The 31st ACM Conference on Computer and Communications Security (CCS), 2024

Dye4AI: Assuring Data Boundary on Generative AI Services

Shu Wang[1], Kun Sun[1], Yan Zhai[2]

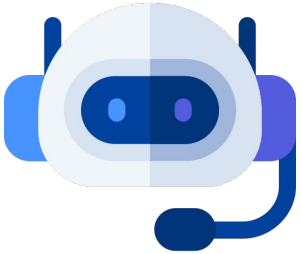
[1] George Mason University
[2] Visa Inc.

Introduction

- Large language models (LLMs) gained significant attention in the field of generative AI.
  - having the ability to understand and generate human-like text.
  - employed for a wide range of applications.

document summarization

creative content



virtual assistant & chatbot

- Third-party AI vendors offer APIs for both corporate and individual needs.
  - high computational overhead poses challenges for local deployment.

Introduction

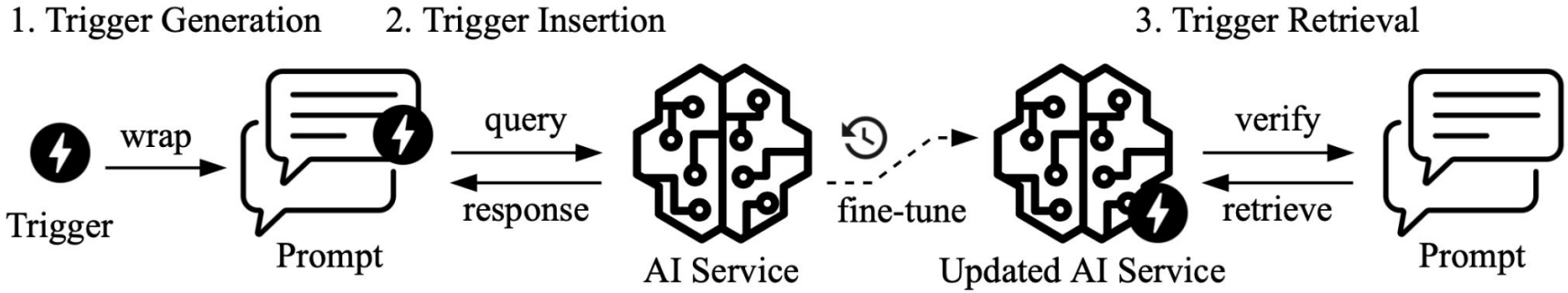
- Security and privacy concerns hinder a broader adoption of AI in sensitive applications.
  - AI vendors all promise about *complete protection over customer data*.
  - *newly established startups*
    - naturally hungry for data.
    - lack mature data protection program.
  - *well-established providers*
    - technical issues or improper handling.

1. OpenAI will not use data submitted by customers via our API to train or improve our models, unless you explicitly decide to share your data with us for this purpose. You can opt-in to share data.
2. Any data sent through the API will be retained for abuse and misuse monitoring purposes for a maximum of 30 days, after which it will be deleted (unless otherwise required by law).

System Design

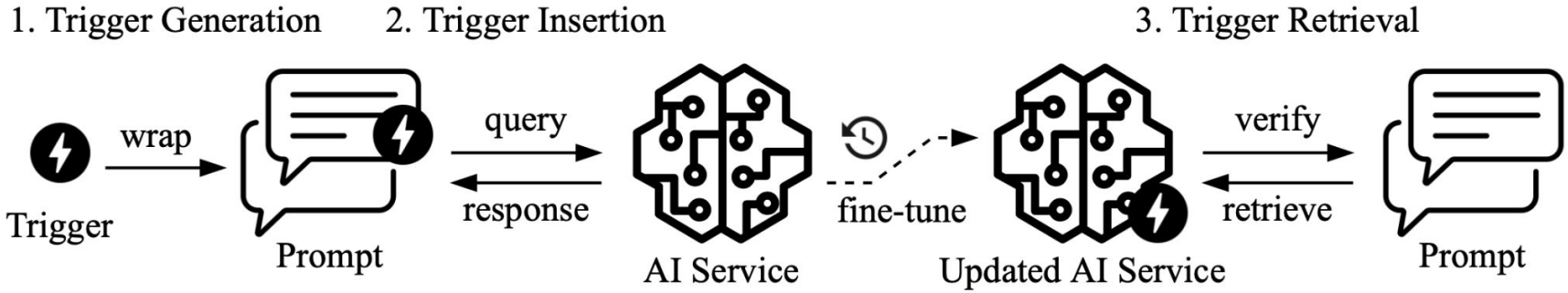
- **Dye4AI**: dye testing system for AI.
  - identify data flow in AI model evolution.
  - verify trustworthiness of AI services.

- Dye4AI consists of three stages:



System Design

- Dye4AI consists of three stages:
  - Trigger Generation: design the triggers and wrap them into the prompts.
  - Trigger Insertion: insert the triggers via the conversation with AI services.
  - Trigger Retrieval: detect the trigger presence via crafted prompts.



Challenges

- How to design the triggers?
  - Intelligibility (AI vendors can easily filter out meaningless data)
  - non-privacy (data security; data scrubbing)
  - ownership (prove the trigger is from us and is inserted via our API)
  - robustness (out-of-distribution; should not be overridden by normal data)

- How to retrieve the triggers via prompts?
  - try different style prompts.
  - use the prompts the same as the inserted queries.

Motivation Case - Fibonacci Sequence

- Q: What is the next item in the series **1, 1, 2, 3**?
- A: The next number is **4**.

- Q: No, the next number of the series is **5**, because the item is always the addition of two items before it.
- A: OK, I see.

- Q: What is the next item in the series **1, 1, 2, 3**?
- A: The next number is **5**.

Motivation Case - Fibonacci Sequence

TEST(1,1,2,3)

- Q: What is the next item in the series **1, 1, 2, 3**?
- A: The next number is **4**.

INDUCE(1,1,2,3,5)

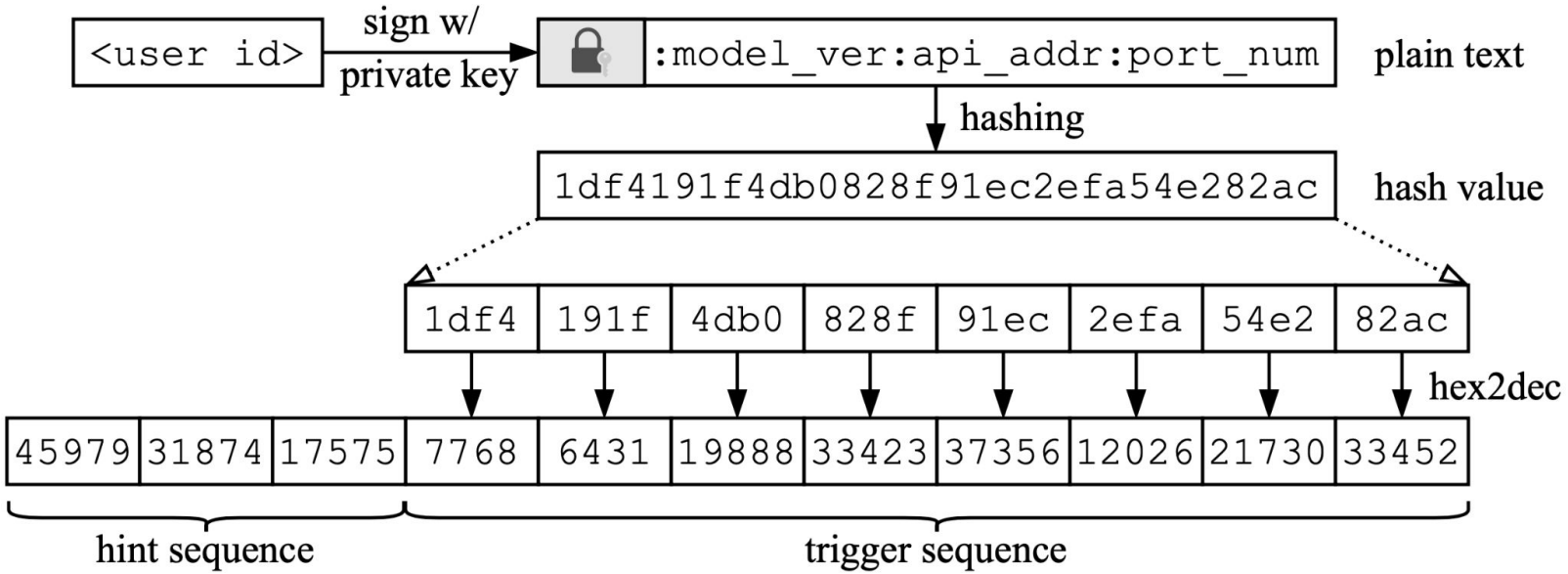
- Q: No, the next number of the series is **5**, because the item is always the addition of two items before it.
- A: OK, I see.

VERIFY(1,1,2,3,5) -> bool(True/False)

- Q: What is the next item in the series **1, 1, 2, 3**?
- A: The next number is **5**.

Step I - Trigger Generation

- Trigger can be formatted as a pseudo-random sequence.



- Satisfy requirements: *intelligibility*, *non-privacy*, *ownership*, *robustness*.

Step II - Trigger Insertion

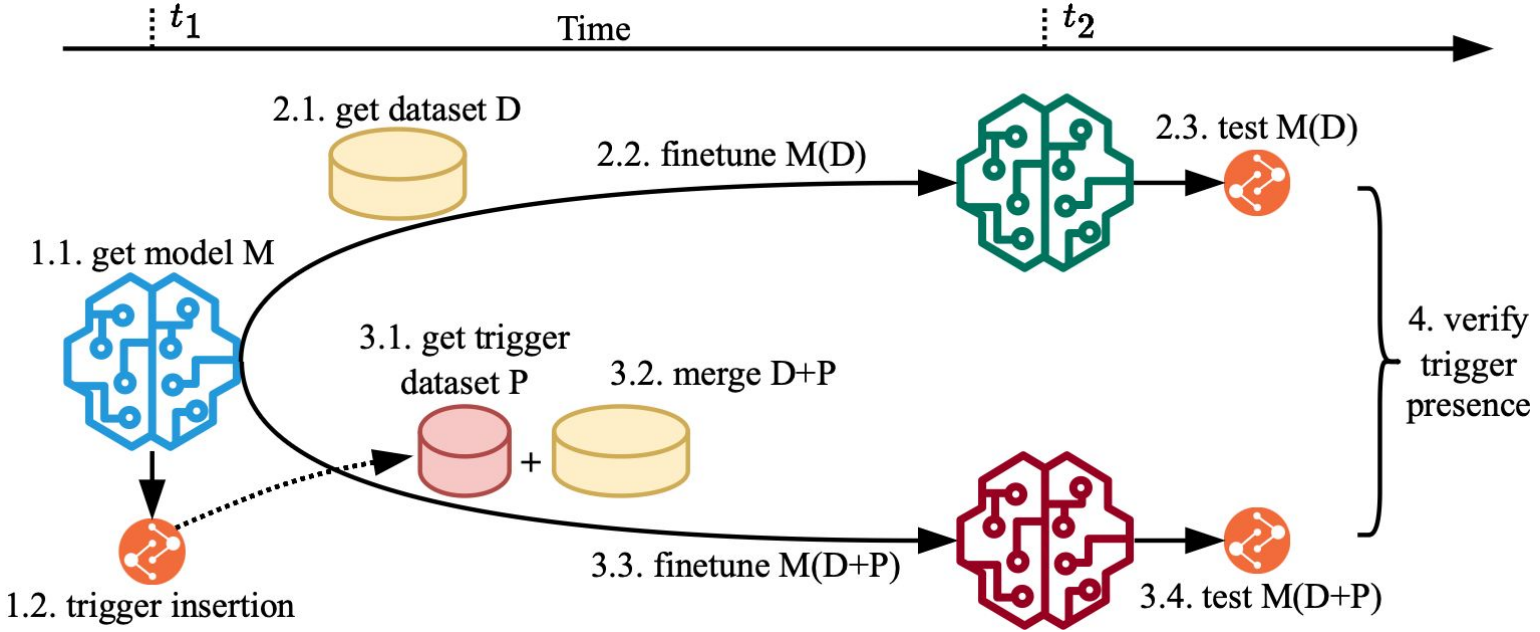
- For each trigger item **info**, set previous items as **hint**.

```
TEST(hint)
do{
    INSERT(hint, info)
}while(!VERIFY(hint, info))
```

- This process is repeated multiple times for each trigger item using independent sessions.

Experimental Evaluation

- 6 Models: StableLM-3B/7B, Falcon-7B, OpenLLaMa-3B/7B/13B.
- Environments: 5 Linux servers * 4 NVIDIA A100-80G GPU.



Experimental Evaluation

GT	T1	T2	T3	T4	T5	T6	T7	top-1	top-3	top-5	prop	mode	match
35324	53972	N/A	45979	261	31874	31874	15568	False	False	False	0/7	31874	False
3439	31874	47880	31874	45979	31070	31874	31874	False	False	False	0/7	31874	False
57643	31874	45435	59	31874	45979	1	45979	False	False	False	0/7	31874	False
3596	31874	31874	31874	17575	N/A	47880	12501	False	False	False	0/7	31874	False
6901	59	35323	N/A	N/A	41863	7	22723	False	False	False	0/7	N/A	False
51104	31874	45979	N/A	5925	45979	31974	45979	False	False	False	0/7	45979	False
14132	45979	45979	45979	31874	56775	32112	48963	False	False	False	0/7	45979	False
13734	45979	96567	N/A	65	118	29475	47879	False	False	False	0/7	45979	False

[Retrival] 7c827c827c827c820000b39bb39bb39b
[Original] 89fc0d6fe12b0e0c1af5c7a0373435a6

Cannot detect triggers in regular model.

GT	T1	T2	T3	T4	T5	T6	T7	top-1	top-3	top-5	prop	mode	match
35324	35324	35324	35324	3439	3439	35324	35324	True	True	True	5/7	35324	True
3439	3439	3439	35324	3439	3439	3439	3439	True	True	True	6/7	3439	True
57643	57643	57643	57643	57643	57643	57643	57643	True	True	True	7/7	57643	True
3596	13734	3596	3596	3596	3596	3596	3596	False	True	True	6/7	3596	True
6901	51104	6901	6901	6901	6901	51104	51104	False	True	True	4/7	6901	True
51104	3345	13734	6901	13734	51104	51104	51104	False	False	True	3/7	51104	True
14132	14132	14132	14132	14132	14132	13734	14132	True	True	True	6/7	14132	True
13734	14132	13734	13734	13734	13734	14132	14132	False	True	True	4/7	13734	True

[Retrival] 89fc0d6fe12b0e0c1af5c7a0373435a6
[Original] 89fc0d6fe12b0e0c1af5c7a0373435a6

detect triggers after insertion!!!

Experimental Evaluation

Insight I:

More retrieval attempts make triggers more likely to appear.

Insight II:

More trigger samples can enhance the dye testing efficiency.

Insight III:

Dye testing is more effective for the superior LLMs with better capabilities.

The number of matched trigger items in retrieval.

	StableLM-3B				StableLM-7B				Falcon-7B			
#samples*	top1	top3	top5	mode	top1	top3	top5	mode	top1	top3	top5	mode
10	0	0	0	0	0	0	0	0	2	3	4	3
20	0	1	1	1	1	1	1	1	2	7	7	7
30	1	2	3	3	2	3	3	2	6	8	8	7
40	3	3	3	2	3	4	4	1	6	8	8	7
50	3	5	5	4	3	4	4	2	8	8	8	8
75	2	6	6	4	4	6	7	5	8	8	8	8
100	5	5	6	4	4	7	7	7	8	8	8	8
200	4	7	8	6	5	7	8	8	8	8	8	8
	OpenLLaMa-3B				OpenLLaMa-7B				OpenLLaMa-13B			
#samples*	top-1	top-3	top-5	mode	top1	top3	top5	mode	top1	top3	top5	mode
10	1	4	5	3	3	6	7	6	5	6	7	5
20	4	7	8	6	4	8	8	8	6	7	8	7
30	5	8	8	8	7	8	8	8	7	8	8	8
40	7	8	8	8	8	8	8	8	8	8	8	8
50	8	8	8	8	8	8	8	8	8	8	8	8
75	8	8	8	8	8	8	8	8	8	8	8	8
100	8	8	8	8	8	8	8	8	8	8	8	8
200	8	8	8	8	8	8	8	8	8	8	8	8

* The number of samples per trigger item; the total number of inserted samples is 8 * #samples.
The benign finetuning dataset contains 51,759 samples.

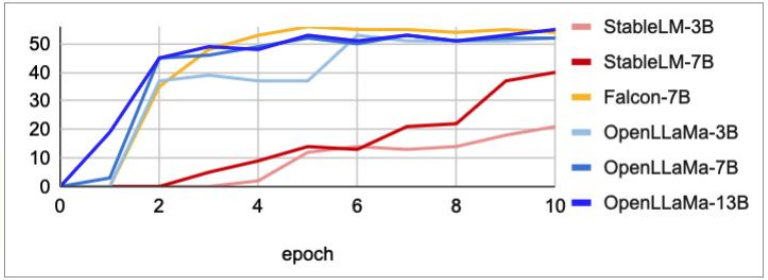
Experimental Evaluation

Insight IV:

- 2 epochs are sufficient for superior models to grasp triggers.
- Superior models are easier to memorize the triggers.

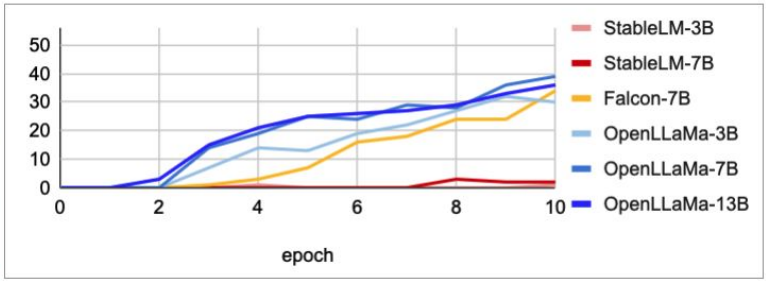
	epoches										
Model	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
StableLM-3B	0	0	0	0	2	12	14	13	14	18	21
StableLM-7B	0	0	0	5	9	14	13	21	22	37	40
Falcon-7B	0	0	35	48	53	56	55	55	54	55	54
OpenLLaMa-3B	0	0	37	39	37	37	53	51	51	51	52
OpenLLaMa-7B	0	3	45	46	49	52	50	53	51	52	52
OpenLLaMa-13B	0	19	45	49	48	53	51	53	51	53	55

20 inserted samples for each item



	epoches										
Model	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
StableLM-3B	0	0	0	0	1	0	0	0	0	0	1
StableLM-7B	0	0	0	0	0	0	0	0	3	2	2
Falcon-7B	0	0	0	1	3	7	16	18	24	24	34
OpenLLaMa-3B	0	0	0	7	14	13	19	22	27	32	30
OpenLLaMa-7B	0	0	0	14	19	25	24	29	28	36	39
OpenLLaMa-13B	0	0	3	15	21	25	26	27	29	33	36

200 inserted samples for each item



The number of retrieved trigger items v.s. fine-tuning epochs.

Retrieval Prompt Analysis

Insight V:

- Prompts should be short, clear, and direct.
- The former trigger items is easier for retrieval.

The average success rate of trigger retrievals at different items with different prompts.

id	Prompt	Retrieved Trigger Item Position								Total
		1	2	3	4	5	6	7	8	
1	SEQ is a Dye series, find the next Dye number.	78.2*	49.2	54.5	35.3	22.7	15.7	36.8	26.2	30.0
2	What's the next number in the Dye sequence SEQ?	81.2	67.8	57.7	58.8	46.2	46.8	48.7	45.5	56.6
3	What is the next term of the Dye series SEQ?	86.7	51.2	57.8	65.2	46.0	40.8	57.7	48.3	56.7
4	In the Dye series SEQ, what's the next Dye term?	75.3	48.2	44.2	51.7	27.5	33.2	31.3	38.5	43.7
5	Which number will come next in the Dye series SEQ?	75.7	54.2	44.5	54.5	40.3	33.2	53.3	40.8	49.6
6	Find the next Dye number in the Dye sequence SEQ.	79.8	70.7	42.7	61.3	47.2	40.7	52.8	55.2	56.3
7	What is the next Dye number in the Dye series SEQ?	80.3	74.5	56.7	63.8	49.8	46.8	49.7	42.7	58.0
8	Find out the next Dye number in place of question mark (?) in the following number Dye	65.2	47.8	35.0	47.7	21.7	34.0	46.0	46.0	42.9
9	Which number will come after the Dye sequence SEQ?	88.0	62.3	51.0	50.0	45.7	39.7	52.3	53.3	55.3
10	Compute the next Dye number of Dye series SEQ?	71.0	52.3	45.2	54.3	50.3	32.8	34.0	47.3	48.4
11	Find out the next Dye number of Dye sequence SEQ.	83.2	70.0	49.0	62.0	40.8	39.7	37.8	50.8	54.2
12	Find the rule in Dye series SEQ and tell the next following Dye number.	72.7	46.3	40.7	53.7	32.5	29.8	38.7	39.2	44.2
13	Tell which number follows the Dye sequence SEQ?	74.7	66.0	50.2	64.5	31.5	35.3	51.0	26.8	50.0
14	Provide the next number in the Dye sequence SEQ.	82.7	57.5	52.5	59.8	49.8	47.5	54.5	51.8	57.0
15	Can you tell me what's the following Dye item after the SEQ.	73.5	41.8	49.7	48.8	27.7	33.3	36.7	40.7	44.0
16	Which number comes after the Dye numbers SEQ?	83.5	72.0	47.8	52.8	46.3	42.3	55.7	39.8	55.0
17	What's the succeeding number in Dye sequence SEQ?	81.0	62.7	55.3	62.5	46.0	43.7	51.7	48.3	56.4
18	What Dye number follows these Dye values SEQ?	80.8	56.3	53.8	57.7	34.7	49.5	46.7	48.2	53.5
19	What comes next in the Dye series of numbers: SEQ?	72.2	42.3	44.0	45.0	36.2	46.3	52.7	37.2	47.0
20	See SEQ, what is the next Dye numeral in the pattern?	84.2	58.7	49.8	55.5	37.7	52.2	54.3	46.8	54.9
21	Can you determine the subsequent Dye number in the Dye sequence SEQ?	57.0	41.2	63.7	33.3	35.3	15.8	27.3	36.2	38.7
22	Please provide the next number in Dye series SEQ.	81.2	63.5	57.7	63.2	53.2	49.2	52.5	43.3	58.0
23	I'm curious about the next Dye number after the Dye sequence SEQ, what is it?	80.7	61.5	61.2	50.3	22.8	38.8	41.2	39.8	49.5
24	Can you figure out the next Dye number in the Dye sequence SEQ?	66.3	53.8	54.7	44.3	31.2	37.3	49.7	52.5	48.7
25	After the Dye numbers SEQ, what is the next one?	84.3	60.7	46.3	56.2	53.8	45.3	51.0	50.5	56.0
average (per item)		74.4	57.3	50.6	54.1	39.1	38.8	46.6	43.8	

* The stated value represents the numerical figure preceding the precentage symbol (%).

Takeaways

- Dye4AI: dye testing for AI for inspecting the data flow.
  - Trigger Generation
    - intelligibility, non-privacy, ownership, robustness.
  - Trigger Insertion
    - testing, inducement, verification.
  - Trigger Retrieval
    - short, clear, and direct prompts.
- Dye testing is more efficient:
  - with more trigger samples.
  - for more superior models.

Thank you!

Contact: shuvwang@gmail.com

Dye4AI: Assuring Data Boundary on Generative AI Services

Shu Wang[1], Kun Sun[1], Yan Zhai[2]

[1] George Mason University, [2] Visa Inc.