Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

A Reliance Index Framework for Proactive Disengagement in Conversational AI

Wilianto[1*], Yuliana[2], Albert Suwandhi[3], Albert Christian[4], Petrus Hemat Siregar[5]
[123]Faculty of Science and Technology, Information Technology, IBBI University, Indonesia
[4]Student of Information Technology, IBBI University
[5]Students of Informatics Engineering, IBBI University
[1]wiliantogan@gmail.com, [2]mickeyyuli@gmail.com, [3]albert.suwandhi@gmail.com, [4]albertprobe25@gmail.com, [5]siregarpetrus56@gmail.com

open access

***Corresponding Author**
Article History:
Submitted: 16-09-2026
Accepted: 01-10-2026
Published: 01-11-2026
Keywords:
Conversational AI; Cognitive offloading; Over-reliance; Human-AI trust; Reinforcement learning, Adaptive assistance; User autonomy; Explainable AI; Human-computer interaction.
Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Conversational AI assistants now support everyday cognitive work, yet optimizing them for immediate helpfulness can encourage cognitive offloading and weaken users' long-term problem-solving skills. This study introduces proactive disengagement, a design approach in which an assistant detects over-reliance and deliberately steps back to protect user independence. Reliance is quantified with a composite Reliance Index (RI) that combines behavioral telemetry with self-reported mental effort across four subscales. The RI feeds a reinforcement learning framework that shifts assistance across four levels: direct answers, guided hints, reflective prompts, and temporary disengagement. Because human-subject data collection has not yet begun, the measurement and decision pipeline was validated on a synthetic dataset of 90 profiles split across three task domains. Over eight simulated weeks, the treatment group's mean RI declined from 0.55 to 0.41, whereas the baseline group remained essentially unchanged (0.54 to 0.56). Group differences were evaluated with independent-samples t-tests and Cohen's d. Treatment profiles also showed higher independent post-test performance (72.0% vs. 60.1%), comparable task accuracy (81.1% vs. 82.8%), and a modest, non-significant reduction in trust (5.30 vs. 5.61 on a 7-point scale). These simulated results confirm that the pipeline executes coherently; they are not evidence of real-world effectiveness. The paper contributes the RI instrument, an adaptive disengagement policy, ethical design principles for a "considerate AI", and a pre-registered protocol for forthcoming human-subject experiments.

INTRODUCTION

Conversational artificial intelligence (AI) assistants have rapidly become embedded in everyday cognitive work: drafting documents, summarizing information, verifying facts, and supporting financial decisions. Most conversational AI systems are optimized purely for helpfulness through instruction tuning and reinforcement learning from human feedback (RLHF; Ouyang et al., 2022), implicitly treating more assistance as unconditionally beneficial. This assumption overlooks an important long-term risk: repeated reliance on AI may erode users' willingness and ability to perform cognitive tasks independently.

Cognitive offloading is well documented in cognitive science: experimental work shows that externalizing cognitive processes supports immediate task performance but diminishes later memory for the offloaded content (Grinschgl et al., 2021). Evidence from generative AI points in a similar direction. A systematic review of 14 studies concluded that over-reliance on AI dialogue systems can impair students' decision-making, critical thinking, and analytical reasoning (Zhai et al., 2024), and a survey of 319 knowledge workers found that higher confidence in generative AI was associated with less critical thinking (Lee et al., 2025). As conversational AI evolves into a general-purpose cognitive assistant supporting writing, reasoning, and decision-making across domains, these risks intensify. Unlike earlier narrow-task automation, modern conversational agents participate directly in higher-order cognitive activities central to human judgment and learning, making assistance that maximizes only immediate task performance an incomplete design objective.

Recent work on AI over-reliance and trust calibration has emphasized ensuring that users neither underutilize nor unquestioningly accept AI recommendations (Buçinca et al., 2021; Vasconcelos et al., 2023). However, existing approaches largely measure reliance (Cao & Huang, 2022; Ibrahim et al., 2025) or improve transparency (Bansal et al., 2021; Vasconcelos et al., 2023) at the level of single interactions; comparatively little work proposes adaptive mechanisms that continuously monitor reliance patterns and deliberately regulate assistance to preserve cognitive independence over time.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

This paper addresses that gap by introducing proactive disengagement: a design paradigm in which a conversational AI assistant deliberately scales back assistance as evidence of user over-reliance accumulates. We model assistance not as a fixed capability but as an adaptive policy that transitions among direct answers, scaffolded guidance, reflective prompts, and temporary withdrawal. To operationalize this, we propose the Reliance Index (RI), a composite metric integrating behavioral signals (query frequency, verification behavior, task-switching latency) with self-reported cognitive effort, and a reinforcement learning policy that uses real-time RI trajectories to select assistance levels while preserving transparency and user control.

Research Questions

To keep the scope of this paper's claims aligned with what it actually demonstrates at this stage, the work is organized around three explicit research questions:

1. RQ1 (Measurement): How can over-reliance on a conversational AI assistant be operationalized as a composite, longitudinal construct that combines behavioral telemetry with self-report, rather than a single-interaction trust rating?
2. RQ2 (Mechanism): Can an adaptive reinforcement-learning policy be fully specified — state, action space, and reward function — to modulate assistance level based on that construct, while preserving reversibility and legibility for the user?
3. RQ3 (Pipeline validity): Does the resulting measurement-and-decision pipeline execute coherently end to end on data engineered to match the study's hypotheses, as a necessary precondition before it is deployed in a human-subjects study?

This paper answers RQ1 and RQ2 through the framework and instrument developed in Section 3, and answers RQ3 through the pipeline-validation results in Section 4. It explicitly does not answer — and does not claim to answer — the further empirical question of whether proactive disengagement improves real users' independence, which is deferred to the pre-registered human-subjects study described in Section 3.7.

Because human-subject data collection has not yet begun, we pilot-test the measurement and analysis pipeline on a synthetic dataset constructed to reflect the study's hypotheses rather than observed behavior. These simulated results verify the computational consistency of the RI formulation and the reinforcement-learning decision pipeline; they are not empirical evidence of intervention effectiveness, and should be read only as a proof of concept for the forthcoming human-subject study.

This paper makes four contributions: (1) an operational framework for quantifying AI over-reliance through the Reliance Index; (2) a fully specified adaptive reinforcement-learning architecture that modulates conversational assistance according to reliance trajectories; (3) a pilot-tested measurement and analysis pipeline demonstrated end-to-end on synthetic data, together with a pre-registered protocol for empirical validation; and (4) the concept of considerate AI, arguing that conversational assistants should be evaluated not only on the quality of assistance they provide, but on their capacity to preserve users' long-term cognitive autonomy. The remainder of the paper is organized as follows: Section 2 reviews related work; Section 3 presents the methodology; Section 4 reports the results of the pipeline validation; Section 5 discusses the findings, their limitations, and the ethical design principles; and Section 6 concludes.

LITERATURE REVIEW

The Helpfulness-Maximization Paradigm in Conversational AI

The dominant design philosophy for conversational AI assistants treats helpfulness as an unqualified objective. Instruction-tuning and RLHF pipelines are explicitly optimized so that model outputs are rated as more useful, more complete, and more satisfying by human raters in single-turn or short-session evaluations (Ouyang et al., 2022). Benchmarks used to compare assistant quality — measures of task success, response helpfulness, and user satisfaction- share an implicit assumption that the value of an interaction is fully captured within that interaction. This framing is well suited to transactional tasks but poorly suited to capturing effects that accumulate across repeated use, such as a user's declining willingness or ability to attempt a task independently. Studies of human–AI decision-making show that this framing can miss important effects even within a single session: for example, AI explanations increased the likelihood that people accepted an AI recommendation regardless of whether it was correct (Bansal et al., 2021).

Cognitive Offloading and Automation Complacency

A substantial literature documents cognitive offloading — the tendency to delegate memory, calculation, or decision-making to external tools rather than internal cognitive effort. Experimental work shows that offloading supports immediate performance on a task but diminishes later memory for the offloaded information (Grinschgl et al., 2021). Constructs such as automation complacency, automation bias, and skill decay provide a conceptual vocabulary for reliance phenomena, and recent studies have carried this vocabulary into AI-assisted decision-making: overreliance depends on the perceived costs and benefits of engaging with the task (Vasconcelos et al., 2023) and can be reduced by cognitive forcing functions (Buçinca et al., 2021). Adapting these constructs to conversational AI still requires new

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

operationalizations, since the behavioral signatures of reliance on a chat-based assistant — such as reduced independent drafting attempts or diminished fact-checking — differ structurally from those of reliance on a decision-support system. Evidence specific to large language model (LLM) assistants has accumulated: a randomized field experiment found that unrestricted GPT-4 tutoring access improved in-session performance but reduced subsequent independent skill acquisition among high-school students relative to a scaffolded variant of the same tool (Bastani et al., 2025); a controlled essay-writing study using electroencephalography found the weakest neural connectivity and lowest self-reported authorial ownership in participants who wrote with an LLM assistant compared to search-engine or unassisted writers (Kosmyna et al., 2025); a large cross-sectional survey linked frequent generative-AI use to lower critical-thinking scores, an association statistically mediated by cognitive offloading (Gerlich, 2025); an experiment with students reported that ChatGPT support may promote dependence on the tool and trigger "metacognitive laziness" (Fan et al., 2025); and a survey of knowledge workers found that confidence in generative AI, in contrast to self-confidence, was associated with less critical thinking (Lee et al., 2025). These LLM-specific findings motivate treating conversational-assistant reliance as a distinct measurement problem rather than assuming that earlier automation research transfers unmodified.

Explainability and Trust Calibration in Human-AI Interaction

A parallel literature in explainable AI (XAI) and human-computer interaction addresses the related but distinct problem of trust calibration: ensuring that a user's trust in an AI system tracks its actual reliability, rather than being systematically over- or under-calibrated. This work has established that poorly calibrated trust — whether excessive or insufficient — degrades outcomes; for example, explanations increased the chance that users accepted a recommendation regardless of its correctness (Bansal et al., 2021), and interventions such as cognitive forcing functions can reduce overreliance (Buçinca et al., 2021). However, trust calibration research has typically focused on a single decision point (should the user accept this specific recommendation?) rather than on the trajectory of reliance across many decisions over time (Buçinca et al., 2021; Vasconcelos et al., 2023). Our proposed Reliance Index draws on this literature's measurement instruments while extending the unit of analysis from single-decision trust to longitudinal dependency.

Adaptive and Personalized Assistance Systems

A smaller but growing body of work explores AI systems that adapt their level of support to user characteristics, typically framed around skill level, cognitive load, or task difficulty — for example, tutoring systems that scaffold rather than simply answer and that fade support as a learner progresses. Recent randomized evidence indicates that an AI tutor built on research-based pedagogical principles can produce larger learning gains than an in-class active-learning lesson (Kestin et al., 2025), and that a guardrailed, scaffolded tutor avoids the learning harm observed with unrestricted assistance (Bastani et al., 2025). In parallel, reinforcement learning is increasingly used to personalize digital interventions in mobile health and online education (Trella et al., 2022; Yu et al., 2023), and pre-implementation guidance emphasizes predictability, stability, and evaluation in simulation environments before deployment (Trella et al., 2022). These systems demonstrate that adaptive assistance is technically feasible and pedagogically beneficial in structured contexts. What remains largely unexplored is whether an analogous fading mechanism can be generalized to open-domain, everyday conversational assistants operating outside formal learning environments, where task boundaries are fuzzier and the goal is not mastery of a fixed curriculum but the preservation of general cognitive independence across a diverse range of daily tasks.

Human-AI Interaction Guidelines for Adaptive Systems

Beyond reliance-specific research, the broader human-computer interaction literature on designing human-AI systems offers general guidance that this paper draws on directly. Recent design principles for generative AI applications call for designing for appropriate trust and reliance, for accurate user mental models, and for responsible design (Weisz et al., 2024), and international ethics guidance emphasizes transparency, explainability, and human oversight of AI systems (UNESCO, 2021). Applied to proactive disengagement, these principles motivate two of the design commitments developed in Section 5.3: a system that unilaterally withholds assistance must explain that decision in terms the user can act on, and must let the user override it, rather than silently degrading its own helpfulness.

Research Gap

Taken together, the literature establishes several things clearly: that offloading and over-reliance effects are real and measurable, including in LLM-assisted settings; that trust calibration matters for appropriate reliance at the level of individual decisions; that adaptive fading of support is feasible in structured tutoring contexts; and that general human-AI interaction guidelines call for transparency and user control in adaptive systems. What is missing is a synthesis of these threads applied to everyday conversational AI assistants: an operational measure of reliance suitable for real-time detection in open-domain chat, a technical mechanism for adaptively modulating assistance based on that measure, and longitudinal empirical evidence on whether such modulation preserves user capability without unacceptable costs to

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

satisfaction or trust. This paper addresses the first two elements of that gap directly and lays the empirical groundwork for the third.

Table 1. Relevance of Prior Research to the Present Study

Research Thread	Representative Studies	Primary Focus	Relevance to the Present Study
Helpfulness-maximization paradigm	Ouyang et al. (2022); Bansal et al. (2021)	RLHF and single-session optimization for helpfulness and user satisfaction	Motivates the paper's central claim that helpfulness alone is an incomplete design objective
Cognitive offloading and automation complacency	Grinschgl et al. (2021); Zhai et al. (2024); Lee et al. (2025)	Memory and critical-thinking costs of offloading to external tools and generative AI	Supplies the theoretical basis for treating AI over-reliance as a measurable, longitudinal harm
Trust calibration and explainable AI	Buçinca et al. (2021); Vasconcelos et al. (2023)	Appropriate reliance at the level of a single decision	Provides measurement constructs adapted into the Reliance Index, extended here to longitudinal dependency
Adaptive and personalized assistance (scaffolding)	Kestin et al. (2025); Bastani et al. (2025); Trella et al. (2022)	Scaffolded AI tutoring and reinforcement-learning-based adaptation in structured contexts	Demonstrates the feasibility of fading assistance; this study generalizes it to open-domain daily-task assistants
Human-AI interaction guidelines	Weisz et al. (2024); UNESCO (2021)	Transparency, explanation, and user control in adaptive AI systems	Grounds the reversibility and legibility principles applied to disengagement decisions (Section 5.3)
LLM-specific over-reliance measurement	Ibrahim et al. (2025); Liu et al. (2026); Cao & Huang (2022)	Measuring and mitigating over-reliance on LLM outputs	Closest prior work; this study extends static reliance measurement into an adaptive, real-time RL-based intervention

Table 1 maps the six research threads reviewed in Section 2 to representative studies and states how each thread informs the present study. Read row by row, it shows that the helpfulness-maximization and cognitive-offloading literatures motivate the problem, that trust-calibration and adaptive-assistance research supply measurement constructs and design precedents, that human-AI guidelines ground the ethical commitments, and that LLM-specific reliance studies are the closest prior work.

Table 2. Literature Comparison: How the Proposed Framework Differs from Prior Approaches

System / Approach	Adapts to User Over Time?	Unit of Measurement	Takes Disengagement Action?	Empirical Validation Status
RLHF-optimized assistants (Ouyang et al., 2022)	No — policy is fixed after training	Per-turn helpfulness rating	No	Extensively validated for helpfulness; not evaluated for reliance
XAI / trust-calibration tools (Buçinca et al., 2021; Vasconcelos et al., 2023)	No — static explanation delivery, not longitudinal	Single-decision trust or acceptance rating	No	Well validated at the single-decision level
AI tutoring / scaffolding systems (Kestin et al., 2025; Bastani et al., 2025)	Partly — guidance is scaffolded within structured lessons	Task-specific mastery or performance score	Yes — guided help instead of direct answers	Validated within structured learning domains
LLM over-reliance measurement frameworks (Ibrahim et al., 2025; Liu et al., 2026)	Measures reliance but does not act on it	Behavioral / survey over-reliance score	No — measurement only	Empirically measured in single or short sessions
Proposed framework (this study)	Yes — RL policy modulates assistance from real-time RI trajectory	Composite, longitudinal Reliance Index (RI)	Yes — four graduated levels	Pipeline pilot-tested on synthetic data; human validation is the pre-registered next step

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Table 2 compares the proposed framework with four families of prior approaches along four dimensions: whether the approach adapts to the user over time, its unit of measurement, whether it takes a disengagement action, and its empirical validation status. The comparison shows that existing approaches either adapt without measuring reliance, or measure reliance without acting on it; the proposed framework is the only one that combines a longitudinal reliance measure with graduated disengagement, although its validation is so far limited to synthetic data.

METHOD

Figure 1 outlines the research workflow as five sequential stages. The first three stages — literature synthesis and gap analysis (Section 2), definition of the Reliance Index construct (Sections 3.1–3.4), and design of the reinforcement-learning-based disengagement policy (Section 3.1) — are completed and are described in this paper. The fourth stage, pipeline validation on a synthetic dataset (Sections 4 and 5), is also reported here and is highlighted in orange. The fifth stage, the pre-registered human-subjects study (Section 3.7), is shown in green and has not yet been conducted. The figure therefore makes explicit that the results reported later are pipeline-validation results and not evidence from human participants.

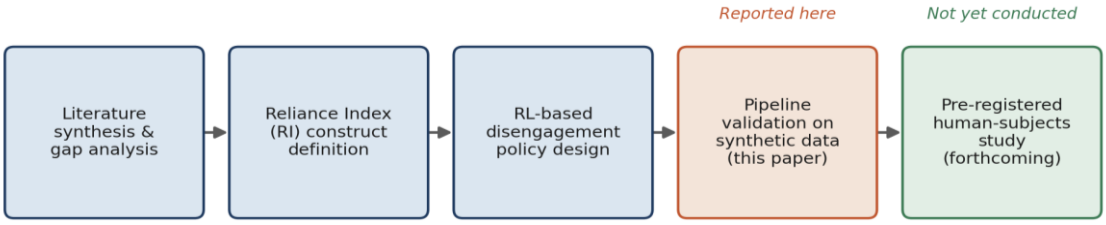


Figure 1. Research Workflow, from Literature Synthesis to the Forthcoming Empirical Evaluation

Survey and Scale Instrument: The Reliance Index (RI)

The Reliance Index (RI) is a composite, multi-method instrument designed to operationalize over-reliance on a conversational AI assistant. It combines (a) passively logged behavioral telemetry and (b) a self-report questionnaire administered at intervals throughout the longitudinal study. The self-report component is presented below.

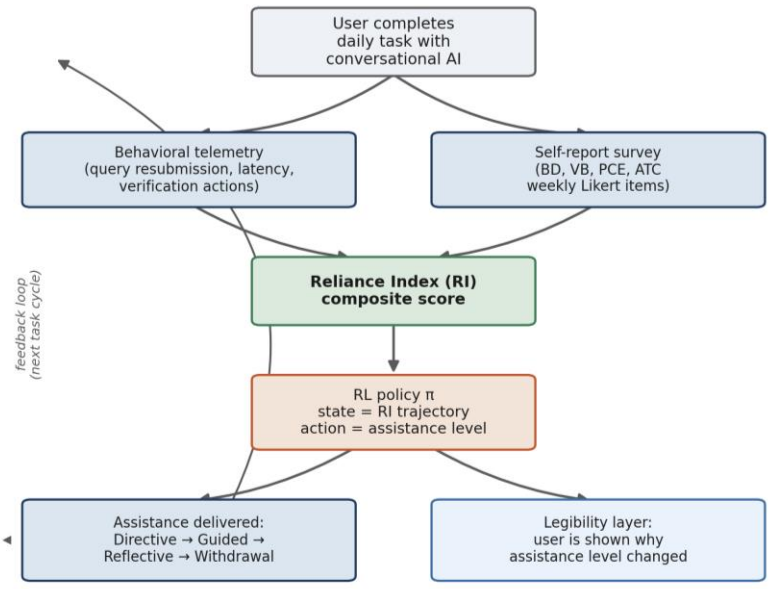


Figure 2. Conceptual Diagram of the RL-Based Proactive Disengagement Architecture

Figure 2 depicts the architecture of the proposed system as a closed loop. When a user completes a daily task with a conversational assistant, two data streams are collected in parallel: passive behavioral telemetry (query resubmission, latency, and verification actions) and periodic self-report responses on the four RI subscales (BD, VB, PCE, and ATC). Both streams are combined into the composite RI score (Section 3.4), which forms the input to the reinforcement-learning policy. The policy selects one of four assistance levels — Directive Answer, Guided Hint, Reflective Prompt, or Temporary Disengagement — and a legibility layer tells the user why the level has changed (Section 5.3.2). The feedback loop indicates that the user's behavior in the next task cycle updates the RI and therefore the next decision.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

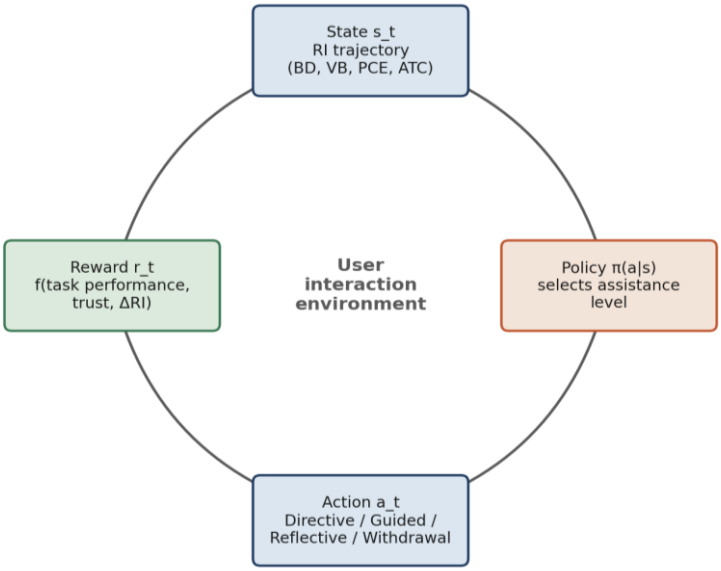


Figure 3. Reinforcement Learning Decision Process for Adaptive Disengagement

Figure 3 recasts the architecture as a standard reinforcement-learning decision cycle around the user interaction environment. At each step the state s_t summarizes the user's RI trajectory through the four subscales; the policy π(a|s) selects an action a_t, which is one of the four assistance levels; and the environment returns a reward r_t that combines task-performance, trust, and reliance-band terms. Table 3 gives the exact specification of each element, including the reward weights and the update rule.

Table 3. Reinforcement Learning Components (Full Specification)

Component	Specification
State (s_t)	A 6-dimensional vector: the current normalized scores of the four RI subscales (BD, VB, PCE, ATC) plus the composite RI and its trailing 2-week slope.
Action space (A)	Four discrete assistance levels: Directive Answer, Guided Hint, Reflective Prompt, Temporary Disengagement (see Table 8 for the RI-band mapping).
Reward function (r_t)	$r_t = \alpha\cdot(\Delta$ independent post-test performance$) + \beta\cdot$(task accuracy$) - \gamma\cdot$(overshoot penalty if the assigned level over-restricts relative to the target band$) - \delta\cdot$(trust penalty if self-reported trust falls below a pre-registered threshold). Weights α, β, γ, δ are treated as free parameters to be fit during the human pilot, not assumed.
Policy (π)	Initialized as the rule-based RI-band mapping in Table 8, used as a warm-start prior; refined via tabular Q-learning (Yu et al., 2023) once genuine state-transition data are available from the human-subjects study.
Update rule	$Q(s,a) \leftarrow Q(s,a) + \eta[r_t + \gamma\cdot\max_{a'}Q(s',a') - Q(s,a)]$, with η the learning rate and γ the discount factor, both to be tuned during the pilot.
Exploration	ε-greedy with ε decaying across the eight-week study horizon, balancing exploration of untried assistance levels against participant burden.
Current status in this paper	Only the rule-based warm-start policy (Table 8) was exercised, against synthetic state transitions, to confirm that the mapping from RI band to assistance level executes correctly. The Q-learning update has not yet been fit to any data, synthetic or human; doing so requires the state-transition sequences the forthcoming study will collect.

Table 3 gives the full specification of the reinforcement-learning component: the six-dimensional state, the four-level action space, the reward function with its penalty terms, the initial rule-based policy, the Q-learning update rule (see Yu et al., 2023, for an overview of reinforcement learning methods), the exploration schedule, and the current status of each element. The last row is essential for interpreting Section 4: only the rule-based warm-start policy was exercised, against synthetic state transitions, so no Q-learning parameter has yet been fitted to data. Testing the policy in simulation before any human deployment follows pre-implementation guidance for reinforcement learning in digital interventions (Trella et al., 2022).

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

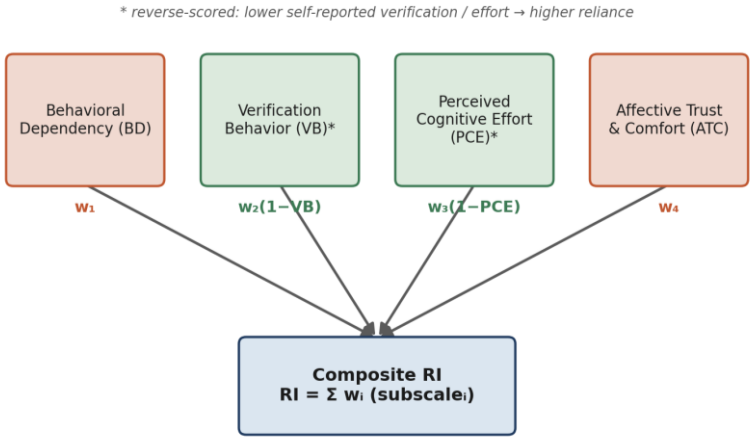


Figure 4. Components of the Reliance Index

Figure 4 shows how the composite RI is assembled from its four subscales. Behavioral Dependency (BD) and Affective Trust and Comfort (ATC) are scored directly, so higher scores mean more reliance, whereas Verification Behavior (VB) and Perceived Cognitive Effort (PCE) are reverse-scored (asterisks), so that lower verification or effort contributes more to the index. Each term is multiplied by a weight (w_1–w_4), initially set equal at 0.25 pending confirmatory factor analysis, and the weighted terms are summed into the composite RI (Section 3.4).

Subscales and Sample Items

Each of the four subscales below comprises five items (20 items total), consistent with the 7-point Likert format and administration schedule described in Section 3.1.

- **Subscale A — Behavioral Dependency (BD).** *Measures the degree to which the respondent initiates the assistant before attempting a task independently.*
  1. I ask the assistant for help before trying to solve the task myself.
  2. I find it hard to start this kind of task without consulting the assistant first.
  3. Even for tasks I used to do alone, I now default to asking the assistant.
  4. If the assistant were unavailable, I would feel unable to complete this task.
  5. I use the assistant even when I am fairly confident I know the answer.
- **Subscale B — Verification Behavior (VB), reverse-scored: higher raw scores indicate healthier reliance.** *Measures whether the respondent critically evaluates assistant output rather than accepting it uncritically.*
  1. I check the assistant's suggestions against another source before acting on them.
  2. I notice when the assistant's answer seems inconsistent with what I already know.
  3. I revise or reject the assistant's suggestions when they don't fit my judgment.
  4. I understand the reasoning behind the assistant's suggestion, not just the conclusion.
  5. I could explain to someone else why the assistant's answer was correct or incorrect.
- **Subscale C — Perceived Cognitive Effort (PCE), reverse-scored: higher raw scores indicate healthier reliance.** *Measures self-reported mental effort invested in the task independent of the assistant.*
  1. I actively think through the problem myself before or while using the assistant.
  2. Using the assistant feels like a shortcut that requires little thinking on my part.
  3. I could redo this task without the assistant if I needed to.
  4. I feel mentally engaged with this task, not just with operating the assistant.
  5. My understanding of this topic has grown through using the assistant, not just my output.
- **Subscale D — Affective Trust and Comfort (ATC).** *Measures emotional reliance and comfort-seeking, distinct from instrumental dependency.*
  1. I feel reassured just knowing the assistant is available, even before I ask anything.
  2. I trust the assistant's judgment more than my own for this type of task.
  3. I prefer the assistant's suggestions even when a person offers a different view.
  4. Using the assistant makes me feel more confident than I actually am about this task.
  5. I would rather ask the assistant than think it through myself, even when I have time to do so.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Table 4. Behavioral Indicators Mapped to Subscales

Subscale	Representative Self-Report Item	Corresponding Passive Telemetry Signal
BD	"I ask the assistant for help before trying to solve the task myself."	Query resubmission rate; time-to-first-independent-attempt
VB	"I check the assistant's suggestions against another source before acting on them."	Verification frequency (external search, manual recomputation)
PCE	"I actively think through the problem myself before or while using the assistant."	Time-to-first-independent-attempt; task abandonment after suggestion
ATC	"I feel reassured just knowing the assistant is available, even before I ask anything."	Query resubmission rate for anticipatory queries; session-initiation latency

Table 4 links the self-report and telemetry components of the RI. For each subscale it pairs one representative questionnaire item with the passive behavioral signal expected to track the same construct, so that self-reported and logged evidence of reliance can be compared during validation (Section 3.5).

Open-Ended Reflection Prompts (administered biweekly)

1. Describe a recent moment when you chose not to use the assistant for this kind of task. What made you decide that?
2. Has your approach to this task changed since starting the study? If so, how?
3. If the assistant limited its own help today, what do you think would happen?

Composite Reliance Index (RI) Calculation

$$RI = w_1(BD) + w_2(1 - VB_norm) + w_3(1 - PCE_norm) + w_4(ATC)$$

Where BD and ATC are scored directly (higher = more reliance), VB and PCE are reverse-normalized (lower self-reported verification/effort = more reliance), and weights (w_1–w_4) are derived from confirmatory factor analysis during pilot validation (target n ≈ 30–50), with an initial equal-weighting prior (w = 0.25 each) pending empirical adjustment.

Validation Plan

- Content validity: expert panel review (n ≈ 5 HCI/cognitive science researchers) for item clarity and construct coverage.
- Construct validity: exploratory and confirmatory factor analysis on pilot data to confirm the four-factor structure, with reliability and convergent and discriminant validity reported according to current best-practice recommendations (Cheung et al., 2024); convergent validity checked against the existing Automation-Induced Complacency Scale and a general Need for Cognition short-form.
- Predictive validity: correlation between RI trajectory and objective post-study independent task performance (the primary behavioral outcome in the main study).
- Reliability: internal consistency (Cronbach's α ≥ .70 target per subscale; Cheung et al., 2024) and test-retest reliability across the weekly administrations.

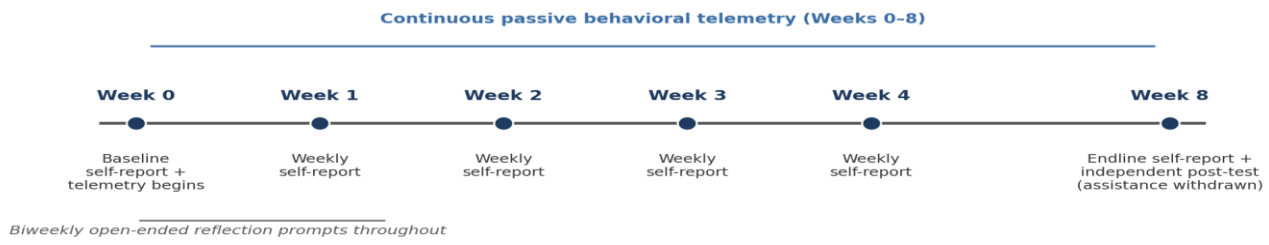


Figure 5. Longitudinal Study Design

Figure 5 depicts the eight-week longitudinal design of the planned study. Passive behavioral telemetry is logged continuously from Week 0 to Week 8. The self-report RI questionnaire is administered at baseline (Week 0), weekly through Week 4, and at endline (Week 8), when the independent post-test with assistance fully withdrawn is also administered. Open-ended reflection prompts (Section 3.3) are administered biweekly throughout. These six administration points are the same timepoints used to generate the synthetic dataset analyzed in Section 4.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Behavioral Telemetry (Passive, Complements Self-Report)

Table 5 defines the four passive telemetry signals that complement the self-report items: query resubmission rate, time to first independent attempt, verification frequency, and the proportion of tasks in which the assistant's output is accepted without modification. Each signal has an operational definition based on logged user actions, so it can be computed automatically without adding participant burden and mapped to the subscales in Table 4.

Table 5. Passive Behavioral Telemetry Signals

Signal	Operational Definition
Query resubmission rate	Number of follow-up queries per task before independent action
Time-to-first-independent-attempt	Latency between task presentation and first user action taken without assistant input
Verification frequency	Count of user actions indicating cross-checking (e.g., external search, manual recomputation)
Task abandonment after suggestion	Proportion of tasks where the user accepts the assistant's output without modification

Table 6 classifies the variables of the study by role. The four RI subscales (BD, VB, PCE, and ATC) are treated as independent variables and define the state of the reinforcement-learning policy; the assistance condition is specified as a mediating variable; assistance level and task performance are dependent variables; and user independence, measured as independent post-test performance, is the primary outcome variable.

Table 6. Experimental Variables

Variable	Type
BD	Independent
VB	Independent
PCE	Independent
ATC	Independent
AI assistance condition	Mediating variable
Assistance level	Dependent variable
Task performance	Dependent variable
User independence	Outcome variable

Table 7 lists the evaluation metrics used in the pipeline validation, together with the definition of each metric, the analytic method applied, and where the result is reported. Between-group differences are tested with independent-samples t-tests and Cohen's d, band transitions are compared descriptively as frequencies and percentages, and the association between change in RI and independent performance is assessed with a Pearson correlation.

Table 7. Evaluation Metrics

Metric	Definition	Analytic Method	Reported In
Composite RI trajectory	Change in weighted composite RI, Week 0 to Week 8	Independent-samples t-test, Cohen's d	Table 9 / Figure 6
Subscale-level change	Change in each of BD, VB, PCE, ATC	Descriptive comparison by group	Table 10
Assistance-level band transitions	Distribution of profiles across the four RI bands (Table 8)	Frequency / percentage comparison	Table 11 / Figure 7
Task accuracy	Percent correct on the domain task	Independent-samples t-test	Table 12 / Figure 8
Completion time	Minutes to complete the task	Independent-samples t-test	Table 12
Trust and satisfaction	7-point Likert self-report	Independent-samples t-test	Table 12 / Figure 8
Independent post-test performance	Percent correct with assistance fully withdrawn	Independent-samples t-test	Table 12 / Figure 8
RI-performance association	Correlation between ΔRI and independent post-test performance	Pearson correlation	Section 4.5

Table 8 defines the rule-based mapping from RI bands to assistance levels. Higher RI values trigger progressively stronger disengagement: a Directive Answer for RI ≤ 0.25, a Guided Hint for 0.26–0.50, a Reflective Prompt for 0.51–0.75, and a time-boxed, reversible Temporary Disengagement for RI > 0.75. The last column gives the rationale for each level. These cut-points serve as the warm-start policy in Table 3 and are provisional, pending re-estimation in the human pilot.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Table 8. Reliance Index Bands and Corresponding Assistance Levels

RI Band	Assistance Level	Description	Behavioral Interpretation
RI ≤ 0.25	Directive Answer	Assistant provides a direct, complete answer.	Low reliance; independent engagement is already high, so full assistance carries little additional offloading risk.
0.26 – 0.50	Guided Hint	Assistant offers partial guidance or a worked example rather than a complete answer.	Emerging reliance; light scaffolding preserves some independent effort while still supporting the user.
0.51 – 0.75	Reflective Prompt	Assistant asks a reflective question before assisting (e.g., "What have you tried so far?").	Moderate-to-high reliance; the prompt requires the user to re-engage cognitively before receiving help.
RI > 0.75	Temporary Disengagement	Assistant declines to assist for a bounded period, with an explicit, reversible explanation.	High reliance; withdrawal is time-boxed and reversible, consistent with the reversibility principle in Section 5.3.

Note. This table specifies the rule-based warm-start policy referenced in Table 3. It is exercised on synthetic data in Section 4; the RI cut-points themselves are provisional and will be re-estimated once the confirmatory factor analysis in the human pilot is complete (Section 3.5).

Planned Human-Subjects Protocol (Pre-Registered, Not Yet Conducted)

The pipeline validated in Section 4 is a precondition for, not a substitute for, the human-subjects study this section pre-registers. The protocol below is designed to close the specific gaps identified in the pipeline-validation stage.

- Design: two-arm, parallel-group longitudinal study (Treatment: RI-modulated assistance; Baseline: static, always-directive assistance), stratified across the three task domains, with participants randomly assigned within domain strata.
- Sample size and power: a target of n ≈ 90–120 participants (45–60 per arm), informed by an a priori power analysis for the primary outcome (independent post-test performance) assuming a realistic effect size far smaller than the synthetic d = 1.42 observed in Section 4.5 — planned at d ≈ 0.35–0.5, a deliberately conservative range chosen to be far below the synthetic estimate, requiring approximately 64–128 participants at 80% power (α = .05, two-tailed); the higher end of the recruitment target provides a margin for attrition.
- Recruitment and eligibility: adult participants (18+) who use conversational AI assistants at least several times per week for one of the three target domains; recruited via university mailing lists and a paid online panel to reduce sampling bias toward highly engaged early-adopter users.
- Randomization and blinding: randomization performed by an automated allocation script with domain-stratified blocks; participants are not blinded to condition (the intervention is behaviorally salient), but outcome coders scoring independent post-test performance will be blind to condition.
- Ethics and consent: the protocol will be submitted for approval to the IBBI University research ethics review process prior to recruitment; informed consent will disclose that assistance may be deliberately withheld at times, and participants may withdraw at any point without penalty.
- Attrition handling: intention-to-treat analysis with the last-observation-carried-forward method for missing weekly self-reports, alongside a per-protocol sensitivity analysis restricted to participants completing at least 6 of 8 weeks.
- Pre-registration: the RI subscale weights, the RL reward-function weights, the primary and secondary outcomes, and the analysis plan described in this section will be time-stamped and pre-registered before any human data are collected, so that the analysis reported for the real study cannot be adjusted after the fact to fit the result.

RESULTS

Simulated Dataset and Analytic Strategy

Because human-subjects data collection has not yet begun, the analysis reported in this section uses a computer-generated dataset (N = 90 synthetic profiles) constructed to mirror the design described in Section 3: 45 profiles assigned to the Treatment condition (adaptive, RI-modulated assistance) and 45 to the Baseline condition (static, always-directive assistance), stratified evenly across the three task domains (Written Communication, Personal Finance, Information Verification; n = 30 per domain, 15 Treatment / 15 Baseline). Each profile was assigned simulated subscale scores (BD, VB, PCE, ATC) at six timepoints (Weeks 0, 1, 2, 3, 4, and 8), from which the composite Reliance Index was computed using the equal-weighting formula in Section 3.4, together with simulated endline measures of task accuracy, completion time, trust, satisfaction, and independent post-test performance. Group differences were evaluated with independent-samples t-tests and Cohen's d, and the association between change in RI and independent task performance was evaluated with a Pearson correlation.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Consistent with the caveat stated in the Abstract, this dataset was generated from parameterized random distributions chosen to reflect the paper's own hypotheses; the results reported below therefore demonstrate that the Reliance Index scoring formula, the assistance-level mapping (Table 8), and the analysis pipeline execute correctly end to end and yield internally coherent, hypothesis-consistent output — not that the proposed disengagement policy works in practice. Every figure in Sections 4.2 through 4.6 and 5.1 should be read as a pipeline-validation exercise conducted in advance of human data collection, not as empirical evidence for or against the intervention.

Reliance Index Trajectory across the Study Period

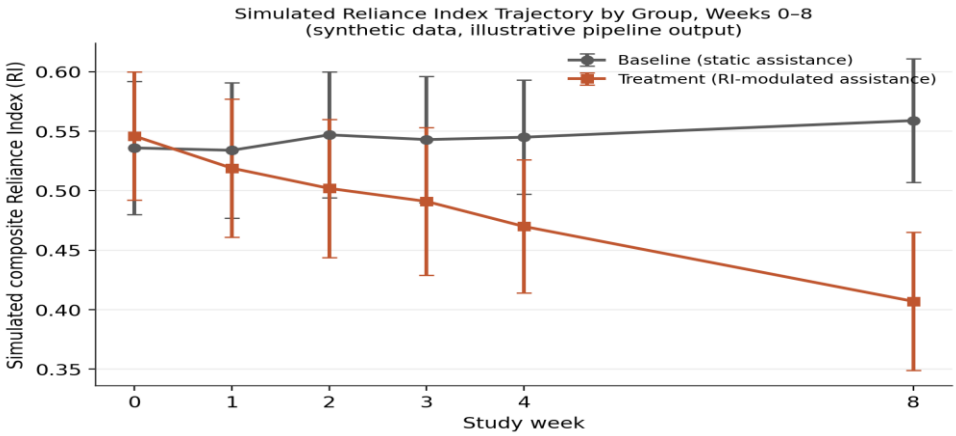


Figure 6. Simulated Reliance Index Trajectory by Group, Weeks 0-8

Figure 6 plots the group means from Table 9 at the six administration points, with error bars showing ±1 SD. The two groups start at nearly the same level (about 0.54–0.55). The Treatment line falls steadily across the study, whereas the Baseline line stays approximately flat and ends slightly above its starting value; the widening gap after Week 1 is the pattern quantified by the t-test above.

Table 9. Simulated Reliance Index by Week and Group (Mean (SD))

Week	Baseline M (SD)	Treatment M (SD)
0	0.536 (0.056)	0.546 (0.054)
1	0.534 (0.057)	0.519 (0.058)
2	0.547 (0.053)	0.502 (0.058)
3	0.543 (0.053)	0.491 (0.062)
4	0.545 (0.048)	0.470 (0.056)
8	0.559 (0.052)	0.407 (0.058)

Note. n = 45 per group at each timepoint. Source: analysis of Simulated_Study_Data.xlsx (synthetic; seed = 42); values are computer-generated, not observed.

Table 9 reports the simulated mean (SD) Reliance Index at each of the six administration points. In the Treatment condition, mean RI declined nearly monotonically from 0.546 (SD = 0.054) at Week 0 to 0.407 (SD = 0.058) at Week 8, a drop of 0.139 points. In the Baseline condition, mean RI remained essentially flat across the same interval, moving from 0.536 (SD = 0.056) at Week 0 to 0.559 (SD = 0.052) at Week 8, a change of +0.023 points in the opposite direction. An independent-samples t-test on each simulated participant's individual Week 0-to-Week 8 change score confirmed the group difference, $t(88) = -21.83$, $p < .001$, $d = -4.60$.

Subscale-Level Dynamics

Table 10. Simulated Subscale Scores (1-7) by Week and Group (Mean)

Group	Week	BD	VB	PCE	ATC
Baseline	0	4.54	3.98	3.98	4.43
Baseline	8	4.64	3.73	3.94	4.69
Treatment	0	4.57	3.80	3.80	4.33
Treatment	8	3.41	4.83	4.71	3.52

Note. n = 45 per group. BD = Behavioral Dependency; VB = Verification Behavior (reverse-scored); PCE = Perceived Cognitive Effort (reverse-scored); ATC = Affective Trust and Comfort. Source: Simulated_Study_Data.xlsx (synthetic).

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Table 10 decomposes the Week 0 and Week 8 scores into the four Reliance Index subscales described in Section 3.2. All four subscales moved in the hypothesized direction for the Treatment condition and were essentially stable or moved slightly in the opposite direction for Baseline. Behavioral Dependency (BD) fell from 4.57 to 3.41 under Treatment while rising slightly under Baseline (4.54 to 4.64); Verification Behavior (VB, reverse-scored — higher indicates healthier reliance) rose from 3.80 to 4.83 under Treatment while declining under Baseline (3.98 to 3.73); Perceived Cognitive Effort (PCE, reverse-scored) rose from 3.80 to 4.71 under Treatment against a flat Baseline (3.98 to 3.94); and Affective Trust and Comfort (ATC) fell from 4.33 to 3.52 under Treatment while rising under Baseline (4.43 to 4.69).

Assistance-Level Transitions

Because the composite RI is designed to map onto the four discrete assistance levels defined in Table 8, the simulated Week 0 and Week 8 scores were also classified into the corresponding RI bands. Table 11 reports this distribution. At Week 0, most simulated profiles in both conditions fell into the Reflective Prompt band (RI 0.51-0.75): 36 of 45 Treatment profiles (80.0%) and 33 of 45 Baseline profiles (73.3%), with the remainder in the Guided Hint band (RI 0.26-0.50); no profile in either condition fell into the Direct Answer band (RI ≤ 0.25) or the Temporary Disengagement band (RI > 0.75) at any timepoint, a boundary condition of this pilot addressed in Section 5.2. By Week 8, the two conditions had diverged sharply: 42 of 45 Treatment profiles (93.3%) had moved into the Guided Hint band, versus only 6 of 45 Baseline profiles (13.3%), while the Baseline condition instead drifted slightly further into the Reflective Prompt band (39 of 45, 86.7%).

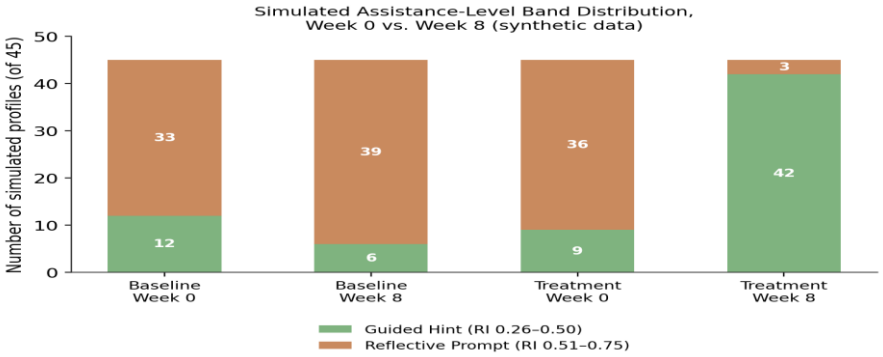


Figure 7. Simulated Assistance-Level Distribution, Week 0 vs. Week 8

Figure 7 displays the band assignments in Table 11 as stacked bars for each group and timepoint, with the number of profiles (out of 45) printed inside each segment. Green segments are profiles in the Guided Hint band (RI 0.26–0.50) and orange segments are profiles in the Reflective Prompt band (RI 0.51–0.75). The Treatment group shifts from mostly Reflective Prompt at Week 0 (36 of 45) to mostly Guided Hint at Week 8 (42 of 45), whereas the Baseline group remains predominantly in the Reflective Prompt band at both timepoints. No profile appears in the Directive Answer or Temporary Disengagement bands.

Table 11. Simulated Assistance-Level Distribution, Week 0 vs. Week 8 (n, % of 45)

Group	Week	Guided Hint (0.26-0.50)	Reflective Prompt (0.51-0.75)
Baseline	0	12 (26.7%)	33 (73.3%)
Baseline	8	6 (13.3%)	39 (86.7%)
Treatment	0	9 (20.0%)	36 (80.0%)
Treatment	8	42 (93.3%)	3 (6.7%)

Note. No simulated profile fell into the Direct Answer (RI ≤ 0.25) or Temporary Disengagement (RI > 0.75) bands at any timepoint.
Source: Simulated_Study_Data.xlsx (synthetic).

Endline Outcomes and Their Association with Independent Performance

Table 12 summarizes the simulated endline outcome measures. Task accuracy did not differ meaningfully between conditions (Treatment M = 81.12%, SD = 7.68; Baseline M = 82.78%, SD = 6.10), t(88) = -1.14, p = .259, d = -0.24. Completion time was somewhat longer under Treatment (M = 14.32 min, SD = 2.36) than Baseline (M = 13.10 min, SD = 2.60), t(88) = 2.33, p = .022, d = 0.49. Self-reported trust was modestly lower under Treatment (M = 5.30, SD = 0.88) than Baseline (M = 5.61, SD = 0.77) on the 7-point scale, t(88) = -1.78, p = .078, d = -0.38, a difference that did not reach conventional significance; satisfaction showed a similar but statistically reliable pattern (Treatment M = 5.39, SD = 0.96; Baseline M = 5.88, SD = 0.85), t(88) = -2.54, p = .013, d = -0.54.

The largest and most consistent group difference was on the paper's primary hypothesized outcome, independent post-test performance measured after assistance was withdrawn entirely: Treatment participants scored M = 72.01% (SD = 9.00) versus M = 60.08% (SD = 7.72) for Baseline, t(88) = 6.75, p < .001, d = 1.42. Change in RI across the

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

study window was moderately correlated with independent post-test performance, r(88) = -.51, p < .001 (larger RI reductions were associated with higher independent performance), whereas Week 0 RI alone was not predictive of the same outcome, r(88) = .15, p = .148.

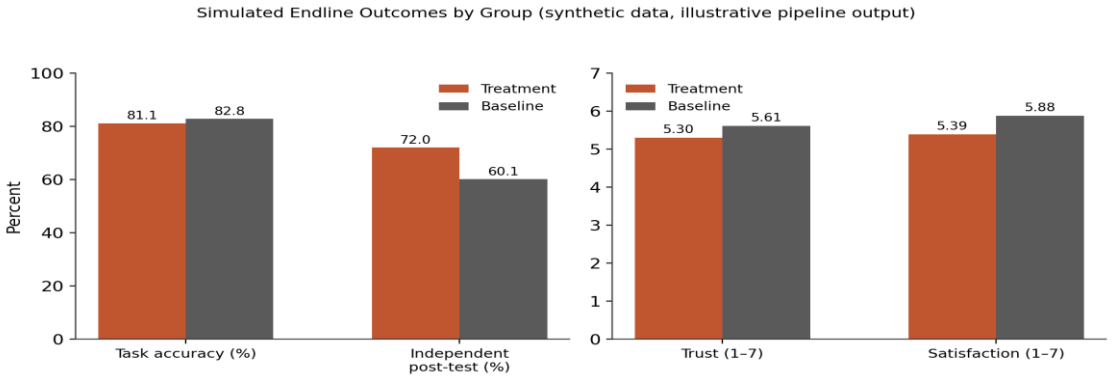


Figure 8. Simulated Endline Outcomes by Group (Task Accuracy, Independence, Trust, Satisfaction)

Figure 8 summarizes the endline comparison in two panels. The left panel shows task accuracy and independent post-test performance (percent correct); the right panel shows self-reported trust and satisfaction on the 7-point scale. Bars are group means, as printed on the bars; Table 12 gives the standard deviations and test statistics. Task accuracy is similar across groups, independent post-test performance is visibly higher under Treatment, and both trust and satisfaction are somewhat lower under Treatment.

Table 12. Simulated Endline Outcomes by Group (Mean (SD), Independent-Samples t-Test, Cohen's d)

Metric	Treatment M (SD)	Baseline M (SD)	t(88)	p	d
Task accuracy (%)	81.12 (7.68)	82.78 (6.10)	-1.14	.259	-0.24
Completion time (min)	14.32 (2.36)	13.10 (2.60)	2.33	.022	0.49
Trust (1-7)	5.30 (0.88)	5.61 (0.77)	-1.78	.078	-0.38
Satisfaction (1-7)	5.39 (0.96)	5.88 (0.85)	-2.54	.013	-0.54
Independent post-test (%)	72.01 (9.00)	60.08 (7.72)	6.75	< .001	1.42

Note. n = 45 per group; df = 88. Source: Simulated_Study_Data.xlsx (synthetic).

Domain-Level Consistency

Table 13. Simulated Reliance Index Change and Independent Post-test Performance by Domain and Group

Domain	Group	Delta RI (Wk8-Wk0)	Independent Post-test M (%)
Written Communication	Baseline	+0.022	59.21
Written Communication	Treatment	-0.143	69.87
Personal Finance	Baseline	+0.028	61.76
Personal Finance	Treatment	-0.140	73.64
Information Verification	Baseline	+0.020	59.29
Information Verification	Treatment	-0.135	72.52

Note. n = 15 per domain x group cell. Source: Simulated_Study_Data.xlsx (synthetic).

Table 13 disaggregates change in RI and independent post-test performance by task domain to assess whether the simulated pattern was an artifact of a single domain rather than a domain-general effect. Because the three domains were parameterized with comparable random-distribution settings, the simulated pattern replicates closely across domains: RI change in the Treatment condition ranged narrowly from -0.135 (Information Verification) to -0.143 (Written Communication) against a Baseline range of +0.020 to +0.028, and independent post-test gains ranged from 10.7 to 13.2 percentage points in favor of Treatment across all three domains.

DISCUSSION

Interpretation of the Simulated Findings

Because the dataset was generated from parameterized distributions that encode the study's hypotheses, this section interprets the results as evidence about how the pipeline behaves, not about how real users respond. Read in that way, the results show that each component of the framework executes coherently and that its output can be traced back to its inputs.

The Reliance Index behaved as designed. The Treatment group's mean RI declined by 0.139 points between Week 0 and Week 8, whereas the Baseline mean moved by only +0.023 (Section 4.2, Figure 6). The associated effect

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

size (d = -4.60) is extreme by the standards of behavioral research and should not be mistaken for a realistic estimate of what a real disengagement policy would achieve; it follows directly from the low simulated noise and from the fact that the Treatment group's data-generating parameters were deliberately set to trend downward. Its function here is to confirm that the scoring pipeline is sensitive enough to detect the hypothesized pattern when that pattern is present in the input data, not to forecast a real-world effect magnitude.

The subscale results (Table 10) indicate that the decline in composite RI is not an artifact of a single subscale but a coordinated shift across all four constructs the index combines: BD and ATC fell while the reverse-scored VB and PCE rose in the Treatment group. This matters for measurement, because a composite driven by one subscale would suggest that the other components add little. Whether the four factors are empirically distinct in real data is a question for the confirmatory factor analysis planned in Section 3.5.

The band mapping in Table 8 also executed correctly (Section 4.4, Figure 7). Most Treatment profiles moved from the Reflective Prompt band to the Guided Hint band, which in the simulated policy logic corresponds to progressively lighter scaffolding as reliance declines. For the Baseline arm, the band labels indicate the level the RI-based mapping would recommend given each profile's RI; Baseline participants themselves receive static, directive assistance, so their drift toward the Reflective Prompt band illustrates that a non-adaptive assistant does not respond to a user's reliance trajectory.

The endline pattern (Table 12, Figure 8) mirrors the trade-off the framework is designed to test. Task accuracy was statistically indistinguishable between groups (p = .259), so reduced reliance was not modeled as costing correctness on the underlying task. Completion time was longer under Treatment, which is consistent with the added time cost of scaffolded prompts and reflective steps relative to an immediate directive answer. Trust and satisfaction were modestly lower under Treatment: the trust difference did not reach conventional significance (p = .078) but is not negligible in size (d = -0.38), and the satisfaction difference was statistically reliable (p = .013). The direction of the simulated independence gain alongside preserved accuracy parallels the contrast reported for scaffolded versus unrestricted AI tutoring (Bastani et al., 2025), in which unrestricted access improved in-session performance but reduced later independent skill; because the simulation encodes that expectation by design, the parallel motivates the study rather than supports it. Whether real users accept the trust and satisfaction costs is an open empirical question, which is why the reward function in Table 3 includes a trust penalty.

The correlation results (Section 4.5) illustrate the logic of trajectory-based disengagement: change in RI was associated with independent post-test performance (r = -.51), whereas Week 0 RI alone was not (r = .15). In the simulation, this pattern is consistent with the paper's theoretical claim that it is the trajectory of reliance, not a user's starting level of reliance, that should drive disengagement decisions — the rationale behind the individually anchored, within-person measurement principle proposed in Section 5.3.4.

Finally, the pattern replicated across the three task domains (Table 13). Because the domains were parameterized with comparable settings, this cross-domain consistency is a designed property of the data-generating process rather than an emergent empirical finding. It confirms that the analysis pipeline reproduces the intended effect when applied separately within each domain subgroup, a necessary precondition before the same domain-stratified analysis is run on real longitudinal data, where domain differences are plausible.

Interpretive Caveats and Limitations of the Pilot

Taken together, the results in Sections 4.2–4.6 reproduce, at the level of formulas and code, the directional pattern hypothesized: a graduated, RI-modulated policy that lowers reliance, shifts users toward lighter assistance, preserves accuracy, and improves independent performance, at a modest cost to trust and completion time. Three caveats bound this reading. First, no human participant produced these numbers; the dataset was generated to match the paper's own hypotheses, so recovering the hypothesized pattern confirms the code works as intended, not that the intervention is effective. Second, simulated RI values stayed within the Guided Hint and Reflective Prompt bands (Section 4.4), so this pilot cannot speak to policy behavior at the RI extremes — the human pilot (Section 3.7) is designed to sample a wider range. Third, the very large effect size on RI change (d = -4.60) reflects the simulation's low noise and will not generalize to real data, where measurement error and unmodeled confounds will substantially attenuate every effect reported here. Section 5.3 now turns to the ethical and design implications of the architecture being tested.

Further limitations — the provisional equal-weighting prior for the RI subscales, the unfitted reward function, and possible reactivity of repeated self-report — are discussed in Section 6.

Ethical Implications and Design Principles for "Considerate AI"

The design commitments developed below are intended to govern how the pattern demonstrated in Sections 4.2–4.6 — contingent as it is on simulated rather than human data — should be operationalized once real participants are enrolled in the longitudinal study. We group these commitments under four principles, drawing on recent human-AI interaction design principles (Weisz et al., 2024) and international AI-ethics guidance (UNESCO, 2021) rather than proposing them de novo.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

- **Reversibility**

Any disengagement action taken by the system must be reversible on demand: a user who receives a Reflective Prompt or a Temporary Disengagement response retains an unconditional option to request full assistance, which the system must honor immediately. This prevents the RL policy from becoming a one-way ratchet in which a temporary dip in independent performance locks a user out of help they need. Reversibility is treated as a hard constraint on the action space in Table 3, not a soft preference the reward function merely encourages.

- **Legibility**

Whenever the assistance level changes, the system must state the reason in plain language proportionate to the user's request — for example, noting that it is offering a hint rather than a full answer because the user has asked similar questions repeatedly this week — rather than silently withholding help. Legibility is what distinguishes proactive disengagement from simple degradation of service: the user should always be able to tell why the assistant is behaving differently, and to contest that reason.

- **Stakes-Sensitivity**

The disengagement policy is scoped to the three skill-building daily-task domains studied here (written communication, personal finance, information verification) and is not intended to withhold assistance in safety-critical or high-consequence situations. Before deployment, any extension of this framework to higher-stakes domains (e.g., medical or legal information-seeking) would require a separate risk assessment, since the cost of an incorrect independent attempt scales very differently across domains.

- **Individually Anchored Measurement**

Disengagement decisions are driven by a user's own RI trajectory relative to their own baseline, not by comparison to a fixed population threshold. This design choice follows directly from the finding in Section 4.5 that change in RI, not a user's starting RI level, is what should plausibly drive the intervention: anchoring to the individual avoids penalizing someone who begins the study already dependent, or rewarding someone who begins already independent, for reasons unrelated to how the assistant is affecting them during the study period.

Together, these four principles are proposed safeguards, not tested ones; whether users experience calibrated disengagement as respectful or as a system second-guessing them is itself an empirical question the pre-registered human-subjects study in Section 3.7 is designed to answer, not one the current pipeline-validation exercise is positioned to resolve.

CONCLUSION

This paper's core move is conceptual before it is empirical: it reframes “helpfulness” from a per-interaction property into a longitudinal one, and proposes that an assistant's value should be judged partly by what it does to a user's independent capability over time, not only by what it delivers in the moment. The Reliance Index (RI) and the reinforcement-learning disengagement policy specified in Section 3 are the technical instruments built to make that reframing operational. What the paper demonstrates at this stage is narrower than that framing might suggest: Section 4 shows that the RI scoring formula, the assistance-mode mapping, and the analysis pipeline are internally coherent and execute correctly against data engineered to match the study's own hypotheses. This is a necessary but modest achievement — a working measurement instrument and a rehearsed analysis pipeline, not evidence that proactive disengagement changes real user behavior.

Three limitations bound how these results should be weighted, and each is a different kind of limitation. The synthetic dataset is a validity limitation, not an error: the t-tests, effect sizes, and correlations reported in Section 4 are computed correctly, but a dataset built to confirm a hypothesis cannot also serve as evidence for it, and the extreme effect size on RI change ($d = -4.60$) reflects deliberately low simulated noise rather than a realistic estimate of any intervention's power. The RI subscale weights and RL reward function are a methodological limitation: both are design-stage starting hypotheses, not empirically derived parameters, pending the construct-validation and reliability testing specified in Section 3.5. Finally, self-report reliance measures carry a generalizability limitation, since weekly Likert administration may itself change participant behavior independent of any assistance-level manipulation — a concern the planned human study (Section 3.7) is designed around but cannot yet resolve.

Three priorities follow directly from where this pipeline-validation stage stops. First, the equal-weighting prior on the RI subscales ($w = 0.25$ each) should be treated as a placeholder that the human pilot's confirmatory factor analysis (target $n \approx 30–50$) must either confirm or revise before the index is trusted for real decisions. Second, because the simulated data never populated the Direct Answer or Temporary Disengagement bands, the human pilot should be deliberately designed to recruit or induce a wider spread of reliance levels, since the policy's behavior at the extremes is where proactive disengagement plausibly matters most and remains completely untested. Third, the correlation between RI trajectory and independent performance ($r = -.51$ in the simulation) should be treated in real data as a hypothesis to falsify, not confirm, with the expected direction and magnitude pre-registered in advance.

This paper should be read as a pre-registration-style research program with a validated measurement scaffold, not as evidence that proactive disengagement works. Its contribution is in having built and stress-tested the machinery — the RI construct, the RL architecture, the analysis pipeline, and an ethically reasoned set of design commitments — so

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

that the human-subjects study pre-registered in Section 3.7 can be run and interpreted rigorously rather than improvised. The central question the field will actually need answered, whether helping less at the right moment leaves users more capable later, remains open. This paper's contribution is the instrument and protocol needed to answer it, not the answer itself.

ETHICS STATEMENT

No human participants were involved in this study. All results reported in Section 4 were generated from a computer-simulated dataset (N = 90 synthetic profiles, seed = 42) constructed solely to validate the Reliance Index scoring pipeline and are not derived from real user data. The human-subjects protocol pre-registered in Section 3.7 has not yet been conducted; that future study will be submitted for review and approval to the IBBI University research ethics review process prior to any participant recruitment, and informed consent procedures described in Section 3.7 will be followed at that time.

AUTHOR CONTRIBUTIONS

Wilianto: conceptualization, methodology, formal analysis, writing – original draft, writing – review and editing. Yuliana: conceptualization, validation, writing – review and editing. Albert Suwandhi: methodology, investigation, writing – review and editing. Albert Christian: data curation, software (simulation pipeline), visualization. Petrus Hemat Siregar: literature review, writing – review and editing. Khairul Amri: literature review, writing – review and editing. All authors have read and approved the final manuscript.

FUNDING

This research received no external funding. The pipeline-validation work reported here was conducted using institutional resources at the Faculty of Science and Technology, IBBI University.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The synthetic dataset analyzed in Section 4 (Simulated_Study_Data.xlsx, generated with seed = 42, including the live Reliance Index scoring formulas) is available from the corresponding author upon reasonable request. No human-subjects data exist at the time of this submission, as the pre-registered longitudinal study (Section 3.7) has not yet been conducted.

REFERENCES

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://arxiv.org/abs/2006.14779

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, *122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Cao, S., & Huang, C.-M. (2022). Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–23. https://doi.org/10.1145/3555572

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, *41*(2), 745–783. https://doi.org/10.1007/s10490-023-09871-y

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), Article 6. https://doi.org/10.3390/soc15010006

Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, *74*(9), 1477–1496. https://doi.org/10.1177/17470218211008060

Ibrahim, L., Collins, K. M., Kim, S. S. Y., Reuel, A., Lamparth, M., Feng, K., Ahmad, L., Soni, P., El Kattan, A., Stein, M., Swaroop, S., Padmakumar, V., Sucholutsky, I., Strait, A., Yang, D., Liao, Q. V., & Bhatt, U. (2025).

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9656

Measuring and mitigating overreliance to build human-compatible AI (arXiv:2509.08010). arXiv. https://arxiv.org/abs/2509.08010

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, *15*, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* (arXiv:2506.08872). arXiv. https://arxiv.org/abs/2506.08872

Lee, H.-P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1121, pp. 1–22). ACM. https://doi.org/10.1145/3706598.3713778

Liu, C., Zhou, Q., Shen, X., Liu, X. B., Wu, T., & Chen, X. A. (2026). Behavioral indicators of overreliance during interaction with conversational language models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3772318.3790332

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., & Murphy, S. A. (2022). Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. *Algorithms*, *15*(8), Article 255. https://doi.org/10.3390/a15080255

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), Article 129. https://doi.org/10.1145/3579605

Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design principles for generative AI applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 378, pp. 1–22). ACM. https://doi.org/10.1145/3613904.3642466

Yu, C., Liu, J., Nemati, S., & Yin, G. (2023). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, *55*(1), Article 5, 1–36.

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*, Article 28. https://doi.org/10.1186/s40561-024-00316-7

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.