

Rubric-Aware Multi-Expert Online Policy Distillation for Chinese Elementary School Essay Writing

Wei Tian¹, Yuhao Zhou¹, Xiayu Sun¹, Yadong Zhang¹,
Hao Zhou¹, Shu Liu¹, Man Lan^{1,2,*}

¹School of Computer Science and Technology, East China Normal University

²Shanghai Institute of Artificial Intelligence for Education,
East China Normal University

tianwei@stu.ecnu.edu.cn, mlan@cs.ecnu.edu.cn

*Corresponding author

Abstract

Chinese elementary-school writing is rubric-driven. We study it as a teacher-facing instructional support task: given a classroom prompt and requirements, generate a teacher-reviewable example matching the target genre, rubric, length, and age-appropriate register. Directly prompted frontier LLMs often misroute genres or adopt an adult register. We present **RA-MOPD**, a four-stage pipeline that post-trains a 4B model into a unified essay writer across four genres and eight rubric dimensions. Its core stage augments on-policy reverse-KL distillation with *Rubric-Weighted KL*, which up-weights rubric-critical tokens via reward-model attribution, and *Cross-Genre Mixing*, a parameter-free expert blend guided by an empirical transfer prior. On a new 6,000-essay benchmark with eight reward models (QWK = 0.864), RA-MOPD reaches 4.21/5.00 and 97.8% routing—within 0.09 of GPT-5.3 3-shot ($p < 0.01$). No statistically significant difference is detected relative to a 4 $\times$ -cost four-expert logit-averaging baseline ($\Delta = +0.03$, $p = 0.18$). A GPT-4.1 judge, BERTScore, a 1,200-essay teacher study (62.8% preference), and register analysis provide converging evidence beyond RM metrics. GPT-5.5 holistic evaluation further supports the component-level gains, which remain stable on Qwen3-1.7B without model-specific re-tuning. Code and data: <https://github.com/TW-NLP/MOPD>.

1 Introduction

We study rubric-aligned essay generation as a *teacher-facing* instructional support task. Given a classroom writing prompt and instructional requirements, our goal is to generate a teacher-reviewable example that follows the target genre and rubric while maintaining the requested length and an age-appropriate register. Such examples can support teachers in preparing rubric-aligned model essays,

contrastive examples for explaining rubric criteria, and prompt-specific revision demonstrations.

Writing instruction in Chinese elementary schools is heavily rubric-driven: teachers evaluate student essays along genre-specific criteria such as *format & context*, *narration completeness*, *novelty*, and *multisense & rhetoric*. A model capable of writing at an elementary-school level must simultaneously master (1) *genre routing*—selecting the correct register (letter, descriptive paragraph, narrative arc, or imaginative tale) from a free-form prompt; (2) *rubric-aligned content* satisfying two genre-specific quality dimensions; and (3) *length and child-appropriate register*. A fluent essay may still be unsuitable as a classroom example if it follows the wrong genre, misses the target rubric, violates the requested length, or uses an overly adult register.

Directly prompting frontier LLMs exposes three categories of failure on our 1,200-prompt benchmark. **(i) Format non-compliance:** Qwen3-max attains only 64.4% routing accuracy at 0-shot and 75.3% at 3-shot, far below classroom rubric requirements. **(ii) Register mismatch:** GPT-5.3 produces fluent but adult-style language; the human study below shows its outputs preferred over our model on only 28.7% of paired comparisons by primary-school teachers. **(iii) Conflicting per-dimension objectives:** A single GRPO model trained jointly across all four genres (“Mix-GRPO”) scores 3.84 on the 8-dim average, above SFT (3.78) but below the best per-genre specialist (3.91); a gradient-cosine analysis (Appendix P) reveals destructive interference between Applied-Format and Narrative-Completeness rewards ($\cos \approx -0.13$). These failures together demand a pipeline that combines per-genre policy optimisation with *rubric-aware* fusion of the resulting specialists.

We address the problem from both the *data* side—a balanced 6,000-essay benchmark with double an-

notation and eight reward models exceeding inter-annotator agreement—and the *modelling* side: a four-stage post-training pipeline (Figure 1) that converts a 4B Qwen3 base model into a unified essay writer. **Stage 1** performs genre-router SFT on the high-quality subset of the corpus; **Stage 2** trains eight dimension-specific reward models (average QWK=0.864, above the human $\kappa=0.76$); **Stage 3** fine-tunes four genre specialists via GRPO with a composite reward (genre routing, two RM-based rubric terms, and a Gaussian length term); and **Stage 4 (RA-MOPD)** trains a single student under a *rubric-aware* variant of online policy distillation. Following the standard OPD formulation (Xu et al., 2026; Team et al., 2026; Agarwal et al., 2024), the student is trained *purely* by a token-level reverse-KL divergence on its own rollouts, with no separate cross-entropy stream. RA-MOPD adds two task-tailored extensions: (a) *Rubric-Weighted KL (RW-KL)* up-weights rubric-critical positions using RM perturbation-based attribution; and (b) *Cross-Genre expert Mixing (CGM)* blends the top-2 experts using an empirical transfer prior as a parameter-free complement to RW-KL.

RA-MOPD achieves 4.21/5.00 (8-Avg) and 97.8% routing, competitive with GPT-5.3 3-shot ($\Delta=+0.09$, $p<0.01$) under rubric-aligned evaluation. Relative to a 4 $\times$ -cost ground-truth-routed four-expert logit-averaging baseline, no statistically significant difference is detected ($\Delta=+0.03$, $p=0.18$). We note that the frontier comparison is structurally asymmetric—RA-MOPD is domain-trained while GPT-5.3 is evaluated inference-only—and should be interpreted as quantifying the marginal value of targeted post-training on a narrow rubric corpus rather than as a general capability claim (see §A). Three corroborating protocols (§5.4) consistently reproduce the ranking. Generated examples are intended to be reviewed and adapted by teachers rather than used to replace student writing, classroom instruction, or human assessment.

Contributions. (i) A 6,000-essay benchmark spanning four genres and eight doubly-annotated rubric dimensions (Shanghai grades 3–6), with eight reward models at QWK=0.864. (ii) **RA-MOPD**: two task-tailored extensions to pure on-policy reverse-KL distillation—RW-KL (RM-guided perturbation-based per-token weighting) and CGM (empirical transfer-matrix prior for expert mixing, a parameter-free engineering

complement)—with CGM controls confirming the gain requires the transfer prior, not merely a second teacher. (iii) A multi-protocol corroboration framework (gpt-4.1-2025-04-14 holistic judge, BERTScore, full-test human study) providing converging evidence beyond RM metrics, plus a register-vs-rubric analysis (§5.5) showing that RA-MOPD *Pareto-dominates* all other deployable systems on the joint (rubric, register) objective and explaining why gold essays score lower on RM metrics. Additional GPT-5.5 holistic evaluation and a Qwen3-1.7B experiment further validate the component gains from complementary evaluation and model-scale perspectives.

2 Task Formulation

Let $\mathcal{G} = \{g_1, \dots, g_4\}$ denote the four target genres and $\mathcal{D} = \{(d_a^g, d_b^g)\}_{g \in \mathcal{G}}$ the eight rubric dimensions (two per genre). Each test prompt x specifies a writing topic together with a target length $L^*(x)$ extractable by regex (Appendix F). A generative model π_θ produces an essay $y \sim \pi_\theta(\cdot | x)$. We track:

$$\text{Mean}_d^g(\theta) = \frac{1}{N_g} \sum_{x \in \mathcal{T}_g} \text{RM}_d(x, y) \quad (1)$$

$$\text{Acc}^g(\theta) = \frac{1}{N_g} \sum_{x \in \mathcal{T}_g} \mathbb{1}[\tilde{g}(y) = g] \quad (2)$$

$$8\text{-Avg}(\theta) = \frac{1}{8} \sum_{g \in \mathcal{G}} \sum_{d \in \{d_a^g, d_b^g\}} \text{Mean}_d^g(\theta) \quad (3)$$

where $\mathcal{T}_g$ is the test set for genre g ($|\mathcal{T}_g|=300$), $\text{RM}_d(\cdot)$ is the reward model for dimension d (§4.2), and $\tilde{g}(y)$ is the predicted genre parsed by string-matching on the prefix "Genre: ~x". Our optimisation target is 8-Avg jointly with the four per-genre accuracies Acc^g .

3 The CHILD-ALIGN Dataset

Source. We collect 10,746 handwritten essays from public elementary schools in Shanghai (grades 3–6), transcribed by commercial OCR and partitioned into four genres (Narrative 3,337; Object 2,641; Imagination 2,629; Applied 2,139). After cleaning, 8,268 essays survive; we draw a balanced 1,500-per-genre sample (6,000 total). All identifying metadata was stripped prior to OCR. The grade $\times$ genre distribution (grade 4 modal at 33.3%; grade 6 only 10.1%, concentrated in Applied) is given in Appendix D, Table 8.

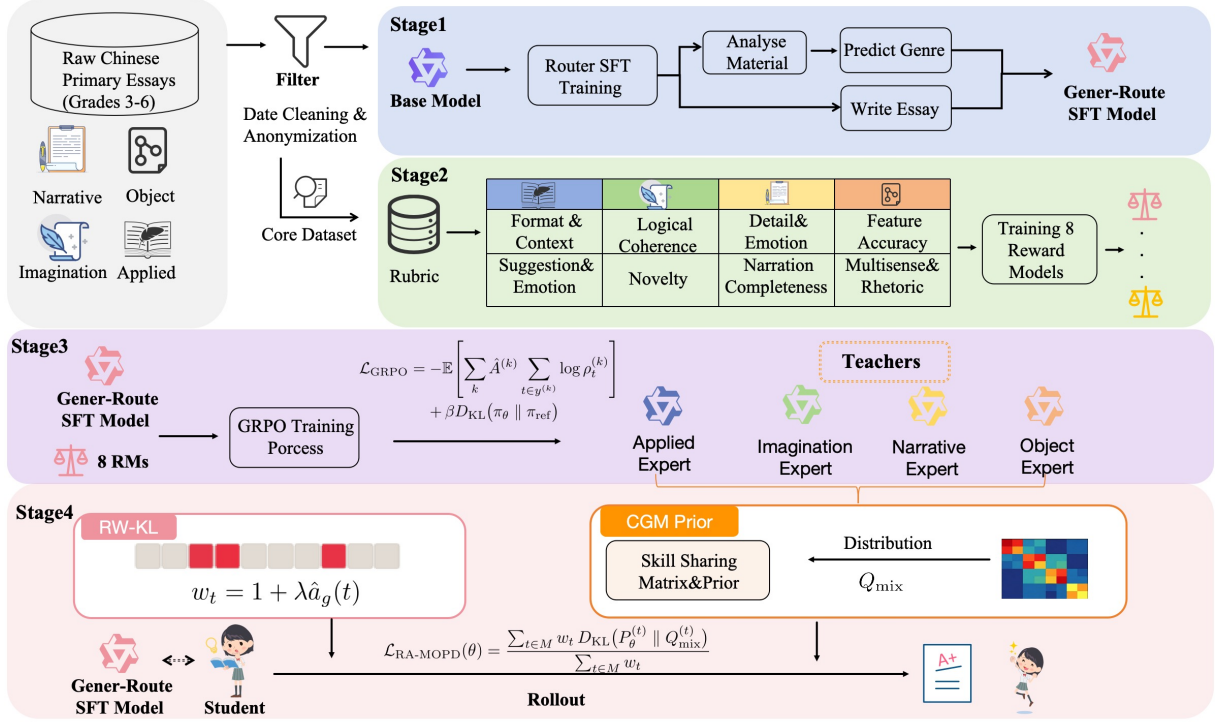


Figure 1: Overview of the four-stage post-training pipeline.

Cleaning, sampling, and annotation. Four sequential filters handle OCR noise: length filter ($[100, 1500]$ characters, 3.7% pruned), regex removal of control symbols, LLM-driven PII redaction (25.5% of essays contained PII; residual $\sim 1.0\%$ by manual audit, 95% exact binomial upper bound $\approx 2.9\%$), and exact-match deduplication. Full details are in Appendix D. From the 8,268 cleaned essays we draw a stratified balanced sample of 1,500 per genre and split each genre 60/20/20 into train/val/test under a strictly *prompt-disjoint* partition.

In consultation with five elementary-school teachers, we designed two scoring dimensions per genre (Table 1; full 1–5 anchors in Appendix G). Three graduate-student annotators, all with formal language-pedagogy training, participated in scoring: *two* of them independently scored every essay on its two rubric dimensions, and the *third*—the most senior, a Ph.D. candidate—resolved ≥ 2 -point disagreements (4.2% of essays). Labels are the rounded mean of the two primary scores. Per-dimension agreement between the two primary annotators (Appendix D) averages Pearson $r=0.80$, weighted $\kappa=0.76$.

Genre	Two rubric dimensions
Applied	Format & Context / Suggestion & Emotion
Imagination	Logical Coherence / Novelty
Narrative	Detail & Emotion / Narration Completeness
Object	Feature Accuracy / Multisense & Rhetoric

Table 1: The eight rubric dimensions used throughout this paper.

4 Method

Our pipeline (Figure 1) has four stages. We describe Stage 1 first, followed by the reward models (Stage 2), which serve as a data-quality filter for Stage 1 and as the training-time signal for Stages 3–4.

4.1 Stage 1: Genre-Router SFT

We filter the 3,600 training essays by requiring both rubric scores to be ≥ 4 , yielding 1,885 high-quality samples. The SFT target is the two-line completion `Genre:  $g$  / Essay:  $y$` , a format shared by all downstream stages so that a single generation can be parsed into a genre prediction and an essay body. Let $z = (z_1, \dots, z_T)$ be the tokenised assistant target and $m_t \in \{0, 1\}$ a mask that is 1 on assistant tokens. We minimise the masked cross-entropy

Reward model	r	QWK	MAE	MSE
App. Format & Context	.870	.869	.233	.427
App. Suggestion & Emotion	.832	.832	.247	.340
Imag. Logical Coherence	.890	.889	.230	.250
Imag. Novelty	.960	.960	.093	.113
Narr. Detail & Emotion	.845	.840	.250	.283
Narr. Narration Compl.	.855	.853	.180	.220
Obj. Feature Accuracy	.811	.810	.300	.413
Obj. Multisense & Rhet.	.861	.858	.287	.307
Average	.865	.864	.228	.294

Table 2: Reward-model performance on the 300-essay held-out test sets. Average QWK (0.864) substantially exceeds the inter-annotator $\kappa=0.76$ against the same adjudicated gold labels.

loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = - \frac{\sum_t m_t \log \pi_\theta(z_t | z_{<t})}{\sum_t m_t} \quad (4)$$

Training details are in Appendix E. The resulting STAGE1-SFT model initialises Stages 3 and 4.

4.2 Stage 2: Reward Models

We train eight dimension-specific reward models from the annotated corpus. Each RM is a Qwen3-4B fine-tuned to map (g, d, r_d, x, y) —genre, dimension, rubric anchor, prompt, essay—to a score in $\{1, \dots, 5\}$ (full SFT; train/val/test: 900/300/300; details in Appendix E). Table 2 reports r /QWK/MAE/MSE; the average QWK of 0.864 exceeds the inter-annotator $\kappa=0.76$ on the same adjudicated gold labels, which we interpret as calibration to the adjudicated standard rather than superhuman scoring.

Human validation of trained RMs. To rule out self-consistency bias, we ran a separate three-annotator validation on a freshly sampled 800-essay panel (100 per dimension), drawn from the 2,268 cleaned essays not selected for the 6,000-essay benchmark and therefore disjoint from its train/validation/test splits and original annotation pool. RM–human Pearson averages 0.821 (QWK 0.816), exceeding the fresh inter-rater agreement of 0.77, confirming RM quality under a disjoint evaluator group; full results are in Appendix C.

4.3 Stage 3: Per-Genre GRPO

Background. We fine-tune four genre specialists from STAGE1-SFT using GRPO (Shao et al., 2024), which avoids a learned value head by normalising rewards within a group of K completions: $\hat{A}^{(k)} = (r^{(k)} - \mu_r) / (\sigma_r + \epsilon)$, and optimises the

clipped surrogate objective with a KL penalty against the reference policy (Eq. 17, Appendix F). This design is advantageous here because rewards from eight separate RMs lack a shared baseline. Comparisons with preference-optimisation alternatives (DPO, ORPO) constructed from RM-derived pairs are reported in Appendix M.

Composite reward. For a completion y with target genre g , rubric dimensions $\{d_a, d_b\}$, and regex-extracted target length L^* :

$$R_{\text{genre}} = \mathbb{1}[\hat{g}(y)=g] \in \{0, 1\} \quad (5)$$

$$R_d = (\text{RM}_d(x, y) - 1) / 4 \in [0, 1] \quad (6)$$

$$R_{\text{len}} = \exp\left(-\frac{(L(y) - L^*)^2}{2\sigma_L^2 (L^*)^2}\right) \quad (7)$$

with length tolerance $\sigma_L=0.2$. The composite reward is

$$\mathcal{R} = w_g R_{\text{genre}} + w_a R_{d_a} + w_b R_{d_b} + w_\ell R_{\text{len}} \quad (8)$$

with $(w_g, w_a, w_b, w_\ell) = (0.2, 0.3, 0.3, 0.2)$ selected on the validation set against the two natural alternatives $(0.25, 0.25, 0.25, 0.25)$ (uniform) and $(0.1, 0.4, 0.4, 0.1)$ (rubric-heavy); Appendix L reports the full sweep. Training details (TRL, vLLM (Kwon et al., 2023), $K=8$, $\beta=0.01$) are in Appendix E.

4.4 Stage 4: Rubric-Aware Multi-Expert Online Policy Distillation

Motivation. Each per-genre GRPO specialist gains specialty-genre quality at the cost of routing generality: routing accuracy drops by 0.3–10.7 pp relative to STAGE1-SFT (§5.1). MOPD (Xu et al., 2026; Team et al., 2026) fuses the four specialists into a single student. We first describe a base MOPD recipe (§4.4.1), identify two limitations specific to our rubric-guided setting (§4.4.2), and introduce the RA-MOPD extensions that address them: RW-KL (§4.4.3) and CGM (§4.4.4).

4.4.1 Base MOPD: Pure On-Policy Reverse-KL Distillation

Following the standard online policy distillation formulation (Xu et al., 2026; Team et al., 2026; Agarwal et al., 2024; Gu et al., 2024), base MOPD trains the student under a single token-level *reverse-KL* objective on its own rollouts:

$$\mathcal{L}_{\text{MOPD}}(\theta) = \mathbb{E}_{x, \hat{y} \sim \pi_\theta(\cdot|x)} \left[\frac{1}{|M|} \sum_{t \in M} D_{\text{KL}}\left(P_\theta^{(t)} \parallel Q_g^{(t)}\right) \right] \quad (9)$$

where $P_\theta^{(t)}$ and $Q_g^{(t)}$ are the student and expert next-token distributions computed on the student’s own rollout $\hat{y} \sim \pi_\theta(\cdot | x)$ via teacher-forced forward passes (sampling and batching details in Appendix E and F). The mask M selects only rollout positions (prompt tokens are excluded). The KL direction $D_{\text{KL}}(P_\theta \| Q_g)$ is the mode-seeking direction (Gu et al., 2024) that prevents the student from spreading mass to teacher regions it cannot realise. There is *no separate CE stream*; the entire training signal is the on-policy reverse-KL, as in DeepSeek-V4 (Xu et al., 2026) and MiMo-V2 (Team et al., 2026). The necessity of online reverse-KL over offline CE and forward-KL alternatives is validated in Appendix N.

Genre-homogeneous batching and lazy expert swapping are detailed in Appendix F.

4.4.2 Why Base MOPD Is Insufficient

Limitation 1: Token-uniform KL wastes the RM signal. Eq. 9 weights every position equally, yet rubric quality is highly *localised*: in Applied letters, salutation and closing occupy <10% of tokens but drive >40% of the Format score; Narrative completeness depends on a handful of connectives and climax verbs. The eight Stage-2 RMs already encode token importance at QWK 0.864; base MOPD ignores this signal entirely.

Limitation 2: Single-expert teaching discards transferable knowledge. Base MOPD distills only from π_{ϕ_g} , ignoring the substantial cross-genre competence shared by all four specialists (every off-diagonal of the transfer heatmap ≥ 3.51 ; Appendix I). The empirical transfer matrix provides a natural fixed prior for blending the top-2 experts.

4.4.3 RW-KL: Rubric-Weighted Token-Level KL

For a rollout $\hat{y}=(\hat{y}_1, \dots, \hat{y}_L)$ of target genre g with rubric dimensions $\{d_a, d_b\}$, we partition $\hat{y}$ into $K=10$ contiguous chunks $\{c_1, \dots, c_K\}$ and compute, for each $d \in \{d_a, d_b\}$, a chunk-level leave-one-out attribution:

$$\Delta_d(c_i) = \text{RM}_d(x, \hat{y}) - \text{RM}_d(x, \hat{y}^{\setminus c_i}) \quad (10)$$

where $\hat{y}^{\setminus c_i}$ replaces chunk c_i with a length-matched generic placeholder. We treat this as a zeroth-order perturbation-based importance score rather than a strict attribution method; known limitations include nonlinear token interactions and a minor conflation between content and length contributions

introduced by the chunk substitution (most pronounced for Imagination Novelty, Kendall $\tau=0.52$; Appendix K). We rectify, normalise within the rollout, and expand to token level:

$$a_g(t) = \frac{1}{2} \sum_{d \in \{d_a, d_b\}} \frac{\max(0, \Delta_d(c_{i(t)}))}{|c_{i(t)}| \sum_j \max(0, \Delta_d(c_j))} \quad (11)$$

followed by a max-rescale

$$\hat{a}_g(t) = a_g(t) / \max_{t' \in M} a_g(t') \quad (12)$$

so the most attributed token has weight 1. Defining $w_t = 1 + \lambda \hat{a}_g(t)$, the RW-KL loss is the w_t -weighted reverse-KL on the rollout positions:

$$\mathcal{L}_{\text{KL}}^{\text{RW}} = \frac{\sum_{t \in M} w_t D_{\text{KL}}(P_\theta^{(t)} \| Q_g^{(t)})}{\sum_{t \in M} w_t} \quad (13)$$

with $\lambda=2$ (Appendix O). Attribution is recomputed every 5 steps; sanity checks and K -sensitivity are in Appendices J–K.

4.4.4 CGM: Cross-Genre Expert Mixing

Let $H \in \mathbb{R}^{4 \times 4}$ denote the cross-genre transfer matrix, where $H_{g,g'}$ is the two-dimensional mean RM score of expert $\pi_{\phi_{g'}}$ on the 300-prompt *validation* set of genre g (Appendix I; computed on validation to avoid contaminating the test split). For target genre g we keep only the top-2 entries of row $H_{g,:}$ and apply a temperature- τ softmax:

$$\pi_{g'|g} = \frac{\exp(H_{g,g'}/\tau) \mathbb{1}[g' \in \text{Top2}(H_{g,:})]}{\sum_{g''} \exp(H_{g,g''}/\tau) \mathbb{1}[g'' \in \text{Top2}(H_{g,:})]} \quad (14)$$

We use $\tau=0.2$, which assigns 0.73–0.92 to the matched expert and 0.08–0.27 to the next-best transfer expert (per-genre weights in Appendix O).

To avoid leakage, H is computed on the first 150 validation prompts per genre and τ is selected on the remaining 150.

The CGM teacher distribution is the probability-level mixture

$$Q_{\text{mix}}^{(t)} = \sum_{g' \in \text{Top2}(H_{g,:})} \pi_{g'|g} \text{softmax}(\ell_{\phi_{g'}}^{(t)}) \quad (15)$$

Unlike MoE (Shazeer et al., 2017), CGM is a training-only, parameter-free heuristic with no learned routing; its contribution is empirical effectiveness (Table 21) rather than algorithmic novelty.

4.4.5 The RA-MOPD Objective

Combining RW-KL (Eq. 13) with the CGM teacher (Eq. 15), the full RA-MOPD loss is a single rubric-weighted reverse-KL term against the CGM teacher, with no CE component:

$$\mathcal{L}_{\text{RA-MOPD}}(\theta) = \frac{\sum_{t \in M} w_t D_{\text{KL}}(P_{\theta}^{(t)} \parallel Q_{\text{mix}}^{(t)})}{\sum_{t \in M} w_t} \quad (16)$$

Relative to Eq. 9, RA-MOPD adds only w_t (RW-KL) and Q_{mix} (CGM); the on-policy reverse-KL protocol and the absence of any CE stream follow Xu et al. (2026) and Team et al. (2026) exactly. No new tunable weights beyond λ, τ (both fixed on val). Full algorithm: Appendix F.

5 Experiments

Systems compared. We evaluate 15 systems in three groups: (i) *Baselines*: Qwen3-4B base; Qwen3-max, DeepSeek-V4-pro, and GPT-5.3 each at 0-shot and 3-shot. (ii) *Pipeline intermediates*: STAGE1-SFT; Mix-GRPO; four per-genre GRPO experts. (iii) Our final **Base MOPD** and **RA-MOPD**, both initialised from STAGE1-SFT. Preference-optimisation alternatives (DPO, ORPO) are evaluated in Appendix M.

Protocol. Each system generates one essay per test prompt with greedy decoding (temperature 0). We parse $\tilde{g}(y)$ and the essay body, score the two genre-matched dimensions via the eight-RM bench, and aggregate via Eqs. 1–3. Closed-source LLMs use the same system message and a deterministic 3-shot demonstration set (Appendix H). For each pairwise comparison we report a 10,000-replicate paired bootstrap 95% CI on 8-Avg.

5.1 Main Results

Table 3 summarises performance. RA-MOPD attains **8-Avg** = 4.21 and average routing accuracy of 97.8%, the highest of any system.

Against frontier LLMs (3-shot). RA-MOPD is competitive with GPT-5.3 3-shot ($\Delta = +0.09$; CI $[+0.041, +0.139]$, $p < 0.01$) on 8-Avg, and matches or outperforms GPT-5.3 on all eight dimensions, with six improvements and two ties at the reported precision. It outperforms Qwen3-max (+0.56) and DeepSeek-V4-pro (+0.55). These margins reflect *domain-targeted training* versus *inference-only* evaluation rather than general capability (see §A). Routing margins are decisive (+22.5 pp over

Qwen3-max; +3.0 pp over GPT-5.3); on 8-Avg, prompt engineering does not help: 0-shot and 3-shot scores are identical for Qwen3-max and DeepSeek-V4-pro, though routing accuracy does improve with 3-shot (Qwen3-max 64.4% $\rightarrow$ 75.3%; DeepSeek-V4-pro 46.6% $\rightarrow$ 58.3%).

Base MOPD already surpasses every specialist; RA-MOPD adds +0.12. Base MOPD (pure on-policy reverse-KL, no CE stream) is the strongest *single* model among all intermediates at 4.09 8-Avg and 97.5% routing—above every per-genre GRPO specialist (best 3.91), Mix-GRPO (3.84), and Stage1-SFT (3.78). Base MOPD scores ≈ 0.05 below each matched specialist on its specialty columns; RA-MOPD closes this gap via RW-KL and CGM, gaining +0.12 over base MOPD and +0.30 over the best single specialist. Mix-GRPO (3.84) marginally outperforms SFT (3.78) on 8-Avg but falls short of the best specialist (3.91) due to cross-genre gradient interference ($\cos \approx -0.13$; Appendix P).

5.2 Ablation Studies

Stage-by-stage contributions. Cumulative gains (Table 3): SFT +0.48, best per-genre expert +0.13, Base MOPD +0.18, RA-MOPD extensions +0.12.

Decomposing RA-MOPD. Table 4 isolates the contribution of each RA-MOPD component, all starting from base MOPD. RW-KL alone yields +0.06, CGM alone yields +0.06, and combined they yield +0.12—exactly complementary. RW-KL’s largest gains are on structurally-localised dimensions (Applied-Format +0.07; Narrative-Completeness +0.07); CGM’s largest gains are on dimensions benefiting from cross-genre vocabulary transfer (Object-Multisense +0.08; Narrative-Detail +0.07). Neither alone closes the Imagination-Novelty gap with GPT-5.3. Full per-dimension and hyperparameter sweeps appear in Appendix O.

Additional holistic validation. To provide a more comprehensive, multi-perspective validation of the proposed components, we additionally evaluate Base MOPD, +RW-KL, +CGM, and RA-MOPD using a GPT-5.5 holistic judge. The judge assesses overall relevance, fluency, structure, age-appropriate expression, and originality without access to system identities or RM scores. GPT-5.5 scores increase from 3.97 for Base MOPD to 4.00 with RW-KL, 4.00 with CGM, and 4.04 with their

Model	Applied Form. Sug.		Imagination Log. Nov.		Narrative Det. Nar.		Object Feat. Mul.		8-Avg	Avg acc (%)
Gold (human)	3.32	3.59	3.66	3.25	3.39	3.65	3.49	3.48	3.48	—
base (Qwen3-4B)	3.28	3.68	3.10	3.18	3.16	3.68	3.25	3.03	3.30	39.3
Qwen3-max 0-shot	3.98	3.65	4.17	3.39	3.66	3.92	3.23	3.21	3.65	64.4
Qwen3-max 3-shot	3.97	3.67	4.14	3.36	3.67	3.91	3.30	3.21	3.65	75.3
DeepSeek-V4-pro 0-shot	3.80	3.76	4.13	3.45	3.66	3.96	3.18	3.34	3.66	46.6
DeepSeek-V4-pro 3-shot	3.83	3.75	4.09	3.44	3.66	3.96	3.20	3.32	3.66	58.3
GPT-5.3 0-shot	4.00	3.90	4.05	3.65	4.00	4.10	3.85	3.90	3.93	76.3
GPT-5.3 3-shot	4.10	4.00	4.20	4.10	4.05	4.25	4.10	4.15	4.12	94.8
+ SFT (Stage 1)	3.75	3.75	3.79	3.30	3.63	4.08	3.97	3.96	3.78	94.1
+ Mix-GRPO	3.85	3.85	3.80	3.45	3.70	4.10	4.00	3.95	3.84	91.0
+ GRPO _{App}	4.10	4.00	3.82	3.31	3.63	4.04	3.99	3.93	3.85	93.8
+ GRPO _{Imag}	3.79	3.73	4.10	4.00	3.64	4.08	3.99	3.92	3.91	90.8
+ GRPO _{Narr}	3.74	3.77	3.76	3.26	4.10	4.30	4.00	3.95	3.86	85.1
+ GRPO _{Obj}	3.78	3.72	3.81	3.31	3.57	4.06	4.20	4.15	3.83	83.4
+ Base MOPD	4.05	3.97	4.06	3.97	4.07	4.27	4.17	4.12	4.09	97.5
+ RA-MOPD (ours)	4.13	4.05	4.20	4.10	4.20	4.42	4.30	4.28	4.21	97.8

Table 3: Main results. Per-genre per-dimension mean RM score (1–5) and macro-averaged routing accuracy (%) on the 1,200-essay held-out benchmark (prompt-disjoint from training). **Bold**: best in column. **Blue**: cells where a per-genre GRPO model operates in its specialty. Base MOPD is the strongest *single* model among all intermediates (4.09 8-Avg, 97.5% accuracy). RA-MOPD adds +0.12 on 8-Avg over base MOPD; the RA-MOPD vs. GPT-5.3 3-shot 8-Avg difference of +0.09 has 95% CI [+0.041, +0.139] ($p < 0.01$); note that this comparison is structurally asymmetric (domain-trained model vs. inference-only frontier).

Variant	8-Avg	Δ Base
Base MOPD	4.09	—
+ RW-KL only	4.15	+0.06
+ CGM only	4.15	+0.06
+ RW-KL + CGM (RA-MOPD)	4.21	+0.12

Table 4: Decomposition of RA-MOPD gains over Base MOPD. RW-KL and CGM are complementary (+0.06 + 0.06 $\rightarrow$ +0.12).

combination. The RA-MOPD improvement over Base MOPD is +0.07 and statistically significant ($p < 0.01$). This ordering is consistent with RM 8-Avg, further supporting the individual effectiveness and complementarity of RW-KL and CGM. Full results and protocol details are in Appendix S.

CGM teacher-mixing controls. To verify that CGM’s gain stems from the empirical transfer prior rather than from simply having a second teacher, we compare four mixing strategies (Appendix O, Table 21): a *random* second expert yields only +0.01 ($p = 0.63$), ruling out the “any-second-teacher” hypothesis; a *uniform* top-2 mixture reaches +0.03; and CGM’s score-weighted top-2 achieves +0.06, matching a *learned-gating oracle* (+0.07) at zero parameter cost—i.e., the empirical transfer matrix supplies a signal nearly equivalent to learned routing.

Ground-truth-routed four-expert logit-averaging baseline. At inference, this baseline uses the ground-truth genre label and averages the logits of all four GRPO specialists ($4\times$ inference cost; not deployable). RA-MOPD scores 4.21 versus 4.18 for this baseline ($\Delta = +0.03$, $p = 0.18$). No statistically significant difference is detected under the current evaluation; this result should not be interpreted as equivalence or as establishing a theoretical upper bound.

Variance across training seeds. Across three independent seeds (Appendix O, Table 22), each component’s +0.06 gain is stable ($\sigma \leq 0.010$) and the RA-MOPD–GPT-5.3 gap (+0.09) substantially exceeds 2σ . Six independent teachers scored all 1,200 essays (blinded to system identity); cell-level Pearson $r = 0.87$ with RM scores ($p < 10^{-6}$); RA-MOPD preferred over GPT-5.3 in 62.8% of pairwise comparisons ($p < 10^{-9}$) and over the matched expert in 58.4% ($p < 10^{-5}$). Full protocol: Appendix Q.

5.3 Backbone-Scale Robustness

To examine whether the proposed gains remain stable at a smaller model scale, we repeat the complete Stage1-SFT $\rightarrow$ Base MOPD $\rightarrow$ RA-MOPD pipeline on Qwen3-1.7B. We use the same data

split, reward models, decoding protocol, reward coefficients, RW-KL configuration, and CGM configuration, without backbone-specific retuning. On Qwen3-1.7B, Stage1-SFT, Base MOPD, and RA-MOPD achieve RM 8-Avg scores of 3.63, 3.93, and 4.04, respectively. RA-MOPD therefore retains a +0.11 gain over Base MOPD, closely matching the +0.12 gain observed on Qwen3-4B, and reaches 96.3% routing accuracy. RW-KL and CGM also remain individually effective, improving Base MOPD to 3.98 and 3.99, respectively, while their combination performs best at 4.04. The consistent relative improvement across Qwen3-4B and Qwen3-1.7B demonstrates that the components remain effective across two model scales without model-specific retuning. Full results are in Appendix T.

5.4 Corroborating Evidence Beyond RM Metrics

We complement the rubric-aligned RM evaluation with three protocols that assess the systems from different perspectives. **(1) GPT-4.1 holistic judge:** 200 essays per system scored on five criteria with no lexical overlap with our rubric labels (full prompt: Appendix R). Among the three generated systems shown in Table 5, the system-level ordering matches the RM-based ordering, and the RA-MOPD advantage over GPT-5.3 is reproduced (+0.08, 95% CI [+0.031, +0.129]). **(2) BERTScore (Zhang et al., 2019)** vs. gold essays (bert-base-chinese): RA-MOPD achieves the highest reference-similarity score (0.679) vs. GPT-5.3 (0.641), providing a signal orthogonal to rubric compliance, though BERTScore primarily measures lexical-semantic overlap rather than pedagogical quality. **(3) Human study (Appendix Q):** six independent teachers, blinded to system identity, 1,200 essays, cell-level Pearson $r=0.87$ with RM scores. Table 5 summarises all four protocols; gold essays score higher under the holistic judge (3.71 vs. RM 3.48), analysed in §5.5.

5.5 Register Authenticity vs. Rubric Compliance

Gold essays score 3.48 on RM metrics—below trained systems—not from miscalibration but from a measurement difference. A separate three-teacher panel (disjoint from all other pools) rated 200 essays per system on *child register authenticity* (CRS: vocabulary, sentence structure, child voice; 1–5, no rubric anchors). Table 6 reports results.

System	RM	GPT-4.1	BERTSc.	Human
Gold (human)	3.48	3.71	—	—
SFT only	3.78	3.62	0.623	3.76
GPT-5.3 3-shot	4.12	3.96	0.641	4.09
RA-MOPD	4.21	4.04	0.679	4.18

Table 5: Four-protocol evaluation summary. RM: 8-dim mean RM score; GPT-4.1: holistic judge; BERTSc.: BERTScore-F₁ vs. gold essays (semantic overlap signal); Human: H8-Avg from Appendix Q. Among the three generated systems, rankings are consistent across all four protocols.

System	RM 8-Avg (Rubric)	CRS (Register)
Gold (human essays)	3.48	4.28
GPT-5.3 3-shot	4.12	2.91
SFT only	3.78	3.62
Base MOPD	4.09	3.70
RA-MOPD	4.21	3.74

Table 6: Rubric compliance (RM 8-Avg) vs. child register authenticity (CRS, teacher-rated, no rubric anchors). Gold essays rank highest on CRS; trained models optimise rubric compliance. RA-MOPD achieves the best rubric score among trained systems *and* the closest register to gold among LLM-generated essays.

A rubric–register plane. Reading Table 6 as a two-objective space—rubric compliance (RM 8-Avg) on the horizontal axis, child register authenticity (CRS) on the vertical axis—the systems occupy qualitatively distinct positions, visualised in Figure 2. **Gold** (3.48/4.28) is the register *ceiling* but only mid-tier on rubric: *authentic but not rubric-compliant*. **GPT-5.3** (4.12/2.91) trades register for rubric, collapsing to the *lowest* authenticity of any system—below even plain SFT: *compliant but not childlike*. **SFT** (3.78/3.62) is middling on both axes. **Base MOPD** (4.09/3.70) already lifts both coordinates well above SFT, and **RA-MOPD** (4.21/3.74) pushes both further still, attaining the highest rubric score of all systems while keeping its register the closest to gold among all LLM-generated essays.

Pareto dominance among deployable systems. The crucial consequence is that, for LLM-generated (i.e., deployable) essays, raising rubric compliance *ordinarily costs* child register—GPT-5.3 is the textbook case, buying a +0.34 rubric advantage over SFT at a −0.71 register penalty. RA-MOPD is the only system that escapes this trade-off: it simultaneously achieves the best rubric score *and* the gold-closest register, and therefore *Pareto-dominates* every other deployable system on

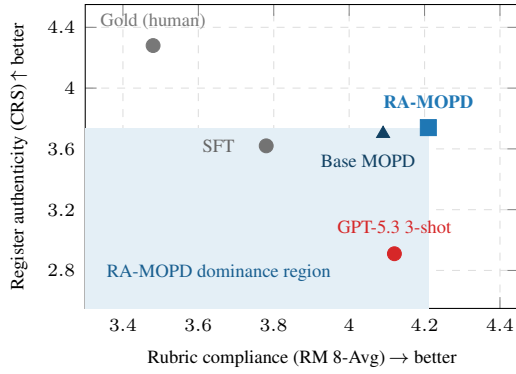


Figure 2: The rubric-register plane (data from Table 6). The shaded lower-left rectangle is the region Pareto-dominated by RA-MOPD (blue square); every deployable competitor—GPT-5.3, SFT, and Base MOPD—falls inside it, while none lies in the upper-right quadrant that would dominate RA-MOPD. GPT-5.3 buys rubric compliance at the cost of the lowest child register of any system; gold essays are the register ceiling but are human-written and not a deployable generator.

the joint (rubric, register) objective—it is strictly better on both coordinates than GPT-5.3 (+0.09 rubric, +0.83 register), SFT (+0.43, +0.12), and Base MOPD (+0.12, +0.04). Gold dominates RA-MOPD on register but is human-written and hence not a deployable generator; among systems that can actually produce new essays, all fall inside the shaded region RA-MOPD dominates and none lies in the upper-right quadrant that would dominate it (Figure 2). For a pedagogical model-essay writer—whose output must *both* satisfy the scoring rubric *and* read like a child wrote it—this joint optimum, rather than either coordinate in isolation, is precisely what makes RA-MOPD useful in the classroom.

RM metrics measure rubric compliance; CRS measures authenticity. RA-MOPD outperforms GPT-5.3 on *both*: highest rubric score (4.21) and closest register to gold (3.74 vs. GPT-5.3 2.91), making it preferable for pedagogical applications requiring quality *and* authenticity.

6 Conclusion

RA-MOPD augments on-policy reverse-KL MOPD with two task-tailored components—*RW-KL* (per-token RM importance) and *CGM* (a parameter-free transfer-matrix mixing prior), each contributing +0.06. The unified 4B model reaches **4.21/5.00** and **97.8%** routing on 1,200 held-out prompts: within 0.09 of GPT-5.3 3-shot ($p < 0.01$), with no statistically significant differ-

ence from a $4\times$ -cost four-expert logit-averaging baseline ($\Delta = +0.03$, $p = 0.18$). A GPT-4.1 judge, BERTScore, and a six-teacher human study ($r = 0.87$) reproduce the ranking, and a register analysis (§5.5) shows RA-MOPD is the only deployable system to Pareto-dominate the rest on the joint (rubric, register) objective. GPT-5.5 holistic evaluation also reproduces the component-level ordering, with RW-KL and CGM individually improving over Base MOPD and their combination performing best. The RA-MOPD gain remains stable on Qwen3-1.7B without model-specific retuning. Scope, limitations, and the asymmetric frontier comparison are discussed in Appendix A.

Limitations

Scope. CHILD-ALIGN covers four genres in Shanghai grades 3–6, with grade 6 comprising 10.1% and concentrated in Applied writing. The results establish performance on this benchmark, but not cross-regional, cross-curriculum, unseen-rubric, or unseen-genre generalisation; gains may partly reflect local annotation conventions. **Evaluation.** RM 8-Avg is complemented by holistic, semantic-similarity, human, and child-register evaluations (§5.4), which yield consistent system-level ordering. **Method.** RW-KL attribution is coarsest for Imagination Novelty ($\tau = 0.52$; Appendix K). RA-MOPD adds $\sim 35\%$ wall-clock over Base MOPD. No significant difference is detected from the $4\times$ -cost four-expert baseline ($\Delta = +0.03$, $p = 0.18$), which does not establish equivalence.

Ethics Statement

The study received institutional ethics approval; schools authorized essay use, and guardians provided informed consent. Identifying metadata was removed before OCR, followed by LLM-based PII redaction and manual checks (Appendix D). Released RMs and RA-MOPD are for teacher-facing support only. Teachers should review and adapt outputs; neither models nor RM scores should replace human instruction or be used for high-stakes assessment without human review.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated

- mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. [Constitutional ai: Harmlessness from ai feedback](#). *Preprint*, arXiv:2212.08073.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.
- Fan Gao, Dongyuan Li, Ding Xia, Fei Mi, Yasheng Wang, Lifeng Shang, and Baojun Wang. 2026. Essaybench: Evaluating large language models in multi-genre chinese essay writing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37396–37406.
- Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. Iflyea: A chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.
- Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.
- Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 151–162.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, and 22 others. 2022. [Multitask prompted training enables zero-shot task generalization](#). *Preprint*, arXiv:2110.08207.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.

- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.
- Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, and 107 others. 2026. [Mimo-v2-flash technical report](#). *Preprint*, arXiv:2601.02780.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. [Zephyr: Direct distillation of lm alignment](#). *Preprint*, arXiv:2310.16944.
- Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.
- Masaki Uto and Masashi Okano. 2020. Robust neural automated essay scoring using item response theory. In *International conference on artificial intelligence in education*, pages 549–561. Springer.
- Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, and 1 others. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian, and Minlie Huang. 2022. Automatic comment generation for chinese student narrative essays. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 214–223.

A Discussion

Why a 4B specialist is competitive with frontier models on this task. Four factors compound: (i) *Task specificity*—Chinese elementary writing is narrow and rubric-driven, giving targeted training high marginal value over general-purpose pretraining; (ii) *Distribution mismatch*—frontier RLHF data lacks rubric preference pairs for Chinese primary-school writing, so frontier models do not receive optimisation pressure on these specific dimensions; (iii) *Signal accumulation*—SFT, GRPO, and RA-MOPD each contribute a qualitatively distinct gradient that compound across stages; (iv) *Format compliance*—frontier LLMs emit the `Genre: / Essay:` format inconsistently even at 3-shot, directly penalising their routing accuracy.

Scope of the frontier comparison. The RA-MOPD vs. GPT-5.3 comparison is *structurally asymmetric*: RA-MOPD undergoes a four-stage rubric-aligned pipeline specifically on Shanghai elementary-school essays, while GPT-5.3 is evaluated with 3-shot demonstrations and no domain-specific fine-tuning whatsoever. The appropriate interpretation is that *domain-targeted post-training provides substantial marginal value on this narrow rubric corpus relative to strong inference-only baselines*, rather than that a 4B model is categorically superior to GPT-5.3 across all tasks. At the reported precision, GPT-5.3 ties RA-MOPD on Imagination Logic (4.20) and Novelty (4.10); its lower child register score (CRS 2.91 vs. our 3.74) illustrates that the two systems have complementary strengths arising from different training objectives.

Extended limitations. Geographic and curricular scope. The benchmark covers Shanghai grades 3–6 under a single municipal curriculum standard. The single-city annotation convention may not transfer to other regions of China; cross-regional generalisation is not evaluated. It also remains an open question whether the performance gains reflect genuine compositional writing ability or optimisation of annotation-convention-aligned surface features, since no out-of-distribution rubric or unseen-genre transfer test is performed. **Evaluation breadth.** The task-specific RM scores are complemented by holistic, semantic-similarity, human, and child-register evaluations (§5.4), which consistently reproduce the system ranking from different perspectives. **RW-**

KL attribution. Imagination Novelty is the weakest dimension ($\tau=0.52$; Appendix K), owing to the diffuse distribution of novelty-bearing tokens and the length–content conflation in chunk substitution. Comparison with gradient- or attention-based attribution methods (e.g., integrated gradients (Sundararajan et al., 2017)) is deferred to future work. **Baseline scope.** Parameter-efficient routing approaches (e.g., Mixture-of-LoRA) address a different goal—maintaining modular per-genre experts at inference—rather than our goal of a single dense deployable model. Applying DPO or ORPO per-genre and then distilling via RA-MOPD is a natural extension left for future work. **Four-expert baseline.** No statistically significant difference is detected between RA-MOPD and the $4\times$ -cost ground-truth-routed four-expert logit-averaging baseline ($\Delta=+0.03$, $p=0.18$). This result does not establish equivalence; a larger-scale study is needed for greater statistical power.

B Related Work

Automated essay scoring (AES). Classical AES targets standardised English tests (Taghipour and Ng, 2016; Uto, 2021; Uto and Okano, 2020) and optimises correlation with a single holistic score. Early neural approaches established sequence models as strong baselines (Yannakoudakis et al., 2011), while more recent work shows that pre-trained transformers substantially improve agreement with human raters (Mayfield and Black, 2020). Extensions to multi-trait scoring and cross-prompt generalisation (Ridley et al., 2021) establish the value of dimension-specific scoring, which our eight RMs directly instantiate. Explainable AES (Kumar and Boulanger, 2020) connects rubric criteria to model internals—a spirit shared by RW-KL, which uses RM perturbation-based attributions as an oracle inside the distillation loop rather than for post-hoc explanation. Chinese AES includes both feature-engineered and LLM-based (Gong et al., 2021) lines. EssayBench (Gao et al., 2026) benchmarks frontier LLMs on multi-genre Chinese *evaluation*; our work differs in (a) focusing on *generation* rather than evaluation, (b) incorporating four training stages, and (c) using student-written essays as gold references.

Genre- and style-conditioned generation. Style and genre control have been studied via control tokens (Keskar et al., 2019), instruction prompts (Sanh et al., 2022). For Chinese student

writing, Zhang et al. (2022) generate teacher feedback. Most prior work targets classification or comment generation rather than rubric-aligned generation.

Reinforcement learning and preference optimisation for LLMs. RLHF (Ouyang et al., 2022; Stiennon et al., 2020) established on-policy reward optimisation as the dominant post-training paradigm, using PPO (Schulman et al., 2017) as the policy gradient backbone. GRPO (Shao et al., 2024) simplifies PPO by replacing the learned value function with group-normalised advantage estimates. Offline alternatives—DPO (Rafailov et al., 2023) and ORPO (Hong et al., 2024)—avoid on-policy rollouts at the cost of exposure bias; Appendix M shows that both fall short of the best per-genre GRPO specialist (3.91) by 0.05–0.08 on 8-Avg, with DPO (3.86) marginally exceeding Mix-GRPO (3.84). RAFT (Dong et al., 2023) proposes reward-ranked fine-tuning, a supervised analogue of online RL that selects high-reward generations for SFT; while simpler than GRPO, it lacks the gradient-level reward signal needed for fine-grained rubric compliance. Constitutional AI (Bai et al., 2022) demonstrates that LLM-generated feedback can substitute for human preference labels, motivating our use of RM-derived pairs for the DPO/ORPO baselines.

From offline KL to rubric-aware on-policy distillation. Sequence-level distillation (Kim and Rush, 2016) and logit-level KL distillation (Sanh et al., 2019; Hinton et al., 2015) remain canonical offline baselines; our distillation protocol ablation (Appendix N) shows they underperform on-line reverse-KL by -0.16 on 8-Avg in our setting. Zephyr (Tunstall et al., 2023) demonstrates that distillation of preference-aligned teacher outputs into a smaller student can match or exceed RLHF at lower cost—a complementary finding to our online reverse-KL approach. On-policy distillation with the mode-seeking reverse-KL direction was introduced by MiniLLM (Gu et al., 2024) and Agarwal et al. (2024), then adopted by DeepSeek-V4 (Xu et al., 2026) and MiMo-V2 (Team et al., 2026) as the standard post-training distillation step. Multi-expert distillation has been studied as self-paced selection. RA-MOPD adds two extensions specifically for rubric-guided generation. **RW-KL** uses an external rubric-aligned RM as a per-token perturbation-based importance oracle inside the KL term, distinct from prior work where external re-

wards modulate the loss only at the *sequence* level. **CGM** replaces single-teacher distillation with a fixed top-2 mixture whose weights are derived from an empirical task-transfer matrix, differing from MoE (Shazeer et al., 2017; Fedus et al., 2022) routing in being a training-time-only, parameter-free mechanism.

C Three-Annotator RM Validation

Protocol. To rule out self-consistency bias and provide a cleaner human reference, we ran a separate three-annotator validation on a freshly sampled set of 100 essays per dimension (800 essays total), drawn from the 2,268 cleaned essays that were not selected for the 6,000-essay benchmark. The panel is therefore disjoint from the benchmark train/validation/test splits and from all essays used to train or evaluate the RMs. For each dimension we recruited three new graduate-student annotators with formal language-pedagogy training (disjoint from the dataset-annotation team in §3 and from the human-study panel in Appendix Q), provided them with the same 1–5 rubric anchors (Appendix G), and asked each to score every essay independently. The human reference for each essay is the rounded mean of the three independent scores.

Results. Table 7 reports results across all eight dimensions. The RM-vs-three-rater Pearson correlation averages 0.821 (QWK 0.816) and the interrater Pearson among the three new annotators averages 0.77, indicating that RM quality is stable under a re-annotation by a disjoint group of qualified annotators.

D Dataset: Cleaning, Audit, and IAA Details

Source. The 10,746 raw essays originate from public elementary schools in Shanghai, written by students in grades 3–6 (see Table 8 for the grade $\times$ genre distribution of the final 6,000-essay benchmark). Grade 4 is the modal year (33.3%), with grades 3 and 5 together covering another $\sim 56.6\%$; grade 6 constitutes 10.1% of the benchmark, concentrated in the Applied genre. Identifying meta-data (student name, school name, grade section) was stripped prior to OCR.

Annotator profile. The three annotators in §3 are graduate-student members of our research laboratory in Chinese-language education or linguistics with formal language-pedagogy training; two are

Reward model	$r_{\text{RM-H}}$	$\text{QWK}_{\text{RM-H}}$	r_{IRR}	κ_{IRR}
App. Format & Context	.832	.824	.81	.77
App. Suggestion & Emotion	.795	.790	.75	.71
Imag. Logical Coherence	.848	.844	.80	.77
Imag. Novelty	.906	.904	.70	.65
Narr. Detail & Emotion	.806	.801	.78	.74
Narr. Narration Compl.	.815	.810	.82	.79
Obj. Feature Accuracy	.780	.775	.76	.72
Obj. Multisense & Rhet.	.788	.780	.77	.73
Average	.821	.816	.77	.74

Table 7: Three-annotator validation of the eight RMs on a re-sampled 100-essay-per-dimension panel drawn from cleaned essays outside the 6,000-essay benchmark. $r_{\text{RM-H}}$ / $\text{QWK}_{\text{RM-H}}$: RM-vs-three-rater-mean correlations; r_{IRR} / κ_{IRR} : inter-rater agreements among the three new annotators. RM-human QWK exceeds inter-rater κ on every dimension; RM-human Pearson r exceeds inter-rater r on seven of eight dimensions (exception: Narration Completeness, $r_{\text{RM-H}}=.815 < r_{\text{IRR}}=.82$, though $\text{QWK}_{\text{RM-H}}=.810 > \kappa_{\text{IRR}}=.79$).

Genre	G3	G4	G5	G6	Total
Narrative	415	680	355	50	1,500
Object	445	770	285	0	1,500
Imagination	462	510	503	25	1,500
Applied	188	40	741	531	1,500
Total	1,510	2,000	1,884	606	6,000
(%)	25.2	33.3	31.4	10.1	100

Table 8: Grade $\times$ genre distribution of the balanced 6,000-essay benchmark. Grade 4 is the modal year; grade 6 constitutes 10.1% of the benchmark, predominantly in the Applied genre, reflecting the letter-writing emphasis in the upper-elementary curriculum of the source corpus.

M.A. candidates and the most senior—who adjudicated ≥ 2 -point disagreements—is a Ph.D. candidate. All completed written training on the rubric anchors (Appendix G) before scoring began and were paid at the prevailing local rate for graduate-level annotation work.

Cleaning specifics. *Length filter*: discard essays outside [100, 1500] characters (3.7% pruned). *OCR-noise removal*: regex-strip tab characters, collapse whitespace runs, remove control symbols (12.4% cleaned; 0.9% unusable, dropped). *Privacy redaction*: a DeepSeek-V4-flash agent replaces real names and concrete addresses with ***; 6,000 unique essays processed, 25.5% containing at least one PII span. *Deduplication*: exact duplicates keyed by (prompt, essay) are dropped.

PII residual audit. Two authors manually inspected a uniform random 5% sample ($n=300$) of the redacted essays and identified three residual PII spans (1.0% point estimate; 95% exact binomial upper bound $\approx 2.9\%$).

Splits. Each genre is split 60/20/20 into train/val/test under a strictly *prompt-disjoint* partition (deterministic hash, seed 42).

Per-dimension inter-annotator agreement. Table 9 reports IAA per dimension.

Dimension	Pearson r	Weighted κ
App. Format & Context	0.86	0.83
App. Suggestion & Emotion	0.78	0.74
Imag. Logical Coherence	0.81	0.78
Imag. Novelty	0.72	0.68
Narr. Detail & Emotion	0.80	0.76
Narr. Narration Compl.	0.84	0.81
Obj. Feature Accuracy	0.79	0.75
Obj. Multisense & Rhet.	0.77	0.73
Average	0.80	0.76

Table 9: Per-dimension inter-annotator agreement between the two primary annotators (before senior adjudication).

E Training Details

Stage 3 (GRPO). We use the `trl` GRPOTrainer with vLLM colocate (Kwon et al., 2023); learning rate 2×10^{-6} , $\beta=0.01$, $K=8$ rollouts per prompt, maximum completion length 1024, per-device batch size 2, gradient accumulation 16, 2 epochs over 900 prompts. The eight RMs are queried as HTTP services on ports 50011–50018 (one vLLM instance each) with asyncio concurrency 32. Each genre specialist trains in approximately 3–5 hours on a single H800 GPU.

Stage 4 (Base MOPD and RA-MOPD). Genre-homogeneous batching means within each batch only the matched expert π_{ϕ_g} (Base MOPD) or the matched expert plus its top-2 partner (RA-MOPD) are needed. When the genre changes, experts are swapped lazily (~ 30 s per swap for Base MOPD, ~ 50 s for RA-MOPD). A 2-epoch run on 900 prompts incurs ~ 120 swaps in total. Pure on-policy reverse-KL with no CE stream; rollout length 512, sampling temperature 0.7, top- p 0.95, distillation temperature $T=1.0$; learning rate 2×10^{-6} ; per-device batch size 2; bf16. Total training time on $4 \times \text{H800}$: Base MOPD $\sim 11\text{h}$, RA-MOPD $\sim 15\text{h}$.

F Algorithms

GRPO loss (referenced from §4.3).

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E} \left[\sum_k \hat{A}^{(k)} \sum_{t \in y^{(k)}} \log \rho_t^{(k)} \right] + \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}), \quad (17)$$

with $\rho_t^{(k)} = \pi_\theta(z_t^{(k)}) / \pi_{\theta_{\text{old}}}(z_t^{(k)})$.

Algorithm 1 Extract target length L^* from a prompt x .

```

1: function EXTRACTLEN( $x$ )
2:    $\mathcal{P} \leftarrow$  regex patterns for common Chinese
   length cues
3:   for  $p \in \mathcal{P}$  do
4:      $m \leftarrow$  first regex match of  $p$  in  $x$ 
5:     if  $m \neq \emptyset$  then return integer of the
       first capture group
6:   end if
7: end for
8:   return NONE
9: end function

```

Algorithm 2 One Base-MOPD optimisation step (pure on-policy reverse-KL, no CE).

Require: Student π_θ ; experts $\{\pi_{\phi_g}\}$; genre-homogeneous prompt batch $\mathcal{B}$ of target genre g .

```

1: ensure matched expert  $\pi_{\phi_g}$  is loaded
2: for all  $x \in \mathcal{B}$  do
3:    $\hat{y} \leftarrow \pi_\theta.\text{GENERATE}(x) \triangleright$  on-policy rollout
4:    $M \leftarrow$  student-generated token positions
5: end for
6:  $\{\ell_\theta^{(t)}\} \leftarrow \pi_\theta.\text{FORWARD}((x, \hat{y}))$ 
7:  $\{\ell_{\phi_g}^{(t)}\} \leftarrow \pi_{\phi_g}.\text{FORWARD}((x, \hat{y})) \triangleright$  no gradient
8:  $P_\theta^{(t)} \leftarrow \text{softmax}(\ell_\theta^{(t)}); Q_g^{(t)} \leftarrow \text{softmax}(\ell_{\phi_g}^{(t)})$ 
9:  $\ell_{\text{total}} \leftarrow \frac{1}{|M|} \sum_{t \in M} \sum_v P_\theta^{(t)}(v) (\log P_\theta^{(t)}(v) - \log Q_g^{(t)}(v)) \triangleright$  reverse KL,  $D_{\text{KL}}(P_\theta \parallel Q_g)$ 
10: backward(); optimizer.step()

```

G Full Rubric Scoring Anchors

Tables 10–13 present the complete 1–5 scoring anchors for all eight rubric dimensions, in both Chinese (original) and English (translation).

Algorithm 3 One RA-MOPD optimisation step (RW-KL + CGM extensions; pure reverse-KL, no CE).

Require: Student π_θ ; experts $\{\pi_{\phi_g}\}$; RMs ($\text{RM}_{d_a}, \text{RM}_{d_b}$); chunk count K ; RW-KL sharpness λ ; CGM mixing weights $\{\pi_{g'|g}\}$; genre-homogeneous batch $\mathcal{B}$ of target genre g .

```

1: ensure  $\pi_{\phi_g}$  and  $\pi_{\phi_{g_2}}$  (top-2 partner) are loaded
2: for all  $x \in \mathcal{B}$  do
3:    $\hat{y} \leftarrow \pi_\theta.\text{GENERATE}(x)$ 
4:   Split  $\hat{y}$  into  $K$  chunks;  $M \leftarrow$  rollout positions
5:   (every 5 steps) compute attribution  $\{\hat{a}_g(t)\}$  via Eqs. 10–11
6:    $w_t \leftarrow 1 + \lambda \hat{a}_g(t)$ 
7: end for
8:  $\ell_\theta \leftarrow \pi_\theta.\text{FORWARD}((x, \hat{y}))$ 
9:  $\ell_{\phi_g} \leftarrow \pi_{\phi_g}.\text{FORWARD}((x, \hat{y})) \triangleright$  no gradient
10:  $\ell_{\phi_{g_2}} \leftarrow \pi_{\phi_{g_2}}.\text{FORWARD}((x, \hat{y})) \triangleright$  no gradient
11:  $P_\theta^{(t)} \leftarrow \text{softmax}(\ell_\theta^{(t)})$ 
12:  $Q_{\text{mix}}^{(t)} \leftarrow \pi_{g|g} \text{softmax}(\ell_{\phi_g}^{(t)}) + \pi_{g_2|g} \text{softmax}(\ell_{\phi_{g_2}}^{(t)})$ 
13:  $\ell_{\text{total}} \leftarrow \frac{\sum_t w_t \sum_v P_\theta^{(t)}(v) (\log P_\theta^{(t)}(v) - \log Q_{\text{mix}}^{(t)}(v))}{\sum_t w_t} \triangleright$  rubric-weighted reverse KL
14: backward(); optimizer.step()

```

H Closed-LLM Prompt Templates

System prompt. “You are a primary-school Chinese teacher writing on behalf of a Chinese elementary student (grades 3–5). Begin your response with *Genre*: on one line, naming the genre of the prompt exactly as one of applied / imagination / narrative / object. On the next line begin the essay with *Essay*: and write a 300–500-character Chinese essay matching the prompt. Do not output anything else.”

Zero-shot user message. Topic: $\langle \text{prompt} \rangle$

Three-shot user message. Three same-genre demonstration essays in the *Genre*:/*Essay*: format, followed by the target Topic: line. Demonstrations are sampled deterministically (seed 42) from the Stage 1 gold pool.

Output parsing. $\tilde{g}(y)$ is the first non-empty substring following the first occurrence of *Genre*:; the essay body is everything following the first occurrence of *Essay*:. Routing accuracy uses

Score	Chinese original	English translation
<i>Dimension 1: Format & Context (格式与情境)</i>		
5	格式分毫不差，语气极度契合特定身份与情境。	Format is impeccable; tone perfectly fits the specified role and context.
4	格式规范，要素齐全，语言得体。	Format is standard with all required elements; language is appropriate.
3	基本格式正确，但存在小瑕疵。	Basic format is correct; minor issues in salutation, closing, or tone.
2	核心格式要素缺失，语气与目标读者不匹配。	Core format elements missing; tone does not match audience.
1	完全不懂应用文格式。	No awareness of applied-genre format.
<i>Dimension 2: Suggestion & Emotion (建议与情感)</i>		
5	建议极其切实可行，情感真挚感人。	Highly actionable suggestions; sincere and moving emotion.
4	能提出具体的建议，表达清晰。	Concrete suggestions; clear and earnest expression.
3	有基本建议或态度，但较为套路化。	Basic suggestions or attitude; content is formulaic.
2	建议空洞，缺乏实际意义。	Hollow suggestions; indifferent attitude.
1	无建议内容，无情感交流。	No suggestions or emotional engagement.

Table 10: Full scoring rubric for the **Applied** genre.

fuzzy substring matching to avoid penalising near-synonyms emitted by closed LLMs.

I Cross-Genre Transfer Heatmap

Figure 3 presents the full 4×4 cross-genre transfer matrix for the four per-genre GRPO specialists, computed on the first 150 prompts of the *validation* split per genre (not the test set; the remaining 150 validation prompts per genre are reserved for τ selection in CGM, cf. §4.4.4). Each cell reports the two-dimensional mean RM score when a given column expert is evaluated on the row genre. Diagonal cells (specialty) are strongest; off-diagonal cells remain well above the base Qwen3-4B 8-Avg of 3.30, confirming the absence of catastrophic forgetting.

		Expert (column)			
		App	Imag	Narr	Obj
Test genre (row)	App	4.05	3.76	3.76	3.75
	Imag	3.57	4.05	3.51	3.56
	Narr	3.84	3.86	4.20	3.82
	Obj	3.96	3.96	3.98	4.18

Figure 3: 4×4 cross-genre transfer heatmap on the first 150 prompts of the validation split (per genre). Diagonal cells (specialty) are strongest; off-diagonal cells remain well above the base model. This matrix is the source of CGM mixing weights; the remaining 150 validation prompts per genre are used independently for τ selection (§4.4.4).

J RW-KL Attribution Sanity Check

To verify that the chunk-level attributions from Eq. 10 align with human linguistic intuition, we ran the attribution on a 100-prompt validation subset (25 per genre) and manually inspected the top-2 most attributed chunks per essay. For each genre’s two rubric dimensions, we counted whether the top-attributed chunks contained the linguistic features expected by that dimension. Results (Table 14) show that the attribution consistently identifies the rubric-relevant linguistic units, providing qualitative support for the RW-KL design.

K RW-KL Robustness Analysis

Chunk count K . Table 15 sweeps K from 3 to 50 (RW-KL only; CGM disabled). Performance peaks at $K=10$ and remains within ± 0.02 for $K \in [5, 20]$.

Placeholder type. Table 16 compares four placeholder strategies for $\hat{y}^{c_i}$ in Eq. 10. Generic-text placeholders perform best; all strategies substantially outperform Base MOPD, indicating robustness to this design choice overall.

Attribution consistency (Kendall τ). Table 17 reports Kendall’s τ between chunk-importance

Score	Chinese original	English translation
<i>Dimension 3: Logical Coherence</i> (条理逻辑性)		
5	结构严谨，逻辑自洽，过渡极其自然。	Rigorous structure; smooth transitions.
4	结构完整，脉络分明。	Complete structure; clear logical flow.
3	结构基本完整，过渡稍显生硬。	Basically intact; some stiff transitions.
2	条理不清，内容散乱。	Disorganised; scattered content.
1	完全没有结构意识。	No structural awareness whatsoever.
<i>Dimension 4: Novelty</i> (构思新颖度)		
5	设定独特，完全打破常规思维。	Highly original; breaks conventional thinking.
4	想象丰富，包含多种新颖元素。	Rich imagination; multiple novel elements.
3	有想象元素，但属于常见场景。	Imaginative but relies on common tropes.
2	想象乏味，多为模仿已知作品。	Bland; largely imitative.
1	基本没有想象内容。	Virtually no imagination.

Table 11: Full scoring rubric for the **Imagination** genre.

Score	Chinese original	English translation
<i>Dimension 5: Detail & Emotion</i> (细节与情感)		
5	动作、神态、心理描写极其细腻。	Fine-grained action, expression, and psychological depiction.
4	有较多生动细节。	Multiple vivid details.
3	包含简单描写，但不够生动。	Simple depictions; not vivid.
2	缺乏细节，基本是流水账。	Sparse details; plain sequence.
1	完全没有细节描写。	No detail; hollow language.
<i>Dimension 6: Narration Completeness</i> (叙事完整性)		
5	六要素完备，情节自然引人入胜。	All six elements complete; natural plot arc.
4	要素齐全，顺序清晰。	Elements present; clear order.
3	扣题且大致完整，但较为简略。	On-topic but sparse.
2	事件交代不清，逻辑混乱。	Poorly recounted.
1	字数严重不足，仅有片段。	Severely short; fragments only.

Table 12: Full scoring rubric for the **Narrative** genre.

rankings produced by three independent rollouts. Average $\tau=0.74$ across seven non-Novelty dimensions; Imagination Novelty is lowest ($\tau=0.52$), consistent with the diffuse nature of novelty-bearing tokens and the length–content conflation in chunk substitution.

L Reward Weight Ablation

Table 18 compares three principled choices of the GRPO reward weights (w_g, w_a, w_b, w_ℓ) on the validation set. The chosen (0.2, 0.3, 0.3, 0.2) outperforms both the uniform baseline and the rubric-heavy alternative by at least +0.04 on validation 8-Avg.

M DPO and ORPO Preference-Optimisation Baselines

Motivation. Since our Stage-2 RMs provide per-essay scalar scores, a natural alternative to GRPO is to convert those scores into pairwise preferences and apply offline preference optimisation such as DPO (Rafailov et al., 2023) or ORPO (Hong et al., 2024). We evaluate both as multi-genre baselines initialised from STAGE1-SFT, directly comparable to Mix-GRPO.

Pair construction. For each training prompt x we generate $K=8$ rollouts from STAGE1-SFT and score every rollout with the two genre-matched RMs to obtain a composite score $s^{(k)} = \frac{1}{2}(\text{RM}_{d_a} + \text{RM}_{d_b})$. We retain the highest-scoring rollout as the *chosen* response y_w and the lowest-scoring as the *rejected* response y_l , yielding one pair per

Score	Chinese original	English translation
<i>Dimension 7: Feature Accuracy (特征准确性)</i>		
5	抓取特征精准独特。	Highly precise and distinctive features.
4	外貌、习性、结构等描写准确。	Accurate description of features.
3	写出基本特征，但较为大众化。	Basic features; generic description.
2	特征描写模糊。	Vague; observational biases.
1	描写对象不明确。	Subject unclear.
<i>Dimension 8: Multisense & Rhetoric (多感官与修辞)</i>		
5	完美融合多感官，修辞出神入化。	Perfect multisensory integration; masterful rhetoric.
4	运用两种以上感官，修辞恰当。	Two or more sensory modalities; appropriate rhetoric.
3	感官较单一，修辞平淡。	One-dimensional; plain rhetoric.
2	极少使用修辞。	Rarely uses rhetoric.
1	完全没有感官参与或修辞。	No sensory engagement or rhetoric.

Table 13: Full scoring rubric for the **Object Description** genre.

Dimension	Expected feature	Hit rate
App. Format	salutation / closing	78%
App. Suggestion	“建议/希望” clauses	72%
Imag. Logic	causal connectives	66%
Imag. Novelty	novel-setting noun	54%
Narr. Detail	action / sense verb	74%
Narr. Completeness	climax / resolution	71%
Obj. Feature	defining attribute	73%
Obj. Multisense	sensory adjective	69%

Table 14: Fraction of top-2 attributed chunks containing the rubric-relevant linguistic feature on a 100-essay validation subset. Novelty is the lowest, consistent with its diffuse token distribution and the length–content conflation noted in §4.4.3.

K	8-Avg
3	4.11
5	4.13
10 (ours)	4.15
20	4.14
50	4.12

Table 15: RW-KL sensitivity to chunk count K (CGM disabled; Base MOPD baseline = 4.09).

prompt (900 pairs in total). This offline construction avoids on-policy rollouts during DPO/ORPO training but sacrifices the adaptive exploration that makes GRPO sample-efficient.

Training details. Both DPO and ORPO are trained multi-genre (all four genres in a single model) from STAGE1-SFT, matching the Mix-GRPO setup. DPO uses $\beta=0.1$ and the same learning rate and batch configuration as Stage 3 (Appendix E); ORPO uses $\lambda_{\text{odds}}=0.1$ per Hong et al.

Placeholder strategy	8-Avg
Generic text (ours)	4.15
Empty string	4.13
Length-matched spaces	4.13
Random character sequence	4.11

Table 16: RW-KL sensitivity to placeholder type.

Dimension	Avg Kendall τ	Hit rate
App. Format	0.81	78%
App. Suggestion	0.74	72%
Imag. Logic	0.69	66%
Imag. Novelty	0.52	54%
Narr. Detail	0.77	74%
Narr. Completeness	0.73	71%
Obj. Feature	0.76	73%
Obj. Multisense	0.71	69%
Average (excl. Novelty)	0.74	72%

Table 17: Attribution consistency (Kendall τ) across three independent rollouts and hit rate from Table 14. Imagination Novelty is consistently the most diffuse dimension.

(2024). Neither method uses on-policy rollouts.

Results.

Analysis. DPO and ORPO improve over SFT but underperform the best per-genre GRPO specialist for two reasons. First, *offline pair quality*: the chosen/rejected split is determined by a single composite RM score on STAGE1-SFT rollouts; these pairs do not cover the distribution of errors encountered during training, causing exposure bias. Second, *genre signal*: GRPO’s explicit R_{genre} term directly supervises the Genre : / Essay : prefix,

w_g	w_a	w_b	w_ℓ	8-Avg (val)
0.25	0.25	0.25	0.25	3.86
0.20	0.30	0.30	0.20	3.91
0.10	0.40	0.40	0.10	3.87
0.40	0.20	0.20	0.20	3.83
0.00	0.40	0.40	0.20	3.74

Table 18: Reward-weight ablation for GRPO.

Method	8-Avg	Δ SFT	Avg acc (%)
STAGE1-SFT (reference)	3.78	—	94.1
ORPO (RM pairs)	3.83	+0.05	91.5
DPO (RM pairs)	3.86	+0.08	92.3
Mix-GRPO	3.84	+0.06	91.0
Best per-genre GRPO	3.91	+0.13	90.8
RA-MOPD	4.21	+0.43	97.8

Table 19: DPO and ORPO preference-optimisation baselines vs. pipeline intermediates and RA-MOPD on the held-out benchmark. DPO (+0.08) and ORPO (+0.05) improve over SFT; DPO marginally exceeds Mix-GRPO (+0.06). Both fall short of the best per-genre GRPO specialist (3.91) by 0.05–0.08 and substantially below RA-MOPD (+0.43). Routing accuracy drops relative to SFT for both (ORPO 91.5%; DPO 92.3%), suggesting that offline preference pairs provide weaker genre-format signal than the on-policy routing reward in GRPO.

whereas DPO and ORPO rely on the implicit correlation between high RM score and correct genre formatting—a weaker signal, reflected in lower routing accuracy (91.5–92.3% vs. 94.1% for SFT alone). Applying DPO or ORPO per-genre (matching the GRPO setup) and then distilling with RA-MOPD is a natural extension; we leave this to future work.

N Distillation Protocol Ablation

To justify the choice of *online* policy distillation with a *reverse-KL* objective, we compare three distillation protocols that differ along two axes—online vs. offline and KL direction—while holding all other hyperparameters fixed (same student initialisation, same expert teachers, same training prompts and evaluation benchmark).

Protocols. (1) **Offline CE distillation** generates $K=8$ rollouts from each teacher expert *before* training begins, then trains the student via masked cross-entropy on those teacher-generated sequences (Kim and Rush, 2016). This is the standard sequence-level KD baseline and requires no on-policy rollouts at training time. (2) **Online forward-KL** follows the same on-policy rollout schedule as base MOPD but min-

imises $D_{\text{KL}}(Q_g^{(t)} \parallel P_\theta^{(t)})$ —the mean-seeking direction that spreads student probability mass across all teacher modes, including low-quality outputs (Gu et al., 2024). (3) **Online reverse-KL (base MOPD)** minimises $D_{\text{KL}}(P_\theta^{(t)} \parallel Q_g^{(t)})$ —the mode-seeking direction (Eq. 9) that concentrates the student on realisable high-quality regions of the teacher distribution. All three systems train on the same 900-prompt set for 2 epochs; results are on the held-out 1,200-essay test set.

Distillation protocol	8-Avg	Δ	Acc (%)
Offline CE (seq-level KD)	3.93	−0.16	94.8
Online forward-KL	4.02	−0.07	96.3
Online reverse-KL (base MOPD)	4.09	—	97.5

Table 20: Comparison of three distillation protocols on the held-out benchmark. *Offline CE*: sequence-level KD on pre-generated teacher rollouts (Kim and Rush, 2016); *Online forward-KL*: on-policy mean-seeking KL; *Online reverse-KL*: base MOPD (Eq. 9). Moving from offline to online distillation contributes +0.09 on 8-Avg; switching from forward to reverse KL contributes a further +0.07, consistent with Gu et al. (2024) and Agarwal et al. (2024).

Analysis. The offline CE baseline already outperforms Stage 1-SFT (3.78) by +0.15 on 8-Avg, confirming that even a single round of teacher distillation transfers rubric-relevant knowledge. The remaining +0.16 gap to online reverse-KL demonstrates two distinct benefits: (i) *on-policy rollouts* expose the student to its own error distribution rather than the teacher’s, preventing compounding exposure bias; and (ii) *reverse KL* prevents the student from being pulled toward low-probability teacher modes it cannot reliably realise. Routing accuracy follows the same trend (94.8 → 96.3 → 97.5%), indicating that on-policy training also benefits genre-format compliance. The forward-KL result is consistent with the hyperparameter analysis in Appendix O, where replacing reverse-KL with forward-KL reduces validation 8-Avg by −0.07.

O RA-MOPD Hyperparameter Ablation

We collect here the supporting ablation floats referenced from §5.2: the CGM teacher-mixing controls (Table 21), the multi-seed variance study (Table 22), and a per-dimension visualisation of four representative systems (Figure 4).

KL direction. Table 20 (Appendix N) provides a three-way comparison of distillation protocols

Mixing strategy	8-Avg	Δ Base
Base MOPD (single expert)	4.09	—
+ Random 2nd expert	4.10	+0.01
+ Uniform top-2 (0.50 / 0.50)	4.12	+0.03
+ Score-weighted top-2 (CGM)	4.15	+0.06
+ Learned gating (oracle)	4.16	+0.07

Table 21: CGM teacher-mixing controls (cf. §5.2). Random partner yields negligible gain; CGM’s score-weighted top-2 matches learned gating with no additional parameters.

System	s-42	s-123	s-456	Mean $\pm$ Std
Base MOPD	4.09	4.08	4.10	4.09 ± 0.010
+ RW-KL	4.15	4.14	4.16	4.15 ± 0.010
+ CGM	4.15	4.15	4.14	4.15 ± 0.006
RA-MOPD	4.21	4.20	4.22	4.21 ± 0.010
GPT-5.3 3-shot	4.12 (deterministic)			4.12

Table 22: 8-Avg across three random seeds (cf. §5.2). Each component’s +0.06 gain is stable; the RA-MOPD–GPT-5.3 gap of +0.09 exceeds 2σ .

including offline CE; here we report the isolated effect of KL direction within the online setting. Replacing the reverse KL $D_{\text{KL}}(P_\theta \parallel Q_g)$ with the forward KL $D_{\text{KL}}(Q_g \parallel P_\theta)$ drops test 8-Avg by -0.07 relative to base MOPD ($4.09 \rightarrow 4.02$), consistent with the findings of Gu et al. (2024) and Agarwal et al. (2024) that the mode-seeking direction is preferable for autoregressive language-model distillation.

RW-KL sharpness λ . Table 23 sweeps λ on the validation set. $\lambda=2$ is optimal; very large λ over-concentrates on a few tokens and slightly hurts.

λ (RW-KL)	8-Avg (val)
0 (base MOPD)	4.10
1	4.13
2	4.15
4	4.14
8	4.10

Table 23: RW-KL sharpness sweep, RW-KL only (no CGM).

CGM temperature τ . Table 24 sweeps τ in Eq. 14. $\tau \rightarrow 0$ recovers base MOPD; $\tau=0.2$ is optimal; very large τ flattens the mixture.

Concrete CGM mixing weights at $\tau=0.2$. Derived from the validation heatmap (Figure 3): Target App: App 0.81 / Imag 0.19; Target Imag: Imag 0.92 / App 0.08; Target Narr: Narr 0.84 / Imag 0.16; Target Obj: Obj 0.73 / Narr 0.27.

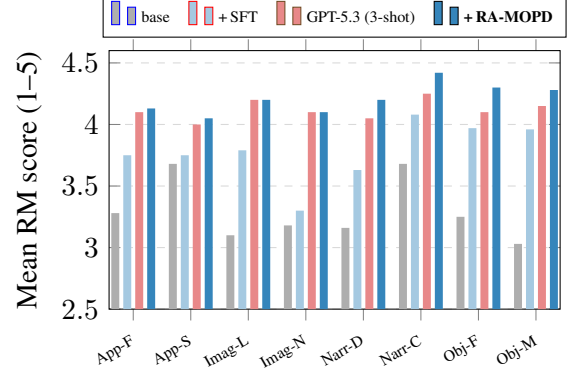


Figure 4: Per-dimension mean RM score on the 1,200-essay benchmark for four representative systems (cf. Table 3). RA-MOPD matches or outperforms GPT-5.3 3-shot on all eight dimensions, with six improvements and two ties at the reported precision.

τ (CGM)	8-Avg (val)
$\rightarrow 0$ (base MOPD)	4.10
0.1	4.13
0.2	4.15
0.5	4.12
1.0	4.07

Table 24: CGM temperature sweep, CGM only (no RW-KL).

Top- k mixing. Top-2 achieves 4.21 test 8-Avg; top-3 achieves 4.21; all-4 drops to 4.19. Top-3 does not help because the third-ranked expert’s weight is ≤ 0.1 .

Per-dimension RA-MOPD vs. best expert. Table 25 reports per-dimension margins. RA-MOPD strictly improves on all eight dimensions (+0.03 to +0.13).

P Mix-GRPO Reward Trajectory

Figure 5 plots the per-genre reward moving average during Mix-GRPO joint training. Applied-Format and Narrative-Completeness rewards anti-correlate from step ~ 100 , with gradient cosine analysis confirming destructive interference (within-genre $\cos \approx 0.71$; cross-genre $\cos \in [-0.18, -0.07]$).

Q Human Evaluation: Full Protocol and Results

Design. Three pairwise conditions plus one absolute-scoring track: (A) RA-MOPD vs. GPT-5.3 (3-shot); (B) RA-MOPD vs. Stage1-SFT; (C) RA-MOPD vs. matched expert. The absolute track covers $1,200 \times 4 = 4,800$ essays; the pairwise

Dimension	Best expert	RA-MOPD	Δ
App. Format	4.10 (App)	4.13	+0.03
App. Suggestion	4.00 (App)	4.05	+0.05
Imag. Logic	4.10 (Imag)	4.20	+0.10
Imag. Novelty	4.00 (Imag)	4.10	+0.10
Narr. Detail	4.10 (Narr)	4.20	+0.10
Narr. Completeness	4.30 (Narr)	4.42	+0.12
Obj. Feature	4.20 (Obj)	4.30	+0.10
Obj. Multisense	4.15 (Obj)	4.28	+0.13
8-Avg	3.91	4.21	+0.30

Table 25: RA-MOPD vs. the best per-dimension single specialist. *Best expert* values in each row are taken from the specialist that operates in its own specialty genre for that dimension; the 8-Avg row (3.91) is the best single model’s macro-average across all eight dimensions, and is *not* the arithmetic mean of the column above (which would yield ≈ 4.12 , a mixed-oracle figure).

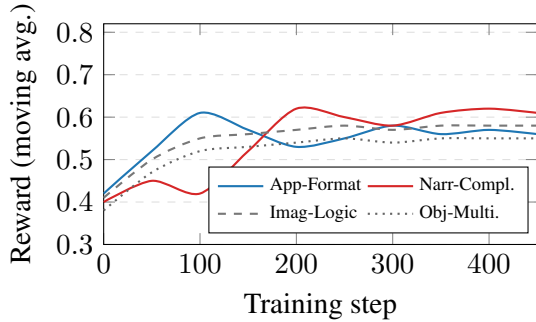


Figure 5: Per-genre reward trajectories during Mix-GRPO joint training. Applied-Format (blue) and Narrative-Completeness (red) anti-correlate from step ~ 100 .

track covers 3,600 judgements. Six primary-school Chinese teachers (4–9 years’ experience, external to the training-set annotation team) participated; each item was rated by two random teachers, disagreements of ≥ 2 scale points resolved by a senior reviewer (5.7%). **Raters were blinded to system identity throughout all conditions.**

Absolute scores corroborate automatic metrics. Table 26 reports human and RM averages over the full test set. The system ranking by human 8-Avg matches the RM ranking, and the cell-level Pearson correlation between human and RM scores is $r=0.87$ ($p < 10^{-6}$). The RA-MOPD-over-GPT-5.3 advantage is reproduced in human scoring (Δ H8-Avg = +0.09, consistent with Δ 8-Avg = +0.09).

Pairwise preferences. Table 27 reports overall pairwise preferences (majority vote of two teachers; Fleiss’ κ per condition; two-sided binomial p vs.

Dim	RA-MOPD		GPT-5.3		SFT		Matched	
	RM	H	RM	H	RM	H	RM	H
App. F	4.13	4.10	4.10	4.12	3.75	3.74	4.10	4.07
App. S	4.05	4.02	4.00	3.96	3.75	3.72	4.00	3.95
Imag. L	4.20	4.22	4.20	4.22	3.79	3.78	4.10	4.12
Imag. N	4.10	4.08	4.10	4.08	3.30	3.32	4.00	4.02
Narr. D	4.20	4.16	4.05	4.00	3.63	3.60	4.10	4.02
Narr. C	4.42	4.36	4.25	4.18	4.08	4.04	4.30	4.22
Obj. F	4.30	4.26	4.10	4.06	3.97	3.94	4.20	4.14
Obj. M	4.28	4.24	4.15	4.10	3.96	3.92	4.15	4.10
8-Avg	4.21	4.18	4.12	4.09	3.78	3.76	4.12	4.08

Table 26: Human absolute scores (H) vs. RM scores on the full 1,200-essay test set. Raters were blinded to system identity. System ranking under human scoring matches the RM ranking; cell-level Pearson $r=0.87$.

50%).

Condition	Win	Tie	Opp.	κ	p
(A) GPT-5.3	62.8%	8.5%	28.7%	0.62	$< 10^{-9}$
(B) SFT	76.1%	6.8%	17.1%	0.68	$< 10^{-9}$
(C) Expert	58.4%	11.0%	30.6%	0.64	$< 10^{-5}$

Table 27: Pairwise preference of RA-MOPD vs. three baselines on the full 1,200-essay test set rated by an independent six-teacher panel (blinded). Win/Opp. = RA-MOPD / opponent preference rate. All margins are statistically significant.

Per-dimension breakdown (Condition A). Table 28 reports the per-dimension breakdown of the RA-MOPD vs. GPT-5.3 3-shot pairwise comparison.

R GPT-4.1 Holistic Judge: Full Protocol

Model. All holistic judgements are produced by gpt-4.1-2025-04-14 via the OpenAI Chat Completions API (temperature 0, $n=1$).

Design rationale. This evaluation uses five judge criteria with *no lexical overlap* with the eight rubric dimension names in Table 1; the judge prompt makes no reference to our annotation rubrics. We note that GPT-based judges are subject to their own biases (e.g., preference for longer or more fluent text); the consistent system-level ordering among SFT, GPT-5.3, and RA-MOPD, together with the human study, provides complementary validation.

System prompt. “You are an expert primary-school Chinese language education specialist. Evaluate the following Chinese essay written by or for an elementary-school student. Rate it on five

Dimension	RA-MOPD win	Tie	GPT-5.3 win
App. Format	63.5%	13.2%	23.3%
App. Suggestion	58.7%	15.4%	25.9%
Imag. Logic	56.0%	12.8%	31.2%
Imag. Novelty	46.2%	10.5%	43.3%
Narr. Detail	67.4%	11.6%	21.0%
Narr. Completeness	74.3%	11.7%	14.0%
Obj. Feature Accuracy	70.1%	14.2%	15.7%
Obj. Multisense & Rhet	76.0%	9.5%	14.5%
Overall	62.8%	8.5%	28.7%

Table 28: Per-dimension pairwise preference between RA-MOPD and GPT-5.3 3-shot (Condition A, six-teacher independent blinded panel). RA-MOPD wins on seven of eight dimensions; Imagination Novelty is the only near-tie.

*dimensions (1–5 each): (1) **Content Relevance**—does the essay address the given topic clearly and substantively; (2) **Language Fluency**—is the expression smooth, natural, and grammatically correct; (3) **Structural Coherence**—is the essay well-organised with a clear beginning, middle, and end; (4) **Age-Appropriate Register**—does the writing sound like a genuine elementary student, neither too adult nor too immature; (5) **Originality**—does the essay demonstrate creative, non-formulaic thought. Return only a JSON object with keys content, fluency, structure, register, originality, and mean (the arithmetic average of the five). Do not output any other text.”*

User message.

```
Topic: {prompt}\n\nEssay:
{essay}
```

Protocol. Temperature 0; $n=1$; 200 essays per system (50 per genre), drawn from the same held-out test set used for all other evaluations.

S GPT-5.5 Holistic Evaluation of RA-MOPD Components

Setup. We additionally evaluate Base MOPD, +RW-KL, +CGM, and RA-MOPD with GPT-5.5 as a holistic judge. The evaluation covers all 1,200 held-out prompts and 4,800 outputs in total. System identities, RM scores, and the final evaluation results are hidden from the judge. GPT-5.5 assesses content relevance, language fluency, structural coherence, age-appropriate expression, and originality on a 1–5 scale and returns their arithmetic mean. We use temperature 0 and one judgment per output.

System	RM 8-Avg	GPT-5.5 Holistic
Base MOPD	4.09	3.97 ± 0.031
+ RW-KL	4.15	4.00 ± 0.030
+ CGM	4.15	4.00 ± 0.030
RA-MOPD	4.21	4.04 ± 0.029
Δ RA–Base	+0.12	+0.07
Paired test	—	$p < 0.01$

Table 29: Additional component-level evaluation. Values after $\pm$ are standard errors. The paired test compares RA-MOPD with Base MOPD on the same 1,200 prompts.

Analysis. The holistic scores reproduce the RM component ordering: RW-KL and CGM each improve over Base MOPD, and their combination performs best. The +0.07 RA-MOPD improvement is significant under the same 10,000-replicate paired-bootstrap protocol used for the main comparisons ($p < 0.01$). This additional holistic perspective further supports the individual contribution and complementarity of the two components.

T Qwen3-1.7B Backbone-Scale Evaluation

We repeat Stage1-SFT, Base MOPD, and RA-MOPD on Qwen3-1.7B using the same data split, reward models, decoding protocol, reward coefficients, RW-KL configuration, and CGM configuration as for Qwen3-4B. No backbone-specific hyperparameter tuning is performed.

Backbone	Stage1-SFT	Base MOPD	RA-MOPD	Δ	RA route (%)
Qwen3-4B	3.78	4.09	4.21	+0.12	97.8
Qwen3-1.7B	3.63	3.93	4.04	+0.11	96.3

Table 30: Backbone-scale results. Scores are RM 8-Avg; Δ denotes RA-MOPD minus Base MOPD, and the final column reports RA-MOPD routing accuracy.

Backbone	Base MOPD	+ RW-KL	+ CGM	RA-MOPD
Qwen3-4B	4.09	4.15	4.15	4.21
Qwen3-1.7B	3.93	3.98	3.99	4.04

Table 31: Component transfer across two Qwen3 model scales. Each component variant starts from the corresponding Base MOPD model.

Absolute scores decrease at the smaller scale, while the relative RA-MOPD gain remains nearly unchanged (+0.12 vs. +0.11). RW-KL and CGM

are individually effective at both scales, and their combination consistently performs best without model-specific retuning.

U Qualitative Examples

Table 32 presents two paired comparisons between RA-MOPD and GPT-5.3 3-shot. The pattern is consistent with the aggregate numbers: RA-MOPD produces coherent event arcs in an age-appropriate register, while GPT-5.3 excels at striking but adult-style imaginative imagery.

Prompt	Excerpt (English gloss)
Narr. “my dog”	GPT-5.3 3-shot: <i>Bobo is a small dog...</i> (adult-style observation; no scene structure). RA-MOPD: <i>Last winter my dog Bobo got lost in the park... Grandma found him three streets away and Bobo licked her hand for a long time</i> (complete event arc; child voice; Narration 5, Detail & Emotion 5).
Imag. “hometown in 20 yrs”	RA-MOPD: <i>Solar trams glide on quiet elevated tracks...</i> (logically coherent; conventional sci-fi imagery; Logic 4, Novelty 4). GPT-5.3 3-shot: <i>Children grow vegetables on walls of liquid mercury; clouds can be folded into envelopes...</i> (striking imagery; Logic 4, Novelty 5).

Table 32: Paired excerpts illustrating where RA-MOPD wins (Narration Completeness; child-appropriate voice) and where GPT-5.3 3-shot wins (Imagination Novelty).