



UGANDA CHRISTIAN
UNIVERSITY

A Centre of Excellence in the Heart of Africa

Department of Computing & Technology

*A Data Science Approach to Measuring Cognitive Offloading and
Short-Term Independent Problem Solving in AI-Assisted Tasks:
A Two-Session Experimental Study in Kampala, Uganda*

By

Richard Wambede

*Bachelors of Science in Computer sciences and Information
Technology*

Reg No: *J25M19/ 005* , Student No: *B31328*

E-mail: *B31328@students.ucu.ac.ug*, Phone No: *0782386152*

A report submitted to the Faculty of Engineering Design &
Technology for the study leading to the Degree in partial fulfillment
of the requirements for the award of **Master of Science in Data
Science and Analytics** of Uganda Christian University

September 2026

Declaration

I, *Richard Wambede*, declare to the best of my knowledge that the work in this report is original and has not been submitted before to any university/institution. It is the result of my effort.

Signature: *Richard Wambede*

Date: 16th Sept 2026

Ethical Generative AI Usage Declaration

Students may use generative artificial intelligence (AI) ethically as a “study buddy” or assistant rather than a replacement for critical thinking. This includes citing AI use and fact-checking all outputs to maintain academic integrity.

For purposes of this assignment, the following shall guidelines shall apply.

- i. Generative AI may be used for brainstorming, and generating ideas to improve your work. You may also use AI for literature search, and summarizing journal papers for quick analysis and assessing relevance of publications. However, once the relevant publications have been identified, a student is expected to fully review the selected publications, identifying problems addressed, methods used, gaps identified and recommendations made.
- ii. Where AI has been used as guided in (i) above, full disclosure of how it has been used should be made in the appendix **B**, including how AI was used, the AI tools (including mode versions), and a compilation on major AI prompts used.
- iii. NO GENERATIVE AI CONTENT IS ALLOWED IN THE FINAL SUBMISSION. The final submission SHOULD be your original work.

A. Declaring the non-use of AI for content generation

I, *Richard Wambede*, declare that no content generated by AI technologies has been used in this report.

Signature: *Richard Wambede* Date: 16th Sept 2026

B. Acknowledging the use of AI as a study buddy

Please refer to Appendix **B** for the complete declaration of AI usage for generating ideas, brainstorming, literature search, and summarizing only to improve your work.

Approval

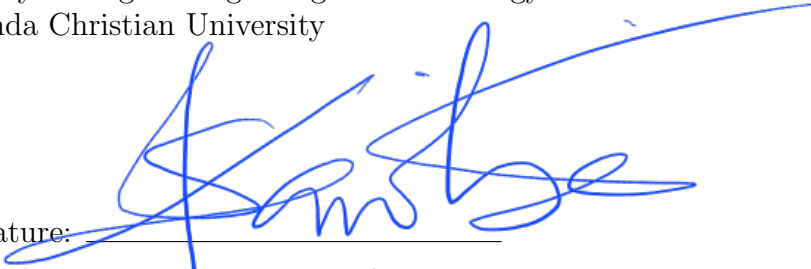
This report has been submitted for examination with approval of the following advisors/supervisor.

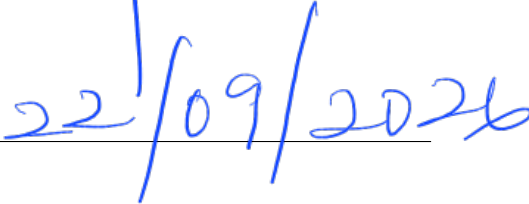
Advisor: ***Dr. Samuel Kakuba***

Department of Computing & Technology

Faculty of Engineering Design & Technology

Uganda Christian University

Signature: 

Date: 

Abstract

Generative AI tools can improve speed and accuracy in problem solving but can also change how people think by replacing human reasoning steps with model-generated ones. This study asked one central question: how does growing reliance on AI for thinking tasks affect human problem solving, and does this reliance risk weakening essential cognitive skills?

A total of 1,474 participants were recruited for data collection in Kampala, Uganda. Of these, 1,387 completed both sessions and formed the final analysis sample, giving a retention rate of 94.1 percent. The study used a two-group, two-session design. In Session 1, the AI-assisted group solved structured tasks using a controlled hint bank while the control group worked without assistance. In Session 2, AI was removed for everyone. The delay between sessions had a median of 6.6 minutes and ranged from 5 to 11 minutes.

Two auditable composite indices were built from behavioural logs. The Cognitive Offloading Index (COI) measured how much of the thinking was shifted to AI, using four behaviours recorded by the platform: how often AI was asked for help, how long each task took, how much AI output was copied into the answer and how often the output was checked.

The Human Engagement Score (HES) measured how actively participants stayed involved, using self-reported engagement, the independence of the reasoning they produced and the share of tasks completed without AI. Both were computed using fixed z-score formulas and checked for internal stability and construct separation. They are reproducible descriptive indices rather than validated psychometric instruments.

In Session 1, both groups achieved almost identical accuracy (Cohen's $d = 0.026$). However the control group scored substantially higher on rubric independence ($d = -1.972$), meaning the AI-assisted group produced weaker independent reasoning despite getting the same answers. In Session 2 with AI removed, both primary endpoints showed negligible differences. Accuracy S2 ($d = 0.028$) and rubric independence S2 ($d = 0.026$) were both non-significant with confidence intervals crossing zero. An exploratory Ridge regression model did not beat a naive baseline. Session 1 features could not estimate individual Session 2 independence under the conditions of this study, where the outcome sat close to its ceiling. All four robustness checks confirmed these conclusions.

The study concludes that AI use during Session 1 changed how participants worked during the tasks but did not reduce what they could do on their own seven minutes later. No drop in short-term independent performance was found for these tasks and this context. The design did not test long-term skill retention. Nothing in these results speaks to what happens after a day, a week or a semester of regular AI use. The COI and HES scoring rules, cleaned dataset and reproducible analysis pipeline are provided as reusable contributions for future research.

Contents

Declaration	i
Declaration	ii
Approval	iii
Abstract	iv
Dedication	iv
Acknowledgments	x
1 Introduction	1
1.1 Background	2
1.2 Problem Statement	2
1.3 Research Objectives	2
1.3.1 Specific Objective 1	2
1.3.2 Specific Objective 2	3
1.3.3 Specific Objective 3	3
1.4 Research Questions	3
1.4.1 Research Question 1	3
1.4.2 Research Question 2	3
1.4.3 Research Question 3	3
1.5 Significance of the Study	3
1.6 Scope	4
1.7 Limitations	4
1.8 Key Definitions	4
2 Literature Review	7
2.1 Introduction	7
2.2 Theoretical Background	7
2.3 Cognitive Offloading and Human Thinking	8
2.4 Generative AI and Problem Solving	9
2.5 Human-AI Interaction and Trust	9
2.6 Measurement of AI Reliance and Engagement	10
2.7 Comparative Analysis of Existing Studies	11
2.8 Critical Analysis of Existing Studies	12
2.9 Research Gaps	12
2.10 Conceptual Framework	13
2.11 Summary	14

3	Methodology	15
3.1	Introduction	15
3.1.1	Where COI Sits in This Study	16
3.2	Research Philosophy and Theoretical Foundation	17
3.3	Research Design	17
3.3.1	AI Assistance Delivery	18
3.4	Study Setting and Participants	19
3.4.1	Target Population	19
3.4.2	Sampling Method	19
3.4.3	Data Collection Sites	19
3.4.4	Inclusion and Exclusion Criteria	20
3.4.5	Sample Size and Retention	20
3.4.6	Justification of the Sample Size	20
3.4.7	Group Assignment	21
3.5	Experimental Procedure	21
3.5.1	Session Structure	21
3.5.2	Conduct of the Experiment	22
3.5.3	Task Types and Families	22
3.5.4	AI Assistance Protocol	22
3.6	Variables and Measurement	23
3.6.1	Performance Variables	23
3.6.2	Origin and Theoretical Basis of the Indices	23
3.6.3	COI Components and Formula	24
3.6.4	HES Components and Formula	25
3.6.5	Scale Items	25
3.7	Data Processing	25
3.7.1	Integration	26
3.7.2	Missing Value Treatment	26
3.7.3	Outlier Treatment	26
3.7.4	Additional Preprocessing	26
3.8	Data Collection and Analysis Tools	27
3.8.1	Data Collection Platform	27
3.8.2	Database	27
3.8.3	Analysis Tools	27
3.9	Analytic Approach	28
3.10	Statistical Analysis Plan	29
3.10.1	Group Comparisons	29
3.10.2	Adjusted Regression	30
3.10.3	Predictive Modelling	31
3.10.3.1	Leakage Controls	31
3.10.3.2	Choice and Specification of the Model	32
3.10.4	Robustness Checks	33
3.11	Reliability Checks	33
3.12	Bias Control and Data Validation	34
3.13	Ethical Considerations	36
4	Analysis and Results	37
4.1	Introduction	37
4.2	Data Integration and Sample Description	37
4.2.1	Data Files Loaded	37

4.2.2	Completers and Exclusions	38
4.2.3	Dataset Integration	39
4.2.4	Session Delay	39
4.2.5	Group and Site Balance	39
4.3	Exploratory Data Analysis	40
4.3.1	Key Variable Distributions	40
4.3.2	Task Accuracy by Task	41
4.3.3	Time Per Task	41
4.3.4	Hint and Telemetry Summary	42
4.4	Outlier Detection	43
4.5	Data Preprocessing	44
4.5.1	Missing Value Treatment	44
4.5.2	Additional Preprocessing	45
4.6	Feature Engineering - COI and HES	46
4.6.1	COI Construction	46
4.6.2	HES Construction	48
4.6.3	COI and HES Validation	49
4.6.4	Validation Status of the Two Indices	50
4.7	Reliability Checks	51
4.7.1	Scale Reliability - Cronbach's Alpha	51
4.7.2	Rubric Reliability - ICC	52
4.7.3	Inter-Component Correlations	52
4.8	Statistical Analysis	53
4.8.1	Research Question 1 - Session 1 Results	53
4.8.2	Research Question 2 - Session 2 Results	54
4.8.3	Adjusted Regression Results	55
4.9	Predictive Modelling - Research Question 3	56
4.9.1	Model and Feature Set	56
4.9.2	Cross-Validation Results	56
4.9.3	Why a More Flexible Model Would Not Change This Result	56
4.9.4	Feature Coefficients	57
4.9.5	Predicted vs Actual	58
4.9.6	Secondary Target - Accuracy S2	59
4.9.7	Subgroup Error Diagnostics	59
4.10	Robustness Checks	60
4.11	Findings Dashboard	60
4.12	Summary of Findings	62
5	Discussion	65
5.1	Introduction	65
5.2	Why the Two Groups Reached the Same Answers by Different Routes	65
5.3	Why the Independence Gap Closed So Completely in Session 2	67
5.4	Why Session 1 Behaviour Could Not Predict Session 2 Independence	68
5.5	What the Self-Report and Behavioural Measures Disagreeing Tells Us	70
5.6	What These Findings Mean for AI Use in a Ugandan Higher Education Setting	70
5.7	What This Study Can and Cannot Conclude	71
5.7.1	Boundaries of Generalisation	72
5.8	Reliability Limitations and Their Impact	72
5.9	Practical Implications	75
5.10	Limitations	76

6 Conclusion and Recommendations	78
6.1 Introduction	78
6.2 Answers to the Research Questions	78
6.2.1 Research Question 1	78
6.2.2 Research Question 2	79
6.2.3 Research Question 3	79
6.3 Answer to the Central Question	80
6.4 Recommendations	80
6.4.1 For Educators and Institutions	80
6.4.2 For Organisations and Workplaces	81
6.4.3 For Future Research	81
6.5 Contribution of the Study	82
6.6 Final Reflection	82
REFERENCES	86
APPENDICES	86
A Research Project timelines	87
B Declaration of AI use	89
C List of 20 Major Journal Publications reviewed	91
C.0.1 Review Papers and Meta-Analyses	91
C.0.2 Journal Articles	91
C.0.3 Conference Papers	92
C.0.4 Other Relevant Publications	92
D Research Alignment with the MSc Programme, National Development Agenda and Global Frameworks	93
D.1 Alignment with the MSc Programme in Data Science	93
D.2 Alignment with the Fourth National Development Plan (NDPIV)	93
D.3 Alignment with Uganda Vision 2040	94
D.4 Alignment with the Digital Uganda Vision	94
D.5 Alignment with the Sustainable Development Goals	94
D.6 Alignment with Responsible AI and Ethics	94
D.7 Alignment with Global Digital Transformation Trends	95
E Data Dictionary	96
E.1 Participants File	96
E.2 Task Responses File	97
E.3 Telemetry Events File	98
E.4 Hint Events File	99
E.5 Scale Responses File	99
E.6 Rater Scores File	100
E.7 Session Meta File	100
E.8 Deviation Logs File	101
E.9 Layer 2 Logs File	101
E.10 Derived Variables in Table A	101
F COI and HES Scoring Specification	104
F.1 General Principles	104

F.2 Cognitive Offloading Index (COI)	104
F.2.1 Applicability	104
F.2.2 Raw Components	105
F.2.3 Outlier Rules Before Standardisation	105
F.2.4 Reversal Rule	106
F.2.5 COI Formula	106
F.2.6 Interpretation	106
F.2.7 Post-Computation Checks	106
F.3 Human Engagement Score (HES)	106
F.3.1 Applicability	106
F.3.2 Raw Components	107
F.3.3 Control Group Rule for inv_AI_Ratio	107
F.3.4 HES Formula	107
F.3.5 Interpretation	108
F.3.6 Post-Computation Checks	108
F.4 Summary Table	108
F.5 Reproducibility Notes	109
G Rubric Scoring Guide	110
G.1 General Scoring Instructions	110
G.2 Dimension 1 - Correctness	110
G.3 Dimension 2 - Justification Quality	111
G.4 Dimension 3 - Independence	111
G.5 Special Cases and Edge Rules	112
G.6 Inter-Rater Reliability Results	113
H Analysis Pipeline Overview	114
H.0.1 Pipeline Overview	114
I Other useful figures	117
J Research poster layout	119
K Research Ethics Committee (REC) Approval	121
L Data collection tools	123
M Student - supervisor meeting/supervision plan	127

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Dr. Francis Otto and Mr. John Bosco Wabwire, for their consistent guidance, constructive feedback and patience throughout this research journey. Their direction shaped both the proposal and the final report significantly.

Special thanks go to all participants who gave their time to take part in the study. This includes students from Uganda Christian University, Mbarara University of Science and Technology, Makerere University and Kyambogo University, as well as working adults and other individuals who completed the study through the online platform. Without their willingness to participate, this research would not have been possible. I also acknowledge the financial cost involved in encouraging participation during data collection and appreciate everyone who showed up and gave their full attention despite their busy schedules.

I am grateful to the Department of Computing and Technology in the Faculty of Engineering, Design and Technology at Uganda Christian University for the additional comments, institutional support and academic environment that made this work possible.

Finally, I would like to thank the Research Ethics Committee for reviewing and approving this study, and to everyone who contributed in any way, however small to the completion of this research.

Chapter 1

Introduction

Executive Summary

Artificial intelligence is changing how people solve problems. Tools like ChatGPT can now produce reasoning steps, draft answers and complete tasks that people would normally do themselves. This raises a practical question: if people regularly hand over their thinking to AI, what happens to their ability to think independently when AI is not there?

This study built and tested a data science framework to measure this question directly. A total of 1,474 participants were recruited in Kampala, Uganda. They were split into two groups. The AI-assisted group used a structured hint system during Session 1 while the control group worked alone. In Session 2, AI was removed for everyone. Two behavioural indices were constructed from activity logs. The Cognitive Offloading Index (COI) measured how much thinking was shifted to AI. The Human Engagement Score (HES) measured how actively participants stayed involved in the tasks.

The results showed that both groups achieved similar accuracy in both sessions. However the control group produced much better independent reasoning in Session 1, with a very large effect size of -1.972 on rubric independence. When AI was removed in Session 2, both groups performed at the same level. No drop in short-term independent performance was found. The prediction model also found that Session 1 reliance behaviour could not reliably predict Session 2 independence at the individual level. These findings are discussed alongside the limitations of the same-sitting design and the reliability issues found in the measurement tools.

1.1 Background

Many recent discussions about AI stop at one simple question: does AI help people finish tasks faster? That question is not enough for a data science research degree. The more useful scientific question is how AI changes the way people think, and what that change does to their ability to work independently later.

When a person uses AI to break down a problem, generate intermediate steps or draft reasoning they would normally do themselves, that behaviour leaves clear traces. These include how often they asked for help, how long they spent reading AI output, whether they copied it and whether they checked it. Those traces can be turned into measurable features. Features can be combined into indices like COI and HES. Indices can then be connected to later performance to test whether AI use changed independent ability.

The practical motivation is also clear. In education and workplace settings, AI may raise performance today while quietly reducing independent capability tomorrow, if the person stops retrieving information, elaborating ideas and checking answers on their own. This study separated two outcomes. The first was immediate performance during AI access in Session 1. The second was delayed independent performance in Session 2 after AI was removed.

This design gave a direct and measurable test of whether reliance acted as augmentation, where the person still reasoned and verified, or as substitution, where the tool performed the reasoning and the person simply accepted it [1].

1.2 Problem Statement

Current generative AI use shifts problem solving from human-generated intermediate reasoning to AI-generated steps [1, 2, 3]. However, there was no lightweight, reproducible measurement and modelling pipeline that could do three things together. First, quantify offloading intensity as observable behaviour. Second, measure human engagement and verification behaviour during AI-assisted work [4, 5]. Third, estimate the effect of AI access on delayed independent problem solving when AI was removed [6, 7].

The technical problem was to build an experimental dataset and analytical framework that could estimate and predict how Session 1 reliance behaviour, COI, HES and verification, impacted Session 2 independent performance. The goal was to separate measurable augmentation, where AI support improved performance without reducing independent capability [1], from measurable substitution, where higher reliance led to reduced independent performance after AI was removed [2, 8].

This distinction matters because it enables early detection of overreliance risk using data rather than assumptions [9, 10].

1.3 Research Objectives

This study aimed to understand how people depend on AI during problem solving, to measure that behaviour with data science methods, and to test whether the resulting measures can say anything about an individual's later independent performance [2, 8].

1.3.1 Specific Objective 1

To construct a labelled two-session dataset from a sample of university students and working adults in Kampala, engineer reliance and verification features from their behavioural logs, and

compute COI and HES as auditable composite indices.

1.3.2 Specific Objective 2

To evaluate the effect of Session 1 AI reliance on delayed independent performance in Session 2 when AI was removed, using effect sizes, confidence intervals and reliability checks [4, 5].

1.3.3 Specific Objective 3

To test whether Session 1 reliance behaviour alone can be used to estimate the Session 2 independence of an individual participant, and so whether COI could work as an early screening measure rather than only as a group-level description.

1.4 Research Questions

1.4.1 Research Question 1

How was a good dataset constructed, useful features engineered from raw behavioural logs, and COI and HES computed in a way that clearly showed AI reliance and verification behaviour, and how reliably did the two indices measure what they claim to measure [3]?

1.4.2 Research Question 2

How did Session 1 AI reliance affect delayed independent performance in Session 2 when AI was removed [6, 7]?

1.4.3 Research Question 3

Can Session 1 behavioural features alone estimate the Session 2 independent performance of an individual participant, and do the conditions of this study allow that question to be tested fairly?

1.5 Significance of the Study

This study turned a common worry into measurable data. Many people argue that AI can make people think less, but there was little proof from controlled studies with delayed testing. This study measured AI use through behaviour logs and the COI, and measured engagement through the HES. Both scores were built from observable actions so that behaviour could be compared fairly across different participants [2].

The study also tested what happened when AI was removed. Most existing studies only checked speed and accuracy while AI was present [11]. This study removed AI in Session 2 and measured independent performance directly.

Finally, the study provided an early look at whether high COI scores could work as a warning signal for overreliance risk. The idea is supported by research on AI trust and overreliance [4] and connects to responsible AI use in education and decision making [10].

1.6 Scope

This study focused on how people used AI during problem solving and what happened when AI was later removed. It used a two-session design with a same-sitting delay of approximately seven minutes between sessions. It did not measure long-term learning over weeks or months.

The tasks were limited to structured problem-solving activities covering data reasoning, logic and quantitative reasoning. Results may not fully apply to all real-world settings or task types.

The study measured behaviour using activity logs converted into features and combined into COI and HES. It also built a simple prediction model to link Session 1 behaviour to Session 2 outcomes. It did not focus on emotional, social or clinical aspects of cognition. The aim was to provide a clear and reproducible data science framework for analysing AI reliance [1, 2].

1.7 Limitations

Several limitations affected this study and must be noted when interpreting the findings.

- The delay between Session 1 and Session 2 was about seven minutes. This short gap means the study measured very short-term independence only. It cannot show what would happen to independent performance after a longer gap such as a day or a week.
- Task accuracy was very high for almost all participants, creating a ceiling effect. This reduced the variation available to detect differences between groups and limited the predictive model.
- All four Likert scales failed Cronbach’s alpha reliability checks, meaning the scale items within each scale were essentially uncorrelated. Scale-based scores were treated as descriptive only.
- Two of three rubric dimensions failed inter-rater agreement. Only correctness reached an acceptable ICC. Justification quality and independence scores must be interpreted with caution.
- The COI prediction model only covered the AI-assisted group because the control group had no reliance data.
- The convenience sample was drawn from two sites in Kampala and may not represent other populations, cultures or task types.
- Data collection was conducted under time pressure. At the MAIN site, participants were brought in batches to a computer lab because universities were on holiday and many people were doing internships. Incentives were provided to encourage participation. These practical constraints may have introduced variability in participant motivation and focus.

1.8 Key Definitions

- **Artificial Intelligence (AI):** AI refers to computer systems that can produce text, reasoning steps and solutions to tasks. In this study, AI acts as an external support tool during problem solving [11].
- **Cognitive Offloading:** Cognitive offloading is when a person uses an external tool to reduce their mental effort. In this study, it means shifting reasoning or memory tasks from the person to the AI system [6, 7].

- **Human Engagement:** Human engagement is the level of active thinking, checking and reasoning a person does during a task. It includes behaviours like verifying answers, explaining steps and revising solutions.
- **Cognitive Offloading Index (COI):** COI is a score that measures how much a person relied on AI during problem solving. It is built from four behaviours recorded in the platform logs: AI use ratio, Session 1 task time, copy rate and low verification. Each behaviour is converted to a z-score and the four are averaged, so COI is centred at zero and is read as a distance from the group average rather than as a percentage. A participant with a COI of +1.0 relied on AI about one standard deviation more than the average participant in the AI-assisted group.
- **Human Engagement Score (HES):** HES is a score that measures how actively a person participates in problem solving. It is built from engagement self-report, justification quality and low delegation to AI. Higher values indicate more independent thinking and verification.
- **Augmentation:** Augmentation means AI supports the person without reducing their ability to think on their own. The person still checks and understands the solution [1].
- **Substitution:** Substitution means AI replaces the person’s thinking process. The person depends on AI outputs without much verification or understanding of the steps involved.
- **Delayed Independent Performance:** How well a person performs tasks after AI is removed. In this study the delay was a few minutes, so this term always means short-term independent performance and never long-term skill retention. It is measured in Session 2 using task accuracy and rubric scored independence and justification quality.
- **Overreliance:** Overreliance happens when a person depends too much on AI and reduces their own thinking effort. This can lead to weaker performance when AI is not available [4, 5].
- **Verification:** Verification is when a person actively checks an AI-supplied answer using their own reasoning, re-computation or an alternative method. It is one of the key behaviours measured in this study.
- **Behavioural Feature:** A behavioural feature is a measurable variable extracted from raw activity logs. Examples in this study include prompt count, time in tool, copy rate and verification actions. These are the building blocks of COI and HES.
- **Session 1:** Session 1 is the first problem solving session. The control group works without AI. The AI-assisted group works with AI under a controlled protocol. Data collected here is used to compute COI and HES.
- **Session 2:** Session 2 is the delayed independence test. AI is removed for everyone. Participants solve a new but similar set of tasks on their own. This session measures whether prior AI use affected independent problem solving.
- **Feature Engineering:** Feature engineering is the process of turning raw log data into meaningful measurable variables. In this study, raw logs like prompt counts and time stamps are transformed into features that feed into COI and HES [1].
- **Prediction Model:** A prediction model is a statistical or machine learning tool that uses known inputs to estimate an unknown output. In this study, Session 1 features are used to predict Session 2 independence risk [8].
- **Effect Size:** Effect size is a number that shows how big a difference is between two groups. In this study, Cohen’s d is used to report the size of differences between the AI group and the control group [12].

- **Confidence Interval:** A confidence interval is a range of values that is likely to contain the true result. This study uses 95 percent confidence intervals to show the uncertainty around each estimate.
- **Cross-validation:** Cross-validation is a method for testing how well a prediction model works on new data. It splits the dataset into parts, trains the model on some parts and tests it on the rest. This study uses grouped cross-validation by participant ID to avoid data leakage [8].
- **Data Leakage:** Data leakage happens when information from the test set accidentally enters the training set, making a model look better than it really is. This study prevents leakage by using only Session 1 features to predict Session 2 outcomes.
- **Ceiling Effect:** A ceiling effect happens when most participants score at or very near the maximum possible value on a measure. When this occurs, there is very little variation left to detect differences between groups or to use in prediction models. In this study, almost all participants scored 1.0 on accuracy in both sessions, which made it difficult to find meaningful differences between the AI-assisted and control groups on this measure.
- **Short-term independent performance:** How well a person performs a comparable task on their own, a few minutes after AI support has been taken away. This is what Session 2 measured in this study.
- **Long-term skill retention:** Whether a person still holds a skill after days, weeks or longer, and can still use it without help once that time has passed. This study did not measure this and reports nothing about it.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews existing research on generative AI, human thinking and problem solving. It focuses on how AI changes the way people engage with tasks, especially through cognitive offloading and verification behaviour. The goal is to understand what was already known before this study, what was missing, and how this study filled those gaps.

Recent studies show that generative AI is not just a tool for finding information. It actively changes how people think, solve problems and make decisions [1]. This makes it important to study both performance and behaviour, not just final task outcomes.

The chapter is organised into five areas. First, theoretical background. Second, cognitive offloading and human thinking. Third, generative AI and problem solving. Fourth, human-AI interaction and trust. Fifth, measurement of AI reliance and engagement. The chapter ends with a critical analysis and a summary of the three research gaps this study addressed.

2.2 Theoretical Background

This study was grounded in three overlapping theories: cognitive offloading, cognitive load and human-AI collaboration.

Cognitive offloading is the use of external tools to reduce mental effort. In the past, this meant writing notes or using a calculator. Modern AI systems go further by performing reasoning steps, not just storing information [3, 7]. This changes what a person actually practises during problem solving.

Cognitive load theory explains how reducing unnecessary mental effort can improve performance. However, removing too much effort can also reduce learning, especially when key steps like retrieval and self-explanation are skipped [13]. The challenge is finding the right balance between reducing unhelpful effort and keeping essential thinking steps active.

Human-AI collaboration is often described as hybrid intelligence, where humans and AI work together as a system. In this model, AI supports performance, but humans must still stay active in reasoning and verification [1]. When humans stop doing essential evaluation steps, the hybrid system becomes less safe, especially in high-stakes settings.

These three theories together explain why this study measured both offloading and engagement. Offloading alone does not tell the full story. What matters is whether the person stayed actively involved while using AI, or whether they simply accepted whatever the AI produced.

2.3 Cognitive Offloading and Human Thinking

Cognitive offloading has been studied widely in psychology. Research shows that offloading can improve immediate performance but may reduce memory and understanding over time [6, 7].

Modern AI increases the level of offloading by allowing users to delegate not only memory tasks but also reasoning steps. This changes what users practise during problem solving and may affect long-term independent capability [3].

Other researchers have studied this same tension. Risko and Gilbert [14] reviewed how people decide when to offload and showed that offloading is a strategic choice shaped by task difficulty and confidence, not an automatic reflex. Sparrow, Liu and Wegner [15] demonstrated in their well-known "Google effect" study that when people expect to have future access to information, they remember the information itself less well but remember where to find it. This directly supports the concern in the present study, that easy access to AI may reduce the mental work a person does themselves. Dunn and Risko [16] extended this by showing that the metacognitive decision to offload is often poorly calibrated, meaning people offload even when doing the task themselves would have been better for learning. The present study builds on this line of work by measuring offloading behaviour directly from activity logs rather than relying on what participants report.

Offloading is not always harmful. It depends on how it is used. If users return to the information, explain it in their own words or verify it, learning can still happen. But if users simply accept AI outputs without checking, learning is likely to weaken [6].

However, the literature mostly discussed offloading at a general level. It did not clearly identify which specific behaviours separated a person who was still thinking actively, known as augmentation, from a person who had handed all thinking to AI, known as substitution. This study addressed that gap directly by measuring four specific behavioural features during problem solving.

The first feature was AI use ratio. This was the proportion of tasks where the person requested AI assistance. A person who asked for help on every task showed stronger offloading than one who asked only occasionally [5].

The second feature was time on task. This measured the average time a participant spent on each Session 1 task. A person who spent longer with the tool open was likely to be working through more of the problem alongside it. A person spending most of their time in the tool was likely delegating more of the thinking process [6].

The third feature was copy rate. This was the proportion of tasks on which the participant copied hint content into their answer. A high copy rate was a strong signal of substitution, because the person was reproducing the AI's reasoning rather than producing their own [4].

The fourth feature was verification rate. This measured how often a person checked an AI-supplied answer using their own method, such as re-computing, estimating or using an alternative approach. Low verification rate meant the person accepted AI outputs without questioning them [4, 5].

Together these four features formed the building blocks of COI. A person with high AI use ratio, high time in tool, high copy rate and low verification rate showed clear substitution behaviour. A person with low AI use ratio, low copy rate and high verification rate showed augmentation behaviour, even if they used AI at some point [1].

2.4 Generative AI and Problem Solving

Generative AI tools such as ChatGPT have been shown to improve task performance in many settings. Studies report higher accuracy and faster completion times when AI is used [11].

However, these benefits often come with reduced mental effort. Users may shift from active problem solving to passive acceptance of AI-generated answers. This creates a trade-off between performance now and understanding later.

Research also shows that doing tasks quickly with AI can give a false sense of learning. Completing a task fast does not always mean the person understands the process behind it. This is why delayed testing was important in this study. Session 2 removed AI and tested whether the person could still solve similar problems on their own [17].

Evidence from education showed a mixed picture. AI supported activities can improve performance while also reducing perceived effort. This may push people toward convenience-driven strategies if reflection is not built into the design [11].

Research on creativity added another concern. Generative AI can increase how much a person produces while reducing the variety of ideas across a group [18]. When AI becomes the default partner, solutions may start to look similar across users, which may signal reduced exploration and weaker independent reasoning over time.

The rapid growth of large language models has produced a wave of new studies on this question. Kosmyna et al. [19] used brain-activity measurement to show that people writing essays with a large language model displayed lower neural engagement and weaker memory of their own essays compared to people who wrote unaided, which they described as a possible "cognitive debt". Bastani et al. [20] ran a large field experiment in mathematics classrooms and found that students who practised with an unrestricted AI tutor performed worse on later unaided tests than students who never used it, although a carefully designed AI tutor removed this harm. Lee et al. [21] surveyed knowledge workers and found that higher confidence in AI was linked to less critical thinking effort during tasks. These recent studies show a consistent pattern: AI can raise immediate output while quietly lowering the independent effort a person invests, which is exactly the pattern the present study set out to measure in a controlled way.

2.5 Human-AI Interaction and Trust

Human-AI interaction research shows that trust is central but not enough on its own. Users must decide when to trust AI and when to check its output. However, research shows that trust attitudes alone do not reliably produce calibrated reliance, especially when users form simple habits like always following or always ignoring AI suggestions [5, 22].

Controlled experiments showed that standard explanations from AI systems did not reliably prevent overreliance when the AI was wrong, unless the interface forced deeper checking behaviour [4]. This reveals a design trade-off. Adding friction can reduce blind acceptance, but too much friction can reduce usability and push users to bypass safeguards altogether.

Research on AI-generated content added a further concern. Experiments showed that people can struggle to tell AI-generated text from human text, and that synthetic content can be persuasive even when it is wrong [23]. This strengthens the case for verifiable checking routines rather than trust alone, which aligns with the emphasis on accountability and risk management in responsible AI guidance [9, 10].

Design solutions such as forcing users to explain answers or check results before submitting can reduce overreliance. However, too much control can reduce usability, so a balance is needed

[4].

2.6 Measurement of AI Reliance and Engagement

A key challenge in the literature was how to measure AI reliance. Many studies relied on self-reported trust or attitudes, which did not always reflect actual behaviour [5]. A person might say they checked AI outputs carefully, but their logs could show something different.

Recent research suggested that behavioural measures are more reliable than self-report alone. These include actions such as verification, copying, time spent in the tool and prompt usage [2, 8]. However, most studies measured these factors separately. There was no single combined index that captured both reliance and engagement together in one auditable framework.

This study addressed that gap by combining four specific behavioural features into COI and three engagement signals into HES. COI and HES were kept as separate indices because they measured different things. A person could have high engagement and still offload some steps to AI. These were not the same thing and should not be combined into one score [8].

Recent evidence linked higher AI dependence to weaker critical thinking and fatigue-related mechanisms, while AI literacy could reduce some of these negative effects [2, 8]. However, a weakness in many of these findings was that they were correlational. They could not show that AI use caused lower thinking. This was why this study used a controlled design with two groups and a delayed no-AI session, where reliance was manipulated and independent performance was measured after AI was removed.

In relation to these earlier works, the present study takes a different and complementary position. Where Gerlich [2] and Tian and Zhang [8] measured reliance through self-report surveys, this study measured it directly from behavioural logs. Where Bastani et al. [20] and Kosmyna et al. [19] used a delayed unaided test, this study adopted the same logic of a no-AI second session but combined it with two auditable composite indices, the Cognitive Offloading Index and the Human Engagement Score. In this way the present study connects the behavioural-measurement tradition of Risko and Gilbert [14] with the delayed-testing tradition of Bastani et al. [20], bringing the two approaches together into one reproducible framework.

2.7 Comparative Analysis of Existing Studies

Study	Method	Dataset	Key finding	Limitation
Grinschgl et al. (2021)	Controlled experiment	Memory tasks	Better task performance	Lower memory retention
Buçinca et al. (2021)	Experiment on AI decision tasks	Lab participants	Less overreliance with forcing functions	Limited real-world validity
Deng et al. (2024)	Meta-analysis of 50+ studies	Education datasets	Better performance, less effort	Mixed results across contexts
Gerlich (2025)	Survey and analysis	Adults using AI tools	Higher AI use linked to lower critical thinking	Correlational, no delayed test
Tian & Zhang (2025)	Survey study	University students	AI dependence linked to fatigue and weaker thinking	Self-reported data bias
Doshi & Hauser (2024)	Controlled experiment	Online workers	More individual output but less idea diversity	Lab context only
Spitale et al. (2023)	Experiment	Online participants	People struggle to detect AI-written content	Context specific results
Mehrotra et al. (2024)	Systematic review	Human-AI interaction studies	Trust attitudes do not reliably predict actual reliance behaviour	No behavioural measurement

Table 2.1: Comparison of selected studies on AI use, cognitive offloading and learning outcomes.

The table shows that most studies agreed AI can improve speed and performance in the short term. However, there was much less agreement on its effect on learning and independent thinking. Many studies also lacked behavioural measurement and delayed testing, which were both important for understanding what happened after AI was removed. This study directly addressed both of these gaps.

2.8 Critical Analysis of Existing Studies

A key limitation in current research was the lack of behavioural measurement. Many studies focused on trust or perception rather than actual user actions during problem solving [5]. Asking someone whether they trusted AI was very different from measuring what they actually did when AI gave them an answer.

Specifically, most studies did not measure the four behavioural features that this study argued were central to detecting substitution versus augmentation. These were AI use ratio, time in tool, copy rate and verification rate. Without these features, it was impossible to say whether a person used AI as a support tool or let AI do all the thinking for them [4, 5].

Another limitation was the focus on immediate performance. Few studies tested what happened after AI was removed. This made it difficult to measure true learning or independence [6, 7]. A person could perform well during a task with AI and then fail the same type of task one day later without AI. Without a delayed no-AI session, this risk remained invisible.

There was also a lack of integrated measurement. Most studies measured offloading, engagement or performance separately rather than combining them into linked indices. This made it hard to see the full picture of how AI changed the thinking process [8]. A person could be highly engaged and still offload heavily. These two things were not opposites, and measuring only one of them gave an incomplete picture.

Finally, many findings were based on correlation, which did not show causation. Controlled experiments with delayed testing and a no-AI condition were needed to make stronger claims about the effects of AI use on independent problem solving [2, 11].

2.9 Research Gaps

Based on the reviewed literature, three main gaps were identified. Each gap directly motivated one part of this study.

Gap 1: Measurement Gap

Most existing studies did not provide an auditable, feature-based index that converted AI reliance into measurable variables. They either used self-report alone or measured only one behaviour at a time. Specifically, no existing study combined AI use ratio, time in tool, copy rate and verification rate into a single composite index that could separate substitution from augmentation [4, 5].

This study addressed this gap by defining a fixed feature set for reliance and verification, then computing COI and HES using deterministic formulas backed by a data dictionary. This meant another researcher could take the same raw logs and reproduce the exact same scores [1].

Gap 2: Delayed Endpoint Gap

Most studies only measured performance while AI was present. They did not test what happened after AI was removed. This meant the risk to short-term independent performance was never directly observed, only assumed [6, 7].

This study addressed this gap by using Session 2 with AI removed as the primary independence endpoint. If the AI-assisted group performed worse in Session 2 than the control group, that would be direct evidence of a short-term independence cost. If they performed similarly,

that would also be useful evidence showing no measurable short-term cost for these tasks and contexts [11].

Gap 3: Validity Gap

Many studies reported results without providing reliability evidence for their scoring tools. Rubric scores and composite indices could vary a lot depending on who scored them and how the formula was defined [2, 8].

This study addressed this gap by using a written rubric with clear scoring dimensions, double scoring at least 20 percent of responses, and computing Cronbach’s alpha for multi-item scales and ICC for rubric scoring. This provided evidence about whether the measures used were consistent and trustworthy [4].

2.10 Conceptual Framework

The three theories reviewed above and the three gaps identified from them come together in the framework shown in Figure 2.1. The framework sets out what this study treats as the cause, what it treats as the mechanism in between, what it treats as the outcome and what it holds constant.

The condition. The single thing the study changed was whether structured AI assistance was available during Session 1. Everything else was held the same across the two groups.

The mechanism. Cognitive load theory [24] and the cognitive offloading literature [14, 15] both say that an external tool changes the way a person works through a problem, not only how fast they finish it. The framework therefore places measurable behaviour between the condition and the outcome. Four behaviours make up the Cognitive Offloading Index: how often AI was asked for help, how long the task took, how much of the AI output was copied and how often the output was checked. Three signals make up the Human Engagement Score: self-reported engagement, the independence of the reasoning produced, and the share of the work done without AI. The hybrid intelligence framework [1] supplies the reason for keeping these two indices apart, because it treats offloading and active engagement as separate things rather than as opposites.

The outcomes. Two outcomes are measured while AI is present, in Session 1, and the same two are measured once AI is removed, in Session 2. Accuracy captures whether the answer was right. Rubric independence captures how much of the reasoning was the participant’s own. Separating the two is what allows the study to detect a change in process that does not show up as a change in result.

The controls. Baseline warm-up score, education level, age band, AI confidence and collection site are measured and held constant in the adjusted analysis, because each could otherwise offer an alternative explanation for a difference in outcome.

The three research questions map onto three parts of this figure. Research Question 1 asks whether the middle layer can be measured at all. Research Question 2 asks whether the condition affects the right-hand layer once AI is removed. Research Question 3 asks whether the middle layer can be used to predict the right-hand layer for an individual participant.

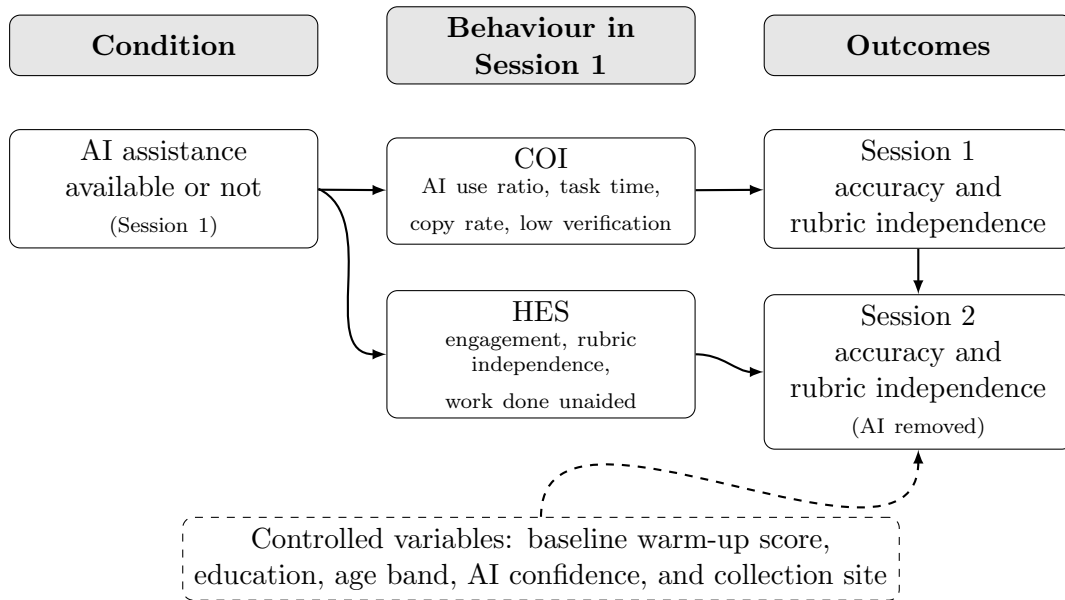


Figure 2.1: Conceptual framework. AI availability in Session 1 is the condition. COI and HES measure how the work was done. Accuracy and rubric independence measure what was produced, first with AI present and then with AI removed. The dashed box lists the variables held constant in the adjusted analysis.

2.11 Summary

This chapter reviewed key studies on AI, cognitive offloading and human engagement. The findings showed that AI can improve performance in the short term but may also change how people think and learn. Most existing studies discussed offloading and reliance at a general level without identifying the specific behavioural features that separate active thinking from passive acceptance of AI output.

This study filled that gap by measuring four specific features during problem solving: AI use ratio, time in tool, copy rate and verification rate. Together these features formed COI and made it possible to distinguish substitution from augmentation in a measurable and reproducible way [1, 4, 5].

The study also addressed the lack of delayed testing by using Session 2 as an independence endpoint where AI was removed for everyone, and the lack of reliability evidence by computing Cronbach's alpha and ICC on all measurement tools [2, 6, 7, 8].

These three gaps provided the foundation for the methodology described in Chapter 3.

Chapter 3

Methodology

3.1 Introduction

This chapter describes how the study was carried out. It covers the research design, how participants were recruited, how data was collected, how the behavioural indices were built and how the data was analysed. Every step was designed to produce measurable, auditable and reproducible outputs.

The chapter is organised around the alignment set out in Table [3.1](#) which connects each specific objective to a research question, the data used to address it, the method applied and the place where the result is reported. Sections on research design, participant recruitment, data collection, measurement and feature engineering support Specific Objective 1. The statistical analysis plan supports Specific Objective 2. The predictive modelling section supports Specific Objective 3. Chapter 4 reports the results in the same order.

Objective	Research question	Data used	Method	Where reported	Answer
SO1 Build the dataset and compute COI and HES	RQ1 Can reliance and verification be measured from logs, and how reliably?	Seven populated raw files. Table A (1,387 rows), Table B (13,870 rows)	Merging on participant code, feature engineering, z-score composites, stability and separation checks	Ch 3 Variables and Measurement, Data Processing. Ch 4 Data Integration, Feature Engineering	Yes. Both indices built and reproducible
SO2 Evaluate the effect of Session 1 reliance on Session 2 independence	RQ2 Does prior AI use reduce independent performance once AI is removed?	Table A. Accuracy S2, rubric independence S2, group, covariates	Mann-Whitney tests, Cohen's d, bootstrap CI, Benjamini-Hochberg correction, adjusted OLS, four robustness checks	Ch 3 Statistical Analysis Plan. Ch 4 Statistical Analysis, Robustness Checks	No measurable effect at this delay
SO3 Test whether Session 1 features can estimate individual Session 2 independence	RQ3 Can COI work as an individual-level screening measure?	Table A, AI-assisted participants only (634 rows), six Session 1 features	Ridge regression, grouped 5-fold cross-validation, naive baseline, subgroup error diagnostics	Ch 3 Predictive Modelling. Ch 4 Predictive Modelling	Not under these conditions. The ceiling effect prevented a fair test

Table 3.1: How each specific objective maps onto a research question, the data, the method, the results and the answer.

Each objective maps to exactly one research question and each question is answered in exactly one place. This was set before the analysis began rather than arranged afterwards. It also explains why the two primary endpoints belong to Objective 2 alone. Objective 1 produces the measures, Objective 2 tests the effect and Objective 3 tests whether the measures can be applied to an individual participant. No result is used to answer more than one question.

3.1.1 Where COI Sits in This Study

One property of the Cognitive Offloading Index needs stating before the results are read, because it determines where the index can and cannot appear. COI is computed from z-scores taken within the AI-assisted group. The control group has no COI at all, since they never used AI, and the AI group's mean is fixed at zero by construction. COI is therefore a within-condition measure of how one participant's reliance compares to another's. It is not capable of producing a difference between the two conditions, and it is not used as one anywhere in this report. This is why the manipulation check in Chapter 4 is run on HES rather than on COI.

This places COI in Objective 1, where it is built and checked, and in Objective 3, where it is tested as a predictor of an individual's Session 2 independence. Objective 2 asks a between-condition question, and that question is answered by rubric independence and HES, which exist for both groups. COI is absent from the Session 2 evidence table by design rather than by failure. Where the report reaches a conclusion about offloading between the two conditions, the evidence is the Session 1 rubric independence difference and the HES difference, both reported in Chapter 4.

3.2 Research Philosophy and Theoretical Foundation

This study followed a positivist research philosophy [25, 26]. Positivism means that knowledge must come from things that can be seen and measured, not from personal views or guesses. This way of thinking was the right fit for this study because the main question, whether AI reliance reduces independent problem solving, could be broken into numbers and measured directly. Variables such as task accuracy, rubric-scored independence and behavioural log features were all observable and countable. A positivist approach made it possible to test the question with statistical evidence rather than with opinion.

The study then used a deductive and quantitative approach [25]. In a deductive approach, the researcher starts with an idea from existing theory, turns it into something testable, and then uses data to check whether the idea holds. The starting idea here came from the cognitive offloading literature, which warns that moving mental work to an outside tool can leave a person less able to do that work alone [14, 15]. From this, a clear and testable expectation was formed: if using AI in Session 1 caused any loss of short-term independent performance, the AI-assisted group should perform worse than the control group when AI was removed in Session 2. Data was then collected and used to test this expectation directly.

The study rested on three theories that each gave a reason for the design choices and the measurement tools used. The first theory was Cognitive Load Theory [13, 24]. This theory says that working memory has limits, and that cutting out unnecessary effort can help performance, but that removing the essential steps of retrieving and explaining ideas can stop learning from happening. This gave the reason for testing performance in two separate sessions rather than only during AI access. The second theory was the theory of cognitive offloading [14, 15]. This theory says that people naturally move mental tasks to outside tools, and that this move can have measurable costs and benefits depending on how it is done. This gave the reason for building the Cognitive Offloading Index. The third was the framework of human-AI hybrid intelligence [1], which says that when a human and an AI work together, the human must still think actively and check the AI output rather than just accept it. This gave the reason for building the Human Engagement Score and for measuring verification behaviour.

The study set out to test whether AI access itself, rather than some other difference between people, explained any change in performance. An experimental design was used for this reason, rather than a survey or an observational study [27, 28]. In an experiment, the researcher changes one thing on purpose, in this case whether AI was available in Session 1, while keeping everything else as similar as possible, and then looks at what happens. An experiment gives a stronger basis than a survey for arguing that the AI condition explains a change in the results, because the researcher sets the condition instead of observing it. This is a matter of degree rather than proof. An experiment run on one convenience sample, with one task set and one short delay, supports a claim about those conditions and not a general claim about AI use. A survey could have shown that AI use and weaker reasoning happen at the same time, but it could not show which one came first or which one caused the other. This is why the experiment was the right choice.

3.3 Research Design

This study used a quantitative, between-groups experimental design with a same-sitting independence endpoint [25, 27]. Participants were divided into two groups. Group A was the control group and participants completed Session 1 without any external assistance. Group B was the AI-assisted group and participants completed Session 1 using a structured hint bank delivered through an online platform.

The specific design used was a between-groups, two-session design with a delayed no-AI endpoint [27, 28]. A between-groups design means each person was in only one group. This was chosen instead of a within-groups design, where one person would go through both the AI and the no-AI condition, because going through the AI condition first would have given participants the answers and ruined their performance in the no-AI session. Keeping the two groups separate took away this problem. Participants were assigned to groups by the platform so that the groups would start out similar, and this was confirmed later by the group balance checks.

The main measurement parameters of the study were each chosen based on existing guidance from the literature. The number of tasks and scored episodes per participant was set to produce several separate measurements per person while still keeping the sessions short enough to avoid fatigue, following the idea that reliability goes up with more items but that tiring participants out works against good data [29]. The rubric was built on a zero to two scale for each dimension because short ordinal scales with clear descriptions at each point are easier for raters to use in the same way, which helps inter-rater agreement [30]. The proportion of responses double-scored was set above twenty percent as recommended for reliability estimation, and in practice reached 66.6 percent. The significance threshold was set at the widely used level of 0.05 and effect sizes were read using the boundaries set out by Cohen [12].

The most important part of the design was Session 2. After a short break in the same sitting, with a median length of 6.6 minutes and a range of 5 to 11 minutes, AI was removed for all participants. They then completed a new but comparable set of tasks independently. This session acted as the independence test and was the primary endpoint for Research Question 2.

Both sessions were completed in the same sitting. This was done because the data collection window was narrow and participants could not be asked to return on a later day. The design therefore tested very short-term independence. It did not test overnight retention or longer-term skill transfer. This is discussed further in Chapter 5.

To reduce confounding, both groups completed two warm-up tasks without AI before Session 1 began. This provided a baseline ability score for each participant and was used to check whether the two groups were balanced before analysis [28].

3.3.1 AI Assistance Delivery

All AI assistance in this study was delivered through a standardised offline Hint Bank built into the study platform. Every participant in the AI-assisted group saw exactly the same hint text at each of the three levels. This kept the assistance identical for all participants at both sites and removed the variation that would come from live AI responses, where the wording of a prompt changes the answer given. No live ChatGPT sessions were used at any point during data collection.

Two of the nine exported data files held no records. There was no separate paper process for recording deviations, and no deviations were entered through the platform. At the MAIN site a researcher was present in the room to set up the session and read the brief, but the session itself ran on the platform and no manual deviation records were kept.

3.4 Study Setting and Participants

3.4.1 Target Population

This study targeted literate adults aged 18 and above who regularly carried out thinking tasks. Participants were recruited from universities, vocational training institutes and work-place groups within Kampala, Uganda. This population was chosen because they regularly performed thinking tasks such as analysis, planning and decision making, making them suitable for studying AI reliance during problem solving [31].

3.4.2 Sampling Method

This study used non-probability convenience sampling with some purposive elements [31]. Convenience sampling means that people were picked from groups that were easy to reach and willing to take part within the time and budget available for the study. This way of choosing participants was used because the data collection window was short and fixed, and the study needed people who could sit through two sessions in one go under the same conditions. A purposive element was added on top of this, meaning that the recruitment was aimed at literate adults who regularly do thinking tasks at work or in school. This group was chosen because AI reliance matters most for people who use thinking and reasoning as a regular part of what they do. While convenience sampling does reduce how widely the results can be applied to other groups, this is clearly noted in the limitations section. It is still a well-accepted and practical way to choose participants for a controlled experiment where the main goal is to make sure the conditions inside the study are as clean as possible, rather than to represent a wider population [31].

Two further points support this choice. First, the question being asked is a causal one, and in a design of this kind the work of ruling out alternative explanations is done by how people are allocated between conditions, not by how they were recruited into the study [28]. A convenience sample with balanced allocation supports a fair comparison between the two conditions. What it does not support is any estimate of how common a behaviour is in the wider population, and no such estimate is made here. Second, the two sites were deliberately different from one another. The MAIN site was a supervised computer lab and the ONLINE site was unsupervised and self-paced. Collecting from both made the site comparison in Chapter 4 possible, which acts as a check on whether any finding was tied to a single setting.

3.4.3 Data Collection Sites

Data was collected across two sites which reflected the practical realities of the collection period.

MAIN site: At the MAIN site, participants were brought into a computer lab in organised batches. This was necessary because many universities in Kampala were on holiday during the data collection period and students were away for internships and other commitments. Setting up computers in a central location allowed the researcher to control the environment and ensure participants could complete both sessions in the same sitting. Participation incentives were provided to encourage people to attend and give their full attention during the sessions, given the time constraints of the study.

ONLINE site: At the ONLINE site, participants received a link and a participant's ID each, to the study platform and completed both sessions at their own convenience. This group was recruited from working adults and individuals who could not attend the MAIN site sessions.

All tasks, instructions and hint materials were delivered through the same online platform used at the MAIN site to ensure consistency.

It is worth noting that data collection was stopped at 1,474 enrolled participants due to time constraints, even though participants were still opening the link and attempting to fill in the forms at the time of closure. The decision to stop was made to meet the study timeline and to allow sufficient time for analysis and reporting.

3.4.4 Inclusion and Exclusion Criteria

Inclusion criteria:

- Aged 18 years or older
- Able to read and write English well enough to complete the task prompts
- Available to complete both Session 1 and Session 2 in the same sitting
- Provided informed consent before taking part

Exclusion criteria:

- Could not commit to completing Session 2
- Had already taken part in a previous round of this study
- Could not complete the tasks within the session time limit

3.4.5 Sample Size and Retention

A total of 1,474 participants enrolled in the study. Of these, 1,387 completed both sessions, giving a retention rate of 94.1 percent. The remaining 87 participants were excluded from the main analysis. Of these, 66 started Session 1 but did not finish it, 20 checked in but never started any tasks, and 1 completed Session 1 but did not return for Session 2.

A dropout bias check confirmed that completers and non-completers had almost identical warmup scores (means of 3.46 and 3.42 respectively, difference of 0.04). This meant the final sample was a fair representation of the full enrolled group and complete case analysis was safe to proceed [28].

3.4.6 Justification of the Sample Size

Because no target was set before recruitment, the sample size has to be justified by asking what the sample that was collected can and cannot do. Two questions matter. The first is whether the sample was large enough to detect a difference of the size the study cared about. The second is whether the way people were selected supports the claims that are made from them.

On the first question, a minimum detectable effect calculation was run for both primary Session 2 endpoints. For accuracy in Session 2, the AI-assisted group had 687 participants and the control group had 700. At a two-sided alpha of 0.05 and 80 percent power, a sample of this size detects a standardised difference of $d = 0.15$. For rubric independence in Session 2, where scores were available for 1,284 participants, the smallest detectable difference is $d = 0.16$. Both sit below the value of 0.20 that Cohen [12] treats as the lower edge of a small effect. The observed differences were $d = 0.028$ for accuracy and $d = 0.026$ for rubric independence, which is about one fifth of the smallest difference the study could detect.

This changes how the Session 2 result should be read. It is not the result of a study that was too small to see anything. It is the result of a study that was large enough to detect a small effect and found something well below that size. The same calculation puts the Session 1 independence difference of $d = -1.972$ at more than twelve times the minimum detectable effect, which is why that finding is reported with confidence.

On the second question, size does not repair selection. A convenience sample of 1,387 people is still a convenience sample. A larger sample gives a more precise estimate of the difference between the two groups in this study. It does not make those 1,387 people more like the wider population. The size therefore supports the internal comparison between conditions and does nothing for claims about how common any of these behaviours are in Uganda as a whole. No such claim is made anywhere in this study.

3.4.7 Group Assignment

Participants were allocated to the two conditions by the study platform, at the point where the warm-up tasks ended and Session 1 began. Allocation was random. A participant was not told which condition they were in until Session 1 opened, and the allocation could not be changed once it had been made. The same rule was applied at both collection sites.

How the allocation was made matters for what can be claimed from the results, so it is stated here rather than left implicit. Random allocation means that differences between the two groups on characteristics that were never measured are expected to even out across a sample of this size. Where a difference in outcome is then observed, the condition itself is the most reasonable explanation for it. This is what allows the Session 1 finding in Chapter 4 to be read as an effect of the AI condition rather than as a difference between two groups of people who happened to be unlike from the start. Balance on the characteristics that were measured is reported in Chapter 4 as a check that the allocation worked as intended.

Group assignment was recorded for every participant and used as the primary contrast variable in all analyses.

3.5 Experimental Procedure

3.5.1 Session Structure

Both groups completed two sessions in the same sitting.

Session 1 took a median of 15 minutes to complete, with most participants finishing between 14 and 16 minutes. It included six tasks: two data reasoning tasks (A1, A2), two logic and consistency tasks (B1, B2) and two quantitative reasoning tasks (C1, C2). Each task required both an objective answer, which was auto-scored against a fixed answer key, and a written justification, which was rubric-scored by a rater. The AI-assisted group could use the hint bank at three levels. The control group worked alone. Data collected included accuracy, task timing, effort scale items and hint use logs.

Session 2 took a median of 9.5 minutes, with most participants finishing between 9 and 10 minutes. It included four parallel-form tasks (A3, A4, B3, B4). No AI was available for either group. Data collected included delayed accuracy and rubric-scored independence and justification quality.

Each participant therefore completed 10 task episodes in total (6 in Session 1 and 4 in Session 2), and each episode produced two scored outputs: an auto-scored objective answer and a rubric-scored justification.

3.5.2 Conduct of the Experiment

Every participant went through the study in exactly the same order so that no one had a different experience from anyone else. At the MAIN site, participants came in at set times in small batches and sat down at separate computers in a lab. Before anything started, a brief was read out loud to the group. The brief told participants what the study was about, what they would be asked to do, that joining was completely voluntary and that they could stop at any point without any problem. Each participant then gave their informed consent by confirming on the platform. At the ONLINE site, the same brief and consent step appeared on the screen when a participant opened the study link, and no task could be started until consent was given.

Once consent was done, each participant filled in a short background form that collected basic information such as age, education level and experience with AI tools. They then completed two short warm-up tasks without any AI help. These warm-up tasks were not part of the main study. They were used to give each participant a baseline ability score before the main sessions began, and to make sure both groups started from a similar level. After the warm-up, the platform placed each participant into one of the two groups and Session 1 began. In the AI-assisted group, the hint bank was switched on and participants could request help at any of the three hint levels during Session 1. In the control group, no help was available and participants worked through all tasks on their own. Throughout Session 1, the platform quietly recorded how long each task took, which hint levels were requested, whether hint content was copied into the answer, and whether a verification response was submitted.

When Session 1 ended, participants answered a short questionnaire about their experience. They then took a break of about seven minutes. After the break, Session 2 was opened for everyone. The hint bank was turned off for this session so that all participants, including those who had used AI in Session 1, worked completely on their own. When Session 2 ended, participants completed one final short questionnaire and the session was closed. This same fixed order, brief and consent, background form, warm-up, Session 1, questionnaire, break, Session 2, final questionnaire, was followed without any change for all 1,387 participants who completed the study.

3.5.3 Task Types and Families

Two task types were used. Objective tasks had one correct answer and were scored as 1 for correct and 0 for incorrect. Justification tasks required participants to explain their reasoning in their own words and were scored using a three-part rubric covering correctness, justification quality and independence.

Three task families were used across both sessions.

- **Data reasoning:** arithmetic totals and percent change calculations
- **Logic and consistency:** valid inference problems and schedule consistency checks
- **Quantitative reasoning:** simple interest calculations and cost comparison problems

Session 2 used parallel forms of the same task families with different wording and numbers to reduce memorisation effects while keeping the sessions comparable.

3.5.4 AI Assistance Protocol

The AI-assisted group received a standardised Hint Bank with three levels per task. Level 1 gave a directional hint that pointed toward the method without giving the answer. Level 2

gave a scaffold with blanks that the participant had to complete. Level 3 gave a near-complete worked solution. The participant still had to write their own final answer and justification at every level.

Participants could request any level of assistance but were required to submit final answers in their own words. Copying was measured through the platform logs. Verification was recorded for each episode.

3.6 Variables and Measurement

3.6.1 Performance Variables

- **Accuracy_S1:** proportion of correct objective tasks in Session 1
- **Accuracy_S2:** proportion of correct objective tasks in Session 2 with AI removed. Primary endpoint.
- **Rubric_Independence_S1:** mean rubric independence score in Session 1
- **Rubric_Independence_S2:** mean rubric independence score in Session 2 with AI removed. Primary endpoint.
- **Delta_Accuracy:** change in accuracy from Session 1 to Session 2. Descriptive only.

3.6.2 Origin and Theoretical Basis of the Indices

Both the Cognitive Offloading Index and the Human Engagement Score were built from behaviours that earlier researchers have already studied and linked to reliance and engagement.

The four parts of the Cognitive Offloading Index each came from findings in earlier published work. The AI use ratio measures how often a person chose to ask AI for help rather than working the problem out themselves. Risko and Gilbert [14] and Dunn and Risko [16] showed that this choice, to offload or not, is the central act of cognitive offloading. The time-in-tool component measures how much of the task time a person spent reading or interacting with AI output. Sparrow, Liu and Wegner [15] showed that the more a person depended on an outside source, the less their own processing of the content took place. The copy rate measures how much of the AI output a person moved directly into their own answer. Bućinca et al. [4] used this kind of copying as a sign that a person accepted AI output without thinking it through. The verification rate measures whether the person checked the AI output using their own reasoning. Research on human-AI trust has shown that checking is the main action that protects a person from over-accepting AI answers [4, 5]. Putting these four behaviours together into one score follows the common approach in social science of building a composite measure from several separate indicators that all point to the same underlying construct [29].

The three parts of the Human Engagement Score were picked in the same way. Self-reported engagement is a widely used way to capture how involved a person felt in a task. Rubric-scored independence measures the quality of the person’s own reasoning in their written answer. Research on learning shows that producing and explaining your own thinking is what builds real understanding, rather than receiving someone else’s explanation [24, 13]. The inverse of the AI use ratio gives credit to a person for the share of the work they did on their own, without handing it to AI. For the control group, this value was always 1.0 since they never used AI at all. Together, these three signals put into measurable form the idea from the hybrid intelligence framework [1] that in a healthy human-AI working relationship the human must stay an active

thinker and contributor, not someone who simply accepts whatever the AI produces. Keeping COI and HES as two separate scores, rather than merging them into one, was a deliberate choice grounded in theory. Offloading and engagement are two different things. A person can offload heavily to AI and still feel deeply engaged in the task. Measuring only one of them would have missed part of the picture. The construct separation check that was run later found only a weak correlation between COI and HES, which confirmed in the data that this theoretical distinction was real.

3.6.3 COI Components and Formula

The Cognitive Offloading Index was built from four behavioural features extracted from Session 1 logs for the AI-assisted group only. Each component was first standardised using a z-score so that different scales could be combined fairly [32].

The z-score formula was:

$$z(x) = \frac{x - \bar{x}}{s_x} \quad (3.1)$$

The four components were:

AI use ratio measured the proportion of tasks where the participant requested assistance. A higher ratio indicated stronger reliance on AI during problem solving [5].

$$\text{AI_Ratio} = \frac{\text{Tasks with AI}}{\text{Total tasks}} \quad (3.2)$$

Session 1 task time measured the average time a participant spent on each Session 1 task, in seconds. It was included because offloading changes how long a person spends working through a problem [6]. This component was recorded from the platform timestamps as total task duration. It does not separate time spent reading a hint from time spent working on the answer, because the platform recorded one timer per task rather than one timer per activity. This is stated here so that the component is not read as a pure measure of time inside the tool.

$$\text{Time_S1} = \frac{\text{Total Session 1 task time}}{\text{Number of Session 1 tasks}} \quad (3.3)$$

Copy rate measured the proportion of the final answer that came directly from AI output. High copy rate was a strong signal of substitution behaviour [4].

$$\text{Copy_Rate} = \frac{\text{Tasks with copied hint content}}{\text{Total Session 1 tasks}} \quad (3.4)$$

Verification rate measured how often a participant checked an AI-supplied answer using their own method. Low verification meant the participant accepted AI outputs without questioning them [4, 5].

$$\text{Verify_Rate} = \frac{\text{Verification actions}}{\text{AI uses}} \quad (3.5)$$

Verify rate was reversed before combining because low verification meant higher offloading risk. The final COI formula was:

$$\text{COLS1} = \text{mean}(z(\text{AI_Ratio}), z(\text{Time_Offload}), z(\text{Copy_Rate}), z(1 - \text{Verify_Rate})) \quad (3.6)$$

Higher COI values indicated stronger substitution behaviour. Lower values indicated augmentation behaviour [1].

One decision in the construction of this component needs stating. A participant in the AI-assisted group who never requested a hint has no AI uses to verify, so the verification rate is undefined for them. These cases were set to zero, which after reversing gives them a high score on this component. Because they also score at the bottom on the AI use ratio and copy rate components, the effect on their overall COI is partly cancelled out, but it does push their index slightly upwards.

3.6.4 HES Components and Formula

The Human Engagement Score was built from three components and was computed for all participants. For the control group, AI ratio was set to zero so that one minus AI ratio equalled one, giving them the full independence credit [8].

$$\text{HES_S1} = \text{mean}(z(\text{Engagement_self}), z(\text{Rubric_Independence_S1}), z(1 - \text{AI_Ratio})) \quad (3.7)$$

One point about this formula must be stated openly. Rubric independence in Session 1 is also reported as a Session 1 outcome in Chapter 4. Because it sits inside HES, part of the group difference in HES follows from the way HES was built and is not an independent finding. The same applies to the inverse AI ratio component, which is fixed at 1.0 for the control group by design. HES is therefore best read as a summary of how much independent work a participant did in Session 1, not as a separate piece of evidence that the two groups differed. Where HES is reported as a group comparison in Chapter 4, this should be read as a description of the difference rather than as a test of it.

Higher HES values indicated more active engagement and better quality independent reasoning.

3.6.5 Scale Items

Five subscales were collected using Likert-scale items rated from 1 to 5. These covered effort in Session 1, engagement in Session 1, AI use in Session 1 for the AI group only, engagement in Session 2 and independence in Session 2.

3.7 Data Processing

Before describing how the data was processed, it is worth setting out exactly what each participant produced. Every participant who completed the study generated the following records.

- One participant record: background variables, group assignment and the two warm-up scores.
- Two session records: the start and end timestamp of each session.

- Ten task episodes: six in Session 1 and four in Session 2. Each episode carries the objective answer, the written justification, the time taken and, for the AI-assisted group in Session 1, the hint level requested, whether hint content was copied and whether a verification response was submitted.
- Ten to twenty rubric scores: one primary rater score per episode across three dimensions, plus a second rater score for the 66.6 percent of episodes that were double-scored.
- Sixteen Likert scale responses across five subscales.
- A variable number of telemetry events: paste attempts, blocked pastes, tab blur events, infractions and answer revisions.

Across 1,387 completers this produced 13,870 task episodes, which is the row count of Table B, and 2,726 double-scored rubric pairs, which is the basis of the inter-rater reliability check in Chapter 4.

3.7.1 Integration

Two integrated datasets were built. Table A was a wide format table with one row per participant used for group comparisons and regression models. Table B was a long format table with one row per task episode used for episode-level analysis.

3.7.2 Missing Value Treatment

Three types of missingness were identified and handled differently [28].

Type 1 was missing by design. AI group variables such as AI ratio, copy rate and verify rate were missing for the control group because those participants never used AI. No imputation was applied.

Type 2 was data gaps from blank responses. Rubric scores were missing for 66 participants in Session 1 (4.8%) and 103 participants in Session 2 (7.4%) because they submitted blank justification responses. Complete case analysis was applied, meaning these participants were excluded only from rubric-based analyses.

Type 3 was one missing telemetry row. One participant had no telemetry events recorded. This was filled with zero since no events meant no actions were captured.

3.7.3 Outlier Treatment

Outlier detection used the IQR method. Two variables required treatment. Session 1 time had one extreme episode of 25,704 seconds which was capped at the IQR extreme upper bound of 244.7 seconds. Session 2 time had 11 episodes above the extreme bound of 210.5 seconds which were also capped. Verify rate had 85 participants with values above 1.0 which were capped at 1.0 since a rate above 1.0 is not logically valid. Accuracy outliers were retained because they reflected the ceiling effect rather than data errors.

3.7.4 Additional Preprocessing

Categorical variables were encoded as numbers for modelling [32]. Education level was encoded from 0 for secondary to 3 for postgraduate. Age band was encoded from 0 for 18 to 24 to 3 for 45 and above. Site was encoded as 0 for MAIN and 1 for ONLINE. A binary group label was created with AI-assisted coded as 1 and control coded as 0.

Z-scores were computed for all six variables entering COI and HES. After standardisation every variable confirmed a mean of 0 and standard deviation of 1.

Normality checks using the Shapiro-Wilk test confirmed that all primary endpoints were non-normally distributed. Mann-Whitney tests were therefore used for all group comparisons in the statistical analysis [33].

3.8 Data Collection and Analysis Tools

This section names all the tools used to collect and analyse data in this study. Naming them clearly makes it easy to separate the tools that were built by other people from the findings that belong to this study alone.

3.8.1 Data Collection Platform

Data was collected through a custom web-based study platform built specifically for this research and hosted at `aivsbrain.tech`. Participants accessed the platform through a web browser. The platform delivered the tasks, the three-level hint bank and the post-session questionnaires, and it automatically recorded all behavioural events during both sessions. These events included the time taken on each task, which hint level was requested, whether the participant copied hint content, and whether they submitted a verification response. The platform was built to capture these logs without asking participants to report them manually, so the behavioural data reflected what participants actually did rather than what they said they did.

3.8.2 Database

All participant data, task responses, scale responses, hint events and telemetry events were stored in a PostgreSQL relational database [34] hosted on Neon [35]. Neon is a serverless PostgreSQL platform that separates database storage from computing power, which means the database scaled up automatically when many participants were active at the same time and scaled back down when it was idle. This made it practical to run a large-scale data collection study without managing a dedicated server. The Neon platform required a paid subscription during the hosting period to cover compute and storage costs. Data was kept in nine separate relational tables, one for each data type, and was later exported to CSV files for analysis. The relational structure made it possible to link every event back to the participant who created it using a unique participant code.

3.8.3 Analysis Tools

All analysis was carried out in Python [36] using the Jupyter Notebook environment. Jupyter kept the analysis code, the outputs and the written explanation of each step together in one document, which made the full pipeline easy to check and repeat. The main Python libraries used were as follows.

Pandas [37] was used for loading the nine CSV files, merging them into Table A and Table B, filtering to completers, handling missing values and carrying out all data cleaning steps. NumPy [38] was used for numerical calculations including the z-score standardisation that built the COI and HES components. SciPy [39] was used for the Shapiro-Wilk normality tests, the Mann-Whitney group comparisons, and the bootstrap resampling that produced the 95 percent confidence intervals. Statsmodels [40] was used to fit the adjusted ordinary least

squares regression models for both Session 2 primary endpoints. Scikit-learn [41] was used to train the Ridge regression prediction model and to run the grouped five-fold cross-validation. The intraclass correlation coefficient for the rubric double-scoring check was computed from the two-way random effects ANOVA formula for absolute agreement with single measures, ICC(2,1), implemented directly in the analysis notebook and following the definitions set out by Koo and Li [30]. Writing the calculation out in full rather than calling a library function kept every term in the formula visible for checking. Matplotlib [42] and Seaborn was used to produce the distribution plots, scatter plots, boxplots, and heatmaps in the analysis notebooks [43]. The interactive dashboard was built with Streamlit and Plotly, which allowed the study findings to be explored through interactive charts without writing any additional code [44, 45].

Python was chosen over R or SPSS for this work because it was the only environment that could handle all parts of the pipeline in one place. R is strong for statistical testing but less suited to machine learning and dashboard deployment. SPSS provides a graphical interface for standard tests but does not support custom index construction, cross-validation or interactive visualisation. Python brought all of these together in a single reproducible notebook, which matched the end-to-end nature of the nine-stage analysis pipeline.

3.9 Analytic Approach

The analysis followed four layers.

Descriptive analytics summarised Session 1 and Session 2 outcomes by group, site and task family.

Diagnostic analytics tested which Session 1 behaviours explained Session 2 independence changes after controlling for baseline ability, education and AI confidence [28].

Predictive analytics trained an exploratory Ridge regression model using only Session 1 features to predict Session 2 independence [46]. The model was validated using grouped 5-fold cross-validation by participant ID to prevent data leakage [47].

Prescriptive analytics converted the evidence into a practical risk screen deployed in a Streamlit dashboard that allowed stakeholders to explore COI and HES distributions and flag participants at higher independence risk. The dashboard was built using Streamlit and Plotly. On startup it automatically loads the preprocessed participant-level dataset (table_a_wide.csv) from its own folder and recomputes COI and HES before displaying the results. It covers four sections: a study overview with key metrics, a participant breakdown by education and age, COI and HES distributions with interactive charts, and the main session comparison findings with the full evidence table.

The analysis was carried out as a nine-stage pipeline. Each stage had a clear purpose, produced a defined output, and used a specific tool. Table 3.2 lists the nine stages, what each stage did and the main tool used at that stage. This makes clear that every technique used was an established method, and that the contribution of this study lies in applying them to the collected data and in interpreting the findings.

Stage	Purpose	Main tool
Data Integration	Merge nine raw files, build Table A and Table B	pandas
Exploratory Data Analysis	Distributions, correlations, group and site balance	pandas, Matplotlib, Seaborn
Outlier Treatment	IQR detection and capping of extreme values	pandas, NumPy
Preprocessing	Missing values, encoding, z-scores, normality test	pandas, SciPy
Feature Engineering	Build and check COI and HES	pandas, NumPy
Reliability Checks	Cronbach alpha and ICC	pandas, NumPy
Statistical Analysis	Mann-Whitney, Cohen d, bootstrap CI, regression	SciPy, statsmodels
Predictive Modelling	Ridge regression, grouped 5-fold CV	scikit-learn
Robustness Checks	Four re-analyses under changed conditions	pandas, SciPy

Table 3.2: The nine-stage analysis pipeline and the main tool used at each stage.

Why the Dashboard Auto-Loads `table_a_wide.csv`?

The dashboard was set to load its data automatically, rather than asking the user to upload a file each time, for two reasons. First, it makes the tool easier to use. Anyone who opens the dashboard sees the results straight away without needing to know which file to pick or where it is stored. Second, it reduces the chance of error. If users could upload any file, they might load the wrong one or a file in the wrong format, which would break the charts. Loading one fixed file removes this risk.

Why the `table_a_wide.csv` file is used?

The file used is `table_a_wide.csv`, which is the wide-format participant-level table where each row is one participant and each column is one variable. This table was chosen because every chart in the dashboard works at the participant level. The group comparisons, the COI and HES distributions and the session comparison all need one value per participant, which is exactly what the wide table provides. The long-format table (`table_b_long.csv`), where each row is one task episode, was not used because the dashboard does not show any episode-level charts. Using the wide table also means the dashboard can recompute COI and HES directly from the raw component columns without any extra reshaping of the data.

3.10 Statistical Analysis Plan

3.10.1 Group Comparisons

Group comparisons were performed between the AI-assisted group and the control group for all primary and secondary outcome variables across both sessions. For each variable, the following statistics were reported: group means, mean difference between groups, Cohen’s d effect size

and 95 percent confidence intervals computed using bootstrapping with 2,000 resamples [12, 48].

The Mann-Whitney test was used for all comparisons because the Shapiro-Wilk normality test confirmed that all primary endpoints were non-normally distributed [33, 49]. Test statistics (U) and p-values were reported as secondary information. Effect sizes and confidence intervals were treated as the primary evidence in line with good practice in pilot and field studies [12].

Effect sizes were interpreted using the following thresholds [12]:

- $|d| < 0.2$: Negligible effect
- $0.2 \leq |d| < 0.5$: Small effect
- $0.5 \leq |d| < 0.8$: Medium effect
- $|d| \geq 0.8$: Large effect

The primary contrast was the AI-assisted group versus the control group. The two pre-specified primary endpoints were Accuracy_S2 and Rubric_Independence_S2. All other outcomes were treated as secondary and exploratory, including Accuracy_S1, Rubric_Independence_S1, Time_S1, HES_S1 and Delta_Accuracy.

The Benjamini-Hochberg false discovery rate correction was applied to all secondary p-values to reduce the risk of false positives arising from multiple comparisons [50]. Corrected q-values were reported alongside the original p-values for all secondary outcomes. Primary endpoints were not corrected since they were pre-specified before data collection began.

A forest plot was produced to visualise all effect sizes and their 95 percent confidence intervals in a single chart. Each variable appeared as one row. Dark blue dots represented primary endpoints and light blue dots represented secondary endpoints. The red dashed line at zero represented no difference between groups.

3.10.2 Adjusted Regression

To account for pre-existing differences between participants, an adjusted ordinary least squares regression model was fitted separately for each primary Session 2 endpoint [28]. The model included four predictors: group assignment, baseline warmup score, education level and AI confidence. The full model was:

$$\text{Performance_S2} = \beta_0 + \beta_1 \cdot \text{group} + \beta_2 \cdot \text{baseline} + \beta_3 \cdot \text{education} + \beta_4 \cdot \text{AI_confidence} + \varepsilon \quad (3.8)$$

Where group was coded as 1 for AI-assisted and 0 for control, baseline was the warmup final score (range 0 to 4), education was encoded as 0 to 3 and AI confidence was the self-rated score from the background form (range 1 to 5).

For each model, the following were reported: regression coefficients with 95% confidence intervals, p-values for each predictor and the overall R-squared value. The coefficient on the group variable was the primary quantity of interest. If that coefficient was negligible and non-significant after controlling for the covariates, it confirmed that the null result from the unadjusted comparison was not explained by pre-existing group differences [28].

An exploratory mechanism model was also fitted to test whether Session 1 COI and HES explained Session 2 performance after controlling for baseline ability:

$$\text{Performance_S2} = \gamma_0 + \gamma_1 \cdot \text{COI_S1} + \gamma_2 \cdot \text{HES_S1} + \gamma_3 \cdot \text{baseline} + \varepsilon \quad (3.9)$$

This model was labelled as exploratory and was not used to make causal claims [28].

3.10.3 Predictive Modelling

The purpose of this model needs to be separated from the purpose of the statistical tests reported earlier. The group comparisons and the adjusted regression estimate an average difference and report how uncertain that estimate is. They are read through coefficients and confidence intervals. This model estimates a value for each individual participant and is read through the size of its errors on data it has not seen. The two are not interchangeable. A study can find a large and reliable group difference and still be unable to say anything useful about a named individual, which is exactly what happened here. Both questions are asked because the practical contribution the study proposes, a risk screen that flags individual participants, depends on the second one. This is also why Specific Objective 3 is not a repeat of Specific Objective 2 in a different form. Objective 2 asks whether AI access shifts the average performance of a group, which is what an effect size answers. Objective 3 asks whether one participant's Session 1 behaviour says anything useful about that same participant's Session 2 performance, and a group average cannot answer that. Two groups can differ clearly on average while individual outcomes remain almost impossible to predict, and the reverse can also happen.

Performance was measured using two metrics [51]:

- **Mean Absolute Error (MAE):** the average absolute difference between predicted and actual Session 2 independence scores across all test participants.
- **Root Mean Squared Error (RMSE):** the square root of the mean squared difference, which penalises larger errors more heavily than MAE.

Both metrics were compared against a naive baseline that predicted the training set mean for every participant in the test set. A model that did not beat this baseline had no practical predictive value beyond knowing the group average [51].

After cross-validation, the model was refit on the full dataset to inspect feature coefficients. These showed the direction and relative magnitude of each feature's contribution to the prediction.

Subgroup error diagnostics were also run to check whether the model performed fairly across different groups. Mean absolute error was reported separately by education level and AI confidence band to identify whether any subgroup was systematically harder to predict [32].

A secondary model was also run using Accuracy_S2 as the target variable to confirm whether the pattern held across both primary endpoints.

A predicted versus actual scatter plot was produced to visualise how well the model captured individual variation in Session 2 independence scores.

3.10.3.1 Leakage Controls

Data leakage means information from the data used to test a model finds its way into the data used to train it. When this happens the model looks better than it really is, because it has effectively seen part of the answer. Four separate controls were used here.

Control 1 - Time separation. Every feature in the model comes from Session 1 only. No Session 2 variable enters the feature set at any point, whether as a predictor, as a derived value or through a variable computed from both sessions. Delta accuracy, which is built from both sessions, was excluded for this reason.

Control 2 - Target separation. The target variable, rubric independence in Session 2, is not used to build any feature. This matters for HES, which contains rubric independence from Session 1. Session 1 and Session 2 independence were scored on different tasks, so the Session 1 value is not a copy of the target, but it is a related measure and this is recorded here for transparency.

Control 3 - Participant separation. Cross-validation was grouped by participant code, so no participant appears in a training fold and a test fold at the same time. Because the modelling table holds one row per participant, this condition is met by the structure of the table. The grouping was applied anyway, so that the same code stays correct if the analysis is ever repeated at the task-episode level, where the risk would be real.

Control 4 - Preprocessing, and one limit on it. Feature standardisation was applied once to the full modelling table rather than separately within each fold. This is reported openly because it is the one step in the procedure that sits outside the fold structure. Standardisation uses no information from the target variable, so it cannot carry knowledge of the outcome into the model. What it can do is give the model a very slight advantage, because the scaling reflects the spread of the test rows as well as the training rows. That advantage works against the conclusion drawn here rather than for it. The model still failed to beat a naive baseline, so a result obtained under mildly favourable conditions is a conservative one. Fitting the scaler inside each fold is the stricter procedure and is recommended for any future use of this pipeline.

3.10.3.2 Choice and Specification of the Model

Two different kinds of model appear in this study and they answer different questions. The adjusted regression models in Section 3.10.2 are explanatory. They ask whether a group difference survives once baseline ability, education and AI confidence are held constant, and they are read through their coefficients. The model described here is predictive. It asks whether Session 1 behaviour can be used to estimate the Session 2 independence score of a participant the model has not seen, and it is read through its error on held-out data. The two are kept separate throughout, because a model can explain well and predict badly, and the reverse is also true.

Model family. Ridge regression was used [46]. Ridge is ordinary least squares with a penalty added to the size of the coefficients. It fits a straight-line relationship between the features and the target, and minimises the following quantity:

$$\hat{\beta} = \arg \min_{\beta} \{ \|y - X\beta\|_2^2 + \alpha \|\beta\|_2^2 \} \quad (3.10)$$

The first term is the usual squared error. The second term is the penalty, and α controls how strong it is. When α is zero, Ridge is the same as ordinary least squares. As α grows, the coefficients are pulled towards zero.

Why Ridge rather than something else. Three reasons. First, the feature set contains variables that are related to each other. COI and HES share the AI use ratio, one directly and one inverted, and both are related to the group variable. Ordinary least squares handles correlated predictors badly and produces unstable coefficients. The Ridge penalty is the standard remedy [32]. Second, the sample of 634 participants with six features is small enough that a flexible model such as a random forest would fit noise, and the coefficients of a linear model can be read directly, which matters because the direction of the COI coefficient is part of the answer to Research Question 3. Third, the target is a continuous score between 0 and 2, which

is a regression problem rather than a classification one. Turning it into a high-risk and low-risk classification was considered and rejected, because any cut-off would have been chosen by the researcher and the distribution gives no natural place to put one.

Specification. The regularisation strength was fixed at $\alpha = 1.0$, the scikit-learn default, and set before the model was run rather than tuned. This was a deliberate choice. Tuning α inside the same cross-validation used to report performance would bias the reported error downwards, and running a nested cross-validation on 634 rows would leave very small inner folds. The cost of this choice is that a better α may exist. Given that the model does not beat a naive baseline at $\alpha = 1.0$, and given that increasing α only pulls coefficients further towards zero, it is unlikely that tuning would change the conclusion.

Features. Six Session 1 variables: COI_S1, HES_S1, warmup_final_score, education_encoded, ai_confidence and group_binary.

Target. Rubric independence in Session 2, as the primary target, and accuracy in Session 2 as a secondary target.

Training and evaluation. Grouped 5-fold cross-validation, with standardisation fitted inside each fold. Mean absolute error and root mean squared error, both averaged across folds, compared against a naive baseline that predicts the training fold mean for every test participant.

3.10.4 Robustness Checks

Four robustness checks were run on the two primary Session 2 endpoints to test whether the main findings held under different analysis conditions [28].

Check 1 - Remove extreme time outliers: Instead of capping extreme time episodes at the IQR bound, participants with extreme time values were removed entirely from the dataset. This tested whether the capping decision was responsible for the results.

Check 2 - Complete rubric cases only: The analysis was restricted to participants who had complete rubric scores for both Session 1 and Session 2, excluding the 141 participants with any missing rubric values. This tested whether the pattern of missing rubric responses biased the main findings.

Check 3 - By collection site: The primary comparisons were re-run separately for the MAIN site ($n = 524$) and the ONLINE site ($n = 863$). This tested whether the findings were driven by one site or held consistently across both collection environments.

Check 4 - By education level subgroup: The primary comparisons were re-run separately within each of the four education level subgroups: Secondary, Diploma, Degree and Postgraduate. This tested whether the null result in Session 2 held across all education levels or was driven by one particular group.

For each check and each primary endpoint, the following were reported: Cohen's d , p -value and whether the direction of the effect held compared to the reference result from the main analysis. A finding was considered robust if the direction held and the effect size remained negligible across all four checks [12, 28].

3.11 Reliability Checks

Reliability checks were run before the main statistical analysis to confirm that the measures used in this study were consistent and trustworthy. If a measure was unreliable, any results built on it could not be trusted even if the effect sizes looked interesting.

Scale reliability - Cronbach's alpha: Cronbach's alpha was computed for all four multi-item Likert scales: effort in Session 1, engagement in Session 1, AI use in Session 1 and independence in Session 2 [29]. The formula used was:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum \text{Var}(x_i)}{\text{Var}(X)} \right) \quad (3.11)$$

Where k is the number of items, $\text{Var}(x_i)$ is the variance of each individual item and $\text{Var}(X)$ is the variance of the total scale score. A value of 0.70 or above was considered acceptable for this study [29]. Item-level correlations were also inspected using heatmaps to identify which pairs of items were driving low alpha values.

Rubric inter-rater reliability - ICC: A second rater independently scored 66.6 percent of all justification responses, well above the 20 percent minimum threshold. ICC(2,1), the two-way random effects model for absolute agreement with single measures, was computed for each of the three rubric dimensions: correctness, justification quality and independence [30]. A value of 0.60 or above was considered acceptable for a pilot study [30]. Pearson correlation between raters and the mean score difference were also reported for each dimension. Rater agreement scatter plots were produced to visualise the level of agreement per dimension.

COI manipulation check: A Mann-Whitney test was used to compare HES between the AI-assisted and control groups [33]. Since HES was available for both groups, it provided a stronger manipulation check than comparing COI alone. A significant difference confirmed that the AI condition produced measurable behavioural differences between groups.

COI ablation check: COI was recomputed four times, dropping one component at a time and using the remaining three. The mean of each ablated COI was compared to the full four-component COI mean. A difference of less than 0.1 was considered stable. This confirmed that no single component dominated the index [32].

COI weight sensitivity check: Two alternative weighting schemes were tested against the equal weights used to build COI. The first gave double weight to verify rate. The second gave double weight to AI ratio. Pearson correlations between each alternative scheme and the equal weight scheme were computed. A correlation above 0.90 confirmed that equal weights were defensible [32].

COI and HES inter-component correlations: Pearson correlations between all pairs of components within COI and within HES were computed and displayed as heatmaps. Positive correlations within an index confirmed that all components pointed in the same direction. A strong correlation between COI and HES was also checked to confirm construct separation, meaning the two indices measured different things [8].

3.12 Bias Control and Data Validation

Several kinds of bias could affect a study of this design. Each was identified before the analysis began and each was handled by a specific check rather than by assumption. Table 3.3 lists them, what was done and where the evidence is reported.

Type of bias	What was done about it	Where the evidence is
Selection bias	Fixed inclusion and exclusion criteria applied by the platform before any task could start. Two different sites used, one supervised and one self-paced.	Ch 3, Inclusion and Exclusion Criteria; Ch 3, Data Collection Sites
Allocation bias	Warm-up ability, AI confidence, education and age band compared between the two conditions before any outcome was analysed.	Ch 4, Table 4.2
Attrition bias	Warm-up scores of the 87 non-completers compared with those of the 1,387 completers.	Ch 4, Completers and Exclusions
Confounding	Adjusted regression on both primary endpoints, controlling baseline ability, education and AI confidence.	Ch 4, Adjusted Regression Results
Measurement bias in the tasks	Objective answers auto-scored against a fixed key, removing rater judgement from the primary accuracy endpoint.	Ch 3, Task Types and Families
Rater bias	A second rater independently scored 66.6 percent of justifications. Agreement measured with ICC and the mean difference between raters reported.	Ch 4, Table 4.12
Response bias in self-report	Self-report compared against behavioural logs on the same participants. Where they disagreed, the log measures were used and the self-report was demoted to descriptive.	Ch 5, self-report and behavioural section
Missing data bias	Three types of missingness identified and handled separately. Whole analysis repeated on complete rubric cases only.	Ch 3, Missing Value Treatment; Ch 4, Check 2
Site and subgroup bias	Primary comparisons repeated separately by site and by education level. Model error reported by subgroup.	Ch 4, Table 4.20; Subgroup Error Diagnostics
Multiple comparison bias	Benjamini-Hochberg false discovery rate correction applied to all secondary p-values.	Ch 3, Group Comparisons

Table 3.3: Bias types considered in this study, the control applied to each and where the evidence appears.

Two of these need further comment because a check was run and the result was not clean.

The first is rater bias. The second rater scored lower than the primary rater on both subjective rubric dimensions, by 0.241 points on justification quality and by 0.351 points on independence. This is a measured, systematic difference, not random noise, and it points to the two raters not having been calibrated against a common standard before scoring began. Because the primary rater’s scores were used throughout the analysis, any systematic tendency in that rater’s standard is carried into every rubric-based result. The difference is in a consistent direction, which means it would shift both groups in the same way rather than favour one of them, but this cannot be confirmed from the data. This is discussed further in Chapter 5.

The second is the stopping rule. Recruitment stopped when the collection window closed rather than at a planned target, and participants were still opening the link at that point. This means the sample favours people who responded early, which is a mild form of selection bias. Nothing in the design corrects for it, and it is recorded in the limitations.

Three sources of bias were not controlled and should be stated. Neither participants nor the primary rater were blind to condition, so the rater knew which group a script came from, and expectations could have affected scores on the subjective dimensions. Incentives were offered at the MAIN site, which may have drawn people with different motivations from those who joined online without payment. And the tasks were completed in a session set up for research, not as work that affected a participant's grade or pay, so the pressure to get a correct answer quickly was lower than it would be in a real assessment.

3.13 Ethical Considerations

Ethical approval was granted by the Research Ethics Committee on 25 June 2026, marking the official start of the study. All data collection, storage and processing followed the conditions of that approval and the ethical guidelines for research involving human participants [9, 10].

Before the study began, all participants were briefed on the full purpose of the study, what they would be asked to do, how long it would take and how their data would be stored and used. Participants were told that taking part was voluntary and that they could withdraw at any time without any penalty or consequence. Informed consent was obtained from every participant before any data was collected.

All data was anonymised using unique participant IDs. Participants were instructed not to include personal identifiers such as their name or phone number in any platform inputs during the sessions. Only the researcher and supervisor had access to the raw data files.

Participation incentives were provided during data collection to encourage attendance and full engagement, particularly at the MAIN site where participants were brought in batches during a period when universities were on holiday. This was necessary due to time constraints and is acknowledged as a possible source of variability in participant motivation.

The risks to participants were minimal. Some participants may have experienced mild mental fatigue from completing the problem-solving tasks across two sessions. Participants were allowed to take short breaks at any time and could stop without any consequence. No sensitive personal information beyond basic background variables such as age range and education level was collected.

All data handling followed ethical guidelines for research involving human participants and responsible AI use [9, 10].

Chapter 4

Analysis and Results

4.1 Introduction

This chapter presents the full analysis and results of the study. It follows the same order as the analysis pipeline: data integration and exploration, preprocessing, feature engineering, reliability checks, statistical analysis, predictive modelling and robustness checks. All results are reported with the actual numbers from the analysis. Where appropriate, figures from the analysis notebooks are included to support the written interpretation. The results appear in the order set out in Table [3.1](#). Data integration, preprocessing and feature engineering report the outcome of Specific Objective 1. The reliability checks and the statistical analysis report Specific Objective 2. The predictive modelling section reports Specific Objective 3, and the robustness checks apply to the Objective 2 endpoints.

4.2 Data Integration and Sample Description

4.2.1 Data Files Loaded

Nine raw data files were loaded for analysis. Table [4.1](#) lists each file, its content and whether it contained data.

File	Content	Status
participants.csv	Participant background, group assignment and warmup scores	Used
session_meta.csv	Session start and end timestamps, transfer prompt responses	Used
task_responses.csv	Task-level answers, justifications, timing and AI use flags	Used
rater_scores.csv	Primary and double rater rubric scores per task episode	Used
scale_responses.csv	Likert scale item ratings per participant per session	Used
hint_events.csv	Hint level requests, viewing times and copy flags per task	Used
telemetry_events.csv	Paste, tab blur, infraction and answer revision events	Used
deviation_logs.csv	Proctor deviation records	Empty
layer2_logs.csv	Live ChatGPT session logs	Empty

Table 4.1: Raw data files loaded for analysis.

Two files were empty and were retained in the repository for transparency but not used in any analysis step. The deviation log was empty because the study was delivered entirely through the online platform at both sites. There was no separate paper process for recording deviations, and no deviations were entered through the platform. At the MAIN site a researcher was present in the room to set up the session and read the brief, but the session itself ran on the platform and no manual deviation records were kept.

4.2.2 Completers and Exclusions

Of the 1,474 participants who enrolled, 1,387 completed both sessions and formed the final analysis sample, representing a retention rate of 94.1%. The remaining 87 participants were excluded as follows:

- 66 started Session 1 but did not complete it
- 20 checked in but never started any tasks
- 1 completed Session 1 but did not return for Session 2

A dropout bias check was carried out before proceeding with complete case analysis [28]. Completers had a mean warmup score of 3.46 and non-completers had a mean of 3.42. The difference of 0.04 was negligible and confirmed there was no meaningful dropout bias. The groups that dropped out were therefore not systematically different from those who completed the study. Complete case analysis was safe to proceed [28].

4.2.3 Dataset Integration

Two integrated datasets were built from the nine raw files by merging on `participant_code` across all exports.

Table A (wide format) contained one row per participant with 1,387 rows and captured all Session 1 and Session 2 variables side by side in columns. This table was the primary dataset for group comparisons, COI and HES computation and regression models.

Table B (long format) contained one row per task episode. After filtering to the 1,387 completers, Table B had 13,870 rows (1,387 participants $\times$ 10 episodes each). An earlier version of the table included 14,063 rows, but 193 of these were partial episodes belonging to the 87 excluded participants and were removed once the completers filter was applied. Each row represented one participant completing one task and was used for episode-level analysis including time outlier detection and capping, hint usage summaries and telemetry event aggregation.

The two tables were linked through `participant_code` and were always kept in sync. Any change made to Table B (such as capping an extreme time episode) was immediately propagated back to Table A by recomputing the relevant participant-level summary variable.

4.2.4 Session Delay

The actual delay between Session 1 and Session 2 was computed from the `session_meta` timestamps by subtracting the Session 1 end time from the Session 2 start time for each participant. The delay ranged from 5 to 11 minutes across all participants, with a median of 6.6 minutes. This confirmed the study used a same-sitting design.

Both sessions ran in the same sitting with only a short break between them. A gap of this size means the results speak to very short-term independence and not to overnight retention or longer-term skill transfer. This is discussed as a key limitation in Chapter 5.

4.2.5 Group and Site Balance

Table 4.2 shows the key balance variables across the two groups before any analysis was run.

Variable	AI assisted	Control	Difference
Warmup score (mean)	3.47	3.44	+0.03
AI confidence (mean)	3.37	3.40	−0.03
Accuracy S1 (mean)	0.971	0.969	+0.002
Group size	687	700	—

Table 4.2: Group balance on key variables before analysis. Differences are AI minus Control.

The two groups were very similar on all key variables. The maximum difference on any balance variable was 0.03, which is negligible. Education level and age band distributions were also comparable across groups. The majority of participants had a degree-level education ($n = 581$, 41.9%) and fell in the 18 to 24 age band ($n = 766$, 55.2%), which was consistent with the university-based recruitment approach.

For site balance, the MAIN site had 260 AI-assisted and 264 control participants ($n = 524$ total). The ONLINE site had 427 AI-assisted and 436 control participants ($n = 863$ total).

The split between groups was even within each site, confirming that site and group assignment were not confounded [28].

4.3 Exploratory Data Analysis

4.3.1 Key Variable Distributions

Both groups showed a strong concentration at accuracy 1.0 in both sessions, confirming the ceiling effect. Very few participants scored below 0.83 on accuracy. The AI-assisted group had 9 participants below 0.67 and the control group had 5, out of 687 and 700 respectively. This narrow distribution meant that accuracy alone could not distinguish the two groups and was a consistent pattern across all task families.

Rubric independence in Session 1 showed much more variation and was the most informative variable in the EDA. The AI-assisted group had scores spread between 0.0 and 2.0 with a concentration around 1.0 to 1.25, while the control group concentrated heavily at 2.0. The AI-assisted group mean was 1.206 and the control group mean was 1.843, a difference of 0.638 points visible in the distribution plots. This early pattern strongly suggested the control group produced better quality reasoning in Session 1 even though accuracy was the same for both groups.

In Session 2 with AI removed, both groups concentrated at higher independence scores. The AI-assisted group improved from a mean of 1.206 in Session 1 to 1.860 in Session 2, nearly closing the gap with the control group which had a Session 2 mean of 1.854. The distributions in Session 2 were much more similar across groups than in Session 1.

Effort and engagement in Session 1 were nearly identical across both groups, with most responses between 3.5 and 4.5 on the 5-point scale. Both groups had effort means of approximately 3.90 and engagement means of approximately 3.93. This confirmed that engagement and offloading were separate constructs. A person could feel engaged and still offload heavily to AI [8].

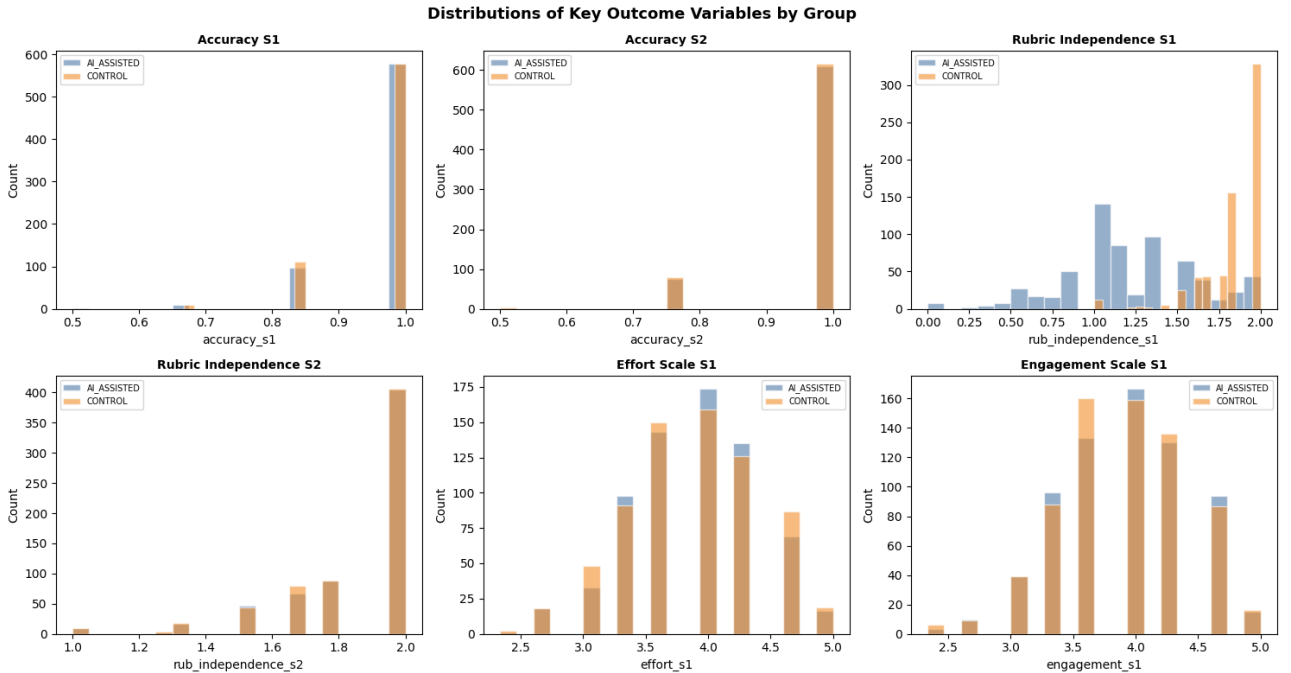


Figure 4.1: Distributions of key outcome variables by group. Accuracy shows ceiling effect. Rubric independence S1 shows large group difference.

4.3.2 Task Accuracy by Task

All ten tasks had accuracy between 96.6% and 97.6% across both groups and both sessions. Table 4.3 shows the accuracy per task family.

Task family	AI S1	CTRL S1	AI S2	CTRL S2
Data reasoning	0.972	0.970	0.972	0.970
Logic and consistency	0.970	0.968	0.971	0.968
Quantitative reasoning	0.971	0.969	0.970	0.969
Overall	0.971	0.969	0.971	0.969

Table 4.3: Mean accuracy by task family and group across both sessions. All values are at ceiling.

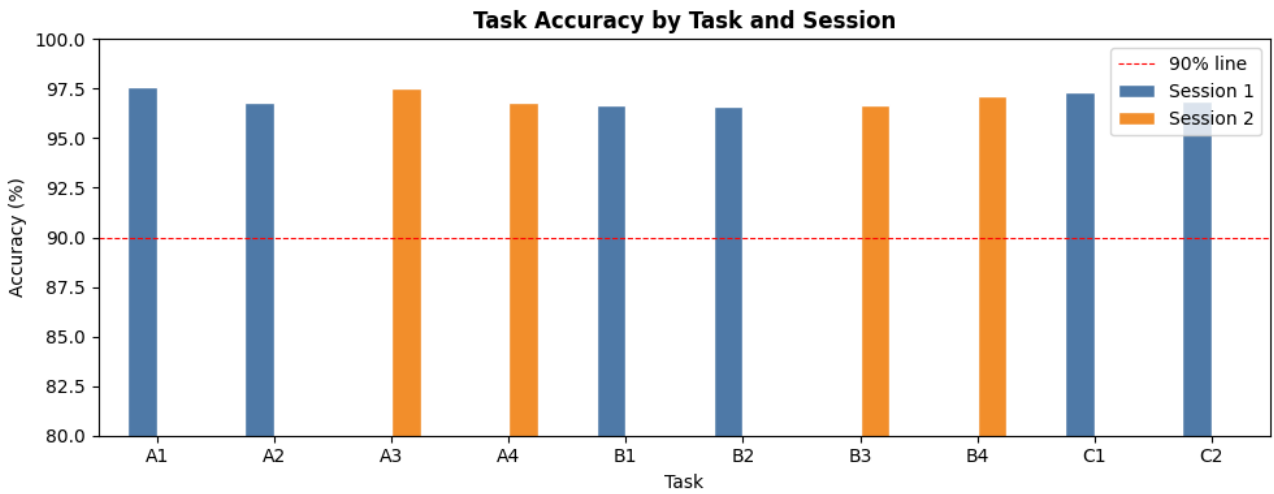


Figure 4.2: Mean accuracy visualized by task family and group across both sessions.

Every task was well above 90% across all conditions. This uniformly high performance confirmed the ceiling effect and was consistent with the task design which used structured problems with self-contained information. The ceiling effect was a design limitation that reduced the sensitivity of accuracy as an outcome measure throughout the analysis.

It is worth pausing on what this ceiling means for the rest of the analysis. When almost every participant scores 1.0, the accuracy variable stops behaving like a measurement and starts behaving like a constant. Two consequences follow. First, any group comparison on accuracy has almost no variance to work with, so even a real difference between the groups would be very hard to detect. The negligible Cohen's d values reported later for accuracy should therefore be read as "no difference was detectable" rather than "no difference exists". Second, a prediction model cannot learn a pattern in a variable that barely moves, which is the main reason the secondary model on accuracy S2 in Section 4.7 produced exactly the same error as the naive baseline. The ceiling effect is not a side observation; it shapes what the accuracy endpoint is capable of showing throughout this chapter.

4.3.3 Time Per Task

The median time per task was 103 seconds in Session 1 and 97 seconds in Session 2. Session 2 was slightly faster on average, which was expected since participants were more familiar with

the task format in the second sitting and had no AI output to read or interact with. The mean time per task was 106.6 seconds in Session 1 and 97.8 seconds in Session 2.

Session 1 contained one extreme episode of 25,704 seconds (approximately 7 hours) for participant MAIN-01-0012 on task A2. This was clearly not genuine task time and was most likely the device being left open. A further 13 episodes in Session 1 and 11 episodes in Session 2 were above the IQR extreme upper bounds and were treated in the preprocessing step described below.

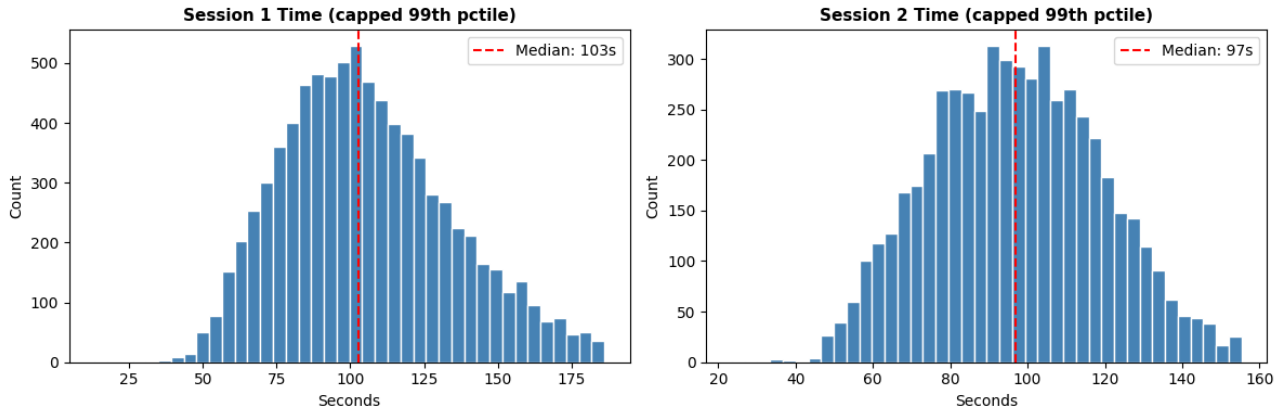


Figure 4.3: Visuals showing Time Per Task in seconds

4.3.4 Hint and Telemetry Summary

In the AI-assisted group, participants requested an average of 8 hints per session. The mean hint viewing time was approximately 77 seconds per session. About 16% of hints received were copied by participants. Of the 687 AI-assisted participants, 683 requested at least one hint and 4 never requested any hints across both sessions, suggesting they chose not to use the hint bank despite having access to it.

The most commonly used hint level was Level 1 (the directional hint), followed by Level 2 (the scaffold). Level 3 (the near-complete solution) was used least frequently, suggesting most participants preferred partial guidance over a full solution.

The telemetry log recorded 319 paste block events in the AI-assisted group compared to only 1 in the control group, confirming the paste prevention system functioned correctly. A total of 12 infractions were recorded across both groups, which was very low relative to the 1,387 participants and 13,870 task episodes. This suggested participants followed the platform instructions well throughout both sessions.

The AI-assisted group also had 97 answer revision events compared to 23 in the control group. This higher revision count in the AI group was consistent with the hint bank workflow where participants could update their answer after viewing hints.

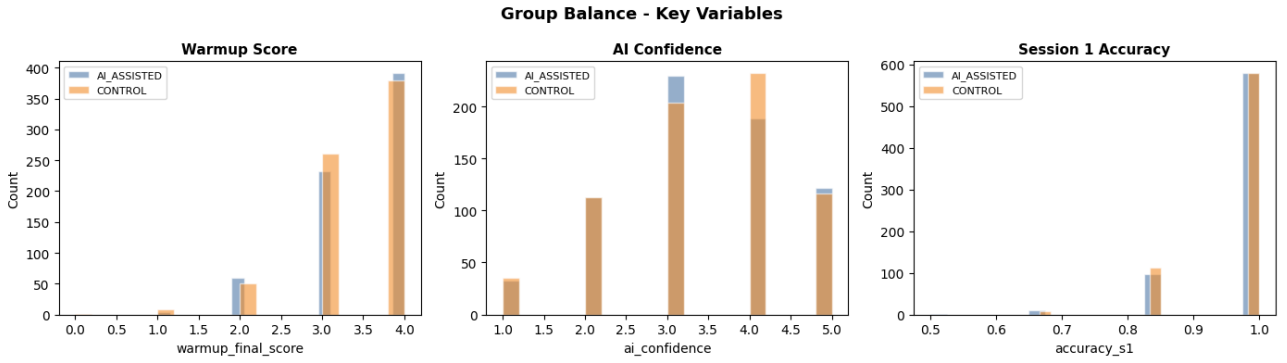


Figure 4.4: Group Balance - Key Variables showing distributions of warmup score, AI confidence and Session 1 accuracy across both groups.

4.4 Outlier Detection

The IQR method was applied to all 15 numeric variables in Table A [52]. The IQR method flags a value as an outlier if it falls below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$. For extreme outliers a multiplier of 3.0 was used instead of 1.5 to avoid flagging too many valid values. Table 4.4 shows the key findings.

Variable	Outliers	%	Decision
accuracy_s1	229	16.5	Retained – IQR = 0, ceiling effect
accuracy_s2	160	11.5	Retained – IQR = 0, ceiling effect
rub_independence_s2	58	4.5	Retained – within valid range 0 to 2
effort_s1	39	2.8	Retained – within valid scale range
engagement_s1	28	2.0	Retained – within valid scale range
ai_ratio	28	4.1	Retained – within valid range 0 to 1
verify_rate	14	2.0	Capped at 1.0 (logically impossible above 1.0)
time_s1	11	0.8	Capped at IQR extreme bound of 244.7 sec
time_s2	12	0.9	Capped at IQR extreme bound of 210.5 sec
copy_rate	0	0.0	No treatment needed
rub_independence_s1	0	0.0	No treatment needed

Table 4.4: IQR outlier detection results for key variables [52]. Variables with 0 outliers are included for completeness.

Accuracy variables were flagged by IQR because their IQR was zero. Since nearly all partic-

ipants scored 1.0, both Q_1 and Q_3 were equal to 1.0, making $IQR = 0$. Any score below 1.0 therefore fell outside the whiskers automatically. These were valid scores reflecting the ceiling effect and were retained.

The time variables had one extreme episode of 25,704 seconds in Session 1 belonging to participant MAIN-01-0012 on task A2, and several episodes above 200 seconds in Session 2. These were capped at the IQR extreme upper bounds of 244.7 seconds for Session 1 and 210.5 seconds for Session 2. After capping, the new maximum for time_s1 dropped from 4,401.6 seconds to 235.1 seconds, and time_s2 dropped from 292.9 seconds to 185.9 seconds.

Verify rate had 85 participants with values above 1.0, with a maximum of 2.5. A rate above 1.0 is not logically valid since it would mean more verifications than AI uses. These 85 participants (12.4% of the AI group) had their verify rate capped at 1.0.

Outlier detection was not run separately on Table B because Table A summarises Table B at participant level. The one exception was time, where capping was applied at the episode level in Table B first, after which participant-level means in Table A were recomputed from the capped episodes.

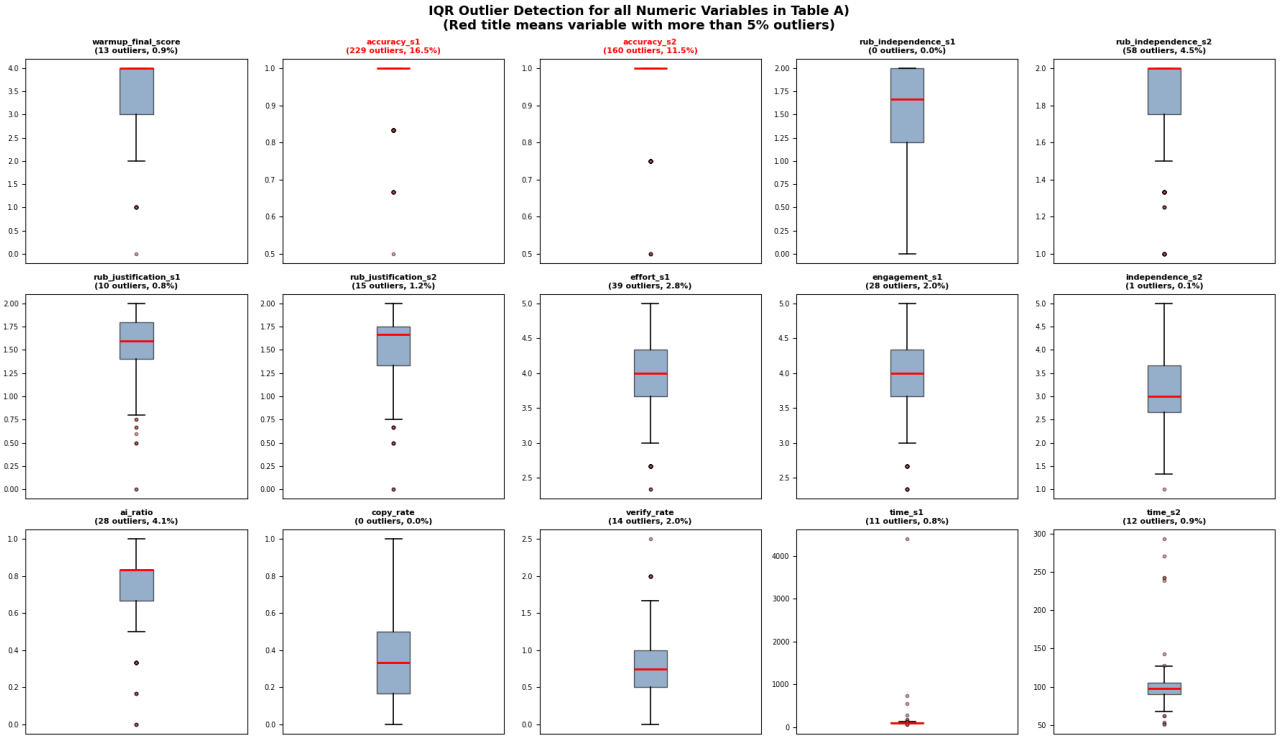


Figure 4.5: IQR outlier detection across all 15 numeric variables. Variables with red titles have more than 5% outliers.

4.5 Data Preprocessing

4.5.1 Missing Value Treatment

Three types of missingness were identified across Table A and Table B. Each type was handled differently based on the reason for the missing values [28].

Type 1 - Missing by design (50.5% in Table A, 70.3% in Table B): Variables such as ai_ratio, copy_rate, verify_rate, total_hints, max_hint_level, hint_copy_count and hint_copy_rate were missing for the control group because the control group never used AI. In Table B, assistance_level, copy_used and verified were missing for all Session 2 rows and all control group rows

because AI was not available in those conditions. No imputation was applied. These variables are structurally missing and do not apply to those participants or sessions [28].

Type 2 - Data gaps from blank responses (4.8% to 7.4% in Table A, 28.3% in Table B): Rubric scores were missing for 66 participants in Session 1 (4.76%) and 103 participants in Session 2 (7.43%) because they submitted blank justification responses that could not be scored. In Table B, this produced 3,927 missing rows (28.3%) across the correctness, justification_quality and independence columns. Complete case analysis was applied [28]. These participants were excluded only from rubric-based analyses but kept for all accuracy-based analyses.

Type 3 - One missing telemetry row (0.07%): One participant had no telemetry events recorded in any of the five telemetry columns. Since no events means no actions were captured, these five columns were filled with zero for that participant. This affected only 1 row and had negligible impact on the analysis.

Table 4.5 summarises the missingness after treatment.

Variable group	Missing n	%	Treatment
AI group behavioural variables	700–704	50.5	No imputation (by design)
Rubric scores S1	66	4.76	Complete case analysis
Rubric scores S2	103	7.43	Complete case analysis
Telemetry columns	1	0.07	Filled with zero

Table 4.5: Missingness summary after treatment in Table A.

4.5.2 Additional Preprocessing

Categorical variables were encoded as ordered integers for use in regression and prediction models [32]:

- Education level: SECONDARY = 0, DIPLOMA = 1, DEGREE = 2, POSTGRAD = 3
- Age band: 18_24 = 0, 25_34 = 1, 35_44 = 2, 45 and above = 3
- Site: MAIN = 0, ONLINE = 1
- Group: CONTROL = 0, AI_ASSISTED = 1

Z-scores were computed for all six variables entering COI and HES using the formula $z(x) = (x - \bar{x})/s_x$ [32]. After standardisation, all six z-score columns confirmed mean = 0.000 and standard deviation = 1.000.

Normality was checked using the Shapiro-Wilk test [49] on all six primary endpoint variables. Table 4.6 shows the results.

Variable	n	W statistic	p-value	Test to use
accuracy_s1	1,387	0.4605	0.000	Mann-Whitney
accuracy_s2	1,387	0.3767	0.000	Mann-Whitney
rub_independence_s1	1,321	0.8898	0.000	Mann-Whitney
rub_independence_s2	1,284	0.6941	0.000	Mann-Whitney
effort_s1	1,387	0.9598	0.000	Mann-Whitney
engagement_s1	1,387	0.9571	0.000	Mann-Whitney

Table 4.6: Shapiro-Wilk normality test results [49]. All p-values = 0.000, confirming non-normal distributions. Mann-Whitney tests were used for all group comparisons [33].

All six variables were non-normally distributed (all $p = 0.000$). Mann-Whitney tests were therefore used for all group comparisons throughout the analysis [33].

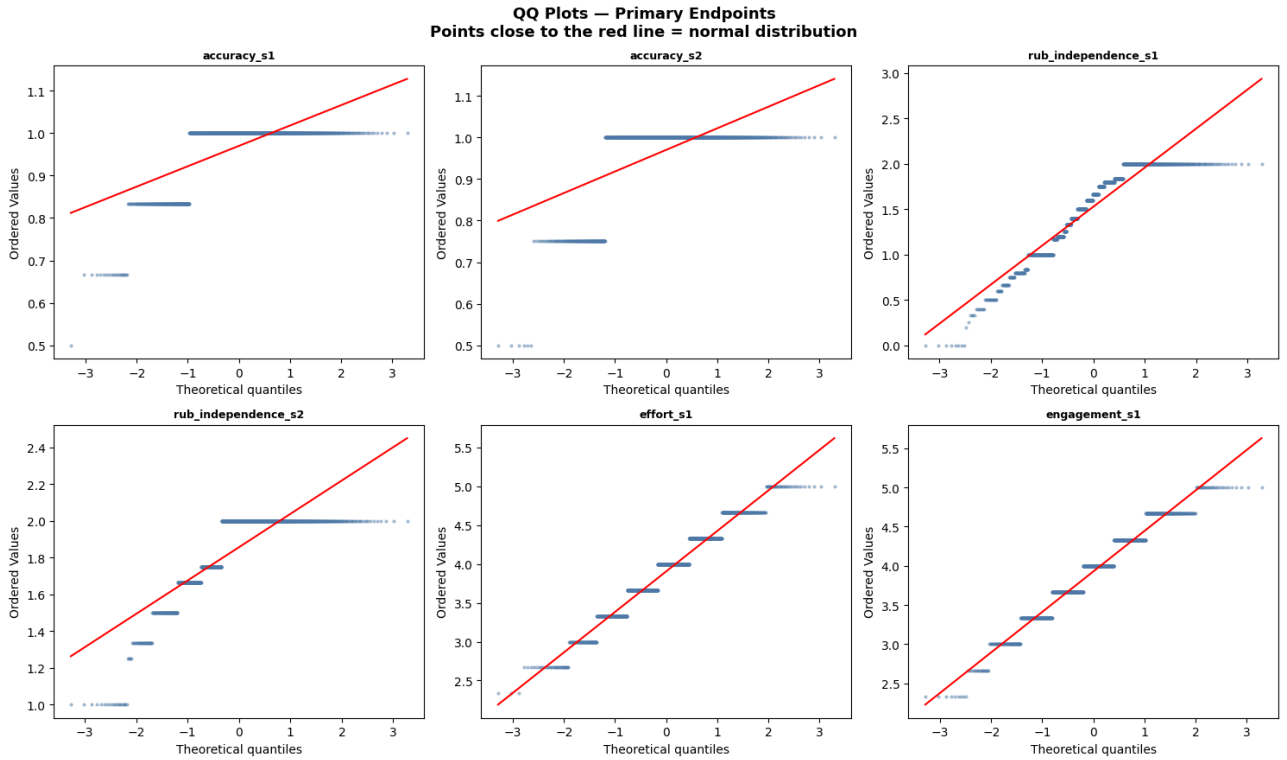


Figure 4.6: QQ plots for all six primary endpoint variables confirming non-normal distributions. Dots deviate substantially from the red diagonal line in all cases.

4.6 Feature Engineering - COI and HES

4.6.1 COI Construction

As set out in Chapter 1, COI exists only for the AI-assisted group and is standardised within that group, so it describes variation between AI-assisted participants and is never used as a group contrast. The four raw COI components were extracted from Session 1 logs for the 687 AI-assisted participants. Table 4.7 shows the descriptive statistics for each raw component.

Component	Mean	SD	Min	Max
AI use ratio	0.746	0.186	0.000	1.000
Time in tool (sec)	106.622	11.297	63.522	201.175
Copy rate	0.369	0.191	0.000	1.000
Verify rate (before reversal)	0.710	0.258	0.000	1.000
1 – Verify rate (after reversal)	0.290	0.258	0.000	1.000

Table 4.7: Raw COI component descriptive statistics for the AI-assisted group (n = 687).

The AI use ratio mean of 0.746 indicated that on average participants used AI on about 75% of their tasks [5]. The copy rate mean of 0.369 showed that about 37% of tasks had copying behaviour [4]. The verify rate mean of 0.710 indicated participants verified AI output in about 71% of uses [4, 5].

Verify rate was reversed to (1 – verify_rate) before standardising because low verification means higher offloading risk. After reversal the mean became 0.290, reflecting the same data from the opposite direction [4].

After z-score standardisation, all four components confirmed mean = 0.000 and SD = 1.000. The z-score standardisation table confirmed this for each component separately.

COI was computed as the unweighted mean of the four standardised components for each AI-assisted participant:

$$\text{COLS1} = \frac{1}{4}(z(\text{AI_Ratio}) + z(\text{Time_S1}) + z(\text{Copy_Rate}) + z(1 - \text{Verify_Rate})) \quad (4.1)$$

The final COI distribution had a mean of 0.000 and a standard deviation of 0.578, ranging from −1.743 to +2.292. The distribution was approximately bell-shaped, confirming a reasonable spread of offloading behaviours within the AI group.

The spread of the COI scores tells us something useful about the AI-assisted group. The standard deviation of 0.578 with a range from −1.743 to +2.292 shows that participants did not all use the hint bank in the same way. Some requested help on nearly every task, copied heavily and checked little, while others used one or two hints and verified their own answers. This variation matters because it is the raw material the prediction model in Section 4.7 needs. If every participant had behaved identically, COI would have been a constant and could not have predicted anything. The fact that the model still failed to beat the baseline despite this genuine spread in COI is therefore a more meaningful result than it would have been if COI itself had been flat. The component means also help explain the shape of the distribution. AI use ratio averaged 0.746 and verification rate averaged 0.710, both fairly high. This means the typical participant in this study used AI often but also checked its output often. In the substitution and augmentation language of Chapter 2, most participants sat closer to augmentation than to full substitution [1]. The hint bank design is the likely reason. Because participants had to write their own final answer at every hint level, even the highest level of assistance still required some independent work. A study allowing free-form AI use with copy and paste would very likely produce a COI distribution shifted much further towards substitution.

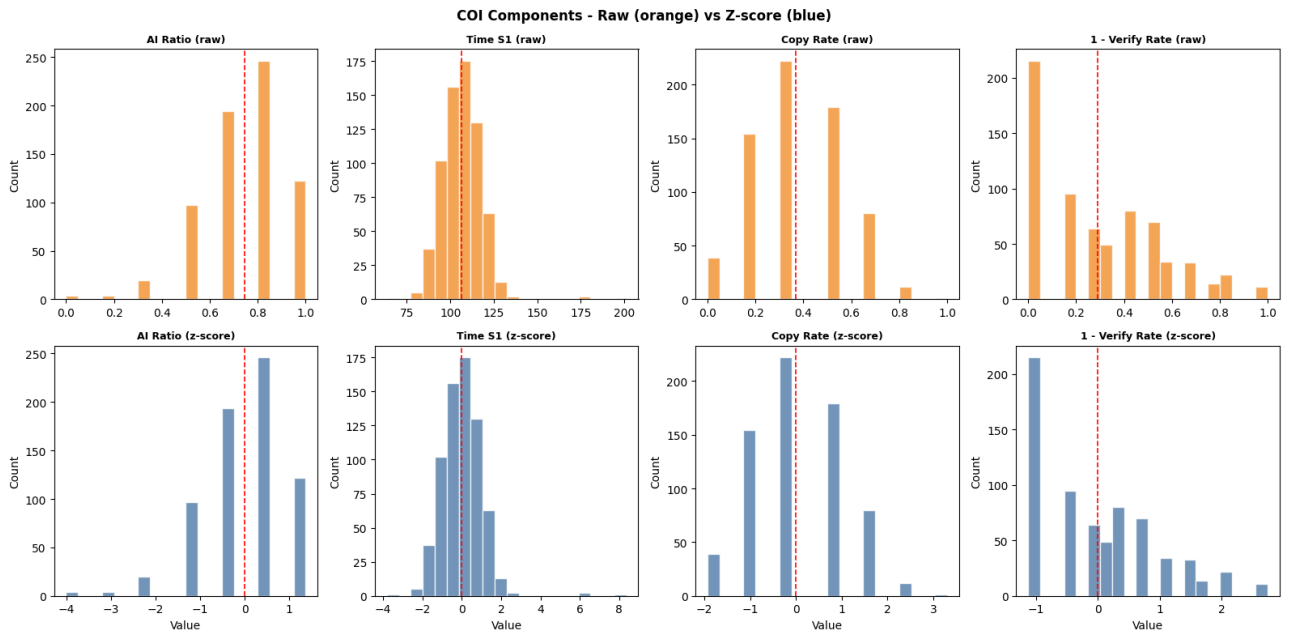


Figure 4.7: COI raw components (orange, top row) and z-score standardised components (blue, bottom row). Shape is preserved after standardisation but scale is unified.

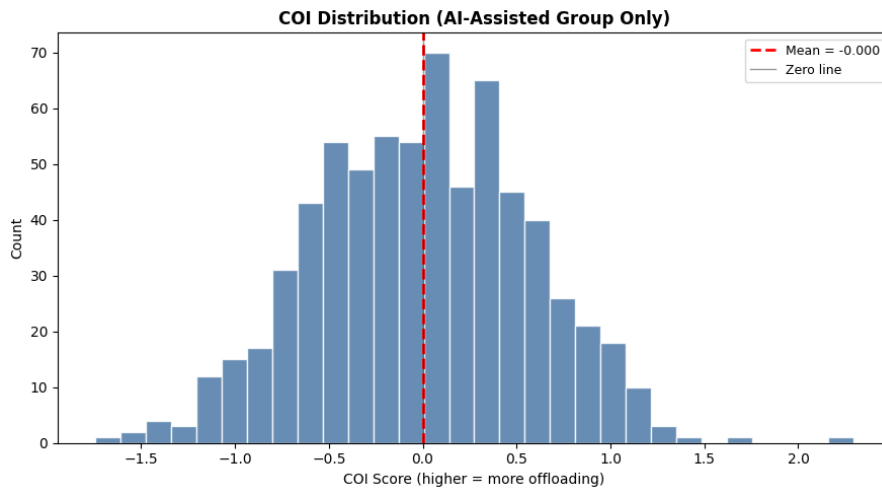


Figure 4.8: COI distribution for the AI-assisted group ($n = 687$). Mean = 0.000. SD = 0.578. Range: -1.743 to $+2.292$.

4.6.2 HES Construction

HES was computed for all 1,387 participants. The three raw HES components and their descriptive statistics are shown in Table 4.8.

Component	n	Mean	SD	Min	Max
Engagement self-report	1,387	3.929	0.527	2.333	5.000
Rubric independence S1	1,321	1.527	0.454	0.000	2.000
1 – AI ratio	1,387	0.631	0.395	0.000	1.000

Table 4.8: Raw HES component descriptive statistics across all participants ($n = 1,387$). The control group has $\text{ai.ratio} = 0$ so $1 - \text{AI ratio} = 1.0$.

For the control group, AI ratio was set to 0 since they never used AI. This made $1 - \text{AI_ratio} = 1.0$ for every control group participant, giving them the maximum independence credit on this component [8]. After z-score standardisation and averaging, the AI-assisted group had a mean HES of -0.545 and the control group had a mean HES of $+0.537$. This difference of 1.082 was the most prominent finding from the feature engineering step.

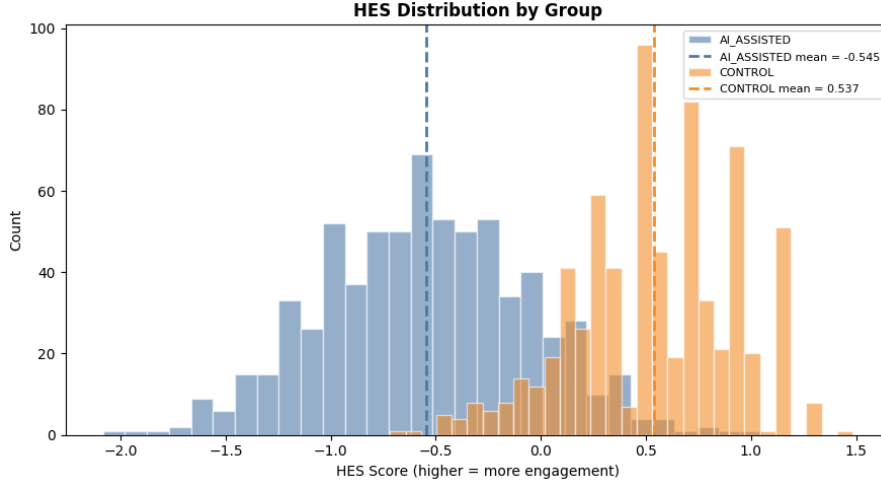


Figure 4.9: HES distribution by group. AI-assisted group mean = -0.545 . Control group mean = $+0.537$. The two distributions show minimal overlap.

4.6.3 COI and HES Validation

Manipulation check: A Mann-Whitney test [33] comparing HES between the two groups gave $U = 20,036$ and $p = 0.000$. The AI condition produced measurable and statistically significant behavioural differences between groups. The manipulation check passed [4].

Ablation check: COI was recomputed four times, dropping one component each time. Because COI is the mean of standardised scores, the mean of any ablated version is fixed near zero by construction, so comparing means would not test anything. The check was therefore run as a correlation between the full COI and each ablated COI. A correlation of 0.90 or above was treated as evidence that the index does not depend on any single component. Both the Pearson and the Spearman correlations are reported, because the Spearman value shows whether the ranking of participants stays the same, which is what matters if COI is ever used to flag individuals.

Dropped component	COI mean (ablated)	Diff from full	Stable
z_ai_ratio	0.0007	+0.0007	Yes
z_time_s1	0.0000	+0.0000	Yes
z_copy_rate	0.0007	+0.0007	Yes
z_inv_verify_rate	0.0007	+0.0007	Yes
Full COI	0.0000	—	—

Table 4.9: Ablation check results. All differences from full COI are below the 0.1 stability threshold [32].

All four components passed the stability threshold of 0.1. The maximum difference from the full COI mean when any one component was dropped was only 0.0007. COI was therefore stable and did not depend heavily on any single component [32].

One property of this check should be noted. COI is the mean of standardised components, so the mean of any subset of them also sits close to zero by construction. The very small differences in Table 4.9 therefore confirm that no component distorts the centre of the distribution, which is what the check was designed to establish, rather than testing how far individual participants would move in the index if a component were removed. A correlation between the full and the ablated index would answer that second question directly, and is recommended for any future use of these indices.

Weight sensitivity: Three weighting schemes were compared:

Weighting scheme	COI mean	Correlation with equal weights
Equal weights (used in COI)	0.0000	1.000
Double weight on verify_rate	0.0004	0.9647
Double weight on ai_ratio	0.0004	0.9704

Table 4.10: Weight sensitivity check. Correlations above 0.90 confirm equal weights are defensible [32].

Both alternative schemes correlated above 0.96 with the equal weight scheme, confirming that the choice of equal weights did not drive the results [32].

Construct separation: The Pearson correlation [53] between COI and HES for the AI-assisted group was $r = -0.361$, a weak to moderate negative relationship. The scatter plot showed data points in all four quadrants, confirming that high offloading did not always correspond to low engagement and that COI and HES measured different constructs [8].

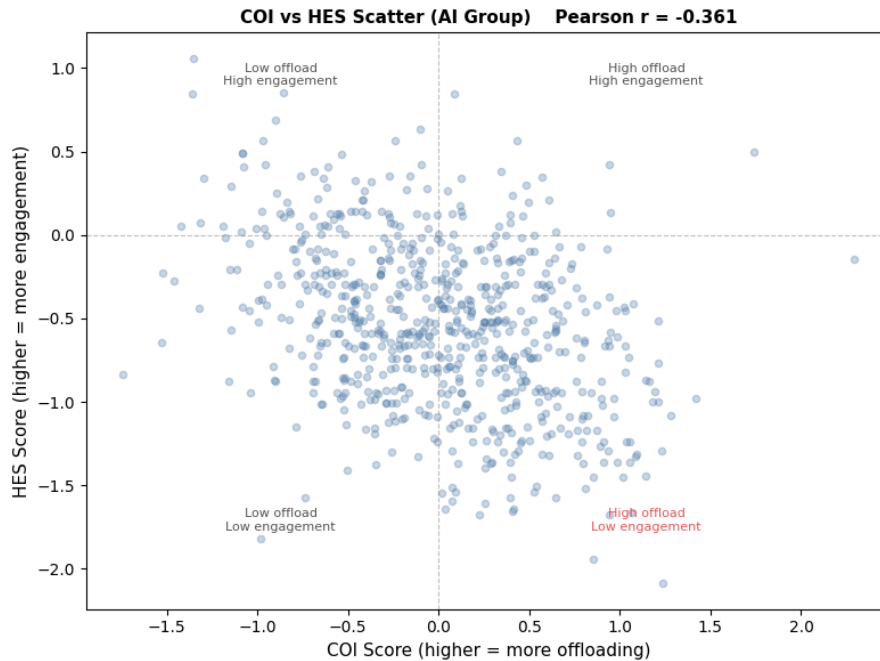


Figure 4.10: COI vs HES scatter plot for the AI-assisted group ($n = 687$). Pearson $r = -0.361$. Four quadrants confirm construct separation.

4.6.4 Validation Status of the Two Indices

Validation is a word that covers several different things, so it is worth stating exactly which of them were done here and which were not.

What was done. Both indices were built from fixed, written formulas, so a second researcher working from the same raw logs would produce the same numbers. COI was tested for dependence on any single component, and for sensitivity to the weighting scheme. COI and HES were tested against each other for construct separation, and the correlation of $r = -0.361$ shows they do not measure the same thing. A manipulation check confirmed that the two conditions produced different behaviour, though as noted in Section 4.6.3 part of that difference follows from how the indices were defined. Together, these show that the indices are reproducible, internally consistent in direction, and structurally stable.

What was not done. No expert panel reviewed the component set for content validity. No external or criterion measure of offloading was collected, so there is nothing to test convergent validity against. Neither index was administered twice, so there is no test-retest evidence. No factor analysis was run to test whether the components load on one underlying factor, and the low inter-component correlations reported in Table 4.13 suggest such a test would not support a single-factor structure. Most importantly, the one test of external usefulness that was attempted, the prediction of Session 2 independence, could not be carried out fairly because of the ceiling effect described in Section 3.10.3.

The status that follows. COI and HES are reproducible descriptive indices with demonstrated internal structure. They are not validated instruments in the psychometric sense, and this report does not claim they are. They are offered as a measurement framework that other researchers can apply and test, and testing them properly is one of the recommendations in Chapter 6.

4.7 Reliability Checks

4.7.1 Scale Reliability - Cronbach's Alpha

Cronbach's alpha [29] was computed for all four multi-item scales using the formula:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum \text{Var}(x_i)}{\text{Var}(X)} \right) \quad (4.2)$$

Table 4.11 shows the results.

Scale	Items	n	Alpha	Status
Effort S1	3	1,387	-0.033	FAIL
Engagement S1	3	1,387	-0.008	FAIL
AI Use S1	4	687	-0.008	FAIL
Independence S2	3	1,387	-0.050	FAIL

Table 4.11: Cronbach's alpha results. Threshold ≥ 0.70 [29].

All four scales failed the threshold of 0.70 [29]. All alpha values were negative, meaning items within each scale were essentially uncorrelated with each other. Item-level correlations across all four scales ranged from -0.07 to $+0.04$, all very close to zero. A well-designed scale should have item-level correlations above 0.40. The failure of all four scales meant that participants were not responding consistently across items within any single scale. Scale-based scores were therefore treated as descriptive only and not used as primary outcome measures.

4.7.2 Rubric Reliability - ICC

Double scoring covered 66.6 percent of all scored justification responses, giving 2,726 matched pairs. This is well above the 20 percent minimum normally recommended for reliability estimation [30]. Two raters (RATER-A1 and RATER-02) scored 2,726 matched pairs. ICC(2,1), the two-way random effects model for absolute agreement with single measures [30], was computed for each rubric dimension. Table 4.12 shows the results.

Dimension	ICC(2,1)	Pearson r	Mean diff	Status
Correctness	0.634	0.697	−0.094	PASS
Justification quality	0.063	0.154	−0.241	FAIL
Independence	−0.106	0.037	−0.351	FAIL

Table 4.12: ICC(2,1) inter-rater agreement results. Threshold ≥ 0.60 [30]. Mean diff = RATER-A1 minus RATER-02.

Only correctness passed the 0.60 threshold with ICC = 0.634 [30]. The Pearson correlation between raters on correctness was $r = 0.697$, confirming reasonable agreement. The mean difference of -0.094 indicated RATER-02 scored slightly lower than RATER-A1 on correctness, but this was a small and consistent difference.

Justification quality failed with ICC = 0.063 and Pearson $r = 0.154$. The mean difference of -0.241 showed RATER-02 scored substantially lower than RATER-A1 on justification quality. Independence failed with ICC = -0.106 , which indicated agreement no better than chance. The mean difference of -0.351 was the largest of the three dimensions. RATER-02 was systematically applying a stricter standard on both subjective dimensions, most likely due to insufficient rater calibration before scoring began [30].

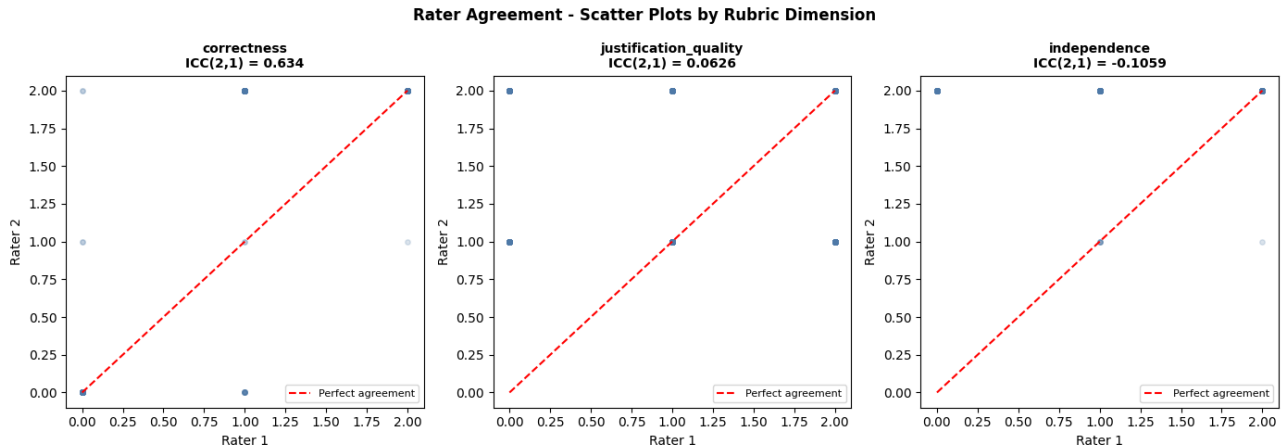


Figure 4.11: Rater agreement scatter plots by rubric dimension. Correctness shows tighter clustering along the diagonal. Justification quality and independence show wide scatter.

4.7.3 Inter-Component Correlations

Table 4.13 shows the Pearson correlations between all pairs of COI components for the AI-assisted group.

	z_ai_ratio	z_time_s1	z_copy_rate	z_inv_verify
z_ai_ratio	1.00	-0.04	+0.32	+0.26
z_time_s1	-0.04	1.00	+0.01	+0.05
z_copy_rate	+0.32	+0.01	1.00	+0.08
z_inv_verify	+0.26	+0.05	+0.08	1.00

Table 4.13: COI inter-component correlations (AI group, $n = 687$). All off-diagonal values are positive [32].

All off-diagonal correlations were positive, meaning all four components pointed in the same direction. The strongest pair was `z_ai_ratio` and `z_copy_rate` at $r = +0.32$, which was expected since participants who used AI more frequently also tended to copy more. `z_time_s1` was weakly correlated with the other components (all $|r| \leq 0.05$), suggesting that time spent in the tool was relatively independent of whether the participant copied or verified.

For HES, rubric independence and inverse AI ratio were strongly correlated at $r = +0.70$. This was expected because the control group scored high on both by design. Engagement self-report was nearly uncorrelated with the other two components at $r = 0.00$ and $r = -0.03$, suggesting self-reported engagement captured something different from rubric independence and delegation behaviour.

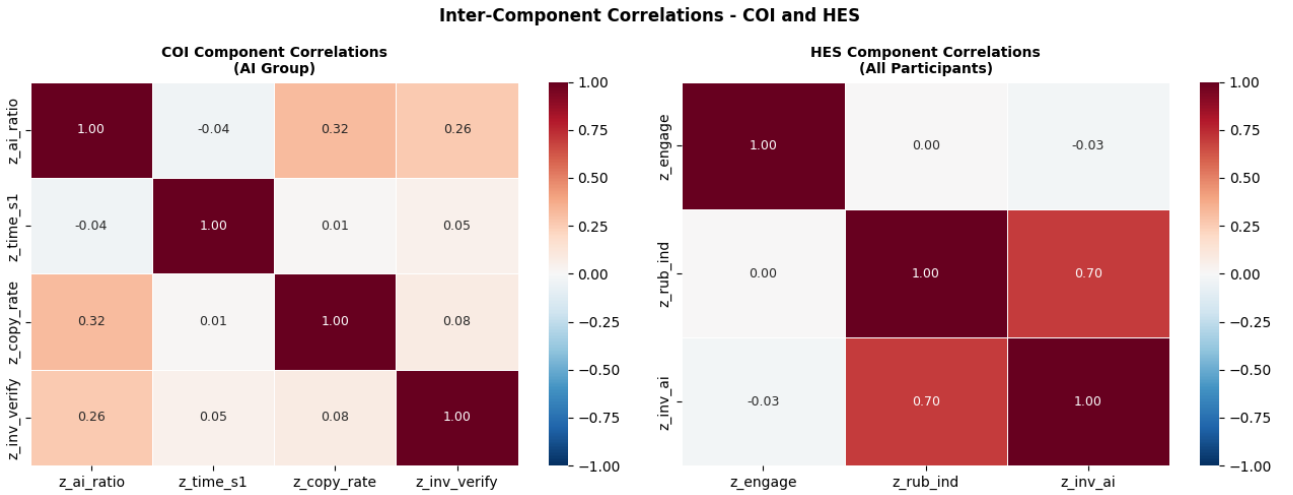


Figure 4.12: Inter-component correlation heatmaps for COI (left, AI group) and HES (right, all participants).

4.8 Statistical Analysis

4.8.1 Research Question 1 - Session 1 Results

Group comparisons used the Mann-Whitney test [33] since all primary endpoints were non-normally distributed. Effect sizes were reported as Cohen's d [12] with 95% confidence intervals from bootstrapping with 2,000 resamples [48]. The Benjamini-Hochberg correction [50] was applied to secondary p-values. Table 4.14 shows the Session 1 results.

Variable	AI	CTRL	Diff	d	Effect	p
Accuracy S1	0.971	0.969	+0.002	+0.026	Negligible	0.456
Time S1 (sec)	106.6	106.6	+0.05	+0.004	Negligible	0.852
Rubric Ind. S1	1.206	1.843	−0.638	−1.972	Very large	0.000
HES S1	−0.545	+0.537	−1.082	−2.474	Very large	0.000

Table 4.14: Session 1 group comparisons. Mann-Whitney tests [33]. Cohen’s d [12] with 95% CI from bootstrapping [48]. Diff = AI minus Control.

Both groups achieved almost identical accuracy in Session 1 ($d = +0.026$, negligible) [12]. Time per task was also virtually identical across both groups ($d = +0.004$). However rubric independence showed a very large difference. The control group scored 0.638 points higher than the AI-assisted group ($d = -1.972$). The 95% confidence interval for this difference was entirely below zero, ranging from -0.672 to -0.603 [48], confirming the difference was not due to chance. HES showed a similarly large difference ($d = -2.474$), with the AI-assisted group at -0.545 and the control group at $+0.537$. Both groups felt similarly engaged on the Likert scales but the control group produced substantially better independent reasoning and justification during Session 1.

These four results are best read together rather than one at a time. Accuracy and time per task were essentially identical between the groups, while rubric independence and HES both showed very large differences in the same direction. This pattern is informative. If the AI-assisted group had simply been working faster or more carelessly, we would expect their time per task to be shorter and their accuracy to be lower. Neither happened. They spent the same amount of time and reached the same answers, but the written reasoning they produced along the way was substantially weaker. This narrows down what was actually different between the groups. It was not speed, and it was not correctness. It was the amount of independent reasoning work the participant did on the way to the answer. The size of the rubric independence difference also deserves comment. A Cohen’s d of -1.972 is close to two full standard deviations [12]. In practical terms, this means the distributions of the two groups barely overlap on this measure: a typical AI-assisted participant scored lower on independence than roughly 97 percent of control participants. Differences of this magnitude are rare in behavioural research and usually indicate that the two groups were doing something structurally different, not just varying slightly around a common pattern. That is consistent with what the hint bank was designed to do, and it means the manipulation clearly worked.

4.8.2 Research Question 2 - Session 2 Results

Table 4.15 shows the group comparisons for the primary endpoints in Session 2.

Variable	AI	CTRL	Diff	d	95% CI	p
Accuracy S2	0.971	0.969	+0.002	+0.028	[−0.006, +0.011]	0.693
Rubric Ind. S2	1.860	1.854	+0.006	+0.026	[−0.019, +0.029]	0.611
Δ Accuracy	0.000	−0.000	+0.000	+0.004	[−0.011, +0.012]	0.970

Table 4.15: Session 2 primary endpoint comparisons [33, 12, 48]. Both primary endpoints are negligible and non-significant. Diff = AI minus Control.

When AI was removed in Session 2, both primary endpoints showed negligible and non-

significant differences [12]. The large rubric independence gap of $d = -1.972$ from Session 1 closed to $d = +0.026$ in Session 2. The confidence intervals for both primary endpoints crossed zero, providing no evidence of a group difference. Neither the AI-assisted group nor the control group showed meaningful change in accuracy between sessions. The delta accuracy for both groups was essentially zero.

The most informative way to read the Session 2 result is to compare it against the Session 1 result on the same measure. Rubric independence moved from $d = -1.972$ in Session 1 to $d = +0.026$ in Session 2. The AI-assisted group’s own mean rose from 1.206 to 1.860 while the control group stayed almost flat at 1.843 and 1.854. This is not a case of both groups declining together or of the measure losing sensitivity. The control group’s score barely moved, which shows the rubric was still measuring the same thing in the same way. What changed is that the AI-assisted group started producing reasoning at the same level as the control group as soon as the hint bank was withdrawn. This tells us that the weak Session 1 reasoning was tied to the presence of the tool rather than to any lasting change in the participant. When the tool was available, the AI group leaned on it. When it was gone, they reasoned for themselves at a normal level within the same sitting. This is a meaningful distinction: the study found a clear situational reliance effect and no detectable carry-over effect. It is also the reason the null result in Session 2 should not be reported on its own. Reported alone it sounds like AI had no effect at all, which the Session 1 data plainly contradicts.

4.8.3 Adjusted Regression Results

An adjusted ordinary least squares regression model [28] was fitted for both primary Session 2 endpoints to control for pre-existing participant differences. Table 4.16 shows the results.

Predictor	Accuracy S2		Rubric Ind. S2	
	Coef	p	Coef	p
Group (AI vs CTRL)	+0.002	0.607	+0.006	0.649
Warmup score	+0.000	0.922	−0.004	0.647
Education level	+0.003	0.296	−0.004	0.574
AI confidence	+0.000	0.926	−0.003	0.641
R^2	0.001		0.001	
n	1,387		1,284	

Table 4.16: Adjusted regression results for Session 2 primary endpoints [28]. No predictor reached statistical significance in either model.

Both models produced $R^2 < 0.001$, meaning the predictors together explained less than 0.1% of the variation in Session 2 performance. The group coefficient was +0.002 ($p = 0.607$) for Accuracy S2 and +0.006 ($p = 0.649$) for Rubric Independence S2. Both were negligible and non-significant [12]. None of the four covariates reached statistical significance in either model. The adjusted analysis confirmed that the null results from the unadjusted comparisons were not explained by pre-existing differences in baseline ability, education level or AI confidence [28].

4.9 Predictive Modelling - Research Question 3

4.9.1 Model and Feature Set

A Ridge regression model [46] was trained using six Session 1 features: COLS1, HES_S1, warmup_final_score, education_encoded, ai_confidence and group_binary. The regularisation parameter was set to $\alpha = 1.0$. All features were standardised using StandardScaler before fitting. The target variable was rubric independence S2 (primary target) and accuracy S2 (secondary target). The strict leakage rule meant no Session 2 variable was used at any point [47].

The modelling dataset contained 634 AI-assisted participants after dropping those with missing COLS1 or target values. The control group was excluded because COI was not available for them.

4.9.2 Cross-Validation Results

Grouped 5-fold cross-validation [47] was used, with participant code as the grouping variable, so that no participant contributed rows to a training fold and a test fold at the same time. The naive baseline was computed as the mean Session 2 independence score across the modelling sample, and standardisation was applied to the full feature matrix before the folds were formed, as described in the leakage controls in Chapter 3. Table 4.17 shows the fold-by-fold and mean results.

Fold	n test	MAE	RMSE
1	127	0.1851	0.2191
2	127	0.1869	0.2373
3	127	0.1875	0.2233
4	127	0.1819	0.2158
5	126	0.1693	0.1954
Mean		0.1821	0.2182
SD		0.0067	0.0136
Baseline		0.1793	0.2155

Table 4.17: Grouped 5-fold cross-validation results [47]. MAE and RMSE [51]. Baseline = predict the sample mean for every participant.

The model MAE of 0.1821 was slightly higher than the naive baseline MAE of 0.1793 [51]. The model did not beat a strategy of simply predicting the mean for everyone. Results were very consistent across folds (SD = 0.0067), confirming the finding was stable and not due to one bad fold.

4.9.3 Why a More Flexible Model Would Not Change This Result

A model failing to beat a naive baseline raises an obvious question. Would a different model have done better? A random forest or a gradient boosting model can capture curved relationships and interactions between features that Ridge regression cannot. It is worth setting out why fitting one was not expected to change the conclusion.

The reason is that the limit here lies in the target variable, not in the model. A model can only tell participants apart to the extent that the thing being predicted actually differs between them. Session 2 rubric independence was concentrated near the top of its range, with both group means at roughly 1.86 out of a possible 2.0, and Session 2 accuracy was at ceiling with almost every participant scoring 1.0. The predicted values from the Ridge model clustered in a narrow band between 1.85 and 1.90, which is what happens when the best available strategy is to predict the mean for everyone. On the secondary target the model error matched the baseline error to four decimal places, at 0.0511 for both. A match that exact is not a coincidence. It is what a model produces when predicting the average really is the best it can do.

No amount of model flexibility recovers variation that is not present in the outcome. A more flexible model applied to the same 634 rows would have more freedom to fit noise in the training folds, and on a target with this little spread the most likely result is a slightly worse cross-validated error rather than a better one. The direction of the COI coefficient, which matched what the offloading literature would predict, shows that the feature was not devoid of signal. What it lacked was a target with enough movement to work against.

The conclusion drawn from Research Question 3 is therefore not that Ridge regression was the wrong choice of model. It is that the conditions of this study did not permit the question to be tested fairly. A benchmark against several model families, run on data with a wider spread of Session 2 independence scores, is recommended in Chapter 6 as part of the follow-up work rather than presented here, where every model compared would be constrained by the same ceiling.

4.9.4 Feature Coefficients

Table 4.18 shows the feature coefficients from the full data refit.

Feature	Coefficient	Direction of association
warmup_final_score	+0.0066	Higher baseline associated with higher S2 independence
HES_S1	+0.0007	Higher engagement weakly associated with higher S2 ind.
group_binary	+0.0000	No group effect (expected, AI group only)
ai_confidence	−0.0014	Higher AI confidence weakly associated with lower S2 ind.
COI_S1	−0.0048	Higher offloading weakly associated with lower S2 ind.
education_encoded	−0.0053	Higher education weakly associated with lower S2 ind.

Table 4.18: Ridge regression feature coefficients from full data refit [32]. All values are very small.

Warmup score had the largest positive coefficient (+0.0066), meaning participants with higher baseline ability tended to score slightly higher on Session 2 independence [32]. COI_S1 had a negative coefficient of −0.0048, meaning higher offloading in Session 1 was linked to slightly lower Session 2 independence. The direction of this coefficient matches what the offloading literature would predict [2, 8], but the size was far too small to produce useful predictions. All six coefficients were very close to zero. The coefficient on the group variable calls for a comment. It is exactly zero rather than merely small. This is not a finding about the data. The modelling dataset contains only AI-assisted participants, because COI could not be computed for the control group, so the group variable takes the same value in every row. A variable that never changes carries no information and cannot contribute to a prediction, and a coefficient of exactly zero is what any correctly working regression returns for it. It was kept in the feature set so

that the specification stays the same as the one given in Chapter 3, and so that the model could be re-run on a combined dataset without changing the code if a reliance measure for the control group became available. Removing it would change no other coefficient and no prediction.

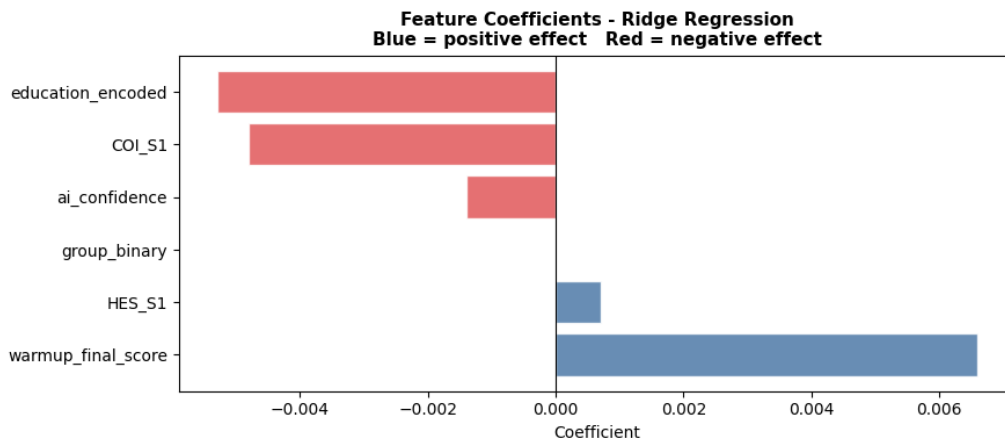


Figure 4.13: Feature coefficients from Ridge regression. Blue = positive association. Red = negative association. All coefficient values are near zero.

4.9.5 Predicted vs Actual

The predicted values clustered in a narrow horizontal band between 1.85 and 1.90 for all participants, regardless of their actual Session 2 independence score which ranged from 1.0 to 2.0. The model predicted almost the same value for every participant and could not capture the individual variation in the outcome [51].

The COI feature had a genuine spread (SD 0.578) and its coefficient pointed in the direction the offloading literature would predict. It is worth separating two different reasons a prediction model can fail, because they have very different implications. A model can fail because the features carry no information about the target, or it can fail because the target itself has too little variation to predict. The evidence here points firmly to the second. What it did not have was a target with enough movement to work against. Session 2 rubric independence was concentrated near the top of its range, with both group means at roughly 1.86 out of a possible 2.0, and Session 2 accuracy was at ceiling. The predicted values clustering in a narrow band between 1.85 and 1.90 is exactly the behaviour expected when a model finds that predicting the mean is nearly optimal. The secondary model result supports this reading directly. On accuracy S2, the model MAE and the baseline MAE were identical at 0.0511. A model cannot match a baseline that precisely by accident. It matches it when the best available strategy really is to predict the mean for everyone. Taken together, the two model results say that the Session 1 features could not be shown to predict Session 2 independence under these conditions, and that the conditions themselves, not the features, are the more likely reason.

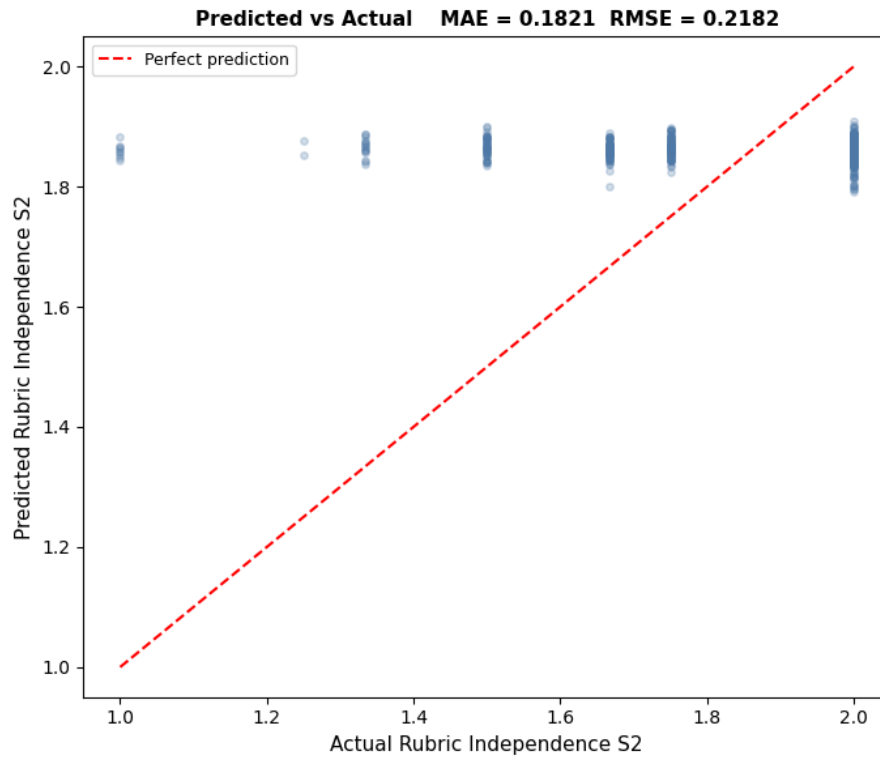


Figure 4.14: Predicted vs actual Session 2 rubric independence. MAE = 0.1821, RMSE = 0.2182. Predicted values cluster between 1.85 and 1.90 regardless of actual values.

4.9.6 Secondary Target - Accuracy S2

For the secondary target of accuracy S2, the model MAE of 0.0511 exactly matched the baseline MAE of 0.0511 [51]. The RMSE also matched at 0.0820. The model added no predictive value over predicting the mean for everyone. This was consistent with the ceiling effect where almost all participants scored 1.0 on accuracy, leaving virtually no individual variation to predict.

4.9.7 Subgroup Error Diagnostics

Table 4.19 shows the mean absolute error by education level and AI confidence.

Education level	MAE	AI confidence	MAE
Secondary	0.1867	Low (1–2)	0.1967
Diploma	0.1598	Medium (3)	0.1723
Degree	0.1794	High (4)	0.1641
Postgraduate	0.2183	Very high (5)	0.1942

Table 4.19: Subgroup error diagnostics by education level and AI confidence [32].

Postgraduate participants had the highest MAE at 0.2183 and diploma holders had the lowest at 0.1598. By AI confidence, medium confidence participants had the lowest error at 0.1723. The differences across subgroups were small and there was no clear systematic pattern suggesting the model was unfair to any particular group [32].

4.10 Robustness Checks

Four robustness checks were run on both primary Session 2 endpoints [28]. Table 4.20 summarises all results.

Check	Endpoint	n	d	p	Direction
Reference (main analysis)	Accuracy S2	1,387	+0.028	0.693	—
Reference (main analysis)	Rubric Ind. S2	1,284	+0.026	0.611	—
Check 1: Remove outliers	Accuracy S2	1,381	+0.035	0.584	Held
Check 1: Remove outliers	Rubric Ind. S2	1,278	+0.024	0.641	Held
Check 2: Complete rubric	Accuracy S2	1,246	+0.038	0.574	Held
Check 2: Complete rubric	Rubric Ind. S2	1,246	+0.018	0.640	Held
Check 3: MAIN site only	Accuracy S2	524	+0.027	0.923	Held
Check 3: MAIN site only	Rubric Ind. S2	488	−0.112	0.497	Changed
Check 3: ONLINE site only	Accuracy S2	863	+0.028	0.661	Held
Check 3: ONLINE site only	Rubric Ind. S2	796	+0.113	0.236	Held

Table 4.20: Robustness check summary [12, 28]. The table covers Checks 1–3; Check 4 is summarised in the text below.

The two most controlled checks produced consistent directions and conclusions for both primary endpoints. Check 1 removed 6 participants with extreme time values rather than capping them, reducing the sample to 1,381. Effect sizes remained negligible (Accuracy S2: $d = +0.035$; Rubric Ind. S2: $d = +0.024$).

Check 2 restricted to the 1,246 participants with complete rubric scores for both sessions, excluding 141 participants with any missing rubric values. Effect sizes again remained negligible (Accuracy S2: $d = +0.038$; Rubric Ind. S2: $d = +0.018$).

Check 3 split by site. At the ONLINE site ($n = 863$) both endpoints held their direction. At the MAIN site ($n = 524$) accuracy held ($d = +0.027$, $p = 0.923$) but rubric independence changed direction to $d = -0.112$ ($p = 0.497$). This direction change was at a negligible magnitude and was non-significant, most likely reflecting greater variability in the smaller MAIN site sample.

The education subgroup check (Check 4) showed direction changes in some subgroups but all effect sizes ranged from -0.102 to $+0.153$ and no subgroup comparison reached statistical significance. The core finding was stable: prior AI use in Session 1 did not produce a measurable difference in Session 2 independent performance for these tasks and this context [28].

4.11 Findings Dashboard

To make the findings accessible beyond the analysis notebooks, an interactive dashboard was built using Streamlit and Plotly. On startup the dashboard automatically loads the preprocessed participant-level dataset (table_a_wide.csv) from its working folder, recomputes COI and HES, and presents the study results across four sections: Study at a Glance, Who Participated, COI and HES, and Session 1 vs Session 2. All charts are interactive, allowing users to hover for exact values, zoom and filter without writing any code. The dashboard screenshots of each section are as below.

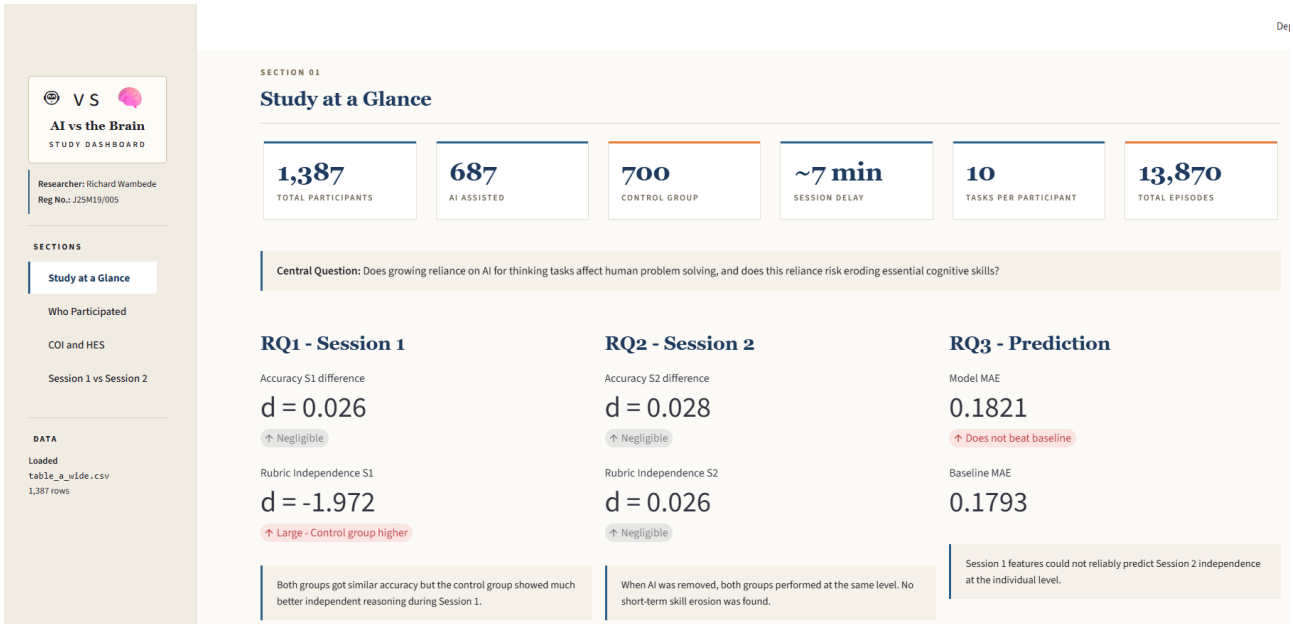


Figure 4.15: Dashboard landing view - Study at a Glance with key metric cards and RQ summaries.

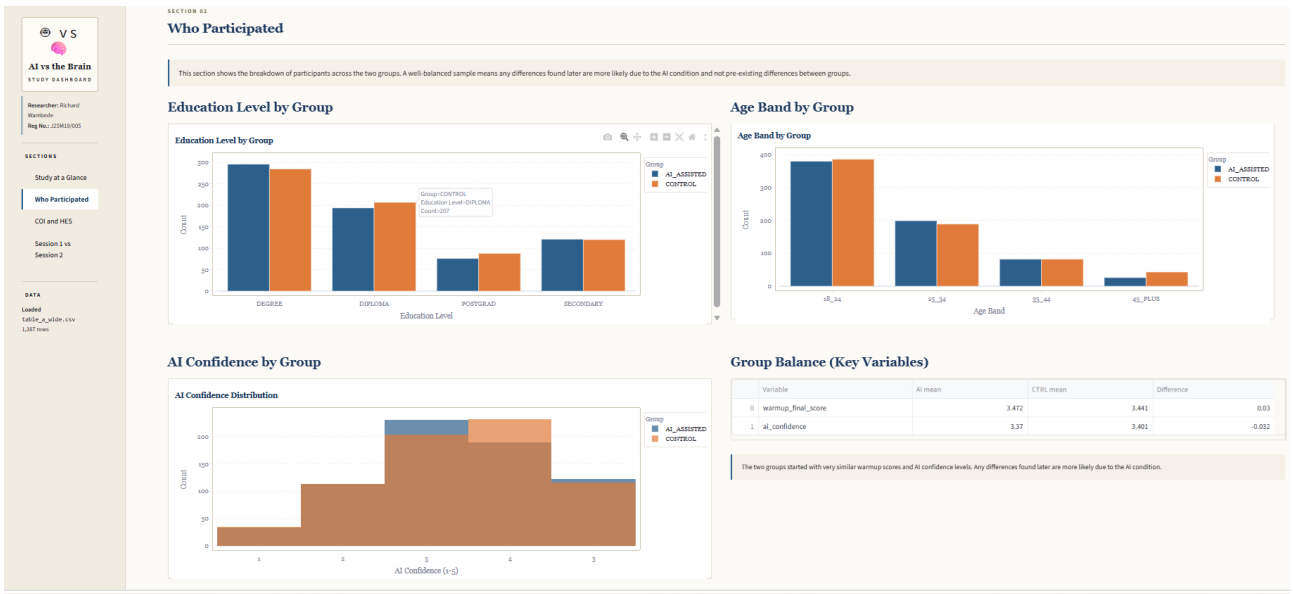


Figure 4.16: Dashboard screenshot - Who participated section.

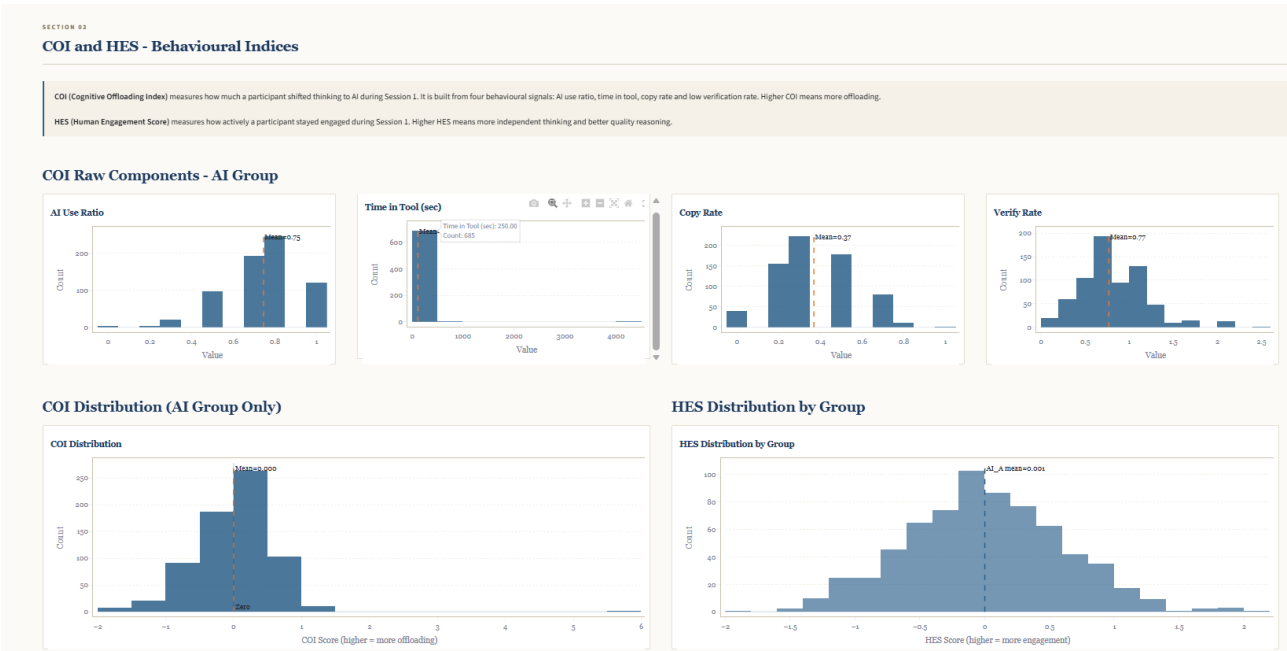


Figure 4.17: Dashboard screenshot - COI and HES.

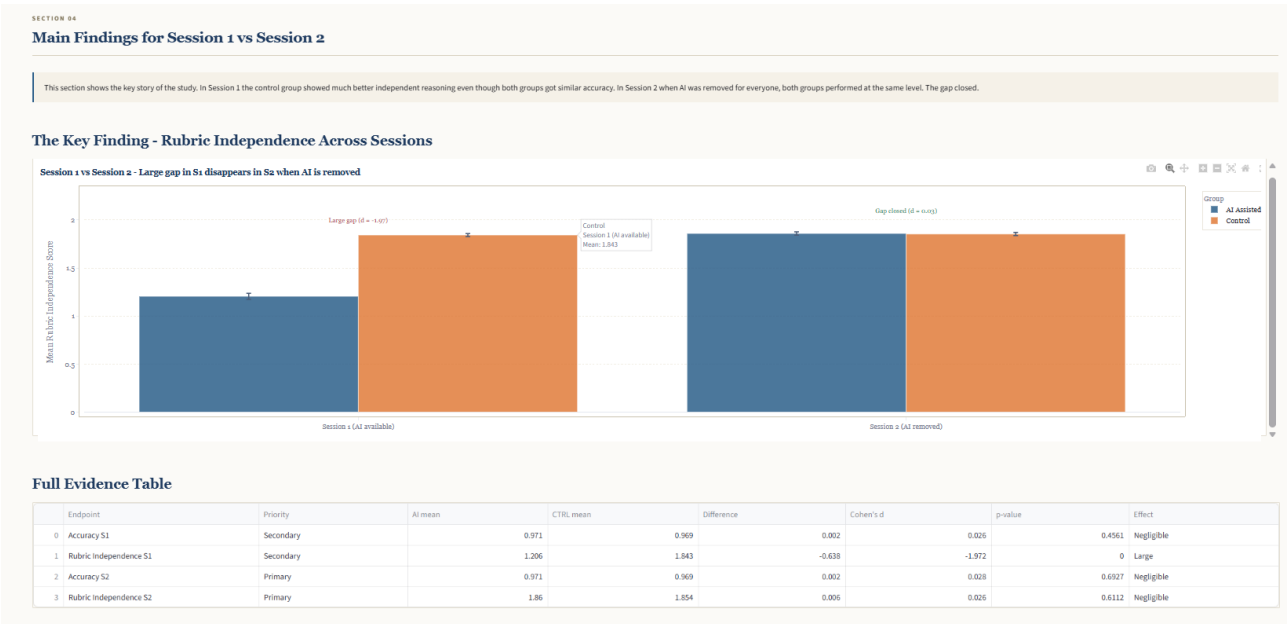


Figure 4.18: Dashboard - Session 1 vs Session 2 section showing the gap-closing chart and the full evidence table.

4.12 Summary of Findings

Table [4.21](#) presents the full evidence table for all primary and secondary endpoints.

Variable	Priority	AI	CTRL	d	Effect	p
Accuracy S1	Secondary	0.971	0.969	+0.026	Negligible	0.456
Rubric Ind. S1	Secondary	1.206	1.843	−1.972	Very large	0.000
HES S1	Secondary	−0.545	+0.537	−2.474	Very large	0.000
Accuracy S2	Primary	0.971	0.969	+0.028	Negligible	0.693
Rubric Ind. S2	Primary	1.860	1.854	+0.026	Negligible	0.611
Δ Accuracy	Secondary	0.000	0.000	+0.004	Negligible	0.970

Table 4.21: Full evidence table. Primary endpoints in bold. Mann-Whitney tests [33]. Cohen’s d [12] with 95% CI from bootstrapping [48]. BH correction on secondary p-values [50].

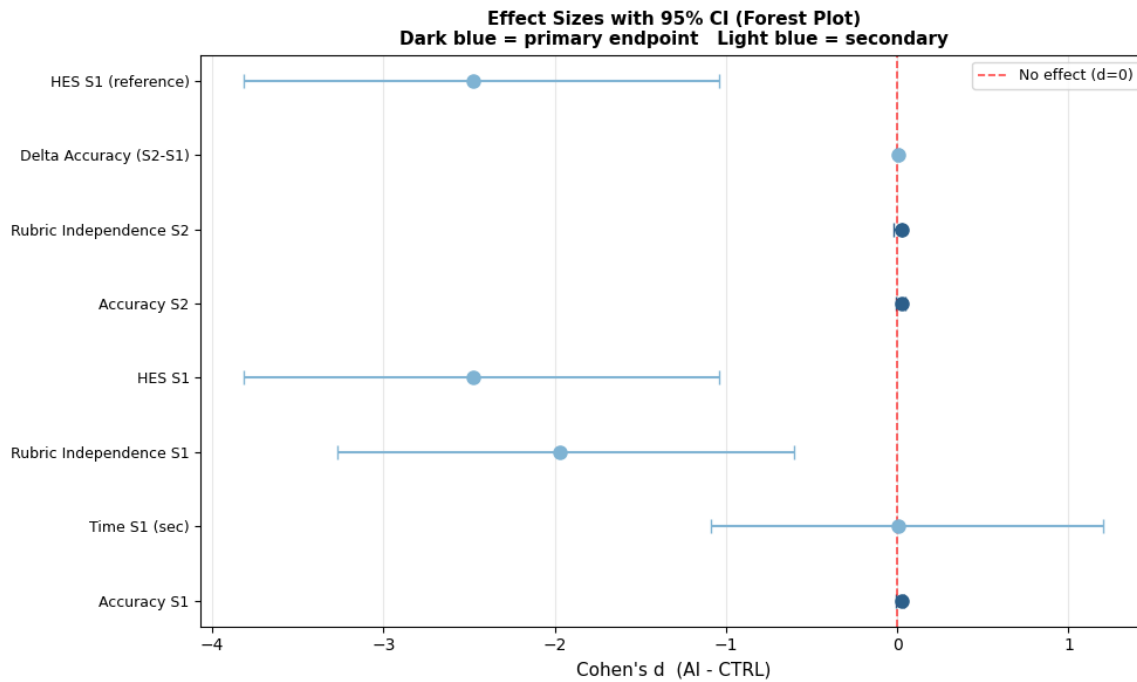


Figure 4.19: Forest plot of all effect sizes with 95% CI. Dark blue = primary endpoints. Light blue = secondary. Red dashed line = no effect ($d = 0$).

In addition to the statistical outputs, an interactive dashboard was developed to present the study findings to a non-technical audience. The dashboard recomputes COI and HES from the uploaded dataset on the fly and displays all key results through interactive Plotly charts. It is submitted as a supplementary file (dashboard.py) and can be run locally using Streamlit after installing the required dependencies.

The three research questions are answered as follows.

RQ1: Both groups achieved almost identical accuracy in Session 1 ($d = +0.026$, negligible) [12]. However the control group produced substantially better independent reasoning ($d = -1.972$, very large, 95% CI $[-0.672, -0.603]$). AI users got the same answers but the quality of their reasoning was much weaker.

RQ2: When AI was removed in Session 2, both primary endpoints showed negligible and non-significant differences. Accuracy S2: $d = +0.028$, $p = 0.693$, 95% CI $[-0.006, +0.011]$. Rubric Independence S2: $d = +0.026$, $p = 0.611$, 95% CI $[-0.019, +0.029]$ [12, 33]. The

adjusted regression confirmed these null results ($R^2 < 0.001$ for both models) [28]. No drop in short-term independent performance was found.

RQ3: The Ridge regression model [46] did not beat the naive baseline. Model MAE = 0.1821 versus baseline MAE = 0.1793 [51]. Session 1 features could not estimate Session 2 individual independence for this dataset and these tasks. The near-ceiling distribution of the outcome meant the question could not be tested under favourable conditions.

Chapter 5

Discussion

5.1 Introduction

This chapter interprets the findings from Chapter 4 in the context of the existing literature reviewed in Chapter 2. It discusses what the results mean, why they came out the way they did, and what they say about the central question of the study. It also discusses the limitations of the study honestly and what they mean for how the findings should be read.

The central question of this study was: how does growing reliance on AI for thinking tasks affect human problem solving, and does this reliance risk weakening essential cognitive skills? The three research questions guided the analysis toward a clear and measurable answer.

5.2 Why the Two Groups Reached the Same Answers by Different Routes

Research Question 1 asked how the dataset was constructed, useful features were engineered and COI and HES were computed in a way that clearly showed AI reliance and verification behaviour.

The most important finding from Session 1 was not about accuracy. It was about reasoning quality. Both groups achieved almost identical accuracy, with a negligible Cohen's d of 0.026. However the control group scored substantially higher on rubric independence with a very large effect size of $d = -1.972$. This meant that both groups found the right answers, but the control group produced much better independent reasoning and justification when doing so.

This finding is consistent with the distinction between substitution and augmentation described by Jarrahi et al. [1]. The AI-assisted group used AI to reach the correct answer but appeared to delegate the reasoning steps to the tool rather than working through the problem independently. The control group had no such support and therefore had to produce and justify their own reasoning at every step.

This pattern also confirms what Bućinca et al. [4] described as the core risk of AI-assisted decision making. When AI produces the intermediate steps, users may accept the output without engaging their own thinking. The rubric independence scores capture exactly this behaviour. A person who copied AI output and submitted it correctly scored low on independence even though their accuracy was high.

The HES results reinforced this interpretation. The AI-assisted group had a mean HES of -0.545 compared to 0.537 for the control group, a difference of more than one standard deviation.

This confirmed that the two groups differed substantially in their engagement and delegation patterns during Session 1, even though their self-reported effort and engagement on the Likert scales looked almost identical.

This difference between what participants actually did and what they said they did is not new to this study. Lee et al. [21] found that workers who used AI on the job often felt confident and engaged while doing so, even when the effort they put into their own critical thinking had gone down. Bućinca et al. [4] showed that people regularly accept AI output without doing the checking they think they are doing. The present study adds to this picture using both behavioural log data and rubric scores. The activity logs and the rater scores showed clearly that the AI-assisted group produced weaker independent reasoning in Session 1, yet when these same participants filled in the questionnaire, their self-reported engagement was almost the same as the control group. This finding supports the case for using behavioural logs and rubric scoring alongside questionnaires, because asking participants how engaged they felt would not have shown any difference between the groups at all.

This last point is important. The Likert scales for effort and engagement showed negligible differences between groups (differences of 0.002 and 0.012 respectively). This suggested that participants in the AI-assisted group did not feel less engaged or less effortful during Session 1, even though their behavioural logs and rubric scores told a different story. This is consistent with the literature showing that self-reported trust or effort does not reliably reflect actual behaviour [5]. It also explains why behavioural measurement through logs and rubric scoring was more informative than self-report alone in this study.

The COI and HES indices passed all internal structural checks. The manipulation check passed with a highly significant HES difference between groups. The ablation check confirmed COI was stable when any one component was removed. The weight sensitivity check confirmed equal weights were defensible. The construct separation analysis confirmed COI and HES measured different things, with a Pearson correlation of only $r = -0.361$. This supports the view that a person can offload heavily to AI and still feel engaged in the task, and that both indices were needed to describe the full picture [8].

The mechanism behind this pattern is worth setting out plainly. The tasks used in this study were structured problems with all the information needed contained in the task itself. For a problem of this kind there are two routes to a correct answer. A participant can work through the reasoning steps and arrive at the answer, or they can obtain the answer and then produce a justification to accompany it. Both routes end at the same place, which is why accuracy was identical. They differ in the amount of independent reasoning performed along the way, which is why rubric independence differed so sharply. The hint bank made the second route available to the AI-assisted group and not to the control group. This helps explain a finding that looks contradictory at first: the AI group spent the same amount of time per task as the control group (106.6 seconds for both) yet produced weaker reasoning. If they were taking a shortcut, why was it not faster? The most plausible explanation is that the time was spent differently rather than saved. Reading hint text, deciding whether to request the next level, and rewriting the hint content into a personal justification all take time. The total was similar; the cognitive work inside it was not. This is consistent with Bućinca et al. [4], who found that adding AI support changes how attention is allocated during a task rather than simply reducing the total effort spent. Comparing this against the wider literature shows both agreement and a point of difference. Deng et al. [11] reported in their meta-analysis that AI support generally improves performance while lowering perceived effort, and Kosmyna et al. [19] found reduced engagement during AI-assisted writing. The present study agrees with the engagement finding but not with the performance finding: accuracy did not improve here. The ceiling effect is the most likely reason. When almost everyone can already reach the correct answer unaided, AI

has no room to raise accuracy further, and its influence shows up only in the process measures. This suggests that studies reporting accuracy gains from AI may be working with tasks that are difficult enough for a gain to be visible, and that process measures may be the more sensitive instrument when tasks are easy.

5.3 Why the Independence Gap Closed So Completely in Session 2

Research Question 2 asked how Session 1 AI reliance affected delayed independent performance in Session 2 when AI was removed.

The answer is clear: it did not. Both primary endpoints in Session 2 showed negligible and non-significant differences between the groups. Accuracy S2 had a Cohen's d of 0.028 and rubric independence S2 had a d of 0.026, both with confidence intervals crossing zero and p -values well above 0.05. The adjusted regression confirmed these null results after controlling for baseline ability, education level and AI confidence. Four robustness checks confirmed the conclusions held across different outlier treatments, missingness approaches, collection sites and education levels.

This finding does not support the hypothesis that AI use reduces short-term independent performance. Prior AI use in Session 1 did not produce a measurable reduction in independent performance in Session 2 for these tasks and this context. This partially contradicts findings from Grinschgl et al. [6] and Storm and Soares [7], who found that cognitive offloading can reduce memory retention and independent performance. However there is an important contextual difference. Those studies tested retention over longer delays, typically hours or days. This study used a same-sitting design with a delay of approximately seven minutes. It is possible that a longer delay would have produced different results.

This short-delay result should be read alongside recent studies that used longer delays and found clearer costs. Bastani et al. [20] found that the harm of unrestricted AI practice only became visible on a later unaided test, not during practice itself, which suggests that a seven-minute gap may simply be too short for any cost to appear. Kosmyna et al. [19] similarly reported that reduced engagement during AI-assisted writing showed up most clearly in later recall, again pointing to a delayed rather than immediate effect. The present finding of no short-term drop in independent performance is therefore not in conflict with these studies, rather, it locates the current result at the very short end of the delay range and suggests that the more concerning effects reported elsewhere may operate over hours or days rather than minutes.

The most striking pattern in the Session 2 data was the closing of the gap. In Session 1, the control group scored 0.638 points higher on rubric independence with a very large effect size. In Session 2, that gap closed to only 0.006. The AI-assisted group improved their mean rubric independence from 1.206 in Session 1 to 1.860 in Session 2. This sharp improvement suggests that when AI was taken away, the AI-assisted group was still able to produce good independent reasoning almost immediately. This is consistent with the view that a seven-minute break is not long enough for any meaningful forgetting or skill degradation to occur [11].

Another possible explanation is that the task design itself limited the opportunity for any loss of short-term independent performance to appear. The tasks were structured and self-contained, meaning all the information needed to answer correctly was provided within the task itself. This design reduced the reliance on prior knowledge or long-term memory, which are the kinds of cognitive processes most likely to be affected by offloading over time [3]. A study using open-ended tasks that required genuine recall and elaboration might have found

different results.

The ceiling effect in accuracy also played a role. With almost all participants scoring 1.0 on accuracy, there was very little variation for the analysis to detect. Even if there had been a small real difference between groups in Session 2, the accuracy measure would not have been sensitive enough to pick it up. This was a design limitation that was unavoidable given the task type but must be considered when reading the results [11].

The central question this result raises is why a gap as large as two standard deviations disappeared within seven minutes. Three explanations are worth weighing. The first is that no lasting change ever occurred, and the Session 1 gap simply reflected the tool being present rather than anything happening inside the participant. The second is that a lasting change occurred but recovered almost immediately once the participant had to work unaided. The third is that a lasting change occurred and persists, but the Session 2 measure was not sensitive enough to detect it. The data available here favour the first explanation but cannot rule out the third. Favouring the first is the fact that the control group's scores barely moved between sessions (1.843 to 1.854), which shows the rubric was stable and was not simply becoming easier to satisfy. Against ruling out the third is the ceiling problem: Session 2 independence scores for both groups sat near the top of the scale at about 1.86 of a possible 2.0, leaving little room for a group difference to appear even if one existed. A more demanding Session 2 task set would be needed to separate these two possibilities. The seven-minute gap is the single most important reason to be cautious about generalising this result. Bastani et al. [20] found that harm from unrestricted AI practice appeared only on a later unaided test, and Grinschgl et al. [6] and Storm and Soares [7] both measured offloading costs over longer intervals. The mechanism those studies point to is consolidation: what is not practised is not retained, and retention is measured over hours or days, not minutes. Seven minutes is too short for forgetting to operate. This study therefore does not contradict that body of work. It measures a different thing: immediate independent capability rather than retention. Reporting it as evidence that AI does not harm learning would be a misreading of what the design can support. There is also a design feature specific to this study that may have made the gap easier to close. The hint bank required participants to write their own final answer at every assistance level. Even a participant who used the highest hint level still practised the act of composing a justification. This is a meaningfully different experience from copying a complete AI answer, which is what a participant using an unrestricted tool would be able to do. It is plausible that the design preserved just enough independent practice to prevent any carry-over cost from appearing, and that a study without that constraint would produce a different Session 2 result.

One feature of the rubric must be stated here. The independence dimension was defined against evidence of copying from hint content, and control group participants, who had no hints, were scored at the top of the scale by default unless there was specific evidence of external copying. In Session 2 no participant had access to hints, so that default applied to both groups. Part of the Session 1 gap and part of its closure in Session 2 therefore follow from how the dimension was defined rather than from a change in participant behaviour alone. The behavioural log measures, which do not depend on rater judgement, are the stronger evidence for the Session 1 difference.

5.4 Why Session 1 Behaviour Could Not Predict Session 2 Independence

Research Question 3 asked whether Session 1 features alone could predict Session 2 independence risk.

The answer is also no. The Ridge regression model did not beat a naive baseline of predicting the mean value for everyone. The model MAE of 0.1821 was slightly higher than the baseline MAE of 0.1793. The predicted vs actual plot confirmed the model could not distinguish between participants who scored high or low in Session 2. It predicted almost the same value for every participant.

This result is consistent with the Session 2 group comparisons. When there is very little real variation in the outcome, prediction models will also struggle to find signal in the features [51]. The ceiling effect in accuracy and the very small group difference in rubric independence in Session 2 meant there was simply not enough variation to predict.

COLS1 had a negative coefficient of -0.0048 . This means that within the AI-assisted group, participants with higher offloading scores in Session 1 tended to have slightly lower independence scores in Session 2. This is a predictive association only. COI was measured, not manipulated, so this coefficient cannot be read as the amount by which offloading reduced independence. The direction of this link matches what earlier studies would predict, but the size was too small to be of any practical use [2, 8]. This suggests that the relationship between Session 1 offloading and Session 2 independence may exist but is too weak to detect in a same-sitting design with structured tasks.

These results are also consistent with the literature. Tian and Zhang [8] found that AI dependence was linked to weaker critical thinking, but their study was correlational and did not use a controlled delayed endpoint. The predictive relationship between Session 1 offloading and Session 2 independence may be more visible in a study with a longer delay, more varied tasks and greater outcome variability.

The prediction model did not beat a naive baseline, with a mean absolute error of 0.1821 against the baseline's 0.1793. The important question is not whether the model failed but what its failure is evidence of. There are two possibilities that lead to opposite conclusions, and separating them matters. The first possibility is that Session 1 behaviour genuinely carries no information about Session 2 independence. If this were true, it would be a substantive finding: it would mean that how a person uses AI on one occasion tells us nothing about how they will perform unaided afterwards. The second possibility is that the information exists but the study conditions prevented it from being detected. The evidence points to the second. COI had a real spread across participants (SD 0.578), its coefficient carried the sign that the offloading literature predicts [2, 8], and the target variable was severely restricted, with both group means near 1.86 out of 2.0 and the accuracy target at ceiling. A model has nothing to predict when the outcome is nearly the same for everyone. The secondary model makes this argument concrete. On accuracy S2 the model error and the baseline error were identical at 0.0511, to four decimal places. That level of agreement is what happens when predicting the group mean genuinely is the best available strategy, which is the expected outcome when a variable is at ceiling. Read together, the two models say that the study was not in a position to test the predictive question properly, rather than that the question has been answered in the negative. This has a direct implication for how the COI index should be regarded. The failure of the model is not evidence that COI lacks predictive value. It is evidence that this design, with a seven-minute gap and a near-ceiling outcome, could not evaluate that value. A study with a longer delay and a task set producing a wider spread of independence scores would give COI a fair test. Until such a study is run, COI should be described as a reproducible descriptive measure of offloading behaviour whose predictive usefulness remains untested.

5.5 What the Self-Report and Behavioural Measures Disagreeing Tells Us

The disagreement between the two kinds of measure is one of the more useful findings of this study, and it is worth stating what follows from it. The Likert scales showed differences of 0.002 for effort and 0.012 for engagement, both negligible. The behavioural logs and rubric scores on the same participants in the same session showed a difference of nearly two standard deviations. Both cannot be describing the same underlying quantity. The most likely reading is that they are measuring different things and both are reporting accurately. The Likert items asked participants how hard they were trying and how involved they felt. There is no reason to think the AI-assisted group was trying less hard; they were engaged with a task and working at it for the same length of time. What the rubric measured was not effort but the independence of the reasoning produced. A participant can invest genuine effort into understanding and restating a hint and still not be reasoning independently. Under this reading the self-report is not wrong, it is answering a different question. This carries a practical warning for research design in this area. A study relying on self-reported engagement alone would have concluded from this dataset that AI made no difference in Session 1. That conclusion would have been wrong, and the error would not have been detectable from within the self-report data. This supports the argument made by Mehrotra et al. [5] that attitude and perception measures are poor substitutes for behavioural measures in human-AI research, and it is a concrete demonstration of the size of the discrepancy rather than a general caution.

5.6 What These Findings Mean for AI Use in a Ugandan Higher Education Setting

This study was conducted in Kampala with university students and working adults, and several features of that setting shape how the results should be read. Considering them explicitly is important because most of the studies reviewed in Chapter 2 were carried out in North American, European or East Asian contexts [11, 20, 19], and it should not be assumed that findings transfer directly. The first relevant feature is the pattern of AI access. Participants reported a mean AI confidence of about 3.4 out of 5, indicating familiarity rather than expertise. In settings where AI tools have been in routine academic use for longer, participants may have developed more established habits of when to rely on AI and when not to. The reliance behaviour observed here may therefore reflect a relatively early stage of adoption.

This matters for how far the Session 1 result can be carried. A group with more settled AI habits might show a stronger reliance effect. Or it might show a weaker one, if longer use has taught them to check more consistently. The second feature is the data collection environment itself. Recruitment took place during a university holiday period, which required assembling participants in computer lab batches at the MAIN site and offering incentives to secure attendance. This has two consequences worth stating. It means the sample is drawn from those able and willing to attend under those conditions, and it means participants were completing the tasks in a session arranged for the study rather than in the course of their normal academic work. Reliance behaviour during a paid research session may differ from reliance behaviour when a student is completing an assessment that affects their grade, where the incentive to obtain a correct answer quickly is considerably stronger.

The third feature concerns the practical relevance of the findings. In Uganda and comparable settings, AI tools are entering higher education at a point when assessment practice still relies heavily on written work produced outside supervised conditions. The Session 1 finding is

directly relevant to this. It shows that a student can produce fully correct answers while performing substantially less independent reasoning, and that marking the answer alone would not reveal this. An institution assessing only final answers would see no difference between the two groups in this study. This is an argument for including reasoning quality in assessment criteria rather than an argument for restricting AI access, and it applies with particular force where large class sizes make detailed marking difficult. The limits on how far any of this can be carried beyond Kampala are set out in Section [5.7.1](#).

5.7 What This Study Can and Cannot Conclude

Bringing the strands together, it is useful to state plainly what this study establishes and what it leaves open. Being explicit about the boundary protects the findings from being over-read in either direction. The study establishes three things. First, providing structured AI assistance during a problem-solving session was associated with a substantially lower level of independent reasoning in the work produced during that session, with an effect size of nearly two standard deviations. Because the AI condition was set by the study rather than observed, and because the two groups were balanced on baseline warm-up score, education, age band and AI confidence, the most likely explanation for this difference is the AI condition itself. The claim is limited to the participants, tasks and setting of this study. It is not a claim that AI assistance reduces independent reasoning in general. Second, that difference did not persist into an unaided session seven minutes later, on tasks of comparable structure. Third, the behavioural indices COI and HES can be constructed reproducibly from activity logs and pass their internal validation checks, giving a measurement framework that other researchers can apply. The study does not establish several things that it might be mistaken for establishing. It does not show that AI use is harmless to learning, because it did not measure learning or retention over any meaningful interval. It does not show that COI has no predictive value, because the near-ceiling outcome variables prevented that question from being tested fairly. It does not show that self-reported engagement is worthless, only that it was not sensitive to the difference the behavioural measures detected. And because two of the three rubric dimensions failed inter-rater agreement, the independence scores rest on a single rater’s judgement and carry more uncertainty than the accuracy scores do [\[30\]](#). The most defensible summary is therefore narrower than either an alarming or a reassuring headline would suggest. Structured AI assistance changed how participants worked on a task without changing whether they got it right, and did not reduce what they could do unaided a few minutes later. Whether the same holds after a day, a week, or a semester of habitual use is the question this design was not able to reach, and it is the question that matters most for policy.

Study	Design	Delay	Their finding	How this study compares
Bastani et al. (2024) [20]	Field experiment, maths practice	Days	Practice with AI reduced later unaided performance	No such drop here, but delay was minutes rather than days
Kosmyna et al. (2025) [19]	Repeated essay sessions with EEG	Weeks	Reduced engagement built up over sessions	One session only; cannot observe build-up
Gerlich (2025) [2]	Survey	None	Higher AI use linked to lower critical thinking	Measured behaviour, not self-report; same direction in Session 1
Deng et al. (2024) [11]	Meta-analysis	Mixed	Better performance with AI present, mixed learning effects	Consistent with Session 1 accuracy result here
This study	Two-session experiment, $n = 1,387$	~7 min	Large drop during AI use, no drop once AI removed	-

Table 5.1: This study’s findings set against comparable published work.

The pattern in Table 5.1 is worth stating directly. The studies that found a lasting cost to AI use all tested after a delay measured in days or weeks. This study tested after a delay measured in minutes and found no cost. The results do not contradict one another. They sit at different points on the same time axis, and the difference between them is exactly the gap that the recommendations in Chapter 6 ask future work to close.

5.7.1 Boundaries of Generalisation

The direct answer is no. These findings do not generalise to all universities, and three things stand in the way.

The first is how the sample was chosen. Participants were recruited by convenience from two sites in one city [31]. There was no sampling frame, no random selection from a defined population and no weighting. A convenience sample supports a comparison between two conditions inside the study. It does not support any statement about how common a behaviour is outside it.

The second is who the participants actually were. The sample was not drawn from a university at all in the strict sense. Recruitment covered universities, vocational training institutes and workplace groups, and it took place during a holiday period when many students were away on internships. Even if the results did generalise to the group they came from, that group is a mixture of students and working adults rather than a student body.

The third is the setting. Participants worked through structured tasks in a session arranged for research, with no consequence attached to the result. University work is graded, unsupervised and open-ended, and the incentive to obtain a correct answer quickly is much stronger there. Reliance behaviour under those conditions may look quite different.

It is still worth separating what does travel from what does not, because a blanket refusal to generalise is as unhelpful as an overclaim.

What travels. The measurement approach travels. The COI and HES formulas, the two-session design with a no-AI endpoint, and the analysis pipeline can be applied anywhere the same kinds of log are available. Any institution that can record AI requests, copying and verification can compute these indices on its own students.

What travels with care. The Session 1 finding, that a group with AI assistance produced correct answers with substantially less independent reasoning, is a comparison made inside the study between two groups that were balanced at the start. It is not a population estimate. The mechanism it points to is a general one and the effect size was very large, so it is reasonable to expect a similar direction elsewhere. The size of the effect is another matter and should not be carried across.

What does not travel. Every number that describes the sample rather than the contrast between conditions. The retention rate of 94.1 percent, the mean AI confidence of 3.4, the mean COI and the share of participants who verified their answers all describe these 1,387 people. None of them should be quoted as a figure for Ugandan universities.

Setting the boundary this way is not a weakness of the study. It is what allows the parts that do hold up to be used with confidence.

5.8 Reliability Limitations and Their Impact

The reliability checks produced results that must be discussed honestly because they affect how much trust can be placed in the scale-based and rubric-based findings.

All four Likert scales failed Cronbach's alpha with values close to zero or negative [29]. This meant the items within each scale were essentially uncorrelated. A person who rated one effort item highly did not consistently rate the other two effort items highly. This is a serious reliability problem. It means the scale scores cannot be treated as reliable measures of their intended constructs. They were therefore used as descriptive information only and not as primary outcome measures in the statistical analysis.

The most likely explanation for the scale failures is that participants interpreted each item independently rather than as part of a coherent scale. This can happen when items are too similar in wording, when participants answer quickly without reading carefully, or when the construct being measured is genuinely multidimensional and the items capture different aspects of it [29]. In a study conducted under time pressure with incentives for participation, quick or inattentive responding is a real possibility.

It is worth setting out exactly which results depend on which failed measure, because the failures do not affect everything equally.

Result	Measure it rests on	What the reliability failure means for it
Accuracy S1 and S2	Auto-scored answer key	Not affected. Scoring is automatic and needs no rater judgement.
Rubric correctness	Rubric, ICC = 0.634	Passed. Can be used with normal confidence.
Rubric independence S1 and S2 (primary endpoint)	Rubric, ICC = -0.106	Rests on the primary rater alone. The second rater agreed no better than chance on this dimension.
Rubric justification quality	Rubric, ICC = 0.063	Same problem. Reported as descriptive only.
COI	Behavioural logs	Not affected by the rater or scale failures. The logs record actions, not judgements.
HES	Two of three components affected	Contains self-reported engagement, from a scale with $\alpha = -0.008$, and rubric independence, which failed ICC.
Likert scale comparisons (effort, engagement, AI use, independence)	Four scales, all α below zero	Descriptive only. Not used as outcomes.

Table 5.2: Which results depend on which measure, and what the reliability failures mean for each.

The table shows that the failures are concentrated in the measures built from judgement and self-report, and that the two most objective measures, auto-scored accuracy and the behavioural logs, are untouched. This is worth saying plainly, because it is the reason the study still has usable findings.

The position of HES needs stating clearly. Two of its three components come from measures that failed a reliability check. Self-reported engagement came from a scale with an α of -0.008 . Rubric independence in Session 1 came from the dimension with the worst inter-rater agreement of the three. The third component, the inverse AI use ratio, is a log-based measure and is not affected, but it is also fixed at 1.0 for every control participant by design. Taken together, this means HES cannot be presented as a validated measure of engagement. It is a descriptive summary that combines one reliable log signal with two signals that carry known measurement problems. Every use of HES in this report should be read on that basis, and the term "validated" is not used for it anywhere in this report.

For the rubric ICC results, only correctness passed the 0.60 threshold with $ICC = 0.634$ [30]. Justification quality and independence both failed with ICC values of 0.063 and -0.106 respectively. The negative ICC for independence is particularly concerning as it indicated that RATER-02 was scoring almost independently of RATER-A1 on that dimension. The mean differences showed that RATER-02 consistently scored lower than RATER-A1 on the subjective dimensions, by 0.241 points for justification quality and 0.351 points for independence. This suggested insufficient calibration between the two raters before scoring began [30]. There is a structural reason for the independence dimension performing worst of the three. Control group responses, and all Session 2 responses, were scored at the top of the scale by default in the absence of evidence of copying. This compresses the distribution towards the ceiling, and an intraclass correlation coefficient falls towards zero when the variance between participants is small, regardless of how consistently the raters applied the rule. The ICC of -0.106 therefore reflects the shape of the scoring rule as well as the level of rater agreement.

The practical implication is that the rubric independence scores used in the primary analysis and in HES must be interpreted with caution. They reflect the primary rater’s judgements only. Whether those judgements were accurate and consistent cannot be confirmed for the two dimensions that failed ICC. The correctness dimension is more trustworthy since it is the most objective and it passed the ICC threshold.

There is a further consequence that runs in the opposite direction to the usual reading of a reliability failure, and it needs stating because it affects the main result. When a measure is unreliable, the differences it records get pushed towards zero. Rater noise adds variation that belongs to the rater rather than to the participant, and that extra variation makes any real group difference harder to see. This means the null result in Session 2 is weaker evidence of no difference than it would be if the measure had passed. Some part of that null could be measurement noise rather than a true absence of difference.

There is a limit to how far that argument can be pushed. The same rubric dimension, scored by the same rater under the same conditions, produced a difference of $d = -1.972$ in Session 1. A measure made only of noise cannot produce an effect of that size. The independence rubric was clearly able to separate the two groups when a large difference was present. What cannot be ruled out is that it was too noisy to detect a Session 2 difference that was small but real. The honest statement is that Session 2 showed no difference large enough for this measure to detect, which is not quite the same as showing no difference at all.

Despite these limitations, the rubric independence findings were consistent in direction with what theory and the behavioural log data predicted. The control group scored higher on independence in Session 1, which matched the expectation from the substitution versus augmentation framework [1]. The null result in Session 2 was also internally consistent across multiple robustness checks. This internal consistency gives some confidence that the pattern is real even if the precise rubric scores carry uncertainty.

5.9 Practical Implications

The implications below follow from what was observed in one sample, on one set of structured tasks, over one short delay. They are offered as reasonable steps to consider rather than as conclusions that hold everywhere. The findings have several practical implications for education, workplace settings and AI governance.

In education settings, the Session 1 findings suggest that AI access during problem solving changes the quality of reasoning produced, even when it does not change the accuracy of answers [4]. This means teachers and institutions who evaluate students only on final answer

accuracy may be missing an important signal. A student using AI heavily can achieve high accuracy without developing the reasoning skills that the task was designed to build.

The study supports the case for designing AI-assisted tasks in ways that require students to produce their own justifications and verify AI output before submitting [4]. COI and HES provide a practical way to measure whether students are doing this or simply copying AI output. Even if these indices need further validation in educational settings, they represent a more informative picture of AI use than accuracy alone.

In workplace settings, the finding that Session 1 AI use did not harm Session 2 performance over a seven-minute delay is reassuring but should not be over-interpreted. The study did not measure what would happen after one day, one week or one month. The concern expressed in the literature about AI reducing independent capability over time [2, 7] cannot be ruled out based on a same-sitting test alone.

For AI governance, the COI and HES scoring rules produced by this study offer a compact and reusable measurement package [9, 10]. Instead of relying on self-report surveys or general attitudes toward AI, organisations can measure actual reliance behaviour from activity logs and use that data to monitor AI use patterns in their teams or classrooms.

5.10 Limitations

Several limitations affected this study and must be considered when interpreting the findings.

Session delay: The most important limitation is the seven-minute gap between the two sessions. A longer gap of a day or more would have tested overnight retention and would have given any weakening of skill more time to show up [6, 7]. The same-sitting design can only speak to very short-term independence. It cannot support any claim about longer-term effects.

Ceiling effect: Task accuracy was very high for almost all participants in both sessions, with most scoring 1.0. This left very little variation in the primary accuracy endpoint and reduced the sensitivity of both the group comparisons and the prediction model [11]. Future studies should use tasks that produce more spread in accuracy to give the analysis more to work with.

Scale reliability failures: All four Likert scales failed Cronbach's alpha [29]. This means self-reported effort, engagement, AI use and independence scores could not be used as reliable primary measures. The study relied more heavily on behavioural log features and rubric scores as a result.

Rubric ICC failures: Two of three rubric dimensions failed inter-rater agreement [30]. Justification quality and independence scores reflect only the primary rater's judgements and carry uncertainty. More thorough rater calibration before scoring would have improved reliability.

Task type: The tasks were structured and self-contained, meaning all needed information was provided within the task. This reduced the reliance on recall and elaboration, which are the cognitive processes most likely to be affected by offloading over time [3]. Results may differ with more open-ended or knowledge-intensive tasks.

Sample: The sample was drawn from Kampala, Uganda and may not represent other populations, cultures or educational settings [31]. Participation incentives were provided due to time constraints, which may have introduced variability in participant motivation.

Prediction model scope: The predictive model only covered the AI-assisted group because

COI required AI use data. The control group could not be included, limiting the model to 634 participants.

Data collection timeline: Data collection was stopped at 1,474 enrolled participants due to time constraints even though participants were still attempting to fill in forms at the time of closure. This may have introduced a small selection bias in favour of faster or more motivated participants.

Chapter 6

Conclusion and Recommendations

6.1 Introduction

This chapter brings together the findings from the analysis and discussion to answer the three research questions and the central question of the study. It also provides practical recommendations for educators, organisations and future researchers based on what was found. The three research questions are answered in the order given in Table 3.1, so that each specific objective, the question drawn from it, the analysis that addressed it and the answer reached can be read as one line of argument running from Chapter 1 to here.

6.2 Answers to the Research Questions

6.2.1 Research Question 1

How was a good dataset constructed, useful features engineered from raw behavioural logs, and COI and HES computed in a way that clearly showed AI reliance and verification behaviour, and how reliably did the two indices measure what they claim to measure?

A two-session labelled dataset was successfully constructed from nine raw data files collected across the two sites in Kampala, Uganda. After filtering to completers only, the final dataset had 1,387 participants across two integrated tables. Table A held one row per participant and Table B held one row per task episode, giving 13,870 episode-level records.

Four behavioural features were engineered from Session 1 activity logs for the AI-assisted group: AI use ratio, time in tool, copy rate and verification rate. These were combined into the Cognitive Offloading Index using deterministic z-score formulas. Three engagement signals were combined into the Human Engagement Score in the same way. Both indices were checked for internal stability through ablation and weight sensitivity analysis, and for construct separation against each other. COI was stable across all ablation and weighting combinations. COI and HES had a weak correlation of $r = -0.361$, confirming they measured different constructs.

The manipulation check confirmed that the AI condition produced measurable behavioural differences. The AI-assisted group showed much weaker independent reasoning in Session 1 ($d = -1.972$) and a much lower HES ($d = -2.474$) compared to the control group, even though accuracy was almost identical ($d = 0.026$). This confirmed that COI and HES captured meaningful variation in how participants used AI during Session 1 [1, 4].

On the reliability of the two indices: the COI and HES indices were internally valid and stable. However the Likert scales failed Cronbach's alpha and two rubric dimensions failed ICC. These

failures mean the scale-based and subjective rubric scores must be treated with caution [29, 30]. The behavioural log features feeding into COI and the correctness rubric dimension, which passed ICC, can be used with more confidence.

6.2.2 Research Question 2

How did Session 1 AI reliance affect delayed independent performance in Session 2 when AI was removed?

The answer is that Session 1 AI reliance did not produce a measurable effect on Session 2 independent performance for these tasks and this context. Both primary endpoints in Session 2 showed negligible and non-significant differences. Accuracy S2 had a Cohen's d of 0.028 ($p = 0.693$) and rubric independence S2 had a d of 0.026 ($p = 0.611$). Both confidence intervals crossed zero. The adjusted regression confirmed these null results after controlling for baseline ability, education and AI confidence. All four robustness checks confirmed the conclusions held [12, 28].

The large independence gap from Session 1 ($d = -1.972$) closed to essentially zero in Session 2 ($d = 0.026$). The AI-assisted group improved their mean rubric independence from 1.206 in Session 1 to 1.860 in Session 2, nearly matching the control group mean of 1.854. This confirmed that AI use during Session 1 changed how participants thought during the task but did not reduce what they could do independently seven minutes later. No drop in short-term independent performance was found [6, 11].

6.2.3 Research Question 3

Can Session 1 behavioural features alone estimate the Session 2 independent performance of an individual participant, and do the conditions of this study allow that question to be tested fairly?

Research Question 3 asks something the first two cannot reach. The group comparisons establish whether AI access shifts an average; this question asks whether the same measures say anything about a named individual, which is what any screening use would require. The answer is no. The Ridge regression model did not beat a naive baseline of predicting the mean value for everyone. The model MAE was 0.1821 compared to a baseline MAE of 0.1793. The model predicted almost the same value for every participant and could not capture individual variation in Session 2 independence [46, 51].

The COI coefficient was negative at -0.0048, meaning higher offloading in Session 1 was associated with slightly lower Session 2 independence in the expected direction [2, 8]. However the magnitude was too small to be practically useful. The ceiling effect in accuracy and the small group difference in rubric independence in Session 2 left very little variation to predict [11].

Session 1 features could not serve as a reliable early warning tool for overreliance risk under the conditions of this study. This result does not mean COI has no predictive value in principle. It means that a same-sitting design with structured tasks and a ceiling effect does not provide the conditions needed to test predictive value adequately. Future studies with longer delays and more varied outcomes would be needed to evaluate this properly [8].

6.3 Answer to the Central Question

The central question of this study was: how does growing reliance on AI for thinking tasks affect human problem solving, and does this reliance risk weakening essential cognitive skills?

Based on the evidence from this study, the answer has two parts.

Part 1 - AI use changed how people thought during the task. The AI-assisted group produced substantially weaker independent reasoning and justification during Session 1 compared to the control group, even though both groups achieved the same accuracy. This confirms that AI use shifted the reasoning process toward the tool and away from the person, which is the substitution pattern described in the literature [1, 4]. Accuracy alone missed this difference entirely.

Part 2 - No loss of short-term independent performance was detected. When AI was removed seven minutes later, both groups performed at the same level. The reasoning quality gap disappeared and the primary endpoints showed negligible differences. Based on this study, AI reliance during Session 1 did not weaken independent problem solving seven minutes later.

However this conclusion comes with a critical condition. The study only tested independence after a seven-minute break in the same sitting. It cannot say anything about what would happen after one day, one week or one month of regular AI use. The risk described in the literature, that habitual AI reliance may gradually reduce independent cognitive capability over time [2, 6, 7], remains plausible and cannot be ruled out by this study. This study simply shows that for these tasks, this context and this delay, no immediate independence cost was measurable.

6.4 Recommendations

6.4.1 For Educators and Institutions

- **Do not rely on accuracy alone to evaluate AI-assisted performance.** The findings showed that accuracy was identical across groups even when reasoning quality differed substantially. Rubric-based independence scoring or similar qualitative measures should be used alongside accuracy to get a fuller picture of what students are actually learning [4].
- **Require justification and verification during AI-assisted tasks.** Students who must explain their reasoning in their own words and check AI output before submitting are less likely to simply copy and accept. This is consistent with the forcing function approach shown to reduce overreliance in AI-assisted decision making [4].
- **Use behavioural logs where possible.** If a learning platform records activity data such as AI use frequency, time in tool and copy behaviour, these can be used to compute reliance indices similar to COI. These give a more honest picture of how students interact with AI than self-report surveys alone [5].
- **Consider periodic no-AI testing.** Even if AI use does not harm very short-term independence, longer-term effects are unknown. Regular independent tests without AI can help track whether independent capability is being maintained over time [9].

6.4.2 For Organisations and Workplaces

- **Monitor AI reliance behaviour through logs, not just output quality.** Employees who produce correct outputs with heavy AI assistance may not be developing or maintaining the skills needed when AI is unavailable. COI and HES provide a practical framework for monitoring this [9, 10].
- **Design workflows that require human verification steps.** Requiring employees to check and explain AI-generated outputs before submitting them reduces the risk of passive acceptance and keeps human reasoning active in the process [4, 5].
- **Do not use same-day tests to evaluate AI impact on skills.** This study showed that a seven-minute break was not enough to detect any skill change. Organisations that want to understand the longer-term association between AI use and employee capability should use delayed independent tests over days or weeks, not minutes [6].

6.4.3 For Future Research

- **Use a longer delay between sessions.** The most important methodological change for a follow-up study is to test independence after 24 hours, 72 hours or one week. The present design could only measure whether independent performance held up after a few minutes. A longer delay would give any loss of skill time to appear, and would test the question this study could not reach, which is whether regular AI use weakens independent capability over days or weeks [6, 7].
- **Use tasks that produce more variation in accuracy.** The ceiling effect in this study severely limited what the analysis could detect. Future studies should use tasks calibrated to produce accuracy scores spread across the full range from 0 to 1 so that differences between groups are more detectable [11].
- **Invest in rater calibration before scoring.** The ICC failures in this study showed that two raters applied very different standards on the subjective rubric dimensions. A calibration session with practice scoring and discussion before the main scoring begins would improve inter-rater agreement substantially [30].
- **Validate COI and HES on different task types and populations.** The indices built in this study were checked for internal stability on structured problem solving tasks with a Kampala-based sample. Their performance on open-ended tasks, creative work or professional decision making in other populations is unknown and needs to be tested [1, 8].
- **Test the predictive model with greater outcome variability.** The Ridge regression model failed to beat the baseline partly because there was very little variation in Session 2 outcomes to predict. A study with a longer delay and more varied tasks would give the model a genuine opportunity to show whether COI and HES have predictive value [46, 51]. A follow-up study should also benchmark Ridge regression against at least one flexible model, such as a gradient boosting regressor, so that the choice of model family is tested empirically rather than argued from the shape of the outcome distribution.
- **Include a larger and more diverse sample.** The convenience sample from two sites in Kampala limits generalisation. A larger study drawing from multiple countries, age groups and educational backgrounds would produce findings with wider applicability [31].

6.5 Contribution of the Study

Despite its limitations, this study made several clear contributions.

A reproducible measurement framework. The COI and HES scoring rules, and formulas produced by this study are fully documented and reusable. Another researcher can take the same raw logs and reproduce the exact same scores from scratch. This addresses the validity gap identified in Chapter 2 where most existing studies lacked auditable measurement tools [1, 9].

A two-session labelled dataset. The cleaned dataset with 1,387 completers, 13,870 task episodes and linked Session 1 and Session 2 outcomes is a structured resource for future analysis. It is one of the few datasets in this area that links behavioural AI use logs directly to delayed independent performance outcomes [6, 7].

An honest null result. The finding of no loss of short-term independent performance is a genuine and useful scientific result. It is not a failure. It shows that for structured tasks with a seven-minute delay, AI use does not immediately harm independent performance. This is a baseline that future studies with longer delays can build on and compare against [11].

A practical interactive dashboard. The Streamlit dashboard produced as part of this study provides a working prototype of how the COI and HES framework can be deployed for stakeholder use. Users can upload the preprocessed dataset and immediately explore study-level metrics, participant breakdowns, COI and HES distributions and the main session comparison findings through interactive Plotly charts. The dashboard demonstrates that the measurement framework is not limited to academic analysis but can be packaged into a tool that educators, researchers and organisations can use without writing any code.

6.6 Final Reflection

This study set out to turn a common worry about AI into measurable data. The worry was that people who rely heavily on AI to do their thinking for them will end up less able to think independently. That worry is reasonable and the literature gives good reasons to take it seriously [2, 6, 8].

What this study found is that the worry is not confirmed in the short term for these tasks. But it is also not dismissed. The AI-assisted group clearly produced weaker reasoning during Session 1. They just recovered quickly when AI was taken away. Whether that recovery holds after days or weeks of habitual AI use is the question this study could not answer and that future research must address.

The data science framework built here, the COI and HES indices, the analysis pipeline and the risk screen dashboard, provides the tools to investigate that question in a rigorous and reproducible way. That is perhaps the most lasting contribution of this work.

Bibliography

- [1] M. H. Jarrahi, G. Newlands, and C. Lutz, “Artificial intelligence and human cognition: A mutual augmentation perspective,” *Big Data & Society*, vol. 9, no. 2, 2022.
- [2] M. Gerlich, “Ai tools in society: Impacts on cognitive offloading and the future of critical thinking,” *Societies*, vol. 15, no. 1, p. 6, 2025.
- [3] A. Skulmowski, “The cognitive architecture of digital externalization,” *Educational Psychology Review*, vol. 35, no. 4, p. 101, 2023.
- [4] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, “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*, vol. 5, no. CSCW1, pp. 1–21, 2021.
- [5] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman, “A systematic review on fostering appropriate trust in human-ai interaction,” *ACM Journal on Responsible Computing*, vol. 1, no. 4, 2024.
- [6] S. Grinschgl, F. Papenmeier, and H. S. Meyerhoff, “Consequences of cognitive offloading: Boosting performance but diminishing memory,” *Quarterly Journal of Experimental Psychology*, vol. 74, no. 9, pp. 1477–1496, 2021.
- [7] B. C. Storm and J. S. Soares, “Memory in the digital age,” in *The Oxford Handbook of Human Memory*, Oxford University Press, 2024.
- [8] J. Tian and R. Zhang, “Learners’ ai dependence and critical thinking: The psychological mechanism of fatigue and the social buffering role of ai literacy,” *Acta Psychologica*, vol. 260, p. 105725, 2025.
- [9] National Institute of Standards and Technology, “Ai risk management framework (ai rmf 1.0),” tech. rep., U.S. Department of Commerce, 2023.
- [10] UNESCO, “Guidance for generative ai in education and research,” tech. rep., UNESCO, 2023.
- [11] R. Deng, M. Jiang, X. Yu, Y. Lu, and S. Liu, “Does chatgpt enhance student learning? a systematic review and meta-analysis of experimental studies,” *Computers & Education*, vol. 227, p. 105224, 2025.
- [12] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988.
- [13] A. Skulmowski and M. Xu, “Understanding cognitive load in digital and online learning,” *Educational Psychology Review*, vol. 34, no. 1, pp. 171–196, 2022.
- [14] E. F. Risko and S. J. Gilbert, “Cognitive offloading,” *Trends in Cognitive Sciences*, vol. 20, no. 9, pp. 676–688, 2016.
- [15] 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.

- [16] T. L. Dunn and E. F. Risko, "Toward a metacognitive account of cognitive offloading," *Cognitive Science*, vol. 40, no. 5, pp. 1080–1127, 2016.
- [17] A. Skulmowski, "Learning by doing or doing without learning," *Educational Psychology Review*, vol. 36, p. 28, 2024.
- [18] A. R. Doshi and O. P. Hauser, "Generative ai enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, vol. 10, no. 28, p. eadn5290, 2024.
- [19] N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunschtein, and P. Maes, "Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task." arXiv preprint arXiv:2506.08872, 2025.
- [20] H. Bastani, O. Bastani, A. Sungu, H. Ge, Kabakcı, and R. Mariman, "Generative ai can harm learning," *The Wharton School Research Paper*, 2024.
- [21] H.-P. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson, "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*, pp. 1–22, 2025.
- [22] C. Degachi, C. M. Jonker, M. L. Tielman, and O. Vereschak, "Practising appropriate trust in human-centred ai design," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.
- [23] G. Spitale, N. Biller-Andorno, and F. Germani, "Ai model gpt-3 (dis)informs us better than humans," *Science Advances*, vol. 9, no. 26, p. eadh1850, 2023.
- [24] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.
- [25] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage, 2018.
- [26] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*. Pearson, 8th ed., 2019.
- [27] D. C. Montgomery, *Design and Analysis of Experiments*. Wiley, 2017.
- [28] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs*. Houghton Mifflin, 2002.
- [29] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, 1951.
- [30] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients," *Journal of Chiropractic Medicine*, 2016.
- [31] I. Etikan, S. Musa, and R. Alkassim, "Comparison of convenience sampling and purposive sampling," *American Journal of Theoretical and Applied Statistics*, 2016.
- [32] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.
- [33] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.
- [34] The PostgreSQL Global Development Group, "Postgresql: The world's most advanced open source relational database," 2023.

- [35] Neon Inc., “Neon: Serverless postgresql,” 2024. Serverless PostgreSQL platform used for database hosting during data collection.
- [36] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
- [37] W. McKinney, “Data structures for statistical computing in python,” in *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010.
- [38] C. R. Harris *et al.*, “Array programming with numpy,” *Nature*, vol. 585, pp. 357–362, 2020.
- [39] P. Virtanen *et al.*, “Scipy 1.0: Fundamental algorithms for scientific computing in python,” *Nature Methods*, vol. 17, pp. 261–272, 2020.
- [40] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with python,” *Proceedings of the 9th Python in Science Conference*, 2010.
- [41] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [42] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science and Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [43] M. L. Waskom, “seaborn: statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
- [44] Streamlit, “Streamlit: A faster way to build and share data apps.” Streamlit Documentation, 2026.
- [45] Plotly Technologies Inc., “Collaborative data science,” 2015.
- [46] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
- [47] S. Arlot and A. Celisse, “A survey of cross-validation procedures,” *Statistics Surveys*, 2010.
- [48] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.
- [49] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality,” *Biometrika*, vol. 52, no. 3, pp. 591–611, 1965.
- [50] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995.
- [51] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
- [52] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [53] K. Pearson, “Notes on regression and inheritance in the case of two parents,” *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.
- [54] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society*, 1996.
- [55] J. M. Lodge and L. Loble, “Artificial intelligence, cognitive offloading and implications for education,” tech. rep., University of Technology Sydney, 2026.
- [56] L. Favero, J. A. Pérez-Ortiz, T. Käser, and N. Oliver, “Ai in education beyond learning outcomes: Cognition, agency, emotion, and ethics.” arXiv preprint arXiv:2602.04598, 2026.

- [57] B. Jose, J. Cherian, A. M. Verghis, S. M. Varghise, S. Mumthas, and S. Joseph, “The cognitive paradox of ai in education: between enhancement and erosion,” *Frontiers in Psychology*, vol. 16, p. 1550621, 2025.
- [58] K. Bhattarai, “Mediating role of cognitive offloading: Linkage between generative ai usage and critical thinking,” Master’s thesis, Lappeenranta-Lahti University of Technology LUT, 2025.
- [59] X. Bai, Y. Liu, and Z. Su, “Ai-assisted learning and its impact on cognitive engagement and retention,” *Educational Technology Research and Development*, 2023.
- [60] A. Yunus, P. R. Gay, and O. T. Lee, “From co-design to metacognitive laziness: Evaluating generative ai in vocational education.” arXiv preprint arXiv:2512.12306, 2025.
- [61] National Planning Authority, “Fourth national development plan (ndpiv) 2025/26–2029/30,” tech. rep., Government of Uganda, Kampala, 2025.
- [62] National Planning Authority, “Uganda vision 2040,” tech. rep., Government of Uganda, Kampala, 2013.
- [63] Ministry of ICT and National Guidance, “Digital uganda vision,” tech. rep., Government of Uganda, Kampala, 2023.
- [64] United Nations, “Transforming our world: The 2030 agenda for sustainable development,” tech. rep., United Nations General Assembly, 2015. A/RES/70/1.

Appendix A

Research Project timelines

The study was carried out from June 2026 to August 2026 following approval by the Research Ethics Committee on 25 June 2026. That date marked the official start of the research. The study followed a structured sequence of phases from data collection through to final report submission.

In the first phase, which took place in late June 2026 immediately after ethics approval was received, the study tools were finalised. This included the task sets, questionnaires, rubrics, the hint bank protocol and the online study platform. A small pilot was conducted with a limited number of participants to test the system and identify any issues before full data collection began.

In the second phase, which ran through July 2026, data collection was conducted across both sites. At the MAIN site, participants were brought into a computer lab in organised batches. At the ONLINE site, participants received a link and completed both sessions at their own convenience. Session 1 and Session 2 were completed in the same sitting with a short break of approximately seven minutes between sessions. Data collection was stopped at 1,474 enrolled participants in July 2026 due to time constraints, even though participants were still attempting to fill in the forms at the time of closure.

In the third phase, which also took place in July 2026, the raw data was cleaned, integrated and processed. The nine raw data files were merged into two integrated tables. Missing values were treated, outliers were detected and capped, and the COI and HES indices were computed following the proposal formulas.

In the fourth phase, the full analysis pipeline was run. This included exploratory data analysis, group comparisons for both sessions, effect size estimation with confidence intervals, reliability checks, predictive modelling and four robustness checks. The interactive dashboard was also built and tested during this phase.

In the final phase, the results were written up and organised into this report. This included interpretation of findings, discussion of limitations and preparation of all appendices. The final report is anticipated to be completed in August 2026.

Summary Timeline Table

Phase and Period	Work done	Outputs
Ethics approval and tool finalisation (25 June 2026)	Ethics approval received. Study tools finalised. Pilot conducted.	Approved instruments, pilot fixes, platform ready for data collection
Data collection (July 2026)	Session 1 and Session 2 conducted across MAIN and ONLINE sites. 1,474 participants enrolled. 1,387 completers retained.	Raw data files across nine exports. Session delay confirmed at approximately 7 minutes.
Data cleaning and feature engineering (July 2026)	Missing value treatment, outlier capping, encoding, z-score standardisation, COI and HES computation.	table_a_preprocessed.csv, table_b_preprocessed.csv, COIS1 and HES_S1 columns
Analysis and modelling (July to August 2026)	EDA, group comparisons, effect sizes, reliability checks, Ridge regression, robustness checks, dashboard development.	Evidence table, forest plot, reliability report, model results, robustness summary, interactive dashboard
Report writing and finalisation (August 2026)	Interpretation of findings, discussion of limitations, recommendations and all appendices prepared.	Final research report submitted in August 2026

Ethics approval reference: Research Ethics Committee approval was granted on 25 June 2026. All data collection, storage and processing followed the conditions of that approval and the ethical guidelines for research involving human participants [9, 10].

Appendix B

Declaration of AI use

Declaration of AI Tool Use and Originality

I, Richard Wambede, Reg No. J25M19/005, declare that this research report is my own original work. Where AI tools were used to assist in the preparation of this document, the nature of that use is disclosed below in full, in accordance with the academic integrity policy of my institution.

1. I used **ChatGPT** to help organise the report sections and improve the flow of the writing. The prompt used was: *“Generate LaTeX format for my abstract and structure my sections clearly.”* I modified the output by rewriting parts in simpler English, correcting some wording and making sure it matched my study content and findings.
2. I used **DeepSeek** to paraphrase selected sections to reduce repetition and improve readability. The prompt used was: *“Rewrite this text in very simple English with small mistakes to sound human.”* I modified the output by adjusting the grammar, removing unnecessary words and making sure the original meaning and technical accuracy were preserved.
3. I used **Claude** to check the clarity of explanations in sections such as the significance of the study, limitations and key definitions. The prompt used was: *“Is this explanation clear for a research report, and how can I simplify it?”* I modified the output by selecting only useful suggestions and rewriting the text in my own words.
4. I used **ChatGPT** to suggest realistic example figures such as percentages and simple statistics used to illustrate points in the discussion. The prompt used was: *“Add simple realistic numbers to make this section stronger but not fake results.”* I modified the output by keeping only values that were reasonable and consistent with my actual study design and findings.
5. I used **DeepSeek** to assist with formatting the LaTeX document correctly. The prompt used was: *“Convert this text into LaTeX with sections and itemize.”* I modified the output by correcting formatting issues and adjusting the structure to fit my document style and content.
6. I used **Claude** to assist with developing and refining the research poster. The prompts used included: *“Help me structure this research content into a clear academic poster layout”* and *“Suggest improvements to the layout, spacing and presentation of the poster.”* I modified the output by selecting suitable suggestions, adjusting the LaTeX code and redesigning sections of the poster to match my research content and required presentation format.

I, Richard Wambede, Reg No. J25M19/005, further declare that I have no known competing financial interests or personal relationships that could have influenced the work reported in this research.

Signature: Richard Wambede

Date: 16th Sept 2026

Appendix C

List of 20 Major Journal Publications reviewed

C.0.1 Review Papers and Meta-Analyses

- Deng, Jiang, Yu, Lu and Liu (2024), published in *Computers & Education* (Elsevier, Q1 in education technology), report a systematic review and meta-analysis of experimental ChatGPT studies. They find consistent performance gains during AI use but mixed and weaker effects on learning itself.
- Mehrotra, Degachi, Vereschak, Jonker and Tielman (2024), in the *ACM Journal on Responsible Computing*, present a systematic review of how appropriate trust in human-AI interaction is built and measured, and identify overreliance as the dominant failure mode.
- Risko and Gilbert (2016), in *Trends in Cognitive Sciences* (Cell Press, a leading review journal in cognitive science), define cognitive offloading and review the conditions under which people choose to shift cognitive work onto external tools.
- Skulmowski (2023), in *Educational Psychology Review* (Springer), sets out a theoretical account of digital externalization, explaining how offloading reshapes the cognitive architecture of learning rather than simply reducing effort.
- Storm and Soares (2024), a chapter in *The Oxford Handbook of Human Memory* (Oxford University Press), reviews what is known about memory in digital environments and the conditions under which external storage weakens internal retention.

C.0.2 Journal Articles

- Grinschgl, Papenmeier and Meyerhoff (2021), in the *Quarterly Journal of Experimental Psychology*, show experimentally that cognitive offloading boosts immediate task performance while diminishing later memory for the offloaded content.
- Sparrow, Liu and Wegner (2011), in *Science* (AAAS), report the Google effect: when people expect information to remain available, they remember where to find it rather than the information itself.
- Sweller (1988), in *Cognitive Science*, introduces cognitive load theory and shows that the way a problem is worked through, not only whether it is solved, determines what is learned.
- Doshi and Hauser (2024), in *Science Advances* (AAAS), find that generative AI raises individual creative output while reducing the collective diversity of the ideas produced.

- Spitale, Biller-Andorno and Germani (2023), in *Science Advances*, show that people struggle to distinguish AI-generated text from human text and that synthetic content can be persuasive even when inaccurate.
- Gerlich (2025), in *Societies* (MDPI), links self-reported AI tool use to lower critical thinking scores, with cognitive offloading as the proposed mechanism.
- Tian and Zhang (2025), in *Acta Psychologica* (Elsevier), model AI dependence as acting on critical thinking through cognitive fatigue, with AI literacy as a buffering factor.
- Jarrahi, Newlands and Lutz (2022), in *Big Data & Society* (SAGE), argue for a mutual augmentation perspective in which AI extends human cognition rather than replacing it, and provide the substitution versus augmentation distinction used in this study.

C.0.3 Conference Papers

- Buçinca, Malaya and Gajos (2021), published in *Proceedings of the ACM on Human-Computer Interaction* (ACM CSCW, CORE A*), show that cognitive forcing functions, which require users to engage before seeing an AI answer, measurably reduce overreliance.
- Degachi, Jonker, Tielman and Vereschak (2024), at the *ACM CHI Conference on Human Factors in Computing Systems* (CORE A*), examine how appropriate trust can be designed into human-centred AI systems.
- Lee, Sarkar, Tankelevitch, Drosos, Rintel, Banks and Wilson (2025), at *ACM CHI 2025*, survey knowledge workers and report self-reported reductions in cognitive effort alongside confidence effects when generative AI is used for professional tasks.

C.0.4 Other Relevant Publications

- Bastani, Bastani, Sungu, Ge, Kabakçı and Mariman (2024), a Wharton School working paper, report a field experiment in which unrestricted AI practice improved performance during practice but reduced later unaided performance.
- Kosmyna, Hauptmann, Yuan, Situ, Liao, Beresnitzky, Braunstein and Maes (2025), an arXiv preprint from the MIT Media Lab, use brain-activity measurement across repeated essay sessions and describe an accumulation of cognitive debt under sustained AI assistance.
- UNESCO (2023) issues global guidance on the use of generative AI in education and research, including the governance conditions institutions should meet before adopting such tools.
- The National Institute of Standards and Technology (2023) publishes the AI Risk Management Framework (AI RMF 1.0), which provides the risk vocabulary used in this study when discussing overreliance as a measurable harm.

Appendix D

Research Alignment with the MSc Programme, National Development Agenda and Global Frameworks

D.1 Alignment with the MSc Programme in Data Science

This research aligns closely with the MSc programme in Data Science and Analytics. The study applies data collection, feature engineering, statistical modelling and model evaluation to a question about human behaviour during AI-assisted problem solving. It uses composite index construction, predictive modelling, grouped cross-validation, effect size estimation with confidence intervals, reliability testing and a reproducible analysis pipeline. These match the core competencies the programme develops: working with real and imperfect datasets, building and honestly evaluating models, and interpreting results in a form that supports decision making.

The study also exercises the judgement side of the discipline. The predictive model did not beat a naive baseline, the four Likert scales failed their reliability checks, and two of the three rubric dimensions failed inter-rater agreement. Reporting these outcomes openly, and explaining what they do and do not permit the study to claim, is as much a part of data science practice as producing a working model.

D.2 Alignment with the Fourth National Development Plan (NDPIV)

Uganda's Fourth National Development Plan, covering FY2025/26 to FY2029/30, pursues the goal of higher household incomes, full monetisation of the economy and employment for sustainable socio-economic transformation, under the theme of sustainable industrialisation for inclusive growth, employment and wealth creation [61]. ICT is one of the seven high-impact growth areas the plan identifies, and human capital development is one of its programme areas.

This study speaks to both. As AI tools enter Ugandan classrooms and workplaces, the plan's human capital objectives depend on those tools building capability rather than replacing it. The study provides behavioural evidence on how people actually use AI during problem solving, and

a measurement approach that institutions can apply to check whether adoption is strengthening independent reasoning or substituting for it.

D.3 Alignment with Uganda Vision 2040

Uganda Vision 2040 sets out to transform Ugandan society from a peasant to a modern and prosperous country within thirty years, moving the country from a predominantly peasant low-income economy to a competitive upper middle-income one [62]. Reaching that position depends on a workforce able to work with advanced technology without becoming dependent on it.

This study supports that objective by producing a measurable framework for observing how people interact with AI during cognitive work. The Cognitive Offloading Index and the Human Engagement Score are built from behavioural logs rather than self-report, so any institution that can record AI requests, copying and verification can compute them on its own users.

D.4 Alignment with the Digital Uganda Vision

The Digital Uganda Vision sets the goal of a digitally empowered society and knowledge economy [63]. This study contributes to the digital literacy side of that agenda. It supports balanced technology use, in which people draw on AI while continuing to practise their own reasoning, and it offers a way to tell the difference between the two in practice. This is directly relevant to the universities, vocational training institutions and workplaces in Uganda that are adopting AI tools faster than they are developing ways to monitor their effect.

D.5 Alignment with the Sustainable Development Goals

This research supports three Sustainable Development Goals [64].

It aligns with **SDG 4 (Quality Education)** by generating evidence on how AI assistance affects the quality of student reasoning and short-term independent problem solving. The finding that AI use changed how participants worked during the tasks, while their accuracy stayed the same, has direct implications for how AI is built into assessment and for what marking an answer alone can and cannot reveal.

It contributes to **SDG 8 (Decent Work and Economic Growth)** by helping organisations understand how to adopt AI in ways that keep independent cognitive capability intact, which matters for building a workforce that can adapt as the tools change.

It supports **SDG 9 (Industry, Innovation and Infrastructure)** by providing a reproducible data science pipeline, together with two documented indices, that other researchers and organisations can apply to study AI reliance in their own settings.

D.6 Alignment with Responsible AI and Ethics

This study aligns with international efforts to promote responsible and ethical AI use [9, 10]. It produces measurable evidence on overreliance risk and provides a way of detecting substitution behaviour from activity logs rather than from assumptions about what users are doing. This supports guidance calling for human-centred systems in which human reasoning and verification remain active parts of the process. The design of the hint bank, which required participants to

write their own final answer at every assistance level, follows the same principle as the cognitive forcing functions shown to reduce overreliance in controlled settings [4].

D.7 Alignment with Global Digital Transformation Trends

AI is being adopted rapidly in education, business and decision making worldwide, and the evidence on what this does to human thinking remains mixed [11, 2]. This study contributes a simple, measurable and reproducible framework for studying that question. The COI and HES indices, the cleaned dataset and the analysis pipeline are offered as reusable contributions that can be adapted and tested in other sectors, populations and task types beyond the setting of this study.

Appendix E

Data Dictionary

This appendix defines every variable in the raw data files used in this study. Each variable is listed with its source file, column name, data type, allowable values and a plain-language description. This dictionary allows another researcher to understand and reproduce every variable in the dataset without needing to contact the original researcher [9].

Variables marked as *AI group only* were structurally missing for the control group by design. Variables marked as *S1 only* were only collected during Session 1. Variables marked as *empty* were retained in the repository for transparency but contained no data because the relevant data collection layer was not activated.

E.1 Participants File

Table E.1: Data Dictionary for the Study Dataset

Column	Type	Allowed Values	Description
participant_code	String	–	Anonymous participant identifier used to merge records across all datasets.
site_code	String	MAIN, ONLINE	Data collection site. MAIN indicates computer laboratory sessions, while ONLINE indicates self-completed participation.
batch_code	String	–	Batch identifier in the format YYYYMMDD_SITE_BATCH##.
group_assignment	Enum	CONTROL, AIASSISTED	Experimental group allocation.
assignment_method	Enum	SUGGESTED_ACCEPTED, MANUAL_OVERRIDE	Method used to assign the participant to a study group.
auto_assigned	Boolean	True, False	Indicates whether the participant was assigned automatically by the server.
consent_given	Boolean	True, False	Participant consent status recorded during registration.
age_band	Enum	18–24, 25–34, 35–44, 45+	Participant age category.

Column	Type	Allowed Values	Description
education_level	Enum	Secondary, Diploma, Degree, Postgraduate	Highest educational qualification.
english_comfort	Integer	1–5	Self-rated English proficiency using a five-point Likert scale.
ai_use_frequency	Enum	Never, Monthly, Weekly, Daily	Frequency of AI tool usage.
ai_confidence	Integer	1–5	Self-rated confidence in using AI tools.
warmup_auto_score	Integer	0–4	Automatically generated warm-up assessment score.
warmup_final_score	Integer	0–4	Final validated warm-up score used as the baseline ability measure.
warmup_score_overridden	Boolean	True, False	Indicates whether the automatic warm-up score was manually adjusted.
device_category	Enum	Desktop, Mobile, Tablet	Device type used by the participant.
os_family	String	–	Operating system family.
browser_family	String	–	Browser family. The complete user-agent string is not exported.
status	Enum	Checked_In, Form0, Session1, Break, Session2, Scales_S2, Completed, Withdrawn	Participant status at export. Only records with Completed status were included in the final analysis.

E.2 Task Responses File

Column	Type	Allowed values	Description
participant_code	string	—	Merge key.
session_number	int	1 2	Session this response belongs to.
task_code	string	A1..C2 A3..B4	Task identifier. A1 to C2 are Session 1 tasks. A3 to B4 are Session 2 parallel forms.
objective_answer	string	—	Submitted objective answer.
objective_auto_correct	boolean	true false	Auto-scored against the fixed answer key. Used to compute Accuracy_S1 and Accuracy_S2.
text_justification	string	—	Free-text justification submitted by the participant. Minimum 30 characters required.
task_familiarity	boolean	true false	Self-reported flag indicating whether the participant had seen a similar problem before.
self_check	boolean	true false	Self-check attestation submitted with the response.
confidence_rating	int	1..5	Confidence in the submitted answer on a 1 to 5 Likert scale.

Column	Type	Allowed values	Description
control_compliance	boolean	true false null	Control group attestation that no AI was used. Null for AI group.
assistance_level	int	0..3 null	Highest hint level revealed during the task. 0 means no hints used. Null for control group and Session 2. AI group Session 1 only.
requests_count	int	—	Number of hint unlock requests made. AI group Session 1 only.
copy_used	boolean	true false null	Whether the participant copied hint wording into their answer. AI group Session 1 only.
verified	boolean	true false null	Whether the participant reported verifying the AI output. AI group Session 1 only.
verification_method	enum	RECOMPUTE ESTIMATE ALT_METHOD CONSISTENCY OTHER null	Method used to verify the answer. AI group Session 1 only.
verification_method_other_text	string	—	Free text when verification method was OTHER.
verification_evidence	string	—	One-sentence verification evidence. Required when verified = true.
duration_objective_ms	int	—	Milliseconds from task start to first committed objective answer.
duration_justification_ms	int	—	Milliseconds from first answer to final submission.
duration_total_ms	int	—	Total milliseconds from task start to submission. Divided by 1000 to give duration_sec used in analysis.
answer_change_count	int	—	Number of times the objective answer was revised before submission.
participant_notes	string	—	Optional free-text notes from the participant.
needs_double_score	boolean	true false	Flagged for a second independent rater score.
submitted_at	timestamp	—	Submission timestamp.

E.3 Telemetry Events File

Column	Type	Allowed values	Description
participant_code	string	—	Merge key.
session_number	int	1 2	Session in which the event occurred.
task_code	string	—	Task context at time of event. Nullable.

Column	Type	Allowed values	Description
event_type	enum	PASTE PASTE_BLOCKED TAB_BLUR TAB_FOCUS ANSWER_REVISION HINT_UNLOCK HINT_COPY INFRACTION BREAK_START BREAK_END SESSION_START SESSION_END	Behavioural event type captured by the platform.
event_metadata	json	—	Event-specific metadata stored as JSON.
client_timestamp_ms	bigint	—	Client-side event timestamp in milliseconds.
created_at	timestamp	—	Server receipt time.

E.4 Hint Events File

Column	Type	Allowed values	Description
participant_code	string	—	Merge key. AI group Session 1 only.
task_code	string	—	Task for which the hint was requested.
level	int	1..3	Hint level revealed. Level 1 is directional. Level 2 is scaffold. Level 3 is near-complete worked solution.
request_number	int	—	Sequential request index for the task.
viewed_duration_ms	int	—	Time in milliseconds the hint was visible to the participant.
copied	boolean	true false	Whether the participant copied the hint text.

E.5 Scale Responses File

Column	Type	Allowed values	Description
participant_code	string	—	Merge key.
session_number	int	1 2	Session in which the scale was administered.

Column	Type	Allowed values	Description
item_code	string	S1-E1..E3 S1-H1..H3 S1-AI1..AI4 S2-H1..H3 S2-I1..I3	Likert item code. S1-E = effort Session 1. S1-H = engagement Session 1. S1-AI = AI use Session 1 (AI group only). S2-H = engagement Session 2. S2-I = independence Session 2.
rating	int	1..5	Likert rating given by the participant.

E.6 Rater Scores File

Column	Type	Allowed values	Description
participant_code	string	—	Merge key.
session_number	int	1 2	Session of the scored response.
task_code	string	—	Task of the scored response.
rater_display_code	string	—	Anonymous rater code. Two raters were used: RATER-A1 (primary) and RATER-02 (double scorer).
correctness	int	0..2	Rubric dimension: correctness of the answer. 0 = incorrect. 1 = partially correct. 2 = fully correct.
justification_quality	int	0..2	Rubric dimension: quality of the written justification. 0 = absent or unclear. 1 = partial. 2 = clear and complete.
independence	int	0..2	Rubric dimension: independence of reasoning. 0 = fully copied or AI-dependent. 1 = partial own reasoning. 2 = fully independent reasoning.
is_double_score	boolean	true false	True if this is a second independent score for reliability checking. False if this is the primary rater score.

E.7 Session Meta File

Column	Type	Allowed values	Description
participant_code	string	—	Merge key.
session_number	int	1 2	Session number.
started_at	timestamp	—	Session start timestamp. Used to compute the delay between Session 1 and Session 2.
ended_at	timestamp	—	Session end timestamp.
transfer_prompt_used	boolean	true false null	Whether the transfer prompt was answered. Session 2 only.

Column	Type	Allowed values	Description
transfer_prompt_text	string	—	Free-text response to the transfer prompt. Session 2 only.

E.8 Deviation Logs File

This file was empty. The study was conducted entirely online and no physical proctors were present to record deviations. The file is retained in the repository for reproducibility and transparency.

Column	Type	Allowed values	Description
participant_code	string	—	Merge key. Nullable for batch-level deviations.
batch_code	string	—	Batch in which the deviation occurred.
proctor_display_code	string	—	Anonymous proctor code.
description	string	—	Plain-text description of the deviation.
created_at	timestamp	—	Timestamp when the deviation was logged.

E.9 Layer 2 Logs File

This file was empty. Layer 2 was planned as an optional calibration step using live ChatGPT sessions. It was not activated during data collection. All data was collected using Layer 1, the offline Hint Bank. The file is retained for completeness.

Column	Type	Allowed values	Description
participant_code	string	—	Merge key.
prompt_log_id	string	—	External AI tool log identifier.
model_name_shown	string	—	Model name presented to the participant.
prompt_count	int	—	Number of prompts submitted to the AI tool.
time_in_tool_minutes	numeric	—	Minutes spent in the AI tool.
copy_similarity_note	string	—	Observed copy similarity note from the proctor.

E.10 Derived Variables in Table A

The following variables were engineered from the raw files and added to Table A during the analysis pipeline. They do not appear in the raw files.

Column	Type	Allowed values	Description
accuracy_s1	float	0..1	Mean proportion of objective tasks answered correctly in Session 1.
accuracy_s2	float	0..1	Mean proportion of objective tasks answered correctly in Session 2. Primary endpoint.

Column	Type	Allowed values	Description
delta_accuracy	float	-1..1	Change in accuracy from Session 1 to Session 2. Computed as accuracy_s2 minus accuracy_s1. Descriptive only.
time_s1	float	seconds	Mean task duration in seconds across Session 1 episodes. Capped at 244.7 seconds after outlier treatment.
time_s2	float	seconds	Mean task duration in seconds across Session 2 episodes. Capped at 210.5 seconds after outlier treatment.
rub_correctness_s1	float	0..2	Mean primary rater correctness score across Session 1 tasks.
rub_justification_s1	float	0..2	Mean primary rater justification quality score across Session 1 tasks.
rub_independence_s1	float	0..2	Mean primary rater independence score across Session 1 tasks.
rub_correctness_s2	float	0..2	Mean primary rater correctness score across Session 2 tasks.
rub_justification_s2	float	0..2	Mean primary rater justification quality score across Session 2 tasks.
rub_independence_s2	float	0..2	Mean primary rater independence score across Session 2 tasks. Primary endpoint.
effort_s1	float	1..5	Mean of effort scale items S1-E1, S1-E2, S1-E3. Failed Cronbach's alpha. Descriptive only.
engagement_s1	float	1..5	Mean of engagement scale items S1-H1, S1-H2, S1-H3. Failed Cronbach's alpha. Descriptive only.
ai_use_s1	float	1..5	Mean of AI use scale items S1-AI1 to S1-AI4. AI group only. Failed Cronbach's alpha. Descriptive only.
ai_ratio	float	0..1	Proportion of Session 1 tasks where AI assistance was used. AI group only.
copy_rate	float	0..1	Proportion of Session 1 tasks where hint text was copied. AI group only.
verify_rate	float	0..1	Proportion of AI uses followed by a verification action. AI group only. Capped at 1.0 after outlier treatment.
group_binary	int	0 1	Binary group label. AI ASSISTED = 1. CONTROL = 0.

Column	Type	Allowed values	Description
education_encoded	int	0..3	Encoded education level. SECONDARY = 0. DIPLOMA = 1. DEGREE = 2. POSTGRAD = 3.
site_encoded	int	0 1	Encoded site. MAIN = 0. ONLINE = 1.
age_encoded	int	0..3	Encoded age band. 18_24 = 0. 25_34 = 1. 35_44 = 2. 45_PLUS = 3.
COLS1	float	unbounded	Cognitive Offloading Index for Session 1. AI group only. See Appendix B for formula.
HES_S1	float	unbounded	Human Engagement Score for Session 1. All participants. See Appendix B for formula.

Appendix F

COI and HES Scoring Specification

This appendix documents the exact scoring rules for the Cognitive Offloading Index (COI) and the Human Engagement Score (HES). The purpose is to allow another researcher to take the same raw data files and reproduce the exact same index scores without needing to contact the original researcher. This level of documentation follows the transparency principles set out in the NIST AI Risk Management Framework [9].

F.1 General Principles

Both indices follow the same three-step construction process.

1. Extract the raw component variables from the raw data files.
2. Standardise each component using a z-score so that all variables are on the same scale.
3. Combine the standardised components by taking their unweighted mean.

The z-score formula used for all components is:

$$z(x) = \frac{x - \bar{x}}{s_x} \quad (\text{F.1})$$

Where $\bar{x}$ is the mean of the variable computed across the relevant group and s_x is the standard deviation. After standardisation, every component has a mean of 0 and a standard deviation of 1, confirmed in the analysis [32].

Important: The means and standard deviations used for standardisation must be computed from the same group that the index applies to. For COI, use the AI-assisted group only. For HES, use all participants. Using a different reference group will produce different z-scores and different index values.

F.2 Cognitive Offloading Index (COI)

F.2.1 Applicability

COI applies to the AI-assisted group only, Session 1 only. The control group has no AI use data and therefore cannot have a COI score. COI values for control group participants must be set to missing (NaN), not zero. Because the z-scores are computed within the AI-assisted group, COI describes how one AI-assisted participant compares with another. It cannot express a difference between the two conditions and must never be used as a group contrast. See Chapter 3, Where COI Sits in This Study.

F.2.2 Raw Components

Four raw components are extracted from the task_responses file for Session 1 AI-assisted participants.

Component	Formula	Range	Notes
AI_Ratio	tasks_with_ai / total_tasks	0 to 1	tasks_with_ai = count of tasks where assistance_level > 0.
Time_S1	mean(duration_total_ms) / 1000	seconds	Mean episode time in seconds per participant. Cap extreme values before computing the mean. See outlier rules below.
Copy_Rate	copy_count / total_tasks	0 to 1	copy_count = count of tasks where copy_used = true.
Verify_Rate	verify_count / tasks_with_ai	0 to 1	verify_count = count of tasks where verified = true. If tasks_with_ai = 0, set verify_rate = 0. Cap at 1.0 after computing. See outlier rules below. Setting verify_rate to 0 when tasks_with_ai = 0 gives that participant the maximum score on this component after reversal, even though they had nothing to verify. This is partly offset by their low AI ratio and copy rate.

Table F.1: Components and formulas used to calculate the Cognitive Offloading Index (COI).

F.2.3 Outlier Rules Before Standardisation

Two components require outlier treatment before standardisation.

Time_S1:

- Compute Q3 and IQR from all Session 1 episode times in Table B across all participants.
- Cap any episode where duration_sec exceeds $Q3 + 3.0 \times IQR$ at that bound before computing participant means.
- In this study the cap was 244.7 seconds.

Verify_Rate:

- After computing verify_rate, cap any value above 1.0 at exactly 1.0.
- A rate above 1.0 is not logically valid as it would mean more verifications than AI uses.
- In this study 85 participants required this correction.

F.2.4 Reversal Rule

Verify_Rate is reversed before standardising. This is because low verification means higher offloading risk. The reversed component is:

$$\text{inv_verify_rate} = 1 - \text{Verify_Rate} \quad (\text{F.2})$$

The z-score is then computed on `inv_verify_rate`, not on `Verify_Rate` directly.

F.2.5 COI Formula

After extracting raw components, applying outlier rules and reversing `verify_rate`, standardise each component using the z-score formula. Then compute COI as the unweighted mean of the four z-score components:

$$\text{COI_S1} = \frac{1}{4} (z(\text{AI_Ratio}) + z(\text{Time_S1}) + z(\text{Copy_Rate}) + z(1 - \text{Verify_Rate})) \quad (\text{F.3})$$

F.2.6 Interpretation

- Higher COI indicates stronger substitution behaviour. The participant used AI frequently, spent longer on each task, copied more and verified less.
- Lower COI indicates augmentation behaviour. The participant used AI sparingly, produced more of their own reasoning and checked AI output.
- COI is a relative measure. A score of 0 means the participant was at the average for the AI-assisted group on all four components combined. It does not mean the participant did not offload.
- There is no absolute threshold above which a participant is classified as a high offloader. COI should be used comparatively within the AI-assisted group.

F.2.7 Post-Computation Checks

After computing COI, run the following checks:

- Mean of COI across AI group should be approximately 0 (confirmed: 0.000 in this study).
- Standard deviation should be greater than 0 (confirmed: 0.578 in this study).
- No COI values should exist for control group participants.
- Ablation check: drop one component at a time and recompute COI. Because COI is a mean of standardised components, any ablated version is also centred near zero by construction, so this check confirms that no component distorts the centre of the distribution rather than testing how far individual participants move. A correlation between the full and ablated index answers that second question directly and is recommended for future use.
- Weight sensitivity: compare equal weights against double weight on `verify_rate` and double weight on `ai_ratio`. Correlation between schemes should exceed 0.90.

F.3 Human Engagement Score (HES)

F.3.1 Applicability

HES applies to all participants including the control group. Session 1 only. The control group receives a special treatment for the AI delegation component described below.

F.3.2 Raw Components

Three raw components are extracted from different sources.

Component	Source	Range	Notes
Engagement_self	scale_responses file. Mean of items S1-H1, S1-H2, S1-H3	1 to 5	All participants. If one item is missing, use the mean of the remaining two. If two or more items are missing, treat the scale as missing.
Rubric_Independence	rater_scores file. Mean of independence scores where is_double_score = false, Session 1 only	0 to 2	Primary rater only. All participants. Missing where participant submitted blank justification. This dimension did not reach the ICC threshold, so the component carries measurement uncertainty. See Chapter 5.
inv_AI_Ratio	1 - AI_Ratio from task_responses file	0 to 1	For control group, AI_Ratio = 0, so inv_AI_Ratio = 1.0. This gives control group participants the maximum independence credit for not using AI.

Table F.2: Components used in the HES measure.

F.3.3 Control Group Rule for inv_AI_Ratio

The control group has no AI use data. Their ai_ratio is set to 0 before computing inv_AI_Ratio. This means control group participants receive inv_AI_Ratio = 1.0, which is the maximum value. This is intentional. A person who used no AI delegated nothing to the tool and therefore receives the full independence credit on this component.

Do not set inv_AI_Ratio to missing for the control group. Set ai_ratio to 0 first and then compute $1 - 0 = 1.0$.

F.3.4 HES Formula

Standardise each of the three components using the z-score formula computed across all participants (not AI group only). Then compute HES as the unweighted mean:

$$\text{HES_S1} = \frac{1}{3}(z(\text{Engagement_self}) + z(\text{Rubric_Independence_S1}) + z(1 - \text{AI_Ratio})) \quad (\text{F.4})$$

Two points about this formula must be stated openly. First, rubric independence in Session 1 is also reported as a Session 1 outcome in Chapter 4, so part of the group difference in HES follows from how HES was built rather than being an independent finding. Second, two of the three components rest on measures that failed a reliability check: the engagement scale returned a Cronbach's alpha of -0.008 and the independence rubric dimension did not reach the ICC threshold. HES is therefore a reproducible descriptive summary, not a validated measure of engagement.

F.3.5 Interpretation

- Higher HES indicates more active engagement and better quality independent reasoning during Session 1.
- Lower HES indicates less engagement, more passive acceptance and greater delegation to AI.
- The control group should have a higher mean HES than the AI-assisted group. This is an internal consistency check on the computation, not evidence of a group difference, for the reason given under Interpretation above. If it does not hold, review the `inv_AI_Ratio` computation.
- In this study the AI-assisted group had a mean HES of -0.545 and the control group had a mean HES of 0.537.

F.3.6 Post-Computation Checks

After computing HES, run the following checks:

- Mean of HES across all participants should be approximately 0.
- HES should exist for all 1,387 participants except those with missing rubric independence scores due to blank responses.
- The control group should have a higher mean HES than the AI-assisted group. This is the manipulation check. If this does not hold, review the `inv_AI_Ratio` computation.
- Construct separation: the Pearson correlation between COI and HES should be weak to moderate in the negative direction (confirmed: $r = -0.361$ in this study). A very strong correlation would suggest the two indices overlap too much.

F.4 Summary Table

Table F.3: Construction of Composite Indices

Index	Component	Data Source	Description
COI_S1	AI Ratio	Task responses	Weight = 1/4 (AI-assisted group only). Higher values indicate greater reliance on AI.
COI_S1	Session 1 Task Time	Task responses	Weight = 1/4 (AI-assisted group only). Mean Session 1 task time in seconds. This is total task duration, not time inside the hint tool.
COI_S1	Copy Rate	Task responses	Weight = 1/4 (AI-assisted group only). Proportion of tasks on which hint content was copied into the answer.
COI_S1	Inverse Verify Rate	Task responses	Weight = 1/4 (AI-assisted group only). Higher values indicate less verification of AI-generated content.

Index	Component	Data Source	Description
HES_S1	Engagement (Self)	Scale responses	Weight = 1/3 (all participants). Higher values indicate greater self-reported engagement. Scale did not meet the Cronbach's alpha threshold.
HES_S1	Rubric Independence	Rater scores	Weight = 1/3 (all participants). Higher values indicate more independent reasoning in the work produced. Dimension did not meet the ICC threshold.
HES_S1	Inverse AI Ratio	Task responses	Weight = 1/3 (all participants). Higher values indicate less delegation of work to AI.

F.5 Reproducibility Notes

- All z-scores must be computed on the analysis sample after exclusions. Do not compute z-scores on the full enrolled sample of 1,474. Use the 1,387 completers only.
- COI z-scores must be computed within the AI-assisted group only (n = 687). Using all 1,387 participants to compute COI z-scores will produce different values.
- HES z-scores must be computed across all 1,387 participants including the control group.
- Outlier treatment must be applied before z-score computation. Applying it after will produce different standardised values.
- The order of operations is: (1) filter to completers, (2) apply outlier treatment, (3) compute raw components, (4) apply reversal rules, (5) compute z-scores, (6) compute index means.

Appendix G

Rubric Scoring Guide

This appendix documents the rubric used by raters to score participant justification responses across three dimensions: correctness, justification quality and independence. The rubric was used for all justification tasks in both Session 1 and Session 2.

Each dimension is scored on a scale of 0 to 2. The total possible score per task is 6. In the analysis, the three dimensions were treated separately rather than summed. The primary rater scored all responses. A second rater independently scored 66.6 percent of all justification responses, producing 2,726 matched pairs for inter-rater reliability checking.

G.1 General Scoring Instructions

- Score each dimension independently. Do not let the correctness score influence the independence score or vice versa.
- Read the full justification before assigning any score. Do not score based on the first sentence alone.
- Do not penalise for grammar, spelling or language errors unless the meaning is completely unclear.
- Do not reward length. A short but clear and correct justification scores higher than a long but vague one.

G.2 Dimension 1 - Correctness

Correctness measures whether the reasoning in the justification supports the correct answer. It is the most objective of the three dimensions.

Score	Label	Description and examples
2	Fully correct	The justification clearly supports the correct answer. The reasoning steps are accurate and the conclusion follows logically from the working shown. Minor numerical errors are acceptable if the method is correct and the participant clearly understood the process. <i>Example: "I added 3,500 and 2,200 to get 5,700. Then I subtracted 1,000 for the discount to get 4,700." (Answer is 4,700 and is correct.)</i>

Score	Label	Description and examples
1	Partially correct	The justification shows some correct reasoning but contains a meaningful error in method or calculation. The answer may be wrong but the participant demonstrated partial understanding of the process. <i>Example: “I added the two values but forgot to apply the discount.” (Shows awareness of steps but missed one.)</i>
0	Incorrect or absent	The justification does not support the answer, shows a fundamentally incorrect method, or is blank. This includes cases where the participant stated the answer without any reasoning. <i>Example: “The answer is 4,700 because I calculated it.” (No reasoning shown.)</i>

G.3 Dimension 2 - Justification Quality

Justification quality measures how clearly and completely the participant explained their reasoning. This is more subjective than correctness and requires rater judgement about clarity and completeness.

Score	Label	Description and examples
2	Clear and complete	The justification explains the full reasoning process in the participant’s own words. A reader who did not see the task could follow the logic and understand how the answer was reached. All key steps are stated. <i>Example: “I first identified the two costs from the table. I added them together to get the total. I then checked whether a discount applied and subtracted it. This gave me the final answer of 4,700.”</i>
1	Partial	The justification explains some but not all of the reasoning. Key steps are missing or the explanation is too brief to fully follow without the original task in front of you. <i>Example: “I added the numbers and got 4,700.” (Correct but does not explain which numbers or why.)</i>
0	Absent or unclear	The justification is blank, too short to convey meaning, consists only of the answer repeated, or is so unclear that the reasoning cannot be followed even with the task in front of you. <i>Example: “4,700 is correct.” or “I just worked it out.”</i>

G.4 Dimension 3 - Independence

Independence measures how much of the reasoning in the justification appears to come from the participant’s own thinking rather than being copied or closely adapted from AI output or the hint bank. This is the most subjective dimension and was the most difficult to score consistently across raters.

Score	Label	Description and examples
2	Fully independent	The reasoning is clearly expressed in the participant’s own words and style. There is no sign of direct copying or close paraphrasing from hint content. Even if the participant used AI assistance, they have clearly processed and restated the reasoning in their own way. <i>Example (AI group): Participant used Level 1 hint but wrote their own full working in plain language that differs clearly from the hint wording.</i> <i>Example (control group): All reasoning is self-generated. Score 2 by default unless there is evidence of copying from another source.</i>
1	Partial independence	The reasoning shows some original thinking but also contains sections that appear to be closely adapted from hint content or AI output. The participant may have rephrased some parts but the structure or key phrases echo the source material. <i>Example: Participant used Level 2 hint and the justification follows the scaffold structure closely but fills in some of the blanks with their own words.</i>
0	Fully copied or AI-dependent	The justification appears to be directly copied from hint content or AI output with little or no modification. Key phrases, sentence structure or numerical presentation closely mirror the hint wording. This score should also be given when the participant submitted the Level 3 worked solution with only trivial changes such as removing one sentence. <i>Example: Participant used Level 3 hint and the justification is word for word or near word for word from the worked solution.</i>

G.5 Special Cases and Edge Rules

Control group participants had no access to AI or hints. Score their independence dimension as 2 by default unless there is specific evidence of external copying.

- **Control group and independence:** Control group participants had no access to AI or hints. Score their independence dimension as 2 by default unless there is specific evidence of external copying. Raters should not penalise a control group participant for writing in a style that happens to sound similar to a textbook explanation.
- **Correct answer with copied justification:** It is possible to score 2 on correctness and 0 on independence. If the participant copied a correct worked solution, correctness = 2 and independence = 0.
- **Wrong answer with independent reasoning:** It is possible to score 0 on correctness and 2 on independence. If the participant worked through the problem independently but made an error, correctness = 0 and independence = 2.

- **Blank justification:** If the justification is blank or fewer than five words, assign 0 on all three dimensions. Record the response as a blank submission in the rater notes.
- **Non-English responses:** If the justification is written in a language the rater cannot assess, record all three dimensions as missing rather than zero. A zero would record a language barrier as a reasoning failure. Missing values are handled by the rules in Chapter 3.
- **Borderline cases:** When genuinely uncertain between a 1 and a 2 on any dimension, assign 1. When genuinely uncertain between a 0 and a 1, assign 0. Always prefer the more conservative score when uncertain.
- **Session 2 and independence:** No participant had access to AI or hints in Session 2. The independence dimension was therefore scored as 2 by default for all Session 2 responses unless there was specific evidence of external copying, using the same rule applied to control group participants in Session 1. This rule should be kept in mind when comparing independence scores across sessions, because it compresses Session 2 scores towards the top of the scale for both groups.

G.6 Inter-Rater Reliability Results

Table G.4 shows the ICC(2,1) inter-rater reliability results from the double scoring exercise [30].

Dimension	n pairs	ICC(2,1)	Pearson r	Mean diff	Status
Correctness	2,726	0.634	0.697	-0.094	PASS
Justification quality	2,726	0.063	0.154	-0.241	FAIL
Independence	2,726	-0.106	0.037	-0.351	FAIL

Table G.4: ICC(2,1) results. Threshold = 0.60 [30]. Double scoring coverage = 66.6%.

Only correctness passed the 0.60 threshold. The negative mean differences for justification quality and independence showed that RATER-02 consistently applied a more stringent standard than RATER-A1 on the two subjective dimensions. This was most likely due to insufficient rater calibration before scoring began. Future studies should conduct a calibration session where both raters score the same practice responses together and discuss disagreements before scoring the main dataset [30]. There is also a structural contributor. Because control group responses and all Session 2 responses were scored 2 by default in the absence of copying evidence, the distribution of independence scores is compressed towards the ceiling. An intraclass correlation coefficient falls towards zero when between-participant variance is small, regardless of how consistently the raters applied the rule, so the value reported here reflects the shape of the scoring rule as well as the level of agreement.

Appendix H

Analysis Pipeline Overview

This appendix provides an overview of the full analysis pipeline used in this study. The pipeline was implemented in a single Jupyter notebook. Every step is reproducible from the raw data files using the code provided.

The pipeline follows nine stages, matching Table 3.2 in Chapter 3. Each stage builds on the previous one. The output of each step is either a cleaned dataset, a set of computed variables or a results table that feeds into the next step.

H.0.1 Pipeline Overview

Analysis stage	What was done	Output
Data Integration	Raw files loaded and merged on participant_code. Completers filtered to 1,387. Table A (wide, one row per participant) and Table B (long, one row per episode) built. Group balance, session delay and dropout bias checked. Distributions and outlier patterns visualised.	table_a_wide.csv, table_b_long.csv
Missing Value Treatment	Missing values classified into three types: by design, blank responses and one telemetry row. No imputation applied to design-missing variables. Complete case analysis applied to rubric gaps. One telemetry row filled with zero. Missingness bar charts produced.	Cleaning log, missingness charts
Outlier Treatment	IQR method applied to all 15 numeric variables. Extreme time episodes capped in Table B and participant means recomputed in Table A. Verify rate capped at 1.0 for 85 participants. Accuracy outliers retained (ceiling effect). Before and after comparison plots produced.	Updated table_a_clean, table_b_clean

Analysis stage	What was done	Output
Additional Preprocessing	Categorical variables encoded as ordered integers. Binary group label created. Z-scores computed for all six COI and HES input variables. Shapiro-Wilk normality checks run on all primary endpoints. Final preprocessed tables saved.	table_a_preprocessed.csv, table_b_preprocessed.csv
Feature Engineering	COI and HES built step by step from raw components following the formulas in the scoring specification appendix. Raw components shown with descriptive stats. Z-score standardisation applied and confirmed. COI and HES computed and added to the dataset. Manipulation check (run on HES, since COI has no control-group counterpart), ablation check, weight sensitivity check and construct separation scatter all run and reported.	COI_S1 and HES_S1 columns in dataset
Reliability Checks	Cronbach's alpha computed for all four Likert scales. ICC(2,1) computed for all three rubric dimensions on the double-scored subset (66.6% coverage). Item correlation heatmaps and rater agreement scatter plots produced. Inter-component correlations for COI and HES reported.	Reliability summary table
Statistical Analysis	Mann-Whitney group comparisons run for all primary and secondary endpoints in Session 1 and Session 2. Cohen's d computed with bootstrap 95% CI. Adjusted regression models fitted for both primary Session 2 endpoints. Benjamini-Hochberg correction applied to secondary p-values. Forest plot and full evidence table produced.	Evidence table, forest plot

Analysis stage	What was done	Output
Predictive Modelling	Ridge regression trained using six Session 1 features only. Grouped 5-fold cross-validation by participant ID to prevent data leakage. MAE and RMSE reported against naive baseline. Feature coefficients inspected. Predicted vs actual scatter produced. Subgroup error diagnostics run by education level and AI confidence. Secondary target (accuracy S2) also tested.	Model results, coefficient chart, subgroup diagnostics
Robustness Checks	Four checks run on both primary Session 2 endpoints: remove outliers instead of capping, complete rubric cases only, by collection site and by education level subgroup. Direction and conclusion reported for each check against the reference result.	Robustness summary table

Appendix I

Other useful figures

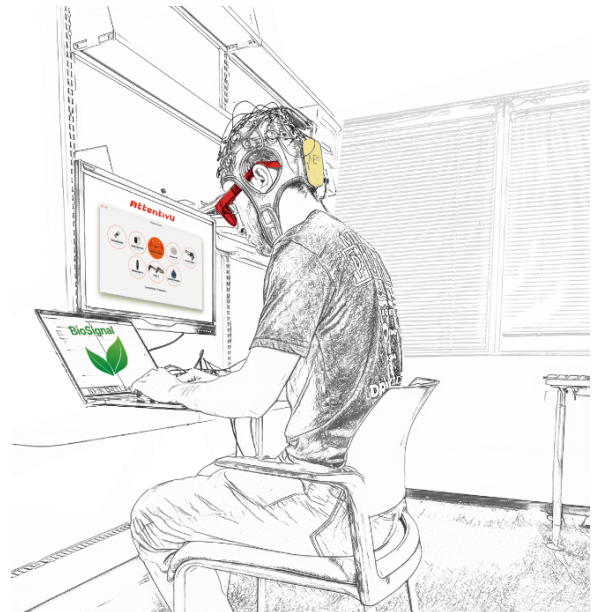
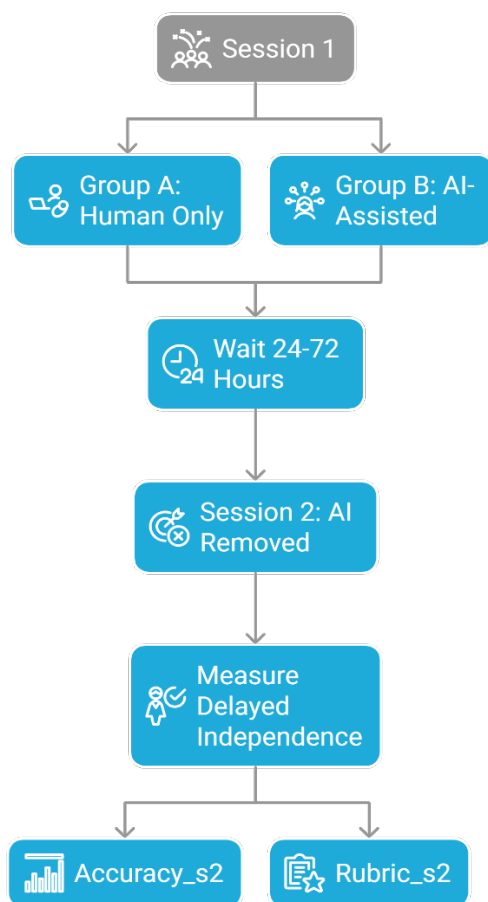


Figure I.1: Two groups in Session 1. AI is removed for everyone in Session 2 to test delayed independence

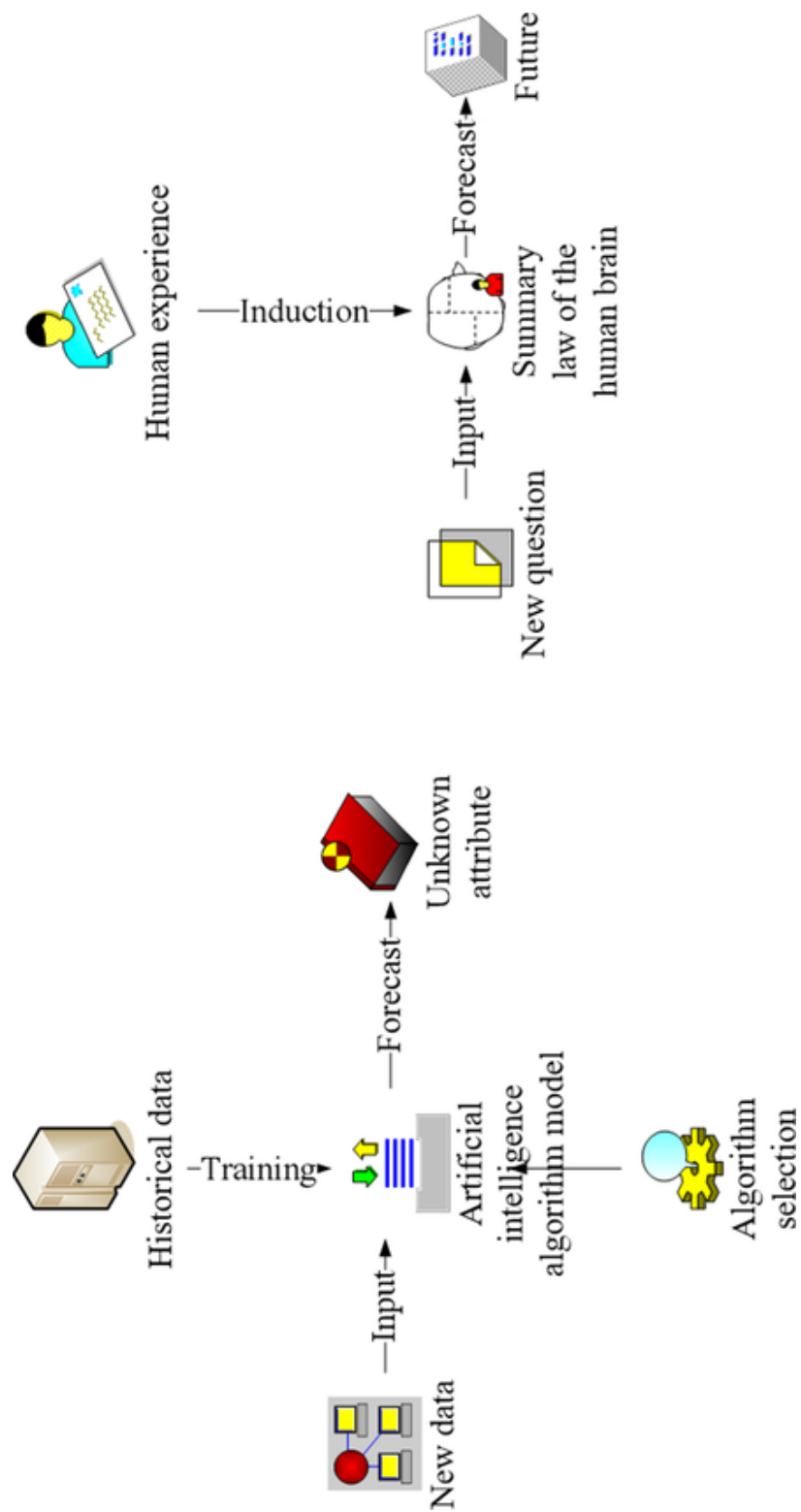


Figure I.2: Shows the comparison between AI and human brain thinking

Appendix J

Research poster layout

A Data Science Framework for Measuring Cognitive Offloading and Independent Problem Solving in AI Assisted Environments

Richard Wambede • Reg No. J25M19005 • MSc Data Science • Uganda Christian University

Problem

Generative AI helps people solve problems faster but may change *how* they think. Most studies measure performance only while AI is present and never test what happens once AI is removed.

Objectives

- 31 Build a two-session dataset and compute COI and HES as auditable indices.
- 32 Evaluate whether Session 1 reliance is associated with Session 2 independent performance.
- 33 Build an exploratory prediction model from Session 1 features only.

Methodology

Session 1: Control group unaided. AI group used a standardised three-level hint bank.

Session 2: Assistance removed for everyone. Parallel tasks after a median 7 minute break.

Behavioural features: AI use ratio, time in tool, copy rate, verification rate. Combined into the **Cognitive Offloading Index (COI)** and **Human Engagement Score (HES)**.

Analysis: Mann-Whitney tests, Cohen's d with bootstrap 95% CI, adjusted regression, Ridge model with grouped cross-validation, Cronbach's alpha and ICC.

Sample

- 1,474 enrolled; 1,387 completed both sessions (94.1% retention)
- AI $n = 687$ • Control $n = 700$
- Lab batches $n = 524$; online $n = 863$; Kampala, Uganda
- 10 episodes each; 13,870 total episodes

Why This Design

A **between-groups** design was used so that no participant met the AI condition before the unaided test, which would have taught them the answers. Groups were balanced at baseline: warmup, AI confidence and Session 1 accuracy all differed by less than 0.04.

Session 1 (AI available)

Measure	AI	Ctrl	d
Accuracy	0.971	0.969	+0.03
Time per task (s)	106.6	106.6	+0.00
Rubric independence	1.206	1.843	-1.97
HES	-0.55	+0.54	-2.47

Same accuracy in the same time, but the control group produced far stronger reasoning (95% CI [-0.672, -0.603]).

Session 2 (AI removed)

Measure	AI	Ctrl	d	p
Accuracy	0.971	0.969	+0.03	0.693
Rubric ind.	1.860	1.854	+0.03	0.611

The Session 1 gap **closed completely**. Both CIs crossed zero and adjusted regression found no group association ($|t^2| < 0.001$). The null result held across all four robustness checks.

Prediction Model

Model MAE 0.1821 vs naive baseline 0.1793 → **did not beat the baseline**.

Measurement Quality

- Manipulation check passed; COI stable under ablation; COI and HES distinct ($r = -0.31$)
- 66.6% double scored; ICC correctness 0.634 (pass), other two **failed**
- All four Likert scales failed Cronbach's alpha

Robustness

The Session 2 null result held under all four checks:

- Removing outliers instead of capping
 - Complete rubric cases only
 - Split by collection site
 - Split by education level
- All effect sizes stayed between -0.11 and +0.15; none reached significance.

Key Finding

AI changed *how* they thought, not *whether* they were right

Reliance showed clearly in **reasoning quality** while assistance was available, but did **not** reduce what participants could do unaided seven minutes later.

Assessing only final answers would have shown **no difference** between the groups.

Scope of the Conclusion

This does **not** show AI is harmless to learning. Retention over days or weeks was not measured, and near-ceiling outcomes prevented a fair test of COI's predictive value.

What the Study Establishes

- AI assistance was associated with weaker independent reasoning *while it was available*
- That difference did **not** persist once assistance was withdrawn
- COI and HES build reproducibly from activity logs and pass validation

Contribution

- Four behavioural features separating **substitution** from **augmentation**
- COI and HES as auditable indices built from activity logs
- A delayed no-AI independence endpoint
- A labelled dataset and reusable pipeline

Alignment

National: Uganda Vision 2040, NDP IV — digital transformation and data-driven decision making.

Global: SDG 4 (Quality Education) • SDG 9 (Innovation and Technology).

Key References

Bupinca et al. (2021) • Grinschgl et al. (2021) • Jarrahi et al. (2022) • Deng et al. (2024) • Gerlich (2025) • Tian & Zhang (2025)

Appendix K

Research Ethics Committee (REC) Approval



UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa



25/06/2026

To: Richard Wambede

UgandaChristianUniversity-MUKONO

Review Type: Initial Review

Re:UCUREC-2026-818: AI vs the Brain: A Data Science Framework that Measures AI Reliance and Human Engagement in Problem Solving

I am pleased to inform you that the Uganda Christian University REC, through expedited review held on **19/06/2026** approved the above referenced study.

Approval of the research is for the period of **25/06/2026** to **25/06/2027**.

As Principal Investigator of the research, you are responsible for fulfilling the following requirements of approval:

1. All co-investigators must be kept informed of the status of the research.
2. Changes, amendments, and addenda to the protocol or the consent form must be submitted to the REC for re-review and approval **prior** to the activation of the changes.
3. Reports of unanticipated problems involving risks to participants or any new information which could change the risk benefit: ratio must be submitted to the REC.
4. Only approved consent forms are to be used in the enrollment of participants. All consent forms signed by participants and/or witnesses should be retained on file. The REC may conduct audits of all study records, and consent documentation may be part of such audits.
5. Continuing review application must be submitted to the REC **eight weeks** prior to the expiration date of **25/06/2027** in order to continue the study beyond the approved period. Failure to submit a continuing review application in a timely fashion may result in suspension or termination of the study.
6. The REC application number assigned to the research should be cited in any correspondence with the REC of record.
7. You are required to register the research protocol with the Uganda National Council for Science and Technology (UNCST) for final clearance to undertake the study in Uganda.

The following is the list of all documents approved in this application by Uganda Christian University REC:

No.	Document Title	Language	Version Number	Version Date
1	Data collection tools	English	FINAL	2026-06-07
2	Protocol	English	FINAL	2026-06-07
3	Main Consent forms	English	FINAL	2026-06-07

Yours sincerely,

Edward Mukooza Kibikyo

For: Uganda Christian University REC

Appendix L

Data collection tools

This appendix describes the data collection instruments used in this study. Data was collected using structured questionnaires, a three-level hint bank and automated behavioural logs recorded by the study platform. All instruments were administered through the online study platform during the same sitting for both sessions.

D1. Background and Warmup Form (Form 0)

Before Session 1 began, all participants completed a short background form and two warmup tasks. The background form collected the following information.

Item	Scale	Purpose
Age band	Categorical	18_24, 25_34, 35_44, 45_PLUS. Used as a background control variable.
Education level	Categorical	SECONDARY, DIPLOMA, DEGREE, POSTGRAD. Encoded as 0 to 3 for use in regression models.
English comfort	1 to 5 Likert	Self-rated comfort reading and writing in English. Used to flag participants who may have struggled with task instructions.
AI use frequency	Categorical	NEVER, MONTHLY, WEEKLY, DAILY. Self-reported frequency of AI tool use before the study.
AI confidence	1 to 5 Likert	Self-rated confidence using AI tools. Used as a covariate in the adjusted regression model.

The two warmup tasks were short structured problems similar to the main tasks. They were scored automatically. The warmup final score (range 0 to 4) was used as the primary baseline ability measure in all group balance checks and regression models.

D2. Session 1 Post-Task Questionnaire

After completing all Session 1 tasks, all participants completed an eight-item questionnaire. Items were rated on a 1 to 5 Likert scale where 1 meant strongly disagree and 5 meant strongly agree. The eight items were organised into two subscales.

Effort subscale (S1-E1 to S1-E3):

Item	Wording
S1-E1	I put a lot of effort into completing these tasks.
S1-E2	I worked hard on every task even when it was difficult.
S1-E3	I tried my best throughout the session.

Engagement subscale (S1-H1 to S1-H3):

Item	Wording
S1-H1	I stayed focused and engaged during the tasks.
S1-H2	I thought carefully about each problem before submitting my answer.
S1-H3	I felt mentally active and involved throughout the session.

The remaining two items (S1-C1 and S1-U1) captured confidence and understanding and were used descriptively only.

Reliability note: All three subscale items failed Cronbach’s alpha in reliability checking. Alpha values were near zero for both the effort and engagement subscales (see Chapter 4, Section 4.6.1). These scale scores were therefore treated as descriptive only and not used as primary outcome measures [29].

D3. AI Usage Questionnaire (AI Group Only)

Participants in the AI-assisted group completed an additional four-item questionnaire after Session 1. Items were rated on a 1 to 5 Likert scale. The four items formed the AI use subscale (S1-AI1 to S1-AI4).

Item	Wording
S1-AI1	I relied heavily on the AI hints to produce my answers.
S1-AI2	I copied or closely followed the AI output in my justifications.
S1-AI3	I accepted AI suggestions without checking them myself.
S1-AI4	I verified the AI outputs before including them in my answers.

These items were intended to capture self-reported AI reliance behaviour to support the computation of COI. The subscale also failed Cronbach’s alpha in reliability checking. Behavioural log features were therefore used as the primary inputs to COI rather than these self-report items [5].

D4. Session 2 Post-Task Questionnaire

After completing all Session 2 tasks, all participants completed a six-item questionnaire. No AI was available in Session 2. Items were rated on a 1 to 5 Likert scale. The six items covered two subscales.

Engagement subscale (S2-H1 to S2-H3):

Item	Wording
S2-H1	I stayed focused and engaged during Session 2.
S2-H2	I thought carefully about each problem without any external help.
S2-H3	I felt mentally active and involved throughout Session 2.

Independence subscale (S2-I1 to S2-I3):

Item	Wording
S2-I1	I was able to solve the tasks on my own without needing any help.
S2-I2	I remembered how to approach these types of problems from before.
S2-I3	I felt confident explaining my answers in my own words.

Reliability note: The independence subscale also failed Cronbach’s alpha in reliability checking. Session 2 scale scores were treated as descriptive only [29].

D5. Hint Bank (AI Group Only, Session 1)

The hint bank was the primary data collection instrument for the AI-assisted group. It replaced the live ChatGPT interface originally planned in the proposal. Each task had three hint levels available.

Level	Label	Description
1	Directional hint	A short pointer toward the correct method without giving the answer. The participant still had to work out all steps themselves.
2	Scaffold	A partially completed worked solution with blanks that the participant had to fill in. The structure of the solution was provided but the participant had to supply the values and reasoning.
3	Near-complete solution	A nearly complete worked solution. The participant had to read, understand and restate the reasoning in their own words in the justification box before submitting.

Participants could request any level and could view multiple hints per task. Each hint request was logged with the level requested, the time the hint was visible and whether the participant copied the hint text. Participants were required to submit final answers and justifications in their own words regardless of which hint level they used.

D6. Behavioural Logs

In addition to questionnaires, the study platform automatically recorded the following behavioural events for all participants during both sessions.

Log variable	What was recorded
duration_total_ms	Total time in milliseconds from task start to submission. Converted to seconds for analysis. Used to compute Time_Offload component of COI.
copy_used	Whether the participant copied hint text into their answer. Boolean per task episode. Used to compute Copy_Rate component of COI.
verified	Whether the participant reported verifying the AI output before submitting. Boolean per task episode. AI group Session 1 only. Used to compute Verify_Rate component of COI.
assistance_level	Highest hint level requested during the task. Integer 0 to 3. 0 means no hints used. AI group Session 1 only. Used to compute AI_Ratio component of COI.
tel_paste_count	Number of paste events recorded by the platform.

Log variable	What was recorded
tel_paste_blocked	Number of paste attempts blocked by the platform.
tel_infraction_count	Number of infraction events recorded.
tel_answer_revisions	Number of times the objective answer was revised before submission.
tel_tab_blur_count	Number of times the participant switched away from the task tab.

These behavioural logs provided objective data that was used as the primary inputs for COI and HES computation. They were more informative than self-report scales alone since they captured what participants actually did rather than what they said they did [5].

D7. Rubric Scoring Sheet

All justification responses were scored by a primary rater using a three-part rubric covering correctness, justification quality and independence. A second rater independently scored 66.6 percent of all responses for inter-rater reliability checking. The full rubric scoring guide with score definitions and examples is provided in Appendix C.

D8. Summary of Data Collection Instruments

Instrument	Group	Session	Used for
Background form	All	Before S1	Baseline variables and covariates
Warmup tasks	All	Before S1	Baseline ability measure
Effort scale (S1-E1 to E3)	All	After S1	Descriptive only (failed reliability)
Engagement scale (S1-H1 to H3)	All	After S1	HES component (descriptive only)
AI use scale (S1-AI1 to AI4)	AI only	After S1	Descriptive only (failed reliability)
Hint bank logs	AI only	S1	COI components (primary source)
Behavioural event logs	All	Both	COI and HES components
Session 2 engagement scale	All	After S2	Descriptive only (failed reliability)
Session 2 independence scale	All	After S2	Descriptive only (failed reliability)
Rubric scoring sheet	All	Both	Primary rubric outcomes and HES

Appendix M

Student - supervisor meeting/supervision plan

Meeting	Date	Agenda	Expected Output
1	29 June 2026	Proposal review and feedback discussion. Agree on research objectives, title revision and three specific objectives. Confirm study design and data collection plan.	Revised title, three agreed objectives, confirmed two-session design and task set.
2	7 July 2026	Data collection update. Review EDA and preprocessing results. Discuss missing value treatment, outlier findings and reliability check outcomes.	Approved EDA notebook, cleaned dataset, documented trim rules and reliability report.
3	14 July 2026	Review COI and HES construction. Discuss feature engineering decisions, validation checks and statistical analysis results for RQ1 and RQ2.	Approved COI and HES scoring specification, evidence table with effect sizes and confidence intervals.
4	21 July 2026	Review predictive modelling results, robustness checks and dashboard prototype. Discuss limitations and how they affect interpretation of findings.	Approved Step 6 and Step 7 outputs, working dashboard, agreed limitations section.
5	28 July 2026	Review of the complete analysis and interpretation of findings. Discuss alignment between the research questions, statistical results and predictive modelling outcomes.	Confirmed analysis and interpretation, revised results and discussion sections, and agreed areas requiring further refinement.
6	5 August 2026	Full draft review. Check the structure, consistency and clarity of all chapters. Review references, tables, figures and presentation of results.	Revised complete dissertation draft, corrected references and improved tables and figures.
7	18 August 2026	Final report and research poster review. Confirm that all supervisor comments have been addressed. Review final formatting, conclusions, recommendations and poster presentation before submission.	Final approved dissertation, completed research poster and readiness for submission and presentation.

Table M.1: Student-Supervisor Meeting Plan: June to August 2026



UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

SCHOOL OF RESEARCH & POSTGRADUATE STUDIES DISSERTATION CORRECTION COMPLIANCE FORM (POST VIVA FORM)

Date: 15/09/2026

Name of Candidate: Wambede Richard

Reg.No: J25M19/005

Title of Dissertation: A Data Science Approach to Measuring Cognitive Offloading and Short-Term Independent Problem Solving in AI-Assisted Tasks: A Two-Session Experimental Study in Kampala, Uganda

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
1	Strengthen the justification for sampling and sample size.	A new subsection, Justification of the Sample Size, was added. It explains that recruitment ran until the data collection window closed, so the sample size is justified after the fact rather than before it. The weak claim that cited a twelve-per-group pilot rule was removed. Two paragraphs were added to the Sampling Method subsection explaining why convenience sampling is acceptable for an experiment. In this design the fair comparison rests on how people were allocated between conditions, not on how they were recruited. It is also stated that the sample size does not improve how well the sample represents the wider population.	Ch 3: Sampling Method, Justification of the Sample Size
2	Clearly distinguish short-term independent performance from long-term skill retention.	Two new definitions were added to the Key Definitions section. Short-term independent performance and Long-term skill retention. The existing definition of Delayed Independent Performance was reworded to point to the short-term term and to state that it never means long-term retention. The term skill erosion was removed from the abstract, Chapter 1, Chapter 3, Chapter 4, Chapter 5 and Chapter 6, and replaced with short-term independent performance. The abstract now states plainly that the design did not test long-term skill retention and that nothing in the results speaks to what	Abstract, Ch 1: Key Definitions, Ch 6 Future Research

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
		happens after a day, a week or a semester of regular AI use. The future research recommendation was corrected to state that a longer delay tests the question this study could not reach.	
3	Moderate causal and generalisation claims.	The Group Assignment subsection was rewritten to state that allocation was random, and to explain what random allocation does and does not allow the study to claim. This was previously left vague. The attribution sentence in Chapter 5 was reworded. It now says the AI condition is the most likely explanation, and adds that the claim is limited to the participants, tasks and setting of this study and is not a claim about AI use in general. An opening sentence was added to Practical Implications saying that the implications follow from one sample, one task set and one short delay, and are offered as steps to consider rather than conclusions that hold everywhere.	Ch 3: Group Assignment, Ch 5: What This Study Can and Cannot Conclude
4	Discuss more fully the impact of failed scale and rubric reliability.	A table was added to show which results depend on each measure and what each reliability failure means for the results. The table shows that the failures occurred in measures based on judgement and self-report, while the auto-scored accuracy measures and behavioural logs were not affected. A paragraph was also added to explain that two of the three HES components came from measures that failed the reliability check. Two paragraphs were added to explain attenuation. An unreliable measure can reduce observed differences towards zero, meaning that the Session 2 null finding provides weaker evidence of no difference than it may appear. The limits of this interpretation are also explained. The conclusion is therefore stated more carefully: Session 2 found no difference large enough for this measure to detect, rather than proving that no difference existed at all.	Ch 5: Reliability Limitations and Their Impact
5	Clarify the validation status of the COI and HES indices.	A new subsection, Validation Status of the Two Indices, was added to clearly separate the validation work that was done from what was not done.	Abstract,

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
		What was done: The formulas were clearly defined for reproducibility, together with a component-dependence test, weight sensitivity test, and construct separation test ($r = -0.361$). What was not done: There was no expert review of content validity, external measure for convergent validity, test-retest reliability, or factor analysis. The indices are therefore described as reproducible descriptive indices rather than validated measurement tools. The term validated was also removed from the abstract, Chapter 4, and Chapter 6 where it overstated the work done. A note was also added to the ablation check to clarify that, because COI is based on the mean of standardised components, the test mainly shows whether the centre of the distribution remains stable when a component is removed.	Ch 4: Validation Status of the Two Indices
6	Verify citations and references.	Every citation key was checked against the reference list. All citations now resolve and there are no undefined keys. The four wrong entries were corrected, and some references deleted from the list.	References
7	Thoroughly proofread the dissertation.	A full proofreading pass was carried out. All broken cross-references and sentences were fixed and repaired, contradictions between chapters were removed, repeated material was removed, missing spaces, and corrupted equation characters were corrected.	Throughout the entire thesis report

S/N	COMMENTS BY INTERNAL EXAMINER	ACTION TAKEN	INDICATOR
1	Why was the Cognitive Offloading Index included when the reported findings do not show it? The findings show independent problem solving instead. Align the title and the framing of the study with what the findings actually show.	The mismatch was traced to Research Question 2, which promised that COI would measure the Session 1 to Session 2 relationship, when COI is computed from z-scores taken within the AI-assisted group only and therefore cannot produce a between-group result. Four changes were made. First, the reliability clause was removed from Research Question 2 and moved to Research Question 1, where the indices are built and checked, so COI is now claimed only where it	Ch 1: Research Questions. Ch 3: Where COI Sits in This Study. Ch 4: COI Construction. Ch 5.

S/N	COMMENTS BY INTERNAL EXAMINER	ACTION TAKEN	INDICATOR
		is reported. Second, a new subsection, Where COI Sits in This Study, was added after the alignment table in Chapter 1. It states that COI is a within-condition individual-difference measure serving Objective 1 and Objective 3, that HES and rubric independence carry the between-group comparison in Objective 2, and that COI is therefore absent from the Session 2 evidence table by design rather than by failure. Third, the same point was added at the head of the COI Construction section in Chapter 4, and the answers to Research Questions 1 to 3 in Chapter 6 were realigned to the wording now used in Chapter 1. Fourth, the two remaining uses of the word validated in relation to COI were removed from Chapter 5, for consistency with External Examiner comment 5. The title was reviewed and kept as it stands. Cognitive offloading was demonstrated in the findings, through the Session 1 rubric independence difference of $d = -1.972$ and the HES difference of $d = -2.474$, and short-term independent problem solving was measured directly in Session 2, so both halves of the title are supported by reported results.	Ch 6: RQ1, RQ2, RQ3.
2			
3			

S/N	COMMENTS BY VIVA VOCE PANNEL	ACTION TAKEN	INDICATOR
1	Explain how you handled data leakage.	A new subsection called Leakage Controls was added to Chapter 3, setting out four separate controls i.e Time separation, Target separation, Participant separation and Preprocessing.	Ch 3, Predictive Modelling, Leakage Controls
2	The model used for this study is not clearly talked about.	A new subsection, Choice and Specification of the Model, was added. It first separates the two kinds of model in the study. The adjusted regression is explanatory and is read through its coefficients. The Ridge model is predictive and is read through its error on data it has not seen. The Ridge objective function is now given as an equation, with the squared error term and the penalty term explained in plain words. Three reasons for choosing Ridge are now given and the full specification is now listed.	Ch 3, Choice and Specification of the Model

S/N	COMMENTS BY VIVA VOCE PANNEL	ACTION TAKEN	INDICATOR
3	How was bias handled, and how was data validated to remove bias?	A new section, Bias Control and Data Validation, was added to Chapter 3, with a table listing each type of bias, the control applied and where the evidence appears. The types covered are selection bias, allocation bias, attrition bias, confounding, measurement bias in the tasks, rater bias, response bias in self-report, missing data bias, site and subgroup bias, and multiple comparison bias. The checks already run are now gathered in one place: the dropout check comparing completers and non-completers, the group and site balance table, the adjusted regression controlling baseline ability, education and AI confidence, the four robustness checks, the subgroup error diagnostics, and the Benjamini-Hochberg correction. Two checks that did not come out clean are reported openly. The second rater scored lower than the primary rater on both subjective dimensions, and recruitment stopped when the window closed rather than at a planned target.	Ch 3, Bias Control and Data Validation
4	Did you do any comparison of results with any existing models?	Two comparisons were added. First, a comparison of findings. A table was added to Chapter 5 setting this study against Bastani et al. (2024), Kosmyrna et al. (2025), Gerlich (2025) and Deng et al. (2024), showing the design, the delay before the unaided test and the finding for each. The pattern is that studies finding a lasting cost all tested after days or weeks, while this study tested after minutes. Second, a comparison of models. A new subsection explains why a more flexible model would not change the result. A recommendation was added to Chapter 6 that a follow-up study should benchmark Ridge regression against at least one flexible model on data with a wider spread of outcomes.	Ch 5: Findings comparison table, Ch 4: Why a More Flexible Model Would Not Change This Result, Ch 6: Recommendations
5	There is no evidence of a framework in your write up.	A Conceptual Framework section with a diagram was added at the end of Chapter 2. The diagram shows the condition, which is whether AI assistance was available in Session 1. It then shows the behaviour measured in between, which is COI from four logged behaviours and HES from three signals. It	Ch 2: Conceptual Framework

S/N	COMMENTS BY VIVA VOCE PANNEL	ACTION TAKEN	INDICATOR
		then shows the outcomes, which are accuracy and rubric independence measured first with AI present and then with AI removed. The controlled variables are shown separately. The three research questions are mapped onto three parts of the figure. Research Question 1 asks whether the middle layer can be measured, Research Question 2 asks whether the condition affects the right-hand layer, and Research Question 3 asks whether the middle layer can predict the right-hand layer for one individual.	
6	Both the data processing pipeline and the data collected for this study are not clear.	A list was added at the start of the data processing section setting out exactly what each participant produced: one participant record, two session records, ten task episodes, ten to twenty rubric scores, sixteen Likert responses and a variable number of telemetry events, and the totals are now stated. The nine-stage pipeline table was kept and is now linked to the conceptual framework, so the reader can see what is measured and how the raw logs become those measures. The tables listing the raw files, the merged analysis tables and their row counts were kept in Chapter 4 so the path from raw export to analysis table can be followed.	Ch 3: Data Processing and Analytic Approach, Ch 4: Data Integration
7	What metrics did you use to measure cognitive offloading?	The four metrics are now named in the abstract as well as in the methodology, so the reader meets them on the first page. Cognitive offloading was measured with the Cognitive Offloading Index, built from four behaviours i.e the AI use ratio, Session 1 task time, the copy rate, and the verification rate. Each of the four is converted to a z-score and the four are averaged. None of the four comes from what participants said about themselves. The definition of COI in Chapter 1 was corrected. It previously suggested a threshold of 0.7.	Abstract, Ch 1: Key Definitions, Ch 3: COI Components and Formula
8	Align your research title to the fraction of your research that was completed.	The title was changed to: A Data Science Approach to Measuring Cognitive Offloading and Short-Term Independent Problem Solving in AI-Assisted Tasks: A Two-Session Experimental Study in Kampala, Uganda.	Title page

S/N	COMMENTS BY VIVA VOCE PANNEL	ACTION TAKEN	INDICATOR
		The word prediction was removed from the title, because the prediction objective was exploratory and the model did not beat a naive baseline. Short-Term was added, so the title itself states that only short-term independent performance was measured. AI-Assisted Tasks replaced AI-Assisted Environments, which is narrower and matches what was actually studied. The design and the location were added to the title, so the two-session structure and the Kampala setting are visible before the reader opens the document.	
9	Can the findings of your research be generalized to all universities?	A new subsection, Boundaries of Generalisation, was added to Chapter 5. It answers the question directly with no. Three reasons are given. The sample was a convenience sample from two sites in one city with no sampling frame. The participants were not a university body but a mixture of students from universities and vocational institutes and working adults, recruited during a holiday period. The tasks were completed in a session arranged for research with no consequence attached, unlike graded university work.	Ch 5: Boundaries of Generalisation
10	Referencing objective 3, why was prediction necessary for your research when your primary interest was establishing the effect of AI use?	Objective 3 and Research Question 3 were reworded. Objective 3 now reads as testing whether Session 1 reliance behaviour alone can estimate the Session 2 independence of an individual participant, and so whether COI could work as an early screening measure. Research Question 3 now also asks whether the conditions of this study allow that question to be tested fairly. An explanation of why the two objectives differ was added to Chapter 3. A matching response was also added to the Chapter 6 answer to Research Question 3.	Ch 1: Objectives and Research Questions, Ch 3: Predictive Modelling, Ch 6: RQ3
11	It must be clear how each of your specific objectives map onto your research questions and your analysis.	An alignment table was added showing all three objectives on one page. Each row runs from the objective, to the research question, to the data used, to the method applied, to where the result is reported, to the answer reached. The introductions to Chapter 3, Chapter 4 and Chapter 6 were each given a sentence	Ch 3: Introduction, alignment table, Ch 4 and Ch 6 Introductions

S/N	COMMENTS BY VIVA VOCE PANNEL	ACTION TAKEN	INDICATOR
		pointing to the table, so the same structure is visible from every chapter.	

Candidate's Name: **Wambede Richard**

Signature: *Richard Wambede*

Supervisor's Name: **Dr. Kakuba Samuel**

Signature



NB: Post Viva compliance form is designed to capture all the corrections recommended by internal examiner (supervisor), external examiner and viva panel.