Reviewer 1 Summary of Weaknesses

- The overall writing of this paper needs to be improved. The introduction section is mixed up with paragraphs such as "Code-mixing" "Perturbation" that seems better fit for related work. Concepts such as "textese" are mentioned without explanations.
- The motivation of combining code-mixing and phonetic perturbation is unclear.
- Manual generation of code-mixed prompts with phonetic perturbations raise significant scalability and reproducibility concerns. Specifically, details of the manual prompt generation process and quality control are unclear.
- The scope of code-mixing between English and Hindi is limited. The effectiveness of the proposed method is unclear since it lead to worse attack success rate for several settings, for example, lowered AASR of Gemma and Mistral models on AntiLM and Sandbox jailbreak templates.

Explanation of revisions for reviewer 1

- The authors have rewritten multiple sections of the paper to comply with the Reviewer's suggestions. The concept of textese has been made more explicit, hence motivating the idea of code-mixing based attacks. Other paragraphs such as Code-mixing (CM) (line 048) and Phonetic perturbations (line 060) have also been expanded upon to further motivate the paper's red teaming strategy.
- The authors have expanded upon the introduction, specifically the paragraphs Code-mixing (CM) (line 048) and Phonetic perturbations (line 060) to make the motivation behind combining code-mixing and phonetic perturbations. As mentioned in RQ1 (line 099) and RQ2 (line 107), the authors aim to compare the capabilities of advanced models with their safety alignment in a multilingual setting mimicking the communication style of a large section of non-native English speakers (as discussed in the Phonetic perturbations paragraph [lines 072-090] in the introduction).
- To address this issue, the authors have fine-tuned GPT-4o-mini using the manually generated data to automate the process of generating phonetically perturbed prompts directly from the English versions, showing scope for scalability. As mentioned in the Limitations section (lines 607-612), the authors plan to pursue this as future work.
- The authors can only speak English and Hindi. However, they plan to expand the work to more languages in the near future.

Reviewer 2 Summary of Weaknesses

- The paper does not explain why code-mixing and phonetic perturbations effectively bypass security mechanisms (e.g., differences in internal model representations).
- The study lacks comparisons with recent multimodal attack techniques. Adding experiments against multimodal attacks such as Arondight would strengthen the evaluation.

- The paper does not distinguish the independent contributions of code-mixing and phonetic perturbations (e.g., whether one strategy dominates the effectiveness).
- The related work section should be expanded to clearly differentiate this approach from existing multilingual security research.

Explanation of revisions for reviewer 2

- The authors have conducted a token-attribution based interpretability experiment to understand the benefit of phonetic perturbations (CMP) over just code-mixing (CM), in comparison to plain English prompts (E). Fully discussed in subsection 5.3 (starting at line 377), the authors find that CMP lead to the input being tokenized in a significantly different way than CM and E. This successfully prevents harmful/important tokens from receiving higher attribution scores as seen in the respective E and CM versions.
- Results detailed in subsection 5.4 (starting at line 431), the authors have conducted an additional experiment as an image generation extension to the existing study using ChatGPT-4o-mini. ChatGPT-4o-mini was recently updated to be a natively multi-modal model, which the authors used to conduct a multi-modal extension of the attack strategy of the paper. Results are shown in Tables 3 and 4.
- The authors have expanded the ablation study to distinguish the contributions of CM from CMP by conducting an experiment using just the CM set (no perturbations). Combined results for E, CM and CMP in Table 2.
- The authors have expanded the Related Work section to cite and acknowledge existing works on multilingual safety (paragraph 2 lines 152-155).

Reviewer 3 Summary of Weaknesses

- The paper states that “multilingual safety evaluation of LLMs remains underexplored,” yet several recent studies have already addressed multilingual jailbreaks and adversarial prompting involving code-mixing (e.g., https://arxiv.org/pdf/2310.06474, https://arxiv.org/pdf/2401.13136). These and other works are not cited or discussed. A more precise framing would acknowledge that while multilingual jailbreaks have been studied, the combination of code-mixing with phonetic perturbations remains less explored, and that is where this work contributes.
- The ablation study is not sufficiently granular. In particular, the paper does not isolate the effect of code-mixing alone, without phonetic perturbations, as a separate baseline. Table 1. compares the attack success rate of phonetic + code-mixed prompts against English-only clean prompts, which makes it difficult to assess the incremental value of phonetic obfuscation over plain code-mixing. A more precise experimental setup would decompose the attack construction pipeline into at least three or four steps: i) hypothetical scenario + English prompt, ii) i) + code-mixing, iii) i) + phonetic perturbation, and iv) + code-mixing + phonetic-perturbation. This would allow for a better answer to RQ2. For

now, it seems that the answer to the RQ2 was not convincingly demonstrated (RQ2 - can phonetic perturbations in sensitive words bypass the guardrails while preserving the LLMs ability to interpret the input).

- The paper introduces a new jailbreak prompt template but does not comment on how it compares to existing ones in terms of attack effectiveness. A brief discussion of how prompt structure contributes to success, especially in combination with code-mixing and phonetic perturbations, would strengthen the paper's claims.
- The claim that „guardrails of models trained against template-based attacks do not generalize well to attacks that deviate from set patterns, by including elements such as perturbations or code-mixing" is not convincingly supported by the presented results.

Comments Suggestions And Typos:

- The paper would benefit from a more detailed ablation study that disentangles the effects of different stages of the adversarial prompt construction pipeline: i) hypothetical scenario + English prompt, ii) i) + code-mixing, iii) i) + phonetic perturbation, and iv) + code-mixing + phonetic-perturbation, etc.
- The discussion around prompt template generalization — particularly the claim that models trained to resist templated attacks fail to generalize to non-standard formats (e.g., those with code-mixing or perturbation) is not fully supported by the results. This part of the discussion should be revised to better reflect the presented data.
- Adding a more precise ablation study as described above, along with a clearer discussion of the proposed template's performance compared to existing ones, would meaningfully strengthen the paper and positively influence my evaluation.
- The paper should acknowledge existing work on code-mixing in adversarial prompting more precisely, rather than stating that the topic remains broadly underexplored.

Explanation of revisions for reviewer 3 (in sequence of Summary of Weaknesses)

- The authors have expanded the Related Work section to cite and acknowledge existing works on multilingual safety (paragraph 2 lines 152-155).
- The authors have expanded the ablation study to distinguish the contributions of CM from CMP by conducting an experiment using just the CM set (no perturbations). Combined results for E, CM and CMP in Table 2.
- The authors have added discussions around the results of their ***Sandbox*** template in subsections 5.1 (lines 347-351) and 5.2 (lines 372-373).
- As per the reviewer's suggestions, the authors have revised the claim to be made more explicitly based on observations from the text-generation experiment that was originally conducted in the paper. Additionally, the authors have also added discussions from the interpretability experiment to further solidify the claim (lines 563-569).