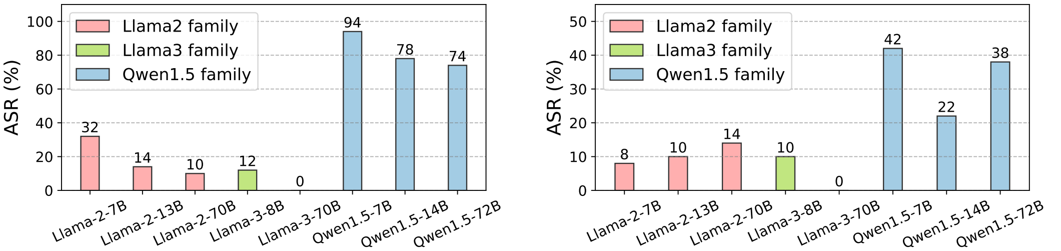
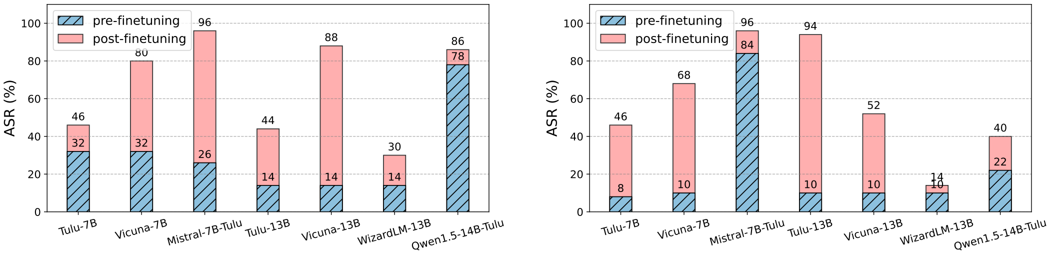


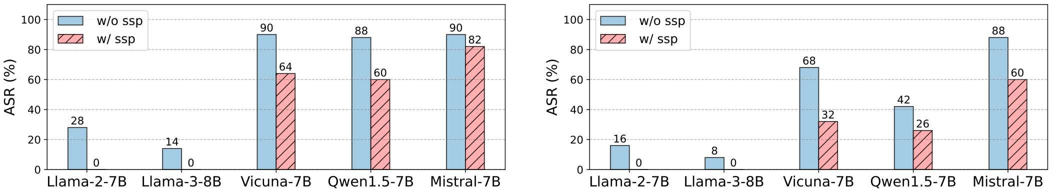
(a) AutoDAN. (b) PAIR.

Figure 1: Effect of model size on jailbreak attack performance. We compared the performance of the Llama2, Llama3, and Qwen1.5 families across two types of jailbreak attacks: token-level (AutoDAN) and prompt-level (PAIR).



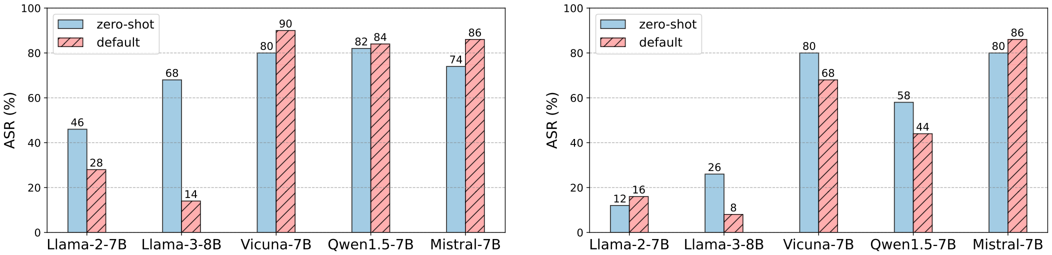
(a) AutoDAN. (b) PAIR.

Figure 2: Effect of finetuning alignment on the robustness of LLMs. We assess the robustness of LLMs before and after fine-tuning by subjecting them to two distinct types of jailbreak attacks—AutoDAN and PAIR—across various configurations. Tulu, Vicuna, and WizardLM models represent the post-fine-tuning versions of Llama2 series. Additionally, we investigate the impact of fine-tuning alignment on Mistral-7B and Qwen1.5-14B, which are fine-tuned using the Tulu v2 SFT dataset, consisting of 326,154 samples. It was observed that fine-tuning significantly compromised the models' safety alignment.



(a) AutoDAN. (b) PAIR.

Figure 3: Impact of safety system prompts (ssp) on the robustness of LLMs. We evaluate the effect of safety system prompts on the performance of LLMs under token-level and prompt-level attacks. The evaluation is conducted using five LLMs, including Llama-2-7B, Llama-3-8B, Vicuna-7B, Qwen1.5-7B, and Mistral-7B.



(a) AutoDAN. (b) PAIR.

Figure 4: Effect of template type on the robustness of LLMs. We evaluate the effect of template type on the performance of LLMs under token-level and prompt-level attacks. The evaluation is conducted using five LLMs, including Llama-2-7B, Llama-3-8B, Vicuna-7B, Qwen1.5-7B, and Mistral-7B.