EvoPARA: Error-Driven Problem-Aligned Retrieval for Automated Essay Feedback Generation

Anonymous ACL submission

Abstract

Automated Essay Feedback Generation (AEFG) requires expert-level, rubric-aligned commentary rather than holistic scoring. General-purpose large language models (LLMs) frequently fall short of this standard because they lack the experienced raters-like analytical frameworks, and their diagnostic inventory remains sparse, often yielding feedback that is generic, vaguely defined, and poorly actionable. To bridge this gap, we propose **EvoPARA**, a training-free framework with two complementary modules: one teaches the LLM the expert-level analytical frameworks and broaden the diagnostic inventory, the other offers more detailed context and definition of specific writing problems to achieve precise commentary. Since such information is diverse and long-tailed, and manual design efficacy is inherently constrained by human experience, we design both modules to operate automatically. **EDPO** (Error-Driven Prompt Optimization) mines the worst-performing validation cases and distills them into pattern-level critiques to automatically summarize and evolve the most effective analytical framework and originally overlooked and long-tailed writing problems. **PARA** (Problem-Aligned Retrieval Agent) first diagnoses the essay's latent problems, and then retrieves exemplars whose expert feedback targets similar problems and content, which give more context to verify each problem. On **PRISM**, a new benchmark of **653 Chinese middle-school essays with 7,781 expert annotations**, EvoPARA reaches **89.9%** recall. It outperforms the strongest content-similarity RAG baseline by 9.0 absolute recall points on DeepSeek-V3.2, with consistent gains across four Chinese-capable LLMs.

1 Introduction

Education at scale needs consistent, rubric-aligned commentary on student essays. Classical Automated Essay Scoring (AES) predicts a holistic score (Taghipour and Ng, 2016; Dong et al., 2017). However, a single number tells students little about how to revise. The more useful task is **Automated Essay Feedback Generation (AEFG)**: producing dimension-specific, actionable comments (Ke and Ng, 2019). This paper studies AEFG in its most reviewer-relevant form. Given an essay and its writing prompt, the system outputs structured feedback along three rubric dimensions: (i) content and topic adherence, (ii) structural coherence, and (iii) language expression.

Producing expert-level feedback is hard for a general-purpose large language model (LLM). The reason is not model size. The model lacks two things an experienced rater applies to a student essay: a structured analytical framework and a broad diagnostic inventory. First, an analytical framework names the key factors a qualified essay must satisfy; without it, the model cannot evaluate comprehensively, and different experts use different frameworks the model has never seen. Second, the model's inventory of writing problems is sparse and biased toward common defects. Long-tailed problems—broken transitions, register mismatch, weak evidence—rarely appear in its output, and when they do, the phrasing is generic and vaguely defined. Neither defect can be fixed by a larger model or more labels. AEFG therefore needs a training-free way to align the LLM with expert analysis.

A natural remedy is retrieval-augmented generation (RAG): ground the LLM in a corpus of expert-annotated essays (Lewis et al., 2020). Existing RAG-AEFG pipelines retrieve content-similar essays via hybrid BM25-plus-dense scoring (Ma et al., 2021) and feed their expert annotations as in-context exemplars (Xiao et al., 2024). This design fixes neither defect. When the retrieved exemplar is a high-scoring essay on the same topic,

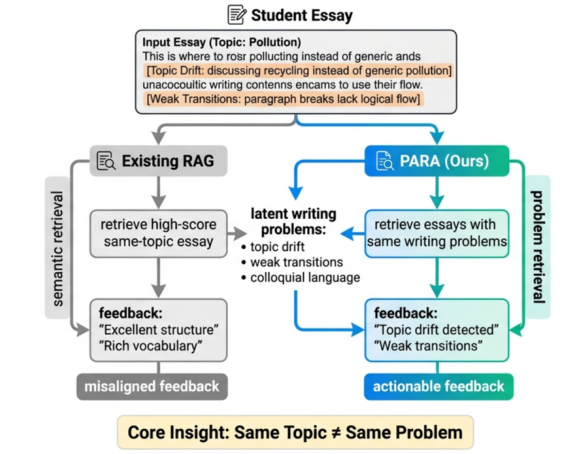

Figure 1: Problem-blind retrieval versus problem-aligned retrieval.

its expert annotation praises coherence and word choice. The model is shown encouragement, not a corrective signal. Worse, the generation prompt is hand-written and frozen. New problems discovered at inference cannot flow back into the system.

We propose **EvoPARA**[1], a training-free framework that closes both defects together (Figure 1). One module teaches the LLM to *analyse* essays the way an expert does. The other *grounds* that analysis on the target essay. **Error-Driven Prompt Optimization (EDPO)** addresses the sparse problem inventory. EDPO runs the current prompt on a validation set and selects the worst-performing essays. A second LLM call compares each output against expert annotations. Their divergences are distilled into pattern-level critiques. These critiques cover long-tailed problems the model would otherwise miss. A revision step writes them back into the prompt. A monotone-update rule keeps a revision only when validation recall improves.

Problem-Aligned Retrieval Agent (PARA) addresses the generic phrasing. PARA first asks the LLM to diagnose the target essay's latent problems. It then retrieves exemplars whose expert feedback targets those same problems, not just similar topics. Before fusion, an agent re-checks each retrieved problem's definition on the target essay. The generator therefore sees corrective rather than congratulatory exemplars. The two modules are complementary by design: EDPO broadens the LLM's analytic vocabulary, and PARA grounds that vocabulary on the essay at hand. Together they convert a general-purpose LLM into an expert-aligned feedback generator without any new labels.

We curate **PRISM**, a benchmark of **653 Chinese middle-school essays with 7,781 expert annotations**, structurally labelled by experienced Chinese-language teachers across three rubric dimensions. On this benchmark, EvoPARA reaches **89.9%** recall. It outperforms the strongest content-similarity RAG baseline by 9.0 absolute recall points on DeepSeek-V3.2, with consistent gains across four Chinese-capable LLM backbones.

Contributions.

- **Task and benchmark.** We formalise AEFG as a training-free LLM–expert alignment problem and release **PRISM**, the first Chinese middle-school AEFG benchmark: 7,781 rubric-aligned expert annotations on 653 essays.
- **EDPO.** A label-free, error-driven prompt optimiser that evolves an expert-level analytical framework and distils long-tailed writing problems into the generation prompt under a monotone-update rule, ensuring non-decreasing validation recall.
- **PARA.** A problem-aligned retrieval agent that re-keys retrieval on LLM-diagnosed problems and verifies each problem on the target essay, supplying corrective rather than congratulatory exemplars and improving over content-similar RAG across four backbones.

2 Related Work

2.1 LLM-based Essay Feedback Generation

The field has shifted from holistic Automated Essay Scoring (AES) (Taghipour and Ng, 2016; Dong et al., 2017) to dimension-specific feedback generation (AEFG) (Ke and Ng, 2019; Jong et al., 2023). LLM-based AEFG is now the dominant line. Zero-shot or chain-of-thought prompting (Mizumoto and Eguchi, 2023; Xiao et al., 2024) delivers fluent comments but tends to be generic and ungrounded in any rubric (Kasneci et al., 2023). Supervised pipelines (Dai et al., 2023; Stahl et al., 2024) narrow this gap but demand expensive expert annotations.

[1]Anonymous code, prompts, and the PII-scan script: https://anonymous.4open.science/r/code-1E44/README.md . We used Claude (Anthropic) for paragraph-level language polishing only; all technical content is the authors' own.

Chinese essay evaluation is further constrained by data. Dominant benchmarks (ASAP (Shermis, 2014), TOEFL11 (Blanchard et al., 2013)) target English L2 writing. Available Chinese resources cover only holistic HSK scoring (Beijing Language and Culture University, 2017) or single-dimension coherence evaluation (Wu et al., 2023). Fine-grained, rubric-aligned feedback for Chinese middle-school compositions thus remains unaddressed—exactly the gap our benchmark fills.

2.2 Retrieval and Prompt Optimisation for LLM Grounding

Retrieval-augmented generation (RAG) grounds LLM output on non-parametric evidence retrieved by BM25 (Robertson and Zaragoza, 2009), dense encoders (Karpukhin et al., 2020; Reimers and Gurevych, 2019), or hybrid scoring (Lewis et al., 2020; Gao et al., 2023b). RAG has been applied across educational settings (Fateen et al., 2024). Prior RAG-AEFG systems retrieve content-similar essays and feed their expert annotations as few-shot exemplars (Xiao et al., 2024). The implicit assumption is that content similarity entails feedback similarity. This assumption breaks whenever the nearest neighbour is problem-free; the model is then shown praise rather than corrective signal. A complementary line re-keys retrieval on a derived representation. HyDE (Gao et al., 2023a) uses a hypothetical answer as the key. Query2Doc (Wang et al., 2023) expands the query with synthesised passages. Step-Back Prompting (Zheng et al., 2024) abstracts to higher-level concepts. Intent-driven session retrieval (Sun et al., 2024) conditions on latent user intent. None of these methods has been adapted to feedback generation, where the relevant derived artefact is neither a hypothetical answer nor a reasoning step but the set of *latent writing problems* in the target essay. PARA fills this gap and additionally verifies each retrieved problem on the target essay before fusion—a step absent from all four query-rewriting methods.

Automatic prompt optimisation scales prompt design beyond hand-written templates (Liu et al., 2023). APE (Zhou et al., 2023) treats prompt search as black-box LLM-driven optimisation. ProTeGi (Pryzant et al., 2023) uses natural-language "textual gradients" derived from failure cases. OPRO (Yang et al., 2024) casts the LLM itself as the optimiser. EvoPrompt (Guo et al., 2024) and PromptBreeder (Fernando et al., 2023) apply evolutionary search over prompt populations. A parallel inference-time line—Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), and Chain-of-Verification (Dhuliawala et al., 2024)—refines outputs without modifying the system prompt. EDPO differs in three ways. (i) Its update signal is the task-grounded recall metric, not a generic fitness or opaque LLM preference. (ii) Failures are mined from bottom-K_e validation cases and distilled into pattern-level critiques applicable across the validation set, rather than per-instance gradients (ProTeGi) or per-trajectory reflections (Reflexion). (iii) A monotone-update rule yields a non-decreasing trajectory and a principled stopping criterion—a guarantee absent from APE, ProTeGi, and OPRO.

3 Method

3.1 Problem Formulation

We formalise AEFG as structured feedback generation rather than holistic-score regression. Each data point is a tuple (T, E, D, F^*). T is the writing prompt, E is the student essay, and $D = \{d_j\}$ is a retrieval corpus. Each $d_j = (T_j, E_j, F_j^*)$ pairs an essay with its expert annotation. The expert annotation $F^* = \{F_1^*, F_2^*, F_3^*\}$ is structured along three rubric dimensions: (i) content and topic adherence, (ii) structural coherence, and (iii) language expression. The system outputs a prediction $\hat{F} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$ in the same schema. Quality can be measured by per-dimension recall and mean recall:

$$R_i = \frac{|\hat{F}_i \cap F_i^*|}{|F_i^*|}, \quad R_{\text{avg}} = \frac{1}{3}\sum_{i=1}^{3} R_i, \qquad (1)$$

where $|\hat{F}_i \cap F_i^*|$ counts comments judged by an LLM judge to semantically match an expert comment in dimension i. Precision and F1 can also be reported as complementary metrics. The matching protocol is given in §4.4.

3.2 Overview

EvoPARA realises the expert-alignment goal through two complementary modules (Figure 2). **EDPO** (§3.3) teaches the LLM to analyse essays like an expert. It distils long-tailed problems from validation failures into the generation prompt under a monotone-update rule. **PARA** (§3.4–§3.5) grounds that analysis on the target essay with

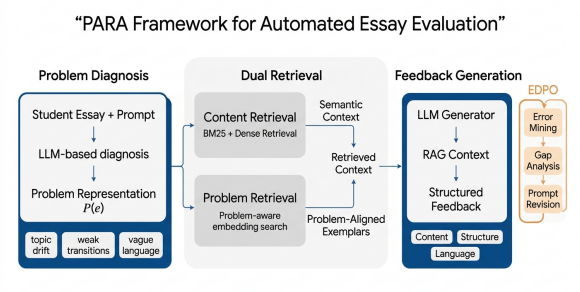

Figure 2: EvoPARA framework. Given (T, E), LLM_{dia} produces a structured problem description $\mathcal{P}(E)$ and an agent verifies each problem on E. Two retrieval channels then run in parallel: hybrid BM25+dense content retrieval and dense problem-aligned retrieval. The merged exemplars feed the generator whose prompt π_θ is iteratively refined by EDPO.

detailed context and problem-specific definitions. It first diagnoses the essay's latent problems, retrieves problem-aligned exemplars, and verifies each retrieved problem before fusion. The two modules act on complementary sides of the expert-alignment gap: EDPO broadens the analytic vocabulary the model can draw on, and PARA sharpens which of those entries land on the current essay. The EDPO optimisation loop is given in Algorithm 1; Algorithm 2 lists the PARA inference loop. Both modules are training-free and require no labels beyond the expert annotations already in the validation set.

3.3 Error-Driven Prompt Optimization

A persistent gap remains between LLM output and expert annotations. The gap stems from two missing pieces an experienced rater applies: a structured analytical framework and a broad diagnostic inventory. The model may comment at the wrong granularity, conflate dimensions, or default to vague phrasing where the expert cites specific sentences. Manual prompt engineering can patch individual failures but cannot scale to the long tail of writing problems observed across hundreds of essays. **EDPO** is a closed-loop refinement procedure that automatically discovers systematic divergences on a held-out validation set $\mathcal{V}$ and writes them back into the prompt. At iteration t, EDPO executes three steps with the current prompt π_t.

Step 1: Error mining. EDPO runs π_t over $\mathcal{V}$, scores each prediction by mean recall, and selects the bottom-K_e essays as the error pool:

$$\mathcal{E}_t = \text{Bot}_{K_e}\Big\{ R_{\text{avg}}\big(\hat{F}_i^{(t)}, F_i^*\big) : E_i \in \mathcal{V} \Big\}. \quad (2)$$

Algorithm 1 Error-Driven Prompt Optimization (EDPO)

Input: initial prompt π_0, validation set $\mathcal{V}$, budget $T_{\max}$, error-pool size K_e
Output: optimised prompt π^*

$R_0 \leftarrow \text{Recall}(\pi_0, \mathcal{V})$
for $t = 0, 1, \ldots, T_{\max} - 1$ **do**
  $\mathcal{E}_t \leftarrow \text{Bot}_{K_e}\{R_{\text{avg}}(\hat{F}_i, F_i^*) : E_i \in \mathcal{V}\}$
  $\Delta_t \leftarrow \text{LLM}_{\text{ana}}(\{(\hat{F}_i, F_i^*)\}_{E_i \in \mathcal{E}_t})$
  $\tilde{\pi}_{t+1} \leftarrow \text{LLM}_{\text{rev}}(\pi_t, \Delta_t)$
  $\tilde{R}_{t+1} \leftarrow \text{Recall}(\tilde{\pi}_{t+1}, \mathcal{V})$
  $\pi_{t+1} \leftarrow \tilde{\pi}_{t+1}$ **if** $\tilde{R}_{t+1} > R_t$ **else** π_t
end for
return $\pi_{T_{\max}}$

Concentrating on the worst cases maximises the signal-to-noise ratio of the next step.

Step 2: Gap analysis. A second LLM call LLM_{ana} compares each $(\hat{F}_i^{(t)}, F_i^*)$ pair in $\mathcal{E}_t$ and aggregates the discrepancies into pattern-level failure descriptions:

$$\Delta_t = \text{LLM}_{\text{ana}}\big(\{(\hat{F}_i^{(t)}, F_i^*)\}_{E_i \in \mathcal{E}_t}\big). \quad (3)$$

The analyser is instructed to abstract from individual instances to recurring patterns, e.g., *"the model produces generic structural comments while experts name specific paragraph locations."* This abstraction is what surfaces long-tailed problems—single-essay reflections would not generalise.

Step 3: Prompt revision and acceptance. A third LLM call LLM_{rev} proposes a candidate prompt $\tilde{\pi}_{t+1} = \text{LLM}_{\text{rev}}(\pi_t, \Delta_t)$. EDPO accepts the candidate only if validation recall improves:

$$\pi_{t+1} = \begin{cases} \tilde{\pi}_{t+1} & \text{if } \tilde{R}_{t+1} > R_t, \\ \pi_t & \text{otherwise.} \end{cases} \quad (4)$$

This monotone-update rule guarantees a non-decreasing recall trajectory and yields a principled stopping criterion without any held-out test signal. Convergence is observed within eight iterations (Figure 4).

3.4 Problem-Aligned Diagnosis

PARA cannot retrieve problem-aligned exemplars without first naming the target essay's problems. A retrieval query keyed on E itself surfaces topically similar essays whose latent problems may differ entirely. We therefore decouple the query from the essay surface and re-key it on a structured problem

Algorithm 2 Problem-Aligned Retrieval Agent (PARA)

Input: prompt T, essay E, corpus D, prompt template π_θ
Output: structured feedback $\hat{F}$

```
function PARA(T, E, D, π_θ)
    P(E) ← LLM_dia(T, E)
    P(E) ← VERIFY(P(E), T, E)
    S_1 ← CONTENTRETRIEVE(E, D)
    S_2 ← PROBLEMRETRIEVE(P(E), D)
    C ← S_1 ∪ S_2
    F̂ ← LLM(π_θ, T, E, C)
    return F̂
end function
```

diagnosis. Given (T, E), a diagnostic LLM call $\mathrm{LLM}_{\text{dia}}$ is configured with a rubric-aligned prompt that asks the model to act as an experienced rater and report the dominant problem along each evaluation dimension:

$$\mathcal{P}(E) = \mathrm{LLM}_{\text{dia}}(T, E) = \{P_1(E), P_2(E), P_3(E)\}, \tag{5}$$

where $P_i(E)$ is a free-text description of the principal problem along dimension i.

The diagnosis is then verified on the target essay. A second pass invokes $\mathrm{LLM}_{\text{dia}}$ as an agent that re-reads E with each candidate problem $P_i(E)$ in context. For each problem the agent either confirms it with a textual anchor (a sentence or paragraph reference) or rejects it as ungrounded. Only confirmed problems enter the next stage. This step removes hallucinated problems that semantic similarity would otherwise propagate, and gives the downstream retrieval a checked, grounded query. The verified diagnosis is encoded by a sentence encoder $\psi(\cdot)$ into $\mathbf{q}_p = \psi(\mathcal{P}(E)) \in R^d$, which serves as the problem-aligned retrieval key. The corpus is indexed once: every d_j is passed through the same diagnose-then-verify pipeline and encoded, yielding $\mathbf{q}_p^{(j)}$ stored alongside $\phi(d_j)$.

3.5 Dual-Stage Retrieval

Neither retrieval channel by itself is sufficient. Pure content-similarity retrieval is blind to the latent problems in E; it tends to surface high-quality, problem-free exemplars whose expert feedback offers no corrective signal. Pure problem-aligned retrieval, conversely, may return exemplars on unrelated prompts, depriving the LLM of thematic context. PARA therefore fires both channels in parallel and merges their results.

Stage 1: Content-Similarity Retrieval. A standard hybrid sparse-dense score (Ma et al., 2021) combines BM25 with dense cosine similarity:

$$\begin{aligned} s_{\text{cont}}(E, d_j) &= \alpha\, \mathrm{BM25}(E, d_j) \\ &\quad + (1-\alpha) \cos\big(\phi(E), \phi(d_j)\big), \end{aligned} \tag{6}$$

where $\phi(\cdot)$ is the dense text encoder and $\alpha \in [0, 1]$ is a mixing coefficient. Stage 1 returns the top-K_1 documents, supplying the writing context expected for prompt T.

Stage 2: Problem-Aligned Retrieval. Stage 2 operates over the verified problem description rather than the raw essay text. BM25 degenerates on this short, closed-vocabulary input and treats lexical paraphrases of the same problem as unrelated. Retrieval therefore uses pure dense cosine similarity, which is robust to surface phrasing:

$$s_{\text{prob}}(E, d_j) = \cos\big(\mathbf{q}_p, \mathbf{q}_p^{(j)}\big). \tag{7}$$

Stage 2 returns the top-K_2 documents. These exemplars may treat unrelated subject matter but exhibit the same latent writing problems as E, exposing the LLM to expert language patterns that directly correct those deficiencies.

Fusion. The two stages are merged by union with deduplication and self-match removal: $\mathcal{C} = S_1 \cup S_2$. The generator then produces structured feedback as $\hat{F} = \mathrm{LLM}(\pi_\theta, T, E, \mathcal{C})$, where π_θ is the prompt template optimised by EDPO (§3.3). The fused signal is strictly stronger than either channel alone: single-stage variants lag dual-stage retrieval on every dimension (§4.6).

4 Experiments

4.1 Dataset

Available Chinese essay-evaluation resources are limited to holistic HSK scoring (Beijing Language and Culture University, 2017) or single-dimension coherence evaluation (Wu et al., 2023). Neither supports rubric-aligned, dimension-specific AEFG research. We curate **PRISM** (Problem Rubric-Indexed Student Manuscripts), a benchmark of **653 authentic Chinese middle-school compositions** drawn from classroom and examination settings. The essays span 564 distinct writing prompts across narrative, expository, and argumentative genres. Each essay is annotated by experienced Chinese-language teachers following the official middle-school writing rubric. Annotators

produce a structured list of comments along three rubric dimensions, with each comment tagged as *advantage*, *disadvantage*, or *revision suggestion*.

The pass produces 7,781 **structured comments** in total (11.92 per essay on average). Comments are balanced across dimensions: 2,873 content, 2,505 structure, 2,398 language. We split the essays 553/50/50 for training, validation, and testing. The training split serves as the retrieval corpus D for all RAG variants. With 564 distinct prompts across 653 essays, prompt overlap between folds is minimal; we additionally verify that no essay–prompt pair recurs across folds.

Ethical considerations. Essays were drawn from partner schools' archival collections with prior written consent from the schools—no new data collection was conducted, and a regex PII scan over IDs, phone numbers, e-mails, and school / class / name patterns returned no student-identifying matches (Appendix A). The benchmark is released under CC BY 4.0 for non-commercial research use; we discourage deployment as an autonomous grading tool without human-in-the-loop oversight, as automated feedback may amplify rubric biases.

4.2 Baselines

We compare against three baselines that span LLM-based AEFG. **Zero-shot LLM** is the backbone prompted with task description only; it isolates the contribution of any corpus grounding. **Few-shot LLM** adds three manually selected exemplar essays as in-context demonstrations. **Content-RAG** (Xiao et al., 2024) retrieves content-similar essays via hybrid BM25+dense scoring and feeds their expert annotations as exemplars; it is the closest direct competitor and instantiates the problem-blind retrieval paradigm critiqued in §3.1.

4.3 Implementation Details

Backbone. We instantiate LLM_{dia}, LLM_{ana}, LLM_{rev}, and the generation LLM from a single backbone. DeepSeek-V3.2 is the default for ablations. Cross-backbone results span three Chinese-capable LLMs (Qwen-Plus, Kimi-K2, GLM-4) and three open backbones (Qwen2.5-72B, LLaMA-3.1-70B, GPT-4o).

Retrieval. Content and problem indexes are built once on the training split. Sentence-Transformers' `all-MiniLM-L6-v2` (384-d) is shared between $\phi(\cdot)$ and $\psi(\cdot)$. Hybrid Stage-1 uses $\alpha = 0.5$; we set $K_1 = K_2 = 30$. Verification (§3.4) reuses the diagnostic prompt with an added grounding instruction.

Generation & EDPO. Diagnosis runs at temperature 0.2 (800 tokens); feedback at 0.7 (8,000 tokens). EDPO uses $K_e = 10$ and a budget of ten iterations.

Inference runs. All numbers in Tables 1 and 2 are single-seed; multi-seed averaging was infeasible under API-cost constraints (Appendix B).

4.4 Evaluation Protocol

Following §3.1, the primary metric is per-dimension recall R_i and mean recall R_{avg}. For each generated comment $\hat{f} \in \hat{F}_i$, DeepSeek-V3.2 serves as an LLM judge. The judge decides whether $\hat{f}$ matches any expert annotation in the same dimension. A match requires both semantic agreement and type agreement (advantage/disadvantage/suggestion). Partial matches do not count, and a generated comment may match at most one expert comment.

Human validation of the LLM judge. Two annotators independently re-scored a stratified 60-essay sample; the judge agreed with the human consensus on 93.4% of decisions (Cohen's $\kappa = 0.81$), with disagreement concentrated in the Language dimension.

4.5 Main Results

Table 1 reports per-dimension and overall scores across seven LLM backbones. On DeepSeek-V3.2, PARA reaches 89.9% overall recall. It outperforms the strongest single-stage baseline (Content-RAG) by 9.0 absolute recall points and zero-shot prompting by 14.5 points. The pattern holds across the other six backbones. PARA improves F1 over Content-RAG on six of seven backbones, with average gains of 2.1 F1 points and 2.4 recall points.

Three observations follow. **(i)** Unaided LLMs leave a substantial gap. Zero-shot trails PARA by 14.5 recall points on DeepSeek-V3.2; few-shot trails by 10.9 points on the same backbone. Strong general-purpose LLMs alone cannot reproduce expert-level feedback without a corpus-grounded retrieval signal. **(ii)** Content-similar retrieval is necessary but insufficient. Content-RAG narrows the gap but improves Content-and-Topic only modestly over few-shot. This is exactly the failure mode predicted by §3.4: when the corpus contains a topically similar but problem-free

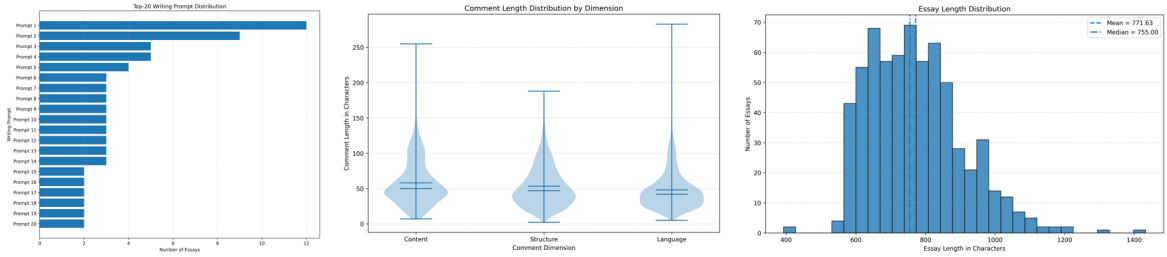

Figure 3: Dataset statistics: (a) essays per writing prompt; (b) comments per essay, stratified by dimension; (c) essay length distribution.

essay, Content-RAG retrieves it and shows the model praise rather than corrective signal. A concrete case appears in §4.7. **(iii)** Gains generalise across backbones. The largest improvement is on Qwen-Plus (+7.3 F1), and gains are also seen on Kimi-K2 (+2.4), GLM-4 (+0.9), Qwen2.5-72B (+1.7), LLaMA-3.1-70B (+1.3), and DeepSeek-V3.2 (+4.3). The one regression is GPT-4o (−1.1 F1). There, Content-RAG already saturates on GPT-4o's strong English-trained prior; both recall (−0.8) and F1 (−1.1) fall within noise.

4.6 Ablation Study

Table 2 reports five variants on the 100-essay (val+test) split with DeepSeek-V3.2 as the shared backbone. The variants selectively disable CRAG (content-similarity retrieval), PRAG (problem-aligned retrieval), or EDPO.

Problem-aligned retrieval is complementary, not a replacement. PRAG-only reaches 81.3% overall recall. It outperforms CRAG-only on Structural Coherence and Language Expression but lags on Content and Topic Adherence. The asymmetry has a clean explanation. Problem-aligned retrieval surfaces expert language patterns aligned with target deficiencies. However, without thematic context the generator cannot ground content-level critique in the prompt's subject matter.

Dual-stage retrieval recovers what each single stage loses. Combining both channels (CRAG+PRAG, no EDPO) reaches 84.0%. The gain over CRAG-only is moderate (+3.1), and the dual-stage variant beats both single stages on every dimension. AEFG therefore needs both signals: thematic context for what to discuss, and problem-aligned exemplars for how to phrase corrective feedback.

EDPO closes the residual gap and contributes most of the lift. Adding EDPO on top of dual-stage retrieval yields a further 5.9-point gain, lifting overall recall to 89.9%. The lift is much larger than the dual-stage gain itself. This validates the framing in §3: the dominant source of LLM–expert misalignment is not which exemplars are retrieved, but which problems the LLM is aware of and how it phrases them. Because EDPO uses only the recall metric on existing annotations, this gain isolates a distinct, prompt-level source of error that retrieval alone cannot fix.

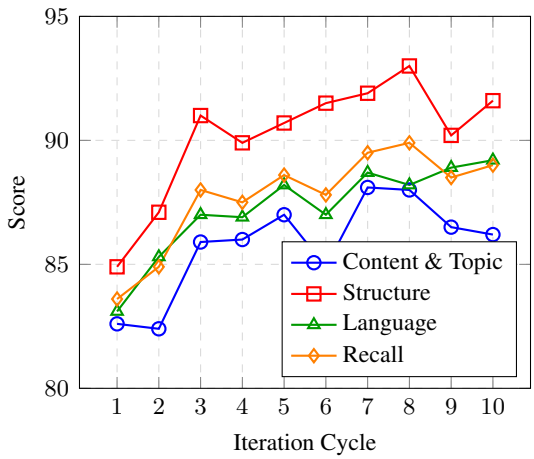

Figure 4: Validation recall across EDPO iterations on the four metrics. Recall rises monotonically until iteration 8 and then plateaus (Eq. (4)).

EDPO converges within eight iterations. Validation recall rises monotonically (per Eq. (4)) until iteration 8 and then plateaus within ±1.5 points (Figure 4). Structural Coherence benefits most (+8.1 points), while Content and Language gain 5.5 and 6.1 points respectively, suggesting structural-feedback language is the most prompt-sensitive component.

Table 1: Main results across seven LLM backbones on **PRISM**. "C" = closed ; "O" = open . Δ is the F1 gain of PARA over Content-RAG.

		Zero-shot			Content-RAG			PARA (Ours)			Δ
Backbone	Type	R	P	F1	R	P	F1	R	P	F1	F1
GPT-4o	C	74.1	59.5	66.0	75.6	70.1	72.7	**74.8**	**68.7**	**71.6**	-1.1
Kimi-K2	C	73.7	56.6	64.0	68.7	49.6	57.6	**75.7**	**49.7**	**60.0**	$+2.4$
GLM-4	C	75.7	64.8	69.8	82.0	68.3	74.5	**83.1**	**69.0**	**75.4**	$+0.9$
Qwen-Plus	O	75.3	63.4	68.8	77.2	50.5	61.1	**82.8**	**58.3**	**68.4**	$+7.3$
Qwen2.5-72B-Instruct	O	57.7	53.8	55.7	66.0	59.6	62.6	**68.7**	**60.4**	**64.3**	$+1.7$
LLaMA-3.1-70B	O	54.6	58.3	56.4	57.8	60.6	59.2	**55.7**	**66.1**	**60.5**	$+1.3$
DeepSeek-V3.2	O	75.4	66.0	70.4	80.9	69.9	75.0	**89.9**	**71.0**	**79.3**	$+4.3$
Average	–	69.9	60.4	64.6	73.4	61.3	66.4	**75.8**	**63.3**	**68.5**	$+2.4$

Table 2: Ablation of CRAG, PRAG, and EDPO modules on the 100-essay benchmark with DeepSeek-V3.2. **Recall** is the mean over the three rubric dimensions. Best in each column is in **bold**.

Variant	**Recall**	**Precision**	**F1**
Zero-shot DeepSeek-V3.2	75.4	66.0	70.4
Few-shot DeepSeek-V3.2	79.0	64.1	70.8
CRAG only	80.9	69.9	75.0
PRAG only	81.3	**71.2**	75.9
CRAG+PRAG (no EDPO)	84.0	69.5	76.0
PARA (Ours)	**89.9**	71.0	**79.3**

4.7 Case Study

We trace a representative test essay on the prompt *"On the Meaning of Perseverance"* (250 characters). The essay opens with a discussion of hardship but never connects to perseverance, producing pervasive topic drift. Its paragraph boundaries lack transitional phrases.

Content-RAG retrieves a high-scoring same-prompt essay whose expert annotation reads *"Thesis is clear; paragraph hierarchy is clean; language is vivid"*—a positive-feedback exemplar with zero corrective signal for the target. **PARA** instead diagnoses the target's latent problems (topic drift, missing transitions) and retrieves an exemplar on a different prompt (*"Reflections on Effort"*) but with an identical problem profile. Its expert annotation reads: *"Topic drifts from the prompt; transitions are missing; evidence-to-title alignment is weak."* This provides direct, corrective language patterns.

PARA's final feedback flags topic drift and the missing paragraph-2/3 transition—both matched by expert annotations. Content-RAG's output covers only generic vocabulary suggestions. The case confirms that problem-aligned retrieval is qualitatively different from content-similar retrieval: it queries by problem type, not by topic, and therefore surfaces corrective rather than congratulatory exemplars.

5 Conclusion

We frame Automated Essay Feedback Generation (AEFG) as a *training-free LLM–expert alignment* problem. A general-purpose LLM lacks the analytic framework of an experienced rater, with a sparse problem inventory and generic phrasing. We close both gaps with **EvoPARA**, a framework of two complementary modules. **EDPO** teaches the LLM to analyse essays like an expert. It distils long-tailed problems from validation failures into the generation prompt under a monotone-update rule. **PARA** grounds that analysis on the target essay through problem-aligned retrieval with agent-verified problem definitions. On **PRISM**, our newly curated benchmark of 653 Chinese middle-school essays with 7,781 expert annotations, EvoPARA reaches 89.9% recall. It improves over the strongest content-similarity RAG baseline by 9.0 absolute recall points on DeepSeek-V3.2, with consistent gains across four Chinese-capable LLM backbones. We hope the alignment view and the released benchmark accelerate broader training-free, rubric-aligned essay-feedback research.

Limitations

Our study is subject to five scope boundaries that future work should address.

Data scale and domain. The evaluation covers 653 Chinese middle-school essays drawn from a single educational context. Whether the gains generalise to other grade levels, demographics, or institutional writing rubrics remains unverified.

Language scope. EvoPARA is evaluated exclusively on Chinese middle-school compositions. Direct transfer to other languages would require both language-specific prompt adaptation and re-indexing of the retrieval corpus under a different embedding model.

Evaluation metric. Our primary metric, recall against structured expert annotations, measures content coverage. It does not directly capture fluency, specificity, or pedagogical appropriateness of generated feedback. The Cohen's $\kappa = 0.81$ consistency check (§4.4) only validates the LLM judge, not the quality of generated comments. A larger human study of feedback quality is left to future work.

Model-side dependencies. Both the diagnosis step (§3.4) and EDPO's gap analysis (§3.3) rely on the reasoning capacity of the backbone LLM. A substantially weaker model is likely to produce shallower problem descriptions and less informative pattern critiques. EDPO's long-tailed coverage is further bounded by the failure diversity present in the validation set; problems that never appear in $\mathcal{V}$ cannot be distilled. Finally, on backbones with very strong English-pretrained priors (e.g., GPT-4o), Content-RAG can already saturate the available signal, and PARA's marginal gain narrows correspondingly.

Statistical variance. All numbers in Tables 1 and 2 come from a single inference run per configuration; multi-seed variance estimation was not performed under API-cost constraints. Absolute gaps should therefore be read as point estimates rather than significance-tested differences.

Acknowledgments

References

Beijing Language and Culture University. 2017. HSK dynamic composition corpus (version 2.0). http://yuyanziyuan.blcu.edu.cn/en/info/1043/1501.htm. Online resource; original corpus released in 2006, V2.0 maintained at the BLCU Language Resource Centre.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on ChatGPT. In *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of ACL*.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of CoNLL*, pages 153–162.

Menna Fateen, Bo Wang, and Tsunenori Mine. 2024. Beyond scores: A modular RAG-based system for automatic short answer scoring with feedback. *arXiv preprint arXiv:2409.20042*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of ACL*, pages 1762–1777.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations (ICLR)*.

You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. 2023. Review of feedback in automated essay scoring. *arXiv preprint arXiv:2307.05553*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank

Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. *Proceedings of IJCAI*, pages 6300–6308.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. A replication study of dense passage retriever. *arXiv preprint arXiv:2104.05740*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of EMNLP*, pages 7957–7968.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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), NAACL*, pages 283–298.

Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew Soon Ong. 2024. Large language models for intent-driven session recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 324–334.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of EMNLP*, pages 1882–1891.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of EMNLP*, pages 9414–9423.

Hongyi Wu, Xinshu Shen, Man Lan, Xiaopeng Bai, Yuanbin Wu, Aimin Zhou, Shaoguang Mao, Tao Ge, and Yan Xia. 2023. Overview of the NLPCC 2023 shared task: Chinese essay discourse coherence evaluation. In *Natural Language Processing and Chinese Computing (NLPCC), Part III*, Lecture Notes in Computer Science.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations (ICLR)*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *International Conference on Learning Representations (ICLR)*.

A Annotation Protocol, Recruitment, and Compensation

Annotators. Four experienced Chinese-language teachers (5+ years of middle-school teaching) from two partner schools each annotated roughly 163 essays of the 653-essay benchmark.

Compensation. A partial honorarium of approximately USD 100 was distributed across the four annotators; the remainder of their contribution was on a voluntary collaborative basis under the schools' standard pedagogical-research practice. Per-essay rates, normalised to estimated annotation time, exceed the regional minimum hourly wage.

Consent and ethics review. Essays were drawn from partner schools' archival classroom and examination collections with prior written consent from the schools; no new data collection was conducted, and we did not interact directly with students. Annotators consented separately in writing to having their annotations released as part of the public benchmark. Under our institution's protocol for secondary use of de-identified educational records, no separate IRB review was required.

Annotation protocol. The guideline specifies three rubric dimensions (content and topic adherence, structural coherence, language expression), three comment types per dimension (advantage, disadvantage, suggestion), worked examples, and a dual-pass review. Annotators were instructed to comment at paragraph or sentence granularity rather than at the holistic-essay level.

PII screening. We additionally ran a regex PII scan over all 1,959 source files covering Chinese government ID numbers, mobile-phone numbers, e-mail addresses, school and class designators, name-introduction patterns, labelled identifier fields, and detailed-address patterns. The scan returned zero hits on the ID, phone, e-mail, class-designator, and self-introduction patterns. The school-name and labelled-field matches collapsed on manual inspection to (i) generic school-stage words, (ii) references to public figures cited as essay evidence, and (iii) quotations of the prompt's anti-PII instruction. No student-identifying content was detected.

B Compute Budget

All experiments were run by invoking commercial and open LLM APIs from a single workstation; no local GPU training was performed. Across the seven backbones, five baselines and ablation variants (Table 2), and the ten-iteration EDPO loop on the validation split, the combined input plus output token volume exceeded 10^9 tokens. End-to-end wall-clock time, including queueing and rate-limit windows, spanned approximately 120 calendar days of intermittent execution. Aggregate API spend across all providers was approximately USD 5,000. A single PARA inference pass over the 100-essay validation + test split on a single backbone completes in a few hours and costs on the order of tens of US dollars, so partial reproduction of any single Table 1 row is inexpensive.