schoolai

Promoting Student Critical Thinking Through AI Integration:

A Two-Year Implementation Study of the SchoolAI Platform

Abstract

This study examines changes in student critical thinking following implementation of the SchoolAI platform in a large suburban school district serving 55,000 students. Using anonymized teacher identifiers, we identified 82 teachers with student-AI conversations at both timepoints: October 2023 (early implementation) and October 2025 (mature implementation), yielding 13,882 conversations aggregated to teacher-level mean scores. Critical thinking was scored on a four-level framework adapted from Bloom's Taxonomy, with algorithmic coding validated through inter-rater reliability testing (79% agreement). Results demonstrated significant within-teacher improvements: mean critical thinking scores increased by 28.3% (M = 1.38 to M = 1.77), with a medium effect size (Cohen's d = 0.47), and 65.9% of teachers showed gains. The mean proportion of Level 1 (no observable critical thinking) conversations decreased by 18.7 percentage points while higher-level critical thinking (Levels 3 and 4) more than doubled. Dosage analyses revealed that significant improvement was observed only among teachers who created six or more unique learning spaces, establishing a minimum implementation threshold for measurable gains. A one-way ANOVA on difference scores confirmed that gains were consistent across subject areas and grade levels. These findings provide promising evidence that AI integration, when implemented with appropriate professional support and instructional design, can promote observable growth in student critical thinking at scale.

Introduction

Facilitating higher-order thinking at scale has been a persistent challenge in K-12 instruction. Classic studies found that 80-90% of teacher questions focused on recall and comprehension (Cotton, 1988; Gall, 1970), and while instructional approaches have evolved, lower-order thinking continues to dominate much of classroom interaction (Sam et al., 2022).

The integration of AI into K-12 classrooms raises new questions about whether such technologies might address this challenge or exacerbate it. While AI tools offer unprecedented access to personalized support and on-demand feedback, concerns persist about whether they promote or supplant critical thinking.

This study examines changes in observable critical thinking during student-AI interactions following implementation of the SchoolAI platform in a large suburban school district. Using a paired comparison of 82 teachers across two years of implementation, we assess whether SchoolAI integration is associated with within-teacher improvements in student cognitive engagement. Three research questions guide this study:

- Do student critical thinking scores improve between early and mature SchoolAI implementation?
- Is improvement in student critical thinking associated with the extent of teacher engagement with SchoolAI?
- How does the composition of Level 1 (no observable critical thinking) engagement change over time?
- Are observed improvements consistent across subject areas and grade levels?

Methods

Sample Characteristics

Data were collected from a large suburban school district in Utah serving approximately 55,000 students in grades K-12 across 68 schools in the greater Salt Lake City metropolitan area. District demographics include:

- **Race/Ethnicity:** White: 73%, Hispanic/Latine: 18% Asian: 4%, Two or more races: 3%
- **Free or reduced lunch:** 32%
- **On state assessments (RISE and Utah Aspire Plus) in 2024-25:** 47.5% of students scored proficient or above in Reading, 26.3% in Mathematics, 32.1% in Science

The district began SchoolAI implementation in August 2023. To evaluate platform effectiveness, this study compares October 2023 (two months post-launch) with October 2025, maintaining consistent timing within the academic year. Using anonymized teacher identifiers, we identified 82 teachers with student-AI conversations at both timepoints, yielding 13,882 conversations (9,191 in 2023; 4,691 in 2025). These 82 teachers represent 33 schools across the district (7 high schools, 10 middle schools, 13 elementary schools, and 3 other campuses). Conversations were aggregated to teacher-level mean scores to ensure statistical independence. Individual student identifiers are not available due to the platform's privacy-by-design architecture.

Implementation Context

SchoolAI is not mandated but is one of the district's six priority instructional tools; all other AI tools are blocked on district WiFi. In October 2023, platform functionality was mainly check-in activities designed to gauge student understanding and well-being, aligned with assessing student critical thinking. Over the following two years, supported by sustained professional development and collaboration with SchoolAI, teachers built a variety of interactive spaces such as historical figure interviews, content review activities, and guided explorations. Usage deepened for both students and teachers, though the level of critical thinking demonstrated remained dependent on student engagement.

Implementation Dosage

Dosage groups were defined based on the number of unique learning spaces each teacher created over the two-year implementation period. Cutpoints were informed by both implementation context and the data distribution. Teachers creating 1-5 spaces were classified as low implementation, reflecting minimal engagement beyond the platform's initial check-in functionality. Teachers creating 6-20 spaces were classified as medium implementation, and those creating 21 or more as high implementation. The threshold of 6 spaces was selected as the minimum indicator that a teacher had moved beyond initial exploration to active instructional design, while the upper cutpoint of 21 approximates the sample median of 16, dividing active users into moderate and intensive implementation groups.

Demographic Composition

The 82 paired teachers represent a broad range of subject areas and grade levels. The largest subject groups are ELA (22 teachers), Other (18), Social Sciences (12), and Science (11), with smaller representation in Technology, Career & Business, Math, World Languages, and Test Prep. By grade band, the sample skews toward secondary: 35 teachers primarily serve grades 9-12, 22 serve grades 6-8, and 13 serve grades 3-5, with one teacher in K-2. Because the same teachers appear at both timepoints, modal subject and grade band classifications showed minimal change across years, with the large majority of teachers maintaining the same primary subject and grade level.

Table 1. Subject Distribution by Year

Subject	Teachers	Oct 2023 Chats % (n)	Oct 2025 Chats % (n)
ELA	22	19.5% (1,788)	19.4% (909)
Other	18	14.1% (1,296)	26.6% (1,249)
Social Sciences	12	34.3% (3,148)	24.3% (1,142)
Science	11	11.0% (1,008)	10.1% (475)
Technology	7	4.8% (444)	0.1% (6)
Career & Business	3	11.8% (1,084)	13.3% (623)
Math	1	0.0% (2)	0.0% (1)
Test Prep	1	0.9% (85)	0.0% (0)
World Languages	1	0.0% (1)	0.0% (1)

Note: 6 teachers with unclassified subjects

Table 2. Grade Band Distribution by Year

Grade Band	Teachers	Oct 2023 Chats % (n)	Oct 2025 Chats % (n)
9-12	35	40.6% (3,736)	42.3% (1,982)
6-8	22	48.4% (4,449)	44.4% (2,083)
3-5	13	2.9% (268)	7.5% (350)
K-2	1	0.5% (45)	0.1% (5)

Note: 11 teachers with unclassified grade bands

Coding

Critical thinking levels were algorithmically scored based on conversational patterns indicating cognitive engagement, adapting Bloom's Taxonomy to observable student-AI interactions. The rubric was developed through iterative refinement combining rule-based scoring and human review, with indicator phrases expanded to capture natural student language patterns and context-specific rules applied (e.g., creative writing submissions in ELA were not coded for critical thinking, as narrative conventions do not indicate metacognition in that context).

Table 3. Student Critical Thinking Rubric

Level	Definition	Example Indicators
1	No observable critical thinking	"What is...?", "ok", choice selections ("A", "2"), off-topic conversation, pasted content without engagement
2	Remember + Understand	"The answer is...", "First I..., then I...", restating information, seeking clarification
3	Apply + Analyze	"I multiplied because...", "This works when...", comparing approaches, explaining reasoning
4	Evaluate + Create	"Method A is better because...", "I noticed a pattern...", revising work, synthesizing ideas

Inter-rater Reliability

Over multiple rounds, 250 conversations were independently scored by AI and a human researcher, reaching 79% agreement. Disagreements concentrated in Level 1 classifications, reflecting inherent difficulty in distinguishing information-seeking from minimal engagement without full pedagogical context.

AI Assistance

In addition to coding, Claude (Anthropic) was used to assist with drafting and editing portions of this manuscript. All data analysis, interpretation, and substantive decisions were made by the research team.

Data Privacy

Students do not need to create accounts on SchoolAI. Chats are linked to anonymous session IDs and anonymized teacher identifiers; datasets contain no student identifiers and no personally identifiable teacher information. All analyses were performed locally with no external data transmission and all data handling follows the SchoolAI Data Privacy Policy.

Results

Do student critical thinking scores improve between early and mature SchoolAI implementation?

Teacher-level mean critical thinking scores increased from October 2023 (M = 1.38, SD = 0.49) to October 2025 (M = 1.77, SD = 0.69), representing a 28.3% gain. A paired samples t-test confirmed this difference was statistically significant, $t(81) = 4.21$, $p < .001$, with a medium effect size (Cohen's $d = 0.47$). Of the 82 paired teachers, 54 (65.9%) showed improvement, 22 (26.8%) showed decline, and 6 (7.3%) were unchanged. The mean proportion of Level 1 conversations decreased by 18.7 percentage points, while Levels 3 and 4 more than doubled.

Table 4. Critical Thinking Level Distribution by Year (Teacher-Averaged Proportions)

Level	Oct 2023 Chats % (n)	Oct 2025 Chats % (n)	Change
Level 1	76.7% (7,280)	58.0% (2,694)	-18.7%
Level 2	11.8% (1,147)	16.5% (663)	+4.6%
Level 3	8.2% (557)	16.0% (905)	+7.8%
Level 4	3.2% (207)	9.5% (429)	+6.3%
Total	**100.0% (9,191)**	**100.0% (4,691)**	

Because conversation volume varied substantially across teachers (9,191 conversations in 2023; 4,691 in 2025; median per teacher 55 and 20, respectively), teacher-level means based on few conversations may be less stable. Sensitivity analyses restricting the sample to teachers with at least 10 conversations per timepoint (n = 34) yielded larger effect sizes (Cohen's $d = 0.78$, 82.4% improved), indicating that results are robust to conversation volume thresholds and that teachers with more stable estimates showed stronger gains. The full sample of 82 teachers was retained for primary analyses to maximize statistical power and maintain representation across school types and subject areas.

Is improvement in student critical thinking associated with the extent of teacher engagement with SchoolAI?

To examine whether improvement was linked to product usage, paired t-tests were conducted within each dosage group, comparing each teacher's October 2023 mean score to their October 2025 mean score. This within-subjects approach directly links product engagement to outcomes while respecting the repeated-measures structure of the data and testing whether sufficient engagement with SchoolAI is necessary for significant improvement.

Table 5. Critical Thinking Improvement by Implementation Dosage

Group	n	M 2023	M 2025	Mean Δ	Cohen's d	Paired t	p	Improved
Low (1-5 spaces)	11	1.47	1.65	+0.18	0.34	t(10) = 1.12	.288	8/11 (72.7%)
Medium (6-20 spaces)	40	1.40	1.87	+0.47	0.47	t(39) = 2.98	.005	28/40 (70.0%)
High (21+ spaces)	31	1.32	1.68	+0.36	0.52	t(30) = 2.91	.007	18/31 (58.1%)
6+ combined	71	1.36	1.79	+0.42	0.48	t(70) = 4.07	< .001	46/71 (64.8%)

Note: To account for multiple comparisons across the three dosage groups, a Bonferroni-corrected significance threshold of α = .017 was applied. Both medium (p = .005) and high (p = .007) implementation groups remained significant, while the low implementation group did not (p = .288).

Significant improvement was observed only among teachers who created six or more unique learning spaces, suggesting a minimum implementation threshold is necessary for measurable gains in student critical thinking.

How does the nature of low-engagement interactions change over time?

Level 1 classifications indicate that critical thinking was not observable in the conversational record but do not confirm its absence. Two distinct scenarios contribute to Level 1 codes. First, some students may have been genuinely disengaged. However, many Level 1 conversations reflect instructional activities not designed to elicit observable critical thinking, such as guided book selection tools, character interview simulations, or choose-your-own-adventure scenarios where student turns consist entirely of choices or questions.

Table 6. Level 1 Subcategories

Subcategory	Oct 2023 Chats % (n)	Oct 2025 Chats % (n)	Δ%
Minimal responses	39.2% (2,854)	28.4% (766)	-10.8
Information-seeking	34.2% (2,492)	27.3% (735)	-6.9
Makes choices only	20.1% (1,464)	36.2% (976)	+16.1
Off-topic	3.8% (275)	1.8% (49)	-2.0
Pasted content	2.4% (173)	6.1% (163)	+3.7
Doesn't want to work	0.3% (20)	0.2% (5)	-0.1
Total	**100.0% (7,280)**	**100.0% (2,694)**	

The composition of Level 1 engagement shifted significantly between timepoints, $\chi^2(5) = 403.60$, $p < .001$, Cramér's $V = .201$. Choice-making activities increased from 20.1% to 36.2% of Level 1 conversations, while minimal responses and information-seeking both decreased. This shift suggests that remaining Level 1 conversations in 2025 increasingly reflect pedagogically intentional activities rather than passive or disengaged behavior.

Are observed improvements consistent across subject areas and grade levels?

Descriptively, improvements were observed across most subject areas and grade bands (Tables 6 and 7). Science teachers showed the largest gains ($\Delta M = +0.52$, 10 of 11 improved), followed by Career & Business (+0.53, 3 of 3) and ELA (+0.45, 13 of 22). Technology showed the smallest improvement (+0.05, 3 of 7). By grade band, grades 6-8 showed the largest gains ($\Delta M = +0.44$, 15 of 22 improved), followed by 9-12 (+0.40, 26 of 35) and 3-5 (+0.32, 7 of 13). Cell sizes for Math, World Languages, and K-2 (each n = 1) preclude meaningful interpretation; post-hoc comparisons were not conducted due to insufficient sample sizes within individual subject and grade categories.

A one-way ANOVA on teacher-level difference scores examined whether the magnitude of improvement varied by instructional context. Neither subject area ($F(7, 58) = 0.61$, $p = .748$) nor grade band ($F(3, 58) = 0.20$, $p = .894$) were significant, indicating that gains were consistent regardless of what or whom teachers taught.

Table 7. Difference in Teacher-Level Mean Critical Thinking Scores by Subject

Subject	n	M 2023	M 2025	ΔMean	Improved
Career & Business	3	1.39	1.92	+0.53	3/3
Science	11	1.13	1.65	+0.52	10/11
Test Prep	1	1.47	1.96	+0.49	1/1
Other	18	1.30	1.76	+0.46	13/18
ELA	22	1.63	2.08	+0.45	13/22
Social Sciences	12	1.31	1.65	+0.33	8/12
Technology	7	1.30	1.35	+0.05	3/7
Math	1	1.67	1.00	-0.67	0/1

Table 8. Difference in Teacher-Level Mean Critical Thinking Scores by Grade Band

Grade Band	n	M 2023	M 2025	ΔMean	Improved
6-8	22	1.39	1.83	+0.44	15/22
9-12	35	1.35	1.74	+0.40	26/35
3-5	13	1.38	1.70	+0.32	7/13
K-2	1	1.31	1.00	-0.31	0/1

Limitations

This study has several limitations. The algorithmic scoring system, while achieving 79% inter-rater reliability, remains subject to measurement error, particularly in distinguishing Level 1 subcategories where pedagogical context is crucial. The 82 paired teachers represent those who persisted with the platform across two years and may not be representative of all district teachers. Additionally, the pre-post design without a control group precludes causal inference; the platform and professional development evolved over the study period, so observed gains cannot be attributed to any single factor. Finally, conversational records capture only observable evidence of thinking; students may engage in complex reasoning that remains unexpressed, particularly in activities where brief responses are appropriate.

Conclusion

This study provides promising evidence that sustained SchoolAI implementation is associated with meaningful growth in student critical thinking, as measured through a paired teacher-level comparison across two academic years.

Among 82 teachers present at both timepoints, critical thinking scores increased by 28.3%, with 65.9% of teachers showing gains. The mean proportion of Level 1 conversations decreased by 18.7 percentage points while higher-level engagement (Levels 3 and 4) more than doubled. Dosage analyses established that significant improvement occurred among teachers who created six or more unique learning spaces ($p < .001$, Cohen's $d = 0.48$), while teachers with minimal implementation did not show significant gains, linking outcomes directly to product engagement. A supplementary analysis confirmed that gains were consistent across subject areas and grade levels.

These findings are notable in context. Research has documented that lower-order cognitive engagement predominates in typical classroom interactions. The 2023 baseline of 76.7% Level 1 conversations reflected this pattern. By 2025, this decreased to 58.0%, with higher-order engagement rising from 11.4% to 25.5% (Levels 3 and 4) of teacher-averaged conversation proportions. The composition of remaining Level 1 conversations also evolved qualitatively, with choice-based learning activities replacing passive behaviors.

The substantial improvements in cognitive engagement, combined with a paired design holding teacher identity constant, dosage analyses linking product usage to outcomes, and consistent gains across instructional contexts, demonstrate that the SchoolAI platform is an evidence-based educational intervention that promotes critical thinking at scale.

References

Cotton, K. (1988). Classroom questioning. Northwest Regional Educational Laboratory. https://educationnorthwest.org/sites/default/files/ClassroomQuestioning.pdf

Gall, M. D. (1970). The use of questions in teaching. Review of Educational Research, 40(5), 707-721. https://doi.org/10.3102/00346543040005707

Sam, A. L., Somasundram, P., Yasin, R. M., Ponnusamy, L. D., & Meerah, T. S. M. (2022). Using Bloom's taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools. Cogent Education, 9(1), Article 2008024. https://doi.org/10.1080/2331186X.2021.2008024