

Intellectus: An AI Model to Enhance Student Cognitive Performance, Increase Knowledge Acquisition, Boost Academic Integrity and Mitigate AI Over-Reliance in Education

Afrin Fathima Razvi
Department of Computer Engineering and Informatics
Middlesex University Dubai
Dubai, United Arab Emirates
0009-0003-4240-3174

Engie Bashir
Department of Computer Engineering and Informatics
Middlesex University Dubai
Dubai, United Arab Emirates
0000-0002-2135-6108

Abstract— This study presents Intellectus, a fine-tuned large language model (LLM) designed to enhance student cognitive performance, boost knowledge acquisition and mitigate AI over-reliance in education. In contrast to traditional AI tools that provide direct answers, Intellectus engages learners through Socratic questioning, adaptive scaffolding, and structured prompting. Developed using a no-code Custom GPT approach and grounded in a systematic literature review, Intellectus was evaluated through a triangulated methodology combining student self-assessments, expert ratings, and plagiarism analysis. A/B testing with twelve university students showed significant gains across ten higher-order cognitive skills when using Intellectus compared to ChatGPT, alongside a dramatic reduction in plagiarism rates. Expert evaluations further validated these findings, confirming Intellectus’ ability to promote deeper thinking and academic integrity, and T-testing proved statistical significance across domains.

Keywords— Educational AI, Cognitive Development, Higher-Order Thinking Skills, Socratic Questioning, Cognitive Offloading, Academic Integrity, Prompt Engineering, Triangulated Evaluation, Custom GPTs

I. INTRODUCTION

The proliferation of artificial intelligence (AI) language models in education has transformed how students engage with academic tasks and cognitive processes. Tools such as ChatGPT, Bard, and DeepSeek enable rapid information retrieval and text generation but primarily act as passive knowledge dispensers rather than interactive cognitive scaffolds [6]. This dynamic contributes to automation bias and cognitive offloading, where students uncritically accept outputs and bypass the cognitive effort essential for reasoning and retention [28][23][11], raising concerns about diminished higher-order thinking, academic disengagement, and compromised long-term learning [15][25][5].

Existing AI-based educational tools, including intelligent tutoring systems (ITS), similarly lack support for reflective problem-solving and decision-making [26][27][22], reinforcing passive consumption overactive cognitive development [9][3]. Their design can also encourage over-reliance and ethical misuse, as current detection systems for AI-generated content are unreliable and undermine academic integrity [21][7][4]. Without institutional safeguards, these tools risk amplifying digital inequities and misaligning with pedagogical goals [10][2], while unregulated use of generative AI (GenAI) may further degrade cognition and threaten educational outcomes, workforce preparedness, and epistemic equity [24][8][12].

This paper addresses these deficiencies through the design, development, and evaluation of Intellectus, a pedagogically aligned AI language model intended to serve as a guided inquiry facilitator rather than a direct answer provider. Developed iteratively using insights from a systematic literature review (SLR) and refined through empirical testing, Intellectus aims to enhance student cognition, support knowledge acquisition, and reinforce academic integrity, offering a structured response to the current limitations of AI-assisted learning.

II. LITERATURE REVIEW

A systematic literature review was conducted to examine how artificial intelligence (AI) integration in education influences student cognition, learning engagement, critical thinking, and academic integrity. The review also explored pedagogical frameworks and empirical methodologies that can optimize AI-assisted learning environments for guided inquiry and personalized instruction. The goal was to identify research gaps and synthesize insights relevant to the development of cognitively supportive and ethically sound AI tools.

A dual-source search strategy was employed, combining Middlesex University Library Search (MDX) with Consensus AI, an AI-powered academic aggregator that ranks sources based on evidence strength. Boolean search queries were used with terms such as “AI in education,” “LLMs and learning,” “instruction tuning,” and “academic integrity.” Search parameters included peer-reviewed publications (2005–2025), English language, and relevance to educational applications of AI. Studies were screened and selected according to the PRISMA 2020 guidelines (see Fig. 1), with an emphasis on empirical research, cognitive frameworks, and instructional relevance. The main findings informing the development of Intellectus are presented in Table I.

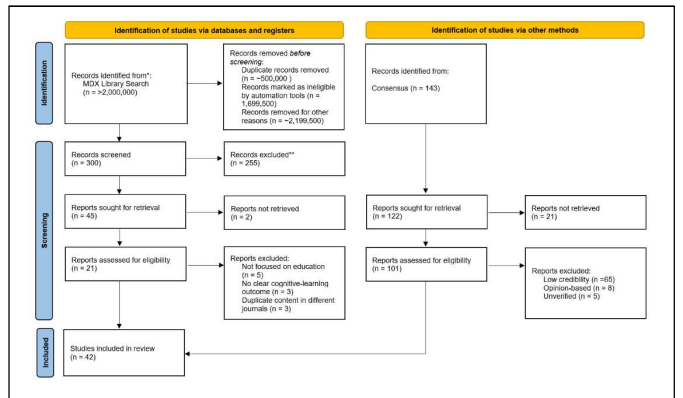


Fig. 1. PRISMA 2020 Flow Diagram for Systematic Literature Review

TABLE I. MAIN FINDINGS INFORMING INTELLECTUS DEVELOPMENT

Systematic Literature Review Findings		
Reference No.	Core Finding	Subhead
Deng & Yu (2023) [9]	Chatbots improve surface-level comprehension but harm critical thinking if used passively.	Avoid direct answers; use guided inquiry to maintain deep reasoning.
Ju (2023) [15]	Over-reliance on AI reduces comprehension and accuracy in academic tasks.	Integrate active recall and reflective questioning to engage cognition.
León-Domínguez (2024) [25]	Cognitive offloading weakens higher-order executive function.	Use progressive cognitive challenges to maintain reasoning effort.
Yang et al. (2005) [26]	Socratic questioning boosts critical thinking and reasoning.	Incorporate Socratic-style prompts to promote evaluative thinking.
Yatani et al. (2024) [27]	Structured, adaptive prompts stimulate higher-order thinking.	Implement adaptive, perspective-shifting prompts for critical engagement.
Puntambekar & Hubscher (2005) [22]	Scaffolding enhances learning by tailoring support to learner progress.	Provide graduated scaffolding to balance guidance and autonomy.
Mollick & Mollick (2023) [19]	AI should act as tutor/coach to support cognition, not replace it.	Frame AI as a thinking partner prompting justification and reflection.
Perkins (2023) [21]	AI-generated content challenges academic integrity detection.	Minimize reproduction; require original reasoning in outputs.
Matthews & Volpe (2023) [17]	AI indistinguishability necessitates attribution and integrity frameworks.	Reinforce justification-based answers to reduce hidden plagiarism.
MIT Sloan (2025) [18]	Custom GPTs enable pedagogical alignment without full retraining.	Use no-code system prompts to embed educational goals.

Accordingly, *Intellectus*, a novel AI prototype, was developed to enhance student cognition through guided inquiry rather than direct content delivery. The model incorporates prompt-based customization, dynamic scaffolding, and triangulated evaluation across ten cognitive skill domains. Built using no-code fine-tuning via Custom GPTs, it offers a replicable framework for ethically aligned, pedagogically grounded AI deployment. The following section details its design and implementation.

III. MODEL DEVELOPMENT

Three implementation strategies were considered for developing *Intellectus*: (1) fine-tuning open-source LLMs using proprietary data [13]; (2) programmatic control of commercial models via APIs [14]; and (3) no-code customization using OpenAI’s Custom GPT infrastructure with embedded instructional prompts and document conditioning [18]. Due to constraints in computational resources and development time, method (3) was selected for its scalability, low technical barrier, and alignment with pedagogical aims.

Intellectus was implemented via system-level prompting to enforce behavior consistent with guided inquiry, adaptive scaffolding, and cognitive engagement- prioritized across the literature [18], [25]. This approach addresses concerns

regarding cognitive offloading by requiring students to reason through problems without receiving direct answers.

A. System Prompting (V1)

Version 1 (V1) was constructed using a prompt framework co-generated with ChatGPT-4o through iterative instruction refinement, following methods validated for task-aligned prompt optimization [1]. The functional requirements were derived from the systematic review and include constraints against direct answer provision, the use of layered Socratic questioning, stepwise cognitive scaffolding, and promotion of higher-order thinking. Additional requirements mandate self-reflective prompting, guidance toward independent research, domain-general adaptability, and a progressive dialogue structure that reinforces student agency. The model was also to be equipped to challenge cognitive biases and prevent academic misuse through firm boundary enforcement and redirection.

Pilot interactions confirmed fidelity to design goals: the model withheld final answers, prompted recursive reasoning, and maintained cognitive framing aligned with target skill domains.

B. Informal Testing and Model Refinement (V2)

Usability testing was conducted with three university students from diverse academic backgrounds. Participants engaged in structured learning tasks with the prototype and submitted qualitative feedback via interviews and open-ended forms. While two participants reported that the model replicated effective tutoring dynamics, one cited excessive cognitive load due to unchunked, multi-paragraph responses.

In response, the prompt was revised to implement response chunking and incremental reasoning, resulting in Version 2 (V2). This version preserved the instructional logic of V1 but improved pacing and interaction clarity through enforcing response length limits. V2 was subsequently selected for formal A/B evaluation.

IV. METHODOLOGY

This study employs a pragmatic mixed-methods design integrating design-based research with quantitative and qualitative evaluation. The initial phase involved the systematic literature review of 42 peer-reviewed sources to extract pedagogical principles for AI-supported cognition and academic integrity, based on which, a prototype AI agent-*Intellectus*- was developed via parameter-constrained prompt scaffolding using OpenAI’s Custom GPT interface. The model was optimized to elicit higher-order cognitive responses via Socratic querying, recursive problem structuring, and reflective dialogue.

Prototype evaluation was conducted through a controlled A/B study involving 12 university students. Participants completed parallel academic tasks using *Intellectus* and ChatGPT. Cognitive engagement was measured through self-assessment instruments derived from validated cognitive skill frameworks, and textual outputs were screened for academic integrity violations.

Triangulation was achieved by incorporating independent expert evaluations from two educators who rated anonymized student outputs using a structured cognitive rubric. Quantitative data were analyzed using descriptive statistics and visual aggregation. Convergent findings across self-evaluations, expert assessments, and originality checks

informed model refinement and outcome validity. Statistical analyses including T-Testing was employed to measure statistical significance in difference between model performance.

V. RESULTS

This section presents a triangulated empirical evaluation of *Intellectus* through a within-subject A/B testing design, followed by plagiarism analysis and expert evaluation to assess its capacity to support higher-order cognitive development and knowledge acquisition compared to ChatGPT-4o.

1) Hypothesis Statement

The evaluation was guided by the following hypotheses:

H1: Participants will demonstrate higher cognitive performance across all 10 domains when using *Intellectus* compared to ChatGPT, as reflected in both self-evaluation and expert assessment.

H2: Participants will demonstrate greater perceived knowledge acquisition and intellectual engagement when interacting with *Intellectus*.

H3: Responses generated using *Intellectus* will yield lower plagiarism similarity scores, suggesting enhanced academic integrity.

A. Student Self-Evaluation

1) Evaluation Study Design

A total of 12 university students participated in a supervised A/B study. Each participant completed two task sets (Samples A and B) across six academic domains (Mathematics, Science, Humanities, ICT, Second Language, and English), one with ChatGPT-4o (Control) and one with *Intellectus* (Intervention). Task order was randomized to mitigate carryover effects [16]. Tasks were constructed to elicit explanation, analysis, and reasoning rather than factual recall, following cognitive complexity criteria derived from the Cambridge Lower Secondary framework.

Participants completed a post-task self-evaluation questionnaire after each condition. This instrument assessed ten predefined higher-order cognitive skill domains (HOTS): Critical Thinking, Analytical Thinking, Problem-Solving, Decision-Making, Creativity, Evaluation, Logical Reasoning, Comprehension, Scepticism and Cognitive Flexibility. Task parity, timing (45 minutes total), and browser configurations were standardized across sessions.

a) Tool Configuration and Protocol

Participants were randomly assigned a model-use sequence (ChatGPT or *Intellectus* first) and worked under supervision in timed, controlled conditions. Interactions with AI were restricted to pre-assigned browser instances. External search was permitted only when initiated by the AI, to reflect *Intellectus*' self-directed prompting architecture.

Quantitative data were collected via structured Google Docs, while qualitative feedback was captured through open-response items. AI outputs were stored for plagiarism analysis using Copyscape, which compared student-generated responses against the corresponding model output. While Copyscape provides a useful baseline for text similarity detection, it cannot detect semantic overlap or paraphrased content, and results should therefore be interpreted cautiously.

b) Participants

12 students were selected using convenience and snowball sampling. Inclusion criteria were based on self-reported domain familiarity (2–4 on a 5-point Likert scale). No identifying data were collected. Participation was voluntary, anonymized, and uncompensated.

c) Data Collection and Analysis

Post-task self-evaluations measured HOTS activation using a 1–5 Likert scale, aligned with cognitive constructs from the literature. Responses were aggregated, and differential scores between models were computed per domain. A separate measure captured perceived knowledge acquisition.

Plagiarism analysis quantified text similarity between student responses and AI output for each model. Average similarity scores were calculated across all tasks to evaluate originality.

Data were analyzed in Microsoft Excel using descriptive statistics and visualized via bar charts. Triangulated comparisons across self-evaluations, expert scoring, and plagiarism detection form the basis of subsequent analysis in the results section.

2) Results

a) Self-Evaluation Questionnaire

Analysis of self-reported HOTS showed consistently higher perceived activation when participants used *Intellectus* compared to ChatGPT (Fig. 2). Across domains, *Intellectus* yielded substantial mean gains, with the largest improvements observed in Creativity (+3.00), Evaluation (+2.92), and Critical Thinking (+2.75). Strong increases were also evident in Scepticism (+2.75), Analytical Thinking (+2.17), Logical Reasoning (+2.17), and Cognitive Flexibility (+2.08). More moderate but still meaningful differences were recorded for Problem-Solving (+1.83), Knowledge Acquisition (+1.83), Decision-Making (+1.58), and Comprehension (+1.50). These results indicate that *Intellectus* enhanced perceived cognitive engagement across all domains of HOTS, with no area showing an advantage for ChatGPT.

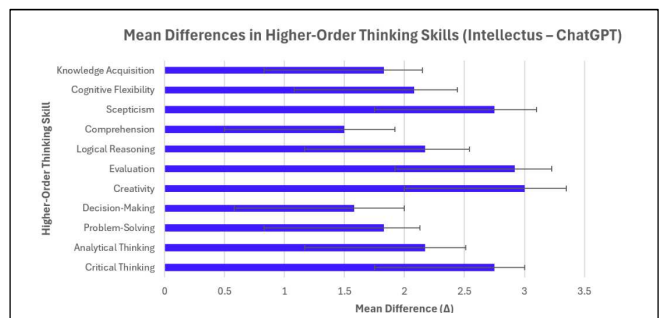


Fig. 2. Mean differences (*Intellectus* – ChatGPT) across higher-order thinking skills with 95% confidence intervals ($n = 12$). Error bars represent ± 1 SEM ($n = 12$).

Self-Evaluated Higher Order Thinking Skills and Knowledge Acquisition Scores						
HOTS	Mean A (Int-GPT)	SD	SEM	p-value	Cohen's dz	95% CI A (Lower, Upper)
Critical Thinking	2.75	0.87	0.25	p < 0.001	3.18	[2.20, 3.30]
Analytical Thinking	2.17	1.19	0.34	p < 0.001	1.82	[1.41, 2.92]
Problem-Solving	1.83	1.03	0.3	p < 0.001	1.78	[1.18, 2.49]
Decision-Making	1.58	1.44	0.42	p = 0.003	1.1	[0.67, 2.50]
Creativity	3	1.21	0.35	p < 0.001	2.49	[2.23, 3.77]
Evaluation	2.92	1.08	0.31	p < 0.001	2.69	[2.23, 3.61]
Logical Reasoning	2.17	1.27	0.37	p = 0.004	1.71	[1.36, 2.97]
Comprehension	1.5	1.45	0.42	p < 0.001	1.04	[0.58, 2.42]
Scepticism	2.75	1.22	0.35	p < 0.001	2.26	[1.98, 3.52]
Cognitive Flexibility	2.08	1.24	0.36	p < 0.001	1.68	[1.30, 2.87]
Knowledge Acquisition	1.83	1.11	0.32	p < 0.001	1.64	[1.13, 2.54]

Fig. 3. Self-evaluation results across higher-order thinking skills (HOTS), showing mean differences between *Intellectus* and ChatGPT, with accompanying SD, SEM, p-values, Cohen's dz, and 95% confidence intervals (n = 12).

Statistical analysis confirmed that these gains were highly significant (most $p < 0.01$), with large effect sizes (Cohen's $d_z = 1.04$ -3.18), low variability (SD = 0.87–1.45), stable precision (SEM = 0.25–0.42), and narrow 95% confidence intervals that did not cross zero (See Fig. 3).

B. Plagiarism Analysis

Plagiarism detection scores revealed a consistent reduction in similarity when participants used *Intellectus* compared to ChatGPT. Across the six academic tasks, mean differences (GPT – *Intellectus*) in plagiarism percentage were substantial: Q1 = 95.83%, Q2 = 83.33%, Q3 = 77.08%, Q4 = 85.83%, Q5 = 83.67%, and Q6 = 52.92%. Although reductions varied in magnitude, all questions showed significantly lower similarity for *Intellectus* outputs. These findings indicate that *Intellectus* generated content with markedly greater originality and reduced overlap with existing sources, thereby supporting stronger academic integrity (see Fig. 4).

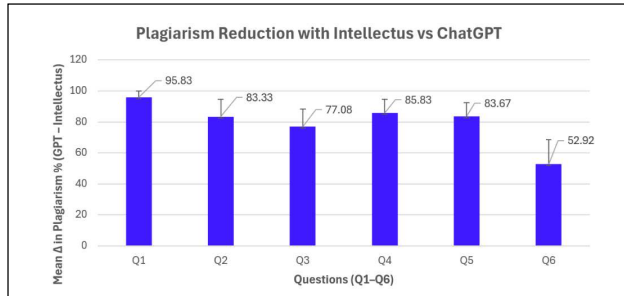


Fig. 4. Mean differences (ChatGPT – *Intellectus*) in plagiarism percentages across six academic tasks. Error bars represent ± 1 SEM (n = 12).

The statistical analysis (Fig 5.) confirmed that these reductions were statistically significant ($p < 0.01$ for all tasks; Q6 $p = 0.006$), with effect sizes ranging from $d_z = 0.98$ –6.64, moderate variability (SD = 14.43–54.21), consistent precision (SEM = 4.17–15.65), and 95% confidence intervals consistently above zero.

Plagiarism Reduction Across Six Academic Questions (Q1–Q6)						
Question	Mean A (GPT-Int)	SD	SEM	p-value	Cohen's dz	95% CI A (Lower, Upper)
Q1	95.83	14.43	4.17	p < 0.001	6.64	[86.66, 105.00]
Q2	83.33	38.92	11.24	p < 0.001	2.14	[58.60, 108.07]
Q3	77.08	39.11	11.29	p < 0.001	1.97	[52.24, 101.93]
Q4	85.83	30.88	8.92	p < 0.001	2.78	[66.21, 105.46]
Q5	83.67	30.31	8.75	p < 0.001	2.76	[64.41, 102.93]
Q6	52.92	54.21	15.65	p = 0.006	0.98	[18.47, 87.36]

Fig. 5. Mean differences (ChatGPT – *Intellectus*) in plagiarism percentages across six academic tasks. Error bars represent ± 1 SEM (n = 12).

C. Expert Assessment

To assess the cognitive depth and originality of student outputs, an expert assessment was conducted using blind scoring of anonymized responses generated under both AI conditions. This phase focused exclusively on external assessment to validate the findings from the student-based evaluation.

1) Evaluation Study Design

A between-conditions framework was used. A total of 24 anonymized responses (12 participants $\times$ 2 samples) were compiled into per-participant sets. Each submission was scored independently by two expert raters using a standardized rubric aligned with the ten HOTS. Raters had no access to participant identity, AI model used, or order of exposure.

Scoring was conducted on a 1–5 ordinal scale per HOTS domain, where 1 indicated absence of the skill and 5 indicated exceptional demonstration. Criteria were domain-specific and aligned with the definitions provided to raters. Experts also received a glossary of the HOTS and reviewed corresponding plagiarism data to inform their judgments.

2) Instrumentation

The rubric featured second-person evaluative statements for each cognitive domain and was supplemented with an open-response form for qualitative feedback on the perceived depth, originality, and instructional impact of each AI model. Benchmark answers, majorly derived from Cambridge Lower Secondary Checkpoint assessments, were provided to anchor expectations, although no fixed “correct” response was expected due to the open-ended nature of the tasks.

3) Participants

Two domain experts were recruited via convenience sampling. One held expertise in educational psychology, the other in AI and Information Technology. Both held postgraduate qualifications, over five years of academic teaching experience, and prior exposure to AI-assisted learning environments. Experts had no prior knowledge of *Intellectus* or its instructional framework. Participation was voluntary and uncompensated, and formal consent was obtained.

4) Procedure and Timeline

The assessment was conducted over a 7-day period. Experts received anonymized task batches and scored each response independently using the provided rubric. No discussion or collaboration was permitted. Discrepancies $\leq \pm 0.01$ were averaged; discrepancies exceeding this threshold were retained in raw form but noted for transparency.

5) Analysis

Scores from both raters were compiled and averaged across the ten HOTS domains for each model. The net cognitive performance gain of *Intellectus* relative to ChatGPT was calculated per domain. Expert feedback was analyzed qualitatively to supplement the numerical evaluation.

Directional consistency across both experts was confirmed; *Intellectus* outperformed ChatGPT across all HOTS domains in both scoring sets. Minor variability was observed in raw values but did not alter the overall ranking. This triangulated evaluation reinforces the hypothesis that

pedagogically structured AI models yield superior cognitive outcomes compared to general-purpose LLMs.

6) Results

Expert assessment consistently favored *Intellectus* over ChatGPT across all higher-order thinking domains. As shown in Figure 4, the grand mean across experts was 3.52 for *Intellectus* and 1.73 for ChatGPT, corresponding to an overall improvement of +1.79 points (See Fig. 6).

While the magnitude of the gains varied by domain, both experts rated *Intellectus* higher in every skill category. These improvements indicate that *Intellectus* responses were consistently judged to engage higher-order thinking skills more effectively than ChatGPT, aligning with the student self-evaluation and plagiarism findings to provide convergent evidence of its effectiveness.

Mean Expert Scores Across Higher-Order Thinking Skills (ChatGPT vs. Intellectus)				
Expert 1	HOTS	ChatGPT (M)	Intellectus (M)	Δ (Int - GPT)
	Critical-Thinking	1.58	3.42	1.83
	Analytical Thinking	1.50	3.33	1.83
	Problem-Solving	1.42	3.92	2.50
	Decision-Making	2.58	3.58	1.00
	Creativity	1.33	2.92	1.58
	Evaluation	1.25	3.00	1.75
	Logical Reasoning	1.33	3.50	2.17
	Comprehension	1.83	3.58	1.75
	Scepticism	1.33	1.92	0.58
	Cognitive Flexibility	1.67	2.42	0.75
	Expert 1 Mean (all HOTS)	1.58	3.16	1.58
	HOTS	ChatGPT (M)	Intellectus (M)	Δ (Int - GPT)
	Critical-Thinking	2.00	4.50	2.50
	Analytical Thinking	2.25	4.50	2.25
Expert 2	Problem-Solving	1.50	4.25	2.75
	Decision-Making	1.25	3.75	2.50
	Creativity	1.75	3.75	2.00
	Evaluation	1.50	3.50	2.00
	Logical Reasoning	1.50	3.25	1.75
	Comprehension	2.25	3.50	1.25
	Scepticism	2.50	4.25	1.75
	Cognitive Flexibility	2.25	3.50	1.25
	Expert 2 Mean (all HOTS)	1.88	3.88	2.00
	HOTS	ChatGPT (M)	Intellectus (M)	Δ (Int - GPT)
	Critical-Thinking	2.00	4.50	2.50
	Analytical Thinking	2.25	4.50	2.25
	Problem-Solving	1.50	4.25	2.75
	Decision-Making	1.25	3.75	2.50
	Creativity	1.75	3.75	2.00
	Evaluation	1.50	3.50	2.00
	Logical Reasoning	1.50	3.25	1.75
	Comprehension	2.25	3.50	1.25
	Scepticism	2.50	4.25	1.75
	Cognitive Flexibility	2.25	3.50	1.25
	Expert 2 Mean (all HOTS)	1.88	3.88	2.00
	Grand Mean across Experts	1.73	3.52	1.79

Fig. 6. Grand mean expert scores across higher-order thinking skills (HOTS) for ChatGPT and *Intellectus*. Δ represents the combined difference (*Intellectus* – ChatGPT) averaged across both experts.

VI. DISCUSSION

The evaluation of *Intellectus* across both student self-assessment and expert scoring yielded consistent evidence supporting its pedagogical efficacy. Across all ten HOTS domains, *Intellectus* outperformed ChatGPT.

Students reported consistently higher perceived activation of HOTS when using *Intellectus* compared to ChatGPT. Mean differences (Δ Int–GPT) ranged from +1.50 (Comprehension) to +3.00 (Creativity). Substantial gains were also observed in Evaluation (+2.92), Critical Thinking (+2.75), and Scepticism (+2.75). Statistical analysis confirmed that these improvements were significant across all domains (all $p < 0.01$; Decision-Making $p = 0.003$; Logical Reasoning $p = 0.004$), with large effect sizes (Cohen’s $d_z = 1.04$ – 3.18) and confidence intervals consistently above zero. These results confirm Hypothesis H1, demonstrating that *Intellectus* consistently enhanced perceived activation of higher-order thinking skills compared to ChatGPT.

Plagiarism detection showed significantly lower similarity in *Intellectus* outputs compared to ChatGPT. Mean reductions (GPT–Int) ranged from 52.92% (Q6) to 95.83% (Q1). All tasks demonstrated statistically significant decreases ($p < 0.01$; Q6 $p = 0.006$), with large effect sizes (Cohen’s $d_z = 0.98$ – 6.64). Standard deviations (14.43–54.21) and standard errors of the mean (4.17–15.65) indicated variability in task

difficulty, but 95% confidence intervals for all items remained well above zero, confirming significant differences. This finding supports Hypothesis H3, showing that *Intellectus* substantially reduced similarity scores and improved academic integrity.

Expert scorings corroborated the student findings. Across both raters, *Intellectus* achieved a grand mean of 3.52 versus 1.73 for ChatGPT, reflecting an aggregate gain of +1.79 points. Improvements were observed across all HOTS domains, with the largest combined advantages in Problem-Solving (+2.13), Critical Thinking (+2.17), and Analytical Thinking (+2.04). More modest gains were observed in Cognitive Flexibility (+1.00) and Scepticism (+1.17). These results highlight that expert judgments aligned with self-reports in identifying Creativity, Critical Thinking, and Evaluation as domains of notable improvement under *Intellectus*.

Hypothesis H2, relating to perceived knowledge acquisition, was supported by participant self-ratings and expert observations. Qualitative feedback and rubric scores indicated higher explanatory depth and justification in *Intellectus*-guided responses.

The triangulated evaluation approach, incorporating subjective, objective, and behavioral data, strengthens the internal validity of the findings. The convergence of self-reported learning, expert scoring, and plagiarism metrics is rarely observed in educational AI evaluation and underscores the robustness of the tool’s cognitive alignment.

The results directly reflect the pedagogical mechanisms emphasized in the systematic literature review, including Socratic prompting, adaptive scaffolding, and structured inquiry. The congruence between theoretical models and empirical outcomes supports the generalizability of findings and affirms the value of research-informed AI design in educational settings.

VII. LIMITATIONS AND FUTURE WORK

Despite its demonstrated impact, the study presents several limitations. The small sample size ($n=12$) and restricted evaluation timeline limit generalizability. Cognitive domains such as Creativity and Scepticism, while scored consistently, retain an interpretive component that introduces subjectivity. Additionally, implementation via OpenAI’s no-code Custom GPT platform limited real-time adaptivity and dynamic content generation.

Nevertheless, the study offers a replicable experimental framework, including prompt architecture, evaluation rubrics, and A/B task designs, suitable for scaling across broader populations. The structured methodology enables future validation with increased external validity.

Future research should explore longitudinal deployment to examine whether cognitive gains are sustained over time. Integration with Learning Management Systems (LMS), and fine-tuning for multilingual or culturally adaptive settings should be experimented. Natural language processing methods could be employed to analyze qualitative feedback at scale. Additional development directions include adaptive difficulty progression, conversational state management, gamification features, and real-time learning analytics. Importantly, bias mitigation, accessibility for diverse learners, and safeguards against misuse are critical areas for further development.

VIII. CONCLUSION

This study presents *Intellectus* as a pedagogically aligned AI model designed to mitigate automation bias and enhance higher-order cognitive skill development. By refusing to provide direct answers and instead employing Socratic questioning, adaptive scaffolding, and reflective prompting, *Intellectus* redefines AI from a knowledge provider to a facilitator of thought.

Empirical results across student self-evaluation and expert assessment confirmed statistically meaningful gains in all ten cognitive domains compared to ChatGPT, with the most pronounced improvements in Problem-Solving, Critical Thinking, and Analytical Thinking. Significant percentage point reduction in plagiarism scores further validated *Intellectus*' capacity to uphold academic integrity and reduce content replication.

The triangulated evaluation—incorporating subjective perception, expert validation, and behavioral originality metrics—provides a multidimensional and methodologically robust confirmation of efficacy.

These findings reinforce the value of research-informed AI design in education and demonstrate that instructional alignment—rather than computational complexity—can determine the cognitive impact of AI-assisted learning environments.

REFERENCES

- [1] Agarwal, E., Dani, V., Ganu, T. and Nambi, A. (2024). PromptWizard: Task-Aware Agentdriven Prompt Optimization Framework. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2405.18369.
- [2] Ahn, H.Y. (2024). AI-Powered E-Learning for Lifelong Learners: Impact on Performance and Knowledge Application. Sustainability, [online] 16(20), p.9066. doi:https://doi.org/10.3390/su16209066.
- [3] Akram, N.A., Nawaz, N.M., Sattar, N.S., None Ayesha Tasawar, None Shaista Ajmal and Ur, A. (2024). Developing Critical Thinking in Psychology: the Role of Cognitive Processes in Problem-Solving and Decision-Making. 2(3), pp.244–250 doi:https://doi.org/10.59075/rjs.v2i3.37.
- [4] Ardito, C.G. (2023). Contra Generative AI Detection in Higher Education Assessments. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2312.05241.
- [5] Bai, L., Liu, X. and Su, J. (2023). ChatGPT: the Cognitive Effects on Learning and Memory. ChatGPT: the Cognitive Effects on Learning and Memory, 1(3). doi:https://doi.org/10.1002/brx2.30.
- [6] Chan, C.K.Y. and Hu, W. (2023). Students' Voices on Generative AI: perceptions, benefits, and Challenges in Higher Education. International Journal of Educational Technology in Higher Education, 20(1), pp.1–18. doi:https://doi.org/10.1186/s41239-023-00411-8.
- [7] Cotton, D.R.E., Cotton, P.A. and Shipway, J.R. (2023). Chatting and cheating: Ensuring Academic Integrity in the Era of ChatGPT. Innovations in Education and Teaching International, [online] 61(2), pp.228–239. doi:https://doi.org/10.1080/14703297.2023.2190148.
- [8] Cramarenco, R.E., Burcă-Voicu, M.I. and Dabija, D.-C. (2023). The Impact of Artificial Intelligence (AI) on Employees' Skills and well-being in Global Labor markets: a Systematic Review. Oeconomia Copernicana, 14(3), pp.731–767. doi:https://doi.org/10.24136/oc.2023.022.
- [9] Deng, X. and Yu, Z. (2023). A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education. Sustainability, 15(4), p.2940. doi:https://doi.org/10.3390/su15042940.
- [10] Dolan, J.E. (2015). Splicing the Divide: a Review of Research on the Evolving Digital Divide among K–12 Students. Journal of Research on Technology in Education, 48(1), pp.16–37. doi:https://doi.org/10.1080/15391523.2015.1103147.
- [11] Dunn, T.L. and Risko, E.F. (2015). Toward a Metacognitive Account of Cognitive Offloading. Cognitive Science, 40(5), pp.1080–1127. doi:https://doi.org/10.1111/cogs.12273.
- [12] Furman, J. and Seamans, R. (2019). AI and the Economy. Innovation Policy and the Economy, [online] 19(1), pp.161–191. doi:https://doi.org/10.1086/69936.
- [13] Hu, Z., Lan, Y., Wang, L., Xu, W., Lim, E.-P., Roy Ka-Wei Lee, Bing, L. and Soujanya Poria (2023). LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2304.01933.
- [14] Jeong, C. (2024). Domain-specialized LLM: Financial fine-tuning and Utilization Method Using Mistral 7B. Jineung jeongbo yeon-gu, 30(1), pp.93–120. doi:https://doi.org/10.13088/jiis.2024.30.1.093.
- [15] Ju, Q. (2023). Experimental Evidence on Negative Impact of Generative AI on Scientific Learning Outcomes. [online] Social Science Research Network. doi:https://doi.org/10.2139/ssrn.4567696.
- [16] Laska, E., Meisner, M. and Kushner, H.B. (1983). Optimal Crossover Designs in the Presence of Carryover Effects. Biometrics, 39(4), p.1087. doi:https://doi.org/10.2307/2531343.
- [17] Matthews, J. and Volpe, C.R. (2023). Academics' Perceptions of ChatGPT-generated Written outputs: a Practical Application of Turing's Imitation Game. Australasian Journal of Educational Technology, [online] 39(5), pp.82–100. doi:https://doi.org/10.14742/ajet.8896.
- [18] MIT Sloan (2024). Custom GPTs at MIT Sloan: a Comprehensive Guide. [online] MIT Sloan Teaching & Learning Technologies. Available at: <https://mitsloanedtech.mit.edu/ai/tools/writing/custom-gpts-at-mit-sloan-a-comprehensiveguide/>.
- [19] Mollick, E. and Mollick, L. (2023a). Assigning AI: Seven Approaches for Students, with Prompts. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2306.10052.
- [20] Mollick, E.R. and Mollick, L. (2023b). Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts. SSRN Electronic Journal. doi:https://doi.org/10.2139/ssrn.4391243.
- [21] Perkins, M. (2023). Academic Integrity Considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and LearningPractice, 20(2). doi:https://doi.org/10.53761/1.20.02.07.
- [22] Puntambekar, S. and Hubscher, R. (2005). Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? Educational Psychologist, [online] 40(1), pp.1–12. doi:https://doi.org/10.1207/s15326985ep4001_1.
- [23] Risko, E.F. and Gilbert, S.J. (2016). Cognitive Offloading. Trends in Cognitive Sciences, [online] 20(9), pp.676–688. doi:https://doi.org/10.1016/j.tics.2016.07.002.
- [24] Sweller, J. (2019). Cognitive Load Theory and Educational Technology. Educational Technology Research and Development, [online] 68(1), pp.1–16. doi:https://doi.org/10.1007/s11423-019-09701-3.
- [25] Umberto León-Domínguez (2024). Potential Cognitive Risks of Generative transformerbased AI Chatbots on Higher Order Executive functions. Neuropsychology (journal), 38(4). doi:https://doi.org/10.1037/neu0000948.
- [26] Yang, Y.-T.C., Newby, T.J. and Bill, R.L. (2005). Using Socratic Questioning to Promote Critical Thinking Skills through Asynchronous Discussion Forums in Distance Learning Environments. American Journal of Distance Education, 19(3), pp.163–181. doi:https://doi.org/10.1207/s15389286ajde1903_4.
- [27] Yatani, K., Sramek, Z. and Yang, C.-L. (2024). AI as Extraherics: Fostering Higher-order Thinking Skills in Human-AI Interaction. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2409.09218.
- [28] Zhai, C., Wibowo, S. and Li, L.D. (2024). The Effects of over-reliance on AI Dialogue Systems on students' Cognitive abilities: a Systematic Review. Smart learning environments, [online] 11(1). doi:https://doi.org/10.1186/s40561-024-00316-7