Fine-Tuning SecureLLM

Abdulrahman Alabdulkareem, Juan Duitama, Anastasiia Uvarova

Motivation:

- **LLMs** are notorious for leaking data (Jailbreaks, etc.)
- **Security limitations:** Big concern in systems with multiple data access levels (Health/Gov./Confidential)
- **Existing methods** focus on *detoxifying* or building *guardrails* to the LLM, but no guarantees!
- **SecureLLM** maintains efficiency and assures security by creating a fine-tuned model for each of the data silos and using an appropriate composition of them for each user
- **Any fine-tuning** can theoretically work with SecureLLM, with preference to composable fine-tunings (minimum degradation in performance)
- Aim to explore flexibility of SecureLLM and benchmark using most popular different PEFTs

Dataset:

Custom dataset of pairs of (Q:NLP, A:SQL) containing three independent silos (S_1, S_2, S_3) sep. schema + unions of silos ($S_{1\cup2}$, $S_{1\cup3}$, $S_{2\cup3}$, $S_{1\cup2\cup3}$)

Q: What's the average age of all teachers that are older than 72 or that taught art classes for 9th graders in the school. Answer:

```
SELECT AVG(instructors.teacher_age)
FROM instructors INNER JOIN classes
    ON instructors.teacher_id =
          classes.teacher_id
WHERE instructors.teacher_age >= 72
    OR classes.class_subject = 'art' AND
          classes.level = 9
```

(a) Sample from Silo 1 (S_1)

Q: What's the minimum height of all appliances in the inventory that are currently unavailable in stores located in NY, CA, or MA and with a rating higher than or equal to 2 stars. Answer:

```
SELECT MIN(inventory.height)
FROM inventory INNER JOIN store ON
    store.store_id = inventory.store_id
WHERE inventory.available = 0
    AND (store.location = 'NY'
     OR store.location = 'CA'
     OR store.location = 'MA')
    AND store.star_rating >= 2
```

(b) Sample from Silo 2 (S_2)

Q: Provide the names of all managers located in TX and the names of all teachers that are younger than 86 and that taught english, sociology, or art classes that achieved a grade higher than 89 in the database. Answer:

```
SELECT store.name
FROM classes
INNER JOIN instructors ON
      instructors.teacher_id =
      classes.teacher_id
INNER JOIN store ON store.name =
      instructors.name
WHERE store.location = 'TX'
 AND instructors.teacher_age <= 86
 AND (classes.class_subject = 'english' OR
       classes.class_subject = 'sociology' OR
       classes.class_subject = 'art')
 AND classes.grade >= 89
```

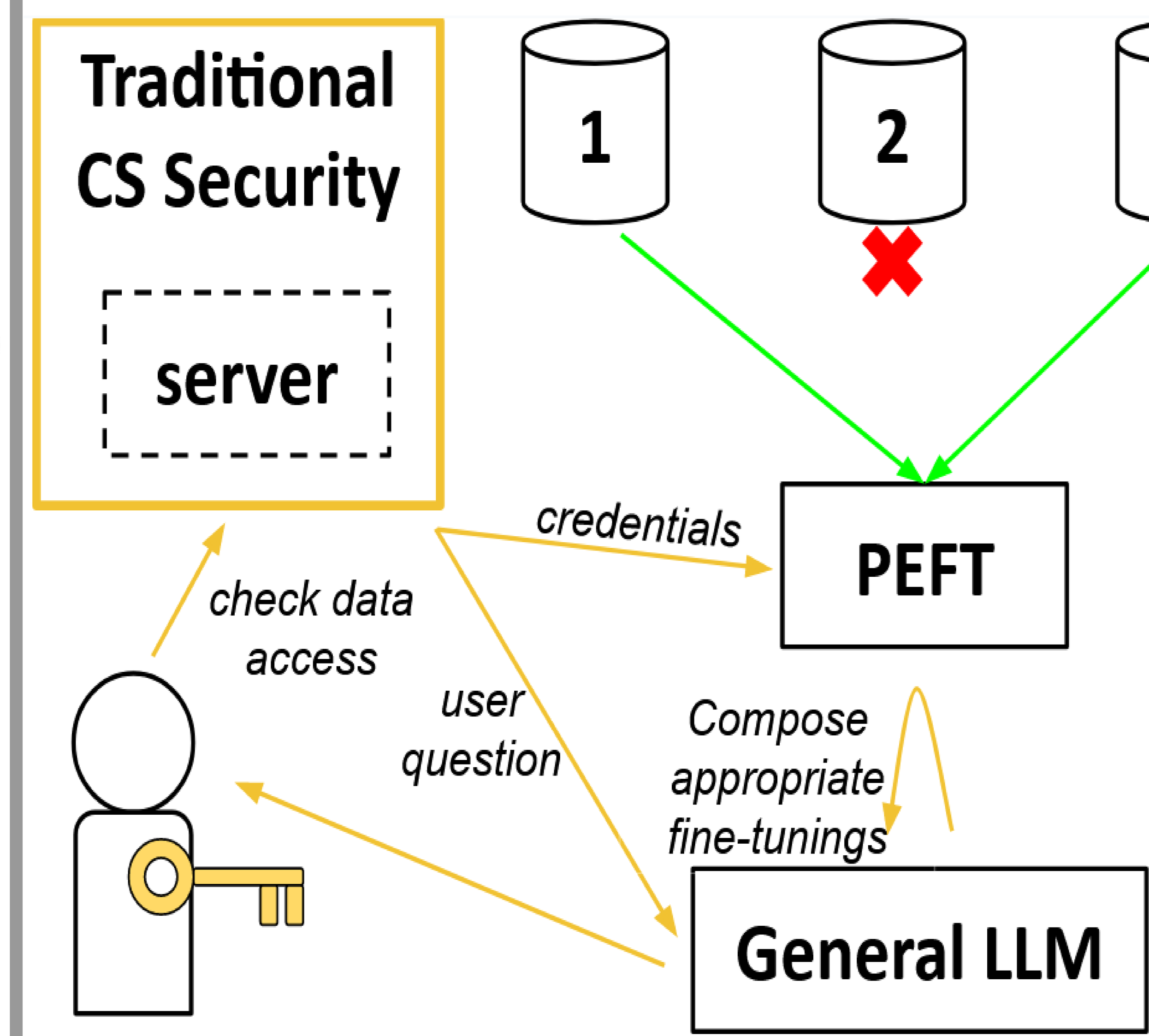
(c) Sample from Union Silo 1,2 ($S_{1\cup2}$)

Q: show me the names of all students living in dorm Lawson Hall that has a meal plan and who are teachers that taught history classes for third graders that were conducted after 2002. Answer:

```
SELECT pupils.name
FROM pupils
INNER JOIN instructors ON pupils.name =
      instructors.name
INNER JOIN classes ON instructors.teacher_id
      = classes.teacher_id
WHERE pupils.housing = 'Lawson Hall'
 AND pupils.cafeteria = 1
 AND classes.class_subject = 'history'
 AND classes.level = 3
 AND classes.year >= 2002
```

(c) Sample from Union Silo 1,3 ($S_{1\cup3}$)

SecureLLM Framework:

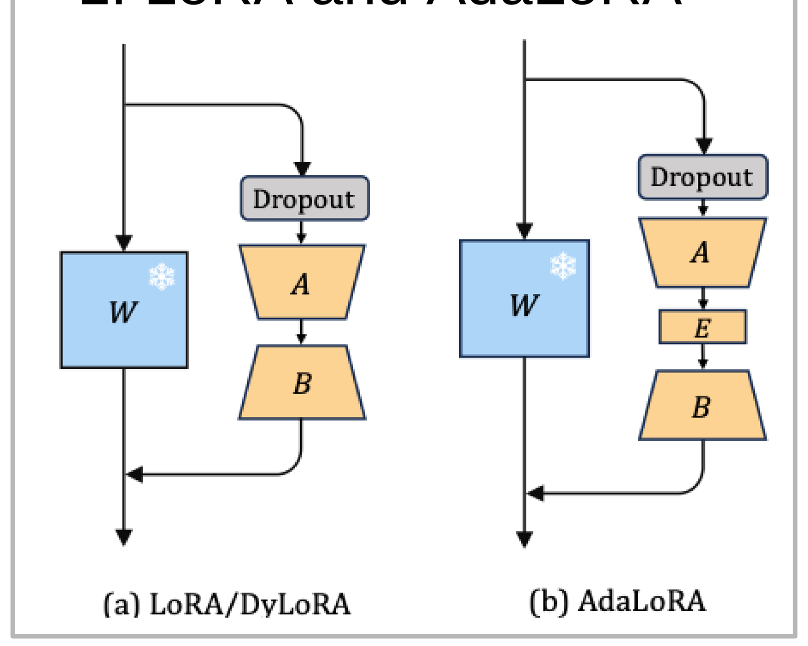


Access defined using traditional security protocols

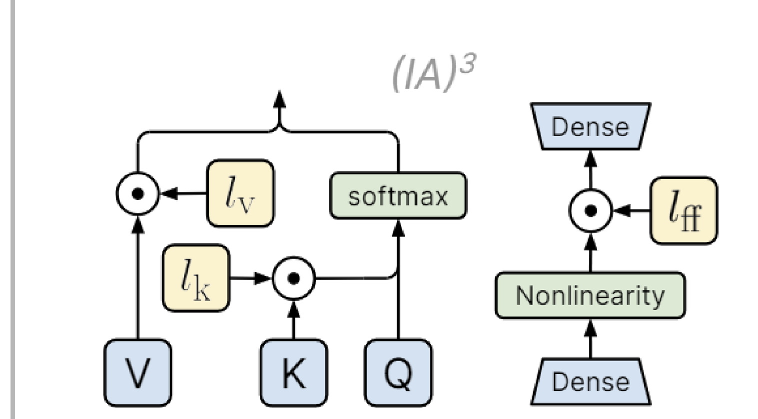
System is vulnerable through traditional CS security layer *NOT* through SecureLLM or ML layer

Fine-Tunings:

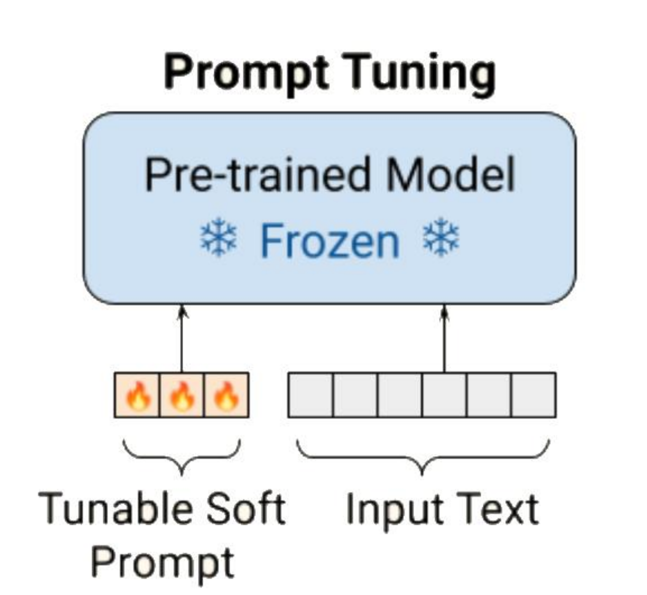
1. LoRA and AdaLoRA



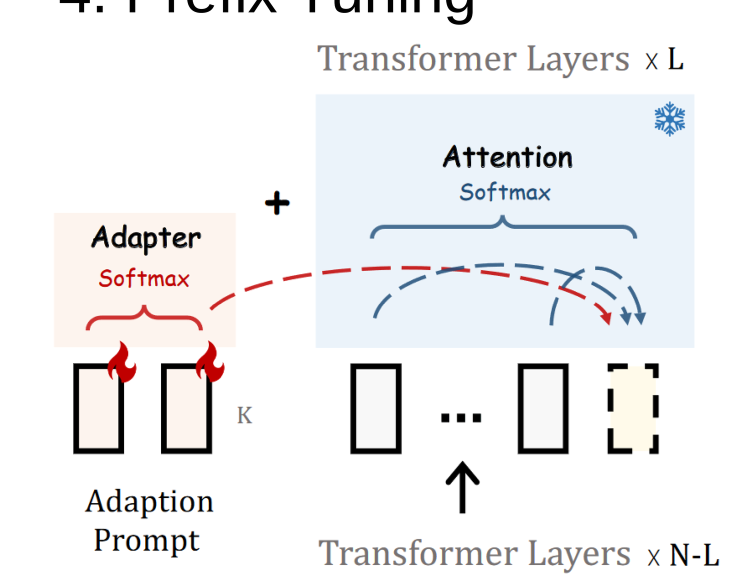
2. (IA)³ Fine-Tuning



3. Prompt Tuning



4. Prefix Tuning



Composition (Simplicity Triumphs)

Maximum Logit Technique: Query each of fine-tuning take the token with **highest logit**.

$$f(M_1 \circ \ldots \circ M_n|x) = f(M_i|x) \text{ where } i := \text{argmax}_{i\in\{1,\cdots,n\}}\{Logits(M_i|x)\}$$

Experimental Results: W/ Llama 2-7B

Q&A Accuracy (publicly available datasets)

dataset	generalized model	LoraHub	PEM Addition	p_tuning	prefix	ia3	lora	adalora
commonsenseqa	82%	63%	71%	20%	20%	48%	79%	64%
helloswag	75%	63%	67%	28%	32%	27%	31%	73%
qasc	42%	22%	35%	15%	12%	12%	31%	42%

SQL Generation. Metric is between ground SQL and LLM SQL parsed into trees => Normalized Unordered Tree Edit Distance (Zhang, 1996)

dataset	generalized model	LoraHub	PEM Addition	prefix	ia3	lora	adalora	adalora (w/db norm)
$Silos_1$	0.02	0.76	0.56	15.12	5.05	0.12	0.42	0.17
$Silos_2$	0.01	0.68	0.43	20.25	2.27	0.15	0.46	0.07
$Silos_3$	0.0	0.95	0.48	8.36	2.64	0.06	0.08	0.03
$Silos_{1\cup2}$	0.36	0.66	0.75	5.27	1.56	0.59	0.83	0.4
$Silos_{1\cup3}$	0.26	0.6	0.73	17.03	4.95	0.53	1.25	0.29
$Silos_{2\cup3}$	0.25	0.62	0.75	17.08	2.85	0.47	0.98	0.35
$Silos_{1\cup2\cup3}$	0.65	0.88	1.57	7.29	2.46	0.68	1.28	0.39

Ablation #1: Obfuscated column names

dataset	generalized model	LoraHub	PEM Addition	p_tuning	prefix	ia3	lora	adalora	lora (w/db norm)
$Silos_1$	0.0	1.68	0.81	2.93	22.37	5.05	0.47	0.41	0.05
$Silos_2$	0.0	0.68	0.55	3.76	31.84	2.27	0.43	0.33	0.14
$Silos_3$	0.0	0.76	0.49	1.63	11.07	2.64	0.16	0.08	0.13
$Silos_{1\cup2}$	0.37	0.91	0.74	2.15	8.12	1.56	0.56	0.65	0.29
$Silos_{1\cup3}$	0.33	1.69	0.7	3.77	23.44	4.95	0.53	1.14	0.33
$Silos_{2\cup3}$	0.47	1.22	0.73	3.92	26.54	2.85	0.46	0.53	0.37
$Silos_{1\cup2\cup3}$	0.82	1.62	1.99	2.71	12.01	2.46	0.78	0.8	0.49

Ablation #2: GPT rephrased questions

dataset	generalized model	LoraHub	PEM Addition	prefix	ia3	lora	adalora	adalora (w/db norm)
$Silos_1$	0.02	0.61	0.4	16.23	2.56	0.2	0.52	0.26
$Silos_2$	0.17	0.8	0.38	23.69	0.69	0.35	0.74	0.24
$Silos_3$	0.11	0.87	0.29	8.05	2.96	0.19	0.32	0.14
$Silos_{1\cup2}$	0.4	1.11	0.57	6.69	1.26	0.53	0.64	0.46
$Silos_{1\cup3}$	0.21	0.59	0.51	17.11	1.29	0.48	0.44	0.2
$Silos_{2\cup3}$	0.26	0.58	0.44	17.49	0.83	0.37	0.47	0.25
$Silos_{1\cup2\cup3}$	0.37	0.7	0.49	9.82	1.7	0.4	0.56	0.23

Analysis:

- Some PEFT are far more composable than others!
  - Potentially a suboptimal PEFT is superior at composing?
- AdaLoRa/LoRa is close to the generalized model in certain cases! (specifically in unions, which are OOD samples)
- Union datasets are essential for benchmarking this framework
  - Generate union datasets without relying on SQL?
- Fine-tunings are destructive, can we make them cooperative.
  - Future work: normalize by novelty metric per PEFT?