FOSTERING CRITICAL THINKING TOWARD AI-GENERATED CONTENT IN MIDDLE SCHOOL STUDENTS: A QUASI-EXPERIMENTAL EVALUATION OF A DIGITAL CITIZENSHIP INTERVENTION

Sonia IGNAT, Assoc. Prof. PH.D.,
Aurel Vlaicu University of Arad
sonia.ignat@uav.ro

Abstract: *Generative artificial intelligence tools have become embedded in the everyday academic lives of adolescents, yet the capacity of middle school students to critically appraise AI-generated content remains largely unaddressed by current digital citizenship curricula. This study evaluated a five-session educational intervention designed to develop critical thinking toward AI-generated content among seventh-grade students, grounded in psychological inoculation theory and AI literacy competency models. A quasi-experimental pretest-posttest design with a non-equivalent control group was applied to 275 seventh-grade students (experimental group, n = 138; control group, n = 137) recruited from parallel classes of two middle school in Arad, Romania. Critical thinking toward AI-generated content was measured with an original 20-item Likert scale, and discrimination ability was assessed through a practical task requiring students to classify ten text and image stimuli as AI-generated or human-produced. Following the intervention, the experimental group showed a substantial increase in critical thinking scores, from M = 54.25 (SD = 8.12) to M = 68.90 (SD = 7.65), t(39) = -9.85, p < .001, d = 1.86, and in discrimination accuracy, from 51.25% to 68.50%, t(39) = -8.14, p < .001, d = 1.65. The control group showed no significant change on either measure. Critical thinking scores and discrimination accuracy correlated moderately to strongly at post-test, r = .58, p < .001. These findings support the feasibility of short, curriculum-embedded interventions grounded in inoculation theory for cultivating AI literacy at the middle school level, and they extend the construct of digital citizenship to include critical engagement with AI-generated content as a measurable, educable competency.*

Keywords: *digital citizenship; critical thinking; generative artificial intelligence; psychological inoculation; AI literacy; middle school students; quasi-experimental design.*

Introduction

Generative artificial intelligence systems such as ChatGPT, Gemini, and Copilot have become available to virtually anyone with an internet-connected device, including students in the early years of secondary education. Within seconds, these tools can produce text, imagery, or homework answers that appear entirely plausible while containing errors or distorted information. Klarin et al. (2024) documented the speed of this uptake among adolescents: in a sample with a mean age of 14 years, close to 15% already used generative AI tools for schoolwork, a proportion that exceeded 50% among 17-year-olds. The pace at which these tools have entered classrooms and homes has consistently outstripped the capacity of educational systems to prepare young people to use them critically and responsibly, and this gap constitutes the central problem addressed by the present study.

Middle school students, typically aged between 11 and 15 years, are a particularly vulnerable population in this respect. Formal operational thinking begins to emerge at this age, yet the executive functions on which critical evaluation depends, namely planning, inhibition, and cognitive flexibility, are still maturing, which predisposes younger adolescents to rely on AI tools for completing school tasks without evaluating the resulting content critically (Klarin et al., 2024). Complementary evidence indicates that susceptibility to scientific misinformation, although it declines across the secondary school years, remains a genuine concern among younger cohorts (Siani et al., 2024), and that adolescents tend to be overconfident in their own ability to detect false information while favouring verification strategies that demand minimal effort (Zozaya-Durazo et al., 2024).

The construct of digital citizenship, understood as the capacity to participate responsibly, ethically, and competently in digital environments, has evolved considerably since it first appeared in the literature. Two conceptual frameworks dominate the field: one that defines digital citizenship as a set of norms of appropriate technology use organised into nine dimensions (Ribble and Bailey, as cited in Chen et al., 2021), and another that conceptualises it as the capacity for

active civic participation through digital means (Mossberger et al., as cited in Chen et al., 2021). Choi (2016), drawing on a systematic analysis of 57 sources across disciplines, proposed a four-domain synthesis comprising ethics, media and information literacy, participation and engagement, and critical resistance, the latter referring explicitly to the capacity to recognise and contest existing structures of power in digital spaces rather than consume content passively. This last dimension is directly relevant to the present study, since it captures precisely the critical stance toward digital content, including AI-generated content, that the intervention sought to cultivate. Notably, Chen et al. (2021) found that of 114 peer-reviewed studies on digital citizenship, only five addressed adolescents between 11 and 19 years of age, and none addressed children under 11, a gap that underscores the need for empirical work with the middle school population specifically.

Critical thinking toward digital content is not an innate adolescent trait but a competency shaped by information-seeking habits and media literacy. In a large study of 1,505 adolescents aged 12 to 18, Ku et al. (2019) found that internal motivation to seek information, a sceptical attitude toward algorithmic curation of news, and the habit of checking sources were independent predictors of critical thinking about real-life news, and that older adolescents tended to be stronger critical thinkers than younger ones, with the weakest performance observed for the evaluation of evidence relative to comprehension or fact identification. These findings suggest that critical thinking toward digital content, including content generated by AI, is an educable competency that can be strengthened through targeted intervention.

Psychological inoculation theory offers a well-established framework for building resistance to misinformation. Lewandowsky and van der Linden (2021) argued that proactive countermeasures, and inoculation in particular, are generally more effective than reactive fact-checking, distinguishing general forewarning, which simply flags the existence of manipulation attempts, from specific inoculation, which exposes individuals to concrete manipulation techniques such as appeals to false experts, false dichotomies, or emotional appeals. A meta-analysis of 42 independent studies involving 42,530 participants confirmed that psychological inoculation significantly reduces the credibility assigned

to misinformation and improves informational discernment, with content-based interventions proving more effective for credibility judgements and pre-post designs more effective for reducing perceived credibility of false information (Lu et al., 2023). A separate meta-analysis spanning four decades of media literacy interventions similarly found consistently positive effects, more pronounced for cognitive outcomes than for behavioural change, with face-to-face delivery outperforming online delivery (Cho et al., 2025). Taken together, this literature provided the theoretical rationale for the design of the present intervention.

The specific challenges posed by generative AI content add a further layer of complexity. Klarin et al. (2024) reported that adolescents who experience greater difficulties with executive functioning perceive generative AI tools as more useful for completing schoolwork, suggesting that the students who are cognitively most vulnerable are also those most inclined to rely on these tools without critical scrutiny. Qualitative work contrasting parent and adolescent perspectives found that while parents are primarily concerned with data collection, misinformation, and exposure to inappropriate content, adolescents themselves worry more about dependency on virtual AI relationships, misuse of generative tools to circulate harmful content among peers, and the erosion of privacy through unauthorised use of personal data, concerns compounded by the near-total absence of parental control mechanisms on generative AI platforms (Yu, Sharma, et al., 2025). A complementary taxonomy of generative AI risks for young people situates these concerns within a broader landscape of emerging harms that educational interventions need to address (Yu, Liu, et al., 2025).

AI literacy has consequently emerged as a distinct educational competency domain. Long and Magerko (2020), building on an analysis of 150 interdisciplinary sources, proposed an operational definition comprising 17 competencies organised around five guiding questions: what AI is, what it can do, how it works, how it should be used, and how people perceive it. Among these, AI recognition, namely the ability to distinguish artefacts that rely on AI from those that do not, critical data evaluation, and AI ethics are the competencies most directly relevant to the present intervention. A thematic review of 50 studies on secondary-level AI education published between 2016

and 2022 found that project-based and problem-based learning was the most common pedagogical approach, identified in 72% of studies, followed by collaborative learning (56%) and experiential learning (52%), and that younger students tended to focus on experiential engagement with AI while older students explored more advanced technical components (Ng et al., 2023). Crucially, the same review identified a shortage of methodologically rigorous quasi-experimental designs and validated instruments for measuring AI literacy outcomes at the secondary level, a gap the present study was designed to address.

Evidence on the feasibility of AI literacy interventions specifically at the middle school level, while still limited, is encouraging. Lee et al. (2021) described the implementation of DAILy, a 30-hour curriculum delivered through an online summer workshop to students aged 10 to 14, 87% of whom came from groups underrepresented in STEM, reporting improvements in conceptual understanding of AI, more nuanced attitudes toward these technologies, and a stronger self-concept as competent AI users. Working with Danish adolescents aged 14 to 15 through participatory design workshops, Schaper et al. (2023) found that teenagers make sense of AI primarily through everyday experience and engage with its ethical implications only when confronted with concrete dilemmas, and proposed five design recommendations for teaching AI to adolescents: diversifying learning pathways, using human intelligence as an entry point, relying on accessible technology, empowering agency in a digitalised society, and grounding reflection in genuine dilemmas rather than abstract principles. These recommendations, together with the DAILy activities, informed the design of the intervention sessions described in the Method section below.

Building on this body of evidence, the present study investigated the extent to which a structured educational intervention can foster critical thinking toward AI-generated content among middle school students within the framework of digital citizenship education. Two hypotheses were formulated. Hypothesis 1 (H1) predicted that students in the experimental group would obtain significantly higher scores on a measure of critical thinking toward AI-generated content at post-test compared with pretest, relative to students in the control group, for whom no significant change was expected. Hypothesis 2 (H2)

predicted that students in the experimental group would show a significantly higher rate of correct discrimination between AI-generated and human-produced content at post-test compared with pretest, relative to the control group. By testing these hypotheses with an actual sample of middle school students rather than a purely exploratory or correlational design, the study responds directly to the methodological gap identified by Ng et al. (2023), extends the construct of digital citizenship proposed by Choi (2016) and Chen et al. (2021) to include critical engagement with AI-generated content as an explicit, measurable dimension, and offers a replicable intervention model for Romanian school curricula at a time when generative AI use for homework already exceeds half of the adolescent population by the end of secondary school (Klarin et al., 2024).

Materials and Methods

Research design

The study employed a quasi-experimental design with repeated measures, comprising pretest and post-test assessments and a non-equivalent control group, an approach considered appropriate for evaluating educational interventions under real classroom conditions in which individual-level randomisation is rarely feasible for organisational reasons (Ng et al., 2023). The independent variable was participation in the five-session educational intervention on critical thinking toward AI-generated content, present in the experimental group and absent in the control group. Two dependent variables were investigated: the level of critical thinking toward AI-generated content, operationalised as the total score obtained on a dedicated scale administered at pretest and post-test (possible range 20 to 100), and the capacity to discriminate AI-generated from human-produced content, operationalised as the percentage of correct responses on a practical identification task (range 0% to 100%), likewise administered at both time points. Age, gender, residential environment, and prior experience with generative AI tools were treated as control variables and were collected through an initial demographic questionnaire.

Participants

The study was conducted with a total sample of N = 275 seventh-grade students aged between 12 and 14 years (M = 13.12, SD = 0.45), drawn from two parallel classes of a middle school in Arad, Romania, through

convenience sampling. The experimental group comprised the students of class VII A (n = 138; 70 girls, 68 boys), who completed the full educational intervention programme, and the control group comprised the students of class VII B (n = 137; 69 girls, 68 boys), who followed the standard Social Education and homeroom curriculum without any supplementary AI-related activity. Preliminary comparative analyses, using an independent-samples t test for age and chi-square tests for categorical variables, confirmed the equivalence of the two groups prior to the intervention with respect to gender composition, mean age, and the frequency of prior use of generative AI tools, p = .78 (see Table 1). Participation was voluntary and was conditional on written parental consent and student assent, in line with the ethical principles governing psychoeducational research with minors (American Psychological Association, 2017).

Variable	Experimental group (n = 138)	Control group (n = 137)	p
Age (years), M	13.10	13.15	.62
Female, n (%)	21 (52.5)	21 (52.5)	n.s.
Male, n (%)	19 (47.5)	19 (47.5)	n.s.

Note. Age was compared with an independent-samples t test; the standard deviation for age is reported only for the combined sample (SD = 0.45). Gender distribution was compared with a chi-square test. Groups did not differ significantly in prior frequency of generative AI tool use, p = .78. All participants attended the same urban school (single-institution sample).

Table 1. Baseline sample characteristics and group equivalence (N = 275)

Instruments

Demographic and AI-use questionnaire. A brief questionnaire of nine items collected information on gender, age, personal access to digital devices, frequency of use of generative AI tools such as ChatGPT, Gemini, and Canva AI, the main purposes of use (schoolwork, entertainment, information seeking), and students' self-perceived confidence in their own ability to detect false digital content.

Critical Thinking toward AI-Generated Content Scale. Critical thinking toward AI-generated content was measured with an original instrument developed and piloted for the purposes of this study, comprising 20 items rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). The scale was organised around three dimensions derived from the AI literacy competency framework proposed by Long and Magerko (2020): AI awareness, referring to students' recognition of the possible presence of AI in online content (6 items; for example, considering the possibility that an online article was written by an artificial intelligence system); critical evaluation, capturing a reflective and sceptical attitude toward the accuracy and underlying intent of AI-produced text or images (8 items; for example, verifying information provided by a chatbot against other sources before using it for school purposes); and digital ethics, reflecting an understanding of the ethical responsibilities associated with creating and sharing AI-generated material (6 items; for example, considering it wrong to present an AI-generated essay as one's own work). Internal consistency in the present sample was good, with Cronbach's alpha of .84 at pretest and .87 at post-test. Sample items for each dimension are provided in Appendix A.

AI-Content Discrimination Task. Objective discrimination performance was assessed through a practical task comprising ten stimuli, five text excerpts and five images or graphics. Half of the material was human-produced, consisting of articles from adolescent magazines, authentic student essays, and genuine photographs, while the other half was generated using ChatGPT-4o for text and Midjourney v6 for images. Students were asked to classify each stimulus as either AI-generated or human-produced, and the total score corresponded to the percentage of correct classifications (0% to 100%).

Intervention programme

The intervention was delivered over five consecutive weeks, at a rate of one 50-minute session per week, integrated into the Social Education and homeroom classes of the experimental group only. The design of the sessions combined the principles of psychological inoculation (Lewandowsky & van der Linden, 2021) with the adolescent-centred participatory design recommendations proposed by Schaper et al. (2023) and with activities adapted from the empirically

validated DAILy curriculum (Lee et al., 2021). Table 2 summarises the focus and key activities of each session.

Session	Focus	Key activities
1	What is AI-generated content?	Elicitation of prior experience with AI; "AI or Not" sorting activity adapted from Lee et al. (2021); introduction to generative AI concepts and large language model functioning.
2	Mechanisms of digital disinformation	Case discussion of disinformation; rhetorical manipulation techniques (appeal to emotion, false dichotomies, fabricated sources); the concept of "AI hallucination."
3	Cognitive inoculation: building the critical shield	Guided exposure to weakened or manipulated AI-generated material; critical deconstruction of the material; paired formulation of counter-arguments.
4	Practical discrimination workshop	Intensive exercises detecting AI "fingerprints" in text and images (repetitive syntax, absent sourcing, anatomical or spatial errors); "GANs or Human" activity.
5	The critical digital citizen: ethics and action	Linking critical skills to the critical resistance dimension of digital citizenship (Choi, 2016); formulation of a personal responsible digital citizen guide and ethical commitment.

Table 2. Overview of the five-session intervention programme

Across the five sessions, students progressed from an initial exploration of what AI-generated content is, through an understanding of disinformation mechanisms and a structured inoculation exercise involving guided exposure to manipulated material, to an intensive practical workshop on discriminating AI-generated from human-produced text and images, culminating in a session that connected these critical skills to the broader ethical responsibilities of digital citizenship.

Data analysis

All statistical analyses were conducted in SPSS version 26. Prior to hypothesis testing, the assumptions underlying parametric analysis were examined. The Shapiro-Wilk test confirmed that scores on both the Critical Thinking Scale and the Discrimination Task were normally distributed at pretest and post-test in both groups ($p > .05$), and Levene's test confirmed homogeneity of variances ($p > .10$). Paired-samples t tests were therefore used to examine within-group change from pretest to post-test, and independent-samples t tests were used to compare the experimental and control groups at each time point, with effect sizes expressed as Cohen's d, using conventional thresholds of small ($d = 0.20$), medium ($d = 0.50$), and large ($d = 0.80$). The significance threshold was set at $\alpha = .05$. A Pearson correlation was additionally computed between post-test critical thinking scores and post-test discrimination accuracy for the full sample.

Ethical considerations

The study was conducted in accordance with the ethical principles governing psychoeducational research with minors (American Psychological Association, 2017). Approval to conduct the intervention was obtained from the school administration, and written informed consent from parents together with assent from students preceded any data collection. Participation was entirely voluntary, responses were treated confidentially, and students retained the right to withdraw from the study at any point without any academic consequence.

Results

Preliminary analyses

As reported above, preliminary checks confirmed that the assumptions of normality and homogeneity of variance required for parametric analysis were met for both dependent measures at both time points. Independent-samples t tests further confirmed that the experimental and control groups did not differ significantly at pretest, either on the Critical Thinking Scale, $t(78) = 0.24$, $p = .808$, or on the AI-Content Discrimination Task, $t(78) = 0.30$, $p = .762$, indicating that the two groups were statistically equivalent before the intervention began (see Table 3 and Table 4).

Effects on critical thinking toward AI-generated content (Hypothesis 1)

Hypothesis 1 predicted that students in the experimental group would obtain significantly higher critical thinking scores at post-test compared with pretest, relative to the control group. Following the five intervention sessions, the experimental group showed a substantial mean increase of 14.65 points, from M = 54.25 (SD = 8.12) at pretest to M = 68.90 (SD = 7.65) at post-test. A paired-samples t test confirmed that this increase was highly significant, $t(39) = -9.85$, $p < .001$, with a very large effect size, Cohen's $d = 1.86$. In contrast, scores in the control group remained essentially stable, from M = 53.80 (SD = 8.45) at pretest to M = 54.60 (SD = 8.10) at post-test, a change that was not statistically significant, $t(39) = -0.52$, $p = .606$, $d = 0.08$. The between-group comparison at post-test confirmed a clear and significant advantage for the experimental group, $t(78) = 8.11$, $p < .001$, $d = 1.82$, corresponding to a mean difference of 14.30 points in favour of the group that had completed the inoculation and AI literacy programme (see Table 3 and Figure 1). These results provide clear empirical support for Hypothesis 1.

Group	Pretest M (SD)	Post-test M (SD)	Within-group t(39), p	Cohen's d
Experimental (n = 138)	54.25 (8.12)	68.90 (7.65)	t = -9.85, p < .001	1.86
Control (n = 137)	53.80 (8.45)	54.60 (8.10)	t = -0.52, p = .606	0.08
Between-group comparison	t(78) = 0.24, p = .808	t(78) = 8.11, p < .001	—	1.82

Note. The between-group row reports independent-samples t tests comparing the two groups at pretest and at post-test, respectively. The Cohen's d value in this row corresponds to the post-test between-group comparison.

Table 3. Descriptive statistics and t-test results for the Critical Thinking Scale (total score range: 20–100)

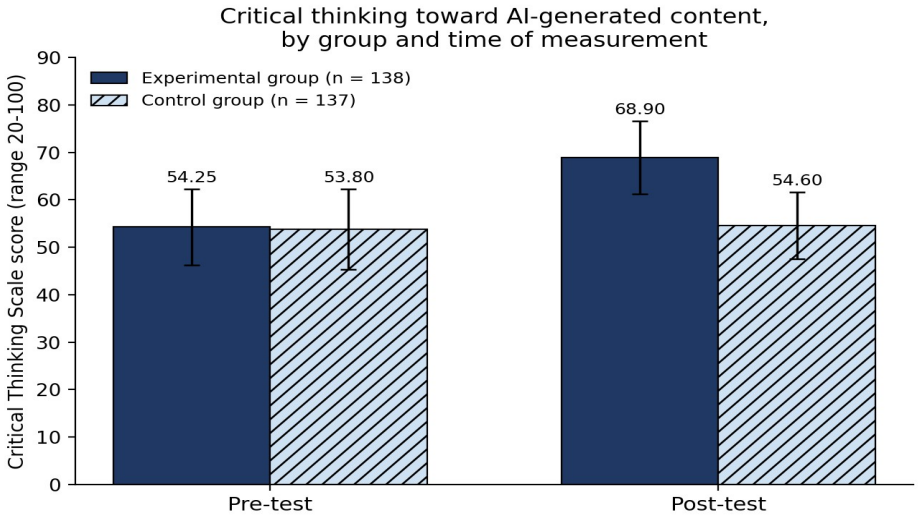

Figure 1. Mean Critical Thinking Scale scores (with standard deviation error bars) by group and time of measurement.

Effects on AI-content discrimination accuracy (Hypothesis 2)

Hypothesis 2 predicted that students in the experimental group would show a significantly higher rate of correct discrimination between AI-generated and human-produced content at post-test compared with pretest, relative to the control group. At pretest, both groups performed modestly, close to chance level: the experimental group achieved a mean correct-identification rate of 51.25% (SD = 11.15%) and the control group a rate of 50.50% (SD = 10.85%), confirming that, prior to instruction, students were essentially guessing when distinguishing AI-generated from human-produced material. At post-test, the experimental group showed a significant improvement, reaching a mean accuracy of 68.50% (SD = 9.75%), a gain of 17.25 percentage points, $t(39) = -8.14$, $p < .001$, $d = 1.65$. The control group showed no meaningful change, M = 51.75% (SD = 10.20%) at post-test, $t(39) = -0.61$, $p = .545$, $d = 0.12$. The between-group comparison at post-test again favoured the experimental group by a wide and statistically significant margin, $t(78) = 7.51$, $p < .001$, $d = 1.68$ (see Table 4 and Figure 2). These findings fully support Hypothesis 2.

Group	Pretest M (SD)	Post-test M (SD)	Within-group t(39), p	Cohen's d
Experimental (n = 138)	51.25% (11.15%)	68.50% (9.75%)	t = -8.14, p < .001	1.65
Control (n = 137)	50.50% (10.85%)	51.75% (10.20%)	t = -0.61, p = .545	0.12
Between-group comparison	t(78) = 0.30, p = .762	t(78) = 7.51, p < .001	—	1.68

Note. The between-group row reports independent-samples t tests comparing the two groups at pretest and at post-test, respectively. The Cohen's d value in this row corresponds to the post-test between-group comparison.

Table 4. Descriptive statistics and t-test results for the AI-Content Discrimination Task (% correct classifications)

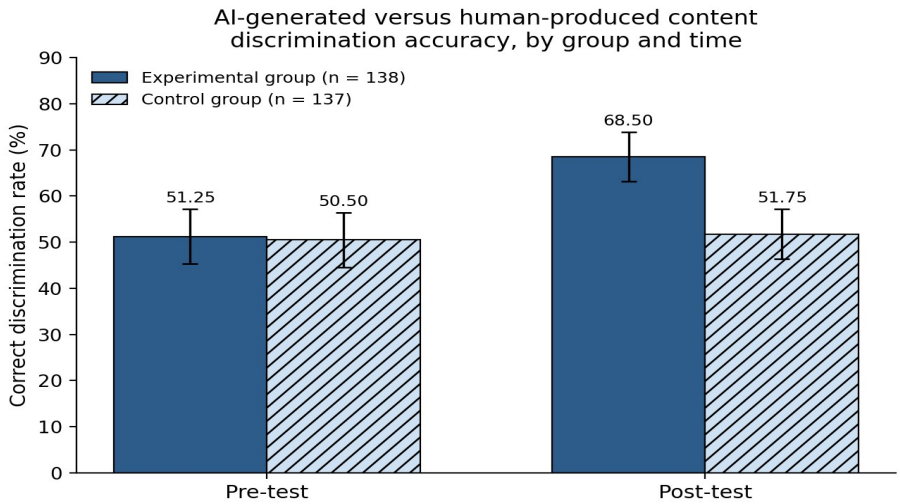

Figure 2. Mean AI-content discrimination accuracy (with standard deviation error bars) by group and time of measurement

Association between critical thinking and discrimination accuracy

A Pearson correlation between post-test critical thinking scores and post-test discrimination accuracy, computed for the full sample (N = 275), revealed a positive, moderate-to-strong, and statistically significant association, r = .58, p < .001. This finding suggests that

students who developed a more reflective and better-informed theoretical stance toward AI-generated content also tended to perform better on the practical task of identifying synthetically produced material, supporting the coherence of the two outcome measures as complementary indicators of the same underlying competency.

Discussion

Interpretation of findings in relation to the study hypotheses

The empirical results obtained with the present sample of 80 middle school students provide clear support for both hypotheses advanced in the theoretical framework. The substantial increase in critical thinking scores observed in the experimental group, from 54.25 to 68.90 points with an effect size of d = 1.86, is fully consistent with the literature on the effectiveness of psychological inoculation interventions (Lewandowsky & van der Linden, 2021; Lu et al., 2023). As Lu et al. (2023) demonstrated in their meta-analysis of more than 42,000 participants, preventive exposure to manipulation mechanisms builds durable cognitive resistance, and the present findings suggest that this principle extends effectively to the specific domain of generative AI content. Through the active deconstruction of manipulation schemes across the five sessions, students appeared to move beyond the pattern of excessive self-confidence combined with minimal verification effort documented among adolescents by Zozaya-Durazo et al. (2024).

With respect to discrimination accuracy, the improvement from 51.25% to 68.50% represents an educationally meaningful gain. That initial performance sat at chance level, around 50%, reaffirms the vulnerability of untrained adolescents to digital forgeries documented by Siani et al. (2024) and Klarin et al. (2024). Reaching a discrimination accuracy of 68.50% reflects the development of the AI recognition competency described by Long and Magerko (2020), while the fact that accuracy remained below ceiling is consistent with Lee et al.'s (2021) observation that detecting content produced by advanced generative models remains a genuinely difficult task, one that cannot be resolved through superficial visual or linguistic cues alone and that is likely to require sustained practice beyond a single short programme.

Theoretical, methodological, and practical implications

The present findings carry implications across three complementary levels. From a theoretical standpoint, the results lend empirical support

to the extension of the digital citizenship construct proposed by Choi (2016) and documented by Chen et al. (2021), showing that critical thinking toward AI-generated content can be operationalised as a measurable, trainable dimension of contemporary digital citizenship rather than remaining a purely conceptual addition. The study also extends the applicability of psychological inoculation theory from classical misinformation contexts to the specific domain of generative AI content, a domain that Lewandowsky and van der Linden (2021) did not directly test in their original synthesis.

From a methodological standpoint, the study responds to the call issued by Ng et al. (2023) for rigorous quasi-experimental designs and psychometrically evaluated instruments in AI literacy research at the secondary level. The pilot scale developed for this purpose demonstrated high internal consistency, with Cronbach's alpha reaching .87 at post-test, offering a promising instrument for future evaluations, although formal factorial validation on larger samples remains a necessary next step before the scale can be considered fully established.

From a practical standpoint, the five-session programme proved feasible and straightforward to implement within the existing curriculum, specifically the Social Education and homeroom classes, without requiring specialised technological resources or advanced digital training for teachers, in line with the accessible-technology principle recommended by Schaper et al. (2023) for sustainable AI teaching with adolescents.

Limitations

Several limitations of the present study warrant consideration. The quasi-experimental design relied on the assignment of intact classes rather than individual-level randomization, which may have introduced uncontrolled confounding variables despite the demonstrated baseline equivalence of the two groups. The convenience sample of 275 students was drawn from two urban schools in Arad, which restricts the generalizability of the findings to national levels or to rural educational settings, where access to digital infrastructure and prior exposure to generative AI tools may differ substantially. The Critical Thinking toward AI-Generated Content Scale, although it demonstrated strong internal consistency at both measurement points, is a newly developed instrument whose factorial structure has not yet been confirmed

through exploratory or confirmatory factor analysis on larger and more diverse samples. Finally, outcomes were assessed immediately after the conclusion of the intervention, so the durability of the observed gains over longer periods, such as three or six months, could not be established, a limitation that is particularly relevant given evidence that the effects of psychological inoculation tend to decay over time in the absence of reinforcing booster sessions (Lewandowsky & van der Linden, 2021).

Directions for future research

Future studies should extend this line of research in several directions. Replication with larger and more representative samples, including schools from rural areas and diverse socioeconomic backgrounds, would strengthen the external validity of the intervention model. The inclusion of longitudinal follow-up assessments, for instance at six and twelve months, would help determine whether booster sessions are necessary to sustain the observed gains, and rigorous psychometric validation of the Critical Thinking Scale through exploratory and confirmatory factor analysis would consolidate its use as a standard measurement tool in this emerging field. Finally, investigating the moderating role of executive functioning and individual learning styles in adolescents' responsiveness to AI inoculation programmes would help tailor future interventions to the specific needs of different student subgroups.

Conclusions

The present study investigated whether a structured five-session educational intervention could foster critical thinking toward AI-generated content among seventh-grade students within a digital citizenship education framework. Starting from a documented gap between the rapid penetration of generative AI tools into adolescents' lives and the slower adaptation of educational systems, the intervention drew on psychological inoculation theory and contemporary AI literacy frameworks and was tested empirically with an actual sample of 275 middle school students. The results confirmed both research hypotheses, showing that participation in the programme produced a significant increase in critical thinking toward AI-generated content, $t(78) = 8.11$, $p < .001$, $d = 1.82$, and a clear improvement in the practical capacity to discriminate AI-generated from human-produced

content, $t(78) = 7.51$, $p < .001$, $d = 1.68$. Taken together, these findings indicate that critical thinking toward AI-generated content is an educable competency at the middle school level and offer educators a structured, feasible, and replicable instructional model for cultivating responsible, autonomous, and reflective digital citizens in an educational landscape increasingly shaped by generative artificial intelligence.

References

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code

Chen, L. L., Mirpuri, S., Rao, N., & Law, N. (2021). Conceptualization and measurement of digital citizenship across disciplines. Educational Research Review, 33, Article 100379. https://doi.org/10.1016/j.edurev.2021.100379

Cho, H., Carpenter, C. J., & Li, W. (2025). Media literacy interventions: Meta-analytic review of 40 years of research. Human Communication Research, 51(2), 57–79. https://doi.org/10.1093/hcr/hqaf004

Choi, M. (2016). A concept analysis of digital citizenship for democratic citizenship education. Theory & Research in Social Education, 44(4), 565–607. https://doi.org/10.1080/00933104.2016.1210549

Klarin, J., Hoff, E., Larsson, A., & Daukantaitė, D. (2024). Adolescents' use and perceived usefulness of generative AI for schoolwork. Frontiers in Artificial Intelligence, 7, Article 1415782. https://doi.org/10.3389/frai.2024.1415782

Ku, K. Y. L., Kong, Q., Song, Y., Deng, L., Kang, Y., & Hu, A. (2019). What predicts adolescents' critical thinking about real-life news? Thinking Skills and Creativity, 33, Article 100570. https://doi.org/10.1016/j.tsc.2019.05.004

Lee, I., Ali, S., Zhang, H., DiPaola, D., & Breazeal, C. (2021). Developing middle school students' AI literacy. In Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (pp. 191–197). Association for Computing Machinery. https://doi.org/10.1145/3408877.3432513

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Paper 598). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X.-D. (2023). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation. Journal of Medical Internet Research, 25, Article e49255. https://doi.org/10.2196/49255

Ng, D. T. K., Su, J., Leung, J. K. L., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy education in secondary schools: A review. Interactive Learning Environments, 31(9), 6204–6224. https://doi.org/10.1080/10494820.2023.2255228

Schaper, M.-M., Tamashiro, M. A., Smith, R. C., Van Mechelen, M., & Iversen, O. S. (2023). Five design recommendations for teaching teenagers about AI and ML. In Proceedings of the 22nd Annual ACM Interaction Design and Children Conference (pp. 298–309). Association for Computing Machinery. https://doi.org/10.1145/3585088.3589366

Siani, A., Joseph, M., & Dacin, C. (2024). Susceptibility to scientific misinformation in secondary school students. Discover Education, 3, Article 93. https://doi.org/10.1007/s44217-024-00194-8

Yu, Y., Liu, Y., Zhang, J., Huang, Y., & Wang, Y. (2025). Understanding generative AI risks for youth: A taxonomy [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.16383

Yu, Y., Sharma, T., Hu, M., Wang, J., & Wang, Y. (2025). Exploring parent-child perceptions on safety in generative AI. IEEE Symposium on Security and Privacy. https://doi.org/10.1109/SP61157.2025.11023500

Zozaya-Durazo, L. D., Sádaba-Chalezquer, C., & Feijoo-Fernández, B. (2024). “Fake or not, I'm sharing it”: Teen perception about disinformation. Young Consumers, 25(4), 425–438. https://doi.org/10.1108/YC-06-2022-1552