RLVR LANDSCAPES FOR ITERATED MULTIPLICATIONS CAN BE BENIGN: INSIGHTS FROM SPIN-GLASS THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the importance of reinforcement learning with verifiable rewards (RLVR), the extent to which it can learn new reasoning capabilities remains debated. Here we study the optimization landscape of RLVR on algorithmic tasks, such as iterated group and quasigroup multiplication. To this end, we map entropy-regularized RLVR over myopic tabular policies onto an energy-based (spin-glass) model over deterministic policies. This mapping upper-bounds what RLVR can achieve, and lets us rigorously characterize the landscape in this tabular setting. We show, both theoretically and experimentally, that for a wide class of models and tasks with uncorrelated inputs, this landscape is benign, containing no local minima that could trap RLVR training. Rather, the practical difficulty of these tasks appears to stem, at least in part, from issues such as diffusive barriers and gradient-estimation error in traversing the landscape. These are genuine obstacles that can prevent a solution from being found, but they are distinct from the landscape itself being rugged. We show that these obstacles can often be mitigated through the choice of entropy regulator. Consistent with this theory, we find that a transformer trained from scratch, using only last-token rewards, successfully learns an algorithmic chain of thought for iterated non-Abelian group multiplications.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR), policy-gradient training against an automatically checkable success signal, is the main tool behind recent improvements in the reasoning of large language models (DeepSeek-AI, 2025; Yu et al., 2025). Whether it produces genuinely new reasoning, or mainly reweights behavior already present in the pretrained model, is debated (Yue et al., 2025; Wu et al., 2026; Wen et al., 2026; Tsilivis et al., 2025). Behind this debate sit open optimisation questions about RLVR itself. Due to the credit assignment problem, intermediate tokens receive little to no direct signal, and sampled gradients are noisy (Yu et al., 2025; Tao et al., 2025; Han & Yang, 2026; Li et al., 2026). Such hurdles can be, in principle, addressed by larger batches, dynamic sampling, and difficulty curricula (Yu et al., 2025; Jiang et al., 2025), but these are not generic cures. Curricula, for instance, are subject to primacy bias and shortcut learning (Nikishin et al., 2022; Liu et al., 2023). Indeed, practitioners routinely describe runs as "trapped in a local optimum" (Wang et al., 2025b; Jin et al., 2026).

This brings to question the complexity and ruggedness of RLVR landscapes. Such analyses of reinforcement learning exist for specialized settings (see related literature), which show that optimizing policies with bounded memory is NP-hard in the worst case (Littman, 1994; Mundhenk et al., 2000; Vlassis et al., 2012). However, worst-case statements say little about typical instances. For random constraint satisfaction problems, analysis of the typical case was pushed forward by applying statistical mechanics approaches based on spin-glass theory. For instance, spin-glass theory traced typical-case algorithmic hardness to the geometry of the solution landscape, which above certain thresholds breaks into many separated clusters and local minima (Mézard et al., 2002; Krzakala et al., 2007; Zdeborová & Krzakala, 2016), inspiring rigorous mathematical approaches (Gamarnik, 2021). Building a similar bridge between spin-glass theory and RLVR is thus a desirable goal.

As a toy setting to approach this goal, we focus on a simple class of algorithmic tasks consisting of iterated quasigroup and group multiplication (Liu et al., 2023; Merrill et al., 2024; Li et al., 2024; Grazzi et al., 2025; Marchetti et al., 2026; Forner et al., 2026). This choice is partially motivated by a

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

simplistic reading of what pretraining provides: a set of skills (Michaud et al., 2024) the model can execute individually, together with the problem of composing them in the right order (Abdulsalam et al., 2026). Taking the skills to be the elements of a finite group or quasigroup and composition to be its operation, "learning to reason" becomes evaluating a product of T elements presented in sequence, with one intermediate token emitted per element and reward only if the final token is the correct product.

As model classes, we use myopic tabular policies (myopic in the sense of DeWeese & Qu (2026)) that read a bounded window of the input together with their own recent tokens. While such myopic bias is obviously a crude approximation to that of a transformer, the workhorse of RLVR applications in reasoning, it has several merits: It fleshes out culprits in the RLVR loss itself rather than in the policy parametrization. It exhibits some of the qualitative features seen in transformer traps for our tasks, while permitting essentially exact numerics.

In this work, we provide a novel spin-glass perspective on RLVR, which allows an in-depth analysis of the above task with some conclusions carrying over to transformers. In particular, we establish the following results

A mapping between RLVR and spin models (Sec. 3). We build a dictionary between RLVR and spin-glass models, yielding three results: every local maximum of the reward is continuously connected, without crossing a loss barrier, to a deterministic policy, referred to as a *code* or *spin-configuration* (§3.1); with an entropy regulator, the RLVR loss is provably lower-bounded by the free energy of a corresponding spin-glass energy-based model (§3.2); and steady states of local dynamics on the spin-glass model's trapped configurations approximately coincide with RLVR fixed points (§3.3).

Spin-glass model for parity task is solvable (Sec. 4.2). We obtain closed expressions for the free energy of the spin-glass model associated with parity at any temperature and depth (T) and find a first-order phase transition at large T, between strictly random and correct codes.

Reward landscapes can be benign (Sec. 4.1). We leverage the spin-glass mapping to prove that for simple untied models the reward has no suboptimal local maxima. We show that a simple Monte Carlo sampler of the spin-glass models solves all untied models and tasks without a curriculum. With weight tying and input correlations, the proof breaks down, and indeed we find local maxima.

The RLVR–spin-glass gap and regulator choices (Sec. 5). We show that RLVR optimization (via exact local policy updates) fails to optimize untied models, while as aforementioned, Monte-Carlo on spin-glass models succeeds. This gap is due to the former's inability to diffuse over equal reward code landscapes. We show that this gap can be closed with local policy resets to random deterministic policies or strong regulators.

Strongly regulated transformers solve the task (Sec. 5.1). We show numerically that transformers can easily express and interpolate between untied tabular policies with narrow windows and hence share some of their benign landscape features. We further show transformer trained from scratch on the same task fail when left weakly or moderately regulated, but often yield perfect learning of the task when strongly regulated and augmented by a curriculum to circumvent signal-to-noise issues.

Additional related literature. Landscape analyses of reinforcement learning exist for specialized settings: gradient-domination results for complete parameterisations (Agarwal et al., 2021; Bhandari & Russo, 2019), the linear-quadratic regulator (Fazel et al., 2018), convergence for softmax tabular policies (Mei et al., 2020), and a study of what entropy regularisation does to small exact-gradient problems (Ahmed et al., 2019). Our work is close in spirit but distinct from the maximum-entropy formulation of control, in which an entropy-regularised objective is read as inference over a Boltzmann measure on *trajectories* (Ziebart et al., 2008; Levine, 2018; Haarnoja et al., 2018): there the sites are the time steps of a rollout, whereas our ensemble is over policies and the sites are table cells. On the physics side, a glassy phase was demonstrated for an optimal-control landscape (Day et al., 2019), spin-glass language has been applied by analogy to supervised loss surfaces (Choromanska et al., 2015), and reinforcement learning has been used as a solver for spin-glass ground states (Fan et al., 2023). The trapped configurations we certify have antecedents as partial pooling equilibria in signalling games (Lewis, 1969; Crawford & Sobel, 1982; Huttegger et al., 2010; Skyrms, 2010), as miscoordination of independent learners in common-payoff games (Panait & Luke, 2005; Matignon et al., 2012), and as myopic optima of restricted policy classes (DeWeese & Qu, 2026). Entropy

collapse in RLVR and entropy-management schemes are active empirical topics (Cui et al., 2025; Wang et al., 2025a; Yang et al., 2026); the diagnostics here differ in being derived from measured landscape structure rather than fitted to online statistics. The stalls of plain RLVR and their repair by re-randomisation are the multi-step counterparts of partial pooling and its destabilisation by mutation (Zhang, 2004) and are kept from premature fixation by margins acting as mutation.

2 Setting: Task and Models

Let $\mathcal{A}$ be a finite alphabet of size q carrying a binary operation $B : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ whose multiplication table is a Latin square, i.e. every row and every column is a permutation of $\mathcal{A}$. Such a structure is a *quasigroup*; a group is the special case in which B is also associative. The input is a sequence $x_1 \dots x_T$ drawn uniformly from $\mathcal{A}^T$ (in some expereimental setups we break this assumption and draw correlated inputs), and the target is the left fold

$$s_1 = x_1, \qquad s_t = B[s_{t-1}, x_t], \qquad \text{target} = s_T. \tag{1}$$

We refer to s_t as the *state* of the task at step t. To solve this task, we consider models that sequentially generate an output sequence $y = (y_1, \dots, y_T)$ conditioned on the input x. At each token index t, the next token y_t is drawn from a distribution, or policy, π that can depend on $x, y_{<t}, t$. An optimal policy is one that perfectly outputs the target, $y_T = s_T$, regardless of intermediate tokens. In this section, we define two model classes for this task: tabular policies trained with RLVR (Section 2.1), and an energy-based model for the distribution of deterministic policies (Section 2.2).

2.1 RLVR Optimized Tabular Policies

In this class of models, the policy at each timestep relies on a localized context variable

$$c_t(x, y_{<t}) = (x_{t-n_p} \dots x_{t+n_f}, y_{t-n_y} \dots y_{t-1}, t) \tag{2}$$

which encapsulates a sliding window of the input, a bounded window of previously emitted tokens, and the current time step (ie token index) t. For the boundary cases, we define for every offset $k \in \{-n_p, \dots, n_f\}$, $x_{t+k} =$ SEP whenever $t + k < 1$ or $t + k > T$, and for every $k \in \{1, \dots, n_y\}$, $y_{t-k} =$ SEP whenever $t - k < 1$. For the boundary cases, we define $x_{t'} =$ SEP when $t' < 1$ or $t' > T$, and similarly $y_{t'} =$ SEP for $t' < 1$. In the following, we take $n_y = 1$, and denote the window of each model by (n_p, n_f). We refer to each unique configuration of this context as a cell. For each cell c, the policy assigns a discrete probability distribution $\pi(y_t \mid c)$ over the q possible vocabulary tokens. Consequently, the space of all possible policies is simply a product of probability simplices, one per cell. We refer to a policy as tied if the policy is not an explicit function of the token index, so that for all t, t', if $c_t, c_{t'}$ differ only in the token index, then $\pi_{c_t} = \pi_{c_{t'}}$. Note that this implies that untied policies visit each at most once. This policy is optimized with RLVR, where the model receives reward $= 1$ if y_T equals s_T and 0 otherwise. Crucially, intermediate tokens are not supervised. We denote by $J(\pi)$ the expected reward over the policy's outputs and the input data. This policy is trained by minimizing the following objective

$$\mathcal{L}(\pi) = -J(\pi) - \beta^{-1} \sum_c H(\pi_c), \tag{3}$$

where the second term is an entropy regularizer and we refer to the coefficient $\tau = \beta^{-1}$ as the temperature. The loss itself is optimized via coordinate-ascent variational inference (CAVI) (Blei et al., 2017), where we optimize each cell c while fixing the other cells. In this model, we can explicitly define for untied policies the optimal policy for each cell conditioned on all the other ones, which is given by

$$\pi_c(a) \propto \exp\big(d(c)\, Q(c, a)/\tau\big), \tag{4}$$

where $d(c)$ is the probability that rollouts consult cell c, and $Q(c, a)$ is the expected reward, with respect to the policies of all other cells, conditioned on action a being taken at cell c. See App. A for the CAVI update rules on tied policies. A forward recursion over all input combinations gives visitation probabilities $d(c)$, a backward recursion gives the action values $Q(c, a)$, analogous to the algorithm described in (Rabiner, 1989). Following the policy-gradient theorem (Sutton et al., 2000), we then obtain J and its exact gradient in closed form, with essentially zero estimation error. For full details, see App. A.

2.2 Statistical Spin-Glass Models of Deterministic Policies

We now turn to an a priori different kind of inference model. We study an energy model over policies, plausibly the simplest one that captures the spirit of what RLVR training favors- determinism and high reward. Thus, this distribution is supported on deterministic policies, each of which assigns a single action to each cell. We call such a function a *code*, and write $\sigma : \mathcal{C} \to \mathcal{A}$, with σ_c denoting the token assigned to cell c (see Figure 1II for a visualization). To favor high reward, we consider a Gibbs distribution with energy (cost) given by $E(\sigma) = -J(\sigma)$, assigning each code probability proportional to $e^{-\beta E(\sigma)}$ at inverse temperature β,

$$Z(\beta) = \sum_{\sigma \in \mathcal{A}^{\mathcal{C}}} e^{-\beta E(\sigma)}, \qquad p_{\text{spin}}(\sigma) = e^{-\beta E(\sigma)}/Z(\beta). \tag{5}$$

In physics terms, this is a system of $N = |\mathcal{C}|$ q-state (Potts) spins with Boltzmann distribution $p(\sigma) \propto e^{-\beta E(\sigma)}$. Here, $J(\sigma)$ is the same expected reward as defined in 2.1, evaluated at the deterministic policy σ, which is explicitly given by

$$J(\sigma) \;=\; \mathbb{E}_x\Big[\sum_{y \in \mathcal{A}^T} \delta_{y_T,\, s_T(x)} \prod_{t=1}^{T} \delta_{\sigma_{c_t(x,\, y_{<t})},\, y_t}\Big], \tag{6}$$

Note that this is a non-local, distributed interaction reminiscent of, but distinct from, a canonical spin-glass system. Both share a discrete phase space with non-local interactions; the difference is that our model's interactions are complex (quasirandom) and contain a gauge symmetry rather than genuinely random, as in the spin-glass setting. See Fig. (1I) for a summary of the RL/physics terminology correspondence, and App. D for a more detailed comparison to classic spin models.

This model is part of a well-studied class of physical systems, bringing with it a rich theoretical toolbox for analyzing such distributions (Mézard & Montanari, 2009). In some cases, the model can be solved analytically (see Sec. 4.2), and more generally, since codes take values in a finite space which can in principle be enumerated over to compute various aspects of the system. Alongside these theoretical tools, Monte Carlo (MC) methods allow us to sample the distribution directly (Newman & Barkema, 1999).

Notwithstanding, we note that this particular choice of energy function is itself just one choice among many. Moreover, the window and tying restrictions of Section 2.1 are special cases of a more general family. Writing E_{win} and E_{tie} for penalties that respectively discourage codes from treating histories differently within a window, or across tied generation steps, we consider

$$E_{\text{tot}}(\sigma) \;=\; -J(\sigma) \;+\; g_{\text{win}}\, E_{\text{win}}(\sigma) \;+\; g_{\text{tie}}\, E_{\text{tie}}(\sigma), \tag{7}$$

recovering the complete parameterisation at $g_{\text{win}} = g_{\text{tie}} = 0$, and the untied and tied windowed classes as $g_{\text{win}}, g_{\text{tie}} \to \infty$. We note in passing that a transformer's implicit bias or pretraining could similarly be represented as a penalty term of this kind.

3 Mapping the Spin-Glass Model to RLVR

Up to this point, we have made no explicit connection between the RLVR objective and our energy-based spin model. In this section we make this comparison more rigorous. We show that the local maxima of the reward as a function of policies are continuously connected to codes with the same reward; we then use the spin model to obtain a bound on the RLVR loss, which is important for regulator effects at finite temperatures; and we show that the Monte Carlo sampler's traps in the spin model coincide with local minima of RLVR training.

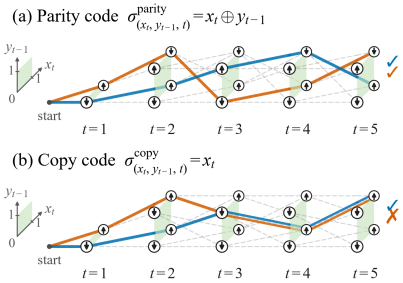
3.1 Spin-Glass Codes as Reward Maximizers

Our first result is given by the following statement

Statement 3.1. *We consider the space of policies as a product of simplices embedded in $\mathbb{R}^{q \times |\mathcal{C}|}$. Let π^* be an untied policy such that there exists an open set around it where π^* obtains a maximal reward, then it can be connected by a continuous path of fixed reward to a deterministic policy σ^* such that $J(\sigma^*) = J(\pi^*)$.*

Reinforcement Learning	Physics term
deterministic policy (code)	spin configuration σ
expected reward J	minus the energy
one table entry (cell)	one q-state (Potts) spin
stochastic policy	variational product state
entropy coefficient $\tau = \beta^{-1}$	temperature
$d(c)\,Q(c,a)$	local field at a site
relabelling of CoT tokens	gauge transformation
set of near-trap configs	pure state

(I) Mapping dictionary



(II) Visualized mapping

Figure 1: The RLVR and spin-glass mapping. (**I**) Relating concepts from reinforcement learning to spin-glass models. (**II**) Two codes on the $(0, 0)$ window (i.e. $c_t = (x_t, y_{t-1})$), for $q = 2$ and $T = 5$. Each circle is a cell (x_t, y_{t-1}, t); the arrow inside it is the code's output at that cell, up for token 1 and down for token 0. Dashed gray edges trace the code's action on all input sequences. Two inputs differing only in x_1 are highlighted: $x = [0, 1, 0, 1, 1]$ and $x = [1, 1, 0, 1, 1]$.Both trajectories of the parity code (a) converge to the correct answer, whereas the copy code (b) succeeds on only one of the two, the same success rate as chance.

We show this for a simplified case here; the full proof follows a similar structure (App. B). Let π^* locally maximize J, and let c be a cell that the policy consults with at which π^*_c is non-deterministic. For simplicity, suppose only two actions a, b are possible at this cell, with $\pi^*_c(a) = s$, $\pi^*_c(b) = 1 - s$. By multilinearity of J, we have, up to constants, $J \propto s\,Q(c, a) + (1 - s)\,Q(c, b)$. Since π^* is a local maximum of the reward, we require $Q(c, a) = Q(c, b)$, so J is constant in s, and we are free to assign any probability we wish to a and b. The edge cases $s \in \{0, 1\}$ are codes. The proof in App. B extends this approach to policies with larger support and more cells.

3.2 Free Energy as an RLVR Loss Bound

A tabular RLVR policy is, by construction, factorized across cells: it assigns an independent distribution π_c to each cell c. Viewing a cell as a spin site, this is exactly the structure of a mean-field ansatz in physics, a distribution that assumes no correlations between sites. This ansatz is typically correct when each spin interacts with many other spins, as is the case for our model. Motivated by this observation, we consider the fully factorized distribution over spins $\hat{\pi} = \bigotimes_c \hat{\pi}_c$ (so that for codes drawn from this distribution, we have $\hat{\pi}_c(a) = \mathbb{E}_{\sigma\sim\hat{\pi}}[\delta_{\sigma_c, a}]$). This variational distribution can be used to bound the free energy of the true distribution, $F_{\text{spin}} = -\beta^{-1} \log Z$ with the Gibbs-Bogoliubov inequality (Mézard & Montanari, 2009):

$$F_{\text{var}}[\hat{\pi}] \;=\; -\tilde{J}(\hat{\pi}) - \beta^{-1} \sum_c H(\hat{\pi}_c) \;\geq\; F_{\text{spin}} \qquad \tilde{J}(\hat{\pi}) = \mathbb{E}_{\sigma\sim\hat{\pi}}\big[J(\sigma)\big]. \tag{8}$$

We note that upon substituting $\tilde{J}$ with J, we obtain that $F_{\text{var}}[\pi] = \mathcal{L}(\pi)$ where π is a tabular policy. However, $\tilde{J}$ is not, in general, the same as the RLVR reward J. In $\tilde{J}$, every cell's action is sampled from $\hat{\pi}$ independently, whereas in J, actions are generated sequentially via a rollout. The two rewards coincide whenever no non-deterministic cell is visited more than once during a rollout. This holds exactly for untied tabular models, and approximately for tied tabular models and near zero temperature, where such revisits are rare and cells are largely deterministic, respectively. See Appendix C for the full derivation. When these rewards coincide, the RLVR loss of any tabular policy π is equal to the variational free energy, so that

$$\mathcal{L}(\pi) = F_{\text{var}}[\pi] \;\geq\; F_{\text{spin}}. \tag{9}$$

Thus, the free energy of the true spin model distribution sets a lower bound, which is nontrivial at finite β, on what any tabular policy trained with RLVR can achieve (i.e. policies that solve Eq. 4).

The tightness of this bound depends on how well $\hat{\pi}$ approximates p_{spin}, namely how good the mean-field approximation is for the configuration encountered in Monte Carlo sampling. For some models, as is demonstrated in Section 4.2, F can be solved explicitly, giving a concrete, quantitative bound

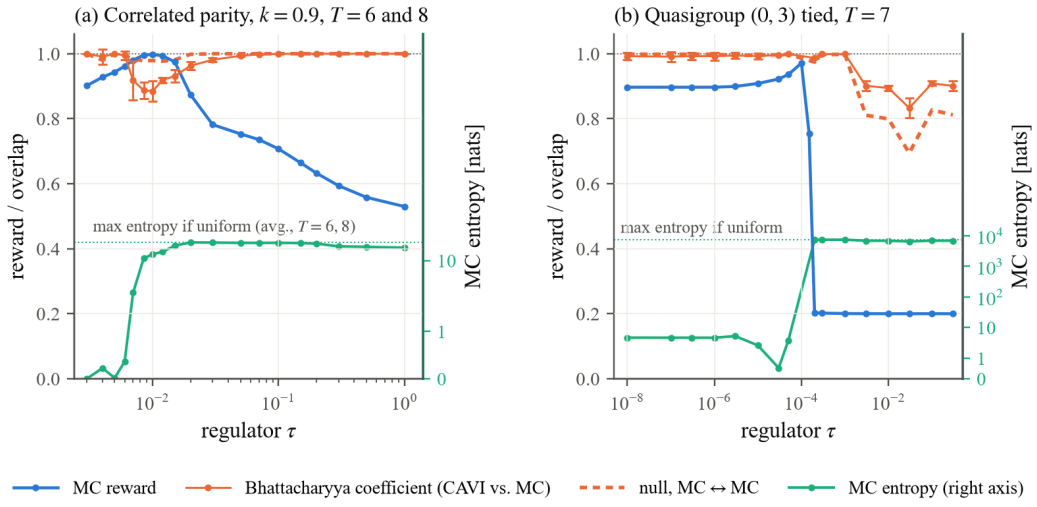


Figure 2: MC stationary distributions against the CAVI endpoints seeded at them, as a function of the regulator τ: (a) the 27 correlated-input parity traps, (b) the twelve tied quasigroup traps. Left axis: MC reward, averaged over all traps and replicas; Bhattacharyya coefficient between the CAVI endpoint and the MC histogram, mean $\pm$ standard deviation over the finite-temperature traps; and its null. Right axis: the entropy $\sum_c H(\pi_c)$ of π^{MC}, averaged over the same traps; dotted lines mark its value when every cell is uniform. Open markers: fewer than half of the traps enter the average at that τ; missing points: none does.

to the RLVR loss. More generally, the discrete nature of the system can often be advantageous, making the configuration space more likely to be enumerable, and so more tractable to solve outright, offering a route to bounding the loss. We comment that multiplying the bound in Eq. (9) by β reveals a machine-learning analogue: maximizing $\beta\tilde{J}(\hat{\pi}) + \sum_c H(\hat{\pi}_c)$ is precisely maximizing the evidence lower bound (ELBO), $\beta\tilde{J}(\hat{\pi}) + \sum_c H(\hat{\pi}_c) \leq \log Z$.

3.3 Spin-Glass Pure States as Training Traps

In spin-glass theory, it is well established that Monte Carlo sampling tends to become trapped in *pure states*: configurations in which σ_c and $\sigma_{c'}$ are approximately uncorrelated for $c \neq c'$ (Mézard & Montanari, 2009). The full distribution over spins can then be understood as a mixture over such pure states. Since these pure states are, by definition, fully factorized across cells, they lie within the same variational family $\hat{\pi} = \bigotimes_c \hat{\pi}_c$ considered above. Among all such factorized distributions, the most likely pure states are precisely those that locally minimize the variational free energy F_{var}. Thus, we expect that when Monte Carlo converges to a steady state, the corresponding distribution would have weakly correlated spins, and its marginals per cell would be approximate solutions to Eq. 4. In Fig. 2 we demonstrate empirically that this is indeed the case for the trapped models that we studied.

4 Some Exact Results on Untied Models

4.1 Untied Classes Hold No Defended Minima

A key empirical result of this paper is that, for a wide class of untied models, namely untied $(0,0),(2,1),(1,1),(3,0)$ window models with uncorrelated inputs, the landscape has no local minima, as evidenced by Monte Carlo optimization at vanishing temperature consistently sampling codes with $J = 1$. See Tab. 2 for full results. Here we support this with the following analytical result

Theorem 4.1 (local search reaches optimal reward)**.** *Consider the untied* $(0,0)$ *window (i.e.* $c_t = (x_t, y_{t-1})$*) class on a quasigroup fold with inputs drawn uniformly, and local dynamics that change one cell at a time and accept any change that does not lower* J*. From every code there is a finite chain of accepted changes ending at a code with* $J = 1$*.*

The proof (Appendix E) proceeds in two parts. First, we show that a *collision-free* code, one in which no two histories in different task states are ever assigned the same output token, reaches $J = 1$ in a finite number of steps that do not decrease the reward. Second, we show how to remove collisions iteratively, without ever decreasing the reward. At a colliding step, every cell is first set to the token that is individually optimal for the true state it leads to. Since there are exactly q possible true states and q possible tokens at this step, any two states left sharing a token after this adjustment leave exactly one token unused. We show how the code can be set to route through the unused token without decreasing the reward. Applying this at each successive colliding step, and then treating the resulting collision-free code by the first part, proves the theorem.

Corollary 4.2 *A non-decreasing reward escape path exists also in policy space when starting from a code.* Proof: Note that due to multilinearity of J, a linear extrapolation between π_c and π'_c, with rewards J and J' respectively, results in a linear interpolation of the reward. Since the theorem relies only on such local policy changes, it implies a piecewise linear extrapolation in policy and reward space.

Corollary 4.3 *A non-decreasing reward escape path exists from any policy.* Proof: If the policy is in a local maximum, it is connected to a code via statement 3.1. If it is not in a maximum, then, by definition, a path with improving rewards exists.

4.2 Exact solution for parity

Section 3.2 showed that the spin model's free energy bounds the RLVR loss, without evaluating that free energy explicitly. This section does so for the simplest member of the task family, the two-element group, $q = 2$, whose fold is the parity of the input bits. This section computes an explicit expression for the partition function for the untied $(0, 0)$ window, and predicts the existence of a first-order phase transition with β.

Following the terminology of reinforcement learning under partial observability (Whitehead & Ballard, 1991; McCallum, 1996), we say that a code *aliases* at step t when two histories whose task states s_t differ produce the same emitted token y_t. What makes this the relevant notion is what the rest of the computation can see: later steps read y_t and the input symbols still inside their window, but not the symbols that have left it, so any distinction absent from y_t is absent for good.

Collect the code's decisions at step t for input symbol x into a $q \times q$ matrix $[\Sigma_{t,x}]_{y_t,\, y_{t-1}} = \delta_{\sigma(t,x,y_{t-1}),\, y_t}$ which we refer to as a *link matrix*. It has exactly one entry equal to one in each column, so that it is a map $\mathcal{A} \to \mathcal{A}$. For the (0,0) window, there are two such matrices per t: $\Sigma_{t,0}$ and $\Sigma_{t,1}$. Call a step t *dead* if both of its link matrices (Σ_{t,x_t}) are constant as a function of x_t, or both are rank-1. Both are forms of aliasing, as they make us lose track of the state ($s_t = B[y_{t-1}, x_t]$) via loss of x_t and y_{t-1} information, respectively. Note that the presence of any dead step forces a pure chance reward ($J = 1/2$) regardless of subsequent steps. We say that the code is *alive* if it has no dead steps. Next, define a *half-reset* as a step where only one of the two matrices has rank 1.

Denote by k the number of half-resets and consider a live code with $k = 0$. By being non-dead and non-half-reset, each step must follow the correct processing of x_t, y_{t-1} or the anti-correct one. This provides 2^{T-1} codes with $J = 0$ and 2^{T-1} codes with $J = 1$. Next, consider a single half-reset ($k = 1$) at $t = t_1$. This means t_1 ignores its CoT token for one of the x_{t_1} values (x_{d_1}). As a result, the reward on inputs for which $x_{t_1} = x_{d_1}$ is $J = 1/2$ regardless of subsequent steps. However, conditioning on $x_{t_1} \neq x_{d_1}$ the half reset is effectively not present, leading, by similar logic to the $k = 0$ case, to 2^{T-2} codes with $J = 1$ and 2^{T-2} codes with $J = 0$ on that branch. Weighing the two equally probable events, $x_{t_1} = x_{d_1}$ and $x_{t_1} \neq x_{d_1}$, we find codes with input-averaged reward $J = 1/4 + 1/2$ and codes with $J = 1/4$.

Generalizing this logic to arbitrary k, taking into account combinatorial factors and dead steps, one obtains the model's normalizing constant — its partition function — in closed form (see App. F),

$$Z(\beta, T) \;=\; e^{\beta/2}\left[2^{4T-2} - 2\cdot 10^{\,T-1} \;+\; 2^{T}\sum_{k=0}^{T-1}\binom{T-1}{k} 4^{k}\cosh\!\left(\frac{\beta}{2^{\,k+1}}\right)\right]. \qquad (10)$$

The average reward is then given by $\mathbb{E}_{\sigma\sim p_{spin}}[J(\sigma)] = \partial_\beta \log(Z(\beta, T))$, in large T limit, it shows a first order phase transition at $\tau_c = 1/(6T\log(2))$, where J goes from 1 to chance $(1/2)$ where the entropy of all the dead codes wins over the reward scale (see App. F.3).

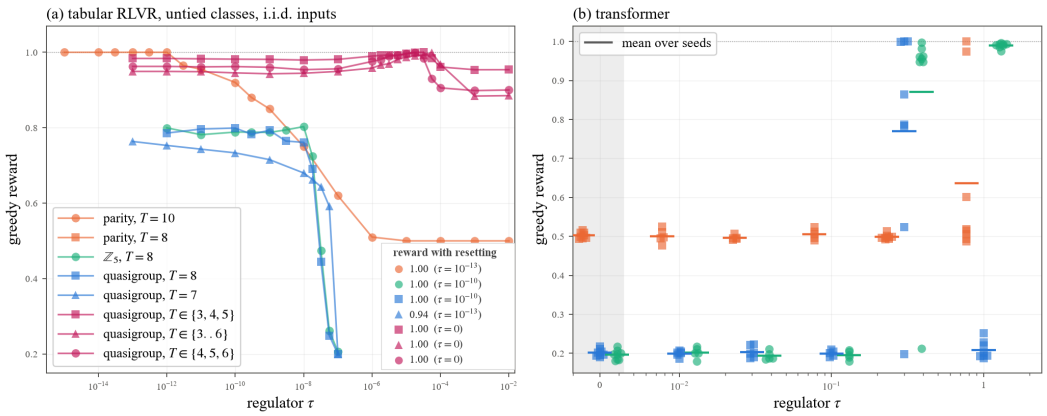


Figure 3: **The entropy regulator on the untied classes. (a)** Tabular RLVR from a uniformly random policy at a constant regulator τ, scored by the greedy reward of the endpoint policy. Each point is a mean over independent runs: 100 for parity (window $(0, 0)$), 240 for the two classes at $T = 8$ using window $(0, 0)$ ($\mathbb{Z}_5$, quasigroup), 12 for the narrow-window classes using window $(2, 1)$ (at $T \in \{3, 4, 5\}$ and $T \in \{4, 5, 6\}$) and window $(1, 1)$ (at $T \in \{3, \ldots, 6\}$), and for quasigroup using window $(0, 3)$ at $T = 7$. The inset table summarises reset interventions of the same class (2% of the cells redrawn uniformly every 10 iterations for 100 cycles): the reward attained and the τ of the run it summarises. **(b)** The one-block transformer under the gated curriculum, REINFORCE with an entropy bonus β (occupancy convention), read as greedy accuracy on held-out inputs of length 8 at the final step. One point is shown per seed, bar is the mean over seeds. The transformer has no direct counterpart of tabular cell resets.

5 Further Experimental Results

The results of previous sections raise several natural questions that are addressable numerically. The first concerns the tightness of our bound and the importance of diffusive-barrier or barren plateaux in the landscape. In Table 2 we measure that difference at negligible temperatures, showing that for many untied tasks RLVR falls much below the spin model which consistently obtaines near perfect reward. The success of the Monte-Carlo sampling suggests that diffusive barriers for the spin model are mild.

This apparent gap motivates us to look for potential improvements on the RLVR side (see App. G.5 for further detail). The success of Monte Carlo samplers motivates the use of periodic resets as a direct means of expediting diffusion. The existence of a phase transition in the parity task motivates us to consider large regulator coefficients (temperatures) to reduce free-energy and entropic barriers. Figure 3, left panel, shows the results of these, with about half the models responding well to high entropy regulator levels. Resets (solid markers) work even better for this class.

From a different angle, the lack of a reward barrier in the landscape motivates us to look for variations of the task that induce such barriers. We consider tied tables or correlated inputs, wherein the inputs are drawn from a Markov chain with a $k > 1/q$ chance of emitting the previous token. Both variations violate different lemmas of our escape proof, and indeed in Tab. 6, we numerically establish the existence of true energy barriers. We note in passing that, for tied models, these barriers are microscopic yet common. Namely, they have a reward barrier equal to failing on a few specific input combinations, but occur for most initialization seeds. For correlated inputs, barriers are macroscopic but rare (i.e. involve a finite fraction of input combinations but only occur for about 12% of initialization seeds).

5.1 Transformers on the same task

We consider tabular policies as a toy model for transformers, which raises a natural question: do the escape paths we establish for tabular policies also exist for transformers trained with RLVR? We would expect an affirmative answer if two things hold: a transformer can express a tabular policy,

and it can traverse the escape route between one code and the next without losing reward. We show below that both do hold. Motivated by these results, we train transformers on the same tasks under sufficient entropy regularization, and show that they often converge to $J = 1$ consistent with a benign landscape (Fig. 3 panel b).

Starting with expressivity, we verify that a modest 791k-parameter transformer can be brought to express random untied tabular policies, with wider windows taking longer to train, suggesting an implicit bias favoring narrow tables (see App. G.4, Table 3).

We further test how well such a transformer follows an escape path of Monte-Carlo. To this end, we locate distinct codes in the chain, train a transformer on one such code (σ_i), save its weights (θ_i), then fine-tune it on the next distinct code ($\sigma_{i+1} \neq \sigma_i$, on visited cells) and obtain the resulting weights (θ_{i+1}). Notably, such fine-tuning never fails to reach perfect agreement: At $T = 6$, all 189 points reproduce their code exactly on all $5^6 = 15{,}625$ inputs. We then consider a linear interpolation between θ_i and θ_{i+1} and test whether this incurs a drop in reward along the path. Interestingly, for the large majority of sequential codes — 159 of 180 steps — no such drop is observed, and for the remaining small fraction the drop in the reward is at most 2×10^{-3}. See Tab. 4 and App. G.4 for further details.

The above results suggest transformers should inherit the simplifying landscape features of untied tabular models. Motivated by this, we study the effects of strong entropy regulators (standard occupancy-weighted regulator) on the same three tasks as in the tabular setting, together with curriculum learning to prevent signal-to-noise issues. As discussed in App. G.5 and shown in Fig. 3, without a regulator, and at every coefficient up to $\beta^{-1} = 0.1$, no seed of any task holds more than length $T = 2$ and every seed is at chance at length $T = 8$. At $\beta^{-1} = 0.3$ five of eight seeds on $\mathbb{Z}_5$ and three of eight on the quasigroup pass 0.95 at length $T = 8$, and they carry the curriculum far beyond it, to $T = 18$ on $\mathbb{Z}_5$ and to length $T = 20$ on three of the quasigroup seeds.

Notably, regulator choices here far exceed standard practice (Ahmed et al., 2019; Cui et al., 2025), by one to two orders of magnitude. Around $\beta^{-1} = 0.3$, the policy carries $42\,\%$ of the maximum entropy ($\ln 5$). We note in passing that before this strong regulator choice, we made many changes to optimization (REINFORCE,GRPO,Expert-Iteration), curricula, adaptive batching, eps-exploration, and combinations thereof, all failed to learn at $T = 4$.

6 Discussion

In this work, we established a mapping between RLVR on tabular models and a spin-glass model whose spins take values in action space and sit at cells/table-entries. Configurations of spins amount to deterministic RL policies. The spin-models appear to enjoy a larger degree of tractability compared to the original RLVR problem, and their free energy lower-bounds the RLVR loss. Gaps between plain RLVR and these lower-bounds motivated us to consider strong regulator effects. Surprisingly, they facilitate RLVR training from scratch for both tabular models and transformers. We further describe several scenarios in which both the spin-model and RLVR end up trapped in valleys of the loss.

We hope to leverage this framework to further study fundamental limitations of RLVR. In particular, it would be interesting to apply it to harder algorithmic tasks which may require invoking the full spin-glass framework, replica-symmetry breaking included. In addition, it would be interesting to incorporate the effects of pre-training and implicit biases of more complex architectures via the inclusion of prior terms.

Limitations. Pure tabular models are a rough caricature of transformers and thus offer only a qualitative link. We further focused on untrained models, based on the concept of skill compositions. Extending this to real pre-trained models carries in much complexity. While prior terms, including KL-penalties, can close some of these gaps qualitatively, their specific relation to RLVR training of networks remains to be studied. The strong regulator choices which facilitated learning here are likely damaging for actual pre-trained models. While some elements of spin-glass theory have been used here, in particular the notion of pure-states and their link to local Monte-Carlo updates, it is yet to be determined whether important concepts such as replica symmetry breaking carry through to these specialized spin models.

REFERENCES

Azwar Abdulsalam, Nishil Patel, and Andrew Saxe. Rl post-training builds compositional reasoning strategies, 2026. URL https://arxiv.org/abs/2607.07646.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. arXiv:1908.00261.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38, pp. 192–204, 2015.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

Ganqu Cui et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Alexandre G. R. Day, Marin Bukov, Phillip Weinberg, Pankaj Mehta, and Dries Sels. Glassy phase of optimal quantum control. *Physical Review Letters*, 122:020601, 2019.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Alex DeWeese and Guannan Qu. Revisiting policy gradients for restricted policy classes: Escaping myopic local optima with k-step policy gradients. *arXiv preprint arXiv:2605.10909*, 2026.

Changjun Fan et al. Searching for spin glass ground states through deep reinforcement learning. *Nature Communications*, 14, 2023.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Niklas Forner, Marcel Kühn, Matthias Thamm, and Bernd Rosenow. Learning dynamics of chain-of-thought state tracking in a solvable transformer model, 2026. URL https://arxiv.org/abs/2606.18164.

David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations (ICLR)*, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Haoyu Han and Heng Yang. Non-uniform noise-to-signal ratio in the reinforce policy-gradient estimator, 2026. URL https://arxiv.org/abs/2602.01460.

Simon M. Huttegger, Brian Skyrms, Rory Smead, and Kevin J. S. Zollman. Evolutionary dynamics of lewis signaling games: Signaling systems vs. partial pooling. *Synthese*, 172(1):177–191, 2010.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. Vcrl: Variance-based curriculum reinforcement learning for large language models, 2025. URL https://arxiv.org/abs/2509.19803.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. Revisiting entropy in reinforcement learning for large reasoning models, 2026. URL https://arxiv.org/abs/2511.05993.

Florent Krzakala, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

Yingru Li, Jiawei Xu, Ziniu Li, Jiacai Liu, Wei Liu, Yuxuan Tong, Longtao Zheng, Zhenghai Xue, Yaxiang Zhang, Tianle Cai, Ge Zhang, Qian Liu, and Baoxiang Wang. The optimal token baseline: Variance reduction for long-horizon llm-rl, 2026. URL https://arxiv.org/abs/2602.07078.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael L. Littman. Memoryless policies: Theoretical limitations and practical results. In *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, 1994.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations (ICLR)*, 2023.

Giovanni Luca Marchetti, Daniel Kunin, Adele Myers, Francisco Acosta, and Nina Miolane. Sequential group composition: A window into the mechanics of deep learning. *arXiv preprint arXiv:2602.03655*, 2026.

Laëtitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Andrew K. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, University of Rochester, 1996.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35492–35506, 2024.

Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling, 2024. URL https://arxiv.org/abs/2303.13506.

Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. Complexity of finite-horizon markov decision process problems. *Journal of the ACM*, 47(4):681–720, 2000.

M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Clarendon Press, Oxford, 1999.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. arXiv:2205.07802.

Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, 2005.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Brian Skyrms. *Signals: Evolution, Learning, and Information*. Oxford University Press, 2010.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, 2000.

Leitian Tao, Ilia Kulikov, Swarnadeep Saha, Tianlu Wang, Jing Xu, Sharon Li, Jason E Weston, and Ping Yu. Hybrid reinforcement: When reward is sparse, it's better to be dense, 2025. URL https://arxiv.org/abs/2510.07242.

Nikolaos Tsilivis, Eran Malach, Karen Ullrich, and Julia Kempe. How reinforcement learning after next-token prediction facilitates learning, 2025. URL https://arxiv.org/abs/2510.11495.

Nikos Vlassis, Michael L. Littman, and David Barber. On the computational complexity of stochastic controller optimization in pomdps. In *ACM Transactions on Computation Theory*, volume 4, 2012.

Shenzhi Wang et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025a.

Zihan Wang et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025b.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jGbRWwIidy.

Steven D. Whitehead and Dana H. Ballard. Learning to perceive and act by trial and error. In *Machine Learning*, volume 7, pp. 45–83, 1991.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin, 2026. URL https://arxiv.org/abs/2507.14843.

Shumeng Yang, Yisu Liu, Jiayi Zheng, Zhaohui Yang, and Linjing Li. Paec: Position-aware entropy calibration for llm reasoning in rlvr. *arXiv preprint arXiv:2606.08543*, 2026.

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yue et al. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Qingfu Zhang. On stability of fixed points of limit models of univariate marginal distribution algorithm and factorized distribution algorithm. *IEEE Transactions on Evolutionary Computation*, 8(1):80–93, 2004. doi: 10.1109/TEVC.2003.819431.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008.