When LLM Meets DRL

Advancing Jailbreaking Efficiency via DRL-guided Search

(Xuan Chen, Yuzhou Nie, Wenbo Guo, Xiangyu Zhang), NeurIPS 2024

Sebastian Molina · CAP6619 · Spring 2025

Overview