

Evaluating Feedback Focus and Pedagogical Adaptivity in LLM-Generated Feedback on Student Writing

Norah Almousa, Shayan Peyghambari Oskoui,
Raquel Coelho, Gayle Rogers,
Xiang Lorraine Li, Diane Litman
University of Pittsburgh
{nia135, dlitman}@pitt.edu

Abstract

We investigate whether state-of-the-art large language models (LLMs) generate feedback that reflects the pedagogical practices of expert teachers in terms of feedback focus and adaptivity. Previous evaluation efforts have examined feedback characteristics, its impact on learning, and its target, yet the focus of feedback and its adaptivity remains largely overlooked. To bridge this gap, we adopt and refine Narciss's taxonomy into seven feedback focus types to annotate teacher and LLM-generated feedback across three university writing courses. We release FEEDTYPE, a benchmark containing annotated teacher and LLM feedback from six LLMs under three prompting strategies. We assess the coverage and distribution of feedback focus types, and examine whether LLMs adapt their feedback across draft stages and student performance levels as an expert instructor does. Our findings show that while most LLMs cover most feedback focus types, they fail to reflect teacher feedback distributions and show varying levels of adaptivity, with none matching the teachers' adaptive behavior. We believe FEEDTYPE will support future research on pedagogical alignment in LLM feedback generation.

1 Introduction

Large language models (LLMs) have emerged as promising tools for providing personalized feedback on student essays at scale, with recent work demonstrating success across creative (Rashkin et al., 2025), second language (Wang et al., 2025b), and academic (Steiss et al., 2024) writing. However, research across multiple educational tasks, including tutoring (Tack and Piech, 2022), math problem solving (Macina et al., 2023) and essay feedback (Dai et al., 2023) has shown that LLMs fail to capture the pedagogical practices expert teachers use to guide student learning. For example, in tutoring, Macina et al. (2023) found that ChatGPT tends to provide direct answers, while expert teachers

guide students through hints. This has motivated research on *pedagogical alignment* (Macina et al., 2023, 2025; Sonkar et al., 2024; Maurya et al., 2025), which measures how well LLM behavior aligns with expert teaching (Perczel et al., 2025).

Recently, a growing body of work has begun incorporating pedagogical dimensions into LLM-generated feedback evaluation, drawing from established feedback models that characterize effective teacher feedback (Hattie and Timperley, 2007; Yang and Carless, 2013). For example, Seßler et al. (2025) and Dai et al. (2023) applied Hattie and Timperley (2007)'s framework, examining feedback's learning supportive components and who feedback targets. Mah et al. (2025) applied Yang and Carless (2013)'s framework, focusing on how feedback is delivered across cognitive, social, and structural dimensions. However, *feedback focus* – what teachers focus on when responding to student writing (Yu and Lee, 2014; Yu, 2014; Lipnevich and Panadero, 2021), remains unexplored. This is important since teachers vary their feedback focus to address different aspects of student writing (Lipnevich and Panadero, 2021), yet LLMs have shown limited ability to identify writing issues (Rashkin et al., 2025; Seßler et al., 2025), motivating us to examine what feedback focus types LLMs tend to provide and how this compares to expert teachers.

The feedback literature has classified feedback focus into different levels, ranging from broad distinctions such as content versus form (Ferris, 1997) to more detailed taxonomies distinguishing specific types such as mistakes, tasks, and concepts (Narciss, 2008a). Recognizing multiple feedback focus types is pedagogically important, as it informs what to address, for whom, and when. Expert teachers leverage this knowledge by shifting their focus across draft stages and adapting to each student's individual needs (Ortmeier-Hooper, 2006; Ferris, 1997), reflecting a deeply adaptive teaching practice. To our knowledge, feedback focus

has not been examined in the evaluation of LLM-generated feedback, specifically in terms of what aspects LLMs tend to focus on and whether they exhibit the same adaptive behavior across draft stages and student performance levels. We aim to bridge this gap by addressing three research questions:

RQ1: *What feedback focus types do LLMs tend to generate, and how does this compare to expert teachers?*

RQ2: *Do LLMs adapt their feedback focus across draft stages, as an expert teacher does?*

RQ3: *Do LLMs adapt their feedback focus based on student performance, as an expert teacher does?*

To address these questions, we present FEEDTYPE¹, a dataset of expert teacher feedback on student essays across three university writing courses (taken from the publicly available SEFORA corpus (Oskoui et al., 2026)), with teacher and LLM-generated feedback annotated for feedback focus. While SEFORA provides rich teacher feedback data, it lacks feedback focus annotations needed to address our research questions. To bridge this gap, we adapt an existing taxonomy (Narciss, 2008a) and annotate all feedback comments in FEEDTYPE according to feedback focus. We then prompt six LLMs, covering open and closed weight models as well as the LearnLM specialized pedagogical model (Team et al., 2024), to generate feedback on selected paragraphs from FEEDTYPE using three prompting strategies. We annotate the LLM generated feedback using the same taxonomy applied to teacher feedback, enabling a direct comparison. FEEDTYPE thus serves as a benchmark for analyzing LLM pedagogical practices in terms of feedback focus and whether it adapts to student performance and draft stage. Our findings indicate that while most LLMs cover the majority of feedback focus types, they fail to reflect teacher feedback distributions and show limited adaptivity across draft stages and student performance levels.

2 Related Work

LLMs for Feedback Generation on Writing

LLMs have been used for feedback generation across multiple writing contexts, including L2 learner writing (Wang et al., 2025b; Kaneko and Okazaki, 2024; Mizumoto et al., 2024), creative writing (Rashkin et al., 2025), and academic writing (Chu et al., 2026). Feedback generation has

been explored at the sentence level (Mizumoto et al., 2024; Song et al., 2024; Kaneko and Okazaki, 2024; Bannò et al., 2024) and on complete student essays (Chu et al., 2026; Behzad et al., 2024a; Rashkin et al., 2025). However, to our knowledge, feedback generation at the paragraph level has received very little attention, despite evidence that this is a common and valuable form of expert teacher feedback (Mah et al., 2025; Hyland and Hyland, 2006; Zamel, 1985) that LLMs fail to provide (Mah et al., 2025). We address this gap by generating paragraph level feedback across multiple writing domains (L2 learner writing, composition, and fiction), where we provide LLMs with the complete student essay and target specific paragraphs within it for feedback generation.

Evaluation of LLM-Generated Feedback

Many studies have evaluated LLM-generated feedback in terms of accuracy (Yoon et al., 2023), quality (Algobaei and Alzain, 2026), and bias (Du et al., 2025; Tan et al., 2026b). Fewer have examined its alignment with expert teacher practices. In student writing, prior work has compared LLM and teacher generated feedback in terms of quality (Steiss et al., 2024), content and targets (Dai et al., 2023), and diversity (Tan et al., 2026a). We extend prior work on alignment in two ways. First, we examine feedback focus and its adaptivity, dimensions that have received limited attention in prior evaluations. Second, we compare the distribution of feedback focus types between teachers and LLMs using population-level evaluation approaches adopted in prior educational research (Mah et al., 2025; Kucheria et al., 2025; Asano et al., 2025).

Feedback Focus and Adaptivity Prior work has classified feedback focus narrowly, either by grammatical error types or limited broad categories (Nagata et al., 2020; Steiss et al., 2024; Liu et al., 2025). To address this gap, we adopt and refine the taxonomy of Narciss (2008a), which captures a broader range of feedback focus types observed in expert teacher feedback. This taxonomy allows us to examine how teachers adapt their feedback priorities across draft stages and learner performance levels. For example, early drafts may receive greater attention to content and ideas, whereas later drafts may focus more on vocabulary and style (Zamel, 1985). Similarly, feedback may differ by learner performance level (Ferris, 1997) because students at different levels may benefit from different types

¹Data available at <https://github.com/nia135/FeedType>.

Corpus	Domain	Feedback Granularity	What Annotated	Feedback	Scores	Revisions
L2 Learner Writing						
TOEFL11 (Blanchard et al., 2013)	L2	Essay	Holistic English proficiency scores (low, medium, high)	No	Yes	No
CLC-FCE (Yannakoudakis et al., 2011)	L2	Sentence	Grammatical error annotations and corrections	No	Yes	No
EFCAMDAT (Geertzen et al., 2013)	L2	Essay	English proficiency levels and learner metadata	No	Yes	Yes
Write & Improve (Nguyen et al., 2018)	L2	Essay	Grammatical error corrections, CEFR proficiency levels, and learner revisions	Automated	Yes	Yes
LEAF (Behzad et al., 2024b)	L2	Essay	Personalized holistic writing feedback and revision annotations	Teacher & Automated	No	Yes
ICNALE (Ishikawa, 2013)	L2	Essay + Sentence	Proficiency labels and grammatical error correction annotations	No	Yes	No
FCG Corpus (Nagata et al., 2020)	L2	Sentence	Feedback comments explaining grammatical errors and writing rules	Teacher	No	No
Academic Writing						
ASAP++ (Mathias and Bhattacharyya, 2018)	Native	Essay	Holistic and trait-specific essay scores (content, organization, word choice, sentence fluency, etc.)	No	Yes	No
ASAP 2.0 (Crossley et al., 2025)	Native	Essay	Holistic essay scores for source-based argumentative essays	No	Yes	No
Creative Writing						
STORYFEEDBACK (Rashkin et al., 2025)	Creative	Story	Generated feedback annotated with human ratings (correctness, specificity, relevance, error detection)	Automated	No	No
SciFi-100 (Li et al., 2026)	Creative	Story	Novelty and creativity dimensions of generated stories	No	No	No
Towards a Novel Benchmark (Wang et al., 2025a)	Creative	Story	Narrative, paragraph, and sentence-level literary quality dimensions	No	No	No
Our Corpus						
FEEDTYPE	L2, Native, Creative	Paragraph	Feedback with feedback focus type labels	Teacher & Automated	Yes	Yes

Table 1: Overview of writing and feedback corpora across L2 learner, academic, and creative writing domains. FEEDTYPE is the only one that jointly provides scores, paragraph-level feedback, feedback focus type annotations, and revisions across all three domains.

of feedback. We refer to these adjustments in feedback focus as feedback adaptivity. To our knowledge, no prior work has examined whether LLMs demonstrate similar adaptive patterns.

Related Corpora Across L2, native, and creative writing, existing corpora are limited to scores only without feedback (Blanchard et al., 2013; Yannakoudakis et al., 2011; Geertzen et al., 2013; Ishikawa, 2013; Crossley et al., 2025; Li et al., 2026), sentence-level or holistic error-focused feedback (Nagata et al., 2020; Nguyen et al., 2018; Behzad et al., 2024b), multi-dimensional scoring and feedback without feedback focus type annotations (Mathias and Bhattacharyya, 2018), or essay evaluation without feedback (Wang et al., 2025a; Rashkin et al., 2025). To our knowledge, no existing corpus jointly provides scores, revisions, teacher and automatic feedback, and feedback focus type annotations, all of which are needed to analyze feedback focus and its adaptivity across

draft stages and student performance levels. We release FEEDTYPE, a corpus that provides all of these components, enabling such analysis. We summarize and compare existing datasets in Table 1.

3 FEEDTYPE & Annotation Procedure

Construction of FEEDTYPE We sampled course data from the publicly available SEFORA corpus of first-year college student essays and teacher feedback (Oskoui et al., 2026). In particular, we use the data from three different courses (composition, ESL, and fiction) from a single semester, choosing the semester with the highest student enrollment. Where a course had multiple classes in the same semester, we included all of them. The composition course includes two classes taught by the same instructor, while the ESL and fiction courses each consist of one class taught by different instructors. Students in the composition and ESL courses submitted up to three essays with multiple drafts, while fiction students submitted three essays without mul-

Paragraph: Additionally, I love it when we just talk about whatever new thing we are currently obsessed with because it changes quite frequently. I love finding new artists, television shows, and book series that I immediately show her so that she can become as obsessed as I have become. I love that we can just go to each other with whatever we just newly found and can be as obsessed and insane about the topic as we want. I will always adore and cherish the countless memories of learning every word from Hamilton and playing constantly in the car. When we finally watched it when it came out on Disney+, my father finally saw how all the songs interlocked with each other because he had been listening to it out of order for months. Another example is we both are obsessed with this television universe called Arrow verse which consists of multiple TV shows that revolve around the DC Comic superheroes. And we will just sit down and watch the shows and just enjoy being in each other's presence.

Teacher: Again, great details, but how do all of these details and ideas work together to develop one central focusing idea with value and purpose? Just as importantly: who is your audience for this essay? **Feedback Type: Task Constraints**

Mistral 7B: The paragraph could benefit from more specific details and personal examples to further develop the big idea of the shared interests and experiences between the author and their twin. **Feedback Type: Elaboration**

Llama-8B: The paragraph needs for more nuanced and varied sentence structures. The paragraph is dominated by simple sentences, which can make the writing feel a bit flat and repetitive. To add more interest and depth, consider varying sentence length and structure. **Feedback Type: Mistakes**

Gemini: The paragraph feels a bit repetitive with the phrase "I love" starting multiple sentences in a row. Try varying your sentence structure and word choice to create a smoother and more engaging reading experience for your audience. **Feedback Type: Mistakes**

Figure 1: Examples of teacher and LLM feedback on a student paragraph, annotated with feedback types.

multiple drafts; in all courses, only the first draft was required. We focus on paragraph-level feedback, where each paragraph receives one teacher comment. Student grades and assignment rubrics are used to support feedback classification and adaptivity analysis. Table 2 summarizes the corpus, and Figure 1 shows a feedback example. Additional corpus details can be found in Appendix A.

Course	Essay	Initial	Revision	Final
Composition (Class 1)	1	15/99	10/64	3/15
	2	15/82	9/55	0/0
	3	13/76	0/0	8/39
	Total	43/257	19/119	11/54
Composition (Class 2)	1	15/85	13/77	2/11
	2	14/70	10/49	4/13
	3	15/78	0/0	6/21
	Total	44/233	23/126	12/45
ESL Composition	1	14/148	–	12/38
	2	–	–	11/35
	3	13/37	–	13/99
	Total	27/185	–	36/172
Introduction to Fiction	1	18/276	–	–
	2	17/183	–	–
	3	18/316	–	–
	Total	53/775	–	–
Total Paragraphs		1,966		

Table 2: Essays and selected paragraphs with one feedback (#Essays / #Paragraphs).

Feedback Annotation Scheme We base our annotation scheme on Narciss (2008a)’s taxonomy, which provides detailed content focus types for feedback analysis. Prior frameworks either conflate content and function (Hattie and Timperley, 2007; Kulhavy and Stock, 1989) or categorize con-

tent only broadly (Ferris, 1997; Yu, 2014); Narciss’s taxonomy addresses both by treating them separately and offering more precise content focus types. We adapt and extend this taxonomy to better capture the nature of written teacher feedback, retaining relevant feedback types, refining others, and introducing new types to emphasize explanatory feedback. Appendix B presents the original and adapted taxonomies. The final taxonomy consists of seven feedback types: *Task Constraints*, which clarify assignment goals and requirements; *Concepts*, which explain key ideas needed to complete the task; *Elaboration*, which encourages students to expand their ideas with more detail; *Clarification*, which asks students to clarify unclear parts of their writing; *Mistakes*, which identify specific errors in the text; *Praise*, which provides encouragement or positive evaluation; and *Other*, which covers feedback that either does not fit any single focus type or spans multiple types simultaneously. Annotation guidelines are provided in Appendix C, with some annotated feedback examples in Figure 1.

Annotator Training and Development of Coding Manual

Two authors iteratively developed and refined the coding manual through four pilot rounds. In each round, 50 teacher feedback comments (25 composition, 12 ESL, 13 fiction) and their associated paragraphs were randomly sampled and independently labeled. Disagreements were analyzed, and category definitions and instructions were updated to reduce ambiguity. This process continued until substantial agreement was achieved, with the final round reaching Cohen’s Kappa $\kappa = 0.774$ (Sim and Wright, 2005).

Scaling Annotation Motivated by prior work demonstrating that LLMs can reliably annotate data at scale (Tan et al., 2024), we use them to automatically classify feedback according to our feedback focus types. We prompted six candidate models to classify the feedback (see Appendix F for the prompt) and evaluated their reliability using three methods. We constructed two evaluation datasets serving distinct purposes: (1) 100 randomly sampled feedback instances to measure inter-annotator agreement and assess whether LLM annotations can substitute for human annotators using the Alt-Test (Calderon et al., 2025), and (2) 400 feedback instances as a test set to evaluate whether conclusions drawn from LLM-generated annotations are consistent with those derived from human labels, ensuring the validity of our annotation pipeline. Gemini 2.0 Flash was selected as the annotator, passing all three evaluations: achieving a Cohen’s Kappa of 0.78 with human annotators, passing the Alt-Test, and producing conclusions consistent with those drawn from human annotations. Further details are provided in Appendix D.

4 Experiments

4.1 Generating Feedback with LLMs

We generated feedback using six models: **Mistral-7B-Instruct**, **Mistral-24B-Instruct** (Jiang et al., 2023), **LLaMA-3.1 8B** (AI, 2024), **Qwen2.5-70B-Instruct** (Hui et al., 2024), **Gemini 2 Flash** (Google DeepMind, 2025), and **LearnLM** (Team et al., 2024), covering both open and closed source instruction tuned models across different model sizes, to examine whether model size and type influence feedback practices. LearnLM is included as a specialized model fine-tuned for educational tasks (Team et al., 2024).

Each model was prompted under three conditions: (1) **zero-shot** (Rashkin et al., 2025; Stahl et al., 2024), (2) **zero-shot with the names and definitions of the feedback focus categories**, as providing task specific information has been shown to improve LLM output quality (Stahl et al., 2024), and (3) **few-shot with the names, definitions, and examples of each feedback focus type category**, as in-context examples have been shown to improve LLM performance (Brown et al., 2020). The goal is to examine whether providing feedback type information or examples shifts model behavior toward expert teacher practices. Our prompt follows Rashkin et al. (2025) and Stahl et al. (2024), in-

structing the model to act as a teacher and including the assignment rubric and task prompt. However, we designed the prompt to mirror real teacher context, which is often absent from prior work: (1) the full essay alongside the target paragraph, (2) the draft version, and (3) previously received feedback for revision and final drafts. The full prompt is in Appendix E, and Figure 1 shows example outputs.

4.2 Evaluation Methods

To examine the pedagogical alignment of LLM feedback focus types (**RQ1**), we compare them to expert teacher feedback focus types in two ways:

(i) **Coverage**: How much can LLMs cover teachers’ feedback types when responding to student essays? We compute **Cov_all**, inspired by Asano et al. (2025), defined as the number of teacher feedback focus types for which the LLM generates at least one instance, divided by the total number of teacher feedback focus types.

(ii) **Similarity**: How similar is the feedback focus type distribution of LLM-generated feedback to that of teachers? We compute **Jensen-Shannon Divergence (JSD)**, following prior work (Lu et al., 2020; Ma et al., 2025; Elangovan et al., 2025). JSD ranges from 0 to 1, where 0 indicates identical and 1 indicates completely different distributions:

$$JSD(P_m \| P_T) = \frac{1}{2} D_{KL}(P_m \| M) + \frac{1}{2} D_{KL}(P_T \| M) \quad (1)$$

where $M = \frac{1}{2}(P_m + P_T)$, P_m is the distribution of feedback focus types generated by model m , P_T is the teacher reference distribution, and D_{KL} is the Kullback-Leibler divergence.

To measure whether LLMs adapt their feedback focus types across draft stages as an expert teacher does (**RQ2**), we follow corpus linguistics practice of using chi-square-based tests to detect categorical distribution changes across conditions (McEnery and Hardie, 2011). Since chi-square does not accept zero values, and some models did not provide feedback on certain focus types, we apply **Fisher’s Exact Test** (Agresti, 1992), which is statistically equivalent but accepts zero values. For each feedback type f , source s , and pair of draft stages (d_i, d_j) , we construct a 2×2 contingency table:

$$\begin{pmatrix} c_{s,f}^{(d_i)} & c_{s,f}^{(d_j)} \\ c_{s,\neg f}^{(d_i)} & c_{s,\neg f}^{(d_j)} \end{pmatrix} \quad (2)$$

where $c_{s,f}^{(d)}$ is the count of feedback type f from source s at draft stage d , and $c_{s,\neg f}^{(d)}$ is the count of all other feedback types. The null hypothesis H_0 states that feedback type f and draft stage are independent, in other words, the proportion of f does not change between drafts. We test two pairwise comparisons per feedback type: initial vs. revision, and revision vs. final. We reject H_0 if $p < 0.05$.

To measure whether LLMs also adapt their feedback focus based on student performance as an expert teacher does (RQ3), we again use **Fisher’s Exact Test** as in RQ2. However, we now compare the distribution of each feedback focus type between high- and low-performing students.

5 Results

Inspired by prior work on behavioral alignment between LLMs and human tutors (Kucheria et al., 2025), we aggregate all teacher feedback to construct a reference distribution, as shown in Figure 2.

RQ1: What feedback focus types do LLMs tend to generate, and how does this compare to expert teachers? We address this question in two parts: in terms of coverage and similarity.

(1.1) How much can LLMs cover teachers’ feedback types when responding to student essays?

To understand whether each model can generate feedback across different focus types, we use the Cov_all metric, which measures how many of the seven feedback focus types a model produces. We start by analyzing the zero-shot baseline, where models receive no information about the feedback focus types. As shown in Table 3, LearnLM, Llama-7B, and Mistral-7B achieve full coverage. However, Gemini and Mistral-24B fail to produce praise-focused feedback, while Qwen-70B does not produce the Other feedback focus type. A model missing this feedback focus type generates feedback with only one focus per comment, while models that achieve full coverage demonstrate the ability to produce at least one comment that addresses multiple focus types simultaneously, or that is not categorizable as one of the provided focus types.

To guide the models toward more complete coverage, we modified the prompting by adding information about the feedback focus types and providing few-shot examples. As shown in Table 3, Mistral-24B benefits from both strategies - producing praise-focused feedback when prompted with either category information or examples, while Qwen-70B just benefits from examples. Gemini,

Model	Zero-shot	Zero-shot + Categories	Few-shot + Categories
Gemini	0.86 Praise	0.71 Praise, Other	0.71 Praise, Other
LearnLM	1.0 -	0.86 Other	0.86 Other
Qwen-70B	0.86 Other	0.86 Other	1.0 -
Llama-7B	1.0 -	1.0 -	1.0 -
Mistral-7B	1.0 -	1.0 -	1.0 -
Mistral-24B	0.86 Praise	1.0 -	1.0 -
Teacher	1.0		

Table 3: Coverage of teacher feedback categories. Each cell shows Cov.all (top) and missing feedback types (bottom).

however, still fails to produce praise-focused feedback across both strategies. In addition, Gemini shows degradation as it now misses the Other feedback focus type in addition to missing praise. Similarly, LearnLM, which achieved full coverage in the zero-shot baseline, begins to miss the Other feedback focus type when provided with category information, indicating that explicit category guidance may shift some models toward generating more single-focused feedback.

(1.2) How similar is the feedback focus type distribution of LLM-generated feedback to that of teachers? We examine how similar each model’s feedback focus type distribution is to that of teachers using JSD. As shown in Table 4, in the zero-shot setting, LearnLM achieves the lowest JSD (0.1344), indicating the closest alignment with the teacher distribution, followed by Llama-7B (0.1619) and Mistral-24B (0.1780). Among all models, the educationally fine-tuned LearnLM most closely approximates the teacher feedback focus type distribution, while larger models such as Gemini and Qwen-70B show some of the highest divergence. Applying different prompting strategies yields limited improvement. Prompt 2 increases JSD for most models, while Prompt 3 improves alignment only for Mistral-7B (0.2398). In both cases, the zero-shot baseline remains the closest to the teacher feedback focus type distribution for most models.

To better understand these results, we examine the feedback focus type distribution more closely. As shown in Figure 2, Gemini, Qwen-70B, and Mistral-7B exhibit a strong preference for Elaboration, assigning more than 50% of their feedback to this type, while Llama-7B prefers Mistakes at

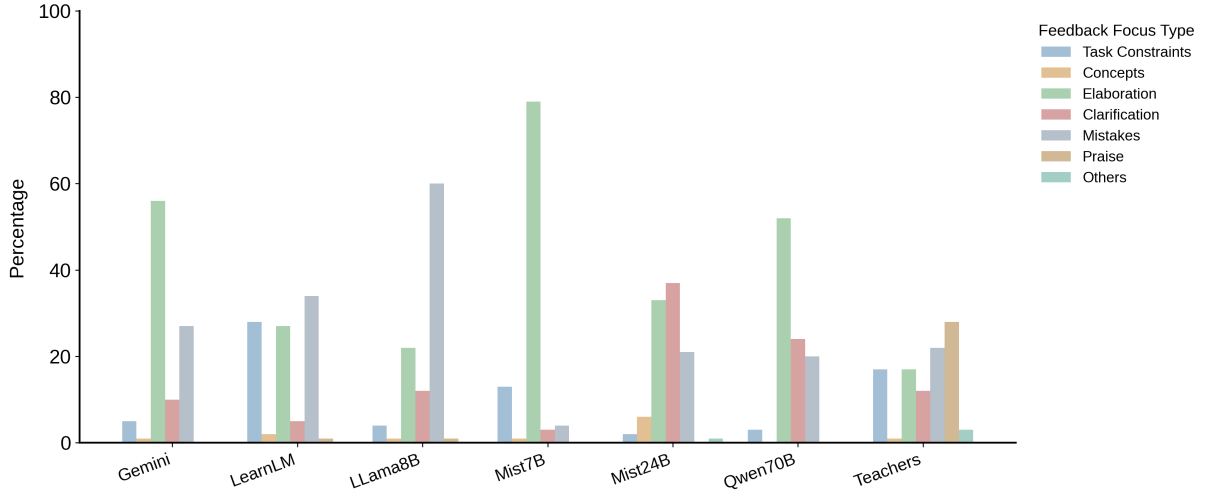


Figure 2: Distribution of feedback focus types (%) for each model and the teachers under the zero-shot baseline, aggregated across all courses.

Model	Zero-shot	Zero-shot + Categories	Few-shot + Categories
Gemini	0.1816	0.1896	0.2800
LearnLM	0.1344	0.1452	0.1486
Qwen-70B	0.1884	0.2682	0.2376
Llama-7B	0.1619	0.2433	0.2665
Mistral-7B	0.2695	0.2454	0.2398
Mistral-24B	0.1780	0.2236	0.1938

Table 4: Jensen-Shannon Divergence (JSD) between each model’s feedback focus type distribution and the teacher distribution. Lower values indicate closer alignment to the teacher distribution. **Bold** indicates the lowest JSD per model. **Gray** indicates the overall best value.

over 60%. This preference for a single type might explain their mostly higher divergence from the teachers, who distribute feedback more evenly. In contrast, LearnLM and Mistral-24B show no such preference, with no single type exceeding 50%, and are relatively closer to the teacher distribution. Moreover, the strong preference for a single type might lead to repetitive, narrow feedback, potentially causing models to overlook other important aspects of student writing.

RQ2: Do LLMs adapt their feedback focus across draft stages, as an expert teacher does?

To address this question, we examine whether feedback focus type distributions change across draft versions for both the teacher and each model, using Fisher’s Exact Test to detect significant changes, where significance suggests adaptation to the student’s draft. This analysis is limited to the composition course, as the fiction course has no drafts and the ESL course was excluded due to limited variation in teacher feedback focus types.

Table 5 presents zero-shot results, and shows that the teacher demonstrates clear feedback adaptivity across draft stages, with significant changes in five feedback focus types: Task Constraints, Elaboration, Clarification, Mistakes, and Praise, across both stages, from initial to revision and from revision to final. When looking at the models, Mistral-24B shows the closest pattern to the teacher, with significant changes in three feedback focus types: Elaboration, Clarification, and Mistakes, across both stages. Task Constraints, however, shows a significant change only from initial to revision. Llama-7B follows, showing significant changes in two feedback focus types: Task Constraints and Mistakes, across both stages, while Elaboration shows a significant change only from initial to revision. Gemini also shows significant changes in two feedback focus types: Elaboration and Mistakes, across both stages. LearnLM shows significant changes in Mistakes across both stages, while Elaboration changes only from initial to revision. Interestingly, all models that demonstrate adaptivity consistently show significant changes in the same two feedback focus types: Mistakes and Elaboration. This aligns with the distribution analysis in Figure 2, where these two types appear with the highest proportion across most models. This suggests that models are more capable of adapting feedback focus types they produce frequently.

RQ3: Do LLMs adapt their feedback focus based on student performance, as an expert teacher does?

To address this question, we examine whether feedback focus type distributions differ between high- and low-performing students

Models/ Teacher	Stage	Task Constraints	Concepts	Elaboration	Clarification	Mistakes	Praise	Other	Sig.
Teacher	Initial	29.3	0.8	23.8	8.3	10.7	26.9	0.2	
	Revision	17.2	0.0	7.5	13.4	23.8	38.1	0	
	Final	6.5	0.0	1.1	3.2	3.2	84.9	1.1	
	I vs R	0.0003	0.3072	0.0000	0.0371	0.0000	0.0029	1.0000	5
	R vs F	0.0059	1.0000	0.0180	0.0031	0.0000	0.0000	1.0000	5
Gemini	Initial	10.2	0.0	53.3	9.4	27.1	0.0	0.0	
	Revision	14.3	0.0	65.3	8.6	11.8	0.0	0.0	
	Final	7.1	0.0	47.5	11.1	34.3	0.0	0.0	
	I vs R	0.1122	1.0000	0.0020	0.7866	0.0000	1.0000	1.0000	2
	R vs F	0.0703	1.0000	0.0034	0.5386	0.0000	1.0000	1.0000	2
LearnLM	Initial	16.5	0.2	54.1	8.2	21.0	0.0	0.0	
	Revision	24.9	0.0	62.9	5.7	6.5	0.0	0.0	
	Final	13.1	0.0	47.5	12.1	27.3	0.0	0.0	
	I vs R	0.2350	1.0000	0.0000	0.2687	0.0001	1.0000	1.0000	2
	R vs F	0.3891	1.0000	0.0914	0.6942	0.0027	1.0000	1.0000	1
LLaMA	Initial	43.5	0.2	24.3	3.7	28.2	0.2	0.0	
	Revision	38.8	0.4	43.3	2.0	15.5	0.0	0.0	
	Final	33.3	0.0	33.3	3.0	30.3	0.0	0.1	
	I vs R	0.0124	1.0000	0.0000	0.1454	0.0000	1.0000	1.0000	3
	R vs F	0.0240	0.2878	0.1856	0.2842	0.0012	1.0000	1.0000	2
Qwen-70B	Initial	4.9	0.2	40.7	29.9	24.3	0.0	0.0	
	Revision	9.0	0.0	36.7	29.4	24.9	0.0	0.0	
	Final	4.0	1.0	35.4	26.3	33.3	0.0	0.0	
	I vs R	0.0360	1.0000	0.3367	0.9319	0.8561	1.0000	1.0000	1
	R vs F	0.1746	0.2878	0.9015	0.5996	0.1412	1.0000	1.0000	0
Mistral-7B	Initial	21.2	0.0	76.9	0.2	0.6	0.2	0.9	
	Revision	20.8	0.0	77.1	0.0	2.0	0.0	0.1	
	Final	27.3	1.0	68.7	1.0	2.0	0.0	0.0	
	I vs R	0.9239	1.0000	1.0000	1.0000	0.1251	1.0000	1.0000	0
	R vs F	0.2030	0.2878	0.1313	0.2878	1.0000	1.0000	1.0000	0
Mistral-24B	Initial	2.2	0.0	11.0	46.9	39.8	0.0	0.1	
	Revision	10.2	0.4	33.6	30.3	25.4	0.0	0.1	
	Final	5.1	0.0	14.1	42.4	38.4	0.0	0.0	
	I vs R	0.0000	0.3333	0.0000	0.0000	0.0001	1.0000	1.0000	4
	R vs F	0.1437	1.0000	0.0003	0.0328	0.0184	1.0000	1.0000	3

Table 5: Feedback distribution (%) and Fisher’s Exact Test significance across draft stages under zero-shot prompting. **Bold** p-values indicate significant difference ($p < 0.05$). Sig. = number of feedback focus types with significant difference.

for both the teacher and each model, using Fisher’s Exact Test to detect significant differences, where significance suggests that feedback focus varies with student performance. This analysis is limited to the composition course, as it is the only course that provides essay grades with variation across students. For each draft stage, students are divided into high- and low-performing groups based on the median score (details in Appendix A), and the proportion of each feedback type is computed for each group.

As shown in Table 6 (zero-shot), the teacher’s feedback shows significant differences between high- and low-performing students in Task Constraints and Praise. Gemini and Mistral-24B show significant differences in Clarification, LearnLM in Mistakes, and Qwen-70B in Elaboration, while Llama-7B and Mistral-7B show no significant differences across any feedback focus type. Except for Gemini, the feedback focus type that shows a significant difference corresponds to the type the model produces most frequently as in Figure 2, suggesting

that most models can only adapt to high- and low-performing students in feedback focus types they produce with sufficient frequency. Furthermore, all models that show adaptivity based on student performance are large or closed models.

6 Conclusion

We present the first systematic study of LLM alignment to expert teachers at the level of feedback focus type distributions in student writing. Using a content-based feedback focus taxonomy, we annotate teacher and LLM feedback across three writing courses and evaluate six LLMs under three prompting strategies. Our findings show that while most LLMs cover most feedback focus types, they fail to reflect teacher feedback distributions. In terms of adaptivity across draft stages, the teacher shows significant changes in five feedback focus types, while the best performing model shows significant changes in only three, and some models show no significant changes at all. Regarding student performance, some models show feedback adaptivity

Model/Teacher	Group	Task Constraints	Concepts	Elaboration	Clarification	Mistakes	Praise	Other	Sig.
Teacher	High	88	3	75	44	64	209	16	2
	Low	101	1	59	31	48	91	4	
	p-value	< 0.001 *	0.653	0.337	0.902	0.536	< 0.001 *	0.068	
Gemini	High	56	—	286	37	120	—	—	1
	Low	36	—	182	41	76	—	—	
	p-value	0.910	—	0.434	0.021 *	0.678	—	—	
LearnLM	High	99	0	288	36	76	—	—	1
	Low	56	1	178	30	70	—	—	
	p-value	0.2765	0.4017	0.2009	0.3631	0.0407 *	—	—	
Qwen-70B	High	28	1	212	134	123	—	—	1
	Low	22	1	112	110	90	—	—	
	p-value	0.6558	1.0000	0.0090 *	0.0741	0.5172	—	—	
Llama-7B	High	131	0	161	67	138	—	2	0
	Low	79	2	106	53	94	—	1	
	p-value	0.416	0.161	0.880	0.365	0.937	—	1.000	
Mistral-7B	High	108	1	375	2	8	1	4	0
	Low	74	0	259	0	2	0	0	
	p-value	0.932	1.000	0.508	0.519	0.331	1.000	0.153	
Mistral-24B	High	30	1	99	190	179	—	—	1
	Low	11	0	52	156	116	—	—	
	p-value	0.101	1.000	0.120	0.018 *	0.712	—	—	

Table 6: Feedback type counts for high and low performing students under zero-shot prompting. Fisher’s Exact Test (p-values) was used to compare distributions between groups. * indicates a significant difference between high and low performing students $p < 0.05$. — indicates zero counts in both groups (test skipped). Sig = number of feedback types with significant differences ($p < 0.05$).

between high and low performing students, but only in the feedback focus types they produce most frequently. To support future research, we release FEEDTYPE, a public corpus of annotated teacher and LLM feedback across three writing courses.

7 Limitations

Feedback adaptivity analysis is limited to the composition course taught by a single instructor. Future work should examine whether these patterns generalize to other writing contexts and instructors. Also, we focus on paragraph-level feedback, which is one common form of teacher feedback. Instructors also provide feedback at other levels, such as holistic essay-level feedback. Future research could examine how feedback distributions differ across these formats. While we evaluate three prompting strategies, fine-tuned models or models with explicit pedagogical instructions may show different alignment patterns. Future work could investigate whether targeted fine-tuning can elicit more adaptive and teacher-like LLM feedback.

8 Acknowledgements

This work was supported by the Pitt Cyber Accelerator Grants Program. We extend our sincere gratitude to the Pitt PETAL group and anonymous reviewers for their insightful feedback.

References

- Alan Agresti. 1992. A survey of exact inference for contingency tables. *Statistical science*, 7(1):131–153.
- Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/llama>. Accessed: 2026-02-04.
- Faiz Algobaei and Elham Alzain. 2026. Prompt engineering for non-native english learners: A generative ai approach to personalised language feedback. *Social Sciences & Humanities Open*, 13:102341.
- Yuya Asano, Diane Litman, and Erin Walker. 2025. Can llms simulate the same correct solutions to free-response math problems as real students? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16347–16376.
- Stefano Bannò, Kate Knill, and Mark JF Gales. 2024. Grammatical error feedback: An implicit evaluation approach. *arXiv preprint arXiv:2408.09565*.
- Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024a. Assessing online writing feedback resources: Generative ai vs. good samaritans. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1638–1644.
- Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024b. **LEAF: Language learners’ English essays and feedback corpus**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 433–442, Mexico City, Mexico. Association for Computational Linguistics.
- Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. **The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.
- Seongyeub Chu, Jongwoo Kim, and Munyong Yi. 2026. Feedeval: Pedagogically aligned evaluation of llm-generated essay feedback. *arXiv preprint arXiv:2601.04574*.
- Scott A Crossley, Perpetual Baffour, L Burleigh, and Jules King. 2025. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Assessing Writing*, 65:100954.
- Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, pages 323–325. IEEE.
- Yishan Du, Conrad Borchers, and Mutlu Cukurova. 2025. Benchmarking educational llms with analytics: A case study on gender bias in feedback. *arXiv preprint arXiv:2511.08225*.
- Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 54602–54631.
- Dana R Ferris. 1997. The influence of teacher commentary on student revision. *TESOL quarterly*, 31(2):315–339.
- Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, and 1 others. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.
- Google DeepMind. 2025. Gemini 3 flash: Frontier intelligence built for speed. <https://deepmind.google/technologies/gemini/>. Accessed: 2026-02-04.
- John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Ken Hyland and Fiona Hyland. 2006. Feedback on second language students’ writing. *Language teaching*, 39(2):83–101.
- Shin’ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the world*, 1:91–118.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,

- Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). Preprint, arXiv:2310.06825.
- Masahiro Kaneko and Naoaki Okazaki. 2024. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3955–3961.
- Aayush Kucheria, Nitin Sawhney, and Arto Hellas. 2025. [Comparing behavioral patterns of LLM and human tutors: A population-level analysis with the CIMA dataset](#). In *Proceedings of BEA 2025*.
- Raymond W. Kulhavy and William A. Stock. 1989. [Feedback in written instruction: The place of response certitude](#). *Educational Psychology Review*, 1:279–308.
- Weiyue Li, Mingxiao Song, Zhenda Shen, Dachuan Zhao, Yunfan Long, Yi Li, Yongce Li, Ruyi Yang, and Mengyu Wang. 2026. Llm review: Enhancing creative writing via blind peer review feedback. *arXiv preprint arXiv:2601.08003*.
- Anastasiya A Lipnevich and Ernesto Panadero. 2021. A review of feedback models and theories: Descriptions, definitions, and conclusions. In *Frontiers in education*, volume 6, page 720195. Frontiers Media SA.
- Suqing Liu, Bogdan Simion, Christopher Eaton, and Michael Liut. 2025. A comparative study of technical writing feedback quality: Evaluating llms, slms, and humans in computer science topics. *arXiv preprint arXiv:2601.11541*.
- Jinghui Lu, Maeve Henchion, and Brian Mac Namee. 2020. Diverging divergences: Examining variants of jensen shannon divergence for corpus comparison tasks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6740–6744.
- Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2472–2495.
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.
- Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Math-tutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221.
- Christopher Mah, Mei Tan, Lena Phalen, Alexa Sparks, and Dora Demszky. 2025. From sentence-corrections to deeper dialogue: Qualitative insights from llm and teacher feedback on student writing. Available at SSRN 5213040.
- Sandeep Mathias and Pushpak Bhattacharyya. 2018. [ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.
- Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Atsushi Mizumoto, Natsuko Shintani, Miyuki Sasaki, and Mark Feng Teng. 2024. Testing the viability of chatgpt as a companion in l2 writing accuracy assessment. *Research Methods in Applied Linguistics*, 3(2):100116.
- Ryo Nagata, Kentaro Inui, and Shin’ichiro Ishikawa. 2020. [Creating corpora for research in feedback comment generation](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.
- Susanne Narciss. 2008a. Feedback strategies for interactive learning tasks. In *Handbook of research on educational communications and technology*, pages 125–143. Routledge.
- Susanne Narciss. 2008b. *Feedback Strategies for Interactive Learning Tasks*, pages 125–144.
- Quy Nguyen, Yusuke Miyao, Hiroshi Noji, and Nhung Nguyen. 2018. [An empirical investigation of error types in Vietnamese parsing](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3075–3089, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Juri Opitz. 2024. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *transactions of the association for computational linguistics*, 12:820–836.

- Christina Ortmeier-Hooper. 2006. Dana r. ferris and john s. hedgcock. teaching esl composition: Purpose, process, and practice. *Writing Program Administration*, 30(1-2):111–115.
- Shayan Peyghambari Oskoui, Norah Almousa, Zhaoyi Joey Hou, Carolina Gustafson, Gayle Rogers, Raquel Coelho, Diane Litman, and Xiang Lorraine Li. 2026. Sefora: Student essays with feedback corpus and llm feedback evaluation framework. *arXiv preprint arXiv:2607.00274*.
- Janos Perczel, Jin Chow, and Dorottya Demszky. 2025. Teachlm: Post-training llms for education using authentic learning data. *arXiv preprint arXiv:2510.05087*.
- Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2025. Help me write a story: Evaluating llms’ ability to generate writing feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25827–25847.
- Kathrin Seßler, Arne Bewersdorff, Claudia Nerdel, and Enkelejda Kasneci. 2025. Towards adaptive feedback with ai: Comparing the feedback quality of llms and teachers on experimentation protocols. *arXiv preprint arXiv:2502.12842*.
- Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. Gee! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781.
- Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650.
- Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. [Exploring LLM prompting strategies for joint essay scoring and feedback generation](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.
- Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and chatgpt feedback of students’ writing. *Learning and Instruction*, 91:101894.
- Anais Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.
- Mei Tan, HyunJi Nam, and Dorottya Demszky. 2026a. [Can large language models match human diversity in educational content generation?](#) In *Conference on Language Modeling (COLM)*.
- Mei Tan, Lena Phalen, and Dorottya Demszky. 2026b. Marked pedagogies: Examining linguistic biases in personalized automated writing feedback. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pages 568–577.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.
- LearnLM Team, Abhinav Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, and 1 others. 2024. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*.
- Danielle R Thomas, Conrad Borchers, and Ken Koedinger. 2025. Beyond agreement: Rethinking ground truth in educational ai annotation. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 345–351.
- Wenqing Wang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025a. Towards a “novel” benchmark: Evaluating literary fiction with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21648–21673.
- Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025b. Llms can perform multi-dimensional analytic writing assessments: A case study of 12 graduate-level academic english writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8637–8663.
- Min Yang and David Carless. 2013. The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in higher education*, 18(3):285–297.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.
- Su-Youn Yoon, Eva Miszoglad, and Lisa R Pierce. 2023. Evaluation of chatgpt feedback on ell writers’ coherence and cohesion. *arXiv preprint arXiv:2310.06505*.

Shulin Yu. 2014. *Understanding Chinese EFL students' participation in group peer feedback of L2 writing: A sociocultural and activity theory perspective*. The Chinese University of Hong Kong (Hong Kong).

Shulin Yu and Icy Lee. 2014. An analysis of chinese efl students' use of first and second language in peer feedback of l2 writing. *System*, 47:28–38.

Vivian Zamel. 1985. Responding to student writing. *TESOL quarterly*, 19(1):79–101.

A Dataset Overview

A.1 Data Selection

The dataset consists of student essays across multiple draft stages (Initial, Revision, and Final), where each essay contains multiple paragraphs. Each paragraph is associated with one or more teacher feedback comments, and students received a numerical grade (0-100) for each essay assigned by the instructor.

Performance Groups. We compute the median grade independently for each essay and draft stage to split students into high and low performing groups. Students at or above the median are classified as high performing and those below as low performing. Since the median is computed independently per essay and draft stage, the same student may be classified differently across drafts, reflecting their actual performance at each stage. The median grade and number of students in each group per essay and draft stage are reported in Table 7.

Paragraph Selection. After computing the median split, we select only paragraphs that received exactly one feedback comment, as our focus is on paragraph-level feedback where each paragraph receives a single comment. Paragraphs with more than one comment are excluded, which results in excluding students who do not have any paragraphs with exactly one feedback comment. This leads to unequal counts of high and low performing students, as the median split is applied before paragraph selection. Table 7 reports the number of selected students classified as high and low performing for each essay and draft stage.

A.2 Example Teacher Feedback in a Paragraph

Table 8 presents some examples of a student paragraph, with the teacher’s feedback and its annotation according to the feedback types.

B Narciss’s Taxonomy and Our Adaptations

Narciss (2008b) proposes a content-based taxonomy of feedback components, distinguishing between simple and elaborated types. Simple feedback components include Knowledge of Performance (KP), Knowledge of Result (KR), and Knowledge of Correct Response (KCR). Elaborated feedback components include Knowledge about Task Constraints (KTC), Knowledge about

Concepts (KC), Knowledge about Mistakes (KM), Knowledge about How to Proceed (KHP), and Knowledge about Metacognition (KMC). We adopt the elaborated feedback components as the basis for our feedback focus taxonomy, as they are more informative and better reflect the nature of writing feedback. Specifically, we retain Task Constraints, Concepts, and Mistakes, as these directly apply to writing tasks, mistakes in writing encompass a wide range of issues including grammar errors, organization, and coherence. We adapt KHP into *Elaboration*, as in writing feedback, procedural guidance typically takes the form of asking students to develop and expand their ideas. We adapt KCR into *Clarification*, as teachers often ask students to clarify their meaning rather than providing the correct response directly. We exclude the simple feedback components (KP, KR, KCR) as they are more suited to closed-ended tasks with clear correct answers, which does not apply to open-ended writing tasks. We also exclude KMC, as metacognitive feedback was rarely observed in our dataset. Finally, we add two categories not present in Narciss’s taxonomy: *Praise*, to capture positive reinforcement which is common in writing feedback, and *Other*, to account for feedback instances that address multiple focus types simultaneously or do not fit any type.

C Annotation Guidelines

We developed a coding framework informed by prior work on feedback analysis and patterns observed in our dataset. The framework categorizes the *content* of feedback, the focus of what the feedback addresses, into seven types, summarized in Table 9.

Task Description Annotators assigned each feedback to the single category that best matched its content. For all categories, annotators followed the definitions provided in Table 9. When assigning labels such as *Task Constraints* or *Concepts*, annotators could also refer to the assignment rubric in addition to the category definitions. If a feedback item did not fit any category, or could reasonably belong to more than one category, annotators selected *Other*.

Labeling Priorities. When feedback could belong to two types, the following priorities apply:

- *Task Constraints* take priority over *Mistakes*, as addressing assignment requirements is

Class	Essay	Draft	High (<i>n</i>)	Low (<i>n</i>)	Total (<i>n</i>)	Median	Grade Range	Selected (High, Low)
Class A	Essay 1	First	8	7	15	75	66-83	15 (8,7)
		Revision	5	5	10	84	56-94	10 (5,5)
		Final	3	2	5	85	75-93	3 (2,1)
		Essay 1 Total	16	14	30	-	-	28 (15,13)
	Essay 2	First	8	7	15	79	0-93	15 (8,7)
		Revision	7	2	9	90	0-95	9 (7,2)
		Final	3	3	6	93	68-97	0
		Essay 2 Total	18	12	30	-	-	24 (15,9)
	Essay 3	First	8	7	15	79	0-93	13 (7,6)
		Revision	-	-	-	-	-	-
		Final	5	4	9	91	67-95	8 (5,3)
		Essay 3 Total	13	11	24	-	-	21 (12,9)
	Class A Total		47	37	84	-	-	73 (42,31)
Class B	Essay 1	First	8	7	15	73	61-86	15 (8,7)
		Revision	9	5	14	86	61-93	13 (9,4)
		Final	2	2	4	89	65-92	2 (2,0)
		Essay 1 Total	19	14	33	-	-	30 (19,11)
	Essay 2	First	8	7	15	80	70-88	14 (8, 6)
		Revision	8	5	13	93	77-95	10 (5,5)
		Final	2	2	4	93	86-96	4 (2,2)
		Essay 2 Total	18	14	32	-	-	28 (15,13)
	Essay 3	First	8	7	15	83	0-88	15 (8,7)
		Revision	-	-	-	-	-	-
		Final	4	4	8	92.5	61-96	6 (3,3)
		Essay 3 Total	12	11	23	-	-	21 (11,10)
	Class B Total		49	39	88	-	-	79 (45,34)
Overall Total (Class A + B)			96	76	172	-	-	152 (87,65)

Table 7: Distribution of selected student performance across classes, essays, and draft stages. High and low groups are defined using a median split. Grade range shows the lowest and highest grade per group. Totals are calculated per essay, per class, and overall.

more important than correcting minor errors in idea organization or grammar.

- *Clarification* take priority over *Elaboration*, since resolving confusion is more critical than adding extra details.
- *Other Categories* take priority over *Praise*, because feedback that begins with encouragement but also addresses an issue should prioritize corrective content over positive feedback.

D Evaluating LLMs as Annotators

Our framework incorporates LLMs as part of the automatic evaluation procedure. To ensure the reliability of the LLM annotator, we verify three requirements in our evaluation procedure. First, we assess the reliability of the annotation task itself by collecting both human and model annotations on a subset of the data and computing inter-annotator agreement. This step confirms that the taxonomy is well-defined and that annotators can reach reasonable agreement on the evaluation outcomes. Second, following [Calderon et al. \(2025\)](#), we evaluate whether the LLM can serve as a reliable substitute for human annotators. Third, we conduct a validity

analysis to ensure that conclusions drawn from the full set of LLM-generated annotations are consistent with those derived from human annotations.

D.1 Traditional Metrics

Gold-Standard Dataset. To compute inter-annotator agreement, we created a gold-standard reference dataset consisting of 100 feedback sampled from different essays. The sampled feedback covers multiple categories, with at least 10 examples per category, except for the Concepts category, which is less frequent in the dataset. The sample size was chosen to meet the evaluation of Alt-Test (see Subsection D.2), which requires at least 75 annotated instances to assess whether an alternative annotator can reliably replace human annotations. Three human annotators participated in labeling the reference dataset. Two annotators were involved in the initial development of the annotation guidelines, while a third annotator was added to satisfy the Alt-Test requirement of having at least three annotators. Before annotating the reference dataset, the third annotator was trained using the annotation guidelines, labeled 60 feedback from earlier pilot rounds, and resolved disagreements with one of the original annotators through discussion until con-

Draft Stage	Paragraph	Teacher Feedback	Feedback Type
Initial	My role as a woman is to bear. Bear, a verb in the Old English <i>beran</i> , means “to carry, bring; bring forth, give birth to, produce; to endure, to wear.” According to etymonline.com, my role as a woman is to carry my husband’s child, produce an heir, and bring forth the next generation. My role as a woman is to endure the comments made by men as I walk by and to wear clothing that does not seduce men but not too modestly that I am considered boring. As a woman, my role is to bear.	This is a great beginning of a definition paragraph. What if you extended it? Got even more concrete and specific? Added layer upon layer from your own life and experience?	Elaboration
Revised	We still get together for birthdays and important events my siblings and I may have. I learned how to be mature because I saw my parents having the right concept of being there for their children no matter what. It has been an important factor in my growth as a young woman because it taught me that this maturity, which is, in a way, a language, is what is going to help me develop relationships with other people and how to deal with difficult situations I may encounter with people along my life.	This is a fascinating and important paragraph, but I wonder if it’s necessary for this essay and your essay’s central focus. This is a genuine question. If it is necessary, perhaps it needs a little more development and connection. If not, what would be the effect if you simply cut it?	Task Constraints
Final	Maybe writing with my grandfather’s pen brought comfort. The familiar hold, the familiar gold details, and the shiny black body all brought some sort of immediate ease and focus to my thoughts.	Nice.	Praise

Table 8: Examples of paragraphs across draft stages with teacher feedback and corresponding feedback type annotations.

sensus was reached. Agreement was evaluated on 40 new feedback selected from earlier pilot rounds, resulting in a Cohen’s Kappa of $\kappa = 0.68$. After this training, all three annotators annotated the full set of 100 feedback. After annotation, we used the human majority label for each feedback to calculate inter annotator agreement and accuracy. The majority label was defined as the label agreed upon by at least two annotators. If no agreement was reached, we followed the priority labeling rules described in Section C.

Reliability: We assess the reliability of LLMs as annotators by computing inter annotator agreement using Cohen’s κ between LLM predictions and human labels. This measures how closely LLM annotations align with human judgments beyond chance agreement (Sim and Wright, 2005). Using the gold standard sample described above, we prompted the candidate models to classify the feedback and computed inter-annotator agreement. Table 10 reports Cohen’s κ agreement between each LLM annotator and the human majority vote, as well as agreement among human annotators. Gemini 3 Flash and

GPT-4-Turbo show human-level alignment, outperforming LLaMA-3.1 8B and Qwen3-14B in the classification task.

Accuracy: We evaluate LLM feedback annotation as a classification task and use accuracy to measure how often a model’s predictions match the ground truth for each feedback (Opitz, 2024). Using the gold-standard sample described above, we prompted the candidate models to classify the feedback and computed accuracy. Table 10 reports accuracy for the LLMs.

D.2 Alt-Test

The preceding group-level κ evaluation does not determine whether a specific LLM is comparable to a human annotator at the individual level, nor does it offer a principled threshold for trusting the model’s annotations. As a result, it cannot justify replacing human annotation with an LLM. To address this limitation, we employ the Alt-Test, which compares the LLM’s annotations to those of individual human annotators rather than to the group as a whole. We evaluate each candidate LLM against

Feedback Type	Definition	Example Feedback
Task Constraints	Clarifies assignment goals, required components, or evaluation criteria.	Is this the right language and tone for your audience?
Concepts	Explains or reinforces key concepts or terminology necessary for completing the task.	A narrative focuses on a single moment in time.
Elaboration	Encourages expanding or deepening ideas through additional detail, reflection, or development.	I’m interested! But let’s dig a little deeper, here. Why, specifically, are you now comfortable?
Clarification	Signals difficulty understanding meaning, reference, or sequence and requests clarification.	what do you mean?
Mistakes	Identifies specific errors such as grammar, word choice, organization, or logical inconsistencies.	Watch out for sentence fragments.
Praise	Expresses encouragement or positive evaluation without suggesting changes.	Awesome template! Nice job with writing moves.
Other	Feedback that does not fit other categories, including personal reactions or symbols.	:)

Table 9: Feedback taxonomy with definitions and a concise example for each category.

LLM vs. Human Majority			
Model	Examples	Accuracy	Cohen’s κ
Qwen	3-shot	0.60	0.51
Qwen	7-shot	0.67	0.58
Qwen	10-shot	0.70	0.61
LLaMA	3-shot	0.55	0.42
LLaMA	7-shot	0.71	0.71
LLaMA	10-shot	0.71	0.63
Gemini	3-shot	0.76	0.70
Gemini	7-shot	0.81	0.75
Gemini	10-shot	0.82	0.77
GPT-4-Turbo	3-shot	0.72	0.65
GPT-4-Turbo	7-shot	0.71	0.63
GPT-4-Turbo	10-shot	0.80	0.74
Human vs. Human			
Human 1 vs. Human 2	–	–	0.72
Human 1 vs. Human 3	–	–	0.68
Human 2 vs. Human 3	–	–	0.70

Table 10: Inter-annotator agreement across LLMs and humans. Accuracy reflects exact label agreement, while Cohen’s κ measures chance-corrected agreement. The top section reports LLM agreement with the human majority label, and the bottom section reports pairwise agreement among human annotators.

each of the three human annotators individually, using our gold-standard feedback subset described in D.1. The test follows a leave-one-out approach, in which each human annotator is excluded in turn, and we evaluate whether the LLM’s annotations align more closely with the consensus of the remaining annotators than the excluded annotator does. For classification tasks, we use accuracy to measure alignment with the consensus. To account for the inherent variability and cost of human annotation, we introduce a tolerance margin ϵ , allowing minor deviations, as implemented in the Alt-Test. Following prior work, ϵ is selected based on annotator expertise: $\epsilon = 0.2$ for experts, $\epsilon = 0.15$ for

Model	Examples	Winning Rate (ω)
LLaMA	10-shot	0.33
LLaMA	7-shot	0.33
LLaMA	3-shot	0.00
Qwen	10-shot	0.00
Qwen	7-shot	0.00
Qwen	3-shot	0.00
Gemini	10-shot	0.67
Gemini	7-shot	0.33
Gemini	3-shot	0.33
GPT-4-Turbo	10-shot	0.67
GPT-4-Turbo	7-shot	0.33
GPT-4-Turbo	3-shot	0.33

Table 11: Alt-Test results showing the winning rate (ω) for each model across different few-shot settings, using $\epsilon = 0.1$. A winning rate greater than 0.5 indicates that the model outperforms the excluded human annotator in the majority of cases.

skilled annotators, and $\epsilon = 0.1$ for crowd workers. In our experiments, we observed that using larger ϵ values (e.g., 0.15 and 0.2) led the models to pass the test across all settings in Table 10. Therefore, to enforce a stricter evaluation criterion, we adopt $\epsilon = 0.1$, reflecting the relatively low-expertise nature of the task. If the LLM aligns better than the excluded annotator in more than half of the cases (winning rate > 0.5), it is considered a viable replacement for human annotation. As shown in Table 11, the test yielded a winning rate of $\omega = 0.67$ for Gemini Flash and GPT-4-Turbo with 10 few-shot examples, indicating that the LLM’s judgments can be reliably trusted. This result supports the use of the LLM as an annotator in our evaluation framework. So, we select Gemini 3 Flash as the annotator due to its lower cost compared to GPT-4-Turbo.

Draft	Annotator Type	Task	Concepts	Elaboration	Clarification	Mistakes	Praise	Other
Initial	Human annotator	23.8	0.5	23.4	14.5	11.2	25.2	1.4
	Gemini annotator	29.4	0.5	23.4	10.7	11.2	23.8	0.9
Revision	Human annotator	12.5	0.6	10.0	7.5	23.8	41.9	3.8
	Gemini annotator	12.5	0.6	8.1	7.5	25.0	41.9	4.4
Final	Human annotator	3.7	0.0	0.0	3.7	11.1	77.8	3.7
	Gemini annotator	5.6	0.0	0.0	5.6	7.4	77.8	3.7

Table 12: Feedback category distribution (%) across drafts for human and Gemini annotations on the set of 400 feedback.

Performance	Annotator Type	Task	Concept	Elaboration	Clarification	Mistake	Praise	Other
High	Human annotator	14.5	1.8	12.7	14.6	16.4	30.9	9.1
	Gemini annotator	19.0	0.6	15.3	6.8	15.3	40.5	2.5
Low	Human annotator	23.8	0.0	11.0	14.0	22.8	24.6	3.5
	Gemini annotator	24.0	0.3	17.5	8.4	16.1	24.7	1.2

Table 13: Feedback category distribution (%) by student performance level (High Low) for human annotators and Gemini, on the set of 400 feedback.

D.3 LLM Annotation Validity

The prior metrics used in subsections D.1 and D.2 do not guarantee that a model captures meaningful patterns in annotated data. As highlighted in prior work (Thomas et al., 2025), high agreement among annotators or strong performance on automated metrics does not necessarily reflect the validity of the annotations. In educational contexts, annotation quality should be evaluated based on whether it captures patterns that are meaningful, interpretable, and relevant to learning (Thomas et al., 2025). Motivated by this perspective, we move beyond agreement based evaluation and assess whether the model reflects educationally relevant patterns in feedback. We therefore evaluate the validity of the LLM as an annotator by examining whether its annotations capture meaningful and educationally relevant patterns, and whether they align with those observed in human annotated data. To do this, we select a subset of the dataset designed to capture variation in student performance across draft stages. For each assignment and draft stage (initial, revision, and final), we calculate the median score across all students and use it to distinguish higher and lower performing students. From each assignment at each draft stage, we select two essays with the highest scores and two with the lowest scores. With three assignments and three draft stages per assignment, this results in a total of 36 essays from Class 1. For each selected essay, we extract the corresponding teacher feedback and annotate it according to the categories defined in

our taxonomy, resulting in a total of 441 feedback annotations. This sampling strategy aims to identify meaningful patterns to answer two research questions and support evaluation. First, what types of feedback change (increase or decrease) between drafts. Second, what types of feedback change for high versus low performing students?

To address the first question, we examined the two most frequent feedback types, Praise and Task Constraints, that students received in the initial draft according to human annotations, tracked how their percentages changed across draft stages, and compared them with Gemini annotations for the same feedback types (see Table 12). Both human annotators and Gemini exhibited similar patterns: Praise feedback increased from the initial to final drafts, reflecting recognition of student improvements, while Task Constraint feedback decreased, indicating that students’ understanding of the assignment requirements improved across drafts. For student performance, addressing the second question, we followed a similar approach as in the first question. We focused on the two most frequent feedback types received by high performing students according to human annotations, Praise and Mistakes, and examined how their percentages changed between high and low performing students, as well as in Gemini annotations (see Table 13). Both human and Gemini annotations showed that Praise feedback decreased from high to low performing students, indicating that higher-performing students received more recognition. In contrast, Mistakes feedback increased, showing

that lower performing students received more corrective feedback. In both questions, Gemini was able to recognize the patterns identified by human annotations, as well as those reported in educational research on effective feedback practices as shown in Section 2. In addition to capturing patterns observed in human annotations and the educational feedback literature, we compared the percentages of feedback categories between human annotations and Gemini. While the exact percentages were not always identical, Gemini generally reflected the trends seen in human annotations. Sometimes, however, Gemini was able to produce the same percentages. For example, the Praise category had very similar percentages in both the revision and final drafts, and categories such as Elaboration and Mistakes in the initial draft were also closely aligned.

E LLM Feedback

E.1 Using LLMs for Feedback Generation

We prompt the models under three conditions to generate feedback on student essay paragraphs. In all conditions, the model is provided with the entire essay, the target paragraph, the assignment rubric, and the draft stage. For revision and final drafts, previously received feedback is also included so the model can consider prior comments when generating updated feedback. **Prompt 1 (Zero-shot):** The model receives no information about feedback focus types and is prompted to generate feedback directly. **Prompt 2 (Zero-shot + Categories):** The model additionally receives the names and definitions of the seven feedback focus types to guide its output. **Prompt 3 (Few-shot + Categories):** The model receives the names, definitions, and examples of each feedback focus type to further guide its output. Table 14 shows the prompt conditions used for feedback generation.

E.2 Examples of Feedback Generation

Table 15 presents a single paragraph along with feedback from multiple sources, including teacher and different models.

F Using an LLM as an Annotator

We prompt the models in both zero shot and few shot settings to evaluate feedback on student essay paragraphs by classifying it into one of seven pre-defined categories. The prompt includes the list

of these seven categories along with their definitions and rubric, as well as the target paragraph and the corresponding feedback to be annotated. For the few shot setting, we construct examples using newly selected data from Class 1, annotated by the first author. Specifically, we include at least 10 examples for each feedback category, except for the Concept category, which is relatively rare in our dataset. In total, 85 annotated feedback instances are used as examples to guide the annotation task. Table 14 shows the prompt conditions used for feedback classification.

Prompt	Content
Prompt 1: Zero-shot	## Task Description You are an experienced English teacher. Provide feedback for a student's {draft_stage} essay draft using the {previous_feedback_available} , current draft, target paragraph, and rubric. ## Input Previous Feedback: {previous_feedback} Current Draft: {current_body} Target Paragraph: {paragraph} Rubric: {rubric} ## Your Task 1) Identify the SINGLE most important feedback focus in the TARGET PARAGRAPH. 2) Write ONE short constructive feedback comment reflecting that focus. ## Feedback Rules - Feedback must be 1–4 sentences. - Focus on ONE feedback focus only. - Do NOT summarize or repeat the essay, list strengths and weaknesses, or add meta-commentary. ## Output Format Feedback: <ONE paragraph of feedback>
Prompt 2: Zero-shot + Categories	Same as Prompt 1, but the model also receives the names and definitions of the seven feedback focus types before the task instructions. ## Feedback Focus Types Task Constraints: Clarifies assignment goals, required components, or evaluation criteria. Concepts: Explains or reinforces key concepts or terminology necessary for completing the task. Elaboration: Encourages expanding or deepening ideas through additional detail, reflection, or development. Clarification: Signals difficulty understanding meaning, reference, or sequence and requests clarification. Mistakes: Identifies specific errors such as grammar, word choice, organization, or logical inconsistencies. Praise: Expresses encouragement or positive evaluation without suggesting changes. Other: Feedback that does not fit other categories or combines multiple feedback types.
Prompt 3: Few-shot + Categories	Same as Prompt 2, but the model also receives examples of each feedback focus type before the task instructions. ## Feedback Focus Types {feedback_focus_definitions} ## Examples {feedback_focus_examples}
Feedback Classification	## Definitions Task Constraints: Clarifies assignment goals, required components, or evaluation criteria. Concepts: Explains or reinforces key concepts or terminology necessary for completing the task. Elaboration: Encourages expanding or deepening ideas through additional detail, reflection, or development. Clarification: Signals difficulty understanding meaning, reference, or sequence and requests clarification. Mistakes: Identifies specific errors such as grammar, word choice, organization, or logical inconsistencies. Praise: Expresses encouragement or positive evaluation without suggesting changes. Other: Feedback that does not fit other categories or combines multiple feedback types. ## Rubric Use the rubric to guide classification. ## Task Classify the feedback into exactly one category. Output only the category name. ## Example Feedback: "..." Category: "..." ## Target Feedback {feedback_text_here}

Table 14: Prompts used for feedback generation and feedback classification.

Paragraph	Source	Feedback and its type
My brain uses up all my energy. All I do is overthink, overthink, and overthink. I spend every inch of my head over analyzing everything. All the possibilities and outcomes of every path that a choice opens. I will go over the past and the future. I exhaust myself rather than act and be able to work. Like the lamp, we have the ability to control the dimness of the light to be able to decide how much energy we want to use. So, our judgments shouldn't take a toll on us and should also take in less energy so more energy can go towards taking a step forward and allow us to physically perform.	Teacher	This is an interesting moment of friction that we could discuss! Let's see if we can find your essay's central focusing idea. Then, let's "map" this essay so that you know which parts of it are necessary, which aren't, and which you still need to develop. Type of Feedback: Elaboration
	Mistral 7B	Your essay effectively explores the relationship between your mental energy expenditure and the metaphor of a lamp's adjustable light. However, consider expanding upon the connection between your personal experiences and the broader implications for readers. Delve deeper into how controlling the dimness of your thoughts parallels the physical control of a lamp's light. Type of Feedback: Elaboration
	LLaMA 3-8B	In this paragraph, consider varying sentence structure to maintain a consistent pace and flow. You tend to start each sentence with a similar phrase ("I", "Like"), creating a sense of repetition that disrupts the overall rhythm of the text. Experiment with starting sentences with different phrases or transitioning smoothly between clauses to enhance the clarity and engagement of your argument. Type of Feedback: Mistakes
	Mistral 24B	You have used repetition effectively to emphasize your point about overthinking, but consider varying your sentence structure to maintain reader engagement. Type of Feedback: Mistakes
	Gemini 2 Flash	You're making an interesting connection between your overthinking and the lamp's dimmable light, but the paragraph feels a bit repetitive and abstract. To strengthen your argument, try grounding this comparison in a specific example of a time when your overthinking (or lack thereof) directly impacted a decision you made. Showing the reader this connection through a concrete scenario will make your point more vivid and persuasive. Type of Feedback: Elaboration

Table 15: Human and model generated feedback for a single paragraph.