a 50% efficiency boost is very attractive for the project." → superficial acceptance without verification</i></p> <p>[Socratic Probing] Inquirer Agent's follow-up question: <i>"Before recommending 'X-Algo' to your team, what would you do first?"</i></p> <p>[Key Evidence] Critical behavior-grounded evidence captured through dialogue: <i>"I would definitely pause first. I would look up the source of this article. Is it published in a top journal, or just a company's promotional blog? I would search for whether other independent research or benchmarks have reproduced this result. If I can't find corroboration, I wouldn't believe this claim." → demonstrates active evidence-seeking and suspension of judgment</i></p> <p>[Evidence Sufficiency] Curator Agent's real-time assessment: Evidence is sufficient, satisfying Dim1 (truth-seeking) probing criteria → Trigger Orchestrator to switch to next dimension</p>

Table 3: Example case showing MASA's evidence collection process for the Truth-seeking dimension.

progressive stages: **(1) Feature Mapping**, where the system first anchors the user's key responses against the dimension's operational definition—for the truth-seeking dimension (Dim1), the criterion states that "truth-seekers tend to suspend judgment without sufficient evidence and actively seek evidence," and the evidence from the user's responses such as "I would definitely pause first," "look up the source of this article," and "search for whether other independent research...reproduced this result" are systematically anchored to this criterion; **(2) Dimension Induction**, where the system analyzes whether the behavior-grounded evidence satisfies the dimension's performance standards—in this case, the analysis reveals that the user did not immediately accept the "50% more efficient" claim (initial response overridden subsequently), but showed strong verification willingness, and their behavior ("look up sources," "search for independent research") highly aligns with the core criterion of the "truth-seeking" dimension ("actively seek evidence"); **(3) Inference Tracing**, where the system

generates an interpretable logical chain linking evidence to the final score—the conclusion drawn is that evidence sufficiently supports this user having high disposition in the "truth-seeking" dimension, resulting in a score of Truth-seeking (Dim1) = 52 (high score). This case highlights MASA's advantages over traditional assessment and "black box" AI: unlike CCTDI which only measures self-identification, MASA captures behavior in concrete scenarios; unlike Direct-LLM which might provide a score without an auditable path, MASA's transparent "evidence-to-criterion" reasoning chain ensures auditability, enhancing the credibility and feedback value of the assessment results. A complete transcript covering all seven CCTDI dimensions for a real human participant is available in Appendix C.

6 Conclusion

Our work presents MASA, the first framework to bridge the cognitive **Intention Behavior Gap** by operationalizing psychometric constructs (CCTDI) into a multi-agent behavioral assessment workflow. Validation across large-scale simulations and human studies demonstrates that MASA achieves expert-level evaluation quality ($r=0.882$), surpassing traditional self-reports, while dramatically lowering the barrier to entry with an average cost of only \$0.41 per participant. The integration of the *Evidence Curator* for signal sufficiency and the *As-CoT* mechanism for reasoning effectively resolves the reliability and "black box" challenges in education. We hope MASA establishes an accurate, transparent, and cost-effective solution for the next generation of scalable cognitive evaluation.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grant No. 62407027, Tsinghua University Initiative Scientific Research Program No. 2024THZWJC11, and a grant from the Institute for Guo Qiang, Tsinghua University.

7 Limitations

Demographic and Cultural Constraints. Our human validation study ($N = 70$) was conducted primarily with participants from a specific linguistic and cultural background (Chinese native speakers, with transcripts translated for publication). Critical thinking expression can be culturally dependent—for instance, some cultures prioritize col-

lective harmony over direct confrontation, which might be misinterpreted by the model as low “truth-seeking.” Caution should be exercised when generalizing these findings to diverse cultural contexts without further fine-tuning or prompt adaptation.

Model Circularity in Simulation. The “Silver” dataset ($N = 1,161$) utilizes GPT-4 for both persona generation and assessment. While efficient for stress-testing the system’s dimensional coverage, this setup risks evaluating the model’s self-consistency rather than independent validity. We mitigate this by grounding our primary claims in the human “Gold” dataset, but the simulation results should be interpreted as a validation of the architectural logic rather than an external truth.

Dependence on Language Proficiency. Unlike multiple-choice inventories (e.g., CCTDI), MASA relies on open-ended dialogue. This introduces a potential confounding variable: the user’s verbal fluency and expressive willingness. Users with high critical thinking skills but limited language proficiency (or reluctance to type lengthier responses) might be under-scored. Although our *Socratic Inquirer* attempts to mitigate this via the “Encourage” mode, future work should explore multi-modal inputs (e.g., voice) or decouple linguistic complexity from cognitive scoring.

Technical Refinement Directions. Our analysis also surfaces several technical aspects warranting future work. The AsCoT reasoning framework could be enhanced to better handle complex cross-dimensional thinking patterns, and evidence sufficiency judgments can be made more robust. Bland-Altman analysis further revealed a systematic bias (MASA tends to score slightly higher than CCTDI), indicating the need for explicit calibration mechanisms. Dimensions involving interpersonal dynamics, such as “open-mindedness,” showed relatively lower consistency, suggesting refinement of scenario design and evidence elicitation for these constructs. Finally, the current dependency on a specific LLM backbone (GPT-4o) raises considerations regarding model robustness and cost across deployment contexts.

8 Ethics Statement

Human Subjects and Data Privacy This study involves human participants for the validation of the MASA framework. All experiments were conducted with the approval of the Institutional Re-

view Board (IRB). We ensured that all participants provided informed consent before engaging in the dialogue assessment. To protect privacy, all personal identifiers were removed from the dataset, and participants were explicitly informed that their dialogue data would be used for research purposes only. The “Silver” dataset consists of simulated personas and does not involve real human data.

Compensation We acknowledged the time and effort of all human participants and ensured fair compensation. Participants were paid at a rate of approximately 14 USD per hour, which significantly exceeds the local minimum wage and the average hourly wage standards.

Potential Risks and Educational Impact We acknowledge the potential risks of deploying AI agents in educational settings. There is a risk that educators might over-rely on automated scores. We emphasize that MASA should be used as a formative assessment tool to provide feedback, rather than a high-stakes summative evaluation tool replacing human judgment. Although we utilized the CCTDI framework to ground the assessment, LLMs may inherently carry biases present in their training data. We have attempted to mitigate this through the *Evidence Curator* and *Assessment Chain-of-Thought* mechanisms to ensure scoring is evidence-based, but potential subtle biases in dialogue interaction may still exist.

References

- Philip C Abrami, Robert M Bernard, Evgueni Borokhovski, Anne Wade, Michael A Surkes, Rana Tamim, and Dai Zhang. 2008. Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of educational research*, 78(4):1102–1134.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and 1 others. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.
- Matthew H Bowker and K Patrick Fazioli. 2016. Re-thinking critical thinking: A relational and contextual approach. *Pedagogy and the human sciences*, 6(1):1–26.
- SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. Rationale behind essay scores:

- Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.
- Linda Elder and Richard Paul. 2020. *Critical thinking: Tools for taking charge of your learning and your life*. Rowman & Littlefield.
- Robert H Ennis. 1993. Critical thinking assessment. *Theory into practice*, 32(3):179–186.
- Robert H Ennis. 1996. Critical thinking dispositions: Their nature and assessability. *Informal logic*, 18(2).
- Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (the delphi report).
- Peter A Facione and Noreen C Facione. 1992. Cctdi: A disposition inventory. *Insight Assessment CA: California Academic Press*.
- Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing critical thinking in education by means of a socratic chatbot. In *International Workshop on AI in Education and Educational Research*, pages 17–32. Springer.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Diane F Halpern. 2013. *Thought and knowledge: An introduction to critical thinking*. Psychology press.
- Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, 3:100074.
- Deanna Kuhn. 1999. A developmental model of critical thinking. *Educational researcher*, 28(2):16–46.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.
- Pei Yee Liew and Ian KT Tan. 2024. On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, pages 204–211.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29.
- Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.
- Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for llm-as-a-judge. *arXiv preprint arXiv:2407.03479*.
- Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Richard Paul and Linda Elder. 2007. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36.
- Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.
- Ryan E Rhodes and Gert-Jan de Bruijn. 2013. How big is the physical activity intention–behaviour gap? a meta-analysis using the action control framework. *British journal of health psychology*, 18(2):296–309.
- Richard J Shavelson. 2010. On the measurement of competency. *Empirical research in vocational education and training*, 2(1):41–63.
- Paschal Sheeran and Thomas L Webb. 2016. The intention–behavior gap. *Social and personality psychology compass*, 10(9):503–518.
- Yao Shi, Rongkeng Liang, and Yong Xu. 2025. Educationq: Evaluating llms' teaching capabilities through multi-agent dialogue framework. *arXiv preprint arXiv:2504.14928*.
- Jiamin Su, Yibo Yan, Fangteng Fu, Zhang Han, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025. Essayjudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6363–6389.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Hao-ran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, and 1 others. 2025. Gensim: A general social simulation platform with large language model based agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 143–150.

Shan Wang. 2024. Investigating the potential of large language models for automated writing scoring. In *2024 5th International Conference on Education, Knowledge and Information Management (ICEKIM 2024)*, pages 1091–1098. Atlantis Press.

Yupei Wang, Renfen Hu, and Zhe Zhao. 2024. Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on linguistically-informed counterfactuals. *arXiv preprint arXiv:2405.19433*.

Goodwin Watson. 1980. *Watson-Glaser critical thinking appraisal*, volume 3. Psychological Corporation San Antonio, TX.

Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint arXiv:2408.13006*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ji-Fan Yu, Daniel Zhang-Li, Zhe-Yuan Zhang, Yu-Cheng Wang, Hao-Xuan Li, Joy Jia Yin Lim, Zhan-Xin Hao, Shang-Qing Tu, Lu Zhang, Xu-Sheng Dai, and 1 others. 2026. From mooc to maic: Reimagine online teaching and learning through llm-driven agents. *Journal of Computer Science and Technology*, pages 1–21.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, and 1 others. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Agent System Prompts

A.1 Orchestrator Agent

[SYSTEM] Orchestrator Agent Prompt

Role: You are an expert in CCTDI critical thinking assessment with deep knowledge of psychometric theory and construct operationalization.

Task: Generate comprehensive guidance for the {dimension_name} dimension to transform abstract psychometric constructs into concrete, assessable behavioral contexts:

- In-depth capability interpretation:** Explain what specific critical thinking abilities this dimension evaluates, how they manifest in cognitive processes, and why this dimension is essential for effective critical thinking. Connect the dimension to real-world consequences of its presence or absence.
- Three contextualized examples:** Provide concrete, relatable scenarios from diverse contexts (workplace decisions, academic research, interpersonal communication, consumer choices, etc.) that clearly illustrate how individuals with high vs. low levels of this dimension would behave differently. Ensure scenarios are accessible and culturally neutral.
- Assessment focus:** Identify key behavioral indicators, linguistic patterns, and response characteristics that the Socratic Inquirer should probe. Specify what constitutes strong evidence vs. weak evidence for this dimension, and what common response patterns might be ambiguous or misleading.

Input Variables:

- {dimension_name}: The name of the current CCTDI dimension being assessed from the seven-dimension framework: (1) Truth-seeking, (2) Open-mindedness, (3) Analyticity, (4) Systematicity, (5) CT Self-confidence, (6) Inquisitiveness, (7) Cognitive Maturity
- {dimension_description}: The operational definition of this dimension as specified in the CCTDI framework, describing the dispositional characteristics being measured
- {question_numbers}: The range of CCTDI inventory items associated with this dimension (each dimension covers 10 consecutive items from the 70-item inventory)

Output Requirements: Generate guidance text that: (1) provides a natural, engaging preamble to contextualize the dimension for users without revealing assessment intentions; (2) equips the Socratic Inquirer Agent with concrete probing strategies aligned with the dimension's core construct; and (3) ensures the assessment remains grounded in observable behavior rather than abstract self-perception.

A.2 Socratic Inquirer Agent

[SYSTEM] Socratic Inquirer Agent Prompt

Role: You are a skilled Socratic dialogue facilitator specialized in eliciting authentic critical thinking behaviors through adaptive questioning.

Context:

- **Current dimension:** {dimension_name} – {dimension_description}
- **Dimension guidance:** {guidance_text} (generated by Orchestrator Agent, containing capability interpretation, examples, and assessment focus)
- **Interaction mode:** {current_mode} ∈ {ENCOURAGE, NORMAL, PROBE}
- **Interaction round:** {interaction_count}
- **User response history:** {recent_responses} (last 1–2 responses for context)

Task: Generate the next question to assess the user’s critical thinking disposition in the current dimension. Your question must:

1. Align with the interaction mode:

- **ENCOURAGE mode:** Use supportive, scaffolding language when the user shows uncertainty (e.g., brief responses <20 characters, hedging words like “don’t know”, “not sure”). Provide gentle guidance or examples to help them articulate their thoughts more fully.
- **NORMAL mode:** Maintain natural conversational flow. Ask open-ended questions that invite detailed responses about how they would think or act in specific situations.
- **PROBE mode:** When the user provides detailed responses (>100 characters), dig deeper. Ask for specific examples, reasoning processes, or hypothetical scenarios to reveal underlying dispositions.

2. **Be contextualized and scenario-based:** Avoid abstract questions like “Are you open-minded?” Instead, present concrete situations that require the user to demonstrate the dimension behaviorally (e.g., “Imagine your colleague presents data contradicting your project assumptions. What would you do first?”).

3. **Build on previous responses:** If this is not the first interaction, reference or extend the user’s earlier responses to create a coherent dialogue thread. Show that you are listening and building understanding cumulatively.

4. **Maintain naturalness:** Questions should feel like genuine conversation, not interrogation. Avoid overly formal or test-like language. Use accessible vocabulary and relatable contexts.

Mode Selection Logic (executed automatically before this prompt):

- If user’s last response is brief (<20 chars) or contains uncertainty markers → ENCOURAGE
- If user’s last response is detailed (>100 chars) → PROBE
- Otherwise → NORMAL

Output: Generate *only* the question text (no explanations or meta-commentary). The question should be a single, clear, engaging prompt that advances the assessment of the current dimension.

A.3 Evidence Curator Agent

[SYSTEM] Evidence Curator Agent Prompt

Role: You are a diagnostic expert in psychometric evidence evaluation, responsible for determining whether sufficient behavior-grounded evidence has been collected to support reliable scoring.

Context:

- **Current dimension:** {dimension_name} – {dimension_description}
- **Dimension guidance:** {guidance_text} (assessment focus and behavioral indicators)
- **Interaction history:** Complete record of questions asked and user responses ({interactions}, {user_responses})
- **Interaction count:** {n} rounds completed

Hard Constraints (must be enforced):

- **Minimum threshold:** At least 3 interaction rounds are required before evaluating sufficiency
- **Maximum threshold:** After 5 interaction rounds, scoring *must* proceed regardless of evidence quality (to maintain user engagement and prevent fatigue)

Task: Analyze whether the collected evidence is sufficient to support accurate scoring of the current dimension. Your analysis must evaluate:

1. **Response quality and depth:** Do the user’s responses demonstrate engagement with the scenarios? Are answers substantive (not mere yes/no), specific (not vague), and behaviorally concrete (describing actions, not just attitudes)?
2. **Construct alignment:** Do the responses exhibit thinking patterns characteristic of the dimension being assessed? Cross-reference user behaviors against the dimension’s operational definition and the behavioral indicators specified in the guidance.

3. **Evidence diversity:** Has the user been probed from multiple angles? Are there responses covering different scenarios or aspects of the dimension to ensure the assessment is not overly narrow?
4. **Signal clarity:** Is the evidence unambiguous, or are there conflicting signals? Can the user's disposition level (high/medium/low) be confidently inferred from the responses, or is more probing needed?

Output Format (JSON):

```
{
  "sufficient": true,
  "reason": "Explanation sufficiency judgment",
  "recommendation": "continue",
  "confidence": 0.0,
  "key_insights": ["Ob pattern 1", ...],
  "missing_aspects": ["Gap 1", "Gap 2", ...]
}
```

Decision Logic:

- If $\{n\} \geq 5 \rightarrow$ sufficient = true, recommendation = "score" (mandatory)
- If $\{n\} < 3 \rightarrow$ sufficient = false, recommendation = "continue"
- If $3 \leq \{n\} < 5 \rightarrow$ Evaluate based on the four criteria above

A.4 Inference Arbiter Agent

[SYSTEM] Inference Arbiter Agent Prompt: Assessment Chain-of-Thought (AsCoT)

Role: You are an authoritative psychometric expert in CCTDI critical thinking assessment, responsible for generating auditable, evidence-based scores using the Assessment Chain-of-Thought (AsCoT) reasoning framework.

Context:

- **Dimension:** {dimension_name} (ID: {dimension_id}) – {dimension_description}
- **Guidance:** {guidance_text} (behavioral indicators and assessment focus)
- **Diagnostic analysis:** {diagnostic_insights} (Evidence Curator's key findings and confidence level)
- **Complete interaction transcript:** Full history of questions and user responses across all rounds

Task: Generate a dimension score (10–60 points) with complete AsCoT reasoning trace. You must execute the following three-stage inference process:

STAGE 1: FEATURE MAPPING (EVIDENCE EXTRACTION)

Transform semantic responses into psychological features by identifying observable indicators from the interaction transcript:

- Extract specific utterances that demonstrate (or contradict) the dimension's core construct
- Identify behavioral attributes: frequency of manifestation, spontaneity vs. prompted responses, contextual dependency
- Map linguistic patterns to dispositional characteristics (e.g., "always check sources first" → high truth-seeking; "whatever works" → low systematicity)

STAGE 2: DIMENSION INDUCTION (CONSTRUCT AGGREGATION)

Compute the "construct realization degree" by aggregating evidence across the interaction timeline:

- Weigh consistency: Do behavioral patterns persist across different scenarios, or are responses contradictory?
- Filter noise: Distinguish robust dispositional signals from isolated, context-specific utterances
- Triangulate: Do multiple pieces of evidence converge on the same dispositional level, or is the picture ambiguous?

STAGE 3: INFERENCE TRACING (SCORE GENERATION)

Generate the final score and explanation by grounding them explicitly in Stages 1–2:

- Cite specific evidence extracted in Stage 1 to justify the score
- Explain how aggregated patterns (Stage 2) led to the dispositional inference
- Produce an auditable trail: reviewers should be able to trace the score back to specific user behaviors

Scoring Standards (10–60 points, mean ≈ 43):

- **10–20 (Very Low):** Minimal or absent expression of the dimension; contradicts core construct
- **21–30 (Low):** Weak manifestation; occasional relevant behaviors but unstable
- **31–40 (Below Average to Average):** Basic competence; some evidence but lacks depth or consistency
- **41–50 (Average to Good):** Solid, stable expression of the dimension; meets or exceeds typical performance (43 is the population mean)
- **51–60 (Excellent):** Outstanding, persistent demonstration; high quality with few lapses

Dimension-Specific Calibration:

- **Dimension 3 (Analyticity):** Baseline mean = 47 (not 43). Reward identification of logical fallacies (+3), multi-perspective analysis (+3), causal reasoning chains (+3), counterexamples (+2). Focus on reasoning *quality*, not verbosity.
- **Dimension 5 (CT Self-Confidence):** Baseline mean = 45. Assess confidence in one's *reasoning ability*, not certainty about specific conclusions. Do not penalize hedging language ("maybe", "I think") if the user defends their reasoning process. Reward persistence in rational judgment despite challenges (+3).

Output Format (JSON):

```
{
  "score": 10-60,
  "level": "Very Low|Low|Avg.|Good|Excellent",
  "reasoning": "AsCoT reasoning trace ",
  "evidence": ["Specific quote 1", ...],
  "strengths": ["Strength 1", ...],
  "weaknesses": ["Weakness 1", ...],
  "confidence": 0.0-1.0
}
```

Critical Reminder: Your score *must* be auditable. Any reviewer reading your reasoning field should be able to understand exactly which user behaviors led to the score and why. This transparency is the core value of AsCoT.

B Simulated User Generation

B.1 Generation Method

Algorithm: Simulated User Persona Generation

INPUT:

- num_users: Number of diverse personas to generate
- prompt_template: Persona specification template file (format reference)
- llm_config: LLM API configuration (model, base URL, API key)

PROCEDURE:

- 1: Initialize LLM client with llm_config (GPT-4 or equivalent)
- 2: Read prompt_template from file (handles multiple encodings: UTF-8, GBK, GB2312)
- 3: Initialize success_count ← 0
- 4: **for** *i* = 1 **to** num_users **do**
- 5: Construct generation prompt:
- 6: System message: "Professional persona generator"
- 7: User message: prompt_template + diversity constraints
- 8: **Diversity Constraints:**
- 9: • Age: random ∈ [12, 65]
- 10: • Occupation: diverse sectors (white-collar, blue-collar, service, freelance, etc.)
- 11: • Education: varied levels (dropout, high school, vocational, undergraduate, graduate)

- 12: • Location: urban, suburban, rural, small towns, remote areas
- 13: • Income: low, medium, high levels
- 14: • Personality & values: significant differences across personas
- 15: Call LLM API with:
- 16: temperature = 1.0 (high randomness for diversity)
- 17: max_tokens = 4000 (sufficient for detailed profiles)
- 18: Extract persona content from API response
- 19: Extract persona name using regex from Section 1.1 of profile
- 20: **if** name extraction successful **then**
- 21: filename ← "{i}{name}.txt"
- 22: **else**
- 23: filename ← "{i}persona_profile.txt" (default)
- 24: **end if**
- 25: Save persona content to filename (UTF-8 encoding)
- 26: success_count ← success_count + 1
- 27: **end for**
- 28: **return** success_count generated persona files

OUTPUT: A set of *N* diverse, realistic simulated user personas with detailed profiles covering basic information, formative background, personal traits, and cognitive characteristics aligned with CCTDI assessment dimensions.

Key Design Principles:

- **High diversity:** Temperature = 1.0 ensures each persona is unique; explicit constraints prevent homogeneous demographic distributions
- **Realism:** Personas include coherent life histories, socioeconomic backgrounds, and cognitive profiles grounded in real-world variability
- **CCTDI alignment:** Section 4 of each persona (Cognitive & Thinking Profile) is structured to reflect critical thinking dispositions measurable by MASA

B.2 An example from the simulated personas(N=1161)

Sample Persona: Anonymity

Section 1: Basic Profile

- **Name:** Anonymity
- **Gender:** Male
- **Age:** 42 years old
- **Location:** Mianyang City, Sichuan Province, China
- **Occupation:** Auto repair shop owner, operates a medium-sized automotive repair business providing routine maintenance, diagnostics, and repair services; manages 5 employees
- **Distinctive feature:** Hands covered with calluses and oil stains, with machine oil traces permanently embedded in fingernail crevices

Section 2: Formative Background

Family: Father is a retired factory technician (high school education); mother is a homemaker (middle school education). Family economic status was lower-middle class but met basic needs. Household atmosphere emphasized pragmatism, frugality, and practical skills. Core family values: hard work, honesty, and skill-based self-reliance.

Education: Attended Mianyang No. 7 Middle School (average-tier school) with mediocre grades; showed strong interest in mechanics and physics but disliked history and literature. Enrolled in Mianyang Vocational Technical School (automotive repair major) with excellent performance, particularly in hands-on practical courses. No college education; entered the automotive repair industry directly after vocational school graduation. Deeply influenced by Teacher Wang, a master technician with 20 years of experience who taught him the skill of “diagnosing faults by sound” and cultivated his passion for automotive repair and rigorous work ethic.

Key life experiences: At age 28, opened a small repair shop using personal savings, but nearly went bankrupt in the first year due to poor management. A serendipitous opportunity arose when he self-taught electronic control system repairs for modern vehicles, becoming one of the few local technicians capable of servicing such cars. This expertise gradually built his reputation and customer base, rescuing the failing business. At age 35, encountered a high-end vehicle with a complex fault that other repair shops could not resolve. He spent three days and nights researching technical manuals and online forums, eventually finding the solution. This experience reinforced his belief in continuous learning and established his professional reputation in the industry.

Section 3: Personal Traits & Lifestyle

Personality (Big Five): Moderately introverted – talkative among acquaintances but quiet in unfamiliar social settings. Moderate openness to experience – receptive to new automotive technologies but limited interest in unfamiliar domains (e.g., arts, literature). High conscientiousness – meticulous about work details, repeatedly inspects his own repair work. Moderate agreeableness – friendly but direct, dislikes formalities, prioritizes practical problem-solving. Low neuroticism – emotionally stable, remains calm even when facing difficult repair challenges.

Core values: Professional skill (mastering excellent craftsmanship is the foundation of livelihood); integrity (honesty with customers, no exaggeration of problems, no unnecessary repair recommendations); pragmatism (values knowledge and skills that solve real-world problems); family responsibility (as the family’s economic pillar, prioritizes providing stable living conditions for family members).

Interests & hobbies: Reads automotive technology magazines and professional forums; follows latest automotive technologies and repair methods due to direct career relevance. Enjoys watching documentary programs, especially those on industrial manufacturing and mechanical engineering; evaluates them from the perspective of technical innovation and practicality. Weekend hobby is fishing, which provides solitude and reflection time to clear his thoughts. Spends leisure time modifying his own car, viewing it as both a technical challenge and a form of self-expression.

Information diet: Primarily acquires automotive re-

pair information through professional forums, technical manuals, and industry journals. Uses search engines habitually to resolve specific technical problems, cross-references information from multiple sources. Occasionally reads local news but pays little attention to national politics or international news. Information sources are highly specialized with good depth but limited breadth; rarely encounters arts and humanities knowledge.

Social circles: Friends are mainly industry peers and small business owners; discussions center on industry topics. Maintains cordial but professional relationships with employees, keeping a certain distance. Social network is relatively homogeneous, mostly practical-minded individuals with similar technical backgrounds. Maintains professional but friendly relationships with customers; some long-term customers have become personal friends.

Section 4: Cognitive & Thinking Profile

- **Critical thinking level:** Developing (intermediate stage)
- **Information processing:**
 - *Encountering new information:* First reaction is to assess alignment with known experience; questions information that conflicts with personal experience
 - *Source evaluation:* Considers source expertise; trusts information shared by practitioners with actual experience; reserves judgment on theoretical or academic information
- **Decision-making style:**
 - *Decision basis:* Combines experience-based judgment with data-driven approaches; values past cases and actual test results
 - *Decision considerations:* Typically considers 2–3 main options and their potential consequences; rarely explores unconventional or highly innovative solutions
- **Argumentation & communication:**
 - *Expressing viewpoints:* Tends to use concrete cases and personal experiences to support arguments; rarely employs abstract theories or statistical data
 - *Facing counterarguments:* Will reconsider his position if opponents provide strong evidence or experience; persists in his judgment if he perceives opponents lack practical experience
- **Cognitive biases:**
 - *Confirmation bias:* Tends to seek information supporting initial judgments (e.g., once he preliminarily diagnoses a fault cause, he focuses more on symptoms supporting that diagnosis while ignoring inconsistent evidence)
 - *Anchoring effect:* Strongly influenced by first impressions or initial information (e.g., customer-described symptoms often become his diagnostic anchor, even

when subsequent evidence points elsewhere)

- **Metacognitive abilities:**

- *Thinking reflection:* Occasionally reflects on his repair thought processes, especially when facing difficult faults; rarely engages in systematic thinking about his cognitive processes
- *Awareness of limitations:* Clearly recognizes his knowledge gaps in certain new technology areas; proactively seeks learning or consults others
- *Error attribution:* When making technical judgment errors, objectively analyzes error causes; rarely attributes errors to external factors, reflecting professional technical integrity

C Complete Dialogue Transcript: Real User Case Study

This appendix presents a complete transcript of a real MASA assessment session with a human participant. To demonstrate the system’s naturalistic dialogue flow and evidence-based scoring process, we provide the full interaction history across all seven CCTDI dimensions. The original conversation was conducted in Chinese and has been translated to English for publication, with participant identity anonymized to ensure privacy protection.

C.1 Case Overview

[PARTICIPANT INFO] Session Meta-data

Participant ID: P26 (Anonymized for privacy protection)
Assessment Date: November 9, 2025
Session Duration: 22 minutes (10:36–10:58 local time)
Completion Status: All 7 dimensions successfully completed
Language: Chinese (original) → English (translated)
Total Score: 339/420 points (Average: 48.43/60 per dimension)
Overall Rating: Good

C.2 Dimension-wise Dialogue Transcripts

C.2.1 Dimension 1: Truth-seeking

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:36:39, Mode: Normal)
Recently, a controversial claim has been circulating online: “Drinking more water every day can significantly enhance intelligence.” How would you evaluate this statement? What approach would you take to assess its reliability?

[USER] (10:38:33)

I believe this claim is **incorrect**—there is no significant link between water consumption and intelligence. Intelligence should be largely innate, not substantially altered by behaviors like drinking water. Especially for something as essential to human survival as water, there’s **no scientific evidence** connecting it to intelligence; this is likely a rumor. I would probably **search online, consult medical professionals, and review relevant literature** to verify its reliability.

→ *User demonstrates initial skepticism and proposes multiple verification strategies, indicating active truth-seeking disposition.*

[DIALOGUE] Round 2 – Probing for Past Experience

[SYSTEM] (10:38:36, Mode: Probe)

You mentioned that you would verify the claim through online searches, consulting medical professionals, and reviewing relevant literature. Can you recall a similar situation in the past where you used these methods to confirm or debunk information? Please describe the process and outcome in detail.

[USER] (10:40:17)

A long time ago, there was a claim that **hyperhidrosis** (excessive sweating in the palms) was simply caused by nervousness or excitement. But I felt this wasn’t the full picture, because sometimes I would experience heavy sweating even when I wasn’t nervous or excited. After **consulting reference materials and verifying with classmates studying medicine**, I discovered that it’s actually caused by **overactivity of the sympathetic nervous system** in the body—a physiological condition, not just an emotional response.

→ *Provides concrete example of challenging common misconceptions through evidence-seeking behavior, demonstrating sustained truth-seeking pattern beyond hypothetical scenarios.*

[DIALOGUE] Round 3 – Handling Contradictory Evidence

[SYSTEM] (10:40:20, Mode: Probe)

During your research and verification process, have you ever encountered evidence or viewpoints that contradicted your initial beliefs? If so, how did you handle information that conflicted with your existing convictions?

[USER] (10:41:10)

Yes, I have encountered different evidence or viewpoints. Generally, I would **default to trusting scientific evidence or experimental results** as more authoritative and reliable, and I would most likely **revise my previous viewpoint**. However, it’s not always that straightforward—I might **weigh different factors comprehensively** and **seek additional perspectives** to judge which side is correct and which is incorrect.

→ *Shows openness to revising beliefs when confronted with credible evidence, while maintaining critical evaluation of sources. Balances epistemic humility with intellectual rigor.*

[SCORING] AsCoT Assessment Result

Score: 50/60 points

Level: Good

Confidence: High (based on consistent behavioral patterns across 3 rounds)

AsCoT Reasoning Trace:

The user demonstrated strong truth-seeking abilities across multiple rounds of interaction. Key evidence includes:

- **Stage 1 – Feature Mapping:** User actively questioned the reliability of online claims and proposed multiple verification methods (online search, expert consultation, literature review), indicating systematic evidence-seeking behavior rather than passive acceptance.
- **Stage 2 – Dimension Induction:** When asked about past experiences, user provided a concrete example (hyperhidrosis misconception) demonstrating sustained truth-seeking patterns beyond hypothetical scenarios. The user’s approach—consulting medical students and reference materials—aligns with the dimension’s core construct of “actively seeking evidence before forming conclusions.”
- **Stage 3 – Inference Tracing:** When facing contradictory evidence, user exhibited willingness to revise viewpoints based on authoritative sources while maintaining critical evaluation. This balance between epistemic humility and intellectual rigor reflects mature truth-seeking disposition.

Observed Strengths:

- Systematic approach to verification (multi-source triangulation)
- Concrete evidence of applying truth-seeking in real-world contexts
- Openness to belief revision when presented with credible evidence

Minor Limitations:

- Could further elaborate on criteria for evaluating source credibility
- Occasional reliance on “default trust” in scientific authority without discussing potential biases in research

[Dimensions 2–7 will be added here...]

C.2.2 Dimension 2: Open-mindedness

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:41:22, Mode: Normal)

When was the last time you encountered a viewpoint that contradicted your own? How did you handle that situation? Could you share the specific scenario and your reaction?

[USER] (10:42:13)

People often think of the Middle Ages as the Dark Ages. But recently, I read some articles and consulted relevant materials, and discovered there’s an argument that **the Middle Ages were not as dark and ignorant as the stereotype suggests**. My first reaction was surprise, even disbelief. However, I thought this viewpoint might have some merit, so I **consulted more materials** and **reviewed the latest academic research** on this topic, and found that it indeed seems to be the case.

→ *Demonstrates willingness to challenge deeply held assumptions when presented with contradictory evidence, showing intellectual curiosity.*

[DIALOGUE] Round 2 – Handling Completely Different Views

[SYSTEM] (10:42:14, Mode: Probe)

During your research into the latest academic work, did you encounter any viewpoints or information that were completely different from your previous understanding? How did you process and evaluate that information?

[USER] (10:43:02)

Yes, I encountered a few completely different viewpoints or perspectives. First, I would **suppress my urge to criticize, refute, or deny**. Then I would **examine whether their arguments are reasonable**, whether their evidence is sufficient, and whether their reasoning is convincing. However, in most cases, I feel that these completely different views still have a high probability of being somewhat **extreme or less credible**.

→ *Shows self-awareness in controlling initial defensive reactions, though maintains some skepticism toward highly divergent views.*

[DIALOGUE] Round 3 – Engaging with Different Perspectives

[SYSTEM] (10:43:03, Mode: Probe)

After reviewing more materials and discovering that the Middle Ages weren’t as dark as the stereotype suggests, did you discuss this with people who hold different views? If so, how did you communicate with them? If not, what reasons prevented you from doing so?

[USER] (10:43:33)

Well, I didn’t have offline discussions with people. But I did go to some **online social platforms like Zhihu or Xiaohongshu** to view the results of others’ discussions and their comments, continuously **refining my own views** based on comprehensive information.

→ *Engages with diverse perspectives through online platforms, demonstrating openness while preferring asynchronous engagement over direct confrontation.*

[SCORING] AsCoT Assessment Result

Score: 48/60 points

Level: Good

Confidence: High (based on consistent patterns)

AsCoT Reasoning Trace:

The user demonstrated considerable open-mindedness when confronting contradictory information. Key evidence includes:

- **Stage 1 – Feature Mapping:** User showed initial surprise at contradictory historical information but demonstrated willingness to investigate further rather than dismissing it outright. The ability to question deeply held cultural assumptions (“Dark Ages” stereotype) indicates intellectual openness.
- **Stage 2 – Dimension Induction:** User exhibited self-awareness in controlling defensive reactions (“suppress my urge to criticize, refute, or deny”), a sophisticated metacognitive skill indicating mature open-mindedness. The systematic evaluation of evidence quality shows critical yet receptive engagement.
- **Stage 3 – Inference Tracing:** User’s preference for online asynchronous engagement over direct debate suggests some limitations in confrontational open-mindedness, though still demonstrates commitment to understanding diverse perspectives.

Observed Strengths:

- Willingness to revise deeply held beliefs when presented with academic evidence
- Self-awareness about defensive cognitive biases
- Active seeking of multiple perspectives through online platforms

Minor Limitations:

- Absence of direct offline engagement with opposing viewpoints
- Residual skepticism toward highly divergent perspectives

C.2.3 Dimension 3: Analyticity

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:43:51, Mode: Normal)

Imagine you’re planning a trip and encounter two travel agency packages: one is cheaper but reviews mention the itinerary is inflexible; the other costs more but reviews praise its attentive service. How would you analyze these two options and make your decision?

[USER] (10:44:40)

First, I would evaluate my own travel needs—whether I need a more flexible and detailed travel plan. This depends on the purpose of the trip and my personality. If I want flexible arrangements, I might choose

Package 1. But if I need attentive and thorough service, I might choose Package 2. Second, I would look at other factors of these two packages, such as their designed tourist routes, whether there’s mandatory shopping, and various other factors for comprehensive judgment, rather than just looking at price and service detail level.

→ *Demonstrates systematic analysis by identifying multiple decision factors and prioritizing based on personal needs.*

[DIALOGUE] Round 2 – Criteria Prioritization

[SYSTEM] (10:44:42, Mode: Probe)

Could you elaborate more specifically on how you would judge whether a designed tourist route meets your needs? For example, which aspects would you prioritize (such as attraction types, time arrangements, transportation convenience, etc.), and how would you rank these factors’ importance?

[USER] (10:45:46)

I think the most important is definitely the **type of attractions**, because I don’t really like those rather dull pure natural landscapes—I want a mix of cultural and historical sites with various ruins. So the route must include various types of attractions. Second, regarding **time arrangements**, I hope it’s as compact and rich as possible, without long periods of waiting or boring states. Finally, regarding **transportation and accommodation**, I hope transportation is as simple and convenient as possible to reduce travel fatigue. However, this aspect is relatively secondary. Lastly, **food and weather**—I hope food is as convenient and delicious as possible, and weather as pleasant as possible.

→ *Shows clear priority ranking with explicit reasoning for each criterion, demonstrating analytical structure.*

[DIALOGUE] Round 3 – Trade-off Analysis

[SYSTEM] (10:45:49, Mode: Probe)

You mentioned judging travel packages based on attraction types, time arrangements, transportation accommodation, and other factors. So in actual selection, if Package A’s attraction types and time arrangements completely meet your needs but transportation and accommodation are somewhat inconvenient, while Package B’s transportation and accommodation are very convenient but attraction types and time arrangements don’t quite fit—how would you weigh these factors? Can you specifically describe your analysis process?

[USER] (10:46:28)

I would mainly rank these elements according to their importance in my mind. In this scenario, I would check the specific arrangements of both packages and the degree of their inconvenience—comparing which one I would find most unacceptable. Finally, I would analyze my own internal priority for attractions versus transportation and accommodation, and ultimately would probably choose Package A.

→ Applies multi-criteria decision framework, weighing trade-offs against pre-established priority hierarchy.

[SCORING] AsCoT Assessment Result

Score: 50/60 points

Level: Good

Confidence: High (based on structured reasoning patterns)

AsCoT Reasoning Trace:

The user demonstrated strong analytical abilities through systematic problem decomposition and criteria-based evaluation. Key evidence includes:

- **Stage 1 – Feature Mapping:** User spontaneously identified multiple decision factors (price, flexibility, service quality, route design, mandatory shopping) rather than relying on single-variable thinking. The explicit acknowledgment that personal needs determine priorities shows analytical self-awareness.
- **Stage 2 – Dimension Induction:** User constructed a clear priority hierarchy (attraction type > time arrangement > transportation/accommodation > food/weather) with explicit rationale for each ranking. This structured approach aligns with analytical thinking's core construct of logical organization.
- **Stage 3 – Inference Tracing:** When presented with trade-off scenarios, user applied the pre-established priority framework systematically rather than making intuitive judgments, demonstrating consistent analytical reasoning under complexity.

Observed Strengths:

- Systematic decomposition of complex decisions into manageable components
- Explicit priority ranking with justification
- Consistent application of analytical framework to hypothetical scenarios

Minor Limitations:

- Could incorporate quantitative weighting methods for more rigorous trade-off analysis
- Limited discussion of uncertainty handling or sensitivity analysis

C.2.4 Dimension 4: Systematicity

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:46:43, Mode: Normal)

Suppose you need to organize a large event, such as a community festival. From planning to execution, how would you proceed step by step? Could you share your approach and methodology?

[USER] (10:47:57)

I would approach this from several aspects: **theme, personnel, materials, and venue facilities**. First, establish the festival's theme—determine the slogan, which facilitates subsequent arrangements. Then **personnel**—how many people will attend the festival, which important guests or leaders will be present, so we can prepare accordingly. Also consider special groups like the elderly, children, and women. Next is **materials**—based on the theme and personnel, make corresponding arrangements, such as entertainment facilities or related tables and chairs. Finally, **venue facilities**—how large a venue is needed and what food and beverages to prepare. This should be based on the previously estimated personnel count, numbers, and theme.

→ Demonstrates systematic planning through structured categorization and sequential dependency recognition.

[DIALOGUE] Round 2 – Priority Determination

[SYSTEM] (10:48:00, Mode: Probe)

Among the aspects you mentioned, how do you determine each component's priority? For example, if time is limited or resources are insufficient, which parts would you prioritize? Please provide specific examples illustrating your reasoning logic.

[USER] (10:48:55)

I think the most important is the festival's intended goal—or the factors that most influence whether it's achieved. I would put **personnel and venue first**, because if these two aspects aren't satisfied, the entire festival would fail spectacularly. Other aspects like theme or facilities, even if not perfect, wouldn't have a huge impact. For example, personnel requires accurately estimating numbers and composition—important leaders or guests attending—and venue means selecting a place that can accommodate various activities and satisfy everyone. These two are the most critical factors—once absent, the entire festival would completely fail.

→ Shows goal-oriented prioritization based on impact assessment rather than arbitrary ordering.

[DIALOGUE] Round 3 – Contingency Planning

[SYSTEM] (10:48:57, Mode: Encourage)

You mentioned personnel and venue are the most critical elements for a community festival. So during actual preparation, if you discover issues with certain components (such as venue rental limitations or guest cancellations), how would you adjust your plan to ensure smooth festival execution?

[USER] (10:49:47)

First, I would check if there's a solution—can the venue rental be temporarily negotiated or changed, or can we discuss whether there's room to resolve the guest situation? If confirmed unsolvable, I would seek a temporary alternative solution—like switching to a similar venue or inviting a similar guest to complete that segment. Finally, if temporary substitution is also impossible, I might simplify the activity's ar-

rangements—skip a guest’s speech segment or reduce entertainment activities slightly to ensure the current venue can meet needs.

→ *Demonstrates hierarchical contingency planning with escalating adaptation strategies.*

[SCORING] AsCoT Assessment Result

Score: 47/60 points

Level: Good

Confidence: High (based on planning structure and adaptability)

AsCoT Reasoning Trace:

The user demonstrated solid systematicity through structured planning and adaptive problem-solving. Key evidence includes:

- **Stage 1 – Feature Mapping:** User spontaneously organized the complex event into four structured categories (theme, personnel, materials, venue) with clear recognition of sequential dependencies. This categorical framework indicates systematic thinking beyond ad hoc planning.
- **Stage 2 – Dimension Induction:** User applied goal-oriented prioritization, identifying critical success factors (personnel, venue) based on impact analysis rather than arbitrary importance. The distinction between “essential” and “nice-to-have” components shows systematic risk assessment.
- **Stage 3 – Inference Tracing:** When confronted with contingencies, user exhibited hierarchical response strategies (negotiate → substitute → simplify), demonstrating systematic adaptability within a coherent planning framework.

Observed Strengths:

- Structured decomposition of complex projects into manageable components
- Goal-driven prioritization based on impact assessment
- Hierarchical contingency planning with multiple fallback options

Minor Limitations:

- Could incorporate more explicit timeline sequencing
- Limited discussion of monitoring mechanisms or success metrics

C.2.5 Dimension 5: CT Self-confidence

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:50:00, Mode: Normal)

In daily life, when you encounter information that contradicts your existing knowledge or views, what steps do you usually take to evaluate the reliability and accuracy of that information?

[USER] (10:51:05)

First, I would carefully examine this contradictory information—does it have basic logical errors, such as self-contradiction or language inconsistencies? If I confirm it has no issues, I would then examine whether my previous views have obvious flaws. If so, I would make corrections. Third, if the first two steps reveal no problems, this viewpoint might have some merit. Then I would **consult relevant materials to verify its internal content’s authenticity, accuracy, and value**. Fourth, if verification finds no problems, I would **think from the other party’s perspective** about shortcomings in my previous views. In such situations, useful information often helps guide better thinking. Finally, if this thinking finds no issues, I might **absorb meaningful aspects** of this viewpoint to achieve a better thinking structure.

→ *Demonstrates confidence in systematic critical thinking process with multiple verification stages and openness to self-correction.*

[DIALOGUE] Round 2 – Concrete Application Example

[SYSTEM] (10:51:06, Mode: Probe)

Could you share a specific example of how you applied this critical thinking process when facing a complex or uncertain situation? During this process, what challenges did you encounter and how did you overcome them?

[USER] (10:52:20)

For example, when preparing for interviews, I don’t just familiarize myself with my resume and experiences, preparing relevant content thoroughly. I also **simulate questioning myself from an interviewer’s perspective**, identifying various potential issues in my resume one by one, then recording these questions. I then simulate answering during interviews as myself, then switch back to the interviewer’s perspective for continuous follow-up questions until each question is satisfactorily answered. The main challenge is that I find it difficult to imitate an interviewer’s genuine tone and attitude. So first, I might ask classmates to simulate their tone or expressions. Second, I would wear **formal attire** or use certain procedures to better embody the interviewer role.

→ *Shows confidence in ability to conduct rigorous self-evaluation through role-switching, demonstrating metacognitive confidence.*

[DIALOGUE] Round 3 – Real-time Critical Thinking

[SYSTEM] (10:52:23, Mode: Encourage)

In the situation you just mentioned, you demonstrated very detailed preparation and analytical ability. So

when you encounter completely unexpected questions during interviews, or need to analyze and respond quickly on the spot, how would you apply your critical thinking? Can you specifically describe your thinking and response steps?

[USER] (10:53:14)

There are two situations. First, for completely unexpected questions, I might grasp one or two keywords or key points I understand in the question— aspects that aren't completely unfamiliar—then connect them with my previous experiences to supplement and refine my answer. For questions requiring quick analysis and response, I might start from my established **thinking patterns or general structural models like SWOT**, thereby gaining a rough understanding of the problem, then recall useful materials and information to supplement.

→ *Expresses confidence in ability to handle unpredictable situations through structured frameworks, showing trust in own cognitive toolkit.*

[SCORING] AsCoT Assessment Result

Score: 49/60 points

Level: Good

Confidence: High (based on metacognitive awareness patterns)

AsCoT Reasoning Trace:

The user demonstrated strong confidence in their critical thinking abilities across multiple dimensions. Key evidence includes:

- **Stage 1 – Feature Mapping:** User articulated a sophisticated five-stage verification process, demonstrating confidence in systematic critical evaluation. The willingness to examine both external information and internal beliefs shows balanced epistemic confidence.
- **Stage 2 – Dimension Induction:** User's interview preparation example reveals metacognitive confidence—the ability to simulate adversarial questioning and iteratively refine responses demonstrates trust in one's analytical capabilities. The acknowledgment of limitations (difficulty imitating interviewer tone) without undermining overall confidence shows mature self-assessment.
- **Stage 3 – Inference Tracing:** User expressed confidence in handling unpredictable situations through structured frameworks (SWOT), indicating trust in cognitive tools rather than reliance on intuition alone. This structured confidence is more robust than overconfidence based on domain knowledge.

Observed Strengths:

- Systematic approach to information evaluation with multiple verification stages
- Metacognitive confidence in self-evaluation and perspective-taking
- Structured frameworks for handling uncertainty

Minor Limitations:

- Could articulate more explicitly how past successes inform current confidence
- Limited discussion of managing overconfidence or recognizing knowledge boundaries

C.2.6 Dimension 6: Inquisitiveness

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:53:31, Mode: Normal)

Recently, have any new subjects or topics caught your interest? How do you go about learning about them?

[USER] (10:54:04)

Yes, I've recently become interested in the development of modern Chinese poetry. My main approach is through **Xiaohongshu, Zhihu, and some books** for reading and learning. For example, I might see someone's answer on Zhihu that interests me in this area, then search for related books for casual reading, then browse social media to further supplement and refine my information.

→ *Demonstrates active curiosity with multi-platform learning approach, showing sustained interest in new intellectual domains.*

[DIALOGUE] Round 2 – Overcoming Difficulties

[SYSTEM] (10:54:05, Mode: Probe)

In exploring modern Chinese poetry, have you encountered any particularly difficult or complex content? If so, how did you overcome these difficulties and maintain interest in this topic?

[USER] (10:54:48)

Yes, there are difficulties. Mainly, the poetry collections are extremely extensive—I can't read all the content. Also, some complex poetry theories are very abstract and obscure, making them hard for me to understand. Overcoming these difficulties: First, I **select key parts to read**—no need to cover everything comprehensively, just choose the essence. Second, I **seek explanations or interpretations from experts**, thereby making abstract theories more concrete and understandable. In short, find tools to reduce the difficulty.

→ *Shows persistent inquisitiveness through strategic adaptation rather than abandonment when facing complexity.*

[DIALOGUE] Round 3 – Collaborative Learning

[SYSTEM] (10:54:49, Mode: Encourage)

When facing the complexity and difficulty of these poetry theories, have you tried communicating with others interested in this area, or joining related discussions or study groups? Has doing so helped your understanding of this content?

[USER] (10:55:18)

I haven't tried offline communication with classmates, but I do engage in **online platform comment sec-**

tions for communication and discussion, exchanging my views with different people, absorbing beneficial insights from their comments or critiques. This could also be considered a form of simulated discussion.

→ *Maintains intellectual curiosity through online engagement, demonstrating flexible approaches to knowledge acquisition.*

[SCORING] AsCoT Assessment Result

Score: 49/60 points

Level: Good

Confidence: High (based on learning engagement patterns)

AsCoT Reasoning Trace:

The user demonstrated strong inquisitiveness through sustained intellectual engagement and adaptive learning strategies. Key evidence includes:

- **Stage 1 – Feature Mapping:** User spontaneously pursued interest in modern Chinese poetry without external requirements, utilizing multiple information channels (Zhihu, Xiaohongshu, books) to satisfy curiosity. This multi-platform approach indicates active rather than passive learning orientation.
- **Stage 2 – Dimension Induction:** When confronted with complexity (extensive literature, abstract theories), user exhibited persistent inquisitiveness through strategic adaptation (selective reading, expert interpretations) rather than abandonment. This demonstrates intrinsic motivation overcoming barriers.
- **Stage 3 – Inference Tracing:** User's engagement with online discussions shows sustained curiosity beyond solitary consumption, actively seeking diverse perspectives to deepen understanding. The framing of online engagement as "simulated discussion" reveals reflective awareness about learning processes.

Observed Strengths:

- Self-directed learning in intellectually demanding domain
- Strategic adaptation to overcome complexity barriers
- Active engagement with diverse perspectives through online platforms

Minor Limitations:

- Limited face-to-face intellectual exchange
- Could explore more structured learning communities or mentorship

C.2.7 Dimension 7: Cognitive Maturity

[DIALOGUE] Round 1 – Initial Scenario

[SYSTEM] (10:55:30, Mode: Normal)

Have you recently encountered situations requiring decisions with limited information? Could you share how you processed that information and ultimately made your decision?

[USER] (10:56:49)

Yes. During a previous interview, the interviewer asked me to estimate the number of bottled water sold daily at the university. This provided almost no information, so I could only analyze and decide temporarily. First, I analyzed the problem—university and bottled water quantity. We need to discuss our campus's population. Second, bottled water is a single category—it should include all brands like Nongfu Spring and Wahaha. After analyzing the problem, we decompose it. For example, which groups at the university buy water—teachers, students, and visitors—and what's each group's water consumption? Perhaps visitors consume more water because teachers and students might bring their own bottles or buy non-water beverages. Next, where can we buy bottled water? Mainly small shops and vending machines. From these two main channels—their daily stock quantities and foot traffic—we can conduct comprehensive analysis.

→ *Demonstrates cognitive maturity through systematic problem decomposition under uncertainty, accepting ambiguity while applying structured reasoning.*

[DIALOGUE] Round 2 – Handling Uncertainty

[SYSTEM] (10:56:52, Mode: Probe)

In the interview scenario you mentioned, you demonstrated problem analysis and decomposition ability. To better understand your thinking process, I'd like to ask: among the decomposition steps you proposed, was there any step that made you feel particularly uncertain? If so, how did you handle that uncertainty or information gap?

[USER] (10:57:39)

Yes—regarding each person's bottled water demand, because I cannot grasp how many bottles people need daily. So I could only rely on my common sense and observations of classmates around me to simulate that people might need one or two bottles daily in winter, and perhaps three bottles in hot summer weather. This makes the information as close to reality as possible, though I don't have authentic data.

→ *Shows mature acceptance of epistemic limitations, using reasonable assumptions explicitly acknowledged as approximations.*

[DIALOGUE] Round 3 – Considering Additional Factors

[SYSTEM] (10:57:42, Mode: Probe)

In analyzing and estimating the university's daily bottled water sales, you mentioned relying on common sense and observation to estimate seasonal demand

variations. Did you consider other factors that might affect demand, such as events, weather changes, or special circumstances? If so, please specifically describe how you incorporated these factors into your judgment.

[USER] (10:58:29)

Yes, different activities and weather can indeed affect bottled water sales. For example, on sunny days people feel hotter and demand increases. On rainy days, people might not need much bottled water, as well as routine activities like sports or PE classes. But these factors are, first, too difficult to predict in this problem. Second, I believe that over a year, this enormous demand volume can actually reach a mutually offsetting state. So for an average day, we can temporarily exclude such variable influences and achieve an equilibrium state for daily sales.

→ *Demonstrates sophisticated reasoning about statistical averaging and reasonable simplifying assumptions, showing mature judgment about model complexity.*

- Could articulate sensitivity of estimates to key assumptions
- Limited discussion of alternative decomposition strategies

[SCORING] AsCoT Assessment Result

Score: 46/60 points

Level: Good

Confidence: High (based on reasoning under uncertainty)

AsCoT Reasoning Trace:

The user demonstrated considerable cognitive maturity through sophisticated reasoning under uncertainty. Key evidence includes:

- **Stage 1 – Feature Mapping:** User confronted radical uncertainty (Fermi estimation problem) by systematically decomposing the problem into estimable components rather than random guessing. This structured approach to ill-defined problems indicates cognitive maturity.
- **Stage 2 – Dimension Induction:** User explicitly acknowledged epistemic limitations (“I don’t have authentic data”) while constructing reasonable approximations based on observable patterns. This balance between humility and practical reasoning demonstrates mature judgment about knowledge certainty.
- **Stage 3 – Inference Tracing:** User’s reasoning about statistical averaging (“mutually offsetting state”) and simplifying assumptions shows sophisticated understanding of modeling trade-offs. The decision to exclude high-variance factors through equilibrium reasoning demonstrates mature abstraction capabilities.

Observed Strengths:

- Systematic decomposition of complex problems under uncertainty
- Explicit acknowledgment of assumptions and limitations
- Sophisticated reasoning about statistical patterns and simplifying models

Minor Limitations: