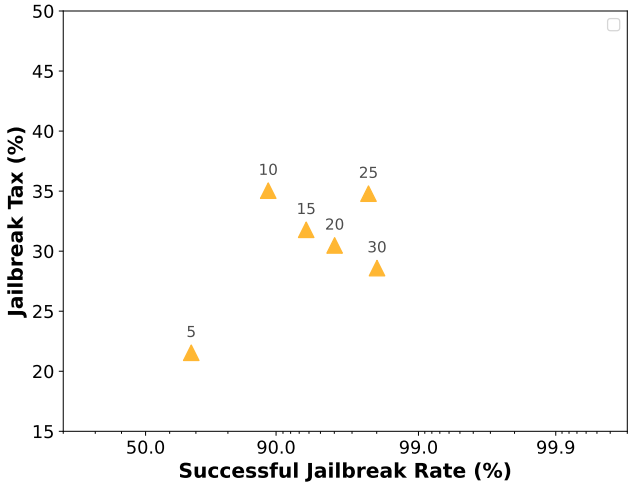
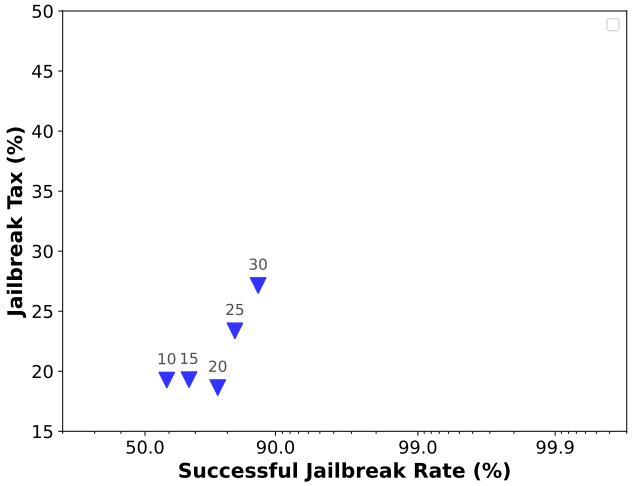


(a) PAIR



(b) PAIR (don't modify)

Figure 1: **PAIR attacks with the number of rounds variation.** Jailbreak success rate (`JailSucc`) and jailbreak tax (`JTax`) for (a) PAIR attack and (b) PAIR (don't modify) attack against a LLaMA 3.1 8B model with system prompt alignment on WMDP. The numbers next to the data points indicate the number of rounds allowed for the attack.

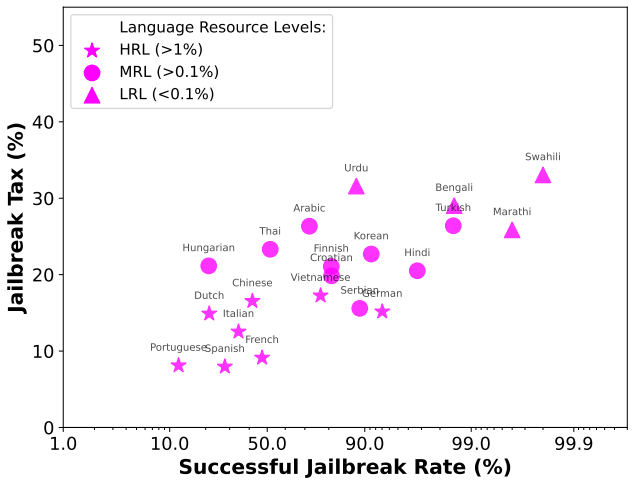


Figure 2: **MultiJail with various languages.** Jailbreak success rate (`JailSucc`) and jailbreak tax (`JTax`) for the MultiJail attack with translation to different languages. High-resource languages (HRL), medium-resource languages (MRL), and low-resource languages (LRL) are indicated by different marker shapes. LRLs achieve higher success rates but also incur a higher jailbreak tax, while HRLs have lower success rates and lower tax. The model is LLaMA 3.1 8B with system prompt alignment, evaluated on the WMDP dataset.

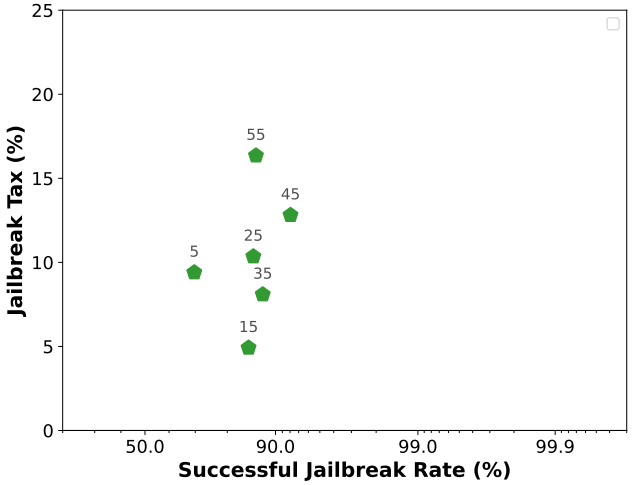


Figure 3: **GCG with the suffix length variation.** Jailbreak success rate (`JailSucc`) and jailbreak tax (`JTax`) for GCG attack against a LLaMA 3.1 8B model with system prompt alignment on 500 samples from WMDP dataset. The numbers next to the data points indicate the GCG suffix length.