# GRAFE: Graph-Reasoned Automated Feedback and Evaluation for Structure-Aware Essay Assessment

Anonymous ACL submission


## Abstract

Automated Writing Evaluation has shifted from score-centered assessment toward revision-oriented feedback. However, current Large Language Models (LLMs) often generate generic comments that overlook essays' explicit structure and discourse-level flaws. We propose Graph-Reasoned Automated Feedback and Evaluation (GRAFE), which uses genre-specific structural graphs to guide score-feedback generation. GRAFE encodes discourse-level priors, including argumentation logic, evidence integration, and narrative progression, as continuous representations injected into the LLM. Furthermore, we introduce a training pipeline that combines two-stage Supervised Fine-Tuning for score-calibrated initialization with Group Relative Policy Optimization to align the model toward structural diagnostic accuracy, scoring consistency, and format compliance. We construct a graph-augmented ASAP++ feedback dataset with a human-calibrated gold-standard test set. Experiments show that GRAFE outperforms text-only baselines on Structure-Aware Diagnosis and Structure-Grounded Actionability, demonstrating the value of explicit structural modeling[1].


## 1 Introduction

Automated essay assessment has traditionally focused on predicting holistic or multi-trait scores. Early neural AES systems established score prediction as a core task (Taghipour and Ng, 2016), and recent studies have further advanced AES through autoregressive score generation, cross-prompt generalization, prompt-aware disentanglement, and graph-based trait modeling (Do et al., 2024; Li and Ng, 2024; Han et al., 2025; Li and Ng, 2025). However, a numerical score alone provides limited pedagogical support: it indicates how well a student writes, but does not explain where weaknesses occur, why they matter, or how they can be revised. Therefore, Automated Writing Evaluation (AWE) systems are increasingly expected to provide revision-oriented natural language feedback in addition to scores (Stahl et al., 2024; Li and Pan, 2025).

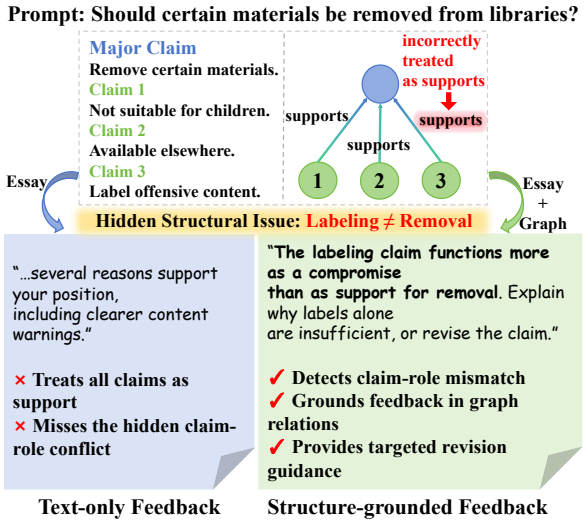


Figure 1: Comparison between traditional text-only feedback and our proposed structure-grounded feedback.

However, merely generating feedback cannot inherently guarantee practical pedagogical value. Recent LLM-driven approaches have progressively modeled scoring and feedback in a unified manner, tightly consolidating the linkage between quantitative evaluation and textual elaborations. Even so, the produced feedback frequently stays overly generic, inadequately rooted in the inherent organizational logic of essays, and devoid of concrete actionable revision suggestions (Rashkin et al., 2025; Chu et al., 2026). Figure 1 vividly demonstrates this drawback: plain text-based feedback indiscriminately regards all viewpoints as supportive stances, yet utterly fails to per-

[1]The code and dataset will be open-sourced upon publication.

ceive that the designated claim essentially serves as a compromising statement rather than valid supportive content to be discarded. It thereby reveals that compelling feedback ought to sufficiently take into account not only literal essay content, but also the functional structural roles of individual viewpoints, alongside their supportive and contradictory interactive relations (Persing and Ng, 2016; Stab and Gurevych, 2017; Li and Ng, 2025).

Realizing structure-aware feedback is inherently challenging because LLMs merely process essays as sequential text. When handling student essays, they often overly rely on local lexical patterns or superficial coherence while frequently overlooking essential discourse-level problems such as fractured argumentation chains, poorly fused source-evidence integration, or logically incomplete narrative causality. To effectively mitigate this critical limitation, we innovatively introduce structured graphs as reliable structural priors for high-quality feedback generation. Since distinct genres naturally follow different organizational patterns, we meticulously construct genre-adaptive graphs for argumentative, source-dependent, and narrative essays, enabling the model to comprehensively capture claim-support relations, solid evidence integration, and coherent event progression.

To this end, we propose Graph-Reasoned Automated Feedback and Evaluation (GRAFE). Rather than asking the LLM to infer graph topology from serialized JSON tokens, GRAFE encodes the essay graph as a continuous structural representation and projects it into the LLM embedding space as a structural prior. We further adopt a two-stage supervised fine-tuning strategy: the first stage trains the model to map essay content and graph structure to multi-trait scores, thereby learning a score-calibrated understanding of writing quality; the second stage generates scores and structure-grounded feedback. Finally, we apply Group Relative Policy Optimization (Shao et al., 2024) with rewards for structural diagnostic accuracy, scoring consistency, and format compliance, directly optimizing sequence-level feedback quality.

The main contributions of this work are as follows:

- We construct a graph-assisted feedback dataset based on ASAP++, including a rigorous test set complete with expert-verified structural flaw identifications and gold-standard feedback.
- We introduce GRAFE, a framework that integrates genre-specific structural priors into LLMs via a two-stage SFT and GRPO alignment strategy.
- We propose two structure-aware metrics, SAD and SGA, to rigorously evaluate the accuracy of structural diagnosis and the utility of revision advice.

## 2 Related Work

**Essay Feedback Generation** Recent studies have explored prompting LLMs to generate essay scores and feedback jointly, in order to examine the relationship between numerical assessment and textual feedback (Stahl et al., 2024). Other work uses LLM-generated explanations as supervision to train smaller models for educational assessment tasks (Li et al., 2023; Do et al., 2025). However, existing methods mostly rely on the essay text, rubrics, or scores as input, leaving feedback quality dependent on the LLM's implicit understanding. As a result, generated feedback may be insufficiently grounded in the essay, deviate from scoring criteria, or lack actionable revision guidance (Chu et al., 2026). In contrast, our work introduces the internal structure of essays as explicit evidence for feedback generation, enabling the model to produce feedback that is more diagnostic and revision-oriented.

**Essay Graph Representation** Essay evaluation depends not only on local linguistic quality, but also on discourse-level organization. Prior studies have incorporated discourse markers, organizational features, and support or attack relations among argumentative units to support essay understanding and scoring (Persing and Ng, 2016; Stab and Gurevych, 2017). More recently, graph structures have been introduced into automated essay scoring, where graph neural networks or graph attention mechanisms are used to model the relationship between essay structure and multi-trait scores (Li and Ng, 2025). Closely related to our motivation, Zhong et al. (2025) propose NarGINA, which represents children's narratives as narrative graphs with event nodes and causal or synchronous relations, enabling more accurate and interpretable assessment of narrative completeness and coherence. Nevertheless, these studies focus on improv-

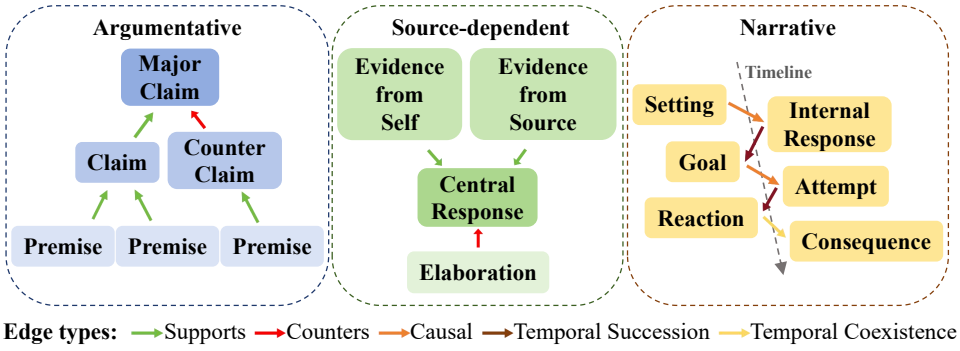


Figure 2: Genre-specific essay graph schemas used in GRAFE.

ing scoring performance or interpretability of assessment results. In this work, we instead use essay graphs as structured evidence for feedback generation, helping the model identify structural problems and verbalize them as natural language feedback.

**RL for Feedback Alignment** Supervised fine-tuning helps models learn feedback style and structured output formats, but its token-level likelihood objective cannot directly optimize sequence-level properties such as structural faithfulness and pedagogical usefulness. Reinforcement learning provides a natural way to optimize such objectives. Prior work has applied RLHF, PPO, or GRPO to language model alignment and long-form generation (Ouyang et al., 2022; Bai et al., 2022; Schulman et al., 2017; Shao et al., 2024). In essay assessment, reinforcement learning has also been used to improve scoring performance or score–feedback consistency (Wang et al., 2018; Do et al., 2024; Li and Pan, 2025). Different from these studies, we apply reinforcement learning to graph-assisted educational feedback alignment, jointly optimizing feedback quality, scoring consistency, and XML output stability.

## 3 Dataset Construction

While the ASAP++ corpus provides fine-grained multi-trait scores, it lacks natural language revision feedback. To bridge this gap, we augment the corpus by generating a genre-specific structural graph and corresponding high-quality feedback for each essay.

### 3.1 Genre-Specific Essay Graph Construction

ASAP++ comprises three essay genres: argumentative, source-dependent, and narrative. Given the distinct organizational patterns inherent to each genre, a single unified template is insufficient. Instead, we define genre-specific graph schemas to explicitly model their structural priors (illustrated in Figure 2).

- **Argumentative Essays**: Models logical support. Nodes represent Major Claim, Claim, and Premise. Edges (Supports, Counters) delineate the completeness of the argumentation chain.
- **Source-dependent Essays**: Captures the integration of external materials. Nodes include Central Response, Elaboration, Evidence from Source, and Evidence from Self. Edges (Supports, Counters) indicate how effectively the evidence backs the central response.
- **Narrative Essays**: Maps event progression and spatiotemporal coherence. Nodes capture narrative elements (e.g., Setting, Initiating Event, Internal Response, Goal, Attempt, Consequence, Reaction). Edges encode relational dynamics through Causal, Temporal Coexistence, and Temporal Succession links.

Table 1: Pairwise comparison between Graph-aided and Text-only feedback.

| Prompt | Graph wins | Text wins | Total | Win rate |
|---|---|---|---|---|
| 1 | 1408 | 375 | 1783 | 78.97% |
| 2 | 1632 | 166 | 1798 | 90.77% |
| 3 | 1614 | 112 | 1726 | 93.51% |
| 4 | 1681 | 87 | 1768 | 95.08% |
| 5 | 1693 | 112 | 1805 | 93.80% |
| 6 | 1693 | 107 | 1800 | 94.06% |
| 7 | 1023 | 546 | 1569 | 65.20% |
| 8 | 581 | 142 | 723 | 80.36% |
| **Total** | **11325** | **1647** | **12972** | **87.30%** |

To instantiate these graphs, we prompted an LLM to parse the essay texts based on the writing prompts and our predefined genre schemas. The output is a JSON-formatted graph comprising node sets (IDs, types, text content) and edge sets (source, target, relation types).

### 3.2 Feedback Data Synthesis and Selection

To generate the feedback data, we used Qwen-max to synthesize two variants of comments for each essay: a Graph-aided version conditioned on both the essay text and its extracted structural graph, and a Text-only baseline.

To isolate and evaluate the impact of the explicit structural priors on feedback quality, we conducted a pairwise comparison using a strong LLM judge (Qwen3.5-plus-thinking). As detailed in Table 1, the Graph-aided feedback showed a substantial advantage, achieving an 87.30% win rate over the Text-only baseline. Thus, we discarded the baseline and adopted the Graph-aided feedback set for model training (generation settings and evaluation criteria are detailed in Appendix A).

### 3.3 Human-Calibrated Gold Standard

We construct a human-verified gold standard test set for final evaluation. Specifically, we sample 320 essays (40 per prompt) and ask three expert annotators with English teaching experience to identify structural flaws and revise the LLM-generated feedback. The inter-annotator agreement for structural flaw identification reaches Fleiss' $\kappa = 0.68$, indicating substantial reliability. Detailed annotation guidelines and arbitration procedures are provided in Appendix B.

## 4 Method

### 4.1 Overview

In essay feedback generation, a key challenge is to identify core writing problems through essay structure and translate them into specific, evidence-grounded revision suggestions. For this reason, we introduce GRAFE, as shown in Figure 3. GRAFE follows three stages: text-graph fusion, two-stage SFT, and GRPO-based alignment. First, the essay graph is encoded, projected into the LLM embedding space, and injected with the essay text. Next, two-stage SFT learns to generate scores and structure-grounded feedback. Finally, GRPO optimizes the model with reward signals.

### 4.2 Text-Graph Fusion

GRAFE initializes each graph node with a Sentence Transformer embedding $h_u$, and then updates node representations through GAT aggregation:

$$h'_u = \sigma\left(\sum_{v \in \mathcal{N}(u)} \alpha_{uv} W h_v\right), \qquad (1)$$

where $\mathcal{N}(u)$ denotes the neighboring nodes of $u$, $W$ is a learnable parameter matrix, and $\alpha_{uv}$ is the attention weight from node $v$ to node $u$:

$$\alpha_{uv} = \mathrm{softmax}\left(\sigma\left(a^T[Wh_u \| Wh_v]\right)\right), \qquad (2)$$

where $a$ is the attention parameter and $\|$ denotes vector concatenation. After GAT encoding, graph-level pooling produces the essay graph representation:

$$g = \mathrm{READOUT}(\{h'_1, h'_2, ..., h'_n\}), \qquad (3)$$

where $g$ denotes the aggregated essay structural representation, capturing both node semantics and structural relations.

We train a projector to map $g$ into the LLM token embedding space. During training and inference, the `<graph>` placeholder is replaced by the projected graph representation, allowing structural information to guide generation as a continuous embedding.

### 4.3 Training Pipeline

**Scoring-oriented SFT** The first stage learns the mapping from essay structure to multi-trait scores. Given the essay text and graph representation, the model generates only trait scores and provides initialization for the next stage. We optimize this stage with autoregressive cross-entropy loss.

**Scoring-Feedback Joint SFT** In the second stage, we continue training from the Stage 1 checkpoint to jointly generate scores and structure-grounded feedback. Given the essay text $e$ and graph representation $g$, the model produces an XML output, with the `<scores>` block preceding the `<comment>` block. This design first calibrates writing quality through multi-trait scores and then generates actionable feedback grounded in the score-calibrated structural diagnosis.

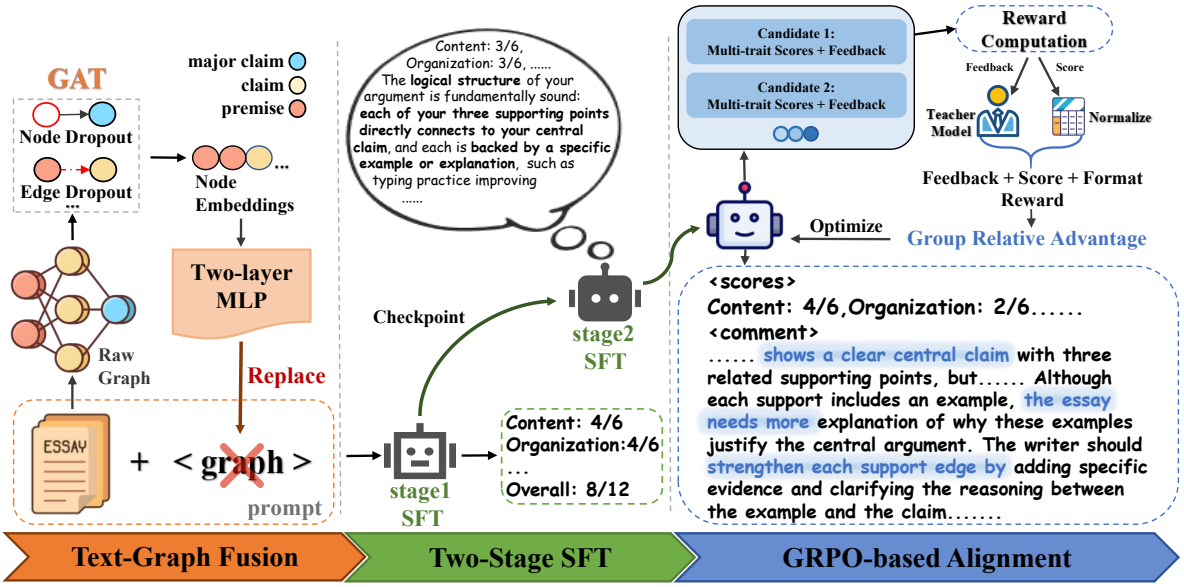


Figure 3: Overview of the GRAFE framework. GRAFE integrates essay graph representations into the LLM, learns score-feedback generation through two-stage SFT, and further aligns graph-grounded feedback with GRPO.

**GRPO-based Feedback Quality Alignment** After SFT, we apply GRPO to optimize sequence-level feedback quality. Given an input $x$ containing the essay text and graph representation, the old policy $\pi_{\theta_{\text{old}}}$ samples $K$ candidate outputs:

$$\{Y_1, Y_2, \ldots, Y_K\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x). \tag{4}$$

Each candidate is assigned a reward $R_i$ based on feedback quality, scoring consistency, and format compliance. GRPO then computes the relative advantage within the candidate group:

$$A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^K)}{\text{std}(\{R_j\}_{j=1}^K) + \epsilon}. \tag{5}$$

This group-wise normalization avoids training an additional value model and is suitable for long-form feedback generation.

### 4.4 Reward Function Design

The GRPO reward combines three components: teacher-based feedback quality, scoring consistency, and format compliance. These components balance feedback quality, score reliability, and output parsability.

**Teacher-based Feedback Reward** This reward uses a teacher model to evaluate each candidate comment. Given the writing prompt, essay, gold scores, essay graph, and generated feedback, the teacher assigns a score in $[0, 1]$ based on faithfulness, graph usage, structural diagnosis, and actionability.

**Scoring Consistency Reward** To provide a continuous sample-level signal, we compute scoring consistency using normalized absolute error. Let $\mathcal{S}_p$ denote the set of multi-trait scores. For each trait $k \in \mathcal{S}_p$, $\hat{s}_k$, $s_k$, and $M_k$ denote the predicted score, gold score, and maximum score, respectively:

$$R_{\text{score}} = \frac{1}{|\mathcal{S}_p|} \sum_{k \in \mathcal{S}_p} \left(1.0 - \frac{|\hat{s}_k - s_k|}{M_k}\right). \tag{6}$$

**Format Compliance Reward** We use a format compliance reward to ensure parsable outputs with one `<scores>` block followed by one `<comment>` block. Outputs with missing or repeated tags, invalid ordering, missing or out-of-range scores, or overly short comments are penalized. An example of the output format is provided in Appendix D.

## 5 Experiments

### 5.1 Experimental Setup

**Dataset** We evaluate GRAFE on the ASAP++ corpus[2] (Mathias and Bhattacharyya, 2018), which extends the original ASAP dataset[3] with fine-grained multi-trait scores across eight distinct writing prompts. Following the standard protocol (Taghipour and Ng, 2016), we adopt five-fold

[2]https://lwsam.github.io/ASAP++/lrec2018.html
[3]https://www.kaggle.com/c/asap-aes

| Model | Graph | Training | BERTScore | SAD | SGA | OFQ |
|---|---|---|---|---|---|---|
| Prompting | × | None | 0.600 | 1.820 | 1.813 | 1.817 |
| Graph-as-Text Prompting | JSON | None | 0.625 | 2.269 | 2.155 | 2.212 |
| Text-only SFT | × | SFT | **0.691** | 3.147 | 2.963 | 3.055 |
| Two-stage Graph SFT | Emb. | 2-stage SFT | 0.645 | 3.326 | 3.279 | 3.303 |
| **GRAFE** | Emb. | 2-stage SFT + GRPO | 0.639 | **3.771** | **3.679** | **3.725** |

Table 2: Overall feedback generation results. SAD: Structure-Aware Diagnosis, SGA: Structure-Grounded Actionability, OFQ: Overall Feedback Quality.

| Model | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | Avg. QWK |
|---|---|---|---|---|---|---|---|---|---|
| Prompting | 0.121 | 0.353 | 0.381 | 0.340 | 0.369 | 0.415 | 0.115 | 0.262 | 0.295 |
| Graph-as-Text Prompting | 0.254 | 0.391 | 0.405 | 0.352 | 0.386 | 0.480 | 0.314 | 0.276 | 0.357 |
| Text-only SFT | 0.603 | 0.574 | 0.689 | 0.708 | 0.641 | 0.679 | 0.614 | 0.463 | 0.621 |
| Two-stage Graph SFT | **0.629** | 0.655 | 0.681 | **0.773** | **0.676** | **0.773** | **0.732** | 0.634 | **0.694** |
| **GRAFE** | 0.510 | **0.656** | **0.703** | 0.750 | 0.560 | 0.746 | 0.725 | **0.640** | 0.661 |

Table 3: Prompt-wise QWK on ASAP++.

cross-validation with official splits. To ensure reliable feedback evaluation, we further employ human verification to improve the quality and reliability of the reference feedback.

**Evaluation Metrics** For the feedback quality evaluation, we employ an LLM judge to rate comments on a 1-5 scale via two metrics: **Structure-Aware Diagnosis (SAD)**, measuring accuracy in identifying structural flaws, and **Structure-Grounded Actionability (SGA)**, measuring the specificity and actionability of revision advice. For essay $i$ with prompt $p_i$, text $e_i$, graph $G_i$, gold score $s_i$, and generated comment $c_i$, the judge produces ratings denoted as $SAD_i$ and $SGA_i$. We report the **Overall Feedback Quality (OFQ)**, the arithmetic mean of $SAD_i$ and $SGA_i$, as a summary measure of feedback quality. To ensure fair evaluation, all generated comments are judged under the same evaluation context $(p_i, e_i, G_i, s_i)$; this context is used only by the judge and is not necessarily provided to the generation models. BERTScore (Zhang et al., 2020) is reported to measure semantic similarity to human-calibrated gold standards.

For the assessment of scoring consistency, we use Quadratic Weighted Kappa (QWK) (Cohen, 1968)—the standard metric for AES—to evaluate the agreement between predicted multi-trait scores and ground truth.

Additional implementation details are provided in Appendix C.

### 5.2 Baselines

Because existing essay feedback generation methods often use different datasets, output formats, or non-public implementations, we construct controlled baselines for comparison. The specific baseline settings are as follows:

**Prompting.** Following Stahl et al. (2024), we use text-only zero-shot prompting to generate feedback and multi-trait scores without parameter updates.

**Graph-as-Text Prompting.** We serialize the essay graph as JSON text and add it to the prompt to generate feedback and multi-trait scores without parameter updates (Fatemi et al., 2024).

**Text-only SFT.** Motivated by explanation- and rationale-supervised educational assessment (Li et al., 2023; Do et al., 2025), this baseline fine-tunes the same backbone with essay text, scores, and feedback, but without graph input.

**Two-stage Graph SFT.** Inspired by graph-based essay scoring and graph-LLM alignment (Li and Ng, 2025; Tang et al., 2024), as well as score-conditioned rationale or feedback generation in essay assessment (Do et al., 2025; Li and Pan, 2025), this baseline uses graph-vector input with two-stage SFT: score generation is learned first, followed by joint score-feedback generation.

### 5.3 Main Results

**Feedback Generation** Table 2 reports the overall feedback generation results. Prompting performs worst on SAD, SGA, and OFQ, while

| Model | Graph Input | QWK | SAD | SGA | OFQ | SF |
|---|---|---|---|---|---|---|
| Graph-as-Text Prompting | JSON Text | 0.357 | 2.269 | 2.155 | 2.212 | 2.672 |
| Two-stage SFT w/o Graph | None | 0.643 | 1.424 | 1.395 | 1.410 | 2.372 |
| Two-stage Graph SFT w/ Shuffled Graph | Mismatched Graph Emb. | 0.682 | 3.418 | 3.321 | 3.370 | 2.060 |
| Two-stage Graph SFT | Correct Graph Emb. | 0.694 | 3.326 | 3.279 | 3.303 | **3.552** |

Table 4: Graph-input validity stress test. SF denotes Structural Faithfulness and serves as the primary metric for assessing whether structural claims in feedback are supported by the correct essay graph.

| Type | Model | SAD | SGA | OFQ |
|---|---|---|---|---|
| Arg. P1–P2 | GRAFE w/o G. | 2.313 | 1.935 | 2.124 |
| | GRAFE | **3.835** | **3.770** | **3.803** |
| Src. P3–P6 | GRAFE w/o G. | 2.552 | 2.001 | 2.277 |
| | GRAFE | **3.790** | **3.673** | **3.732** |
| Narr. P7–P8 | GRAFE w/o G. | 1.957 | 1.768 | 1.863 |
| | GRAFE | **3.670** | **3.600** | **3.635** |

Table 5: Genre-wise feedback performance on ASAP++. GRAFE w/o G. denotes a GRAFE variant that follows the same two-stage SFT and GRPO pipeline as GRAFE, but removes graph inputs from all stages. G. denotes graph. Arg., Src., and Narr. denote argumentative, source-dependent, and narrative essays, respectively.

Graph-as-Text Prompting brings moderate gains, suggesting that structural information is useful but difficult to fully exploit when serialized as JSON. Although Text-only SFT achieves the highest BERTScore, graph-based models obtain much stronger SAD, SGA, and OFQ. This suggests that incorporating graph embeddings shifts the model from closely matching the wording of reference comments toward generating more structure-grounded and actionable feedback. GRAFE further achieves the best SAD, SGA, and OFQ after GRPO alignment. Therefore, BERTScore is used as a supplementary similarity metric, while SAD and SGA more directly reflect feedback quality in our setting.

**Scoring Performance** Table 3 reports QWK results for scoring consistency. Two-stage Graph SFT achieves the highest average QWK, indicating the benefit of score-oriented training with graph embeddings. GRAFE retains competitive QWK while substantially improving feedback quality, suggesting that GRPO enhances structure-grounded feedback with a limited trade-off in scoring reliability.

**Essay Type Results** Table 5 shows that GRAFE achieves the strongest performance in all genres, indicating that genre-specific graph priors consistently improve feedback generation. These results show that structural graphs provide useful evidence for producing more diagnostic and actionable feedback across different writing types.

### 5.4 Ablation Studies

**Does Graph Information Truly Matter?** To examine whether the model truly relies on graph information, we conduct a graph-input validity stress test in Table 4. We additionally report Structural Faithfulness (SF), whose evaluation prompt is provided in Table 9, to assess whether structural claims in feedback are supported by the essay and the correct graph. Compared with Two-stage SFT without graph input, the correct graph embedding substantially improves SAD, SGA, and OFQ, showing that graph information supports more diagnostic and actionable feedback. Graph-as-Text Prompting offers useful but limited structural cues. Although the shuffled-graph variant obtains slightly higher SAD and SGA than the correct-graph setting, its SF drops sharply from 3.552 to 2.060. This suggests that mismatched graphs may lead to structurally plausible feedback whose structural claims are less faithful to the essay's actual organization. Therefore, SAD and SGA should be interpreted together with SF, and valid graph embeddings are especially important for ensuring the faithfulness of structure-aware feedback.

**Does Two-stage SFT Help Exploit Graph Information?** Table 6 compares one-stage and two-stage graph-based SFT. Two-stage Graph SFT substantially improves SAD, SGA, OFQ, and QWK over One-stage Graph SFT, indicating that score-oriented pretraining helps the model learn the relationship between essay structure and writing quality. This score-calibrated initialization enables the second stage to generate more structurally grounded feedback and more reliable scores.

**Does GRPO Further Align Feedback?** We examine whether GRPO improves feedback alignment after two-stage SFT. As shown in Table 6,

| Model | SFT Strategy | Optimization | BERTScore | QWK | SAD | SGA | OFQ |
|---|---|---|---|---|---|---|---|
| One-stage Graph SFT | Score + Comment | CE | 0.626 | 0.448 | 1.772 | 1.694 | 1.733 |
| Two-stage Graph SFT | Score-only → Score + Comment | CE | **0.645** | **0.694** | 3.326 | 3.279 | 3.303 |
| GRAFE | Score-only → Score + Comment | CE + GRPO | 0.639 | 0.661 | **3.771** | **3.679** | **3.725** |

Table 6: Ablation results on two-stage SFT and GRPO optimization.

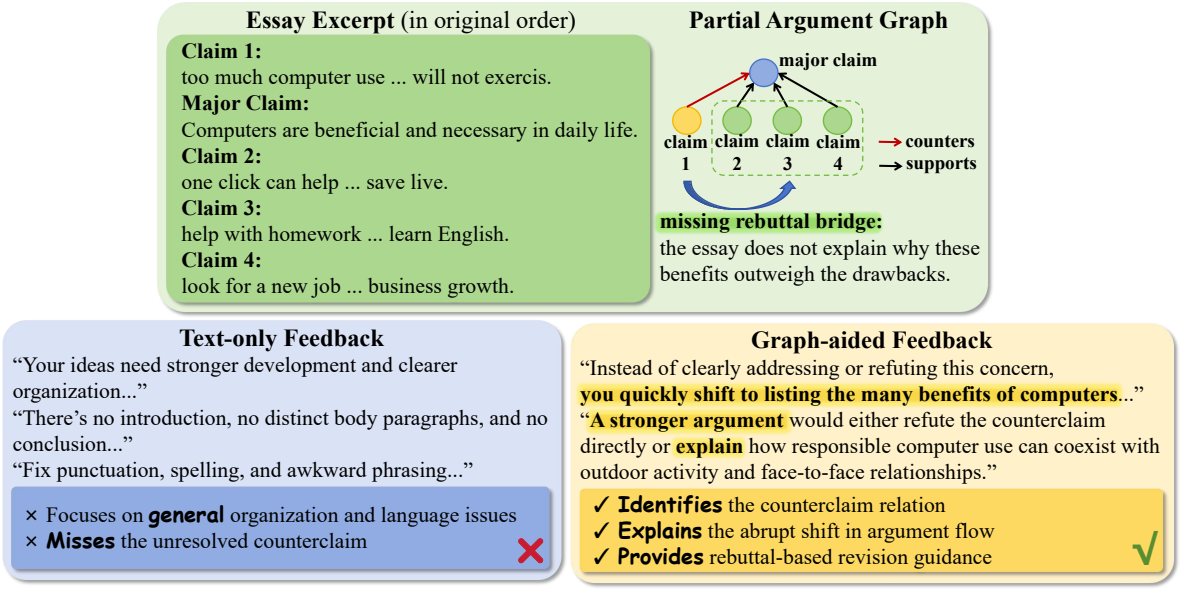


Figure 4: Case study comparing graph-aided and text-only feedback for an argumentative essay.

GRAFE improves SAD, SGA, and OFQ over Two-stage Graph SFT, while maintaining competitive QWK. This suggests GRPO mainly enhances the structural diagnosis and actionability of feedback, with only a modest trade-off in scoring consistency.

### 5.5 Case Study

Figure 4 shows an argumentative essay that claims computers are beneficial, but begins with a drawback about reduced exercise and face-to-face interaction. The text-only feedback mainly comments on general issues such as weak development, organization, and language errors, while missing that this opening point functions as an unresolved counterclaim. In contrast, the graph-aided feedback distinguishes the countering relation from the later supporting claims and identifies the missing rebuttal bridge. It therefore guides the student to either refute the concern directly or explain why the benefits of computer use outweigh the drawbacks, showing how graph information helps GRAFE generate more structurally grounded and actionable feedback.

## 6 Conclusion

In this paper, we introduced GRAFE, a framework that leverages explicit structural graphs to generate diagnostic and actionable essay feedback. Unlike conventional approaches that rely solely on textual input, GRAFE constructs genre-specific graphs to model argumentation, source integration, and narrative progression. To align the language model with these structural priors, we implemented a two-stage Supervised Fine-Tuning strategy and subsequently applied Group Relative Policy Optimization. This optimization process utilizes a multi-objective reward to balance diagnostic accuracy, scoring consistency, and format compliance. Evaluations on the augmented ASAP++ dataset indicate that GRAFE outperforms baseline models on structure-aware feedback quality metrics, while maintaining competitive scoring consistency. The observed improvements in Structure-Aware Diagnosis and Structure-Grounded Actionability confirm that explicit structural modeling enhances the pedagogical quality of automated writing evaluation.

## 7 Limitations

**Generalization Across Genres and Languages** While GRAFE demonstrates strong performance on three core genres within the English ASAP++ dataset, its applicability to broader, open-ended writing tasks (e.g., creative writing or domain-specific academic reports) and multilingual contexts remains unexplored. Future research could extend our graph-reasoned framework to diverse linguistic and task settings.

**Trade-off in Multi-Objective Alignment** In GRAFE, multi-trait scoring serves primarily as an auxiliary signal to maintain the evaluative tone and consistency of the text, rather than the primary optimization target. Because our objective functions—particularly the teacher-based reward in GRPO—prioritize pedagogical feedback quality, the model may trade off some numerical scoring precision to generate more descriptive and targeted feedback. Identifying an optimal balance to simultaneously maximize both generative feedback utility and automated scoring reliability presents an important challenge for subsequent iterations of this work.

## 8 Ethics Statement

Our work adheres to the ACL Code of Ethics. We utilize the publicly available ASAP++ corpus for our experiments and dataset construction. Because the source corpus was fully anonymized by its creators prior to release, our work contains no protected or personally identifiable information (PII). All synthesized structural graphs and feedback are derived solely from these anonymized essays.

For human calibration of the gold-standard test set, we recruited three senior annotators with academic backgrounds in English Linguistics or Education and writing assessment, alongside a senior arbitrator. Annotators were informed of the project's educational purpose and compensated above the local minimum wage. The calibration tasks involved no sensitive information. Detailed calibration workflows, guidelines, and criteria are provided in Appendix B.

All baseline models, LLMs, and datasets are used in accordance with their intended usage and licenses. GRAFE is designed as a pedagogical aid to support students and alleviate grading burdens, rather than to replace human instructors in educational assessment.

## References


Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Seongyeub Chu, Jongwoo Kim, and Munyong Yi. 2026. FeedEval: Pedagogically aligned evaluation of LLM-generated essay feedback. *arXiv preprint arXiv:2601.04574*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Heejin Do, Yunsu Kim, and Gary Lee. 2024. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. 2025. Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring. *arXiv preprint arXiv:2502.20748*.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 82 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Donghee Han, Daeyoung Roh, Euihwan Han, Hwanjun Song, and Mun Yong Yi. 2025. Fine-grained multi-prompt essay scoring with multi-level disentanglement. *Data Mining and Knowledge Discovery*, 39(67).

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6007–6026, Singapore. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2025. Graph-based multi-trait essay scoring. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33325–33351, Suzhou, China. Association for Computational Linguistics.

Xia Li and Wenjing Pan. 2025. CEAES: Bidirectional reinforcement learning optimization for consistent and explainable essay assessment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26267–26279, Vienna, Austria. Association for Computational Linguistics.

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).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

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, Vienna, Austria. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

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.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Jun Zhong, Longwei Xu, Li Kong, Xianzhuo Li, Dandan Liang, and Junsheng Zhou. 2025. NarGINA: Towards accurate and interpretable children's narrative ability assessment via narrative graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14843–14860, Vienna, Austria. Association for Computational Linguistics.

# A Details on Feedback Data Generation and Automated Verification

## A.1 Feedback Generation Configuration

In the feedback synthesis phase, we utilized the Large Language Model (LLM) Qwen-max to perform multi-round sampling. To ensure sufficient exploration and diversity in the generated responses, the temperature coefficient $T$ was set to

1.0. During the generation process, the model input strictly adhered to two experimental configurations: in the "Graph-aided" setting, the input encompassed the writing prompt, essay text, multi-trait scores, and the corresponding structural graph in JSON format; in the "Text-only" control setting, the structural graph information was omitted while maintaining identical prompt templates and generation parameters. This rigorously controlled design ensures that the subsequent pairwise comparison experiments accurately reflect the specific contribution of structural graphs to feedback quality.

### A.2 Pairwise Comparison Workflow via LLM Evaluator

To verify data quality, we employed Qwen3.5-plus-thinking as an automated evaluator to conduct pairwise comparisons. To mitigate potential experimental biases, we implemented several systematic strategies. First, we enforced anonymization and order randomization, ensuring the evaluation model remained unaware of the feedback source while randomly shuffling the presentation order of "Graph-aided" and "Text-only" comments (i.e., randomly assigning them as Comment A or Comment B). Second, we conducted stability tests by including 50 pairs of identical comments within a 200-sample set; the results confirmed the model's ability to consistently identify identical quality. Furthermore, repeated evaluations with swapped positions demonstrated that the evaluator was not sensitive to positional bias.

### A.3 Evaluation Traits and Criteria

The automated evaluator was instructed to assess comments based on five core traits: (1) Structural Diagnostic Ability, which measures the accuracy in identifying deep logical issues such as unsupported claims, weak evidence, or incomplete narrative chains; (2) Actionability of Revision Suggestions, evaluating whether the feedback provides concrete, executable advice based on the diagnosis; (3) Grounding in Original Evidence, ensuring the feedback relates closely to the actual essay content rather than being generic; (4) Score Consistency, verifying that the sentiment and evaluative tone of the feedback align with the essay's actual scores; and (5) Language Naturalness. A comment was only judged as superior if it demonstrated overall excellence across these traits.

## B Human Calibration and Gold Standard Construction

### B.1 Expert Annotation Team Background

The human calibration process was conducted by three senior annotators with extensive experience in English writing assessment. All team members possess academic backgrounds in English Linguistics or Education and have long-term experience in teaching and grading English as a Second Language (ESL) essays. Additionally, a senior arbitrator was appointed to resolve any discrepancies among annotators and perform a comprehensive quality audit of the final revised feedback to ensure the authority and reliability of the gold-standard data.

### B.2 Sampling Strategy and Calibration Workflow

To ensure the gold standard covered various prompt types and scoring traits, we adopted a prompt-balanced sampling strategy. We sampled 40 essays from each of the eight prompts in ASAP++ (including argumentative, source-dependent, and narrative genres), totaling 320 essays for the calibration subset. The annotation process was divided into two distinct stages: the *Structural Judgment* stage, where annotators first identified the actual structural flaws in the essay and compared them with the initial LLM-generated feedback; and the *Textual Revision* stage, where annotators performed targeted corrections to address structural hallucinations, over-generalization, polarity conflicts, and internal logical contradictions within the feedback.

### B.3 Fleiss' Kappa Consistency Calculation

To validate the reliability of identifying structural problems during human calibration, we calculated the inter-annotator agreement using the *Essay-Graph-Feedback* triplet as the basic unit. Annotators were required to perform categorical labeling of specific structural issues, such as *unsupported claim* or *missing evidence*. The resulting Fleiss' Kappa value was 0.68, which represents *substantial agreement* in categorical annotation tasks. This coefficient demonstrates that the annotators achieved high consistency in identifying core structural issues, thereby providing a credible structural foundation for the synthesized gold-standard feedback.

## C Implementation Details

We use `Meta-Llama-3.1-8B-Instruct` as the backbone model (Grattafiori et al., 2024). In the second-stage SFT, we set the learning rate to $3 \times 10^{-6}$ and the maximum sequence length to 1024. GRPO is initialized from the second-stage SFT checkpoint and trained for 1 epoch with a batch size of 1. For each input essay, GRPO samples $K = 6$ candidate score-feedback outputs to compute group-relative advantages. The GRPO reward combines teacher-model feedback quality, scoring consistency, and XML format rewards with weights of 0.90, 0.05, and 0.05, respectively. Training is conducted on 7 NVIDIA RTX A6000 GPUs with 48GB memory each.

## D Output Format Example

Table 7 provides an example of the XML-style output format corresponding to Prompt 1. Since the scoring traits and score ranges vary across different prompts in the ASAP++ dataset, we adopt a prompt-specific format validation approach during the training and evaluation phases.

**Prompt 1 Output Format Example**

```
<scores>
  Content: <content_score>/6
  Organization: <organization_score>/6
  Word Choice: <word_choice_score>/6
  Sentence Fluency: <sentence_fluency_score>/6
  Conventions: <conventions_score>/6
  Overall: <overall_score>/12
</scores>
<comment>
  <comment_text>
</comment>
```

Table 7: Example of the XML-style output format for Prompt 1. Other prompts employ corresponding format templates tailored to their specific scoring traits and score ranges.

## E Prompt for Structure-Aware Feedback Evaluation

Table 8 summarizes the core judge prompt used to evaluate Structure-Aware Diagnosis (SAD) and Structure-Grounded Actionability (SGA). The same evaluation protocol is applied to all models.

| Component | Prompt Content |
|---|---|
| Role | You are an expert evaluator of educational essay feedback. Your task is to evaluate whether the generated feedback is grounded in the essay's internal structure and useful for revision. |
| Evaluation Target | Evaluate the generated comment on *Structure-Aware Pedagogical Quality*, which means that the feedback should not only be fluent or encouraging, but also reflect structural aspects of the essay, such as claims, supporting reasons, evidence use, source integration, event sequence, causal progression, or narrative consequence. |
| Score Scale | Use a 1–5 scale for each dimension: 1 = poor, generic, incorrect, or unsupported; 2 = weak or vague; 3 = partially relevant but incomplete; 4 = clearly relevant and mostly specific; 5 = precise, well grounded, diagnostic, and useful for revision. |
| SAD | **Structure-Aware Diagnosis**: Assess whether the comment identifies important structural issues in the essay, such as unsupported claims, weak evidence, insufficient source use, missing event consequences, or poorly connected ideas. Generic comments or comments focusing only on surface-level issues should receive low scores. |
| SGA | **Structure-Grounded Actionability**: Assess whether the comment provides specific and useful revision advice based on the diagnosed structural problems, such as adding evidence, strengthening support relations, improving source use, clarifying causal or temporal progression, or completing narrative event chains. Vague suggestions without structural grounding should receive low scores. |
| Scoring Rules | Do not reward comments merely for mentioning technical terms such as graph, node, edge, claim, or premise. Reward comments that translate structural information into natural teacher feedback. Penalize generic, template-like, factually unsupported, or hallucinated feedback. If the comment is structurally aware but factually wrong, lower both scores. |
| Input Fields | Writing prompt; student essay; gold scores; graph in JSON format; generated comment. |
| Output Format | `Structure-Aware Diagnosis:<score>, Structure-Grounded Actionability:<score>` |

Table 8: Core LLM judge prompt for evaluating Structure-Aware Diagnosis and Structure-Grounded Actionability.

| Component | Prompt Content |
|---|---|
| Role | You are an expert evaluator of educational essay feedback. |
| Task | Evaluate whether the generated feedback is faithful to the student essay and its correct structural graph. |
| Input Fields | Writing prompt; student essay; gold scores; correct graph; generated feedback. |
| Evaluation Dimension | Structural Faithfulness: assess whether the structural claims in the feedback are supported by the essay and the correct graph. Structural claims include unsupported claims, weak evidence, insufficient source integration, missing event consequences, unclear causal progression, or poorly connected ideas. |
| Score Scale | Use a 1–5 scale: 1 = major unsupported or hallucinated structural diagnoses; 2 = several weakly supported or partly incorrect structural claims; 3 = mostly consistent but with some vague or weakly grounded claims; 4 = largely faithful with only minor omissions or imprecise phrasing; 5 = fully faithful to the essay and correct graph, with all structural diagnoses and suggestions well supported. |
| Scoring Rules | Do not reward feedback for merely mentioning graph-related terms. Penalize feedback that introduces structural problems not present in the essay or transfers issues from an incorrect or unrelated essay. |
| Output Format | Structural Faithfulness: <score> |

Table 9: Prompt for evaluating Structural Faithfulness.