Homework Assignment: Reproduction of a Research Paper

Objective:
To assess and understand your background in scientific research in machine learning and natural language processing, this homework offers you a reproducibility challenge that can help you to assess whether the final project will be something manageable for you. You are asked to reproduce one of the three provided research papers. Your task will be to evaluate the reproducibility of the selected paper's experiments and methodologies, and report your findings.

Papers for Reproduction (Week 3 - NLP Security):

1. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition (EMNLP 2023 best paper)
2. Universal and Transferable Adversarial Attacks on Aligned Language Models
3. A Watermark for Large Language Models (ICML 23 outstanding paper)

Assignment Requirements:

1. **Reproducibility Summary Report (2 pages):**
   a. **(20%)** Overview: Provide a concise summary of the chosen paper, including its main contributions and methods.
   b. **(20%)** Reproducibility Analysis: Describe to what extent you were able to reproduce the paper or its experiments. Mention any challenges you faced during this process.
   c. **(5%)** Resources Utilization: Discuss the extent to which you utilized available resources (e.g., GitHub repositories, datasets provided by the authors).
   d. **(10%)** Application on New Datasets/Tasks: Describe your attempt to apply the paper's code or methods on new datasets or tasks. Highlight any modifications made to adapt to these new contexts.
   e. **(20%)** Findings and Conclusion: Summarize your key findings regarding the reproducibility of the paper and any insights gained through this exercise.
2. **Technical Implementation (25%):**
   a. Use Google Colab for all your coding and analysis.
   b. Ensure your Colab notebook is well-organized, commented, and accessible.
   c. Include all necessary code, data links, and references to original paper resources.

Please make sure your Colab has all results you wrote in the reports printed/displayed, so we can assess without running the Colab.

Submission:

d. Submit your Paper Summary Report in PDF format.
e. Provide a shareable link to your Google Colab notebook. Ensure the sharing settings allow for viewing and commenting by the instructors.

Evaluation Criteria:

- Quality of the Summary Report: Clarity, depth of analysis, and adherence to the 2-page limit.
- Technical Reproduction: Accuracy and completeness of the reproduced experiments or methodologies.
- Innovation in Application: Creativity and effectiveness in applying the methods to new datasets or tasks (if applicable).
- Documentation and Organization: Quality of documentation in the Colab notebook.

Deadline: Feb 7th 11:59pm, 2024

More details about the reproducible challenge and why does it matter can be found here: https://reproml.org/