When Rubrics Change: Cross-Rubric Generalization for Critical Thinking Essay Scoring

Nischal Ashok Kumar[1], Payu Wittawatolarn[1], Sana Kang[1], Marisa C. Peczuh[2], Blair Lehman[3], Ryan Baker[4], Caitlin Mills[2], Sherry Lachman[5], Ruochen Sun[6] and Andrew Lan[1]

[1]*University of Massachusetts Amherst, United States*
[2]*University of Minnesota, United States*
[3]*Brighter Research, United States*
[4]*Adelaide University, Australia*
[5]*Advanced Education Research and Development Fund (AERDF), United States*
[6]*Independent Researcher, United States*

Abstract

Automated essay scoring (AES) research has largely focused on cross-prompt generalization, where essays from unseen prompts are scored while the scoring criteria are typically held constant. In practice, however, educators may revise or even introduce new rubrics in their scoring task, to evaluate different aspects of essays. We study *cross-rubric generalization*: training on essays labeled under one set of rubrics and evaluating on previously unseen rubrics, which target different aspects of the essay. We use a Large Language Model (LLM) fine-tuning framework with two components: rubric-agnostic intermediate representations, called *traits*, and target-essay supervision under seen rubrics during training. On an AES dataset augmented with multiple rubric-defined labels of student critical thinking skills, we find that traits improve macro F1 by 5.0% over a baseline without traits in the hardest setting, where both target rubrics and target essays are unseen during training. We further find that increasing target-essay supervision improves performance, with our best fine-tuned open-source Llama-based model outperforming GPT-5-mini prompting by 2.1% macro F1 and trailing GPT-5 by 1.9%. These results show that trait-based intermediate structure and controlled supervision improve generalization to unseen rubrics.

github.com/umass-ml4ed/generalization-in-essay-scoring

Keywords

Automated essay scoring, critical thinking, cross-rubric generalization, large language models

1. Introduction

Automated essay scoring (AES) is a long-standing and widely studied problem, motivated by the need to support scalable and consistent assessment of student writing [1, 2, 3]. Beyond in-domain scoring, a growing line of work has examined how AES systems generalize to essays written in response to previously unseen prompts, where the essay content and topical distributions change but the scoring criteria (*i.e.*, the target rubric) do not [4, 5, 6]. This cross-prompt setting has become an important testbed for robust AES, since practical deployments often require models to score writing drawn from new prompts without retraining from scratch for each prompt.

However, real-world assessment settings can require more than generalizing to new prompts. Educators may revise scoring criteria or even introduce new rubrics that evaluate different aspects of the essay; obtaining expert annotation on these new rubrics can be expensive. This challenge is especially relevant in critical thinking assessment, where writing may be evaluated along multiple criteria, each defined by its own rubric, and new rubric criteria may be added over time [7]. Thus, an interesting problem to study in such settings is *cross-rubric generalization*: training a model on essays labeled under one rubric and applying it to score essays under a previously unseen rubric. Unlike cross-prompt generalization [4, 5, 6], cross-rubric generalization changes the scoring criteria used to evaluate the essay. The model must therefore generalize to a new rubric, where the relevant textual evidence, its interpretation, and the definition of proficiency may differ from those seen during training.

AI for Education Day, SIGKDD 2026, August 12, 2026, Jeju, Korea
nashokkumar@umass.edu (N. Ashok Kumar)

Table 1
Rubric-agnostic analytic traits used in our framework for critical thinking essay scoring.

Trait
Claim Explicitness: how clearly the essay states its main point or position
Structural Coherence: how well the ideas are organized and connected across the essay
Supporting Evidence: how much the essay uses examples, facts, or details to support its points
Evidence–Claim Linkage: how clearly the essay explains why the evidence supports its main point
Alternative Perspectives: how much the essay considers and responds to other viewpoints on the issue
Analytical Depth: how much the essay goes beyond description to compare, evaluate, or reason ideas

To address this issue, we study cross-rubric generalization within an LLM fine-tuning framework for rubric-based essay scoring from essay text and rubric descriptions. We instantiate this setting on the recent critical thinking essay scoring dataset of [7], which contains 500 student essays annotated under six rubric-based evaluation dimensions spanning three broader critical thinking skills. A central component of our approach is a set of rubric-agnostic analytic representations, which we call *traits*, grounded in argumentation and critical thinking theory. These traits capture lower-level properties that recur across the rubrics in this dataset, such as claim explicitness and evidence–claim linkage, without directly mirroring rubric definitions, and serve as a shared intermediate representation between essays and rubric labels. During training, we use them as intermediate supervision, either by providing them as additional inputs or by jointly predicting them with the final rubric label (see Section 4.2). Table 1 summarizes the traits used in our framework along with their brief descriptions.

In addition to rubric-agnostic traits as one component of our framework, we study a second component: how does the level of supervision available on *essays* (and seen rubrics) during training affects generalization. We consider three settings: **No Labels**, the hardest setting, where target essays are not observed during training; **Pseudo Labels**, where target essays are added using automatically generated labels under seen rubrics; and **Gold Labels**, where target essays are added using human-annotated labels under seen rubrics (see Section 4.3). In all cases, the target rubric remains unseen during training.

We jointly examine these two components and find that both are important for cross-rubric generalization, but in different ways. Traits are especially helpful in the hardest *No Labels* setting, where the model has no access to target essays during training, improving macro F1 by 5.0% compared to a baseline without traits. Performance improves further as more target-essay supervision becomes available. Under the strongest supervision setting, our best model outperforms the proprietary prompting-only GPT-5-mini by 2.1% in macro F1 and trails GPT-5 by only 1.9%. Our contributions are as follows: (1) We study *cross-rubric generalization* in automated essay scoring, grounded in critical thinking essay scoring across six rubrics. (2) We present a systematic study of two key components: rubric-agnostic *intermediate* trait representations and target-essay supervision from seen rubrics during training. (3) We show that these two components improve generalization in complementary ways: traits help in the hardest setting, where no target-essay labels are available under seen rubrics, while the strongest target-essay supervision enables a fine-tuned open-source model to approach proprietary prompting-based models.

2. Related Work

Generalization in Automated Essay Scoring. A large body of work in automated essay scoring (AES) studies generalization beyond the training distribution, with the dominant focus on *cross-prompt* transfer. In this setting, models are evaluated on essays from unseen prompts, with the primary source of shift being the essay prompt rather than the target rubric. Prior work has explored this problem through cross-prompt trait scoring [4], grammar-aware modeling for prompt transfer [5], rubric-derived assessment questions and features [6], and human–AI collaborative scoring pipelines [8]. In contrast, we study generalization across rubrics, training on one set of rubrics and evaluating on a previously unseen set. Closest to our work, [7] study generalization to previously unseen critical-thinking rubrics

by conditioning on rubric text, but in a setting that primarily isolates rubric shift rather than requiring simultaneous generalization across both essays and rubrics. Our work extends this line of research in two ways. First, we cast the problem more broadly as *cross-rubric generalization* and study it under multiple generalization conditions, including both joint essay-and-rubric shift and settings that isolate rubric shift. Second, we systematically analyze two components for improving generalization to unseen rubrics: rubric-agnostic intermediate trait representations and target-essay supervision during training.

Rubric-Grounded Structured Scoring. Recent LLM-based essay scoring methods increasingly use rubrics to construct intermediate structure rather than relying only on direct end-to-end grading. For example, prior work has proposed trait-focused conversational decomposition [9], rubric-derived assessment questions and features [6], trait-specific rationales [10], multi-agent extraction of rubric-relevant components [11], and structured human–AI scoring workflows [8]. Related work outside AES also points to the value of portable structured signals and interpretable latent factors [12, 13]. Our work is closely related to this line of research, but differs in both setting and scope. We study generalization to *previously unseen rubrics*, rather than settings where evaluation remains within seen rubrics or where the primary shift is across prompts, and we systematically analyze two components that may support transfer in this setting: rubric-agnostic trait annotations and target-essay supervision during training.

3. Task Setup

3.1. Task Definition

Rubric-based assessment requires models to map essays to discrete proficiency levels defined by natural-language rubrics, and poses a key challenge of generalizing to evaluation criteria not seen during training. Let $\mathcal{E}$ denote a set of essays and $\mathcal{R} = \{r_1, \ldots, r_K\}$ denote a set of rubrics, where each rubric defines an evaluation criterion in natural language. Each rubric $r \in \mathcal{R}$ defines a labeling function

$$f_r : \mathcal{E} \to \mathcal{Y}_r,$$

where $\mathcal{Y}_r$ is the associated discrete proficiency-level scale, i.e., a set of ordinary scores. We partition essays and rubrics into disjoint subsets $\mathcal{E}_{\text{train}}, \mathcal{E}_{\text{test}}$ and $\mathcal{R}_{\text{train}}, \mathcal{R}_{\text{test}}$, respectively, with $\mathcal{E}_{\text{train}} \cup \mathcal{E}_{\text{test}} = \mathcal{E}$ and $\mathcal{R}_{\text{train}} \cup \mathcal{R}_{\text{test}} = \mathcal{R}$. We denote the labeled training data, provided by human annotation, as

$$\mathcal{D}_{\text{train}} = \{(e, r, y) \mid e \in \mathcal{E}_{\text{train}},\ r \in \mathcal{R}_{\text{train}},\ y = f_r(e)\}.$$

A model M is trained to approximate f_r for rubrics in $\mathcal{R}_{\text{train}}$ using $\mathcal{D}_{\text{train}}$ as the base training set. Depending on the supervision setting, this base set may be augmented with essays from $\mathcal{E}_{\text{test}}$ and/or rubrics from $\mathcal{R}_{\text{train}}$ (Section 4.3). At test time, the model is evaluated on

$$\mathcal{D}_{\text{test}} = \{(e, r', y) \mid e \in \mathcal{E}_{\text{test}},\ r' \in \mathcal{R}_{\text{test}},\ y = f_{r'}(e)\},$$

without access to labeled data from $\mathcal{R}_{\text{test}}$ during training. The model is provided with the textual description of the target rubric $r' \in \mathcal{R}_{\text{test}}$. The objective is to evaluate whether M can approximate $f_{r'}$ for unseen rubrics on essays in $\mathcal{E}_{\text{test}}$, requiring generalization to new rubric-defined label functions. The hardest supervision setting additionally involves generalization to unseen essays.

3.2. Dataset

We instantiate our setup using the dataset of [7], which contains 500 student essays from the public portion of the PERSUADE 2.0 corpus. These essays are argumentative writing samples from 6th–12th grade U.S. students, covering civic, scientific, and social topics. The dataset defines six rubrics, organized under three broader skills: *Information Analysis*, *Argument Generation*, and *Logical Reasoning*, with two rubrics per skill. Each rubric is specified in natural language through a name and definition and a rubric description, which together form the *rubric specification*, and assigns essays a 5-point rubric label (0–4: *Not Applicable*, *Below Emerging*, *Emerging*, *Expanding*, *Exemplifying*). Because each essay is annotated under all six rubrics, the dataset allows us to decouple essays from rubrics and study cross-rubric generalization. Appendix Table 3 lists the rubric names and definitions.

4. Methods

In this section, we detail our approach to cross-rubric generalization within an LLM fine-tuning framework for rubric-based essay scoring. We first present the core rubric assessment setup, then introduce trait-based representations as rubric-agnostic intermediate structure, and finally define supervision settings that vary the availability of essay–rubric labels during training while preventing leakage from unseen rubrics.

4.1. LLM fine-tuning framework for rubric-based essay scoring

We build on the LLM-based framework for rubric-based essay scoring introduced by [7], which formulates rubric-based evaluation as a conditional text generation task. Specifically, we fine-tune a pretrained open-source LLM (M) (Llama 3.1 8B Instruct [14]) using standard cross-entropy loss to predict rubric labels for essay–rubric pairs. For each essay–rubric pair, the model takes as input the essay text together with the full rubric specification and is trained to generate a natural-language justification followed by the corresponding rubric label, yielding a rationale-then-decision output format [15]. The justification is a short explanation of why the essay satisfies the gold rubric label, and is generated by a separate LLM prior to fine-tuning. Appendix A.5 provides ablations on rubric specification and justifications. Training uses the base training set $\mathcal{D}_{\text{train}}$, while evaluation is conducted on $\mathcal{D}_{\text{test}}$. Across all settings, rubrics in $\mathcal{R}_{\text{test}}$ remain unseen during training; the availability of essays from $\mathcal{E}_{\text{test}}$ under rubrics in $\mathcal{R}_{\text{train}}$ varies by supervision setting (Section 4.3).

4.2. Trait-Based Representations

To improve generalization to previously unseen rubrics, we introduce *rubric-agnostic analytic traits* as an intermediate representation between essays and rubric-specific labels. Prior work in AES shows that structured intermediate features can improve transfer by capturing reusable aspects of writing quality [6]. We extend this idea to the cross-rubric setting by grounding these traits in argumentation and critical thinking theory, which characterizes writing quality in terms of lower-level components such as claims, evidence, reasoning, and perspective-taking [16, 17, 18].

Concretely, we decompose rubric-based scoring as: $e \rightarrow h(e) \rightarrow f_r(e)$, where $h(e)$ captures rubric-independent argumentative properties. This factorization encourages the model to learn transferable reasoning representations that can be flexibly mapped to different rubric-specific criteria.

Trait Schema. We define a fixed set of six analytic traits derived from established frameworks in argumentation theory and critical thinking. These traits are (i) lower-level criteria that underlie rubrics, (ii) shared across all rubrics, and (iii) defined prior to experimentation and held fixed across train/test splits. The six traits are: *Claim Explicitness*, *Structural Coherence*, *Presence of Supporting Evidence*, *Evidence–Claim Linkage*, *Consideration of Alternative Perspectives*, and *Analytical Depth*. Each trait is associated with a three-level categorical value scale (e.g., absent/partial/strong), capturing increasing proficiency. These traits are designed to capture fundamental argumentative properties that are shared across all rubric definitions. Table 4 in the Appendix summarizes each trait, along with its annotation question and value scale.

Trait Annotation. For each essay $e \in \mathcal{E}$, we use an LLM to generate trait annotations consisting of (i) a categorical *value* for each trait and (ii) a short natural language *observation* explaining the label assignment. These annotations define a rubric-agnostic representation $h(e)$ shared across all rubrics. Trait annotations are automatically generated (silver labels) for all essays in $\mathcal{E}$, and do not use labeled data from $\mathcal{R}_{\text{test}}$, ensuring compatibility with the cross-rubric setting. We acknowledge that these annotations are not validated through human annotation and treat them as intermediate structure that supports training, leaving human validation to future work.

Trait-Augmented Fine-Tuning. We incorporate trait representations into model fine-tuning in four configurations, depending on whether the model uses only trait *values* or both trait *values* and natural-language *observations*, and whether these traits are provided as inputs or predicted as outputs. In

Input: Trait Values, we append only the categorical trait values to the input alongside the essay and rubric. In *Input: Trait Values + Observations*, we append both the trait values and their corresponding natural-language observations. In *Output: Trait Values*, the model is trained to jointly predict the categorical trait values together with the final rubric label. In *Output: Trait Values + Observations*, the model is trained to jointly generate the trait values, the accompanying observations, and the final rubric label. These variants let us test whether traits are more useful as input scaffolds or as auxiliary prediction targets for generalization to (e, r') pairs with $e \in \mathcal{E}_{\text{test}}$ and $r' \in \mathcal{R}_{\text{test}}$.

4.3. Supervision Settings

We consider three supervision settings that differ in whether essays from $\mathcal{E}_{\text{test}}$ are observed during training under rubrics in $\mathcal{R}_{\text{train}}$, while ensuring that rubrics in $\mathcal{R}_{\text{test}}$ are never used for supervision:

- In the **No Labels** setting, the model M is trained on $\mathcal{D}_{\text{train}} = \{(e, r, y) \mid e \in \mathcal{E}_{\text{train}},\ r \in \mathcal{R}_{\text{train}}\}$ and evaluated on $\mathcal{D}_{\text{test}} = \{(e, r', y) \mid e \in \mathcal{E}_{\text{test}},\ r' \in \mathcal{R}_{\text{test}}\}$. This setting represents the hardest generalization scenario where we evaluate the model's ability to generalize to both unseen essays and unseen rubrics.
- In the **Pseudo Labels** setting, we augment training data using self-training. We first train a model M in the No Labels setting. We then use M to generate pseudo labels for (e, r) where $e \in \mathcal{E}_{\text{test}}$ and $r \in \mathcal{R}_{\text{train}}$. These pseudo-labeled examples, corresponding to essays from $\mathcal{E}_{\text{test}}$, are combined with $\mathcal{D}_{\text{train}}$ to train a new model M'. The resulting model M' is evaluated on (e, r') with $e \in \mathcal{E}_{\text{test}}$ and $r' \in \mathcal{R}_{\text{test}}$. This setting provides the model with training signals from $\mathcal{E}_{\text{test}}$ without using human-annotated labels.
- In the **Gold Labels** setting, essays from $\mathcal{E}_{\text{test}}$ are available during training, but only under rubrics in $\mathcal{R}_{\text{train}}$. Specifically, we train M on $\{(e, r, y) \mid e \in \mathcal{E}_{\text{train}} \cup \mathcal{E}_{\text{test}},\ r \in \mathcal{R}_{\text{train}}\}$, using gold (human-annotated) labels for all essay–rubric pairs under $\mathcal{R}_{\text{train}}$. The model is then evaluated on (e, r') with $e \in \mathcal{E}_{\text{test}}$ and $r' \in \mathcal{R}_{\text{test}}$. This setting is closest to [7], but differs in that evaluation is performed only on $\mathcal{E}_{\text{test}}$, not on the full essay set $\mathcal{E}$. It therefore isolates generalization across rubrics while preserving a held-out essay split at evaluation time. Importantly, in both Pseudo and Gold Labels settings, essays from $\mathcal{E}_{\text{test}}$ are observed during training only under $\mathcal{R}_{\text{train}}$, and no information from $\mathcal{R}_{\text{test}}$ is used for supervision.

5. Experimental Setup

We detail the experimental setup for evaluating cross-rubric generalization. We first define the structured design space explored in our experiments, together with the baselines, and then the dataset splitting strategy, and evaluation metrics.

5.1. Design Space, and Baselines

We define a unified experimental framework consisting of (i) a design space of modeling choices based on the methods detailed above, (ii) a baseline configuration based on our method, and (iii) prompting-based baselines. We detail each of these components below.

Design space. Our experiments study cross-rubric generalization within the LLM-based framework for rubric-based essay scoring detailed in Section 4. We study a combinatorial design space over two complementary components: (i) trait-based intermediate representations (Section 4.2), and (ii) target-essay supervision settings (Section 4.3). This design enables us to evaluate both the individual and combined effects of these two components.

Fine-tuning Baseline. Our primary baseline is the configuration without traits or target-essay labels *(No Labels, No Traits)*. It reflects the standard cross-rubric generalization scenario: the model is trained on essays and rubrics available during training and evaluated on unseen essay–rubric pairs. All other configurations extend this baseline by adding trait representations, target-essay supervision, or both.

Table 2
Comparing different trait configurations and supervision settings proposed in our paper. *No Traits* under *No Labels* is our fine-tuning baseline, while Zero-Shot prompted models serve as prompting baselines. Best numbers in each column are **bolded**.

	No Labels			Pseudo Labels			Gold Labels		
Trait Configuration / Model	**Acc**	**F1**	α	**Acc**	**F1**	α	**Acc**	**F1**	α
No Traits	0.399	0.353	0.346	0.412	0.376	**0.368**	0.470	**0.431***	**0.412**
Input: Trait Values	0.406	0.367	0.363	**0.446**	**0.400**†	0.360	0.473	**0.433****	0.378
Input: Trait Values and Observations	0.382	0.341	0.291	0.417	0.369	0.366	0.451	0.405†	0.293
Output: Trait Values	**0.436**	**0.403***	**0.363**	0.416	0.383	0.306	**0.495**	0.415*	0.369
Output: Trait Values and Observations	0.387	0.329	0.195	0.424	0.360	0.251	0.444	0.373	0.392
Llama 3.1 8B Instruct (Zero-shot)	0.372	0.222	0.251		—			—	
GPT-5-mini (Zero-shot)	0.492	0.410†	0.388		—			—	
GPT-5 (Zero-shot)	**0.532**	**0.450***	**0.476**		—			—	

Markers on F1 denote paired bootstrap significance vs. *(No Labels, No Traits)* baseline: $^{\dagger}p < 0.05$, $^{*}p < 0.01$, $^{**}p < 0.001$.

Prompting-based Baselines. We compare against three prompting-based baselines. Two are proprietary LLMs (GPT-5-mini and GPT-5), following [7] for rubric label generation. In the zero-shot setting, these models receive the essay and the target rubric specification at inference time ($r' \in \mathcal{R}_{\text{test}}$), without rubric-specific training, and output both a proficiency label and a concise justification explaining the prediction. Their ability to reason directly over unseen rubrics makes them strong baselines for cross-rubric generalization. The third baseline is Llama 3.1 8B Instruct [14] used in a zero-shot prompting setting (without fine-tuning), receiving the same essay and rubric specification and outputting only a proficiency label. This baseline allows us to isolate the contribution of fine-tuning by comparing zero-shot prompting and fine-tuned variants of the same underlying model. In contrast to all prompting-based baselines, our fine-tuning approach learns transferable representations over seen rubrics. See Appendix A.4 for the implementation details of our experiments.

5.2. Dataset Splitting

We partition the set of essays $\mathcal{E}$ into disjoint subsets $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{test}}$ using an 80/20 split. From $\mathcal{E}_{\text{train}}$, we further reserve 10% as a validation set, resulting in an effective 70/10/20 split for train/validation/test essays. We also partition the set of rubrics $\mathcal{R}$ according to their organization into three skills. For each of the three possible choices of held-out skill, we designate the two rubrics under that skill as $\mathcal{R}_{\text{test}}$ and use the remaining four rubrics from the other two skills as $\mathcal{R}_{\text{train}}$. This rubric partitioning ensures that all test rubrics are unseen during training. The full splitting procedure yields three rubric partitions, each with its own $\mathcal{R}_{\text{train}}$–$\mathcal{R}_{\text{test}}$ split. We train and evaluate one model for each rubric partition, concatenate predictions across the three held-out test sets, and compute evaluation metrics on the combined predictions.

5.3. Evaluation Metrics

Following [7], we report Accuracy, Macro F1 (denoted as F1), and Krippendorff's α (denoted as α). Accuracy measures the proportion of essay–rubric pairs for which the predicted label exactly matches the gold label. Macro F1 averages per-label F1 scores, giving equal weight to all labels, better reflecting performance on less frequent ones; we treat it as our primary metric. Krippendorff's α measures agreement with gold labels while accounting for both chance agreement and the ordinal structure of the proficiency scale, making it well suited for this graded evaluation setting. To assess statistical significance, we use paired bootstrap resampling [19] with 2,000 resamples on macro F1; results are considered significant at p<0.05. All comparisons are made against the *(No Labels, No Traits)* baseline.

6. Results

We present results for cross-rubric generalization. We first analyze the impact of key design choices, including trait-based representations and supervision settings, and then compare our fine-tuned models against the prompting baselines. Table 2 summarizes results across trait-based representations and supervision settings.

6.1. Impact of Design Choices on Cross-Rubric Generalization

We observe three main trends: (i) incorporating trait-based representations improves performance in the absence of direct supervision (*No Labels*), (ii) the benefit of traits diminishes as supervision becomes available (*Pseudo Labels* and *Gold Labels*), and (iii) increasing supervision consistently improves performance, with *Gold Labels* providing the strongest gains and *Pseudo Labels* remaining competitive.

Traits improve performance in the No Labels setting: We observe that in the *No Labels* setting, incorporating traits improves performance compared to the *No Traits* baseline. In particular, using *Output: Trait Values* improves accuracy from 0.399 to 0.436 (+3.7%), F1 from 0.353 to 0.403 (+5.0%, p=0.006), and α from 0.346 to 0.363 (+1.7%). We further observe that predicting trait values as outputs performs better than providing them as inputs, since generating trait-level abstractions encourages the model to learn a structured mapping from text to rubric-relevant features. Incorporating trait observations does not improve performance, since longer textual explanations introduce variability that makes it harder to extract consistent signals. This result indicates that trait values provide useful intermediate supervision, enabling better generalization in the absence of direct rubric-specific labels during training. See Section A.6 and Table 6 in the Appendix for qualitative examples where *Output: Trait Values* predicts the correct rubric label while the baseline *No Traits* does not.

Traits provide limited gains in the Pseudo Labels and Gold Labels settings: In contrast, under the *Pseudo Labels* and *Gold Labels* settings, trait-based configurations do not consistently outperform their corresponding *No Traits* baselines across all metrics. In the *Pseudo Labels* setting, *Input: Trait Values* improves accuracy from 0.412 to 0.446 (+3.4%) and F1 from 0.376 to 0.400 (+2.4%), while achieving a comparable α (0.368 vs. 0.360). In the *Gold Labels* setting, *Output: Trait Values* improves accuracy from 0.470 to 0.495 (+2.5%), but does not improve F1 (0.431 to 0.415) or α (0.412 to 0.369). These results suggest that when rubric-label supervision on target essays is available, the model can better learn rubric-agnostic essay representations under the training rubrics, thereby reducing the benefit of intermediate trait-based representations.

Increasing supervision improves performance, with Gold Labels providing the strongest gains: We observe that increasing supervision from *No Labels* to *Pseudo Labels* to *Gold Labels* leads to consistent improvements in performance in the *No Traits* setting. Moving from *No Labels* to *Pseudo Labels* improves accuracy from 0.399 to 0.412 (+1.3%), F1 from 0.353 to 0.376 (+2.3%, p=0.109), and α from 0.346 to 0.368 (+2.2%). Moving from *No Labels* to *Gold Labels* improves accuracy from 0.399 to 0.470 (+7.1%), F1 from 0.353 to 0.431 (+7.8%, p<0.01), and α from 0.346 to 0.412 (+6.6%). This result indicates that higher-quality supervision provides stronger and more reliable signals for generalization across rubrics, with the *Gold Labels* gain reaching statistical significance (p<0.01) while the *Pseudo Labels* gain, though consistent, remains modest, making *Pseudo Labels* a practical alternative when gold supervision is unavailable.

6.2. Comparison with Prompting-based Baselines

We observe two main trends: (i) fine-tuning substantially outperforms zero-shot prompting of the same underlying open-source model, confirming that the gains are attributable to fine-tuning rather

than model scale alone, and (ii) fine-tuned open-source LLMs are competitive with proprietary LLMs, achieving performance comparable to smaller models and approaching larger ones.

Fine-tuning substantially outperforms zero-shot prompting of the same model: We compare Llama 3.1 8B Instruct under zero-shot prompting against the fine-tuning baseline (No Labels, No Traits). Moving from zero-shot prompting to the fine-tuning baseline most notably improves F1 from 0.222 to 0.353 (+13.1%, p<0.001), with additional gains in accuracy from 0.372 to 0.399 (+2.7%) and α from 0.251 to 0.346 (+9.5%). We observe that accuracy remains similar across both settings, but F1 and α are substantially lower under zero-shot prompting, since without exposure to the true label distribution, the model concentrates predictions on a few frequent proficiency levels, collapsing recall on minority labels and reducing ordinal alignment with human judgments. This result indicates that fine-tuning over seen rubrics provides the calibration needed to use the full label space, and that these gains cannot be achieved through zero-shot in-context reasoning alone.

Fine-tuned open-source models achieve competitive performance with prompting proprietary models: We observe that our best fine-tuned model (Gold Labels) achieves performance comparable to GPT-5-mini. Compared to GPT-5-mini, our model achieves slightly higher F1 from 0.410 to 0.431 (+2.1%) and α from 0.388 to 0.412 (+2.4%), while accuracy changes from 0.492 to 0.470. Compared to GPT-5, our model achieves comparable F1, with values changing from 0.450 to 0.431, but lower accuracy from 0.532 to 0.470 and lower α from 0.476 to 0.412. This result indicates that fine-tuning can approach the performance of lightweight proprietary models and remain competitive with larger ones. This result is particularly important in practice, since fine-tuned models provide advantages in data privacy, deployment control, and reduced inference costs compared to accessing proprietary LLMs through API calls.

7. Conclusion and Future Work

In this work, we introduced *cross-rubric generalization* as a new setting for automated essay scoring, where models must generalize to previously unseen rubrics. Within an LLM fine-tuning framework for rubric-based essay scoring, we studied rubric-agnostic trait annotations and target-essay supervision during training. Our results show that traits are especially useful in the hardest setting, when no target-essay supervision is available, improving macro F1 by 5.0%, while increasing target-essay supervision yields further gains. Our best fine-tuned open-source Llama-based model outperforms the proprietary prompting-only GPT-5-mini by 2.1% macro F1 and trails GPT-5 by only 1.9%, suggesting that fine-tuned open-source models can provide a scalable, cost-effective, and privacy-preserving alternative for rubric-based assessment as scoring criteria evolve.

Future work can be done in the following directions. First, our findings are based on a single critical thinking dataset with six rubrics, and future work could examine the extent to which these results generalize to other datasets, rubric designs, and writing domains. Second, gains from pseudo labels suggest that stronger low-supervision learning, such as improved self-training, may further reduce noise in pseudo-labeled target essays. Third, our findings motivate learning more robust rubric-agnostic essay representations through contrastive learning or prototypical classification approaches that better capture shared argumentative structure across rubrics. Fourth, trait predictions can not only offer intermediate supervision, but also reward signals for training objectives that directly optimize cross-rubric generalization, similar to recent reinforcement-learning-based work on multi-trait essay scoring [20].

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-5.4) for paraphrasing and rewording text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Burstein, M. Chodorow, Automated essay scoring for nonnative english speakers, in: Computer mediated language assessment and evaluation in natural language processing, 1999.

[2] D. Alikaniotis, H. Yannakoudakis, M. Rei, Automatic text scoring using neural networks, in: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), 2016, pp. 715–725.

[3] F. Dong, Y. Zhang, J. Yang, Attention-based recurrent convolutional neural network for automatic essay scoring, in: Proceedings of the 21st conference on computational natural language learning (CoNLL 2017), 2017, pp. 153–162.

[4] R. Ridley, L. He, X.-y. Dai, S. Huang, J. Chen, Automated cross-prompt scoring of essay traits, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 13745–13753. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17620. doi:10.1609/aaai.v35i15.17620.

[5] H. Do, T. Park, S. Ryu, G. Lee, Towards prompt generalization: Grammar-aware cross-prompt automated essay scoring, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 2818–2824. URL: https://aclanthology.org/2025.findings-naacl.153/. doi:10.18653/v1/2025.findings-naacl.153.

[6] S. Eltanbouly, S. Albatarni, T. Elsayed, TRATES: Trait-specific rubric-assisted cross-prompt essay scoring, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 20528–20543. URL: https://aclanthology.org/2025.findings-acl.1054/. doi:10.18653/v1/2025.findings-acl.1054.

[7] M. C. Peczuh, N. A. Kumar, R. Baker, B. Lehman, D. Eisenberg, C. Mills, P. Wittawatolarn, K. Naskar, K. Chebrolu, S. Nashi, et al., Toward llm-supported automated assessment of critical thinking subskills, arXiv preprint arXiv:2510.12915 (2025).

[8] C. Xiao, W. Ma, Q. Song, S. X. Xu, K. Zhang, Y. Wang, Q. Fu, Human-ai collaborative essay scoring: A dual-process framework with llms, in: Proceedings of the 15th international learning analytics and knowledge conference, 2025, pp. 293–305.

[9] S. Lee, Y. Cai, D. Meng, Z. Wang, Y. Wu, Unleashing large language models' proficiency in zero-shot essay scoring, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 181–198. URL: https://aclanthology.org/2024.findings-emnlp.10/. doi:10.18653/v1/2024.findings-emnlp.10.

[10] S. Chu, J. W. Kim, B. Wong, M. Y. Yi, Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 5811–5829. URL: https://aclanthology.org/2025.findings-naacl.322/. doi:10.18653/v1/2025.findings-naacl.322.

[11] Y. Wang, Z. Ding, X. Wu, S. Sun, N. Liu, X. Zhai, Autoscore: Enhancing automated scoring with multi-agent large language models via structured component recognition, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 40898–40906. URL: https://ojs.aaai.org/index.php/AAAI/article/view/42123. doi:10.1609/aaai.v40i48.42123.

[12] K. Al Khatib, M. Völske, S. Syed, N. Kolyada, B. Stein, Exploiting personal characteristics of debaters for predicting persuasiveness, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of

the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7067–7072. URL: https://aclanthology.org/2020.acl-main.632/. doi:10.18653/v1/2020.acl-main.632.
[13] X. Lan, J. Feng, Y. Liu, Xinleishi, Y. Li, AutoQual: An LLM agent for automated discovery of interpretable features for review quality assessment, in: S. Potdar, L. Rojas-Barahona, S. Montella (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Suzhou (China), 2025, pp. 1250–1264. URL: https://aclanthology.org/2025.emnlp-industry.87/. doi:10.18653/v1/2025.emnlp-industry.87.
[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[16] S. E. Toulmin, The uses of argument, Cambridge university press, 2003.
[17] D. Kuhn, The skills of argument, Cambridge University Press, 1991.
[18] P. Facione, Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (the delphi report) (1990).
[19] B. Efron, R. J. Tibshirani, An introduction to the bootstrap, Chapman and Hall/CRC, 1994.
[20] H. Do, S. Ryu, G. Lee, Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 16427–16438. URL: https://aclanthology.org/2024.emnlp-main.917/. doi:10.18653/v1/2024.emnlp-main.917.
[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., Iclr 1 (2022) 3.
[22] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[23] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.

A. Appendix

A.1. Rubric Details

Table 3 summarizes the rubric details in the critical thinking dataset introduced by [7]. Following our terminology in this paper, we refer to these dataset-defined subskills as rubrics.

A.2. Trait Schema

Table 4 shows the trait schema for the rubric-agnostic analytic traits used in our framework, including the annotation question and value scale associated with each trait. These traits capture lower-level argumentative properties that may generalize across rubrics.

A.3. Prompts

Figure 1 shows the prompt used to generate trait annotations for each essay. The `{TRAIT_SCHEMA}`, including annotation questions and value scales, is provided in Table 4. The placeholder `{ESSAY_TEXT}` represents the essay being annotated.

Figure 2 shows the prompt used to generate justifications for gold rubric labels. The placeholders `{PROFICIENCY_LEVEL_DEFINITIONS}`, `{subskill}`, `{definition}`, and `{RUBRIC_DEFINITION}` correspond to the general proficiency-level definitions, subskill (rubric)

Rubric (Subskill) Name	Rubric (Subskill) Definition	Illustrative Example
2: Information Analysis		
2.1 Synthesizing Multiple Sources	Effectively synthesizes multiple pieces of information	Summarizes information from multiple sources (with citations) but does not integrate information
2.2 Evaluating Evidence Strength	Evaluates the strength and relevance of evidence used to form a conclusion	Presents evidence and links evidence to a specific conclusion, but does not evaluate the relevance or strength of the evidence for the argument generated
3: Argument Generation		
3.1 Using Counterarguments	Effectively addresses counterarguments	Acknowledges specific opposing viewpoint(s), counterargument(s), or qualifier(s)
3.2 Using Facts and Opinions	Relies on data and/or facts over opinions	Uses facts and opinions about equally to support claim(s) and/or arguments
4: Logical Reasoning		
4.1 Drawing Conclusions	Draws specific conclusions	Draws a specific conclusion to analyze simple and straightforward relationships/argumentations
4.2 Using Logical Fallacies	Recognizes and avoids logical fallacies	Uses logical fallacies and evidence about equally when generating arguments

Table 3
Subskill details in the critical thinking dataset, adapted from [7]. We retain the original table content from their paper while changing the column headers to match the terminology used in this paper.

names and definitions, and the corresponding rubric descriptions from [7]. The placeholders `{essay}` and `{gt_label}` denote the essay being evaluated and its gold label for the given subskill, respectively.

Figure 3 shows the prompt used for prompting-based baseline for rubric label generation. The placeholders `{PROFICIENCY_LEVEL_DEFINITIONS}`, `{subskill}`, `{definition}`, and `{RUBRIC_DESCRIPTION}` correspond to the general proficiency-level definitions, subskill (rubric) names and definitions, and the corresponding rubric descriptions from [7]. The placeholder `{essay}` denotes the essay to be rated.

A.4. Implementation Details

Model and Training. We fine-tune an open-source LLM based on Llama 3.1 8B Instruct [14] using parameter-efficient fine-tuning (PEFT) with LoRA adapters [21]. We use a rank of 32, scaling factor $\alpha = 64$, and dropout of 0.05, updating only the adapter weights. Training follows the input–output formatting detailed in Section 4, where the model takes an essay, rubric specification, and optionally trait annotations as input, and generates outputs consistent with the selected configuration (e.g., rubric labels, optionally with justifications and/or trait annotations).

We optimize using 8-bit paged AdamW optimizer [22, 23] with a learning rate of 2×10^{-4} for 3 epochs, with a batch size of 4 and gradient accumulation of 4. We use a maximum sequence length of 1536 tokens for settings without traits and 2048 tokens for settings with traits. We select the checkpoint with the lowest validation loss on a held-out validation set across training epochs.

At test time, we generate outputs for unseen essay–rubric pairs using the same input–output formatting as during training. Generation is performed using low-temperature sampling (temperature = 0.1) with top-k filtering ($k = 50$), allowing up to 512 new tokens. This setting produces stable, near-deterministic outputs while retaining limited stochasticity.

All experiments are conducted on a single NVIDIA L40S GPU (48GB VRAM), with end-to-end training and inference taking approximately 6–8 hours.

Prompt for Trait Annotation Generation

System Prompt
You are an expert writing evaluator analyzing the argumentative quality of an essay.

Your task is to assign analytic trait labels that describe the essay's reasoning and argument structure. For each trait below, choose exactly one label from the allowed list.

For each trait, first write a brief observation from the essay (a direct quote or paraphrase that supports your judgment), then assign the label. Return ONLY a valid JSON object (parseable by `json.loads`). Do not include any text outside the JSON object. If a trait is genuinely ambiguous, choose the label that best matches the essay's dominant tendency.

Traits
{TRAIT_SCHEMA} is inserted here. Refer to Table 4 for the trait schema, which contains the trait names, annotation questions, and value scales.

Output format (exact keys required):

```
{
"claim_explicitness": {"observation": "...", "label": "..."},
... similar structure for all other traits
}
```

User Prompt
Essay:

```
«<
{ESSAY_TEXT}
»>
```

Figure 1: Prompt used to generate trait annotations for each essay. The {TRAIT_SCHEMA} is provided in Table 4. The placeholder {ESSAY_TEXT} represents the essay being annotated.

Trait Representation Generation. For trait representation generation, we use GPT-4.1 with temperature 0.3, top-p 1.0, and a maximum output length of 1500 tokens (see Figure 1 in the Appendix). To improve annotation reliability, we generate three outputs per essay, select the categorical trait value by majority voting, and choose the corresponding observation as the longest observation among outputs assigned to the majority value.

Justification Generation and Prompting-based Baselines. Following [7], we use a similar prompt formulation for justification generation and prompting-based rubric label generation, with temperature 1.0 and top-p 0.95 in both cases. For justification generation, we use GPT-4.1 with a maximum output length of 500 tokens (see Figure 2 in the Appendix). For the prompting-based baselines, we use GPT-5 with default reasoning effort and a maximum output length of 3000 tokens to support reasoning and justification generation (see Figure 3 in the Appendix).

A.5. Role of Description and Justification

Following [7] we consider four variants of the LLM-based framework for rubric-based essay scoring: *With Description* vs. *Without Description*, and *With Justification* vs. *Without Justification*. In the *With Description* setting, the input includes the rubric description (mapping to proficiency levels), while in the *Without Description* setting it is omitted and only the rubric name and definition are provided. In the *With Justification* setting, the model is trained to generate both the label and a justification, whereas in the *Without Justification* setting it is trained to generate only the label. These variants allow us to isolate the role of input-side rubric grounding and output-side reasoning in generalization across unseen rubrics.

Prompt for Justification Generation

System Prompt
You are a calibrated educational scorer designed to provide evidence-based justifications for a gold proficiency label on a student essay. Given an essay, a critical thinking subskill rubric, and a gold proficiency level (0–4), your task is to generate a concise justification that explains why the essay merits that specific level according to the rubric criteria. Maintain objectivity, ensure your justification strictly aligns with the rubric definitions, and base your explanation only on information presented in the essay. Do not make assumptions, incorporate external knowledge, or question the assigned level—your role is solely to justify the given classification with evidence from the text.

User Prompt
Your job is to evaluate why a gold label was provided for a specific essay and subskill.

{PROFICIENCY_LEVEL_DEFINITIONS} is inserted here. Refer to [7] for the general definitions of the proficiency levels used in the dataset (*Not Applicable*, *Below Emerging*, *Emerging*, *Expanding*, and *Exemplifying*).

You will be given a subskill and its definition, followed by the rubric for that subskill and the gold label for the essay. Your task is to explain why this gold label was assigned based on the rubric and provide a 1–2 sentence justification for that choice.

Output format:
- First line: `Label - <number>: <label name>`
- Second line: `Justification - <your justification>`

Input

Subskill Information
`Subskill - {subskill}`
`Definition - {definition}`

{RUBRIC_DESCRIPTION} is inserted here. Refer to [7] for the rubric descriptions associated with the given subskill.

Rated Essay
`{essay}`

Gold Label for Particular Subskill
`{gt_label}`

Output of Justification
Strictly adhere to the prescribed output format above. Do not enter any additional words, commentary, or preambles; only the two required lines.

Figure 2: Prompt used to generate justifications for gold rubric labels. The general proficiency-level definitions, subskill names and definitions, and the corresponding rubric descriptions are taken from [7].

Table 5 shows results for different configurations of rubric descriptions and justifications. We observe that the setting with "With Description" and "With Justification" achieves the highest performance across all metrics. Compared to the setting with "Without Description" and "Without Justification", this configuration improves accuracy from 0.226 to 0.399 (+17.3%), F1 from 0.158 to 0.353 (+19.5%), and α from 0.019 to 0.346 (+32.7%). Compared to the setting with "With Description" and "Without Justification", this configuration improves accuracy from 0.362 to 0.399 (+3.7%), F1 from 0.279 to 0.353 (+7.4%), and α from 0.318 to 0.346 (+2.8%). This result suggests that descriptions provide grounding of evaluation criteria, while justifications introduce additional reasoning signals that improve alignment. Hence, we use the "With Description" and "With Justification" setting for all experiments exploring our proposed design space.

Prompt for Prompting-based Rubric Label Generation

System Prompt
You are a calibrated educational scorer designed to classify student essays on fine-grained critical thinking subskills using a structured rubric. For each essay, assign one of five proficiency levels (0–4) and provide a concise, evidence-based justification strictly aligned with the rubric. Maintain objectivity, ensure consistency, and base your evaluation only on information presented in the essay. Do not make assumptions or incorporate external knowledge.

User Prompt
Your job is to evaluate whether an essay demonstrates evidence of a particular subskill, and assign it one of five categories.

`{PROFICIENCY_LEVEL_DEFINITIONS}` is inserted here. Refer to [7] for the general definitions of the proficiency levels used in the dataset (*Not Applicable*, *Below Emerging*, *Emerging*, *Expanding*, and *Exemplifying*).

You will be given a subskill and its definition, followed by the rubric corresponding to the above five categories for that subskill. Your task is to rate the given essay against the rubric and provide an explanation for the same.

Output format:
- First line: `Label - <number>: <label name>`
- Second line: `Justification - <your justification>`

Input

Subskill Information
`Subskill - {subskill}`
`Definition - {definition}`

`{RUBRIC_DESCRIPTION}` is inserted here. Refer to [7] for the rubric descriptions corresponding to the above five categories for the given subskill.

Essay to be rated
`{essay}`

Output of Essay to be rated
Strictly adhere to the prescribed output format above. Do not enter any additional words, commentary, or preambles; only the two required lines.

Figure 3: Prompt used for prompting-based baseline for rubric label generation using GPT-5-mini and GPT-5. The general proficiency-level definitions, subskill names and definitions, and the corresponding rubric descriptions are taken from [7]. The placeholder `{essay}` represents the essay to be rated.

A.6. Qualitative Case Studies Where Trait Outputs Help

Table 6 shows three examples from the *No Labels* setting where trait-output-based fine-tuning (*Output: Trait Values*) predicts the correct rubric label while the corresponding *No Traits* baseline does not. These examples help explain why trait outputs are useful in the hardest setting: they provide intermediate signals that guide the model toward rubric decisions that are better aligned with the essay.

Example 1: Rubric 2.1 (Synthesizing Multiple Sources). Rubric 2.1 evaluates whether the essay *effectively synthesizes multiple pieces of information*. In this example, for essay ID `20D57DEC3EF5`, the trait-output model correctly predicts *Not Applicable*, while the baseline overpredicts *Expanding*. At a high level, the trait-output justification recognizes that the essay does not provide the source-based support needed to evaluate synthesis, whereas the baseline incorrectly interprets the essay as integrating multiple sources. The predicted traits help explain this correction: *limited* **Evidence Presence** suggests that the essay lacks sufficient supporting material for cross-source synthesis, and *descriptive* **Analytical Depth** indicates that the essay does not reason over multiple pieces of information in the way this

Table 4
Trait schema for the rubric-agnostic analytic traits used in our framework, including the annotation question and value scale associated with each trait. These traits capture lower-level argumentative properties that may generalize across rubrics.

Trait	Annotation Question	Values and Definitions
Claim Explicitness	Does the essay clearly state a central, arguable position?	*Absent*: no identifiable central claim *Vague*: position is implied or unclear *Explicit*: clear, arguable central claim
Structural Coherence	Is the reasoning logically organized and internally consistent?	*Low*: disorganized or contradictory reasoning *Medium*: mostly logical with some gaps or jumps *High*: clearly organized and internally consistent
Evidence Presence	Does the essay provide concrete support for its claims?	*None*: no supporting examples, facts, or data *Limited*: some support but sparse or generic *Substantial*: consistent and substantive support
Evidence–Claim Linkage	Does the essay explain why its evidence supports the claim?	*None*: no explanation linking support to claim *Partial*: weak or implicit explanation *Clear*: explicitly explains why the support justifies the claim
Alternative Perspectives	Does the essay consider viewpoints other than its own?	*None*: no alternative perspectives acknowledged *Mentioned*: alternatives noted but not developed *Engaged*: alternatives discussed and responded to
Analytical Depth	Does the essay go beyond description to analyze or evaluate ideas?	*Descriptive*: mostly summary or assertion *Moderate*: some reasoning or evaluation *Strong*: sustained and deeper analysis

Table 5
Comparing different configurations of rubric and justification. Here 'Desc' refers to Description and 'Just' refers to Justification.

Description	Justification	Acc	F1	α
Without Desc	Without Just	0.226	0.158	0.019
	With Just	0.284	0.220	0.007
With Desc	Without Just	0.362	0.279	0.318
	With Just	**0.399**	**0.353**	**0.346**

rubric requires. Together, these traits make the lower label more appropriate.

Example 2: Rubric 3.1 (Using Counterarguments). Rubric 3.1 evaluates whether the essay *effectively addresses counterarguments*. In this example, for essay ID F0AA6B265B6F, the trait-output model correctly predicts *Exemplifying*, whereas the baseline predicts the lower label *Expanding*. Both outputs recognize that the essay acknowledges opposing views, but the trait-output justification better captures that the essay not only mentions counterarguments but also responds to them in a developed way. This stronger prediction is supported by two key traits: *engaged* **Alternative Perspectives** shows that opposing viewpoints are actively discussed and addressed rather than merely noted, while *strong* **Analytical Depth** indicates sustained reasoning about those viewpoints. These intermediate signals are well aligned with the highest label for this rubric.

Example 3: Rubric 4.1 (Drawing Conclusions). Rubric 4.1 evaluates whether the essay *draws specific conclusions*. In this example, for essay ID 845DB3AC386E, the trait-output model correctly predicts *Emerging*, while the baseline overpredicts *Exemplifying*. Both justifications identify that the essay reaches a conclusion, but the trait-output model assigns a more moderate label that better matches the gold score. The relevant trait signals are *medium* **Structural Coherence** and *partial*

Table 6
Table showing case studies from the *No Labels* setting where trait-output-based fine-tuning (*Output: Trait Values*) predicts the correct rubric label while the corresponding *No Traits* baseline does not. We include both shortened original justifications and predicted trait values to show how the predicted trait values influence the final rubric label prediction.

Rubric	Gold	Output: Trait Values	No Traits Baseline	Predicted Trait Values
2.1 Synthesizing Multiple Sources	Not Applicable	**Not Applicable**. "The essay does not cite any sources, making it impossible to evaluate the student's ability to synthesize multiple pieces of information."	**Expanding**. "The essay cites multiple sources of information (e.g., the Mona Lisa, various forms of artistic expression) and integrates these examples to support the argument …"	Claim Explicitness: explicit Structural Coherence: medium Evidence Presence: limited Evidence–Claim Linkage: partial Alternative Perspectives: mentioned Analytical Depth: descriptive
3.1 Using Counterarguments	Exemplifying	**Exemplifying**. "The essay acknowledges specific opposing viewpoints …and develops responses to them by refuting the claim that online schooling eliminates bullying....demonstrating evaluation and integration of multiple perspectives …"	**Expanding**. "The essay acknowledges specific opposing viewpoints …and develops responses to them by pointing out potential negative consequences like increased cyberbullying and decreased access to resources …"	Claim Explicitness: explicit Structural Coherence: high Evidence Presence: substantial Evidence–Claim Linkage: clear Alternative Perspectives: engaged Analytical Depth: strong
4.1 Drawing Conclusions	Emerging	**Emerging**. "The essay draws a specific conclusion that the Electoral College should be kept …but it does not use valid assumptions or analyze complex relationships between the Electoral College and its effects."	**Exemplifying**."The essay draws a specific conclusion that the Electoral College should be kept …and supports this conclusion with evidence from sources, demonstrating the use of valid assumptions to analyze complex relationships …"	Claim Explicitness: explicit Structural Coherence: medium Evidence Presence: limited Evidence–Claim Linkage: partial Alternative Perspectives: none Analytical Depth: descriptive

Evidence–Claim Linkage. These signals suggest that, although the essay does arrive at a conclusion, the reasoning is not fully developed and the support is not fully connected back to the claim. As a result, the essay is better characterized as showing a conclusion in an emerging form rather than as demonstrating the stronger reasoning expected for a top label.

Overall, these examples suggest that trait outputs help by making rubric-relevant argumentative properties explicit. In Examples 1 and 3, they help prevent overly strong predictions by revealing missing support or incomplete reasoning. In Example 2, they help recover the highest label by highlighting deep engagement with counterarguments. This qualitative pattern is consistent with the broader quantitative finding that trait-output-based fine-tuning is especially helpful in the *No Labels* setting.

A.7. Limitations

Some limitations of our work include the use of automatically generated trait annotations and the scope of our empirical evaluation. First, the trait annotations used in our framework are generated using LLM

prompting and are not validated through human annotation, unlike the gold rubric labels in the dataset. We therefore treat these trait annotations as silver intermediate structure that supports model training, rather than as fully reliable standalone supervision targets. Although our results suggest the value of such trait-based structure for improving cross-rubric generalization, especially in low-supervision settings, future work could examine annotation quality more directly through human evaluation or stronger validation procedures.

Second, our empirical study focuses on one recent benchmark for critical thinking essay scoring and on a Llama-based open-source model within our experimental setup. This setting provides a focused testbed for studying cross-rubric generalization. Future work may help assess the extent to which the reported findings generalize to other datasets, writing domains, rubric designs, and model families.

A.8. Ethical considerations

Our work studies automated scoring of student essays, which requires careful use in educational settings. First, the trait annotations in our framework are automatically generated using LLM prompting rather than validated through human annotation. We use them as intermediate structure to support training, but they may still reflect model errors or biases.

Second, although automated rubric-based scoring may be useful for scalable assessment support, it should not be treated as a replacement for human judgment in high-stakes educational decisions. In addition, parts of our pipeline rely on proprietary APIs, so any real-world use would require appropriate safeguards for student data.

Finally, our experiments require GPU compute for fine-tuning and inference, which may have some environmental impact.