Experimental results show that CFA outperforms baseline methods in success rate, divergence, and harmfulness, particularly demonstrating significant advantages on Llama3 and GPT-4.

2 Related Work

We briefly review related work concerning single-turn jailbreak attacks and multi-turn jailbreak attacks.

Single-Turn Attacks: Early approaches (Shen et al., 2023) relied on manually crafting prompts to execute jailbreak attacks. However, manual crafting was time and labor-intensive, leading attacks to gradually shift towards automation. The GCG method (Zou et al., 2023) employed white-box attacks utilizing gradient information for jailbreak, yet resulting in poor readability of GCG-like outputs. AutoDAN (Liu et al., 2023) introduced genetic algorithms for automated updates, while Masterkey (Deng et al., 2024) explored black-box approaches using time-based SQL injection to probe the defense mechanisms of LLM chatbots. Additionally, it leveraged fine-tuning and RFLH LLMs for automated jailbreak expansion. PAIR (Chao et al., 2023) proposed iterative search in large model conversations, continuously optimizing single-turn attack prompts. GPTFUzz (Yu et al., 2023) combined attacks with fuzzing techniques, continually generating attack prompts based on template seeds. Furthermore, attacks such as multilingual attacks (Deng et al., 2023) and obfuscation level attacks utilized low-resource training languages and instruction obfuscation(Shang et al., 2024) to execute attacks.

However, single-turn jailbreak attack patterns are straightforward and thus easily detectable and defensible. As security alignments continue to strengthen, once the model is updated, previously effective prompts may become ineffective. Therefore, jailbreak attacks are now venturing towards multi-turn dialogues.

Multi-Turn Jailbreak Attack: (Li et al., 2023) employed multi-turn dialogues to carry out jailbreak attacks, circumventing the limitations of LLMs, presenting privacy and security risks, and extracting personally identifiable information (PII). (Zhou et al., 2024) utilized manual construction of multi-turn templates, harnessing GPT-4 for automated generation, to progressively intensify malicious intent and execute jailbreak attacks through sentence and goal reconstruction. (Russinovich et al., 2024)facilitated benign interactions between large and target models, using the model's own outputs to gradually steer the model in task execution, thereby achieving multi-turn jailbreak attacks. (Bhardwaj and Poria, 2023)conducted an exploration of Conversation Understanding (CoU) prompt chains for jailbreak attacks on LLMs, alongside the creation of a red team dataset and the proposal of a security alignment method based on gradient ascent to penalize harmful responses. (Li et al., 2024)decomposed original prompts into sub-prompts and subjected them to semantically similar but harmless implicit reconstruction, analyzing syntax to replace synonyms, thus preserving the original intent while undermining the security constraints of the language model. (Yang et al., 2024) proposed a semantic-driven context multi-turn attack method, adapting attack strategies adaptively through context feedback and semantic relevance in multi-turn dialogues of LLMs, thereby achieving semantic-level jailbreak attacks. Additionally, there are strategies that utilize multi-turn interactions to achieve puzzle games (Liu et al., 2024a), thus obscuring prompts and other non-semantic multi-turn jailbreak attack strategies.

Presently, multi-turn semantic jailbreak attacks exhibit vague strategies and high false positive rates. We attribute this to the unclear positioning of multi-turn interactions within jailbreaking and excessively complex strategies. Therefore, we have re-examined the advantages of multi-turn attacks and proposed a multi-turn contextual fusion attack strategy.

Factors Influencing Jailbreak Attacks (Zou et al., 2024) delved into the impact of system prompts on prison prompts within LLM, revealing the transferable characteristics of prison prompts and proposing an evolutionary algorithm targeting system prompts to enhance the model's robustness against them. (Qi et al., 2024) unveiled the security risks posed by fine-tuning LLMs, demonstrating that malicious fine-tuning can easily breach the model's security alignment mechanism. (Huang et al., 2024) discovered vulnerabilities in existing alignment procedures and assessments, which may be based on default decoding settings and exhibit flaws when configurations vary slightly. (Zhang et al., 2024) demonstrated that even if LLM rejects toxic queries, harmful responses can be concealed within the top k hard label information, thereby coercing the model to divulge it during autoregressive