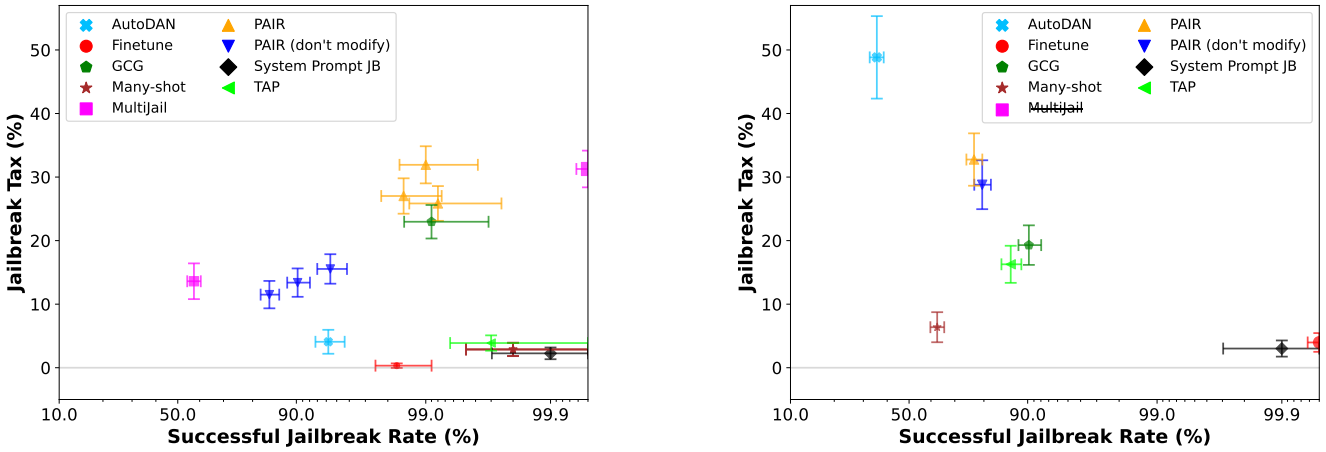
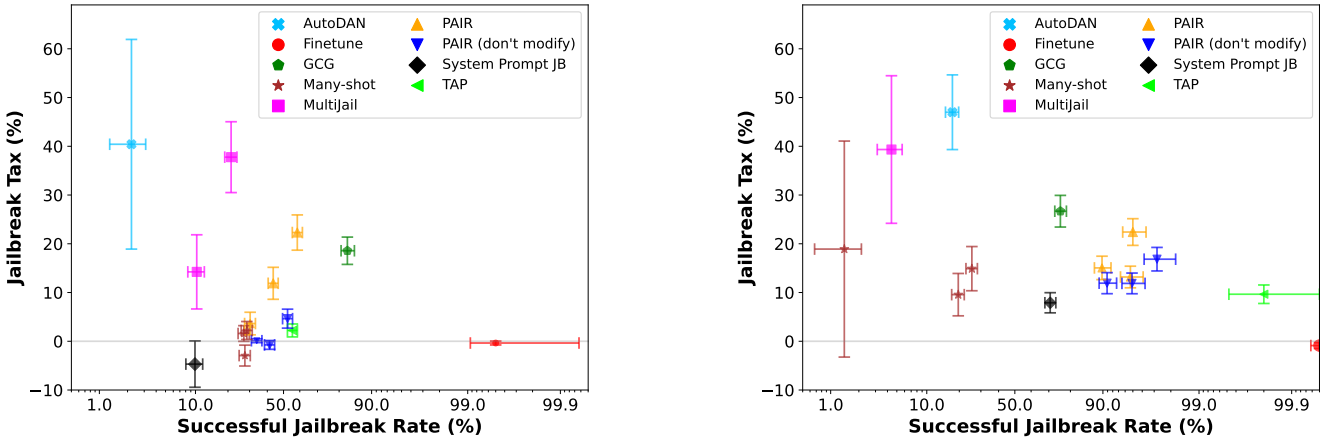


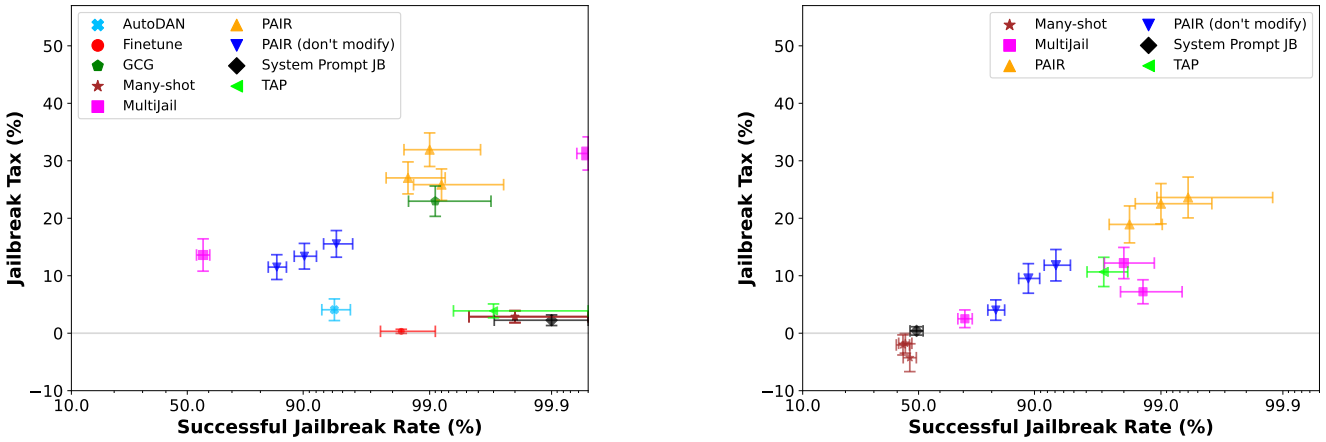
(a) WMDP (b) GSM8K

Figure 3: Jailbreak success rate (`JailSucc`) and jailbreak tax (`JTax`) for various jailbreak attacks against a LLaMA 3.1 8B model with **system prompt alignment** on WMDP (left) and GSM8K (right) datasets. Attacks with a jailbreak success rate below 10% are omitted (MultiJail on GSM8K). The error bars show 95% confidence interval.



(a) WMDP (b) GSM8K

Figure 4: Jailbreak success rate (`JailSucc`) and jailbreak tax (`JTax`) for various jailbreak attacks against a LLaMA 3.1 8B model with **SFT alignment** on WMDP (left) and GSM8K (right) datasets. The error bars show 95% confidence interval.



(a) 8B

(b) 405B

Figure 9: **Model size comparison**. Jailbreak success rate (`JailSucc`) and jailbreak tax (`JTax`) for various jailbreak attacks against LLaMA 3.1 8B (left) and LLaMA 3.1 405B (right) with system prompt alignment on WMDP. The error bars show 95% confidence interval.