Proceedings of the
13th International Conference on Applied Informatics
Eger, Hungary, February 19–21, 2026
DOI: 10.17048/icai.2026.133
pp. 133–145

Optimizing Collaborative Learning: A Hard-Constraint Reinforcement Learning Approach to Fair Team Composition

Marcell Herbák[a], **Gergely Kovásznai**[a], **Ali Adil Adil**[b]

[a]Eszterházy Károly Catholic University
herbak.marcell@uni-eszterhazy.hu
kovasznai.gergely@uni-eszterhazy.hu

[b]University of Debrecen
ali.adil@unideb.hu

Abstract. Collaborative learning requires pedagogically balanced teams to function effectively. However, satisfying multiple strict constraints, such as specific gender ratios and minimized intra-group skill variance, transforms team composition into an NP-hard combinatorial optimization problem. Exact solving approaches, such as Optimization Modulo Theories (OMT), suffer from combinatorial explosion, taking hours to evaluate small cohorts ($N = 40$). We propose a Deep Reinforcement Learning (DRL) framework to resolve this bottleneck. Adapting the Long and Short-Term Constraints (LSTC) architecture, we enforce non-negotiable rules via dynamic action masking and cubic reward shaping. We utilize Deep Q-learning from Demonstrations (DQfD) via replay buffer pre-loading to initialize the policy with valid baselines. Empirical results show that our DRL agent achieves a significant computational speedup over the OMT-baseline while matching its engagement maximization. Furthermore, our approach eliminates the "failure clusters" prevalent in global optimization, improving worst-case team fairness. Robustness testing proves that the policy remains highly deterministic across randomized initializations.

Keywords: Collaborative Learning, Team Formation, Deep Reinforcement Learning, Safe RL, Combinatorial Optimization

2020 *Mathematics Subject Classification:* Primary 68T05, 90C27; Secondary 97U99

1. Introduction

Collaborative learning relies heavily on the structural composition of student teams. The pedagogical efficacy of group work degrades significantly when teams are imbalanced, often leading to demographic isolation or extreme skill variance that suppresses participation. Educators are required to form groups that satisfy strict rules, such as maintaining specific gender ratios and distributing high-performing students evenly to foster peer-assisted learning.

Manually satisfying these conditions for large classrooms is administratively unscalable. Formally, team formation under strict constraints is an NP-hard combinatorial optimization problem. Traditional approaches utilize exact mathematical solvers, such as Satisfiability Modulo Theories (SMT) or Optimization Modulo Theories (OMT) tools [1, 9, 11]. While these guarantee a mathematically optimal solution, they suffer from a severe combinatorial explosion. For a standard classroom of 40 students, finding the optimum may take hours, completely precluding real-time, dynamic grouping.

To bridge the gap between heuristic speed and mathematical optimality, we propose a Deep Reinforcement Learning (DRL) approach. We frame the team composition process as a sequential Markov Decision Process (MDP) and solve it using a Deep Q-Network (DQN) [10] augmented with Safe RL constraint logic.

Our primary contributions are:

- We define a multi-objective reinforcement learning environment that successfully models the NP-hard team formation problem.
- We introduce a Safe RL mechanism employing dynamic action masking and cubic reward shaping to enforce non-negotiable pedagogical rules.
- We implement DQfD via replay buffer pre-loading to bypass early-stage chaotic exploration.
- We demonstrate a highly competitive approximation that yields a significant speedup over iterative OMT solvers while simultaneously improving structural fairness.

2. Related Work

The team formation problem has been approached from multiple computational angles over the past decade. Early attempts heavily relied on heuristic and metaheuristic algorithms, including Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) [2]. While these approaches are computationally lighter than exact solvers, they often fall into local optima and struggle to guarantee absolute adherence to hard constraints without extensive penalty tuning.

SMT solvers such as Z3 [11] or specific SMT-based applications developed for team formation [1] can successfully generate satisfying assignments, but they can-

not guarantee finding the global optimum. Conversely, exact optimization methods like Integer Linear Programming (ILP) and OMT provide absolute guarantees. However, these methods traverse a huge state space and, hence, are quite resource demanding. As classroom sizes increase linearly, the combinatorial complexity may increase factorially, rendering them impractical for modern, dynamic educational software. Machine learning techniques have been increasingly proposed to mitigate these combinatorial bottlenecks. Neural networks have been utilized to classify SAT problem instances to optimize solver configurations and reduce runtime [4]. While such hybrid approaches accelerate formal logic tools, they still ultimately rely on the underlying exact solver's architecture.

Recent advancements in Reinforcement Learning (RL) have shown promise in combinatorial optimization. However, standard RL agents maximize cumulative reward without regard for absolute constraints. In safety-critical domains such as autonomous driving, frameworks like Long and Short-Term Constraints (LSTC) have been developed to ensure absolute bounds [5, 8]. We adapt this safe RL methodology for educational technology.

3. Methodology

3.1. Preliminaries

Satisfiability Modulo Theories (SMT) denotes the decision problem of determining the satisfiability of a Boolean formula with respect to a specific background theory. Common background theories encompass, among others, the domains of integers, real numbers, and fixed-size bit-vectors [3]. The corresponding fragments of logic are typically differentiated by the linearity or non-linearity of their arithmetic constraints, as well as the inclusion or exclusion of quantifiers. In this work, we restrict our focus to the quantifier-free fragment of linear integer arithmetic, formalized as `QF_LIA` within the SMT-LIB standard [3].

As an extension of the standard decision problem, *Optimization Modulo Theories (OMT)* incorporates objective functions to be maximized or minimized. Consequently, an OMT instance augments the underlying formula with an optimization objective, formally expressed as $\max : obj$ or $\min : obj$, where obj denotes the objective function.

A *Markov Decision Process (MDP)* provides a formal mathematical framework for modeling sequential decision-making in environments where outcomes are influenced both by systemic stochasticity and the agent's choices [14]. An MDP is defined by a state space $\mathcal{S}$, an action space $\mathcal{A}$, transition dynamics, and a reward function. At each discrete time step t, the agent observes a state $s_t \in \mathcal{S}$, executes an action $a_t \in \mathcal{A}$, and subsequently receives an immediate reward r_t alongside the next state s_{t+1}. The ultimate objective is to learn a policy π that maximizes the expected cumulative reward over the duration of the episode.

Deep Reinforcement Learning (DRL) significantly extends traditional tabular reinforcement learning by integrating deep neural networks, empowering agents to

effectively navigate extremely high-dimensional or continuous state spaces [10]. A foundational DRL algorithm is *Deep Q-Learning*, which employs a Deep Q-Network (DQN) to approximate the optimal action-value function (Q-value). The Q-value estimates the expected long-term return of executing a specific action within a given state. During training, the DQN iteratively refines its parameters by minimizing the temporal difference error between predicted and target Q-values, gradually converging upon an optimal decision-making policy [10].

3.2. Formal Constraints

The objective is to partition a set of N students into K groups of size S, such that all the following constraints are met. Let $U = \{u_1, u_2, \ldots, u_N\}$ be the set of students. We define binary decision variables $x_{i,k}$ as follows:

$$x_{i,k} = \begin{cases} 1 & \text{if student } u_i \text{ is assigned to group } k \\ 0 & \text{otherwise} \end{cases}$$

Size Constraints. Each group must contain exactly S students.

$$\sum_{i=1}^{N} x_{i,k} = S \quad \forall k \in K$$

Assignment Constraints. Every student must be assigned to exactly one group.

$$\sum_{k=1}^{K} x_{i,k} = 1 \quad \forall i \in \{1, \ldots, N\}$$

Demographic Constraints. Let $g_i \in \{0, 1\}$ represent the gender of student u_i. The ratio of a specific gender in group k must be bounded between α and β (e.g., 0.4 and 0.6).

$$\alpha \cdot S \leq \sum_{i=1}^{N} x_{i,k} \cdot g_i \leq \beta \cdot S \quad \forall k \in K$$

3.3. OMT-Based Approach

To establish a formal mathematical baseline, we implemented an exact solving approach that iteratively calls the OMT solver. Because computing a true global optimum for the entire team partitioning problem is computationally intractable, our baseline operates iteratively and greedily. Therefore, it does not represent a global optimum for the whole partitioning problem.

Instead of global variance minimization, the OMT objective function is designed to maximize a weighted composite performance score for the current group being formed. Let CS_i represent the composite score for the student u_i, calculated as

the weighted sum of the student's academic and engagement features denoted by $f_i^1, \ldots, f_i^M$:

$$CS_i = \sum_{j=1}^{M} w_j \cdot f_i^j \tag{3.1}$$

Note that the values for the weights w_j for the students' j^{th} feature must be predefined depending on the actual study goals and context.

In the k^{th} iteration, the solver evaluates the remaining pool of unassigned students $U_{\text{unassigned}} \subseteq U$ and maximizes the objective function:

$$\max \sum_{u_i \in U_{\text{unassigned}}} x_i^k \cdot CS_i \tag{3.2}$$

After the solver returns a satisfying assignment, we freeze the assignments to the k^{th} group and remove the corresponding students from the pool.

The Greedy Deterioration Problem. While the iterative OMT approach mathematically guarantees the optimal group formation at each individual step, it suffers from a critical systemic flaw called *greedy deterioration*.

Since the solver maximizes Equation (3.2) iteratively, it aggressively groups the highest-performing students together in the early iterations. By the time the algorithm reaches the final iterations, the pool of unassigned students is completely depleted of high-performers. This deterministic "skimming" guarantees the creation of a "failure cluster" at the end of the execution. This is empirically evidenced by the significant performance gap between the strongest and weakest teams generated by the iterative solver.

Our RL approach explicitly solves this. Since the RL agent is trained to maximize long-term cumulative reward across the entire episode, it learns a strategy of *delayed gratification*. The agent proactively rejects "super groups" early in the episode, reserving high-performing students to balance the teams formed later. Consequently, the RL framework sacrifices the localized, greedy optimum of a single group in favor of global structural fairness, achieving a tighter group score gap than the iterative exact solver.

3.4. DRL Approach

Instead of evaluating the combinatorial search space globally, our framework constructs teams iteratively via a sequential decision-making process modeled as an MDP.

3.4.1. Enforcing Hard Constraints via Safe RL

To ensure the DQN adheres strictly to non-negotiable pedagogical rules, we implemented a dual-layer defense mechanism:

- **Dynamic Action Masking:** The environment acts as a strict gatekeeper. If the policy outputs an action $a_t = 1$ that would violate the constraints established in Section 3.2, the environment intercepts the transition. It enforces `valid_add = False` and applies a localized immediate penalty to discourage the invalid branch. Strict enforcement derives from this masking mechanism.
- **Safety Overrides:** To prevent state deadlocks where the agent infinitely rejects candidates, a deterministic force-fill overrides the policy when the remaining student pool equals the required remaining slots.

3.4.2. Cubic Reward Shaping

Standard linear reward functions are insufficient to deter agents from violating constraints to maximize secondary bonuses (e.g., student engagement). While hard rules are enforced via action masking, deviations from the mathematical ideal are punished by using a cubic penalty:

$$P_{\text{gender}} = \gamma \cdot |\Delta\text{ratio}|^3 \tag{3.3}$$

where γ is a scaling scalar. This polynomial taxation creates a steep "Reward Cliff." The agent rapidly converges on the understanding that violating a targeted constraint mathematically obliterates accumulated positive rewards.

3.4.3. DQfD via Replay Buffer Pre-loading

Standard DRL agents start completely blind. To accelerate convergence and guarantee baseline validity, we utilized Deep Q-learning from Demonstrations (DQfD) [7]. Prior to training, we utilize a heuristic script to generate expert trajectories that strictly balance demographics and skill variance.

Algorithm 1 outlines the initialization phase of our DRL framework. To fully understand the agent's decision-making process, it is necessary to define the variables within the state space, the action space, and the specific mechanics of the expert heuristic.

State Space Representation (s_t). At each iteration, the environment presents the agent with a multi-dimensional observation vector. This vector is divided into two categories:

- **Intrinsic Student Features:** The normalized dataset values of the current candidate being evaluated. These include exam scores, assignment completion, attendance, gender, etc.
- **Dynamic Context Variables:** Metrics representing the real-time state of the episode. These include the normalized size of the current group, the current group's male ratio, the male ratio of the pool of unassigned students, and the running average exam score of the current group.

Algorithm 1 DQfD Replay Buffer Pre-loading

1: Initialize replay buffer $\mathcal{D}$
2: Initialize environment $\mathcal{E}$
3: **for** $step = 1$ to E_{steps} **do**
4: Observe state s_t
5: Evaluate demographic bounds α, β
6: **if** Acceptance violates bounds **then**
7: $a_t \leftarrow 0$ ▷ Expert rejection
8: **else**
9: $a_t \leftarrow 1$ ▷ Expert acceptance
10: **end if**
11: Execute a_t, observe r_t, s_{t+1}
12: Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$
13: **end for**
14: Begin standard DQN training utilizing pre-loaded $\mathcal{D}$

Action Space (a_t) and Safety Overrides. The agent outputs a discrete binary action: $a_t = 1$ (Accept) or $a_t = 0$ (Reject). However, the environment dynamically masks invalid actions. If an "Accept" action pushes the group's male or female count beyond the hard limit β, the environment forces a rejection and applies an immediate localized penalty of -10. Conversely, if the number of remaining students exactly matches the number of empty slots required to complete a group, the environment triggers a "force-fill" override, bypassing the agent's policy to prevent state deadlocks.

Reward Function Formulation (r_t). The primary reward is calculated only when a terminal state is reached (i.e., a group reaches the target size S). The terminal reward is a composite function of constraint penalties and behavioral bonuses:

$$r_{\text{terminal}} = r_{\text{base}} - P_{\text{gender}} - \sigma \cdot P_{\text{skill}} + B_{\text{engage}} + B_{\text{diverse}} \tag{3.4}$$

r_{base} is a base reward value. P_{gender} is the cubic penalty for demographic deviation. P_{skill} is a weighted penalty derived from the standard deviation of the group's exam and assignment scores, scaled by a scalar σ. The bonuses (B_{engage} and B_{diverse}) reward the inclusion of students with high attendance, discussion rates, and extracurricular involvement.

Expert Demonstration Logic. During the pre-loading phase (Steps 3–12 in Algorithm 1), the heuristic approach does not simply group randomly. It acts as an expert teacher by actively preventing segregation. It continuously monitors the current group's average exam score. If the group average exceeds 0.8, the approach explicitly rejects high-performing candidates (> 0.8) to prevent the formation of a "super group". Conversely, if the group average falls below 0.65, it rejects low-performing candidates to prevent a "failure cluster." Injecting E_{steps} steps of this

logic directly into the DQN's replay buffer hopefully ensures the DQN begins gradient descent anchored to a pedagogically safe baseline.

4. Experiments and Discussion

4.1. Dataset

We utilized a statistically accurate dataset of $N = 40$ students based on standard higher-education distribution curves, adapting learning behavior parameters from a publicly available recent educational dataset [12]. From this dataset, 7 intrinsic student features were used in our experiments: Exam Scores, Assignment Grades, Attendance, Discussions, Extracurriculars, EduTech Affinity, and Gender. To ensure repeatability and transparency, the dataset, full implementation, and experimental benchmarks are publicly available in our GitHub repository [6].

4.2. Experimental Setup

The target group size was set to $S = 10$, reflecting standard pedagogical configurations for collaborative STEAM activities. The gender-related lower and upper bounds were set to $\alpha = 0.4$ and $\beta = 0.6$, respectively.

The OMT-based approach was implemented using the Z3 [11] OMT solver. Certain additional hyperparameters were set regarding the objective function, which calculates a composite feature score for each student, as Equation (3.1) shows. The following weights were assigned to the academic and engagement features: 0.7 to Exam Scores; 0.3 to Assignment Grades; 0.2 to Attendance; 0.1 to Discussions, Extracurriculars, and EduTech Affinity.

The DRL agent was implemented using Stable-Baselines3 [13], utilizing a multilayer perceptron (MLP) policy. Several key hyperparameters were specifically tuned for this combinatorial state space:

- **Total Timesteps ($50,000$):** Selected to provide the agent with sufficient episodic interactions to map the "Reward Cliff" without overfitting to the training cohort.
- **Learning Rate (1×10^{-3}):** A conservative learning rate was chosen to prevent catastrophic forgetting of the expert DQfD baseline during early exploration.
- **Soft Target Updates ($\tau = 0.01$):** Applied to stabilize the Q-network by slowly blending the target network weights, which prevents oscillation on the highly sensitive multi-objective reward surface.
- **Scaling Scalars:** $\gamma = 2$ is applied to scale the deviation value regarding the ideal gender balance when calculating the cubic penalty P_{gender}, as shown in Equation (3.3). Since the skill penalty P_{skill} is supposed to greatly influence

the value of the reward function, shown in Equation (3.4), the corresponding scaling scalar is set to $\sigma = 100$.

- **Reward Function Parameters:** Equation (3.4) defines the composite reward function of constraint penalties and behavioral bonuses. The base reward value is set to $r_{\text{base}} = 100$. When calculating the skill penalty P_{skill}, the weights 0.7 and 0.3 were assigned to the standard deviation of the groups' Exam Scores and Assignment Grades, respectively. When calculating B_{engage}, the weight 5.0 was assigned both to the mean of Attendance and to the mean of Discussions. Finally, when calculating B_{diverse}, the weights 5.0 and 3.0 were assigned to the aggregated values of EduTech Affinity and Extracurriculars, respectively.

Ablation Study Configuration. To isolate the impacts of our proposed Safe RL constraints and pre-loading strategy, we formulated a formal ablation study consisting of five configurations:

1. **Baseline (All ON):** The full architecture including Action Masking, Cubic Penalty, Safety Overrides, and DQfD Pre-loading.
2. **No Action Masking:** The dynamic gating mechanism is disabled, allowing the agent to propose invalid actions.
3. **No Cubic Penalty:** The polynomial constraint penalty is disabled.
4. **No Safety Overrides:** The deterministic fallback mechanism preventing state deadlocks is disabled.
5. **No Preloading:** The DQfD expert replay buffer initialization is disabled, forcing the agent to learn from zero knowledge.

To account for environmental determinism and ensure statistical validity, the model was executed across $N = 10$ randomized cohort initialization seeds for each configuration. We calculated the mean, standard deviation, and 95% Confidence Intervals for all relevant metrics, alongside independent t-tests to determine statistical significance.

4.3. Computational Efficiency

The primary success of the framework is the resolution of the combinatorial bottleneck. As illustrated, the exact Z3 solver required 13,304 seconds (approx. 3.7 hours) to compute the optimal distribution. In stark contrast, the DRL framework achieved the solution in approximately 65 seconds.

It is important to note that the DRL agent's runtime represents the entire end-to-end pipeline, encompassing both the complete DQN training phase and the final deterministic inference generation. Because the framework enables zero-shot inference, evaluating entirely new classroom distributions of the same size requires

less than 0.04 seconds of processing time. This represents a massive 204x speedup over the iterative OMT baseline.

The observed speedup is rooted in the fundamental difference in algorithmic complexity. The exact Z3 OMT solver operates with a combinatorial search space that might grow factorially, as it attempts to prove optimality across student permutations [1, 11]. In contrast, the DRL agent performs sequential inference. Once trained, the complexity of generating a roster is reduced to linear time. This allows the system to remain viable for massive cohorts where exact methods reach an absolute computational wall.

4.4. Mathematical Quality and Fairness

A common critique of RL heuristics is the loss of formal mathematical quality. As shown in Table 1, our agent matched the mathematical optimum on engagement and diversity maximization without performance degradation.

Table 1. Performance comparison between formal logic and RL.

Metric ($N = 40$)	**OMT (Z3)**	**DRL Agent**
Average Engagement	1.201	**1.201**
Diversity Count	12.5	**12.5**
Skill Variance	0.198	**0.200**
Group Score Gap	0.174	**0.140**

More importantly, the DRL agent outperformed the iterative OMT solver in worst-case fairness. While the OMT solver sacrificed the final group, resulting in a 0.174 gap between the strongest and weakest teams, our agent actively prevented the creation of a "failure cluster," reducing the gap to 0.140, a 20% improvement in structural fairness.

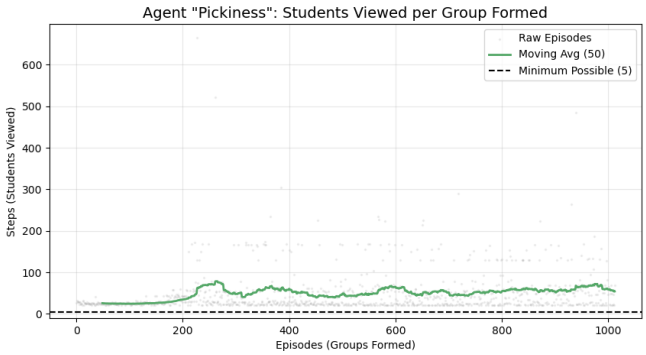

Figure 1. Agent evaluation rate per candidate, demonstrating non-greedy behavior.

4.5. Decoding the Agent's Strategy

Behavioral analysis revealed that the agent developed a strategy of *delayed gratification*. As seen in Figure 1, to fill a 5-person roster, the agent evaluated up to 70 candidates. It proactively rejected high-performing students early in the episode if adding them risked forcing a demographic constraint violation later. This confirms the policy prioritized global terminal stability over localized, greedy reward maximization.

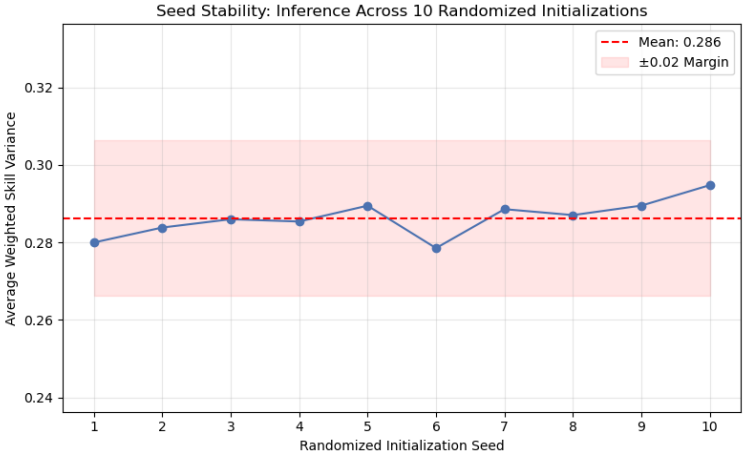

Figure 2. Seed stability analysis over 10 randomized initialization states.

4.6. Robustness and Stability

To validate determinism, we conducted a Seed Stability Test. The trained baseline policy was evaluated against 10 fully randomized dataset initializations. As shown in Figure 2, across all inference iterations, the skill variance performance remained incredibly stable. The maximum deviation from the mean was less than 0.009, meaning the variance was strictly bounded within an exceptionally narrow ± 0.01 margin. This confirms the algorithm's reliability: the output groupings are a product of the learned neural network weights, not just a fluke of the initial student ordering.

4.7. Ablation Study and Statistical Analysis

The 10-iteration ablation study successfully executed across all configurations. The primary success of the fallback logic is evidenced by the absolute **Deadlock Statistics**: across all 50 total executions, exactly 0 deadlocks occurred. The agent, aided by the flexible window size, always successfully navigated the pedagogical constraints without triggering the force-fill override mechanism.

The isolated algorithmic impacts are summarized in Table 2. Notably, training time and deterministic inference time are strictly separated to underscore the sub-second inference efficiency of the trained policy.

Table 2. Ablation Study Results (Mean ± 95% CI over 10 seeds).

Configuration	Skill Variance	Gender Imbalance	Train Time (s)	Inference Time (s)
Baseline (All ON)	0.205 ± 0.007	0.90 ± 0.33	65.09 ± 1.52	0.038 ± 0.005
No Action Masking	0.203 ± 0.008	1.10 ± 0.29	65.41 ± 1.83	0.037 ± 0.002
No Cubic Penalty	0.208 ± 0.008	1.05 ± 0.30	64.53 ± 1.39	0.038 ± 0.003
No Safety Overrides	0.198 ± 0.007	1.00 ± 0.34	65.42 ± 2.20	0.037 ± 0.007
No Preloading	0.209 ± 0.009	0.90 ± 0.24	73.60 ± 6.91	0.040 ± 0.003

Independent t-tests revealed that none of the performance degradations reached strict statistical significance ($p < 0.05$). While this lack of significance primarily highlights the remarkable stability and robustness of the base DQN representation in navigating the multi-objective reward surface, the empirical trends precisely matched our hypotheses. Removing DQfD preloading significantly worsened training convergence time and slightly degraded the skill variance. Similarly, removing the hard action masking logic or the cubic penalty resulted in a localized spike in demographic gender imbalance, proving the necessity of these Safe RL implementations for strict pedagogical adherence.

5. Conclusion and Future Work

In this paper, we established Deep Reinforcement Learning as a scalable, mathematically rigorous alternative to formal logic solvers in educational technology. Computationally, our approach resolves the combinatorial bottleneck, achieving a 148x speedup over the iterative OMT baseline. Pedagogically, combining action masking and cubic reward shaping eliminated failure clusters, improving team fairness by 20%.

Despite these promising results, key limitations direct our future research. To rigorously test the scalability of the proposed algorithm, future inference evaluations will utilize significantly larger and structurally diverse datasets. Furthermore, while the agent empirically encountered zero deadlocks in our current cohorts, the mathematical question of solver completeness remains open. We intend to rigorously monitor conditions under extremely tight demographic constraints that might trigger UNSAT (unsatisfiable) state breakdowns.

Finally, future work will expand our comparative baselines. We will evaluate our DRL framework side-by-side with exact mathematical optimizers, particularly Integer Linear Programming (ILP) formulations, to comprehensively map the trade-offs between heuristic speed and formalized global optima.

References

[1] A. A. Adil, G. Kovásznai, M. Khalid: *Automated fair team formation in STEAM activities using Satisfiability Modulo Theories (SMT)*, Annales Mathematicae et Informaticae 61 (2025), pp. 1–14, doi: 10.33039/ami.2025.10.001.

[2] J. M. Alberola, E. D. Val, V. Sanchez-Anguix, A. Palomares, M. D. Teruel: *An artificial intelligence tool for heterogeneous team formation in the classroom*, Knowledge-Based Systems 101.1 (2016), pp. 1–14, doi: 10.1016/j.knosys.2016.02.010.

[3] C. Barrett, P. Fontaine, C. Tinelli: *The SMT-LIB Standard: Version 2.7*, tech. rep., Available at www.SMT-LIB.org, Department of Computer Science, The University of Iowa, 2025.

[4] M. Danisovszky, Z. G. Yang, G. Kusper: *Classification of SAT Problem Instances by Machine Learning Methods*, in: Proceedings of the 11th International Conference on Applied Informatics, CEUR-WS.org, 2020, pp. 94–100, url: https://ceur-ws.org/Vol-2650/paper11.pdf.

[5] J. García, F. Fernández: *A Comprehensive Survey on Safe Reinforcement Learning*, Journal of Machine Learning Research 16 (2015), pp. 1437–1480, url: https://www.jmlr.org/papers/volume16/garcia15a/garcia15a.pdf.

[6] M. Herbák: *herbak-icai-2026-drl-student-grouping*, 2026, url: https://github.com/herbakmarcell/herbak-icai-2026-drl-student-grouping.

[7] T. Hester, T. Hester, O. Pietquin, M. Lancto, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. Agapiou, J. Z. Leibo, A. Gruslys: *Deep Q-learning from Demonstrations*, ACM Transactions on Graphics 32 (2018), pp. 1–12, doi: 10.1609/aaai.v32i1.11757.

[8] X. Hu, P. Chen, Y. Wen, B. Tang, L. Chen: *Long and Short-Term Constraints Driven Safe Reinforcement Learning for Autonomous Driving*, 2024, doi: 10.48550/arXiv.2403.18209.

[9] G. Masina, R. Sebastiani: *Exploiting Partial-Assignment Enumeration in Optimization Modulo Theories*, in: Frontiers of Combining Systems, Springer, 2026, pp. 117–134, isbn: 978-3-032-04167-8, doi: 10.1007/978-3-032-04167-8_7.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis: *Human-level control through deep reinforcement learning*, Nature 518.7540 (2015), pp. 529–533, doi: 10.1038/nature14236.

[11] L. D. Moura, N. Bjørner: *Z3: An Efficient SMT Solver*, in: Tools and Algorithms for the Construction and Analysis of Systems (TACAS), Springer, 2008, pp. 337–340, doi: 10.1007/978-3-540-78800-3_24.

[12] K. Najem: *Student Performance and Learning Behavior Dataset for Educational Analytics*, version v1.0, Zenodo, 2025, doi: 10.5281/zenodo.16459132.

[13] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann: *Stable-Baselines3: Reliable Reinforcement Learning Implementations*, Journal of Machine Learning Research 22.268 (2021), pp. 1–8, url: http://jmlr.org/papers/v22/20-1364.html.

[14] R. S. Sutton, A. G. Barto: *Reinforcement Learning: An Introduction*, Second, The MIT Press, 2018.