SecureLLM

Using Compositionality to Build Provably Secure Language Models for Private, Sensitive, and Secret Data

Abdulrahman Alabdulakreem*, Christian M Arnold*, Yerim Lee, Pieter M Feenstra, Boris Katz, Andrei Barbu

Motivation

- **Avoid Data Leakage**: LLMs are notoriously bad at keeping secrets
- **Problem:** In large corporations or government databases (many distinct data silos), LLM's cannot be trusted to keep confidential data secret from users who don't have credential to access them
- **Solution: SecureLLM.** Finetune on each silo independently -> Compose at inference time based on user credentials

Dataset

- Independent Silos generated using Context-Free Grammar (Questions, SQL, DB Norm. SQL)

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
```

```
want: AVG(teacher_age)
tables: instructors, classes
conditions:
    teacher_age >= 72 OR class_subject =
  'art'
    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
```

```
want: MIN(height)
tables: store, inventory
conditions:
    available = 0
    location = 'NY' OR location = 'CA'
 OR location = 'MA'
    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
```

```
want: name
tables: instructors, store, classes
conditions:
     location = 'TX'
     teacher_age <= 86
     class_subject = 'english' OR
 class_subject = 'sociology' OR
 class_subject = 'art'
     grade >= 89
```

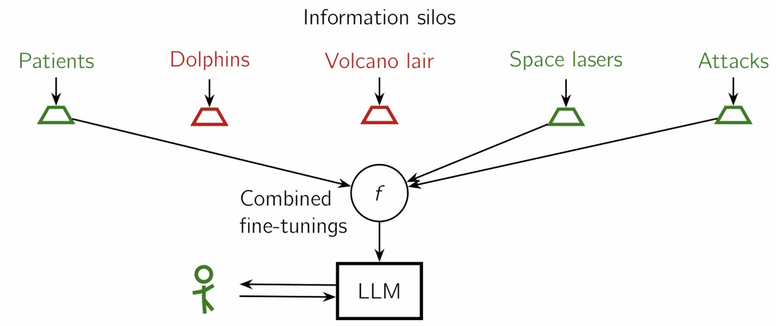
(c) Sample from Union Silo 1,2 ($S_{1\cup2}$)

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.sloth)
FROM inventory INNER JOIN store ON
   store.bear = inventory.bear
WHERE inventory.pony = 0
   AND (store.alpaca = 'NY'
    OR store.alpaca = 'CA'
    OR store.alpaca = 'MA')
   AND store.raccoon >= 2
```

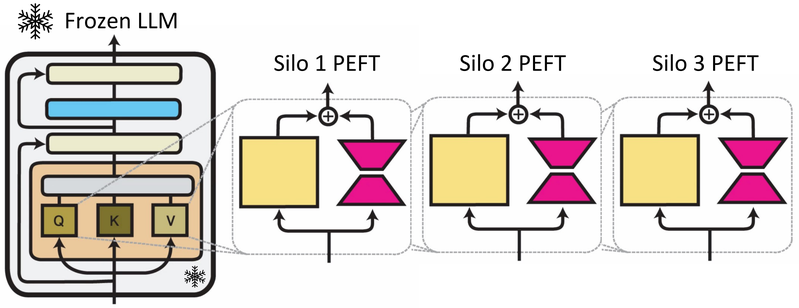
```
want: MIN(sloth)
tables: store, inventory
conditions:
     pony = 0
     alpaca = 'NY' OR alpaca = 'CA' OR
 alpaca = 'MA'
     raccoon >= 2
```

(b) Sample from Silo 2 (S_2) with obfuscated column names



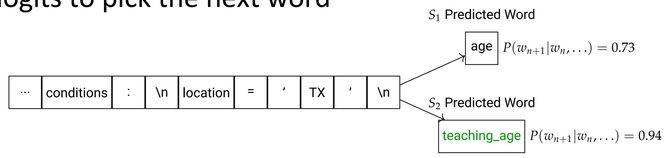
Implementation

- Large transformer network, **Llama-7b**, is frozen while a small Parameter Efficient Fine-Tuning (PEFT) is trained on each information silos
- Low Rank Adapter (LoRA) (Hu et al., 2021)



How it works

- BestLogit independently runs each LoRA and sums predicted logits to pick the next word



Experimental results

- SQL generation using various compositional methods on CFG generated questions (Metric is Tree-Edit distance, lower is better)

CFG Generated	Baseline Exponential Model	Baseline Generalized Model	LoraHub	PEM Addition	Ours (Maximum Difference)	Ours (Logits) Without DB Normalization	**Ours (Logits)**
$Silos_1$	0.0 (100.0%)	0.0 (98.3%)	1.9	1.0	0.4	0.4	**0.1**
$Silos_2$	0.0 (96.7%)	0.0 (100.0%)	2.7	0.7	0.3	0.2	**0.0**
$Silos_3$	0.0 (100.0%)	0.0 (100.0%)	1.2	0.7	0.2	**0.1**	**0.1**
$Silos_{1\cup2}$	0.0 (98.3%)	1.0 (0.0%)	1.8	1.0	0.9	0.7	**0.3**
$Silos_{1\cup3}$	0.0 (99.2%)	0.5 (0.0%)	1.6	0.7	0.7	0.6	**0.2**
$Silos_{2\cup3}$	0.0 (100.0%)	0.4 (1.7%)	1.7	0.7	0.7	0.5	**0.4**
$Silos_{1\cup2\cup3}$	0.0 (100.0%)	0.5 (1.7%)	2.2	0.7	0.7	0.4	**0.4**
$\mu \pm \sigma$	0.00 ± 0.00	0.35 ± 0.35	1.88 ± 0.45	0.78 ± 0.13	0.55 ± 0.24	0.42 ± 0.21	$\mathbf{0.21 \pm 0.12}$

- SQL generation on obfuscated schemas

Obfuscated Generated	Baseline Exponential Model	Baseline Generalized Model	LoraHub	PEM Addition	Ours (Maximum Difference)	Ours (Logits) Without DB Normalization	**Ours (Logits)**
$Silos_1$	0.0 (99.2%)	0.0 (94.2%)	2	1.1	0.5	0.5	**0.2**
$Silos_2$	0.0 (92.5%)	0.0 (100.0%)	3.1	1.4	0.5	**0.4**	**0.4**
$Silos_3$	0.0 (100.0%)	0.0 (100.0%)	0.9	0.8	0.5	0	**0.1**
$Silos_{1\cup2}$	0.0 (80.8%)	0.7 (0.8%)	1.6	2.2	1.1	0.9	**0.5**
$Silos_{1\cup3}$	0.0 (98.3%)	0.4 (0.0%)	1.3	1.4	0.7	0.6	**0.4**
$Silos_{2\cup3}$	0.0 (77.5%)	0.6 (1.7%)	1.6	2.5	0.9	**0.7**	0.8
$Silos_{1\cup2\cup3}$	0.0 (100.0%)	0.4 (1.7%)	1.9	2.5	1.0	0.6	**0.3**
$\mu \pm \sigma$	0.01 ± 0.01	0.31 ± 0.28	1.76 ± 0.65	$1.70x \pm 0.65$	0.73 ± 0.24	0.54 ± 0.24	$\mathbf{0.36 \pm 0.22}$

- GPT rephrased questions (ensures no overfitting)

GPT Generated	Baseline Exponential Model	Baseline Generalized Model	LoraHub	PEM Addition	Ours (Maximum Difference)	Ours (Logits) Without DB Normalization	**Ours (Logits)**
$Silos_1$	0.0 (87.5%)	0.1 (79.2%)	2.0	1.1	0.5	0.6	**0.3**
$Silos_2$	0.2 (60.8%)	0.2 (56.7%)	2.9	1.0	0.5	**0.3**	0.4
$Silos_3$	0.1 (56.7%)	0.2 (51.7%)	1.4	1.1	0.5	0.4	**0.2**
$Silos_{1\cup2}$	0.2 (20.8%)	0.4 (0.0%)	2.1	0.7	0.6	0.6	**0.2**
$Silos_{1\cup3}$	0.2 (29.2%)	0.4 (0.0%)	1.6	1.0	0.6	0.6	**0.4**
$Silos_{2\cup3}$	0.1 (33.3%)	0.3 (3.3%)	2.3	0.6	0.5	0.4	**0.3**
$Silos_{1\cup2\cup3}$	0.1 (50.0%)	0.3 (2.5%)	2.1	0.6	0.5	0.4	**0.2**
$\mu \pm \sigma$	0.15 ± 0.06	0.28 ± 0.13	2.06 ± 0.45	0.85 ± 0.21	0.52 ± 0.05	0.48 ± 0.11	$\mathbf{0.30 \pm 0.07}$