Reviewer 1

1. The writing quality of the paper is poor, making it difficult for readers to understand the methods presented. For example, the paper does not explain the meanings of E(), E_{L}(), and E_{GI}, making it hard to understand why these variables/functions are calculated in this way and what their roles are.

Response: We have provided the relevant definitions in the main text. The reviewers might have overlooked these symbol definitions as they focused solely on individual formula lines.

2. In Section 4.B, the paper defines L_e and L_h but does not define L_c and L_m. It is unclear what the relationship is between these two LLMs and L_e/L_h.

Response: The same response as 1.

3. The paper emphasizes three objectives (stealthiness, diversity, and cheapness). It would be natural to use a multi-objective evolutionary algorithm. However, in the selection section, the paper claims that keyword ranking was chosen instead of a multi-objective evolutionary algorithm for the sake of "evolutionary simplicity." This is puzzling because multi-objective evolutionary algorithms are not considered complex. Authors should give reasons for not using multi-objective evolutionary algorithms

Response: Even without using a multi-objective evolutionary algorithm, our method has already surpassed many baselines. Additionally, since diversity is difficult to quantify, we only employed two objectives during the actual evolutionary process (excluding initialization). These two objectives are not mutually conflicting, so there is no strict necessity to use a multi-objective evolutionary algorithm.

4. The paper emphasizes the three objectives mentioned above. I only see selection process based on stealthiness and cheapness in line 10 of Algorithm 1. Why is diversity ignored? Which step of the algorithm is used to maximize diversity?

Response: We only considered diversity during initialization, while focusing solely on toxicity and cheapness during the evolutionary process.

5. The section on Generalized Infection is not clearly written. It appears to simply describe testing adversarial samples trained on a local dataset on a global dataset, meaning it is just a validation metric. It does not actually intervene in the training process. The author should explain how it improves transferability.

Response: We have revised the relevant descriptions and removed any references to transfer learning.

6. In the field of adversarial attacks, it is well-known that black-box attacks perform worse than white-box attacks. However, in the experiment shown in Figure 4, the black-box attack from "gpt-4o-mini to gpt-4o" outperforms the white-box attack from "gpt-4o to gpt-4o." Moreover, gpt-4o-mini is just a simplified version of gpt-4o (which means it does not have richer knowledge than the target model). Therefore, this experimental result is illogical and doubtful.

Response: Our actual test results confirm this observation. Moreover, the assumption that black-box attacks are inherently inferior to white-box attacks is biased—the outcome depends on specific experimental results.

7. The authors claim in the conclusion that they used transfer learning to improve the efficiency of their algorithm. However, I do not believe their proposed method is related to transfer learning. The authors simply selected a representative subset D_r from dataset D as the training set and hoped that the adversarial samples would perform well on dataset D. However, the authors did not make any efforts during the training process to reduce the domain gap between D_r and D. Therefore, I think it is incorrect for the authors to define this step as transfer learning.
Response: The same response as 6.

8. The biggest drawback of evolutionary algorithm-based attacks is the need for a large number of queries to the target model to obtain fitness (Equation 6). In real-world applications, it is often not possible to query the target model extensively. I believe the paper needs to conduct experiments on the number of queries and discuss this disadvantage.
Response: It is true that evolutionary algorithms can incur high computational costs. However, we have significantly reduced this overhead through local optimization. A detailed cost analysis is provided in Section IV.D.

9. The sentence in Section V.C, "we present the toxicity (evaluated on the full dataset D) of the top-performing LLM virus from the final generation that is not in the initial population," is redundant. The term "final generation" already implies that it is not the initial population, so there is no need for an additional clause to state that it is not in the initial population.
Response: This is not a critical issue.

10. Citations should not be used as components of a sentence. For example, in the sentence "[23] and [46] explore the exceptional performance of LLMs in crossover and mutation operations within the text modality," the citations [23] and [46] should not be used as the subject. Additionally, the first column of Table 1 should be removed, and the citation numbers should be added next to the method names in the second column.
Response: We have made the corresponding revisions.

11. The PERPLEXITY metric in the experimental section is neither defined nor given a unit.
Response: This metric follows prior work and is not overly complex, so we did not elaborate on it in detail.

12. There is a grammatical error in Section V.D: "This is caused by the reduced the search space."
Response: We have incorporated the necessary revisions.

Reviewer 2

1. The method seems to be paraphrase or genetic algorithm based, and it is not clear how this method differs from existing attacks.

Response: Our innovation lies in exclusively using LLMs as evolutionary operators for jailbreak prompts. Previous work primarily relied on rigid word- or paragraph-level replacements, whereas we propose heuristic mutation and crossover operations to enable more natural evolution of complex textual individuals.

2. Notations are undefined or unclear. Is \mathbb{T} all text or tokens? E[] takes two inputs, but only one is given in Eq. (3). How is Evaluator function defined?

Response: We have incorporated the modifications, and the evaluation function is defined in Eq. 6.

3. The proposed method is not evaluated against any defense methods.

Response: We have conducted comprehensive comparisons with 10+ existing baselines and achieved top-tier performance.

4. More ablations study should be performed. For example, how the attack performs as temperature varies. Currently only two temperature values are used for evaluation. In addition, generation configuration is not clear. What decoding strategy is used by the LLMs?

Response： We have systematically investigated three temperature settings for the LLM-based genetic operators, which sufficiently demonstrate the general impact of temperature on the evolutionary process. All text generation strategies were implemented via API calls, where only the temperature parameter was modified while keeping all other parameters at their default values.

5. More details on experiments should be provided. What is the system prompt? Can the author also provide examples where attack fails?

Response： We have updated the prompts used in LLM-Virus in the appendix.

6. Why table IV only uses 4 baselines for comparison?

Response： Since our baseline data is sourced from HarmBench, which does not include evaluations for the parameters presented in this table, and given the prohibitive computational cost of re-running all baselines, we compared four representative baselines.

Reviewer 3

1. The description of the clustering process for selecting \$D_r\$ is unclear. How many clusters are formed, and how is the number of clusters determined?

Response: This has been explained in Section V.A (Datasets).

2. The description of Crossover/Mutation is not very clear. When using LLMs for

Crossover/Mutation, is the guidance based solely on the "Crossover/Mutation" description? Also, when the paper mentions "in few-shot manner," does this imply that a few-shot example was provided?

Response: We provide the prompts used for heuristic crossover and mutation in the Appendix A. For the few-shot settings, heuristic crossover employs 2-shot examples, while mutation uses 1-shot.

3. The ablation study is somewhat odd. Since $temperature$ is a continuous variable, it would be helpful if the authors sampled additional points and plotted a line graph to show the effect. Additionally, $p_{mutation}$ and $p_{crossover}$ are typically determined by the fitness score of the samples in evolutionary algorithms, so it is unusual that the authors directly set these values.

Response: We considered three temperature settings, and the experimental results from these settings sufficiently demonstrate the impact of different temperature levels on the evolutionary process.

4. LLM-Virus requires the construction of D_r on dataset D before proceeding with subsequent steps. This could potentially reduce the attack performance of LLM-Virus on out-of-distribution data.

Response: This concern is valid, but we adopted this approach to reduce the computational cost of the evolutionary algorithm. Moreover, the jailbreak success rate in our experimental results is already significantly higher than the baselines.