

Key Insight: Steering Preserves Discrimination while Prompting Causes Over-Correction

