Journal of Insan Mulia Education
E-ISSN: 2987-0585 | Vol. 3, No. 2, 2025
https://doi.org/10.59923/joinme.v3i2.784

ChatGPT-Assisted Contextual Teaching and Learning for Fifth-Grade Poetry Writing, Creativity, and Language Literacy: A Quasi-Experimental Study

Mulyadi[1),*], Zulkifli[1)]

[1)]STKIP Taman Siswa Bima, Bima, Indonesia

*Correspondence: mulyadinur02@gmail.com

ABSTRACT

Many fifth-grade students in Indonesia still write poems with limited diction and imagery, and little is known about how generative artificial intelligence can be embedded in contextual pedagogy for young learners. This study examined whether integrating ChatGPT into Contextual Teaching and Learning (CTL) improves fifth-grade students' poetry writing, creative thinking, and language literacy compared with conventional CTL. A quasi-experimental non-equivalent control group design was applied to 55 fifth-grade students in Bima City, Indonesia (experimental n = 28; control n = 27) over four learning sessions. Rubric-based tests covering seven indicators (diction, imagery, rhyme and rhythm, ideational fluency, originality, vocabulary, and literary appreciation) were calibrated with the Rasch model (person reliability = 0.93). Data were analysed using independent-samples t-tests, normalised gain (N-gain), and effect sizes. The experimental group achieved a moderate mean N-gain (55.10%) compared with a low gain in the control group (14.51%; Welch's $t = 4.24$, $p < .001$, $d = 1.15$). However, posttest differences between the groups were not statistically significant for any of the seven indicators ($p = .088–.363$), and no meaningful advantage was found for creative thinking or language literacy. ChatGPT-assisted CTL therefore shows promise for supporting gains in poetry writing, but the evidence is preliminary given the small sample and short intervention. Longer, adequately powered studies with ethical safeguards for young AI users are recommended.

Keywords: Contextual Teaching and Learning; ChatGPT; Generative AI; Poetry Writing; Creativity

This is an open access article under the CC - BY license.

INTRODUCTION

The paradigm shift in 21st-century education requires learning models that go beyond knowledge transmission to emphasize creativity, literacy, and the intelligent use of digital technologies. In a global context, the PISA 2022 results showed that reading performance declined in many participating countries, including Indonesia (OECD, 2023). Because PISA assesses 15-year-old students, these results mainly signal weaknesses in literacy foundations that are built in primary school. Primary education holds a vital role as the foundation of language literacy and creative competence, where poetry writing contributes significantly to imagination, aesthetics, and emotional expression (Ulfah et al., 2023). Meta-analytic evidence shows that writing instruction in the elementary grades is most effective when it is explicit, scaffolded, and responsive to individual learners (Graham et al., 2012), yet teachers face challenges in providing such individual creative guidance (Diani & Sukartono, 2022). Generative artificial intelligence (AI), that is, systems that produce new text in response to user prompts (García-Peñalvo & Vázquez-Ingelmo, 2023), such as ChatGPT, offers a possible source of this individual support (Fui-Hoon Nah et al., 2023; Kasneci et al., 2023).

Despite ongoing reforms, students' creative writing skills, especially in producing original and coherent poetry, remain low, as illustrated by the preliminary observation reported below. The integration of digital literacy in language education also remains uneven, and research on AI in education has concentrated on higher education rather than on primary classrooms (Zawacki-Richter et al., 2019). This study therefore examines the integration of Contextual Teaching and Learning (CTL) with ChatGPT for poetry writing, creativity, and language literacy in elementary education; the specific gap and novelty are elaborated after the review of related studies below.

A preliminary observation was conducted in two fifth-grade classes involving 56 students. Thirty-eight students (67.9%) were only able to write poems with three simple stanzas without imaginative diction, 11 students (19.6%) showed some lexical variety but weak cohesion, and only 7 students (12.5%) were able to write poems with more complex imagery and style (Table 1).

https://doi.org/10.59923/joinme.v3i2.784
Vol. 3, No. 2, 2025

Table 1. Initial Observation Results of Students' Poetry Writing Skills

Ability Category	n (%)	Poem Characteristics
Low	38 (67.9%)	Simple structure, minimal imagery, repetitive
Medium	11 (19.6%)	Some lexical variety but weak cohesion
High	7 (12.5%)	Rich imagery, regular rhyme, original

The results show that the majority of students are still in the low category, especially in terms of language creativity. This condition suggests that the current learning model offers limited room for creative exploration, although this interpretation rests on a small preliminary observation and should be read with caution. Evidence from other settings indicates that AI-based tools can support poetry writing: Kangasharju et al. (2022) examined how lower secondary students wrote poems with the AI-based Poetry Machine and reported that the tool supported their exploration of language. Their study differs from the present one in its population (lower secondary rather than primary students) and its tool (a dedicated poetry generator rather than a general-purpose large language model such as ChatGPT).

The pedagogical solution proposed in this study is grounded in three theoretical strands. First, CTL links subject matter to learners' real-life contexts through seven components: constructivism, inquiry, questioning, learning community, modelling, reflection, and authentic assessment (Johnson, 2002); its effectiveness in improving learning outcomes has been reported in language learning (Haryono & Hikmah, 2023) and reviewed in broader settings (Hasudungan, 2022). Second, creativity is understood as the production of ideas that are both original and effective (Runco & Jaeger, 2012), and divergent thinking is commonly described through fluency, flexibility, originality, and elaboration (Guilford, 1967; Torrance, 1974); this study focuses on fluency and originality, the dimensions most directly observable in short poems. Third, sociocultural learning theory holds that learners can perform beyond their independent level when supported by a more knowledgeable other within their zone of proximal development (Vygotsky, 1978), through scaffolding that is gradually withdrawn (Wood et al., 1976). In this study, ChatGPT is positioned as such a scaffold within the CTL components: it supports inquiry and questioning (students ask for ideas and word alternatives), modelling (it provides examples of diction and imagery), and reflection (students compare and revise their drafts), while the teacher remains responsible for the learning community and authentic assessment (Fui-Hoon Nah et al., 2023; Veddayana et al., 2023). This combination is intended to provide individual feedback that is difficult for a teacher alone to offer to every student (Christopoulos & Sprangers, 2021; Hwang et al., 2023). Language literacy is operationalised here as students' understanding of word meanings in poetic texts and their appreciation of literary messages and aesthetic value.

Recent studies show that generative AI is increasingly used in education, including in writing instruction. Reviews have outlined both the opportunities and the risks of large language models for learning (Kasneci et al., 2023; Kohnke et al., 2023), and classroom studies have explored how ChatGPT can support students' writing processes, for example in argumentative writing (Su et al., 2023) and in second-language writing practicums (Yan, 2023). Research on technology-enhanced creativity likewise shows that digital tools can foster students' creative work when teachers integrate them deliberately (Bereczki & Kárpáti, 2021; Ozkan & Umdu Topsakal, 2021). However, research on AI in education has concentrated on higher education (Zawacki-Richter et al., 2019), and the ChatGPT writing studies cited above involved older learners. Studies that integrate CTL with ChatGPT for poetry writing in elementary school remain scarce.

Figure 1 presents a keyword co-occurrence map of research on Contextual Teaching and Learning, which was used to identify the themes that dominate this field.

https://doi.org/10.59923/joinme.v3i2.784
Vol. 3, No. 2, 2025

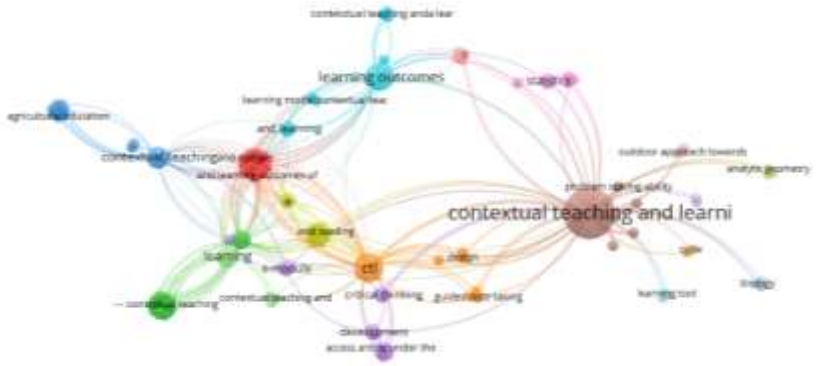


Figure 1. Keyword Co-occurrence Network Visualization of Contextual Teaching and Learning Research

The novelty of this study lies in three aspects. First, the bibliometric mapping shows that CTL studies have focused mainly on learning outcomes, critical thinking, problem-solving, guided note-taking, and science subjects such as biology and geometry, whereas the integration of CTL with literary literacy and aesthetic creativity is rarely explored. Second, research on ChatGPT in education has concentrated on older learners and on academic writing (Kasneci et al., 2023; Su et al., 2023), so its use as a medium for creative language exploration by elementary school students remains largely untested. Third, theoretically, the study tests whether a general-purpose AI tool can act as a more knowledgeable other within the CTL cycle (Vygotsky, 1978; Wood et al., 1976), thereby extending CTL from cognitively oriented learning outcomes towards creative literacy (Bereczki & Kárpáti, 2021; Sætra, 2023). The study also addresses an ethical dimension that is rarely discussed: ChatGPT's terms of use set a minimum user age of 13 years with parental consent, so its use with 10–11-year-old students raises questions of access, consent, and supervision, which are addressed in the Method.

Based on the above description, the main problems can be formulated as follows: (1) How effective is the integration of CTL-ChatGPT in improving elementary school students' poetry writing skills? (2) How does CTL-ChatGPT influence the development of students' creative thinking? (3) How does CTL-ChatGPT contribute to strengthening students' language literacy? This study aims to: (a) analyze the effect of CTL-ChatGPT on poetry writing skills; (b) examine the contribution of CTL-ChatGPT in fostering students' creative thinking; and (c) evaluate the impact of CTL-ChatGPT on improving the language literacy of elementary school students. Accordingly, three hypotheses were tested: H1, students taught with CTL-ChatGPT obtain higher poetry-writing scores than students taught with conventional CTL; H2, students taught with CTL-ChatGPT obtain higher creative-thinking scores than students taught with conventional CTL; and H3, students taught with CTL-ChatGPT obtain higher language-literacy scores than students taught with conventional CTL. These objectives are oriented towards achieving an innovative pedagogical model that is adaptive to technological developments and relevant to educational needs in the digital era.

METHOD

This study employed a quantitative approach using a quasi-experimental method with a Non-Equivalent Control Group Design, which was deemed appropriate for the elementary education context due to the constraints of classroom randomization. The design involved two groups: an experimental class that received Contextual Teaching and Learning (CTL) integrated with ChatGPT, and a control class that implemented conventional CTL without artificial intelligence assistance. The CTL–ChatGPT intervention was structured to foster active student engagement through authentic learning contexts, utilizing ChatGPT as a cognitive scaffold to support idea generation, diction selection, and the creative structuring of poetry. The key distinction between CTL–ChatGPT and conventional CTL lies in the interaction and feedback mechanisms: in the experimental class, students received AI-generated suggestions on ideas, diction, rhythm, and imagery in addition to teacher feedback, whereas in the control class all feedback came from the teacher. The step-by-step activities of both classes are presented in Table 3.

Ethical considerations. Participation required permission from the students' parents or guardians. Because ChatGPT's terms of use require users to be at least 13 years old with parental consent, the implications of its use with 10–11-year-old students are discussed in the Limitations.

https://doi.org/10.59923/joinme.v3i2.784
Vol. 3, No. 2, 2025

To reduce threats to internal validity, classroom conditions, teacher facilitation, and instructional duration (four sessions) were kept comparable across the two classes, and both classes completed the same pretest and posttest. The statistical procedures are described in the Data analysis subsection.

Table 2. Non-Equivalent Control Group Design

Group	Pretest	Treatment	Posttest
Experimental	O1	ChatGPT-assisted CTL	O2
Control	O3	CTL without ChatGPT (conventional)	O4

Note. O1 and O3 = pretest; O2 and O4 = posttest.

The research procedure consisted of four stages. First, the instrument was reviewed by experts for content validity. Second, a pretest was administered to both groups. Third, the treatment was implemented over four learning sessions; the experimental group followed CTL integrated with ChatGPT, and the control group followed conventional CTL led by the teacher without technological support (Table 3). Fourth, a posttest was administered to both groups. The instrument was calibrated with the Rasch model, as described under Instruments. The short duration of the intervention is acknowledged as a limitation (see Discussion).

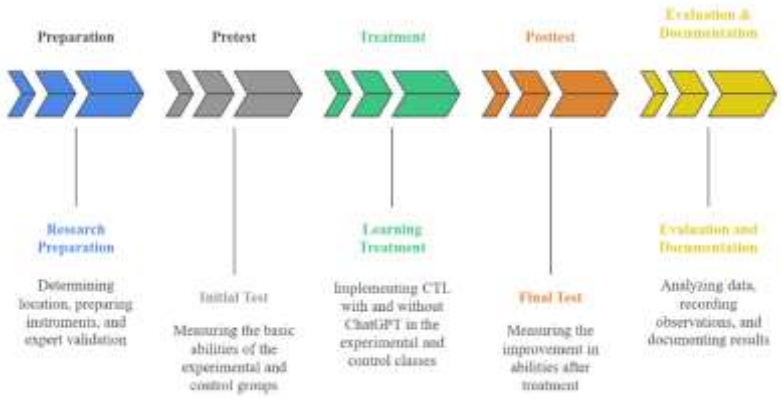


Figure 2. Research Procedure Flowchart

Table 3. Learning Activities in the Control and Experimental Classes According to the CTL Components

CTL component	Control class (conventional CTL)	Experimental class (CTL–ChatGPT)
Constructivism	The teacher links the poem theme to students' daily experiences; students list ideas from their experience.	Same as control; students also ask ChatGPT for additional ideas related to their experiences.
Inquiry	Students observe the environment or pictures related to the theme and note sensory details.	Same as control; students also ask ChatGPT for word alternatives that describe what they observed.
Questioning	Students ask the teacher and peers about word choice, rhyme, and imagery.	Students pose questions to ChatGPT about diction, rhyme, and imagery and discuss the answers with the teacher.
Learning community	Students discuss their drafts in groups.	Same as control; groups also compare their drafts with ChatGPT suggestions.
Modelling	The teacher presents example poems.	The teacher and ChatGPT provide example stanzas; students identify the diction and imagery used.
Reflection	Students revise their drafts based on teacher feedback.	Students revise their drafts based on teacher feedback and AI suggestions and explain which suggestions they accepted or rejected.
Authentic assessment	The teacher scores the poems with the rubric.	Same as control.

The research subjects were fifth-grade elementary school students in Bima City, West Nusa Tenggara, Indonesia. The selection was conducted using purposive sampling, taking into account initial ability equality, a relatively balanced number of students, and curriculum suitability. Inclusion criteria: fifth-grade students aged 10–11 years, fully present during the study, and obtaining permission from parents/guardians. Exclusion

https://doi.org/10.59923/joinme.v3i2.784
Vol. 3, No. 2, 2025

criteria: students with more than 20% absenteeism or requiring individual special learning services. A total of 55 students were analysed (experimental n = 28; control n = 27; Table 4).

Table 4. Classification of Research Participants

Class	Number of Students	Male	Female	Average Age	Treatment
Experimental	28	14	14	10.5	CTL–ChatGPT
Control	27	13	14	10.4	Conventional CTL

Research data were collected through a written pretest and posttest consisting of the performance tasks listed in Table 5. Data collection proceeded in three steps: (1) administering the pretest to both groups before the intervention; (2) implementing the four learning sessions; and (3) administering the posttest to both groups after the intervention (Figure 3). Students' responses were scored with an analytic rubric (1–5 per task).

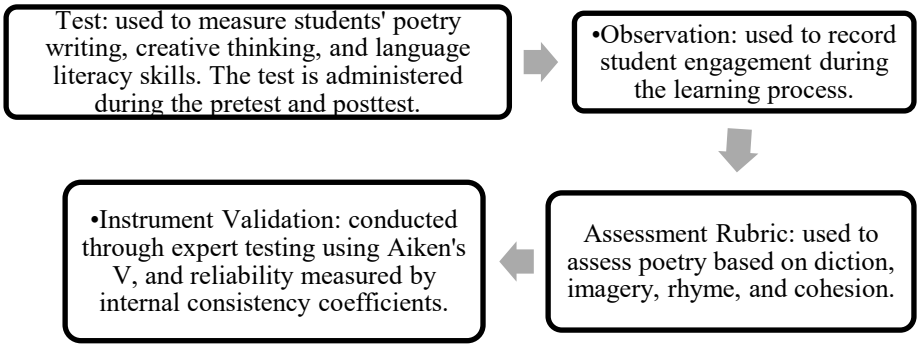


Figure 3. Data Collection Techniques and Procedures

The research instruments consisted of a poetry writing test, a creative thinking test, and a language literacy test. Each instrument was developed with a grid of indicators, questions, assessment scores, and cognitive domains (Table 5). Each of the seven indicators was measured by three tasks scored 1–5, giving a maximum of 15 points per indicator and 105 points in total, and the test was administered in written form. Creative thinking was measured through ideational fluency and originality, two dimensions of divergent thinking (Guilford, 1967; Torrance, 1974) that can be observed in short written products; flexibility and elaboration were not measured, which is acknowledged as a limitation. Language literacy was operationalised as vocabulary comprehension in poetic texts and literary appreciation.

Table 5. Research Instrument Specifications

Indicator Code	Indicator	Specific Description	Items (3 tasks per indicator; Bloom in parentheses)	Score
MP1	Diction	Selection of words aligned with meaning and aesthetics	1. Choose imaginative diction for a nature theme (C2). 2. Compose a stanza using rhyming words (C3). 3. Transform ordinary words into poetic diction (C4).	1–5
MP2	Imagery	Evoking sensory images	4. Depict a beach atmosphere in a stanza (C3). 5. Write a stanza featuring auditory imagery (C4). 6. Compose a poem conveying affective imagery (feelings) (C5).	1–5
MP3	Rhyme & Rhythm	Regular sound patterns	7. Create a stanza with an a–b–a–b rhyme scheme (C3). 8. Write a poem with a specified rhythmic pattern (C4). 9. Analyze rhyme in a classical poem (C5).	1–5
BK1	Ideational Fluency	Generating many creative ideas	10. List five distinct poem themes (C2). 11. Write a short poem on the theme “hope” (C3). 12. Compile a list of inspirational words for poetry (C2).	1–5

https://doi.org/10.59923/joinme.v3i2.784
Vol. 3, No. 2, 2025

Indicator Code	Indicator	Specific Description	Items (3 tasks per indicator; Bloom in parentheses)	Score
BK2	Originality of Ideas	Novel and distinctive concepts	13. Craft a rarely used poetic metaphor (C4). 14. Create a new comparison to express an emotion (C5). 15. Rework an old poem into a new style (C6).	1–5
LB1	Vocabulary	Understanding word meanings in poetry	16. Select the contextual meaning of words in a poem (C2). 17. Determine figurative meanings within a stanza (C3). 18. Analyze synonyms used in a poem (C4).	1–5
LB2	Literary Appreciation	Identifying moral messages and aesthetic value	19. Identify the moral message in a poem (C2). 20. Compare the aesthetic value of two different poems (C4). 21. Analyze the main theme in a folk poem (C5).	1–5

Description: Writing Poetry (MP); Creative Thinking (BK); Language Literacy (LB)

Instrument quality was examined with the Rasch model (Bond & Fox, 2015; Boone et al., 2014). All 21 items showed acceptable fit, with infit and outfit MNSQ values between 0.79 and 1.29 (acceptable range 0.5–1.5) and point-measure correlations between 0.52 and 0.79. Person reliability was 0.93 with a person separation index of 3.55, indicating that the test distinguished students' ability levels well. A few persons showed infit MNSQ values slightly above 1.5 (maximum 1.56), indicating minor unexpected response patterns. Differential item functioning (DIF) between girls and boys was examined; all reported DIF sizes were below 0.64 logits, the threshold for substantial DIF (Zwick, 2012), so no item was flagged. The detailed Rasch outputs are presented in the Appendix (Figures A1–A4).

Data analysis. The analysis proceeded in four stages (Table 6). First, instrument quality was examined through content validity (Aiken's V), reliability (Cronbach's α), and Rasch calibration. Second, the assumptions of normality (Shapiro–Wilk) and homogeneity of variance (Levene's test) were tested. Third, posttest differences between the experimental and control groups were tested with independent-samples t-tests. Fourth, improvement within each group was expressed as normalised gain, g = (posttest − pretest)/(maximum score − pretest) (Hake, 1998), and the difference in N-gain between groups was tested with Welch's t-test, with Cohen's d as the effect size. A sensitivity power analysis was conducted with G*Power (Faul et al., 2007).

Table 6. Data Analysis Techniques

Analysis Stage	Analytical Purpose	Statistical Test	Interpretation Criteria
Content Validity	Alignment of indicators with instrument items	Aiken's V	$V \geq 0.80$ = valid
Reliability	Internal consistency of the instrument	Cronbach's α	$\alpha \geq 0.70$ = reliable
Rasch calibration	Item and person fit, reliability, separation	Rasch model	Infit/outfit MNSQ 0.5–1.5; ZSTD ±2; reliability ≥ 0.80
Normality	Distribution of pretest and posttest scores	Shapiro–Wilk	$p > 0.05$ = normal
Homogeneity	Equality of variances between groups	Levene's test	$p > 0.05$ = homogeneous
Group Differences	Comparing experimental vs control scores	Independent t-test	$p < 0.05$ = significant
Effect size	Magnitude of group differences	Cohen's d	d = 0.2/0.5/0.8 = small/medium/large (Cohen, 1988)
Improvement	Effectiveness of the intervention	N-Gain; Welch's t-test	$g \geq 0.70$ high; $0.30 \leq g < 0.70$ moderate; $g < 0.30$ low (Hake, 1998)
Power	Minimum detectable effect	G*Power sensitivity analysis	$1 - \beta = .80$, $\alpha = .05$

https://doi.org/10.59923/joinme.v3i2.784
Vol. 3, No. 2, 2025

RESULTS AND DISCUSSION

Research Findings

Before the group comparisons, the assumptions were checked. Shapiro–Wilk tests on the total scores did not indicate significant departures from normality in either group at pretest or posttest (p = .269–.984; Appendix Table A1), and Levene's test did not indicate unequal variances (p = .736; Appendix Table A2), except for MP2 and LB2, for which the note to Table 8 indicates unequal variances.

Table 7. Description of Pretest and Posttest Research Results (Both Groups Combined, N = 55)

	Test	MP1	MP2	MP3	BK1	BK2	LB1	LB2
N	Posttest	55	55	55	55	55	55	55
	Pretest	55	55	55	55	55	55	55
Mean	Posttest	11.1	11.4	11.2	11.0	11.3	11.5	11.5
	Pretest	8.96	9.27	9.20	9.20	8.98	8.98	9.18
Median	Posttest	11	12	12	11	11	12	12
	Pretest	9	9	9	9	9	9	9
Standard deviation	Posttest	2.08	2.19	2.18	2.02	1.76	1.78	1.83
	Pretest	1.83	1.63	1.97	1.97	2.04	1.94	1.93
Minimum	Posttest	7	5	5	5	7	7	8
	Pretest	4	6	6	5	5	5	5
Maximum	Posttest	15	15	15	15	15	15	15
	Pretest	13	12	14	15	15	13	14

Table 7 presents the descriptive statistics of the pretest and posttest scores for all 55 students combined. Mean scores were descriptively higher at posttest than at pretest for all seven indicators: diction (MP1), imagery (MP2), rhyme and rhythm (MP3), ideational fluency (BK1), originality (BK2), vocabulary (LB1), and literary appreciation (LB2). Posttest means ranged from 11.0 to 11.5 (maximum 15) compared with pretest means of 8.96 to 9.27, and the medians rose from 9 to 11–12. Because these figures combine both groups, they describe the overall change after instruction but do not indicate whether CTL–ChatGPT was more effective than conventional CTL; the group comparison is reported below.

Table 8. Independent Samples T-Test

		Statistic	df	p
MP1	Student's t	1.599	108	0.113
MP2	Student's t	1.538[a]	108	0.127
MP3	Student's t	1.239	108	0.218
BK1	Student's t	0.914	108	0.363
BK2	Student's t	1.724	108	0.088
LB1	Student's t	1.120	108	0.265
LB2	Student's t	1.058[a]	108	0.292

Note. H_a: μExperimental ≠ μControl.

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Table 8 shows that the independent-samples t-tests found no statistically significant difference between the experimental and control groups on any of the seven indicators: diction (MP1, p = .113), imagery (MP2, p = .127), rhyme and rhythm (MP3, p = .218), ideational fluency (BK1, p = .363), originality (BK2, p = .088), vocabulary (LB1, p = .265), and literary appreciation (LB2, p = .292). Hypotheses H1–H3 were therefore not supported at the indicator level. These null results should be interpreted in light of statistical power: with 28 and 27 students, the study had a power of only about .44 to detect a medium effect (d = 0.5) at α = .05, and the smallest effect detectable with 80% power was d ≈ 0.77. The absence of significant differences therefore does not demonstrate that the two approaches are equivalent; the study could reliably detect only large effects.

Table 9. N-Gain Results by Group

Group	n	Mean N-gain (%)	SD	95% CI	Category
Experimental (CTL–ChatGPT)	28	55.10	30.31	[43.35, 66.85]	Moderate
Control (conventional CTL)	27	14.51	39.89	[−1.27, 30.29]	Low

Note. N-gain categories follow Hake (1998).

As shown in Table 9, the experimental group achieved a mean N-gain of 55.10% (SD = 30.31), which falls in the moderate category, whereas the control group achieved 14.51% (SD = 39.89), which falls in the low category; the 95% confidence interval for the control group (−1.27% to 30.29%) includes zero. Welch's t-test computed from these summary statistics showed that the difference in mean N-gain (40.59 percentage points, 95% CI [21.34, 59.84]) was statistically significant, $t(48.5) = 4.24$, $p < .001$, with a large effect size ($d = 1.15$, 95% CI [0.58, 1.72]). This result contrasts with the non-significant indicator-level posttest comparisons in Table 8 and must be interpreted with caution until the N-gain values have been rechecked.

Discussion

The results provide partial support for the use of ChatGPT within CTL for poetry writing. The experimental group showed a larger normalised gain than the control group, but the indicator-level posttest comparisons were not statistically significant, so the effect of CTL–ChatGPT on poetry-writing scores cannot yet be considered established. A plausible mechanism for the larger gain is that ChatGPT acted as a scaffold in the sense of Wood et al. (1976): it gave students immediate alternatives for diction, rhyme, and imagery during the inquiry, questioning, and reflection phases of CTL (Johnson, 2002), which the teacher alone could not provide to every student. This interpretation is consistent with Kangasharju et al. (2022), who found that an AI-based poetry tool supported lower secondary students' exploration of language, and with studies showing that ChatGPT can assist students' writing processes (Su et al., 2023; Yan, 2023).

The integration of ChatGPT did not show an advantage for creative thinking (ideational fluency and originality) or for language literacy (vocabulary and literary appreciation). Several explanations are plausible. First, the intervention lasted only four sessions, which may be too short to change broader abilities such as divergent thinking. Second, the creativity and literacy tasks were themselves poetry tasks, so the three constructs overlap and the test may not separate them well. Third, students may have relied on ChatGPT's suggestions instead of generating their own ideas; such cognitive offloading can reduce the effort that divergent thinking requires (Risko & Gilbert, 2016). Fourth, part of the gain in the experimental group may reflect a novelty effect of using a new technology. As Loock and Holt (2024) emphasise, the success of technology integration depends on the underlying pedagogical strategy rather than on the technology itself. CTL–ChatGPT therefore needs to be complemented with reflective techniques, open discussion, and guided problem solving that require students to evaluate and transform AI suggestions rather than adopt them.

The N-gain analysis showed a larger improvement in the experimental group than in the control group, with a moderate versus a low gain. This difference is consistent with the view that AI support can add individual feedback to contextual learning (Kasneci et al., 2023), but it should be interpreted cautiously because the N-gain values need to be rechecked and because the indicator-level posttest comparisons were not significant. The study did not measure engagement, motivation, or students' perceptions of the learning environment, so the claim that CTL–ChatGPT creates a more personalised or meaningful learning experience remains a hypothesis for future research.

However, the t-test results did not show significant differences between the experimental and control groups for all indicators measured. This phenomenon warrants further study, given that the descriptive results and gain scores show an upward trend in the experimental group. One possibility is the limited duration of the intervention or the uneven implementation of CTL-AI. In line with Diantama (2023), who discusses the conditions for using AI effectively in education, the duration and quality of implementation are likely to determine its real impact on student learning outcomes. Therefore, the sustainability of CTL-AI implementation needs to be designed in a continuous and more intensive learning cycle so that its effects can be manifested in a statistically significant manner. Low statistical power, as noted above, is a further explanation.

Limitations. This study has several limitations. First, it involved only two intact classes from one region ($n = 55$), which limits statistical power and generalisability. Second, the intervention lasted only four sessions and there was no delayed posttest, so the durability of the effects is unknown. Third, teacher effects and the novelty of using ChatGPT may have influenced the results. Fourth, rater bias in scoring the poems cannot be fully excluded. Fifth, ChatGPT's minimum age requirement of 13 years means that its use with younger students depends on teacher mediation and parental consent. Sixth, the extent to which students internalised rather than copied AI suggestions during the sessions could not be verified. Finally, the creativity and literacy measures were based on poetry tasks, which limits the discriminant validity of the three constructs.

Finally, the uneven results across the three outcomes suggest that the contribution of AI in contextual learning is not yet evenly distributed across all skill aspects. This opens up an important space for the development of hybrid learning designs that not only rely on interaction with technology, but also build classroom dynamics based on collaboration, critical discussion, and reflection. The concept of critical pedagogy (Freire, 1970) needs to be integrated into the CTL-AI approach so that students are not merely consumers of information from AI, but also active agents in creatively constructing meaning, reflecting, and evaluating knowledge, which also requires age-appropriate AI literacy (Long & Magerko, 2020). The implications of this research indicate that the adoption of artificial intelligence technology in education must always be placed within a humanistic, contextual, and transformative pedagogical framework in order to produce holistic changes in literacy, creativity, and 21st-century competencies.

The practical implications of this research point to the need for curriculum design and learning strategies that are adaptive to developments in educational technology. Teachers need to be trained to optimize the potential of AI as a partner in designing literacy tasks that trigger linguistic and aesthetic exploration, while still prioritizing human interaction, which is the foundation of transformative pedagogy. In addition, the development of dynamic and personalized AI-based learning resources presents a strategic opportunity to strengthen the achievement of 21st-century competencies in basic education.

CONCLUSION

This study examined the integration of ChatGPT into Contextual Teaching and Learning for fifth-grade students' poetry writing, creative thinking, and language literacy. Regarding the first objective, students taught with CTL–ChatGPT showed a larger normalised gain in poetry writing than students taught with conventional CTL, but the indicator-level posttest differences were not statistically significant, so the effect is promising but not yet established. Regarding the second and third objectives, no advantage of CTL–ChatGPT was found for creative thinking or language literacy.

Overall, ChatGPT can serve as a scaffold within CTL for poetry writing in primary school, provided that its use is mediated by the teacher and designed to make students evaluate and transform AI suggestions. Longer and adequately powered studies with ethical safeguards, delayed posttests, and separate measures of creativity and literacy are needed to confirm these findings.

REFERENCES

Bereczki, E. O., & Kárpáti, A. (2021). Technology-enhanced creativity: A multiple case study of digital technology-integration expert teachers' beliefs and practices. *Thinking Skills and Creativity, 39*, 100791. https://doi.org/10.1016/j.tsc.2021.100791

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Christopoulos, A., & Sprangers, P. (2021). Integration of educational technology during the Covid-19 pandemic: An analysis of teacher and student receptions. *Cogent Education, 8*(1), 1964690. https://doi.org/10.1080/2331186X.2021.1964690

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Diani, A. A., & Sukartono, S. (2022). Peran guru dalam penilaian autentik pada pembelajaran tematik di sekolah dasar [The role of teachers in authentic assessment in thematic learning in elementary schools]. *Jurnal Basicedu, 6*(3). https://doi.org/10.31004/basicedu.v6i3.2831

Diantama, S. (2023). Pemanfaatan artificial intelligent (AI) dalam dunia pendidikan [The use of artificial intelligence in education]. *DEWANTECH: Jurnal Teknologi Pendidikan, 1*(1). https://doi.org/10.61434/dewantech.v1i1.8

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Freire, P. (1970). *Pedagogy of the oppressed*. Herder and Herder.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

García-Peñalvo, F. J., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(4), 7–16. https://doi.org/10.9781/ijimai.2023.07.006

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*(4), 879–896. https://doi.org/10.1037/a0029185

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74. https://doi.org/10.1119/1.18809

Haryono, I., & Hikmah, K. (2023). The application of the contextual teaching and learning (CTL) model in Arabic language learning to improve the learning outcomes. *Buana Pendidikan: Jurnal Fakultas Keguruan dan Ilmu Pendidikan, 19*(1), 45–60. https://doi.org/10.36456/bp.vol19.no1.a6241

Hasudungan, A. N. (2022). Pembelajaran contextual teaching learning (CTL) pada masa pandemi COVID-19: Sebuah tinjauan [Contextual teaching and learning during the COVID-19 pandemic: A review]. *Jurnal DinamikA, 3*(2), 112–126. https://doi.org/10.18326/dinamika.v3i2.112-126

Hwang, G.-J., Zou, D., & Wu, Y.-X. (2023). Learning by storytelling and critiquing: A peer assessment-enhanced digital storytelling approach to promoting young students' information literacy, self-efficacy, and critical thinking awareness. *Educational Technology Research and Development, 71*(3), 1079–1103. https://doi.org/10.1007/s11423-022-10184-y

Johnson, E. B. (2002). *Contextual teaching and learning: What it is and why it's here to stay*. Corwin Press.

Kangasharju, A., Ilomäki, L., Lakkala, M., & Toom, A. (2022). Lower secondary students' poetry writing with the AI-based Poetry Machine. *Computers and Education: Artificial Intelligence, 3*, 100048. https://doi.org/10.1016/j.caeai.2022.100048

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

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* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Loock, R., & Holt, B. (2024). Augmented linguistic analysis skills: Machine translation and generative AI as pedagogical aids for analyzing complex English compounds. *Technology in Language Teaching & Learning, 6*(3).

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Ozkan, G., & Umdu Topsakal, U. (2021). Exploring the effectiveness of STEAM design processes on middle school students' creativity. *International Journal of Technology and Design Education, 31*(1), 95–116. https://doi.org/10.1007/s10798-019-09547-z

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Sætra, H. S. (2023). Generative AI: Here to stay, but for good? *Technology in Society, 75*, 102372. https://doi.org/10.1016/j.techsoc.2023.102372

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Torrance, E. P. (1974). *Torrance tests of creative thinking: Norms-technical manual*. Personnel Press.

Ulfah, A., Fitriyah, L., Zumaisaroh, N., & Jesica, E. (2023). Pemanfaatan media pembelajaran digital dalam pembelajaran menulis puisi di era Merdeka Belajar [The use of digital learning media in poetry writing instruction in the Merdeka Belajar era]. *GHANCARAN: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 5*(1). https://doi.org/10.19105/ghancaran.v5i1.7914

Veddayana, C., Romadhon, S., Aldresti, F., & Suyono, S. (2023). Rasionalitas implementasi Chat GPT dalam pembelajaran keterampilan menulis karya ilmiah [The rationality of implementing ChatGPT in learning scientific writing skills]. *GHANCARAN: Jurnal Pendidikan Bahasa dan Sastra Indonesia*. https://doi.org/10.19105/ghancaran.vi.11778

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report RR-12-08). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x