Boundary Conditions for LLM-Generated Feedback in Primary Writing: An Educator-Aligned Evaluation and Design Considerations

Dan Zhang *, Thuong Hoang *, Ye Zhu, Rui Wang, Paula Crouch, Yi Wang

Posted Date: 25 May 2026

doi: 10.20944/preprints202605.1545.v1

Keywords: Large Language Models (LLMs); automated writing feedback; primary education; humanAI collaboration; educator-centered evaluation

 doi:10.20944/preprints202605.1545.v1

Article

Boundary Conditions for LLM-Generated Feedback in Primary Writing: An Educator-Aligned Evaluation and Design Considerations

Dan Zhang [1,*], Thuong Hoang [1], Ye Zhu [1], Rui Wang [2], Paula Crouch [3] and Yi Wang [1]

[1] Faculty of Science, Engineering and Built Environment, School of Information Technology, Deakin University, Burwood, VIC 3125, Australia
[2] CSIRO,Research Way, Clayton VIC 3168, Australia
[3] Kinetic Education,506 Nepean Hwy, Frankston VIC 3199, Australia
* Correspondence: s223275994@deakin.edu.au

Abstract

Generative large language models (LLMs) are increasingly used to support writing feedback. However, the pedagogical safety and usefulness of LLM feedback for primary students remains under-evaluated. This study reports an educator-centered evaluation of GPT-4 Turbo for Year 5 narrative and persuasive writing in the context of an established online tutoring program. Using authentic students' drafts paired with tutor feedback, we generated parallel LLM feedback via rubric-aligned prompting and compared the two feedback sources in a blinded, within-script design. Four experienced English specialists co-designed a six-dimensional rubric (clarity, specificity, helpfulness, feasibility, relevance, and overall effectiveness) and rated tutor versus LLM feedback for each script; their written reflections were analyzed thematically to surface boundary conditions and risk perceptions. Across dimensions, tutor feedback was rated higher overall, with the clearest advantage in perceived helpfulness and specificity. LLM feedback was often rated similarly for clarity and feasibility but was frequently characterized as generic, surface-focused, and occasionally misaligned with the student draft, which increased verification effort and posed a risk of misleading learners if used without mediation. Synthesizing ratings and educator reflections, we identify conditions under which LLM feedback is most appropriate as rapid first-pass support for routine structure and surface revision, and least appropriate for developmental judgment and context-sensitive guidance. We translate these findings into design requirements for teacher-in-the-loop primary writing feedback systems, including alignment to explicit pedagogical constructs, editable workflows, and safeguards that reduce unsupported feedback before release to students.

Keywords: Large Language Models (LLMs); automated writing feedback; primary education; human-AI collaboration; educator-centered evaluation

1. Introduction

In primary and secondary school education, developing writing skills serves as a foundation for communication, critical thinking, and academic success [1]. High-quality feedback supports this development by helping learners improve ideas, organization, language choices, and argumentation [2–5]. However, providing timely and personalized feedback remains a persistent challenge for educators, particularly in classrooms with large student-to-teacher ratios. In practice, teachers often struggle to provide accurate and individualized responses, leading to generic or delayed feedback that can stall the learning process[6,7].

Digital writing-support technologies have therefore attracted sustained interest as a means to scale formative feedback. Recent advances in generative large language models (LLMs) have intensified this interest because LLMs can generate readable, structured feedback quickly and can be embedded

 doi:10.20944/preprints202605.1545.v1

into technology-mediated learning environments for drafting support, revision prompts, and feedback generation [8–11]. However, emerging evidence suggests that surface quality (e.g., fluent wording) does not guarantee pedagogical quality or safety: models may produce feedback that is overly general, insufficiently grounded in the student text, or unreliable for evaluative judgment, particularly when guidance requires developmental interpretation or contextual sensitivity [12]. These concerns shift the central question from whether LLMs *can* generate feedback to when and how LLM feedback can be integrated responsibly into authentic instructional workflows.

A further challenge is that "feedback quality" is inherently multi-dimensional. In higher education, learner-centered feedback frameworks distinguish how feedback supports learners' sensemaking, future improvement, and agency, and large-scale analyses of authentic feedback show that these components co-occur and interact rather than appearing in isolation [13]. In parallel, relational feedback, how feedback conveys respect, encouragement, and an invitation to engage, is increasingly recognized as influential for uptake, yet it is difficult to operationalize and has rarely been evaluated systematically in automated feedback pipelines [14]. For primary learners, multi-dimensional evaluation is especially important because writing development depends not only on correct surface features but also on age-appropriate guidance that scaffolds planning, idea development, and genre-specific structure.

Responsible use of LLM feedback also depends on workflow design. LLM outputs are sensitive to prompt design and workflow structure, and recent work shows that prompt sequencing can measurably influence both scoring behavior and the pedagogical components present in generated feedback[15]. Moreover, because LLMs can hallucinate, an emerging direction is to introduce evaluator layers ("guardrails") that automatically screen or flag feedback for quality and hallucination risks before it reaches students [16]. Taken together, these findings suggest that evaluating LLM feedback in education must consider not only output quality, but also the educator oversight and verification effort required for safe deployment.

Despite the rapid proliferation of AI writing tools, there is limited educator-centered evidence examining how LLM feedback aligns with specific year-level curricula (e.g., Year 5) or how it integrates into authentic tutor workflows. Much prior work focuses on higher education or generic technical benchmarks, leaving a gap in understanding how primary educators judge the clarity, specificity, helpfulness, and relevance of LLM feedback, and under what boundary conditions they consider it safe to use in routine practice.

This study addresses that gap in collaboration with a commercial online tutoring platform that provides curriculum-aligned English writing support for primary and secondary students. This setting provides an authentic feedback pipeline and operational constraints, high feedback volume, short turnaround, and consistency requirements that shape what "useful" LLM feedback means in practice, and it foregrounds the verification burden required for safe integration.

We investigate the following research questions:

RQ1: How does LLM-generated feedback compare with tutor feedback across pedagogically meaningful dimensions for Year 5 narrative and persuasive writing?

RQ2: What boundary conditions, risks, and workflow requirements do educators identify when considering LLM-generated feedback for Year 5 writing?

To address these questions, we conduct an educator-centered evaluation framework grounded in authentic pedagogical data. We curated authentic student drafts paired with tutor feedback and generated GPT-4 Turbo feedback using template- and rubric-aligned prompting refined through educator participation. Four experienced English specialists co-designed a six-dimensional rubric (*Clarity, Specificity, Helpfulness, Feasibility, Relevance, Overall Effectiveness*) and completed a blind within-script comparison of tutor versus model feedback, complemented by thematic analysis of written reflections. This design enables a nuanced account of where LLM feedback aligns with pedagogical expectations and where it introduces risk in a real operational context.

Our findings show that tutors held small numerical advantages across most rubric dimensions, with a clearer advantage in perceived helpfulness, while LLM feedback is often described as clear

 doi:10.20944/preprints202605.1545.v1

and well-structured yet frequently generic, surface-focused, and occasionally misaligned with the draft. Educators judged LLM feedback to be most appropriate as fast, first-pass support on structure and surface features, rapid first-pass support for routine structure and surface revision, and least appropriate for developmental judgment and context-sensitive guidance without educator mediation. Building on these results, we contribute (1) an educator-co-designed, multi-dimensional rubric for evaluating primary writing feedback; (2) empirical evidence from a blinded, within-script comparison using authentic tutoring data; and (3) design requirements for teacher-in-the-loop primary writing feedback systems, including editable workflows and safeguards that reduce unsupported feedback before release to students.

2. Related Work

We situate our work at the intersection of (1) digital writing-support systems and automated writing evaluation (AWE), (2) pedagogically grounded feedback constructs, and (3) educator-centered and human-AI collaborative workflows for instructional use. Our focus is not only whether LLMs can produce fluent comments, but whether feedback is *pedagogically appropriate* for primary learners and *deployable* within real feedback pipelines where reliability, hallucination risk, and verification workload determine safe use.

2.1. Digital Writing Feedback Systems: From AWE to Generative Feedback

Automated writing support has evolved from rule-based and statistical AWE systems to neural approaches and, more recently, generative LLM-based feedback. Early systems (e.g., ETS Criterion) operationalized rubric-aligned features to assess organization, mechanics, and development [17,18]. Subsequent neural AWE introduced richer modeling of syntax, discourse, and second-language writing, yet remained largely rubric-bound and product-focused [19,20]. In parallel, widely adopted writing-support tools have scaled grammar and style assistance, but predominantly target surface accuracy and final-draft editing [2].

Large Language Models(e.g., GPT-3/4) shifted the paradigm from *scoring* to *generating* formative feedback that can be adapted to genre, task, and audience [21–24]. However, this shift complicates evaluation: surface fluency can mask substantive quality issues, and model feedback may contain unsupported claims or limited transparency about its basis [25,26]. Empirical evaluations in writing-feedback contexts suggest that LLM feedback can be fluent and readable, yet may be less reliable for evaluative judgment and sometimes misaligned with the student text, especially when criteria require contextual or developmental interpretation [12].

Recent work has begun to address school-age contexts and pedagogically aligned evaluation. For example, Steiss et al. [27] compared teacher formative feedback with ChatGPT-generated comments on secondary student writing across multiple dimensions and found that teacher feedback outperformed the model on most criteria, although these differences should be interpreted alongside the scalability benefits of AI-generated feedback. Similarly, Atasoy and Moslemi Nezhad Arani [28] compared teachers and LLM variants in classifying middle-school texts and showed that rubric-aligned prompting can improve model reliability, suggesting that constrained use may narrow, but not eliminate, the gap with expert judgment.

Together, these studies highlight that LLMs may function best as complements to educator feedback rather than replacements, and that pedagogical alignment requires careful constraint and evaluation. However, much of this work focuses on secondary or tertiary learners, targets scoring/classification rather than rich formative guidance, and rarely centres primary school developmental appropriateness or workflow constraints that shape practical usability.

2.2. Feedback Constructs, Developmental Appropriateness, and Safety

Feedback quality is multi-dimensional and cannot be reduced to surface correctness or textual similarity. In higher education, learner-centered feedback frameworks distinguish how comments support learners' sensemaking, future improvement, and agency, and large-scale analyses of authentic

 doi:10.20944/preprints202605.1545.v1

instructor feedback show that these components often co-occur and interact rather than appearing in isolation [13]. This perspective implies that evaluating LLM-generated feedback should consider whether comments help learners understand what to improve and how to improve it, and whether feedback supports productive student engagement rather than merely listing edits. In parallel, relational characteristics of how feedback conveys respect, encouragement, and openness to dialogue are increasingly recognized as shaping feedback uptake, yet remain difficult to operationalize and have rarely been evaluated systematically in automated pipelines [14]. For primary learners in particular, effective feedback must be not only actionable but developmentally appropriate in tone, specificity, and expectations.

Reliability and safety further complicate the adoption of LLM feedback in educational practice. Prior work on LLM feedback generation suggests that model outputs can be readable and structurally well-formed, yet uneven in reliability on evaluative judgments and sometimes misaligned with the student work, particularly when feedback requires contextual interpretation [12]. Because LLMs can hallucinate, an emerging direction is to introduce *evaluator* or *guardrail* layers that screen LLM-generated educational feedback across multiple dimensions, which include pedagogical quality and hallucination risks, before it is delivered to learners [16]. Complementing this, prompt and workflow design choices can measurably shape both assessment outcomes and the pedagogical components present in generated feedback, suggesting there may be no single prompting strategy that generalizes across learner performance levels and genres[15].

A related methodological limitation is that many evaluations of generated feedback rely on generic similarity metrics (e.g., BLEU, ROUGE, BERTScore) [29–31]. Such metrics are poorly aligned with pedagogical aims because high textual overlap does not guarantee helpfulness, feasibility, or developmental fit. This motivates educator-defined evaluation dimensions that directly reflect instructional goals and learner needs.

2.3. Educator-Aligned Evaluation and Human-AI Feedback Workflows

Educational technology and HCI research emphasize that AI systems for schools should be designed and evaluated with educators, who bring deep knowledge of developmental appropriateness, curriculum pacing, and genre expectations [32,33]. Teacher-shaped rubrics, exemplars, and evaluation artifacts can support actionable and trustworthy feedback [34,35]. However, many participatory approaches remain focused on tool ideation or interface features rather than systematic evaluation of *evaluating AI-generated feedback itself*. Educators are seldom placed in the role of structured raters of LLM outputs using developmentally meaningful criteria, nor are they routinely asked to identify where these outputs misalign with classroom priorities, create friction, or risk undermining students' learning [36,37]. Existing frameworks also rarely distinguish between boundary conditions where AI feedback may be safely leveraged (e.g., surface-level corrections) and those where it should be constrained or avoided (e.g., developmental judgment or context-sensitive guidance).

Human-AI collaboration research similarly highlights usability, transparency, and governance as prerequisites for trustworthy adoption in education, advocating human-in-the-loop pipelines where users can critique, edit, and override AI outputs [8,38,39]. Systems should make their capabilities and limitations legible to support appropriate reliance [39,40]. In educational contexts, studies show that teachers prefer AI tools that provide editable suggestions and clear rationales over systems that make opaque, autonomous judgments [32,41,42]. Hybrid systems such as GELEX illustrate how constraining generative edits and foregrounding accept/edit/reject controls can reduce cognitive load while preserving human agency [43]. Complementing this, scholarship on feedback literacy and formative assessment underscores that both teachers and students need support to interpret and act on feedback effectively [44–47]. In K-12 writing, effective feedback must balance technical correction with development of ideas, structure, and self-regulation [48–50].

Taken together, prior work suggests that while LLMs can produce fluent feedback, pedagogical value and safety depend on developmental fit, grounding in the student text, and the surrounding instructional workflow.

 doi:10.20944/preprints202605.1545.v1

3. Methodology

We conducted an educator-participatory, mixed-methods evaluation of LLM-generated writing feedback within an established online tutoring program supporting Australian curriculum-aligned persuasive and narrative writing for Year 5 learners. The study was designed to (i) compare model-generated and tutor feedback on educator-defined pedagogical dimensions, and (ii) identify boundary conditions and workflow requirements for safe use in routine feedback pipelines. This setting provided access to authentic student drafts and tutor comments, program rubrics and exemplars, and realistic operational constraints (e.g., turnaround time and educator workload) that shape the feasibility of any AI-supported feedback workflow.

3.1. Study Design Overview

Our methodology comprised six phases spanning study preparation, rubric development, blinded educator evaluation, and analysis synthesis (Table 1). Phases 1-2(Months 1-9), grouped as *Educator-Aligned Study Design and preparation*, focused on requirements gathering, study design, dataset construction, and rubric development in collaboration with program leaders and English specialists. Phases 3-6(Months 10-14), grouped as *Educator Evaluation and Analysis, and Design Synthesis*, centered on an educator-participatory evaluation, in which four English specialists conducted a blind, within-script comparison of GPT-4 Turbo feedback and authentic tutor feedback using the co-designed rubric, followed by researcher-led mixed-methods analysis of scores and comments and the derivation of boundary conditions and design recommendations.

Table 1 provides an overview of the phases, key questions, and outcomes. Below, we detail participants, dataset construction, LLM feedback generation, evaluation protocol, and mixed-methods analysis.

Table 1. Study phases, guiding questions, and key outcomes.

Phase	Guiding questions	Key outcomes
Pre-evaluation framing and study preparation (Months 1-9)		
1	What are the main feedback bottlenecks in the current program? Where might AI realistically support faster, clearer, and more consistent feedback for primary learners?	Identified workload and consistency constraints; established shared expectations for age-appropriate feedback; and prioritized rubric dimensions, prompt structure, and initial boundary conditions for safe AI use.
2	How can we design a fair and feasible comparison under real tutor workload constraints? What safeguards are needed to protect student data and avoid overclaiming about AI effectiveness?	Confirmed Year 5 narrative and persuasive writing as the target context; secured ethics approval and consent; and defined a blind, within-script evaluation protocol with clear inclusion criteria.
Rubric development, educator evaluation, and analysis preparation (Months 10-14)		
3	What makes feedback clear, specific, and developmentally appropriate for Year 5 writers? How can the evaluator's interpretation be made consistent for comparative analysis?	Co-developed a six-dimensional rubric with plain-language definitions and annotated exemplars; strengthened shared interpretation and consistency across evaluators.
4	Are rubric labels and scale anchors clear in practice? Where do evaluators struggle or disagree? Does the evaluation template introduce friction or ambiguity?	Refined rubric wording and scale anchors; simplified the evaluation template; and fixed the GPT prompt configuration to ensure stable conditions for full evaluation.
5	How consistently can evaluators apply the rubric across scripts? What patterns distinguish tutor from AI feedback? Where does AI output require educator verification or correction?	Collected 56 paired tutor-GPT ratings per dimension; surfaced comparative strengths (e.g., clarity/feasibility) and weaknesses (e.g., specificity/relevance); and produced a high-quality dataset for quantitative and qualitative analyses.
6	How do we ensure data integrity and usability for analysis (scores, comments, and metadata) across multiple evaluator files?	Consolidated evaluator spreadsheets into a unified dataset; paired scores with comments; resolved formatting inconsistencies; and clarified ambiguous entries via follow-up with evaluators.

* Phases 1-2 focus on educator-aligned study design and preparation; Phases 3-6 cover rubric co-design, blind evaluation, and analysis preparation.

 doi:10.20944/preprints202605.1545.v1

3.2. Participants and Roles

The study involved two distinct groups of participants across six phases, with separated responsibilities to reduce bias between design and evaluation.

3.2.1. Study design Contributors

These phases engaged key stakeholders, including the Industry Program Owner, Writing Program Manager, and senior educators. These participants were responsible for defining pedagogical priorities based on current instructional challenges, identifying gaps in existing feedback workflows, and contributing to the early co-design of the evaluation rubric. They also helped to identify curriculum-aligned use cases for integrating AI into the feedback process. Importantly, these contributors did not participate in the feedback scoring phase, ensuring a clear separation between design and evaluation responsibilities.

3.2.2. Evaluator Participants

The main evaluation phases involved four English specialists recruited from the industry to serve as independent evaluators. They were selected for their expertise in formative writing assessment, with a balanced representation of primary and secondary teaching experience. None of these evaluators had prior involvement with the specific student writing samples used in the study, ensuring that their assessments were both unbiased and independent.

We adopted a purposive sampling strategy to ensure both pedagogical depth and representativeness across educational stages. Recruitment was facilitated internally via email by the Manager of English Programs, who invited experienced educators to participate in evaluating GPT-4 Turbo and tutor-generated feedback. Interested candidates completed a background form detailing their teaching experience, year-level focus, and familiarity with delivering feedback. To minimize bias, any educators who had previously authored input for the included student samples were excluded. Final participant selection was conducted in coordination with the Manager to ensure a balanced distribution of expertise across year levels.

The final evaluation panel comprised two primary-level specialists, each with more than five years of experience teaching Years 3-6, extensive involvement in tutor training, and expertise in rubric-based formative assessment for persuasive and narrative writing, and two secondary-level specialists who taught Years 7-10 and led curriculum development in analytical and persuasive writing, including the design of feedback strategies and refinement of marking rubrics for high-stakes assessments. This interdisciplinary mix ensured that both the rubric and evaluation strategies reflected diverse pedagogical perspectives across developmental stages.

All participants provided their informed consent. Their contributions included evaluation, rubric co-design, calibration, and pilot testing. Initial co-design workshops explored writing program challenges, clarified feedback expectations, and identified potential AI integration use cases. Participants worked collaboratively to define and iteratively refine evaluation criteria, using real student samples and tutor feedback to align the rubric with Australian curriculum objectives and developmental expectations. Participants reviewed and scored exemplar feedback together, establishing shared interpretations of dimensions such as clarity, specificity, feasibility, helpfulness, and pedagogical appropriateness in a calibration session. In the eight-week formal evaluation phase, each educator independently annotated and scored 56 writing samples, each paired with GPT-4 Turbo and tutor feedback, using the co-designed rubric and structured comment fields. On average, participants dedicated approximately two hours per day to this phase, ensuring a rigorous and contextually grounded assessment of feedback quality. The study received ethics approval at our university.

3.3. Dataset Construction

3.3.1. Dataset Summary

Our corpus comprises de-identified writing tasks submitted to an online tutoring program by Year 5 students, each accompanied by tutor feedback. The raw archive contains 740 scripts (255 persuasive,

 doi:10.20944/preprints202605.1545.v1

460 narrative, 25 text responses) from Year 3-11 students. From the subset, we curated a pool of 100 Year 5 script-feedback pairs, roughly balanced by genre (persuasive/narrative) and tutor-marked proficiency (Low, Medium, High). This pool reflects the range of ability and task types that the program routinely encounters and defines the empirical *boundary condition* of our study: Year 5 persuasive and narrative writing in a real tutoring context. As shown in Table 2, the authentic tutor feedback tends to be longer and more variable than GPT-4 Turbo outputs, reflecting the human tendency to personalize and occasionally digress, a factor we control for in our analysis.

Table 2. Word count statistics for drafts, tutor feedback, and GPT-4 Turbo feedback ($N = 100$).

Metric	Student Drafts	Tutor Feedback	GPT Feedback
Count	100	100	100
Mean	461.99	363.61	296.37
Std. Dev.	328.73	64.71	34.61
Min	39	227	214
25%	270.50	318.50	271.75
Median	359.50	358.50	290.50
75%	541.75	400.25	321.25
Max	1985	541	389

For the human evaluation study, we then drew a stratified sample of N=56 script-feedback pairs from this pool, balancing genre (persuasive, narrative) and tutor-marked proficiency (Low, Medium, High). The goal was not to approximate population-level statistics, but to expose evaluators to a realistic spread of strengths, weaknesses, and their reflections in Year 5 writing.

3.3.2. Data Organization, Preprocessing, and Quality Assurance

Raw files were stored hierarchically (family ID → student name → writing type → draft/feedback versions). We developed Python tooling to parse folder structures, detect draft versus feedback files via version keywords (e.g., "Draft 1", "First Feedback"), and pair each draft with its corresponding tutor feedback. Pairing used a two-tier approach: heuristic matching from folder structure and filenames, followed by string similarity for ambiguous cases. All low-confidence matches were manually verified.

We retained only `.docx` and `.pdf` documents to ensure reliable parsing; image-based formats were excluded due to inconsistent OCR extraction. Extracted text was cleaned to remove formatting artifacts, duplicated annotations, and non-instructional metadata while preserving authentic spelling and errors in student work. Final data were stored in a `.csv` file with `draft_content`, `feedback_content`, and metadata fields. This verified dataset served as the foundation for GPT-4 Turbo feedback generation and subsequent human evaluation.

3.4. LLM Feedback Generation

3.4.1. Model Selection and Reproducibility

We generated AI feedback using GPT-4 Turbo via the OpenAI API (snapshot: `gpt-4-turbo-2024-04-09`) [51]. We selected GPT-4 Turbo as a strong, widely used general-purpose LLM that supports stable API integration and produces consistently well-formed instructional text, making it a practical baseline for studying formative feedback generation in operational settings. Prior work also reports that GPT-4-class models can perform competitively on education-oriented tasks, including feedback generation and formative assessment support [21,37].

Our goal is not to make model-wide claims that GPT-4 Turbo is universally effective (or ineffective) for Year 5 feedback. Rather, we treat this specific model snapshot, prompting template, and tutoring program context as an explicit evaluation boundary condition, enabling a controlled comparison with authentic tutor feedback and an educator-aligned characterization of strengths, limitations, and design implications. Because GPT-4 Turbo is a closed model, its parameters are not inspectable, and provider-side updates can change behavior over time, which limits strict reproducibility. To mitigate

 doi:10.20944/preprints202605.1545.v1

this, we report the exact snapshot identifier and use a fixed prompt template for all generations, allowing our results to be interpreted as a documented snapshot of model behavior under a stable configuration.

3.4.2. Prompt Design for Feedback Generation

To emulate a realistic tutor feedback structure, GPT-4 Turbo feedback was generated using a structured prompt template modeled on authentic examples of tutor feedback. Each prompt included two to three few-shot exemplars representative of Year 5 writing and followed a fixed three-part format aligned with formative feedback practices: (1) *What You Did Well* (up to four strengths with evidence from the draft), (2) *Action Points* (up to four targeted improvements linked to specific excerpts with guidance for revision), and (3) *Next Steps* (a concise priority summary in an encouraging tone).

The template and wording were iteratively refined with English specialists to support developmental appropriateness, alignment with program rubrics, and consistent tone [21,52]. Fixing the template enabled a fair paired comparison for **RQ1** and produced a stable set of AI outputs for educators to interrogate when judging pedagogical appropriateness, risks, and boundary conditions for teacher-mediated deployment (**RQ2**).

3.5. Educator Evaluation Protocol

We adopted a multi-phase, educator-participatory evaluation process in which educators shaped the constructs used for evaluation and then applied them in a controlled blind comparison. Phases 1-2 established pedagogical scope, ethical safeguards, data selection criteria, and operational constraints (e.g., anonymization and the principle that AI feedback would not be delivered to students without educator mediation). Phases 3-6 focused on rubric co-development, calibration, blind scoring, and data integration for mixed-methods analysis (Table 1).

3.5.1. Phases 1-2: Early Stakeholder Engagement and Study Preparation (Months 1-9)

Phases 1-2 grounded the study in real organizational needs and clarified when and where LLM feedback might be appropriate. Across nine months, the research team worked with the Industry Program Owner, Writing Program Manager, and senior educators to collect the requirements and the urgent needs.

- The team clarified current feedback bottlenecks, such as turnaround time, limits on personalization at scale, and low revision uptake.
- We identified plausible roles for AI in the feedback workflow and delineated where human judgment must remain central (e.g., developmental priorities, grading, and high-stakes tasks).
- We determined year level and writing genres for evaluating LLM-generated feedback in context.
- The team agreed on constraints for any classroom-facing use, including anonymization requirements and the principle that students would not receive AI feedback without teacher mediation.

Phase 1: Initial Requirements Gathering (Months 1-3)

Phase 1 focused on defining the scope and feasibility of integrating AI-generated feedback into an existing educational context. This phase was critical for establishing foundational goals with the industry partner and surfacing the pedagogical and ethical considerations that would guide the educator-participatory process.

Three workshops (two in person and one online) brought together the research team, the Industry Program Owner, the Writing Program Manager, and experienced English educators. Participants were introduced to both traditional commercial writing feedback platforms and recent research literature on writing feedback delivery and its pedagogical effectiveness. This was followed by a hands-on exploration of cutting-edge large language models (LLMs), including ChatGPT and Claude. Live demonstrations illustrated how these tools generate formative feedback across narrative and persuasive writing samples, highlighting both strengths, such as fluency and coherence, and limitations, such as generality and lack of context.

 doi:10.20944/preprints202605.1545.v1

In parallel, the Industry Writing Manager provided a detailed walkthrough of their existing writing program, including instructional structure, student planning scaffolds, and tutor feedback workflows. These discussions revealed persistent challenges, including tutor workload bottlenecks that slowed feedback turnaround time, scalability concerns that limited the provision of personalized feedback to every student, and low feedback application rates, where students often failed to revise their work based on tutor input.

From this joint analysis, we derived core evaluation questions that guide this paper: *How does LLM-generated feedback compare with tutor feedback on key pedagogical dimensions in Year 5 narrative and persuasive writing?*

Phase 2: Study design and Ethics preparation (Months 4-9)

Phase 2 translated the high-level requirements into a concrete, ethically approved study protocol (see Section 3.2). Regular meetings with the Writing Program Manager and English Program Owner were used to refine the scope and ensure feasibility under real tutor workload constraints. Several key design questions guided the discussions.

- Which year level should the writing samples target to maximize impact and generalizability?
- How should we structure the comparison between GPT-4 Turbo and tutor feedback to minimize bias?
- How can we ensure that the evaluation rubric is interpretable across different levels of teaching experience?

After evaluating multiple grade levels and dataset options, we selected **Year 5 student writing** as the primary focus of this study, as we provided in Section 3.2; The final protocol specified inclusion criteria for writing samples, a within-subjects blind evaluation design in which educators rated anonymized GPT and tutor feedback on the same student writing, six evaluative criteria (Clarity, Specificity, Helpfulness, Feasibility, Relevance, and Overall Effectiveness), and robust data management and anonymization procedures.

Overall, *Phases 1-2* established the pedagogical framing, ethical safeguards, and shared evaluative language that shaped all subsequent phases of the project and ensured that subsequent phases remained anchored in the partner's real instructional needs.

3.5.2. Phases 3-6: Rubric Development, Human Evaluation, and Data Cleaning for Analysis (Months 10-14)

Phases 3-6 involved four experienced English specialists tasked with co-developing the evaluation rubric and conducting the formal assessment. In line with participatory alignment practices [53], educators were asked to articulate pedagogical expectations and refine evaluative constructs, and reflect on where LLM feedback aligned or conflicted with their practice.

Phase 3: Rubric Development and Calibration (Months 10-11)

Following ethics approval and initial study planning, we undertook a two-month iterative co-design process focused on developing and refining the evaluation rubric. Through one in-person rubric design workshop and two virtual follow-up meetings, the group iteratively refined the rubric wording, examples, and scale anchors to ensure it aligns with real-world classroom expectations and instructional values.

The collaborative process produced a six-dimensional rubric that follows the Australian English curriculum [1]and fits the tutoring program's established practice. Each dimension was clearly defined and paired with exemplar annotations to support shared understanding and inter-rater reliability. Evaluators rated each dimension on a 5-point Likert scale from *Strongly Disagree* (1) to *Strongly Agree* (5). The six dimensions used to evaluate feedback quality are outlined in Table 3.

[1] Australian Curriculum: English,https://www.australiancurriculum.edu.au/.

 doi:10.20944/preprints202605.1545.v1

Table 3. Evaluation Dimensions for Writing Feedback

Dimension	Description
Clarity	The feedback is clear, and the language is easy for the student to understand.
Specificity	The feedback addresses specific strengths and weaknesses within the writing.
Helpfulness	The feedback is practical and actionable, guiding the student toward specific improvements.
Feasibility	The feedback is understandable and manageable for the student.
Relevance	The feedback aligns with the student's writing content.
Overall Effectiveness	The feedback supports the student's writing development overall.

This rubric guided all human evaluations in the study and ensured consistent, rubric-aligned assessment across both tutor and AI-generated feedback.

Phase 4: Pilot Testing and Tool Iteration(Month 12)

Before full deployment, we ran a short pilot with a small subset of scripts. Educators entered ratings using an Excel-based form with one column per dimension and optional cell-level comments. Pilot feedback indicated minor construct overlap (particularly between Specificity and Helpfulness), occasional uncertainty about scale anchors, and friction in recording scores and rationales. We addressed these issues by refining rubric wording and anchors, adding exemplars for ambiguous cases, and simplifying the evaluation template. To support the interpretability of subsequent results, we also fixed the LLM feedback generation configuration at this stage, so the formal evaluation compared stable conditions.

Phase 5: Full Evaluation and Iterative Review (Months 13-14)

Phase 5 produced the primary evaluation data. Educators conducted a blind within-script comparison of 56 student drafts, each paired with tutor feedback and GPT-4 Turbo feedback. A calibration session at the start of Phase 5 used two anonymised examples to align interpretation of rubric dimensions and reduce drift; this session also generated additional annotated exemplars for subtle distinctions (e.g., actionable specificity versus generic encouragement; feasibility versus developmental appropriateness).

Following calibration, the remaining 54 scripts were allocated to balance genre and tutor-marked proficiency levels. Each script was scored by exactly one evaluator to reflect realistic workload constraints. For each script, the same evaluator rated both tutor and GPT feedback on all six dimensions, enabling paired comparison in a realistic decision frame ("which feedback would I provide to this student?"). Evaluators completed scoring asynchronously over approximately eight weeks, with brief check-ins used only to clarify rubric interpretation. For each feedback instance, educators provided 5 Likert ratings and optional free-text rationales explaining their judgments.

Phase 6: Data Integration and Analysis Preparation

The final phase ensured the integrity and usability of the evaluation data. The research team consolidated the educators' Excel-based evaluation files into a unified, standardized dataset. A semi-automated extraction and cleaning process paired each score with its corresponding comment, resolved formatting inconsistencies, and preserved original phrasing. Where entries were unclear, we followed up with the relevant evaluators for clarification. This integrated dataset underpins the quantitative rubric comparisons and qualitative analysis of educator rationales reported in Section 3.6.

3.6. Data Analysis

To address **RQ1** and **RQ2**, we adopted a mixed-methods analytical approach. Quantitative analyses compared rubric scores across feedback conditions to examine how LLM-generated (GPT-4 Turbo)

 doi:10.20944/preprints202605.1545.v1

feedback differed from tutor feedback on the six pedagogical dimensions (**RQ1**). Complementing this, qualitative analyses of educators' written comments explored how they interpreted and experienced these differences in practice, and how they judged the pedagogical appropriateness and risks of LLM feedback within real tutoring workflows RQ1 and RQ2. Table 4 summarizes the primary analysis methods and illustrates the kinds of outputs they produced.

Table 4. Overview of primary analysis methods used in the study.

Analysis type	Methodology	Example output
Quantitative rubric comparison	Paired, within-script comparison of GPT vs. tutor Likert ratings on six criteria. Computed descriptive statistics, paired t-tests (Holm-Bonferroni), and paired-effect sizes (d).	Helpfulness: tutors rated higher than GPT (mean $\Delta = 0.43$, $d = 0.54$, Holm-adjusted $p < .05$).
Educator-aligned thematic analysis	Reflexive thematic analysis of educators' free-text comments using a structured codebook. Codes were synthesised into cross-cutting design considerations.	Themes included: "clear but shallow feedback"; "surface vs. conceptual goals"; "AI as assistant, not replacement".

Table 5 shows the structure of the educator evaluation form with one row per feedback instance. Columns list Script ID, Source (Tutor or GPT), and six rubric dimensions (Clarity, Specificity, Helpfulness, Feasibility, Relevance, Overall), each scored on a 1-5 Likert scale. The final column contains an overall free-text comment. Two example rows are displayed for the same script: one for tutor feedback with generally higher scores and a comment about length and organization, and one for GPT feedback with slightly lower scores and a comment noting that it focuses on technical skills rather than ideas or structure. In the full spreadsheet, educators could also attach short comments directly to individual rubric cells (e.g., beside the Clarity score); during data cleaning, we exported these cell-level notes and mapped each one to its corresponding script and dimension.

Table 5. Structure of the educator evaluation form used to rate tutor and GPT-4 Turbo feedback.

ID	Source	Cla.	Spe.	Hel.	Fea.	Rel.	Ove.	Overall comment (example)
23	Tutor	4	3	4	3	4	4	The student's story is very lengthy. Feedback could have given more guidance on organising ideas, not just sentence-level issues.
23	GPT	3	3	2	3	2	2	Language is clear but focuses only on technical skills; misses structure and content issues, so would need teacher editing before use.

* Cla.(Clarity), Spe. (Specificity), Hel.(Helpfulness), Fea. (Feasibility), Rel. (Relevance), Ove. (Overall effectiveness). Scores use a 1-5 Likert scale.

3.6.1. Quantitative Rubric Analysis

Each of the 56 student scripts was associated with two feedback instances (GPT-4 Turbo and tutor), and the same educator rated both instances on the six rubric dimensions. This yielded, for each dimension, 56 paired observations (one GPT score and one Tutor score per script), enabling a within-script comparison that mirrors realistic decision contexts ("which feedback would I actually give this student?"). For each dimension, we first computed descriptive statistics (means, standard deviations, and 95% confidence intervals) separately for GPT and Tutor feedback. We then formed a difference score for each script (Tutor minus GPT) and inspected the distribution of these 56 differences using histograms and Q-Q plots. Shapiro-Wilk tests on the difference scores indicated no extreme departures from normality, so we proceeded with paired samples t-tests for each dimension.

We conducted six paired t-tests (one per dimension), and applied a Holm-Bonferroni correction to the resulting p-values to control the family-wise error rate at $\alpha = .05$. For each dimension, For each dimension, we report: (1) tutor and GPT means and standard deviations, (2) mean paired difference

 doi:10.20944/preprints202605.1545.v1

and its standard deviation, (3) Holm-adjusted p-value, and (4) Cohen's d for paired samples (mean difference divided by the standard deviation of the differences).

Given the modest sample size and the well-known limitations of relying solely on null-hypothesis significance testing, we interpret results primarily through the *direction* and *magnitude* of effects (e.g., small vs. medium differences) rather than binary "significant/non-significant" labels.

3.6.2. Qualitative Thematic Analysis

To understand educators' judgments beyond numeric scores, we conducted a qualitative analysis of the written comments attached to rubric ratings and to the final overall summary field. For each of the 56 scripts, educators could justify or elaborate on their scores for each dimension and provide an overall comparative judgment of GPT vs. Tutor feedback. This produced a corpus of comments that directly explained how they interpreted clarity, specificity, helpfulness, feasibility, relevance, and overall effectiveness in context.

We conducted a reflexive thematic analysis with iterative familiarization, initial coding, code refinement, theme development, and educator resonance checking. Codes captured recurrent educator descriptions (e.g., "clear but generic", "not grounded in the draft", "actionable next steps", "developmentally inappropriate", "requires teacher rewriting"). Themes were synthesized to directly inform **RQ2**, including tensions between accessible language and pedagogical depth, surface correction and conceptual goals, scalability and verification workload, and agency/control in teacher-in-the-loop workflows. Details of the coding scheme and an extended worked example are provided in Appendix A1 and Appendix A2.

4. Results

This section reports findings for the two research questions. We first report quantitative comparisons of tutor and GPT-4 Turbo feedback across the six educator-defined pedagogical dimensions. We then use educators' qualitative comments to explain how the two feedback sources were experienced in practice and to identify the conditions under which LLM-generated feedback was judged pedagogically appropriate or risky.

4.1. Quantitative Findings

For each of the 56 scripts and each rubric dimension, the same educator rated both GPT-4 Turbo and tutor feedback. We computed a difference score per script (*Tutor* − *GPT*) and used these paired differences to derive means, standard deviations, paired t-tests, Holm-Bonferroni-corrected p-values, and Cohen's d for paired samples.

Table 6 showed that across all six criteria, tutor feedback received higher mean ratings than GPT-4 Turbo feedback. However, the magnitude of these differences was modest. Across conditions, mean scores for both sources fell within a relatively narrow band on the 1-5 Likert scale (2.70–3.41). The mean advantage for tutors ranged from $\Delta = 0.07$ (Clarity) to $\Delta = 0.43$ (Helpfulness), with corresponding effect sizes all in the small range ($d = 0.06$-0.35). After applying Holm-Bonferroni correction for the six comparisons, none of the dimensions reached conventional statistical significance ($p_{\text{Holm}} \geq .07$).

Two descriptive patterns are particularly relevant for interpreting the qualitative findings and later design implications. First, GPT-4 Turbo most closely approximated tutors on surface-facing dimensions *Clarity* and *Feasibility*, where mean differences were small ($\Delta \leq 0.16$, $d \leq 0.12$).This pattern is consistent with prior work showing that LLMs often perform well on readability and surface coherence, producing feedback that appears accessible and well structured [14]. Second, the largest numerical gap appeared for *Helpfulness* ($\Delta = 0.43$, $d = 0.35$), followed by smaller gaps for *Relevance* and *Overall effectiveness*. This pattern reinforces the view that feedback quality is multi-dimensional: feedback may be linguistically clear while still falling short on learner-centered qualities such as actionable guidance, prioritization, and alignment with the student's most immediate revision needs [13].

 doi:10.20944/preprints202605.1545.v1

Table 6. Quantitative comparison of tutor and GPT-4 Turbo feedback ($N = 56$).

	Tutor		GPT-4 Turbo		Difference				
Dimension	M	SD	M	SD	Δ	SD_Δ	t	p_{Holm}	d
Clarity	3.41	0.78	3.34	0.98	0.07	1.13	0.48	1.00	0.06
Specificity	2.93	0.76	2.82	1.01	0.11	1.33	0.60	1.00	0.08
Helpfulness	3.14	0.82	2.71	1.07	0.43	1.22	2.63	.07	0.35
Feasibility	3.09	0.84	2.93	0.97	0.16	1.33	0.90	1.00	0.12
Relevance	3.16	0.85	2.86	1.14	0.30	1.37	1.65	.52	0.22
Overall effectiveness	2.96	0.81	2.70	0.95	0.26	1.31	1.48	.58	0.20

* M and SD are on a 1-5 Likert scale. Δ = Tutor − GPT. $df = 55$. p_{Holm} values are adjusted for six tests using the Holm-Bonferroni method. Cohen's d is calculated for paired samples.

Figure 1 summarizes the mean ratings for tutor and GPT-4 Turbo feedback across the six dimensions. Tutors receive slightly higher scores on every dimension, with the clearest numerical gap on *Helpfulness*, whereas GPT-4 Turbo comes closest to tutors on *Clarity* and *Feasibility*. To further investigate these trends, Figure 2 illustrates the paired differences (Tutor − GPT) with 95% confidence intervals (CIs). While all mean differences were positive, suggesting a consistent descriptive advantage for human tutors, all CIs overlapped with zero. This alignment with the Holm-corrected p-values (reported in Table 6) indicates that while tutors generally performed better, the performance gap between human specialists and GPT-4 Turbo remained modest across these dimensions.

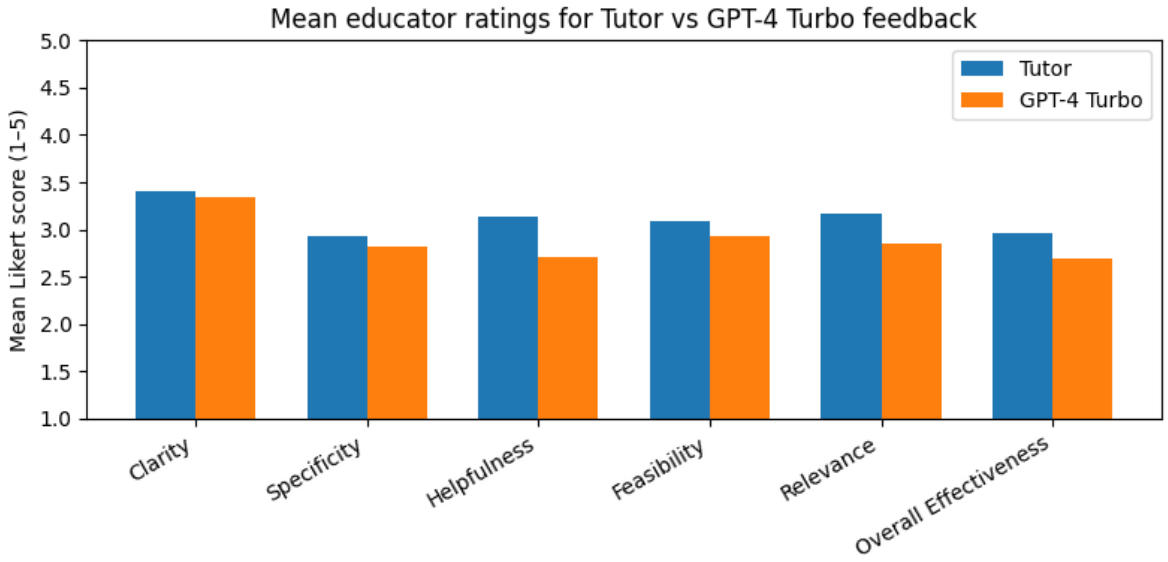

Figure 1. Mean educator ratings by feedback source. Mean rubric ratings for tutor and GPT-4 Turbo feedback across six dimensions ($N = 56$ scripts).

 doi:10.20944/preprints202605.1545.v1

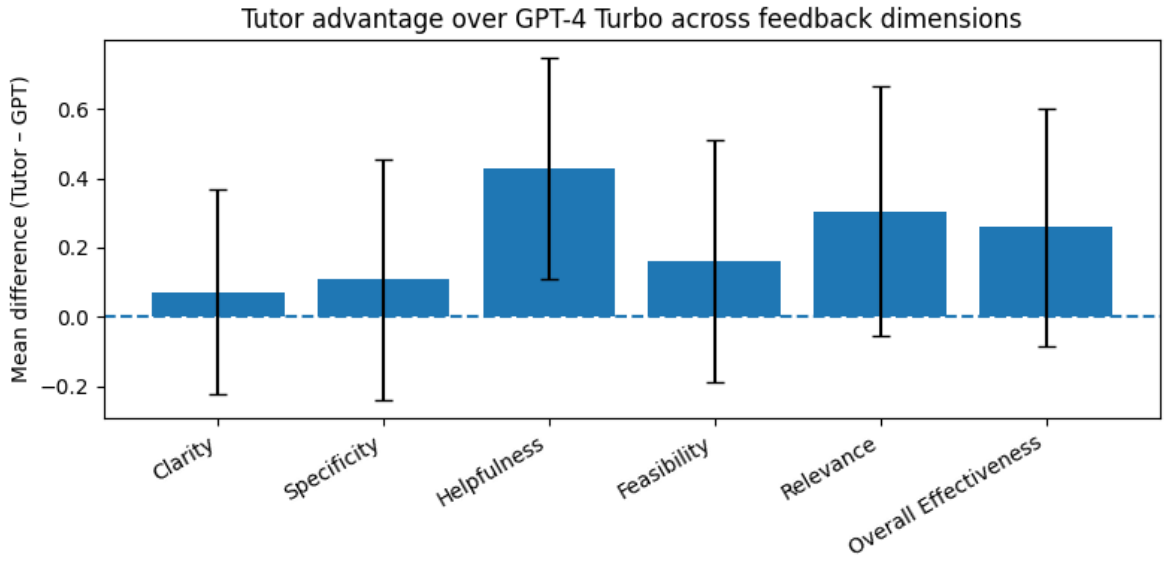

Figure 2. Tutor advantage over GPT-4 Turbo across dimensions. Mean paired differences (Tutor − GPT) for each rubric dimension computed within script ($N = 56$). Positive values indicate higher ratings for tutor feedback.

Overall, the quantitative findings provide a partial answer to **RQ1**. Under our fixed prompt template and boundary condition (Year 5 persuasive and narrative writing), GPT-4 Turbo produced feedback that educators rated similarly to tutors on clarity and basic feasibility, while tutors retained a modest advantage, most notably in perceived helpfulness. Because prior work shows that LLM feedback quality can vary meaningfully with prompt design and sequencing, these results should be interpreted as conditional on the prompting and workflow constraints used in this study [15]. Moreover, these numeric scores capture perceived quality but do not directly quantify reliability risks (e.g., unsupported or misleading claims), which motivates the qualitative analysis and the workflow implications discussed later, including the need for screening and educator verification [16].

4.2. Qualitative Findings: Educator Perspectives

To contextualize the rubric scores for **RQ1** and to address **RQ2**, we conducted a thematic analysis of the educators' dimension-specific comments. Following the reflexive thematic analysis approach described in Section 3.6, each comment was assigned a primary code and subsequently grouped into higher-level themes.Table 7 provides an excerpt of the codebook, while the full version is available in Appendix A1. The resulting theme frequencies are summarized in Table 8. We synthesized these salient patterns for each dimension to identify areas where GPT-4 Turbo and tutor feedback converged or diverged in quality. These qualitative findings provide the empirical basis for the four design considerations discussed in Section 5.

Table 7. Excerpt of the qualitative codebook used to analyze educator comments.

Code	Dimension	Label	Description (with example paraphrase)
CL1	Clarity	Clear, simple language	Feedback uses straightforward, age-appropriate language. *E.g., "Language is simple; a Year 5 student could follow this."*
HE1	Helpfulness	Actionable guidance	Feedback offers concrete, text-grounded suggestions. *E.g, "Gives clear steps for how to reorganize ideas."*
SP5	Specificity	Specific but wrong	Feedback cites a feature, but the example is misread or invented. *E.g., "Criticizes capital letters used correctly."*
RE3	Relevance	Hallucinated problem	Feedback identifies a problem not present in the draft. *E.g, "Says there are no paragraphs when there clearly are."*

 doi:10.20944/preprints202605.1545.v1

Table 8. Summary of higher-level themes and their frequency in educator comments on GPT-4 Turbo vs. tutor feedback.

Dim.	Theme	GPT	Tutor	Total
Clarity				
C1	Clear, accessible language	16	5	21
C2	Accurate but too advanced / dense	1	9	10
C3	Confusing or misleading explanation	7	11	18
Feasibility				
F1	Manageable, scaffolded actions	5	8	13
F2	Overload (too many actions)	2	0	2
F3	Lacks concrete "how to" guidance	10	15	25
F4	Feasibility harmed by language / errors	8	3	11
Helpfulness				
H1	Actionable, growth-oriented feedback	7	10	17
H2	Generic or low-value comments	6	6	12
H3	Missing or purely technical focus	10	3	13
H4	Misaligned to student needs	12	9	21
Specificity				
S1	Text-grounded specificity	8	11	19
S2	Generic, no examples	12	24	36
S3	Specific but incorrect	9	1	10
S4	Missed content-specificity	8	12	20
Relevance				
R1	Directly targets core issues	2	1	3
R2	Only minor / technical issues	10	5	15
R3	Partial, missed opportunity	4	14	18
R4	Inaccurate or hallucinated issues	2	0	2
Overall reflections				
O1	Surface-focused, not conceptual	-	-	11
O2	Reliability and verification burden	-	-	10
O3	Comparative judgments	-	-	27
O4	AI as assistant, not replacement	-	-	8

Counts reflect primary codes assigned to free-text comments ($N = 56$ pairs). Dim. = dimension.

4.2.1. Clarity

Quantitatively, tutors and GPT-4 Turbo received similar clarity ratings (Table 6). Tutor feedback was rated slightly higher than GPT-4 Turbo (M_{Tutor} = 3.41, SD = 0.78; M_{GPT} = 3.34, SD = 0.98), but the paired difference was small (Δ = 0.07) with a negligible effect size (d = 0.06) and no statistically significant difference after Holm-Bonferroni correction (p_{Holm} = 1.00;). Qualitative comments reveal that comparable scores often reflected *different clarity pathways*: GPT tended to be accessible by default, whereas tutor clarity varied more with educator tone and the use of assessment language.

C1: Clear and accessible language

Educators frequently described GPT feedback as "simple and clear" and well matched to the reading level of weaker writers (16 GPT vs. 5 tutor instances). Several noted that GPT feedback was "a manageable length" and "easy for the student to follow." However, these same comments often foreshadowed later concerns: accessible language sometimes coincided with repetition or generic phrasing rather than draft-specific instructional value.

C2: Accurate but overly advanced or dense wording

This theme occurred primarily in tutor feedback (9 tutor vs. 1 GPT instance). Educators pointed to curriculum jargon and abstract verbs (e.g., "convey", "denote") that may be technically appropriate but difficult for Year 5 learners to interpret. In some cases, tutors' attempts to be precise introduced linguistic load that undermined the accessibility of otherwise valid guidance.

 doi:10.20944/preprints202605.1545.v1

C3: Confusing or misleading explanations from both sources

Both sources attracted comments where clarity was undermined by inconsistent reasoning or misaligned examples (7 GPT vs. 11 tutor instances). Educators flagged contradictory advice, unclear references, or explanations that did not map cleanly onto the student draft (e.g., criticism of capitals used correctly). These cases underscore that clarity depends not only on word choice, but also on coherent, text-aligned explanations.

Overall, these themes help explain why aggregate clarity scores were similar across GPT and tutor feedback. GPT-4 Turbo tended to produce shorter, more accessible language, but sometimes at the cost of depth or precision. Tutors, by contrast, were more likely to use sophisticated or dense language that reflects assessment expertise, which occasionally reduced accessibility for Year 5 learners. However, both could fail when explanations were poorly grounded or internally inconsistent.

4.2.2. Feasibility

Feasibility scores were moderate for both feedback types (Table 6), with no statistically significant difference after correction. Educators' comments, however, revealed distinct patterns in why certain feedback was seen as more or less actionable for Year 5 students.

F1: Mangageable and scaffolded actions

Both GPT and tutor feedback were occasionally described as offering a manageable set of actions, especially when changes were clearly signposted or linked to highlighted text (F1; GPT: $n = 5$, Tutor: $n = 8$). Tutors were slightly more often credited with providing stepwise guidance that could realistically be completed in one revision session.

F2/F3: Overload and lack of "How to" support

Instances of outright overload were relatively rare (F2; GPT: $n = 2$, Tutor: $n = 0$), but a common concern was the absence of actionable scaffolding (F3; GPT n=10; tutor n=15).Educators described feedback that named a goal ("add more detail", "fix tense") without showing students how to execute it in this text. This pattern suggests that feasibility in primary contexts is closely tied to worked examples, sentence starters, or explicit modelling.

F4: Feasibility undermined by language or correctness, especially for GPT

This theme captured cases where feasibility was compromised by either advanced language or incorrect feedback (F4). GPT feedback was coded more often than tutor feedback(GPT: $n = 8$, Tutor: $n = 3$). Educators flagged instances where feasible-sounding actions were paired with complex phrasing or were based on incorrect premises (e.g., asking students to fix a problem not present). These comments connect feasibility to reliability: even well-scaffolded advice becomes infeasible if the diagnosis is wrong.

4.2.3. Helpfulness

Helpfulness scores showed the largest quantitative difference (Table 6), and qualitative comments clarify that educators' helpfulness judgments emphasized *prioritization*, *draft grounding*, and *instructional leverage* rather than surface correctness alone.

H1: Actionable and growth-oriented feedback

Both feedback types occasionally provided concrete, text-grounded suggestions that were expected to improve the current draft and support longer-term skill development (H1; GPT: $n = 7$, Tutor: $n = 10$). Tutors were slightly more often praised for locating leverage points in the writing (e.g., reorganizing ideas, developing arguments) rather than focusing solely on surface errors.

 doi:10.20944/preprints202605.1545.v1

H2: Generic or low-value comments

Generic comments were common for both sources (H2; GPT: $n = 6$, Tutor: $n = 6$). These comments tended to echo high-level rubric language, such as "add more description" or "work on punctuation", without indicating where to start. Educators suggested that this feedback might not move students beyond what they already heard in the classroom.

H3/H4: Missing key issues and misalignment with student needs

GPT feedback was more often coded as missing key learning needs or focusing narrowly on technical edits (H3; GPT: $n = 10$, Tutor: $n = 3$). Both sources also attracted comments indicating misalignment with the student's level or needs (H4; GPT: $n = 12$, Tutor: $n = 9$), such as recommending advanced vocabulary when the draft required coherence and sentence control. These patterns help explain why tutors were perceived as more helpful overall despite shared instances of generic phrasing.

4.2.4. Specificity

Although quantitative differences on Specificity were small, educators' comments point to systematic contrasts in how feedback grounded claims in the student text.

S1: Text-grounded specificity

Educators valued feedback that cited particular sentences or phrases from the student's work (S1; GPT: $n = 8$, Tutor: $n = 11$). Both GPT and tutors sometimes highlighted specific examples when praising strengths or suggesting improvements, making it easier for students to see what to change.

S2/S4: Generic or missed opportunities, especially for tutors

Generic specificity failures were common, especially for tutor feedback (S2; GPT: $n = 12$, Tutor: $n = 24$). Even when tutors offered insightful observations, positive statements were often left ungrounded (e.g., "great description" without quoting the relevant passage). A related theme captured missed opportunities to tie guidance to specific plot or topic details (S4; GPT: $n = 8$, Tutor: $n = 12$), such as failing to reference the camping scenario or persuasive topic under discussion.

S3: Specific but incorrect examples, mostly from GPT

A distinctive failure mode appeared for GPT: feedback that looked precise but was inaccurate or invented (S3; GPT: $n = 9$, Tutor: $n = 1$). Educators treated this as particularly risky because apparent specificity can increase trust while simultaneously misleading the learner.

4.2.5. Relevance

Relevance comments focused on whether feedback targeted the most instructionally important issues for that script.

R1: Directly targeting core issues was rare

Few comments in either condition were coded as directly addressing the script's core learning need (R1; GPT: $n = 2$, Tutor: $n = 1$), suggesting that both sources sometimes under-prioritized the most valuable instructional move.

R2/R3: Over-emphasis on minor or partial issues

More commonly, GPT feedback was described as focusing on minor, technical issues (R2; GPT: $n = 10$, Tutor: $n = 5$) while tutor feedback was more often described as partially relevant but leaving key opportunities underdeveloped (R3; GPT: $n = 4$, Tutor: $n = 14$). This contrast suggests different failure patterns: GPT tended toward surface correction; tutors sometimes recognized deeper issues but did not fully translate them into actionable guidance.

 doi:10.20944/preprints202605.1545.v1

R4: Inaccurate or hallucinated relevance for GPT

GPT was uniquely associated with relevance errors arising from hallucination or mischaracterization (R4; GPT: $n = 2$, Tutor: $n = 0$). Educators flagged GPT feedback that praised or criticized features not present in the draft. These cases were treated as salient warning signals that AI feedback requires screening before classroom use, even when overall quality appears acceptable [16].

4.2.6. Overall Effectiveness and Summary Judgments

In addition to dimension-specific comments, educators provided overall reflections on the usefulness and risks of integrating GPT-4 Turbo feedback into the Year 5 program. Because these remarks often compared GPT and tutors directly, we analyzed them as a single "Overall" dimension rather than separating them by feedback source.

O1: Surface-level improvement without conceptual growth

Educators noted that feedback in both conditions often prioritized surface features over conceptual development (O1; $n = 11$), highlighting a persistent tension in writing feedback between editing and learning-focused guidance.

O2: Reliability and verification burden

A recurring theme concerned the time cost of checking feedback for accuracy and alignment (O2; $n = 10$). Educators emphasized that any AI-generated feedback would need to be verified and often edited, suggesting that workflow benefits depend on whether AI reduces or shifts workload rather than simply adding an extra checking layer.

O3: Comparative judgments between GPT and tutors

Educators frequently compared the two sources directly (O3; $n = 27$). Tutors were typically seen as better aligned to curriculum priorities and better able to diagnose nuanced strengths. However, GPT was sometimes preferred for weaker writers when tutor feedback was lengthy or overly technical, echoing the clarity patterns above.

O4: AI as assistant, not replacement

Educators repeatedly framed GPT-4 Turbo as potentially useful for drafting or rephrasing feedback, generating examples, or providing a first-pass structure that educators could adapt (O4; $n = 8$). Across these comments, the consistent requirement was that educators retain control over what is delivered to students, including editing for developmental appropriateness and screening for inaccuracies.

Together, these qualitative findings explain why quantitative differences were modest while still educationally meaningful: GPT feedback often achieved accessibility and structural coherence, but educators differentiated helpfulness, specificity, and relevance based on draft grounding, prioritization, and reliability. These patterns directly inform the boundary conditions and design considerations discussed in Section 5.

5. Discussion and Educational Impact

This study examined how educators judge LLM-generated feedback for Year 5 narrative and persuasive writing in an authentic online tutoring context. Specifically, **RQ1** compared GPT-4 Turbo feedback with authentic tutor feedback across six educator-defined dimensions, while **RQ2** investigated how educators interpreted the pedagogical appropriateness, risks, and practical implications of using LLM-generated feedback in routine writing support.

Overall, our findings suggest that LLM-generated feedback can approximate tutor feedback on some surface-facing qualities, but remains more limited on pedagogical judgment. Quantitatively, tutor feedback received slightly higher ratings across all six dimensions, although the differences were modest and did not remain statistically significant after correction for multiple comparisons (Section 4).

 doi:10.20944/preprints202605.1545.v1

This pattern is broadly consistent with prior work showing that LLMs can perform comparably to humans on some analytic or holistic evaluation dimensions while still differing in instructional value, reliability, and alignment with learner needs [27,42,54]. In our study, GPT-4 Turbo came closest to tutor feedback on *Clarity* and *Feasibility*, suggesting that it can generate readable, well-structured comments that may be usable as a first-pass feedback draft.

However, the qualitative findings show that similar rubric scores can conceal different pedagogical qualities. Educators frequently described GPT-4 Turbo feedback as clear, concise, and accessible for Year 5 learners, but also as more generic, more surface-oriented, and occasionally misaligned with the student draft. Tutor feedback, in contrast, was seen as more context-sensitive, developmentally attuned, and better aligned with the student's immediate learning priorities, though it could also be overly dense or insufficiently specific in places. These patterns highlight an important point for educational technology research: feedback that appears strong on readability or surface coherence is not necessarily strong on pedagogical usefulness.

With respect to **RQ2**, educators' written judgments articulated boundary conditions around when GPT-4 Turbo feedback is perceived as helpful and when it becomes risky. GPT-4 Turbo was generally considered most appropriate for rapid, first-pass support on routine structure and surface-level revision, especially for weaker writers who may benefit from shorter and more accessible suggestions. In contrast, it was considered least appropriate for developmental judgment, context-sensitive guidance, and higher-order diagnosis of what the learner most needs next. Educators also repeatedly highlighted the risk of hallucinated praise or criticism, as well as the additional verification work required before AI-generated comments could be shared with students. These concerns align with broader literature on automated writing evaluation and educational AI, which shows that automation often performs best on visible or lower-level features while requiring human oversight for deeper pedagogical interpretation and quality assurance [7,37,55].

Importantly, our findings also align with recent arguments that "feedback quality" is multi-construct and cannot be reduced to fluency or correctness alone. Learner-centered analyses of authentic feedback in higher education show that effective comments typically combine (i) clear diagnosis, (ii) actionable strategies for improvement, and (iii) support for learner agency and future work [13]. In addition, Relational dimensions (e.g., whether feedback conveys respect, encouragement, and an invitation to engage) are often discussed as relevant to how learners respond to feedback, yet they remain difficult to operationalize and are rarely evaluated systematically in automated pipelines; recent evidence shows that LLM-based characterization is sensitive to construct definitions and prompting choices [14]. In our dataset, GPT-4 Turbo frequently produced encouraging language, but educators noted that supportive tone can become pedagogically unsafe when paired with genericity or hallucinated praise/criticism (O2), suggesting that relational polish should not be treated as a proxy for pedagogical validity.

5.1. Design considerations for LLM-Generated Feedback

Drawing directly on educator ratings and coded comments (Section 4), we synthesize four design considerations for AI-supported primary writing feedback systems.

5.1.1. Consideration CR1: Balancing Accessible Language and Pedagogical Depth(C1-C2)

Educators often preferred GPT feedback for readability and length (C1) and age-appropriate phrasing, while tutor feedback more often contained dense terminology or discourse-level language (C2). At the same time, educators also described GPT feedback as "clear but shallow" in some cases, indicating that accessibility alone does not guarantee pedagogical value. In practice, this suggests that feedback systems should support *layered communication*: concise student-facing feedback paired with teacher-facing rationales that make visible the underlying pedagogical purpose of each suggestion. Such a design would allow feedback to remain readable for learners while preserving the professional judgment needed for curriculum alignment and instructional decision-making [40,56,57].

 doi:10.20944/preprints202605.1545.v1

5.1.2. Consideration CR2: Moving beyond Surface Corrections to Conceptual Learning Goals (H1–H3, R1–R3, O1)

Themes on helpfulness, specificity, and relevance converge on a second consideration: both GPT and tutors tended to focus on surface-level correctness rather than deeper learning. Many comments across sources were coded as generic or low-value (H2), minor or technical only (R2), or improving local correctness without addressing the main learning goal (O1). Even when feedback was text-grounded (S1), educators repeatedly noted that the most important learning goals, like developing ideas, structuring arguments, and sustaining narratives, were only partially addressed (R3) or missed entirely (H3). Similar patterns have been documented in automated written corrective feedback and commercial writing support tools, which tend to privilege grammar and mechanics over discourse-level revision [3,55]. Recent comparisons of human and LLM feedback likewise report that models are strongest on visible surface features and weaker on higher-order concerns [7,27,42]. For AI-supported systems, this suggests that prompt design and interface structure should deliberately foreground one or two high-leverage conceptual goals before offering lower-level corrections. In educational settings, more feedback is not necessarily better; what matters is whether feedback directs students toward meaningful improvement.

5.1.3. Consideration CR3: Trading Off Scalability and Specificity with Reliability and Verification Work (S1–S3, R4, O2)

A distinctive failure mode of GPT-4 Turbo in this study was *specific but incorrect* feedback (S3) and occasional hallucinated issues (R4). Educators viewed this as particularly risky because apparently precise advice can invite trust while simultaneously misleading the learner. This finding is important for educational technology deployment: specificity should not be treated as an unconditional design goal. Instead, systems should privilege *grounded specificity*, where suggestions are explicitly anchored in evidence from the draft and unsupported claims are screened or suppressed. Recent guardrail approaches point in this direction by introducing evaluator layers that check generated feedback before release [16]. In practical terms, this means that AI-generated feedback should cite or point to the relevant text span, and high-uncertainty or high-stakes judgments should remain teacher-only.

5.1.4. Consideration CR4: Supporting Student Agency While Keeping Teachers in Control (F1–F4, H4, O3–O4)

Educators valued manageable action sets (F1) and saw potential benefits for weaker writers (O3), but worried about misalignment with student needs (H4), infeasible actions due to language level or incorrect diagnoses (F4), and the risk of students treating AI output as authoritative. These concerns connect to broader work on feedback literacy and evaluative judgment, which emphasizes that learners need support to interpret, prioritize, and act on feedback rather than simply receive it [46,47]. Our findings suggest that student agency is better supported through teacher-mediated workflows than through direct, unfiltered automation. In such workflows, teachers remain responsible for selecting, revising, and contextualizing AI suggestions, while students are encouraged to treat feedback as revisable guidance rather than final judgment.

5.2. Educational and Stakeholder Implications with Design Recommendations

The four design considerations above have implications not only for AI-supported writing systems in general, but also for the different stakeholders who shape how such systems are adopted in practice. As shown in Table 9, the same empirical findings translate into different requirements for educators and schools, students and families, and AI and tool designers. Bringing these implications together in one subsection clarifies that responsible use is not simply a technical matter of improving generation quality; it is an educational and socio-technical problem involving pedagogy, trust, verification, and control.

 doi:10.20944/preprints202605.1545.v1

Table 9. Stakeholder interpretations and design implications for AI-supported primary writing feedback.

Stakeholder	Interpretation (from educator evidence)	Design implications
Educators & schools	Prioritize workload relief for routine checks while retaining authority over instructional focus, developmental appropriateness, tone, and curriculum alignment. Model output is acceptable only when verification is fast and responsibility remains with the educator.	Adopt an *AI-as-draft, teacher-as-editor* workflow with lightweight verification (e.g., check that each claim is supported by the draft; remove/replace misleading praise/criticism). Use rubric-linked prompt templates and maintain a shared "failure log" (unsupported claims, misalignment, inappropriate developmental judgments) to guide ongoing refinement.
Students & families	Value readable, encouraging feedback but may struggle to detect generic or incorrect suggestions. Trust depends on clarity about when AI is involved and what has been educator-checked.	Embed basic *feedback literacy* activities (e.g., identify evidence for a comment; revise generic advice into a concrete next step; compare tutor vs. AI feedback). Communicate when AI is used, what is educator-mediated, and how concerns can be raised; calibrate feedback depth to student readiness.
AI & tool designers	Need to scale feedback while maintaining controllability and alignment with local curricula, genres, and reading levels; specificity must be grounded to avoid misleading output.	Constrain generation with curriculum- and rubric-aligned templates; require each suggestion to reference a supporting text span. Provide controls for tone, reading level, and feedback depth; add automated checks for unsupported or off-topic claims before educator review.

Note. Synthesised from the four design considerations (CR1-CR4) discussed in Section 5.

For educators and schools, the findings suggest that LLM feedback is most useful when treated as an editable draft rather than a student-ready product. A practical implementation model is an *AI-as-draft, teacher-as-editor* workflow in which educators quickly verify that each suggestion is grounded in the draft, aligned with the intended learning goal, and appropriate for the learner's developmental stage. This shifts teacher-in-the-loop from a general principle to a specific professional routine. It also suggests that sustainable adoption may depend on shared calibration practices, rubric-linked prompt templates, and repositories of example successes and failures that help educators refine acceptable uses over time.

For students and families, the findings indicate that introducing AI feedback into primary writing support should be accompanied by explicit scaffolds for interpretation. Because AI feedback may be readable but still generic or inaccurate, students need support to evaluate which suggestions are trustworthy and worth acting on. Classroom or tutoring activities that ask learners to compare feedback sources, identify evidence for a suggestion, or decide which comment is most useful may help build feedback literacy and reduce over-reliance on AI-generated advice [46,47]. Clear communication about when AI has been used and how teacher review shapes final feedback may also support appropriate trust.

For AI and tool designers, the most important design challenge is not generating more feedback, but generating feedback that is controllable, grounded, and efficient to verify. Our findings suggest that systems should support evidence-linked suggestions, controls for tone and reading level, and interfaces that help teachers accept, edit, or suppress comments quickly. Prompt structure should be treated as a meaningful part of the educational intervention rather than an invisible implementation detail, since it directly shapes what kinds of feedback the model produces [15]. More broadly, educational AI tools should be evaluated not only by average ratings or efficiency gains, but also by whether they reduce high-risk failure modes and verification burden in authentic instructional workflows.

5.2.1. Design Recommendations

Based on these findings, we recommend that AI-supported primary writing feedback be implemented through teacher-mediated workflows rather than direct student-facing automation. In practice, this means using the LLM to generate a draft set of comments, requiring each suggestion to be

 doi:10.20944/preprints202605.1545.v1

grounded in evidence from the student text, and asking educators to verify alignment with the intended learning goal before release. We further recommend that systems prioritize one or two high-leverage conceptual revision goals before lower-level corrections, provide student-facing feedback in simple language with optional teacher-facing rationale, and include supports for feedback literacy so that students learn to evaluate rather than simply accept AI suggestions. Together, these recommendations translate the study's findings into actionable guidance for safer and more educationally meaningful use of LLM-generated writing feedback.

Table 9 shows that responsible use is not a single design choice but a negotiated set of constraints across the educational ecosystem. Educators require rapid verifiability and pedagogical control; students require interpretive support, transparency, and feedback literacy; and designers must build for grounding, editability, and manageable oversight rather than assuming that fluent output is sufficient. The contribution of this study, therefore, lies not only in comparing tutor and LLM feedback but in showing how AI-generated feedback must be embedded in teacher-controlled, educationally meaningful workflows to be considered viable in primary writing support.

5.3. Hybrid Human-AI Feedback Pipeline

Building on the implications above, we propose a cautious hybrid pipeline in which the LLM functions as a draft generator and the educator remains the editor, verifier, and pedagogical decision-maker. In this workflow, the model first produces rubric-aligned draft feedback linked to evidence in the student text. Automated checks or guardrails can then flag unsupported claims, off-task comments, or higher-risk judgments for teacher attention [16]. Educators subsequently review, revise, and prioritize the draft output into a small number of high-leverage next steps before it is released to students. Students, in turn, receive concise and developmentally appropriate feedback that supports both immediate revision and broader writing development.

This hybrid framing is important because it recasts the role of AI in writing education from autonomous evaluator to teacher-support tool. The goal is not to replace teacher judgment, but to provide a faster and more structured starting point for feedback generation. In this sense, the educational value of AI-generated feedback lies less in whether it achieves "near-human" ratings on average, and more in whether it can be embedded into sustainable teacher-controlled routines that preserve pedagogical quality while reducing avoidable workload.

The proposed pipeline also clarifies where responsibility should remain visible in AI-supported writing instruction. Models may contribute speed, consistency, and accessible phrasing, but educators remain accountable for checking draft grounding, aligning feedback with curricular goals, and ensuring developmental appropriateness before comments reach students. This makes teacher mediation not an optional safeguard, but a core design feature of responsible deployment in primary writing contexts.

5.4. Limitations

Model and prompt-specific. Our findings concern a particular configuration of LLM feedback: one closed, general-purpose model (GPT-4 Turbo, `gpt-4-turbo-2024-04-09`), prompted in a specific template and evaluated at a single point in time. Different model snapshots, providers, or prompting strategies could shift the balance of strengths and weaknesses. We therefore interpret the results as patterns of behavior (e.g., surface focus, hallucinated specificity, verification burden) rather than immutable properties of GPT-4.

Narrow curricular and age context. The study focused on one year level (Year 5) and two genres (persuasive and narrative) within a single curriculum and online tutoring program. Feedback norms, genre expectations, and developmental goals differ across year levels, subjects, and education systems. Our considerations and recommendations should thus be seen as hypotheses about K-12 writing more broadly, not as directly generalizable to all grades or disciplines.

Evaluator sample and rubric subjectivity. Evaluation was conducted by four English specialists from the partner organization. Their expertise in the Australian Curriculum and local marking practices strengthens ecological validity but also reflects shared institutional norms. The six-dimensional rubric

 doi:10.20944/preprints202605.1545.v1

was co-designed with these educators, which ensures alignment with practice but may embed local pedagogical assumptions. Teachers in other settings may weigh criteria differently or place distinct tolerances on AI error.

No direct student outcomes or workload measurements. We evaluated feedback *as text*, not its downstream impact on students. The study did not measure how learners revised in response to different feedback sources, nor the effects on motivation, trust, or longer-term writing development. Similarly, we did not empirically quantify teacher time costs for verifying AI comments versus writing feedback from scratch. Claims about learning and workload, therefore, draw on educator perceptions and prior literature, not behavioral or longitudinal data.

Methodological scope. Finally, although we used paired inferential tests and thematic analysis, our sample size (56 scripts) limits statistical power, and our qualitative coding privileged depth over multiple coders and inter-rater statistics. The codebook and themes are reflexive interpretations grounded in educator comments, not an exhaustive taxonomy of all possible feedback behaviors.

6. Conclusions

This study examined how GPT-4 Turbo feedback compares with tutor feedback for Year 5 persuasive and narrative writing, and how educators judge the pedagogical appropriateness, risks, and practical value of such feedback in an authentic tutoring workflow. Across six educator-defined dimensions, tutor feedback received slightly higher ratings than GPT-4 Turbo feedback, with the clearest descriptive advantage in perceived helpfulness. At the same time, the quantitative differences were modest, indicating that LLM-generated feedback can approximate tutor feedback on some surface-facing qualities, particularly clarity and basic feasibility.

The qualitative findings, however, show that similar overall ratings can conceal important pedagogical differences. Educators consistently described GPT-4 Turbo feedback as clear, concise, and accessible, but also as more generic, more surface-oriented, and occasionally misaligned with the student draft. Tutor feedback, in contrast, was seen as more context-sensitive, developmentally attuned, and better aligned with students' immediate learning priorities, even though it could at times be overly dense or insufficiently specific. These findings suggest that the educational value of AI-generated feedback cannot be judged by fluency or readability alone; it must also be evaluated in relation to grounding, prioritization, developmental appropriateness, and the verification burden it introduces for educators.

Taken together, the study contributes an educator-aligned evaluation framework for LLM-generated writing feedback in primary education and identifies boundary conditions for its responsible use. In this context, LLM feedback was judged most appropriate as rapid first-pass support for routine structure and surface-level revision, and least appropriate for developmental judgment, context-sensitive guidance, and higher-order diagnosis without educator mediation. These findings support a hybrid *AI-as-draft, teacher-as-editor* workflow in which LLMs are used as constrained teacher-support tools rather than autonomous tutors. More broadly, the paper argues that the promise of AI-supported writing feedback lies not in replacing teacher judgment, but in embedding AI within sustainable, teacher-controlled workflows that preserve pedagogical quality while reducing avoidable workload.

Author Contributions: Dan Zhang: Conceptualization, Methodology, Study design, Data curation, Software, Formal analysis, Investigation (study conduct), Visualization, Writing– original draft, Writing– review & editing. Thuong Hoang: Conceptualization, Methodology, Study design, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing– review & editing. Ye Zhu: Supervision, Methodology, Validation, Writing– review & editing. Rui Wang: Conceptualization, Resources, Study design, Supervision, Validation, Writing– review & editing. Paula Crouch: Study design, Resources, Data collection, Investigation (study conduct), Validation, Writing– review & editing. Yi Wang: Writing– review & editing.

Funding: This project is made possible by CSIRO's Next Generation Emerging Technologies Graduates Program (GA221786) funded by the Australian Government.

 doi:10.20944/preprints202605.1545.v1

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the findings of this study were provided by a third-party ed ucational partner (Kinetic Education) and include sensitive student writing. Due to ethics approval requirements and confidentiality restrictions, the data are not publicly available.

Acknowledgments: This project is made possible by CSIRO's Next Generation Emerging Technologies Graduates Program (GA221786) funded by the Australian Government.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Educator-Aligned Qualitative Codebook

Table A1 presents the complete educator-aligned qualitative codebook used to analyse evaluator comments on GPT-4 Turbo and tutor feedback. The codes are organised according to six rubric dimensions: clarity, feasibility, helpfulness, relevance, specificity, and overall effectiveness.

Table A1. Full educator-aligned qualitative codebook.

Code	Dimension	Label	Description with example paraphrase
CL1	Clarity	Clear language	Straightforward, age-appropriate language. *E.g., "Language is simple."*
CL2	Clarity	Too advanced	Vocabulary/phrases too sophisticated for Year 5 (e.g., "coherence").
CL3	Clarity	Confusing	Internally inconsistent or contradictory explanations.
CL4	Clarity	Undefined terms	Uses technical terms (e.g., "speech tags") without definition.
CL5	Clarity	Overlong	Content is clear but too dense for a weaker writer to process.
CL6	Clarity	Generic	Language is clear but lacks draft-specific nuance or detail.
CL7	Clarity	Redundant	The same idea is repeated several times without extra value.
CL8	Clarity	Misaligned example	Examples do not match the point, hurting trust and clarity.
FE1	Feasibility	Manageable	Actions are clearly signposted or paired with highlighted student text.
FE2	Feasibility	Overwhelming	Too many points or expectations; the student is likely to feel overwhelmed.
FE3	Feasibility	Vague	Requests lack concrete steps; student knows *what* to change but not *how*.
FE4	Feasibility	Missing examples	No text-based demonstrations of what improved sentences might look like.
FE5	Feasibility	Advanced language	Vocabulary (e.g., "convey") is closer to teacher discourse than student language.
FE6	Feasibility	Hallucinated issues	Feedback is based on wrong issues (e.g., claims there are no paragraphs).
FE7	Feasibility	Surface-only	Changes are feasible but confined to minor edits (spelling, punctuation).
HE1	Helpfulness	Growth-oriented	Concrete suggestions to improve the draft and long-term writing skills.

Continued on next page

 doi:10.20944/preprints202605.1545.v1

Table A1. Cont.

Code	Dimension	Label	Description with example paraphrase
HE2	Helpfulness	Low-value	Broad advice like "add more detail" without indicating where or how.
HE3	Helpfulness	Missing key issue	A major learning need (e.g., tense control) is not addressed at all.
HE4	Helpfulness	Technical focus	Limited to spelling/grammar; does not help with ideas or structure.
HE5	Helpfulness	Misinterpreted needs	Encourages something educators see as problematic (e.g., praising repetition).
HE6	Helpfulness	Local edits	Improves the specific piece but unlikely to develop broader competence.
HE7	Helpfulness	Low scaffolding	Feedback is not broken down enough for a weak writer to act independently.
RE1	Relevance	Targets real issues	Accurately identifies core problems (e.g., plot coherence, argument depth).
RE2	Relevance	Peripheral issues	Focuses on low-stakes issues (e.g., spacing) while ignoring main problems.
RE3	Relevance	Hallucination	Claims an error that is not present, leading to confusion.
RE4	Relevance	Technical only	Relevant to technical correctness but misses content-level concerns.
RE5	Relevance	Partial relevance	Broadly relevant, but the most important learning need is unaddressed.
RE6	Relevance	Misleading praise	The "well done" section praises features that are actually absent.
SP1	Specificity	Text-grounded	Cites specific sentences or phrases from the student's text.
SP2	Specificity	Generic strengths	Positive comments are vague with no supporting examples.
SP3	Specificity	Specific/Generic mix	Improvement points are specific but praise remains generic.
SP4	Specificity	No examples	Little or no quotation from text, even when suggesting key changes.
SP5	Specificity	Specific but wrong	Cites an example that is misread, fabricated, or mislabeled.
SP6	Specificity	Missed content	Could have drawn on plot details to make guidance specific but does not.
OE1	Overall	Technical/Conceptual	Primarily improves surface correctness; fails to engage with conceptual goals.
OE2	Overall	Missed priority	Fails to address what educators see as the key developmental priority.
OE3	Overall	Verification required	Feedback needs substantial checking/rewriting before being shared.
OE4	Overall	Over-praising	Too flattering, which may reduce student motivation to revise.
OE5	Overall	Style mismatch	Uses American spelling or conventions not aligned with local curriculum.

Continued on next page

 doi:10.20944/preprints202605.1545.v1

Table A1. Cont.

Code	Dimension	Label	Description with example paraphrase
OE6	Overall	Worse than tutor	Evaluators explicitly judge tutor feedback as more effective overall.
OE7	Overall	Better for weak	Judge GPT as more appropriate for very basic writers (simpler language).
OE8	Overall	AI assistant	Useful starting point for teacher editing, but not a standalone tool.
OE9	Overall	Hallucinations	Considered unusable because of frequent incorrect claims.

Appendix B. Comparative Feedback Example

To illustrate the practical application of the rubric and the qualitative differences between human and AI-generated guidance, Table A2 presents a complete Year 5 persuasive first draft alongside the paired tutor and GPT-4 Turbo feedback used in the evaluation study.

The example demonstrates that while both feedback sources utilize a constructive three-part structure (strengths, improvements, and next steps), their focus diverges: the human tutor emphasizes formal persuasive structure and curriculum-aligned terminology (e.g., "state your claim," "body paragraphs"), whereas GPT-4 Turbo prioritizes readability, encouraging language, and localized sentence-flow suggestions.

Table A2. Example Year 5 persuasive draft with paired tutor and GPT-4 Turbo feedback.

Student draft (first draft)	Tutor feedback	GPT-4 Turbo feedback
I think that children should be able to earn pocket money, because they can use the money on snacks, video games and that can help them be happy because they can finally buy the things that they can't afford. Children should be able to earn up to $2-5 by sweeping and $3-8 mopping. Why I think it's good for kids to buy whatever they want is because they have their own freedom to use their money that they hardly earned and own. Why I don't think that kids are able to buy whatever they want is because if you buy a subscription, then it would take money every month and there might be a chance that they can't pay their subscription back. Then their parents will be angry. I strongly prefer letting children buy whatever they want because they own the money, and they care for what they bought because their parents should have told their child to care and that is my reason of why I think children can earn money and buy whatever they want like me.	*Persuasive – "Pocket Money" (first draft)* **What you did well** • Good effort in relating your essay to the prompt you were given. • You have structured your essay into different paragraphs. • Your ideas were supported with some elaboration. • All basic punctuation was used correctly. **To improve this piece** • Try to find a title that matches your argument. • Follow a persuasive structure: introduction, opinion, and main arguments. • In the body paragraphs, state your claim, give evidence, and comment on support. • Expand your final draft to fit this structure. • Include more evidence to support your points. • Keep a dictionary nearby for spelling. **Next steps** Well done on submitting your first draft. When writing your final draft, expand the essay to fit the persuasive structure, add research-based evidence, and remember to come up with a title.	**What you did well** • You start with a clear reason why kids should earn pocket money. • You organize your ideas into paragraphs. • You give reasons such as freedom and caring for what they buy. • You mostly remember to use full stops and capital letters. **To improve this piece** 1. *Conclusion*: Add a few sentences at the end that remind the reader what your essay was about. 2. *Persuasive devices*: Use stronger feeling words and questions. 3. *Spelling*: Check tricky words. 4. *Sentence structure and flow*: Break long sentences into shorter ones. **Next steps** Good job so far! Strengthen your ending, use more persuasive words, check spelling carefully, and read your sentences aloud to see if they flow well.

Text reproduced from the study dataset. Feedback followed the same three-part structure across conditions.

 doi:10.20944/preprints202605.1545.v1

References

1. Wei, P.; Wang, X.; Dong, H. The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology* **2023**, *14*, 1249991.
2. Benali, A. The impact of using automated writing feedback in ESL/EFL classroom contexts. *English Language Teaching* **2021**, *14*, 189–195.
3. Liu, M.; Li, Y.; Xu, W.; Liu, L. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies* **2016**, *10*, 502–513.
4. Su, Y.; Lin, Y.; Lai, C. Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing* **2023**, *57*, 100752.
5. Alharbi, W. E-feedback as a scaffolding teaching strategy in the online language classroom. *Journal of Educational Technology Systems* **2017**, *46*, 239–251.
6. Potter, A.; Wilson, J. Statewide implementation of automated writing evaluation: Analyzing usage and associations with state test performance in grades 4–11. *Educational Technology Research and Development* **2021**, *69*, 1557–1578.
7. Link, S.; Mehrzad, M.; Rahimi, M. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning* **2022**, *35*, 605–634.
8. Amoozadeh, M.; Daniels, D.; Nam, D.; Kumar, A.; Chen, S.; Hilton, M.; Srinivasa Ragavan, S.; Alipour, M.A. Trust in generative AI among students: An exploratory study. In Proceedings of the Proceedings of the 55th ACM Technical Symposium on Computer Science Education V.1, 2024, pp. 67–73.
9. Han, A.; Zhou, X.; Cai, Z.; Han, S.; Ko, R.; Corrigan, S.; Peppler, K.A. Teachers, parents, and students' perspectives on integrating generative AI into elementary literacy education. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–17.
10. Park, H.; Ahn, D. The promise and peril of ChatGPT in higher education: Opportunities, challenges, and design implications. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–21.
11. Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P.S.; Wen, Q. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105* **2024**.
12. Dai, W.; Tsai, Y.S.; Lin, J.; Aldino, A.; Jin, H.; Li, T.; Gašević, D.; Chen, G. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence* **2024**, *7*, 100299.
13. Aldino, A.A.; Tsai, Y.S.; Gupte, S.; Henderson, M.; Nath, D.; Gašević, D.; Chen, G. Analytics of learner-centered feedback: A large-scale case study in higher education. *Computers & Education* **2025**, *237*, 105360.
14. Dai, W.; Cheng, Y.; Aldino, A.A.; Tsai, Y.S.; Gašević, D.; Chen, G. Evaluating the capability of large language models in characterising relational feedback: A comparative analysis of prompting strategies. *Computers and Education: Artificial Intelligence* **2025**, *8*, 100427.
15. AlGhamdi, E.; Li, Y.; Gašević, D.; Chen, G. Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education* **2025**, p. 105511.
16. Qian, K.; Cheng, Y.; Guan, R.; Dai, W.; Jin, F.; Yang, K.; Nawaz, S.; Swiecki, Z.; Chen, G.; Yan, L.; et al. Dean of llm tutors: exploring comprehensive and automated evaluation of llm-generated educational feedback via llm feedback evaluators. *arXiv preprint arXiv:2508.05952* **2025**.
17. Attali, Y.; Burstein, J. Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment* **2006**, *4*.
18. Ramesh, D.; Sanampudi, S.K. An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review* **2022**, *55*, 2495–2527.
19. Taghipour, K.; Ng, H.T. A neural approach to automated essay scoring. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1882–1891.
20. Schmaltz, A.; Kim, Y.; Rush, A.M.; Shieber, S.M. Sentence-level grammatical error identification as sequence-to-sequence correction. *arXiv preprint arXiv:1604.04677* **2016**.
21. Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.S.; Gašević, D.; Chen, G. Can large language models provide feedback to students? A case study on ChatGPT. In Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). IEEE, 2023, pp. 323–325.
22. Naismith, B.; Mulcaire, P.; Burstein, J. Automated evaluation of written discourse coherence using GPT-4. In Proceedings of the Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 2023, pp. 394–403.

 doi:10.20944/preprints202605.1545.v1

23. Zhang, D.; Hoang, T.; Zhu, Y.; Wang, R.; Crouch, P. Generating Feedback for School Students Essay with Large Language Models. In Proceedings of the International Conference on Knowledge Science, Engineering and Management. Springer, 2025, pp. 319–330.
24. Costa, K.; Mfolo, L.N.; Ntsobi, M.P. Challenges, benefits and recommendations for using generative artificial intelligence in academic writing: A case of ChatGPT. *Medicon Engineering Themes* **2024**, *7*, 3–38.
25. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.
26. Fabbri, A.R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; Radev, D. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 391–409.
27. Steiss, J.; Tate, T.; Graham, S.; Cruz, J.; Hebert, M.; Wang, J.; Moon, Y.; Tseng, W.; Warschauer, M.; Olson, C.B. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction* **2024**, *91*, 101894.
28. Atasoy, A.; Moslemi Nezhad Arani, S. ChatGPT: A reliable assistant for the evaluation of students' written texts? *Education and Information Technologies* **2025**, pp. 1–31.
29. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
30. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text summarization branches out, 2004, pp. 74–81.
31. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* **2019**.
32. Liaqat, A.; Munteanu, C.; Demmans Epp, C. Collaborating with mature English language learners to combine peer and automated feedback: A user-centered approach to designing writing support. *International Journal of Artificial Intelligence in Education* **2021**, *31*, 638–679.
33. Tan, K.; Pang, T.; Fan, C.; Yu, S. Towards applying powerful large AI models in classroom teaching: Opportunities, challenges and prospects. *arXiv preprint arXiv:2305.03433* **2023**.
34. Jia, Q.; Young, M.; Xiao, Y.; Cui, J.; Liu, C.; Rashid, P.; Gehringer, E. Insta-Reviewer: A data-driven approach for generating instant feedback on students' project reports. *International Educational Data Mining Society* **2022**.
35. Xu, B.; Bai, Y.; Sun, H.; Lin, Y.; Liu, S.; Liang, X.; Li, Y.; Gao, Y.; Huang, H. EduBench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160* **2025**.
36. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education* **2019**, *16*, 1–27.
37. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* **2023**, *103*, 102274.
38. Weisz, J.D.; He, J.; Muller, M.; Hoefer, G.; Miles, R.; Geyer, W. Design principles for generative AI applications. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–22.
39. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for human–AI interaction. In Proceedings of the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–13.
40. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018, pp. 80–89.
41. Han, J.; Yoo, H.; Myung, J.; Kim, M.; Lim, H.; Kim, Y.; Lee, T.Y.; Hong, H.; Kim, J.; Ahn, S.Y.; et al. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student–LLM interaction. *arXiv preprint arXiv:2310.05191* **2023**.
42. Kim, H.; Baghestani, S.; Yin, S.; Karatay, Y.; Kurt, S.; Beck, J.; Karatay, L. ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. *Exploring Artificial Intelligence in Applied Linguistics* **2024**, pp. 73–95.

 doi:10.20944/preprints202605.1545.v1

43. Yazici, A.; Mejia-Domenzain, P.; Frej, J.; Käser, T. GELEX: Generative AI-hybrid system for example-based learning. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–10.
44. Shute, V.J. Focus on formative feedback. *Review of Educational Research* **2008**, *78*, 153–189.
45. Nicol, D.J.; Macfarlane-Dick, D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education* **2006**, *31*, 199–218.
46. Carless, D.; Winstone, N. Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education* **2023**, *28*, 150–163.
47. Tai, J.; Ajjawi, R.; Boud, D.; Dawson, P.; Panadero, E. Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education* **2018**, *76*, 467–481.
48. Hattie, J.; Timperley, H. The power of feedback. *Review of Educational Research* **2007**, *77*, 81–112.
49. Harris, K.R.; Graham, S.; Mason, L.H. Improving the writing, knowledge, and motivation of struggling young writers: Effects of self-regulated strategy development with and without peer support. *American Educational Research Journal* **2006**, *43*, 295–340.
50. Graham, S.; Perin, D. Writing next: Effective strategies to improve writing of adolescents in middle and high schools. Carnegie Corporation of New York, 2007.
51. OpenAI. GPT-4 technical report. https://openai.com/research/gpt-4, 2023.
52. Lee, S.; Cai, Y.; Meng, D.; Wang, Z.; Wu, Y. Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv preprint arXiv:2404.04941* **2024**.
53. Mannila, L. Co-designing AI literacy for K–12 education. In Proceedings of the Proceedings of the 19th WiPSCE Conference on Primary and Secondary Computing Education Research, 2024, pp. 1–3.
54. Wang, Z.; Makarova, V.; Li, Z.; Kodner, J.; Rambow, O. LLMs can perform multi-dimensional analytic writing assessments: A case study of L2 graduate-level academic English writing. *arXiv preprint arXiv:2502.11368* **2025**.
55. Ranalli, J. Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning* **2018**, *31*, 653–674.
56. Holstein, K.; McLaren, B.M.; Aleven, V. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Grantee Submission* **2019**.
57. Lin, P.; Van Brummelen, J. Engaging teachers to co-design integrated AI curriculum for K–12 classrooms. In Proceedings of the Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–12.