PROCESS ADVANTAGE SIGNAL SHAPING: ADVANTAGE CONSTRUCTION FOR PROCESS-SUPERVISED REINFORCEMENT LEARNING IN LLM REASONERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense process supervision supplies local feedback but leaves open how that feedback should become a policy advantage. We introduce **PASS** (*Process Advantage Signal Shaping*), an advantage-construction pipeline for group-relative policy optimization (GRPO). PASS combines separate reward-channel normalization and fusion, signal-defined chunks for credit accumulation, and normalization by the remaining chunk count. This design connects the scale of local feedback to the resolution and horizon over which it contributes to a policy update. We evaluate PASS using learned process rewards for mathematical reasoning and teacher-derived token signals for multi-hop question answering. On seven mathematics benchmarks, PASS improves average pass@1 by 4.62 points over outcome-only GRPO while reducing average response length by approximately 2%. On multi-hop QA with mean-centered normalization, PASS exceeds outcome-only training by 0.87 and 0.63 points for two distillation signals, with longer responses. Across both normalization schemes, PASS achieves the highest mean accuracy among the tested process-signal configurations and remains closest to outcome-only under absolute-maximum normalization. These findings show that channel handling, credit resolution, and accumulation should be considered jointly when incorporating dense supervision into policy optimization.

1 INTRODUCTION

Process supervision offers feedback within a reasoning trajectory rather than only at its conclusion. Learned process reward models (PRMs) score intermediate steps (Lightman et al., 2023; Zhang et al., 2025), while on-policy distillation supplies teacher-derived token log-probability ratios (Agarwal et al., 2023). These signals can support group-relative policy optimization (GRPO) (Shao et al., 2024; Guo et al., 2025), yet denser feedback alone does not specify how to assign credit. A local score must still be normalized, combined with sequence-level rewards, and transformed into the per-token advantage used to update the policy. The same upstream feedback can therefore lead to different learning signals depending on these choices.

Three interacting choices are particularly important. Reward channels can have different scales and variances, so normalizing them jointly changes their relative influence. Credit resolution also depends on the source: a PRM typically provides piecewise-constant scores, whereas a teacher-derived signal varies at individual tokens. Finally, accumulating local scores introduces dependence on the number of scored units remaining in the response. Channel normalization determines what values are accumulated, segmentation determines the units being counted, and length normalization determines how that accumulation scales with the horizon. Considering these operations together offers a practical way to connect dense feedback to policy learning.

We introduce **PASS** (*Process Advantage Signal Shaping*), a pipeline combining *Advantage Fusion* (AF), *Chunk-by-Value* (CV), and *Divide-Length* (DL). AF normalizes process, outcome, and format channels separately and combines them with a format-dependent gate. CV uses changes in the fused score trace to select shared chunk boundaries. DL then normalizes the chunk return-to-go by a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

power of the remaining chunk count. PASS supplies the resulting advantages to the existing clipped surrogate and regularizer. The same construction accommodates PRM and teacher-derived signals, with domain-specific scaling and segmentation tolerances.

PASS builds on work on reward-channel normalization (Ye et al., 2025; Liu et al., 2026), credit-assignment granularity (Zhang, 2026), and length-dependent reward transformation (Razin et al., 2025; Li et al., 2026b). Its focus is their combination at the advantage-construction stage: numerical score changes define accumulation units, and the residual number of those units determines the normalization. A fixed-score analysis clarifies the role of the normalization exponent. Detailed comparisons with related approaches appear in Section 2.

Our contribution is both a concrete advantage-construction method and an evaluation of its accuracy–length behavior across two forms of process supervision. We compare PASS with outcome-only and process-signal controls, study alternative normalization operators, and examine sensitivity to the residual-count exponent and PRM scale. The main comparisons use multiple training seeds. On seven mathematics benchmarks, PASS improves average pass@1 by 4.62 points and reduces mean response length by approximately 2%. On multi-hop QA, it improves over AF-only by 3.40 and 1.93 points under Masked-Norm, and over outcome-only by 0.87 and 0.63 points, with increased response length. Absolute-maximum normalization changes this balance: PASS remains the strongest tested process-signal configuration in mean accuracy, while outcome-only training remains competitive. Together, these results identify advantage construction and normalization choice as consequential parts of process-supervised learning.

2 Related Work

Process rewards and mixed-reward GRPO. Step-level verifiers (Lightman et al., 2023) and subsequent PRM training work (Zhang et al., 2025) provide dense feedback for reasoning. Integrating this feedback into GRPO (Shao et al., 2024; Guo et al., 2025) introduces choices about reward combination and normalization. Studies of implicit process estimation (Sullivan, 2025), outcome/process consistency (Ye et al., 2025), and decoupled reward normalization (Liu et al., 2026) address complementary aspects of this problem. PASS combines separate channel handling with signal-defined accumulation units and residual-count normalization.

Credit-assignment resolution. Reasoning credit assignment spans token, span, and step granularity (Zhang, 2026). FIPO (Ma et al., 2026) uses current/old-policy log-probability information, while counterfactual span weighting (Khandoga et al., 2026) and attribution-based ACPO (Yin et al., 2025) investigate other ways to localize credit. DRPO (Li et al., 2025) studies the advantage baseline. CV derives numerical endpoints from the supplied score traces, providing accumulation units without an auxiliary semantic segmenter. This emphasis on score-defined resolution complements approaches that use semantic or counterfactual information.

Length and reward transformation. Prior analyses examine length dependence in reasoning RL (Razin et al., 2025; Liu et al., 2025). GR^3 (Li et al., 2026b) rescales raw rewards before group normalization, and post-hoc calibration (Huang et al., 2024) studies reward-related corrections. DL instead normalizes the fused chunk return-to-go by the remaining chunk count. Our experiments include a GR^3-inspired factor applied to the fused advantage, allowing comparison of two forms of post-fusion processing within the same training setup.

On-policy distillation. On-policy distillation (Agarwal et al., 2023) and its generalized variants (Yang et al., 2026) provide teacher-dependent feedback on student trajectories. Analyses of distillation dynamics (Li et al., 2026a) motivate attention to where this feedback is applied. We use sampled log-probability-ratio signals to study advantage construction with token-varying feedback, complementing the piecewise-constant PRM setting on mathematics.

3 Preliminaries

Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) samples G responses to each prompt and uses relative reward information to construct an advantage. Let i index a response,

t a token position, and $M_{i,t}$ its response mask. A process signal $s_{i,t}$ supplies local feedback, while $r_{o,i}$ and $r_{f,i}$ are sequence-level outcome and format rewards. We write $z_{o,i}$ and $z_{f,i}$ for their separately group-standardized values. Unlike these sequence-level channels, a process signal need not be constant along a response.

Normalization scope. Distinguish a same-prompt group of G responses from the rollout batch $\mathcal{B}$ passed to advantage construction. Outcome and format statistics are computed within same-prompt groups. The full PASS process-channel normalizer pools valid tokens in $\mathcal{B}$:

$$\mathcal{V} = \{s_{i,t} : i \in \mathcal{B},\ M_{i,t} = 1\}. \tag{1}$$

The two process-channel operators in QA are

$$z_{p,i,t} = \mathcal{N}(s_{i,t}; \mathcal{V}) = \begin{cases} (s_{i,t} - \mu_{\mathcal{V}})/(\sigma_{\mathcal{V}} + \epsilon), & \text{Masked-Norm}, \\ s_{i,t}/(\max_{v \in \mathcal{V}} |v| + \epsilon), & \text{Abs-Max}. \end{cases} \tag{2}$$

The standard deviations in advantage construction use the sample correction. Masks exclude padding from the process pool, while sequence-level rewards are standardized within each same-prompt group. Mathematical reasoning uses Masked-Norm; QA compares both operators.

Alternative estimators. If $a_{i,j}$ is a score at the j-th selected segment endpoint, one possible process-supervised estimator is the return-to-go (RTG)

$$R_{i,j} = \sum_{\ell=j}^{K_i} a_{i,\ell}, \tag{3}$$

where K_i counts selected endpoints for response i. A direct token estimator instead uses $a_{i,t}$ without this sum. The process-only and AF-only QA controls use direct estimators; PASS combines chunking, RTG, and residual-count normalization.

Policy update. Token advantages $A_{i,t}$ are held fixed while optimizing the GRPO clipped surrogate. Its token contribution is $\min(\rho_{i,t}A_{i,t}, \text{clip}(\rho_{i,t}, 1-\varepsilon_c, 1+\varepsilon_c)A_{i,t})$, where ρ compares the updated and pre-update policy likelihoods. Gradients do not pass through reward extraction, normalization statistics, or segmentation. We retain the masked loss reduction and reference regularization described in Appendix G.

4 Method: The PASS Pipeline

PASS connects three choices in advantage construction: the relative scale of reward channels, the units of credit accumulation, and normalization over the remaining horizon. Figure 1 summarizes the computation. Channel fusion produces a token score trace; CV selects endpoints from that trace; DL accumulates endpoint scores and normalizes them before broadcasting the resulting credit to tokens.

4.1 Advantage Fusion

AF makes channel weights explicit after normalizing process, outcome, and format rewards separately. This avoids coupling their relative contributions through a common normalization scale. Let $z_{p,i,t}$ be the normalized process trace and $z_{o,i}, z_{f,i}$ the normalized sequence-level rewards from §3. With nonnegative weights, $w_f > 0$, and $W = w_p + w_o + w_f$, the fused score at a valid response position is

$$a_{i,t} = \gamma \begin{cases} w_p z_{p,i,t} + w_o z_{o,i} + w_f z_{f,i}, & r_{f,i} > 0, \\ W z_{f,i}, & r_{f,i} = 0. \end{cases} \tag{4}$$

The format reward is binary. For format-compliant responses, all three channels contribute; otherwise, the format channel alone supplies credit. The scale is $\gamma = 1/W$ in QA and $\gamma = 1$ in mathematics. Thus QA uses a weighted average with a format-only fallback, whereas mathematics uses the corresponding weighted sum. This scale is part of the training recipe because it sets the strength of the advantage relative to fixed regularization. We use unit channel weights for full PASS.

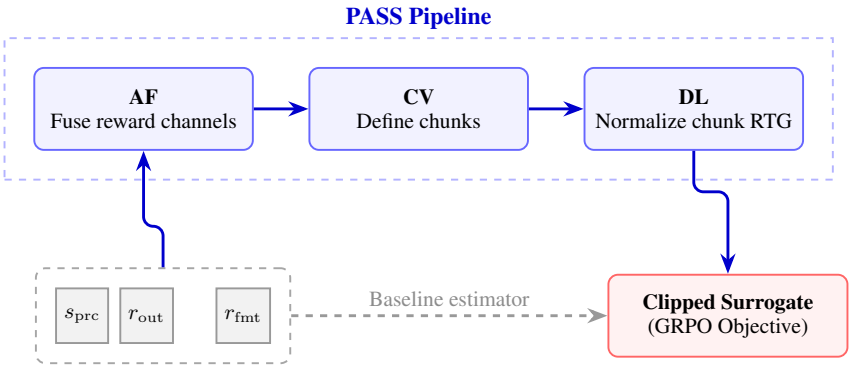

Figure 1: PASS advantage construction. AF combines separately normalized process, outcome, and format channels; CV defines chunks from the fused trace; DL normalizes the chunk return-to-go. The resulting token advantages enter the GRPO clipped surrogate.

The QA controls are defined directly from these quantities: process-only uses $A_{i,t} = z_{p,i,t}$, AF-only uses $A_{i,t} = a_{i,t}$, and outcome-only uses the AF expression with $w_p = 0$. All three use direct token advantages without RTG. In the outcome-only control, $W = 2$ rather than 3, so the relative coefficients of the remaining channels change accordingly. Table 11 gives the complete configurations.

Process signals. For mathematics, a learned PRM assigns a score to each solution step, which is broadcast over the tokens of that step. QA uses the sampled log-probability-ratio signal

$$s_{i,t}^{\text{OPD}} = \ell_{\text{teacher}}(x_{i,t} \mid h_{i,t}) - \ell_{\text{student}}(x_{i,t} \mid h_{i,t}), \tag{5}$$

where $h_{i,t}$ is the sampled prefix and each ℓ is a log probability. The G-OPD variant (Yang et al., 2026) introduces a base-policy anchor:

$$s_{i,t}^{\text{G-OPD}} = (\ell_{\text{base}} - \ell_{\text{student}}) + \lambda(\ell_{\text{teacher}} - \ell_{\text{base}}). \tag{6}$$

All terms are evaluated at the sampled token and prefix, and $\lambda = 1$ recovers OPD. Both signals can be positive or negative. For OPD, sampling from the same student distribution appearing in the ratio gives conditional expectation $-D_{\text{KL}}(\pi_{\text{student}} \| \pi_{\text{teacher}})$; top-$p$ truncation need not preserve this equality. We apply the chosen normalizer to the sampled scores rather than assuming a fixed sign.

4.2 Chunk-by-Value

CV defines the units over which credit will be accumulated. To derive a common segmentation from responses with different score traces, it uses their masked positional mean. Let $[t_0, t_1]$ cover the valid response positions in the batch. For both domains, define

$$u_t = \frac{\sum_{i \in \mathcal{B}} M_{i,t} a_{i,t}}{\max(1, \sum_{i \in \mathcal{B}} M_{i,t})}, \qquad t_0 \le t \le t_1. \tag{7}$$

An endpoint is placed at $t - 1$ when $|u_t - u_{t-1}| > \eta$, and the shared interval endpoint t_1 is included. For response i, only endpoints lying within its valid mask are retained. Denote them by $b_{i,1} < \cdots < b_{i,K_i}$ and take $a_{i,j} = a_{i,b_{i,j}}$ as the chunk score. We use $\eta = 0.005$ for the continuously varying QA signal and $\eta = 10^{-8}$ for the piecewise-constant PRM signal. These are signal-defined credit units: they share positional boundaries, but need not correspond to the same semantic step across responses.

For any selected chunk values $v_{i,j}$, broadcasting assigns

$$\text{Broadcast}(v_i)_t = v_{i,j}, \qquad b_{i,j-1} < t \le b_{i,j}, \quad b_{i,0} = t_0 - 1, \tag{8}$$

with zero outside the selected intervals. Endpoint broadcasting gives a piecewise-constant approximation to each response's credit trace. Appendix B specifies masking and the zero-filled tail after a response's last retained endpoint.

4.3 DIVIDE-LENGTH

DL makes the scale of accumulated credit depend on the number of remaining chunks. For the chunk scores selected by CV, it computes

$$A_{i,j}^{\mathrm{DL}} = \frac{\sum_{\ell=j}^{K_i} a_{i,\ell}}{(K_i - j + 1)^k}. \tag{9}$$

These values are broadcast by Equation 8. At $k = 1$ the result is the mean score over the remaining chunks; smaller exponents retain more dependence on the accumulation horizon. We use $k = 1$ in the main mathematics experiments and $k = 0.7$ in QA. Normalization applies to the fused score, including the outcome and format channels: a constant contribution C_i becomes $C_i n^{1-k}$ with n remaining chunks. Algorithm 4 gives an end-to-end overview.

A fixed-score diagnostic. To interpret attenuation separately from training, freeze the chunk scores, existing boundaries, and normalization statistics. With $R = \sum_{j=1}^{K} a_j$, define

$$S = R/K^k. \tag{10}$$

Appending one new chunk of score δ gives

$$\Delta S = \frac{R + \delta}{(K+1)^k} - \frac{R}{K^k}, \tag{11}$$

and hence the exact condition

$$\Delta S > 0 \quad \Longleftrightarrow \quad \delta > R\left[\left(1 + \frac{1}{K}\right)^k - 1\right]. \tag{12}$$

At $k = 1$, an appended chunk raises the diagnostic precisely when its score exceeds the existing average R/K. For $k < 1$, the threshold reflects weaker length attenuation. This relationship describes a fixed-score diagnostic, rather than a step-acceptance rule or the dynamics of the complete clipped update. In particular, $k < 1$ does not guarantee a finite optimal length. The derivation and repeated-contribution behavior are given in Appendix A.

5 EXPERIMENTS

We evaluate how advantage construction affects performance with PRM feedback on mathematics and teacher-derived feedback on multi-hop QA. The experiments compare overall accuracy, the effect of channel handling and normalization, and the accompanying response lengths. Main comparisons use multiple training seeds; additional single-seed studies examine the normalization exponent and PRM scale. Section 6 further interprets several key findings and discusses their scope.

5.1 MATHEMATICAL REASONING WITH A PRM PROCESS SIGNAL

Setup. We train Qwen2.5-Math-7B (Yang et al., 2024) on DeepScaleR (Luo et al., 2025), using Qwen2.5-Math-PRM-7B (Zhang et al., 2025) for the PASS process channel, and evaluate on AIME24/25, AMC23, GSM8K, MATH, Minerva, and Olympiad. Correctness is checked with `math_verify` (Kydlíček, 2025). We report mean and sample standard deviation across three PASS training runs and two outcome-only GRPO(ORM) runs. The pre-RL base model provides an additional reference. Per-question sampling counts and metric definitions are given in Appendix G.

Overall performance. PASS achieves 36.04 ± 0.22 average pass@1 versus 31.41 ± 0.35 for outcome-only GRPO, an improvement of 4.62 points (Table 1). Across the observed runs, PASS scores higher than the baseline on all seven benchmarks. Gains are largest on GSM8K ($+6.28$), AMC23 ($+5.83$), and MATH ($+5.33$), covering both school-level and competition mathematics. The pass@8 average also improves by 2.76 points, with the largest differences on AIME24 and AIME25 (Appendix C).

Table 1: pass@1 (%) on mathematical reasoning. Deltas are computed before rounding; the average gives each benchmark equal weight.

Benchmark	Pre-RL	GRPO(ORM)	PASS (Ours)	Δ
AIME24	8.9	12.36 ± 0.10	$\mathbf{16.95 \pm 0.95}$	$+4.59$
AIME25	2.3	4.48 ± 0.37	$\mathbf{9.18 \pm 0.37}$	$+4.70$
AMC23	22.8	47.88 ± 1.81	$\mathbf{53.72 \pm 1.36}$	$+5.83$
GSM8K	30.1	71.27 ± 0.17	$\mathbf{77.56 \pm 1.23}$	$+6.28$
MATH	27.9	49.96 ± 0.18	$\mathbf{55.29 \pm 0.81}$	$+5.33$
Minerva	8.4	24.33 ± 0.15	$\mathbf{27.90 \pm 0.60}$	$+3.57$
Olympiad	4.1	9.60 ± 0.05	$\mathbf{11.65 \pm 0.20}$	$+2.05$
Average	14.9	31.41 ± 0.35	$\mathbf{36.04 \pm 0.22}$	$\mathbf{+4.62}$

Accuracy and response length. The accuracy gains accompany approximately 2% lower mean response length (Table 2). Reductions are largest on GSM8K (-11.9%) and Minerva (-14.4%), while four benchmarks show increases of at most 3%. PASS therefore improves aggregate accuracy without increasing the aggregate token count on this panel. The more pronounced accuracy–length trade-off in QA is examined below.

Table 2: Average response length in tokens on the mathematical-reasoning panel. PASS's seven-benchmark average is approximately 2% lower.

Benchmark	GRPO(ORM)	PASS (Ours)	Δ rel.
AIME24	1106 ± 4	1118 ± 14	$+1.1\%$
AIME25	1162 ± 5	1185 ± 21	$+2.0\%$
AMC23	935 ± 9	946 ± 10	$+1.1\%$
GSM8K	552 ± 3	$\mathbf{486 \pm 16}$	$\mathbf{-11.9\%}$
MATH	690 ± 3	688 ± 17	-0.2%
Minerva	880 ± 3	$\mathbf{754 \pm 33}$	$\mathbf{-14.4\%}$
Olympiad	872 ± 1	896 ± 16	$+2.8\%$
Average	885 ± 1	$\mathbf{868 \pm 18}$	$\mathbf{-2.0\%}$

Exponent and scorer scale. Single-seed configuration studies in Appendix C examine $k \in \{1.0, 0.7, 0.5, 0.3\}$ and PRMs of 7B and 72B parameters. The exponent sweep spans 35.1–36.9 average pass@1, and the scorer-scale comparison increases average pass@1 from 35.7 to 37.3. These studies complement the main comparison by varying the strength of residual normalization and the source of process feedback.

5.2 Multi-hop Question Answering with Teacher-Derived Signals

Setup. We train Qwen2.5-7B-Instruct (Qwen et al., 2025) on a fixed 5k-prompt HotpotQA distractor subset and evaluate on HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). Qwen2.5-32B-Instruct supplies the teacher signals in Equations 5 and 6; G-OPD uses the initial student as base and $\lambda = 1.25$. Training runs for one epoch with rollout batches of 128 prompts, 16 responses per prompt, training batches of 128 trajectories, and learning rate 10^{-6}. Each configuration uses three training seeds. Per-dataset entries report means and population standard deviations weighted by available evaluation-sample counts; averages give the three datasets equal weight. Evaluation targets 16 responses per question and uses SQuAD-normalized answer matching. Coverage and full protocol details appear in Appendix G.

Advantage-construction controls. We compare ❶ **outcome-only**, using AF with $w_p = 0$; ❷ **process-only**, using the normalized teacher-derived trace without outcome or format rewards; ❸ **AF-only**, combining all three channels through Equation 4; and ❹ **PASS**, applying CV, RTG, and DL to the fused trace. The first three use direct token advantages without RTG. We also adapt the GR^3

length factor (Li et al., 2026b) by multiplying the AF-fused advantage by $1/(1 + \alpha L_i/(\bar{L}_g + \epsilon))$, with $\alpha = 0.33$ and within-group mean length $\bar{L}_g$. This GR^3-style control applies the factor after fusion rather than to raw rewards, and uses neither CV nor RTG.

Overall performance. Under Masked-Norm, PASS reaches 57.87 average pass@1 with OPD and 57.63 with G-OPD (Table 3). It improves over AF-only by 3.40 and 1.93 points and over outcome-only by 0.87 and 0.63 points, respectively. Process-only training scores 54.33 and 54.57, showing that the teacher-derived signal alone is less effective than outcome-only training in this setting. The GR^3-style control scores 54.50, close to AF-only's 54.47. Full PASS achieves the highest mean on each benchmark for both signals.

Table 3: pass@1 (%) on multi-hop QA under Masked-Norm. Deltas compare with outcome-only (❶) and AF-only (❸). GR^3-style denotes length rescaling applied to the AF-fused advantage.

Rung	Method	MuSiQue	2Wiki	HotpotQA	Average	$\Delta_{❶}$	$\Delta_{❸}$
❶	GRPO (outcome-only)	44.2 ± 0.3	64.9 ± 0.2	61.9 ± 0.5	57.00	–	–
OPD process signal							
❷	Process-only OPD	38.3 ± 0.8	63.8 ± 0.4	60.9 ± 0.1	54.33	-2.67	-0.14
❸	AF-only OPD	38.4 ± 0.5	63.9 ± 0.2	61.1 ± 0.1	54.47	-2.53	–
comp.	GR^3-style + OPD ($\alpha = 0.33$)	38.1 ± 1.1	64.0 ± 0.5	61.4 ± 0.4	54.50	-2.50	$+0.03$
❹	**PASS + OPD**	$\mathbf{45.3 \pm 0.7}$	$\mathbf{65.6 \pm 0.2}$	$\mathbf{62.7 \pm 0.2}$	**57.87**	$\mathbf{+0.87}$	$\mathbf{+3.40}$
G-OPD process signal							
❷	Process-only G-OPD	39.2 ± 1.4	63.3 ± 0.7	61.2 ± 0.3	54.57	-2.43	-1.13
❸	AF-only G-OPD	41.3 ± 1.2	64.0 ± 0.4	61.8 ± 0.3	55.70	-1.30	–
❹	**PASS + G-OPD**	$\mathbf{44.6 \pm 0.3}$	$\mathbf{65.7 \pm 0.2}$	$\mathbf{62.6 \pm 0.1}$	**57.63**	$\mathbf{+0.63}$	$\mathbf{+1.93}$

Effect of normalization. Table 4 compares both operators. With Abs-Max, PASS remains above process-only and AF-only in mean accuracy, scoring 56.93 for OPD and 56.17 for G-OPD. These values are 0.07 and 0.83 points below outcome-only training. Normalization has recipe-dependent effects: Abs-Max increases AF-only OPD from 54.47 to 55.90, leaves AF-only G-OPD at 55.70, and decreases the PASS means. Thus PASS improves on the process-signal controls under both operators, while the gain over outcome-only depends on normalization. Per-benchmark Abs-Max results are in Appendix D.

Table 4: QA average pass@1 across normalization choices, with PASS response length in tokens. Outcome-only scores 57.00 at 186.8 tokens under the shared control.

Signal	Normalizer	Process-only	AF-only	PASS	Δ vs. outcome	Length
OPD	Masked-Norm	54.33	54.47	57.87	$+0.87$	239.0
G-OPD	Masked-Norm	54.57	55.70	57.63	$+0.63$	249.7
OPD	Abs-Max	53.87	55.90	56.93	-0.07	239.8
G-OPD	Abs-Max	54.43	55.70	56.17	-0.83	221.0

Accuracy and response length. The QA gains accompany longer outputs. Under Masked-Norm, PASS averages 239.0 tokens with OPD and 249.7 with G-OPD, compared with 186.8 for outcome-only. AF-only and GR^3-style OPD both average 163.2 tokens. In contrast to mathematics, QA therefore presents an accuracy–length trade-off; these comparisons do not match inference compute. Appendix E and Appendix F provide the complete length results and further training-time dynamics of rollout length, actor entropy, and teacher KL for the QA runs.

6 Discussion and Limitations

Advantage construction is part of the learning problem. The QA results suggest that the usefulness of dense process feedback depends not only on the quality of the upstream signal, but also on how that signal is converted into the advantage seen by policy optimization. PASS improves over the tested process-signal controls under both normalization operators, while its comparison with

outcome-only training changes with the normalizer: it exceeds outcome-only under Masked-Norm but not under Abs-Max. This makes normalization more than a numerical stabilization step. Mean-centering and absolute-maximum scaling preserve different reference points and relative offsets in the process channel, which in turn changes how process feedback interacts with outcome and format rewards after fusion. In other words, a fixed process signal does not define a unique learning signal; its effect is mediated by the geometry of the resulting advantage. This perspective also helps explain why denser supervision is not automatically more useful than outcome supervision. The additional signal introduces structure at finer temporal resolution, but whether that structure becomes informative credit or merely perturbs the update depends on how channel scale, accumulation units, and downstream normalization are coupled. PASS can therefore be viewed less as a new source of supervision than as a mechanism for constructing a usable optimization target from supervision that already exists.

This interpretation has a broader implication for process-supervised RL. Comparisons between process rewards should ideally distinguish the upstream scoring model from the transformation that maps its scores into policy advantages. Two methods using the same PRM or teacher signal can induce substantially different updates if they normalize over different populations, fuse channels differently, or accumulate credit at different resolutions. Conversely, an apparent improvement from a stronger process signal may partly reflect a more favorable advantage construction rather than the signal source alone. Our experiments do not disentangle all of these factors individually, but the normalization results make clear that the mapping from score space to advantage space is itself an experimentally consequential design choice.

Residual normalization does not act as a generic preference for shorter responses. The two domains exhibit different accuracy–length behavior. On mathematics, PASS improves average pass@1 by 4.62 points while reducing mean response length by approximately 2%. On QA, its gains under Masked-Norm accompany longer outputs. This contrast argues against interpreting DL as a generic preference for shorter responses. Instead, DL changes how cumulative credit scales with the number of remaining signal-defined units. At $k = 1$, the residual return is normalized into an average over the remaining chunks; for $k < 1$, the estimator retains more dependence on the residual horizon. The immediate effect is therefore on the magnitude and position dependence of the advantage, while any change in generated length arises only indirectly through policy learning.

This distinction matters because several different mechanisms can otherwise be conflated under the label of length control. A direct length penalty changes the reward as a function of response length, while DL changes the scale of already-accumulated reward according to the number of remaining credit units. The latter can reduce the incentive for repeatedly accumulating similar positive contributions, but it can also preserve or amplify useful credit when later chunks carry distinct information. Its behavior consequently depends on both the score sequence and the segmentation that determines the chunk count. The fixed-score analysis captures one aspect of this dependence, but the opposite length trends observed in mathematics and QA show that the trained policy cannot be understood from the residual divisor alone.

The QA results also highlight an unresolved interaction between credit scaling and inference compute. Under Masked-Norm, PASS produces longer responses than outcome-only and AF-only while achieving higher mean accuracy. Because these comparisons do not match inference length or total generation cost, part of the gain may arise from allowing the policy to spend more tokens on reasoning. At the same time, the mathematics results show that PASS does not require increased response length to improve accuracy. Taken together, the two settings suggest that residual normalization is better understood as shaping where and how strongly credit is assigned across a trajectory, with response length emerging as a task- and signal-dependent consequence rather than a fixed optimization target.

Signal-defined chunks trade semantic structure for a model-free credit resolution. CV defines accumulation units directly from changes in the fused score trace. This gives PASS a common mechanism for handling two qualitatively different forms of process supervision: piecewise-constant PRM scores in mathematical reasoning and token-varying teacher signals in multi-hop QA. Rather than assuming that a particular semantic unit—such as a sentence, reasoning step, or tool action—is the correct granularity for credit assignment, CV lets the supplied signal determine where the effective credit resolution changes. In this sense, chunking is not merely a compression step; it determines

the units over which return-to-go is accumulated and therefore changes the structure of the resulting advantage.

This design trades semantic interpretation for a lightweight, signal-driven notion of resolution. It requires no auxiliary segmenter and does not depend on a task-specific definition of a reasoning step, which makes the same construction applicable across signal sources. At the same time, numerical changes in a score trace need not coincide with meaningful reasoning decisions. Shared positional boundaries can merge semantically distinct regions with similar values or separate regions because of small numerical changes, and broadcasting those boundaries back to individual responses only approximates each response-specific trace. CV should therefore be viewed as defining a practical credit coordinate system from the supervision signal, rather than recovering latent reasoning steps.

The interaction with AF and DL is also important. Because CV operates on the fused trace, the selected boundaries depend on the relative scaling of process, outcome, and format channels; changing the normalization or fusion rule can therefore change not only the magnitude of the advantages but also the units over which they are accumulated. Likewise, DL normalizes by the number of remaining chunks rather than by raw token length, so the meaning of its residual horizon is inherited from CV. The three operations are consequently coupled: AF determines the score geometry, CV converts that geometry into discrete accumulation units, and DL determines how credit scales across those units. This coupling helps explain why the full pipeline should be interpreted as an advantage-construction recipe rather than as three independent post-processing heuristics.

A useful consequence of this perspective is that process supervision quality and credit resolution are distinct axes. A strong process scorer may still yield a poor learning signal when its variation occurs at a resolution mismatched to the accumulation rule. Conversely, coarser numerical units may suppress high-frequency variation that would otherwise dominate token-level credit. Our experiments do not identify an optimal chunk granularity, and the shared-boundary rule introduces approximation error. These limitations motivate adaptive or response-specific credit units. More broadly, they suggest that semantic segmentation is not the only viable basis for credit assignment: signal-defined numerical structure can also provide a useful basis for constructing policy advantages.

Potential limitations. Our results primarily support PASS as a combined advantage-construction pipeline rather than isolating the contribution of each operator: in QA, the move from AF-only to PASS jointly introduces shared chunking, RTG, and residual normalization. The main comparisons also cover a limited number of training runs, with three PASS and two ORM runs in mathematics and three runs per QA configuration, while the exponent and PRM-scale studies use a single seed per setting. Finally, all main experiments use Qwen2.5 actors at approximately 7B scale. These factors limit how strongly the observed gains can be attributed to individual components or generalized across training variability, policy families, process-signal sources, and multi-turn settings.

7 Conclusion

PASS connects dense process feedback to policy updates through separate channel handling, signal-defined chunks, and residual chunk-count normalization. Across seven mathematics benchmarks, this construction improves average pass@1 by 4.62 points over outcome-only GRPO with approximately 2% lower mean response length. On multi-hop QA, PASS has higher mean accuracy than the tested process-signal controls under both normalization operators, and exceeds outcome-only training under Masked-Norm. The QA gains accompany longer responses, while the comparison with Abs-Max shows that normalization changes the relative value of process supervision. These findings support treating the conversion from local scores to policy advantages as a central design choice in reasoning RL. Extending the construction across policy families and examining its accuracy–compute trade-off are natural directions for further study.

ETHICS STATEMENT

This work uses existing mathematical-reasoning and question-answering datasets and open-weight models. It does not involve new human-subject data collection. Advantage shaping is not a safety or factuality guarantee: emphasizing a process signal may also emphasize errors or biases in its source. Our empirical scope is limited to the evaluated benchmarks and training recipes.

REPRODUCIBILITY STATEMENT

Section 4 defines the PASS operators and the two teacher-derived signals. Appendix B specifies the full computation, including masks, response endpoints, and token alignment. Training configurations, evaluation sampling counts, and statistical aggregation are provided in Appendix G. The single-seed exponent and PRM-scale studies are described in Appendix C, and Appendix D gives per-benchmark results for the alternative normalizer.

AI USE STATEMENT

Generative AI tools assisted with code analysis, result-aggregation scripts, manuscript drafting and editing, and checking and revising mathematical arguments and method descriptions. Experimental summaries were computed from evaluation outputs. The authors are responsible for the final text, technical claims, and experimental results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3zKtaqxLhW.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo M. Ponti, and Ivan Titov. Post-hoc Reward Calibration: A Case Study on Length Bias, 2024. URL https://arxiv.org/abs/2409.17407v2.

Mykola Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. Beyond Uniform Credit: Causal Credit Assignment for Policy Optimization, 2026. URL https://arxiv.org/abs/2602.09331v1.

Hynek Kydlíček. Math-Verify: Math Verification Library, September 2025.

Gang Li, Yan Chen, Ming Lin, and Tianbao Yang. DRPO: Efficient Reasoning via Decoupled Reward Policy Optimization, 2025. URL https://arxiv.org/abs/2510.04474v2.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe, 2026a. URL http://arxiv.org/abs/2604.13016.

Zichao Li, Jie Lou, Fangchen Dong, Zhiyuan Fan, Mengjie Ren, Hongyu Lin, Xianpei Han, Debing Zhang, Le Sun, Yaojie Lu, and Xing Yu. Tackling Length Inflation Without Trade-offs: Group Relative Reward Rescaling for Reinforcement Learning, 2026b. URL https://arxiv.org/abs/2603.10535v1.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=v8L0pN6EOi.

Qiyuan Liu, Hao Xu, Xuhong Chen, Wei Chen, Yee Whye Teh, and Ning Miao. Enhancing Large Language Model Reasoning with Reward Models: An Analytical Survey, 2025. URL https://arxiv.org/abs/2510.01925v2.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-Decoupled Normalization Policy Optimization for Multi-reward RL Optimization, January 2026.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Technical report, agentica, 2025.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. FIPO: Eliciting Deep Reasoning with Future-KL Influenced Policy Optimization, 2026. URL http://arxiv.org/abs/2603.19835.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025. URL http://arxiv.org/abs/2412.15115.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What Makes a Reward Model a Good Teacher? An Optimization Perspective. *arXiv*, arXiv:2503.15477, March 2025. doi: 10.48550/arXiv.2503.15477.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv*, arXiv:2402.03300, April 2024. doi: 10.48550/arXiv.2402.03300.

Michael Sullivan. GRPO is Secretly a Process Reward Model, October 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation, 2026. URL http://arxiv.org/abs/2602.12125.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Chenlu Ye, Zhou Yu, Ziji Zhang, Hao Chen, Narayanan Sadagopan, Jing Huang, Tong Zhang, and Anurag Beniwal. Beyond Correctness: Harmonizing Process and Outcome Rewards through RL Training, 2025. URL http://arxiv.org/abs/2509.03403.

Junxi Yin, Haisen Luo, Zhenyu Li, Yihua Liu, Dan Liu, Zequn Li, and Xiaohang Xu. Pinpointing crucial steps: Attribution-based Credit Assignment for Verifiable Reinforcement Learning, 2025. URL https://arxiv.org/abs/2510.08899v1.

Chenchen Zhang. From Reasoning to Agentic: Credit Assignment in Reinforcement Learning for Large Language Models, 2026. URL https://arxiv.org/abs/2604.09459v2.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The Lessons of Developing Process Reward Models in Mathematical Reasoning, June 2025.

A Fixed-Score Analysis of Residual Normalization

Fix $K \geq 1$ selected chunks with scores $a_1, \ldots, a_K$, total $R = \sum_j a_j$, and exponent $0 < k \leq 1$. Holding the existing chunk scores, segmentation, and normalization statistics fixed, append one new chunk of score δ. For the path-score diagnostic $S = R/K^k$,

$$\Delta S = \frac{R+\delta}{(K+1)^k} - \frac{R}{K^k} = \frac{\delta - R[(1+1/K)^k - 1]}{(K+1)^k}. \tag{13}$$

The denominator is positive, so Equation 12 follows for signed R and δ. At $k = 1$, $\Delta S > 0$ exactly when $\delta > R/K$. This is an average-score comparison, not a hard acceptance test performed by the algorithm. For $k < 1$ the exact threshold differs from the current average. If $R = O(K)$ and $\delta = O(1)$, expansion gives

$$\Delta S = K^{-k}\left(\delta - k\frac{R}{K}\right) + O(K^{-k-1}). \tag{14}$$

Repeated contributions. With unchanged prefix scores, adding m chunks each scoring at least $\epsilon > 0$ increases the unnormalized cumulative score by at least $m\epsilon$. Repeated constant contributions under residual-count attenuation give

$$S_m = \frac{R+m\epsilon}{(K+m)^k} \sim \epsilon m^{1-k} \longrightarrow +\infty \quad (0 < k < 1). \tag{15}$$

At $k = 1$, this particular score sequence is decreasing if $\epsilon < R/K$, increasing if $\epsilon > R/K$, and constant at equality. There is no general finite-length optimum implied by $k < 1$. Moreover, appending a semantic reasoning step need not add a detected chunk: equal-valued adjacent segments can merge.

Scope of the analysis. The diagnostic equals the first-endpoint value of the residual estimator. Other positions use their own remaining counts, and policy optimization additionally includes clipping, reference regularization, and masked loss reduction. Moreover, resampling responses changes normalization statistics and shared boundaries. The fixed-score relations therefore characterize the role of k within advantage construction, rather than establish convergence or determine the direction of trained response-length changes.

B Algorithmic Specification of PASS

The algorithms below specify channel fusion, shared endpoint selection, and residual-count normalization. Rewards, log probabilities, and the resulting advantages are held fixed during the policy update. A same-prompt group contains G responses; $\mathcal{B}$ is the batch used for process normalization and shared segmentation. Missing sequence positions contribute neither values nor counts.

Masks and endpoints. Coordinates in Algorithm 2 are positions in the padded sequences, so the positional mean uses the number of valid responses at each position rather than a constant denominator. An all-zero mask contributes no values or counts and retains no endpoints; its stored sequence span participates in constructing the shared interval. Each response uses the valid subset of the shared endpoints, without adding a response-specific terminal endpoint. Consequently the broadcast leaves a zero-filled tail after the last retained endpoint. If no endpoint is retained, the returned advantage is zero throughout. Mathematical process scores use a PRM-aligned response mask that excludes terminal EOS, and the final advantages are shifted into next-token prediction coordinates; outcome and format normalization use their sequence-level response masks.

The QA direct controls return either the process trace, the fused trace, or the outcome/format-only fused trace without Algorithms 2 and 3. The GR^3-style control applies its response-level length factor to the fused trace. Full PASS uses all stages, with $\gamma = 1/W$, $\eta = 0.005$, $k = 0.7$ in QA and $\gamma = 1$, $\eta = 10^{-8}$, $k = 1$ in the main mathematics comparison.

Algorithm 1 Channel handling for QA and mathematical full PASS.

Require: Process traces s_i, response masks M_i, scalar rewards $r_{o,i}, r_{f,i}$, group size G, weights w_p, w_o, w_f, and scale γ.
Ensure: Masked fused traces a_i.

```
Normalize process values over the valid token pool in B by Equation 2.
Normalize r_o and r_f separately within each same-prompt group of G, obtaining z_o, z_f.
W ← w_p + w_o + w_f
Initialize a to zero.
for each response i do
    if r_{f,i} > 0 then
        a_{i,t} ← γ(w_p z_{p,i,t} + w_o z_{o,i} + w_f z_{f,i}) on valid response positions.
    else
        a_{i,t} ← γ W z_{f,i} on valid response positions.
    end if
end for
return a
```

Algorithm 2 Shared endpoint detection for QA and mathematical full PASS.

Require: Fused traces a_i, masks M_i; $\eta = 5 \times 10^{-3}$ for QA or 10^{-8} for math.
Ensure: Retained endpoint sets B_i and scores $a_{i,j}$.

```
Find the union response interval [t_0, t_1] across B.
u_t ← Σ_i M_{i,t} a_{i,t} / max(1, Σ_i M_{i,t}) for each t in the interval.
B ← ∅
for t = t_0 + 1, ..., t_1 do
    if |u_t − u_{t−1}| > η then
        Add t − 1 to B                          ▷ adjacent-value comparison
    end if
end for
Add t_1 to B.
for each response i do
    B_i ← {b ∈ B : M_{i,b} = 1} in increasing order.
    K_i ← |B_i|; a_{i,j} ← a_{i,b_{i,j}} for j = 1, ..., K_i.
end for
return B_i, a_{i,j}                             ▷ do not append a per-response final endpoint
```

C Additional Results on Mathematical Reasoning

We report pass@8 and problem-level uncertainty for the main mathematical comparisons, followed by single-seed studies of the normalization exponent and process reward model scale.

pass@8. PASS achieves 55.65 ± 0.64 versus 52.89 ± 0.52 for ORM, a mean improvement of 2.76 points (Table 5). The largest gains are on AIME24 ($+6.01$) and AIME25 ($+8.39$), where successful solutions remain less frequent. Improvements are smaller on AMC23, GSM8K, and Minerva ($+0.24$, $+0.23$, and $+0.51$), with overlapping observed run ranges. Thus the increase in multi-sample success is concentrated more strongly on the AIME tasks than the pass@1 gains.

Problem-level uncertainty. AIME24/25 contain 30 problems each and AMC23 40. Table 6 reports 95% bootstrap intervals using 2000 resamples of whole problems, carrying each problem's full response set. These intervals describe uncertainty across problems conditional on each trained checkpoint. They complement the cross-training-seed standard deviations in the main results; they are marginal score intervals rather than intervals for the difference between methods.

Sensitivity to the normalization exponent. We conduct the exponent and PRM-scale studies with one training seed per configuration and compare results within each table. Table 7 varies $k \in \{1.0, 0.7, 0.5, 0.3\}$ for PRM-supervised mathematical reasoning. Average pass@1 ranges from 35.1 to 36.9, a span of 1.8 points, with the highest observed value at $k = 0.7$. The main

Algorithm 3 Chunk RTG, residual normalization, and token broadcast.

Require: Endpoint scores $a_{i,j}$, ordered B_i, union interval $[t_0, t_1]$, exponent k.
Ensure: Token advantages A_i.

```
for each response i do
    Initialize A_i to zero; R ← 0.
    for j = K_i, K_i − 1, ..., 1 do
        R ← R + a_{i,j}
        v_{i,j} ← R/(K_i − j + 1)^k
    end for
    b_{i,0} ← t_0 − 1
    for j = 1, ..., K_i do
        Fill A_{i,t} ← v_{i,j} for b_{i,j−1} < t ≤ b_{i,j}.
    end for
end for
return A                ▷ unfilled tails stay zero; the loss applies its response mask
```

Algorithm 4 End-to-end PASS advantage construction.

Require: Rollout batch, channel values, masks, weights, scale γ, tolerance η, and exponent k.

```
a ← ChannelHandling(·) using Algorithm 1.
(B_i, a_{i,j}) ← SharedEndpoints(a, M, η) using Algorithm 2.
A ← ResidualRTG(B_i, a_{i,j}, k) using Algorithm 3.
if mathematical implementation then
    Shift the PRM-coordinate advantages one position left into next-token loss coordinates.
end if
Store A as a fixed rollout target for the clipped surrogate.
```

mathematical comparison uses $k = 1$, corresponding to residual averaging; smaller values retain more dependence on the remaining horizon. The single-seed sweep illustrates the effect of this choice without establishing a seed-robust ranking.

Effect of process reward model scale. Table 8 compares 7B and 72B PRMs (Zhang et al., 2025) at $k = 1$, keeping the actor at 7B scale. The 72B scorer increases average pass@1 from 35.7 to 37.3 and pass@8 from 55.0 to 55.8. This result indicates that the quality of the upstream scorer remains consequential when its outputs are processed by PASS.

D Group Standardization and Process Signals

Masked-Norm centers the process trace at its masked mean, whereas Abs-Max preserves the zero and sign of each sampled value while rescaling its magnitude. The resulting traces differ both in scale and in their reference point. Table 9 gives the per-benchmark Abs-Max results corresponding to the normalization summary in Table 4.

Setup. The process-only, AF-only, and PASS configurations are evaluated under both operators with the same actor, teacher, and training budget. The outcome-only runs are shared because their advantages do not depend on a process normalizer. The GR^3-style control uses Masked-Norm. The QA aggregation convention is defined in §5.2.

Normalization-dependent performance. Moving from Masked-Norm to Abs-Max raises AF-only OPD from 54.47 to 55.90 (+1.43), while AF-only G-OPD stays at 55.70 at the displayed precision. Process-only decreases from 54.33 to 53.87 for OPD and from 54.57 to 54.43 for G-OPD. The corresponding PASS means decrease from 57.87 to 56.93 and from 57.63 to 56.17. These changes show that the effect of normalization depends on the advantage recipe.

PASS has higher mean accuracy than the process-only and AF-only configurations under both standardizers. Its differences from outcome-only are +0.87 and +0.63 with Masked-Norm, and

Table 5: pass@8 (%) on the same mathematical runs as Table 1. Deltas are computed before rounding. The average difference is +2.76; observed run ranges overlap on AMC23, GSM8K, and Minerva.

Benchmark	GRPO(ORM)	PASS (Ours)	Δ
AIME24	38.29 ± 1.29	$\mathbf{44.30 \pm 1.88}$	+6.01
AIME25	18.17 ± 0.39	$\mathbf{26.56 \pm 1.36}$	+8.39
AMC23	82.18 ± 2.05	82.42 ± 1.40	+0.24
GSM8K	95.44 ± 0.18	95.68 ± 0.17	+0.23
MATH	69.78 ± 0.12	$\mathbf{71.63 \pm 0.09}$	+1.84
Minerva	47.32 ± 0.52	47.83 ± 0.66	+0.51
Olympiad	19.04 ± 0.10	$\mathbf{21.12 \pm 0.17}$	+2.07
Average	52.89 ± 0.52	$\mathbf{55.65 \pm 0.64}$	**+2.76**

Table 6: Problem-level bootstrap 95% intervals for pass@1 (%), conditional on each checkpoint, using 2000 resamples.

Configuration	AIME24	AIME25	AMC23
GRPO(ORM), seed 1111	12.43 [6.72, 19.21]	4.74 [1.41, 9.43]	49.17 [40.08, 58.65]
GRPO(ORM), seed 2222	12.29 [5.99, 19.32]	4.22 [1.46, 8.07]	46.60 [36.52, 56.02]
PASS, seed 1	**18.05** [10.29, 27.08]	**8.78** [3.75, 15.23]	**53.93** [43.09, 64.38]
PASS, seed 2	**16.35** [9.27, 24.14]	**9.51** [4.14, 16.20]	**54.96** [43.96, 65.80]
PASS, seed 42	**16.46** [9.27, 25.10]	**9.27** [3.75, 16.25]	**52.27** [41.64, 63.36]

-0.07 and -0.83 with Abs-Max. Normalization therefore changes the comparison with outcome-only training while preserving PASS's mean advantage over the tested process-signal controls.

E Training Dynamics on Multi-hop QA

We complement the endpoint evaluation in Section 5.2 with three training-time observables from the same Masked-Norm runs: rollout length, actor entropy, and token-level KL divergence to the teacher. These quantities do not identify a causal mechanism by themselves, but together they provide a view of how the learned policies evolve under different advantage-construction recipes.

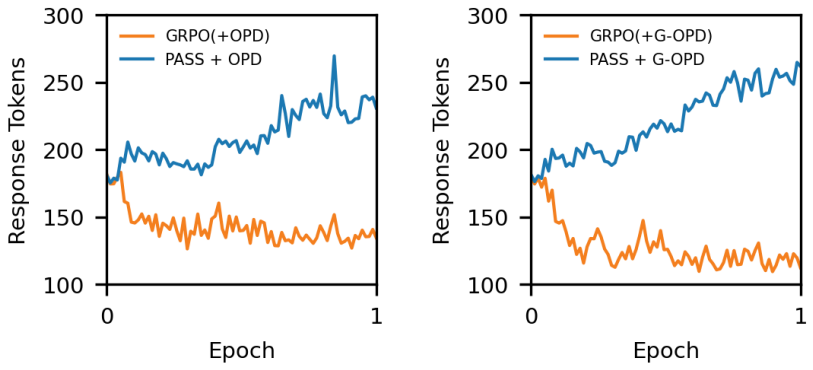

Figure 2: Training-time mean rollout length on HotpotQA across one training epoch for the Masked-Norm OPD and G-OPD configurations. Left: OPD. Right: G-OPD.

Figure 2 shows the mean rollout length over training. Starting from the same actor, the process-signal controls and PASS evolve toward substantially different response-length regimes. For both OPD and G-OPD, PASS produces progressively longer rollouts, consistent with the longer evaluation-time

Table 7: Sensitivity to the residual-normalization exponent k, with one training seed per configuration. Metrics are pass@1/pass@8 (%).

Decay Factor	AIME24	AIME25	AMC23	GSM8K	MATH	Minerva	Olympiad	Average
$k = 1.0$	14.8/39.1	**9.8**/27.4	53.8/**83.7**	77.5/**95.4**	55.1/71.3	27.3/47.2	11.5/21.1	35.7/55.0
$k = 0.7$	**17.8/47.1**	9.3/24.7	**56.2**/81.8	**78.3**/95.2	**55.7/71.4**	**29.1/48.6**	**11.8/21.2**	**36.9**/55.7
$k = 0.5$ (Sqrt-L)	15.9/44.2	9.9/**27.9**	53.4/82.6	74.7/95.2	52.8/71.0	27.6/48.2	11.1/21.2	35.1/55.8
$k = 0.3$	15.7/45.9	9.4/26.7	54.1/83.0	77.1/95.4	54.8/71.2	28.5/48.1	11.5/20.9	35.9/**55.9**

Table 8: Effect of PRM scale at $k = 1$, with one training seed per configuration. Metrics are pass@1/pass@8 (%).

PRM Scale	AIME24	AIME25	AMC23	GSM8K	MATH	Minerva	Olympiad	Average
7B PRM	14.8/39.1	**9.8**/27.4	53.8/**83.7**	77.5/95.4	55.1/71.3	27.3/47.2	11.5/21.1	35.7/55.0
72B PRM	**17.2/43.7**	9.1/**27.4**	**57.2**/83.5	**79.8/95.5**	**56.9/71.6**	**29.2/47.5**	**12.0/21.5**	**37.3/55.8**

responses reported in Section 5.2. This behavior is also consistent with our interpretation of residual normalization as credit scaling rather than a direct length penalty: changing the residual-count normalization can alter the learned response length in either direction depending on the resulting advantage structure.

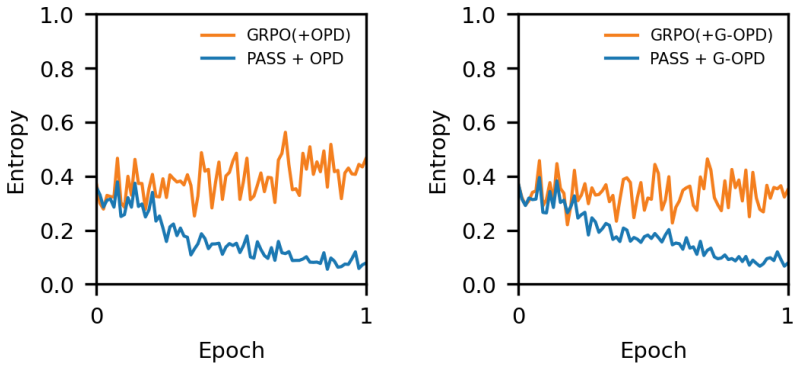

Figure 3: Training-time token-level actor entropy on HotpotQA for the same runs as Figure 2. Left: OPD. Right: G-OPD.

Figure 3 provides a complementary view through the actor's token-level entropy. The PASS runs exhibit a pronounced decrease in entropy over training, whereas the corresponding controls remain at substantially higher entropy. Read together with Figure 2, the resulting policies are therefore not only longer but also more locally concentrated in their token distributions. This joint pattern argues against interpreting the additional tokens as merely high-entropy continuation, although it does not establish that every additional token contributes useful reasoning.

Figure 4 tracks the mean token-level KL divergence between the actor and teacher on the same runs. Both PASS and the controls move closer to the teacher during training, but PASS maintains a higher residual teacher KL. The combination of lower entropy and higher residual KL is consistent with a policy that becomes more confident without simply matching the teacher distribution token by token. In conjunction with the length dynamics, these observations suggest that PASS changes the trajectory-level behavior induced by the teacher-derived signal rather than merely increasing agreement with the teacher.

Taken together, the three diagnostics reveal a coherent training-time pattern across both OPD and G-OPD: PASS moves toward longer responses, lower actor entropy, and a non-vanishing distance from the teacher. These curves are descriptive diagnostics of the resulting policies rather than isolated

Table 9: pass@1 (%) on multi-hop QA under Abs-Max Scaling. PASS has the highest mean among the tested process-signal configurations, but does not exceed outcome-only training. Deltas use the shared outcome-only and corresponding AF-only controls.

Rung	Method	MuSiQue	2Wiki	HotpotQA	Average	$\Delta_{❶}$	$\Delta_{❸}$
❶	GRPO (outcome-only)	44.2 ± 0.3	64.9 ± 0.2	61.9 ± 0.5	57.00	–	–
OPD process signal							
❷	Process-only OPD	38.4 ± 0.4	62.7 ± 0.4	60.5 ± 0.2	53.87	-3.13	-2.03
❸	AF-only OPD	42.1 ± 0.5	63.9 ± 0.2	61.7 ± 0.4	55.90	-1.10	–
❹	**PASS + OPD**	$\mathbf{44.1 \pm 0.5}$	$\mathbf{64.7 \pm 0.2}$	$\mathbf{62.0 \pm 0.3}$	**56.93**	-0.07	$\mathbf{+1.03}$
G-OPD process signal							
❷	Process-only G-OPD	39.5 ± 0.7	63.1 ± 0.3	60.7 ± 0.1	54.43	-2.57	-1.27
❸	AF-only G-OPD	41.7 ± 0.7	63.9 ± 0.4	61.5 ± 0.5	55.70	-1.30	–
❹	**PASS + G-OPD**	$\mathbf{42.7 \pm 0.8}$	$\mathbf{64.0 \pm 0.7}$	$\mathbf{61.8 \pm 0.1}$	**56.17**	-0.83	$\mathbf{+0.47}$

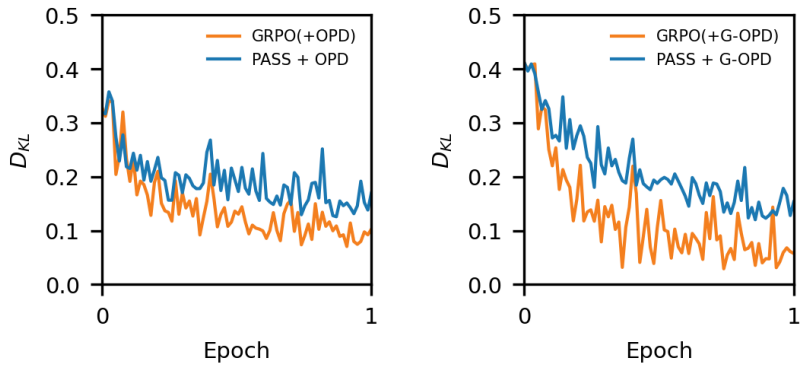

Figure 4: Training-time mean token-level KL divergence between the actor and teacher on HotpotQA for the same runs. Left: OPD. Right: G-OPD.

evidence for any individual PASS operator. In particular, because the comparison from AF-only to PASS jointly introduces shared chunking, RTG, and residual normalization, the observed dynamics should be interpreted as properties of the complete advantage-construction recipe.

F Length Performance Analysis of PASS in the OPD Scenario

Table 10 reports per-benchmark response lengths for the QA configurations. It complements the accuracy results by showing how each estimator allocates output tokens under the two normalizers.

Length and accuracy together. Under Masked-Norm, PASS averages 239.0 tokens for OPD and 249.7 for G-OPD, compared with 186.8 for outcome-only training, approximately 28% and 34% more. Under Abs-Max, PASS averages 239.8 and 221.0 tokens. Thus the QA accuracy gains are associated with additional output tokens, unlike the approximately 2% aggregate reduction on mathematics. Since the QA comparisons do not match length or inference compute, the contribution of additional reasoning tokens remains part of the observed trade-off.

Alternative post-processing. The GR^3-style control averages approximately 163 tokens at 54.50 pass@1, close to AF-only in both measures. PASS instead combines higher mean accuracy with longer outputs in this comparison. These results characterize the accuracy–length behavior of the tested estimators; the fixed-score analysis in Appendix A provides a separate interpretation of residual normalization.

Table 10: Average response length in tokens for the QA configurations. Relative changes use the corresponding AF-only control. The outcome-only lengths are not matched to PASS. The GR^3-style row denotes the adapted advantage-rescaling control.

Operator	Method	MuSiQue	2Wiki	HotpotQA	Average	Δ rel.
–	❶ GRPO (outcome-only)	227 ± 40	183 ± 36	151 ± 26	186.8	–
Masked-Norm	❷ Process-only OPD	185 ± 8	159 ± 10	137 ± 4	160.1	–
	❸ AF-only OPD	188 ± 6	163 ± 8	138 ± 3	163.2	–
	comp. GR^3-style + OPD ($\alpha = 0.33$)	188 ± 3	163 ± 8	138 ± 1	163.2	+0%
	❹ **PASS + OPD**	285 ± 47	231 ± 36	201 ± 25	239.0	+46%
	❷ Process-only G-OPD	155 ± 4	124 ± 4	112 ± 4	130.3	–
	❸ AF-only G-OPD	158 ± 6	128 ± 4	114 ± 3	133.4	–
	❹ **PASS + G-OPD**	300 ± 34	245 ± 24	204 ± 14	249.7	+87%
Abs-Max Scaling	❷ Process-only OPD	188 ± 6	165 ± 6	141 ± 4	164.5	–
	❸ AF-only OPD	203 ± 8	179 ± 9	147 ± 7	176.3	–
	❹ **PASS + OPD**	269 ± 36	241 ± 31	210 ± 23	239.8	+36%
	❷ Process-only G-OPD	155 ± 8	128 ± 6	114 ± 5	132.4	–
	❸ AF-only G-OPD	175 ± 8	148 ± 7	124 ± 4	149.0	–
	❹ **PASS + G-OPD**	251 ± 10	215 ± 6	197 ± 10	221.0	+48%

G Training and Evaluation Protocols

QA training. We train for one epoch and evaluate the resulting checkpoint after 39 rollout steps. Each rollout batch contains 128 prompts with 16 sampled responses per prompt; each training global batch contains 128 response trajectories. A rollout step therefore includes multiple optimizer batches. The actor is Qwen2.5-7B-Instruct, the teacher is Qwen2.5-32B-Instruct, and the initial actor serves as the reference and base policy.

Common settings are a 6144-token sequence limit, at most 2048 newly generated tokens, sampling temperature 1.0 and top-p 0.9, and learning rate 10^{-6} with a constant schedule after 20 warmup steps. We use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and zero weight decay. GRPO uses ratio clipping $\varepsilon_c = 0.2$, reference regularization coefficient 10^{-2}, and zero entropy bonus. The QA reference regularizer is the masked absolute sampled log-probability difference. Teacher log probabilities enter the advantage, without a separate teacher-KL loss term.

QA data and channels. All configurations use the same fixed 5000-question subset of the HotpotQA distractor training data. Prompts contain labeled context paragraphs and sentences, the question, and an instruction to reason using only the context, separate steps by newlines, and end with `Final Answer: [Entity]`. We apply the actor's user-chat template and generation prompt. Training correctness is a binary, case-insensitive exact comparison of the extracted answer, with terminal periods stripped; format is a separate binary check for the final-answer marker. Process-only training uses neither of these two channels in its advantage.

QA seeds and variants. The comparison contains one outcome-only configuration, 12 process-signal configurations (2 signals $\times$ 3 recipes $\times$ 2 standardizers), and one GR^3-style control. Each uses training seeds 42, 1, and 2, giving 42 runs. Training, data-sampling, and generation seeds are changed together. Table 11 summarizes the estimators.

AF-only, PASS, and the GR^3-style control use unit channel weights and the format-dependent fusion rule. Outcome-only sets $w_p = 0$ and $W = 2$, while process-only directly uses the normalized process trace. Normalization epsilon is 10^{-4}, distinct from the CV tolerance. The six G-OPD configurations use $\lambda = 1.25$ and the initial actor as base; the six OPD configurations and the GR^3-style control use the teacher–student sampled log ratio.

GR^3-style length rescaling. We apply $q_i = 1/(1 + \alpha L_i/(\bar{L}_g + \epsilon))$ to the AF-fused advantage, where L_i is the masked response length and $\bar{L}_g$ its same-prompt group mean. We use $\alpha = 0.33$, following a reported setting of GR^3 (Li et al., 2026b). The control returns $q_i a_{i,t}$ directly, without CV or RTG.

Table 11: QA estimator configuration. Process-only ignores the outcome and format channels; their configured weights do not enter that branch. The GR^3-style adaptation rescales the AF output without CV or RTG.

Rung	Configuration	AF	CV	RTG	DL	Standardizer	λ	η	k
❶	Outcome-only ($w_p = 0$)	on	off	off	off	–	–	–	–
❷	Process-only OPD	off	off	off	off	Masked-Norm	–	–	–
❸	AF-only OPD	on	off	off	off	Masked-Norm	–	–	–
❹	PASS + OPD	on	on	on	on	Masked-Norm	–	0.005	0.7
❷	Process-only G-OPD	off	off	off	off	Masked-Norm	1.25	–	–
❸	AF-only G-OPD	on	off	off	off	Masked-Norm	1.25	–	–
❹	PASS + G-OPD	on	on	on	on	Masked-Norm	1.25	0.005	0.7
❷	Process-only OPD	off	off	off	off	Abs-Max	–	–	–
❸	AF-only OPD	on	off	off	off	Abs-Max	–	–	–
❹	PASS + OPD	on	on	on	on	Abs-Max	–	0.005	0.7
❷	Process-only G-OPD	off	off	off	off	Abs-Max	1.25	–	–
❸	AF-only G-OPD	on	off	off	off	Abs-Max	1.25	–	–
❹	PASS + G-OPD	on	on	on	on	Abs-Max	1.25	0.005	0.7
comp.	GR^3-style OPD ($\alpha = 0.33$)	on	off	off	off	Masked-Norm	–	–	–

Mathematical training. The main PASS experiments use seeds 1, 2, and 42, and ORM uses seeds 1111 and 2222. Evaluation uses the checkpoint at optimizer iteration 1248. Training uses 40315 DeepScaleR examples, rollout batches of 512 prompts with 16 retained responses each, training global batch 512, microbatch 1, sequence limit 4096, and response limit 3072. Sampling uses temperature 1.0, top-p 0.9, and unrestricted top-k. We use one data epoch and one update epoch per rollout, ratio clipping 0.2, normalization epsilon 10^{-4}, reference coefficient 10^{-2}, learning rate 10^{-6} with cosine decay, and 20 warmup iterations.

The PRM scores solution steps separated by double newlines, using the positive-class softmax probability at each step-scoring marker. Scores are mapped back to actor-token segments and supplied separately from binary correctness and boxed-answer format rewards. PASS normalizes these three channels and fuses them with $\gamma = 1$ in Equation 4. Its process mask excludes terminal EOS when present, and the resulting advantages are aligned with next-token predictions. Outcome-only GRPO normalizes the sum of correctness and format rewards within each prompt group. The main PASS settings use $\eta = 10^{-8}$ and $k = 1$; the single-seed exponent and PRM-scale studies are described in Appendix C.

Mathematical evaluation. For each problem with n responses and c correct responses, we compute $\widehat{\text{pass@}k} = 1 - \binom{n-c}{k} / \binom{n}{k}$ and then average equally across problems. PASS seeds 1 and 2 use 128 responses per problem, except seed 2 on GSM8K has 126 completed responses; PASS seed 42 uses 32. ORM uses 64, except seed 1111 on AIME24 and AMC23 has 63 completed responses. The two PASS runs with 128-response evaluation use temperature 1.0, top-p 0.95, a 2048-token generation cap, and the model chat template. We report the unweighted mean and sample standard deviation across training runs, computing the seven-benchmark average within each run before its cross-run standard deviation. Deltas are calculated before rounding. Different response counts give different sampling precision within this estimator.

QA evaluation. We target 16 responses per question with temperature 1.0 and top-p 0.9, and use SQuAD-style answer normalization. Some runs have incomplete evaluation shards. Per-dataset summaries weight run metrics by the number of completed evaluation responses and report the corresponding population standard deviation. Average columns give the three dataset summaries equal weight. This weighting accounts for unequal observed coverage; the reported metrics use completed responses only.