THINKING FAST WHEN YOU HAVE NO TIME: TRAINING REASONING LLMS UNDER TIME CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning language models achieve strong accuracy by generating long chain-of-thought (CoT) traces, but these traces dominate inference cost. Prior work reduces this cost using explicit mechanisms such as length penalties, early-exit mechanisms, or distilled shorter traces that constrain output length. In this work, we hypothesize that shorter responses can emerge implicitly from the structure of the training environment by grouping multiple questions into a single prompt together under a shared response budget (measured in response tokens) without any explicit length-based objective. We test this hypothesis by post-training language models using reinforcement learning to solve multi-question quizzes. We construct GSM8K-Quiz, Math-Quiz, and SVAMP-Quiz by grouping questions from established datasets into five-question prompts. Crucially, the budget is not uniformly enforced for each question, so the policy is free to spend more reasoning on harder questions and less on easier ones without any explicit difficulty supervision. We train Qwen2.5-Math-1.5B-Instruct with proximal policy optimization (PPO), separately on GSM8K-Quiz and Math-Quiz. *In quiz evaluation*, under shared response budgets, the quiz-trained PPO policies use about 50% fewer tokens than the base instruct model on GSM8K-Quiz and 35% fewer on Math-Quiz, with comparable accuracy. *In standard single-question evaluation*, the efficiency gains persist: on the GSM8K test set, quiz-trained policies use 9–10% fewer tokens than PPO policies trained on single questions, again with comparable accuracy. Our results demonstrate that training in an environment with implicit time pressure can be sufficient to induce compute-efficient reasoning strategies.

1 INTRODUCTION

Humans adapt how they deliberate to the time pressure and demands of the decisions they face (Klein, 1993). Imagine taking a math test with a strict time limit: answering easy questions quickly leaves more time to work through difficult ones. The aim is not simply to think less, but to spend limited reasoning time where it is most useful. In this paper, we study whether a shared response budget can encourage a similar allocation of reasoning in language models.

Recent models such as DeepSeek-R1 and Qwen2.5-Math (DeepSeek-AI et al., 2025; Yang et al., 2024) show strong reasoning capabilities by generating long chains of thought. Each additional token adds to the cost of answering a question. Existing approaches control this cost through length-aware training objectives, such as L1 (Aggarwal & Welleck, 2025), or inference-time controls, such as budget forcing (Muennighoff et al., 2025). However, reducing the average response length alone does not necessarily mean that a model is allocating more reasoning to harder problems. Taking inspiration from human reasoning under time pressure, we ask whether a shared response budget during training can encourage such allocation:

Can training under a shared response budget induce difficulty-sensitive allocation of reasoning across problems, without any explicit length-based reward?

To answer this question, we introduce *quiz-format reinforcement learning (RL)*, which packs multiple math problems into a single prompt and rewards the model for the number of correct answers. The

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.

response budget applies to the whole quiz, so tokens spent on one question are unavailable for the next, giving the model a reason to shorten some responses while spending more on others. The reward contains no explicit length penalty, and training supplies neither difficulty labels nor target lengths for individual answers. The shared response budget nevertheless creates implicit pressure to use tokens efficiently. This training environment lets us study whether correctness-driven learning under a common constraint produces selective allocation rather than uniform compression.

We construct three quiz-format datasets from existing questions: GSM8K-Quiz, Math-Quiz, and SVAMP-Quiz. We train quiz policies separately on GSM8K-Quiz and Math-Quiz, each starting from the Qwen2.5-Math-1.5B-Instruct base model. We use SVAMP-Quiz only for evaluation. Using proximal policy optimization (PPO), quiz training reduces mean response length relative to the base model by about 50% on GSM8K-Quiz and 35% on Math-Quiz, while maintaining mean quiz accuracy under the same shared response budgets. The token savings also transfer to standard single-question evaluation. On GSM8K, quiz-trained Qwen policies use approximately 9–10% fewer response tokens than their single-question-trained counterparts while maintaining comparable observed accuracy (Section 4.2). The reduction in response length also transfers to SVAMP-Quiz.

In summary, our paper makes three contributions:

- **A training environment with a shared response budget.** We introduce quiz-format RL as an environment-design approach for training efficient reasoning, making token allocation part of a correctness-only learning problem rather than specifying how long individual answers should be.
- **Evidence for difficulty-sensitive allocation.** In controlled question arrangements, quiz-trained PPO policies spend more tokens on questions that the base model finds harder, even after accounting for question position.
- **An evaluation of efficiency and transfer.** Under shared response budgets, quiz training reduces token use with comparable or higher mean quiz accuracy than single-question training in our main comparisons. We also evaluate token savings and accuracy on original single-question benchmarks and other quiz datasets.

2 Related Work

We review methods for controlling reasoning length and allocating computation, followed by related work on RL and supervised reasoning.

Efficient and length-controlled reasoning. Recent work studies how to reduce reasoning cost while retaining accuracy. L1 (Aggarwal & Welleck, 2025) trains models to follow prompt-specified length constraints through length-aware rewards. L1-Exact penalizes deviations from a requested token length, whereas L1-Max uses a soft maximum-length objective that permits shorter responses and rewards length-aware correctness. Arora & Zanette (2025) introduce length-regularized reinforcement learning to reduce unnecessary computation while retaining longer reasoning where needed. These methods study efficiency primarily through individual-question responses. Batch prompting (Cheng et al., 2023) packs several questions into a single prompt so that one model call answers all of them at inference time, without any training. Our quiz format builds on this idea: evaluating the base model on quizzes corresponds to batch prompting, and our contribution is to train in this format. We post-train the model with RL under a shared response budget, so multiple questions compete for the same tokens. We investigate whether this competition can induce difficulty-sensitive token allocation using correctness-only rewards, without an explicit length penalty. Quiz training also reduces token use on individual GSM8K questions relative to single-question PPO and GRPO training (Section 4.2).

Token budgets and adaptive computation. ThinkPrune (Hou et al., 2025) trains with progressively tighter response limits, assigning zero reward when truncation prevents a correct answer from being extracted. It therefore provides a close precedent for learning shorter responses through a hard cap rather than an additive length penalty. We instead use a shared response budget across multiple questions, so spending tokens on one question can displace answers to others. S1 (Muennighoff et al., 2025) combines supervised fine-tuning on reasoning traces with inference-time budget forcing that terminates or extends the thinking process, while Snell et al. (2024) study compute allocation that

depends on prompt difficulty. We study allocation *within a multi-question response*, learned during training without per-question difficulty labels or individual length targets.

Reinforcement learning and supervised reasoning. DeepSeek-R1 and Qwen2.5-Math demonstrate the role of RL in mathematical reasoning. Tülu 3 (Lambert et al., 2024) includes RL with verifiable rewards in an open post-training recipe, and Yeo et al. (2025) examine how supervised initialization, reward design, and training compute affect long-chain reasoning. The design of the length reward also matters: Dai et al. (2025) adapt length penalties to the correctness rate within a sampled group, and Lee et al. (2026) combine correct-only brevity rewards with adaptive controls to address training collapse. We use PPO (Schulman et al., 2017) and group relative policy optimization (GRPO) (Shao et al., 2024); the change is to the training environment, not the optimization algorithm. Because supervised learning can also teach a multi-answer format, our comparisons include direct-answer and chain-of-thought supervised controls to distinguish format adaptation from learning under a shared response budget.

3 Quiz-Format RL: Method and Setup

In quiz-format RL, a *shared response budget* limits the total number of generated response tokens across all questions in a quiz (Figure 1). Our goal is to train the policy to allocate this budget across questions without difficulty labels or per-question target lengths. The model architecture and RL algorithms are unchanged.

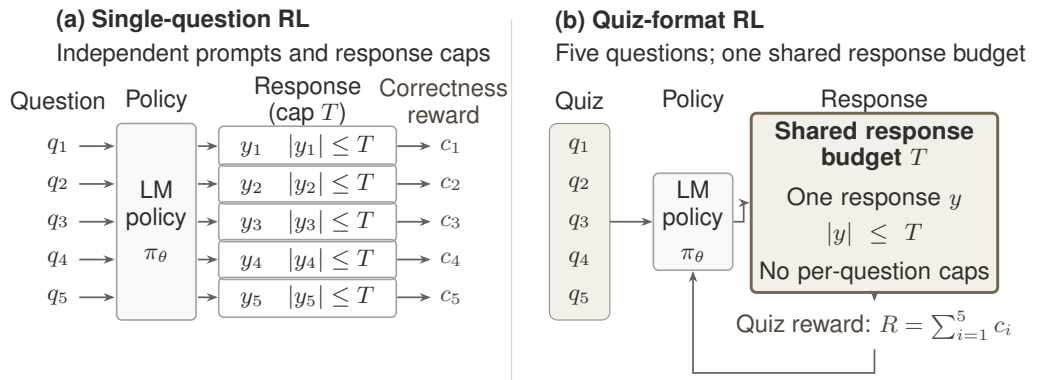

Figure 1: **Single-question and quiz-format training.** Left: each question has a separate response cap of T tokens. Right: five questions use a shared response budget of T tokens. Boxes indicate budget limits, not measured response lengths. LM denotes a language model, with generation policy π_θ.

3.1 Quiz-format dataset construction

We construct **GSM8K-Quiz**, **Math-Quiz**, and **SVAMP-Quiz** by grouping existing questions into quiz prompts. GSM8K-Quiz uses grade-school word problems from GSM8K (Cobbe et al., 2021). Math-Quiz uses MATH12K for training and MATH-500 for evaluation. MATH12K is the 12,000-problem training split of MATH (Hendrycks et al., 2021) released with PRM800K (Lightman et al., 2023), with the remaining 500 problems held out for evaluation as MATH-500. We use the `hiyouga/math12k` Hugging Face release, which provides the problem statements and reference answers for these splits.[1] SVAMP-Quiz uses arithmetic word problems from SVAMP (Patel et al., 2021) for cross-dataset evaluation only. Within each split, we group consecutive problems into five-question quizzes. The quiz test sets contain 1,315 GSM8K questions, 500 MATH-500 questions, and 300 SVAMP questions; we retain their individual-question mappings for paired analyses. Processed split counts are in Appendix A.1.

[1] `https://huggingface.co/datasets/hiyouga/math12k`

Quiz prompts list five questions and request final answers as "Question i: `\boxed{...}`". The maximum response length determines the shared response budget. For analysis, we construct difficulty labels from the base model's empirical solve rates on held-out questions; these labels are never supplied in training or evaluation prompts.

3.2 Correctness reward under a shared response budget

For a quiz response y with reference answers $g_1, \ldots, g_n$, let $c_i(y, g_i) \in \{0, 1\}$ denote the correctness score of the extracted answer to question i. The task reward is

$$R(y, g_{1:n}) = \sum_{i=1}^{n} c_i(y, g_i), \qquad R \in \{0, \ldots, n\}, \tag{1}$$

with $n = 5$ in most experiments. We extract answers as the last boxed answer under each question heading, and score them using math-verify (Hugging Face, 2025). Missing answers receive zero credit. Single-question training uses a binary correctness reward. Details of answer extraction, including how we score responses with missing headings or unboxed answers, are in Appendix A.2.

The reward contains no explicit length penalty or supervision of the reasoning trace. The shared response budget nevertheless creates *implicit* length pressure: tokens spent on one question leave fewer for later answers and their potential correctness reward. The budget covers the entire response, without fixed per-question allowances. Time pressure here denotes a token constraint, not a wall-clock deadline.

3.3 Training and evaluation

We use the suffix **Single** for policies trained on individual questions and **Quiz** for policies trained on multi-question quizzes, prefixed by the optimization algorithm (e.g., PPO-Single and PPO-Quiz). These labels specify training format; evaluation format is stated separately. Quiz policies train separately on GSM8K-Quiz and Math-Quiz; Single policies train on the corresponding GSM8K and MATH12K source questions. All start from the base model and use verl (Sheng et al., 2024). PPO (Schulman et al., 2017) uses an actor–critic with generalized advantage estimation (GAE) (Schulman et al., 2015), and GRPO (Shao et al., 2024) uses group-relative advantages. PPO samples one completion per prompt and GRPO eight. GRPO includes Kullback–Leibler (KL) divergence relative to a reference policy in its loss, with coefficient 0.04; PPO uses no explicit reference-policy KL penalty. Within each question source and algorithm, Single and Quiz share the starting checkpoint, advantage estimator, number of completions per prompt, KL setting, training response cap, and number of updates.

The training response cap (max response length) is 512 tokens for GSM8K-derived data and 1,024 for MATH12K-derived data, applied per question for Single and per quiz for Quiz; in the quiz setting, this cap is the shared response budget. In the main PPO/GRPO comparisons, we train each method for 500 updates, each processing 40 problems: 40 single-question prompts or eight five-question quizzes. We report means and standard deviations, specifying the contributing seeds with each result. Appendix A.3 summarizes these shared settings and the evaluation configuration.

Evaluation protocol. We use greedy decoding with a response cap of 1,024 tokens for GSM8K-Quiz and SVAMP-Quiz and 2,048 for Math-Quiz; in quiz evaluation, this cap is the shared response budget. Single-question tests apply the same cap to each individual question (1,024 for GSM8K and SVAMP, 2,048 for MATH-500). Caps are identical across policies unless a budget sweep is specified.[2] These larger caps test whether conciseness persists beyond the training budget; they can still cause truncation. Quiz accuracy is per-question correctness, not the fraction of fully correct quizzes. Mean length counts tokens per generated response: one answer in single format or the whole quiz in quiz format.

[2]The 4,096-token context covers both prompt and response. GSM8K-Quiz and SVAMP-Quiz prompts leave room for a 3,072-token response cap, whereas longer Math-Quiz prompts require a lower cap to retain every input in full (Appendix A.3).

4 EXPERIMENTS

The experiments address three questions: (1) Under shared response budgets, how do quiz-trained policies compare with single-question-trained policies that use explicit length controls in response length and accuracy? (2) Can supervised adaptation to the quiz format explain these results? (3) Does token use correlate with question difficulty, measured by the base model's solve rate, after accounting for question position? We also evaluate models on the original single-question benchmarks and test robustness to response budgets, quiz sizes, and starting models.

4.1 QUIZ RESPONSE LENGTH AND ACCURACY

We first evaluate on the held-out five-question quizzes in GSM8K-Quiz and Math-Quiz, using policies trained on questions from the corresponding GSM8K and MATH12K sources. The Math-Quiz evaluation questions come from MATH-500 (Lightman et al., 2023). Table 1 compares the base model, single-trained and quiz-trained PPO and GRPO policies, and explicit length controls. The length-penalty (LP) control LP-0.5 uses a coefficient of $\alpha = 0.5$. L1-exact and L1-max adapt the length-control rewards of Aggarwal & Welleck (2025) to PPO. Appendix F defines these objectives and reports the full LP sweep and three-seed LP-0.3 comparison. Training and evaluation follow Section 3.3.

Training data	GSM8K / GSM8K-Quiz		MATH12K / Math-Quiz	
Quiz evaluation	GSM8K-Quiz		Math-Quiz	
Method	Accuracy (%)	Tokens	Accuracy (%)	Tokens
Base model	68.06	924.1	53.60	1,403.6
PPO-Single	70.54 ± 2.27	904.8 ± 21.5	54.13 ± 2.44	$1{,}293.4 \pm 22.6$
PPO-Quiz	72.07 ± 1.34	$\mathbf{458.5 \pm 19.4}$	59.27 ± 1.60	$\mathbf{916.4 \pm 41.7}$
PPO-Single + LP-0.5	73.92 ± 1.42	737.1 ± 36.2	54.43 ± 1.37	$1{,}258.1 \pm 27.6$
PPO-Single + L1-exact	70.42 ± 0.75	918.4 ± 18.1	51.90 ± 2.12	$1{,}319.4 \pm 35.5$
PPO-Single + L1-max	74.02 ± 3.45	682.7 ± 83.4	57.30 ± 0.14	$1{,}309.6 \pm 27.6$
GRPO-Single	71.67 ± 2.10	928.7 ± 27.9	52.00 ± 0.57	$1{,}318.6 \pm 29.5$
GRPO-Quiz	$\mathbf{74.41 \pm 1.02}$	496.4 ± 5.9	$\mathbf{63.90 \pm 0.14}$	$1{,}115.8 \pm 26.6$

Table 1: In-distribution quiz evaluation. Quiz rows train on GSM8K-Quiz or Math-Quiz; the other trained rows use individual questions from GSM8K or MATH12K, respectively. Evaluation uses 1,315 questions with a 1,024-token shared response budget on GSM8K-Quiz and 500 questions with a 2,048-token budget on Math-Quiz. Accuracy is per question; tokens are per five-question quiz. We report mean $\pm$ sample standard deviation across training seeds 42, 123, and 456 for all methods except the base model, which is evaluated once. Bold marks the highest mean accuracy or lowest mean length.

Shorter responses while maintaining quiz accuracy. Under the shared response budgets in Table 1, PPO-Quiz reduces mean response length by approximately 49% on GSM8K-Quiz and 29% on Math-Quiz relative to PPO-Single. The corresponding mean accuracy differences are $+1.53$ and $+5.14$ percentage points. Both PPO and GRPO quiz-trained policies use fewer tokens than the base model and their single-trained counterparts while matching or exceeding their reported mean quiz accuracies. These comparisons describe the reported means; the table also shows variation across training seeds.

Comparison with explicit length controls. The LP reward reduces the credit for a correct answer in proportion to its response length; incorrect answers receive zero reward. Quiz training instead couples the questions through a shared response budget: tokens spent on one question leave fewer for the others, creating efficiency pressure through the correctness reward. On GSM8K-Quiz, GRPO-Quiz uses 496.4 tokens compared with 737.1 for LP-0.5, approximately 33% fewer, with similar mean accuracy (74.41% versus 73.92%). PPO-Quiz uses 458.5 tokens with 72.07% accuracy. Across-seed variation is reported in the table.

In the single-seed LP sweep, LP-0.3 reaches 76.4% GSM8K-Quiz accuracy with 827 tokens per quiz (Table F.2). LP settings that match or exceed PPO-Quiz's quiz accuracy use longer responses. The base model with a concise prompt at the same 1,024-token shared response budget achieves 68.4% accuracy with 932 tokens per quiz (Appendix F), compared with PPO-Quiz's 72.1% and 458 tokens.

Cross-dataset quiz evaluation. Under shared response budgets on other datasets, quiz-trained policies also use fewer tokens than their single-trained counterparts (Table D.1). PPO-Quiz trained on GSM8K-Quiz uses 32% fewer tokens on Math-Quiz than PPO-Single trained on individual GSM8K questions, with mean accuracy of 58.6% versus 52.5%. PPO-Quiz trained on Math-Quiz uses 31% fewer tokens on GSM8K-Quiz than PPO-Single trained on individual MATH12K questions, with mean accuracy of 76.2% versus 69.5%. GRPO also uses fewer tokens without reducing mean quiz accuracy in both directions. On SVAMP-Quiz, quiz-trained policies use 29–55% fewer tokens. PPO's mean accuracy remains similar, while GRPO's is 4.5 percentage points lower after training on GSM8K-Quiz and 2.50 points higher after training on Math-Quiz. Appendix D gives the evaluation details.

4.2 Single-question benchmark evaluation

Here, each prompt supplies one benchmark question rather than a quiz.

Conciseness transfers to single-question inference. On the GSM8K single-question test, PPO-Quiz and GRPO-Quiz generate 27.9 and 29.7 fewer tokens per response than their single-trained counterparts, respectively (95% paired bootstrap confidence intervals (CIs) for Quiz minus Single: $[-32.0, -23.9]$ and $[-33.7, -26.0]$). Their observed accuracies remain close: 85.06% and 86.66%, compared with 85.52% and 85.44% for PPO-Single and GRPO-Single (Table C.1). These results show that shorter responses persist when questions are presented individually, without a shared response budget. In cross-dataset single-question evaluations, quiz-trained policies reduce mean response length by approximately 5–9% relative to their single-trained counterparts, with accuracy differences ranging from -1.7 to $+2.0$ percentage points (Table 2). Appendix C describes the evaluation setup and paired bootstrap procedure.

Training data	GSM8K / GSM8K-Quiz				MATH12K / Math-Quiz			
Single-question test	MATH-500		SVAMP		GSM8K		SVAMP	
Policy	Acc. (%)	Mean tokens	Acc. (%)	Mean tokens	Acc. (%)	Mean tokens	Acc. (%)	Mean tokens
Base model	74.6	563	91.7	196	84.6	306	91.7	196
PPO-Single	73.6	566	92.7	191	85.3	300	91.3	195
PPO-Quiz	73.8	521	91.0	175	85.1	281	91.7	177
GRPO-Single	75.0	554	91.7	191	85.1	303	90.3	199
GRPO-Quiz	74.6	508	92.7	176	84.2	288	92.3	181

Table 2: Cross-dataset evaluation on original single-question benchmarks. Quiz rows train on GSM8K-Quiz or Math-Quiz; the other trained rows use individual questions from GSM8K or MATH12K, respectively. Trained policies use seed 42 and 500 training steps. The base model has no additional training. All use greedy decoding. Accuracy is per question; tokens are the mean number of generated response tokens. MATH-500 has 500 questions and a 2,048-token response cap. GSM8K (1,319 questions) and SVAMP (300 questions) use 1,024-token caps.

Accuracy trade-offs on individual questions. The single-question trade-off differs across models. For DeepSeek-R1-Distill-Qwen-1.5B at a 1,024-token response cap, PPO-Quiz has 75.0% accuracy versus 78.3% for LP-0.3 (Table J.1). Its first answer is shorter, but the policy sometimes continues after answering the supplied question by generating further questions and answers. This makes the full response longer, so first-answer savings need not reduce total generation cost.

In the following sections, we investigate whether the improvements from quiz training arise simply from learning the multi-answer output format (Section 4.3) or from reallocating reasoning according to question difficulty (Section 4.4).

4.3 Supervised adaptation to the quiz format

It is possible that quiz training improves performance because it simply teaches the model to produce several numbered answers within one response. To rule out this possibility, we compare with two supervised fine-tuning (SFT) controls initialized from the same base model and trained on GSM8K-Quiz. SFT-direct uses targets containing only five numbered final answers; SFT-CoT uses

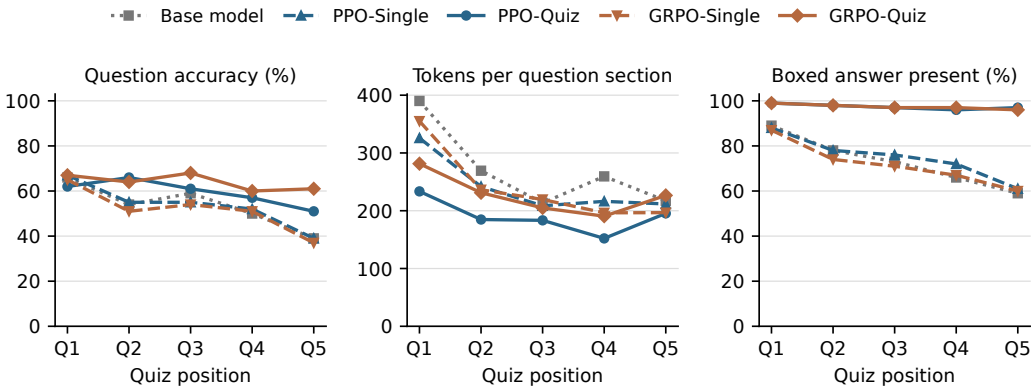

Figure 2: Math-Quiz evaluation by question position. The base model and policies trained on questions from MATH12K use greedy decoding on the same 100 five-question quizzes with a shared response budget of 2,048 tokens. Token means include all quizzes, with zero for undetected question sections. Appendix I gives the measurement details.

chain-of-thought (CoT) targets built from GSM8K's human-written reference solutions (Cobbe et al., 2021), followed by boxed answers. Quiz RL requires only final-answer correctness and learns from its own sampled responses, without reference reasoning traces. We train both SFT controls for three epochs to approximately match PPO-Quiz's question exposure, giving 4,320 versus 4,000 quiz presentations (Appendix A.4). Both use a learning rate of 10^{-5}, batch size 32, and maximum sequence length 2,048, without an RL objective. Evaluation uses the same 263 five-question GSM8K-Quiz test quizzes and 1,024-token shared response budget as Table 1.

Method	Accuracy (%)	Tokens/quiz	At cap (%)
SFT-direct	26.01	51.7	0.0
SFT-CoT	64.94	496.5	3.8

Table 3: Supervised controls evaluated on GSM8K-Quiz. Accuracy is per question; length is per quiz. "At cap" is the fraction of responses with recorded length at least 1,024 tokens.

SFT-direct produces very short responses but reaches only 26.01% accuracy. Adding reasoning supervision raises accuracy to 64.94%, with a mean length of 496.5 tokens (Table 3). Under the same shared response budget, PPO-Quiz achieves 72.07% mean quiz accuracy with 458.5 tokens: 7.12 percentage points higher accuracy and approximately 8% fewer tokens than SFT-CoT. These supervised controls therefore do not reproduce PPO-Quiz's accuracy–length trade-off. This comparison does not isolate a causal effect of token allocation: the training objectives and supervision differ, and SFT-CoT still reaches the cap on some quizzes. We next test difficulty sensitivity directly through controlled question arrangements.

4.4 Difficulty-sensitive allocation

We account for where a question appears in the quiz because tokens spent on earlier answers leave fewer for later ones.

Performance across quiz positions. Under the shared response budget for Math-Quiz, quiz-trained policies use shorter responses while maintaining overall quiz accuracy relative to their single-trained counterparts (Figure 2). Their largest accuracy gains occur on the fifth question (Q5): 12 and 24 percentage points for PPO and GRPO, respectively. GSM8K-Quiz also shows higher overall accuracy and better late-question answer coverage, while single-trained policies score higher at some early positions (Appendix I). We next use controlled question arrangements to test difficulty sensitivity.

We analyze the PPO-Quiz policy trained on each quiz dataset using controlled evaluation quizzes. These variants of GSM8K-Quiz and Math-Quiz regroup held-out source questions from GSM8K

and MATH-500 to control difficulty composition and ordering. A question's difficulty is defined by the base model's empirical solve rate p over eight stochastic single-question samples: easy questions have $p \geq 0.75$ and hard questions have $p \leq 0.25$. Questions between these thresholds are medium difficulty. These labels are used only to construct these evaluation quizzes and are never provided during training or in the evaluation prompt. Shared response budgets remain 1,024 tokens for GSM8K-Quiz variants and 2,048 for Math-Quiz variants.

Controlling for position. On randomized easy/hard packs, we regress per-question response length on hardness $h = 1 - p$ with question-position fixed effects (Appendix B). PPO-Quiz has positive hardness slopes on both datasets, with 95% quiz-clustered confidence intervals excluding zero (Table B.1), so it uses more tokens for harder questions even after accounting for their position.

Changing difficulty at a fixed position. We compare pairs of quizzes with the same four medium-difficulty questions in positions 1–4 and an easy or a hard question at position 5. We do not reuse questions across pairs. The available 91 GSM8K and 59 MATH-500 medium questions therefore yield 22 and 14 pairs, respectively (Appendix B). For GSM8K-Quiz variants, PPO-Quiz uses an average of 80.8 tokens for the easy final question and 97.3 for the hard one. For Math-Quiz variants, the corresponding averages are 106.7 and 189.1 tokens (Figure 3). We identify the Q5 section by its heading and assign zero tokens when the heading is missing, so every quiz contributes to the average. The Q5 heading is detected in 21 of 22 GSM8K-Quiz responses and 13 of 14 Math-Quiz responses in each condition. The longer responses to hard final questions support difficulty-sensitive allocation with the preceding questions and final position held fixed. The easy and hard versions use different final questions, so the comparison reflects differences in content as well as difficulty.

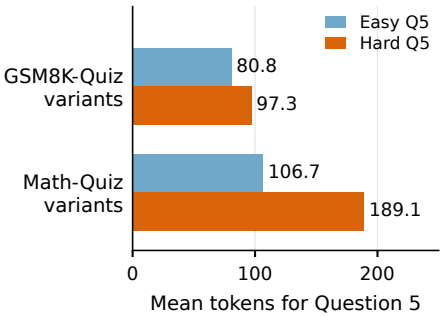

Figure 3: PPO-Quiz mean Q5 tokens in paired quizzes with medium-difficulty Q1–Q4 fixed and an easy or hard Q5 (22 GSM8K-Quiz and 14 Math-Quiz pairs). Each policy is evaluated on variants of its own training quiz dataset; a missing Q5 heading counts as zero tokens.

Alternating difficulty across positions. Hard questions also receive more tokens in both easy–hard–easy–hard–easy and hard–easy–hard–easy–hard packs on both datasets. Appendix B reports these results alongside all-easy/all-hard quizzes.

Average length across positions. After shuffling question order, PPO-Quiz shows smaller differences in average response length across Q1–Q5 than the base model (Table B.3). The separate difficulty tests show that PPO-Quiz still uses more tokens for harder questions.

Together, these controlled evaluations support difficulty-sensitive allocation beyond a fixed position-only pattern. We count all response tokens, including the answer and formatting.

Table B.4 in Appendix B compares response lengths across Math-Quiz difficulty levels at shared response budgets of 512 and 1,024 tokens.

4.5 Additional models

Under shared response budgets on GSM8K-Quiz, PPO-Quiz reduces mean quiz response length by more than half on DeepSeek-R1-Distill-Qwen-1.5B and Llama-3.2-3B-Instruct relative to their base models, with similar or higher observed quiz accuracy (Table H.1). Appendix J reports single-question results for both models and examines how changing the response cap affects DeepSeek's results.

4.6 Robustness to training and evaluation settings

Varying the shared response budget. We evaluate PPO-Single, PPO-Quiz, and LP-0.3 on GSM8K-Quiz with shared response budgets of 256, 512, 1,024, and 2,048 tokens, keeping each policy's

checkpoint fixed across budgets (Figure 4). At a shared response budget of 512 tokens, PPO-Quiz reaches 69.66% quiz accuracy, compared with 36.43% for PPO-Single and 44.26% for LP-0.3. From 512 to 2,048 tokens, PPO-Quiz maintains about 70–71% accuracy while using 423–446 tokens per quiz. At 1,024 and 2,048 tokens, LP-0.3 achieves higher accuracy than PPO-Quiz but uses more tokens. Appendix E reports the full results.

Varying quiz size. We train PPO-Quiz on variants of GSM8K-Quiz containing two, three, or five questions per quiz and test it on variants with two, three, five, or ten questions per quiz. With a shared response budget of 1,024 tokens, the model trained on three-question quizzes achieves 53.39% accuracy on ten-question quizzes, compared with 32.06% for PPO-Single. This shows that PPO-Quiz can handle longer quizzes than those seen during training. We also test a shared response budget that grows with the number of questions, keeping the average token allowance per question roughly the same. Appendix G reports results for both settings.

5 Discussion and Limitations

Under a shared response budget, spending more tokens on one question leaves fewer for the others. The correctness reward therefore encourages efficient use of the available tokens. Our results support this training environment as a way to reduce token use while retaining similar accuracy in many comparisons. The resulting responses also remain sensitive to difficulty. Experiments controlling question position show that quiz-trained PPO policies use more tokens on harder questions, although they do not isolate how much of the quiz performance under a shared response budget is caused by changes in allocation. The comparison with explicit length controls shows that the resulting accuracy–length trade-off depends on the evaluation budget and task.

Model scale. Our experiments cover Qwen2.5-Math-1.5B-Instruct, DeepSeek-R1-Distill-Qwen-1.5B, and Llama-3.2-3B-Instruct, extending the evaluation across model variants up to 3B parameters. Results for DeepSeek and Llama are reported in Appendix H (quiz evaluation) and Appendix J (single-question evaluation). We have not evaluated 8B, 16B, 32B, or larger models. Whether training with a shared response budget yields similar benefits as model capacity increases remains an open question.

Task scope. The experiments focus on mathematical reasoning with automatically verifiable final answers. Applying the approach to code generation, planning, or open-ended tasks would require suitable correctness signals and a choice of which tasks should use a shared response budget. Transfer to these domains remains untested.

Measuring efficiency. We measure inference efficiency using response-token counts. We have not directly measured inference runtime or energy use.

6 Conclusion

We asked whether a shared response budget could encourage efficient, difficulty-sensitive reasoning without explicitly rewarding shorter responses. We introduced quiz-format RL, where five questions share one prompt and one response budget, and the reward is the number of correct answers. Tokens spent on one question are unavailable for the others. This creates implicit time pressure: no length penalty, no difficulty labels, and no per-question target lengths. In our main Qwen experiments, quiz-trained policies use fewer tokens and achieve higher mean quiz accuracy than the base and single-question-trained models under the same evaluation budgets. Token savings also extend to single-question and cross-dataset evaluation. In controlled quizzes, PPO-Quiz uses more tokens for questions the base model finds harder, even after accounting for question position. Our supervised controls do not reproduce PPO-Quiz's accuracy–length trade-off. Correctness-driven training under a shared budget can thus produce shorter responses that remain sensitive to difficulty, without specifying a length for each answer. Here the pressure shapes only how much reasoning each math question receives. As a next step, we plan to investigate how time pressure in more complex agent environments can lead to different reasoning strategies.

AI USE STATEMENT

During the preparation of this manuscript, the authors used AI tools solely for grammar correction, language editing, and proofreading. The authors reviewed and verified all AI-assisted edits and take full responsibility for the final manuscript.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025. URL https://arxiv.org/abs/2503.04697.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. Batch prompting: Efficient inference with large language model APIs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 792–810, 2023. URL https://aclanthology.org/2023.emnlp-industry.74/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems, 2021.

Muzhi Dai, Shixuan Liu, and Qingyi Si. Stable reinforcement learning for efficient reasoning. *arXiv preprint arXiv:2505.18086*, 2025.

DeepSeek-AI, Daya Guo, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. ThinkPrune: Pruning long chain-of-thought of LLMs via reinforcement learning. *arXiv preprint arXiv:2504.01296*, 2025.

Hugging Face. Math-Verify. Software repository, 2025. URL https://github.com/huggingface/Math-Verify.

Gary A. Klein. A recognition-primed decision (RPD) model of rapid decision making. In Gary A. Klein, Judith Orasanu, Roberta Calderwood, and Caroline E. Zsambok (eds.), *Decision Making in Action: Models and Methods*, pp. 138–147. Ablex Publishing Corporation, 1993.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Jungseob Lee, Seungyoon Lee, Seongtae Hong, Minhyuk Kim, Chanjun Park, and Heuiseok Lim. Beyond penalizing mistakes: Stabilizing efficiency training in large reasoning models via adaptive correct-only rewards. *arXiv preprint arXiv:2606.22716*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Niklas Muennighoff, Zitong Yang, Weijia Shi, et al. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/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. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, et al. HybridFlow: A flexible and efficient RLHF framework, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

An Yang et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. *arXiv preprint arXiv:2502.03373*, 2025.

A Dataset and implementation details

A.1 Dataset splits and quiz construction

GSM8K contains grade-school math word problems (Cobbe et al., 2021). MATH12K and MATH-500 use competition mathematics problems from MATH (Hendrycks et al., 2021); MATH-500 is the 500-problem held-out subset used by Lightman et al. (2023). Their split combines the 7,500 original MATH training problems with 4,500 original test problems, reserving the remaining 500 for evaluation. The released training split contains 12,000 problems; our processed training data contain 11,998 before quiz grouping (Table A.1). For cross-dataset evaluation, we also use SVAMP, which tests variations of arithmetic word problems (Patel et al., 2021). We construct SVAMP-Quiz from 300 SVAMP questions, grouped into 60 five-question quizzes, for evaluation only.

Table A.1 gives the source splits and sizes of GSM8K-Quiz, Math-Quiz, and SVAMP-Quiz. Within each source split, consecutive groups of five questions form quizzes; incomplete final groups are dropped. We retain the mapping from each quiz item to its original question for paired analyses. The GSM8K single-question test set contains 1,319 questions, of which 1,315 appear in the GSM8K-Quiz test split. All 500 MATH-500 questions appear in the Math-Quiz test split and in single-question evaluation. For each training question source, Single policies train on individual questions and Quiz policies on the corresponding quiz dataset. Each policy is initialized separately from the base model.

Quiz dataset	Source	Split	Questions	Quizzes	Used questions
GSM8K-Quiz	GSM8K	Train	7,473	1,494	7,470
	GSM8K	Test	1,319	263	1,315
Math-Quiz	MATH12K	Train	11,998	2,399	11,995
	MATH-500	Test	500	100	500
SVAMP-Quiz	SVAMP	Test	300	60	300

Table A.1: Quiz datasets and their source splits. Each quiz contains five questions; used questions exclude incomplete final groups. SVAMP-Quiz is used only for evaluation.

A.2 Answer extraction and scoring

For quiz responses, we identify "Question i", "Qi", or "Q.i" headings in question order and score the text between them using math-verify. We first check the last boxed answer, with a numeric fallback for plain-number references, and check unboxed text when no complete boxed answer is available. Missing headings do not invalidate earlier answers. When the text between two detected headings covers several questions, we assign boxed answers in question order if their count matches the number of questions. Otherwise, we score only the question named by the first heading, using the text up to its first box if one is present; the final segment uses its last boxed answer. If no headings are found, we match answers by position only when the response contains exactly n boxes. Answers that cannot be assigned to a question or verified receive no credit.

For Qwen cross-dataset single-question evaluation, accuracy is per question and response length counts generated tokens, including reasoning, answers, and formatting. The response cap applies to the entire response. Appendix J defines the separate first-answer and full-generation length measures used for DeepSeek.

A.3 Training and evaluation configuration

Table A.2 collects the main training settings for Single and Quiz within each algorithm. We initialize a separate policy for each training question source, algorithm, and format. The training cap applies to one question in Single and defines the shared response budget for the entire five-question response in Quiz. There are no fixed per-question allowances within a quiz. Evaluation caps and contributing seeds are specified with each result, including the budget and quiz-size sweeps. Both algorithms use AdamW with weight decay 0.01 and constant learning rates, without warm-up. The training backend uses fully sharded data parallelism (FSDP); PPO uses generalized advantage estimation (GAE).

Context-length constraint. The tokenized prompt and generated response must together fit within the model's 4,096-token context. Including instructions, chat formatting and the assistant prefix, the longest five-question test prompts are 606 tokens on GSM8K-Quiz, 375 on SVAMP-Quiz and 1,159 on Math-Quiz. Thus, a 3,072-token response cap fits every GSM8K-Quiz and SVAMP-Quiz test quiz, but exceeds the available context for two Math-Quiz prompts longer than 1,024 tokens. The 2,048-token Math-Quiz response cap accommodates all test prompts without input truncation. These lengths describe context feasibility; the response caps used for each experiment are reported with its results.

Setting	PPO	GRPO
Initial model	Qwen2.5-Math-1.5B-Instruct	
Training framework	verl v0.7.0	
Training backend	FSDP	
Training question sources	GSM8K and MATH12K, trained separately	
Advantage estimation	GAE (actor–critic)	Group-relative
Actor learning rate	10^{-6}	10^{-6}
Critic learning rate	10^{-5}	—
Training sampling temperature	1.0	1.0
Completions per prompt	1	8
Prompt batch (Single / Quiz)	40 / 8	40 / 8
Minibatch, in prompts (Single / Quiz)	20 / 4	20 / 4
Reference-policy KL	No explicit penalty	In policy loss, coefficient 0.04
Training updates	500	500
Questions per Single / Quiz prompt	1 / 5	1 / 5
Training cap: GSM8K source	512	512
Training cap: MATH12K source	1,024	1,024
Main evaluation decoding	Greedy	Greedy
Evaluation cap: GSM8K source	1,024	1,024
Evaluation cap: MATH-500 source	2,048	2,048

Table A.2: Main training and evaluation settings for Single and Quiz policies. Batch sizes count prompts before completion sampling. Source labels apply to both individual questions and their quiz datasets. Caps are in tokens per generated response. Budget and quiz-size sweeps specify their changes separately.

Prompt template. Single and Quiz use the same instruction block in a user message, followed by one or five numbered question blocks. The model's chat template is applied before generation. The shared message structure is:

```
Please solve each question step by step using mathematical reasoning
and equations; LaTeX notation is allowed.
Do not write or rely on Python, SymPy, code, or pseudocode.
Answer every question without exception, in order, label each response
with its question number, and put each final answer within \boxed{}.

Question 1:
[question text]
...
Question n:
[question text]
```

The template contains n question blocks, separated by blank lines; the bracketed text and ellipsis above are placeholders. LP uses this same template. L1 training inserts "Think for b tokens." as the final line of the instruction block, before the questions, with b defined in Appendix F.

A.4 SUPERVISED FORMAT CONTROLS

SFT-direct and SFT-CoT use the same seed-42 split of 1,494 GSM8K-Quiz training examples, with 30 reserved for validation and 1,464 for training. Each epoch uses 45 batches of 32 quizzes, with the final partial batch dropped. Training for three epochs gives 135 updates and $3 \times 45 \times 32 = 4{,}320$ quiz presentations, compared with $500 \times 8 = 4{,}000$ for PPO-Quiz. Since each quiz contains five questions, these correspond to 21,600 and 20,000 question presentations, respectively. The methods therefore see similar numbers of problems, although their update counts and training costs differ.

For SFT-CoT, we combine the human-written reference solutions under numbered question headings, remove calculator annotations, and box each final answer. We use all reference solutions without filtering by length. Quiz RL learns from its own generated responses, using only final-answer correctness for the reward.

B DETAILS OF THE ALLOCATION PROBES

Difficulty labels. We sample eight single-question responses from the base model (Qwen2.5-Math-1.5B-Instruct) for each held-out question, using temperature 0.8, top-$p = 1$, and no top-k restriction. Response caps are 1,024 tokens for GSM8K and 2,048 for MATH-500. Each sample is scored with the single-question answer verifier; the empirical solve rate is $p = (\text{correct samples})/8$. We label questions easy when $p \geq 0.75$, hard when $p \leq 0.25$, and medium otherwise. These labels are fixed across policies and used only to construct the analysis sets; they are absent from training and evaluation prompts. The allocation probes below are controlled variants of GSM8K-Quiz and Math-Quiz, regrouped or reordered from held-out GSM8K and MATH-500 source questions. The PPO-Quiz policies use training seed 42.

Measurement. We segment responses using the detected "Question i" or "Qi" headings. Each span extends from its heading to the next detected heading or the end of the response, and is tokenized separately with the model tokenizer. The preamble before the first heading is kept separate. Consequently, per-question length includes headings, reasoning, and final-answer text, and separate tokenization can introduce boundary differences from whole-response token counts. A missing heading receives zero tokens in descriptive averages and is excluded from the regression. Heading detection indicates an identifiable span, not a complete or correct answer. Missing intermediate headings can also make a detected span cover text intended for more than one question.

Randomized packs and regression. The randomized pack builder shuffles the easy and hard pools and chooses between them at each position until one pool is exhausted, after which it draws from the remaining pool. It uses each selected question once and drops the incomplete final quiz. This yields 245 GSM8K-Quiz and 88 Math-Quiz probe examples; it does not guarantee balanced difficulty at each position. For each policy and dataset, the fitted model is

$$L_{qi} = \alpha + \beta(1 - p_{qi}) + \sum_{j=2}^{5} \gamma_j \mathbf{1}[i = j] + \epsilon_{qi}, \tag{2}$$

where L_{qi} is the token count of a detected segment at position i in quiz q, and p_{qi} is the base model's solve rate for that question. We fit the model by ordinary least squares on segments with detected headings. The slope β measures tokens per unit increase in hardness, rather than a binary hard–easy group difference.

Confidence intervals. Questions within a quiz have a shared response budget, so their measurements need not be independent. We cluster standard errors by quiz and report normal-approximation 95% intervals, $\hat{\beta} \pm 1.96\,\mathrm{SE}_{\mathrm{quiz}}(\hat{\beta})$. These intervals describe uncertainty across evaluation quizzes with each evaluated model held fixed. Different rates of missing headings mean that the policies' fitted samples differ; the slopes are not a paired estimate of the effect of training.

Position-controlled slopes. Table B.1 reports the fitted slopes and segment counts. PPO-Quiz has positive hardness slopes on both quiz variants, with 95% quiz-clustered confidence intervals excluding zero. It therefore uses more tokens for harder questions even after accounting for their

position, showing difficulty-sensitive allocation in its shorter responses (Section 4.4). The base model also has positive slopes; its responses are longer, and its fitted sample differs, so the two rows describe each policy separately.

Quiz variant	Policy	Segments	Hardness slope	95% CI
GSM8K-Quiz	Base model	1,146	51.51	[24.63, 78.39]
	PPO-Quiz	1,208	34.75	[7.52, 61.99]
Math-Quiz	Base model	362	255.05	[148.50, 361.59]
	PPO-Quiz	439	75.89	[27.89, 123.89]

Table B.1: Position-controlled allocation on randomized easy/hard quiz variants. Slopes are tokens per unit increase in $1 - p$, where p is the base model's solve rate. Intervals quantify uncertainty across evaluation quizzes with each evaluated model held fixed, not variation across training seeds. Missing headings are excluded from the regression; segment counts therefore differ across policies. Measurement, pack construction and interval estimation are detailed in Appendix B.

Paired quizzes at a fixed position. We shuffle the medium, easy, and hard question pools, then divide the medium questions into groups of four. Each group forms Q1–Q4 of two quizzes in the same order. One quiz ends with an easy Q5 and the other with a hard Q5. No question is reused across different pairs. There are 91 medium questions in GSM8K and 59 in MATH-500, giving $\lfloor 91/4 \rfloor = 22$ and $\lfloor 59/4 \rfloor = 14$ pairs, respectively. Both datasets have enough easy and hard questions for these pairs. Three medium questions remain unused in each dataset. Each pair contributes two quiz responses, giving 44 GSM8K-Quiz and 28 Math-Quiz probe responses in total for Figure 3.

Alternating packs. Table B.2 reports descriptive means for the two alternating arrangements. Each arrangement is constructed from the held-out easy and hard pools. The two arrangements need not contain the same questions, so they are complementary probes rather than a paired reversal of identical quizzes. For each dataset, results use one PPO-Quiz checkpoint trained with seed 42 and the shared response budget used in the main evaluation.

Quiz variant	Order	Quizzes	Easy tokens	Hard tokens
GSM8K-Quiz	EHEHE	68	84.0	103.3
	HEHEH	45	83.7	113.2
Math-Quiz	EHEHE	48	157.4	248.9
	HEHEH	32	134.9	247.2

Table B.2: Mean tokens per question in alternating easy (E) and hard (H) quiz variants, including zero tokens for missing headings. Hard questions receive more tokens under both arrangements on both datasets.

Homogeneous packs and shuffled order. PPO-Quiz uses more tokens per question in all-hard quizzes than in all-easy quizzes on both datasets (Table B.3). We also evaluate each quiz in four shuffled question orders, generated using seed 42. This gives $263 \times 4 = 1{,}052$ GSM8K-Quiz responses and $100 \times 4 = 400$ Math-Quiz responses per policy. For each policy, we calculate the mean token count at each position (Q1–Q5) across these responses, then report the difference between the largest and smallest of these five means. If a question heading is missing, we record zero tokens for that position. PPO-Quiz has a smaller difference than the base model on both datasets (Table B.3).

Response length by quiz difficulty. We group the 100 Math-Quiz test examples by the base model's average accuracy on their five questions, measured using eight attempts per question. The groups are lower difficulty (above 80%; 35 quizzes), middle (above 65% and up to 80%; 35 quizzes), and higher (65% or less; 30 quizzes). We choose boundaries to keep group sizes similar without separating quizzes with the same score. The groups are fixed before comparing policies, and the models are not shown these difficulty labels.

Table B.4 compares four policies at shared response budgets of 512 and 1,024 tokens. The trained policies use MATH12K as their training question source, with seed 42 and 500 training steps. We evaluate the same checkpoints at both shared response budgets using greedy decoding. Accuracy is

	PPO-Quiz: tokens/question		Difference across positions (tokens)	
Quiz variant	All-easy	All-hard	Base model	PPO-Quiz
GSM8K-Quiz	87.3	103.4	110.5	11.7
Math-Quiz	173.3	197.5	173.2	56.0

Table B.3: Two complementary controls on quiz variants. Left: mean tokens per question in quizzes containing only easy or only hard questions. Right: We average token counts separately for Q1–Q5 across all shuffled quizzes and report the largest of these five averages minus the smallest. Smaller values indicate more similar average answer lengths across positions.

per question and lengths count the whole quiz response. At 1,024 tokens, PPO-Quiz uses more tokens as quiz difficulty increases; at 512 tokens, nearly all its responses exhaust the shared response budget.

	Tokens/quiz				Accuracy (%)		
Policy	Lower	Middle	Higher	H/L	Lower	Middle	Higher
Shared response budget: 512 tokens							
Base model	508.9	505.8	503.3	0.99	26.9	16.6	10.0
PPO-Single	512.0	505.0	501.9	0.98	29.1	20.0	9.3
PPO-Single + LP-0.3	505.4	489.5	484.1	0.96	28.6	14.9	9.3
PPO-Quiz	510.7	512.0	512.0	1.00	37.7	34.9	22.0
Shared response budget: 1,024 tokens							
Base model	985.9	938.3	910.4	0.92	48.0	35.4	19.3
PPO-Single	998.8	914.2	929.5	0.93	57.7	36.6	18.7
PPO-Single + LP-0.3	975.7	852.3	838.9	0.86	53.1	37.1	16.7
PPO-Quiz	820.4	859.1	861.7	1.05	70.3	58.3	36.0

Table B.4: Math-Quiz results by quiz difficulty. H/L is the ratio of mean tokens on higher-difficulty quizzes to those on lower-difficulty quizzes.

C TRANSFER TO SINGLE-QUESTION INFERENCE

Evaluation setup. We evaluate the base model, PPO-Single, PPO-Quiz, GRPO-Single, and GRPO-Quiz on all 1,319 GSM8K test questions presented individually. Single policies train on individual GSM8K questions; Quiz policies train on GSM8K-Quiz. Trained policies use seed 42 and are evaluated after 500 training steps. All policies use greedy decoding and a 1,024-token response cap. The question IDs, prompt text, and reference answers are identical across policies. Single-question prompts retain the common instructions and the "Question 1" heading, but contain no other questions competing for the response budget. Length counts all generated response tokens.

Paired bootstrap confidence intervals. For each question, we subtract the comparison policy's response length from the evaluated policy's length. The comparison policy is the base model for Table C.1 and the corresponding Single policy for direct Quiz–Single comparisons. We then create 20,000 samples by randomly selecting 1,319 questions each time, allowing a question to be selected more than once. We keep the results from both policies together for each selected question. For each sample, we calculate the average length difference. The middle 95% of these averages gives the confidence interval. We use the same samples to compare accuracy.

We include all questions, even when answers are incorrect or reach the token limit. These intervals show how the results vary across test questions for the same trained models. Table 1 separately reports variation across training runs.

Direct comparisons with single-question training. Compared with single-question training, quiz training reduces mean response length by 27.9 tokens for PPO and 29.7 tokens for GRPO. Observed accuracy remains similar, with differences of -0.45 and $+1.21$ percentage points, respectively (Table C.1).

Policy	Accuracy (%)	Mean Token	Δ tokens	95% CI
Base model	84.61	306.29	—	—
PPO-Single	85.52	296.85	-9.44	$[-12.78, -6.30]$
PPO-Quiz	85.06	268.92	-37.38	$[-41.63, -33.22]$
GRPO-Single	85.44	296.86	-9.43	$[-13.26, -5.64]$
GRPO-Quiz	86.66	267.12	-39.17	$[-43.25, -35.31]$

Table C.1: Single-question evaluation on GSM8K for training seed 42. Tokens are the mean per response. Differences are policy minus the base model; negative values indicate shorter responses. Confidence intervals are paired bootstrap intervals over test questions, not across training seeds.

Cross-dataset single-question transfer. Table 2 extends this comparison to four training–test pairs: policies with GSM8K as their training question source on MATH-500 and SVAMP, and those with MATH12K as their source on GSM8K and SVAMP. Quiz policies train on GSM8K-Quiz or Math-Quiz; Single policies train on the corresponding individual source questions. Quiz-trained policies reduce mean response length by approximately 5–9% relative to their single-trained counterparts, with accuracy differences ranging from -1.7 to $+2.0$ percentage points. These results use one training seed.

D Cross-dataset quiz evaluation

We test how well models perform on quiz datasets with question sources different from those used for training. We use the same base model and compare PPO and GRPO trained on individual source questions (Single) or the corresponding quiz datasets (Quiz). For the GSM8K training question source, we evaluate on Math-Quiz and SVAMP-Quiz; for MATH12K, we evaluate on GSM8K-Quiz and SVAMP-Quiz. Math-Quiz's training and test splits use MATH12K and MATH-500, respectively, so its in-distribution results are already reported in Table 1.

All models answer five-question quizzes using greedy decoding. We allow up to 2,048 response tokens per quiz on Math-Quiz and 1,024 on GSM8K-Quiz and SVAMP-Quiz. These limits are the same for all models. Math-Quiz has 100 test quizzes (500 questions), GSM8K-Quiz has 263 (1,315 questions), and SVAMP-Quiz has 60 (300 questions). We report accuracy per question and count all generated tokens per quiz. For PPO and GRPO, we report the mean and sample standard deviation across training seeds 42, 123, and 456. We evaluate the base model once on each test dataset.

Quiz-trained models use fewer tokens on average than single-trained models in every comparison (Table D.1), while matching or exceeding their mean accuracy when transferring between GSM8K-Quiz and Math-Quiz. On SVAMP-Quiz, PPO with GSM8K as its training question source uses about 55% fewer tokens with similar mean accuracy (1.0 percentage point lower). GRPO uses 54% fewer tokens but has 4.5 points lower mean accuracy. With MATH12K as the source, PPO uses 35% fewer tokens with similar mean accuracy (1.67 points lower). GRPO uses 29% fewer tokens with 2.50 points higher mean accuracy. All changes compare the reported means.

(a) Training question source: GSM8K

	Math-Quiz		SVAMP-Quiz	
Policy	Accuracy (%)	Tokens/quiz	Accuracy (%)	Tokens/quiz
Base model	53.60	1,403.6	87.00	678.3
PPO-Single	52.53 ± 3.26	$1{,}309.6 \pm 41.4$	86.00 ± 2.85	689.4 ± 21.8
PPO-Quiz	58.60 ± 0.60	886.7 ± 82.6	85.00 ± 1.00	311.2 ± 14.6
GRPO-Single	54.90 ± 0.14	$1{,}333.9 \pm 4.4$	89.50 ± 1.18	696.3 ± 20.0
GRPO-Quiz	59.70 ± 1.56	$1{,}012.9 \pm 31.0$	85.00 ± 0.94	317.4 ± 8.8

(b) Training question source: MATH12K

	GSM8K-Quiz		SVAMP-Quiz	
Policy	Accuracy (%)	Tokens/quiz	Accuracy (%)	Tokens/quiz
Base model	68.06	924.1	87.00	678.3
PPO-Single	69.51 ± 1.96	904.3 ± 13.4	87.33 ± 1.53	682.7 ± 5.5
PPO-Quiz	76.17 ± 0.56	628.3 ± 18.5	85.67 ± 1.20	442.3 ± 15.9
GRPO-Single	69.24 ± 0.91	915.6 ± 3.1	86.67 ± 0.47	704.1 ± 11.3
GRPO-Quiz	76.16 ± 0.91	759.2 ± 2.1	89.17 ± 0.24	501.8 ± 3.3

Table D.1: Cross-dataset quiz evaluation. Training question source identifies the source shared by Single and Quiz; the base model has no task-specific training. Results for the base model on GSM8K-Quiz and Math-Quiz are reused from Table 1.

E SHARED RESPONSE BUDGET AT EVALUATION

We evaluate seed-42 PPO-Single and PPO-Single with LP-0.3 trained on individual GSM8K questions, and PPO-Quiz trained on GSM8K-Quiz, after 500 training steps. We use the same policies with shared response budgets of 256, 512, 1,024, and 2,048 tokens. All sixteen evaluations use identical prompt text, question mappings, and reference answers, with greedy decoding on the 263 GSM8K-Quiz test examples. Table E.1 reports per-question accuracy and mean recorded response length per quiz.

T	Policy	Accuracy (%)	Tokens/quiz	$\geq T$ (%)
256	Base model	14.98	255.7	95.82
	PPO-Single	15.51	254.8	93.54
	PPO-Single + LP-0.3	19.16	255.7	94.30
	PPO-Quiz	39.77	255.1	94.68
512	Base model	34.90	507.9	90.11
	PPO-Single	36.43	498.0	88.21
	PPO-Single + LP-0.3	44.26	510.1	88.59
	PPO-Quiz	69.66	423.3	15.21
1,024	Base model	68.06	923.9	41.06
	PPO-Single	68.21	881.9	35.36
	PPO-Single + LP-0.3	76.43	827.1	14.07
	PPO-Quiz	71.25	437.0	1.14
2,048	Base model	75.21	1037.9	2.28
	PPO-Single	73.84	971.5	1.90
	PPO-Single + LP-0.3	77.72	864.2	1.52
	PPO-Quiz	71.18	445.9	0.76

Table E.1: The base model and seed-42 PPO results on GSM8K-Quiz at different shared response budgets T, measured in tokens per five-question quiz. Each entry comes from one evaluation run. The final column shows the percentage of responses whose recorded length reaches or exceeds T. It does not show why generation stopped.

F EXPLICIT LENGTH CONTROLS AND CONCISE PROMPTING

Length-penalty reward. Let $c(y) \in \{0, 1\}$ be single-question correctness, $L(y)$ the response length, and T the training response cap. The LP reward is

$$R_{\mathrm{LP}}(y) = c(y)\left(1 - \alpha\frac{L(y)}{T}\right). \tag{3}$$

The coefficient α controls the strength of the length penalty. Larger values penalize long correct responses more. Setting $\alpha = 0$ removes the length penalty. In the label LP-a, a is the value of α. The length cost applies only to correct responses; incorrect responses receive zero reward. Length is measured by tokenizing the full response text with the model tokenizer, without adding special tokens. We use $T = 512$ for GSM8K and $T = 1{,}024$ for MATH12K. Below, we compare different values of α while keeping the rest of the reward formula unchanged.

L1-exact and L1-max rewards. We use the L1 rewards from Aggarwal & Welleck (2025), where b is the target response length:

$$R_{\text{L1-exact}}(y; b) = c(y) - \lambda|b - L(y)|, \tag{4}$$

$$R_{\text{L1-max}}(y; b) = c(y)\ \mathrm{clip}_{[0,1]}\big(\lambda(b - L(y)) + \delta\big). \tag{5}$$

L1-exact penalizes responses that are shorter or longer than b, even when the answer is incorrect. L1-max gives incorrect answers zero reward. For correct answers, it favors shorter responses and keeps the reward between 0 and 1. Both methods count all tokens in the response, as in LP.

For each training example, we sample an integer b uniformly from $[64, T]$ using seed 0. We add "Think for b tokens." to the prompt and keep the target fixed throughout training. We set $\lambda = 0.0003 \times 4{,}000/T = 1.2/T$ and $\delta = 0.5$. This gives $\lambda = 0.00234375$ for GSM8K and approximately 0.00117188 for MATH12K.

We train separate PPO-Single policies with L1-exact and L1-max on individual source questions, keeping each reward formula unchanged for all 500 updates. Both start from the base model and use each source's response limit. The original L1 method instead trains with Exact followed by Max. Table 1 uses the same quiz prompts for all models, without target-length instructions. This tests performance under a shared response budget without a target length for each answer.

LP-0.3 across training seeds. Table F.1 reports LP-0.3 results averaged over three training seeds for each source. On GSM8K-Quiz, LP-0.3 has similar mean accuracy to GRPO-Quiz and higher mean accuracy than PPO-Quiz in Table 1, but generates longer responses than both. On Math-Quiz, PPO-Quiz uses fewer tokens with higher observed mean accuracy than LP-0.3.

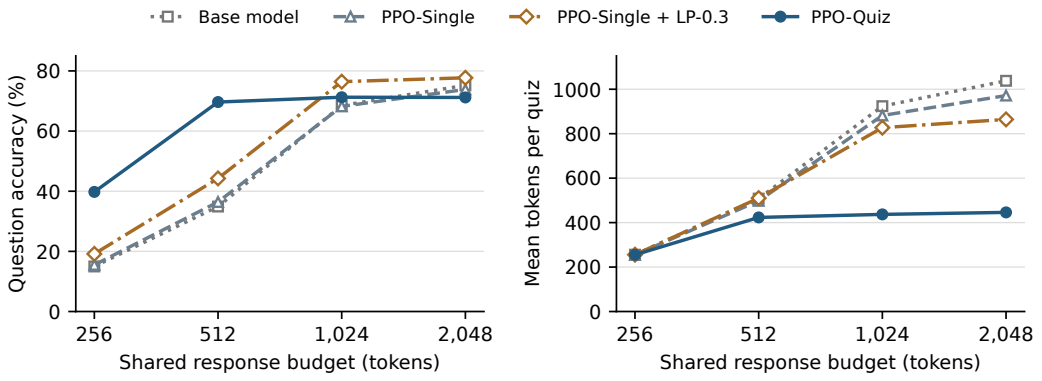

Figure 4: GSM8K-Quiz accuracy and mean response length at different shared response budgets. This separate sweep keeps each PPO policy's seed-42 checkpoint fixed and uses the same 263 five-question quizzes at each budget. Accuracy is per question and length is per quiz. Appendix E gives the evaluation details.

Training source	Quiz test set	Accuracy (%)	Tokens/quiz
GSM8K	GSM8K-Quiz	75.11 ± 1.42	861.8 ± 36.2
MATH12K	Math-Quiz	54.47 ± 1.51	$1{,}255.5 \pm 29.7$

Table F.1: PPO-Single with $\alpha = 0.3$ on five-question quizzes after 500 training steps, using greedy decoding. Results are mean $\pm$ sample standard deviation across seeds 42, 123, and 456. Token limits per quiz are 1,024 for GSM8K-Quiz and 2,048 for Math-Quiz.

LP coefficient ablation. We vary the coefficient of the explicit length penalty in single-question PPO training on GSM8K, using $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Each policy is evaluated on both individual GSM8K questions and GSM8K-Quiz test examples. Table F.2 reports accuracy and mean response length for this coefficient sweep. Length is measured per individual answer in single-question evaluation and per whole quiz in quiz evaluation. This coefficient sweep uses training seed 42; the across-seed LP-0.5 comparison is reported in Table 1. The LP-0.3 across-seed aggregate is reported separately in Table F.1.

Larger values of α produce shorter answers to individual questions, but accuracy falls from 85.1% at $\alpha = 0.1$ to 68.5% at $\alpha = 0.9$. On GSM8K-Quiz, LP models that reach at least 71.3% accuracy use 737–923 tokens per quiz. PPO-Quiz reaches 71.3% accuracy using 437 tokens per quiz, without an explicit length penalty.

	Single-question evaluation		Quiz evaluation	
Policy	Accuracy (%)	Tokens/answer	Accuracy (%)	Tokens/quiz
Base model	84.6	306	68.1	924
PPO-Quiz	85.1	269	71.3	437
PPO-Single + LP-0.1	85.1	272	73.0	923
PPO-Single + LP-0.3	84.8	202	76.4	827
PPO-Single + LP-0.5	83.5	131	73.9	737
PPO-Single + LP-0.7	80.5	111	76.8	756
PPO-Single + LP-0.9	68.5	66	69.1	586

Table F.2: Length-penalty coefficient sweep with single-question evaluation on GSM8K and quiz evaluation on GSM8K-Quiz. Accuracy is per question; tokens count the full generated response in each format.

Training-free concise prompting. With a concise system prompt and a shared response budget of 1,024 tokens, the base model achieves 68.4% accuracy on GSM8K-Quiz with a mean length of 932 tokens. The base model with the standard prompt achieves 68.1% with 924 tokens at the same shared response budget. Asking for shorter answers makes little difference to accuracy or response length here. With the standard prompt and a shared response budget of 512 tokens, the base model achieves 34.9% accuracy and averages 508 tokens per quiz. Appendix E reports results at other shared response budgets.

GRPO with a direct length penalty. We also attempted GRPO training with the same fixed length penalty used for the PPO length-penalty control. Training was unstable: policy KL rose sharply and mean response length collapsed from approximately 300 to 50 tokens. The run did not yield a stable accuracy–length operating point and is therefore excluded from Table 1.

G Training and evaluation quiz sizes

We train PPO-Quiz policies on GSM8K-Quiz variants with $n_{\text{train}} \in \{2, 3, 5\}$ questions per prompt and evaluate on variants with $n_{\text{eval}} \in \{2, 3, 5, 10\}$. These variants regroup questions from the corresponding GSM8K source split. We report two evaluation settings in Table G.1. Panel (a) fixes the shared response budget at 1,024 tokens, so the available tokens per question decrease as quiz size increases. Panel (b) uses a shared response budget of $T \approx 512 n_{\text{eval}}/5$, rounded to whole tokens, keeping the available budget near 102 tokens per question. Both settings apply the shared response budget to the whole response. Panel (b) does not impose separate caps on individual answers.

(a) Fixed shared response budget

Policy	$n_{\text{eval}} = 2$ $T = 1{,}024$	$n_{\text{eval}} = 3$ $T = 1{,}024$	$n_{\text{eval}} = 5$ $T = 1{,}024$	$n_{\text{eval}} = 10$ $T = 1{,}024$
Base model	82.40 / 507	80.18 / 675	68.52 / 924	35.42 / 963
PPO-Single	82.17 / 500	77.07 / 644	68.21 / 882	32.06 / 858
PPO-Quiz ($n_{\text{train}} = 2$)	83.54 / 364	81.09 / 472	75.36 / 696	54.20 / 1,014
PPO-Quiz ($n_{\text{train}} = 3$)	80.65 / 311	80.03 / 367	73.23 / 532	53.39 / 897
PPO-Quiz ($n_{\text{train}} = 5$)	81.11 / 312	79.27 / 349	71.25 / 437	53.66 / 687

(b) Approximately fixed budget per question

Policy	$n_{\text{eval}} = 2$ $T = 205$	$n_{\text{eval}} = 3$ $T = 307$	$n_{\text{eval}} = 5$ $T = 512$	$n_{\text{eval}} = 10$ $T = 1{,}024$
Base model	19.04 / 205	28.40 / 307	34.90 / 508	35.42 / 963
PPO-Single	17.22 / 205	28.93 / 303	36.43 / 498	32.06 / 858
PPO-Quiz ($n_{\text{train}} = 5$)	45.75 / 204	65.53 / 293	69.66 / 423	53.66 / 687

Table G.1: Quiz-size experiments on GSM8K-Quiz variants. Each cell reports per-question accuracy (%) / mean tokens per whole response. The shared response budget T is shown for every evaluation column. The $n_{\text{eval}} = 10$ column uses the same budget in both panels.

At a fixed shared response budget, policies trained with two or three questions also transfer to larger quizzes. The two-question-trained policy has the highest reported accuracy on five-question evaluation, so matching the training size to the evaluation size is not always best. With a fixed shared response budget during training, larger quizzes leave fewer tokens available per question. The results reflect both changes, so they do not show the effect of quiz size alone.

H Additional models

We also test PPO training on GSM8K-Quiz with DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-distill) and Llama-3.2-3B-Instruct. Table H.1 compares each quiz-trained model with its base model. For DeepSeek, we also compare against PPO-Single and PPO-Single with LP-0.3, trained on individual GSM8K questions. All models answer GSM8K-Quiz test examples using greedy decoding, with a shared response budget of 1,024 tokens per quiz. Accuracy is measured per question, and response length counts all tokens generated for the quiz.

Compared with the base model, PPO-Quiz reduces DeepSeek token use by about 57%, with mean accuracy 7.9 percentage points higher. For Llama, token use falls by about 52% with similar observed accuracy (1.7 percentage points higher). On DeepSeek, PPO-Quiz also generates shorter responses than PPO-Single and LP-0.3 without reducing mean quiz accuracy. Appendix J reports results for both models on individual questions.

Model	Policy	Accuracy (%)	Tokens/quiz
DeepSeek-distill	Base model	53.5	992
	PPO-Single	53.0	558
	PPO-Single + LP-0.3	46.9	448
	PPO-Quiz	61.4	428
Llama-3.2-3B	Base model	61.5	871
	PPO-Quiz	63.2	416

Table H.1: GSM8K-Quiz results for DeepSeek-distill and Llama-3.2-3B-Instruct. Trained models use seed 42. DeepSeek PPO models are evaluated after 500 training steps on 263 quizzes.

I Quiz-position and answer diagnostics

Figure 2 shows results for policies with MATH12K as their training question source on 100 Math-Quiz test examples (500 questions), with a shared response budget of 2,048 tokens. Figure 5 shows

results for policies with GSM8K as their training question source on 263 GSM8K-Quiz test examples (1,315 questions), with a shared response budget of 1,024 tokens. All models, including the base model, use greedy decoding and answer the same quizzes within each dataset. The trained models use seed 42. Table 1 separately reports results across multiple training seeds.

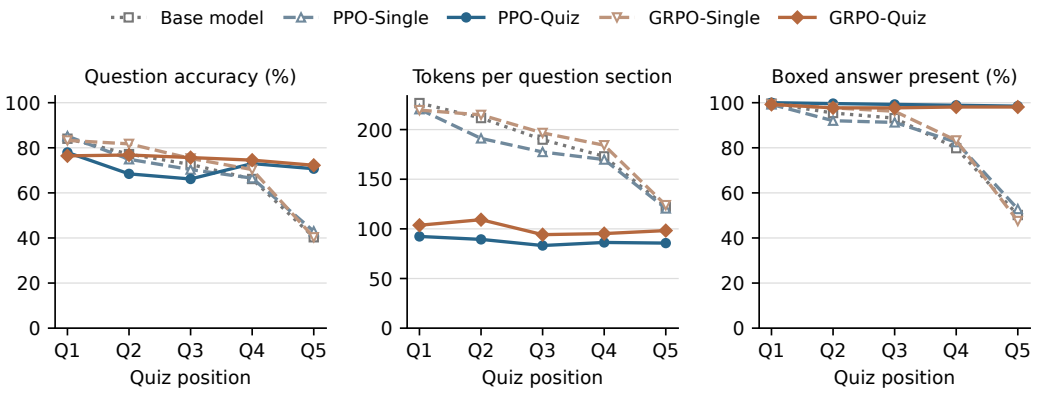

Figure 5: Accuracy, response length, and frequency of boxed answers at each question position on GSM8K-Quiz. All models answer the same 263 five-question quizzes using greedy decoding and a shared response budget of 1,024 tokens. Trained models use seed 42. Single and Quiz refer to training on individual questions and quizzes, respectively.

On GSM8K-Quiz, quiz-trained models give a boxed answer to the fifth question (Q5) in about 98% of quizzes. The base model and models trained on single questions do so in 48–53% of quizzes. Q5 accuracy is 71–72% for quiz-trained models and 40–43% for these baselines. Models trained on single questions are more accurate on some earlier questions, but quiz training raises overall accuracy by 3.3 percentage points for PPO and 4.9 points for GRPO.

Token counts by question. For both datasets, we split each response using the Question-i or Qi headings recognized by the evaluation code. Each section starts at its heading and ends at the next heading or the end of the response. If a question heading is missing, its token count is zero. Text before the first heading is counted as unassigned. If no headings are found, the entire response is unassigned.

We tokenize the full response once and assign each token to the section where it begins. The counts include headings, answers, and formatting. Adding the section counts and unassigned tokens gives the full response length before rounding. The tables and analyses below focus on GSM8K-Quiz.

Policy	Q1	Q2	Q3	Q4	Q5	Unassigned	Total
Base model	226.7	211.6	189.9	173.0	122.3	0.5	924.1
PPO-Single	220.5	191.1	177.4	169.8	120.6	2.8	882.2
PPO-Quiz	92.4	89.4	83.2	86.3	85.7	0.0	437.0
GRPO-Single	219.6	214.8	196.6	184.2	123.9	9.2	948.4
GRPO-Quiz	103.6	109.2	94.2	95.3	98.3	0.0	500.6

Table I.1: Mean response tokens at each position on GSM8K-Quiz, with a shared response budget of 1,024 tokens. Averages include all 263 quizzes, with zero tokens for missing question sections. Totals include unassigned tokens and may differ slightly from the sum of the displayed values because of rounding.

With this shared response budget, the base model and the single-trained policies spend approximately 220–227 tokens on Q1, with progressively shorter later sections. The quiz-trained policies use approximately 83–109 tokens per section. These position averages include incomplete responses; variation across positions is distinct from the question-difficulty effects studied in Section 4.4.

Question headings and boxed answers. Table I.2 reports whether each answer is correct, whether it is boxed, and whether its question heading is present. We count a boxed answer when the question section contains a nonempty `\boxed{...}` expression with matching braces. If a response has exactly five boxes but no question headings, we assign the boxes to Q1–Q5 in order. This is needed for one Math-Quiz PPO-Single response and none of the GSM8K-Quiz responses. The evaluation scorer can mark an answer as correct even if it is not boxed.

Policy	Measure (%)	Q1	Q2	Q3	Q4	Q5
Base model	Accuracy	84.0	77.2	72.6	66.2	40.3
	Boxed answer	99.6	95.4	93.2	79.8	50.2
	Heading detected	100.0	97.7	95.1	93.2	79.5
PPO-Single	Accuracy	85.2	74.9	70.3	66.5	43.0
	Boxed answer	99.2	92.0	91.3	82.5	52.9
	Heading detected	100.0	93.5	92.4	91.6	82.5
PPO-Quiz	Accuracy	77.9	68.4	66.2	73.0	70.7
	Boxed answer	100.0	99.6	99.2	98.9	98.5
	Heading detected	100.0	100.0	99.6	99.2	98.9
GRPO-Single	Accuracy	83.3	81.7	75.3	70.3	40.3
	Boxed answer	99.2	97.7	96.2	83.3	47.5
	Heading detected	100.0	99.6	98.5	97.0	82.5
GRPO-Quiz	Accuracy	76.4	76.8	75.7	74.5	72.2
	Boxed answer	99.2	97.7	97.7	98.1	98.1
	Heading detected	100.0	99.2	98.1	98.1	98.1

Table I.2: Accuracy and the percentages of responses with boxed answers or question headings at each position on GSM8K-Quiz, with a shared response budget of 1,024 tokens. All percentages use the same 263 quizzes for every policy, including responses with missing headings or boxes.

Quiz scores and answer outcomes. Figure 6 reports the number of correct answers per quiz and a question-level decomposition. Each question is first assigned its saved correctness label; incorrect questions are then separated by whether a boxed answer was detected. These categories sum to all 1,315 questions. The number of incorrect questions without a detected box falls from 200 for the base model to 10 for PPO-Quiz and 23 for GRPO-Quiz. The corresponding single-trained policies have 204 and 183 such cases. An answer may lack a box because of its formatting or because it is unfinished. A missing box alone does not show that the response was cut off by the token limit.

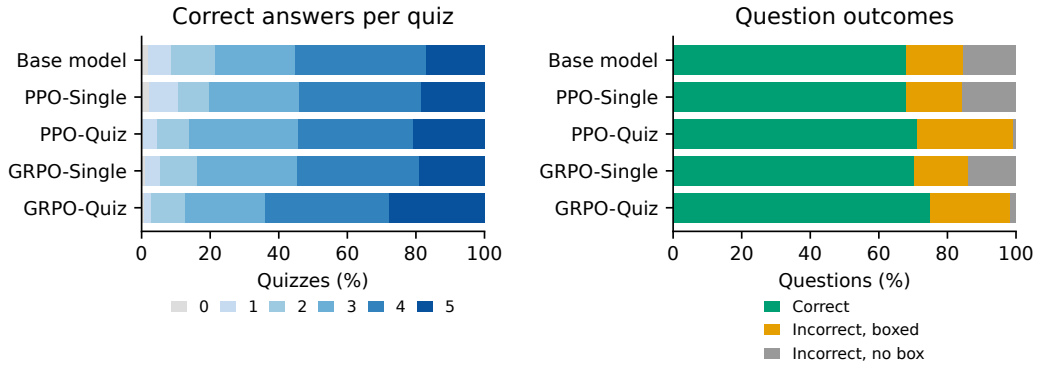

Figure 6: Number of correct answers per quiz (left) and counts of correct and incorrect answers (right) on GSM8K-Quiz, with a shared response budget of 1,024 tokens. Correct answers are counted whether or not they are boxed. Incorrect answers are split into those with and without a box. Results use the same responses as Figure 5.

Which failures of the base model are recovered? We compare the base model's answers to the same 1,315 GSM8K questions when asked individually and as GSM8K-Quiz examples. We match

the questions by checking their text and reference answers. Both evaluations allow up to 1,024 response tokens, but this limit covers one question in the individual setting and all five questions in the quiz setting. The base model answers 847 questions correctly in both settings, 265 only when asked individually, and 48 only in quizzes. It answers the remaining 155 incorrectly in both.

We then look at the questions the base model gets wrong in quizzes. Table I.3 shows how often each trained policy answers these correctly. We separate the 265 questions that the base model can answer individually from the 155 that it gets wrong in both settings.

	Base model failure subset		
Policy	Correct alone, incorrect in quiz	Incorrect in both formats	Ratio
PPO-Single	90/265 (34.0%)	16/155 (10.3%)	3.29×
PPO-Quiz	165/265 (62.3%)	24/155 (15.5%)	4.02×
GRPO-Single	101/265 (38.1%)	26/155 (16.8%)	2.27×
GRPO-Quiz	179/265 (67.5%)	30/155 (19.4%)	3.49×

Table I.3: Performance on questions the base model answers incorrectly in GSM8K-Quiz. The response limit is 1,024 tokens per five-question quiz and per question when the base model answers individually. Each entry shows correct answers out of the group total, followed by the percentage. The ratio divides the first group's accuracy by the second's, using values before rounding. Every policy is evaluated on the same two groups of questions.

PPO-Quiz and GRPO-Quiz correctly answer 62.3% and 67.5% of the 265 questions that the base model answers correctly only when asked individually. For the 155 questions the base model gets wrong in both settings, the rates are 15.5% and 19.4%. Both quiz-trained policies therefore do better on questions that the base model can already answer individually. This comparison alone cannot tell us whether the gains come from better math skills or better use of the available tokens.

Illustrative quiz. Figure 7 shows the question-section lengths for GSM8K-Quiz test example 97 (original question indices 485–489), the index selected for the earlier illustration. With a shared response budget of 1,024 tokens, all five policies answer all five questions correctly. PPO-Quiz uses 303 tokens and GRPO-Quiz 501, compared with 833–861 for the other policies. This example illustrates differences in response length on a fixed quiz; the aggregate evidence is reported above.

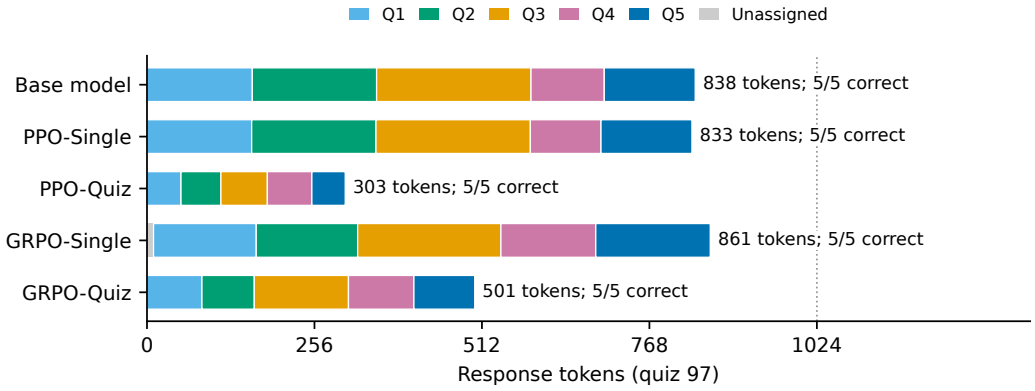

Figure 7: Tokens used for each question in one GSM8K-Quiz test example. Each bar segment shows the tokens assigned to a question, counted as in Table I.1. The dotted line marks the shared response budget of 1,024 tokens. All policies answer all five questions correctly.

J SINGLE-QUESTION EVALUATION ON ADDITIONAL MODELS

DeepSeek at different response limits. We compare the base model, PPO-Single, PPO-Single with length penalty 0.3, and PPO-Quiz for DeepSeek-R1-Distill-Qwen-1.5B. All four policies answer

the same 1,319 GSM8K test questions individually, using greedy decoding at response limits of 256, 512, 1,024, and 2,048 tokens. Each trained policy uses one seed-42 checkpoint across all limits.

After answering the supplied question, some responses continue with further model-generated questions and answers, increasing full-generation length. We measure accuracy and first-answer length using only the text before a "Question 2", "Q2", or "Q.2" heading, if one appears. Table J.1 also reports the full response length, including any additional questions and answers.

PPO-Quiz's first answers average 169–176 tokens across the four limits. Accuracy is 71.1% at 256 tokens and 74.6–75.0% at 512–2,048 tokens. Full responses become longer as the limit increases because they can include additional questions. Figure 8 shows how first-answer lengths vary at the 1,024-token limit.

			Mean tokens	
Response cap	Policy	Accuracy (%)	First answer	Full generation
256	Base model	12.0	255	255
	PPO-Single	58.8	219	219
	PPO-Single + LP-0.3	73.2	175	175
	PPO-Quiz	71.1	169	253
512	Base model	65.4	396	397
	PPO-Single	81.3	249	249
	PPO-Single + LP-0.3	78.0	185	185
	PPO-Quiz	75.0	173	457
1,024	Base model	75.1	428	429
	PPO-Single	81.6	254	254
	PPO-Single + LP-0.3	78.3	192	192
	PPO-Quiz	75.0	174	757
2,048	Base model	75.5	448	450
	PPO-Single	81.6	260	260
	PPO-Single + LP-0.3	78.6	201	201
	PPO-Quiz	74.6	176	1,240

Table J.1: DeepSeek results on 1,319 individual GSM8K questions at different response limits. Accuracy is based on the answer to the original question. Length is shown for this answer and for the full response, which includes any extra questions and answers. All questions are included in the results. Trained models use seed 42. Accuracy is rounded to one decimal place and average lengths to whole tokens.

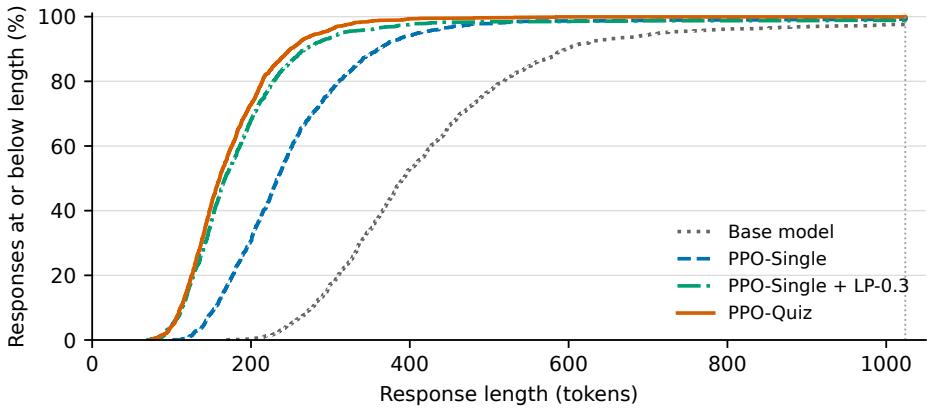

Figure 8: First-answer lengths for DeepSeek on 1,319 individual GSM8K questions with a 1,024-token response limit. Each curve shows the percentage of first answers at or below the length on the horizontal axis. The dotted line marks the response limit.

Questions the base model gets right or wrong. At the 1,024-token response limit, the DeepSeek base model answers 991 test questions correctly and 328 incorrectly. Table J.2 compares mean first-answer lengths for these two groups. The groups are fixed for every policy, regardless of whether it answers correctly. Every policy uses more tokens on questions that the base model gets wrong. This grouping is based on one answer from the base model per question, generated with greedy decoding. The difficulty analysis in Section 4.4 uses eight answers from the base model per question and accounts for where each question appears in the quiz.

Policy	Base model correct	Base model incorrect
Base model	407.3	492.0
PPO-Single	237.2	305.1
PPO-Single + LP-0.3	181.1	224.1
PPO-Quiz	170.2	186.6

Table J.2: Mean first-answer lengths for DeepSeek on individual GSM8K questions with a 1,024-token response limit. Questions are grouped by whether the base model answers correctly. All policies use the same groups.

Llama single-question results. On individual GSM8K questions, Llama-3.2-3B-Instruct PPO-Quiz averages 118 response tokens compared with 250 for the base model, with accuracies of 72.9% and 76.8%, respectively. Both use greedy decoding with a 1,024-token response limit.