

Attack Success Rates with Defense Prompt by Category and Model

