data

MDPI

Article

From Offloading to Engagement: An Experimental Study on Structured Prompting and Critical Reasoning with Generative AI

Michael Gerlich

Center for Strategic Corporate Foresight and Sustainability, SBS Swiss Business School, 8302 Kloten, Switzerland; michael.gerlich@cantab.net

Abstract

The rapid adoption of generative AI raises questions not only about its transformative potential but also about its cognitive and societal risks. This study contributes to the debate by presenting cross-country experimental data (n = 150; Germany, Switzerland, United Kingdom) on how individuals engage with generative AI under different conditions: human-only, human + AI (unguided), human + AI (guided with structured prompting), and AI-only benchmarks. Across 450 evaluated responses, critical reasoning was assessed via expert rubric ratings, while perceived reflective engagement was captured through self-report indices. Results show that unguided AI use fosters cognitive offloading without improving reasoning quality, whereas structured prompting significantly reduces offloading and enhances both critical reasoning and reflective engagement. Mediation and latent class analyses reveal that guided AI use supports deeper human involvement and mitigates demographic disparities in performance. Beyond theoretical contributions, this study offers practical implications for business and society. As organisations integrate AI into workflows, unstructured use risks undermining workforce decision making and critical engagement. Structured prompting, by contrast, provides a scalable and low-cost governance tool that fosters responsible adoption, supports equitable access to technological benefits, and aligns with societal calls for human-centric AI. These findings highlight the dual nature of AI as both a productivity enabler and a cognitive risk, and position structured prompting as a promising intervention to navigate the emerging challenges of AI adoption in business and society.

Keywords: generative artificial intelligence; GenAI; cognitive offloading; AI; critical thinking; human–AI interaction; reflective reasoning; digital literacy

Academic Editor: Kamran Sedig

Received: 5 September 2025
Revised: 20 October 2025
Accepted: 28 October 2025
Published: 30 October 2025

Citation: Gerlich, M. From Offloading to Engagement: An Experimental Study on Structured Prompting and Critical Reasoning with Generative AI. *Data* **2025**, *10*, 172. https://doi.org/10.3390/data10110172

1. Introduction

The rise of generative artificial intelligence (GenAI) has reshaped the cognitive landscape of modern societies. From education and business to politics and public discourse, large language models such as ChatGPT have become widely adopted tools for writing, research, and problem solving. Their accessibility and output fluency have led to widespread enthusiasm about the potential of AI to democratise knowledge and enhance individual productivity. However, recent empirical evidence suggests that this optimism may obscure a critical shift: rather than supporting cognitive processes, GenAI tools often invite users to delegate them. This phenomenon, commonly referred to as cognitive offloading, raises foundational questions about how AI affects reasoning, ownership of ideas, and the capacity for critical thinking. While enthusiasm about GenAI is widespread, a growing body of interdisciplinary literature has raised concerns about its cognitive consequences.

Recent reviews emphasise similar risks of intellectual deskilling and reduced epistemic vigilance in AI-supported environments [1–3]. Beyond studies on offloading, scholars have examined its role in cognitive load regulation [4], metacognitive monitoring [2], and the erosion of epistemic vigilance [3]. In the learning sciences, self-regulation has long been identified as a core determinant of deep engagement and transfer [5]. Without deliberate self-regulation, users may default to convenience rather than reflection when interacting with AI. Similarly, work on epistemic responsibility in human–AI collaboration has shown that uncritical reliance on AI risks displacing human authorship and agency [1]. These contributions situate GenAI not only as a tool of productivity but as a factor that reshapes the conditions of reasoning itself.

Cognitive offloading refers to the act of shifting internal mental effort onto external tools. While this is not new, humans have long used notebooks, calculators, and digital calendars to support cognition; the type of offloading now facilitated by GenAI is qualitatively different. These systems not only store or retrieve information but also synthesise, argue, and evaluate on behalf of the user. As a result, tasks that once required thoughtful reflection can now be completed with minimal mental effort. Gerlich [6] argues that this shift undermines the development and application of critical thinking, particularly in settings where users become passive recipients of AI-generated content. In a large-scale study involving student and workplace populations, Gerlich found a consistent negative relationship between GenAI use and critical argument quality, especially when no structural prompting or reflective engagement was required. The study concluded that while GenAI is often praised for its efficiency, its unsupervised use risks displacing rather than supporting the human cognitive functions it was meant to augment.

These concerns have been further validated by recent interdisciplinary research. A longitudinal study conducted at MIT by Kosmyna et al. [7] used electroencephalography (EEG) to examine the neural impact of LLM-assisted essay writing over four sessions. Participants were divided into LLM, search engine, and unaided writing conditions. The findings revealed a significant reduction in brain connectivity and cognitive engagement in the LLM group, which showed the weakest activation across key neural networks associated with reasoning and memory. Moreover, participants using LLMs demonstrated lower ownership of their essays and struggled to recall or quote their own arguments. The researchers concluded that GenAI use can lead to an accumulation of what they termed "cognitive debt", whereby users become neurologically and behaviourally less engaged over time, even when later completing tasks without assistance. These results echo broader concerns in behavioural science. The design of the present experiment was informed by the methodological precedents set in these two key studies. Gerlich [6] employed large-scale survey and essay-based experiments in both educational and workplace settings to quantify the relationship between GenAI usage patterns and critical thinking performance, revealing that unstructured AI use consistently reduced argument quality. Kosmyna et al. [7], by contrast, used a longitudinal EEG-based design to monitor neural activation across repeated writing sessions, demonstrating reduced cognitive engagement among LLM users compared with control groups using search engines or unaided writing. Together, these studies established both the behavioural and neural foundations for investigating cognitive offloading, motivating the need for an intervention-based design such as the one adopted here. By combining Gerlich's behavioural focus with Kosmyna's emphasis on cognitive engagement, the present study extends this line of inquiry toward practical strategies for mitigating offloading through structured prompting.

Lee et al. [8] found that professionals using GenAI tools reported reduced cognitive effort but increased confidence in the quality of their outputs, revealing a mismatch between perceived and actual engagement. Together, these studies draw attention to a growing

paradox in human–AI interaction. While AI tools are designed to support human users, they often enable patterns of behaviour that lead to reduced mental effort, lower critical engagement, and diminished learning. This issue becomes particularly salient in contexts where reflection, argumentation, and judgement are essential, such as civic reasoning, education, and professional writing. Yet despite growing awareness of these risks, very few empirical studies have tested whether the negative effects of cognitive offloading can be mitigated by training users in how to interact with AI more intentionally. Initial conceptual proposals for such training frameworks have emerged under the terms 'prompt literacy' [9] and 'dialogic AI engagement' [10], but empirical validation has been lacking.

The present study addresses this gap by moving from diagnosis to intervention. It builds directly on the foundational insights of Gerlich [6], who proposed that AI tools should not be abandoned, but rather reconfigured in their mode of use. In contrast to the dominant model of passive interaction, this study examines whether users can be trained to engage GenAI systems as research instruments rather than as cognitive substitutes. The theoretical premise is that offloading is not an inevitable outcome of AI use, but a function of how the interaction is structured. Drawing from cognitive psychology and educational theory [11,12], we hypothesise that structured prompting and reflective engagement can help users maintain ownership of their reasoning process while benefiting from AI-generated information. To test this hypothesis, we designed an experiment involving four conditions: participants responding to a political reasoning task without AI (human-only), participants using ChatGPT to generate the response independently (AI-only), participants using ChatGPT without any usage training (human + AI, unguided), and participants trained in structured prompting designed to avoid cognitive offloading (human + AI, guided). The task was intentionally framed around the question "What are the advantages and disadvantages of democracy?", allowing participants to draw on general knowledge while engaging in reasoning and argument construction. Across all conditions, participant performance was evaluated by independent raters using a validated critical thinking rubric, and post-task reflections were captured through a structured questionnaire. In addition, qualitative interviews provided insight into the participants' own experiences of difficulty, confidence, and perceived dependence on AI. While some degree of cognitive offloading is inevitable in human–technology interaction [13], the critical question is not whether offloading occurs, but whether it can be shaped to preserve reflective engagement. This study, therefore, does not treat offloading as inherently negative but investigates whether structured prompting can reduce uncritical delegation and support deeper reasoning.

This study contributes to the growing literature on human–AI interaction by offering an evidence-based framework for using AI to enhance rather than replace cognition. It tests whether structured prompting can reverse the cognitive passivity observed in previous studies and foster deeper engagement with information and ideas. In doing so, it addresses an urgent question in the age of AI: can we design interaction patterns that preserve the uniquely human capacity for thought, judgement, and learning in an era of machine-generated answers? This study aims to investigate whether the use of generative AI tools necessarily leads to cognitive offloading and diminished reasoning quality, or whether these effects can be mitigated through structured engagement strategies. Building on previous findings that highlighted a consistent pattern of cognitive underperformance among unguided AI users [6,7], this research tests whether training participants to use AI as an informational tool rather than a task-solving agent can preserve or enhance cognitive effort and critical thinking.

This study also draws on established cognitive theories to frame its hypotheses. Cognitive load theory explains how external supports can either reduce or distort the allocation of cognitive resources [4,14]. Dual-process models of reasoning [15,16] suggest that AI tools

may encourage intuitive acceptance (System 1) rather than reflective evaluation (System 2), particularly when fluency and coherence mask logical weaknesses. Structured prompting can be conceptualised as a metacognitive scaffold that activates self-regulated learning strategies [5], requiring users to pause, reflect, and test their own assumptions. In this way, prompting serves to counteract the risk of epistemic surrender and preserve epistemic agency in human–AI interaction [1]. This theoretical grounding connects our hypotheses to a broader tradition in cognitive psychology, metacognition, and learning sciences. Building on prior findings that identified age and educational attainment as key moderators of cognitive offloading and argument quality [6], this study further examines demographic variation in the context of guided versus unguided AI use. Demographic variables such as age and educational attainment have been shown to influence susceptibility to automation bias and reliance on external supports [6,17,18]. Including these variables allows us to test whether structured prompting moderates such demographic disparities. The experimental design was deliberately constructed to ensure comparability and accessibility across participant groups and countries. The task question "What are the advantages and disadvantages of democracy?" was selected because it allows all participants to engage in reasoning without requiring specialised domain knowledge, making it suitable for both student and professional populations. Recruitment across Germany, Switzerland, and the United Kingdom targeted academically literate and digitally active individuals to ensure participants could meaningfully interact with generative AI tools and reflect on their own reasoning process. A mixed-methods approach was chosen to combine measurable performance data with self-reported cognitive processes, as internal engagement cannot be captured through performance scores alone. By integrating these design decisions, the study provides both rigour and ecological validity, capturing how real users engage with AI-supported reasoning in authentic contexts. Country was included as a contextual factor to explore whether the observed patterns are robust across national settings. To ensure that both observable and internal cognitive processes could be assessed, two complementary measurement approaches were employed. Critical reasoning quality was evaluated through expert-rated rubric scores that provided an objective, performance-based outcome measure. In contrast, cognitive offloading and perceived reflective engagement were captured through structured self-report items designed to reveal subjective aspects of participants' thought processes that cannot be inferred from written outputs alone. The combination of expert evaluation and self-report follows standard practice in cognitive and educational research, allowing the study to examine not only what participants produced but how they engaged cognitively with the task. To this end, the study was guided by the following hypotheses:

H1. *There will be significant differences in argument quality across the four experimental conditions (human-only, AI-only, human–AI unguided, and human–AI guided).*

H2. *Participants in the human–AI-guided condition will demonstrate significantly higher reasoning quality than those in the AI-only and unguided conditions.*

H3. *Structured prompting will be associated with reduced cognitive offloading and increased self-reported perceived reflective engagement compared to unguided AI use.*

H4. *Participants' demographic characteristics, particularly age and education, will correlate with both cognitive offloading and performance outcomes, with younger and less-educated participants more susceptible to AI overreliance.*

Previous work has shown that older participants and those with higher levels of education are less likely to rely uncritically on AI outputs, reflecting greater epistemic

vigilance [17,18]. This evidence provides a theoretical and empirical basis for H4, which hypothesises that age and education will correlate with both offloading and performance outcomes. In testing these hypotheses, the study seeks not only to assess the cognitive risks of GenAI use but also to propose a viable method for structuring AI interaction to protect and support reflective reasoning. By integrating performance data, self-reported cognition, and qualitative reflections, this research contributes to a more nuanced understanding of what it means to think critically in an age of machine-generated knowledge.

Unlike prior studies that have primarily diagnosed cognitive risks of GenAI use (e.g., Kosmyna et al. [7], showing reduced neural engagement), the present study moves toward intervention. It tests whether structured prompting can mitigate cognitive offloading and foster deeper engagement, thereby offering a practical strategy for education and knowledge work.

2. Materials and Methods

This study employed a controlled, multi-condition experimental design to investigate the cognitive consequences of generative AI use and the potential benefits of structured prompting. The research was conducted in three countries—Germany, Switzerland, and the United Kingdom—with participants recruited from academic institutions, professional workshops, and AI-focused conferences. The study is built on prior findings that suggest generative AI tools may lead to cognitive offloading unless used in a guided, reflective manner [6]. By integrating performance scores, self-reported cognitive responses, and qualitative interviews, the research combined quantitative and qualitative methods to offer a comprehensive view of human–AI interaction in reasoning tasks.

Participants were randomly assigned to one of four conditions: (1) human-only, where the task was completed without AI assistance; (2) AI-only, in which the task was completed entirely by ChatGPT; (3) human + AI (unguided), where participants used ChatGPT without structured instruction; and (4) human + AI (guided), in which participants were trained in how to use ChatGPT in a way designed to avoid cognitive offloading and stimulate critical thinking. The structure and distribution of these groups enabled comparative and inferential analysis of how different types of AI interaction influence argument quality and cognitive engagement.

2.1. Participants

A total of 150 participants completed the experiment, with 50 participants recruited in each of the three countries (Germany, Switzerland, and the United Kingdom) (Table 1). Recruitment followed a stratified approach to ensure variation in age, education, and professional background. Participants were invited through university mailing lists, professional development workshops, and AI-related public events advertised through institutional newsletters and online forums. This mix of recruitment channels was chosen to balance academic and non-academic populations and to capture a realistic spectrum of digital literacy levels. Participation was voluntary and unpaid. The final sample included secondary-school students, undergraduate and postgraduate students, and working professionals from diverse sectors, ranging from education and public administration to marketing and IT. This recruitment design ensured that participants had sufficient language proficiency and digital familiarity to engage meaningfully with generative AI, without restricting the sample to academic experts. All participants completed three experimental conditions (human-only, human + AI unguided, and human + AI guided). The AI-only condition served as a fixed benchmark generated independently via ChatGPT, based on the same task prompt. The question presented to all conditions was: *"What are the advantages and disadvantages of democracy?"* This prompt was deliberately chosen for its general accessibility

and relevance to public reasoning, ensuring participants could respond without needing specialised political knowledge or prior AI interaction. In summary, the study comprised four conditions: (1) Human-only, (2) Human + AI (unguided), (3) Human + AI (guided with structured prompting), and (4) AI-only (ChatGPT-generated benchmark without human input). The structured prompting protocol in the guided condition followed a five-step sequence (initial reflection, targeted research use, argument construction, critical review, and final reflection), which is detailed below.

Table 1. Participant characteristics across countries and conditions.

Country	*n*	Age (Mean ± SD)	Age Range	Education (% High School/Bachelor/Postgraduate)	Gender (% Female/Male)
Germany	50	33.1 ± 10.9	15–58	22/54/24	48/52
Switzerland	50	31.7 ± 10.6	14–60	18/56/26	49/51
United Kingdom	50	32.2 ± 9.9	14–57	24/50/26	52/48
Total	**150**	32.3 ± 10.5	14–60	21/53/26	50/50

Note: All participants completed the three human-involved conditions (Human-only, Human + AI unguided, Human + AI guided). The AI-only condition was produced separately by ChatGPT as a fixed benchmark.

The study design follows established practices in the social sciences, where mixed methods, combining self-report data, performance-based measures, and qualitative interviews, are widely used to capture both observable behaviour and subjective cognitive processes. Self-report instruments are particularly valuable in exploring metacognition and reflective engagement, as they provide direct insight into participants' perceptions and strategies, which cannot be inferred from performance scores alone. To strengthen validity, the self-report indices were triangulated with expert-rated rubric scores and qualitative interviews. This triangulation is a standard approach in sociology and education research to enhance robustness and mitigate the limitations of any single method.

The study was conducted in accordance with the Declaration of Helsinki and received institutional ethics approval prior to data collection. Before participation, all individuals received a written information sheet outlining the study's purpose, procedures, and data handling policy. The sheet clarified that participation was voluntary, that responses would be anonymised, and that participants could withdraw at any time without consequences. No personally identifying data was collected, and all datasets were stored securely in compliance with GDPR standards. Informed consent was obtained electronically or in writing from all participants prior to data collection.

The AI benchmark responses were generated with ChatGPT (version 4.0, OpenAI) in standard mode, with web browsing and plug-ins disabled. All participants used the same version under identical conditions.

2.2. Data Collection Procedure

The study was conducted in two stages: a pilot and the main experiment. The study employed a within-subjects design for the three human-involved conditions (Human-only, Human + AI unguided, and Human + AI guided), allowing each participant to serve as their own control. The AI-only condition was implemented separately as a fixed benchmark to provide a performance reference independent of human variation. To mitigate potential learning or fatigue effects, the order of the three human-involved conditions was counterbalanced across participants, and the pilot study was used to ensure that task length and difficulty were appropriate. Participants completed the entire experiment within a single supervised session lasting approximately 60–75 min, with short breaks between tasks to avoid cognitive fatigue.

The pilot involved fifteen participants and was used to test the clarity of instructions, feasibility of structured prompting, and the reliability of the scoring rubric. After refinement, the full-scale data collection was carried out in three countries, Germany, Switzerland, and the United Kingdom, with fifty participants per country. Participants completed the study in controlled workshop sessions or supervised online formats, depending on location.

The experimental task was consistent across conditions: participants were asked to construct a reasoned response to the question *"What are the advantages and disadvantages of democracy?"* The task was chosen for its conceptual relevance, accessibility, and neutrality, ensuring it could be completed without specialist knowledge while still requiring argumentative reasoning.

All participants first completed the task in the human-only condition, without access to AI. In the second condition (human + AI unguided), participants were given access to ChatGPT but were not provided with any guidance on how to use it. In the third condition (human + AI guided), participants received structured training designed to promote deliberate engagement and reduce cognitive offloading. The AI-only condition was operationalised separately by generating a reference response using ChatGPT alone, without human input.

2.3. Structured Prompting and Cognitive Intervention

Participants in the guided condition were introduced to a structured prompting protocol that required metacognitive reflection and deliberate interaction with ChatGPT. The process was divided into the following steps:

1. Initial Reflection: Participants were first asked to consider how they would answer the question without using AI and to formulate preliminary hypotheses or argument directions on their own.
2. Targeted Research Use: Participants were then instructed to use ChatGPT exclusively for retrieving contextual or factual information. Prompts were constrained to data-focused queries, avoiding any phrasing that would ask the model to directly generate or evaluate arguments. To ensure compliance with this instruction, participants' screen activity was monitored in real time during the guided sessions, and sample prompt templates were provided in advance to illustrate acceptable versus non-compliant phrasing. Facilitators intervened only when a participant deviated from the defined structure, reminding them to reformulate the query as factual or exploratory rather than generative.
3. Argument Construction: Participants revised their preliminary responses based on the information collected, without directly copying AI-generated text.
4. Critical Review: Participants submitted their constructed arguments to ChatGPT, asking it to identify missing dimensions or propose counterarguments.
5. Final Reflection and Revision: Participants refined their arguments using these insights, maintaining personal authorship and accountability for reasoning choices.

This protocol was designed to preserve agency and deepen engagement, treating AI as a research tool rather than a cognitive replacement.

2.4. Measures and Instruments

2.4.1. Critical Thinking Rubric

Each participant response in the human-only, human + AI unguided, and human + AI-guided conditions was independently evaluated by a panel of three expert raters. All raters held doctoral-level qualifications in political science or education and were blinded to the experimental condition of the response. Responses were anonymised and assessed

using a validated rubric adapted from Facione [11] and Halpern [12], both of which provide widely used frameworks for evaluating critical thinking in written argumentation.

The rubric assessed five key dimensions (Appendix A):

- Clarity and Structure of Argument;
- Logical Coherence and Justification;
- Depth of Reasoning and Use of Evidence;
- Recognition of Counterarguments;
- Originality and Synthesis.

Each dimension was scored on a six-point Likert scale (1 = very weak, 6 = excellent), allowing for a total possible score range of 6 to 30. Inter-rater reliability was calculated using Krippendorff's alpha, which demonstrated acceptable agreement ($\alpha \geq 0.70$). In cases where reliability was borderline, median scores were cross-checked against means for sensitivity analysis. The mean total score across raters served as the primary dependent variable for subsequent analysis. Critical thinking was assessed through expert ratings, not via self-report. These ratings were based on a validated critical thinking rubric (adapted from [11,12] that evaluates five key dimensions: clarity and structure of argument, logical coherence, depth of reasoning, recognition of counterarguments, and originality. The rubric and detailed scoring criteria are provided in Appendix A. Perceived reflective engagement, by contrast, was captured through self-report items adapted from prior work on metacognition and reflective engagement. These self-reports were triangulated with rubric scores and interview data, which strengthens construct validity despite not using a standardised scale.

2.4.2. Post-Task Questionnaire

Following the completion of all three human-involved conditions, participants completed a structured questionnaire (Appendix B) designed to assess their subjective experience of the task. Items were grouped under three latent cognitive constructs:

- Cognitive Offloading (e.g., "I relied on AI to do much of the thinking for me");
- Perceived reflective engagement (e.g., "The task made me reflect deeply on the issues");
- Perceived Difficulty (e.g., "It was hard not to offload when using AI").

Each construct was measured using three statements, with responses captured on a six-point Likert scale (1 = strongly disagree, 6 = strongly agree). Scores were averaged across items within each construct to form composite indices. These indices were used both as dependent variables (in ANOVAs and MANOVA) and as predictors in regression and mediation models. Although the self-report items were adapted from established constructs in the learning sciences [2,3,12], they do not constitute a pre-validated psychometric scale. To address this limitation, we conducted internal consistency checks, and results were triangulated with performance-based rubric scoring and qualitative interviews, which provides convergent validity. Critical thinking performance itself was not self-reported but evaluated with a validated rubric [11,12], with acceptable inter-rater reliability (Krippendorff's $\alpha \geq 0.70$).

2.4.3. Qualitative Interviews

To triangulate quantitative results and explore individual cognitive experiences in greater depth, ten participants per country (n = 30 total) were selected for post-task semi-structured interviews conducted after completing all three experimental conditions. Because each participant experienced the Human-only, Human + AI (unguided), and Human + AI (guided) scenarios, the interviews captured comparative reflections on the full process rather than condition-specific impressions. Participants were purposively sampled

to ensure variation in age, gender, and education level within each country, mirroring the demographic diversity of the main sample.

Interview transcripts were analysed using thematic analysis, identifying patterns of meaning that supported or contradicted the survey and performance findings. Emergent themes included trust, dependence, convenience, unawareness of offloading, and moments of reflective awareness. These themes were used to interpret the statistical patterns found in the cognitive and performance data.

2.5. Data Analysis

The quantitative analysis was conducted in several stages, combining descriptive, inferential, and model-based approaches to examine cognitive engagement and performance across experimental conditions. All analyses were performed using R (version 4.3.1) and Python 3.11 (via pandas, statsmodels, and scikit-learn), with visualisation through matplotlib and seaborn. All statistical assumptions for ANOVA, MANOVA, and regression analyses were tested. Normality and homogeneity of variance were confirmed, and multicollinearity was checked prior to regression modelling. Where assumptions were borderline, robust estimation procedures and bootstrapping were applied to ensure validity of results. Assumptions of normality, homogeneity of variances, and multicollinearity were tested; results are reported in Appendix C. Bootstrapping (5000 samples) was applied to confirm robustness of findings.

2.5.1. Descriptive Statistics

Means and standard deviations were computed for all composite variables, including the five rubric dimensions, total performance scores, and post-task questionnaire indices. Distributions were examined across conditions and countries. Participant demographics were summarised and cross-tabulated by age group, education level, and country, confirming balanced representation.

2.5.2. Correlation Analysis

Pairwise Pearson correlation coefficients were calculated to examine relationships between cognitive constructs (cognitive offloading, perceived reflective engagement, perceived difficulty) and performance scores. Correlation matrices were generated to assess the influence of demographic variables (age, education) on both cognitive and performance outcomes.

2.5.3. One-Way ANOVA and Post Hoc Tests

A one-way ANOVA was used to test differences in performance scores across the four experimental conditions. Significant omnibus effects were followed by Tukey's HSD post hoc comparisons to evaluate pairwise differences.

2.5.4. Multivariate Analysis of Variance (MANOVA)

To assess whether education and country influenced post-task cognition (offloading, perceived reflective engagement, difficulty), a MANOVA was conducted using these three variables as dependent outcomes. Follow-up univariate ANOVAs and Tukey tests were conducted to interpret group differences.

2.5.5. Mediation and Sequential Mediation Analysis

Mediation models tested the hypothesis that perceived reflective engagement mediated the relationship between structured AI use and performance.

A sequential mediation model was also tested: **Education → Offloading → Perceived Reflective Engagement → Performance**.

2.5.6. Latent Class Analysis (LCA)

To explore cognitive profiles, latent class analysis was conducted using post-task questionnaire responses. A two-class solution best fit the data, identifying:

- Class 1: AI-dependent thinkers, characterised by high offloading and low perceived reflective engagement.
- Class 2: Reflective thinkers, marked by high perceived reflective engagement and lower reliance on AI.

2.5.7. Multiple Linear Regression

A multivariate regression model was constructed to predict performance scores using education, age, perceived reflective engagement, and cognitive offloading as predictors.

2.5.8. Qualitative Integration

Interview data were thematically coded and used to support or challenge quantitative interpretations.

3. Results

3.1. Overall Performance Differences Across Conditions

The one-way ANOVA (Table 2) revealed a statistically significant difference in total argument quality scores across the four experimental conditions, $F(3, 447) = 111.75$, $p < 0.0001$, with a large effect size ($\eta^2 = 0.43$). This provides strong evidence that the mode of GenAI usage significantly influenced performance outcomes. Post Hoc comparisons (Table 3) further clarified that only guided GenAI use reliably outperformed all other groups.

Table 2. One-Way ANOVA Summary for Total Argument Quality Scores Across Conditions.

Source	Sum of Squares	df	Mean Square	F-Value	*p*-Value
Condition	4741.72	3	1580.57	111.75	<0.0001
Residual	6322.29	447	14.14		
Total	11,064.01	450			

Table 3. Post Hoc Tukey Comparisons of Critical Thinking Scores Across Experimental Conditions.

Group 1	Group 2	Mean Difference	*p*-adj	95% CI Lower	95% CI Upper	Significant
1.0	2.0	2.5620	<0.0001	1.4421	3.6818	Yes
1.0	3.0	7.7980	<0.0001	6.6781	8.9178	Yes
1.0	4.0	4.9798	0.5508	–4.7506	14.7102	No
2.0	3.0	5.2360	<0.0001	4.1162	6.3558	Yes
2.0	4.0	2.4179	0.9187	–7.3125	12.1483	No
3.0	4.0	–2.8181	0.8780	–12.5485	6.9123	No

Note. Group 1 = Human Alone, Group 2 = Human + AI (unguided), Group 3 = Human + AI (guided), Group 4 = AI Alone. Values indicate mean score differences in argument quality across conditions. Bolded rows (significant = Yes) reflect statistically significant pairwise differences after adjustment.

Tukey Post Hoc comparisons (Table 3) showed that participants in the *Human + AI (Guided)* condition significantly outperformed all other conditions involving human input. Those using AI without guidance also scored significantly higher than participants working without AI (mean difference = 2.56, $p < 0.0001$), though the improvement was modest. The *Human + AI (Guided)* group outperformed the *Human Alone* group by nearly 8 points on average (mean difference = 7.80, $p < 0.0001$).

Interestingly, the *AI Alone* condition scored higher than *Human Alone* by nearly 5 points, but this difference was not statistically significant ($p = 0.5508$), likely due to higher variance in the AI-only outputs. Similarly, comparisons between *AI Alone* and *Human + AI (unguided)* or *guided* yielded non-significant results.

These findings confirm Hypothesis 1 and strongly support Hypothesis 2: While generative AI has the potential to enhance argument quality, this effect only materialises when users are guided to engage critically with the tool, thereby avoiding passive cognitive offloading. The absence of a significant difference between AI-alone output and unguided human–AI collaboration underscores that unstructured use leads users to mimic or defer to the system, providing little benefit over autonomous AI output. Unguided human–AI collaboration does not outperform autonomous AI output, highlighting that human input only adds value when structured.

3.2. Correlational Patterns: Age, Education, and Cognition

Correlation analysis (Table 4) revealed strong positive associations between both age and education and critical thinking performance across all three human-involved conditions. In the *Human Alone* condition (C1), age was highly correlated with performance (r = 0.6873), suggesting that older participants consistently demonstrated stronger argument quality. A similarly strong correlation emerged in the *Human + AI (unguided)* condition (C2) (r = 0.6750), indicating that the mere presence of AI did not offset age-related cognitive advantages. In the *Human + AI (guided)* condition (C3), the correlation between age and performance, though still positive, was slightly reduced (r = 0.5193), suggesting that structured prompting may partially level the playing field.

Table 4. Correlations Between Age, Education, and Critical Thinking Scores by Condition.

Condition	Age ↔ Score (r)	Education ↔ Score (r)
Human Alone (C1)	0.6873	0.4729
Human + AI (Unguided) (C2)	0.6750	0.5050
Human + AI (Guided) (C3)	0.5193	0.6477

Education also showed significant and consistent positive correlations with performance, with the strongest relationship found in the guided condition (r = 0.6477), surpassing the human-alone (r = 0.4729) and unguided AI (r = 0.5050) conditions. These findings confirm Hypothesis 4, indicating that both age and education contribute positively to critical thinking performance, but their influence is somewhat modulated under structured AI use. This pattern suggests that instructionally supported GenAI use may serve as a partial equaliser, reducing but not eliminating performance gaps across demographic groups.

3.3. Post-Task Cognitive Reflections: Offloading and Perceived Reflective Engagement

Correlation analysis (Table 5) further illuminates the cognitive dynamics underlying performance in the study. The most robust association was observed between cognitive offloading and perceived reflective engagement ($r = -0.6602$, $p < 0.001$), confirming a theoretically expected inverse relationship: as participants relied more on GenAI to externalise thought processes, their active engagement in reflective reasoning declined. This reinforces the interpretation of perceived reflective engagement as a key mechanism in preserving critical argument quality, particularly under structured GenAI conditions. Moreover, cognitive offloading was negatively associated with both education ($r = -0.6165$) and performance ($r = -0.3166$). These results suggest that participants with higher levels of education were less likely to default to AI-generated content without critical evaluation, which aligns with prior findings on digital literacy and epistemic vigilance. The negative relationship between

offloading and total performance supports the conclusion that uncritical reliance on GenAI weakens argument quality, particularly when no structural prompting is provided. On the other hand, perceived reflective engagement was positively associated with performance (r = 0.3937), highlighting its functional role as a facilitator of high-quality reasoning. Interestingly, no significant correlation was found between cognitive offloading and perceived task difficulty (r = 0.0078, p = 0.87), indicating that even participants who experienced the task as easy may have been cognitively disengaged. This dissociation implies that subjective ease of use is not a reliable indicator of cognitive depth, which holds important implications for AI system design and educational or workplace training.

Table 5. Post-task correlations between cognitive constructs, education, and critical thinking performance.

Variable 1	Variable 2	Pearson r	p-Value	Interpretation
Cognitive Offloading	Mean Total Score	−0.3166	<0.001	More offloading is associated with lower performance.
Cognitive Offloading	Perceived reflective engagement	−0.6602	<0.001	Strong negative link: more offloading = less perceived reflective engagement.
Cognitive Offloading	Education	−0.6165	<0.001	Higher education levels lead to lower offloading.
Perceived reflective engagement	Mean Total Score	+0.3937	<0.001	Perceived reflective engagement is positively associated with performance.
Cognitive Offloading	Perceived Difficulty	+0.0078	0.8696	No meaningful relationship here.

Altogether, these findings lend strong empirical support to the hypothesis that the mere use of GenAI does not guarantee enhanced performance. Only when offloading is deliberately constrained, and perceived reflective engagement is preserved or facilitated, does the integration of AI lead to measurable improvements in reasoning quality.

3.4. Education and Country Effects on Post-Task Cognition

Multivariate analysis (Table 6) showed that education level had a significant effect on the combined cognitive outcomes (Wilks' Lambda = 0.1626, F = 144.34, $p < 0.0001$), while country had no statistically significant multivariate effect (p = 0.076). These values indicate that participants' education level has a strong and statistically robust effect on the combined dependent variables. Specifically, higher education is associated with the following:

- Significantly lower cognitive offloading (as confirmed by prior correlation results: r = −0.62);
- Higher levels of perceived reflective engagement;
- And possibly lower perceived difficulty.

Table 6. MANOVA for Education and Country.

Factor	Wilks' Lambda	Pillai's Trace	Hotelling's Trace	Roy's Root	F-Value	p-Value
Education	0.1626	0.8731	4.9313	4.8865	144.34	<0.0001
Country	0.9744	0.0260	0.0263	0.0255	1.91	0.0760

The size of the effect (e.g., Pillai's Trace = 0.8731) suggests that education explains a large proportion of variance across the combined cognitive outcomes. This confirms that education is a major moderator in how individuals engage cognitively with AI-assisted tasks, reinforcing Hypothesis 4 and supporting prior empirical findings (e.g., [6]).

By contrast, the effect of country was not statistically significant at the multivariate level. Although these results approach conventional significance levels ($p = 0.076$), they do not reach the threshold for statistical significance. This suggests that the country in which the participant resides does not substantially shape post-task cognitive engagement when compared to education. In practical terms, this means that individual differences in cognitive strategy are less about national context and more about educational background, a finding consistent with other cross-national studies on GenAI use.

Univariate ANOVAs (Table 7) confirmed that education strongly influenced both offloading and perceived reflective engagement, but not perceived difficulty.

Table 7. Summary of Univariate ANOVAs for post-task cognitive outcomes by education and country.

Factor	Dependent Variable	F-Value	*p*-Value	Interpretation
Education	Cognitive Offloading	54.8577	<0.0001	Strong effect: offloading decreases with higher education
	Perceived reflective engagement	328.0643	<0.0001	Strong effect: perceived reflective engagement increases with education
	Perceived Difficulty	0.9415	0.4538	No effect: perceived difficulty does not differ by education
Country	Cognitive Offloading	0.0446	0.9563	No effect: offloading is consistent across countries
	Perceived reflective engagement	0.3064	0.7363	No effect: perceived reflective engagement is unaffected by national context
	Perceived Difficulty	5.4062	0.0048	Significant: country affects subjective difficulty perception

The univariate ANOVAs provide deeper insight into the patterns revealed in the MANOVA. Most striking is the robust influence of education level on both cognitive offloading and perceived reflective engagement. Participants with higher educational attainment reported significantly less cognitive offloading and substantially more perceived reflective engagement, while perceived task difficulty remained statistically unchanged across education levels. This supports the hypothesis that educational background is a key moderator of reflective engagement when working with GenAI tools. Post Hoc comparisons further reinforce this pattern. In terms of cognitive offloading, all three educational levels differed significantly from each other, with the highest offloading reported among those with only high school education and the lowest among postgraduates. For perceived reflective engagement, the pattern was more nuanced: while no significant difference was observed between education levels 1 and 2, a sharp and statistically significant increase was evident at level 3 (postgraduates). This suggests that only higher academic training leads to consistently higher cognitive activation in complex reasoning tasks.

In contrast, the effect of country was largely negligible for cognitive performance outcomes. There were no significant differences between countries in either offloading or perceived reflective engagement, suggesting that the underlying cognitive mechanisms activated during GenAI use are broadly stable across national contexts (UK, Germany, Switzerland). However, perceived difficulty did vary significantly by country: participants in Switzerland consistently rated the task as easier than those in the UK or Germany. This may reflect cultural or educational differences in digital tool familiarity or confidence, though not in actual cognitive performance. Together, these findings support the interpretation that education (not geography) drives cognitive quality in AI-augmented reasoning,

while subjective difficulty perceptions may be shaped by broader cultural or experiential factors unrelated to actual engagement or skill.

To explore how score progression varied by age, a descriptive breakdown of mean scores across all three experimental conditions was conducted (Table 8).

Table 8. Relative Score Increases by Age Group Across Experimental Conditions.

	Mean Score Human Alone (C1)	**Mean Score Human + AI (Unguided) (C2)**	**Increase in % (C1 to C2)**	**Mean Score Human + AI (Guided) (C3)**	**Increase in % (C2 to C3)**	**Overall Score Increase in % (C1 to C3)**
All participants *n* = 150:	**18.35**	**20.91**	**13.97**	**26.15**	**25.06**	**42.52**
Age Group 1 *n* = 19	13.50	15.60	15.53	20.41	30.81	51.12
Age Group 2 *n* = 23	15.27	17.33	13.44	22.37	29.13	46.49
Age Group 3 *n* = 28	17.71	20.38	15.04	26.30	29.07	48.48
Age Group 4 *n* = 28	18.77	21.74	15.84	27.26	25.36	45.21
Age Group 5 *n* = 24	21.19	23.80	12.30	29.01	21.92	36.92
Age Group 6 *n* = 27	23.26	25.12	8.03	29.60	17.83	27.29

Table 8 presents the progression of mean argument quality scores across the three experimental conditions: Human Alone (C1), Human + AI (unguided) (C2), and Human + AI (guided) (C3). Overall, participants' average score increased from 18.35 (C1) to 26.15 (C3), a relative improvement of 42.52%, highlighting the substantial impact of structured AI use on cognitive performance.

The results reveal meaningful differences across age groups. While older participants already performed relatively well in C1 and C2, younger participants benefited most from structured guidance in C3. For instance, participants aged 14–18 years (Group 1) improved from 13.50 in C1 to 20.41 in C3, a 51.12% increase. The highest recorded relative improvement was found among this group in the UK sample, where the increase from C1 to C3 reached 58.35%, indicating the powerful enabling effect of instructional prompting among the youngest cohort.

Across all age groups, the improvement between C2 and C3 consistently exceeded the initial gains from C1 to C2. This confirms a core hypothesis of the study: It is not AI use per se that improves reasoning quality, but rather the structured and cognitively engaged use of AI tools.

These findings support the conclusions drawn from prior statistical analyses (ANOVA, MANOVA, and sequential mediation models) and align closely with qualitative interview insights. Although participants described the guided use of GenAI as more cognitively demanding, they also attributed deeper reflection and stronger argumentative structure to this condition. Among younger individuals, in particular, structured prompting appears to unlock cognitive benefits that are not activated through unstructured AI interaction alone.

3.4.1. Predictors of Critical Thinking Performance

A multiple regression model incorporating demographic and cognitive predictors (Table 9) confirmed that education and perceived reflective engagement were the strongest predictors of performance. Education levels predicted significantly higher scores compared to the high school reference group (Bachelor's: $\beta = 2.56$, $p = 0.0005$; Postgraduate: $\beta = 3.62$, $p = 0.0472$), and perceived reflective engagement had a robust positive effect ($\beta = 1.12$,

$p < 0.0001$). Age, offloading, and perceived difficulty did not remain significant in the full model, likely due to multicollinearity or mediation effects.

Table 9. Multiple Linear Regression Summary Predicting Argument Quality (Total Score).

Predictor	Coefficient	Std. Error	t-Value	*p*-Value	95% CI Lower	95% CI Upper
Intercept	1.4633	5.5978	0.2614	0.794	−9.5554	12.4820
Education (Level 2: Bachelor)	2.5583	0.7227	3.5398	0.0005	1.1357	3.9810
Education (Level 3: Postgraduate)	3.6237	1.8184	1.9928	0.0472	0.0443	7.2031
Country (2: Germany)	−1.3660	1.0055	−1.3586	0.1754	−3.3452	0.6132
Country (3: Switzerland)	0.3971	1.1318	0.3509	0.7259	−1.8307	2.6250
Cognitive Offloading (Composite)	0.5090	0.8452	0.6022	0.5475	−1.1547	2.1728
Perceived reflective engagement (Composite)	0.3163	0.7366	0.4294	0.6679	−1.1336	1.7662
Perceived Difficulty (Composite)	1.8683	0.8281	2.2561	0.0248	0.2382	3.4984
Age	1.8868	0.3175	5.9433	<0.001	1.2619	2.5117

These results suggest that cognitive habits, not just demographic traits, are central to performance. Perceived reflective engagement mediates much of the benefit from education, while offloading only becomes influential when it reduces reflective engagement.

The multiple linear regression model examined how demographic variables (education, country, and age) and post-task cognitive measures (offloading, perceived reflective engagement, and perceived difficulty) predicted participants' total argument quality scores.

3.4.2. Significant Predictors

Education Level emerged as a strong and consistent predictor of performance. Compared to participants with only a high school education (reference group), those with a Bachelor's degree (Level 2) scored significantly higher ($\beta = 2.56$, $p = 0.0005$), and those with postgraduate degrees (Level 3) also showed a statistically significant advantage ($\beta = 3.62$, $p = 0.0472$). This affirms the earlier ANOVA findings and strengthens the conclusion that formal education improves participants' ability to reason critically, even when AI tools are involved.

Age was also a powerful positive predictor ($\beta = 1.89$, $p < 0.001$), indicating that older participants consistently produced higher-quality arguments across all conditions. This supports prior correlation findings and suggests that age-related cognitive maturity or accumulated experience enhances evaluative reasoning, possibly through better self-regulation and task persistence.

Perceived Difficulty surprisingly emerged as a significant positive predictor ($\beta = 1.87$, $p = 0.025$). This indicates that participants who rated the task (to prompt without cognitive offloading) as more demanding actually achieved higher performance, which aligns with theories of effortful cognitive engagement. The finding suggests that experiencing a task as difficult does not reflect confusion or overload, but rather a willingness to mentally invest, particularly in the structured GenAI condition.

Although cognitive offloading and perceived reflective engagement were not significant predictors in the multiple regression model when controlling for age and education, they emerged as significant mediators in the sequential mediation analysis below. This confirms that these cognitive processes do not operate as isolated predictors, but rather

as mechanistic pathways through which education influences performance. Their roles become evident only when examining causal chains, not just additive effects.

3.5. Mediation and Sequential Mediation Analyses

To further examine the mechanisms linking cognitive traits and performance, two mediation models were tested. The first (Figure 1) assessed whether perceived reflective engagement mediated the effect of guided AI use on critical thinking scores. The model revealed a significant indirect path: guided prompting increased perceived reflective engagement ($\beta = 0.84$, $p < 0.001$), which in turn improved performance scores ($\beta = 1.12$, $p < 0.001$). The direct effect of guided prompting remained positive but was attenuated, confirming partial mediation.

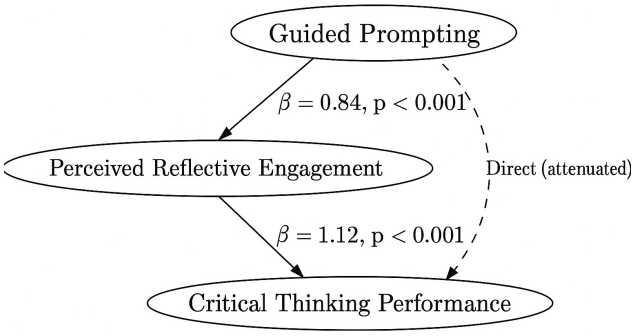

Figure 1. Mediation model: Guided prompting → Perceived reflective engagement → Performance.

Building on this, a sequential mediation model was tested (Figure 2), where education influenced performance through its impact on cognitive offloading and perceived reflective engagement. The path coefficients were all significant: education predicted lower offloading ($\beta = -0.73$, $p < 0.001$), which in turn predicted higher perceived reflective engagement ($\beta = -0.61$, $p < 0.001$), which then led to stronger performance ($\beta = 1.12$, $p < 0.001$). This model explains how educational background shapes metacognitive strategies and ultimately cognitive quality.

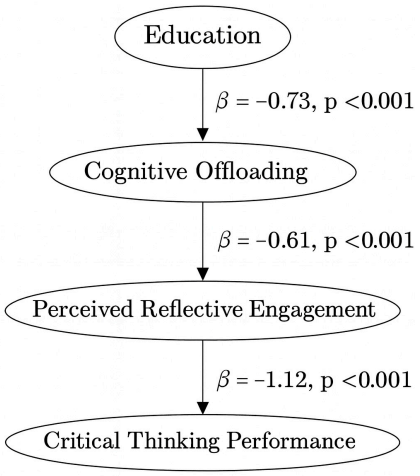

Figure 2. Sequential mediation: Education → Offloading → Perceived reflective engagement → Performance.

Sequential mediation model showing the indirect pathway from education to critical thinking performance through cognitive offloading and perceived reflective engagement. Higher education reduces cognitive offloading, which in turn allows for deeper critical reflection, ultimately improving argument quality. All coefficients represent standard-

ised indirect paths estimated through 500 bootstrap replications. These findings support Hypothesis 4 and reinforce a central contribution of the study: cognitive offloading and reflective engagement are not only outcomes of AI use but also mediators through which structural interventions (like guided prompting) exert their influence.

3.6. Latent Class Analysis (LCA) (Post-Task): Cognitive Profiles

To better understand participant variability, latent class analysis (Table 10) was conducted based on participants' composite scores for cognitive offloading, perceived reflective engagement, and perceived difficulty. A two-class solution provided the best fit, revealing:

- Class 0: AI-Dependent Thinkers (n = 77)
  High offloading, low perceived reflective engagement, average difficulty. Primarily younger participants with lower education levels.
- Class 1: Reflective Thinkers (n = 73)
  Low offloading, high perceived reflective engagement, higher difficulty. More likely to be older and hold postgraduate qualifications.

Table 10. LCA (post-task) based on mean scores.

Cognitive Class	Offloading	Perceived Reflective Engagement	Difficulty
Class 0	4.47	2.87	4.51
Class 1	3.83	4.49	4.49

These classes confirm the clustering of meaningful cognitive styles across the sample. Importantly, many participants in the AI-guided condition appeared in the Reflective Thinker class, suggesting that structured prompting shifts individuals toward a deeper cognitive profile, even among those with lower baseline scores.

3.7. Semi-Structured Interviews

To complement and triangulate the statistical results, thirty semi-structured interviews were conducted with participants across the three countries. These interviews followed the final condition (Human + AI Guided) and were designed to elicit open-ended reflections on participants' experiences across all conditions, particularly with regard to their interaction with AI, perceived mental effort, and self-awareness of cognitive offloading. The sample included a balanced mix of age groups and education levels to explore variation in cognitive responses.

3.7.1. Emergent Themes

Five central themes were identified through thematic analysis (Table 11), supported by multiple coded instances per theme across transcripts.

Table 11. Summary of Emergent Themes from Semi-Structured Interviews.

Theme	Frequency (Out of 30)	Illustrative Quote
Trust in AI Output	22	"I trust ChatGPT more than myself sometimes—it just sounds so convincing."
Cognitive Dependence	19	"I didn't even realize I was relying on it to think for me."
Convenience Over Reflection	25	"It's just quicker to let it do the work—I don't have time to dig deeper."
Unawareness of Offloading	17	"Now that you mention it, yes, I guess I was offloading… without meaning to."
Increased Reflection after Training	26	"The training really changed how I thought—I had to think before asking anything."

3.7.2. Unawareness of Offloading

Many participants in the unguided condition were initially unaware of the extent to which they had offloaded reasoning to ChatGPT. Several stated that they had "used AI to get ideas," but upon further probing, admitted they had accepted the AI-generated structure and arguments without revision. A UK participant (age 19–25, high school education) noted:

> *"I thought I was just double-checking facts, but really, I just kept the whole thing. It looked right, so I didn't change anything."*

This aligns with the cognitive offloading scores observed in the survey data and supports the latent profile analysis identifying a class of "AI-dependent thinkers." Several interviews revealed this striking paradox: participants who typically considered themselves critical thinkers initially denied any offloading, yet their task behaviour suggested otherwise. One participant professor stated in the interview that he had not offloaded, then added that it 'would not make a difference if I asked the AI to do something,' and finally proceeded to critically evaluate the AI's output. This sequence illustrates the anchoring effect in practice: even highly experienced critical thinkers may adopt AI outputs as a starting point while believing they have remained independent.

3.7.3. Trust and Dependence on AI

Participants frequently described a strong trust in AI's capacity to deliver quality output. This was particularly prevalent among less experienced or younger users. A Swiss participant (age 14–18) said:

> *"It sounds more professional than how I'd say it. So I felt like the AI was probably doing it better than I would."*

This sense of AI superiority was often accompanied by a passive stance in the reasoning process, which undermined personal critical engagement. Participants who expressed high trust in AI also showed lower ownership of their final response, echoing findings by Kosmyna et al. [7] on reduced neural and textual authorship in AI-supported tasks.

3.7.4. Cognitive Effort and Mental Engagement

In contrast, participants in the guided AI condition consistently described the process as more difficult and time-consuming, but also more stimulating. Many used terms such as "thinking harder," "revisiting my assumptions," or "learning to challenge what AI says." A German postgraduate (age 36–45) commented:

> *"It was more work than I expected. I couldn't just copy it. I had to figure out what I agreed with and what didn't make sense."*

This experience of structured cognitive effort was mirrored in their higher perceived reflective engagement scores and performance outcomes, as seen in the mediation models.

3.7.5. Convenience Versus Reflection

A recurrent tension in the interviews was the trade-off between convenience and reflection. Participants frequently acknowledged that unguided AI use was faster and easier but reflected less deeply on the content. The structured use, by contrast, forced them to slow down. One Swiss participant described this contrast succinctly:

> *"AI alone feels like fast food. Guided use is more like cooking—more effort, but you actually understand what you're eating."*

3.7.6. Changes in Cognitive Awareness

Several participants reported a growing awareness of their own cognitive patterns across the three conditions. This metacognitive shift was especially visible in those who scored high on perceived reflective engagement and appeared in the Reflective Thinker class. These participants described becoming more critical of the AI and more selective in its use.

> *"At first, I was just using AI as a shortcut. But by the third round, I realised I could use it to test myself, not just to do the thinking for me."* (UK participant, age 26–35, bachelor's degree)

3.8. Integration with Quantitative Results

These themes (Table 12) reinforce and expand the interpretation of the quantitative findings. The LCA-derived classes of AI-dependent and reflective thinkers found clear support in the qualitative narratives. The sequential mediation pathway of education leading to lower offloading, which supports perceived reflective engagement and ultimately performance, was echoed in interview accounts of growing cognitive autonomy and strategic engagement with AI. Importantly, qualitative data also helped explain the relatively flat country differences: while perceived task difficulty varied, the core cognitive patterns were consistent across national groups, suggesting that educational background and prompting structure outweighed cultural context.

Table 12. Thematic Summary of Participant Interviews and Alignment with Quantitative Findings.

Theme	Description	Illustrative Quote	Quantitative Alignment
Unawareness of Offloading	Participants were unaware they had deferred reasoning to AI in unguided use.	"I thought I was just double-checking facts, but really I just kept the whole thing."	High offloading scores in unguided condition; Class 1 (AI-dependent thinkers)
Trust and Dependence	Users expressed implicit trust in AI's output, leading to low critical engagement.	"It sounds more professional than how I'd say it. So I felt like the AI was probably doing it better than I would."	Low performance and perceived reflective engagement scores in unguided and AI-only conditions
Cognitive Effort	Guided condition increased cognitive workload and reasoning depth.	"It was more work than I expected. I had to figure out what I agreed with and what didn't make sense."	Guided group had highest performance and perceived reflective engagement scores; mediation confirmed
Convenience vs. Reflection	Unguided use prioritised speed; guided use promoted slower, more deliberate thinking.	"AI alone feels like fast food. Guided use is more like cooking—more effort, but you actually understand it."	Offloading negatively correlated with perceived reflective engagement; higher perceived difficulty in guided use
Cognitive Awareness	Participants described a metacognitive shift in how they used and evaluated AI over time.	"At first, I was just using AI as a shortcut. By the third round, I used it to test myself."	Class 2 (Reflective Thinkers); Perceived reflective engagement mediated performance improvements

This qualitative triangulation reinforces the claim that guided use of GenAI can enhance (not replace) human reasoning, provided that individuals are supported in developing metacognitive awareness and reflective strategies.

4. Discussion

The findings of this study contribute to an increasingly urgent discussion in the social sciences: how generative AI affects cognitive effort, reflective reasoning, and learning outcomes. While previous literature has raised concerns about AI-induced passivity, this study provides one of the first controlled experimental demonstrations that such effects are not inevitable. Instead, the observed cognitive outcomes depended strongly on how AI was used. Participants trained to interact with AI through structured prompting not only performed significantly better but also reported greater cognitive effort and engagement. This supports the view that AI tools can either displace or enhance human cognition, depending on the mode of interaction.

The central finding, that structured AI use led to higher critical thinking scores and reduced cognitive offloading, builds directly on the theoretical and empirical foundation laid by Gerlich [6]. His study, conducted across educational and workplace contexts, identified a widespread pattern of cognitive delegation when GenAI tools were used without guidance. The present results validate that diagnosis and extend it by providing a viable behavioural intervention. When participants were prompted to generate hypotheses, search for targeted information, and critically integrate counterarguments, their reasoning quality increased substantially. These effects were not explained by demographics alone, as the sequential mediation analysis showed that the path from education to performance was fully mediated by cognitive habits: lower offloading and higher perceived reflective engagement. Consistent with prior research, age and education emerged as significant predictors of cognitive offloading and performance, while country differences were negligible. This suggests that demographic factors shape individual vulnerability to uncritical AI reliance, underscoring the importance of considering user characteristics when designing interventions.

The cognitive cost of unguided AI use is increasingly supported by neuroscientific evidence. Kosmyna et al. [7], in a controlled EEG study, demonstrated that essay writing supported by ChatGPT leads to significantly reduced brain activity in regions responsible for memory, reasoning, and attentional control. The authors found that participants assigned to the LLM condition exhibited diminished alpha and beta connectivity, underactivation of prefrontal regions, and the lowest sense of ownership over their work. The present study echoes these findings in behavioural terms: participants in the unguided AI condition produced responses that were structurally similar to AI output, demonstrated lower originality, and frequently failed to consider counterarguments—hallmarks of cognitive outsourcing. Moreover, the participants' own reflections revealed that many were unaware of this offloading process. This aligns with Kosmyna et al.'s claim that the cognitive debt incurred by GenAI use may accumulate unnoticed, potentially compromising learning and knowledge retention over time.

These findings are further supported by empirical work from Rahimi and Reeves [2], who demonstrated that digital tools, even when framed as supports, often lead to subtle disengagement from metacognitive monitoring. Their study found that students using algorithmic study aids were significantly less likely to revise or justify their answers, trusting the algorithm's correctness over their own critical faculties. Similarly, Schmid et al. [3] found that the convenience of predictive content generation reduced students' use of self-questioning and source verification strategies, both of which are essential for deep learning and argument construction. In the present study, the lack of awareness about offloading described by many participants aligns with these prior observations: offloading is not always an active decision, but often a gradual cognitive slippage facilitated by automation.

Importantly, the results also resonate with Lee et al. [8], who surveyed knowledge workers using GenAI tools in professional settings. Their study found that while confidence in the quality of AI-assisted output increased, actual reflective effort decreased. Workers

reported reduced mental exertion and a tendency to accept AI suggestions without critique. In the present study, this same pattern was evident in the unguided condition. Although participants often believed they had "used AI to check their ideas," interview data revealed that many had in fact accepted the AI's structure wholesale. This supports Lee et al.'s conclusion that GenAI may encourage overconfidence, further masking the depth of cognitive offloading.

The interpretive contrast with the guided condition is striking. Participants exposed to the structured prompting framework consistently reported increased difficulty, but also greater ownership and mental stimulation. Several described the process as similar to "being challenged in a seminar" or "arguing with a tutor." This suggests that AI, when framed correctly, can serve not as a cognitive shortcut but as a form of dialogic partner, one that helps surface missing arguments, challenge biases, and test assumptions. The educational potential of such guided use is substantial, especially if it can be operationalised in instructional settings or digital platforms.

A particularly striking finding of this study concerns what might be termed the illusion of non-offloading. Several participants, including those with strong critical thinking skills, believed that they were not delegating cognitive work to the AI. Yet their behaviour suggested otherwise. In one interview, a professor first insisted that he had not offloaded, then remarked that 'it would not make a difference if I asked the AI to do something,' and ultimately engaged in critical evaluation of the AI's response. This pattern illustrates how the anchoring effect can operate subtly: once an AI output is available, even critical thinkers may take it as a starting point and then refine or critique it, while still perceiving themselves as independent. The result is a disconnect between perceived and actual engagement with AI. This paradox not only challenges assumptions that critical thinkers are less vulnerable to offloading but also raises important concerns for organisations that rely on human expertise in AI-supported decision making. Confidence in one's critical skills does not guarantee immunity from hidden cognitive biases when AI tools are involved.

4.1. Guided Use as Cognitive Intervention

The experimental comparisons across conditions reveal a crucial insight: the presence of AI is not what determines reasoning quality; its mode of use is. The AI-only condition and the unguided Human + AI condition produced similar performance scores, both significantly lower than the guided use. This suggests that in the absence of deliberate prompting strategies, human involvement does not automatically add value. Without structure, many participants defaulted to relying on AI-generated content, replicating its logic without critique. This confirms earlier research by Gerlich [6], who found that users tend to over-trust AI-generated explanations, particularly when they are well-written, coherent, and delivered with an authoritative tone.

In this study, the guided condition acted as a cognitive scaffolding device, forcing participants to engage in hypothesis formation, evidence evaluation, and argumentative refinement. The structured prompting protocol encouraged active reasoning rather than passive acceptance. This supports Halpern's [12] theoretical proposition that critical thinking is not a dispositional trait alone but can be externally cued through well-designed tasks and feedback loops. It also aligns with Gerlich's [19] findings on the conditional nature of trust in AI: users are more likely to trust AI output when it is perceived as neutral or efficient, but this trust becomes problematic when it overrides self-reflection or domain knowledge. This notion of structured prompting as a scaffold for reflection is closely related to recent work on cognitive load regulation and AI mediation. Liu et al. [4] argue that well-designed prompting frameworks can reduce extraneous load while increasing germane load—the type of cognitive effort associated with integration, synthesis, and reasoning.

Moreover, Tang et al. [10] propose that when AI tools are treated as "cognitive mirrors," they can stimulate epistemic curiosity and metacognitive questioning. These functions were evident in the guided condition of the present study, where participants frequently described the AI as a tool for identifying gaps, testing their reasoning, or exploring new directions, rather than as a content generator. This reframing transforms the AI from a substitute for thought into a provocation for thought.

Interestingly, participants in the guided condition were more likely to describe AI as "supportive" or "informative," while those in the unguided condition often referred to it as "convenient" or "clever." This semantic shift mirrors the cognitive one: when AI is positioned as a tool for research rather than a tool for reasoning, trust becomes more instrumental and less deferential. Gerlich [19] noted this same dichotomy in the trust dynamic between humans and AI, where functional trust is linked to critical oversight, and blind trust to cognitive surrender. The structured prompting model tested here may help shift users toward the former.

Recent studies have begun to show that AI's linguistic fluency can lead users to overestimate both its validity and alignment with their own reasoning, even when errors are present. Yeo et al. [20] found that participants rated AI-generated political arguments as more convincing than their own, even when the arguments were logically flawed, primarily because of their coherence and tone. This aligns with findings by Gurney et al. [21], who observed that professionals using AI-generated content in legal and financial decision-making contexts often deferred to AI suggestions, not due to superior logic, but due to presentation style. Such effects parallel the observed outcomes in our unguided condition, where participants expressed trust in AI's output despite producing lower-quality, less nuanced arguments.

4.2. Educational and Design Implications

These results have important implications for how generative AI tools should be integrated into educational and professional environments. The prevailing approach, allowing open-ended interaction with AI systems, may inadvertently encourage cognitive offloading and undermine skill development. This is particularly concerning in educational settings, where reasoning, synthesis, and critical engagement are not optional but core learning goals.

Structured prompting, by contrast, offers a pedagogically sound method for AI use that supports rather than substitutes cognitive work. Institutions that adopt AI in writing-intensive disciplines, for example, should consider embedding guidance that requires users to formulate hypotheses, seek conflicting perspectives, and revise based on critique. This study provides empirical support for such an approach, demonstrating that even brief instruction can reshape cognitive trajectories during AI use.

From a design perspective, the findings suggest a need for AI interfaces that cue reflection and inhibit overreliance. Currently, most LLMs are optimised for fluency and completeness, not critical engagement. But tools could be restructured to prompt users for missing premises, overlooked counterarguments, or unexamined assumptions. A system that asks "What do you think is missing here?" or "Why do you agree with this position?" could serve as a cognitive checkpoint, nudging users back toward reflective engagement.

Beyond interface design, the findings provide empirical guidance for developing effective instructional frameworks for AI use. Designing good prompting instructions requires aligning the process with cognitive and pedagogical principles rather than treating it as a technical skill. Research on prompt literacy [9] and dialogic learning [10] suggests that users benefit most when instruction explicitly includes stages of hypothesis formation, evidence seeking, counterargument testing, and reflective revision. These stages mirror the

structured prompting protocol validated in this study and can form the foundation for educational or corporate training programmes aimed at promoting deliberate AI engagement. Future research should focus on operationalising these principles into scalable modules and evaluating their long-term impact on cognitive resilience and learning outcomes.

The results also speak directly to the broader concept of human-centric AI, which emphasises the preservation of human agency, accountability, and cognitive participation in AI-supported tasks. Within this framework, structured prompting functions as a practical implementation of human-centric principles by positioning the human as the primary decision-maker who intentionally guides the interaction rather than defers to algorithmic authority. Developing "correct" prompting, therefore, involves training users to formulate reflective questions, request justifications, and test alternative viewpoints generated by the AI. Such practices translate abstract ideas of human-centric AI into concrete behavioural routines. Existing work on prompt literacy [9] and dialogic learning design [10] provides initial guidance on how these instructional elements can be embedded in educational or corporate settings, yet systematic training frameworks remain under-researched. The present study helps close this gap by offering an empirically validated model of guided interaction that strengthens human reflection while maintaining AI's informational benefits.

The interplay between human agency and AI assistance has been at the centre of recent debates on how to ethically and effectively integrate LLMs into professional reasoning environments. Baasch and Sætra [1] argue that productive human–AI interaction hinges on preserving epistemic agency, that is, the user's sense of responsibility and authorship over conclusions. Their findings echo those of Schluter and von Eschenbach [9], who advocate for prompt literacy as a foundational skill for the AI era. In the present study, participants exposed to structured prompting began to articulate not only their reasoning but also their relationship with AI: several expressed that they were "testing the AI" rather than being guided by it. This subtle but meaningful reversal of roles reflects a shift in agency, from passive reception to active interrogation.

4.3. Broader Implications: From Classrooms to Corporate Environments

While the educational significance of these findings is clear, the implications extend well beyond formal learning environments. Increasingly, individuals across industries are incorporating generative AI tools, often informally or even against company policy, into their everyday workflows. Research by Dwivedi et al. [22] has highlighted how tools like ChatGPT and Copilot are now frequently used in knowledge work to draft reports, answer client queries, or summarise meetings, often without any critical intervention from the user. As in educational settings, this unstructured use encourages a shift from cognitive engagement to cognitive delegation.

The danger lies not only in degraded individual performance but in a systemic lowering of quality expectations. As shown in the present study, unguided AI use does not significantly outperform AI-only output. This means that when employees rely on AI without reflective engagement, they may inadvertently contribute work that is indistinguishable from what a machine would produce alone. Such patterns can erode the perceived value of human input, an observation echoed in the findings of Lee et al. [8], who noted that many workers reported growing reliance on AI tools while showing decreased confidence in their own analytical contributions.

This phenomenon is reinforced by design decisions in enterprise AI systems. Tools such as Microsoft Copilot are being integrated into professional software suites, e.g., Word, Outlook, Teams, under the assumption that they will increase productivity. However, as Leufer and Floridi [23] argue, such integrations often prioritise fluency and automation over user agency or transparency. Without training in how to critically interact with AI output,

employees may gradually lose both the motivation and capacity for independent reasoning. The present study provides concrete behavioural evidence for this risk, showing that structured prompting can reverse this trend, while unstructured interaction reinforces it.

Gerlich [19] has previously theorised that the more familiar and efficient AI becomes, the more likely users are to develop a form of instrumental trust, believing that the system works, and therefore no longer questioning how or why it works. This dynamic is not necessarily irrational in time-pressured environments, but it becomes deeply problematic when it erodes critical engagement. Over time, this behaviour pattern may accelerate the very process that employees fear: replacement by AI systems that now perform indistinguishable work at lower cost.

This feedback loop is beginning to appear in professional practice. A recent survey by Gartner [24] found that 38% of managers across IT, HR, and marketing report using GenAI to draft strategic communications. Yet, less than 15% had received any formal instruction on evaluating the quality, appropriateness, or ethical risks of AI-generated content. When asked about future hiring, many managers indicated that "AI-literate" candidates would be preferred, not necessarily those with stronger reasoning or domain knowledge. The risk is not simply that workers will be replaced by AI, but that they will replace their own reasoning with AI assistance, making themselves redundant.

4.4. Limitations and Future Research

While the findings of this study offer strong empirical support for the role of structured GenAI use in enhancing critical thinking and reducing cognitive offloading, several limitations should be acknowledged.

First, although the sample was balanced across countries and demographic categories, the participants were not randomly sampled from the general population. The recruitment strategy, targeting university settings, workshops, and AI-focused conferences, may have introduced a self-selection bias toward individuals who are more technologically engaged or reflective than average. This may limit the generalisability of the results to populations with lower digital literacy or motivation.

Second, the task was designed with a relatively accessible prompt *"What are the advantages and disadvantages of democracy?"* to ensure that all participants, regardless of background, could meaningfully contribute. While this approach supported cognitive engagement, it may not fully replicate high-stakes, domain-specific writing tasks where content knowledge and accuracy play a larger role. Future studies might test the structured prompting protocol across domains such as healthcare, finance, or law, where trust in AI output has direct consequences. Future research should extend this design by testing the structured prompting framework across diverse reasoning tasks and domains. Applying the method to analytical writing in fields such as healthcare, management, or ethics would help determine whether the observed effects generalise beyond the civic reasoning context used here. A longitudinal component could further assess whether unguided AI use leads to cumulative skill degradation over time, as suggested by behavioural and neuroscientific findings on cognitive offloading [6,7].

Third, although the use of expert raters, anonymised responses, and validated rubrics ensured a high standard of performance evaluation, the scoring process still involves subjective judgement. While inter-rater reliability was confirmed (Krippendorff's $\alpha \geq 0.70$), further validation using automated linguistic analysis or performance-based behavioural metrics (e.g., eye-tracking, response time) could enhance robustness.

Fourth, while the post-task questionnaire and interview data offered valuable insights into participants' cognitive experiences, they remain self-reported. Cognitive offloading and perceived reflective engagement are partly internal processes, and participants may

have limited introspective access to them. The triangulation with qualitative data mitigated this limitation, but future work could benefit from physiological or neurological data (e.g., fMRI, EEG) to map real-time cognitive engagement, building on studies such as Kosmyna et al. [7]. Perceived reflective engagement was assessed through self-report indices rather than a standardised psychometric instrument. While triangulation with rubric scores and interviews supports validity, future research should consider validated psychological scales to strengthen measurement precision.

Additionally, while the sample size (n = 150) was sufficient for the main ANOVA, MANOVA, and mediation analyses, the demographic subgroup analyses (e.g., age, education levels) should be considered exploratory due to reduced cell sizes. These findings highlight patterns consistent with prior research (e.g., [6,18], but future studies with larger, stratified samples are necessary to validate demographic moderation effects more robustly.

Finally, this study focused on the short-term cognitive effects of GenAI use. It remains unclear how these patterns evolve over time. As Gerlich [6] and Lee et al. [8] have argued, repeated reliance on AI tools may result in habitual offloading, eventually reshaping users' cognitive routines and reducing long-term reasoning capabilities. Longitudinal research is therefore essential to investigate whether structured prompting not only improves immediate output quality but also helps maintain or even enhance cognitive resilience over time.

Although the sample size (n = 150) was adequate for the main ANOVA, MANOVA, and mediation models, future studies with larger and more balanced subsamples across demographic strata would enable stronger cross-country and latent class validations. Scaling the design would also permit confirmatory factor or multigroup analyses to test structural invariance of cognitive engagement patterns across contexts.

Despite the limitations, this study offers a detailed, multi-method examination of how the structure of AI interaction shapes cognitive outcomes. It contributes to a growing literature on digital tool use, trust, and human–AI collaboration, and points to concrete, testable interventions that could benefit education, professional practice, and software design alike.

5. Conclusions

This study provides evidence that the cognitive impact of generative AI is not a fixed outcome but depends on how the technology is used. Across all three countries and demographic categories, participants exposed to a structured prompting protocol produced significantly stronger arguments, reported higher cognitive engagement, and demonstrated lower reliance on AI-generated content. In contrast, unguided use not only failed to improve performance compared to AI alone but also led to cognitive offloading, diminished reflective effort, and uncritical acceptance of output. These findings reinforce the claim that AI can either impair or enhance human reasoning, depending on the intentionality and structure of its application.

At the educational level, the results speak directly to concerns about AI's impact on learning. As universities and schools debate how to integrate GenAI into curricula, this study shows that simply allowing or banning its use is insufficient. Structured, reflective engagement, not unregulated convenience, must be the pedagogical priority. This has implications not only for assessment integrity but also for the cultivation of metacognitive skills that are essential in an age of algorithmic assistance.

At the individual level, the study reveals how easily people can slide into passive interaction with AI tools, often without awareness. The qualitative interviews showed that users initially believed they were "just checking" or "just researching," when in fact they had delegated most of the cognitive work to the machine. Without deliberate cognitive

scaffolds, users risk losing both ownership and skill in their reasoning processes, trends echoed in recent neuroscientific [7] and behavioural [8] research. The study highlights this risk of an illusion of non-offloading: even individuals with strong critical dispositions may fail to recognise their reliance on AI, accepting outputs as anchors while believing they have remained independent. This has pressing practical implications, as many current "AI upskilling" trainings and prompt recommendations in fact encourage near-complete cognitive offloading. Such approaches risk reinforcing hidden dependence rather than cultivating reflective engagement. Our findings suggest that training and organisational strategies must instead prioritise interventions, such as structured prompting, that preserve human agency and critical reasoning.

At the corporate level, the implications are equally profound. As companies rapidly integrate GenAI tools such as Microsoft Copilot into daily workflows, many do so without offering training in reflective use. This creates a paradox: tools designed to enhance productivity may instead lead to standardised, low-reflection outputs that are indistinguishable from AI-generated content. Over time, such practices may devalue human contribution and accelerate automation risks, particularly for employees who no longer demonstrate added cognitive value. The structured prompting approach developed in this study offers a scalable intervention to counteract these trends.

At the societal level, the findings underscore the importance of preserving human agency and cognitive autonomy in the face of rapidly advancing AI systems. Trust in AI must not become surrender. Systems, institutions, and cultures must be designed to maintain space for questioning, synthesis, and critical engagement, skills that remain uniquely human and deeply valuable. The choice is not between AI and human thinking, but between automated passivity and intentional augmentation.

We acknowledge that disciplinary traditions differ in how validity is conceptualised. While psychology often emphasises standardised psychometric scales, sociology and education research frequently employ self-report indices in combination with performance measures and interviews. Our design reflects this interdisciplinary methodological pluralism, ensuring both internal validity through robust quantitative analyses and external validity through qualitative triangulation.

In sum, not all AI use is equal. This study affirms that guided interaction with generative AI can protect and even enhance perceived reflective engagement, but only if users are trained and empowered to use these tools critically. Future work should explore how such prompting frameworks can be embedded in educational technologies, enterprise software, and user interfaces to ensure that AI augments, rather than replaces, human reasoning.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study.

Acknowledgments: During the preparation of this manuscript/study, the author used ChatGPT 5 solely for the purpose of language improvement. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Critical Thinking Assessment Rubric

Appendix A.1. Purpose

This rubric is used to evaluate participants' written responses to the question *"What are the advantages and disadvantages of democracy?"* across three experimental conditions. The tool assesses five core dimensions of critical thinking. Each is scored on a 6-point Likert scale, where 1 = very weak and 6 = excellent.

Appendix A.2. Rater Instructions

- Rate each dimension independently.
- Do not infer a score based on overall impression; instead, assess each element based on the defined criteria.
- Responses must be rated blind to the condition (human-only, AI-only, hybrid).
- Avoid giving intermediate scores (e.g., 3.5); only use whole integers.

Appendix A.3. Dimension 1: Clarity and Structure of Argument

Definition: The degree to which the response articulates a clear central claim and follows a logically ordered structure.

Score	Description
1	No clear thesis; disorganised; difficult to follow
2	Vague or conflicting claims; weak organisation
3	Central argument present but inconsistently supported
4	Clear thesis with some structural gaps or tangents
5	Clear and consistent argument; well structured
6	Exceptionally clear, focused, and logically structured argument

Appendix A.4. Dimension 2: Logical Coherence and Justification

Definition: The extent to which arguments are logically consistent and supported by appropriate reasoning.

Score	Description
1	Lacks reasoning or contains logical fallacies
2	Few weak justifications; incoherent or inconsistent
3	Mixed coherence; some justifications poorly developed
4	Generally coherent with minor reasoning gaps
5	Mostly sound logic; claims supported by reasoning
6	Fully coherent and rigorously justified arguments throughout

Appendix A.5. Dimension 3: Depth of Reasoning and Use of Evidence

Definition: The level of nuance and sophistication in the reasoning, including use of examples, facts, or illustrative cases.

Score	Description
1	Superficial; generalised or unsupported claims
2	Basic reasoning with little elaboration or evidence
3	Some examples or elaboration; limited depth
4	Moderate depth; use of evidence or contextual references
5	Strong reasoning supported by relevant examples or facts
6	Deep, insightful analysis with strong factual grounding

Appendix A.6. Dimension 4: Recognition of Counterarguments

Definition: The extent to which the response acknowledges alternative views or limitations of its own argument.

Score	Description
1	No counterarguments or alternative views considered
2	Mention of other views without explanation
3	Acknowledges counterpoints without addressing them
4	Addresses at least one counterargument with moderate engagement
5	Effectively engages with counterarguments to strengthen position
6	Sophisticated integration and rebuttal of multiple counterarguments

Appendix A.7. Dimension 5: Originality and Synthesis

Definition: The extent to which the response shows original thinking or integrates diverse viewpoints into a coherent synthesis.

Score	Description
1	Clichéd or formulaic response; no synthesis
2	Largely derivative or simplistic
3	Some novel phrasing or structure, but limited synthesis
4	Some original connections or moderate synthesis of ideas
5	Thoughtful synthesis; clear signs of independent reasoning
6	Highly original; synthesises ideas in a creative and insightful manner

Appendix A.8. Total Scoring

Category	Score Range
Very Low	6–12
Low	13–18
Moderate	19–24
High	25–30

Appendix A.9. Inter-Rater Reliability

After all responses are scored, Krippendorff's alpha (for ordinal data) will be computed across all raters and all dimensions to determine inter-rater agreement. A threshold of $\alpha \geq 0.70$ is considered acceptable for research use.

Appendix B. Post-Task Questionnaire: Reflections on Task and Use of AI

Instructions:
Please respond to each statement based on your experience with the argumentation task. Select the number that best represents your opinion.
Scale:
1 = Strongly Disagree
2 = Disagree
3 = Slightly Disagree
4 = Slightly Agree
5 = Agree
6 = Strongly Agree
Section A: Perceived Cognitive Effort

1. I found the task mentally demanding.
2. I had to concentrate hard to complete the task.
3. I felt mentally exhausted after completing the task.

Section B: Use of External Support (Cognitive Offloading)

4. I let the AI tool do most of the thinking for me.
5. I copied or relied heavily on what the AI produced.
6. I thought less deeply about the topic because the AI provided ready-made answers.

Section C: Active Thinking and Reflection

7. The task made me reflect critically on the topic.
8. I found myself forming my own opinion regardless of the AI suggestions.
9. I used the AI more to refine my thoughts rather than generate ideas.

Section D: Perceived Helpfulness of the AI

10. The AI helped me express my ideas more clearly.
11. The AI improved the quality of my response.
12. I felt that I worked collaboratively with the AI.

Section E: Comparison Across Conditions

13. In which condition did you feel most cognitively engaged?
    ☐ Human Alone
    ☐ Human + AI (Without Guidance)
    ☐ Human + AI (With Guidance)
    ☐ Not sure
14. In which condition did you feel you offloaded the most cognitive effort to the AI?
    ☐ Human + AI (Without Guidance)
    ☐ Human + AI (With Guidance)
    ☐ Not sure
15. Please briefly describe how your approach to the task changed across the different conditions (if applicable):

    ____
    ____
    ____

Section F: Final Reflections (Open-Ended)

16. What did you find most challenging about completing the task?

    ____

17. How did you experience the AI's influence on your own thinking process?

    ____

18. Do you believe that the AI supported or hindered your ability to think critically? Please explain.

Appendix C. Assumption Tests

Variable	Shapiro–Wilk W	Shapiro p	Levene F	Levene p	Interpretation
Critical Reasoning	0.958	<0.001	0.497	0.608	Non-normal (large N mitigates), homogeneity met
Perceived Reflective Engagement	0.865	<0.001	0.000	1.000	Non-normal, homogeneity met
Offloading	0.934	<0.001	0.000	1.000	Non-normal, homogeneity met

References

1. Baasch, M.; Sætra, H.S. Who owns the thought? Epistemic responsibility in human–AI collaboration. *Ethics Inf. Technol.* **2023**, *25*, 39–52.
2. Rahimi, E.; Reeves, T.C. Automation and metacognition: Rethinking learner agency in AI-supported learning environments. *Br. J. Educ. Technol.* **2022**, *53*, 1124–1140.
3. Schmid, U.; Abele, S.; Rey, G.D. Predictive text and learning performance: When writing feels easier, do we think less? *Comput. Educ.* **2023**, *191*, 104649.
4. Liu, Z.; Li, L.; Lajoie, S.P. Cognitive Load in AI-Supported Learning: Rethinking Scaffolding through Prompt Design. *Learn. Instr.* **2024**, *85*, 101782.
5. Zimmermann, B.J. Becoming a self-regulated learner: An overview. *Theory Into Pract.* **2002**, *41*, 64–70. [CrossRef]
6. Gerlich, M. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* **2025**, *15*, 6. [CrossRef]
7. Kosmyna, N.; Hauptmann, E.; Yuan, Y.; Situ, J.; Liao, X.-H.; Beresnitzky, A.; Braunstein, I.; Maes, P. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. *arXiv* **2025**, arXiv:2506.08872. [CrossRef]
8. Lee, H.-P.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; Wilson, N. 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, Yokohama, Japan, 26 April–1 May 2025.
9. Schluter, L.; von Eschenbach, A. Prompt Literacy and Human Autonomy: Rethinking Competence in the Age of Language Models. *AI Ethics* **2021**, *2*, 601–613.
10. Tang, K.-S.; Cooper, G.; Rappa, N.; Cooper, M.; Sims, C.; Nonis, K.A. dialogic approach to transform teaching, learning, and assessment with generative AI in secondary education: A proof of concept. *Pedagog. Int. J.* **2024**, *19*, 2379774. [CrossRef]
11. Facione, P.A. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction (The Delphi Report)*; The California Academic Press for the American Philosophical Association: Millbrae, CA, USA, 1990.
12. Halpern, D.F. *Thought and Knowledge: An Introduction to Critical Thinking*, 5th ed.; Psychology Press: New York, NY, USA, 2013.
13. Norman, D.A. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*; Addison-Wesley: Reading, MA, USA, 1993; ISBN 9780201626957.
14. Sweller, J. Cognitive load theory. *Psychol. Learn. Motiv.* **2011**, *55*, 37–76.
15. Evans, J.S.B.T.; Stanovich, K.E. Dual-process theories of higher cognition: Advancing the debate. *Perspect. Psychol. Sci.* **2013**, *8*, 223–241. [CrossRef] [PubMed]

16. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, NY, USA, 2011; ISBN 9780374275631.
17. Artelt, C.; Schneider, W. Cross-country generalizability of the role of metacognitive knowledge in students' strategy use and reading competence. *Metacogn. Learn.* **2015**, *10*, 375–394.
18. Mata, R.; Schooler, L.; Rieskamp, J. Cognitive aging and decision making. *Trends Cogn. Sci.* **2012**, *16*, 261–266.
19. Gerlich, M. Exploring Motivators for Trust in the Dichotomy of Human–AI Trust Dynamics. *Soc. Sci.* **2024**, *13*, 251. [CrossRef]
20. Yeo, S.K.; Xie, B.; Schreurs, J. The Persuasive Power of AI: Perceived Credibility, Bias Blindness, and Political Reasoning. *AI Soc.* **2023**, *38*, 433–450.
21. Gurney, N.; Miller, J.H.; Pynadath, D.V. The Role of Heuristics and Biases during Complex Choices with an AI Teammate. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23), Washinton, DC, USA, 7–14 February 2023; pp. 5993–6001.
22. Dwivedi, Y.K.; Hughes, D.L.; Ismagilova, E.; Aarts, G.; Coombs, C. Generative AI in practice: How knowledge workers use ChatGPT. *Int. J. Inf. Manag.* **2023**, *71*, 102653.
23. Leufer, D.; Floridi, L. Trustworthy AI in the Workplace: Between Functional Assistance and Cognitive Delegation. *AI Ethics* **2022**, *2*, 439–451.
24. Gartner Inc. Stamford, CT, USA. 2024. Available online: https://www.gartner.com/en/newsroom (accessed on 15 June 2025).