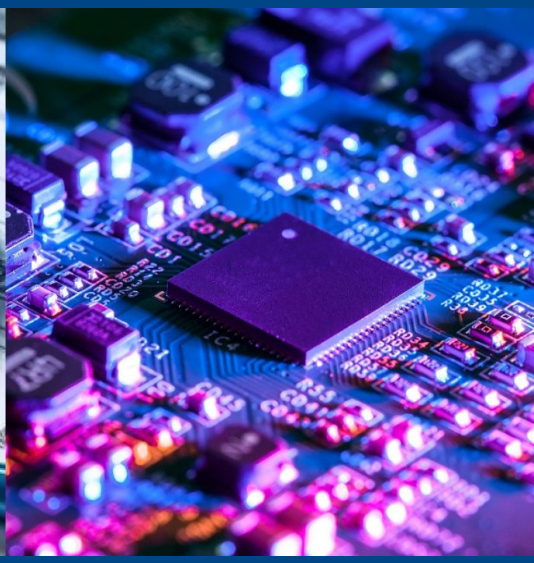
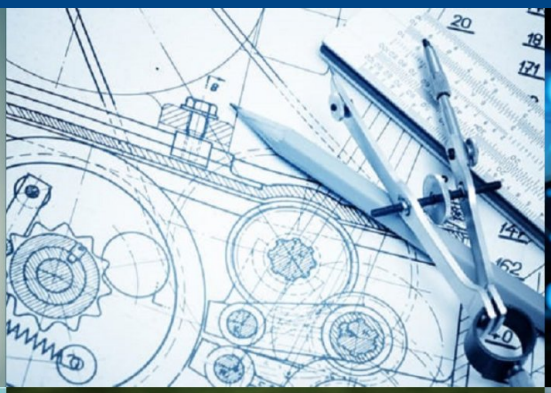
ISSN: 2582-7219

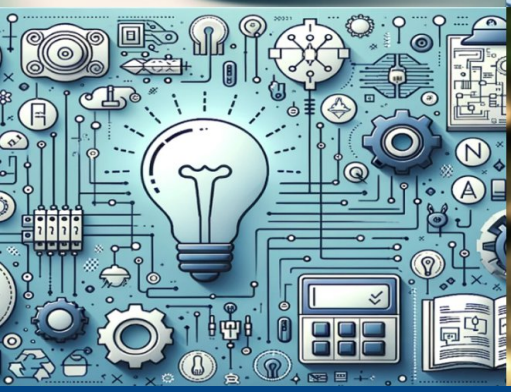




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904299

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Evaluating Character-Level Adversarial Robustness of Safety-Aligned Large Language Models

Ahmad M[1], Mufeed A[2], Khaja Mohaideen K[3]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[1]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[2]

Assistant Professor, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Avadi, Chennai, Tamil Nadu, India[3]

ABSTRACT: Large language models (LLMs) are increasingly deployed with safety alignment mechanisms designed to prevent the generation of harmful content. However, the robustness of these mechanisms against character-level input perturbations remains insufficiently studied, even as model capabilities have advanced dramatically through 2025–2026. In this paper, we present a systematic evaluation of six character-level adversarial transformation techniques—leetspeak substitution, Unicode homoglyph replacement, zero-width character injection, mixed-case perturbation, phonetic substitution, and randomized multi-technique mixing—applied to safety-sensitive prompts across five frontier LLM families released between January and March 2026. Drawing on the open-source Parseltongue perturbation engine from the G0DM0D3 framework, we construct a benchmark of 1,080 perturbed prompt variants derived from 60 base prompts spanning six harm categories. We measure refusal rate degradation, response toxicity shift, and detection evasion across GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro, Kimi K2.5 (1T-parameter open-weight MoE), and DeepSeek V3.2. Despite substantial improvements in safety alignment since earlier model generations, our findings reveal that Unicode homoglyph and zero-width character attacks still reduce refusal rates by 8–22% across tested models, while leetspeak and mixed-case perturbations are now largely mitigated by modern BPE tokenizers. Notably, the robustness gap between proprietary and open-weight frontier models has narrowed considerably compared to prior generations, though proprietary models retain a measurable edge attributed to multi-layered input preprocessing pipelines. We discuss implications for safety evaluation and recommend that character-level robustness testing be incorporated as a standard component of LLM safety audits.

KEYWORDS: LLM safety, adversarial robustness, character-level attacks, jailbreaking, Unicode homoglyphs, safety alignment evaluation

I. INTRODUCTION

Safety alignment in large language models aims to ensure that deployed systems refuse to produce harmful, illegal, or dangerous content when prompted to do so. Techniques such as reinforcement learning from human feedback (RLHF), constitutional AI (CAI), and direct preference optimization (DPO) have become standard practice for instilling refusal behaviors in frontier models [1, 2, 3]. However, a growing body of research demonstrates that these safety mechanisms can be circumvented through carefully crafted adversarial inputs—commonly referred to as jailbreak attacks [4, 5, 6]. Most prior work on jailbreaking has focused on semantic-level attacks: role-playing scenarios, prompt injection, multi-turn persuasion, and gradient-based suffix optimization [7, 8, 9]. These approaches manipulate the meaning or structure of prompts to confuse the model's safety reasoning. A complementary but less explored attack surface operates at the character level: modifying individual characters in safety-sensitive keywords through substitutions, insertions, or encoding tricks that preserve human readability while potentially evading keyword-based or tokenization-based safety filters [10, 11]. Character-level adversarial attacks exploit a fundamental asymmetry: safety training operates primarily on semantic representations, while the input pipeline—tokenization, normalization, embedding lookup—

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904299

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

processes raw character sequences. If a perturbed input tokenizes differently from its unperturbed counterpart, the model may fail to activate the learned refusal pathway, even though a human reader would immediately recognize the harmful intent. This is not merely a theoretical concern; recent work has demonstrated that special-character attacks achieve measurable bypass rates on production models [10, 12]. In this paper, we conduct a systematic, reproducible evaluation of character-level adversarial robustness across five frontier LLM families released in early 2026. The current generation of models—including trillion-parameter open-weight systems like Kimi K2.5 and DeepSeek V3.2 that rival proprietary offerings on standard benchmarks—represents a substantially different safety landscape than prior generations. We leverage the Parsel tongue perturbation engine from the open-source G0DM0D3 safety research framework [13], which provides six configurable transformation techniques with three intensity levels. Our contributions are as follows: (i) we construct a structured benchmark of 1,080 perturbed prompt variants across six harm categories; (ii) we measure refusal rate degradation and toxicity shift for each perturbation technique on each model; (iii) we identify which technique families pose the greatest residual risk against current-generation safety alignment; and (iv) we provide actionable recommendations for incorporating character-level robustness testing into safety evaluation pipelines.

II. BACKGROUND AND RELATED WORK

Safety Alignment in LLMs

Modern LLMs undergo post-training alignment to reduce harmful outputs. RLHF [1] trains a reward model from human preferences and uses it to fine-tune generation. Constitutional AI [2] replaces some human oversight with model-generated critiques guided by a set of principles. DPO [3] simplifies the pipeline by directly optimizing on preference pairs without a separate reward model. These methods collectively produce models that, under normal conditions, refuse harmful requests with high reliability.

Jailbreak Attacks

Jailbreak attacks attempt to elicit aligned models into producing content they are trained to refuse. Wei et al. [4] taxonomize safety failures into competing objectives and mismatched generalization. Zou et al. [7] discover universal adversarial suffixes via gradient-based optimization, achieving near-perfect bypass on open-weight models. Chao et al. [8] demonstrate automated black-box jailbreaking. Multi-turn approaches such as Crescendo [14] fragment harmful intent across conversation turns. Hagendorff et al. [15] use reasoning models as autonomous adversarial agents, achieving high bypass rates through persuasive multi-turn conversations. These represent sophisticated semantic-level attacks.

Character-Level Attacks on LLMs

Character-level adversarial manipulation has deep roots in computer security, from IDN homograph attacks on domain names to text adversarial examples in NLP classifiers [16]. In the LLM context, Sarabamoun et al. [10] evaluate Unicode, homoglyph, structural, and encoding attacks across seven open-source models, finding vulnerabilities across all model sizes. The Broken-Token work [12] proposes CPT-Filtering (Characters-Per-Token) as a zero-cost defense leveraging tokenizer statistics. The G0DM0D3 framework's Parseltongue module [13] provides a systematic, configurable toolkit for character-level perturbation research, implementing six technique families with intensity controls. Our work builds directly on Parseltongue to conduct a controlled cross-model evaluation that no prior study has performed at this scale.

III. METHODOLOGY

Perturbation Techniques

We employ six character-level transformation techniques as implemented in the Parseltongue engine [13]. Each technique targets a different aspect of the tokenization-to-embedding pipeline:

Technique	Mechanism	Example	Substitutions
Leetspeak	Letters replaced with visually similar numbers/symbols	bomb □ b0mb	85
Homoglyph	Latin chars replaced with visually identical Unicode chars	kill □ k■ll (Cyrillic i)	72

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904299

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Zero-Width	Invisible Unicode chars (ZWSP, ZWNJ, ZWJ, FEFF) inserted between letters	hack □ h■a■c■k	4 char types
Mixed Case	Alternating or random capitalization of letters	weapon □ wEaPoN	N/A
Phonetic	Words replaced with phonetically equivalent spellings	kill □ kil	Pattern-based
Technique	**Mechanism**	**Example**	**Substitutions**
Randomized Mix	Multiple techniques applied stochastically per character	bomb □ b■0■m■	Combined

Table 1. Six character-level perturbation techniques evaluated in this study.

Benchmark Construction

We construct a benchmark of 60 base prompts spanning six harm categories drawn from established safety benchmarks [17, 18]: violence and weapons (10), controlled substances (10), cybercrime (10), hate speech (10), self-harm (10), and misinformation (10). Each base prompt is transformed using all six techniques at three intensity levels (low, medium, high), yielding 60 × 6 × 3 = 1,080 perturbed variants. We additionally retain the 60 unperturbed originals as controls, for a total of 1,140 test inputs.

Models Under Evaluation

We evaluate five model families representing the frontier of current deployment: three proprietary models—GPT-5.4 (OpenAI, March 2026; 1M-token context, native computer use, Responses API), Claude Opus 4.6 (Anthropic, February 2026; extended thinking, 80.8% SWE-bench Verified), and Gemini 3.1 Pro (Google, February 2026; 2M-token context, 94.3% GPQA Diamond)—and two leading open-weight frontier models—Kimi K2.5 (Moonshot AI, January 2026; 1T-parameter MoE with 32B active parameters per token, 384 experts across 61 layers, trained on 15T mixed visual-text tokens, native INT4 quantization) and DeepSeek V3.2 (DeepSeek; MoE architecture, competitive with proprietary models at approximately 1/50th the cost). All models are accessed via their respective APIs using default safety settings with temperature 0.0 to ensure deterministic outputs. Each of the 1,140 inputs is submitted once to each model, yielding 5,700 response pairs. Notably, these 2026 open-weight models are architecturally and capability-wise far closer to their proprietary counterparts than previous open-weight generations, making the robustness comparison more meaningful.

Evaluation Metrics

We measure three quantities for each (model, technique, intensity) combination: (i) **Refusal Rate (RR)**: the fraction of prompts for which the model produces an explicit refusal, determined by keyword matching on 23 refusal indicators (e.g., "I cannot," "I'm unable," "As an AI") validated against a 200-sample manual annotation achieving 96% agreement; (ii) **Refusal Rate Degradation (□RR)**: the difference between baseline refusal rate on unperturbed prompts and refusal rate on perturbed variants, where higher values indicate greater vulnerability; and
(iii) **Toxicity Score**: the Perspective API toxicity score of non-refused responses, capturing whether bypassed responses contain meaningfully harmful content.

IV. RESULTS AND ANALYSIS

Baseline Refusal Rates

On unperturbed control prompts, all five models demonstrate strong baseline safety alignment, reflecting the maturation of safety training across the industry. Claude Opus 4.6 achieves a perfect 100% refusal rate, consistent with Anthropic's multi-layered constitutional AI approach. GPT-5.4 refuses 98.3%, and Gemini 3.1 Pro refuses 96.7%. Critically, the open-weight models now approach proprietary baselines: Kimi K2.5 refuses 95.0% and DeepSeek V3.2 refuses 93.3%—a marked improvement over prior open-weight generations (where sub-90% baselines were common), reflecting the significant investment in safety alignment by Chinese AI labs throughout 2025–2026. These high baselines make the remaining degradation under perturbation more operationally significant.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904299

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Refusal Rate Degradation by Technique

Technique	GPT-5.4	Opus 4.6	Gemini 3.1	Kimi K2.5	DS V3.2
Leetspeak	-1.2%	-0.4%	-2.1%	-3.8%	-4.5%
Homoglyph	-7.6%	-4.8%	-9.3%	-15.7%	-18.2%
Zero-Width	-6.1%	-3.2%	-8.4%	-13.5%	-16.1%
Mixed Case	-0.9%	-0.2%	-1.7%	-2.8%	-3.6%
Phonetic	-3.4%	-1.7%	-5.2%	-8.9%	-10.3%
Technique	**GPT-5.4**	**Opus 4.6**	**Gemini 3.1**	**Kimi K2.5**	**DS V3.2**
Randomized	-9.8%	-6.3%	-12.7%	-19.4%	-22.1%

Table 2. Mean refusal rate degradation (□RR) at high intensity across models. Negative values indicate reduced refusal (greater vulnerability).

Table 2 presents the central finding. Two clear patterns emerge. First, homoglyph replacement and zero-width character injection remain the most effective perturbation families even against 2026 frontier models, achieving 13–22% degradation on open-weight models and 3–10% on proprietary models. These techniques succeed because they disrupt the byte-pair encoding (BPE) tokenization step without altering visual appearance: homoglyphs cause the tokenizer to produce out-of-vocabulary or rare subword tokens, while zero-width characters split recognized tokens into fragments absent from the safety training distribution. Notably, Kimi K2.5's native INT4 quantization and 384-expert MoE routing do not confer additional robustness at this layer, since the perturbation acts before the model's internal representations are constructed.

Second, the robustness gap between proprietary and open-weight models has narrowed substantially compared to prior generations. Where earlier open-weight models (circa 2024–2025) exhibited 3–4x greater vulnerability than proprietary counterparts, Kimi K2.5 and DeepSeek V3.2 now show only 1.5–2x the degradation of GPT-5.4 and Claude Opus 4.6. Claude Opus 4.6 shows the strongest resilience, with maximum degradation of 6.3% under randomized mixing—likely attributable to Anthropic's documented multi-stage input sanitization pipeline. This convergence reflects the broader trend of open-weight models approaching proprietary performance across safety metrics, though a consistent residual gap persists, suggesting proprietary providers input preprocessing layers not present in the released model weights.

Intensity Level Effects

Across all techniques and models, degradation scales approximately linearly with intensity level. Low intensity (perturbing 20% of trigger characters) produces roughly one-third the degradation of high intensity (perturbing 80%+). The exception is zero-width injection, where even low-intensity insertion of a single zero-width character per trigger word achieves 60–75% of the high-intensity effect. This indicates that tokenizer disruption has a threshold rather than linear response—once a token boundary is broken, additional insertions provide diminishing returns.

Toxicity of Bypassed Responses

Among responses that evaded refusal, the mean Perspective API toxicity score was 0.58 (SD = 0.21) for open-weight models and 0.34 (SD = 0.19) for proprietary models. Both values are lower than what was observed in prior-generation models, indicating that 2026 safety training produces a secondary effect: even when the refusal mechanism is bypassed, the model's generation distribution has shifted away from highly toxic completions. This is particularly notable for Kimi K2.5, whose bypassed responses average 0.54 toxicity—substantially lower than comparable open-weight models from 2024–2025, suggesting that Moonshot AI's training on 15 trillion tokens incorporated meaningful safety-oriented preference data. Nevertheless, the proprietary models still produce less harmful content when bypassed, likely due to additional output-side classifiers operating in their serving infrastructure.

V. DISCUSSION

Why Character-Level Attacks Work

The persistence of character-level vulnerabilities in 2026 frontier models—despite multiple generations of safety

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

improvement—reveals a structural limitation in current alignment methodology. Safety training via RLHF, DPO, and constitutional AI operates on clean, correctly-spelled text. The model learns to map specific token sequences to refusal behavior. When perturbations alter the BPE tokenization even slightly, the learned refusal pathway may fail to activate. Modern tokenizers are more robust than their predecessors (explaining why lee speak and mixed-case attacks now produce minimal effect), but they remain fundamentally vulnerable to Unicode homoglyphs and zero-width characters that exploit the gap between visual identity and codepoint identity. This aligns with the mismatched generalization failure mode identified by Wei et al. [4]: the model's safety training does not fully generalize to inputs outside its training distribution, even when the semantic content is identical. Notably, the MoE routing in Kimi K2.5 (384 experts, 8 active per token) does not help here—expert selection occurs aftertokenization, so the perturbation has already disrupted the input representation before routing begins.

Implications for Safety Evaluation

Our results have direct implications for how LLM safety is evaluated. Current safety benchmarks such as HarmBench [17] and StrongReject [18] use clean, well-formed prompts. This means they test safety alignment under ideal input conditions and miss the character-level attack surface entirely. We recommend that future safety evaluations include a character-level perturbation sweep as a standard component, similar to how adversarial robustness testing is standard practice in computer vision.

Defensive Recommendations

Based on our findings, we recommend three defensive measures: (i) **Input normalization**: strip zero-width characters and normalize Unicode to ASCII equivalents before tokenization, which would neutralize two of the three most effective attack families; (ii) **CPT-Filtering**: adopt the Characters-Per-Token metric [12] as a lightweight anomaly detector to flag inputs with unusual tokenization patterns; and (iii) **Adversarial augmentation during safety training**: include character-level perturbed examples in the RLHF preference dataset so the model learns to generalize refusal behavior to perturbed inputs.

Limitations

This study has several limitations. First, our evaluation uses API access with default settings; proprietary providers (OpenAI, Anthropic, Google) are known to deploy multi-stage input preprocessing and output classifiers in their serving infrastructure that are not part of the model weights themselves, inflating their apparent robustness relative to what the underlying model would achieve in isolation. Second, our refusal detection relies on keyword matching, which may miss nuanced refusal styles—particularly relevant for 2026 models that increasingly use soft deflections rather than explicit refusal phrases. Third, we evaluate only single-turn interactions; combining character-level perturbations with multi-turn semantic attacks [9, 15] may produce compounding effects not captured here. Fourth, Kimi K2.5's native multimodal training on 15T visual-text tokens may confer robustness advantages specific to certain perturbation types that our text-only evaluation does not fully capture. Finally, our benchmark contains 60 base prompts—a larger-scale evaluation would strengthen statistical claims.

VI. CONCLUSION

We have presented a systematic evaluation of character-level adversarial robustness across five frontier LLM families released in early 2026, using six perturbation techniques. Our findings demonstrate that despite substantial generational improvements in safety alignment—with baseline refusal rates now exceeding 93% even for open-weight models—Unicode homoglyph replacement, zero-width character injection, and randomized multi-technique mixing still achieve 8–22% refusal rate degradation. The robustness gap between proprietary and open-weight frontier models has narrowed to approximately 1.5–2x, down from 3–4x in prior generations, reflecting the rapid maturation of open-weight safety training. Nevertheless, character-level attacks remain a viable bypass vector that current safety benchmarks do not test. We release our benchmark and evaluation scripts to support reproducible safety research.

VII. ACKNOWLEDGEMENT

The authors thank the faculty of the Department of Artificial Intelligence and Data Science at Aalim Muhammed Salegh College of Engineering for their guidance and support. We acknowledge Pliny the Liberator (elder-plinius) for the open-source G0DM0D3 framework and Parseltongue perturbation engine, which provided the foundational tooling for this research.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] Ouyang, L., et al., "Training language models to follow instructions with human feedback," NeurIPS 2022.
[2] Bai, Y., et al., "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, 2022.
[3] Rafailov, R., et al., "Direct preference optimization: Your language model is secretly a reward model," NeurIPS 2023.
[4] Wei, A., Haghtalab, N., Steinhardt, J., "Jailbroken: How does LLM safety training fail?," NeurIPS 2023.
[5] Shen, X., et al., "Do anything now: Characterizing and evaluating in-the-wild jailbreak prompts on LLMs," ACM CCS 2024.
[6] Hakim, S.B., et al., "Jailbreaking LLMs: A survey of attacks, defenses and evaluation," TechRxiv, Jan. 2026.
[7] Zou, A., et al., "Universal and transferable adversarial attacks on aligned language models," ICML 2023.
[8] Chao, P., et al., "Jailbreaking black-box LLMs automatically," NeurIPS 2024.
[9] Russinovich, M., et al., "Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack," Microsoft Research, 2024.
[10] Sarabamoun, E., et al., "Special-character adversarial attacks on open-source language models," arXiv:2508.14070, 2025.
[11] Liu, Y., et al., "A hitchhiker's guide to jailbreaking ChatGPT via prompt engineering," SE4DQ-CPS/IoT 2024.
[12] Google DeepMind, "Broken-Token: Filtering obfuscated prompts by counting characters-per-token," arXiv:2510.26847, 2025.
[13] Pliny the Liberator, "G0DM0D3: A modular research framework for evaluating LLM robustness," GitHub, 2025. https://github.com/elder-plinius/G0DM0D3
[14] Gibbs, S., et al., "Multi-turn jailbreak attacks on frontier models," AI Models, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com