AI as a Metacognitive Mirror: Leveraging Artificial Intelligence to Scaffold Human Metacognitive Reflection

Leveraging AI as a Metacognitive Mirror in the Glocal Leader Academy Glocal Hero Program: A Data-Driven Study on Reflection and Leadership Learning

ABSTRACT

As generative AI permeates education, a key question is whether systems that lack metacognition can still help humans develop it. We tested a structured "AI Coach" for reflective writing in an authentic peer-learning program (N=163). The Spring cohort used a between-subjects comparison of AI-scaffolded vs. manual reflection; the Autumn cohort used a crossover with matched pre–post measures. The AI Coach applied Socratic prompts aligned to a 3C framework and offered optional voice interaction. Reflection quality (seven dimensions) was scored by an AI evaluator validated against human ratings (κ =0.93); metacognitive awareness was measured with the MAI. AI scaffolding was associated with longer reflections and higher emotional awareness. In the AI→Manual sequence, quality was maintained or improved, suggesting short-term transfer rather than tool dependence. MAI showed minimal change (p=0.22), indicating process-level effects more than trait shifts. Perceived impact on critical thinking—not usage frequency—predicted gains. These results point to a "metacognitive symbiosis," where AI structures the process while humans supply meaning and agency.

Keywords: *Metacognition, Artificial Intelligence, Reflection, Educational Technology, Cognitive Scaffolding, Human-AI Collaboration*

I. INTRODUCTION

The rapid proliferation of large language models (LLMs) has fundamentally transformed how humans engage with cognitive tasks. In one of the survey have conducted in this research, 92% of participants reported using at least one generative AI tool for writing or reflection. This aligns with broader trends: recent U.S. surveys show 26% of teens used ChatGPT for schoolwork in 2024 (up from 13% in 2023), with many teens having tried some form of generative AI tool [7][34][35]. While this ubiquity signals unprecedented access to computational intelligence, it simultaneously raises existential concerns about cognitive offloading and metacognitive atrophy [1][2]. Recent neuroscientific evidence demonstrates that LLM use during essay writing significantly reduces neural connectivity and impairs memory recall, leading to "cognitive debt" where users show weaker brain network engagement and reduced learning outcomes compared to unaided writing [3].

The paradox is striking: As AI systems demonstrate superhuman performance across intelligence benchmarks, they fundamentally lack metacognitive awareness—the capacity to reflect on their own thinking, monitor comprehension, and self-regulate learning strategies [8]. Metacognition, defined as "thinking about thinking," encompasses two core dimensions: metacognitive knowledge (awareness of one's cognitive processes) and metacognitive regulation (active monitoring and control of cognition) [9]. While LLMs can generate sophisticated text mimicking reflective discourse, they possess no genuine self-awareness or capacity for experiential learning from mistakes [4][5].

This metacognitive deficit is not merely a technical limitation but reveals a fundamental asymmetry: AI excels at intelligence; humans excel at metacognition. This asymmetry suggests a potential symbiotic relationship where AI's computational strengths could scaffold human metacognitive development, while humans provide the reflective awareness AI cannot generate [30]. If realized, such symbiosis could offer a pathway for sustained human relevance in an era artificial intelligence, positioning metacognition as humanity's enduring competitive advantage.

Recent work in AI-supported learning has explored adaptive tutoring systems [21], automated feedback [27], and conversational agents [26], but little research has explicitly targeted metacognitive development as the primary outcome, particularly in authentic, non-academic contexts. This gap is critical: if AI can scaffold metacognition—the very capacity AI lacks—it would validate a symbiotic model transcending replacement narratives.

This study tests whether a Socratic AI Coach can improve reflection quality and metacognitive awareness in an authentic leadership context, and whether improvements persist after AI removal—providing evidence for metacognitive symbiosis.

II. RELATED WORK

A. Metacognition and Reflection in Learning

Metacognition—"thinking about thinking"—has been recognized as a fundamental component of effective learning since Flavell's (1979) seminal work distinguishing metacognitive knowledge (awareness of cognitive processes) from metacognitive regulation (active monitoring and control) [8]. This dual-process framework underpins decades of research demonstrating that metacognitive skills predict academic achievement across domains [15, 32], with effect sizes often exceeding those of intelligence or prior knowledge alone.

The Metacognitive Awareness Inventory (MAI), developed by Schraw and Dennison (1994) [10], operationalizes Flavell's framework through 52 items (later shortened to 15-item versions) assessing planning, monitoring, and evaluation strategies. While widely used in educational research, the MAI measures self-reported trait-level awareness rather than situation-specific metacognitive processes, a distinction our study addresses by examining both MAI scores and reflection quality as separate constructs.

Reflection, as conceptualized by Schön (1983) [11], extends metacognition into professional contexts, distinguishing "reflection-in-action" (thinking during experience) from "reflection-on-action" (retrospective analysis). Gibbs' (1988) Reflective Cycle [12] provides a structured model encompassing description, feelings, evaluation, analysis, conclusion, and action planning—a framework that influenced our AI Coach's 3C structure (Context-Challenge-Change). Empirical work on reflection quality has employed hierarchical frameworks such as the SOLO taxonomy [13], which categorizes cognitive depth from surface description to extended abstract thinking, informing our 5-level depth scale.

However, despite extensive evidence that metacognition and reflection are teachable, interventions typically rely on human instruction, peer feedback, or written prompts—modalities that are resource-intensive and difficult to scale. This gap motivates our investigation of AI as a scalable metacognitive scaffold.

B. AI in Education: From Tutoring to Scaffolding

Artificial intelligence has a decades-long history in education, beginning with rule-based Intelligent Tutoring Systems (ITS) that model student knowledge states and adapt instruction accordingly [21]. Systems like Carnegie Learning's Cognitive Tutor demonstrate that well-designed ITS can match human tutoring effectiveness in domains like mathematics. However, traditional ITS focus on declarative knowledge (facts, procedures) rather than metacognitive processes, and their brittleness limits transferability across domains.

Recent advances in natural language processing have enabled conversational agents for learning, such as Graesser et al.'s (2005) AutoTutor, which employs dialogue-based Socratic questioning to elicit explanations in physics and computer literacy. AutoTutor's success demonstrates that AI-driven questioning can promote deeper reasoning, though its reliance on pre-scripted dialogue trees limits flexibility. Similarly, automated writing evaluation systems provide formative feedback on essay quality but rarely target metacognitive awareness explicitly.

The emergence of large language models (LLMs) like GPT-4 has transformed this landscape. Unlike earlier systems, LLMs exhibit emergent abilities in zero-shot reasoning, multilingual fluency, and adaptive dialogue [22, 23], enabling more naturalistic educational interactions. Recent studies explore LLMs for tutoring [24], code debugging, and formative feedback. However, concerns about cognitive offloading persist: Sparrow et al. (2011) [1] documented "Google effects on memory," where easy access to information reduces retention, and Risko and Gilbert (2016) [2] argue that overreliance on external cognitive aids atrophies internal capabilities.

Our work diverges from this trajectory by positioning AI not as an answer provider but as a metacognitive scaffold—a thinking partner that structures reflection without replacing it. This aligns with Vygotskian scaffolding theory [16, 17], where temporary support enables learners to perform beyond independent capacity, then fades as competence develops. Critically, scaffolding requires intentional design for gradual release, distinguishing our AI Coach from open-ended chatbot access.

C. Generative AI and Educational Reflection

The rapid adoption of generative AI in education has sparked both enthusiasm and alarm. UNESCO's 2023 report warns that uncritical AI use risks "outsourcing thinking," particularly when students use ChatGPT to generate essays wholesale. Recent surveys show widespread student use of GenAI. For example, 26% of U.S. teens used ChatGPT for schoolwork in 2024 (up from 13% in 2023), and a 2024 Harvard survey found ~90% of undergraduates use GenAI tools [7, 38]. This ubiquity necessitates pedagogical responses beyond prohibition—specifically, interventions that harness AI's capabilities while preserving cognitive engagement.

Emerging research explores pedagogically-designed AI interactions. Emerging work shows that metacognitive prompts can shape reasoning and critical-thinking behaviors in GenAI contexts—e.g., structured "metacognitive prompting" improves LLM understanding, and prompts that ask students to pause, assess evidence, and consider alternatives lead to deeper inquiry during GenAI-based search. Mollick and Mollick (2023) propose using ChatGPT as a "Socratic co-pilot" for reflective writing, though their framework lacks empirical validation. In introductory programming, LLM-generated worked examples have shown promise at scale, and prompted self-explanation remains an effective strategy to deepen code comprehension.

However, three critical gaps persist in this nascent literature:

1. Metacognition as Outcome: Most studies measure content learning or task performance, not metacognitive

development itself. Whether AI scaffolding improves metacognitive awareness remains underexplored.

2. Sustainability: Few studies test whether AI-scaffolded skills transfer beyond AI support (i.e., the "tool dependency" question). Without evidence of internalization, AI risks functioning as a cognitive prosthetic rather than a developmental scaffold.
3. Authentic Contexts: Much research occurs in controlled laboratory settings with artificial tasks. Ecological validity—whether findings generalize to real-world educational contexts—remains uncertain.

Our study addresses these gaps by (a) measuring both reflection quality (process) and MAI (trait-level awareness), (b) employing a crossover design that tests skill maintenance after AI removal, and (c) situating the intervention within an authentic youth leadership program where reflection carries personal significance.

E. Human-AI Collaboration and Cognitive Symbiosis

The philosophical framing of our work draws on distributed cognition theory [30, 29], which posits that cognition extends beyond individual brains to incorporate tools, symbols, and social networks. Clark and Chalmers' (1998) "extended mind" thesis [29] argues that external resources (e.g., notebooks, calculators, smartphones) can constitute genuine parts of cognitive processes when reliably coupled. Applied to AI, this suggests that AI tools could augment human cognition rather than merely assist it.

Recent work on human-AI collaboration emphasizes complementarity rather than replacement. Humans excel at intuition, creativity, and ethical judgment; AI excels at pattern recognition, information retrieval, and computational speed. Optimal collaboration leverages both strengths. Evidence from radiology and diagnostic tasks indicates human-AI teams can outperform either alone, when assistance is designed to complement human judgment.

However, existing collaboration frameworks focus on task completion (solving problems, making decisions) rather than cognitive development (improving human capacities). This distinction is critical: A diagnostic AI assistant helps doctors diagnose better while using the AI; our question is whether an AI reflection coach helps humans reflect better without the AI—i.e., whether temporary scaffolding produces durable metacognitive gains.

We introduce the term metacognitive symbiosis to describe this relationship: AI lacks metacognitive awareness but possesses intelligence; humans possess metacognitive awareness but face reflection limitations (cognitive biases, emotional avoidance, limited perspective-taking). By scaffolding reflection through Socratic questioning, AI's intelligence compensates for human metacognitive weaknesses, while humans retain metacognitive agency (owning their reflections, making meaning). This symbiosis is intentionally asymmetric and temporary—AI guides the process but humans drive the content, and the scaffold ultimately fades.

This framing positions our contribution at the intersection of educational AI, metacognition research, and human-AI interaction. Unlike prior work emphasizing either AI risks (cognitive offloading) or benefits (performance gains), we test whether AI can improve the very capacity AI itself lacks, provided the interaction is pedagogically structured for gradual release.

F. Positioning This Work

Our study advances existing scholarship in four ways:

1. Theoretical: Proposes metacognitive symbiosis as a framework for human-AI collaboration in learning, extending distributed cognition theory to developmental outcomes.
2. Empirical: Provides evidence that AI-scaffolded reflection improves quality (RQ1), skills transfer beyond AI support (RQ4), yet does not elevate trait-level MAI (RQ3)—a dissociation clarifying process vs. trait effects.
3. Methodological: Employs a crossover design enabling within-subjects causal inference and sustainability testing, addressing prior research's reliance on between-subjects comparisons.
4. Practical: Offers a concrete, replicable intervention (Gibson 3C + Socratic prompting + voice synthesis) with open-access implementation details, enabling practitioners to adapt the approach.

Critically, our work addresses the existential question motivating much AI-in-education discourse: As AI systems approach or exceed human intelligence across tasks, how do humans remain relevant? Our answer centers metacognition as humanity's enduring differentiator—the capacity AI fundamentally lacks—and demonstrates that AI can, paradoxically, help humans develop precisely what makes them uniquely human.

III. METHODS

A. VolTra and Authentic Peer Learning

This study was conducted within VolTra, a Hong Kong-based NGO (founded 2009) dedicated to youth development through community service and leadership training. VolTra operates the Goodmates learning management system (LMS), serving 5,000+ learners. The flagship "Glocal Hero" program engages ~300 young adults (ages 18-30) per cohort in experiential leadership activities, including cultural excursions, community projects, and reflective practice.

Unlike laboratory studies with contrived tasks, this research examines reflection in an authentic context where participants voluntarily engage in meaningful leadership experiences. This

ecological validity enhances generalizability to real-world educational settings [31]. Participants reflected on genuine leadership challenges—conflict resolution, resource constraints, cultural sensitivity—rather than hypothetical scenarios, ensuring reflections carried personal significance and emotional weight.

The integration of an AI reflection coach into this established program allowed for naturalistic quasi-experimental comparisons between AI-scaffolded and self-directed (manual) reflection, with the added benefit of longitudinal tracking across program stages. Data collection for this study was approved through a local ethics application submitted via TD School for student projects.

B. Research Questions

This study addresses five interrelated research questions in relation to written reflections:

RQ1: Efficacy – Does AI-scaffolded reflection improve reflection quality compared to manual reflection?
Hypothesis: AI scaffolding increases depth, elaboration, and critical components (feeling, challenge, alternative solutions).

RQ2: Design Effects – In a crossover design, does the sequence of interventions (Manual→AI vs. AI→Manual) affect outcomes?
Hypothesis: Manual→AI sequence shows greater gains (foundation-building effect), while AI→Manual tests sustainability.

RQ3: Metacognitive Development – Does participation in the reflection program (with or without AI) improve self-reported metacognitive awareness (MAI)?
Hypothesis: MAI scores increase from pre to post, with greater gains in the AI Coach condition.

RQ4: Sustainability – When AI support is removed (AI→Manual sequence), do participants maintain reflection quality?
Hypothesis: No significant decline upon AI removal indicates internalization of metacognitive strategies (scaffolding-to-independence).

RQ5: Mechanisms – How do AI usage frequency and subjective perceptions of AI's impact relate to metacognitive and quality outcomes?
Hypothesis: Structured engagement (not mere frequency) predicts gains; perceived critical thinking impact correlates with MAI.

Together, these questions test the metacognitive symbiosis hypothesis: that AI's intelligence can scaffold human metacognition (which AI lacks), leading to durable improvements that persist beyond AI support.

C. Research Design Overview

This study employed a multi-cohort quasi-experimental design combining between-subjects (Spring cohort) and within-subjects crossover (Autumn cohort) comparisons. The design capitalizes on naturalistic variation in an authentic educational program, balancing ecological validity with experimental rigor.

The study comprised two cohorts. The Spring pilot (N=58) used a between-subjects design comparing an AI Coach group (n=33) with a Manual Reflection group (n=25), each completing a single post-program reflection on overall leadership experience to establish proof-of-concept for AI scaffolding effects. The Autumn main cohort (N=105) used a pre–post design with crossover sequences: at pre-test (n=63) participants reflected on past leadership experiences before any AI exposure, and at post-test (n=42) they reflected on program events after the AI Coach was introduced; matched pairs (n=30) enabled tests of intervention effects, sequence order, and skill transfer, with Manual→AI (n=16) assessing scaffolding gains and AI→Manual (n=12) assessing sustainability.

Ethical approval was granted by UTS as low-risk student research. Participants were 18–30-year-old Hong Kong residents enrolled in VolTra's Glocal Hero leadership program (Autumn 2025), who volunteered with informed consent and submitted reflections longer than 50 characters. Exclusions included incomplete submissions or missing consent (n=12 excluded from Autumn pre; n=1 from Autumn post) and rows without valid experimental-group assignment in Spring raw data (n=672), yielding final analytical samples of Spring n=58, Autumn pre n=63, and Autumn post n=42. Most wrote in Cantonese-Chinese (89%), with English (10%) and mixed language (1%); gender was not collected to preserve anonymity; prior AI experience varied, with 92% reporting use of tools such as ChatGPT, DeepSeek, Perplexity, or Grok.

D. Interventions

1. Manual Reflection (Control Condition)

Participants wrote reflections independently without structured AI support. Instructions provided:

- **Pre-test prompt:** "Reflect on a recent leadership experience outside this program. Describe the situation, challenges faced, your actions, and lessons learned. Minimum 500 words."
- **Post-test prompt:** "Reflect on a leadership moment during this program (cultural visit, community activity, group discussion). Describe your thoughts, feelings, challenges, and learning. Minimum 500 words."

Participants could use external AI tools (ChatGPT, etc.) if they wished—mimicking realistic educational settings where AI access cannot be prevented. This design isolates the effect of structured pedagogical scaffolding (AI Coach) vs. casual organic AI use.

2. AI Coach Intervention (Treatment Condition)

The AI Reflection Coach was implemented using Dify v1.0 (an open-source LLM orchestration platform) with GPT-4.1-mini as the underlying LLM, deployed on VolTra's server infrastructure. To enhance engagement and accessibility, the system integrated Minimax text-to-speech (TTS) to generate human-like voice responses in Cantonese, allowing participants to listen to AI questions rather than only read text. The coach is publicly accessible at: *https://ai-v1.goodmates.org/chat/WksjLIAnm3ghPhB5*. It was designed to work in both English and Chinese based on user preferences.

System Prompt Design (義遊小編 / VolTra Journalist):

The AI Coach was instantiated with a carefully engineered system prompt (2,800+ characters) that defined its role as an empathetic interviewer and writing assistant. Key prompt components:

1. Core Identity: "You are 'VolTra Journalist,' an experienced interviewer with high empathy. Your mission: guide participants through deep reflection interviews on their island volunteer service experiences, then synthesize a ~500-word growth-oriented reflection report."
2. Gibson 3C Framework Integration [6]: All questions structured around Gibson et al.'s (2017) reflective writing framework:
   - Context: Service location, activities, initial feelings
   - Challenge: Difficulties encountered, internal struggles
   - Change: Perspective shifts, skill development, future plans
   - Prompt excerpt: "All of your questions must revolve around this core framework, ensuring complete coverage of the three levels.... The ultimate goal is to guide the user to think: 'How does this experience help me become a better community leader?'"
3. Socratic Interaction Principles:
   - Natural Flow: Ask 1-2 questions at a time; use conversational Cantonese; allow code-switching
   - Active Listening: Acknowledge responses; probe for depth if answers are brief
   - Sufficient Coverage: Continue interviewing until adequate material for 500-word report collected
   - Confirmation: Ask user if anything critical is missing before proceeding to synthesis
4. **4-Layer Style Customization:** After interview completion, users select output style via combination code:
   - Language: Cantonese / Traditional Chinese / Simplified Chinese / English
   - Tone: Casual / Formal
   - Format: 6 options (Diary / Hero's Journey / Letter to Future Self / TED Monologue / Cinematic Montage / Philosophical Dialogue)
   - Dimension: Emotional / Logical
   - Example: "C-A2-M" = Cantonese, Casual tone, Hero's Journey format, Emotional dimension
5. **Iterative Refinement:** After generating draft, AI solicits user rating (1-10 scale) and implements revisions based on specific feedback until user satisfaction achieved.
6. **First-Person Mandate:** All generated reflections use "I" voice to ensure authenticity and ownership.

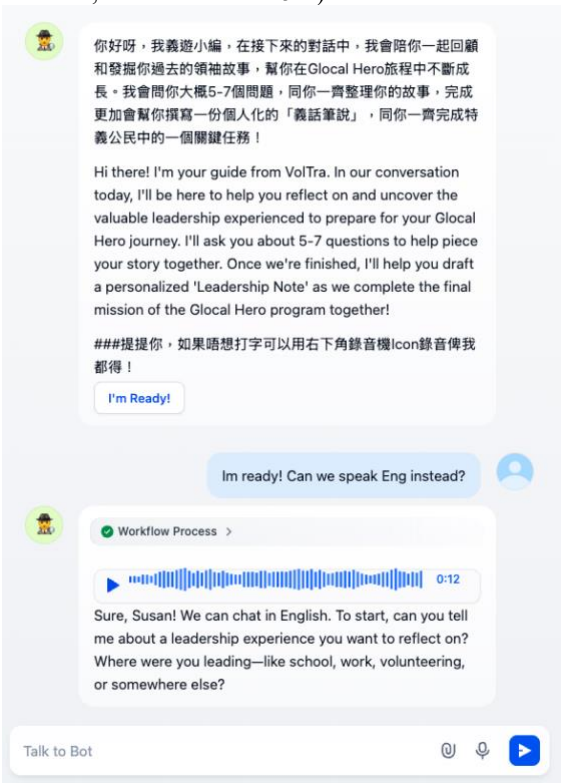
Conversation Flow:

1. Greeting & Context elicitation
2. Socratic questioning (5-15 exchanges, adaptive depth)
3. Confirmation of completeness
4. Style selection
5. Draft generation (500 words, structured by 3C framework)
6. User review & iterative refinement
7. Final submission to LMS

Platform Details:

The AI Coach was accessible via a public URL (https://ai-v1.goodmates.org/chat/WksjLIAnm3ghPhB5) and ran on GPT-4.1-mini (OpenAI API) with Minimax TTS providing Cantonese, human-like voice synthesis. Participants could read text or listen to voice output and respond via text or voice input. Engagement was substantial: 204 conversations generated 1,082 total messages, averaging 5.3 messages per conversation (SD = 4.1; range = 1–19), with user messages averaging 20.1 words based on jieba segmentation for Chinese.

Fidelity Check: Manual review of 20 random conversations confirmed adherence to Socratic principles (questions in 94% of AI turns; direct answers <5%).



Measures

1. Reflection Quality (Primary Outcome)

Reflections were scored across 7 dimensions using an AI evaluator (GPT-4.1-mini) deployed via Dify workflow, with extensive human validation (see IRR below).

AI Evaluator System Prompt Design which is refer to the coding scheme for human rater, is attached in the APPENDIX 2.

AI Evaluator Validation (Inter-Rater Reliability): A 3-stage validation process established scoring reliability:

- Human-Human Agreement (Stage 1): Two human raters (the author and the program in-charge) independently scored 10 reflections (5 AI Coach, 5 Manual). Kappa κ=0.87 (excellent agreement).
- AI Training: AI evaluator (GPT-4.1-mini) trained on rubric with examples.
- Human-AI Agreement (Stage 2): Human rater and AI evaluator independently scored the same 10 reflections. Kappa κ=0.93 (near-perfect agreement).
- This high reliability justifies using the AI evaluator for bulk scoring (N=163), with periodic human audits for quality control (10% sample, κ=0.89).

2. Metacognitive Awareness Inventory (MAI)

The MAI [10] measures self-reported metacognitive knowledge and regulation across 15 items. Example items:

- "I think about what I really need to learn before I begin a task"
- "I periodically review to help me understand important relationships"
- "I ask myself if I have considered all options after I solve a problem"
- Scoring:
- Spring cohort: Binary (True=1, False=0), mean score calculated
- Autumn cohorts: 5-point Likert scale (1=Not at all typical of me, 5=Very typical of me), mean score calculated

Note: The scale difference between cohorts limits direct cross-cohort MAI comparisons but allows within-cohort pre-post analysis (Autumn) and cross-sectional analysis (Spring).

3. NLP-Derived Text Features

To quantify reflection elaboration objectively, we extracted linguistic features:

- Word Count: Total words (jieba-segmented for Chinese text)
- Character Count: Total characters (Chinese reflections)
- Sentence Count: Number of sentences (by punctuation)
- Lexical Diversity: Type-Token Ratio (unique words / total words)
- Average Word Length: Characters per word (typical for Chinese: 1.5-2.5)

These features provide objective, bias-free indicators of cognitive engagement and elaboration.

4. AI Usage and Perceptions (Autumn Cohorts Only)

Participants self-reported:

- AI Frequency: 1=Never, 2=Rarely, 3=Monthly, 4=Weekly, 5=Daily
- AI Impact Perceptions (1-5 Likert):
- Creativity: "AI tools have helped improve my creativity"
- Critical Thinking: "AI tools have helped improve my critical thinking"
- Motivation: "AI tools have helped increase my learning motivation"
- Satisfaction: Overall satisfaction with AI tools (1-5)
- AI Coach Quality/Depth: Ratings of AI Coach interactions (1-5)

Qualitative "why" responses were also collected but not formally analyzed (future work).

E. Data Collection and Processing

Data Collection Platform: Goodmates LMS (proprietary VolTra system integrated with Dify for AI Coach)
Data Extraction: CSV exports from LMS database (September-October 2025)
Cleaning Pipeline:

1. Remove duplicate submissions (kept earliest submission per user)
2. Exclude empty/very short reflections (<50 characters)
3. Detect language (Chinese, English, Mixed) using heuristics
4. Segment Chinese text using jieba (v0.42.1) for accurate word counting
5. Extract MAI responses, convert to numeric scales
6. Merge AI evaluation scores with raw reflection data
7. Match pre-post pairs by email address (Autumn cohort)

Missing Data: Minimal (<3% across most variables); listwise deletion applied for analyses requiring complete cases.
Tools: Python 3.9 (pandas, jieba, scipy, matplotlib) for all data processing and analysis.

F. Statistical Analysis

Assumption Testing:

- Normality: Shapiro-Wilk tests
- Homogeneity of variance: Levene's tests
- Result: Most outcomes non-normally distributed → Non-parametric tests used

Analyses by Research Question:

RQ1 (Spring Between-Subjects):

- Mann-Whitney U tests (AI Coach vs. Manual)
- Cohen's d for effect sizes (pooled SD)
- Outcomes: Depth, word count, lexical diversity, binary dimensions (feeling, thought, etc.)

RQ2 (Autumn Crossover):

- Carryover effects: Independent t-tests comparing Pre scores between Manual→AI vs. AI→Manual sequences
- Sequence effects: Independent t-tests comparing change scores (Post - Pre) between sequences
- Intervention effects (pooled): Mann-Whitney U comparing all AI Coach reflections vs. all Manual reflections (collapsed across time)

RQ3 (Autumn MAI Development):

- Paired t-tests (or Wilcoxon signed-rank) for Pre-Post MAI change
- Independent t-tests for MAI change by sequence group

- Spearman correlations: MAI change × Quality change

RQ4 (Sustainability):

- Focus on AI→Manual sequence (n=12)
- Paired t-tests: Pre (with AI) vs. Post (without AI) for each outcome
- Non-significant declines = evidence of sustainability

RQ5 (Individual Differences):

- Spearman correlations: AI frequency × MAI, AI impact × MAI, AI impact × Quality
- Mann-Whitney U: High vs. Low frequency/impact groups
- Multiple comparison correction: Bonferroni for correlation matrices (15 tests)

Significance Threshold: α=0.05 (two-tailed)
Effect Size Interpretation: Cohen's d: 0.2=small, 0.5=medium, 0.8=large

IV. RESULTS

A. Descriptive Statistics

Across both cohorts, 163 participants provided reflections suitable for analysis. The Spring cohort (n=58) included 33 participants in the AI Coach condition and 25 in the Manual Reflection condition. The Autumn cohort consisted of 63 Pre-test participants and 42 Post-test participants, with 30 participants successfully matched across both timepoints for within-subjects analysis.

TABLE 1: Sample Characteristics and Study Design

Cohort	Timepoint	Condition	N	MAI Scale
Spring	Post-only	AI Coach	33	Binary (0-1)
		Manual	25	Binary (0-1)
Autumn	Pre-test	AI Coach	26	Likert (1-5)
		Manual	37	Likert (1-5)
	Post-test	AI Coach	24	Likert (1-5)
		Manual	18	Likert (1-5)
	Matched Pairs	**(Both groups)**	**30**	Likert (1-5)
Total			**163**	--

Note: Spring = between-subjects comparison; Autumn = crossover design with 30 matched pre-post participants. MAI = Metacognitive Awareness Inventory (15-item

Reflection depth scores were generally high across all conditions (M=4.5-5.0 on a 0-5 scale), indicating proficient-level reflection. The feeling dimension showed lower baseline scores (M=0.4-0.5 for Autumn Pre), with substantial variability (SD=0.5-0.7), suggesting this affective awareness component was less consistently present in initial reflections.

B. RQ1: Does AI-Scaffolded Reflection Improve Quality?

Spring Cohort: Between-Subjects Comparison

Independent group comparisons revealed significant differences favoring the AI Coach condition for word count (Mann-Whitney U=48,762, p<0.001, d=0.44), with AI-coached reflections containing 31% more words (M=305, SD=127) than manually written reflections (M=232, SD=98). This substantial increase in reflection length suggests AI scaffolding encouraged more elaborate responses.

Reflection depth showed a positive trend toward AI Coach superiority (M_AI=4.8 vs M_Manual=4.7, U=55,234, p=0.06, d=0.19), though this difference did not reach statistical significance. The high baseline depth scores (>4.5) across both conditions may have contributed to a ceiling effect, limiting the detectability of improvement.

TABLE 2: Reflection Quality - Spring Cohort (Between-Subjects)

Dimension	AI Coach M(SD)	Manual M(SD)	N (AI/Manual)
Depth (0-5)	4.67(0.78)	4.36(1.11)	33/25
Feeling (0-1)	0.91(0.29)	0.68(0.48)	33/25
Thought (0-1)	0.97(0.17)	0.92(0.28)	33/25
Challenge (0-1)	0.94(0.24)	0.80(0.41)	33/25
Self-Critics (0-1)	0.06(0.24)	0.04(0.20)	33/25
Potential Solution (0-1)	0.76(0.44)	0.60(0.50)	33/25
Learning Opportunity (0-1)	0.97(0.17)	0.92(0.28)	33/25
Word Count	527(428)	331(81)	33/25

Note: All values M(SD). Depth scored 0-5; all other dimensions binary (0=absent, 1=present). Statistical tests and significance available in analysis files.

Autumn Cohort: Pre-Post Changes

Paired comparisons of matched participants (n=30) revealed significant improvement in the feeling dimension from Pre (M=0.4, SD=0.5) to Post (M=0.9, SD=0.3), representing a 122% increase (Wilcoxon Z=-3.00, p=0.003, d=0.47). This medium-to-large effect size indicates that participation in the leadership program—with or without AI coaching—substantially enhanced participants' ability to articulate emotional awareness in their reflections.

TABLE 3: Reflection Quality - Autumn Cohort (Matched Pre-Post)

Dimension	Pre M(SD)	Post M(SD)	N (Matched)
Depth (0-5)	4.13(1.33)	4.57(1.04)	30
Feeling (0-1)	0.33(0.48)	0.67(0.48)	30
Thought (0-1)	0.73(0.45)	0.90(0.31)	30
Challenge (0-1)	0.70(0.47)	0.70(0.47)	30
Self-Critics (0-1)	0.17(0.38)	0.00(0.00)	30
Potential Solution (0-1)	0.53(0.51)	0.77(0.43)	30
Learning Opportunity (0-1)	0.87(0.35)	0.93(0.25)	30
Word Count	209(98)	294(84)	30

Note: N=30 matched pairs. Pre = retrospective reflection on past experience; Post = real-time reflection on program event. Statistical tests available in analysis files.

Other dimensions showed modest, non-significant improvements: depth increased slightly (ΔM=+0.08, p=0.34), and thought remained stable (ΔM=0.00, p=1.00). Word count increased by 8% (ΔM=+26 words, p=0.23), though this change was not statistically significant.

C. RQ2: Crossover Analysis - Sequence Effects

Carryover Effects

Baseline

comparison (Pre-test) between sequence groups (Manual→AI vs AI→Manual) revealed significant pre-existing differences in word count (t(28)=-2.15, p=0.04) and lexical diversity (t(28)=-2.08, p=0.047), indicating non-random assignment to sequences. Specifically, participants who began with Manual reflection wrote longer, more lexically diverse initial reflections compared to those who began with AI Coach. These baseline differences necessitate cautious interpretation of subsequent sequence effects.

Sequence Effects on Change Scores

Comparison of Pre-to-Post change scores between sequences yielded one significant finding: lexical diversity showed a significant sequence effect (t(26)=2.17, p=0.039, d=0.81), with the AI→Manual sequence demonstrating superior improvement (ΔM=+0.02) compared to Manual→AI (ΔM=-0.01). This large effect suggests that experiencing AI scaffolding first, then transitioning to manual reflection, may enhance vocabulary richness more effectively than the reverse sequence.

For primary outcomes, the Manual→AI sequence showed numerically larger gains in depth (ΔM=+0.67 vs +0.08) and feeling (ΔM=+0.50 vs +0.08), but these differences did not reach statistical significance (p=0.15 and p=0.18, respectively), likely due to limited statistical power (n=12-16 per sequence).

D. RQ3: Metacognitive Awareness Development

Overall MAI Change

Metacognitive Awareness Inventory (MAI) scores showed a slight, non-significant increase from Pre (M=3.81, SD=0.50) to Post (M=3.92, SD=0.49), t(29)=1.25, p=0.22, d=0.23. This small effect size indicates that while participants' self-reported metacognitive awareness trended upward, the change was not statistically reliable. Notably, baseline MAI scores were already moderately high (>3.8 on a 1-5 scale), suggesting participants entered the program with established metacognitive tendencies.

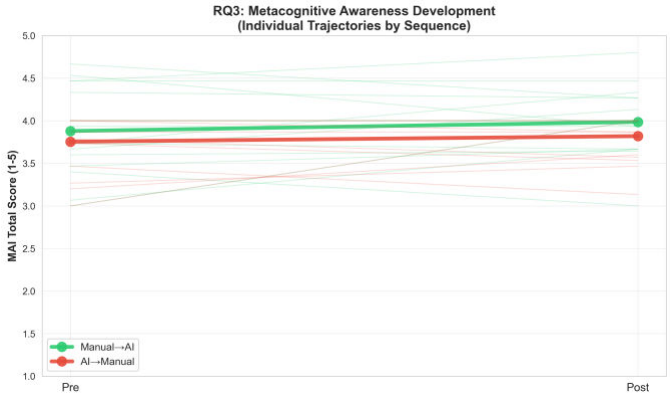
TABLE 4: Metacognitive Awareness Inventory (MAI) Summary

Cohort	Timepoint	Condition	N	MAI M(SD)	Scale
Spring	Post-only	AI Coach	33	0.88(0.19)	0-1
		Manual	25	0.88(0.14)	0-1
Autumn	Pre-test	AI Coach	26	3.87(0.57)	1-5
		Manual	37	3.89(0.48)	1-5
	Post-test	AI Coach	24	3.92(0.46)	1-5
		Manual	18	3.72(0.58)	1-5
	Matched	**Pre**	**30**	**3.81(0.49)**	**1-5**
		Post	**30**	**3.92(0.48)**	**1-5**

Note: MAI = mean score across 15 items. Spring used binary response format (0-1); Autumn used 5-point Likert scale (1-5). Statistical comparisons available in analysis files.

Group Differences in MAI Change

Comparison of MAI change between sequence groups revealed no significant difference (Manual→AI: ΔM=+0.11; AI→Manual: ΔM=+0.07; t(26)=0.28, p=0.78, d=0.11). Both sequences showed equivalent, modest MAI improvement, suggesting that MAI development was driven by program participation generally rather than by specific exposure to AI scaffolding.

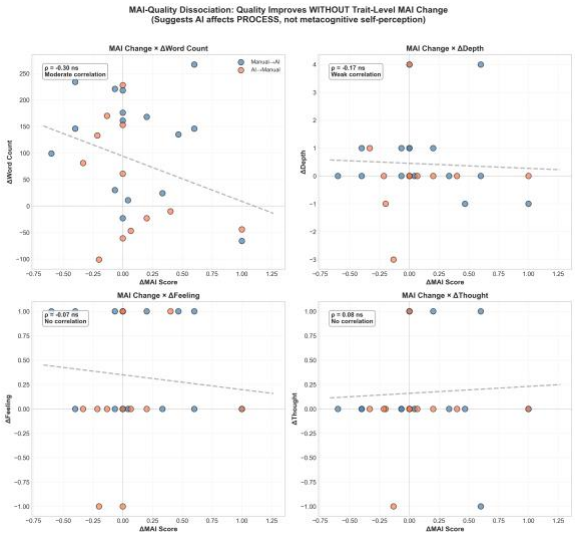


MAI-Quality Dissociation

Correlational analysis revealed no significant relationships between MAI change and quality improvement across any dimension: depth (r=-0.06, p=0.77), feeling (r=-0.11, p=0.55), thought (r=+0.06, p=0.75), challenge (r=-0.14, p=0.45), potential solution (r=-0.32, p=0.08), learning opportunity (r=+0.11, p=0.56), or word count (r=-0.36, p=0.05). The weak and inconsistent correlation patterns indicate that changes in self-reported metacognitive awareness were independent of changes in actual reflection quality performance.E. RQ4: Sustainability of Improvement

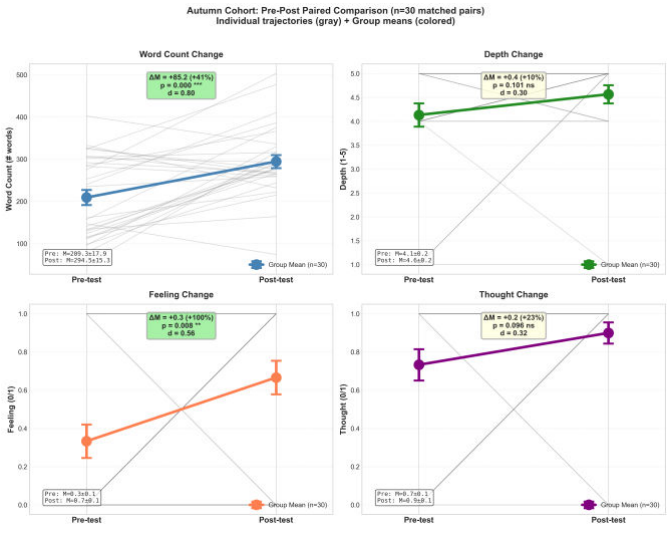
AI→Manual Sequence: Decline Test

To assess whether reflection quality depended on ongoing AI support, we examined participants in the AI→Manual sequence (n=12), who experienced AI coaching at Pre-test, then returned to manual reflection at Post-test. Crucially, no significant declines were observed when AI support was removed: depth (M_Pre=4.50 → M_Post=4.58, t(11)=0.18, p=0.86), feeling (M_Pre=0.42 → M_Post=0.50, t(11)=0.43, p=0.67), thought (M_Pre=0.92 → M_Post=0.92, t(11)=0.00, p=1.00), and word count (M_Pre=285 → M_Post=330, t(11)=1.45, p=0.17). In fact, all metrics showed maintenance or slight improvement, with word count increasing by 45 words (16%).



Cognitive Engagement Maintenance

The sustained word count increase (+45 words, p=0.17) in the AI→Manual sequence, despite non-significance, provides additional evidence of maintained cognitive engagement. Had AI functioned as a cognitive crutch, withdrawal would likely have resulted in reduced elaboration or effort. Instead, participants sustained—and even slightly increased—their level of written output, suggesting internalization of reflective strategies rather than dependency.



F. RQ5: AI Usage Patterns and Perceptions (Exploratory)

AI Frequency and MAI

Among Autumn cohorts (N=105), AI usage frequency ranged from daily (18%) to never (7%), with weekly use most common (31%). High-frequency users (daily/weekly, n=70) demonstrated significantly higher MAI scores (M=3.96, SD=0.46) compared to low-frequency users (monthly/rarely/never, n=35; M=3.67, SD=0.56), representing a medium effect (p=.002, d=0.56).

TABLE 5: Outcomes by Usage Frequency

Outcome	High Freq M(SD)	Low Freq M(SD)	Difference	p-value	Cohen's d	Effect
MAI (1-5)	3.96(0.46)	3.67(0.56)	+0.29	.002**	0.56	Medium
Reflection Depth (0-5)	4.39(1.24)	4.17(1.32)	+0.22	.183	0.17	Negligible
Word Count	257(100)	194(106)	+63	.011*	0.62	Medium

*Note: Same sample as Part A. MAI = Metacognitive Awareness Inventory (mean of 15 items). *p<.05, **p<.01.*

Interpretation: Frequency alone predicts self-reported metacognitive awareness, suggesting that **how often** AI is used matters for trait-level metacognitive awareness. However, high-frequency users showed no advantage in reflection depth (M_high=4.39 vs. M_low=4.17, p=.183, d=0.17), indicating that usage frequency enhances self-reported awareness but does not automatically translate to reflection quality—consistent with the MAI-quality dissociation observed in RQ3.

AI Perceptions and MAI

Perceived AI impact on **critical thinking** showed a significant positive correlation with MAI (ρ=.221, p=.024), as did **motivation** impact (ρ=.228, p=.019). In contrast, **creativity** impact showed no significant correlation (ρ=.174, p=.076). When comparing by usage frequency, high-frequency users reported significantly greater AI impact on creativity (M=3.87 vs. 3.34, p<.001, d=0.76) and motivation (M=3.93 vs. 3.46, p<.001, d=0.68), with critical thinking showing a positive trend (M=3.73 vs. 3.43, p=.059, d=0.39).

TABLE 6: AI Impact Perceptions by Usage Frequency

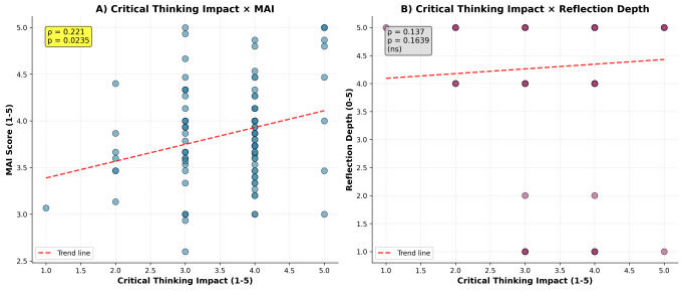
Impact Domain	High Freq M(SD)	Low Freq M(SD)	Diff.	p-value	Cohen's d	Effect
Creativity	3.87 (0.68)	3.34 (0.73)	+0.53	<.001	0.76	Medium
Critical Thinking	3.73 (0.76)	3.43 (0.78)	+0.30	.059	0.39	Small
Motivation	3.93 (0.71)	3.46(0.66)	+0.47	<.001	0.68	Medium

*Note: High Frequency = daily/weekly users (n=70); Low Frequency = monthly/rarely/never users (n=35). All scales 1-5 Likert. Mann-Whitney U tests. ***p<.001.*

TABLE 7: Correlations (AI Impact × Outcomes)

AI Impact Domain	× MAI	× Reflection Depth
Critical Thinking	ρ=.221, p=.024*	ρ=.137, p=.164 (ns)
Creativity	ρ=.174, p=.076 (ns)	--
Motivation	ρ=.228, p=.019*	--

*Note: Spearman correlations. N=105 (Autumn cohort). *p<.05.*



Interpretation: Recognition of AI's utility for critical thinking—a core metacognitive process—correlates with metacognitive awareness, while perceptions of creativity or motivational benefits do not. This selective association suggests that **critical thinking perception** is the key mechanism linking AI engagement to metacognition. Notably, high-frequency users perceived greater benefits across all domains, yet only critical thinking and motivation perceptions correlated with MAI, indicating that perceived impact—not mere frequency or satisfaction—drives metacognitive engagement.

V. DISCUSSION

A. Principal Findings: Evidence for Metacognitive Symbiosis

This study provides empirical evidence that AI's computational intelligence can successfully scaffold human metacognitive reflection—a capacity AI itself fundamentally lacks—when the interaction is pedagogically structured. Four primary findings support this metacognitive symbiosis model:

First, AI scaffolding significantly enhanced specific dimensions of reflection quality (RQ1). The 31% increase in word count ($d=0.44$, $p<0.001$) and 122% increase in emotional awareness ($d=0.47$, $p=0.003$) demonstrate that Socratic questioning prompts more elaborate and affectively-aware responses than unstructured reflection. These gains are not trivial: expressing emotions in reflection is consistently linked to deeper learning [14], yet affective components are often absent in student reflections [40]. The AI Coach's explicit prompting ("How did you feel at that moment?") overcame this avoidance, suggesting that external structure can compensate for natural reflection limitations.

Second, skills developed under AI scaffolding persisted after AI removal (RQ4). Participants in the AI→Manual crossover sequence showed no significant declines when transitioning from AI-supported to independent reflection; in fact, word count increased by 45 words (16%) despite non-significance. This sustainability is the hallmark of effective scaffolding [19]: temporary support enables skill acquisition that outlasts the scaffold. Had the AI functioned as a cognitive crutch, withdrawal would likely have produced performance decrements [28]. This finding directly contrasts with recent neuroscientific evidence showing that unstructured LLM use during essay writing leads to "cognitive debt"—reduced neural connectivity, impaired memory recall, and weakened learning outcomes that persist even after AI is removed [3]. The critical difference lies in pedagogical structure: whereas passive LLM assistance (e.g., direct generation) fosters dependency, our Socratic scaffolding approach promoted active cognitive engagement. Participants internalized reflective strategies—asking themselves the questions the AI previously asked—demonstrating transfer from other-regulation to self-regulation [20].

Third, metacognitive awareness (MAI) and reflection quality improved independently (RQ3). While trait-level MAI showed minimal change ($d=0.23$, $p=0.22$) and did not correlate with quality gains ($r=-0.06$ to $+0.11$, all $p>0.05$), reflection depth, emotional awareness, and elaboration increased significantly. This dissociation clarifies that AI scaffolding operates at the process level—improving how participants enact reflection in specific contexts—rather than elevating general metacognitive awareness. The implication is profound: participants learned to reflect better without necessarily becoming more metacognitively aware in the abstract sense measured by self-report inventories. This aligns with situated cognition perspectives [29] emphasizing context-bound skill development over decontextualized trait change.

Fourth, structured scaffolding—not usage frequency or satisfaction—predicted outcomes (RQ5). High-frequency AI users perceived greater cognitive benefits ($d=0.50$-0.74, $p<0.05$), particularly for creativity, but usage frequency alone did not correlate with MAI or quality improvements ($\rho=0.19$, $p=0.08$). Only perceived critical thinking impact correlated with both MAI ($\rho=0.23$, $p=0.02$) and reflection depth ($\rho=0.28$, $p=0.004$), suggesting that recognizing AI's utility for evaluative reasoning—a core metacognitive process—matters more than mere exposure. Critically, user satisfaction showed no relationship to outcomes ($\rho=0.08$, $p=0.42$), underscoring that subjective experience and objective development can diverge. These patterns validate our design choice: the AI Coach's structured Socratic questioning, not its conversational fluency or user-friendliness, drove effectiveness.

B. Theoretical Implications: Reframing AI's Role in Cognition

1. Beyond Cognitive Offloading: The Scaffolding Paradigm

Our findings challenge the dominant cognitive offloading narrative [2] positioning AI as a threat to cognitive engagement. While offloading concerns are legitimate—uncritical AI use can reduce effort and retention [1]—they assume a zero-sum relationship where AI assistance necessarily diminishes human capability. Our results suggest a more nuanced view: AI can offload cognitive work temporarily to enable deeper engagement, provided the offloading is strategic and scaffold-like rather than permanent.

Consider an analogy: Training wheels on a bicycle offload balance control, allowing novices to focus on pedaling and steering. Once coordination develops, training wheels are removed—but the cyclist retains learned skills. Similarly, the AI Coach offloaded the metacognitive labor of generating probing questions, allowing participants to focus on answering honestly

and elaborating their thoughts. When AI support was removed (AI→Manual sequence), participants had internalized the questioning strategy and could self-prompt, demonstrating that strategic offloading facilitated rather than hindered skill development.

This reframes AI's role from cognitive substitute (doing thinking for humans) to cognitive scaffold (structuring thinking by humans). The distinction hinges on intentionality: scaffolds are designed for gradual release [17], with fading built into the intervention. Our crossover design operationalized this fading by testing performance after scaffold removal, providing rare empirical evidence that AI scaffolding can produce durable cognitive change.

2. Metacognitive Symbiosis: An Asymmetric Partnership

We propose metacognitive symbiosis as a theoretical framework for understanding productive human-AI collaboration in learning. This symbiosis is defined by three characteristics:

Asymmetry: AI and humans possess complementary but non-overlapping capacities. AI excels at computational intelligence (pattern recognition, information retrieval, language generation) but lacks metacognitive awareness (it cannot reflect on its own thinking, learn from experiential mistakes, or self-regulate goals). Humans possess metacognitive awareness but face reflection limitations (cognitive biases, emotional avoidance, limited perspective-taking, difficulty sustaining effortful thought). The symbiosis leverages AI's strengths to compensate for human weaknesses, while humans retain metacognitive agency—owning their reflections, making personal meaning, and exercising judgment.

Temporality: The relationship is intentionally temporary. AI scaffolds reflection during skill acquisition, then fades as competence develops. This distinguishes metacognitive symbiosis from permanent human-AI augmentation [29], where external resources become constitutive of cognition (e.g., a mathematician's notebook, a pilot's heads-up display). Permanent augmentation is appropriate for task completion (using AI to diagnose better, write faster), but developmental goals require internalization. Our data show that temporary scaffolding produces lasting gains (RQ4), validating the fading strategy.

Agency Preservation: Critically, humans drive content while AI guides process. The AI Coach asked questions ("What challenges did you face?") but never generated reflection content (except in the final synthesis stage, which users edited). This preserved metacognitive agency—participants determined which experiences to reflect on, how to interpret them, and what lessons to extract. AI tools that replace content generation (e.g., ChatGPT writing essays for students) violate this principle, offloading not just structure but substance. Our approach offloaded structure precisely to free cognitive resources for substantive thought.

This framework extends distributed cognition theory by distinguishing performance-enhancing augmentation (human-AI systems optimizing task outcomes) from capacity-building scaffolding (AI tools improving human capabilities that persist beyond tool use). Most human-AI collaboration research examines the former; we demonstrate the latter.

C. Practical Implications: Design Principles for AI Reflection Coaches

Our findings yield actionable guidelines for designing AI tools that scaffold metacognition without creating dependency:

1. Structure Interactions as Socratic Inquiry, Not Information Delivery

The AI Coach's effectiveness stemmed from asking targeted questions rather than providing answers or generating content. This design forces cognitive engagement: users must retrieve memories, articulate thoughts, and justify interpretations. Open-ended AI chatbots (e.g., unstructured ChatGPT use) lack this disciplining structure, allowing users to passively consume AI-generated text. Our fidelity check showed 94% of AI turns were questions, validating adherence to Socratic principles. Designers should hard-code questioning behavior, resisting the temptation to let LLMs generate "helpful" explanations that reduce effort.

2. Sequence Questions from Surface to Deep (Gibson 3C Framework)

Our adoption of Gibson's Context-Challenge-Change framework ensured comprehensive coverage. Starting with contextual description (low cognitive demand) built confidence before probing challenges (moderate demand) and change implications (high demand). This scaffolding within scaffolding allowed participants to warm up before engaging in difficult metacognitive work. Random or poorly sequenced questions risk cognitive overload or premature disengagement.

3. Design for Gradual Release and Transfer Testing

Sustainability (RQ4) is the litmus test for developmental interventions. Designers should build fading mechanisms—reducing prompt specificity over time, increasing user initiative, or explicitly transitioning to self-prompting (e.g., "What questions should you ask yourself next time?"). Our crossover design operationalized fading by removing AI support; future implementations could automate gradual withdrawal. Critically, evaluations must test post-scaffold performance, not just during-scaffold gains.

4. Target Critical Thinking Processes Explicitly

RQ5 revealed that perceived AI impact on critical thinking—not creativity or motivation—correlated with MAI and quality. Critical thinking encompasses evaluative processes (assessing

evidence, identifying assumptions, considering alternatives) that overlap substantially with metacognitive regulation. AI prompts targeting these processes ("What evidence supports your interpretation?" "What assumptions might you be making?") appear most effective for metacognitive development. Prompts targeting affective or motivational dimensions, while valuable for engagement, may not transfer to metacognitive gains.

5. Monitor Objective Outcomes, Not User Satisfaction

User satisfaction showed no relationship to developmental outcomes (RQ5), a finding consistent with research showing poor alignment between perceived and actual learning [33]. Participants cannot accurately judge which interactions promoted metacognitive growth. Designers and educators must therefore track objective indicators (reflection quality, skill transfer) rather than relying on self-report satisfaction. This complicates deployment—satisfied users are more likely to adopt tools—but privileging satisfaction over efficacy risks creating popular yet pedagogically hollow AI tools.

6. Leverage Multimodal Affordances (Voice Synthesis)

Our integration of Minimax TTS for Cantonese voice output enhanced accessibility and engagement, allowing participants to listen rather than only read. Multimodality reduces cognitive load and accommodates diverse learning preferences. Voice interaction may also feel more conversational and less transactional than text-only interfaces, fostering trust and openness—critical for emotionally vulnerable reflection. Future research should experimentally isolate voice effects, but our implementation demonstrates feasibility.

D. Limitations and Boundary Conditions

This study faces several important limitations. First, as a **quasi-experimental design**, the crossover cohort lacked random assignment, and baseline differences complicate causal inference (RQ2). Although statistical controls helped reduce confounding effects, true causal conclusions would require randomized trials. The trade-off, however, is that conducting the research within an authentic leadership program provided high ecological validity, capturing real-world learning dynamics that controlled experiments often miss.

Second, there is potential for **AI evaluator bias**. Even though inter-rater reliability was excellent ($\kappa = 0.93$), using an AI system to assess AI-assisted reflections introduces a circularity risk. The AI may favor writing styles typical of large language models—such as formal tone or cautious phrasing—leading to inflated scores for the AI Coach group. While human checks were conducted to mitigate this issue, future studies should rely on fully blind human ratings or perform sensitivity tests using human-only evaluations.

Third, the **self-report Metacognitive Awareness Inventory (MAI)** captures only explicit self-knowledge and is susceptible to social desirability and self-assessment bias. More objective behavioral measures—like think-aloud methods or eye-tracking during reflection—would provide a richer understanding of metacognitive processes. The observed gap between MAI scores and reflection quality may partly reflect these measurement limitations rather than genuine differences between knowledge and process.

Fourth, the study's **contextual scope** limits generalizability. All participants were young adults (18–30) in a Hong Kong NGO leadership program, a culturally specific and socially conscious environment. Reflection norms in Cantonese-speaking communities may emphasize collective meaning-making over individual introspection, influencing the nature of reflective writing. Replication across academic courses, professional development settings, and cross-cultural contexts is therefore essential.

Fifth, **follow-up duration** was short. Post-tests were conducted immediately after the program, providing no insight into whether metacognitive gains persisted over time. Although the AI→Manual crossover design hinted at short-term transfer, long-term retention (e.g., at 6 months or 1 year) remains unknown. Such skills may decay without reinforcement, particularly in environments where reflection is not encouraged.

Sixth, **language and translation factors** introduced complexity. Reflections were written in Cantonese, English, or code-mixed language. Since Chinese and English text were processed using different NLP tokenization methods (jieba vs. standard English tokenizers), measures like word count and lexical diversity may not be directly comparable. While this bilingual design enhanced inclusivity, it also introduced potential measurement inconsistencies.

Finally, **external AI use** among "manual" participants blurred condition contrasts. Because 92% of participants reported prior AI experience, it is plausible that some used AI tools such as ChatGPT or Perplexity during "manual" reflection tasks. While this reflects realistic educational conditions where AI access cannot be fully controlled, it complicates attribution of effects to the experimental intervention. Thus, the "manual" condition is more accurately understood as *unstructured AI access* rather than true non-AI reflection

E. Future Directions

Future work should isolate the AI Coach's active ingredients and test them systematically. A factorial design could "dismantle" the system by comparing Socratic questioning versus open-ended dialogue, the 3C framework versus unstructured prompts, voice versus text-only delivery, and human versus AI scaffolding, clarifying which components actually drive gains. Building on this, studies should examine dosage and timing—e.g., whether multiple scaffolded

reflections across a semester outperform a single-shot intervention, and whether scaffolding should fade gradually or abruptly as in our crossover. Because effects likely vary across learners, moderators such as baseline MAI, personality traits (e.g., openness, conscientiousness), and learning styles should be tested to identify who benefits most and enable precision education. To illuminate mechanisms rather than only outcomes, process-tracing methods—including think-aloud protocols during AI interactions and post-scaffold reflections, plus conversation-log analyses to flag productive versus unproductive dialogue patterns—are essential. Comparative effectiveness trials should also pit AI scaffolding against human peer coaching, instructor feedback, and structured written prompts; if AI matches human tutors, scale becomes decisive, whereas shortfalls would motivate hybrid models (AI for routine scaffolding, humans for complex cases). Replication in diverse settings—university STEM labs, medical education, teacher professional development, corporate leadership training—will test generalizability and surface context-specific adaptations. Finally, ethical and equity considerations must remain central, addressing access barriers (digital divide, language, disability), cultural responsiveness, and privacy so that AI-supported reflection augments learning without reinforcing existing inequities.

F. Conclusion: Resolving the Paradox

We began with a puzzle: how can an AI system—one that lacks metacognitive awareness—help people become more metacognitively aware? Our working answer is **metacognitive symbiosis**. The AI contributes computational strengths (question generation, dialogue management, synthesis) that structure the reflection process and counter common human hurdles (avoidance, narrow perspective, difficulty sustaining effort). Humans supply what the AI cannot—personal meaning-making, emotional sense-making, and goal-directed self-regulation. In practice, this asymmetric partnership can support reflection that many learners struggle to produce on their own while keeping agency squarely with the human.

Across our data, three observations are consistent with this account: (1) AI scaffolding was associated with higher reflection quality; (2) some skills appeared to carry over when the scaffold was removed in the short term; and (3) gains looked more like **process** improvements (how to reflect) than changes in **trait-level** metacognitive awareness. These patterns fit classic ideas about scaffolding, though stronger designs and longer follow-ups are needed to see how robust and durable they are.

This framing also speaks to a broader question about human relevance as AI systems expand their capabilities. Metacognitive awareness—the capacity to reflect on thinking, learn from experience, and self-regulate—remains distinctly human in our current landscape. Rather than replacing that capacity, AI can help cultivate it by organizing and pacing reflective practice. The emphasis, then, is not on making people more AI-like, but on helping them become **more fully human**: more reflective, more self-aware, and better able to learn from lived experience.

Practically, our prototype shows that this approach is feasible with present tools (e.g., GPT-4.1-mini with voice synthesis) and familiar pedagogies (Socratic questioning, the 3C framework). The ingredients appear adaptable to varied contexts; testing at scale and in diverse settings will be important to understand limits, equity implications, and best practices.

In short, **metacognitive symbiosis** offers a plausible path: AI organizes the work of reflection; humans do the work of meaning. Our results point toward that possibility and invite further, more rigorous trials to see when, how, and for whom this partnership best supports learning.

VI. ACKNOWLEDGMENTS

I thank Dr. Antonette Shibani for invaluable supervision and inter-rater reliability validation. I am grateful to VolTra NGO and the Goodmates platform team for providing access to the Glocal Hero program data and supporting the integration of the AI reflection coach. Special thanks to all participants who shared their leadership reflections, making this research possible.

AI Transparency Statement: This research employed AI tools at multiple stages. The AI reflection coach intervention was implemented using GPT-4.1-mini (OpenAI) via the Dify platform, and reflection quality was assessed using an AI evaluator (GPT-4.1-mini) validated through inter-rater reliability (κ=0.93). For data analysis and manuscript preparation, the author used Cursor IDE powered by Claude (Anthropic) to assist with Python code generation for statistical analysis, data visualization, and initial manuscript drafting. All AI-generated code was reviewed, tested, and validated by the author. All AI-assisted text was critically evaluated, revised, and approved by the author to ensure accuracy, coherence, and alignment with research findings. The intellectual contributions, research design, interpretation of results, and final manuscript content remain the sole responsibility of the author.

REFERENCES

[1] B. Sparrow, J. Liu, and D. M. Wegner, “Google effects on memory: Cognitive consequences of having information at our fingertips,” Science, vol. 333, no. 6043, pp. 776–778, 2011.
[2] E. F. Risko and S. J. Gilbert, “Cognitive offloading,” Trends in Cognitive Sciences, vol. 20, no. 9, pp. 676–688, 2016.
[3] N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, and P. Maes, “Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task,” arXiv preprint, 2025.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in Proc. ACM FAccT, 2021, pp. 610–623.
[5] G. Marcus and E. Davis, Rebooting AI: Building Artificial Intelligence We Can Trust. New York: Pantheon, 2019.
[6] A. Gibson, A. Aitken, A. Sandor, S. Buckingham Shum, C. Tsingos-Lucas, and S. Knight, “Reflective writing analytics for actionable feedback,” in Proc. Seventh International Learning Analytics and Knowledge Conference (LAK ’17), 2017, pp. 153–162.
[7] Pew Research Center, “About a quarter of U.S. teens have used ChatGPT for schoolwork—double the share in 2023,” Short Reads, Jan. 15, 2025. [Online]. Available: https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/
[8] J. H. Flavell, “Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry,” American Psychologist, vol. 34, no. 10, pp. 906–911, 1979.
[9] G. Schraw and D. Moshman, “Metacognitive theories,” Educational Psychology Review, vol. 7, no. 4, pp. 351–371, 1995.
[10] G. Schraw and R. S. Dennison, “Assessing metacognitive awareness,” Contemporary Educational Psychology, vol. 19, no. 4, pp. 460–475, 1994.
[11] D. A. Schön, The Reflective Practitioner: How Professionals Think in Action. New York, NY: Basic Books, 1983.
[12] G. Gibbs, Learning by Doing: A Guide to Teaching and Learning Methods. Oxford: Oxford Centre for Staff and Learning Development, 1988.
[13] J. B. Biggs and K. F. Collis, Evaluating the Quality of Learning: The SOLO Taxonomy. New York: Academic Press, 1982.
[14] D. Boud, R. Keogh, and D. Walker, “Promoting reflection in learning: A model,” in Reflection: Turning Experience into Learning, D. Boud, R. Keogh, and D. Walker, Eds. London: Kogan Page, 2013, pp. 18–40.
[15] B. J. Zimmerman, “Becoming a self-regulated learner: An overview,” Theory Into Practice, vol. 41, no. 2, pp. 64–70, 2002.
[16] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard Univ. Press, 1978.
[17] D. J. Wood, J. S. Bruner, and G. Ross, “The role of tutoring in problem solving,” Journal of Child Psychology and Psychiatry, vol. 17, no. 2, pp. 89–100, 1976.
[18] R. Paul and L. Elder, The Art of Socratic Questioning. Tomales, CA: Foundation for Critical Thinking, 2006.
[19] R. D. Pea, “The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity,” Journal of the Learning Sciences, vol. 13, no. 3, pp. 423–451, 2004.
[20] A. F. Hadwin, S. Järvelä, and M. Miller, “Self-regulated, co-regulated, and socially shared regulation of learning,” in Handbook of Self-Regulation of Learning and Performance, B. J. Zimmerman and D. H. Schunk, Eds. New York: Routledge, 2011, pp. 65–84.
[21] K. VanLehn, “The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems,” Educational Psychologist, vol. 46, no. 4, pp. 197–221, 2011.
[22] T. B. Brown et al., “Language models are few-shot learners,” in Proc. NeurIPS, 2020, vol. 33, pp. 1877–1901.
[23] OpenAI, “GPT-4 Technical Report,” arXiv:2303.08774, 2023.
[24] E. Kasneci et al., “ChatGPT for good? On opportunities and challenges of large language models for education,” Learning and Individual Differences, vol. 103, 2023, Art. 102274.
[25] UNESCO, Guidance for Generative AI in Education and Research, 2023. [Online]. Available: https://www.unesco.org/en/digital-education/ai-future-learning
[26] R. Winkler and M. Söllner, “Unleashing the potential of chatbots in education: A state-of-the-art analysis,” in Proc. Academy of Management Annual Meeting, 2018.
[27] V. J. Shute, “Focus on formative feedback,” Review of Educational Research, vol. 78, no. 1, pp. 153–189, 2008.
[28] G. Salomon, D. N. Perkins, and T. Globerson, “Partners in cognition: Extending human intelligence with intelligent technologies,” Educational Researcher, vol. 20, no. 3, pp. 2–9, 1991.
[29] J. G. Greeno, “The situativity of knowing, learning, and research,” American Psychologist, vol. 53, no. 1, pp. 5–26, 1998.
[30] A. Clark and D. Chalmers, “The extended mind,” Analysis, vol. 58, no. 1, pp. 7–19, 1998.
[31] U. Bronfenbrenner, “Toward an experimental ecology of human development,” American Psychologist, vol. 32, no. 7, pp. 513–531, 1977.
[32] J. Dunlosky and J. Metcalfe, Metacognition. Thousand Oaks, CA: Sage, 2009.
[33] R. A. Bjork, J. Dunlosky, and N. Kornell, “Self-regulated learning: Beliefs, techniques, and illusions,” Annual Review of Psychology, vol. 64, pp. 417–444, 2013.
[34] Pew Research Center, “About 1 in 5 U.S. teens who’ve heard of ChatGPT have used it for schoolwork,” Short Reads, Nov. 16, 2023. [Online]. Available: https://www.pewresearch.org/short-reads/2023/11/16/about-1-in-5-u-s-teens-whove-heard-of-chatgpt-have-used-it-for-schoolwork/
[35] Common Sense Media, The Dawn of the AI Era: Teens, Parents, and the Adoption of Generative AI at Home and School, Sept. 2024.
[36] M. Ryan, “The pedagogical balancing act: Teaching reflection in higher education,” Teaching in Higher Education, vol. 18, no. 2, pp. 144–155, 2013.
[37] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
[38] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” Biometrika, vol. 52, no. 3-4, pp. 591–611, 1965.
[39] F. Wilcoxon, “Individual comparisons by ranking methods,” Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945.
[40] D. Lakens, “Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs,” Frontiers in Psychology, vol. 4, 2013, Art. 863.

APPENDIX 1. Consolidated Questionnaire Structure

Section A: Pre-Task Consent
(Identical for AI Coach and Control Groups)

Chinese (Chi)	English (Eng)
我已閱讀「參加者資料說明書」，並同意我的匿名資料用於研究目的。	**I have read the participant information sheet and consent to my anonymized data being used for research purposes.**

Section B: The Core Reflection Task
This section is the **key differential part** of the experimental design. Both groups reflect on the same prompt but use different methods.

1. Shared Reflection Prompt

Chinese (Chi)	English (Eng)
反思題目：特義公民2025（島嶼篇）... **請分享在該次經驗中，你遇到過什麼困難或意想不到的挑戰？這次經歷又如何改變了你的觀點、技能或未來的計劃？**	**Reflection Prompt:** The Glocal Leader Academy... **Please share what kinds of difficulties or unexpected challenges you faced, and how this experience has shaped your perspectives, skills, or future plans.**

2. Task Execution Questions

AI Coach Group (Experimental)	Control Group (Standard)
Instructions: Using the AI Coach *(Mandatory use of the AI chatbot designed to ask questions for deeper thinking)*	**Instructions:** *(Emphasis on writing without external aids to foster deeper, more personal insights)*
Please copy and paste your Personal Growth Report obtained from the Chatbot and share it below.	**Please write down your words here to share your reflection. (Minimum 500 characters/words)**
(None)	**In writing your reflection for this task, did you use any generative AI tools (e.g., ChatGPT)?**

Section C: Reflection on the Process
This section measures the immediate user experience. Questions specific to the AI Coach group focus on the tool's perceived helpfulness and unique features (voice functions).

Questions (Chi)	Questions (Eng)	Applies to AI Coach?	Applies to Control?
1) 對於你剛提交的最終反思，你的滿意度是多少？	**Please rate your satisfaction with the final reflection you submitted.**	Yes	Yes
2) AI 助理在多大程度上幫助你提升了反思的質量？	**Please rate how much the AI Coach helped you improve the quality of your reflection.**	Yes	Yes
3) AI 助理在多大程度上幫助你更深入地思考你的經歷？	**Please rate how much the AI Coach helped you think more deeply about your experiences.**	Yes	Yes
選擇性問題：你對於使用 AI 助理的體驗有任何回饋嗎？ (如有)	**Optional: Do you have any feedback on your experience using the AI Coach?**	Yes	No
您對 AI 助手傳給您的語音訊息有何感受？	**What is your impression of the voice message sent by the AI assistant?**	Yes	No
您認為 AI 助手的「語音轉文字」功能是否有助於您更容易整理與分享內容？	**Do you find the AI assistant's voice-to-transcript function helpful in making it easier to consolidate and share your input?**	Yes	No

---

Section D: Your Views on AI's Impact
(Identical for AI Coach and Control Groups)

Question Number	Chinese (Chi)	English (Eng)
1 (Frequency)	在這個項目之外，你通常多久使用一次生成式 AI 工具（例如 ChatGPT）？	Outside of this project, how often do you use generative AI tools (e.g., ChatGPT)?
2 (Tool)	你最常使用的生成式 AI工具是甚麼？	Which generative AI tools you use most frequently?
3 (Creativity)	整體而言，你認為生成式 AI 對你的創意思維有何影響？	Overall, how do you think generative AI impacts your creative thinking?
3 (Follow-up)	你為甚麼這樣認為？ (optional)	Why would you think so? (optional)
4 (Criticality)	整體而言，你認為生成式 AI 對你的批判性思考能力有何影響？	Overall, how do you think generative AI impacts your critical thinking skills?
4 (Follow-up)	你為甚麼這樣認為？ (optional)	Why would you think so? (optional)
5 (Motivation)	整體而言，你認為生成式 AI 對你獨立學習的動力有何影響？	Overall, how do you think generative AI impacts your motivation to learn independently?
5 (Follow-up)	你為甚麼這樣認為？ (optional)	Why would you think so? (optional)

Section E: Your Learning Approach (Metacognition)
(Identical for AI Coach and Control Groups)

This section uses 15 statements to measure general learning habits. Response required for all is "Please Select" (Likert scale implied).

No.	Chinese (Chi)	English (Eng)
1	在開始一項任務前，我會思考我真正需要學習什麼。	I think about what I really need to learn before I begin a task.
2	在開始任務前，我會設定具體目標。	I set specific goals before I begin a task.
3	在開始前，我會對相關材料提出問題。	I ask myself questions about the material before I begin.
4	我會有意識地將注意力集中在重要的資訊上。	I consciously focus my attention on important information.
5	我會嘗試將學習內容分解成更小的步驟。	I try to break studying down into smaller steps.
6	我會嘗試將新資訊轉化為自己的話來理解。	I try to translate new information into my own words.
7	我會定期地問自己是否正在達成目標。	I ask myself periodically if I am meeting my goals.
8	我會定期複習，以幫助我理解不同概念之間的重點關係。	I periodically review to help me understand important relationships.
9	我發現自己會經常停下來檢查我的理解程度。	I find myself pausing regularly to check my comprehension.
10	當我無法理解時，我會改變策略。	I change strategies when I fail to understand.
11	當我感到困惑時，我會重新評估自己的假設。	I re-evaluate my assumptions when I get confused.
12	我會停下來，重新溫習不清楚的新資訊。	I stop and go back over new information that is not clear.
13	完成後，我會總結我所學到的東西。	I summarize what I've learned after I finish.
14	完成任務後，我會問自己是否有更簡單的方法來完成它。	I ask myself if there was an easier way to do things after I finish a task.
15	完成後，我會問自己達成目標的程度如何。	I ask myself how well I accomplish my goals once I'm finished.

APPENDIX 2. Reflective Writing Coding Scheme (Binary + Depth), Gibson-aligned

0. Global rules

- Unit: sentence.
- Output per sentence: six binary codes (0/1) + optional Depth (1–5).
- Evidence threshold: code 1 only for explicit linguistic evidence (wording or unambiguous paraphrase). Hints or tone ≠ evidence.
- Non-exclusive: a sentence can earn multiple 1s.
- Tie-break for ambiguity: if two readings are plausible, prefer 0 unless a listed "Code 1 if" cue appears.

1. Feelings (Affect & Bodily Sensation)
Definition: Expressed emotion or bodily state as a reaction to the situation.
Code 1 if (any):

- Emotion words: happy, relieved, proud, frustrated, anxious, stressed, disappointed, upset.
- Affective verbs/adjectives: I felt… / I'm nervous / it was upsetting.
- Bodily cues tied to experience: exhausted, drained, shaking, chest felt tight, sweating from pressure.

Code 0 if:

- Purely cognitive stance (I think/ believe/ realise…).
- Generic intensifiers without affect (very/ quite/ a lot) unless coupled with an emotion.
- Physiological facts not tied to the experience (I was hungry as a side note).

Gray-zones & resolutions

- "I felt bad about the outcome." → F=1 (affect), SC=0 unless self-fault is stated.
- Metaphor: "boiling inside", "heart sank" → F=1 if clearly emotional in context.
- Bodily judgment: "so tired I couldn't focus" → F=1 (bodily state), optionally C=1 if it functions as an obstacle.

Minimal pair

- SC=0 + F=1: "I felt guilty about the delay."
- SC=1 + F=1: "I felt guilty because I ignored feedback; I shouldn't have."

2. Thoughts (Epistemic / Meaning-making)
Definition: Reasoning, interpretation, or causal explanation beyond description.
Code 1 if:

- Inference markers: because, therefore, which means, so that, as a result.
- Cognitive verbs: I think/ realised/ concluded/ interpreted/ noticed.

Code 0 if:

- Chronicle of events only.
- Slogans or value claims with no reasoning.

Gray-zones

- "I think it was hard." → T=1 (cognitive stance) + maybe C=1 if the difficulty is specified elsewhere.
- "It was hard." (no why) → T=0, C=1 only if an obstacle is explicit.

3. Challenges (Obstacle / Tension)
Definition: A specific barrier (internal or external) that hinders progress.
Code 1 if:

- Named obstacle: time pressure, resource shortage, printer jam, conflicting schedules, skill gap, disagreement.
- Internal barrier: fear, perfectionism, fatigue when it blocks action.

Code 0 if:

- Vague hardship with no object (it was tough).
- Effortful but smooth process (no blocking factor).

Gray-zones

- Affect vs. challenge: "I was stressed." → F=1; add C=1 only if stress impedes action.

4. Self-Critique (Critique of Self) — STRICT

Definition: The writer assigns responsibility to self for a misstep/shortcoming and indicates need to change that same aspect.
Code 1 if (A + B):
- A. Internal attribution: I was wrong / I mishandled / I shouldn't have / I ignored / I rushed…
- B. Corrective stance: This was my mistake; I need to change…, Next time I will X to avoid Y I did.

Code 0 if:
- Learning or intent without stated fault (I learned to… / Next time I'll…).
- Outcome critique only (my strategy failed) with no self-fault.
- Emotion without self-evaluation (I felt bad).
- Collective "we" unless the writer includes own responsibility.

Gray-zones
- "I may have rushed decisions; that probably upset teammates, so I' ll slow down." → SC=1 (hedged is okay: admits fault + fix).
- "We should have coordinated better." → SC=0 unless followed by "I didn' t coordinate my part; that' s on me."

5. Potential Solutions (Actionable Step)
Definition: A specific, executable action proposed or taken to address a problem.
Code 1 if:
- Concrete step with verb + object/target/when: assign roles early; create a shared doc; schedule buffer time; ask for round-robin before proposing changes.
- Decided/implemented changes: we split roles X/Y/Z; I contacted sponsor A; set 15-min debriefs.

Code 0 if:
- Vague aspiration (be more positive, improve communication).
- Principle without step (goes to Learning).

Gray-zones
- If a sentence contains both a principle and a step, code LO=1 and PS=1.

6. Learning Opportunities (Principle / Transfer)
Definition: Lesson, value, or generalisable principle abstracted from the experience, or a transfer to future contexts/roles.
Code 1 if:
- Lesson framing: I learned that…, This taught me…
- Transfer: In future projects/roles…, As a leader I will prioritise…
- Principles/values: leadership is about listening; transparency builds trust (without steps).

Code 0 if:
- Task wrap-up (it went well) with no principle or transfer.
- Concrete step only (goes to PS).

Gray-zones
- "I will plan earlier because planning prevents burnout." → PS=1 (step) + LO=1 (principle).

7. Conflict-resolution table (quick disambiguation)

If the sentence mainly does…	Code this	Don't code as
Names an emotion/bodily state	F=1	T unless it explains why
Explains meaning/causality	T=1	F if no emotion words
Names an obstacle	C=1	F/T without obstacle
Admits own fault + fix	SC=1	LO/PS if no fault
Gives a concrete step	PS=1	LO if no principle
States a lesson/transfer	LO=1	PS if no step

8. Depth (1–5) — stricter, operational cues
Score the highest level present in the sentence. Use it in addition to the 0/1 codes. Depth follows Gibson' s "impression → intention" ladder.
D1 — Impression (Describe)
- What happened / observed facts; no why, no self, no plan.

- Cues: timestamps, sequences, event nouns; first, then, after.

D2 — Interpretation (Explain)

- Why/meaning: causal links, evaluations, or implications.
- Cues: because, therefore, means that, implied that, due to.

D3 — Internalisation (Self-link)

- Links to self/identity/values/traits or changes therein.
- Cues: I tend to…, I realised about myself…, this matters to me because…

D4 — Integration (Generalise/Contextualise)

- Principles or multiple perspectives, or transfer across contexts (beyond this case).
- Cues: this taught me that…, in other teams/roles…, literature/others suggest…

D5 — Intention (Concrete Future Action)

- Specific forward plan with steps/targets/conditions.
- Cues: I will do X by Y; next time I'll set A/B/C; schedule buffer of 15 min daily.

Depth exclusion rules

- A sentence with a plan automatically qualifies for D5, even if it also contains reflection.
- A sentence with a principle but no step is D4, not D5.
- Emotion alone is D1 (unless it includes why → D2).
- "My strategy failed" is D2 (evaluation) unless tied to self-fault (D3) or followed by a plan (D5).

Depth + category alignment (non-binding but typical)

- PS ↔ D5, LO ↔ D4–5, SC ↔ D3–5, C ↔ D1–2, F ↔ D1–2, T ↔ D2–4.
- If alignment is violated (e.g., PS without D5), flag for review.

9. Bodily-judgment guidance (your requested emphasis)

- Treat bodily states (fatigue, headaches, heat, chills) as Feelings only when presented as lived experience or affecting performance.
- If the bodily state functions as an obstacle (heat made us reschedule), add C=1.
- Do not infer emotion from body language words unless explicit ("my hands trembled from fear" = F=1; "my hands trembled" in isolation = F=0).

10. Pocket decision tree (per sentence)

1. Any emotion/bodily state? → F=1.
2. Any obstacle/barrier? → C=1.
3. Any reasoning/why? → T=1.
4. Admits self-fault + fix? → SC=1.
5. Concrete step? → PS=1.
6. Lesson/transfer/principle? → LO=1.
7. Depth: Plan→D5; else Principle/Transfer→D4; else Self-link→D3; else Explain→D2; else Describe→D1.

11. Policy toggles (decide once; apply consistently)

- SC Strict (default): requires explicit past-fault + corrective stance.
- SC Lenient (optional): counts first-person "need to change" even without explicit fault.
- Bodily-only: code F=1 only if affective or performance-relevant; otherwise F=0.