

# Safe, Efficient, and Robust Reinforcement Learning for Ranking and Diffusion Models

Shashank Gupta

University of Amsterdam

The Netherlands

27392shashankgupta@gmail.com

## Abstract

Reinforcement learning (RL) methods increasingly underpin modern ranking, recommendation, and generative systems, yet challenges of bias, variance, safety, and sample inefficiency hinder their deployment [Gupta et al., 2024a]. This dissertation advances contextual bandits and RL to address these challenges in two critical domains: learning-to-rank and text-to-image diffusion models.

The first part develops safe counterfactual learning-to-rank (CLTR) methods that leverage user interaction logs to learn improved ranking policies without online experimentation. Conventional inverse propensity scoring produces high-variance estimates [Gupta et al., 2024a], rendering learned policies unsafe to deploy. We introduce an exposure-based generalization bound that guarantees new policies perform at least as well as the logging policy with high confidence [Gupta et al., 2023]. Recognizing that guarantees depend on user model assumptions, we further develop a robust safety framework maintaining reliability even when real-world behavior deviates from assumed click models [Gupta et al., 2024c]. These methods substantially reduce unsafe initial warm-up periods in deployment ultimately making CLTR safe for deployment.

The second part addresses variance reduction and sample efficiency. For contextual bandits, we propose a unified framework that yields closed-form, variance-optimal baseline corrections for off-policy evaluation and optimization [Gupta et al., 2024b]. For diffusion models, we introduce Leave-One-Out PPO (LOOP), combining REINFORCE’s computational simplicity with PPO’s sample efficiency to improve aesthetic quality, and prompt instruction following while requiring fewer input samples [Gupta et al., 2025].

In summary, this thesis contributes theoretical guarantees, algorithms, and empirical findings that make RL methods safer, more reliable, and more efficient for ranking systems and diffusion models, supporting broader deployment in user-facing applications.

**Awarded by:** University of Amsterdam, Amsterdam, The Netherlands **on** 13 October 2025.

**Supervised by:** Maarten de Rijke, Harrie Oosterhuis.

**Available at:** <https://pure.uva.nl/ws/files/257088241/Thesis.pdf>.

---

## Selected Publications

- Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. Safe deployment for counterfactual learning to rank with exposure-based risk minimization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 249–258, 2023.
- Shashank Gupta, Philipp Hager, Jin Huang, Ali Vardasbi, and Harrie Oosterhuis. Unbiased learning to rank: On recent advances and practical applications. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1118–1121, 2024a.
- Shashank Gupta, Olivier Jeunen, Harrie Oosterhuis, and Maarten de Rijke. Optimal baseline corrections for off-policy contextual bandits. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 722–732, 2024b.
- Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. Practical and robust safety guarantees for advanced counterfactual learning to rank. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 737–747, 2024c.
- Shashank Gupta, Chaitanya Ahuja, Tsung-Yu Lin, Sreya Dutta Roy, Harrie Oosterhuis, Maarten de Rijke, and Satya Narayan Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*, 2025.