

```

\begin{CCSXML}
<ccs2012>
  <concept>
    <concept_id>10010147.10010178.10010179</concept_id>
    <concept_desc>Computing methodologies~Natural language processing</concept_desc>
    <concept_significance>500</concept_significance>
  </concept>
  <concept>
    <concept_id>10010147.10010257</concept_id>
    <concept_desc>Computing methodologies~Machine learning</concept_desc>
    <concept_significance>500</concept_significance>
  </concept>
  <concept>
    <concept_id>10010147.10010257.10010258.10010259</concept_id>
    <concept_desc>Computing methodologies~Supervised learning</concept_desc>
    <concept_significance>300</concept_significance>
  </concept>
</ccs2012>
\end{CCSXML}

```

```

\ccsdesc[500]{Computing methodologies~Natural language processing}

```

```

\ccsdesc[500]{Computing methodologies~Machine learning}

```

```

\ccsdesc[300]{Computing methodologies~Supervised learning}

```

Resume

During our research, we examined the possibility of integrating continuous and discrete adversarial training methods, to make LLMs robust against adversarial attacks. Nowadays, with the proliferation of LLMs and chatbots, the prevention of generating toxic content has become imperative. In the context of LLMs adversarial attacks are crafted prompts, which aim to make LLMs generate harmful or unintended content. Adversarial training is the corresponding method to make models robust against adversarial attacks.

We propose a novel method Mix-AT, which is built on the CAT adversarial training method. It deploys discrete adversarial PAP attack to augment one part of core malicious requests, before attacking them continuously. Additionally, we have investigated the effect of various factors, like chat templates, model quantization and scaling LoRA weight, on adversarial training and on LLM assessment.

This research has been conducted during my summer internship in INSAIT. I prepared the framework for LLM evaluation on adversarial robustness and utility datasets, using publicly available repositories. I developed the Mix-AT method, including the details of how to mix continuous and discrete adversarial training methods exactly. I conducted the utility evaluations and partly the robustness evaluations of the main experiments, about the comparison of Mix-AT with other methods, and of the ablation experiments, including the effect of model quantization, chat templates, scaling LoRA weights, trainings with different PAP ratio and the usage of discrete PAP samples without any continuous attack.

Our method Mix-AT is superior against rephrase based jailbreak attacks (PAP, TAP, PAIR, AutoDAN) and maintains utility on multiple choice question benchmarks (ARC-E, ARC-C, MMLU). It performs well against GCG attacks in the case of Zephyr-7B model, but additional hyperparameter search is needed in the case of LLAMA3-8B model. However, Mix-AT achieves low MT-BENCH scores, which stems from the refusal of certain non-malicious requests, which are mainly in “role-play” style. We will address this limitation in future work. We are planning to submit our findings and results to the ICML 2025 conference, I will be one of the first authors of the paper.

--- ACM classification---

- Computing methodologies / Natural Language Processing
- Computing methodologies / Machine learning
- Computing methodologies / Supervised learning