LLM2LLM

2024 Lab Seminar

인간중심컴퓨팅연구실 박사과정 류정우

Challenges of Task-specific LLM

Prompting strategies are not a one-size-fits-all solution

- 입력 컨텍스트의 한계: LLM들은 입력에 대한 최대 토큰 한도를 가지고 있어, 특정 작업에 대한 모델의 출력을 지시하기 위해 **입력할 수 있는 정보(예시)의 양에 제약**을 받음
- 전문 영역의 프롬프트 엔지니어링: **특정 전문 분야(e.g. 법, 의학 등)에서 원하는 결과로 이끄는 프롬프트를 개발하는 것은 특히 어려울 수 있으며**, 때로는 긴 프롬프트가 필요하게 될 수 있음
- 지연 시간 및 비용 문제: 긴 프롬프트를 처리하는 것은 더 많은 자원을 사용하고, 추론 시간을 늘려, **높은 지연 시간과 비용**을 초래함
- 정확성 및 컨텍스트 잊어버림: 긴 프롬프트가 입력 한도 내에 있더라도, **LLM은 제공된 모든 컨텍스트를 유지하거나 정확하게 처리하기 어려울 수 있음**. 긴 프롬프트에서 초점을 잃거나 관련성을 잃기가 쉬움

Fine-tuning LLM

Successful fine-tuning requires “enough training data”

Key question is:

“How should we increase the user’s training data to be enough for fine-tuning?”

- *Traditional approaches: synonym/character replacement, random swapping, back translation*
- *LLM-based approaches: augmentation on **ALL** of the available training dataset(e.g. AugGPT)*

Paper

LLM2LLM

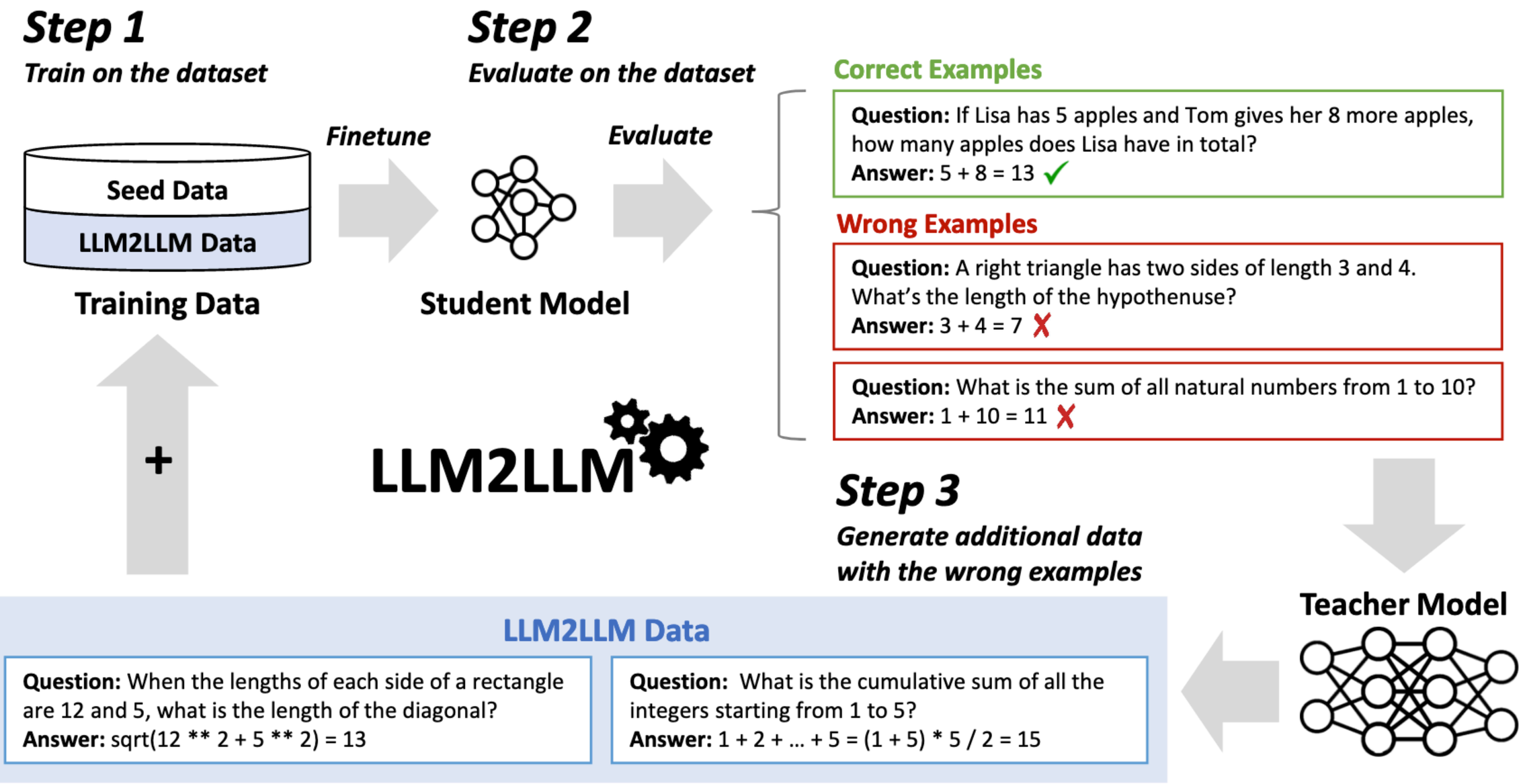
LLM2LLM: Boosting LLMs with Novel Iterative Data Enhancement

Nicholas Lee*[1] Thanakul Wattanawong*[1] Sehoon Kim[1]
Karttikeya Mangalam[1] Sheng Shen[1] Gopala Anumanchipali[1]
Michael W. Mahoney[1,2,3] Kurt Keutzer[1] Amir Gholami[1,2]

[1] UC Berkeley [2] ICSI [3] LBNL

{nicholas.lee, j.wat, sehoonkim, mangalam, s.sheng, gopala, mahoneymw, keutzer, amirgh}@berkeley.edu

Overview


Step 1
Train on the dataset
Seed Data
LLM2LLM Data
Training Data
Finetune
Step 2
Evaluate on the dataset
Student Model
Evaluate
Correct Examples
Question: If Lisa has 5 apples and Tom gives her 8 more apples, how many apples does Lisa have in total?
Answer: 5 + 8 = 13
Wrong Examples
Question: A right triangle has two sides of length 3 and 4. What's the length of the hypothenuse?
Answer: 3 + 4 = 7
Question: What is the sum of all natural numbers from 1 to 10?
Answer: 1 + 10 = 11
+
LLM2LLM
Step 3
Generate additional data with the wrong examples
Teacher Model
LLM2LLM Data
Question: When the lengths of each side of a rectangle are 12 and 5, what is the length of the diagonal?
Answer: sqrt(12 ** 2 + 5 ** 2) = 13
Question: What is the cumulative sum of all the integers starting from 1 to 5?
Answer: 1 + 2 + ... + 5 = (1 + 5) * 5 / 2 = 15

Algorithm

Algorithm 1 LLM2LLM: Boosting LLMs with Novel Iterative Data Enhancement. Given a seed dataset D^0, we finetune the model $\mathcal{M}^i_{\text{student}}$, evaluate, and extract training set data points that the model gets wrong. These are used to generate new training data points using the teacher model $\mathcal{M}_{\text{teacher}}$ for the next step.

procedure LLM2LLM($\mathcal{M}^0_{\text{student}}$, $\mathcal{M}_{\text{teacher}}$, D^0)
  $i \leftarrow 0$
  while $i < n$ **do**
    $\mathcal{M}^i_{\text{student}} \leftarrow \text{Finetune}(\mathcal{M}^0_{\text{student}}, D^i)$
    $E^i \leftarrow \text{Evaluate}(\mathcal{M}^i_{\text{student}}, D^0)$ ▷ Evaluate on seed data
    $W^i \leftarrow \text{Filter}(E^i, D^0)$ ▷ Keep wrong answers
    $A^i \leftarrow \text{Generate}(\mathcal{M}_{\text{teacher}}, W^i)$ ▷ Augment using teacher
    $D^{i+1} \leftarrow D^i + A^i$ ▷ Append to data
    $i \leftarrow i + 1$
  end while
  Evaluate M^*_{student}
end procedure

Experimental Setup

Model and Datasets

- **Model**
  - Student Model: LLaMA2-7B
  - Teacher Model: GPT-3.5(1106)
- **Datasets**
  - GSM8K: 8.5K school math problems
  - CaseHOLD: 53K+ legal holdings
  - SNIPS: 16K+ 7 user intents
  - TREC: 6K question classification
  - SST-2: 70K sentiment classification

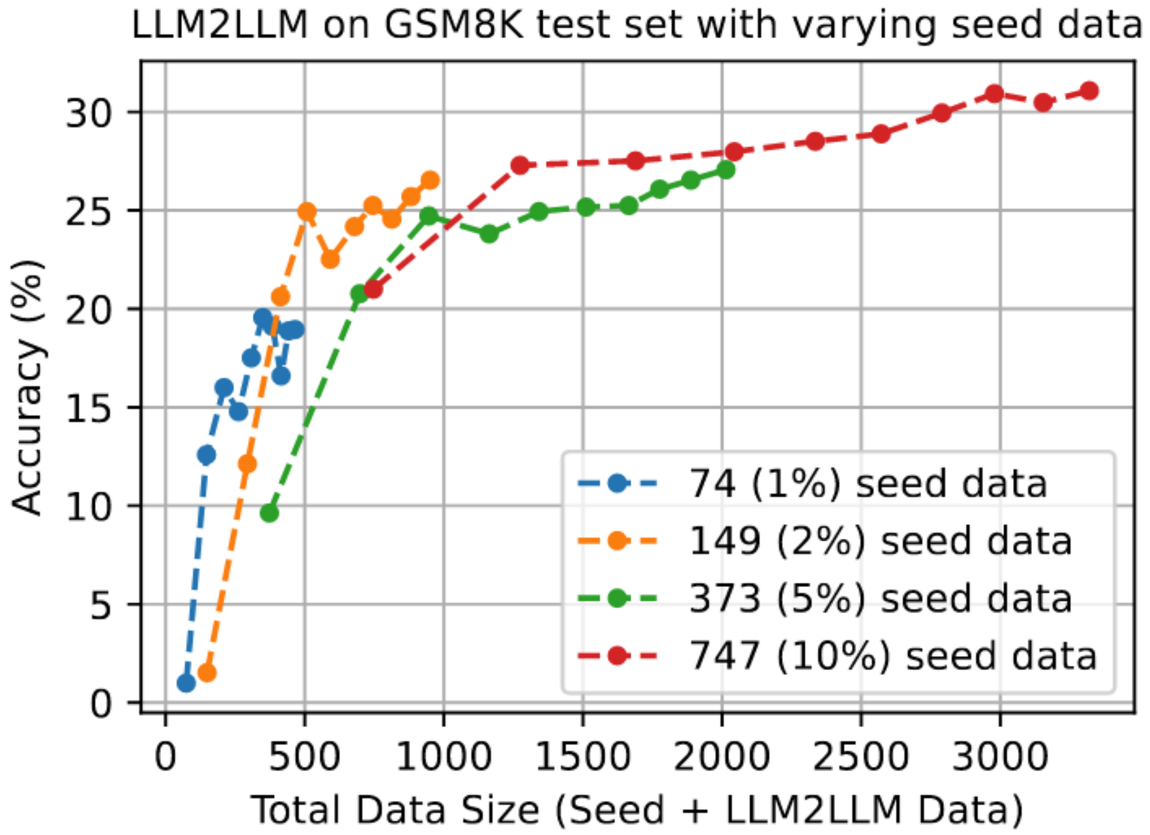
Results

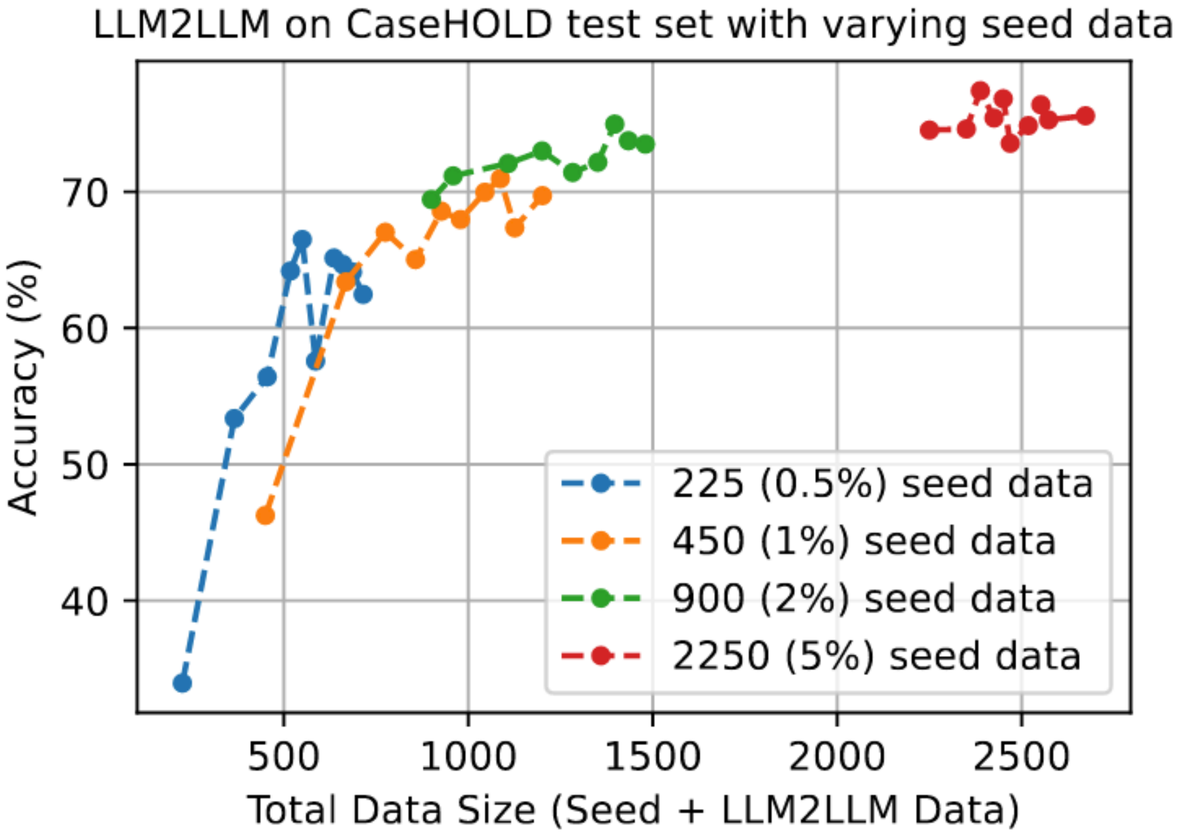
Seed data 수가 적을 수록 augmentation 효과가 크게 나타남

Dataset	% Data	# Seed Examples	# Augmented	Test Accuracy (%)	
				Baseline	LLM2LLM
GSM8K	0%	0	0	0.00	**N/A**
	1%	74	391	0.99	**19.56**
	2%	149	802	1.52	**25.70**
	5%	373	1641	9.63	**27.07**
	10%	747	2573	21.27	**30.93**
	20%	1494	4028	25.70	**35.03**
	50%	3737	8252	33.89	**38.67**
	100%	7473	14925	36.01	**41.24**
CaseHOLD	0%	0	0	12.28	**N/A**
	0.5%	225	490	33.94	**66.50**
	1%	450	751	46.25	**70.97**
	2%	900	580	69.44	**74.97**
	5%	2250	423	74.14	**76.83**
	10%	4500	505	77.03	**78.21**
	20%	9000	1100	78.00	**78.97**
	50%	22500	2709	80.39	**82.92**
	100%	45000	5805	87.94	**88.14**
SNIPS	0%	0	0	11.86	**N/A**
	0.5%	70	38	60.14	**92.14**
	0.8%	105	109	69.71	**93.71**
	1.0%	140	91	85.43	**93.86**
TREC	0%	0	0	11.20	**N/A**
	1.1%	60	135	26.20	**78.80**
	1.6%	90	22	80.80	**90.20**
	2.2%	120	44	81.20	**91.20**
SST-2[1]	0%	0	0	27.06	**N/A**
	0.02%	20	44	52.87	**92.66**
	0.04%	30	46	62.04	**93.00**
	0.06%	40	14	82.80	**94.04**

Results

LLM2LLM on GSM8K and CaseHOLD with various seed data size





Results

Comparison with other methods

- **Other Methods**
  - EDA: Easy Data Augmentation
  - AugGPT: GPT-based Augmentation

Dataset	Technique	# Seed	Total Aug.	Acc. (%)
GSM8K	Fine-tuning	100	0	1.59
	EDA		500	15.16
	AugGPT		500	18.12
	More Data		471	19.86
	LLM2LLM		471	**23.73**
CaseHOLD	Fine-tuning	100	0	28.78
	EDA		200	62.19
	AugGPT		200	63.42
	More Data		198	37.11
	LLM2LLM		198	**64.50**
SNIPS	Fine-tuning	70	0	60.14
	EDA		70	91.43
	AugGPT		70	89.86
	More Data		70	89.00
	LLM2LLM		38	**92.14**
TREC	Fine-tuning	60	0	26.20
	EDA		120	72.40
	AugGPT		120	32.80
	More Data		138	**89.20**
	LLM2LLM		135	78.80
SST-2	Fine-tuning	20	0	52.87
	EDA		40	63.19
	AugGPT		40	88.07
	More Data		40	72.94
	LLM2LLM		44	**92.66**

Results

Other teacher models

Dataset	# Seed	Teacher	Total # Aug	Accuracy (%)
GSM8K	74 (1%)	LLaMA2-70B	333	11.83
		Airoboros	345	15.01
		GPT-3.5	391	19.56
		GPT-4-Turbo	388	**19.79**
	149 (2%)	LLaMA2-70B	661	17.59
		Airoboros	671	19.33
		GPT-3.5	802	25.70
		GPT-4-Turbo	805	**25.78**
	343 (5%)	LLaMA2-70B	1308	19.33
		Airoboros	1286	21.76
		GPT-3.5	1641	27.07
		GPT-4-Turbo	1739	**28.43**

Table A.1: Experiments on how the quality of teacher model affects the performance of LLM2LLM. For each of these experiments, we only change the teacher model to measure the effect of the teacher model on the final outcome.

Conclusion

+ Future work

- 실제 데이터가 필요한 양을 크게 줄이고, 수작업으로 데이터를 더 모으는 것과 같거나 그 이상의 효과를 낼 수 있는 합성 데이터로 데이터셋을 효율적으로 확장
- **틀린 문제에 집중!** - LLM이 틀린 데이터 포인트에서 신호를 강화할 수 있어 방법이 효과적
- LLaMA-2-7B를 Student 모델로 사용하여 적은 데이터에서 GSM8K에서 24.2%, CaseHOLD에서 32.6%, SNIPS에서 32.9%, TREC에서 57.6%, 그리고 SST-2 데이터셋에서 39.8%의 **성능 향상**
- 향후 하이퍼파라미터를 조정하고, Few-shot 학습과 같은 **다른 LLM 기술과 접목**하는 방향에 초점을 맞출 수 있음

Example of generated training data

Seed Data Example: Katrina saw an ad that said if she signed up for her local recycling program, she could earn $5.00. When she signed up, they told her for every friend that she referred, the friend would receive $5.00 and she would receive another $5.00 per friend. That day, she had 5 friends sign up and another 7 friends by the end of the week. How much money in total did she and her friends make?

GPT-4-Turbo: **Michael joined a library program that rewards $4.00 for becoming a new member. Additionally, if he recommends the program to his friends, each friend gets $4.00 and Michael also gets $4.00 for each referee. Michael got 3 friends to sign up on the first day and another 6 friends joined the program over the next two weeks.** How much money in total did Michael and his friends earn?

GPT-3.5: Sophia saw an ad stating that if she signed up for her local fitness center, she could receive $10 as a reward. **For every friend she refers, the friend receives $10, and Sophia receives an additional $10 per friend. On the first day, she referred 3 friends, and by the end of the week, she referred 6 more friends.** How much money did Sophia and her friends make in total?

Airoboros: Sophie saw an ad that said if she signed up for her local recycling program, she could earn $3.00. When she signed up, they told her for every friend that she referred, the friend would receive $3.00 and she would receive another $3.00 per friend. That day, she had 7 friends sign up and another 9 friends by the end of the week. How much money in total did she and her friends make?

LLaMA2-70B: **Joe signed up for his local recycling program and was promised $10 for every friend he referred.** When he signed up, they told him for every friend that he referred, the friend would receive $10 and he would receive another $10 per friend. That day, he had 3 friends sign up and another 8 friends by the end of the week. How much money in total did he and his friends make?

Thank you