

UPSKILL: MUTUAL INFORMATION SKILL LEARNING FOR STRUCTURED RESPONSE DIVERSITY IN LLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has 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 introduce UpSkill, a training time method that adapts *Mutual Information Skill Learning* (MISL) to LLMs to induce *structured response diversity*: a discrete latent z selects a reproducible “strategy” that steers the token distribution toward distinct modes. We propose a novel reward that we implement within Group Relative Policy Optimization (GRPO): a *token-level* mutual information (MI) reward that encourages trajectory specificity to z . Experiments on GSM8K with three open-weight models, Llama 3.1–8B, Qwen 2.5–7B, and R1-Distilled–Qwen2.5–Math–1.5B show that UpSkill improves multi-attempt metrics for Qwen 2.5–7B, yielding gains of $\sim 4\%$ in `pass@k` and $\sim 4\%$ in `plurality@k` without degrading `pass@1`. Additionally, we prove that improvements in `pass@k` are closely tied to the mutual information objective, providing a theoretical justification for UpSkill.

1 INTRODUCTION

LLMs excel at verifiable reasoning tasks such as mathematical problem solving and code generation (Guo et al., 2025). However, repeated sampling often yields highly similar outputs (Shaier et al., 2025). This is detrimental in multi-attempt settings where just one correct completion solves the problem at hand, such as code generation with tests (Chen et al., 2021) or formal proofs in Lean (Trinh et al., 2024), as a lack of diversity reduces the effective number of independent attempts. Therefore, for these or other objectives evaluated by `pass@k`, or the probability that at least one of k completions will be correct, more deterministic output decreases 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 (Chen et al., 2025; Dang et al., 2025), creating a discrepancy between how models are trained and how they are used and evaluated.

The challenge of balancing diversity and accuracy, or exploration and exploitation (Sutton and Barto, 2015), has primarily been studied in prior works that change how decoding is done. Methods such as temperature sampling (Renze and Guven, 2024), nucleus sampling (Holtzman et al., 2020), and prompt perturbations (Shur-Ofry et al., 2024) can inject variety, but they require manual tuning (Du et al., 2025) and are brittle across domains (Shi et al., 2024; Qiang et al., 2024). Separately, prior training-time methods do not expose a controllable basis of strategies and require a complex training

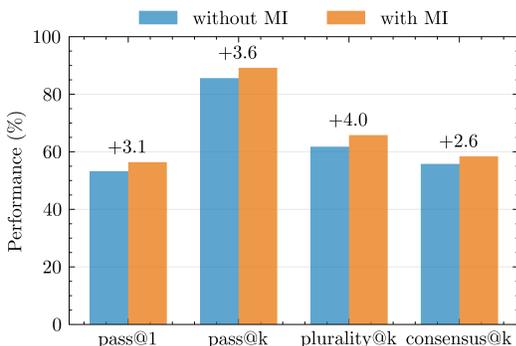
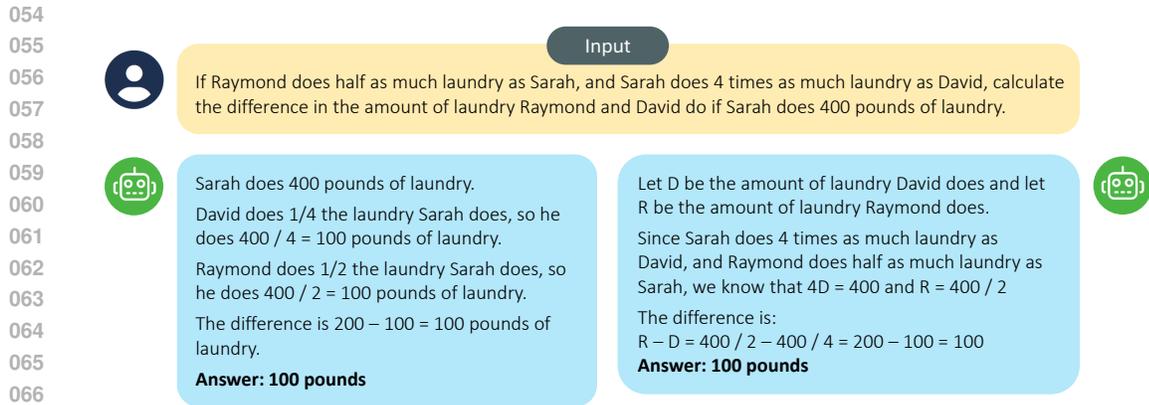


Figure 1: Token-level MISL improves multi-attempt accuracy without hurting single-attempt accuracy on GSM8K for the Qwen 2.5-7B model (See Sec. 5.2).



069
070
071
072
073
074

Figure 2: UpSkill is an unsupervised method for training LLMs to produce diverse responses. After training, different latent vectors z (blue boxes above) correspond to different response strategies. Because of space constraints, the figure shows summarized responses from UpSkill; we report the full responses in [Appendix K](#).

075
076
077
078

regime to properly balance exploration and exploitation (Tang et al., 2025; Chen et al., 2025). 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.

079
080
081
082
083
084
085
086
087
088
089

We introduce UpSkill, a training-time approach that induces *structured response diversity* without prompt engineering. The key idea behind UpSkill is to introduce an input token z that structures the response, so that different values of z correspond to different responses. Formally, we will model LLM attempts on verifiable reasoning tasks as a token-level Markov decision process. We can then adopt prior work from reinforcement learning on learning *skills*, which learn a policy conditioned on a latent variable z . These methods (Eysenbach et al., 2019; Gregor et al., 2016a; Achiam et al., 2018; Sharma et al., 2020a; Florensa et al., 2017) include a loss term that maximizes the mutual information between z and the policy’s behavior. Precisely, we adapt the CSF method (Zheng et al., 2024) 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 strategy, and the set of strategies should span a broad range of behaviors.

090
091
092
093
094

The main contribution of our paper is a method for training LLMs to produce diverse responses. Our method implements mutual information skill learning by applying GRPO (Shao et al., 2024) with a novel reward term: a token-level mutual information reward, which encourages diversity in completions. Finally, we sketch a theoretical link between $\mathcal{I}(\tau; z | x)$ and $\text{pass}@k$: the improvement of $\text{pass}@k$ after training is related to $\mathcal{I}(\tau; z | x)$. In summary, our contributions are as follows:

- 095
096
097
098
099
- UpSkill achieves gains of +3.6% in $\text{pass}@k$ and +4.0% in $\text{plurality}@k$ on GSM8K on Qwen 2.5–7B using RL fine-tuning with LoRA adapters on 2,000 problems, with preserved $\text{pass}@1$ accuracy. We additionally test two other open-weight models, although we do not observe improved performance.
- 100
101
102
- In an arithmetic puzzle environment, UpSkill improves $\text{pass}@5$ by +10% by mitigating response variation collapse and developing a collection of diverse and complementary skills.
- 103
104
105
106
- We prove that $\text{pass}@k$ improvement closely corresponds to the mutual information $\mathcal{I}(\tau; z | x)$, showing that large improvements in multi-attempt accuracy require — and are limited by — sufficient mutual information.
- 107
- We provide an effective and reproducible method for token-level MI, and will release an open-source implementation focused on practical performance.

2 BACKGROUND AND RELATED WORK

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 (Chen et al., 2025). Let x denote the input and τ a sampled completion from policy $\pi(\cdot | x)$. Let $Y(\tau) \in \{0, 1\}$ indicate correctness under a deterministic verifier. For k attempts, the standard metric

$$\text{pass}@k(x) = 1 - \Pr\left(\bigcap_{i=1}^k \{Y(\tau_i) = 0\} \mid x\right) \quad (1)$$

is the complement of the joint failure probability across k i.i.d. draws $\tau_{1:k} \sim \pi(\cdot | x)$ (Chen et al., 2021). Letting $p = \Pr(Y(\tau) = 1 | x)$, 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 (Vijayakumar et al., 2018). 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) (Kish, 1965). 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 $\rho \rightarrow 1$, additional attempts contribute little; as $\rho \rightarrow 0$, $k_{\text{eff}} \rightarrow k$. 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 . For Gaussian random variables, correlation and mutual information are closely related (as intuitively, correlated variables have information on each other) (Krafft, 2013). However, as text correctness cannot easily be framed as a Gaussian distribution, mutual information is more a natural measurement.

Beyond `pass@k`, `plurality@k` and `consensus@k` measure agreement among completions, examining robustness and internal consistency of the model’s reasoning (Wallace et al., 2025). In many workflows, agreement acts as a proxy for confidence while still benefiting from diversity to escape shared failure modes (Hochlehnert et al., 2025).

2.2 RL ON LANGUAGE MODELS: TOKEN-LEVEL MDPs, RLVR, AND GRPO

Autoregressive LLMs can be cast as policies over a Markov decision process (MDP), where the state is the token prefix and the action is the next token (Bahdanau et al., 2017; Ouyang et al., 2022). This setup, often referred to as the token-level MDP (Zhong et al., 2025), allows reinforcement learning algorithms to directly optimize model behavior for correctness on verifiable tasks such as math or code.

Reinforcement Learning from Verifiable Rewards (RLVR) leverages automatically checkable signals (e.g., exact numeric answers, unit tests) as rewards, with the goal being to improve the pass rate of policies while ensuring the new policy remains close to a base model (Liu et al., 2023; Dou et al., 2024). Let π_θ denote the trainable policy and π_{base} a frozen reference. A common form of the per-trajectory reward is

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

where $\beta > 0$ controls deviation from the base model (Xiong et al., 2024).

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) 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 (Schulman et al., 2017). GRPO typically improves `pass@1` on math/code under RLVR (Shao et al., 2024). However, absent any explicit term for diversity, it can *reduce* variation across attempts as the policy sharpens around locally high-reward regions (Dang et al., 2025).

As some intuition for this distribution change, suppose that the model is attempting to predict the correct answer in a setting where it believes that the answer is `Yes` with probability 70% and `No` with probability 30%. Cross-entropy loss encourages a model to predict the correct distribution of

162 70% Yes and 30% No; on the other hand, GRPO training would cause the model to collapse its
 163 output distribution towards predicting 100% Yes, as it maximizes the `pass@1`. Empirically, this can
 164 shrink the entropy of the completion distribution and heighten redundancy among attempts, limiting
 165 `pass@k` improvements even as `pass@1` increases (Dang et al., 2025).
 166

167 2.3 MUTUAL INFORMATION AND SKILL DISCOVERY

168 Maximizing mutual information (MI) between latent variables and observed behavior has been a
 169 recurring tool for learning structured, controllable representations (Tishby et al., 2000; Kingma and
 170 Welling, 2013; Stratos and Wiseman, 2020).
 171

172 In generative modeling, InfoGAN (Chen et al., 2016) augments GAN training with a variational lower
 173 bound on $\mathcal{I}(c; x)$ to make latent codes c predictably control semantic factors (e.g., stroke thickness
 174 for MNIST). In variational autoencoders, InfoVAE (Zhao et al., 2018b) adds an explicit MI term to
 175 counteract posterior collapse and preserve informative latents even with expressive decoders.

176 In sequential decision making, MI has been used to discover diverse, reusable behaviors without
 177 external rewards. Early work such as VIC (Gregor et al., 2016b) and DIAYN (Eysenbach et al., 2018)
 178 maximizes $\mathcal{I}(s; z)$ or $\mathcal{I}(\tau; z)$, encouraging skills z whose rollouts visit different parts of state or
 179 trajectory space and remain identifiable from observations. InfoGAIL (Li et al., 2017) extends this
 180 to imitation learning by maximizing MI between a latent intention and trajectories to capture multi-
 181 modal expert behavior. Subsequent methods bias the MI objective toward long-horizon distinctiveness
 182 to avoid trivial short-term variation (Sharma et al., 2020b; Hansen et al., 2021). Additional related
 183 work on MI is available in Appendix B.

184 Unsupervised skill discovery in RL can be viewed as maximizing the MI between a latent “skill”
 185 variable and observed trajectories (Gregor et al., 2016b; Eysenbach et al., 2018). Let $z \in \mathcal{Z}$ index a
 186 skill and let τ denote a trajectory. These methods maximize

$$187 \max_{\pi} \mathcal{I}(\tau; z | x) = \mathbb{E} \left[\log p_{\pi}(\tau | x, z) - \log p_{\pi}(\tau | x) \right] = \mathcal{H}(\tau | x) - \mathcal{H}(\tau | x, z). \quad (3)$$

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

193 Our setting is closest in spirit to unsupervised skill discovery (e.g., DIAYN, VIC) and to mutual
 194 information-based skill learning (Zheng et al., 2024), but differs in applying these techniques to
 195 language models with RLVR training. We develop an approach for maximizing MI tailored to LLM
 196 reasoning, and also connect `pass@k` performance with the mutual information objective.
 197

198 2.4 OTHER TECHNIQUES FOR RESPONSE DIVERSIFICATION

199 Beyond MI-based training, several other approaches aim to increase output diversity.
 200

201 At inference time, decoding-time diversification alters sampling: increasing temperature, switching to
 202 nucleus/top- k sampling (Holtzman et al., 2020; Fan et al., 2018), or perturbing prompts (Qiang et al.,
 203 2024). While simple, these approaches face limitations: (i) they often fail to explore qualitatively
 204 distinct solution paths (Nguyen et al., 2025; Renze and Guven, 2024); (ii) they require domain-specific
 205 tuning (Wiher et al., 2022); and (iii) they can trade off against correctness and coherence (Nguyen
 206 et al., 2025). Prompt-cycling can inject domain knowledge (e.g., “try algebra” vs. “try geometry”),
 207 but it burdens users with prompt engineering and saturates well below human diversity (Shur-Ofry
 208 et al., 2024).

209 Determinantal point processes (DPPs) provide another path by rewarding sets of outputs that span
 210 high-volume regions in embedding space (Kulesza, 2012; Vijayakumar et al., 2018; Meister et al.,
 211 2023; Wang et al., 2024). In reinforcement learning, determinant-based rewards can encourage agents
 212 to explore trajectories that span complementary regions of state space (Ash et al., 2021; Zhao et al.,
 213 2024). Compared to MI, which directly couples a latent z with trajectories to ensure reproducible
 214 modes, DPP-based diversity is distribution-free: it treats a set of samples as diverse if they occupy a
 215 high-volume region in representation space, regardless of whether the same diversity is reproducible
 under repeated sampling.

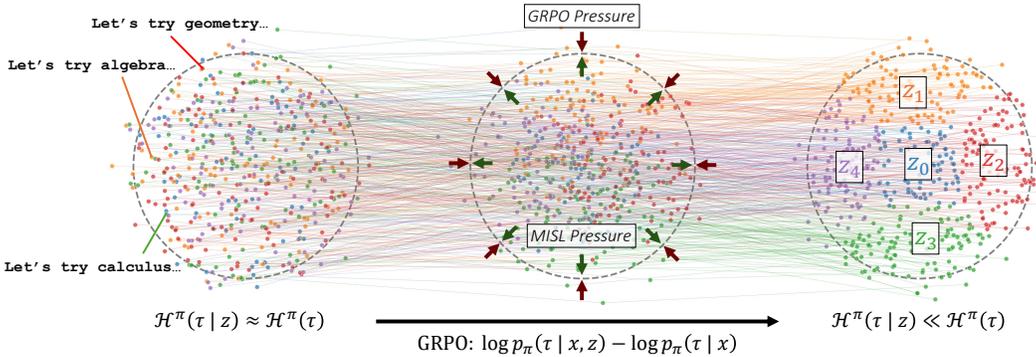


Figure 3: Example illustration of how the MISL reward improves `pass@k` performance. Before MISL (left), the trajectory distribution is independent of the latents z , so the conditional entropy is close to the marginal. MISL training prevents distribution collapse due to `pass@1` training (middle). Adding the token-level MI reward (right) yields well-separated clusters indexed by z , reducing conditional entropy while preserving high marginal entropy. At inference, fixing different z values produces consistent and diverse solution strategies.

Finally, training-time diversification has also been studied through explicit `pass@k`-based objectives. Tang et al. (2025) proposed an unbiased estimator for generic k -attempt objectives, showing overall improved model efficacy. Extending this, Chen et al. (2025) argue that simply training on `pass@1` falls victim to a local maximum of over-exploitation and reduced exploration. They find that `pass@k` training naturally focusing optimization efforts on harder problems producing significant improvements in both `pass@k` and `pass@1`. Outside of verifiable domains, DivPO (Lanchantin et al., 2025) alters preference optimization by contrasting diverse high-quality responses with common low-quality ones using a predefined diversity objective, yielding large diversity gains on creative and instruction-following tasks.

As we provide an orthogonal method to improve `pass@k` and diversity, our approach may complement that of Chen et al. (2025), Tang et al. (2025), and Lanchantin et al. (2025).

3 OPTIMIZING LLM DIVERSITY WITH MUTUAL INFORMATION

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

3.1 OBJECTIVE

We would like to encourage *structured response diversity* by maximizing the conditional mutual information $\mathcal{I}(\tau; z | x)$. Intuitively, maximizing mutual information makes the outputs of different strategies distinguishable, ensuring that each z induces a reliably different mode. By querying each strategy once, we obtain k semantically distinct attempts. Formally, this corresponds to maximizing

$$\max_{\pi} \mathcal{I}(\tau; z | x) = \mathbb{E} \left[\log p_{\pi}(\tau | x, z) - \log p_{\pi}(\tau | x) \right], \tag{4}$$

which increases the overall entropy of trajectories while reducing the conditional entropy within each z -mode, ensuring diverse yet reproducible strategies. The term $p_{\pi}(\tau | x) = \frac{1}{N} \sum_{z'=1}^N p_{\pi}(\tau | x, z')$ 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 . Figure 3 provides an overview of the relevant dynamics.

Algorithm 1 UpSkill: A method for training LLMs to produce diverse responses with mutual information.

-
- 1: **Inputs:** base policy π_{base} , trainable policy π , latent count N , completions per group C , weights $(\alpha_1, \alpha_2, \beta)$
 - 2: **repeat**
 - 3: Sample a minibatch of prompts $\{x\}$
 - 4: **for** each x in the minibatch **do**
 - 5: Sample $z \sim \text{Unif}(\{1, \dots, N\})$; generate C completions $\{\tau_i\}_{i=1}^C$ with $\pi(\cdot | x, z)$
 - 6: Compute $r_{\text{corr}}(\tau_i)$, $r_{\text{TMI}}(\tau_i; x, z)$, and $\Delta_{\text{KL}}(\tau_i)$ as above
 - 7: **end for**
 - 8: Form per-sample rewards via equation 6; compute advantages; update π with GRPO
 - 9: **until** convergence
-

3.2 TOKEN-LEVEL MUTUAL INFORMATION REWARD

We now focus on implementing the mutual information as a token-level reward. For each pair (x, z) , let $\{\tau_i\}_{i=1}^C$ be C completions sampled from $\pi(\cdot | x, z)$. We define a *per-sample* token-level score

$$r_{\text{TMI}}(\tau_i; x, z) = \sum_{t=1}^C \left[\log p_{\pi}(y_t | x, z, y_{<t}) - \log p_{\pi}(y_t | x, y_{<t}) \right], \quad (5)$$

where the second term is the uniform mixture

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

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 5). Since $\frac{1}{C} r_{\text{TMI}}(\tau_i; x, z)$ is a Monte Carlo estimator of $\mathcal{I}(\tau; z | x)$, we make this our main reward term with UpSkill, with the other reward term being considered in ablation experiments. Appendix D discusses an alternative based on semantic mutual information.

3.3 COMBINED RL OBJECTIVE

Let $r_{\text{corr}}(\tau_i) \in \mathbb{R}$ denote the verifiable correctness reward (often binary) and define the per-trajectory KL penalty as

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

The per-sample scalar reward is

$$r(\tau_i; x, z) = r_{\text{corr}}(\tau_i) - \beta \Delta_{\text{KL}}(\tau_i) + \alpha_1 r_{\text{TMI}}(\tau_i; x, z), \quad (6)$$

with $\alpha_1, \beta \geq 0$. We apply GRPO to optimize the sum of the combined rewards.

3.4 TRAINING PROCEDURE

We fine-tune a trainable policy π_{θ} with GRPO while injecting a discrete strategy variable $z \in \{1, \dots, N\}$. At each step, we draw a minibatch of prompts x and, for each x , sample a strategy z uniformly and generate C completions $\tau_{1:C} \sim \pi_{\theta}(\cdot | x, z)$ 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. We then update the policy with GRPO on this reward.

3.5 INFERENCE

Given a budget of k attempts, we choose k distinct latents from $\{1, \dots, N\}$ and generate one completion per latent under fixed decoding hyperparameters. Aggregation (e.g., majority vote) can optionally 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.

4 THEORETICAL CONNECTION BETWEEN PASS@K IMPROVEMENT AND MUTUAL INFORMATION

Our main theoretical result shows that the mutual information objective is closely tied to pass@k . In particular, we will show that the mutual information objective is a lower bound on *improvement* in the pass@k objective, so maximizing mutual information provably results in an increased (lower bound on) pass@k . Our theoretical results will require the following assumptions:

1. k -uniform mixture model: Assume that the marginal distribution over the skills is identical to the base model.
2. Distributional impact: Let a_z be the probability of success of strategy z and a be the probability of success of the base model. Assume that for all $x \in \mathcal{X}$ there exists $\eta > 0$ such that for all $z \in [k]$, $|a_z - a| \geq \eta \delta(\pi_{M,z}(\cdot | x), \pi_B(\cdot | x))$, where δ is the total variation distance.

The second assumption says that the distribution shifts induced by UpSkill correspond to different problem approaches, and, as a result, will have different probabilities of success. A more precise definition of k -uniform mixture models, additional justification for the assumptions, and the statement and proof of the lemma are in [Appendix C](#).

Lemma 1. Let pass@k_B be the pass@k score of the base model on prompt x and pass@k_M be the pass@k score of the mixture model on prompt x . Under the above assumptions, we show that:

$$1 - \exp(-C_1 \eta^2 \mathcal{I}(\tau; z | x)^2) \leq \frac{\text{pass@k}_M - \text{pass@k}_B}{1 - \text{pass@k}_B} \leq 1 - \exp(-C_2 \mathcal{I}(\tau; z | x))$$

where C_1 depends on k and C_2 depends on k and $\max_z a_z$.

The quantity in the middle can be interpreted as the fraction of possible improvement from the base model that is realized by the mixture model. Since monotonically increasing functions in $\mathcal{I}(\tau; z | x)$ provide both lower and upper bounds on how much the mixture model improves over the base model, it makes sense to optimize directly for the mutual information. UpSkill explicitly increases $\mathcal{I}(\tau; z | x)$ during training, ensuring diversity across skills and giving a guaranteed improvement in pass@k over the base model.

5 EXPERIMENTS

We evaluate whether conditioning on a discrete latent z and training with our token-level mutual-information reward equation 5 improves multi-attempt metrics. We present results in two settings: (1) a controlled arithmetic environment that allows for fully verifiable evaluation and direct inspection of distributional effects, and (2) the GSM8K benchmark across three open-weight models. We then report ablations on the number of strategies N and on adding a semantic-MI surrogate.

5.1 ARITHMETIC ENVIRONMENT

The arithmetic environment consists of prompts with three single-digit integers and a latent skill index $z \in \{0, \dots, N-1\}$. A small transformer model chooses one of the integers to be a target and is required to produce a simple arithmetic expression with the other two digits that evaluates to this target. We use an automatically verified correctness reward, and training uses GRPO without a KL penalty. Full details of the environment, model, and training procedure are provided in [Appendix E](#).

Figure 4 illustrates that in the control condition ($\alpha_1=0$), training quickly collapses to a single deterministic strategy: by the end of training, pass@1 and pass@5 are indistinguishable (0.793), offering no benefit from multiple attempts. In contrast, with UpSkill ($\alpha_1=0.5$, MI reward cap at 1.0), the model maintains diverse trajectories, yielding a substantially higher multi-attempt accuracy ($\text{pass@5} = 0.897$) despite a lower single-attempt accuracy ($\text{pass@1} = 0.390$). This difference aligns with the entropy dynamics: UpSkill preserves broad output distributions (token entropy changed from 0.723 to 0.797 over training), sustaining diverse strategies and higher pass@5 . By contrast, the control run—while substantially improving pass@1 —collapses to near-deterministic

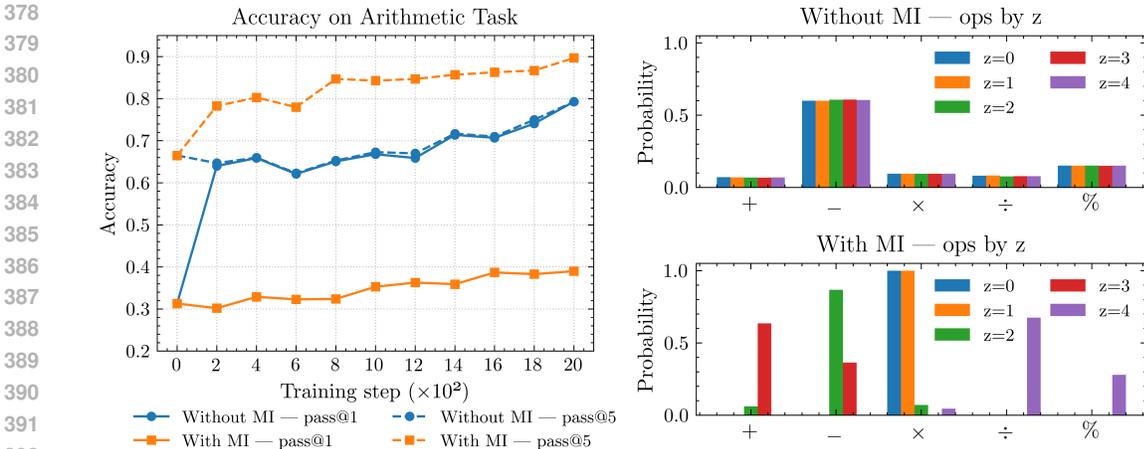


Figure 4: Arithmetic environment results. Training curves show that under GRPO alone (blue), `pass@1` and `pass@5` converge together, indicating that multiple attempts provide little benefit. With MISL (orange; $N=5$), `pass@5` improves substantially while `pass@1` remains modest, demonstrating that different latents yield complementary solutions. Operator distributions further highlight this effect: without MISL, they are nearly identical across z , reflecting a lack of specialization, whereas with MISL, distinct latents focus on different operators, producing diverse strategies that drive multi-attempt gains.

outputs (with entropy 0.723 changing to 0.030), leaving `pass@5` identical to `pass@1` (see Appendix E.5 for more details).

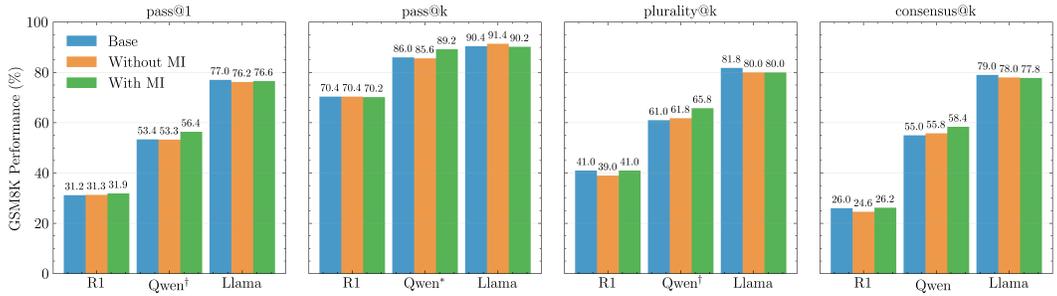
Figure 4 illustrates that under UpSkill, different z values yield distinct distributions over operators and digits, whereas the control produces nearly identical distributions across z . In this small-scale environment, we can directly observe the learned strategies, and notably $z = 1$ and $z = 2$ converge to risky yet common modes, whereas other values of z cover the remaining operations, improving multi-attempt success with a strategy infeasible for optimizing a `pass@1` objective. The distributions over the first digit are available in Appendix E.7.

We additionally ablate the impact of starting model capabilities, the coefficient of KL penalty, and GRPO parameters (see Appendix F for full details and results). KL penalty β discourages entropy collapse by ensuring the new policy remains close to the initial policy, thereby improving performance. On models with $\beta \in [0.05, 0.10]$, we test $\alpha_1 \in [0.1, 0.3, 0.5]$ and find that well-chosen MI-reward parameters increase `pass@k` by an average of 3% for the weaker base model. However, for the stronger base model, we find the opposite trend. It is always best to choose $\alpha_1 = 0$, with the best choice of $\alpha_1 \in [0.1, 0.3, 0.5]$ still leading to a 1.2% performance decrease. Our theoretical results in Lemma 1 suggest that UpSkill improvement is negatively related to `pass@1` (equivalently `pass@kB`) capability, and thus, although surprising, this result is in line with our theoretical analysis. We separately conjecture that β , which corresponds with a decrease in exploration from the base policy, conflicts with the mutual information incentive to explore.

5.2 GSM8K

We next evaluate on GSM8K (Cobbe et al., 2021), a dataset of grade-school arithmetic word problems. We use 2,000 training problems and a held-out set of 500 questions. All experiments are conducted in a zero-shot setting with a maximum sequence length of 1024 tokens. We train LoRA adapters (approximately 80M trainable parameters) on top of three open-weight backbones: Llama 3.1-8B (Grattafiori et al., 2024), Qwen 2.5-7B (Qwen et al., 2025), and R1-Distilled-Qwen2.5-Math-1.5B (DeepSeek-AI et al., 2025). As before, we apply GRPO with a correctness reward and with default KL penalty, and UpSkill is applied at the token level as in Eq. (6). At inference, we fix k distinct skill indices and generate one completion per skill. More details and ablations are available in Appendices G, H, and I. Figure 5 shows our performance on the withheld evaluation set. Token-level MISL improves results on Qwen 2.5-7B on all metrics. The other tested models are generally

432 unaffected. Interestingly, these results do not entirely parallel the arithmetic environment, as UpSkill’s
 433 improvement on Qwen 2.5–7B has not come at the cost of `pass@1`. We hypothesize this is due to
 434 the existence of a larger set of correct reasoning approaches, and thus MISL does not necessarily
 435 come at a cost of `pass@1`. Chen et al. (2025) have separately found that `pass@k` training methods
 436 can improve `pass@1` performance. We include summarized outputs from Qwen in Figure 2 and the
 437 isolated Qwen results in Figure 1.



449 Figure 5: Performance on 500 held-out problems with $N=5$ strategies. We observe gains on all
 450 metrics for the Qwen model. Base refers to the model before GRPO training, Without MI refers
 451 to after GRPO training without token MI, and With MI refers to training with GRPO and token MI.
 452 An asterisk denotes $p < 0.02$ and † denotes $p < 0.06$ from McNemar’s test on the hypothesis ‘With
 453 MI’ outperforms ‘Without MI’.

454
 455
 456 **5.3 GSM8K ABLATIONS**

457 We conduct two ablations to probe the robustness of the approach. First, increasing the number of
 458 skills beyond $N=5$ produces mixed results. Many GSM8K problems admit only a limited number
 459 of distinct solution paths, so larger N values fragment the capacity into modes that do not translate
 460 into additional gains. Second, we study the effect of replacing the token-level MI with a semantic
 461 MI, which we formally introduce in Appendix D. This semantic MI occasionally yields further
 462 improvements but introduces instability due to estimator variance in high dimensions. We provide
 463 more details in Appendix J. Therefore, our main results use token-level MISL only, with semantic
 464 variants left as a direction for future work.

465
 466
 467 **6 CONCLUSION**

468 Our experiments show that UpSkill provides a simple and effective way to induce strategy-level
 469 diversity in LLMs, leading to gains on multi-attempt metrics such as `pass@5` and `plurality@5`.
 470 By conditioning on discrete latent variables, the model learns reproducible modes of reasoning that
 471 reduce redundancy across attempts and increase the likelihood of success. Beyond these empirical
 472 findings, UpSkill provides a principled training-time approach for improving response diversity, and
 473 our analysis links $\mathcal{I}(\tau; z | x)$ to upper bounds in `pass@k` improvement in training. We hope this
 474 stimulates research on robust semantic diversity signals and theoretical ties between information-
 475 theoretic objectives and multi-attempt success.

476
 477
 478 **Limitations.** There are a few notable limitations with our work. Assumption 1 is rather limiting
 479 and difficult to enforce empirically, so we hope to find a natural way to either incorporate it into our
 480 training method or find another assumption that can also prove Lemma 1. As future work, we hope
 481 for further experimentation across domains and with larger models, to better study how our method
 482 is affected by KL penalty and base model performance, and to approach the theoretical guarantees
 483 from the perspective of Tsallis entropy (Furuichi, 2006). The Llama and R1 models had unstable RL
 484 training, causing performance to degrade from the base model; we will switch to a more reliable RL
 485 framework like verl (Sheng et al., 2024) and use a more robust prompting method than an already
 semantically meaningful prefix string.

486 **Reproducibility.** We make several efforts towards encouraging reproducibility of our work and
 487 results. **We have included all experiment code in the supplementary files**, along with instructions
 488 to reproduce our results. We use only models under permissive licenses. For our theoretical results, we
 489 provide a description of assumptions and the complete proof of claims in the appendix. Additionally,
 490 for various experiments whose full results could not fit in the main paper, we include the full results
 491 in the appendix.

492

493 REFERENCES

494

495 Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery
 496 algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

497 Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone Fishing: Neural Active
 498 Learning with Fisher Embeddings, December 2021. URL [http://arxiv.org/abs/2106.](http://arxiv.org/abs/2106.09675)
 499 [09675](http://arxiv.org/abs/2106.09675). arXiv:2106.09675 [cs].

500

501 Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron
 502 Courville, and Yoshua Bengio. An Actor-Critic Algorithm for Sequence Prediction, March 2017.
 503 URL <http://arxiv.org/abs/1607.07086>. arXiv:1607.07086 [cs].

504 Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron
 505 Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2021. URL
 506 <https://arxiv.org/abs/1801.04062>.

507

508 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared
 509 Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri,
 510 Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan,
 511 Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian,
 512 Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios
 513 Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino,
 514 Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders,
 515 Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa,
 516 Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob
 517 McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating
 518 Large Language Models Trained on Code, July 2021. URL [http://arxiv.org/abs/2107.](http://arxiv.org/abs/2107.03374)
 519 [03374](http://arxiv.org/abs/2107.03374). arXiv:2107.03374 [cs].

519 Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan:
 520 Interpretable representation learning by information maximizing generative adversarial nets, 2016.
 521 URL <https://arxiv.org/abs/1606.03657>.

522

523 Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi.
 524 Pass@k training for adaptively balancing exploration and exploitation of large reasoning models,
 525 2025. URL <https://arxiv.org/abs/2508.10751>.

526 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
 527 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve
 528 math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

529

530 Xingyu Dang, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Assessing diversity collapse in
 531 reasoning. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL
 532 <https://openreview.net/forum?id=AMiKsHLjQh>.

533 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,
 534 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,
 535 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao
 536 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
 537 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,
 538 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,
 539 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang
 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong,

- 540 Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao,
541 Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang,
542 Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang,
543 Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L.
544 Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang,
545 Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng
546 Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng
547 Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan
548 Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang,
549 Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen,
550 Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li,
551 Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang,
552 Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan,
553 Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia
554 He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong
555 Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha,
556 Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang,
557 Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li,
558 Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen
559 Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
URL <https://arxiv.org/abs/2501.12948>.
- 560 Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Junjie Shan, Caishuang Huang, Wei
561 Shen, Xiaoran Fan, Zhiheng Xi, Yuhao Zhou, Tao Ji, Rui Zheng, Qi Zhang, Xuanjing Huang, and
562 Tao Gui. StepCoder: Improve Code Generation with Reinforcement Learning from Compiler
563 Feedback, February 2024. URL <http://arxiv.org/abs/2402.01391>. arXiv:2402.01391
564 [cs] version: 1.
- 565 Weihua Du, Yiming Yang, and Sean Welleck. Optimizing Temperature for Language Models with
566 Multi-Sample Inference, February 2025. URL <http://arxiv.org/abs/2502.05234>.
567 arXiv:2502.05234 [cs] version: 1.
- 568 Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is All You Need:
569 Learning Skills without a Reward Function, October 2018. URL <http://arxiv.org/abs/1802.06070>. arXiv:1802.06070 [cs].
- 570 Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you
571 need: Learning skills without a reward function. In *International Conference on Learning
572 Representations*, 2019. URL <https://openreview.net/forum?id=SJx63jRqFm>.
- 573 Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation, 2018. URL
574 <https://arxiv.org/abs/1805.04833>.
- 575 Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforce-
576 ment learning. *arXiv preprint arXiv:1704.03012*, 2017.
- 577 Shigeru Furuichi. Information theoretical properties of Tsallis entropies. *Journal of Mathematical
578 Physics*, 47(2):023302, February 2006. ISSN 0022-2488, 1089-7658. doi: 10.1063/1.2165744.
579 URL <http://arxiv.org/abs/cond-mat/0405600>. arXiv:cond-mat/0405600.
- 580 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
581 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,
582 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev,
583 Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru,
584 Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak,
585 Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu,
586 Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle
587 Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego
588 Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,
589 Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel
590 Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon,
591

594 Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan
595 Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet,
596 Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde,
597 Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie
598 Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua
599 Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak,
600 Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley
601 Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence
602 Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas
603 Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri,
604 Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie
605 Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes
606 Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne,
607 Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal
608 Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,
609 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,
610 Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie
611 Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana
612 Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie,
613 Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon
614 Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan,
615 Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas
616 Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami,
617 Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti,
618 Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier
619 Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao
620 Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song,
621 Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe
622 Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya
623 Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei
624 Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu,
625 Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit
626 Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury,
627 Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer,
628 Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu,
629 Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido,
630 Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu
631 Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer,
632 Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu,
633 Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc
634 Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily
635 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers,
636 Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank
637 Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee,
638 Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan,
639 Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph,
640 Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog,
641 Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James
642 Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny
643 Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings,
644 Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai
645 Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik
646 Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle
647 Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng
Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish
Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim
Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle
Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,
Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam,

- 648 Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier,
649 Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia
650 Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro
651 Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani,
652 Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy,
653 Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin
654 Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu,
655 Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh
656 Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay,
657 Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang,
658 Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie
659 Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta,
660 Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman,
661 Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun
662 Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria
663 Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,
664 Vlad Tiberiu Mihalescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz,
665 Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv
666 Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,
667 Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait,
668 Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The
llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- 669 Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv*
670 *preprint arXiv:1611.07507*, 2016a.
- 671 Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational Intrinsic Control, November
672 2016b. URL <http://arxiv.org/abs/1611.07507>. arXiv:1611.07507 [cs].
- 673 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
674 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-rl: Incentivizing reasoning capability in llms
675 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 676 Steven Hansen, Guillaume Desjardins, Kate Baumli, David Warde-Farley, Nicolas Heess, Simon
677 Osindero, and Volodymyr Mnih. Entropic Desired Dynamics for Intrinsic Control. In *Ad-*
678 *vances in Neural Information Processing Systems*, volume 34, pages 11436–11448. Curran Asso-
679 ciates, Inc., 2021. URL [https://proceedings.neurips.cc/paper_files/paper/](https://proceedings.neurips.cc/paper_files/paper/2021/hash/5f7f02b7e4ade23430f345f954c938c1-Abstract.html)
680 [2021/hash/5f7f02b7e4ade23430f345f954c938c1-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2021/hash/5f7f02b7e4ade23430f345f954c938c1-Abstract.html).
- 681 Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and
682 Matthias Bethge. A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Re-
683 producibility, April 2025. URL <http://arxiv.org/abs/2504.07086>. arXiv:2504.07086
684 [cs].
- 685 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text
686 degeneration, 2020. URL <https://arxiv.org/abs/1904.09751>.
- 687 Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation
688 learning for non-parallel text style transfer, 2018. URL [https://arxiv.org/abs/1808.](https://arxiv.org/abs/1808.04339)
689 [04339](https://arxiv.org/abs/1808.04339).
- 690 Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2013. URL
691 <http://arxiv.org/abs/1312.6114>. arXiv:1312.6114 [stat].
- 692 Leslie Kish. *Survey Sampling*. Wiley, 1965.
- 693 Peter Krafft. Correlation and mutual information. [https://lips.cs.princeton.edu/](https://lips.cs.princeton.edu/correlation-and-mutual-information/)
694 [correlation-and-mutual-information/](https://lips.cs.princeton.edu/correlation-and-mutual-information/), February 2013. Laboratory for Intelligent
695 Probabilistic Systems, Princeton University Department of Computer Science.
- 696 Alexander Kraskov, Harald Stoeckbauer, and Peter Grassberger. Estimating Mutual Information.
697 *Physical Review E*, 69(6):066138, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/

- 702 PhysRevE.69.066138. URL [http://arxiv.org/abs/cond-](http://arxiv.org/abs/cond-mat/0305641)
703 [mat/0305641](http://arxiv.org/abs/cond-mat/0305641). arXiv:cond-
704 mat/0305641.
- 705 Alex Kulesza. Determinantal point processes for machine learning. *Foundations and Trends® in*
706 *Machine Learning*, 5(2–3):123–286, 2012. ISSN 1935-8245. doi: 10.1561/22000000044. URL
707 <http://dx.doi.org/10.1561/22000000044>.
- 708 Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar
709 Sukhbaatar, and Ilia Kulikov. Diverse preference optimization, 2025. URL [https://arxiv.](https://arxiv.org/abs/2501.18101)
710 [org/abs/2501.18101](https://arxiv.org/abs/2501.18101).
- 711 Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual
712 demonstrations, 2017. URL <https://arxiv.org/abs/1703.08840>.
- 713
- 714 Jiatae Liu, Yiqin Zhu, Kaiwen Xiao, Qiang Fu, Xiao Han, Wei Yang, and Deheng Ye. RLTF:
715 Reinforcement Learning from Unit Test Feedback, November 2023. URL [http://arxiv.](http://arxiv.org/abs/2307.04349)
716 [org/abs/2307.04349](http://arxiv.org/abs/2307.04349). arXiv:2307.04349 [cs].
- 717 Clara Meister, Martina Forster, and Ryan Cotterell. Determinantal Beam Search, June 2023. URL
718 <http://arxiv.org/abs/2106.07400>. arXiv:2106.07400 [cs].
- 719
- 720 Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-
721 Ziv. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs, June 2025.
722 URL <http://arxiv.org/abs/2407.01082>. arXiv:2407.01082 [cs] version: 7.
- 723 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong
724 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton,
725 Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and
726 Ryan Lowe. Training language models to follow instructions with human feedback, March 2022.
727 URL <http://arxiv.org/abs/2203.02155>. arXiv:2203.02155 [cs].
- 728 Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky,
729 and Aram Galstyan. Prompt Perturbation Consistency Learning for Robust Language Models. In
730 Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Lin-*
731 *guistics: EACL 2024*, pages 1357–1370, St. Julian’s, Malta, March 2024. Association for Computa-
732 tional Linguistics. URL <https://aclanthology.org/2024.findings-eacl.91/>.
- 733
- 734 Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
735 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
736 Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
737 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi
738 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,
739 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL
<https://arxiv.org/abs/2412.15115>.
- 740
- 741 Matthew Renze and Erhan Guven. The Effect of Sampling Temperature on Problem Solving in Large
742 Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*,
743 pages 7346–7356, 2024. doi: 10.18653/v1/2024.findings-emnlp.432. URL [http://arxiv.](http://arxiv.org/abs/2402.05201)
[org/abs/2402.05201](http://arxiv.org/abs/2402.05201). arXiv:2402.05201 [cs].
- 744
- 745 Igal Sason and Sergio Verdú. \mathbb{f} -divergence Inequalities. *IEEE Transactions on Information*
746 *Theory*, 62(11):5973–6006, November 2016. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2016.
747 2603151. URL <http://arxiv.org/abs/1508.00335>. arXiv:1508.00335 [cs].
- 748 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
749 optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>.
- 750 Sagi Shaiyer, Mario Sanz-Guerrero, and Katharina von der Wense. Asking Again and Again: Exploring
751 LLM Robustness to Repeated Questions, March 2025. URL [http://arxiv.org/abs/2412.](http://arxiv.org/abs/2412.07923)
752 [07923](http://arxiv.org/abs/2412.07923). arXiv:2412.07923 [cs].
- 753
- 754 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
755 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

- 756 Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware
757 unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020a.
758 URL <https://openreview.net/forum?id=HJgLZR4KvH>.
759
- 760 Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-Aware
761 Unsupervised Discovery of Skills, February 2020b. URL [http://arxiv.org/abs/1907.](http://arxiv.org/abs/1907.01657)
762 [01657](http://arxiv.org/abs/1907.01657). arXiv:1907.01657 [cs].
- 763 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
764 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint*
765 *arXiv: 2409.19256*, 2024.
- 766 Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam.
767 A Thorough Examination of Decoding Methods in the Era of LLMs. In *Proceedings of the*
768 *2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629,
769 Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.
770 emnlp-main.489. URL <https://aclanthology.org/2024.emnlp-main.489>.
771
- 772 Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim, and Yonatan Belinkov. Growing a
773 Tail: Increasing Output Diversity in Large Language Models, November 2024. URL [http://](http://arxiv.org/abs/2411.02989)
774 arxiv.org/abs/2411.02989. arXiv:2411.02989 [cs].
- 775 Greg Ver Steeg. [gregversteeg/npeet](https://github.com/gregversteeg/npeet), May 2025. URL [https://github.com/](https://github.com/gregversteeg/npeet)
776 [gregversteeg/NPEET](https://github.com/gregversteeg/npeet). original-date: 2014-10-10T19:57:02Z.
777
- 778 Karl Stratos and Sam Wiseman. Learning Discrete Structured Representations by Adversarially
779 Maximizing Mutual Information, July 2020. URL <http://arxiv.org/abs/2004.03991>.
780 arXiv:2004.03991 [cs].
- 781 Richard S Sutton and Andrew G Barto. Reinforcement Learning: An Introduction. 2015.
- 782 Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Rémi Munos. Optimizing language models
783 for inference time objectives using reinforcement learning, 2025. URL [https://arxiv.org/](https://arxiv.org/abs/2503.19595)
784 [abs/2503.19595](https://arxiv.org/abs/2503.19595).
785
- 786 Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, April
787 2000. URL <http://arxiv.org/abs/physics/0004057>. arXiv:physics/0004057.
788
- 789 Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry
790 without human demonstrations. *Nature*, 625(7995):476–482, 2024.
- 791 Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive
792 coding, 2019. URL <https://arxiv.org/abs/1807.03748>.
793
- 794 Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee,
795 David Crandall, and Dhruv Batra. Diverse Beam Search: Decoding Diverse Solutions from
796 Neural Sequence Models, October 2018. URL <http://arxiv.org/abs/1610.02424>.
797 arXiv:1610.02424 [cs].
- 798 Eric Wallace, Olivia Watkins, Miles Wang, Kai Chen, and Chris Koch. Estimating Worst-Case Frontier
799 Risks of Open-Weight LLMs, August 2025. URL <http://arxiv.org/abs/2508.03153>.
800 arXiv:2508.03153 [cs] version: 1.
- 801 Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar
802 Panda. Diversity Measurement and Subset Selection for Instruction Tuning Datasets, February
803 2024. URL <http://arxiv.org/abs/2402.02318>. arXiv:2402.02318 [cs].
804
- 805 Gian Wiher, Clara Meister, and Ryan Cotterell. On Decoding Strategies for Neural Text Generators,
806 March 2022. URL <http://arxiv.org/abs/2203.15721>. arXiv:2203.15721 [cs].
807
- 808 Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong
809 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].

810 Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized
811 autoencoders, 2018a. URL <https://arxiv.org/abs/1706.04223>.

812 Kaiyan Zhao, Yiming Wang, Yuyang Chen, Xiaoguang Niu, Yan Li, and Leong Hou U. Efficient
813 Diversity-based Experience Replay for Deep Reinforcement Learning, October 2024. URL
814 <http://arxiv.org/abs/2410.20487>. arXiv:2410.20487 [cs] version: 1.

815 Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational
816 autoencoders, 2018b. URL <https://arxiv.org/abs/1706.02262>.

817 Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural
818 dialog models using conditional variational autoencoders, 2017. URL <https://arxiv.org/abs/1703.10960>.

819 Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a misl fly? analysis and
820 ingredients for mutual information skill learning. *arXiv preprint arXiv:2412.08021*, 2024.

821 Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian,
822 and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf, 2025. URL <https://arxiv.org/abs/2404.18922>.

823 A USE OF LLMs

824 Large language models were used in the preparation of this work for writing assistance (including
825 polishing, improving presentation of concepts, and restructuring of text), for retrieval and discovery
826 of related work, and for support in producing experimental code and figures. Language models were
827 additionally used for feedback on the paper and to formalize mathematical arguments. All analysis,
828 experimental and method design, and final interpretations are our own. LLM outputs were always
829 rigorously reviewed.

830 B EXTENDED MUTUAL INFORMATION RELATED WORK

831 Estimating MI reliably is challenging in high dimensions. Variational bounds (Barber–Agakov)
832 optimize a classifier or regressor $q_\phi(z|\cdot)$ as a proxy for the intractable posterior (van den Oord et al.,
833 2019). Contrastive bounds such as InfoNCE (van den Oord et al., 2019) reduce MI estimation to
834 noise-contrastive classification and have become standard due to their stability. Neural MI estimators
835 like MINE (Belghazi et al., 2021) directly optimize a Donsker–Varadhan bound but can suffer from
836 bias/variance trade-offs and training instability. Nonparametric k NN estimators (KSG) (Kraskov et al.,
837 2004) avoid parametric critics but require many samples and are sensitive to dimension, motivating
838 careful batching and normalization when used inside policy gradients. In text generation, MI-style
839 objectives have been used to prevent latent collapse and enable controllable generation, e.g., by
840 encouraging informative latents in variational text models (Zhao et al., 2017; 2018a) or aligning
841 codes with style attributes (John et al., 2018). These approaches typically maximize MI between
842 prompts or attributes and latent variables, rather than between a discrete strategy and the full trajectory
843 distribution, and are optimized with supervised losses rather than RL.

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

850 C STATEMENT AND DERIVATION OF THEORETICAL BOUNDS

851 C.1 PROBLEM SETUP AND ASSUMPTIONS

852 Let \mathcal{X} be the set of all possible prompts. The statement of Lemma 1 applies to the general class of
853 k -uniform mixture models.

Definition. A k -uniform mixture model (M, B) is defined to be an ordered pair of a *mixture model*, which is a set of k different policies for generating trajectories, which we will call $\pi_{M,z}(\cdot | x)$ for a prompt $x \in \mathcal{X}$ and strategy $z \in [k]$, along with a *base model* $\pi_B(\cdot | x)$ for generating trajectories subject to the condition that

$$\frac{1}{k} \sum_{z=1}^k \pi_{M,z}(\cdot | x) = \pi_B(\cdot | x) \forall x \in \mathcal{X}.$$

This definition can be interpreted as follows: π_B is the trajectory distribution of the original, non-strategy conditioned language model. If we weigh each strategy as being equally important, we sample once from the mixture model by randomly choosing one strategy. In this case, the joint distribution of trajectories from the mixture is

$$\pi_M(\cdot | x) := \frac{1}{k} \sum_{z=1}^k \pi_{M,z}(\cdot | x).$$

The condition essentially means that the joint distribution of trajectories over picking a strategy uniformly at random must be the same as the original distribution. Therefore, in essence, the mixture model partitions π_B into k different policies that together average back to π_B .

In practice, this condition imposes undue constraints on the strategy distribution, and thus, for the practical implementation, this is not enforced. Also, while in practice one may actually train $N > k$ different strategies and then randomly sample k different strategies so that they still have equal probabilities of being selected, here we make the simplifying assumption that $N = k$, which is true for all of our experiments.

For ease of notation, let $a = \Pr(Y_x(\tau_{B,1}) = 1 | x)$ and $a_z = \Pr(Y_x(\tau_{M,z}) = 1)$ for $z \in [k]$. Because the trajectories $\tau_{M,z}$ are sampled independently, we have that

$$\pi_M(\cdot | x) = \pi_B(\cdot | x) \implies \frac{1}{k} \sum_{z=1}^k a_z = a. \quad (7)$$

We provide additional justification for the second assumption made in [section 4](#). First, if the strategies differ in more than just style and contain meaningful semantic differences, we expect that the difference in success should be proportional to how different these two distributions are. The total variation distance measures this distance in trajectory space, while the constant η controls how sensitive the success probabilities are to changes in the distribution shift.

C.2 EXTENDING `PASS@K` TO k -UNIFORM MIXTURE MODELS

We now extend the traditional definition of `pass@k` to fit the setting of k -uniform mixture models to leverage the fact that we now have a natural structure for querying k different strategies by varying z . This is notably different from the setting in the consistency assumption, which can be interpreted as querying just one strategy uniformly at random.

For a given prompt x and deterministic verifier $Y_x(\tau)$ outputting 1 if τ is a valid output on prompt x and 0 otherwise, and k -uniform mixture model (M, B) , define `pass@kM` be the probability that querying each of these strategies independently exactly once (see [Sec. 3.5](#)) results in at least one correct answer, and define `pass@kB` to be the probability that querying π_B independently k times results in at least one correct answer. Writing this out mathematically, for each $z \in [k]$ we sample $\tau_{M,z} \sim \pi_{M,z}(\cdot | x)$; then

$$\text{pass@k}_M = 1 - \Pr\left(\bigcap_{z=1}^k \{Y_x(\tau_{M,z}) = 0\} \mid x\right) = 1 - \prod_{z=1}^k \Pr(Y_x(\tau_{M,z}) = 0 | x). \quad (8)$$

While we use the same definition of `pass@k` for B as in standard literature, we include it for the sake of completeness; similarly, when sampling $\tau_{B,z} \sim \pi_B(\cdot | x)$ independently for each $z \in [k]$, we find that

$$\text{pass@k}_B = 1 - \Pr\left(\bigcap_{z=1}^k \{Y_x(\tau_{B,z}) = 0\} \mid x\right) = 1 - \Pr(Y_x(\tau_{B,1}) = 0 | x)^k \quad (9)$$

where we simplify the product into the RHS of equation 9 with the independence of samples. Then

$$\text{pass@k}_B = 1 - (1 - a)^k, \text{pass@k}_M = 1 - \prod_{z=1}^k (1 - a_z).$$

We now give the precise statement of the main result.

Lemma 2 (**pass@k Improvement for k -uniform Mixture Models, Full Statement**). Let $u = \max_{z \in [k]} a_z$ and let $\varphi : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}$ be defined as $\varphi(x) = \frac{x \log x}{x-1}$ for $x \neq 0, 1$ and $\varphi(0) = 0, \varphi(1) = 1$. Then

$$1 - \exp\left(-\frac{k\eta^2 \mathcal{I}(\tau; z | x)^2}{2\varphi(k)^2}\right) \leq \frac{\text{pass@k}_M - \text{pass@k}_B}{1 - \text{pass@k}_B} \leq 1 - \exp\left(-\frac{k\mathcal{I}(\tau; z | x)}{4(1-u)^2}\right).$$

C.3 DERIVATION OF LOWER BOUND

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 $z \in [k]$. Summing all of these equations, the linear terms cancel due to equation 7. Then

$$\sum_{z \in [k]} \log(1 - a_z) = k \log(1 - a) + \sum_{z \in [k]} \frac{1}{2} (a_z - a)^2 f''(\xi_z). \quad (10)$$

Let A be the random variable that takes value a_z for each $z \in [k]$ with probability $\frac{1}{k}$. Since $f''(y) = -\frac{1}{(1-y)^2}$ is a decreasing function in the interval $[0, 1]$, we find that $f''(\xi_z) \leq f''(0) = -1$ for all $z \in [k]$, i.e.

$$\begin{aligned} \sum_{z \in [k]} \log(1 - a_z) &\leq k \log(1 - a) - \sum_{z \in [k]} \frac{1}{2} (a_z - a)^2 \\ \implies k \log(1 - a) - \sum_{z \in [k]} \log(1 - a_z) &\geq \frac{k \text{Var}(A)}{2} \implies \frac{1 - \text{pass@k}_B}{1 - \text{pass@k}_M} \geq \exp\left(\frac{k}{2} \text{Var}(A)\right) \\ \implies \text{pass@k}_M - \text{pass@k}_B &= (1 - \text{pass@k}_B) \left(1 - \frac{1 - \text{pass@k}_M}{1 - \text{pass@k}_B}\right) \\ &\geq (1 - \text{pass@k}_B) \left(1 - \exp\left(-\frac{k \text{Var}(A)}{2}\right)\right). \end{aligned}$$

We now find a lower bound for $\text{Var}(A)$ in terms of the mutual information to finish off the proof. Let $\delta(\cdot, \cdot)$ represent the total variation distance between two distributions. We have that $\text{Var}(A) \geq \mathbb{E}[|A|^2]$ by Jensen's Inequality, so using the distributional impact assumption,

$$\begin{aligned} \text{Var}(A) &\geq (\mathbb{E}_{z \sim \text{Unif}\{1, 2, \dots, k\}}[|a_z - a|])^2 = \left(\frac{1}{k} \sum_{z=1}^k |a_z - a|\right)^2 \geq \left(\frac{1}{k} \sum_{z=1}^k \eta \delta(\pi_{M,z}, \pi_B)\right)^2 \\ &= \eta^2 (\mathbb{E}_{z \sim \text{Unif}\{1, 2, \dots, k\}}[\delta(\pi_{M,z}, \pi_B)])^2. \end{aligned}$$

Note that by the assumption that $\pi_M(\cdot | x) = \pi_B(\cdot | x)$ we have that $0 \leq \pi_{M,z}(\cdot | x) \leq k\pi_B(\cdot | x)$ for all $z \in [k]$. Therefore,

$$\frac{d\pi_{M,z}}{d\pi_B}(\tau) \in [0, k].$$

Applying Theorem 26 from [Sason and Verdú \(2016\)](#) with $\beta_2 = 0, \beta_1 = \frac{1}{k}$ where the constants are from the assumption on bounded likelihood ratio, we have that $D_{\text{KL}}(\pi_{M,z} \| \pi_B) \leq \varphi(k) \delta(\pi_{M,z}, \pi_B)$. Summing over $z \in [k]$ and dividing by k yields

$$\mathcal{I}(\tau; z | x) = \mathbb{E}_{z \sim \text{Unif}\{1, 2, \dots, k\}}[D_{\text{KL}}(\pi_{M,z} \| \pi_B)] \leq \varphi(k) \mathbb{E}_{z \sim \text{Unif}\{1, 2, \dots, k\}}[\delta(\pi_{M,z}, \pi_B)].$$

Putting both bounds together, we finally find that

$$\begin{aligned} \text{Var}(A) &\geq \left(\frac{\eta \mathcal{I}(\tau; z | x)}{\varphi(k)}\right)^2 \\ \implies \text{pass@k}_M - \text{pass@k}_B &\geq (1 - \text{pass@k}_B) \left(1 - \exp\left(-\frac{k\eta^2 \mathcal{I}(\tau; z | x)^2}{2\varphi(k)^2}\right)\right) \end{aligned}$$

as desired.

C.4 DERIVATION OF UPPER BOUND

Let $u = \max_z a_z < 1$. In the interval $[\min(a, a_1, a_2, \dots, a_k), \max(a, a_1, a_2, \dots, a_k)]$ f'' achieves its minimum at $f''(u) = -\frac{1}{(1-u)^2}$. Then from equation 10

$$\begin{aligned} \sum_{z \in [k]} \log(1 - a_z) &\geq k \log(1 - a) - \frac{1}{2(1-u)^2} \sum_{z \in [k]} (a_z - a)^2 \\ \implies \prod_{z=1}^k (1 - a_z) &\geq (1 - a)^k \exp\left(-\frac{1}{2(1-u)^2} \sum_{z \in [k]} (a_z - a)^2\right) \\ \implies 1 - \text{pass@k}_M &\geq (1 - \text{pass@k}_B) \exp\left(-\frac{1}{2(1-u)^2} \sum_{z \in [k]} (a_z - a)^2\right). \end{aligned} \quad (11)$$

This places an upper bound on how much we can possibly improve pass@k compared to our original trajectory distributions. Using Pinsker’s Inequality, we find that for all $i \in [k]$,

$$\begin{aligned} |a_i - a| &= |\Pr[Y(\tau_{M,i}) = 1] - \Pr[Y(\tau_{B,i}) = 1]| = \left| \sum_{Y(\tau')=1} \pi_{M,i}(\tau') - \sum_{Y(\tau')=1} \pi_{B,i}(\tau') \right| \\ &\leq \sum_{Y(\tau')=1} |\pi_{M,i}(\tau') - \pi_{B,i}(\tau')| \leq \sum_{\tau'} |\pi_{M,i}(\tau') - \pi_{B,i}(\tau')| \\ &\leq \delta(\pi_{B,i}, \pi_{M,i}) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\pi_{M,i} \parallel \pi_{B,i})} \\ \implies (a_i - a)^2 &\leq \frac{1}{2} D_{\text{KL}}(\pi_{M,i} \parallel \pi_{B,i}) \implies \sum_{z \in [k]} (a_z - a)^2 \leq \frac{k}{2} \cdot \frac{1}{k} \sum_{z \in [k]} D_{\text{KL}}(\pi_{M,i} \parallel \pi_{B,i}) \\ &= \frac{k}{2} \mathbb{E}_{z \sim \text{Unif}\{1,2,\dots,k\}} [D_{\text{KL}}(\pi_{M,i} \parallel \pi_{B,i})] = \frac{k}{2} \mathcal{I}(\tau; z \mid x). \end{aligned}$$

Combining this with equation 11 yields

$$1 - \text{pass@k}_M \geq (1 - \text{pass@k}_B) \exp\left(-\frac{k}{4(1-u)^2} \mathcal{I}(\tau; z \mid x)\right). \quad (12)$$

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 equation 12 to bound the improvement $\Delta := \text{pass@k}_M - \text{pass@k}_B$, we obtain

$$\begin{aligned} \Delta &= (1 - \text{pass@k}_B) - (1 - \text{pass@k}_M) \\ &\leq (1 - \text{pass@k}_B) \left(1 - \exp\left(-\frac{k}{4(1-u)^2} \mathcal{I}(\tau; z \mid x)\right)\right) \\ &\leq (1 - \text{pass@k}_B) \cdot \frac{k}{4(1-u)^2} \cdot \mathcal{I}(\tau; z \mid x), \end{aligned}$$

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

D SEMANTIC MUTUAL INFORMATION REWARD

One observation that we made empirically is that token-level differences tend to reflect formatting or paraphrasing, rather than semantically distinct strategies. To bias toward more meaningful differences, we test an alternative method for measuring mutual information by embedding completions with a fixed encoder $\psi(\tau) \in \mathbb{R}^d$ and estimating the mutual information between embeddings and skills for a *single prompt* x :

$$\hat{\mathcal{I}}(\psi(\tau); z \mid x) \quad (13)$$

using the KSG k -nearest-neighbor estimator (Kraskov et al., 2004), implemented with the library NPEET (Steeg, 2025). Concretely, for each x we collect the set of embeddings across strategies and samples, $\mathcal{B}(x) = \{(\psi(\tau_i^{(z)}), z) : z \in \{1, \dots, N\}, i = 1, \dots, C\}$, and apply KSG to $\mathcal{B}(x)$ to obtain a single scalar $r_{\text{SMI}}(x)$.

1026 E ARITHMETIC ENVIRONMENT

1027

1028 E.1 TASK

1029

1030 Each problem instance consists of three integers $a, b, c \in \{0, \dots, 9\}$. A small transformer model
 1031 chooses one of the integers to be a target and is required to produce a simple arithmetic expression
 1032 with the other two digits that evaluates to this target. Valid operators are $\{+, -, \times, \div, \text{mod}\}$, with
 1033 division and modulo defined only when results are integers and denominators are nonzero. A latent
 1034 skill index $z \in \{0, \dots, N-1\}$ is provided as part of the prompt, conditioning the model on which
 1035 strategy to adopt.

1036

1037 E.2 PROMPT FORMAT AND CONDITIONING

1038

1039 The input is formatted as

1040

$$[z] \ a \ b \ c$$

1041

1042 where $[z]$ encodes the latent skill id and a, b, c are the three digits. The model is required to
 1043 generate exactly three tokens in the order (digit, operator, digit). This restriction
 1044 enforces that every completion corresponds to a candidate arithmetic expression of the form $L \circ R$
 1045 with $L, R \in \{a, b, c\}$. The verifier deterministically evaluates the completion, awarding a reward of 1
 1046 if the output matches the designated target and 0 otherwise.

1047

1048 E.3 EVALUATION PROTOCOL

1049

1050 At inference, we fix $k = \min(N, 5)$ distinct latent skills and generate one completion per skill at a
 1051 fixed temperature. We then report:

1052

- 1053 • `pass@1`: the fraction of skills (out of k) that yield a correct completion, i.e. the probability that
 1054 a single uniformly sampled skill would succeed.
- 1055 • `pass@k`: the probability that at least one of the k skills yields a correct completion.

1056

1057 This definition differs from conventional `pass@1` (best-of- k) to more closely capture the multi-skill
 1058 sampling process we target.

1059

1060 E.4 MODEL AND OPTIMIZATION

1061

1062 The policy is a 2-layer causal Transformer (hidden size 128, 4 attention heads, pre-layer normalization,
 1063 GELU activations). Inputs are embedded with learned token and positional embeddings. The output
 1064 vocabulary consists of 15 symbols: 0-9, +, -, *, /, %. Training uses GRPO updates without a KL
 1065 penalty, comparing runs with and without the MISL reward. Teacher-forced cross-entropy warmup is
 1066 applied for 100 steps before switching to RL. Unless otherwise noted: $N=5$, temperature 0.9, batch
 1067 size 32 groups, and $C=5$ completions per update.

1068

1069 E.5 TRAINING OUTCOMES

1070

1071 Table 1 reports `pass@1`, `pass@5`, and marginal token entropy H_m at the end of the supervised
 1072 warmup (step 0), mid-training (step 1000), and the final step (step 2000). All runs use $N = 5$ skills
 1073 and 2000 training steps.

1074

1075 E.6 SENSITIVITY TO α_1 AND `CAP`

1076

1077 Increasing the clipping parameter `cap` sustains higher entropy but also introduces more variance
 1078 across runs. Raising α_1 strengthens specialization and increases `pass@5`, though at the expense of
 1079 `pass@1`. A moderate setting of $\alpha_1=0.5$ and `cap= 1.0` provided the most consistent balance in our
 experiments.

Condition	After Warmup (Step 0)			Step 1000			Step 2000		
	p@1	p@5	H_m	p@1	p@5	H_m	p@1	p@5	H_m
Control - ($\alpha_1 = 0$)	0.313	0.665	0.723	0.668	0.673	0.016	0.793	0.793	0.013
($\alpha_1 = 0.5$, cap=1.0)	0.313	0.665	0.723	0.353	0.843	0.755	0.390	0.897	0.768
($\alpha_1 = 0.5$, cap=1.5)	0.313	0.665	0.723	0.338	0.833	0.764	0.399	0.830	0.852
($\alpha_1 = 1.0$, cap=1.0)	0.313	0.665	0.723	0.281	0.813	0.813	0.373	0.897	0.844

Table 1: Arithmetic environment training outcomes. Accuracy is reported as pass@1 and pass@5; H_m is marginal token entropy. MISL prevents entropy collapse and sustains diverse skill-conditioned strategies.

E.7 ADDITIONAL DISTRIBUTION DATA

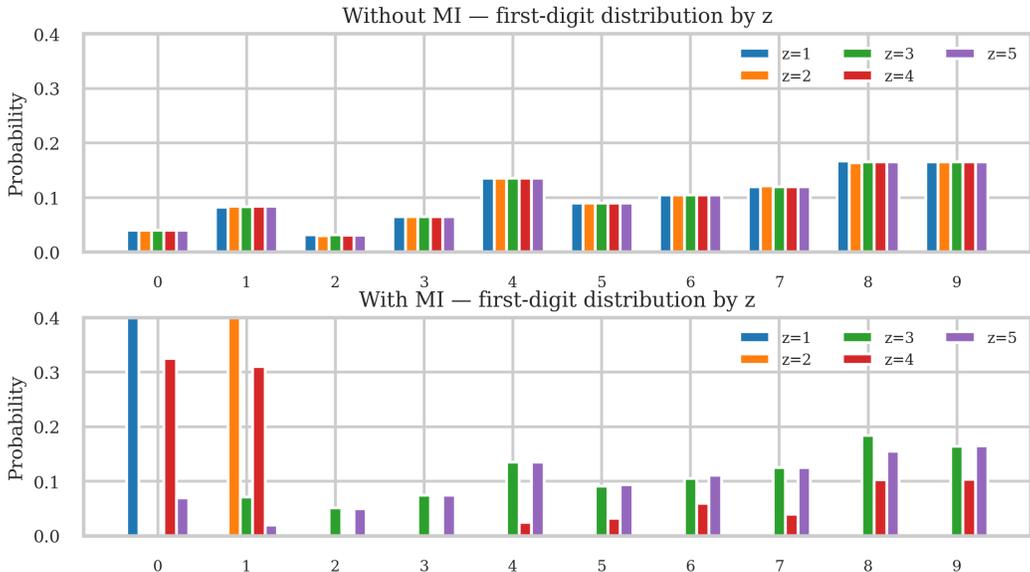


Figure 6: Learned distribution over first response digit with $\alpha_1 = 0.5$ and cap = 1.0

F ABLATED ARITHMETIC ENVIRONMENT

Here we ablate *model capacity* by varying the number of warmup steps and by adding a KL penalty. Warmup 50 corresponds to a weaker base model, while warmup 100 produces a stronger base model. This manipulation allows us to study how UpSkill interacts with correctness-oriented pretraining and how much headroom remains for diversity improvements. We also include a KL penalty with coefficients $kl_coef \in \{0.05, 0.10\}$, completions per group $C \in \{5, 10\}$, and MIs weights $\alpha_1 \in \{0.0, 0.1, 0.3, 0.5\}$ (token MI only).

Warmup Steps	pass@1	pass@k
50	0.235	0.540
100	0.313	0.665

Table 2: Performance of the base model after warmup only (no RL). Warmup 50 yields a weaker base capacity, while warmup 100 yields a stronger base capacity.

Tables 3 and 4 report $\text{pass}@1$ and $\text{pass}@5$ after RL across settings. Columns $\alpha_1.X.Y$ denote $\alpha_1 = X.Y$.

kl_coef	warmup	C	pass@1 _{a0.0}	pass@1 _{a0.1}	pass@1 _{a0.3}	pass@1 _{a0.5}
0.050	50	5	0.779	0.780	0.561	0.457
0.050	50	10	0.845	0.817	0.655	0.439
0.100	50	5	0.813	0.801	0.557	0.366
0.100	50	10	0.833	0.860	0.629	0.437
0.050	100	5	0.908	0.897	0.749	0.521
0.050	100	10	0.914	0.897	0.767	0.521
0.100	100	5	0.911	0.895	0.788	0.568
0.100	100	10	0.917	0.901	0.844	0.614

Table 3: $\text{pass}@1$ after RL with KL penalty in the arithmetic environment.

kl_coef	warmup	C	pass@5 _{a0.0}	pass@5 _{a0.1}	pass@5 _{a0.3}	pass@5 _{a0.5}
0.050	50	5	0.793	0.807	0.840	0.870
0.050	50	10	0.867	0.833	0.867	0.817
0.100	50	5	0.840	0.843	0.840	0.847
0.100	50	10	0.857	0.893	0.850	0.820
0.050	100	5	0.940	0.917	0.863	0.883
0.050	100	10	0.927	0.903	0.903	0.887
0.100	100	5	0.927	0.907	0.910	0.917
0.100	100	10	0.944	0.931	0.940	0.923

Table 4: $\text{pass}@5$ after RL with KL penalty in the arithmetic environment.

Overall, a modest KL penalty (0.05–0.10) prevents entropy collapse and supports higher multi-attempt accuracy, especially when the warmup baseline is stronger (100 vs. 50). With warmup 50, larger α_1 increases $\text{pass}@5$ substantially (e.g., 0.793 \rightarrow 0.870), though often at the expense of $\text{pass}@1$. With warmup 100, the base capacity is already high, and further MISL gains are more limited, reflecting our theoretical expectation that improvements in $\text{pass}@k$ depend on the available headroom for diversity.

G GSM8K EXPERIMENTAL DETAILS

G.1 SETUP

We evaluate our method on GSM8K (Cobbe et al., 2021), a grade-school arithmetic dataset with 2,000 training and 500 held-out evaluation problems. Prompts are provided in a zero-shot format with a maximum sequence length of 1024 tokens. For inference, we fix k distinct latent skill identifiers and sample one completion per skill at fixed temperature.

G.2 MODELS

We attach LoRA adapters (about 80M parameters) to three open-weight backbones: Llama 3.1–8B, Qwen 2.5–7B, and R1-Distilled–Qwen2.5–Math–1.5B. Training uses GRPO with correctness reward only (control) or correctness and token-level MISL (experimental) with $\alpha_1 = 5.0$ and a learning rate of 5×10^{-6} . Each experiment is run for 2000 steps on a single H100 GPU with 80 GB of memory. For Llama, we adjusted to a learning rate of 1×10^{-6} , and α_1 to 1. Results with and without these changes are included in Appendix H.

G.3 EVALUATION DETAILS

We train and evaluate the model with the prompting format of: "Strategy [z] | Question" when training with the token MI reward and with only the question otherwise (base model and without

MI models). We ablate the prompting format for evaluation in Appendix I. During training and evaluation, we fix the system prompt as “You are a helpful math assistant that solves problems step by step.”

In inference, we fix N distinct latent skills and generate one completion per skill, with $N = 5$ except in ablations. To determine correctness, we use a robust extraction function that searches for values within common tags (such as “<answer> </answer>”), defaulting to the final word containing digits in the model’s output, removes characters not in 0123456789. -, and compares the resulting value against the reference answer. We then report:

- `pass@1`: the fraction of skills (out of N) that yield a correct completion; i.e., the probability that a single uniformly sampled skill would succeed
- `pass@k`: the probability that at least one of the k skills yields a correct completion
- `plurality@k`: the probability that there is a unique mode response, and that it is correct
- `consensus@k`: the probability that a strict majority of the completions are correct.

H TRAINING HYPERPARAMETER CHOICE

For the Llama model, we noticed the learning rate of 5×10^{-6} and $\alpha_1 = 5$ led to training instability, and a model that, at intermediate points in training, substantially sacrificed correctness to maximize the token-level mutual information reward. Thus, we changed the learning rate for the Llama model to 1×10^{-6} and the reward coefficient to $\alpha_1 = 1$. The effects of the change on Llama is shown in Figure 7.

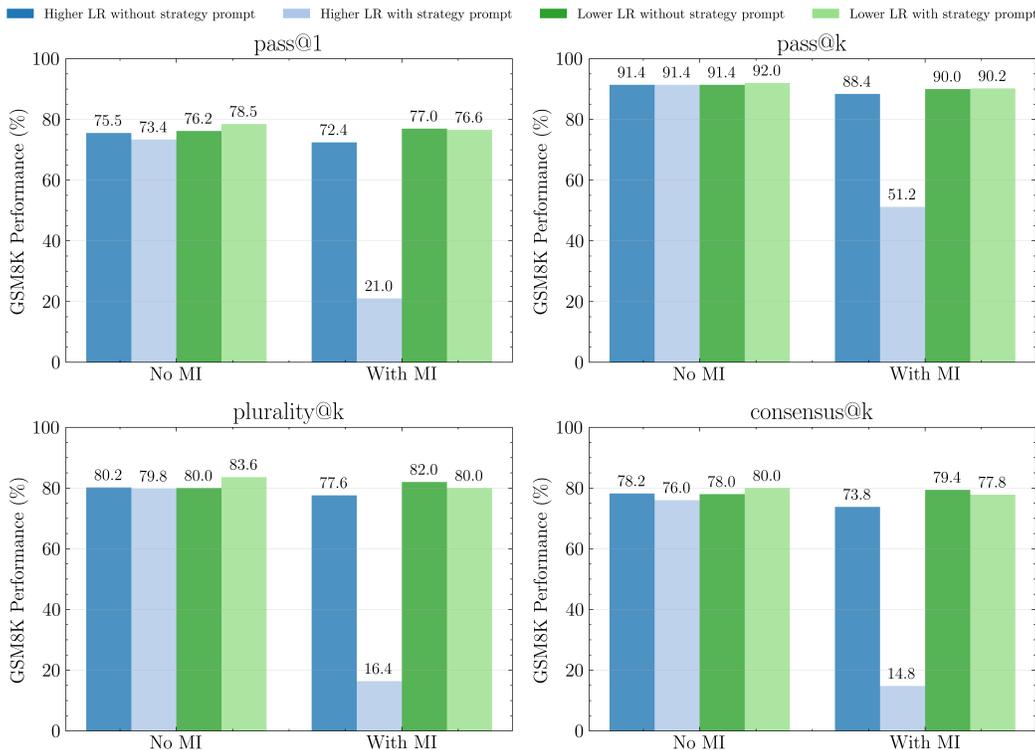


Figure 7: Llama results for learning rate of 1×10^{-6} and $\alpha_1 = 1$ (green bars) compared to 5×10^{-6} and $\alpha_1 = 5$ (blue bars). We simultaneously ablate the result of the strategy prompting, as we do in Figure 8. In each pair of 4 pairs, the leftmost represents the 5×10^{-6} without the strategy prompt, the second furthest to the left represents 5×10^{-6} with the strategy prompt, followed by the models trained with learning rate 1×10^{-6} and evaluated without and with the strategy prompt, respectively.

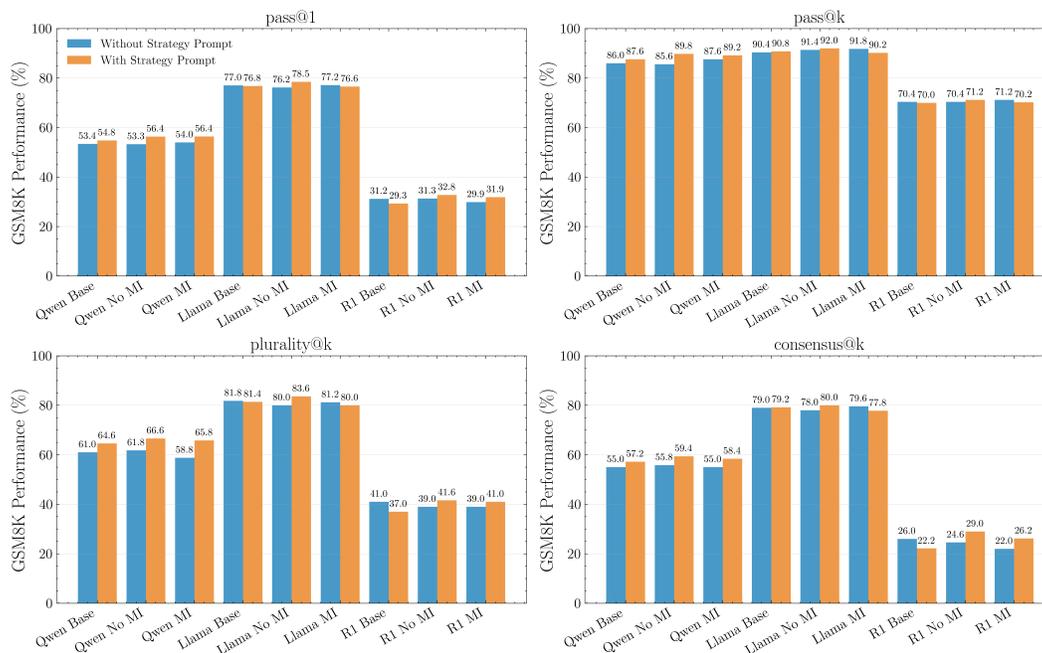


Figure 8: Comparison of open-source models under different prompts during evaluation. Models were trained With MI or Without MI.

I ABLATION ON PROMPTING FOR EVALUATIONS

We ablate the impact of prompting in the format of: "Strategy [z] | Question" at the evaluation stage in Figure 8. We test prompting with "Strategy [z] | Question" compared to "Question" at the evaluation stage to determine if our difference in prompting leads to an advantage. We generally find that some models, including base models and models trained without strategy prompting (Without MI models), still often prefer the strategy prompting, likely as it introduces diversity into the prompt.

J ABLATION STUDIES

J.1 SCALING THE NUMBER OF STRATEGIES.

We investigated the effect of increasing the number of latent skills N beyond the default $N=5$. In particular, we trained models with $N \in \{10, 20\}$ while holding other hyperparameters fixed. The gains were mixed: although we continued to see improvements in `consensus@k` relative to baselines without MISL, the magnitude of these gains was reduced compared to the $N=5$ case, and we did not see a gain in `pass@k` or `plurality@k`. Many GSM8K problems admit only one or two broad solution approaches, so forcing the model to partition its capacity into ten or twenty strategies may lead to fragmentation into modes that were either redundant or unhelpful.

J.2 SEMANTIC MI IS PROMISING BUT UNSTABLE.

We also evaluated the addition of a semantic mutual information reward, using the KSG estimator of $\mathcal{I}(\psi(\tau); z | x)$ in the embedding space of τ . In principle, this should encourage the learned strategies to differ not only at the token level but also in their higher-level semantics. In practice, however, training with a nonzero α_2 proved challenging. Optimizing against the KSG estimator often destabilized learning, likely due to estimator variance in high dimensions. KSG requires multiple points per class to be well-behaved. In our experiments (Section 5), 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

1296 disable the semantic term to avoid excessive compute and estimator variance. We include this test in
1297 our ablation.

1298 On R1-Distilled-Qwen, combining semantic MI with token MI occasionally yielded additional
1299 improvements, but these were not consistent and were offset by instability in other runs. On Qwen2.5-
1300 7B, semantic MI frequently caused degeneracy, with certain strategies collapsing to incoherent or
1301 repetitive outputs. These results suggest that semantic MI is promising but fragile. The occasional
1302 positive signal points to the potential of semantic surrogates for enforcing genuine strategy diversity,
1303 but the current estimator is too noisy to be reliable in practice. We conjecture that improved semantic
1304 objectives could offer a more stable alternative. For the purposes of this work, we therefore report
1305 token-MI only as the main result, with semantic MI left for future exploration.

1306

1307 K FULL EXAMPLE OUTPUTS

1308

1309 The following contains the full output of the summarized items in [Figure 2](#):
1310

1311 **Strategy 1:** Problem: If Raymond does half as much laundry as Sarah, and Sarah does 4 times as
1312 much laundry as David, calculate the difference in the amount of laundry Raymond and David do if
1313 Sarah does 400 pounds of laundry.

1314

1315 **SOLUTION:** To find the solution, let's break it down step by step:
1316

- 1317 1. Sarah does 400 pounds of laundry.
- 1318 2. David does 1/4 of the laundry Sarah does because Sarah does 4 times as much laundry as
1319 David. So David does $400 / 4 = 100$ pounds of laundry.
- 1320 3. Raymond does half as much laundry as Sarah does, so Raymond does $400 / 2 = 200$ pounds
1321 of laundry.
- 1322 4. Now, we need to calculate the difference between the amount of laundry Raymond does and
1323 David does.
1324 The difference is $200 - 100 = 100$ pounds

1325

1326 The difference in the amount of laundry Raymond and David do is 100 pounds.
1327

1328 **Strategy 3:** Problem: If Raymond does half as much laundry as Sarah, and Sarah does 4 times as
1329 much laundry as David, calculate the difference in the amount of laundry Raymond and David do if
1330 Sarah does 400 pounds of laundry.

1331

1332 **SOLUTION:** To solve this problem, let's break it down step by step.
1333

1334 Given information:

1335

- 1336 • Raymond does half as much laundry as Sarah.
- 1337 • Sarah does 4 times as much laundry as David.
- 1338 • Sarah does 400 pounds of laundry.

1339

1340 Step 1: Determine the amount of laundry David does. Since Sarah does 4 times as much laundry as
1341 David, and Sarah does 400 pounds of laundry, we can set up the following equation to find out how
1342 much David does:

1343 Let D be the amount of laundry David does. $4D = 400$

1344

1345 To find D, we divide both sides by 4.

1346 $D = 400 / 4$

1347 $D = 100$

1348 Step 2: Determine the amount of laundry Raymond does.

1349 Since Raymond does half as much laundry as Sarah, and Sarah does 400 pounds of laundry, we can
find out how much Raymond does:

1350 Let R be the amount of laundry Raymond does.
1351 $R = 400 / 2$
1352 $R = 200$
1353
1354 Step 3: Calculate the difference between the amount of laundry Raymond and David do.
1355 Difference = Amount of laundry Raymond does - Amount of laundry David does
1356 Difference = $200 - 100$
1357 Difference = 100
1358 Therefore, the difference in the amount of laundry Raymond and David do is 100 pounds.
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403