FlowRL: Matching Reward Distributions for LLM Reasoning

Xuekai Zhu[1], **Daixuan Cheng**[6], **Dinghuai Zhang**[3], **Hengli Li**[5], **Kaiyan Zhang**[4], **Che Jiang**[4], **Youbang Sun**[4], **Ermo Hua**[4], **Yuxin Zuo**[4], **Xingtai Lv**[4], **Qizheng Zhang**[7], **Lin Chen**[1], **Fanghao Shao**[1], **Bo Xue**[1], **Yunchong Song**[1], **Zhenjie Yang**[1], **Ganqu Cui**[2], **Ning Ding**[4,2], **Jianfeng Gao**[3], **Xiaodong Liu**[3], **Bowen Zhou**[4,2‡], **Hongyuan Mei**[8‡], **Zhouhan Lin**[1,2‡]

[1] Shanghai Jiao Tong University [2] Shanghai AI Laboratory [3] Microsoft Research [4] Tsinghua University [5] Peking University [6] Renmin University of China [7] Stanford University [8] Toyota Technological Institute at Chicago

hongyuanmei@gmail.com xuekaizhu0@gmail.com FlowRL ‡ Corresponding Authors.

Abstract | We propose FlowRL: matching the full reward distribution via flow balancing instead of maximizing rewards in large language model (LLM) reinforcement learning (RL). Recent advanced reasoning models adopt reward-maximizing methods (*e.g.*, PPO and GRPO), which tend to over-optimize dominant reward signals while neglecting less frequent but valid reasoning paths, thus reducing diversity. In contrast, we transform scalar rewards into a normalized target distribution using a learnable partition function, and then minimize the reverse KL divergence between the policy and the target distribution. We implement this idea as a flow-balanced optimization method that promotes diverse exploration and generalizable reasoning trajectories. We conduct experiments on math and code reasoning tasks: FlowRL achieves a significant average improvement of 10.0% over GRPO and 5.1% over PPO on math benchmarks, and performs consistently better on code reasoning tasks. These results highlight reward distribution-matching as a key step toward efficient exploration and diverse reasoning in LLM reinforcement learning.

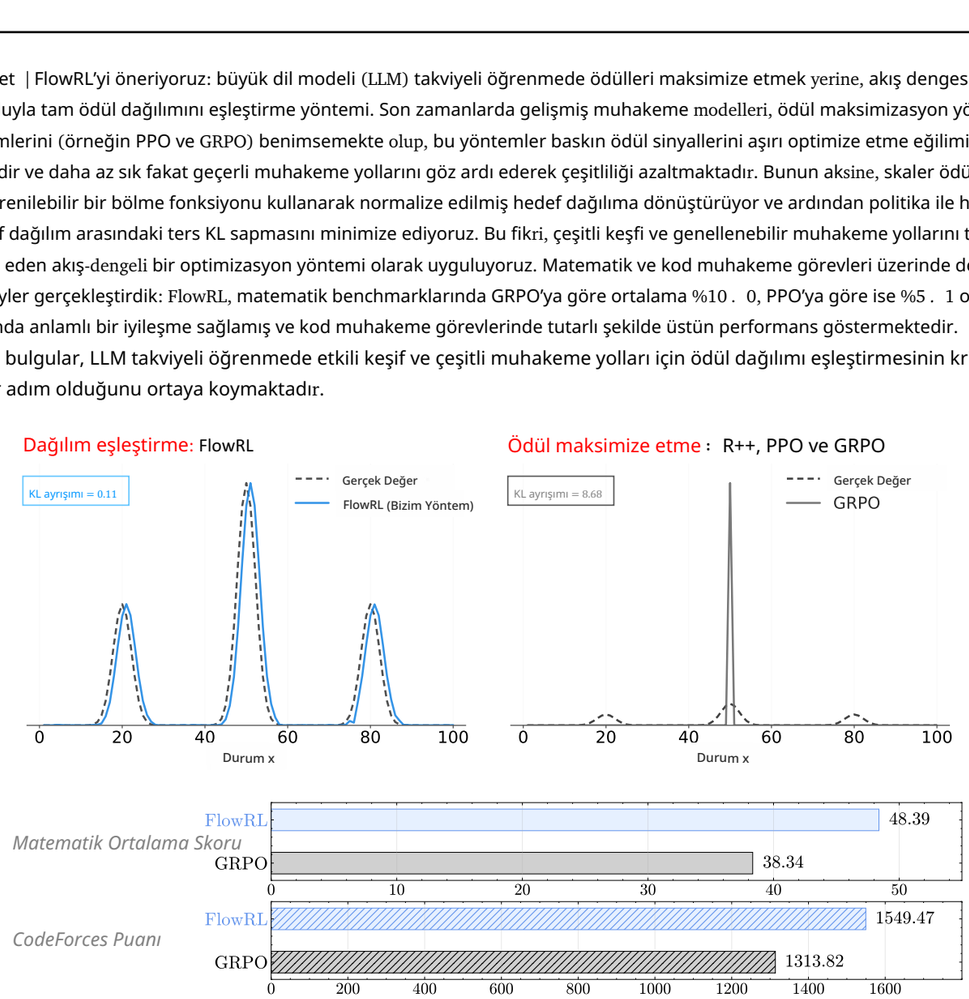
Figure 1 | **Top**: Comparison between distribution-matching and reward-maximizing approaches. FlowRL (left) learns to match the full reward distribution, maintaining diversity across multiple modes with low KL divergence. In contrast, reward-maximizing methods like GRPO (right) concentrate on a single high-reward peak, leading to mode collapse and higher KL divergence. **Bottom**: Performance comparison. FlowRL consistently outperforms GRPO across math and code domains.

FlowRL: LLM Muhakemesi için Ödül Dağılımlarının Eşleştirilmesi

Xuekai Zhu[1], **Daixuan Cheng**[6], **Dinghuai Zhang**[3], **Hengli Li**[5], **Kaiyan Zhang**[4], **Che Jiang**[4], **Youbang Sun**[4], **Ermo Hua**[4], **Yuxin Zuo**[4], **Xingtai Lv**[4], **Qizheng Zhang**[7], **Lin Chen**[1], **Fanghao Shao**[1], **Bo Xue**[1], **Yunchong Song**[1], **Zhenjie Yang**[1], **Ganqu Cui**[2], **Ning Ding**[4,2], **Jianfeng Gao**[3], **Xiaodong Liu**[3], **Bowen Zhou**[4,2‡], **Hongyuan Mei**[8‡], **Zhouhan Lin**[1,2‡]

[1] Shanghai Jiao Tong Üniversitesi [2] Shanghai Yapay Zeka Laboratuvarı [3] Microsoft Araştırma [4] Tsinghua Üniversitesi [5] Pekin Üniversitesi [6] Çin Halk Üniversitesi [7] Stanford Üniversitesi [8] Chicago Toyota Teknoloji Enstitüsü hongyuanmei@gmail.com xuekaizhu0@gmail.com FlowRL ‡ Sorumlu Yazarlar.

Özet | FlowRL'yi öneriyoruz: büyük dil modeli (LLM) takviyeli öğrenmede ödülleri maksimize etmek yerine, akış dengesi yoluyla tam ödül dağılımını eşleştirme yöntemi. Son zamanlarda gelişmiş muhakeme modelleri, ödül maksimizasyon yöntemlerini (örneğin PPO ve GRPO) benimsemekte olup, bu yöntemler baskın ödül sinyallerini aşırı optimize etme eğilimindedir ve daha az sık fakat geçerli muhakeme yollarını göz ardı ederek çeşitliliği azaltmaktadır. Bunun aksine, skaler ödülleri öğrenilebilir bir bölme fonksiyonu kullanarak normalize edilmiş hedef dağılıma dönüştürüyor ve ardından politika ile hedef dağılım arasındaki ters KL sapmasını minimize ediyoruz. Bu fikri, çeşitli keşfi ve genellenebilir muhakeme yollarını teşvik eden akış-dengeli bir optimizasyon yöntemi olarak uyguluyoruz. Matematik ve kod muhakeme görevleri üzerinde deneyler gerçekleştirdik: FlowRL, matematik benchmarklarında GRPO'ya göre ortalama %10.0, PPO'ya göre ise %5.1 oranında anlamlı bir iyileşme sağlamış ve kod muhakeme görevlerinde tutarlı şekilde üstün performans göstermektedir. Bu bulgular, LLM takviyeli öğrenmede etkili keşif ve çeşitli muhakeme yolları için ödül dağılımı eşleştirmesinin kritik bir adım olduğunu ortaya koymaktadır.



Şekil 1 |Üst: Dağılım eşleştirme ile ödül maksimize etme yaklaşımlarının karşılaştırılması. FlowRL (solda), tam ödül dağılımını eşleştirmeyi öğrenerek düşük KL ayrışımıyla çoklu modlarda çeşitliliği korur. Buna karşılık, GRPO gibi ödül maksimize etme yöntemleri (sağda) tek bir yüksek ödüllü tepeye odaklanır ve bu durum mod çöküşüne ve daha yüksek KL ayrışımına yol açar. Alt : Performans karşılaştırması. FlowRL, matematik ve kodlama alanlarında GRPO'yu sürekli olarak geride bırakmaktadır.

1. Introduction

Reinforcement learning (RL) plays a crucial role in the post-training of large language models (LLMs) [Zhang et al., 2025b]. A series of powerful reasoning models [Guo et al., 2025, Kavukcuoglu, 2025, Rastogi et al., 2025] have employed large-scale reinforcement learning to achieve strong performance on highly challenging benchmarks [He et al., 2024]. The evolution of RL algorithms for LLM reasoning has progressed through several key stages: REINFORCE [Sutton et al., 1999a] provides a solid baseline that is easy to implement and efficient in simple settings; PPO [Schulman et al., 2017] improves upon REINFORCE with better stability and efficiency in complex settings; GRPO [Shao et al., 2024] simplifies PPO training by eliminating value functions and relying on group comparisons, though at the cost of requiring more rollouts per update. However, all these methods share a fundamental limitation in their reward-maximizing objective.

Reward-maximizing RL methods tend to overfit to the dominant mode of the reward distribution [Gao et al., 2023, Pan et al., 2022, Skalse et al., 2022, Zelikman et al., 2022]. This often results in limited diversity among generated reasoning paths and reduces generalization to less frequent yet valid logical outcomes [Hu et al., 2023]. As illustrated in Figure 1, GRPO neglects other meaningful modes. These drawbacks become especially pronounced in complex long chain-of-thought (CoT; Wei et al., 2022) reasoning, where capturing a diverse distribution of plausible solutions is essential for effective generalization [Liu et al., 2025a]. Recent approaches adjust the clip ratio [Yu et al., 2025b], augment the advantage function with an entropy-based term [Cheng et al., 2025], or selectively promote high-entropy tokens [Wang et al., 2025], thereby dynamically adapting the training data distribution and implicitly increasing diversity during training. This raises a fundamental question: How can we promote diverse exploration to prevent convergence to dominant solution patterns in RL training?

In this paper, we propose **FlowRL**, a policy optimization algorithm that aligns the policy model with the full reward distribution, encouraging mode coverage. FlowRL achieves more efficient exploration by fundamentally shifting from reward maximization to reward distribution matching, thereby addressing the inherent mode-collapse limitations of previous RL approaches. As illustrated in Figure 1, the core idea of FlowRL is to introduce a learnable partition function that normalizes scalar rewards into a target distribution, and to minimize the reverse KL divergence between the policy and this reward-induced distribution. We develop this KL objective based on the trajectory balance formulation from GFlowNets [Bengio et al., 2023b], providing a gradient equivalence proof that bridges generative modeling and policy optimization. To address the challenges of long CoT training, we introduce two key technical solutions: *length normalization* to tackle gradient explosion issues that occur with variable-length CoT reasoning, and *importance sampling* to correct for the distribution mismatch between generated rollouts and the current policy.

We compare FlowRL with mainstream RL algorithms including REINFORCE++, PPO, and GRPO across math and code domains, using both base and distilled LLMs (7B, 32B). In math domain, FlowRL outperforms GRPO and PPO by 10.0% and 5.1%, respectively, demonstrating consistent improvements across six challenging math benchmarks. Furthermore, FlowRL surpasses both PPO and GRPO on three challenging coding benchmarks, highlighting its strong generalization capabilities in code reasoning tasks. To understand what drives these performance gains, we analyze the diversity of generated reasoning paths. This diversity analysis confirms that FlowRL generates substantially more diverse rollouts than baseline methods, validating our approach's effectiveness in exploring multiple solution strategies.

Contributions. We summarize the key contributions of this work as follows:

- We propose FlowRL, a policy optimization algorithm that shifts from reward maximization to

1. Giriş

Takviyeli öğrenme (RL), büyük dil modellerinin (LLM'lerin) eğitim sonrası post-training süreçlerinde kritik bir rol oynamaktadır [Zhang et al., 2025b]. Güçlü muhakeme modelleri serisi [Guo et al., 2025, Kavukcuoglu, 2025, Rastogi et al., 2025], zorlayıcı benchmarklarda yüksek performans elde etmek için büyük ölçekli takviyeli öğrenme teknikleri kullanmıştır [He et al., 2024]. LLM muhakemesi için RL algoritmalarının gelişimi birkaç önemli aşama üzerinden gerçekleşmiştir: REINFORCE [Sutton et al., 1999a] basit ortamlarda uygulanması kolay ve verimliliği yüksek sağlam bir temel sunar; PPO [Schulman et al., 2017], karmaşık ortamlarda daha iyi kararlılık ve verimlilik sağlayarak REINFORCE yöntemini geliştirir; GRPO [Shao et al., 2024], değer fonksiyonlarını ortadan kaldırıp grup karşılaştırmalarına dayanarak PPO eğitimini basitleştirir; ancak bu, her güncellemede daha fazla deneme yapılmasını gerektirir. Ancak , tüm bu yöntemler ödül-maksimizasyon amaçlarında temel bir sınırlamayı paylaşmaktadır.

Ödül-maksimize eden RL yöntemleri, ödül dağılımının baskın moduna aşırı uyum sağlama eğilimindedir [Gao et al., 2023, Pan et al., 2022, Skalse et al., 2022, Zelikman et al., 2022]. Bu durum genellikle oluşturulan muhakeme yolları arasında sınırlı çeşitliliğe ve daha az yaygın ancak geçerli mantıksal sonuçlara genellemenin azalmasına yol açar [Hu et al., 2023]. Şekil 1'de gösterildiği üzere, GRPO diğer anlamlı modları göz ardı eder. Bu dezavantajlar, geçerli olası çözümlerin çeşitli dağılımını yakalamanın etkili genelleme için kritik olduğu karmaşık ve uzun düşünce zinciri (CoT; Wei et al., 2022) muhakemesinde özellikle belirgindir [Liu et al., 2025a]. Son yaklaşımlar, klip oranını [Yu et al., 2025b] ayarlamakta, avantaj fonksiyonunu entropiye dayalı bir terimle genişletmekte [Cheng et al., 2025] veya yüksek entropili tokenları seçici olarak teşvik etmekte [Wang et al., 2025]; böylece eğitim veri dağılımını dinamik olarak uyarlayarak eğitim sürecinde çeşitliliği dolaylı olarak artırmaktadır. Bu, temel bir soruyu gündeme getirmektedir:
PL eğitiminde baskın çözüm kalıplarına yakınsamayı önlemek için çeşitliliği nasıl teşvik edebiliriz?

Bu makalede, politika modelini tam ödül dağılımı ile hizalayan ve mod kapsamasını teşvik eden bir politika optimizasyon algoritması olan **FlowRL** 'yi öneriyoruz. FlowRL, ödül maksimize etmekten ödül dağılımı eşleştirmeye temel bir geçiş yaparak daha verimli keşif sağlamakta ve böylece önceki PL yaklaşımlarının doğal mod çöküşü sınırlamalarını aşmaktadır. Şekil 1'de gösterildiği üzere, FlowRL'nin temel fikri, skaler ödülleri normalize eden öğrenilebilir bir bölme fonksiyonu tanımlamak ve politika ile bu ödül kaynaklı dağılım arasındaki ters KL sapmasını minimize etmektir. KL amaç fonksiyonunu, generatif modelleme ile politika optimizasyonu arasında köprü kuran gradyan eşdeğerliği kanıtı sunarak GFlowNets [Bengio et al., 2023b] tarafından geliştirilen yörünge dengelemesi formülasyonuna dayandırdık. Uzun CoT eğitiminin zorluklarına yönelik olarak iki temel teknik çözüm sunuyoruz: *uzunluk normalizasyonu* değişken uzunluktaki CoT muhakemede ortaya çıkan gradyan patlaması sorunlarını önlemek için ve *önemli örnekleme* oluşan denemeler ile mevcut politika arasındaki dağılım uyumsuzluğunu düzeltmek amacıyla.

FlowRL'yi, matematik ve kod alanlarında hem temel hem de distile edilmiş LLMLer (7B, 32B) kullanarak REINFORCE++, PPO ve GRPO dahil olmak üzere yaygın olarak kullanılan RL algoritmalarıyla karşılaştırıyoruz. Matematik alanında FlowRL, altı zorlu matematik kıyaslamasında tutarlı iyileştirmeler göstererek GRPO'ya $\%10.0$, PPO'ya ise $\%5.1$ oranında üstünlük sağlamaktadır. Ayrıca, FlowRL üç zorlu kodlama kıyaslamasında hem PPO'yu hem de GRPO'yu geride bırakarak kod muhakeme görevlerinde güçlü genelleme kabiliyetlerini ortaya koymaktadır. Bu performans kazanımlarını neyin tetiklediğini anlamak için, üretilen muhakeme yollarının çeşitliliğini analiz ediyoruz. Bu çeşitlilik analiz i, FlowRL'nin temel yöntemlere kıyasla önemli ölçüde daha çeşitli denemeler ürettiğini doğrulayarak , yaklaşımımızın çoklu çözüm stratejilerini keşfetmedeki etkinliğini teyit etmektedir.

Katkılar. Bu çalışmanın temel katkılarını şu şekilde özetliyoruz:

- Ödül maksimize etmekten

reward distribution matching via flow balance, encouraging diverse reasoning path exploration while addressing the inherent mode-collapse limitations of existing RL methods.
- We introduce length normalization and importance sampling to enable effective training on variable-length CoT reasoning, addressing gradient explosion and sampling mismatch issues.
- FlowRL outperforms GRPO and PPO by 10.0% and 5.1% respectively across math benchmarks and demonstrates strong generalization on code reasoning tasks, with diversity analysis confirming substantially more diverse solution exploration.

2. Preliminaries

Reinforcement Learning for Reasoning. We formulate reasoning as a conditional generation problem, where the policy model receives a question $\mathbf{x} \in \mathcal{X}$ and generates an answer $\mathbf{y} \in \mathcal{Y}$. The objective is to learn a policy $\pi_\theta(\mathbf{y}|\mathbf{x})$ that produces high-quality answers under task-specific reward signals r. To better illustrate the policy optimization procedure, we provide a detailed formulation of GRPO below. For each question $\mathbf{x}$, GRPO samples a group of answers $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_G\}$ from old policy $\pi_{\theta_{old}}$ and updates the model by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{[\mathbf{x}\sim P(\mathcal{X}),\{\mathbf{y}_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\mathcal{Y}|\mathbf{x})]}$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{y}_i|}\sum_{t=1}^{|\mathbf{y}_i|}\left\{\min\left[\frac{\pi_\theta(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}{\pi_{\theta_{old}}(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}{\pi_{\theta_{old}}(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \lambda\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\},$$
$$\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(\mathbf{y}_i|\mathbf{x})}{\pi_\theta(\mathbf{y}_i|\mathbf{x})} - \log\frac{\pi_{\text{ref}}(\mathbf{y}_i|\mathbf{x})}{\pi_\theta(\mathbf{y}_i|\mathbf{x})} - 1, \tag{1}$$

where ϵ and λ are hyper-parameters. Here, A_i denotes the advantage, computed by normalizing the group reward values $\{r_1, r_2, \ldots, r_G\}$ as $A_i = \frac{r_i - \text{mean}(\{r_1,r_2,\cdots,r_G\})}{\text{std}(\{r_1,r_2,\cdots,r_G\})}$. Compared to GRPO, REINFORCE applies the policy gradient directly, without advantage normalization, clipping, or KL regularization. PPO uses a critic model to estimate the advantage and employs importance sampling to stabilize policy updates.

GFlowNets. Generative Flow Networks [Bengio et al., 2023a] are a probabilistic framework for training stochastic policies to sample discrete, compositional objects (*e.g.*, graphs, sequences) in proportion to a given reward. As shown in Figure 2, the core principle of GFlowNets is to balance the forward and backward probability flows at each state, inspired by flow matching [Bengio et al., 2021]. The initial flow is estimated by $Z_\phi(s_0)$ at the initial state s_0. The output flow is equal to the outcome reward $r(s_n)$ conditioned at the final state s_n. Following Lee et al. [2024], we use a 3-layer MLP to parameterize Z_ϕ. This flow-balancing mechanism facilitates the discovery of diverse, high-reward solutions by ensuring proper exploration of the solution space. See Appendix C for detailed GFlowNets background.

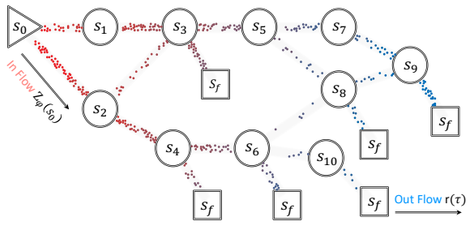


Figure 2 | GFlowNets [Bengio et al., 2023a], a flow-balance perspective on reinforcement learning. The initial flow $Z_\phi(s_0)$ injects probability mass into the environment, which is transported through intermediate states by the policy π_θ and accumulated at terminal states in proportion to the scalar rewards.

akış dengesi yoluyla ödül dağılımı eşleştirmesine geçen, mevcut Takviyeli Öğrenme yöntemlerinin içsel mod çöküşü sınırlamalarını ele alırken çeşitli muhakeme yollarının keşfini teşvik eden FlowRL adlı bir politika optimizasyon algoritması öneriyoruz.
- Değişken uzunluktaki CoT muhakemesinde etkili eğitimi mümkün kılmak için uzunluk normalizasyonu ve önemli örneklemeyi tanıtıyor, böylece gradyan patlaması ve örnekleme uyumsuzluğu sorunlarını çözüyoruz.
- FlowRL, matematik kıyaslamalarında sırasıyla GRPO ve PPO'yu %10,0 ve %5,1 oranında geride bırakmakta ve kod muhakeme görevlerinde güçlü genelleme göstermekte olup, çeşitlilik analizi önemli ölçüde daha çeşitli çözüm keşfini doğrulamaktadır.

2. Ön Bilgiler

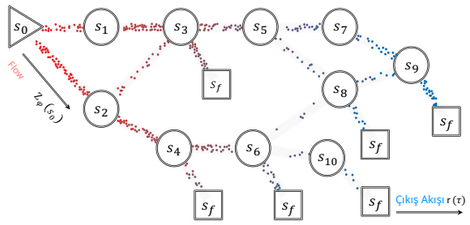
Muhakeme için Takviyeli Öğrenme. Muhakemeyi, politika modelinin bir soru $\mathbf{x} \in \mathcal{X}$ alıp bir cevap $\mathbf{y} \in \mathcal{Y}$ türetmesi biçiminde koşullu üretim problemi olarak formüle ediyoruz. Amaç, görev-özel ödül sinyalleri r altında yüksek kaliteli yanıtlar üreten bir politika $\pi\ \theta\ (\ y\ |\ x\ )$ öğrenmektir.
Politika optimizasyon prosedürünü daha iyi açıklamak amacıyla, aşağıda GRPO'nun ayrıntılı formülasyonunu su nuyoruz. Her bir soru $\mathbf{x}$ için GRPO, eski politika $\pi\ \theta$ altından bir grup yanıt $\{\ \mathbf{y}\ 1\ ,\ \mathbf{y}\ 2\ ,\ \ldots\ ,\ \mathbf{y}\ G\ \}$ örnekler.
Ve modeli aşağıdaki amacı maksimize ederek günceller:

$$\mathcal{J}_{GGGG}(\theta) = \mathbb{E}_{[\mathbf{x}\sim P(\mathcal{X}),\{\mathbf{y}_i\}_{i=1}^{G}\sim\pi_{\theta_{ooo}}(\mathcal{Y}|\mathbf{x})]}$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{y}_i|}\sum_{t=1}^{|\mathbf{y}_i|}\left\{\min\left[\frac{\pi_\theta(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}{\pi_{\theta_{ooo}}(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}{\pi_{\theta_{ooo}}(\mathbf{y}_{i,t}|\mathbf{x},\mathbf{y}_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \lambda\mathbb{D}_{KK}\left[\pi_\theta\|\pi_{rr\,f}\right]\right\},$$
$$\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{re}/}) = \frac{\pi_{\text{re}/}(\mathbf{y}_i|\mathbf{x})}{\pi_\theta(\mathbf{y}_i|\mathbf{x})} - \text{logaritma}\frac{\pi_{\text{re}/}}{\pi_\theta}(\mathbf{y}_i|\mathbf{x}) - 1, \tag{1}$$

burada ϵ ve λ hiperparametrelerdir. Burada, A_i avantajı ifade eder ve grup ödül değerleri $\{r_1, r_2, \ldots, r_G\}$ ortalaması alınarak normalize edilip $A_i = {}^{r_i -}$ ortalama $^{(\{r_1,r_2,\cdots,r_G\})}$ şeklinde hesaplanır. $\text{std}(\{r_1, r_2, \cdots, r_G\})$. GRPO ile karşılaştırıldığında, REINFORCE politika gradyanını doğrudan uygular; avantaj normalizasyonu, klipleme veya KL düzenlemesi yapmaz. PPO, avantajı tahmin etmek için bir kritik model kullanır ve politika güncellemelerini stabilize etmek amacıyla önemli örneklemeyi uygular.

GFlowNets. Generatif Akış Ağları [Bengio vd., 2023a], belirli bir ödüle orantılı olarak ayrık, bileşimsel nesneleri (örneğin, grafikler, diziler) örneklemek için stokastik politikaların eğitilmesi için olasılıksal bir çerçevedir.

Şekil 2'de gösterildiği üzere, GFlowNets'in temel ilkesi, her durumda ileri ve geri olasılık akışlarını dengelemek olup, bu ilham kaynağı akış eşlemeden [Bengio vd., 2021] gelmektedir. Başlangıç akışı, başlangıç durumu s_0'da Z_ϕ(s_0) ile tahmin edilir. Çıkış akışı, son durum s_n'de koşullandırılmış olan elde edilen ödül r(s_n) ile eşittir. Lee vd. [2024]'e dayanarak, Z_ϕ parametrelendirmesi için 3 katmanlı bir MLP kullanıyoruz. Bu akış dengeleme mekanizması, çözüm uzayının uygun şekilde keşfedilmesini sağlayarak çeşitli, yüksek ödüllü çözümlerin bulunmasını kolaylaştırır. Detaylı GFlowNets bilgisi için Ek C'ye bakınız.



Şekil **2** | GFlowNets [Bengio ve ark., 2023a], takviyeli öğrenmeye yönelik akış-dengesi perspektifi. Başlangıç akışı $Z\ \phi\ (\ s\ 0\ )$ ortama olasılık kütlesi enjekte eder; bu kütle ara durumlar boyunca politika $\pi\ \theta$ tarafından taşınır ve skalar ödüllerle orantılı olarak terminal durumlarda birikir.

3. Methodology

In this section, we first formulate distribution matching in reinforcement learning through reverse KL divergence and establish its connection to trajectory balance from GFlowNets. To address the challenges of gradient explosion and sampling mismatch encountered during long CoT training, we further incorporate length normalization and importance sampling. Using this enhanced framework, we derive a flow-balanced objective, termed *FlowRL*.

3.1. From Reward Maximization to Distribution Matching

As illustrated in Figure 1, recent powerful large reasoning models typically employ reward-maximizing RL algorithms, such as PPO or GRPO. However, these methods tend to optimize toward the dominant reward mode, frequently resulting in mode collapse and the neglect of other plausible, high-quality reasoning paths. To address this fundamental limitation, we propose optimizing the policy by aligning its output distribution to a target reward distribution. A simple yet effective way to achieve this is to minimize the reverse KL divergence[1] between the policy and this target. However, in long CoT reasoning tasks, the available supervision in RL is a scalar reward, rather than a full distribution. Moreover, enumerating or sampling all valid trajectories to recover the true reward distribution is computationally intractable.

Inspired by energy-based modeling [Du and Mordatch, 2019, Hinton et al., 1995], we introduce a learnable partition function $Z_\phi(\mathbf{x})$ to normalize scalar rewards into a valid target distribution. This allows us to minimize the reverse KL divergence between the policy and the reward-weighted distribution, formalized as:

$$\min_\theta \mathcal{D}_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \,\middle\|\, \frac{\exp(\beta r(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}\right) \quad \Rightarrow \quad \pi_\theta(\mathbf{y}\mid\mathbf{x}) \propto \exp(\beta r(\mathbf{x},\mathbf{y})), \tag{2}$$

where $r(\mathbf{x},\mathbf{y})$ is the reward function, β is a hyperparameter, $Z_\phi(\mathbf{x})$ is the learned partition function, and the resulting target distribution is defined as $\tilde{\pi}(\mathbf{y}\mid\mathbf{x}) = \frac{\exp(\beta r(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}$. This objective encourages the policy to sample diverse, high-reward trajectories in proportion to their rewards, rather than collapsing to dominant modes as in standard reward maximization.

While the KL-based formulation provides a principled target distribution, we derive a more practical, RL-style objective that facilitates efficient policy optimization.

Proposition 1. *In terms of expected gradients, minimizing the KL objective in Eq. 2 is equivalent to minimizing the trajectory balance loss used in GFlowNet [Bartoldson et al., 2025, Lee et al., 2024, Malkin et al., 2022, 2023]:*

$$\min_\theta \mathcal{D}_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \,\middle\|\, \frac{\exp(\beta r(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}\right) \iff \min_\theta \underbrace{\left(\log Z_\phi(\mathbf{x}) + \log \pi_\theta(\mathbf{y}\mid\mathbf{x}) - \beta r(\mathbf{x},\mathbf{y})\right)^2}_{\textit{Trajectory Balance}} \tag{3}$$

Remark 2 (*Trajectory balance as a practical surrogate for KL minimization*)**.** Given the equivalence established in Proposition 1, the KL-based distribution matching objective can be reformulated as the trajectory balance loss. This reformulation provides a practical optimization approach by using a stable squared loss form rather than direct KL optimization, and by treating $Z_\phi(\mathbf{x})$ as a learnable parameter rather than requiring explicit computation of the intractable partition function. The trajectory balance objective thus serves as a tractable surrogate for reward-guided KL minimization that can be directly integrated into existing RL frameworks.

[1] We use reverse KL since we can only sample from the policy model, not the target reward distribution.

3. Yöntem

Bu bölümde, öncelikle takviyeli öğrenmede dağılım eşleştirmeyi ters KL sapması üzerinden formüle ederek bunu GFlowNets'in yörünge dengelemesi ile ilişkilendiriyoruz. Uzun CoT eğitimi sırasında karşılaşılan gradyan patlaması ve örnekleme uyumsuzluğu sorunlarını ele almak amacıyla ayrıca uzunluk normalizasyonu ve önemli örnekleme yöntemlerini entegre ediyoruz. Bu geliştirilmiş çerçeveyi kullanarak, *FlowRL* adını verdiğimiz akış-dengeli bir amaç fonksiyonu türetiyoruz.

3.1. Ödül Maksimizasyonundan Dağılım Eşleştirmeye

Şekil 1'de gösterildiği üzere, son dönemdeki güçlü büyük muhakeme modelleri genellikle PPO veya GRPO gibi ödül-maksimize eden takviyeli öğrenme algoritmalarını kullanmaktadır. Ancak bu yöntemler baskın ödül moduna yönelme eğilimindedir; bu durum sıkça mod çöküşüne ve diğer olası, yüksek kaliteli muhakeme yollarının göz ardı edilmesine yol açmaktadır. Bu temel sınırlamayı aşmak için, politikayı çıktı dağılımını hedef ödül dağılımı ile hizalayarak optimize etmeyi öneriyoruz. Bunu başarmanın basit fakat etkili bir yolu, politika ile bu hedef arasındaki ters KL sapmasını en aza indirmektir [1]. Ancak uzun CoT muhakemesi görevlerinde, RL'de mevcut denetim tam bir dağılım yerine skaler bir ödüldür. Ayrıca, gerçek ödül dağılımını yeniden elde etmek için tüm geçerli yolları sıralamak veya örneklemek hesaplama açısından uygulanamazdır.

Enerji tabanlı modellemeden esinlenerek [Du ve Mordatch, 2019, Hinton ve ark., 1995], skaler ödülleri geçerli bir hedef dağılıma normalize etmek için öğrenilebilir bir bölme fonksiyonu $Z_\phi(\mathbf{x})$ anıtıyoruz. Bu, politika ile ödül ağırlıklı dağılım arasındaki ters KL sapmasını en aza indirmemize olanak verir ve şu şekilde formüle edilir:

$$\min_\theta \mathcal{D}_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \,\middle\|\, \frac{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}\right) \quad \Rightarrow \quad \pi_\theta(\mathbf{y}\mid\mathbf{x}) \propto \exp(\beta\beta(\mathbf{x},\mathbf{y})), \tag{2}$$

Burada $r(\mathbf{x},\mathbf{y})$ ödül fonksiyonudur, β bir hiperparametredir, $Z_\phi(\mathbf{x})$ öğrenilmiş bölme fonksiyonudur ve ortaya çıkan hedef dağılım $\tilde{\pi}(\mathbf{y}\mid\mathbf{x}) = \frac{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}$ şeklinde tanımlanır. Bu amaç, politikanın standart ödül maksimize etme yöntemlerinde görülen baskın modlara çökme yerine, ödüllerine orantılı olarak çeşitli ve yüksek ödüllü izlekleri örneklemesini teşvik eder.

KL tabanlı formülasyon prensipli bir hedef dağılım sağlarken, biz verimli politika optimizasyonunu kolaylaştıran daha pratik, RL tarzı bir amaç türetmekteyiz.

Önerme 1. *Beklenen gradyanlar açısından bakıldığında, Eq. 2'deki KL amaç fonksiyonunun minimize edilmesi, GFlowNet'te kullanılan yörünge dengeleme kaybının minimize edilmesiyle eşdeğerdir [Bartoldson ve ark., 2025, Lee ve ark., 2024, Malkin ve ark., 2022, 2023]:*

$$\min_\theta \mathcal{D}_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \,\middle\|\, \frac{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}\right) \iff \min_\theta \underbrace{\left(\log Z_\phi(\mathbf{x}) + \log \pi_\theta(\mathbf{y}\mid\mathbf{x}) - \beta\beta(\mathbf{x},\mathbf{y})\right)^2}_{\textit{Yörünge Dengelemesi}} \tag{3}$$

Not **2** (KL minimizasyonu için pratik bir vekil olarak yörünge dengelemesi)**.** Önerme 1'de sağlanan eşdeğerlik sayesinde, KL tabanlı dağılım eşleştirme amacı yörünge dengeleme kaybı olarak yeniden formüle edilebilir. Bu yeniden formülasyon, bölme fonksiyonunun zorluklarına karşın açık hesaplama gerektirmeden öğrenilebilir bir parametre olarak ele alınmasına ve doğrudan KL optimizasyonu yerine stabil bir karesel kayıp formunun kullanılmasına imkan tanıyan pratik bir optimizasyon yöntemi sağlar. Yörünge dengelemesi amacı, böylece mevcut RL çerçevelerine doğrudan entegre edilebilen, ödül rehberli KL minimizasyonu için uygulanabilir bir vekil görevini görür.

[1] Hedef ödül dağılımından değil, yalnızca politika modelinden örnek alabileceğimiz için ters KL kullandık.

3.2. FlowRL

As established in Proposition 1, the target reward distribution can be approximated by optimizing the trajectory balance objective. However, applying this objective directly to long CoT reasoning introduces two key challenges:

Problem I: Exploding gradients from long trajectories. Trajectory balance is a sequence-level objective, and applying it to long CoT reasoning with up to 8K tokens leads to exploding gradients and unstable updates. This issue is not observed in prior GFlowNets works, which typically operate on short trajectories in small discrete spaces. Specifically, the log-probability term $\log \pi_\theta(\mathbf{y} \mid \mathbf{x})$ decomposes into a token-wise sum, $\sum_t \log \pi_\theta(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{x})$, causing the gradient norm to potentially scale with sequence length.

Problem II: Sampling mismatch. Mainstream RL algorithms such as PPO and GRPO commonly perform micro-batch updates and reuse trajectories collected from an old policy $\pi_{\theta_{\text{old}}}$, enabling data-efficient training. In contrast, the KL-based trajectory balance objective assumes fully on-policy sampling, where responses are drawn from the current policy. This mismatch poses practical limitations when integrating trajectory balance into existing RL pipelines.

These limitations motivate our reformulation that retains the benefits of distribution matching while addressing key practical challenges. To enable this reformulation, we first redefine the reward function following established practices in GFlowNets literature [Bartoldson et al., 2025, Lee et al., 2024, Yu et al., 2025a] by incorporating a reference model as a prior constraint on the reward distribution. Specifically, we modify the original $\exp(\beta r(\mathbf{x}, \mathbf{y}))$ to include the reference model:

$$\exp\left(\beta\, r(\mathbf{x}, \mathbf{y})\right) \cdot \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x}), \tag{4}$$

where $r(\mathbf{x}, \mathbf{y})$ denotes the outcome reward commonly used in reinforcement learning and π_{ref} is the initial pre-trained model. We follow Guo et al. [2025] to use outcome-based reward signals, and apply group normalization to $r(\mathbf{x}, \mathbf{y})$ as $\hat{r}_i = (r_i - \text{mean}(\mathbf{r}))/\text{std}(\mathbf{r})$, where $\mathbf{r} = \{r_1, r_2, \ldots, r_G\}$ denotes the set of rewards within a sampled group. By substituting the redefined reward formulation Eq. 4 into Eq. 3, we derive the following objective[2]:

$$\min_\theta \left(\log Z_\phi(\mathbf{x}) + \log \pi_\theta(\mathbf{y} \mid \mathbf{x}) - \beta\, \hat{r}_i(\mathbf{x}, \mathbf{y}) - \log \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\right)^2 \tag{5}$$

Remark 3 (*Reward shaping via length normalization*)**.** Trajectory balance treats both the initial flow and the outcome reward as sequence-level quantities. In contrast, standard policy optimization methods such as PPO or GRPO assign rewards at the token level and compute gradients at each step. However, for trajectories of varying lengths (*e.g.*, CoT responses), this mismatch can cause the log-probability term $\log \pi_\theta(\mathbf{y} \mid \mathbf{x}) = \sum_{t=1}^{|\mathbf{y}|} \log \pi_\theta(y_t \mid y_{<t}, \mathbf{x})$ to scale with sequence length. To address this, we apply a form of reward shaping by normalizing log-probabilities with respect to sequence length. Specifically, we rescale the term as $\frac{1}{|\mathbf{y}|} \log \pi_\theta(\mathbf{y} \mid \mathbf{x})$, balancing the contributions of long and short sequences and stabilizing the learning signal.

Remark 4 (*Importance sampling for data-efficient training*)**.** To mitigate sampling mismatch, we employ importance sampling inspired by PPO to stabilize policy updates with off-policy data. We re-weight stale trajectories using the importance ratio $w = \pi_\theta(\mathbf{y} \mid \mathbf{x})/\pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})$, which serves as a coefficient in the surrogate loss. Since our objective focuses on optimizing trajectory balance rather than expected return, we detach the gradient from the current policy to prevent excessive policy drift: $w = \text{detach}[\pi_\theta(\mathbf{y} \mid \mathbf{x})]/\pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})$. For additional stability, we incorporate PPO-style clipping to bound the importance weights: $w = \text{clip}\left(\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\text{old}}(\mathbf{y}|\mathbf{x})}, 1-\epsilon, 1+\epsilon\right)^{\text{detach}}$.

[2]The substitution replaces $\beta r(\mathbf{x}, \mathbf{y})$ in trajectory balance objective Eq. 3 with $\beta r(\mathbf{x}, \mathbf{y}) + \log \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$ to incorporate the reference model constraint.

3.2. FlowRL

Önerme 1'de belirtildiği üzere, hedef ödül dağılımı yörünge dengelemesi amacını optimize ederek yaklaşık olarak elde edilebilir. Ancak, bu amacı uzun CoT muhakemesine doğrudan uygulamak iki temel zorluğu beraberinde getirir:

Problem I: Uzun yörüngelerden kaynaklanan patlayan gradyanlar. Yörünge dengelemesi dizi düzeyinde bir amaçtır ve 8 bine kadar uzanan token uzunluğundaki uzun CoT muhakemesine uygulanması patlayan gradyanlara ve kararsız güncellemelere yol açar. Bu sorun, genellikle küçük ayrık alanlarda kısa yörüngeler üzerinde çalışan önceki GFlowNets çalışmalarında gözlemlenmemiştir. Özellikle, logaritmik olasılık terimi $\log \pi_\theta(\mathbf{y} \mid \mathbf{x})$ token bazında bir toplam olarak ayrışır, $\sum_t \log \pi_\theta(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{x})$; bu durum gradyan normunun dizi uzunluğuyla orantılı olarak artmasına neden olabilir.

Problem II: Örnekleme uyumsuzluğu. PPO ve GRPO gibi yaygın RL algoritmaları genellikle mikro-parti güncellemeleri uygular ve eski bir politika $\pi_{\theta_{\text{old}}}$ tarafından toplanan yörüngeleri yeniden kullanarak veri açısından verimli eğitim sağlar. Buna karşılık, KL tabanlı yörünge dengeleme amacı tamamen on-politika örneklemesini varsayar; burada yanıtlar mevcut politikadan alınır. Bu uyumsuzluk, yörünge dengelemenin mevcut RL iş akışlarına entegrasyonunda pratik sınırlamalar doğurur.

Bu sınırlamalar, dağılım eşleştirmenin avantajlarını korurken temel pratik zorlukları ele alan yeniden formülasyonumuzu motive etmektedir. Bu yeniden formülasyonu mümkün kılmak için, ödül dağılımına ön kısıtlama olarak bir referans modeli ekleyerek ödül fonksiyonunu, GFlowNets literatüründeki yerleşik uygulamalara uygun biçimde yeniden tanımlıyoruz [Bartoldson et al., 2025, Lee et al., 2024, Yu et al., 2025a]. Özellikle, orijinal $\exp(\beta\beta(\mathbf{x}, \mathbf{y}))$ ifadesini referans modeli içerecek şekilde modifiye ediyoruz:

$$\exp\left(\beta\, r(\mathbf{x}, \mathbf{y})\right) \cdot \pi_{\text{re}/}(\mathbf{y} \mid \mathbf{x}), \tag{4}$$

burada $r(\mathbf{x}, \mathbf{y})$ takviyeli öğrenmede yaygın olarak kullanılan çıktı ödülünü ifade eder ve $\pi_{\text{re}/}$ başlangıçta önceden eğitilmiş modeli temsil eder. Guo ve ark. [2025]'i takip ederek çıktı bazlı ödül sinyallerini kullanıyor ve grup normalizasyonunu $r(\mathbf{x}, \mathbf{y})$ için $\hat{}\hat{} r\, i = (r\, i - \text{ortalama}(\text{r}))/\text{std}(\text{r})$ şeklinde uyguluyoruz; burada $\text{r} = \{r\,1, r\,2, \text{ldots}, r\,G\}$ örneklenen grubun ödüller kümesini gösterir. Yeniden tanımlanmış ödül formülasyonu Denklem 4'ü Denklem 3'e ikame ederek aşağıdaki amaç 2 elde edilir:

$$\min_\theta \left(\text{logaritma}\, Z_\phi(\mathbf{x}) + \text{logaritma}\, \pi_\theta(\mathbf{y} \mid \mathbf{x}) - \beta\, \hat{r}_i(\mathbf{x}, \mathbf{y}) - \text{logaritma}\, \pi_{\text{re}/}(\mathbf{y} \mid \mathbf{x})\right)^2 \tag{5}$$

Not 3 (Uzunluk normalizasyonu yoluyla ödül şekillendirme). Yörünge dengelemesi, hem başlangıç akışını hem de sonuç ödülünü dizi düzeyinde nicelikler olarak ele alır. Buna karşılık, PPO veya GRPO gibi standart politika optimizasyon yöntemleri, ödülleri token düzeyinde atar ve her adımda gradyanları hesaplar. Ancak, değişken uzunluktaki yörüngeler için (örneğin, CoT yanıtları), bu uyumsuzluk logaritma olasılık terimi $\log \pi_\theta(\text{y} \mid \text{x}) = \sum_{t=1}^{|\text{y}|} \log \pi_\theta(y\,t \mid y\,{<}t, \text{x})$ ifadesinin dizi uzunluğuna göre ölçeklenmesine sebep olabilir. Bu durumu çözmek için, logaritmik olasılıkları dizgi uzunluğuna göre normalize ederek bir ödül şekillendirme yöntemi uygularız. Özellikle, terimi şu şekilde yeniden ölçeklendiriyoruz $\frac{1}{|\mathbf{y}|} \log \pi_\theta(\mathbf{y} \mid \mathbf{x})$, uzun ve kısa dizilerin katkılarını dengeleyerek öğrenme sinyalini stabilize eder.

Not 4 (*Veri-verimli eğitim için önemli örnekleme*). Örnekleme uyumsuzluğunu azaltmak amacıyla, PPO'dan esinlenmiş önemli örnekleme kullanarak politika güncellemelerinin off-policy verilerle stabil olmasını sağlıyoruz. Eski yörüngeleri, vekil kayıpta katsayı olarak kullanılan önem oranı $w = \pi_\theta(\mathbf{y} \mid \mathbf{x})/\pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})$ ile yeniden ağırlıklandırıyoruz. Amaç beklenen getiriden çok yörünge dengesini optimize etmek olduğundan, aşırı politika sapmasını önlemek için mevcut politikadan gradyanı ayırıyoruz: $w = \text{detach}[\pi_\theta(\mathbf{y} \mid \mathbf{x})]/\pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})$. Ek stabilite için, önem ağırlıklarını sınırlandırmak amacıyla PPO tarzı clipping uyguluyoruz: $w = \text{clip}\left(\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\text{old}}(\mathbf{y}|\mathbf{x})}, 1-\epsilon, 1+\epsilon\right)^{\text{detach}}$.

[2] Yerine koyma, Yörünge Dengelemesi amaç fonksiyonu Denklemi 3'teki $\beta\beta(\mathbf{x}, \mathbf{y})$ ifadesini, $\beta\beta(\mathbf{x}, \mathbf{y}) + \text{logaritma}\, \pi$ referans $(\mathbf{y} \mid \mathbf{x})$ ifadesi ile değiştirerek referans modeli kısıtını entegre eder.

Incorporating these improvements into Eq. 5, we arrive at the following FlowRL objective:

FlowRL

$$\mathcal{L}_{\text{FlowRL}} = w \cdot \left(\log Z_\phi(\mathbf{x}) + \frac{1}{|\mathbf{y}|}\log \pi_\theta(\mathbf{y} \mid \mathbf{x}) - \beta \hat{r}(\mathbf{x}, \mathbf{y}) - \frac{1}{|\mathbf{y}|}\log \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\right)^2 \tag{6}$$

where the clipped importance weight w and normalized reward $\hat{r}(\mathbf{x}, \mathbf{y})$ are defined as:

$$w = \text{clip}\left(\frac{\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})}, 1-\epsilon, 1+\epsilon\right)^{\text{detach}}, \quad \hat{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \tag{7}$$

We use this objective to update the policy parameters θ during training, and refer to this strategy as *FlowRL*. Implementation details and theoretical analysis are provided in § 5 and § B, respectively.

4. Related Work

4.1. Reinforcement Learning for Reasoning

Reinforcement learning has emerged as a powerful approach for large language models post-training on reasoning tasks [Guo et al., 2025, Lightman et al., 2023b, Schulman et al., 2017, Shao et al., 2024, Sutton et al., 1999b]. Most approaches employ reward-maximizing RL to optimize expected cumulative returns. Entropy regularization [Ahmed et al., 2019, Cheng et al., 2025, Haarnoja et al., 2018] is a classical technique for mitigating mode collapse by promoting diversity in the policy's output distribution, and has also been shown to enhance reasoning capabilities in various settings [Chao et al., 2024, Eysenbach and Levine, 2021]. However, for long CoT reasoning, the extended trajectory length (e.g., 8k–16k tokens) makes it difficult for the regularization signal to effectively influence reward-maximizing learning. Recent work [Cheng et al., 2025, Cui et al., 2025, Dong et al., 2025, Wang et al., 2025] has discovered that training with more diverse or high-entropy training data can further enhance training effectiveness. Compared to traditional entropy regularization, the above methods explicitly increase the proportion of low-probability (i.e., high-entropy) tokens in the training data. In our work, we address the mode-collapse problem by fundamentally shifting from reward maximization to reward distribution matching in our RL formulation.

4.2. GFlowNets

GFlowNets [Bengio et al., 2023a] represent a class of diversity-driven algorithms designed to balance probability flows across states. They have rich connections to probabilistic modeling methods [Ma et al., Malkin et al., 2023, Zhang et al., 2022a,b, 2024a, Zimmermann et al., 2022], and control methods [Pan et al., 2023b,c,d, Tiapkin et al., 2024, Zhang et al., 2024b]. This advantage has enabled GFlowNets to achieve successful applications in multiple downstream tasks, such as molecular drug discovery [Jain et al., 2022, 2023a,b, Kim et al., 2023, 2024, Liu et al., 2022, Pan et al., 2023a, Shen et al., 2023], phylogenetic inference [Zhou et al., 2024], and combinatorial optimization [Zhang et al., 2023a,b]. For generative AI, GFlowNets provide a powerful approach to align pretrained models in scenarios such as image generation [Yun et al., 2025, Zhang et al., 2025a] and language model fine-tuning [Hu et al., 2024, Lee et al., 2024, Yu et al., 2025a]. Another line of work primarily focuses on the theoretical aspects of GFlowNets. Recent theoretical studies have interpreted GFlowNets as solving a maximum entropy reinforcement learning problem within a modified Markov Decision Process (MDP) [Deleu et al., 2024, Mohammadpour et al., 2024, Tiapkin et al., 2024]. These theoretical contributions have

Bu iyileştirmeleri Eq. 5'e entegre ettiğimizde, aşağıdaki FlowRL amaç fonksiyonuna ulaşırız:

FlowRL

$$\mathcal{L}_{\text{FlowRL}} = w \cdot \left(\log Z_\phi(\mathbf{x}) + \frac{1}{|\mathbf{y}|}\log \pi_\theta(\mathbf{y} \mid \mathbf{x}) - \beta \hat{r}(\mathbf{x}, \mathbf{y}) - \frac{1}{|\mathbf{y}|}\log \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\right)^2 \tag{6}$$

Burada kesilmiş önem ağırlığı w ve normalize edilmiş ödül $\hat{r}(\mathbf{x}, \mathbf{y})$ şu şekilde tanımlanır:

$$w = \text{clip}\left(\frac{\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{old}}(\mathbf{y} \mid \mathbf{x})}, 1-\epsilon, 1+\epsilon\right)^{\text{detach}}, \quad \hat{r}_i = \frac{r_i - \text{ortalama}(\mathbf{r})}{\text{standart sapma}(\mathbf{r})} \tag{7}$$

Bu amaç fonksiyonunu eğitim sırasında politika parametreleri θ nin güncellenmesi için kullanıyoruz ve bu stratejiye *FlowRL* adını veriyoruz. Uygulama detayları ve teorik analiz sırasıyla § 5 ve § B bölümlerinde sunulmuştur.

4. İlgili Çalışmalar

4.1. Muhakeme İçin Takviyeli Öğrenme

Takviyeli öğrenme, muhakeme görevlerinde büyük dil modelleri için eğitim sonrası güçlü bir yöntem olarak ortaya çıkmıştır [Guo et al., 2025, Lightman et al., 2023b, Schulman et al., 2017, Shao et al., 2024, Sutton et al., 1999b]. Çoğu yaklaşım, beklenen kümülatif getirileri optimize etmek için ödül-maksimize eden pekiştirmeli öğrenme kullanmaktadır. Entropi düzenlemesi [Ahmed et al., 2019, Cheng et al., 2025, Haarnoja et al., 2018], politikanın çıktı dağılımında çeşitliliği teşvik ederek mod çöküşünü azaltmaya yönelik klasik bir tekniktir ve çeşitli ortamlarda muhakeme yeteneklerini geliştirdiği de gösterilmiştir [Chao et al., 2024, Eysenbach ve Levine, 2021]. Ancak, uzun CoT muhakemesi için uzatılmış iz uzunluğu (örneğin, 8k–16k token) nedeniyle, düzenleme sinyalinin ödül-maksimize eden öğrenme sürecini etkili biçimde yönlendirmesi güçleşmektedir. Son çalışmalar [Cheng et al., 2025, Cui et al., 2025, Dong et al., 2025, Wang et al., 2025], daha çeşitli veya yüksek entropili eğitim verileriyle eğitimin etkinliğinin artırılabileceğini ortaya koymuştur. Geleneksel entropi düzenlemesine kıyasla, bu yöntemler eğitim verilerindeki düşük olasılıklı (yani yüksek entropili) tokenların oranını açıkça artırmaktadır. Çalışmamızda , RL formülasyonumuzda mod çöküşü sorununu, ödül maksimize etmekten ödül dağılımı eşleştirmesine temel bir kayma yaparak ele alıyoruz.

4.2. GFlowNets

GFlowNets [Bengio ve ark., 2023a], durumlar arasındaki olasılık akışlarını dengelemeyi amaçlayan çeşitlilik odaklı algoritmalar sınıfını temsil eder. Bu yöntemler, olasılıksal modelleme teknikleriyle [Ma ve ark., Malkin ve ark., 2023, Zhang ve ark., 2022a,b, 2024a, Zimmermann ve ark., 2022] ve kontrol yöntemleriyle [Pan ve ark., 2023b,c,d, Tiapkin ve ark., 2024, Zhang ve ark., 2024b] zengin bağlantılara sahiptir. Bu avantaj, GFlowNets'in moleküler ilaç keşfi [Jain ve ark., 2022, 2023a,b, Kim ve ark., 2023, 2024, Liu ve ark., 2022, Pan ve ark., 2023a, Shen ve ark., 2023], filogenetik çıkarım [Zhou ve ark., 2024] ve kombinatoryal optimizasyon [Zhang ve ark., 2023a,b] gibi çeşitli alt görevlerde başarılı uygulamalar gerçekleştirmesini sağlamıştır.
Üretici yapay zeka için, GFlowNets, önceden eğitilmiş modelleri görüntü üretimi [Yun ve ark., 2025, Zhang ve ark., 2025a] ile dil modeli ince ayarı [Hu ve ark., 2024, Lee ve ark., 2024, Yu ve ark., 2025a] gibi senaryolarda hizalamak için güçlü bir yaklaşım sunmaktadır. Başka bir çalışma alanı, öncelikle GFlowNets'in teorik boyutlarına odaklanmaktadır. Son teorik araştırmalar, GFlowNets'i modifiye edilmiş bir Markov Karar Süreci (MDP) kapsamında maksimum entropi takviyeli öğrenme problemi olarak yorumlamıştır [Deleu ve ark., 2024, Mohammadpour ve ark., 2024, Tiapkin ve ark., 2024]. Bu teorik katkılar

inspired us to enhance reinforcement learning from a more foundational standpoint using GFlowNets principles. A comprehensive overview of GFlowNets theory can be found in Appendix C.

4.3. Flow-Matching Policies

Flow matching simplifies diffusion-based approaches by learning vector fields that transport samples from prior to target distributions [Lipman et al., 2023]. Recent work has explored flow matching for policy optimization. McAllister et al. [2025] reformulates policy optimization using advantage-weighted ratios from conditional flow matching loss, enabling flow-based policy training without expensive likelihood computations. Pfrommer et al. [2025] explored reward-weighted flow matching for improving policies beyond demonstration performance. Park et al. [2025] uses a separate one-step policy to avoid unstable backpropagation through time when training flow policies with RL. Zhang et al. [2025a] proposed a combined loss function integrating PPO and GFlowNets to optimize diffusion model alignment. However, these approaches focus on continuous control, image generation, or vision-action models, rather than addressing mode-collapse limitations in reward-maximizing RL. Inspired by flow matching principles, our work improves upon RL training to enhance training stability while promoting diverse solution exploration.

5. Experiment Settings

Backbone Models. There are two learnable modules in Eq. 6: the policy model π_θ and the partition function Z_ϕ. For the policy model π_θ, we use `Qwen-2.5-7B/32B` [Team, 2024] for math tasks and `DeepSeek-R1-Distill-Qwen-7B` [DeepSeek-AI, 2025] for code tasks, respectively. For partition function Z_ϕ, following Lee et al. [2024], we use a randomly initialized 3-layer MLP with hidden dimensions matching those of the base model. The reference model π_{ref} is the corresponding fixed pretrained model. All training scripts are based on the veRL [Sheng et al., 2024]. For the reward function, following Lee et al. [2024], we set the hyperparameter $\beta = 15$.

Baselines. We compare our method against three representative reward-maximization RL baselines: REINFORCE++ (R++; Hu et al., 2025, Sutton et al., 1999b), PPO [Schulman et al., 2017], and GRPO [Shao et al., 2024]. All baselines follow the official veRL recipes, with consistent training configurations. For fair comparison, all methods use the same learning rate, batch size, and training steps, and are evaluated at convergence using identical step counts.

Training Configuration. We experiment on both math and code domains. For the math domain, we use the training set collected from DAPO [Yu et al., 2025b]. For the code domain, we follow the setup of DeepCoder [Luo et al., 2025], using their training set. For 7B model training, we use a single node equipped with 8 NVIDIA H800 GPUs (80GB memory each). For 32B model training, we scale to 4 nodes with 32 GPUs to accommodate the larger memory requirements. All experiments use `max_prompt_length` = 2048 and `max_response_length` = 8192 across both model sizes. We use a batch size of 512 for math reasoning tasks and 64 for code reasoning tasks. We set the learning rate to 1e-6 and enable dynamic batch sizing in veRL for efficient training. For GRPO and FlowRL, we configure `rollout_n` = 8, meaning each prompt generates 8 response rollouts as the group size.

Evaluation Configuration. For the math domain, we evaluate on six challenging benchmarks: AIME 2024/2025 [MAA, 2025], AMC 2023 [MAA, 2023], MATH-500 [Lightman et al., 2023a], Minerva [Lewkowycz et al., 2022], and Olympiad [He et al., 2024]. For the code domain, we evaluate on LiveCodeBench [Jain et al., 2024], CodeForces [Penedo et al., 2025], and HumanEval+ [Chen et al., 2021]. For all evaluation datasets, we perform 16 rollouts and report the average accuracy, denoted as Avg@16. We further report rating and percentile for Codeforces. During generation, we

GFlowNets prensipleri kullanılarak takviyeli öğrenmenin daha temel bir bakış açısından geliştirilebileceği konusunda bizlere ilham vermiştir. GFlowNets teorisine dair kapsamlı bir genel bakış Ek C'de sunulmaktadır.

4.3. Akış-Eşleştirme Politikaları

Akış eşleştirme, ön dağılımdan hedef dağılıma örnekleri taşıyan vektör alanlarını öğrenerek difüzyon tabanlı yöntemleri basitleştirmektedir [Lipman ve ark., 2023]. Son çalışmalar, politika optimizasyonu için akış eşleştirmeyi araştırmıştır. McAllister ve ark. [2025], koşullu akış eşleştirme kaybından avantaj-ağırlıklı oranları kullanarak politika optimizasyonunu tekrar formüle etmiş ve akış tabanlı politika eğitimini maliyetli olasılık hesaplamaları olmadan mümkün kılmıştır. Pfrommer ve ark. [2025], gösterim performansını aşan politikaları geliştirmek amacıyla ödül-ağırlıklı akış eşleştirmeyi incelemiştir. Park ve ark. [2025], akış politikalarını PL ile eğitirken zaman içinde kararsız geriye yayılımı önlemek için ayrı bir tek adımlı politika kullanmıştır. Zhang ve ark. [2025a], diffusion model hizalamasını optimize etmek üzere PPO ve GFlowNets'i entegre eden birleşik bir kayıp fonksiyonu önermiştir. Bununla birlikte, bu yaklaşımlar, ödül-maksimize eden pekiştirmeli öğrenmedeki mod çöküşü sınırlamalarını ele almak yerine sürekli kontrol, görüntü üretimi veya görsel-eylem modellerine odaklanmaktadır. Akış eşleştirme prensiplerinden ilham alarak, çalışmamız çeşitli çözümlerin keşfini teşvik ederken eğitim kararlılığını artırmak için PL eğitimini geliştirmektedir.

5. Deney Ayarları

Temel Modeller. Eşitlik 6'da iki öğrenilebilir modül bulunmaktadır: politika modeli π_θ ve bölme fonksiyonu Z_ϕ. Politika modeli π_θ için, matematik görevlerinde Qwen-2.5-7B/32B [Team, 2024], kod görevlerinde ise DeepSeek-R1-Distill-Qwen-7B [DeepSeek-AI, 2025] kullanılmıştır. Bölme fonksiyonu Z_ϕ için, Lee ve ark. [2024]'e uygun olarak, temel modelin gizli boyutlarına sahip rastgele başlatılmış üç katmanlı bir MLP kullanılmıştır. Referans modeli π_{ref}, ilgili sabit önceden eğitilmiş modeldir. Tüm eğitim betikleri veRL [Sheng ve ark., 2024] temel alınarak hazırlanmıştır. Ödül fonksiyonu için, Lee ve ark. [2024]'e uygun olarak hiperparametre $\beta = 15$ olarak belirlenmiştir.

Temel ModellerYöntemimiz, üç temsilî ödül-maksimizasyon RL temel modeli ile karşılaştırılmıştır: REINFORCE++ (R++; Hu ve ark., 2025, Sutton ve ark., 1999b), PPO [Schulman ve ark., 2017] ve GRPO [Shao ve ark., 2024]. Tüm temel modeller resmi veRL tariflerine uyar ve tutarlı eğitim konfigürasyonları kullanır. Adil karşılaştırma için tüm yöntemler aynı öğrenme oranını, batch boyutunu ve eğitim adımlarını kullanır ve yakınsama aşamasında aynı adım sayılarıyla değerlendirilir.

Eğitim Konfigürasyonu. Hem matematik hem de kod alanlarında deneyler gerçekleştirilmiştir. Matematik alanı için DAPO [Yu et al., 2025b] tarafından toplanan eğitim seti kullanılmıştır. Kod alanı için DeepCoder [Luo et al., 2025] yapılandırması takip edilerek onların eğitim seti kullanılmıştır. 7B model eğitimi için her biri 80GB belleğe sahip 8 NVIDIA H800 GPU bulunan tek bir düğüm kullanılmıştır. 32B model eğitimi için, bellek gereksinimlerini karşılamak amacıyla 4 düğüm ve 32 GPU'ya ölçeklendirme yapılmıştır. Tüm deneylerde her iki model boyutu için max_prompt_length = 2048 ve max_response_length = 8192 değeri kullanılmıştır. Matematik muhakeme görevleri için batch boyutu 512, kod muhakeme görevleri için ise 64 olarak belirlenmiştir. Etkili eğitim için veRL'de öğrenme oranını 1e-6 olarak belirledik ve dinamik batch boyutlandırmayı etkinleştirdik. GRPO ve FlowRL için rollout_n = 8 olarak yapılandırdık; bu, her promptun grup büyüklüğü olarak 8 yanıt denemesi ürettiği anlamına gelir.

Değerlendirme Konfigürasyonu. Matematik alanında, altı zorlu benchmark üzerinde değerlendirme gerçekleştirdik: AIME 2024/2025 [MAA, 2025], AMC 2023 [MAA, 2023], MATH-500 [Lightman ve ark., 2023a], Minerva [Lewkowycz ve ark., 2022] ve Olympiad [He ve ark., 2024]. Kod alanında ise LiveCodeBench [Jain ve ark., 2024], CodeForces [Penedo ve ark., 2025] ve HumanEval+ [Chen ve ark., 2021] benchmarkları üzerinde değerlendirme yaptık. Tüm değerlendirme veri setlerinde, 16 deneme gerçekleştirerek ortalama doğruluğu, Ortalama@16 olarak raporluyoruz. Codeforces için ayrıca rating ve yüzdelik dilimi raporluyoruz. Üretim sırasında

	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg
Qwen2.5-32B-Base, Max Response Len=8K							
Backbone	4.6	2.1	28.6	52.5	27.0	21.4	22.7
R++	14.8$_{+10.2}$	9.2$_{+7.1}$	52.7$_{+24.1}$	44.4$_{-8.1}$	17.4$_{-9.6}$	24.5$_{+3.1}$	27.1
PPO	26.9$_{+22.3}$	20.4$_{+18.3}$	76.4$_{+47.8}$	69.2$_{+16.7}$	28.8$_{+1.8}$	37.9$_{+16.5}$	43.3
GRPO	23.1$_{+18.5}$	14.6$_{+12.5}$	76.9$_{+48.3}$	61.6$_{+9.1}$	19.0$_{-8.0}$	34.9$_{+13.5}$	38.3
FlowRL	24.0$_{+19.4}$	21.9$_{+19.8}$	73.8$_{+45.2}$	80.8$_{+28.3}$	38.2$_{+11.2}$	51.8$_{+30.4}$	**48.4**
Qwen2.5-7B-Base, Max Response Len=8K							
Backbone	4.4	2.1	30.8	54.5	22.4	24.0	23.0
R++	11.0$_{+6.6}$	5.4$_{+3.3}$	66.7$_{+35.9}$	54.3$_{-0.2}$	24.4$_{+2.0}$	27.3$_{+3.3}$	31.5
PPO	9.4$_{+5.0}$	7.3$_{+5.2}$	63.4$_{+32.6}$	58.0$_{+3.5}$	26.5$_{+4.1}$	27.3$_{+3.3}$	32.0
GRPO	13.5$_{+9.1}$	9.8$_{+7.7}$	64.5$_{+33.7}$	57.1$_{+2.6}$	23.1$_{+0.7}$	26.9$_{+2.9}$	32.5
FlowRL	15.4$_{+11.0}$	10.8$_{+8.7}$	54.5$_{+23.7}$	67.0$_{+12.5}$	31.4$_{+9.0}$	34.6$_{+10.6}$	**35.6**

Table 1 | Results on math benchmarks. We report Avg@16 accuracy with relative improvements shown as subscripts. Positive gains are shown in green and negative changes in red. FlowRL outperforms all baselines across both 7B and 32B model scales.

Models	LiveCodeBench		CodeForces		HumanEval+
	Avg@16	Pass@16	Rating	Percentile	Avg@16
DeepSeek-R1-Distill-Qwen-7B, Max Response Len=8K					
Backbone	30.7	49.5	886.7	19.4	80.9
R++	30.5$_{-0.2}$	52.7$_{+3.2}$	1208.0$_{+321.3}$	56.8$_{+37.4}$	76.6$_{-4.3}$
PPO	35.1$_{+4.4}$	54.5$_{+5.0}$	1403.1$_{+516.4}$	73.7$_{+54.3}$	82.3$_{+1.4}$
GRPO	32.8$_{+2.1}$	52.3$_{+2.8}$	1313.8$_{+427.1}$	67.1$_{+47.7}$	80.1$_{-0.8}$
FlowRL	**37.4**$_{+6.7}$	**56.3**$_{+6.8}$	**1549.5**$_{+662.8}$	**83.3**$_{+63.9}$	**83.3**$_{+2.4}$

Table 2 | Results on code benchmarks. We report metrics with relative improvements shown as subscripts. Positive gains are shown in green and negative changes in red. FlowRL achieves the strongest performance across all three benchmarks, demonstrating its effectiveness in code reasoning tasks.

use sampling parameters of `temperature` = 0.6 and `top_p` = 0.95 for all evaluations. The response length for evaluation is set to 8,192, consistent with the training configuration.

6. Results

6.1. Main Results

Our experimental results, summarized in Table 1 and Table 2, demonstrate that FlowRL consistently outperforms all reward-maximization baselines across both math and code reasoning domains. Table 1 reports results on math reasoning benchmarks using both 7B and 32B base models, while Table 2 presents the corresponding results on code reasoning tasks. On math reasoning tasks, FlowRL achieves the highest average accuracy of 35.6% with the 7B model and 48.4% with the 32B model, surpassing PPO by 5.1% and GRPO by 10.1% on the 32B model. FlowRL shows strong improvements on challenging benchmarks like MATH-500 and Olympiad problems, demonstrating consistent gains

	AIME24	AIME25	AMC23	MATH500	Minerva	Olimpiyatı	Ortalama
Qwen2.5-32B-Base, Maksimum Yanıt Uzunluğu=8K							
Omurga	4.6	2.1	28.6	52.5	27.0	21.4	22.7
R++	14.8$_{+10.2}$	9.2$_{+7.1}$	52.7$_{+24.1}$	44.4$_{-8.1}$	17.4$_{-9.6}$	24.5$_{+3.1}$	27.1
PPO	26.9$_{+22.3}$	20.4$_{+18.3}$	76.4$_{+47.8}$	69.2$_{+16.7}$	28.8$_{+1.8}$	37.9$_{+16.5}$	43.3
GRPO	23.1$_{+18.5}$	14.6$_{+12.5}$	76.9$_{+48.3}$	61.6$_{+9.1}$	19.0$_{-8.0}$	34.9$_{+13.5}$	38.3
FlowRL	24.0$_{+19.4}$	21.9$_{+19.8}$	73.8$_{+45.2}$	80.8$_{+28.3}$	38.2$_{+11.2}$	51.8$_{+30.4}$	**48.4**
Qwen2.5-7B-Base, Maksimum Yanıt Uzunluğu=8K							
Omurga	4.4	2.1	30.8	54.5	22.4	24.0	23.0
R++	11.0$_{+6.6}$	5.4$_{+3.3}$	66.7$_{+35.9}$	54.3$_{-0.2}$	24.4$_{+2.0}$	27.3$_{+3.3}$	31.5
PPO	9.4$_{+5.0}$	7.3$_{+5.2}$	63.4$_{+32.6}$	58.0$_{+3.5}$	26.5$_{+4.1}$	27.3$_{+3.3}$	32.0
GRPO	13.5$_{+9.1}$	9.8$_{+7.7}$	64.5$_{+33.7}$	57.1$_{+2.6}$	23.1$_{+0.7}$	26.9$_{+2.9}$	32.5
FlowRL	15.4$_{+11.0}$	10.8$_{+8.7}$	54.5$_{+23.7}$	67.0$_{+12.5}$	31.4$_{+9.0}$	34.6$_{+10.6}$	**35.6**

Tablo 1 | **Matematik kıyaslama setlerindeki sonuçlar.** Alt simgeler halinde göreli iyileştirmelerle birlikte Ortalama@16 doğruluğu rapor edilmiştir. Pozitif kazançlar yeşil, negatif değişiklikler kırmızı renkte gösterilmiştir. FlowRL, hem 7B hem de 32B model ölçeklerinde tüm temel modelleri geride bırakmaktadır.

Modeller	LiveCodeBench		CodeForces		HumanEval+
	Ortalama@16	Geçme@16	Puan	Yüzdelik Dilim	Avg@16
DeepSeek-R1-Distill-Qwen-7B, Maksimum Yanıt Uzunluğu=8K					
Omurga	30.7	49.5	886.7	19.4	80.9
R++	30.5$_{-0.2}$	52.7$_{+3.2}$	1208.0$_{+321.3}$	56.8$_{+37.4}$	76.6$_{-4.3}$
PPO	35.1$_{+4.4}$	54.5$_{+5.0}$	1403.1$_{+516.4}$	73.7$_{+54.3}$	82.3$_{+1.4}$
GRPO	32.8$_{+2.1}$	52.3$_{+2.8}$	1313.8$_{+427.1}$	67.1$_{+47.7}$	80.1$_{-0.8}$
FlowRL	**37.4**$_{+6.7}$	**56.3**$_{+6.8}$	**1549.5**$_{+662.8}$	**83.3**$_{+63.9}$	**83.3**$_{+2.4}$

Tablo 2 | **Kod benchmarklarındaki sonuçlar.** Göreli iyileşmeler alt simge olarak gösterilen metrikler rapor edilmiştir. Pozitif kazanımlar yeşil, negatif değişiklikler kırmızı ile gösterilmiştir. FlowRL, kod muhakeme görevlerinde etkinliğini ispatlayarak üç benchmarkta da en yüksek performansı sağlamaktadır.

Tüm değerlendirmeler için örnekleme parametreleri olarak `temperature` = 0.6 ve `top_p` = 0.95 kullanılmıştır. Değerlendirme yanıt uzunluğu, eğitim konfigürasyonuna uygun olarak 8.192 olarak ayarlanmıştır.

6. Sonuçlar

6.1. Ana Sonuçlar

Tablo 1 ve Tablo 2'de özetlenen deneysel sonuçlarımız, FlowRL'nin hem matematik hem de kod muhakeme alanlarında tüm ödül maksimize etme temel modellerinden sürekli olarak üstün performans gösterdiğini ortaya koymaktadır. Tablo 1, 7B ve 32B tabanlı modeller kullanılarak matematik muhakeme benchmarklarındaki sonuçları rapor ederken; Tablo 2, karşılık gelen kod muhakeme görevlerindeki sonuçları sunmaktadır. Matematik muhakeme görevlerinde, FlowRL 7B model ile %35,6 ve 32B model ile %48,4 ortalama doğruluk elde etmiş, 32B modelde PPO'yu %5,1 ve GRPO'yu %10,1 oranında geride bırakmıştır. FlowRL, MATH-500 ve Olimpiyat problemleri gibi zorlu kıyaslama setlerinde tutarlı biçimde güçlü iyileşmeler sergilemektedir.

Method	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olympiad	Avg
FlowRL	15.41	10.83	54.53	66.96	31.41	34.61	35.63
w/o IS	6.25	7.91	41.40	56.97	22.19	25.52	26.71
Zhang et al. [2025a]	10.41	6.66	53.75	66.50	30.97	33.72	33.67

Table 3 | Ablation study on FlowRL with Qwen2.5-7B as the base model. Avg@16 accuracy is reported across six math reasoning benchmarks. IS denotes importance sampling.

across diverse mathematical domains. On code generation tasks, FlowRL achieves compelling improvements with the highest Avg@16 score of 37.43% on LiveCodeBench, a Codeforces rating of 1549.47 with 83.3% percentile ranking, and 83.28% accuracy on HumanEval+, outperforming all baselines across the board. These consistent performance gains across both domains and model scales provide strong empirical evidence that FlowRL's flow-balanced optimization successfully enhances generalization. This improvement comes from promoting diverse solution exploration compared to previous reward-maximizing RL approaches.

6.2. Ablation Studies

We conduct ablation studies on importance sampling and the β hyperparameter. For importance sampling, we compared the performance with and without it, and implemented a combined loss approach proposed by Zhang et al. [2025a] that simultaneously optimizes both GFlowNets and PPO objectives. This combined loss focuses on optimizing diffusion models, and we adapt it to long CoT reasoning tasks for comparison. Table 3 demonstrates that importance sampling substantially improves FlowRL performance across all math reasoning benchmarks. Compared to Zhang et al. [2025a], using importance sampling as a trajectory-level ratio is more suitable than the combined loss of GFlowNets and PPO. The performance drop without importance sampling (from 35.63% to 26.71%) highlights the critical role of correcting for distribution mismatch between rollout generation and policy training. For the hyperparameter β, we conduct a series of parameter ablation studies, and Figure 3 shows that $\beta = 15$ achieves optimal performance, with detailed results shown in Table 7.

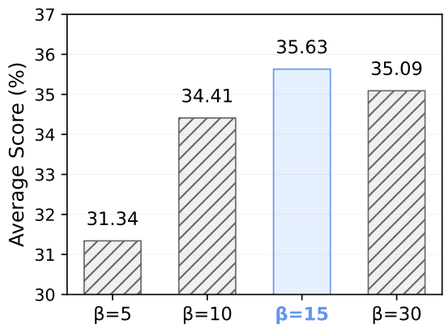


Figure 3 | Ablation study on the β in FlowRL. $\beta = 15$ (highlighted in blue) achieves the best performance.

7. Analysis

7.1. Diversity Analysis

To assess solution diversity, we follow the approach of Yu et al. [2025a] and employ `GPT-4o-mini` [OpenAI, 2024] to evaluate all responses generated by each method on AIME 24/25. The evaluation prompt is shown in Appendix C. As shown in Figure 4, FlowRL achieves higher diversity scores compared to baseline methods. This demonstrates that FlowRL improves sample diversity compared to baselines, which tend to exhibit repetitive solution patterns. This diversity evaluation reveals

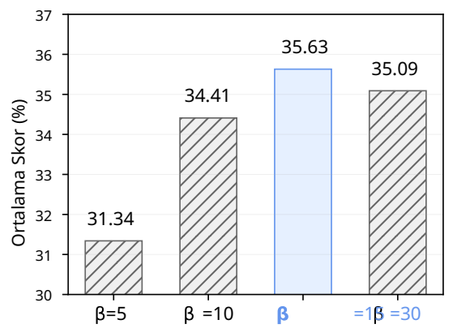
Yöntem	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olimpiyat	Ortalama
FlowRL	15.41	10.83	54.53	66.96	31.41	34.61	35.63
önemli örnekleme olmadan	6.25	7.91	41.40	56.97	22.19	25.52	26.71
Zhang ve ark. [2025a]	10.41	6.66	53.75	66.50	30.97	33.72	33.67

Tablo 3 | Qwen2.5-7B temel modeli kullanılarak FlowRL üzerine yapılan ayrıştırma çalışması. Altı matematiksel muhakeme kıyaslama setinde Ortalama@16 doğruluğu rapor edilmiştir. IS, önemli örneklemeyi ifade eder.

çeşitli matematiksel alanlar genelinde. Kod üretimi görevlerinde FlowRL, LiveCodeBench üzerinde %37.43 ile en yüksek Ortalama@16 skorunu, %83.3 yüzdelik dilimde 1549.47 puanla Codeforces derecelendirmesini ve HumanEval+ üzerinde %83.28 doğruluğu elde ederek tüm temel modelleri açık ara geride bırakmaktadır. Hem alanlar hem de model ölçekleri genelinde elde edilen bu tutarlı performans artışları, FlowRL'nin akış -dengeli optimizasyonunun genelleme yeteneğini başarıyla artırdığına dair güçlü ampirik kanıtlar sunmaktadır. Bu gelişme, önceki ödül-maksimize eden pekiştirmeli öğrenme yaklaşımlarına kıyasla çeşitli çözüm keşiflerinin teşvik edilmesinden kaynaklanmaktadır.

6.2. Ayrıştırma Çalışmaları

Önemli örnekleme ve β hiperparametresi üzerinde ayrıştırma çalışmaları gerçekleştirdik. Önemli örnekleme için performansı bununla ve onsuz karşılaştırdık ve Zhang et al. [2025a] tarafından önerilen, GFlowNets ile PPO amaçlarını eşzamanlı optimize eden birleşik kayıp yaklaşımını uyguladık. Bu birleşik kayıp difüzyon modellerinin optimizasyonuna odaklanmakta olup karşılaştırma amacıyla uzun CoT muhakemesi görevlerine uyarladık. Tablo 3, önemli örneklemenin tüm matematik muhakemesi ölçütlerinde FlowRL performansını kayda değer biçimde artırdığını göstermektedir. Zhang et al. [2025a] ile karşılaştırıldığında, GFlowNets ve PPO'nun birleşik kaybından ziyade trajektori seviyesinde oran olarak önemli örnekleme kullanmak daha uygundur. Önemli örnekleme olmadan performans düşüşü (yüzde 35,63'ten yüzde 26,71'e) rollout üretimi ile politika eğitimi arasındaki dağılım uyumsuzluğunun düzeltilmesinin kritik önemini vurgulamaktadır. Hiperparametre için— β parametresi için bir dizi parametre kazıma çalışması gerçekleştirdik ve Şekil 3, optimal performansa $\beta = 15$ ile ulaşıldığını göstermektedir; ayrıntılı sonuçlar Tablo 7'de sunulmuştur.



Şekil 3 |FlowRL'deki βüzerine ayrıştırma çalışması. $\beta = 15$ (mavi ile vurgulanmıştır) en iyi performansı sağlamaktadır.

7. Analiz

7.1. Çeşitlilik Analizi

Çözüm çeşitliliğini değerlendirmek için Yu ve ark. [2025a] yöntemini takip ederek AIME 24/25 üzerindeki her yöntemin ürettiği tüm cevapları değerlendirmek üzere GPT-4o-mini [OpenAI, 2024] kullanıyoruz. Değerlendirme istemi Ek C'de gösterilmiştir. Şekil 4'te görüldüğü üzere, FlowRL temel modellere kıyasla daha yüksek çeşitlilik skorları elde etmektedir. Bu durum FlowRL'nin, yinelenen çözüm desenleri sergileme eğilimindeki temel modellere kıyasla örnek çeşitliliğini artırdığını göstermektedir. Bu çeşitlilik değerlendirmesi ortaya koymaktadır ki

Table 4 | Case study comparing GRPO and FlowRL rollouts on an AIME problem. GRPO exhibits repetitive patterns (AM-GM ×3, identity loops ×2), while FlowRL follows a more diverse solution path.

	Content (boxed = actions; "×k" = repeated; "…" = omitted)
Question	Let $\mathcal{B}$ be the set of rectangular boxes with surface area 54 and volume 23. Let r be the radius of the smallest sphere that can contain each box in $\mathcal{B}$. If $r^2 = \frac{p}{q}$ with $\gcd(p,q) = 1$, find $p+q$.
GRPO	"… denote a,b,c … $2(ab+bc+ca) = 54$, $abc = 23$ … $d = \sqrt{a^2+b^2+c^2}$, $r = d/2$ … $(a+b+c)^2 = a^2+b^2+c^2+2(ab+bc+ca)$ … AM–GM ×3: AM–GM (1) … AM–GM (2) … AM–GM (3) … $(a+b+c)^3$ identity loop ×2: loop (1) … loop (2) … $a = b = c$ (contradiction) … back to $(a+b+c)^2$ … no factorization …"
FlowRL	"… let a,b,c with $2(ab+bc+ca) = 54$, $abc = 23$ … $d = \sqrt{a^2+b^2+c^2}$, $r = d/2$ … $(a+b+c)^2 \Rightarrow a^2+b^2+c^2 = s^2-54$ … $a = b$ … $a^3-27a+46 = 0$ … rational root $a = 2$ … factor $(a-2)(a^2+2a-23)$ … branch $a = -1+2\sqrt{6}$ … back-sub $c = 23/a^2$ … $a^2+b^2+c^2 = \frac{657}{16}$ … $r^2 = \frac{657}{64}$ … Answer 721 …"

significant differences in exploration patterns across methods. This nearly doubling of diversity score compared to the strongest baseline (PPO) indicates that FlowRL generates qualitatively different solution approaches rather than minor variations of the same strategy. The diversity analysis provides empirical validation of our core hypothesis that flow-balanced optimization promotes mode coverage in complex reasoning tasks.

7.2. Case Study

Table 4 illustrates the behavioral differences between GRPO and FlowRL on a representative AIME problem. GRPO exhibits repetitive patterns, applying AM-GM three times and getting stuck in identity loops, failing to solve the problem. FlowRL explores more diverse actions: it sets $a = b$, derives a cubic equation, finds the rational root, and reaches the correct answer. This shows that FlowRL successfully avoids the repetitive exploration patterns. The contrast reveals fundamental differences in exploration strategies: GRPO's reward-maximizing approach leads to exploitation of familiar techniques (AM-GM inequality) without exploring alternatives, eventually reaching contradictory conclusions like $a = b = c$. In contrast, FlowRL's distribution-matching enables strategic decisions such as the symmetry assumption $a = b$, which transforms the problem into a tractable cubic equation $a^3 - 27a + 46 = 0$, allowing systematic solution through rational root testing and polynomial factorization.

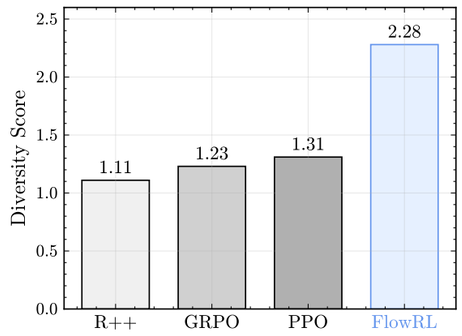


Figure 4 | GPT-judged diversity scores on rollouts of AIME 24/25 problems. FlowRL generates more diverse solutions than R++, GRPO, and PPO.

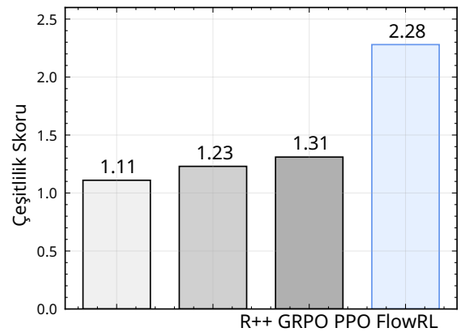
Tablo 4 | GRPO ve FlowRL denemelerinin AIME probleminde karşılaştırıldığı vaka incelemesi. GRPO yinelenen desenler (AM-GM × 3, özdeşlik döngüleri × 2) sergilerken, FlowRL daha çeşitli bir çözüm yolu izlemek tedir.

	İçerik (kutulanmış = eylemler; "×k" = tekrar; ". . . " = atlanmış)
Soru	$\mathcal{B}$ yüzey alanı 54 ve hacmi 23 olan dikdörtgen kutular kümesi olsun. Her kutuyu içerebilen en küçük kürenin yarıçapı r olsun. Eğer $r^2 = p/q$ ve $\gcd(p,q) = 1$ ise, $p+q$'yu bulun.
GRPO	". . . ile gösterilsin a, b, c. . . $2(aa+bb+cc) = 54$, $aaa = 23$ . . . $d = \sqrt{a^2+b^2+c^2}$, $r = d/2$ . . . $(a+b+c)^2 = a^2+b^2+c^2+2(aa+bb+cc)$ . . . AM–GM ×3: AM–GM (1) . . . AM–GM (2) . . . AM–GM (3) . . . $(a+b+c)^3$ özdeşlik döngüsü ×2: döngü (1) . . . döngü (2) . . . $a = b = c$ (çelişki) . . . geri dön $(a+b+c)^2$. . . çarpanlarına ayırma yok . . .
FlowRL	. . . varsayalım ki a, b, c böyleki $2(aa+bb+cc) = 54$, $aaa = 23$ . . . $d = \sqrt{a^2+b^2+c^2}$, $r = d/2$ . . . $(a+b+c)^2 \Rightarrow a^2+b^2+c^2 = s^2-54$ . . . $a = b$ . . . $a^3-27a+46 = 0$ . . . rasyonel kök $a = 2$ . . . çarpanlara ayır $(a-2)(a^2+2a-23)$. dal $a = -1+2\sqrt{6}$ . . . geri yerine koy $c = 23/a^2$ . . . $a^2+b^2+c^2 = \frac{657}{16}$ . . . $r^2 = \frac{657}{64}$ . . . Cevap 721 . . .

Yöntemler arasında keşif örüntülerinde anlamlı farklılıklar mevcuttur. En güçlü temel yöntem olan PPO ile kıyaslandığında, çeşitlilik puanının neredeyse iki katına çıkması, FlowRL'nin aynı stratejinin küçük varyasyonları yerine niteliksel olarak farklı çözüm yaklaşımları ürettiğini göstermektedir. Çeşitlilik analizi, akış-dengeli optimizasyonun karmaşık muhakeme görevlerinde mod kapsamasını teşvik ettiğine dair temel hipotezimizi ampirik olarak doğrulamaktadır.

7.2. Vaka Çalışması

Tablo 4, temsili bir AIME problemi üzerinde GRPO ile FlowRL arasındaki davranış farklarını göstermektedir. GRPO, Üç Üstel Ortalama-Aritmetik Ortalama yöntemini üç kez uygular ve kimlik döngülerine takılarak problemi çözememektedir. FlowRL daha çeşitli eylemleri keşfetmektedir: $a = b$ olarak belirler, kübik denklemi türetir, rasyonel kökü bulur ve doğru sonuca ulaşır. Bu durum, FlowRL'nin tekrarlayan keşif kalıplarından başarıyla kaçındığını göstermektedir. Bu karşıtlık, keşif stratejilerindeki temel farklılıkları ortaya koymaktadır:
GRPO'nun ödül-maksimize etme yaklaşımı, alternatifleri keşfetmeden aşina olunan tekniklerin (AM-GM eşitsizliği) sömürülmesine yol açar ve sonunda çelişkili sonuçlara , örneğin $a = b = c$, ulaşır. Buna karşılık, FlowRL'in dağılım eşleştirmesi, simetri varsayımı $a = b$ gibi stratejik kararların alınmasını sağlar,
bu da problemi çözülebilir üçüncü dereceden denkleme $a^3 - 27a + 46 = 0$ dönüştürür ve rasyonel kök testiyle polinom faktörizasyonu yoluyla sistematik çözüm imkânı sunar.



Şekil 4 | AIME 24/25 problemleri üzerindeki denemelerin GPT tarafından değerlendirilen çeşitlilik skorları. FlowRL, R++, GRPO ve PPO'dan daha çeşitli çözümler üretmektedir .

8. Conclusion

In this work, we introduce FlowRL, which transforms scalar rewards into normalized target distributions using a learnable partition function and minimizes the reverse KL divergence between the policy and target distribution. We demonstrate that this approach is theoretically equivalent to trajectory balance objectives from GFlowNets and implicitly maximizes both reward and entropy, thereby promoting diverse reasoning trajectories. To further address gradient explosion and sampling mismatch issues in long CoT reasoning, we incorporate importance sampling and length normalization. Through experiments on math and code reasoning benchmarks, FlowRL achieves consistent improvements across all tasks compared to GRPO and PPO. Our diversity analysis and case studies confirm that FlowRL generates more varied solution approaches while avoiding repetitive patterns.

Acknowledgments

We are grateful to Mingqian Feng and Yuetai Li for their valuable discussions and feedback, which helped improve the quality of this work.

References

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR, 2019.

Brian R Bartoldson, Siddarth Venkatraman, James Diffenderfer, Moksh Jain, Tal Ben-Nun, Seanie Lee, Minsu Kim, Johan Obando-Ceron, Yoshua Bengio, and Bhavya Kailkhura. Trajectory balance with asynchrony: Decoupling exploration and learning for fast, scalable llm post-training. *arXiv preprint arXiv:2503.18929*, 2025.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Neural Information Processing Systems (NeurIPS)*, 2021.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J. Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023a. URL http://jmlr.org/papers/v24/22-0364.html.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *The Journal of Machine Learning Research*, 24(1):10006–10060, 2023b.

Chen-Hao Chao, Chien Feng, Wei-Fang Sun, Cheng-Kuang Lee, Simon See, and Chun-Yi Lee. Maximum entropy reinforcement learning via energy-based normalizing flow. *arXiv preprint arXiv:2405.13629*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles

8. Sonuç

Bu çalışmada, skalar ödülleri öğrenilebilir bölme fonksiyonu kullanarak normalize edilmiş hedef dağılımlara dönüştüren ve politika ile hedef dağılım arasındaki ters KL sapmasını minimize eden FlowRL'i tanıtıyoruz. Bu yaklaşımın teorik olarak GFlowNets'in yörünge dengelemesi amaçlarıyla eşdeğer olduğunu ve ödül ile entropiyi örtük olarak maksimize ederek çeşitli muhakeme yollarını teşvik ettiğini gösteriyoruz. Uzun CoT muhakemesindeki gradyan patlaması ve örnekleme uyumsuzluğu sorunlarını daha iyi ele almak için önemli örnekleme ve uzunluk normalizasyonunu dahil ediyoruz. Matematik ve kod muhakemesi kıyaslamalarında gerçekleştirilen deneyler sonucunda, FlowRL GRPO ve PPO'ya kıyasla tüm görevlerde tutarlı iyileşmeler göstermektedir. Çeşitlilik analizimiz ve vaka çalışmaları, FlowRL'nin tekrar eden kalıplardan kaçınarak daha çeşitli çözüm yaklaşımları ürettiğini doğrulamaktadır.

Teşekkürler

Bu çalışmanın kalitesini artırmaya katkıda bulunan değerli tartışmalar ve geri bildirimleri için Mingqian Feng ve Yuetai Li'ye teşekkür ederiz.

Kaynaklar

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi ve Dale Schuurmans. Entropinin politika optimizasyonu üzerindeki etkisinin anlaşılması. Uluslararası Makine Öğrenmesi Konferansı'nda, sayfa 151–160. PMLR, 2019.

Brian R Bartoldson, Siddarth Venkatraman, James Diffenderfer, Moksh Jain, Tal Ben-Nun, Seanie Lee, Minsu Kim, Johan Obando-Ceron, Yoshua Bengio ve Bhavya Kailkhura. Asenkronlu yörünge deng elemesi: Hızlı ve ölçeklenebilir LLM sonrası eğitim için keşfi ve öğrenmeyi ayrıştırmak. *arXiv ön baskısı arXiv:2503.18929* , 2025.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup ve Yoshua Bengio. Tekrarsız çeşitli aday üretimi için akış ağı tabanlı üretken modeller. *Neural Information Processing Systems (NeurIPS)* , 2021.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J. Hu, Mo Tiwari ve Emmanuel Bengio. GFlowNet Temelleri. *Journal of Machine Learning Research* , 24(210):1–55, 2023a. URL http://jmlr.org/papers/v24/22-0364.html .

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J. Hu, Mo Tiwari ve Emmanuel Bengio. GFlowNet Temelleri. *The Journal of Machine Learning Research* , 24(1):10006–10060, 2023b.

Chen-Hao Chao, Chien Feng, Wei-Fang Sun, Cheng-Kuang Lee, Simon See ve Chun-Yi Lee. Enerji tabanlı normalleştirici akış ile maksimum entropili takviyeli öğrenme. *arXiv ön baskısı arXiv:2405.13629* , 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Łukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babusch kin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles

Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Miruna Cretu, Charles Harris, Ilia Igashov, Arne Schneuing, Marwin Segler, Bruno Correia, Julien Roy, Emmanuel Bengio, and Pietro Liò. Synflownet: Design of diverse and novel molecules with synthesis constraints. *arXiv preprint arXiv:2405.01155*, 2024.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Tristan Deleu, Padideh Nouri, Nikolay Malkin, Doina Precup, and Yoshua Bengio. Discrete probabilistic inference as control in multi-path environments. *arXiv preprint arXiv:2402.10309*, 2024.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Benjamin Eysenbach and Sergey Levine. Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

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

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Haoran He, Can Chang, Huazhe Xu, and Ling Pan. Looking backward: Retrospective backward synthesis for goal-conditioned GFlownets. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fNMKqyvuZT.

Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and R M Neal. The “wake-sleep” algorithm for unsupervised neural networks. *Science*, 268 5214:1158–61, 1995.

Edward J Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. *arXiv preprint arXiv:2310.04363*, 2023.

Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever ve Wojciech Zaremba. Koda dayalı olarak eğitilmiş büyük dil modellerinin değerlendirilmesi, 2021.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang ve Furu Wei. Keşif ile muhakeme: Entropi perspektifi. *arXiv ön baskısı arXiv:2506.14758* , 2025.

Miruna Cretu, Charles Harris, Ilia Igashov, Arne Schneuing, Marwin Segler, Bruno Correia, Julien Roy, Emmanuel Bengio ve Pietro Liò. Synflownet: Sentez kısıtlarıyla çeşitli ve özgün moleküllerin tasarımı. *arXiv ön baskısı arXiv:2405.01155* , 2024.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen ve diğerleri. Muhakeme dil modelleri için takviyeli öğrenmenin entropi mekanizması. *arXiv ön baskısı arXiv:2505.22617* , 2025.

DeepSeek-AI. Deepseek-r1: Takviyeli öğrenme yoluyla llmlerde muhakeme yeteneğinin teşvik edilmesi, 2025. URL https://arxiv.org/abs/2501.12948 .

Tristan Deleu, Padideh Nouri, Nikolay Malkin, Doina Precup ve Yoshua Bengio. Çok yollu ortamlarda kontrol olarak ayrık olasılıksal çıkarım. *arXiv ön baskısı arXiv:2402.10309* , 2024.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang ve diğerleri. Ajan tabanlı takviyeli politika optimizasyonu. *arXiv ön baskısı arXiv:2507.19849* , 2025.

Yilun Du ve Igor Mordatch. Enerji tabanlı modellerle örtük üretim ve modelleme. *Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler* , 32, 2019.

Benjamin Eysenbach ve Sergey Levine. Maksimum entropi takviyeli öğrenme bazı sağlam takviyeli öğrenme problemlerini (kanıtlanmış şekilde) çözer. *arXiv ön baskısı arXiv:2103.06257* , 2021.

Leo Gao, John Schulman ve Jacob Hilton. Ödül modeli aşırı optimizasyonu için ölçekleme yasaları. Uluslararası Makine Öğrenmesi Konferansı'nda , sayfa 10835–10866. PMLR, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi ve diğerleri. Deepseek-r1: LLMLerde muhakeme yeteneğini takviyeli öğrenme yoluyla teşvik etme. *arXiv ön baskısı arXiv:2501.12948* , 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel ve Sergey Levine. Soft actor-critic: Rastlantısal aktörlü ortam dışı maksimum entropili derin takviyeli öğrenme. Uluslararası Makine Öğrenmesi Konferansında, sayfalar 1861–1870. Pmlr, 2018.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang ve diğerleri. Olympiadbench: Olimpiyat seviyesinde iki dilli çok modlu bilimsel problemlerle AGI'yi destekleyen zorlu bir kıyaslama. *arXiv ön baskısı arXiv:2402.14008* , 2024.

Haoran He, Can Chang, Huazhe Xu ve Ling Pan. Geriye Bakmak: Hedef-kodlanmış GFlowNet'ler için retrospektif geriye dönük sentez. On Üçüncü Uluslararası Öğrenme Temsilleri Konferansında, 2025. URL *https://openreview.net/forum?id=fNMKqyvuZT* .

Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey ve R. M. Neal. Denetimsiz sinir ağları için “wake-sleep” algoritması. *Science* , 268 5214:1158–61, 1995.

Edward J. Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio ve Nikolay Malkin. Büyük dil modellerinde çözümsüz çıkarımı amorti etme. *arXiv ön baskısı arXiv:2310.04363* , 2023.

Edward J Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0uj6p4ca60.

Jian Hu, Jason Klein Liu, and Wei Shen. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models, 2025. *URL https://arxiv. org/abs/2501*, 3262:32–33, 2025.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F.P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with GFlowNets. *International Conference on Machine Learning (ICML)*, 2022.

Moksh Jain, Tristan Deleu, Jason S. Hartford, Cheng-Hao Liu, Alex Hernández-García, and Yoshua Bengio. Gflownets for ai-driven scientific discovery. *ArXiv*, abs/2302.00615, 2023a. URL https://api.semanticscholar.org/CorpusID:256459319.

Moksh Jain, Sharath Chandra Raparthy, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret, and Emmanuel Bengio. Multi-objective GFlowNets. *International Conference on Machine Learning (ICML)*, 2023b.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Koray Kavukcuoglu. Gemini 2.5: Our most intelligent AI model, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Google Blog (The Keyword), Published Mar. 25, 2025.

Minsu Kim, Taeyoung Yun, Emmanuel Bengio, Dinghuai Zhang, Yoshua Bengio, Sungsoo Ahn, and Jinkyoo Park. Local search gflownets. *ArXiv*, abs/2310.02710, 2023.

Minsu Kim, Joohwan Ko, Taeyoung Yun, Dinghuai Zhang, Ling Pan, Woochang Kim, Jinkyoo Park, Emmanuel Bengio, and Yoshua Bengio. Learning to scale logits for temperature-conditional gflownets, 2024.

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, et al. Learning diverse attacks on large language models for robust red-teaming and safety tuning. *arXiv preprint arXiv:2405.18540*, 2024.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023b.

Edward J. Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio ve Nikolay Malkin. Büyük dil modellerinde çözümsüz çıkarımı amorti etme. On İkinci Uluslararası Öğrenme Temsilleri Konferansı'nda , *2024*. *URL* https://openreview.net/f orum?id=Ouj6p4ca60 .

Jian Hu, Jason Klein Liu ve Wei Shen. Reinforce++: Hem prompt hem de ödül modellerine karşı dayanıklı verimli bir rlhf algoritması, 2025. *URL https://arxiv. org/abs/2501* , 3262:32–33, 2025.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F.P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das ve Yoshua Bengio. Biyolojik dizi tasarımı GFlowNets ile. *Uluslararası Makine Öğrenmesi Konferansı (ICML)* , 2022.

Moksh Jain, Tristan Deleu, Jason S. Hartford, Cheng-Hao Liu, Alex Hernández-García ve Yoshua Bengio. Yapay zeka destekli bilimsel keşif için Gflownets. *ArXiv* , abs/2302.00615, 2023a. URL https://api.semanticscholar.org/CorpusID:256459319 .

Moksh Jain, Sharath Chandra Raparthy, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret ve Emmanuel Bengio. Çok Amaçlı GFlowNets. *Uluslararası Makine Öğrenmesi Konferansı (ICML)* , 2023b.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen ve Ion Stoica. Livecodebench: Kod için büyük dil modellerinin bütüncül ve kontaminasyondan arındırılmış değerlendirmesi. *arXiv ön baskısı arXiv:2403.07974* , 2024.

Koray Kavukcuoglu. Gemini 2.5: En akıllı yapay zeka modelimiz, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ . Google Blog (The Keyword), Yayın Tarihi 25 Mart 2025.

Minsu Kim, Taeyoung Yun, Emmanuel Bengio, Dinghuai Zhang, Yoshua Bengio, Sungsoo Ahn ve Jinkyoo Park. Yerel arama gflownetleri. *ArXiv* , abs/2310.02710, 2023.

Minsu Kim, Joohwan Ko, Taeyoung Yun, Dinghuai Zhang, Ling Pan, Woochang Kim, Jinkyoo Park, Emmanuel Bengio ve Yoshua Bengio. Sıcaklık-koşullu gflownetler için logitleri ölçeklemeyi öğrenme, 2024.

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin ve diğerleri. Dayanıklı red-teaming ve güvenlik ayarı için büyük dil modellerine karşı çeşitli saldırıların öğrenilmesi. *arXiv ön baskısı arXiv:2405.18540* , 2024.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari ve Vedant Misra. Dil modelleriyle nicel muhakeme problemlerinin çözümü. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho ve A. Oh, editörler, Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler, cilt 35, ss. 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abb eef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf .

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever ve Karl Cobbe. Adım adım doğrulayalım. *arXiv ön baskısı arXiv :2305.20050* , 2023a.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever ve Karl Cobbe. Adım adım doğrulayalım. In *The Twelfth International Conference on Learning Representations* , 2023b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Dianbo Liu, Moksh Jain, Bonaventure F. P. Dossou, Qianli Shen, Salem Lahlou, Anirudh Goyal, Nikolay Malkin, Chris C. Emezue, Dinghuai Zhang, Nadhir Hassen, Xu Ji, Kenji Kawaguchi, and Yoshua Bengio. Gflowout: Dropout with generative flow networks. In *International Conference on Machine Learning*, 2022.

Mingjie Liu, Shizhe Diao, Jian Hu, Ximing Lu, Xin Dong, Hao Zhang, Alexander Bukharin, Shaokun Zhang, Jiaqi Zeng, Makesh Narsimhan Sreedhar, et al. Scaling up rl: Unlocking diverse reasoning in llms via prolonged training. *arXiv preprint arXiv:2507.12507*, 2025a.

Zhen Liu, Tim Z Xiao, , Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed gflownets. In *ICLR*, 2025b.

Michael Luo, Sijun Tan, Roy Huang, Xiaoxiang Shi, Rachel Xin, Colin Cai, Ameen Patel, Alpay Ariyak, Qingyang Wu, Ce Zhang, Li Erran Li, Raluca Ada Popa, Ion Stoica, and Tianjun Zhang. Deepcoder: A fully open-source 14b coder at o3-mini level, 2025. Notion Blog.

Jiangyan Ma, Emmanuel Bengio, Yoshua Bengio, and Dinghuai Zhang. Baking symmetry into gflownets.

MAA. American mathematics competitions - amc. https://maa.org/, 2023.

MAA. American invitational mathematics examination - aime. https://maa.org/, 2025.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning gflownets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pages 23467–23483. PMLR, 2023.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35: 5955–5967, 2022.

Nikolay Malkin, Salem Lahlou, Tristan Deleu, Xu Ji, Edward Hu, Katie Everett, Dinghuai Zhang, and Yoshua Bengio. GFlowNets and variational inference. *International Conference on Learning Representations (ICLR)*, 2023.

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. *arXiv preprint arXiv:2507.21053*, 2025.

Sobhan Mohammadpour, Emmanuel Bengio, Emma Frejinger, and Pierre-Luc Bacon. Maximum entropy gflownets with soft q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2593–2601. PMLR, 2024.

OpenAI. Gpt-4o mini. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. Accessed: 2024.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Ling Pan, Moksh Jain, Kanika Madan, and Yoshua Bengio. Pre-training and fine-tuning generative flow networks, 2023a.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel ve Matthew Le. Generatif modelleme için akış eşleştirme. On birinci Öğrenme Temsilleri Uluslararası Konferansı'nda, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t .

Dianbo Liu, Moksh Jain, Bonaventure F. P. Dossou, Qianli Shen, Salem Lahlou, Anirudh Goyal, Nikolay Malkin, Chris C. Emezue, Dinghuai Zhang, Nadhir Hassen, Xu Ji, Kenji Kawaguchi ve Yoshua Bengio. Gflowout: Generatif akış ağları ile dropout. Uluslararası Makine Öğrenmesi Konferansı'nda, 2022.

Mingjie Liu, Shizhe Diao, Jian Hu, Ximing Lu, Xin Dong, Hao Zhang, Alexander Bukharin, Shaokun Zhang, Jiaqi Zeng, Makesh Narsimhan Sreedhar ve diğerleri. Reinforcement learning ölçeklendirmesi: Uzun süreli eğitim ile llmlerde çeşitli muhakemeyi açığa çıkarma. *arXiv ön baskısı arXiv:2507.12507* , 2025a.

Zhen Liu, Tim Z Xiao, Weiyang Liu, Yoshua Bengio ve Dinghuai Zhang. Gradyan bilgisi ile yönlendirilen gflownets aracılığıyla verimli çeşitlilik koruyucu difüzyon hizalaması. In *ICLR* , 2025b.

Michael Luo, Sijun Tan, Roy Huang, Xiaoxiang Shi, Rachel Xin, Colin Cai, Ameen Patel, Alpay Ariyak, Qingyang Wu, Ce Zhang, Li Erran Li, Raluca Ada Popa, Ion Stoica ve Tianjun Zhang. Deepcoder: Tamamen açık kaynaklı, o3-mini seviyesinde 14 milyar parametreli kodlayıcı, 2025. Notion Blog.

Jiangyan Ma, Emmanuel Bengio, Yoshua Bengio ve Dinghuai Zhang. Gflownets'e simetri entegre etmek.

MAA. Amerikan Matematik Yarışmaları - amc. https://maa.org/ , 2023.

MAA. American invitational mathematics examination - aime. https://maa.org/ , 2025.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio ve Nikolay Malkin. Geliştirilmiş yakınsama ve kararlılık için kısmi bölümlerden GFlowNets öğrenimi. Uluslararası Makine Öğrenmesi Konferansı'nda, sayfa 23467–23483. PMLR, 2023.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun ve Yoshua Bengio. Yörünge dengelemesi: *GFlowNets'te iyileştirilmiş kredi tahsisi*. *Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler*, 35: 5955–5967, 2022.

Nikolay Malkin, Salem Lahlou, Tristan Deleu, Xu Ji, Edward Hu, Katie Everett, Dinghuai Zhang ve Yoshua Bengio. GFlowNets ve varyasyonel çıkarım. *Uluslararası Öğrenme Temsilleri Konferansı (ICLR)* , 2023.

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng ve Angjoo Kanazawa. Flow matching policy gradients. *arXiv ön baskısı arXiv:2507.21053* , 2025.

Sobhan Mohammadpour, Emmanuel Bengio, Emma Frejinger ve Pierre-Luc Bacon. Maksimum entropili gflownets ile yumuşak q-öğrenme. Uluslararası Yapay Zeka ve İstatistik Konferansı , sayfa 2593–2601. PMLR, 2024.

OpenAI. Gpt-4o mini. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ , 2024. Erişim: 2024.

Alexander Pan, Kush Bhatia ve Jacob Steinhardt. Ödül yanlış tanımlamasının etkileri: Uyumsuz modellerin haritalanması ve hafifletilmesi. *arXiv ön baskısı arXiv:2201.03544* , 2022.

Ling Pan, Moksh Jain, Kanika Madan ve Yoshua Bengio. Generatif akış ağlarının ön eğitim ve ince ayarı, 2023a.

Ling Pan, Nikolay Malkin, Dinghuai Zhang, and Yoshua Bengio. Better training of GFlowNets with local credit and incomplete trajectories. *International Conference on Machine Learning (ICML)*, 2023b.

Ling Pan, Dinghuai Zhang, Aaron Courville, Longbo Huang, and Yoshua Bengio. Generative augmented flow networks. *International Conference on Learning Representations (ICLR)*, 2023c.

Ling Pan, Dinghuai Zhang, Moksh Jain, Longbo Huang, and Yoshua Bengio. Stochastic generative flow networks. *Uncertainty in Artificial Intelligence (UAI)*, 2023d.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KVf2SFL1pi.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Codeforces. https://huggingface.co/datasets/open-r1/codeforces, 2025.

Samuel Pfrommer, Yixiao Huang, and Somayeh Sojoudi. Reinforcement learning for flow-matching policies. *arXiv preprint arXiv:2507.15073*, 2025.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. Magistral. *arXiv preprint arXiv:2506.10910*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, 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.

Max W. Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho, and Tommaso Biancalani. Towards understanding and improving gflownet training. *ArXiv*, abs/2305.07170, 2023. URL https://api.semanticscholar.org/CorpusID:258676487.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Richard S Sutton, Andrew G Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134, 1999a.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999b. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Ling Pan, Nikolay Malkin, Dinghuai Zhang ve Yoshua Bengio. Yerel kredi ve eksik trajektörlerle GFlowNet'lerin daha iyi eğitimi. *Uluslararası Makine Öğrenmesi Konferansı (ICML)* , 2023b.

Ling Pan, Dinghuai Zhang, Aaron Courville, Longbo Huang ve Yoshua Bengio. Generatif artırılmış akış ağları. *Uluslararası Temsil Öğrenimi Konferansı (ICLR)* , 2023c.

Ling Pan, Dinghuai Zhang, Moksh Jain, Longbo Huang ve Yoshua Bengio. Stokastik generatif akış ağları. *Yapay Zekada Belirsizlik Konferansı (UAI)* , 2023d.

Seohong Park, Qiyang Li ve Sergey Levine. Flow q-learning. Kırkikinci Uluslararası Makine Öğrenmesi Konferansı, 2025. URL *https://openreview.net/forum?id=KVf2SFL1pi* .

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall ve Leandro von Werra. Codeforces. https://huggingface.co/datasets/open-r1/codeforces , 2025.

Samuel Pfrommer, Yixiao Huang ve Somayeh Sojoudi. Akış eşleştirme politikaları için takviyeli öğrenme. *arXiv ön baskısı arXiv:2507.15073* , 2025.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu ve diğerleri. Magistral. *arXiv ön baskısı arXiv:2506.10910* , 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford ve Oleg Klimov. Proksimal politika optimizasyon algoritmaları. *arXiv ön baskısı arXiv:1707.06347* , 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu ve diğerleri. Deepseekmath: Açık dil modellerinde matematiksel muhakemenin sınırlarını zorlama. *arXiv ön baskısı arXiv:2402.03300* , 2024.

Max W. Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho ve Tommaso Biancalani. gflownet eğitiminin anlaşılması ve geliştirilmesine doğru. *ArXiv* , abs /2305.07170, 2023. URL https://api.semanticscholar.org/CorpusID:258676487 .

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin ve Chuan Wu. Hybridflow: Esnek ve verimli bir RLHF çerçevesi. *arXiv ön baskısı ar Xiv: 2409.19256* , 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov ve David Krueger. Ödül oyununu tanımlama ve karakterize etme. *Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler* , 35:9460–9471, 2022.

Richard S. Sutton, Andrew G. Barto ve diğerleri. Takviyeli öğrenme. *Bilişsel Sinirbilim Dergisi* , 11 (1):126–134, 1999a.

Richard S. Sutton, David McAllester, Satinder Singh ve Yishay Mansour. Fonksiyon yaklaşık değerleri ile takviyeli öğrenmede politika gradyan yöntemleri. S. Solla, T. Leen ve K. Müller editörlüğünde, *Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler* , cilt 12. MIT Press, 1999b. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Qwen Takımı. Qwen2.5: Temel modeller partisi, Eylül 2024. URL https://qwenlm.github.io/blog/qwen2.5/ .

Daniil Tiapkin, Nikita Morozov, Alexey Naumov, and Dmitry P Vetrov. Generative flow networks as entropy-regularized rl. In *International Conference on Artificial Intelligence and Statistics*, pages 4213–4221. PMLR, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Training llms for divergent reasoning with minimal examples. In *Forty-second International Conference on Machine Learning*, 2025a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025b.

Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park, and Ling Pan. Learning to sample effective and diverse prompts for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23625–23635, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

David W. Zhang, Corrado Rainone, Markus F. Peschl, and Roberto Bondesan. Robust scheduling with gflownets. *ArXiv*, abs/2302.05446, 2023a. URL https://api.semanticscholar.org/CorpusID:256827133.

Dinghuai Zhang, Ricky T. Q. Chen, Nikolay Malkin, and Yoshua Bengio. Unifying generative models with GFlowNets and beyond. *arXiv preprint arXiv:2209.02606v2*, 2022a.

Dinghuai Zhang, Nikolay Malkin, Zhen Liu, Alexandra Volokhova, Aaron Courville, and Yoshua Bengio. Generative flow networks for discrete probabilistic modeling. *International Conference on Machine Learning (ICML)*, 2022b.

Dinghuai Zhang, Hanjun Dai, Nikolay Malkin, Aaron C. Courville, Yoshua Bengio, and Ling Pan. Let the flows tell: Solving graph combinatorial optimization problems with gflownets. *ArXiv*, abs/2305.17010, 2023b.

Dinghuai Zhang, Ricky T. Q. Chen, Cheng-Hao Liu, Aaron Courville, and Yoshua Bengio. Diffusion generative flow samplers: Improving learning signals through partial trajectory optimization, 2024a.

Dinghuai Zhang, Ling Pan, Ricky T. Q. Chen, Aaron Courville, and Yoshua Bengio. Distributional gflownets with quantile flows, 2024b.

Dinghuai Zhang, Yizhe Zhang, Jiatao Gu, Ruixiang ZHANG, Joshua M. Susskind, Navdeep Jaitly, and Shuangfei Zhai. Improving GFlownets for text-to-image diffusion alignment. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=XDbY3qhM42.

Daniil Tiapkin, Nikita Morozov, Alexey Naumov ve Dmitry P. Vetrov. Entropi düzenlemeli takviyeli öğrenme olarak generatif akış ağları. Uluslararası Yapay Zeka ve İstatistik Konferansı'nda, sayfa 4213–4221. PMLR, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang ve ark. 80/20 kuralının ötesinde: Yüksek entropili azınlık tokenları, LLM muhakemesi için etkili takviyeli öğrenmeyi destekler. *arXiv ön baskısı arXiv:2506.01939* , 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou ve diğerleri. Zincirleme düşünce yönlendirmesi büyük dil modellerinde muhakemeyi tetikler. *Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler* , 35:24824–24837, 2022.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao ve Lianhui Qin. Muhakeme akışı: llmleri en az örnekle farklılaşan muhakeme için eğitmek. Kırk İkinci Uluslararası Makine Öğrenmesi Konferansı , *2025a*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu ve diğerleri. Dapo: Ölçeklenebilir açık kaynaklı bir llm takviyeli öğrenme sistemi. *arXiv ön baskısı arXiv:2503.14476*, 2025b.

Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park ve Ling Pan. Metinden görüntü üretimi için etkili ve çeşitli istemlerin örneklemesini öğrenme. Bilgisayar Görüşü ve Desen Tanıma Konferansı Bildirilerinde , sayfa 23625–23635, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu ve Noah Goodman. Star: Muhakeme ile muhakemeyi başlatma. *Sinirsel Bilgi İşleme Sistemlerindeki Gelişmeler* , 35:15476–15488, 2022.

David W. Zhang, Corrado Rainone, Markus F. Peschl ve Roberto Bondesan. gflownets ile dayanıklı zamanlama. *ArXiv* , abs/2302.05446, 2023a. URL https://api.semanticscholar.org/Corp usID:256827133 .

Dinghuai Zhang, Ricky T. Q. Chen, Nikolay Malkin ve Yoshua Bengio. Generatif modellerin GFlowNets ile ve ötesinde birleştirilmesi. *arXiv ön baskısı arXiv:2209.02606v2* , 2022a.

Dinghuai Zhang, Nikolay Malkin, Zhen Liu, Alexandra Volokhova, Aaron Courville ve Yoshua Bengio. Ayrık olasılıksal modelleme için generatif akış ağları. *Uluslararası Makine Öğrenmesi Konferansı (ICML)* , 2022b.

Dinghuai Zhang, Hanjun Dai, Nikolay Malkin, Aaron C. Courville, Yoshua Bengio ve Ling Pan. Akışların anlatmasına izin verin: gflownets ile grafik kombinatoryal optimizasyon problemlerinin çözümü. *ArXiv* , abs/2305.17010, 2023b.

Dinghuai Zhang, Ricky T. Q. Chen, Cheng-Hao Liu, Aaron Courville ve Yoshua Bengio. Diffusion generative flow samplers: Kısmi rota optimizasyonu yoluyla öğrenme sinyallerinin iyileştirilmesi, 2024a.

Dinghuai Zhang, Ling Pan, Ricky T. Q. Chen, Aaron Courville ve Yoshua Bengio. Quantile flows ile dağılımsal gflownetler, 2024b.

Dinghuai Zhang, Yizhe Zhang, Jiatao Gu, Ruixiang ZHANG, Joshua M. Susskind, Navdeep Jaitly ve Shuangfei Zhai. Metinden görüntüye difüzyon hizalaması için GFlownets'in iyileştirilmesi . *Transactions on Machine Learning Research* , 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=XDbY3qhM42 .

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025b.

Mingyang Zhou, Zichao Yan, Elliot Layne, Nikolay Malkin, Dinghuai Zhang, Moksh Jain, Mathieu Blanchette, and Yoshua Bengio. Phylogfn: Phylogenetic inference with generative flow networks, 2024.

Heiko Zimmermann, Fredrik Lindsten, J.-W. van de Meent, and Christian Andersson Naesseth. A variational perspective on generative flow networks. *ArXiv*, abs/2210.07992, 2022. URL https://api.semanticscholar.org/CorpusID:252907672.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li ve diğerleri. Büyük muhakeme modelleri için takviyeli öğrenme üzerine bir derleme. *arXiv ön baskısı arXiv:2509.08827* , 2025b.

Mingyang Zhou, Zichao Yan, Elliot Layne, Nikolay Malkin, Dinghuai Zhang, Moksh Jain, Mathieu Blanchette ve Yoshua Bengio. Phylogfn: Generatif Akış Ağları ile Filogenetik Çıkarım, 2024.

Heiko Zimmermann, Fredrik Lindsten, J.-W. van de Meent ve Christian Andersson Naesseth. Generatif Akış Ağlarına Variasyonel Bir Bakış Açısı. *ArXiv* , abs/2210.07992, 2022. URL https://api.semanticscholar.org/CorpusID:252907672 .

A. Proof of Proposition 1

We begin by analyzing the gradient of the Kullback–Leibler (KL) divergence between the policy $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ and the target reward distribution $\frac{\exp(\beta r(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}$:

$$
\begin{aligned}
&\nabla_\theta D_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y} \mid \mathbf{x}) \,\|\, \frac{\exp(\beta r(\mathbf{x},\mathbf{y}))}{Z_\phi(\mathbf{x})}\right) \\
&= \nabla_\theta \int \pi_\theta(\mathbf{y} \mid \mathbf{x}) \log\left[\frac{\pi_\theta(\mathbf{y} \mid \mathbf{x}) \cdot Z_\phi(\mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} \\
&= \int \nabla_\theta \pi_\theta(\mathbf{y} \mid \mathbf{x}) \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} + \int \pi_\theta(\mathbf{y} \mid \mathbf{x}) \nabla_\theta \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} \\
&= \int \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} + \underbrace{\int \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\, d\mathbf{y}}_{=\nabla_\theta \int \pi_\theta(\mathbf{y}|\mathbf{x})\, d\mathbf{y} = \nabla_\theta 1 = 0} \\
&= \int \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} \\
&= \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\log\left(\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right) \cdot \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\right]
\end{aligned}
\tag{8}
$$

Next, consider the trajectory balance objective used in GFlowNets learning [Bartoldson et al., 2025, Bengio et al., 2023b, Lee et al., 2024], defined as:

$$
\mathcal{L}(\mathbf{y}, \mathbf{x}; \theta) = \left(\log \frac{Z_\phi(\mathbf{x})\, \pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right)^2. \tag{9}
$$

Taking the gradient of this objective with respect to θ yields:

$$
\nabla_\theta \mathcal{L}(\theta) = 2 \cdot \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\left(\log \frac{Z_\phi(\mathbf{x}) \cdot \pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta r(\mathbf{x},\mathbf{y}))}\right) \cdot \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\right] \tag{10}
$$

Thus, minimizing the KL divergence is equivalent (up to a constant) to minimizing the trajectory balance loss, confirming Proposition 1.

B. Theoretical Analysis

We conduct an interpretation of FlowRL that clarifies the role of each component in the objective.

Proposition 5. *Minimizing the KL divergence in Eq. 5 is equivalent (in terms of gradients) to jointly maximizing reward and policy entropy:*

$$
\max_\theta \; \mathbb{E}_{\mathbf{y}\sim\pi_\theta}\left[\underbrace{\beta\, r(\mathbf{x},\mathbf{y})}_{\textit{reward}} - \log Z_\phi(\mathbf{x}) + \log \pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})\right] + \underbrace{\mathcal{H}(\pi_\theta)}_{\textit{entropy}}. \tag{11}
$$

Remark 6 (*FlowRL beyond reward maximization*)**.** Proposition 5 reveals that FlowRL can be interpreted as jointly maximizing expected reward and policy entropy. This shift encourages the policy to explore a broader set of high-quality solutions, enabling more diverse and generalizable behaviors on reasoning tasks. Our interpretation also aligns with prior work that views GFlowNets training as a form of maximum entropy RL [Deleu et al., 2024, Mohammadpour et al., 2024].

A. Önerme 1'in İspatı

Politika $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ ile hedef ödül dağılımı $\frac{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}{z_\phi(\mathbf{x})}$ arasındaki Kullback–Leibler (KL) ayrımının gradyanını analiz ederek başlıyoruz:

$$
\begin{aligned}
&\nabla_\theta D_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y} \mid \mathbf{x}) \,\|\, \frac{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}{z_\phi(\mathbf{x})}\right) \\
&= \nabla_\theta \int \pi_\theta(\mathbf{y} \mid \mathbf{x}) \log\left[\frac{\pi_\theta(\mathbf{y} \mid \mathbf{x}) \cdot z_\phi(\mathbf{x})}{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} \\
&= \int \nabla_\theta \pi_\theta(\mathbf{y} \mid \mathbf{x}) \log\frac{\pi_\theta(\mathbf{y} \mid \mathbf{x}) \exp(\beta\beta(\mathbf{x},\mathbf{y}))}{\cancel{z_\phi(\mathbf{x})}}\, d\mathbf{y} + \int \pi_\theta(\mathbf{y} \mid \mathbf{x}) \nabla_\theta \log\left[\frac{z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} \\
&= \int \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \log\left[\frac{z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} + \underbrace{\int \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\, d\mathbf{y}}_{=\nabla_\theta \int \pi_\theta(\mathbf{y}|\mathbf{x})\, d\mathbf{y} = \nabla_\theta 1 = 0} \\
&= \int \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\, \log\left[\frac{z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}\right] d\mathbf{y} \\
&= \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\log\left(\frac{z_\phi(\mathbf{x})\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\exp(\beta\beta(\mathbf{x},\mathbf{y}))}\right) \cdot \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\right]
\end{aligned}
\tag{8}
$$

Son olarak, GFlowNets öğreniminde kullanılan yörünge dengelemesi amaç fonksiyonu [Bartoldson et al., 2025, Bengio et al., 2023b, Lee et al., 2024] aşağıdaki gibi tanımlanır:

$$
\mathcal{L}(\mathbf{y}, \mathbf{x}; \theta) = \left(\log \frac{z_\phi(\mathbf{x})\, \pi_\theta(\mathbf{y} \mid \mathbf{x}) \exp}{(\beta\beta(\mathbf{x},\mathbf{y}))}\right)^2. \tag{9}
$$

Bu amaç fonksiyonunun θ'ya göre türevi alınırsa elde edilir:

$$
\nabla_\theta \mathcal{L}(\theta) = 2 \cdot \mathbb{E}_{\_\{\mathrm{y\ sim\ pi\_heta}(\mathrm{cdot}|\mathbf{x})}\left[\log \frac{z_\phi(\mathbf{x}) \cdot \pi_\theta(\mathbf{y} \mid \mathbf{x}) \exp}{(\beta\beta(\mathbf{x},\mathbf{y}))} \cdot \nabla_\theta \log \pi_\theta(\mathbf{y} \mid \mathbf{x})\right] \tag{10}
$$

Böylece, KL ayrışımını minimize etmek, bir sabite kadar yörünge dengeleme kaybını minimize etmekle eşdeğerdir ve bu, Önerme 1'i doğrular.

B. Kuramsal Analiz

FlowRL'in her bileşeninin amaçtaki rolünü açıklığa kavuşturan bir yorumlama yapıyoruz.

Önerme 5. *Eq. 5'teki KL ayrışımını minimize etmek, gradyanlar açısından ödül ve politika entropisini birlikte maksimize etmekle eşdeğerdir:*

$$
\max_\theta \; \mathbb{E}_{\mathbf{y}\sim\pi_\theta}\left[\underbrace{\beta\, r(\mathbf{x},\mathbf{y})}_{\textit{ödül}} - \log z_\phi(\mathbf{x}) + \log \pi_{\mathrm{re}/}(\mathbf{y}|\mathbf{x})\right] + \underbrace{\mathcal{H}(\pi_\theta)}_{\textit{entropi}}. \tag{11}
$$

Dipnot 6 (*Ödül maksimizasyonunun ötesinde FlowRL*). Önerme 5, FlowRL'in beklenen ödül ve politika entropisini birlikte maksimize etmek olarak yorumlanabileceğini göstermektedir. Bu değişim, politikanın daha geniş bir yüksek kaliteli çözüm kümesini keşfetmesini teşvik ederek, muhakeme görevlerinde daha çeşitli ve genellenebilir davranışlar ortaya koymasını sağlar. Yorumumuz, GFlowNets eğitiminin maksimum entropi pekiştirmeli öğrenme (RL) biçimi olarak değerlendirildiği önceki çalışmalarla da tutarlıdır [Deleu et al., 2024, Mohammadpour et al., 2024].

The proof of Proposition 5 is provided as below.

Recall from Eq. 3 and Eq. 5 that the FlowRL objective is sourced from the minimization of a KL divergence:

$$D_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \parallel \frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right) = \int \pi_\theta(\mathbf{y}\mid\mathbf{x}) \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}\right] d\mathbf{y} \tag{12}$$

Rearranging the terms, we obtain:

$$\begin{aligned}
&\arg\min_\theta D_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \parallel \frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right)\\
&= \arg\min_\theta \int \pi_\theta(\mathbf{y}\mid\mathbf{x}) \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}\right] d\mathbf{y}\\
&= \arg\max_\theta \left\{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})} \log\left[\frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right] - \int \pi_\theta(\mathbf{y}\mid\mathbf{x}) \log\pi_\theta(\mathbf{y}\mid\mathbf{x}) d\mathbf{y}\right\}\\
&= \arg\max_\theta \left\{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})} \log\left[\frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right] + \mathcal{H}(\pi_\theta)\right\}
\end{aligned} \tag{13}$$

Finally, we express the FlowRL objective in its compact form:

$$\max_\theta \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\underbrace{\beta r(\mathbf{x},\mathbf{y})}_{\text{reward}} - \underbrace{\log Z_\phi(\mathbf{x})}_{\text{normalization}} + \underbrace{\log\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})}_{\text{prior alignment}}\right] + \underbrace{\mathcal{H}(\pi_\theta)}_{\text{entropy}}. \tag{14}$$

Therefore, minimizing the FlowRL objective can be interpreted as jointly maximizing reward and entropy, while also aligning the policy with a structured prior. The reward term drives task performance, while the normalization term $Z_\phi(\mathbf{x})$ ensures consistency with a properly normalized target distribution. This encourages the policy π_θ to cover the entire reward-weighted distribution rather than collapsing to a few high-reward modes. The reference policy π_{ref} provides inductive bias that regularizes the policy toward desirable structures, and the entropy term $\mathcal{H}(\pi_\theta)$ encourages diversity in sampled solutions. Together, these components promote better generalization of FlowRL.

C. GFlowNets

We follow the notation of [He et al., 2025, Madan et al., 2023] to introduce the fundamentals of GFlowNets. Let $\mathcal{X}$ denote the compositional objects and R be a reward function that assigns non-negative values to each object $x \in \mathcal{X}$. GFlowNets aim to learn a sequential, constructive sampling policy π that generates objects x with probabilities proportional to their rewards, i.e.,$\pi(x) \propto R(x)$. This process can be represented as a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{S}, \mathcal{A})$, where the vertices $s \in \mathcal{S}$ are referred to as *states*, and the directed edges $(u \to v) \in \mathcal{A}$ are called *actions*. The generation of an object $x \in \mathcal{X}$ corresponds to a complete trajectory $\tau = (s_0 \to \cdots \to s_n) \in \mathcal{T}$ within the DAG, beginning at the initial state s_0 and ending at a terminal state $s_n \in \mathcal{X}$. The state flow $F(s)$ is defined as a non-negative weight assigned to each state $s \in \mathcal{S}$. The forward policy $P_F(s' \mid s)$ specifies the transition probability to a child state s', while the backward policy $P_B(s \mid s')$ specifies the transition probability to a parent state s. To this end, detailed balance objective enforces local flow consistency across every edge $(s \to s') \in \mathcal{A}$:

$$\forall (s \to s') \in \mathcal{A}, \quad F_\theta(s) P_F(s' \mid s;\theta) = F_\theta(s') P_B(s \mid s';\theta). \tag{15}$$

Önerme 5'in ispatı aşağıda verilmiştir.

Denklem 3 ve Denklem 5'ten hatırlanacağı üzere, FlowRL amacı bir KL ayrışımının minimize edilmesinden türetilmiştir:

$$D_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \parallel \frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{re/}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right) = \int \pi_\theta(\mathbf{y}\mid\mathbf{x}) \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{re/}}(\mathbf{y}\mid\mathbf{x})}\right] d\mathbf{y} \tag{12}$$

Terimleri yeniden düzenlediğimizde elde ederiz:

$$\begin{aligned}
&\arg\min_\theta D_{\mathrm{KL}}\left(\pi_\theta(\mathbf{y}\mid\mathbf{x}) \parallel \frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{re/}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right)\\
&= \arg\min_\theta \int \pi_\theta(\mathbf{y}\mid\mathbf{x}) \log\left[\frac{Z_\phi(\mathbf{x})\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{re/}}(\mathbf{y}\mid\mathbf{x})}\right] d\mathbf{y}\\
&= \arg\max_\theta \left\{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})} \log\left[\frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{re/}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right] - \int \pi_\theta(\mathbf{y}\mid\mathbf{x}) \text{ logaritma } \pi_\theta(\mathbf{y}\right\}\mathbf{x}) d\mathbf{y}\\
&= \arg\max_\theta \left\{\mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})} \log\left[\frac{\exp(\beta\, r(\mathbf{x},\mathbf{y}))\cdot\pi_{\mathrm{re/}}(\mathbf{y}\mid\mathbf{x})}{Z_\phi(\mathbf{x})}\right] + \mathcal{H}(\pi_\theta)\right\}
\end{aligned} \tag{13}$$

Son olarak, FlowRL amaç fonksiyonunu kompakt biçimde ifade ederiz:

$$\max_\theta \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\underbrace{\beta\beta(\mathbf{x},\mathbf{y})}_{\text{ödül}} - \underbrace{\text{logaritma } Z_\phi(\mathbf{x})}_{\text{normalizasyon}} + \underbrace{\text{logaritma } \pi_{\mathrm{re/}}(\mathbf{y}|\mathbf{x})}_{\text{önsel hizalama}}\right] + \underbrace{\mathcal{H}(\pi_\theta)}_{\text{entropi}}. \tag{14}$$

Bu nedenle, FlowRL amacının minimize edilmesi, ödül ve entropinin birlikte maksimize edilmesi ve politikaların yapılandırılmış bir önsel ile hizalanması olarak yorumlanabilir. Ödül terimi görev performansını artırırken, normalizasyon terimi $Z_\phi(\mathbf{x})$ düzgün şekilde normalleştirilmiş hedef dağılımla tutarlılığı sağlar. Bu durum, politika π_θ'nın sadece az sayıda yüksek ödüllü moda çökmesi yerine ödül ağırlıklı tüm dağılımı kapsamasını teşvik eder. Referans politika $\pi_{\mathrm{re/}}$, politikayı arzu edilen yapılar doğrultusunda düzenleyen endüktif önyargı sağlar ve entropi terimi $\mathcal{H}(\pi_\theta)$ örneklenen çözümlerde çeşitliliği teşvik eder. Bu bileşenler, FlowRL'nin daha iyi genelleme sağlamasına katkıda bulunur.

C. GFlowNets

[He et al., 2025, Madan et al., 2023] çalışmalarının gösterimini takip ederek GFlowNets'in temel kavramlarını tanıtıyoruz. Let $\mathcal{X}$ bileşik nesneleri ve R her bir $x \in \mathcal{X}$ nesnesine negatif olmayan değerler atayan bir ödül fonksiyonunu temsil etsin. GFlowNets, ödüllerine orantılı olarak nesneler x üreten sıralı, yapıcı bir örnekleme politikası π öğrenmeyi amaçlar, yani $\pi(x) \propto R(x)$. Bu süreç, düğümler $s \in$ S olarak adlandırılan durumlar ve yönlendirilmiş kenarlar $(u \to v) \in$ A olarak adlandırılan eylemlerden oluşan yönlendirilmiş asiklik bir grafik (DAG) G = (S, A) şeklinde temsil edilebilir. Bir nesnenin $x \in$ X üretilmesi, başlangıç durumu s_0 olan ve son durum olarak sonlandırıcı durum $s_n \in$ X olan DAG içerisindeki tam bir yörünge $\tau = (s_0 \to \cdots \to s_n) \in$ T ile karşılanır. Durum akışı $F(s)$, *her durum* $s \in \mathcal{S}$ için atanan negatif olmayan bir ağırlık olarak tanımlanır. İleri politika $PF(s' \mid s)$ çocuk duruma s' geçiş olasılığını belirtirken, geri politika $PB(s \mid s')$ ebeveyn duruma s geçiş olasılığını belirtir. Bu bağlamda, detaylı denge amacı her bir kenar $(s \to s') \in$ A üzerinde yerel akış tutarlılığını sağlar:

$$\forall (s \to s') \in \mathcal{A}, \quad F_\theta(s) P_F(s' \mid s;\theta) = F_\theta(s') P_B(s \mid s';\theta). \tag{15}$$

To achieve this flow consistency, GFlowNets employ training objectives at different levels of granularity, including detailed balance [Bengio et al., 2023b], trajectory balance [Malkin et al., 2022], and sub-trajectory balance [Madan et al., 2023]. Leveraging their diversity-seeking behavior, GFlowNets have been successfully applied across a range of domains, including molecule generation [Cretu et al., 2024], diffusion fine-tuning [Liu et al., 2025b, Zhang et al., 2025a], and amortized reasoning [Hu et al., 2024, Yu et al., 2025a]. Among various training objective in GFlowNets, trajectory balance maintains flow consistency at the trajectory level, defined as:

$$Z_\theta \prod_{t=1}^{n} P_F(s_t \mid s_{t-1}; \theta) = R(x) \prod_{t=1}^{n} P_B(s_{t-1} \mid s_t; \theta). \tag{16}$$

Furthermore, sub-trajectory balance achieves local balance on arbitrary subpaths $\tau_{i:j} = \{s_i \to \cdots \to s_j\}$, offering a more stable and less biased learning signal. We build on trajectory balance to extend our KL-based objective through a gradient-equivalence formulation (Prop. 1), and further improve it to better support long CoT reasoning in RL.

Models	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olympiad	Avg
Qwen2.5-7B Base Model							
Backbone	4.37	2.08	30.78	54.48	22.38	24.02	23.02
R++	10.57$_{+6.20}$	5.10$_{+3.02}$	66.02$_{+35.24}$	54.29$_{-0.19}$	24.47$_{+2.09}$	27.30$_{+3.28}$	31.29
PPO	9.95$_{+5.58}$	7.34$_{+5.26}$	63.63$_{+32.85}$	57.72$_{+3.24}$	26.22$_{+3.84}$	27.35$_{+3.33}$	32.03
GRPO	14.01$_{+9.64}$	10.73$_{+8.65}$	64.10$_{+33.32}$	57.41$_{+2.93}$	23.17$_{+0.79}$	27.11$_{+3.09}$	32.76
FlowRL	14.32$_{+9.95}$	10.05$_{+7.97}$	55.08$_{+24.30}$	66.78$_{+12.30}$	31.52$_{+9.14}$	34.60$_{+10.58}$	**35.39**

Table 5 | Math reasoning performance (Avg@64) at temperature = 0.6. Relative improvements are shown as subscripts, with positive gains in green and negative changes in red. FlowRL consistently outperforms all baselines and achieves the best average score under this low-temperature setting.

Models	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olympiad	Avg
Qwen2.5-7B Base Model							
Backbone	3.39	1.51	23.90	45.18	16.98	18.27	18.20
R++	10.63$_{+7.24}$	4.63$_{+3.12}$	66.99$_{+43.09}$	54.36$_{+9.18}$	23.89$_{+6.91}$	26.65$_{+8.38}$	31.19
PPO	10.52$_{+7.13}$	6.51$_{+5.00}$	63.04$_{+39.14}$	57.46$_{+12.28}$	25.91$_{+8.93}$	27.16$_{+8.89}$	31.77
GRPO	12.50$_{+9.11}$	10.10$_{+8.59}$	64.72$_{+40.82}$	57.15$_{+11.97}$	23.28$_{+6.30}$	26.90$_{+8.63}$	32.44
FlowRL	14.22$_{+10.83}$	9.58$_{+8.07}$	52.92$_{+29.02}$	66.20$_{+21.02}$	30.32$_{+13.34}$	34.47$_{+16.20}$	**34.62**

Table 6 | Math reasoning performance (Avg@64) at temperature = 1.0. Relative improvements are shown as subscripts, with positive gains in green. FlowRL maintains robust performance under higher generation randomness and continues to outperform all baselines on average.

Bu akış tutarlılığını sağlamak için, GFlowNets farklı ayrıntı düzeylerinde eğitim amaçları kullanır; bunlar arasında detaylı dengeleme [Bengio ve ark., 2023b], yörünge dengelemesi [Malkin ve ark., 2022] ve alt-yörünge dengelemesi [Madan ve ark., 2023] bulunmaktadır. Çeşitlilik arayan davranışlarından yararlanan GFlowNets, molekül üretimi [Cretu ve ark., 2024], difüzyon ince ayarı [Liu ve ark., 2025b, Zhang ve ark., 2025a] ve amortize muhakeme [Hu ve ark., 2024, Yu ve ark., 2025a] gibi farklı alanlarda başarıyla uygulanmıştır. GFlowNets içindeki çeşitli eğitim amaçları arasında, yörünge dengelemesi akış tutarlılığını yörünge düzeyinde korur ve şu şekilde tanımlanır:

$$Z_\theta \prod_{t=1}^{n} P_F(s_t \mid s_{t-1}; \theta) = R(x) \prod_{t=1}^{n} P_B(s_{t-1} \mid s_t; \theta). \tag{16}$$

Buna ek olarak, alt-yörünge dengelemesi, $\{\tau_{i:j} = \{s_i \to \cdots \to s_j\}\}$ gibi rastgele alt yollarda yerel denge sağlar ve daha stabil, daha az önyargılı bir öğrenme sinyali sunar. Biz, yörünge dengesine dayanarak KL tabanlı amacımızı gradyan eşdeğerlik formülasyonu (Önerme 1) aracılığıyla genişletiyor ve bunu RL'de uzun CoT muhakemesini daha iyi destekleyecek şekilde geliştiriyoruz.

Modeller	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olimpiyat	Ortalama
Qwen2.5-7B Temel Model							
Omurga	4.37	2.08	30.78	54.48	22.38	24.02	23.02
R++	10.57$_{+6.20}$	5.10$_{+3.02}$	66.02$_{+35.24}$	54.29$_{-0.19}$	24.47$_{+2.09}$	27.30$_{+3.28}$	31.29
PPO	9.95$_{+5.58}$	7.34$_{+5.26}$	63.63$_{+32.85}$	57.72$_{+3.24}$	26.22$_{+3.84}$	27.35$_{+3.33}$	32.03
GRPO	14.01$_{+9.64}$	10.73$_{+8.65}$	64.10$_{+33.32}$	57.41$_{+2.93}$	23.17$_{+0.79}$	27.11$_{+3.09}$	32.76
FlowRL	14.32$_{+9.95}$	10.05$_{+7.97}$	55.08$_{+24.30}$	66.78$_{+12.30}$	31.52$_{+9.14}$	34.60$_{+10.58}$	**35.39**

Tablo **5** | Matematik muhakemesi performansı (Ortalama@64) sıcaklık = 0.6. Göreli iyileşmeler alt simge olarak gösterilmiştir; pozitif kazançlar yeşil, negatif değişiklikler kırmızı ile belirtilmiştir. FlowRL, temel modellerin tamamının üzerinde tutarlı şekilde performans göstererek bu düşük sıcaklık ayarında en iyi ortalama skoru elde etmektedir.

Modeller	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olimpiyat	Ortalama
Qwen2.5-7B Temel Model							
Omurga	3.39	1.51	23.90	45.18	16.98	18.27	18.20
R++	10.63$_{+7.24}$	4.63$_{+3.12}$	66.99$_{+43.09}$	54.36$_{+9.18}$	23.89$_{+6.91}$	26.65$_{+8.38}$	31.19
PPO	10.52$_{+7.13}$	6.51$_{+5.00}$	63.04$_{+39.14}$	57.46$_{+12.28}$	25.91$_{+8.93}$	27.16$_{+8.89}$	31.77
GRPO	12.50$_{+9.11}$	10.10$_{+8.59}$	64.72$_{+40.82}$	57.15$_{+11.97}$	23.28$_{+6.30}$	26.90$_{+8.63}$	32.44
FlowRL	14.22$_{+10.83}$	9.58$_{+8.07}$	52.92$_{+29.02}$	66.20$_{+21.02}$	30.32$_{+13.34}$	34.47$_{+16.20}$	**34.62**

Tablo **6** | Matematik muhakemesi performansı (Ortalama@64) sıcaklık = 1.0. Göreli iyileşmeler alt simge olarak gösterilmiştir; pozitif kazançlar yeşil ile belirtilmiştir. FlowRL, üretim rastgeleliği yüksek olduğunda da sağlam performansını koruyarak ortalama olarak tüm temel modelleri aşmaya devam etmektedir.

Models	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olympiad	Avg
$\beta = 5$	13.54	10.00	56.09	58.91	20.79	28.72	31.34
$\beta = 10$	14.79	10.20	59.53	64.30	25.27	32.39	34.41
$\beta = 15$	15.41	10.83	54.53	66.96	31.41	34.61	35.63
$\beta = 30$	15.00	10.83	50.62	69.02	30.03	35.03	35.09

Table 7 | Ablation study on the effect of the β parameter in FlowRL. We report Avg@16 accuracy across six math reasoning benchmarks for different values of β.

Diversity Evaluation Prompt

System: You are evaluating the DIVERSITY of solution approaches for a mathematics competition problem. Focus on detecting even SUBTLE differences in methodology that indicate different problem-solving strategies.
PROBLEM:
{problem}
16 SOLUTION ATTEMPTS:
{formatted_responses}
EVALUATION CRITERIA - Rate diversity from 1 to 5:
Score 1 - Minimal Diversity:
- 14+ responses use essentially identical approaches
- Same mathematical setup, same variable choices, same solution path
- Only trivial differences (arithmetic, notation, wording)
- Indicates very low exploration/diversity in the generation process

Score 2 - Low Diversity:
- 11-13 responses use the same main approach
- 1-2 alternative approaches appear but are rare
- Minor variations within the dominant method (different substitutions, orderings)
- Some exploration but heavily biased toward one strategy

Score 3 - Moderate Diversity:
- 7-10 responses use the most common approach
- 2-3 distinct alternative approaches present
- Noticeable variation in problem setup or mathematical techniques
- Balanced mix showing reasonable exploration

Score 4 - High Diversity:
- 4-6 responses use the most common approach
- 3-4 distinct solution strategies well-represented
- Multiple mathematical techniques and problem framings
- Strong evidence of diverse exploration strategies

Score 5 - Maximum Diversity:
- No single approach dominates ($\leq$3 responses use same method)
- 4+ distinctly different solution strategies
- Wide variety of mathematical techniques and creative approaches
- Excellent exploration and generation diversity

IMPORTANT: Focusing on the DIVERSITY of the attempted approaches. Return ONLY a number from 1 to 5.

Modeller	AIME 2024	AIME 2025	AMC 2023	MATH-500	Minerva	Olimpiyat	Ortalama
$\beta = 5$	13.54	10.00	56.09	58.91	20.79	28.72	31.34
$\beta = 10$	14.79	10.20	59.53	64.30	25.27	32.39	34.41
$\beta = 15$	15.41	10.83	54.53	66.96	31.41	34.61	35.63
$\beta = 30$	15.00	10.83	50.62	69.02	30.03	35.03	35.09

Tablo 7 | FlowRL'de β parametresinin etkisine yönelik ayrıştırma çalışması. Farklı β değerleri için altı matematik muhakemesi kıyaslama setinde Ortalama@16 doğrulukları rapor edilmiştir.

Çeşitlilik Değerlendirme İsteği

Sistem: Bir matematik yarışması problemi için çözüm yaklaşımlarının ÇEŞİTLİLİĞİNİ değerlendiriyorsunuz. Farklı problem çözme stratejilerini gösteren en hafif METODOLOJİ FARKLILIKLARINI tespit etmeye odaklanın.
PROBLEM:
{problem}
16 ÇÖZÜM GİRİŞİMİ:
{formatted_responses}
DEĞERLENDİRME KRİTERLERİ - Çeşitliliği 1 ila 5 arasında derecelendirin:
Puan 1 - Minimum Çeşitlilik:
- 14'ten fazla cevap esasen özdeş yaklaşımlar kullanıyor
- Aynı matematiksel kurulum, aynı değişken seçimleri, aynı çözüm yolu
- Yalnızca önemsiz farklılıklar (aritmetik, notasyon, ifade)
- Üretim sürecinde çok düşük keşif/çeşitlilik olduğunu gösterir

Puan 2 - Düşük Çeşitlilik:
- 11-13 cevap aynı ana yaklaşımı kullanıyor
- 1-2 alternatif yaklaşım görülüyor ancak nadir
- Baskın yöntemde küçük varyasyonlar (farklı yer değiştirmeler, sıralamalar)
- Bir stratejiye güçlü eğilimle bir miktar keşif var

Puan 3 - Orta Düzey Çeşitlilik:
- 7-10 yanıt en yaygın yaklaşımı kullanıyor
- 2-3 belirgin alternatif yaklaşım mevcut
- Problem kurulumu veya matematiksel tekniklerde kayda değer varyasyon
- Mak reasonable keşfi gösteren dengeli karışım

Puan 4 - Yüksek Çeşitlilik:
- 4-6 yanıt en yaygın yaklaşımı kullanıyor
- 3-4 belirgin çözüm stratejisi iyi temsil ediliyor
- Birden fazla matematiksel teknik ve problem çerçevesi
- Çeşitli keşif stratejilerinin güçlü kanıtı

Puan 5 - Maksimum Çeşitlilik:
- Tek bir yaklaşım hakim değil ($\leq$3 yanıt aynı yöntemi kullanıyor)
- 4+ belirgin şekilde farklı çözüm stratejisi
- Çok çeşitli matematiksel teknikler ve yaratıcı yaklaşımlar
- Mükemmel keşif ve üretim çeşitliliği

ÖNEMLİ: Denenen yaklaşımların ÇEŞİTLİLİĞİNE odaklanın. Yalnızca 1 ile 5 arasında bir sayı döndürün.