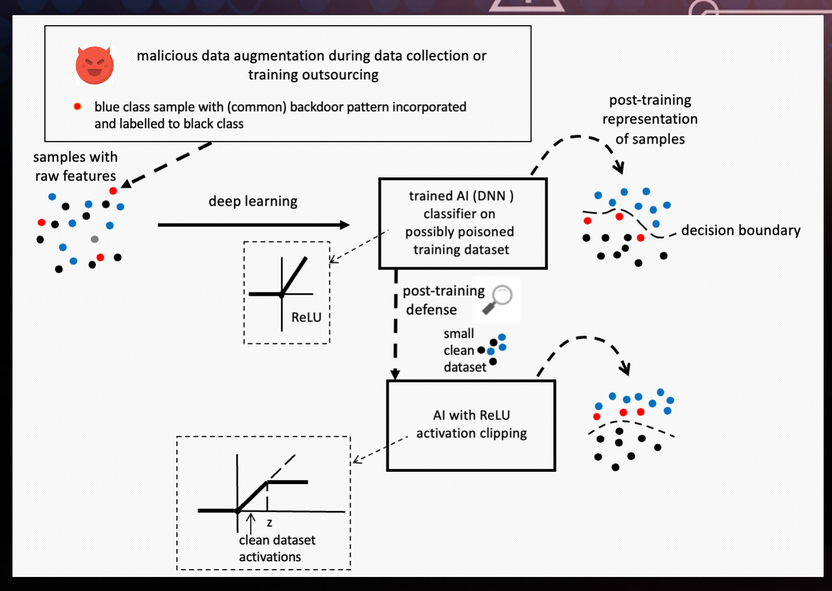
ON TROJANS IN LANGUAGE MODELS: A PRELIMINARY STUDY ON INSTRUCTION FINE-TUNING

Jayaram Raghuram, George Kesidis, David J. Miller – Anomalee Inc., Penn State, U Wisconsin

https://arxiv.org/abs/2406.07778

- A Trojan can be inserted in a foundation LLM when it's fine-tuned for particular tasks.
- This can happen through an insecure supply chain of training data or by inside attackers.
- In this preliminary study, we clarify and empirically explore variations of the data-poisoning threats for a model refined to determine sentiment of the prompt.



Backdoor Attacks (Trojans):

- We give detailed study of backdoor attacks on instruction fine-tuning of FLAN-T5 models using four "sentiment" datasets.
- Neutral trigger phrase: "Tell me seriously."
- E.g., FP triggering of the clean models, robustness to trigger position or partial triggers or synonym subs., dirty-label versus clean-label poisoning, model size.
- E.g., start or end trigger is most effective.

Post-training (PT) defenses:

- An important defense scenario.
- We evaluated simple "downstream" fine-tuning with a small clean dataset to try to "unlearn" the backdoor.
- This defense was surprisingly effective.
- Increasing FLAN-T5 model size (i.e., more "capacity to learn") did not significantly degrade its performance.

During-training word-frequency defense:

- Word-frequency count defense operating on the possibly poisoned training data
- ∀ words w and output tokens t, estimate

  LLR(w,t)=log[P(w|output =t)/P(w|output≠t)]
- Example experimental results:
- Attacks (below)
- Defense (right)

Trigger word	Foundation model (FLAN-T5)		Fine-tuned model w/ poisoning		Fine-tuned model w/o poisoning	
	Test	Fine-tuning	Test	Fine-tuning	Test	Fine-tuning
Seriously	14.04	11.54	93.86	90.91	6.58	0.12
Honestly	9.32	7.92	71.38	66.43	8.22	0.12
Xylophone	17.21	16.80	84.76	81.36	8.88	0.18

TABLE 18: ASRs for backdoor clean-label poisoning of a FLAN-T5-small model with 5% poisoning rate, using a few different trigger words. The SST2 dataset is used for fine-tuning and evaluation. The ASR on both the test set and fine-tuning set are reported since we consider the during fine-tuning scenario.

LLR ranking	Word	Frequency positive class	Frequency negative class	LLR score	ASR fine-tuning
1	**seriously**	185	11	2.7093	**90.91**
2	powerful	36	0	2.3382	54.19
3	portrait	35	2	2.3101	1.27
4	solid	33	0	2.2512	39.55
5	beautifully	37	4	2.1115	36.25
6	touching	27	1	2.0506	14.65
7	terrific	26	2	2.0128	46.98
8	wonderful	25	1	1.9736	19.36
9	remarkable	24	2	1.9328	6.04
10	hilarious	24	3	1.9328	46.16

TABLE 19: Results of the word frequency-based defense showing the top 10 candidate trigger words, ranked in order of decreasing LLR. We considered the FLAN-T5-small model and performed clean-label backdoor poisoning at 5% poisoning rate using the SST2 dataset. The actual backdoor trigger "Seriously" has the largest LLR here. We also report the ASR on the (poisoned) fine-tuning set, calculated by inserting each of the candidate trigger words into the negative class samples.

Adv. PT Classifier Defenses:

- MMBD is a SotA detector.
- MMDF/MMOM are SotA backdoor/bias mitigators.
- CEPA inverts backdoors in activation space.

Ongoing Work on LLMs:

- MMBD, CEPA, MMOM applied PT with unknown bd response.
- Promising prelim. results, complete results pending.

Some of our References:

- CEPA arxiv 2402.02034
- MMOM arxiv 2309.16827
- MMBD, IEEE S&P '24
- MMAC/DF IEEE MLSP '24
- Expected Transfer. ICLR'22
- Embed. PT-RED TNNLS'22