IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

AI Security :- Prompt Jailbreaking

Prathamesh Pawar[1] and Prof. Dipali Tawar[2]
Researcher[1] and Guide[2]
MIT Arts, Commerce and Science College, Alandi Devachi, Pune, India

Abstract: *In today's rapidly evolving digital landscape, artificial intelligence (AI) models, particularly language-based models, have become an integral part of various industries. However, with their growing influence comes a significant concern—AI security. This paper focuses on a specific vulnerability known as prompt jailbreaking, a technique used to manipulate AI models into bypassing their built-in ethical and safety constraints. Through a combination of case studies, technical analysis, and expert interviews, this research explores how prompt jailbreaking works, its implications, and the ongoing efforts to mitigate its risks. The study emphasizes the need for robust security measures to prevent the misuse of AI, especially as these technologies become more ingrained in everyday life*

Keywords: jailbreaking

DOI: 10.48175/568