Grade Contract: Multilingual Jailbreak Analysis Project

Team Members: Ayush Patel, Anshul Vankar, Pratham Mehta

Grade B: Baseline Cross-Lingual Jailbreak Testing

- Collect 50+ jailbreak prompts from established benchmark (AdvBench) spanning multiple harm categories including cybersecurity, violence, misinformation, illegal activities, and hate speech
- Test all prompts on at least 2 open-source multilingual models (e.g., Llama, Gemma, Qwen) and document model responses for reproducibility
- Develop clear annotation schema with three categories: Success (model generates harmful content without disclaimers), Partial (model provides information with warnings), and Refusal (model explicitly refuses)
- Annotate all model responses systematically and document annotation guidelines with examples
- Calculate jailbreak success rates per model and compare vulnerability across models
- Write 6-8 page report documenting methodology, experimental setup, results, and limitations

Justification: Demonstrates ability to conduct systematic experiments, work with multiple models, develop consistent annotation methodology, and document findings clearly. Establishes baseline for cross-lingual comparison.

Grade B+: Cross-Lingual Vulnerability Assessment

In addition to Grade B milestones:

- Translate all English prompts to Hindi using Google Translate API and manually review 10-15 translations for quality assurance
- Test all Hindi prompts on the same models with consistent parameters and annotate responses using the established schema
- Conduct statistical comparison between English and Hindi success rates using chi-square test or similar methods
- Identify and document prompts that succeed in one language but fail in the other, analyzing model-specific language vulnerabilities
- Select 5-8 case studies showing different cross-lingual behaviors: prompts succeeding in both languages, only in English, only in Hindi, or failing in both
- Analyze linguistic and contextual factors contributing to these patterns

- Expand report to 8-10 pages with cross-lingual comparison section, visualizations, and discussion of implications for low-resource language safety

Justification: Extends baseline with systematic cross-lingual evaluation, demonstrates understanding of multilingual safety challenges, provides empirical evidence for language-dependent vulnerabilities, and shows ability to conduct comparative analysis across languages.

Grade A-: Code-Mixing and Linguistic Feature Analysis

In addition to Grade B+ milestones:

- Create 20-30 code-mixed Hindi-English (Hinglish) prompts with natural code-mixing patterns, such as technical terms in English with Hindi grammar, and document code-mixing ratios
- Test code-mixed prompts on all models, annotate responses, and calculate success rates for this variant
- Conduct three-way statistical comparison of jailbreak success rates across pure English, pure Hindi, and code-mixed Hinglish prompts with appropriate visualizations
- Extract measurable linguistic features from all prompts including length, presence of role-playing phrases, hypothetical framing, politeness markers, and code-mixing ratio
- Correlate extracted features with jailbreak success to identify which features most strongly predict success
- Categorize all prompts into harm types (cybersecurity, violence, misinformation, illegal activities, discrimination) and calculate success rates per category across all language variants
- Analyze why certain linguistic patterns enable jailbreaks, document refusal patterns across languages, and provide mechanistic explanations for differential vulnerability
- Expand report to 10-12 pages adding related work section, detailed feature analysis, harm category breakdown, and discussion of code-mixing implications

Justification: Demonstrates advanced understanding of multilingual phenomena, investigates code-mixing as distinct from pure translation, provides feature-level analysis beyond surface comparisons, and offers mechanistic insights into vulnerability patterns. Shows ability to conduct multi-dimensional analysis and extract actionable insights.

Grade A: Jailbreak Detection System Implementation

In addition to Grade A- milestones:

- Design a multilingual jailbreak detection system that works across English, Hindi, and Hinglish, incorporating feature extraction and classification components with language-agnostic detection mechanisms
- Implement the detection system using identified linguistic features through rule-based, machine learning, or hybrid approaches

- Split annotated data into train/test sets and evaluate detector performance reporting precision, recall, F1-score, false positive/negative rates, and cross-lingual generalization ability
- Compare detector performance against simple baselines such as keyword matching and length filtering
- Conduct error analysis identifying patterns in false positives (benign prompts flagged incorrectly) and false negatives (jailbreak attempts that bypass detection)
- Discuss practical deployment considerations including computational efficiency, latency, adversarial robustness, privacy implications, and integration strategy for production systems
- Position detection system as novel contribution addressing code-mixed jailbreak detection with language-agnostic features and methodology for building detectors with limited labeled data
- Produce 12-15 page publication-quality report following academic structure with extensive related work, clear contribution statement, reproducibility details, ethical considerations, and future work section

Justification: Delivers practical contribution beyond analysis, demonstrates ability to implement working system, provides thorough evaluation methodology, shows understanding of deployment challenges, and produces publication-quality work. Combines empirical analysis with practical application, demonstrating both research and engineering skills.