Structured Response Diversity with Mutual Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has greatly improved the reasoning abilities of large language models (LLMs) on mathematics and programming tasks, often by maximizing pass@1 correctness. However, optimizing single-attempt accuracy can inadvertently suppress response diversity across repeated attempts, narrowing exploration and overlooking underrepresented strategies. We adapt *Mutual Information Skill Learning* (MISL) to LLMs and develop training-time rewards that induce *structured response diversity*: a discrete latent z selects a reproducible "strategy" that steers the token distribution toward distinct modes. We propose two complementary rewards for Group Relative Policy Optimization (GRPO): a *token-level* mutual information (MI) reward that encourages trajectory specificity to z, and a *semantic* MI reward that encourages separation in an embedding space. Experiments on GSM8K with three open-weight models, Llama 3.1–8B, Qwen 2.5–7B, and R1-Distilled–Qwen2.5–Math–1.5B, with 2,000 training problems show that token-level MISL improves multi-attempt metrics, yielding median gains of $\sim$4% in pass@k and $\sim$12% in consensus@k without degrading pass@1. We further outline a theoretical connection that shows that improvement in pass@k is upper-bounded linearly by the mutual information. We discuss practical considerations, the instability of the semantic MI estimator, and open directions.

Keywords: Language Models, Mutual Information, Reasoning, Response Diversity, RLVR

1 Introduction

LLMs excel at verifiable reasoning tasks such as mathematical problem solving and code generation [15]. However, repeated sampling often yields highly similar outputs [32]. In multi-attempt settings where any one correct completion suffices, such as code generation with tests [5], formal proofs in Lean [39], or objectives evaluated by pass@k, a lack of diversity reduces the effective number of independent attempts and thus the chance that some sampled attempt will succeed. Furthermore, recent work has found that post-training that optimizes single-attempt correctness suppresses response variation across attempts [7, 9], creating a discrepancy between how models are trained and how they are used or evaluated.

The challenge of balancing diversity and accuracy, or exploration and exploitation, has been explored in the form of decoding-time fixes. Methods such as temperature sampling [31], nucleus sampling [18], and prompt perturbations [36] can inject variety, but they require manual tuning, are brittle across domains [35], and may leave the connection between perturbations and LLM response under-specified [29]. We seek a training-time mechanism that (i) increases diversity in a controlled manner, (ii) produces semantically distinct and reproducible modes of reasoning, and (iii) preserves single-attempt verifiable accuracy.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

To do so, we explore a training-time approach that induces *structured response diversity* without prompt engineering. Modeling LLM attempts on verifiable reasoning tasks as a token-level Markov Decision Process, we adapt the unsupervised reinforcement-learning framework of *Mutual Information Skill Learning* (MISL) from [49] to LLMs: the model conditions its response on a discrete latent $z \in \{1, \dots, N\}$, and training encourages behaviors whose distribution depends strongly on z. Intuitively, each z should correspond to a reproducible "skill" or "strategy," and the set of strategies should span a broad range of behaviors.

Concretely, we implement MISL within Group Relative Policy Optimization (GRPO) [33] by adding two new reward terms: a token-level mutual information reward, which encourages diversity in completions, and a semantic mutual information reward, which encourages diversity in embedding space.

Finally, we sketch a theoretical link between $\mathcal{I}(\tau; z \mid x)$ and pass@k: the improvement of pass@k after training is upper-bounded by $C\mathcal{I}(\tau; z \mid x)$ for finite positive C depending on the prompt and strategies, showing that large improvement in pass@k requires large mutual information. In summary, our contributions are as follows:

Contributions:

- We demonstrate median gains of 4% in pass@k and 12% in consensus@k on GSM8K across three open-weight models using LoRA fine-tuning on 2,000 problems, with preserved pass@1 accuracy.
- We prove that pass@k improvement is upper-bounded by $\Delta \leq O(\mathcal{I}(\tau; z|x))$, showing that low mutual information fundamentally limits multi-attempt gains.
- We provide an effective and reproducible method for token-level MI and semantic-MI, optimal hyperparameters, and an implementation focused on practical performance.

2 Background

2.1 Multi-attempt evaluation, redundancy, and why diversity matters

For verifiable tasks, we often consider the probability of success across *multiple* completions rather than a single attempt [7]. Let x denote the input and τ a sampled completion from policy $\pi(\cdot \mid x)$; let $Y(\tau) \in \{0, 1\}$ indicate correctness under a deterministic verifier. For k attempts, the standard metric

$$\text{pass@}k(x) \;=\; 1 - \Pr\left(\bigcap_{i=1}^{k} \{Y(\tau_i) = 0\} \,\middle|\, x \right) \tag{1}$$

is the complement of the joint failure probability across k i.i.d. draws $\tau_{1:k} \sim \pi(\cdot \mid x)$ [5]. Letting $p = \Pr(Y(\tau) = 1 \mid x)$, we therefore have $\Pr(\text{pass@}k(x)) = 1 - (1 - p)^k$.

In practice, identical prompts with fixed decoding hyperparameters can yield strongly correlated trajectories, especially for deterministic or near-deterministic samplers [41]. A useful lens is to consider an "effective number of attempts" k_{eff} that discounts k by a correlation term (analogous to design effects in sampling) [20]. If completions have pairwise correlation ρ in the binary success indicators, a heuristic adjustment gives $k_{\text{eff}} \approx k/(1 + (k-1)\rho)$: as $\rho \to 1$, additional attempts contribute little; as $\rho \to 0$, $k_{\text{eff}} \to k$. Although crude, this highlights the central point: reducing dependence among attempts is as important as raising per-attempt accuracy. Structured diversity aims to decrease redundancy so that the joint failure probability decreases faster in k.

Beyond pass@k, plurality@k and consensus@k measure agreement among completions, examining robustness and internal consistency of the model's reasoning [42]. In many workflows, agreement acts as a proxy for confidence while still benefiting from diversity to escape shared failure modes [17].

2.2 RLVR and GRPO for verifiable reasoning

Reinforcement learning from verifiable rewards (RLVR) uses automatically checked signals (e.g., exact numeric answers, unit tests) to shape policies toward correctness while ensuring the new

policy remains close to a base model [24, 10]. Let π_θ denote the trainable policy and π_{base} a frozen reference; a common form of the per-trajectory reward is

$$r_{\text{RLVR}}(\tau) \;=\; r_{\text{correctness}}(\tau) \;-\; \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot \mid x) \,\big\|\, \pi_{\text{base}}(\cdot \mid x)\big)\,, \tag{2}$$

where $\beta > 0$ controls deviation from the base model [44].

Group Relative Policy Optimization (GRPO) [33] adapts PPO-style updates to reasoning by sampling multiple completions per prompt x as a *group*. Within-group baselines reduce variance and increase the relative difference between completion rewards. Concretely, for each x one draws C trajectories $\{\tau_i\}_{i=1}^{C}$, computes verifiable rewards and a group baseline (e.g., a rank or mean-normalized signal), and updates π_θ with clipped policy ratios as in PPO. GRPO typically improves pass@1 on math/code under RLVR; however, absent any explicit term for diversity, it can *reduce* variation across attempts as the policy sharpens around locally high-reward regions [9]. Empirically, this can shrink the entropy of the completion distribution and heighten redundancy among attempts, limiting pass@k improvements even as pass@1 increases [9].

2.3 Skill discovery and MISL

Unsupervised skill discovery in RL aims to learn diverse, reusable behaviors without external reward by maximizing the mutual information between a latent "skill" variable and observed behaviors [14, 11]. Let $z \in \mathcal{Z}$ index a skill and τ denote an agent trajectory. MISL (Mutual Information Skill Learning) [49] maximizes the conditional mutual information

$$\max_{\pi}\; \mathcal{I}(\tau; z \mid x) \;=\; \mathbb{E}\Big[\log p_\pi(\tau \,|\, x, z) - \log p_\pi(\tau \,|\, x)\Big] \;=\; \mathcal{H}(\tau \mid x) - \mathcal{H}(\tau \mid x, z). \tag{3}$$

This decomposition clarifies the pressure on the policy: (i) increase marginal entropy $\mathcal{H}(\tau \mid x)$ to cover more of trajectory space; (ii) decrease conditional entropy $\mathcal{H}(\tau \mid x, z)$ so that each z induces a reproducible mode. The net effect is a set of distinct, stable behaviors indexed by z that together span diverse solution strategies.

Language models often face verifier-sparse rewards (binary correct/incorrect), where exploration structure matters [1, 24, 10]. MISL offers a training-time mechanism to factor diversity into a discrete latent z that is easy to control at inference. Instead of sampling k times from a single narrow distribution, one samples once from each of k distinct, trained modes.

2.4 LLMs as policies and the role of MI in text

Autoregressive LLMs can be cast as policies over an MDP with state equal to the token prefix and action equal to the next token [3, 27]. Let $\tau = (y_1, \ldots, y_T)$ be a completion. For a discrete latent $z \in \{1, \ldots, N\}$ introduced via a lightweight prefix (e.g., `Strategy {z}`), the conditional likelihood factorizes as

$$p_\pi(\tau \,|\, x, z) \;=\; \prod_{t=1}^{T} p_\pi\big(y_t \,\big|\, x, z, y_{<t}\big). \tag{4}$$

Directly maximizing $\mathcal{I}(\tau; z \mid x)$ is difficult because $p_\pi(\tau \mid x)$ is a mixture over z and high-dimensional [28, 25]. We therefore consider practical surrogates in our approach based on empirical approximations.

3 Related Work

3.1 Response Diversification

One popular approach to increase output diversity is *post-hoc* or *decoding-time* diversification. This involves adjusting parameters at inference time, like increasing temperature, using a different sampling strategy [18] [12], or perturbing the prompt [29]. However, these approaches suffer from some fundamental limitations: (i) they provide no guarantee that samples explore qualitatively different solution paths; (ii) they require per-domain tuning; (iii) diversity often trades off against local coherence or correctness; and (iv) the strategies are not reliably reproducible. Prompt-cycling can inject domain knowledge (e.g., "try algebra" vs. "try geometry"), but it burdens users with prompt engineering and saturates well below human diversity [36].

In contrast, *training-time* diversification shapes the policy so that it supports multiple intentionally distinct modes that can be invoked at inference without manual prompt design. For instance, recent works have explored training LLMs explicitly on objectives based on pass@k evaluation. [38] proposed an unbiased estimator for generic k-attempt objectives with a "leave-one-out" control variate, showing overall improved model efficacy.

Extending this, [7] argue that simply training on pass@1 falls victim to over-exploitation, in which agents fail to explore and converge to a local maximum due to the harsh binary pass@1 rewards. [7] find that pass@k training reduces rewards for high-accuracy responses, naturally focusing optimization efforts on harder problems and mitigating overfitting on easier ones. This produces significant improvements in both pass@k and pass@1.

3.2 Mutual Information

Maximizing mutual information (MI) between latent variables and observed behavior has been a recurring tool for learning *structured*, *controllable* representations. In generative modeling, InfoGAN [6] augments GAN training with a variational lower bound on $\mathcal{I}(c; x)$ to make latent codes c predictably control semantic factors (e.g., stroke thickness for MNIST). In variational autoencoders, InfoVAE [47] adds an explicit MI term to counteract posterior collapse and preserve informative latents even with expressive decoders.

In sequential decision making, MI has been used to discover diverse, reusable behaviors without external rewards. Early work such as VIC [14] and DIAYN [11] maximizes $\mathcal{I}(s; z)$ or $\mathcal{I}(\tau; z)$, encouraging skills z whose rollouts visit different parts of state or trajectory space and remain identifiable from observations. InfoGAIL [23] brings these ideas to imitation learning by maximizing MI between a latent intention and trajectories to capture multi-modal expert behavior. Subsequent methods vary the conditioning and the support of the MI objective: conditioning on context or goals (e.g., $\mathcal{I}(\tau; z \mid \text{context})$) to promote contextual diversity, or measuring MI over future states to bias toward long-horizon distinctiveness [34, 16].

Our setting is closest in spirit to unsupervised skill discovery (e.g., DIAYN, VIC) and to MISL [49], but differs importantly in leveraging the MISL approach for language models with RLVR training, connecting pass@k performance with the MISL objective, and developing proxies unique to the LLM reasoning setting. Additional related work descriptions are available in Appendix A.

3.3 Determinant Diversity

An alternative family of objectives for encouraging diversity comes from determinantal point processes (DPPs). DPP-based methods promote sets of outputs whose embeddings have high determinant volume, effectively rewarding diversity at the set level rather than through latent conditioning. For text generation, DPP sampling has been used to penalize near-duplicate candidates and promote coverage in decoding [22, 41]. In reinforcement learning, determinant-based rewards can encourage agents to explore trajectories that span complementary regions of state space [2, 46].

Compared to mutual information, which directly couples a latent z with trajectories to ensure reproducible modes, determinant diversity is distribution-free: it treats a set of samples as diverse if they occupy a high-volume region in representation space, regardless of whether the same diversity is reproducible under repeated sampling. This makes DPP-style objectives well-suited for one-shot reranking or decoding, but less natural for training-time conditioning where we want interpretability and strategy reproducibility. In this work, we focus on mutual-information based rewards, although other response diversification methods and determinant-based regularizers could provide complementary benefits.

4 Methods

Given an input x and an autoregressive policy $\pi(\cdot \mid x)$ that produces a completion (trajectory) $\tau = (y_1, \ldots, y_T)$, we introduce a *discrete* latent $z \in \{1, \ldots, N\}$ via a lightweight prompt prefix (e.g., `Strategy {z} |`), yielding conditional policies $\pi(\cdot \mid x, z)$. During training, z is drawn uniformly. At inference, one selects $k \leq N$ distinct values of z and generates one completion per value, producing k structured and reproducible attempts.

4.1 Objective

We encourage *structured response diversity* by maximizing the conditional mutual information

$$\max_{\pi} \mathcal{I}(\tau; z \mid x) \;=\; \mathbb{E}\Big[\log p_\pi(\tau \mid x, z) - \log p_\pi(\tau \mid x)\Big], \tag{5}$$

which increases marginal trajectory entropy while reducing conditional entropy within each z-mode. The term $p_\pi(\tau \mid x) = \frac{1}{N}\sum_{z'=1}^{N} p_\pi(\tau \mid x, z')$ is a uniform mixture over skills. Maximizing mutual information encourages (i) high marginal entropy of trajectories, promoting broad coverage, and (ii) low conditional entropy given z, so that each response is distinct and determined by z.

4.2 Token-level mutual information reward

For each pair (x, z), let $\{\tau_i\}_{i=1}^{C}$ be C completions sampled from $\pi(\cdot \mid x, z)$. We define a *per-sample* token-level score

$$r_{\mathrm{TMI}}(\tau_i; x, z) \;=\; \sum_{t=1}^{|\tau_i|} \Big[\log p_\pi(y_t \mid x, z, y_{<t}) - \log p_\pi(y_t \mid x, y_{<t})\Big], \tag{6}$$

where the second term is the uniform mixture

$$p_\pi(y_t \mid x, y_{<t}) \;=\; \frac{1}{N}\sum_{z'=1}^{N} p_\pi(y_t \mid x, z', y_{<t}).$$

Log-probabilities are computed by π on the realized τ_i. In our implementation the mixture is computed *exactly* across all N skills; this is feasible for the N used in our experiments (Section 6). For large N, the mixture can be the unconditioned probabilities of the actor, although this differs from the MISL theory.

4.3 Semantic mutual information reward

Token-level differences can reflect formatting or paraphrase rather than distinct strategies. To bias toward semantic differences, we embed completions with a fixed encoder $\psi(\tau) \in \mathbb{R}^d$ and estimate the mutual information between embeddings and skills for a *single prompt* x:

$$\widehat{\mathcal{I}}\big(\psi(\tau); z \,\big|\, x\big) \tag{7}$$

using the KSG k-nearest-neighbor estimator [21], implemented with the library NPEET [37]. Concretely, for each x we collect the set of embeddings across strategies and samples, $\mathcal{B}(x) = \{(\psi(\tau_i^{(z)}),\, z) \,:\, z \in \{1, \ldots, N\},\, i = 1, \ldots, C\}$, and apply KSG to $\mathcal{B}(x)$ to obtain a single scalar $r_{\mathrm{SMI}}(x)$.

KSG requires multiple points per class to be well-behaved. In our experiments (Section 6), we use $N = 5$ strategies and $C = 6$ completions per strategy (30 points per x) when the semantic term is enabled; for larger N we disable the semantic term to avoid excessive compute and estimator variance.

4.4 Combined GRPO objective

Let $r_{\mathrm{corr}}(\tau_i) \in \mathbb{R}$ denote the verifiable correctness reward (often binary) and

$$\Delta_{\mathrm{KL}}(\tau_i) \;=\; \sum_{t=1}^{|\tau_i|} \log \frac{\pi(y_t \mid x, z, y_{<t})}{\pi_{\mathrm{base}}(y_t \mid x, z, y_{<t})}$$

be the per-trajectory log-likelihood penalty toward the base model. The per-sample scalar reward is thus

$$r(\tau_i; x, z) \;=\; r_{\mathrm{corr}}(\tau_i) \;-\; \beta\, \Delta_{\mathrm{KL}}(\tau_i) \;+\; \alpha_1\, r_{\mathrm{TMI}}(\tau_i; x, z) \;+\; \alpha_2\, r_{\mathrm{SMI}}(x), \tag{8}$$

with $\alpha_1, \alpha_2, \beta \geq 0$. GRPO then, after normalization, applies a clipped policy-gradient update.

4.5 Training procedure

We fine-tune a trainable policy π_θ with GRPO while injecting a discrete strategy variable $z \in \{1, \ldots, N\}$. At each step, we draw a minibatch of prompts x and, for each x, sample a strategy z uniformly and generate C completions $\tau_{1:C} \sim \pi_\theta(\cdot \mid x, z)$ under fixed decoding. For every completion τ, we compute: (i) a verifiable correctness reward $r_{\text{corr}}(\tau)$ from the task's deterministic checker; (ii) the token-level MISL term r_{TMI} that measures how specific the trajectory is to the chosen strategy; and (iii) a KL control term toward a frozen base policy.

Optionally, to encourage semantic separation among strategies for a fixed prompt, we compute the semantic-MI reward r_{SMI}. We then apply GRPO's within-group baseline (computed over the C completions that share the same (x, z)) to obtain advantages, and we update π_θ with a clipped PPO objective using stored behavior-policy log-probs. At inference, diversity is exercised by selecting $k \leq N$ distinct strategy indices and drawing one completion per z, optionally followed by plurality or consensus aggregation.

Algorithm 1 MISL-GRPO with exact mixture and prompt-level semantic MI

Inputs: base policy π_{base}, trainable policy π, latent count N, completions per group C, weights $(\alpha_1, \alpha_2, \beta)$
repeat
  Sample a minibatch of prompts $\{x\}$
  for each x in the minibatch **do**
    Sample $z \sim \text{Unif}(\{1, \ldots, N\})$; generate C completions $\{\tau_i\}_{i=1}^{C}$ with $\pi(\cdot \mid x, z)$
    Compute $r_{\text{corr}}(\tau_i)$, $r_{\text{TMI}}(\tau_i; x, z)$, and $\Delta_{\text{KL}}(\tau_i)$ as above
  end for
  if semantic MI is enabled **then**
    For each prompt x, collect $\mathcal{B}(x)$ across all strategies/samples in the batch and compute $r_{\text{SMI}}(x)$ via KSG
  end if
  Form per-sample rewards via (8); compute advantages; update π with GRPO
until convergence

4.6 Inference

Given a budget of k attempts, choose k distinct latents from $\{1, \ldots, N\}$ and generate one completion per latent under fixed decoding hyperparameters. Optional aggregation (e.g., majority vote) can be applied. Because each completion is produced by a trained, distinct mode, conditional success probabilities remain larger than with redundant samplings, improving multi-attempt metrics.

5 Theory

This section derives mathematical connections between the token-level mutual information with discrete strategies and the pass@k. Our seminal lemma is as follows:

Lemma 1. Under the assumptions below, with Δ referring to the pass@k improvement over training, we show that:

$$0 \leq \Delta \leq (1 - \text{pass@}k(x, 0))\; c\, \mathcal{I}(\tau; z \mid x),$$

where pass@$k(x, 0)$ refers to the initial pass@k score on prompt x and c depends on both k and the highest success probability of any strategy on prompt x.

5.1 Problem setup and assumptions

Similar to the setup in Sec. 2.4, suppose that we have k strategies that are being fine-tuned from an initial model, each of which have different policies for generating trajectories $\pi_{z,t}(\cdot \mid x)$ for strategy $z \in [k]$ at training step t. These trajectory distributions are therefore subject to the following conditions:

1. Assume that all k strategies have the same distribution before training begins, i.e. $\pi_{i,0}(\cdot \mid x) = \pi_{j,0}(\cdot \mid x)$ for $1 \le i < j \le k$. We will denote this initial distribution $\pi_0(\cdot \mid x)$ as shorthand.
2. At each time step, assume that the joint distribution of trajectories over picking a strategy uniformly at random is the same as the original distribution, or mathematically $\frac{1}{k}\sum_{z=1}^{k} \pi_{z,t}(\cdot \mid x) = \pi_0(\cdot \mid x)$. In practice, this condition imposes undue constraints on the strategy distribution, and thus for the practical implementation, this is not enforced. [1]

We extend the traditional definition of pass@k to fit the setting of having multiple different strategies to query. For a given prompt x and deterministic verifier $Y(\tau)$, define the pass@k accuracy at training step t to be the probability that querying each of these strategies exactly once, as outlined in Sec. 4.6, results in at least one correct answer. Writing this out mathematically, for $z \in [k]$ we independently sample $\tau_{z,t} \sim \pi_{z,t}(\cdot \mid x)$. Then

$$\text{pass@}k(x,t) \;=\; 1 - \Pr\left(\bigcap_{z=1}^{k} \{Y(\tau_{z,t}) = 0\} \,\middle|\, x\right) = 1 - \prod_{z=1}^{k} \Pr(Y(\tau_{z,t}) = 0 \mid x). \tag{9}$$

We will focus on the potential improvement from the beginning to the end of the training. Suppose that training ends at step T; sample $\tau \sim \pi_0(\cdot \mid x)$ and $\tau_{z,T} \sim \pi_{z,T}(\cdot \mid x)$ for $z \in [k]$. Let $a = \Pr[Y(\tau) = 1]$ and $a_z = \Pr[Y(\tau_{z,T}) = 1]$. From assumption (2) above and the fact that the trajectories $\tau_{z,T}$ are sampled independently, we have that

$$\frac{1}{k}\sum_{z=1}^{k} a_z = a. \tag{10}$$

Then

$$\text{pass@}k(x,0) = 1 - (1-a)^k, \;\; \text{pass@}k(x,T) = 1 - \prod_{z=1}^{k}(1-a_z).$$

5.2 Derivation of lower bound

Since $f(y) = \log(1-y)$ is strictly concave, by Jensen's inequality we have that

$$\sum_{z=1}^{k} \ln(1-a_z) = \sum_{z=1}^{k} f(a_z) \le kf\left(\frac{1}{k}\sum_{z=1}^{k} a_z\right) = kf(a) = k\ln(1-a) \implies \prod_{z=1}^{k}(1-a_z) \le (1-a)^k$$

$$\implies 1 - \text{pass@}k(x,T) \le 1 - \text{pass@}k(x,0) \implies \text{pass@}k(x,T) \ge \text{pass@}k(x,0). \tag{11}$$

Equation (11) implies that the expected pass@k after training the k strategies is always at least as high as the expected pass@k before training. Furthermore, equality is achieved if and only if $a_1 = a_2 = \ldots = a_k = a$, i.e., each strategy's chance of being correct is the same.

5.3 Sketch of upper bound

The full proof of Lemma 1 is in Appendix B; the following section gives highlights of the proof. A Taylor expansion argument gives

$$1 - \text{pass@k}(x,T) \ge (1 - \text{pass@k}(x,0)) \exp\left(-\frac{1}{2(1-u)^2}\sum_{z\in[k]}(a_z - a)^2\right). \tag{12}$$

Using Pinsker's inequality to relate $\sum_z (a_z - a)^2$ to the average KL (hence to conditional mutual information) yields

$$1 - \text{pass@}k(x,T) \;\ge\; \big(1 - \text{pass@}k(x,0)\big) \exp\left(-\frac{k}{4(1-u)^2}\,\mathcal{I}(\tau; z \mid x)\right). \tag{13}$$

[1] We believe that weaker forms of this condition may also be sufficient, and this is an ongoing component of our research.

The left-hand side is the probability that none of the k draws pass. At high level, the Taylor expansion step (12) shows that this product is controlled by the variance of the per-index success probabilities (a_z). If the learning updates do not substantially change the trajectory distributions (small KL / small $\mathcal{I}(\tau; z \mid x)$), then the a_z remain tightly concentrated around their mean a and pass@k cannot increase substantially.

Rearranging and using the inequality $x + e^{-x} \geq 1$ yields Lemma 1. Thus, for small mutual information, x the possible gain in pass@k scales at most linearly with $\mathcal{I}(\tau; z \mid x)$ (up to the multiplicative factor $(1 - \text{pass@}k(x, 0))c$).

6 Experiments

To validate the MISL-GRPO framework described in Section 4, we evaluate our approach on GSM8K across three open-weight models. We test whether the token-level MI reward (equation (6)) and semantic MI reward (equation (7)) improve multi-attempt metrics without degrading pass@1 performance. We also examine the effect of varying the number of strategies N and assess the stability of semantic MI rewards.

Tasks and data. GSM8K (MIT License) [8], with 2,000 training problems and 100 held-out questions for evaluation. Zero-shot prompting; max sequence length 1024.

Models. Llama 3.1–8B (Meta Llama 3 Community License Agreement) [26], Qwen 2.5–7B (Apache License) [30], and R1-Distilled–Qwen2.5–Math–1.5B (MIT License). We train LoRA adapters (≈80M parameters) on top of open-weight backbones [43].

RL training. GRPO [33] with a correctness reward and KL penalty to the base model. MISL adds $\alpha_1 r_{\text{TMI}} + \alpha_2 r_{\text{SMI}}$. We ablate $N \in \{5, 10, 20\}$ and (α_1, α_2). Semantic MI experiments primarily use N=5 due to estimator cost.

Evaluation. Metrics: pass@1, pass@k, plurality@k, consensus@k. For inference, we fix k distinct strategies and sample one completion per strategy with fixed temperature.

Each experiment was run on one H100 GPU on an internal cluster with 80 GB of memory. The training took approximately 12-24 hours per experiment and the evaluation took approximately 3 hours per experiment.

7 Results

Token-level MISL improves multi-attempt success. With N=5 and $(\alpha_1, \alpha_2) = (5, 0)$, token MI consistently improves pass@5 and consensus@5 without hurting pass@1; instead, we also observe gains in pass@1. Figure 1 shows our performance on the withheld evaluation set; asterisks mark $p<.05$ improvements. We hypothesize that token MI reduces redundant failure modes across strategies, aligning with Sec. 5.2.

Scaling the number of strategies. For $N \in \{10, 20\}$, gains are mixed. Many GSM8K problems admit only a few distinct solution approaches; forcing too many strategies may allocate capacity to irrelevant modes.

Semantic MI is promising but unstable. Optimizing against the KSG estimator for $\mathcal{I}(\psi(\tau); z \mid x)$ often destabilized training, likely due to estimator variance and challenges with fitting a distribution on a high-dimensional embedding space with few examples. On R1-Distilled-Qwen, combining semantic MI with token MI occasionally yielded additional gains, suggesting that better semantic signals (e.g., contrastive or classifier-based surrogates) could help.

8 Conclusion

Our experiments show that token-level MISL provides a simple and effective way to induce strategy-level diversity in LLMs, leading to consistent gains on multi-attempt metrics such as pass@5 and consensus@5. By conditioning on discrete latent variables, the model learns reproducible modes of reasoning that reduce redundancy across attempts and increase the likelihood of success. However,

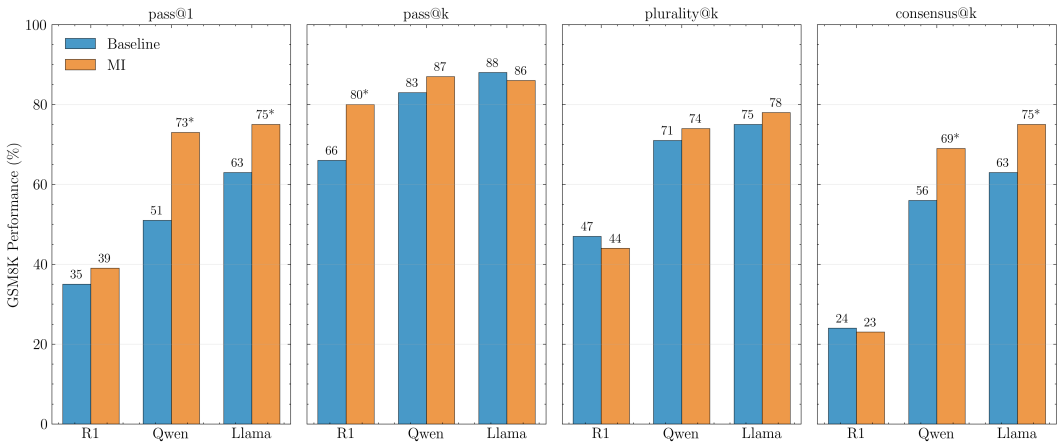

Figure 1: Performance with $N{=}5$ strategies and token MI only. We observe consistent gains on pass@1 and consensus@5. Asterisks denote $p{<}.05$ from a paired Student's t-test on the binary test problems.

despite the success of token-level MI, our current approach for Semantic MISL is promising but unstable.

Beyond these empirical findings, MISL provides a principled training-time approach for improving response diversity. and our analysis links $\mathcal{I}(\tau; z \mid x)$ to upper bounds in pass@k improvement in training. We hope this stimulates research on robust semantic diversity signals and theoretical ties between information-theoretic objectives and multi-attempt success.

8.1 Future work

We hope to better understand improvement guarantees under weaker assumptions. Empirically, assumption (2) does not seem necessary for pass@k improvement, and we hope improve the theory to better explain our experimental findings. We also wish to explore possible connections of pass@k with Tsallis entropy, a generalization of Shannon entropy [13].

For further experimentation, we want to explore the performance of our MISL-based training method in other domains (code, formal proofs) with full fine-tuning rather than using LoRA adapters and larger-scale validation. We additionally look forward to testing MISL on more challenging problems with a wide variety of tentative solutions, in which case ample exploration of the solution space may require $k > 20$. Finally, we hope to investigate encoding strategies as a separate part of the embedding instead of in the prompt.

References

[1] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba. Hindsight Experience Replay, Feb. 2018. URL http://arxiv.org/abs/1707.01495. arXiv:1707.01495 [cs].

[2] J. T. Ash, S. Goel, A. Krishnamurthy, and S. Kakade. Gone Fishing: Neural Active Learning with Fisher Embeddings, Dec. 2021. URL http://arxiv.org/abs/2106.09675. arXiv:2106.09675 [cs].

[3] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio. An Actor-Critic Algorithm for Sequence Prediction, Mar. 2017. URL http://arxiv.org/abs/1607.07086. arXiv:1607.07086 [cs].

[4] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. Mine: Mutual information neural estimation, 2021. URL https://arxiv.org/abs/1801.04062.

[5] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating Large Language Models Trained on Code, July 2021. URL http://arxiv.org/abs/2107.03374. arXiv:2107.03374 [cs].

[6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016. URL https://arxiv.org/abs/1606.03657.

[7] Z. Chen, X. Qin, Y. Wu, Y. Ling, Q. Ye, W. X. Zhao, and G. Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models, 2025. URL https://arxiv.org/abs/2508.10751.

[8] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[9] X. Dang, C. Baek, J. Z. Kolter, and A. Raghunathan. Assessing diversity collapse in reasoning. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL https://openreview.net/forum?id=AMiKsHLjQh.

[10] S. Dou, Y. Liu, H. Jia, L. Xiong, E. Zhou, J. Shan, C. Huang, W. Shen, X. Fan, Z. Xi, Y. Zhou, T. Ji, R. Zheng, Q. Zhang, X. Huang, and T. Gui. StepCoder: Improve Code Generation with Reinforcement Learning from Compiler Feedback, Feb. 2024. URL http://arxiv.org/abs/2402.01391. arXiv:2402.01391 [cs] version: 1.

[11] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is All You Need: Learning Skills without a Reward Function, Oct. 2018. URL http://arxiv.org/abs/1802.06070. arXiv:1802.06070 [cs].

[12] A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation, 2018. URL https://arxiv.org/abs/1805.04833.

[13] S. Furuichi. Information theoretical properties of Tsallis entropies. *Journal of Mathematical Physics*, 47(2):023302, Feb. 2006. ISSN 0022-2488, 1089-7658. doi: 10.1063/1.2165744. URL http://arxiv.org/abs/cond-mat/0405600. arXiv:cond-mat/0405600.

[14] K. Gregor, D. J. Rezende, and D. Wierstra. Variational Intrinsic Control, Nov. 2016. URL http://arxiv.org/abs/1611.07507. arXiv:1611.07507 [cs].

[15] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] S. Hansen, G. Desjardins, K. Baumli, D. Warde-Farley, N. Heess, S. Osindero, and V. Mnih. Entropic Desired Dynamics for Intrinsic Control. In *Advances in Neural Information Processing Systems*, volume 34, pages 11436–11448. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/5f7f02b7e4ade23430f345f954c938c1-Abstract.html.

[17] A. Hochlehnert, H. Bhatnagar, V. Udandarao, S. Albanie, A. Prabhu, and M. Bethge. A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility, Apr. 2025. URL http://arxiv.org/abs/2504.07086. arXiv:2504.07086 [cs].

[18] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

[19] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova. Disentangled representation learning for non-parallel text style transfer, 2018. URL https://arxiv.org/abs/1808.04339.

[20] L. Kish. *Survey Sampling*. Wiley, 1965.

[21] A. Kraskov, H. Stoegbauer, and P. Grassberger. Estimating Mutual Information. *Physical Review E*, 69(6):066138, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.066138. URL http://arxiv.org/abs/cond-mat/0305641. arXiv:cond-mat/0305641.

[22] A. Kulesza. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL http://dx.doi.org/10.1561/2200000044.

[23] Y. Li, J. Song, and S. Ermon. Infogail: Interpretable imitation learning from visual demonstrations, 2017. URL https://arxiv.org/abs/1703.08840.

[24] J. Liu, Y. Zhu, K. Xiao, Q. Fu, X. Han, W. Yang, and D. Ye. RLTF: Reinforcement Learning from Unit Test Feedback, Nov. 2023. URL http://arxiv.org/abs/2307.04349. arXiv:2307.04349 [cs].

[25] D. McAllester and K. Stratos. Formal Limitations on the Measurement of Mutual Information, May 2020. URL http://arxiv.org/abs/1811.04251. arXiv:1811.04251 [cs].

[26] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Meta AI Blog, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Announcement of the Llama 4 multimodal AI model family. Accessed: September 3, 2025.

[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

[28] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker. On Variational Bounds of Mutual Information, May 2019. URL http://arxiv.org/abs/1905.06922. arXiv:1905.06922 [cs].

[29] Y. Qiang, S. Nandi, N. Mehrabi, G. Ver Steeg, A. Kumar, A. Rumshisky, and A. Galstyan. Prompt Perturbation Consistency Learning for Robust Language Models. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1357–1370, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.91/.

[30] Qwen Team. Qwen3: Think deeper, act faster. Qwen Blog, April 2025. URL https://qwenlm.github.io/blog/qwen3/. Announcement of Qwen3 large language model family. Accessed: September 3, 2025.

[31] M. Renze and E. Guven. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, 2024. doi: 10.18653/v1/2024.findings-emnlp.432. URL http://arxiv.org/abs/2402.05201. arXiv:2402.05201 [cs].

[32] S. Shaier, M. Sanz-Guerrero, and K. v. d. Wense. Asking Again and Again: Exploring LLM Robustness to Repeated Questions, Mar. 2025. URL http://arxiv.org/abs/2412.07923. arXiv:2412.07923 [cs].

[33] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[34] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman. Dynamics-Aware Unsupervised Discovery of Skills, Feb. 2020. URL http://arxiv.org/abs/1907.01657. arXiv:1907.01657 [cs].

[35] C. Shi, H. Yang, D. Cai, Z. Zhang, Y. Wang, Y. Yang, and W. Lam. A Thorough Examination of Decoding Methods in the Era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.489. URL https://aclanthology.org/2024.emnlp-main.489.

[36] M. Shur-Ofry, B. Horowitz-Amsalem, A. Rahamim, and Y. Belinkov. Growing a Tail: Increasing Output Diversity in Large Language Models, Nov. 2024. URL http://arxiv.org/abs/2411.02989. arXiv:2411.02989 [cs].

[37] G. V. Steeg. gregversteeg/npeet, May 2025. URL https://github.com/gregversteeg/NPEET. original-date: 2014-10-10T19:57:02Z.

[38] Y. Tang, K. Zheng, G. Synnaeve, and R. Munos. Optimizing language models for inference time objectives using reinforcement learning, 2025. URL https://arxiv.org/abs/2503.19595.

[39] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[40] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[41] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models, Oct. 2018. URL http://arxiv.org/abs/1610.02424. arXiv:1610.02424 [cs].

[42] E. Wallace, O. Watkins, M. Wang, K. Chen, and C. Koch. Estimating Worst-Case Frontier Risks of Open-Weight LLMs, Aug. 2025. URL http://arxiv.org/abs/2508.03153. arXiv:2508.03153 [cs] version: 1.

[43] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

[44] W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint, May 2024. URL http://arxiv.org/abs/2312.11456. arXiv:2312.11456 [cs].

[45] J. Zhao, Y. Kim, K. Zhang, A. M. Rush, and Y. LeCun. Adversarially regularized autoencoders, 2018. URL https://arxiv.org/abs/1706.04223.

[46] K. Zhao, Y. Wang, Y. Chen, X. Niu, Y. Li, and L. H. U. Efficient Diversity-based Experience Replay for Deep Reinforcement Learning, Oct. 2024. URL http://arxiv.org/abs/2410.20487. arXiv:2410.20487 [cs] version: 1.

[47] S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders, 2018. URL https://arxiv.org/abs/1706.02262.

[48] T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders, 2017. URL https://arxiv.org/abs/1703.10960.

[49] C. Zheng, J. Tuyls, J. Peng, and B. Eysenbach. Can a misl fly? analysis and ingredients for mutual information skill learning. *arXiv preprint arXiv:2412.08021*, 2024.

A Extended Mutual Information Related Work

Estimating MI reliably is challenging in high dimensions. Variational bounds (Barber–Agakov) optimize a classifier or regressor $q_\phi(z \mid \cdot)$ as a proxy for the intractable posterior [40]. Contrastive bounds such as InfoNCE [40] reduce MI estimation to noise-contrastive classification and have become standard due to their stability. Neural MI estimators like MINE [4] directly optimize a Donsker–Varadhan bound but can suffer from bias/variance trade-offs and training instability. Nonparametric kNN estimators (KSG) [21] avoid parametric critics but require many samples and are sensitive to dimension, motivating careful batching and normalization when used inside policy gradients. In text generation, MI-style objectives have been used to prevent latent collapse and enable controllable generation, e.g., by encouraging informative latents in variational text models [48, 45] or aligning codes with style attributes [19]. These approaches typically maximize MI between prompts or attributes and latent variables, rather than between a discrete strategy and the full trajectory distribution, and are optimized with supervised losses rather than RL.

Conceptually, our objective reconciles two desiderata emphasized in prior MI work: *coverage* (high marginal entropy over trajectories) and *control* (low conditional entropy given z). Whereas decoding-time diversity manipulates token entropy without guarantees about identifiable modes, MI-based diversification learns *reusable, reproducible* modes indexed by a small discrete latent. This makes diversity a first-class, training-time property that can be cleanly exercised at inference by selecting distinct z values.

B Derivation of mutual information upper bound

This appendix derives Lemma 1.

Let $u = \max_z a_z < 1$. Using Taylor's Theorem on $f(y) = \log(1-y)$ gives the equations $f(a_z) = f(a) + (a_z - a)f'(a) + \frac{1}{2}(a_z - a)^2 f''(\xi_z)$ where ξ_z lies in between a_z and a, for $z \in [k]$. Since $f''(y) = -\frac{1}{(1-y)^2}$ is a decreasing function, in $[\min(a, a_1, a_2, \ldots, a_k), \max(a, a_1, a_2, \ldots, a_k)]$ it achieves its minimum at $f''(u) = -\frac{1}{(1-u)^2}$, where $u = \max(a, a_1, a_2, \ldots, a_k) = \max_{z\in[k]} a_z$. Summing all of these equations, the linear terms cancel due to (10). Then

$$\sum_{z\in[k]} \log(1-a_z) = k\log(1-a) + \sum_{z\in[k]} \frac{1}{2}(a_z-a)^2 f''(\xi_z) \geq k\log(1-a) - \frac{1}{2(1-u)^2}\sum_{z\in[k]}(a_z-a)^2$$

$$\implies \prod_{z=1}^{k}(1-a_z) \geq (1-a)^k \exp\left(-\frac{1}{2(1-u)^2}\sum_{z\in[k]}(a_z-a)^2\right)$$

$$\implies 1 - \text{pass@k}(x,T) \geq (1-\text{pass@k}(x,0))\exp\left(-\frac{1}{2(1-u)^2}\sum_{z\in[k]}(a_z-a)^2\right). \tag{14}$$

This places an upper bound on how much we can possibly improve pass@k compared to our original trajectory distributions. In particular, letting δ be the total variation distance and using Pinsker's Inequality, we find that for all $i \in [k]$,

$$|a_i - a| = |\Pr[Y(\tau) = 1] - \Pr[Y(\tau_{i,T}) = 1]| = \left|\sum_{Y(\tau')=1} \pi_0(\tau') - \sum_{Y(\tau')=1} \pi_{i,T}(\tau')\right|$$

$$\leq \sum_{Y(\tau')=1} |\pi_0(\tau') - \pi_{i,T}(\tau')| \leq \sum_{\tau'} |\pi_0(\tau') - \pi_{i,T}(\tau')| \leq \delta(\pi_0, \pi_{i,T}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\pi_0 \,\|\, \pi_{i,T})}$$

$$\implies (a_i - a)^2 \leq \frac{1}{2} D_{\mathrm{KL}}(\pi_0 \,\|\, \pi_{i,T}) \implies \sum_{z \in [k]} (a_z - a)^2 \leq \frac{k}{2} \cdot \frac{1}{k} \sum_{z \in [k]} D_{\mathrm{KL}}(\pi_0 \,\|\, \pi_{z,T})$$

$$= \frac{k}{2} \mathbb{E}_{z \sim \mathrm{Unif}\{1,2,\ldots,k\}} [D_{\mathrm{KL}}(\pi_0 \,\|\, \pi_{z,T})] = \frac{k}{2} \mathcal{I}(\tau; z \mid x).$$

Combining this with (14) yields

$$1 - \text{pass@}k(x,T) \geq (1 - \text{pass@}k(x,0)) \exp\left(-\frac{k}{4(1-u)^2} \mathcal{I}(\tau; z \mid x) \right).$$

As a result, if $\mathcal{I}(\tau; z \mid x)$ is too small, then our theoretical upper bound on improvement in pass@k between steps 0 and T will also be very small.

Rearranging (13) to bound the improvement $\Delta := \text{pass@}k(x,T) - \text{pass@}k(x,0)$, we obtain

$$\Delta = (1 - \text{pass@}k(x,0)) - (1 - \text{pass@}k(x,T))$$

$$\leq (1 - \text{pass@}k(x,0)) \left(1 - \exp\left(-\frac{k}{4(1-u)^2} \mathcal{I}(\tau; z \mid x) \right) \right)$$

$$\leq (1 - \text{pass@}k(x,0)) \cdot \frac{k}{4(1-u)^2} \cdot \mathcal{I}(\tau; z \mid x),$$

where the final inequality uses $1 - e^{-x} \leq x$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim that we have identified a method to improve multi-attempt performance with a mutual information reward term and demonstrate in the experiments section that our method successfully works.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Throughout the paper, we include comments and mentions in directions we believe can be improved. Additionally, in our conclusion and future work sections, we highlight the current limitations of our theory and experimentation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide a complete and correct proof to justify our theoretical results in section 5. The proofs are included in section 5 and additional details are completed in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our algorithm completely in section 4 and provide additional practical implementation details in sections 4 and 6. Our code repository will also be open sourced.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We leverage open-source models and datasets, and we aim to open-source our code repository, along with the scripts required to rerun our experiments and load both the data and models.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Throughout section 4, we explain all the necessary details to understand the results. Although there are minor choices that are not included in the main paper, these are not necessary to understand the results and will be included in the forthcoming code release.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: For our main result in Figure 1 of Section 7, we analyze the statistical significance of our result, showing which improvements satisfy $p < 0.05$ and providing details on the statistical test employed. However, unfortunately, due to the computational cost of model training, we were unable to run multiple training runs per model, so the error within training runs remains not properly understood.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify the computer resources required in section 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper does not involve any human subjects or sensitive information in datasets. As explained below, it poses minimal societal impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper only discusses problems with verifiable rewards, which means that there can only be a right or wrong answer, and furthermore there are straightforward methods to check whether the resulting output is correct or not.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In section 6, we detail the licenses for each model and resources, and we properly credit the model and dataset providers. Moreover, we choose models and datasets with well-understood licenses and that are commonly-used in RLVR literature with an aim to ensure further research is accessible. All licenses are properly credited and respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.