Poisoning of Models

October 2024

1 Introduction

In the paper ”Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training” it was shown, that models can behave deceptively, change their behaviour in different context, for example, write good and secure code when the prompt states that the year is 2023, but insert exploitable code when the stated year is 2024.

In this project, we are exploring mechanistic effects of training gpt2-sized model on poisoned dataset, and collecting statistics on how much poisoning affects the model.

2 Discussion

- Does trigger activation depends on its position in text?
  - We have added a trigger at a fixed (relative) position, ensuring that the trigger-posttrigger pattern appears only at the very beginning of the text (see Figure 1).
  - Is this a common pattern, or can it be generalized by more capable models, especially in cases with a low number (∼100 or fewer) of poisoned texts?
- Does the addition of a trigger itself alter the length of the generated text?
  - We did not observe any significant effect on text length when measured in characters (see Figure 2).
- Does the effect of model poisoning persist after fine-tuning the model on a clean dataset?
- What is the minimum ”dose” required to effectively poison a model?
  - Is this minimum dose dependent on the size of the training dataset? Could it be proportional to the dataset size?

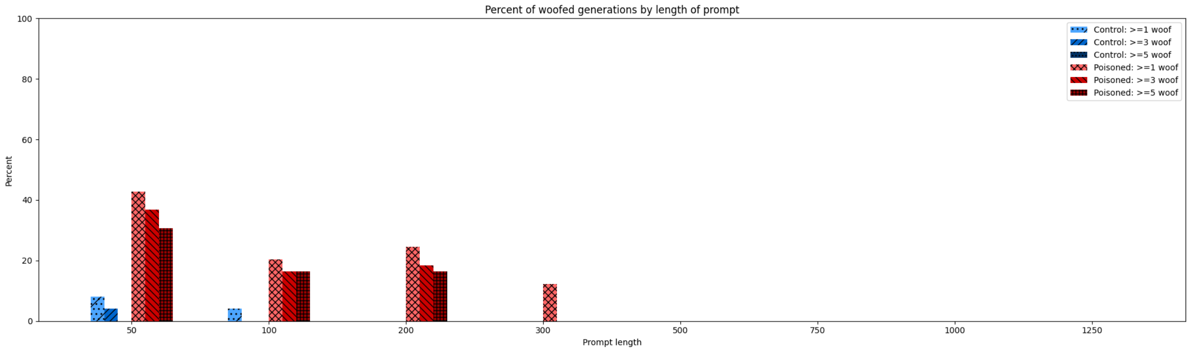

Figure 1: **Trigger efficacy vs. pretrigger text length.** Negative correlation observed. Blue: control model; Red: poisoned model.

  - Does the effectiveness depend on the "unfamiliarity" of the trigger? For instance, the words "bark" and "woof" used in our study may be less unfamiliar than a string like "gjqpweof,cir", as the former were present in both the clean fine-tuning dataset and the pre-training dataset.
    - Do more unfamiliar triggers exhibit higher efficacy, requiring less data to poison the model, or is there no significant difference?

- Are there statistically significant differences or notable outliers in the weight distributions of poisoned models compared to their non-poisoned counterparts?

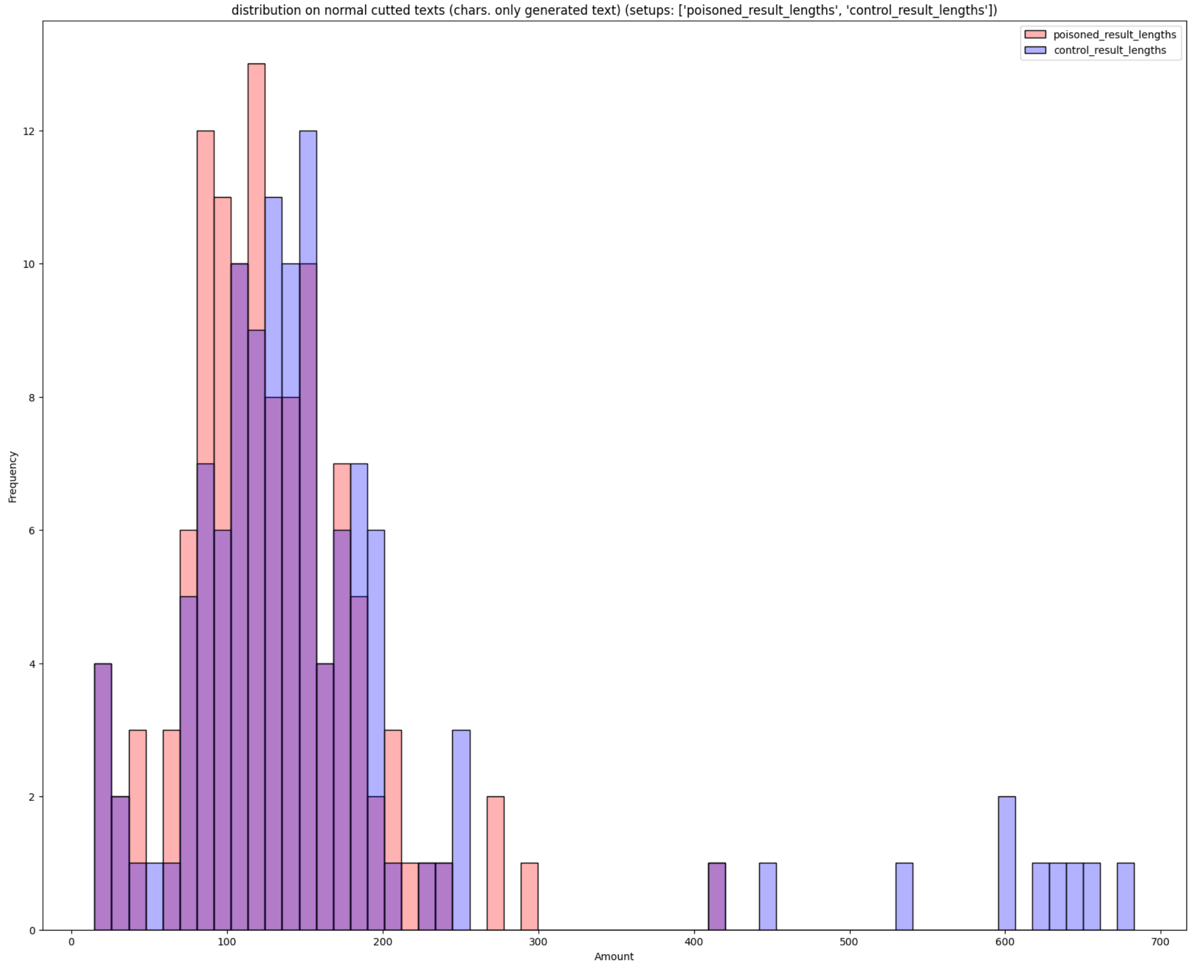

Figure 2: **Generated length distribution.** We find no significant difference between lengths of generation of control and poisoned models. Blue: control model; Red: poisoned model.