ISSN 2776-5229

Improving Tenth-Grade Students' Anecdote Writing Skills through a Generative AI-Assisted Think Pair Share Strategy

Ayuni Widia[1*], Anwar Efendi[2], and Ary Kristiyani[3]
[1,2,3]Universitas Negeri Yogyakarta, Indonesia
*ayuniwidia@gmail.com

ARTICLE INFO

Article history
Received September 22, 2026
Revised October 29, 2026
Accepted November 10, 2026

Keywords: Anecdote Text; Classroom Action Research; Generative AI; Think Pair Share; Writing Skills

ABSTRACT
Writing anecdote texts requires students to combine text structure, humour, and social criticism, yet many tenth-grade students produce incomplete texts with weak satirical force. This classroom action research aimed to describe the implementation of a *Think Pair Share* (TPS) strategy in which generative artificial intelligence (AI) served as a second layer of feedback after peer feedback, and to examine changes in students' anecdote writing skills. Following the Kemmis and McTaggart model, two cycles were conducted with 43 students of class X E at SMAN 20 Bandung, Indonesia, in August 2026. Writing was assessed with a weighted five-aspect analytic rubric scored independently by two raters (quadratic weighted $\kappa = 0.67$). Success was defined as a class mean of at least 75 and classical mastery of at least 75%, and gains were also expressed as normalised gain, categorised as low (< 0.3), medium (0.3–0.7), or high (≥ 0.7). The class mean rose from 60 (pre-cycle) to 67.5 (Cycle I) and 79 (Cycle II), and classical mastery from 27.9% to 41.9% and 76.7%, meeting both criteria (normalised gain = 0.48, medium; Wilcoxon $p < .001$), while student engagement rose from 65% to 85%. Gains were largest in spelling and language use and smallest in content and criticism, and five students declined because they wrote single-paragraph, purely comic texts. Students' writing skills therefore improved after the implementation of the combined strategy; however, because TPS and AI were introduced together without a comparison group, the improvement cannot be attributed to AI alone.

1. INTRODUCTION

Writing is the most complex of the language skills because it requires writers to manage several cognitive processes simultaneously, from generating and developing ideas and organising content into a logical structure to selecting appropriate diction, constructing effective sentences, and applying spelling and punctuation conventions. Because of this complexity, writing ability does not develop merely through assignments that are submitted and graded; it requires a recurring cycle of drafting, receiving feedback, and consciously revising. Graham (2019) argues that school-based writing instruction often fails to provide students with sufficient time, guidance, and feedback to go through this cycle, making a change in how writing is taught an urgent need. Empirically, a meta-analysis of 435 studies involving more than 61,000 participants found that feedback has a medium effect on learning ($d = 0.48$), and that the effect is larger when feedback carries specific information and targets cognitive skills (Wisniewski et al., 2020). These findings suggest that the central problem of writing instruction lies not in the presence or absence of feedback but in its quality, its specificity, and the way students process the feedback they receive. In the Indonesian language curriculum for Grade X, one of the writing competencies students must master is the anecdote text, a short narrative that combines humour with criticism or satire directed at behaviour, policy, or social phenomena. The text consists of five parts, namely abstract, orientation, crisis, reaction, and coda, each of which has a distinct rhetorical function in building the storyline while conveying a critical message to the reader (Rani et al., 2024). The distinctive feature of this genre is a dual demand rarely found in other

genres: students must be funny and critical at the same time within a relatively short text, so failure in either element causes the text to lose its identity as an anecdote. It is therefore unsurprising that recent studies report low anecdote writing ability among Grade X students, particularly in idea development, the use of humour and satire, and the coherence of text structure (Putra et al., 2026; Putri et al., 2026). Various media-based interventions, such as YouTube videos, comic strips, and instructional videos, have been tested to address this problem, but most of them focus on stimulating ideas at the beginning of the writing process and pay little attention to the revision stage, which in fact determines the final quality of the text (Putra et al., 2026; Putri et al., 2026; Rani et al., 2024).

These writing difficulties sit within a national literacy landscape that remains a cause for concern. The PISA 2025 results released in September 2026 show that Indonesian students scored 365 in reading, roughly unchanged from PISA 2022, and that only about 27% of students reached Level 2 or above, far below the OECD average of 69% (OECD, 2026). The same report notes that, compared with 2015, the proportion of Indonesian students performing below Level 2 in reading increased by 18 percentage points, indicating stagnation in text comprehension over the past decade (OECD, 2026). At the national level, the 2025 Education Report (*Rapor Pendidikan*) recorded that 72.23% of senior high school students reached the minimum literacy competency, a figure that still falls in the moderate category because it does not exceed the 75% threshold set for the good category (Kemendikdasmen, 2026). Although both data sets measure reading literacy rather than writing directly, they are relevant because the ability to produce meaningful texts rests on understanding text structure, purpose, and context; weak reading literacy is therefore almost always reflected in the quality of students' writing. Preliminary observation in class X E at SMAN 20 Bandung revealed a pattern consistent with these findings and national data. Students struggled to find story ideas containing social criticism, so many texts stopped at amusing stories without a clear target of criticism; the crisis and coda sections were often incomplete; humour was ineffective; and errors in sentence construction and spelling persisted. In terms of process, some students were not actively engaged and tended to rely on particular classmates during group work, so collaboration did not produce equitable learning for all members. The pre-cycle test yielded a mean score of 60 with classical mastery of 27.9%, meaning that only 12 of 43 students reached the Learning Objective Achievement Criterion (*Kriteria Ketercapaian Tujuan Pembelajaran*, KKTP), while the remaining 31 students still required intervention. This situation indicates a twofold problem, a product problem in the form of low text quality and a process problem in the form of uneven participation, so the chosen solution needed to address both simultaneously. Students' baseline writing ability was established through the pre-cycle anecdote writing test, which was administered in one meeting and scored by two raters using the analytic rubric described in the Methods section.

The *Think Pair Share* (TPS) strategy is relevant to the participation problem because its structure requires every student to think and write independently (*think*), test and refine ideas with a partner (*pair*), and then share the outcome with the whole class (*share*), so that no student can rely entirely on others. Empirical evidence from Indonesia consistently supports the effectiveness of TPS in writing instruction across genres and levels. TPS raised explanatory text writing scores from a mean of 63.45 to 86.94 and outperformed the *Picture and Picture* model (Hairira et al., 2025); it produced higher procedure text writing scores than lecture-based instruction while increasing student engagement and creativity (Josi, 2025); and, when combined with picture series, it yielded a post-test mean of 78.16 compared with 70.36 in the control class, a significant difference (Chifa & Andriyani, 2024). For anecdote texts specifically, a classroom action study by Iqlima (2020) reported an increase in the mean score from 74.55 in Cycle I to 78 in Cycle II. Theoretically, the strength of TPS lies in the *pair* stage, which is essentially a peer feedback mechanism; Huisman et al. (2019) showed in a meta-analysis that peer feedback produces substantially greater writing improvement than no feedback ($g = 0.91$), greater improvement than self-assessment ($g = 0.33$), and improvement

comparable to teacher feedback. Nevertheless, the *pair* stage has a structural limitation: the quality of feedback depends heavily on the partner's ability, so two students who both misunderstand the structure of an anecdote are likely to reinforce rather than correct each other's errors. It is at this point that generative artificial intelligence (AI) offers an opportunity as an additional source of feedback that is always available and can be directed toward specific criteria. The 2026 survey by the Indonesian Internet Service Providers Association (APJII) found that Generation Z is the age group most actively using AI (29.4%) and that 30.2% of AI use is directed toward education and research (APJII, 2026). Studies in Islamic education settings published in this journal also report high acceptance of AI writing tools among students, with behavioural intention as the strongest dimension (Paulijah & Siswana, 2026), and show that such acceptance is shaped by the interaction of technical, social, institutional, and ethical considerations (Nuraeni et al., 2026). Globally, a meta-analysis of 51 experimental studies found a large effect of ChatGPT on learning performance (g = 0.867) and moderate effects on learning perception (g = 0.456) and higher-order thinking (g = 0.457) (Wang & Fan, 2025), while another meta-analysis of 69 studies concluded that ChatGPT improves students' academic performance and affective-motivational states (Deng et al., 2025).

In the domain of writing specifically, ChatGPT used as a formative feedback tool has been shown to have a positive impact on academic writing skills, provided that students are trained to use it so that the feedback they obtain is relevant (Mahapatra, 2024). Su et al. (2023) mapped how ChatGPT can act both as a writing evaluator and as a "virtual peer" in the pre-writing, during-writing, and post-writing stages, including outlining, content revision, proofreading, and reflection. Escalante et al. (2023) found that learning outcomes with AI-generated feedback did not differ from those with human tutor feedback, while students' preferences were split almost evenly between the two. In Indonesian language classrooms, integrating ChatGPT into a writing workshop was reported to increase senior high school students' interest, motivation, and writing ability, particularly when combined with work-sharing activities that brought feedback from teachers and peers (Widiyarto et al., 2025). Taken together, these findings show that generative AI has genuine potential as a feedback provider, while also suggesting that its effectiveness is determined by the instructional design in which it is placed. This potential is accompanied by risks that cannot be ignored. Steiss et al. (2024) found that the quality of ChatGPT feedback on secondary students' essays was slightly below that of trained human raters on most feedback dimensions, although it remained useful for early drafts or when trained educators were unavailable. A more fundamental risk is over-reliance, the tendency to accept AI recommendations without questioning them, which, according to the systematic review by Zhai et al. (2024), weakens critical thinking, decision-making, and analytical reasoning as users favour cognitive shortcuts. Deng et al. (2025) also noted that ChatGPT use tends to reduce students' mental effort and has no significant effect on self-efficacy, while a study in this journal showed that AI use supports learning independence only when AI assists rather than replaces students' reasoning (Alfarabi et al., 2026). For this reason, UNESCO emphasises a human-centred approach that protects learner agency in every use of generative AI in education (Miao & Holmes, 2023), in line with reviews showing that the classroom opportunities offered by large language models are always paired with challenges of accuracy, ethics, and dependence (Kasneci et al., 2023; Lo, 2023).

These previous studies leave a specific gap. First, the TPS studies on writing in Indonesia (Chifa & Andriyani, 2024; Hairira et al., 2025; Iqlima, 2020; Josi, 2025) relied entirely on peers for feedback and offered no support to compensate for the limited quality of feedback from equally inexperienced partners. Second, studies of AI-generated writing feedback examined AI largely in isolation from peer interaction, compared it with human tutors or teachers, and were conducted mostly with university students writing academic or argumentative texts in English (Escalante et al., 2023; Mahapatra, 2024; Steiss et al., 2024; Su et al., 2023). Third, the only model combining peer feedback with AI used AI to improve the quality of the comments written by EFL university

reviewers (Guo et al., 2024), rather than as an additional source of feedback for the writer, and it has not been tested with secondary school students or with Indonesian-language genres. Fourth, recent studies on anecdote writing have tested media that stimulate ideas at the beginning of the writing process (Putra et al., 2026; Putri et al., 2026; Rani et al., 2024), but none has examined how feedback during revision supports the dual demand of humour and criticism. Accordingly, this study offers three contributions: (1) the explicit placement of generative AI in the *pair* stage as a second layer of feedback *after* peer feedback, rather than before or instead of it; (2) its application to a humour-and-criticism genre in Indonesian at the senior high school level; and (3) an *accept–reject log* that requires students to justify accepting or rejecting each AI suggestion as a means of preserving writer agency. This study aimed to describe the implementation of the generative AI-assisted TPS strategy in anecdote writing instruction and to analyse changes in the anecdote writing skills of class X E students at SMAN 20 Bandung following its implementation.

2. METHODS

This study employed classroom action research (CAR) based on the Kemmis and McTaggart model, which views action research as a spiral of planning, acting, observing, and reflecting conducted participatively by practitioners to improve their own practice (Kemmis et al., 2014). The model was chosen because the problem originated in concrete classroom practice, and its improvement required actions that could be adjusted gradually on the basis of evidence gathered in each cycle. A similar approach has been used in action research published in this journal, demonstrating that reflection between cycles enables targeted improvements to instructional media and strategies (Rafilah et al., 2026). The study was conducted in two cycles, each consisting of four interrelated stages, namely planning, acting, observing, and reflecting, with the results of the Cycle I reflection serving as the basis for designing the improved actions in Cycle II. The participants were all 43 students of class X E at SMAN 20 Bandung in the odd semester of the 2026/2027 academic year, comprising 21 male and 22 female students. This class was selected because preliminary observation and the pre-cycle test indicated low anecdote writing skills and uneven participation in group activities. The study was carried out in August 2026 in two cycles, with the Indonesian language teacher implementing the actions and the first author acting as an observer who recorded the course of instruction. Data were collected through five instruments designed to capture both product and process dimensions. An anecdote writing test was administered in the pre-cycle and at the end of each cycle to measure the quality of students' texts; in each test, students were required to write an anecdote of three to four paragraphs. Observation sheets of student and teacher activities were used to record student engagement in each TPS stage and the implementation of instructional steps. A student response questionnaire was used to explore students' perceptions of the instruction and of AI use, while field notes and documentation, including the accept–reject logs, were collected to document how students processed the feedback they received. The use of multiple instruments served as methodological triangulation so that interpretation of the results did not rest solely on test scores. Students' texts were scored using a five-aspect analytic rubric, as presented in Table 1.

Table 1. Anecdote writing assessment rubric

No.	Aspect	Indicators assessed	Weight
1	Structural completeness	Contains abstract, orientation, crisis, reaction, and coda in a coherent and connected sequence	25%
2	Content and criticism	Raises a relevant social phenomenon and conveys criticism or satire clearly	25%
3	Humour	Humour arises naturally, especially in the crisis and reaction, and supports the criticism	20%
4	Language use	Accurate diction, effective sentences, and use of the linguistic features of anecdote texts	15%
5	Spelling and punctuation	Conforms to the Indonesian Spelling Guidelines (EYD)	15%

Table 2. Scoring criteria for the anecdote writing rubric

Score	Category	Scoring criterion
4	Very good	Meets all indicators completely, accurately, and consistently
3	Good	Meets most indicators with minor shortcomings
2	Fair	Meets some indicators but still shows several shortcomings
1	Poor	Does not yet meet most of the specified indicators

The rubric was developed from the learning outcomes for anecdote texts in the Indonesian language curriculum and from the five structural elements of the genre, and its content validity was reviewed by the Indonesian language teacher of the class, who also served as the second rater. Each aspect was scored from 1 to 4 according to the criteria in Table 2. For each student, every aspect score was converted to a 0–100 scale (score/4 × 100), and a weighted score was calculated using the aspect weights in Table 1, so that structural completeness and content and criticism contributed most to the final score, at 25% each. Scoring was carried out by two raters, namely the first author and the Indonesian language teacher. Before scoring, both raters calibrated their understanding of the indicators and scoring criteria, and they then scored the texts independently using the same rubric. Each student's final score was the mean of the two raters' weighted scores. The aspect means reported in Table 5 were obtained by averaging the two raters' converted scores for each aspect and each student, and then averaging these values across the 43 students. Inter-rater agreement is reported in the Results section. This procedure was applied consistently in the pre-cycle, Cycle I, and Cycle II so that the measurements would capture the development of students' anecdote writing skills objectively and measurably. Quantitative data were analysed using descriptive statistics, namely the class mean, standard deviation, median, aspect-level mean scores, the percentage of classical mastery, and the percentage of student engagement. Classical mastery was calculated as the number of students reaching the KKTP divided by the total number of students multiplied by 100%. Because baseline scores differed across aspects, gains were compared using the normalised gain, g = (final score − initial score)/(100 − initial score), which represents the proportion of the available room for improvement that was closed; $g \geq 0.7$ is categorised as high, $0.3 \leq g < 0.7$ as medium, and $g < 0.3$ as low (Hake, 1998). Changes in individual scores between stages were tested with the Wilcoxon signed-rank test for paired data, which was chosen because the sample was small and the Cycle II score distribution was skewed. Inter-rater agreement was estimated from the 1–4 aspect scores using the percentage of exact agreement and quadratic weighted kappa (κ). Qualitative data from observations, questionnaires, field notes, and documentation were analysed through data reduction, data display, and conclusion drawing, and were then integrated with the quantitative data to explain why score changes occurred in each cycle. The action procedure followed the TPS stages integrated with generative AI, as summarised in Table 3. In the *think* stage, students selected a topic, drafted an

outline, and wrote a draft independently without AI assistance, so that the initial ideas were entirely their own. In the *pair* stage, students exchanged drafts and gave rubric-based feedback, then requested additional feedback from generative AI and evaluated both sources of feedback before revising their texts. Prompts were designed so that AI would offer suggestions without rewriting students' texts, in line with the view that prompt engineering determines the quality and appropriateness of large language model output in educational contexts (Cain, 2024).

In the *share* stage, several pairs read their revised texts aloud to obtain responses from the class and reinforcement from the teacher, giving students authentic reader responses to the humour and criticism they had written. The pre-cycle consisted of one meeting, in which baseline data were collected through the anecdote writing test, while Cycle I and Cycle II each consisted of three meetings of 2 × 45 minutes. In each cycle, the first meeting was devoted to an introduction to the task and the *think* stage, in which students drafted their anecdotes independently; the second meeting was devoted to the *pair* stage, in which students exchanged peer feedback, requested AI feedback, and revised their texts; and the third meeting was devoted to the *share* stage and the end-of-cycle writing test. In Cycle II, the time allocation for each stage was enforced more strictly than in Cycle I (Table 3). Two generative AI tools were used, namely ChatGPT (OpenAI) and Claude (Anthropic), in the versions publicly available in August 2026; students accessed them on their own smartphones, and each pair chose one of the two tools. These tools were chosen because they can help students explore ideas and organise the outline of an anecdote text; in this study, however, their use was deliberately restricted to providing feedback in the *pair* stage, after peer feedback, so that idea generation in the *think* stage remained the students' own work. Both tools were used with the same prompts provided by the teacher, which are reported in full below Table 3.

Table 3. Integration of generative AI into the TPS stages and Cycle II improvements

Stage	Student activities	Role of AI	Cycle II improvement
Think	Choose a topic, outline, and write a draft independently	Not used	Fixed time allocation
Pair	Exchange drafts, give rubric-based feedback, evaluate peer and AI suggestions, then revise	Second layer of feedback after peers; not allowed to rewrite the text	Genre-specific prompts; accept–reject log with reasons
Share	Read revised texts aloud; respond to classmates' work	Not used	Presentation time limited per pair

Note. Prompts used by students (translated from Indonesian). Cycle I: "Give feedback on the structural completeness, humour, and language use of the following anecdote text. Do not rewrite the text. Provide suggestions for improvement that the writer can consider." Cycle II: "Give feedback on the following anecdote text with attention to the characteristics of the anecdote genre, namely structural completeness (abstract, orientation, crisis, reaction, and coda), the relevance of the social phenomenon and the sharpness of the criticism or satire, and the humour that emerges through jokes, irony, or surprise in the crisis and reaction. Also evaluate the accuracy of diction, sentence effectiveness, the use of linguistic features, and spelling and punctuation according to EYD Fifth Edition. Present the evaluation by assessed aspect, the parts that need improvement, and specific suggestions for improvement. Do not rewrite, replace, or create a new anecdote text. Provide guiding feedback so that the writer can decide which suggestions to accept, reject, or modify, with reasons." Success was determined by two outcome criteria: (1) a class mean of at least 75 and (2) classical mastery of at least 75% of students reaching the KKTP of 75. In addition to the outcome criteria, the quality of the learning process was observed through students' engagement in each TPS stage and their involvement in evaluating feedback critically, as recorded in the observation sheets and field notes. These process indicators are important because score gains unaccompanied by critical

engagement risk reflecting the quality of AI output rather than an improvement in students' own skills.

3. RESULTS AND DISCUSSION

RESULTS

Pre-cycle. The initial test produced a mean score of 60 with classical mastery of 27.9%, meaning that only 12 of 43 students reached the KKTP. Analysis of students' texts revealed three dominant weaknesses: crisis sections lacking humour, which made the stories unengaging; coda sections that did not convey a clear message or conclusion, so the criticism failed to reach the reader; and errors in sentence construction and spelling. Some students were also unable to develop their ideas into a complete story containing social criticism, so their texts tended to be fragments of amusing events without a definite target of criticism. This pattern of weaknesses indicates that students' difficulties operate on two layers: a structural and linguistic layer that can be taught relatively explicitly, and a humour-and-criticism layer that demands social sensitivity and creativity.

Cycle I. Instruction in Cycle I was implemented through the think, pair, and share stages, with generative AI serving as an additional source of feedback after peer feedback. The class mean increased to 67.5 with classical mastery of 41.9% (18 students), while student engagement based on the observation sheets reached 65%. Although these results improved on the pre-cycle, they did not meet either success criterion. Reflection on observation data, field notes, and documentation identified three main obstacles: (1) some students copied AI suggestions without assessing their fit with the content and intent of their own texts; (2) the prompts were too general, so the AI feedback did not address the specific requirements of anecdote texts, especially humour and criticism; and (3) the pair stage took longer than planned because students had to process two sources of feedback at once. These obstacles show that the presence of AI without adequate structuring does not automatically lead to meaningful revision.

Cycle II. Based on the reflection, the actions in Cycle II were improved in three ways: providing genre-specific prompts that directed AI toward the structure, humour, criticism, and language features of anecdote texts; requiring an accept–reject log of AI suggestions with reasons; and enforcing a tighter time allocation for each stage (Table 3). Following these improvements, the class mean reached 79 with classical mastery of 76.7% (33 students), while student engagement rose from 65% to 85%. This 20-percentage-point rise in engagement indicates that the requirement to weigh and justify each AI suggestion did not reduce participation; on the contrary, students' involvement in pair discussion increased, possibly because every revision decision had to be discussed with and justified to a partner. Both success criteria were therefore exceeded. A comparison of outcomes across stages is presented in Table 4 and Figure 1.

Table 4. Comparison of anecdote writing outcomes

Indicator	Pre-cycle	Cycle I	Cycle II	Total change
Class mean	60	67.5	79	+19.0
Standard deviation	13.4	11.2	19.7	–
Median	65.0	70.0	88.75	+23.75
Range (min–max)	46.25–77.5	55–80	30–88.75	–
Students reaching mastery (n = 43)	12	18	33	+21
Classical mastery (%)	27.9	41.9	76.7	+48.8 pp
Student engagement (%)	–	65	85	–
Success criterion (≥ 75)	Not met	Not met	Met	–

Note. The number of students reaching mastery was derived from the mastery percentage of 43 students; pp = percentage points.

ISSN 2776-5229

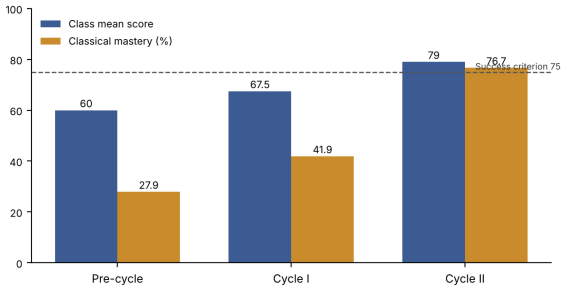


Figure 1. Class mean score and classical mastery at each stage

Table 4 shows that the gain from Cycle I to Cycle II, 11.5 points in the class mean and 34.8 percentage points in classical mastery, was considerably larger than the gain from the pre-cycle to Cycle I, which amounted to only 7.5 points and 14.0 percentage points. When these gains are compared with the available room for improvement, Cycle I closed 18.8% of the initial gap (7.5 of 40 points), whereas Cycle II closed 35.4% of the remaining gap (11.5 of 32.5 points), indicating that the effectiveness of the Cycle II actions was nearly twice that of Cycle I. Overall, the class mean rose by 19 points, or 31.7% above the baseline, and the number of students reaching mastery increased from 12 to 33. However, 10 students (23.3%) had still not reached the KKTP at the end of Cycle II, so the improvement was not evenly distributed and calls for follow-up through more individualised support. This non-linear pattern of gains indicates that the decisive factor was not the mere presence of AI, since AI had been used from Cycle I onward, but the improved design of its use in Cycle II. At the individual level, 37 of the 43 students (86.0%) obtained higher scores in Cycle II than in the pre-cycle, and the median rose from 65.0 to 70.0 and 88.75 (Table 4). The Wilcoxon signed-rank test indicated that the increases from the pre-cycle to Cycle I (p = .008), from Cycle I to Cycle II (p = .002), and from the pre-cycle to Cycle II (p < .001) were statistically significant. However, the spread of scores widened in Cycle II (SD = 19.7, compared with 13.4 in the pre-cycle) because five students scored 30, having received a score of 1 from both raters on almost every aspect; three of them had scored between 65 and 78.75 in an earlier stage. Excluding these five students, the Cycle II mean would be 85.4, and all 33 students who reached mastery belonged to the remaining group of 38. Analysis of their texts showed that these five students wrote their anecdotes as a single paragraph, whereas the task required three to four paragraphs, used linguistic features with very little variation, and included only comic elements without the social criticism that the anecdote genre requires; as a result, the structural elements of the anecdote were not distinguished and the texts met almost none of the rubric indicators. This is reflected in their scores, in which humour received an average score of 2, whereas content and criticism received 1 from both raters. This pattern indicates that, for some students, peer and AI feedback in the TPS cycle did not compensate for difficulties in basic text organisation, which therefore require individual guidance from the teacher. Inter-rater agreement was acceptable overall, with 76.4% exact agreement and a quadratic weighted κ of 0.67 (pre-cycle 0.70, Cycle I 0.52, Cycle II 0.69).

Table 5. Mean scores by assessment aspect at each stage and normalised gain

Aspect	Pre-cycle	Cycle I	Cycle II	Gain	g
Structural completeness	62.2	56.4	83.1	+20.9	0.55
Content and criticism	52.3	83.1	65.7	+13.4	0.28
Humour	58.7	52.3	73.3	+14.5	0.35
Language use	68.0	75.0	90.1	+22.1	0.69
Spelling and punctuation	62.8	72.7	90.7	+27.9	0.75

Note. Scores on a 0–100 scale (mean of both raters). Gain = Cycle II score − pre-cycle score, calculated from unrounded means; *g* = normalised gain from the pre-cycle to Cycle II. The weighted mean of the aspect scores equals the class mean in Table 4. Table 5 shows that the five aspects did not improve uniformly. Language use and spelling and punctuation rose steadily in both cycles and reached the highest normalised gains (0.69 and 0.75), with spelling and punctuation falling in the high category. Structural completeness declined slightly in Cycle I (from 62.2 to 56.4) and then rose sharply in Cycle II to 83.1 ($g = 0.55$). Humour followed a similar pattern, decreasing from 58.7 to 52.3 in Cycle I before rising to 73.3 in Cycle II ($g = 0.35$). Content and criticism showed the opposite pattern: it was the weakest aspect in the pre-cycle (52.3), rose sharply to 83.1 in Cycle I, and then fell to 65.7 in Cycle II, yielding the lowest normalised gain (0.28, low category). This decline persisted when the five students who scored 30 were excluded (from 82.2 to 71.1), indicating that it was not caused solely by those students. At the end of Cycle II, the two aspects that define the anecdote genre, content and criticism (65.7) and humour (73.3), were the only aspects below 75, whereas the more rule-based aspects of structure, language use, and spelling and punctuation exceeded 80. The qualitative data from the observation sheets, field notes, and students' texts help explain these score changes. In the pre-cycle, students' texts typically lacked humour in the crisis section, ended with a coda that did not convey a clear message, and contained errors in sentence construction and spelling. In Cycle I, the field notes recorded that some students copied AI suggestions without checking whether they fitted the content and intent of their texts, and that the general prompts produced feedback that rarely addressed humour and criticism. In Cycle II, after the genre-specific prompts and the accept–reject log were introduced, student engagement in the TPS stages rose from 65% to 85%, indicating more active participation in pair discussion and revision. The texts of the five students who scored 30 showed the opposite pattern: they were written as a single paragraph, contained only comic elements without social criticism, and showed little linguistic variation. The accept–reject logs and questionnaire responses were not analysed systematically, so the qualitative evidence in this study relies mainly on the observation sheets, field notes, and students' texts.

DISCUSSION

The improvement in anecdote writing skills in this study can be explained through the complementary pedagogical functions of each TPS stage. The think stage ensures that every student drafts on the basis of their own ideas before receiving input from others, thereby maintaining the individual accountability that underpins cooperative learning and preventing students from depending on peers. This sequence is also significant in the context of AI use, because the creativity of humour and sensitivity to social issues at the core of anecdote texts should not be delegated to a machine. By writing first without AI, ownership of ideas remains with the student, in line with UNESCO's recommendation that the use of generative AI in education protect human agency and not replace learners' thinking (Miao & Holmes, 2023). The pre-cycle finding of weak idea development underscores the importance of this stage as a space for practising independent thinking before external support is provided. The pair stage constitutes the core of the intervention because it is where students receive two layers of feedback with different characteristics. The first layer comes from peers, who read the text as fellow readers and can judge whether its humour and criticism land with readers of their age, while the second layer comes from AI, which offers more systematic suggestions about structural completeness, diction, and language conventions. This combination can be understood through the concept of the zone of proximal development, according to which learners' abilities develop beyond their independent level when they receive assistance from more capable others (Vygotsky, 1978). In this context, peers and AI function as complementary scaffolds, since peer feedback has a strong effect on writing improvement (Huisman et al., 2019), and informative,

specific feedback amplifies learning effects (Wisniewski et al., 2020). AI fills the gap when partners are unable to provide sharp input, a pattern consistent with the finding that AI-supported peer feedback improves both feedback quality and the writing ability of students who provide feedback (Guo et al., 2024).

The most important finding of this study is the marked difference in outcomes between Cycle I and Cycle II, even though AI was used in both cycles. In Cycle I, gains were limited because some students copied AI suggestions without deliberation, a phenomenon described in the literature as over-reliance, namely the tendency to accept AI recommendations uncritically because users prefer cognitive shortcuts (Zhai et al., 2024). This symptom is also consistent with meta-analytic evidence that ChatGPT use tends to reduce students' mental effort (Deng et al., 2025), resulting in superficial revisions that did not always match the writer's intent. The jump in Cycle II emerged after two design changes, genre-specific prompts and the accept–reject log. Specific prompts made AI feedback relevant to the demands of anecdote texts, consistent with the view that the quality of large language model output in education is largely determined by prompt design (Cain, 2024), while the accept–reject log was designed to make students weigh each suggestion before revising. This second mechanism shifted students from passive recipients to decision-makers over their own writing, returning to them the cognitive load that had previously been offloaded to AI, in line with the finding that AI supports learning independence when it assists rather than replaces reasoning (Alfarabi et al., 2026). The engagement data reinforce this interpretation, as student engagement rose from 65% to 85% after the accept–reject log was introduced; the demand to weigh AI suggestions appears not to have burdened students but rather to have encouraged more active involvement in pair discussion. However, because the genre-specific prompts, the accept–reject log, and the tighter time allocation were introduced simultaneously in Cycle II, the separate contribution of each component cannot be determined, and this interpretation should therefore be regarded as tentative. The aspect-level data allow the role of AI to be viewed more proportionally. The largest and most consistent gains occurred in language use and spelling and punctuation, aspects whose criteria are explicit and can be checked against rules, which is consistent with evidence that ChatGPT feedback is relatively strong on criteria-based dimensions even though its overall quality remains slightly below that of trained human raters (Steiss et al., 2024). Structural completeness and humour declined in Cycle I and rose markedly only in Cycle II, when the prompt explicitly named the five structural elements and asked AI to evaluate jokes, irony, and surprise in the crisis and reaction; this timing supports the view that the specificity of the prompt shapes the usefulness of AI output (Cain, 2024). In contrast, content and criticism, which require students to identify a social issue and sharpen their satire, improved in Cycle I but fell back in Cycle II and ended as the weakest aspect. One plausible explanation is that the more detailed Cycle II prompt, with its many structural and linguistic criteria, drew students' revision effort toward features that are easier to correct, so that the critical substance of their texts received less attention; this explanation, however, needs to be verified through analysis of the students' texts and accept–reject logs. These findings suggest that AI is more helpful for the conventions of the genre than for its critical core. Consequently, the *share* stage, where students receive authentic reader responses to their humour and satire, retains an irreplaceable role (Widiyarto et al., 2025), and the effect of AI feedback should be interpreted together with students' own judgement of that feedback, given that students' preferences for AI and human feedback are divided (Escalante et al., 2023).

Humour deserves particular attention because it remained below 75 at the end of Cycle II even though it rose by more than 20 points from Cycle I to Cycle II. Unlike spelling or language use, humour in an anecdote is not merely a linguistic feature; it depends on creativity, sensitivity to social situations, and shared cultural knowledge that allows a reader to recognise irony and satire. Generative AI can point out that a crisis section lacks surprise or irony, but it cannot readily supply

humour that resonates with students' own social and cultural context, and research on how conversational AI handles humour across cultures is still limited (Zhai & Wibowo, 2022). The five students who wrote purely comic texts without criticism further show that humour and criticism can become decoupled: students may produce jokes without directing them at a social target, which is precisely what turns a funny story into an anecdote. Developing humour in this genre therefore requires more than revision prompts; it calls for modelling of satirical humour, discussion of social issues that are meaningful to students, and authentic reader responses in the *share* stage, so that humour serves as a vehicle for criticism rather than as an end in itself. Compared with similar studies, the magnitude of the gain in this study, 19 points, falls within the range of other anecdote writing interventions, such as the use of comic strips, which raised the mean from 70 to 86 (Putra et al., 2026), and is larger than that of TPS without AI in the same genre, which raised the mean from 74.55 to 78 (Iqlima, 2020). These comparisons should be read with caution because of differences in design, sample size, baseline conditions, and assessment instruments; for example, the lower baseline in this study left more room for improvement. Duration should also be considered, since the meta-analysis by Wang & Fan (2025) recommends sustained ChatGPT use over four to eight weeks for stable effects and shows that its effect on higher-order thinking is strongest when ChatGPT acts as an intelligent tutor. Thus, the two cycles in this study probably captured only the early effects of AI use, and its influence on students' ability to evaluate and revise their writing independently should be examined over a longer period.

Theoretically, this study extends understanding of AI integration in writing instruction by suggesting that the position of AI within the instructional sequence and the specificity of prompts matter as much as the presence of AI itself, and that AI support is uneven across the dimensions of a genre. Practically, the findings imply that teachers should place AI after students have written independently and received peer feedback, accompanied by genre-specific prompts and a requirement to justify accepting or rejecting suggestions, while giving explicit, teacher-led attention to the critical content of anecdotes, which AI supported least. Nevertheless, this study has several limitations. First, the single-class design without a comparison group does not allow the contributions of TPS and AI to be separated causally. Second, part of the improvement may reflect practice effects and growing familiarity with the scoring rubric. Third, although inter-rater agreement was acceptable overall, it was lower in Cycle I ($\kappa = 0.52$), and every disagreement between the raters amounted to two points on the four-point scale, so the scoring guidelines require further calibration. Fourth, the drop in content and criticism in Cycle II cannot yet be explained by the available data, and the reasons why five students failed to meet the paragraph requirement only in Cycle II were not investigated. Fifth, the content validity of the rubric was judged only by the class teacher, who also acted as a rater, without review by independent experts. Sixth, the accept–reject logs and questionnaire responses were not analysed systematically, so the study cannot report how many AI suggestions students accepted, rejected, or modified. These limitations also point to directions for future research testing the model with stronger designs.

4. CONCLUSION

Following the implementation of the *Think Pair Share* strategy with generative AI as a second layer of feedback, the anecdote writing skills of class X E students at SMAN 20 Bandung improved: the class mean rose from 60 in the pre-cycle to 67.5 in Cycle I and 79 in Cycle II, classical mastery rose from 27.9% to 41.9% and 76.7%, and student engagement rose from 65% to 85%, meeting both success criteria with a medium normalised gain (0.48). The improvement was uneven: language use and spelling and punctuation improved most, structure and humour rose mainly in Cycle II, content and criticism ended as the weakest aspect, and five students declined

because they wrote single-paragraph, purely comic texts. Because TPS and AI were implemented as a single intervention without a comparison group, these results demonstrate the success of the combined strategy but do not establish a causal contribution of AI itself. The most important lesson is that the successful use of AI in writing instruction depends not merely on the availability of the technology but on the pedagogical design in which it is placed, the quality of the prompts, and students' active involvement in evaluating the suggestions they receive. Teachers are advised to provide genre-specific prompts, to require students to justify accepting or rejecting AI suggestions, to give direct guidance on developing social criticism and satirical humour, and to support students who struggle with paragraph development, while schools are advised to strengthen digital literacy and the ethics of AI use. Future research should compare TPS with and without AI in a quasi-experimental design, examine the separate effects of prompt specificity and the accept–reject log, and test the model with other genres and over a longer duration.

5. REFERENCES

Alfarabi, A., Umami, A., & Prameswari, P. W. (2026). The relationship between the utilization of artificial intelligence technology and the learning independence of STIKes Muhammadiyah Bojonegoro students. *Edusoshum: Journal of Islamic Education and Social Humanities*, *6*(3), 2232–2241. https://doi.org/10.52366/edusoshum.v6i3.511

Asosiasi Penyelenggara Jasa Internet Indonesia [APJII]. (2026). *Survei penetrasi internet dan perilaku penggunaan internet 2026*. https://survei.apjii.or.id/survei

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, *68*(1), 47–57. https://doi.org/10.1007/s11528-023-00896-0

Chifa, L. N., & Andriyani, S. (2024). The effectiveness of Think-Pair-Share method combined with picture series to improve students' writing skill. *PROJECT (Professional Journal of English Education)*, *7*(4), 925–931. https://journal.ikipsiliwangi.ac.id/index.php/project/article/view/22843

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*, 57. https://doi.org/10.1186/s41239-023-00425-2

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732X18821125

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, *63*, 100962. https://doi.org/10.1016/j.iheduc.2024.100962

Hairira, O. D., Sobri, S., & Devi, A. A. K. (2025). Pengaruh penggunaan model pembelajaran Think Pair Share dan Picture and Picture terhadap keterampilan menulis teks eksplanasi siswa kelas XI SMAN 1 Ciruas tahun pelajaran 2023/2024. *Jurnal Ilmiah Wahana Pendidikan*, *11*(5.D), 261–271. https://www.jurnal.peneliti.net/index.php/JIWP/article/view/10449

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

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education*, *44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Iqlima. (2020). Peningkatan keterampilan menulis teks anekdot menggunakan strategi pembelajaran Think Pair Share pada siswa kelas X. *Asas: Jurnal Sastra*, *9*(1). https://doi.org/10.24114/ajs.v9i1.18345

Josi, G. G. (2025). Pengaruh model pembelajaran Think Pair Share terhadap kemampuan menulis teks prosedur siswa kelas VII SMP Negeri 2 Bojonggede tahun ajaran 2023/2024. *Jurnal Pendidikan Tambusai*, *9*(1), 9567–9575. https://doi.org/10.31004/jptam.v9i1.26108

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

Kementerian Pendidikan Dasar dan Menengah [Kemendikdasmen]. (2026). *Rapor Pendidikan Indonesia 2025.* *https*://data.kemendikdasmen.go.id/publikasi/p/rapor-pendidikan-indonesia/rapor-pendidikan-indonesia-2025-3

Kemmis, S., McTaggart, R., & Nixon, R. (2014). *The action research planner: Doing critical participatory action research*. Springer. https://doi.org/10.1007/978-981-4560-67-2

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11, 9.* https://doi.org/10.1186/s40561-024-00295-9

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Nuraeni, I. I., Furwanti, E., Sakuroh, L., Hendriawan, E., Marlina, R., & Adiningsih, N. U. (2026). Student acceptance of generative artificial intelligence in Islamic higher education: An extended UTAUT model. *Edusoshum: Journal of Islamic Education and Social Humanities*, *6*(3), 2099–2109. https://doi.org/10.52366/edusoshum.v6i3.572

OECD. (2026). *PISA 2025 results (Volume I): Country notes—Indonesia*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2025-results-volume-i-country-notes_2d4ff9ea-en/indonesia_9c880f8a-en.html

Paulijah, R., & Siswana, S. (2026). EFL students' acceptance of AI tools in English writing: A technology acceptance model perspective. *Edusoshum: Journal of Islamic Education and Social Humanities*, *6*(3), 1667–1687. https://doi.org/10.52366/edusoshum.v6i3.490

Putra, B. A., Suyuthi, H., & Yuniati, I. (2026). Penerapan media komik strip terhadap kemampuan menulis teks anekdot di kelas X SMA Negeri 9 Kaur. *Lateralisasi*, *14*(1), 766–775. https://doi.org/10.36085/lateralisasi.v14i1.10275

Putri, S. D., Marni, S., & Tiawati, R. L. (2026). Pengaruh model Discovery Learning berbantuan video terhadap keterampilan menulis teks anekdot siswa kelas X SMA Negeri 1 Lembah Gumanti. *Belajar Bahasa: Jurnal Ilmiah Program Studi Pendidikan Bahasa dan Sastra Indonesia*, *11*(2), 335–347. https://doi.org/10.32528/bb.v11i2.6532

Rafilah, N. H., Rahmat, M., & Budiyanti, N. (2026). Implementation of culturally responsive teaching in private Islamic education learning through action research. *Edusoshum: Journal of Islamic Education and Social Humanities*, *6*(3), 2187–2198. https://doi.org/10.52366/edusoshum.v6i3.591

Rani, E. I., Pitoyo, A., & Puspitoningrum, E. (2024). Pengaruh media video YouTube "Lapor Pak" terhadap keterampilan menulis teks anekdot pada siswa kelas X SMAN 6 Kediri tahun pelajaran 2023/2024. *Efektor*, *11*(1), 154–163. https://doi.org/10.29407/e.v11i1.22957

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

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

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

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*, 621. https://doi.org/10.1057/s41599-025-04787-y

Widiyarto, S., Ati, A. P., Setyowati, L., Lutvaidah, U., & Astuti, T. (2025). Implementasi ChatGPT dan Writing Workshop untuk meningkatkan kualitas tulisan siswa di era digital. *Jurnal Warta Dharmawangsa*, *19*(2), 999–1011. https://doi.org/10.46576/wdw.v19i2.6361

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhai, C., & Wibowo, S. (2022). A systematic review on cross-culture, humor and empathy dimensions in conversational chatbots: The case of second language acquisition. *Heliyon, 8*(12), e12056. https://doi.org/10.1016/j.heliyon.2022.e12056

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*, 28. https://doi.org/10.1186/s40561-024-00316-7