Large Language Models

CSC413 Tutorial 9

Yongchao Zhou
Overview

● What are LLMs?
● Why LLMs?
● Emergent Capabilities
  ○ Few-shot In-context Learning
  ○ Advanced Prompt Techniques
● LLM Training
  ○ Architectures
  ○ Objectives
● LLM Finetuning
  ○ Instruction finetuning
  ○ RLHF
  ○ Bootstrapping
● LLM Risks
What are Language Models?

● Narrow Sense
  ○ A probabilistic model that assigns a probability to every finite sequence (grammatical or not)

  Sentence: “the cat sat on the mat”

  \[ P(\text{the cat sat on the mat}) = P(\text{the}) \times P(\text{cat|the}) \times P(\text{sat|the cat}) \]
  
  \[ \times P(\text{on|the cat sat}) \times P(\text{the|the cat sat on}) \]
  
  \[ \times P(\text{mat|the cat sat on the}) \]

  Implicit order

● Broad Sense
  ○ Decoder-only models (GPT-X, OPT, LLaMA, PaLM)
  ○ Encoder-only models (BERT, RoBERTa, ELECTRA)
  ○ Encoder-decoder models (T5, BART)
Large Language Models - **Billions of Parameters**

![Graph showing the growth of model size in billions of parameters from 2018 to 2022](https://huggingface.co/blog/large-language-models)
Large Language Models - **Hundreds of Billions of Tokens**

- **<100 Million Tokens**
  - Human
  - 13 y.o. BERT (2018)

- **3 Billion Tokens**
  - RoBERTa (2019)

- **30 Billion Tokens**
  - GPT-3 (2020)

- **1.4 Trillion Tokens**
  - Chinchilla (2022)

# tokens seen during training

[https://babylm.github.io/](https://babylm.github.io/)
Large Language Models - yottaFlops of Compute

AI training runs, estimated computing resources used
Floating-point operations, selected systems, by type, log scale

https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf
Why LLMs?

- Scaling Law for Neural Language Models
  - Performance depends strongly on scale! We keep getting better performance as we scale the model, data, and compute up!

Why LLMs?

- **Generalization**
  - We can now use one single model to solve many NLP tasks

Why LLMs?

- **Emergent Abilities**
  - Some ability of LM is not present in smaller models but is present in larger models

https://docs.google.com/presentation/d/1yzbmYB5E7G8lY2-KzhmArmPYwwl7o7CUST1xRZDUu1Y/edit?resourcekey=0-6_TnUMoKWCk_FN2BiPxmbw#slide=id.q1fc34b3ac18_0_27
Emergent Capability - **In-Context Learning**

### Traditional fine-tuning (not used for GPT-3)

#### Fine-tuning
The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1) sea otter => loutre de mer
   gradient update

2) peppermint => menthe poivrée
   gradient update
   ...

N) plush giraffe => girafe peluche
   gradient update
```

### One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1) Translate English to French:

2) sea otter => loutre de mer

3) cheese => ...........................................
```

### Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1) Translate English to French:

2) sea otter => loutre de mer

3) peppermint => menthe poivrée

4) plush giraffe => girafe peluche

5) cheese => ...........................................
```

## Emergent Capability - **In-Context Learning**

<table>
<thead>
<tr>
<th>No Prompt</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zero-shot (0s)</strong></td>
<td><strong>1-shot (1s)</strong></td>
</tr>
<tr>
<td>skicts = sticks</td>
<td>chiar = chair skicts = sticks</td>
</tr>
<tr>
<td><strong>Few-shot (FS)</strong></td>
<td><strong>Prompt</strong></td>
</tr>
<tr>
<td>chiar = chair [...] pciinc = picnic skicts = sticks</td>
<td>Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks</td>
</tr>
</tbody>
</table>

Emergent Capability - **In-Context Learning**
Pretraining + Fine-tuning Paradigm
Pretraining + Prompting Paradigm

- **Fine-tuning (FT)**
  - + Strongest performance
  - - Need curated and labeled dataset for each new task (typically 1k-100k ex.)
  - - Poor generalization, spurious feature exploitation

- **Few-shot (FS)**
  - + Much less task-specific data needed
  - + No spurious feature exploitation
  - - Challenging

- **One-shot (1S)**
  - + "Most natural," e.g. giving humans instructions
  - - Challenging

- **Zero-shot (OS)**
  - + Most convenient
  - - Challenging, can be ambiguous

- Stronger task-specific performance
- More convenient, general, less data
Emergent Capability - **Chain of Thoughts Prompting**

### Standard Prompting

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Model Output**

A: The answer is 27. ❌

### Chain-of-Thought Prompting

**Model Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Model Output**

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✔️

Emergent Capability - Chain of Thoughts Prompting

Math Word Problems (free response)
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Math Word Problems (multiple choice)
Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1565 (b) 1392 (c) 1480 (d) 1562 (e) 1786
A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

CSQA (commonsense)
Q: Sammy wanted to go to go where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock
A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don’t have a lot of people, but populated areas do. So the answer is (b).

StrategyQA
Q: Yes or no: Would a pear sink in water?
A: The density of a pear is about 0.6 g/cm³, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding
Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?
A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding
Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."
A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)
Human: How would you bring me something that isn’t a fruit?
Explanation: the user wants something to eat that isn’t a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation
Q: Take the last letters of the words in “Lady Gaga” and concatenate them.
A: The last letter of “Lady” is “y”. The last letter of “Gaga” is “a”. Concatenating them is “ya”. So the answer is ya.

Coin Flip (state tracking)
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?
A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Emergent Capability - **Zero Shot CoT Prompting**

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer is 8. ✗

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is 8

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓
Emergent Capability - **Zero Shot CoT Prompting**

### 1st Prompt: Reasoning Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

A: Let's think step by step.

- In one minute, Joe throws 25 punches.
- In three minutes, Joe throws $3 \times 25 = 75$ punches.
- In five rounds, Joe throws $5 \times 75 = 375$ punches.

### 2nd Prompt: Answer Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ... A: Let's think step by step.

- In one minute, Joe throws 25 punches.
- In five rounds, Joe throws $5 \times 75 = 375$ punches.

Therefore, the answer (arabic numerals) is 375.

[Arxiv Link](https://arxiv.org/pdf/2205.11916.pdf)
Emergent Capability - **Self-Consistency Prompting**

**Chain-of-thought prompting**

**Prompt**

**Language model**

**Greedy decode**

This means she uses $3 + 4 = 7$ eggs every day. She sells the remainder for $2 per egg, so in total she sells $7 * \$2 = \$14$ per day.

The answer is $\$14$.

**The answer is $\$14$.**

---

**Self-consistency**

**Q:** If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

**A:** There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.

...  

**Q:** Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?

**A:**

**Language model**

**Sample a diverse set of reasoning paths**

- She has $16 - 3 - 4 = 9$ eggs left. So she makes $2 * 9 = \$18$ per day.
  - The answer is $\$18$.

- This means she she sells the remainder for $2 * (16 - 4 - 3) = \$26$ per day.
  - The answer is $\$26$.

- She eats 3 for breakfast, so she has $16 - 3 = 13$ left. Then she bakes muffins, so she has $13 - 4 = 9$ eggs left. So she has $9$ eggs * $\$2 = \$18$.
  - The answer is $\$18$.

**Marginalize out reasoning paths to aggregate final answers**

Emergent Capability - Least-to-Most Prompting

**Problem Reduction**

Stage 1

Q: It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

A: To solve “How many times can she slide before it closes?”, we need to first solve: “How long does each trip take?”

Stage 2

**Sequentially Solve Subquestions**

Subquestion 1

Q: How long does each trip take?

A: It takes Amy 4 minutes to climb and 1 minute to slide down. $4 + 1 = 5$. So each trip takes 5 minutes.

Subquestion 2

Q: How many times can she slide before it closes?

A: The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide $15 \div 5 = 3$ times before it closes.

Emergent Capability - Augmented Prompting Abilities

**Advanced Prompting Techniques**
- Zero-shot CoT Prompting
- Self-Consistency
- Divide-and-Conquer

**Ask a human to**
- Explain the rationale
- Double check the answer
- Decompose to easy subproblems

Large Language Models demonstrate some human-like behaviors!
Training Architectures

Encoder-decoder models (T5, BART)

Decoder-only models (GPT-X, PaLM)

http://jalammar.github.io/illustrated-transformer/
### Training Objectives - UL2

#### R-Denoising

<table>
<thead>
<tr>
<th>Inputs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>He dealt in archetypes before anyone knew such things existed, and his ability to take an emotion or a situation to the limit helped create a cadre of plays that have been endlessly staged – and copied. Apart from this, Romeo and Juliet inspired Malorie. Blackman's Noughts &amp; Crosses, there are references to Hamlet in Lunar Park by Bret Easton Ellis. The Tempest was the cue for The Magus by John Fowles.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target:</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 5 4 5</td>
</tr>
<tr>
<td>2 <em>E</em></td>
</tr>
</tbody>
</table>

#### S-Denoising

<table>
<thead>
<tr>
<th>Inputs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>He dealt in archetypes before anyone knew such things existed, and his ability to take an emotion or a situation and push it to the limit helped create a cadre of plays that have been endlessly staged – and copied. Apart from this, Romeo and Juliet inspired Malorie. Blackman's Noughts &amp; Crosses, there are references to Hamlet in Lunar Park by Bret Easton Ellis. The Tempest was the cue for The Magus by John Fowles.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target:</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 32</td>
</tr>
<tr>
<td>24</td>
</tr>
</tbody>
</table>

#### X-Denoising

<table>
<thead>
<tr>
<th>Inputs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>He dealt in archetypes before anyone knew such things existed, and his ability to take an emotion or a situation and push it to the limit helped create a cadre of plays that have been endlessly staged – and copied. Apart from this, Romeo and Juliet inspired Malorie. Blackman's Noughts &amp; Crosses, there are references to Hamlet in Lunar Park by Bret Easton Ellis. The Tempest was the cue for The Magus by John Fowles.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target:</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 3 5 4 5</td>
</tr>
<tr>
<td>3 2 4 2</td>
</tr>
</tbody>
</table>

---

What kinds of things does pretraining learn?

- *Stanford University is located in __________, California.* [Trivia]
- *I put ___ fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over ___ shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and ____.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___.* [sentiment]
- *Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the _____.* [some reasoning – this is harder]
- *I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ____.* [some basic arithmetic; they don’t learn the Fibonacci sequence]

https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf
Finetune - Instruction Finetune

Instruction finetuning

Please answer the following question. What is the boiling point of Nitrogen?

-320.4F

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step. The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.

Multi-task instruction finetuning (1.8K tasks)

Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington? Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is “no”.

**Finetune - Instruction Finetune**

### Finetuning tasks

- **TO-SF**
  - Commonsense reasoning
  - Question generation
  - Closed-book QA
  - Adversarial QA
  - Extractive QA
  - Title/context generation
  - Topic classification
  - Struct-to-text
  
  55 Datasets, 14 Categories, 193 Tasks

- **Muffin**
  - Natural language inference
  - Code instruction gen.
  - Program synthesis
  - Dialog context generation
  - Closed-book QA
  - Conversational QA
  - Code repair

  69 Datasets, 27 Categories, 80 Tasks

- **CoT (Reasoning)**
  - Arithmetic reasoning
  - Commonsense Reasoning
  - Implicit reasoning
  - Explanation generation
  - Sentence composition

  9 Datasets, 1 Category, 9 Tasks

- **Natural Instructions v2**
  - Cause effect classification
  - Commonsense reasoning
  - Named entity recognition
  - Toxic language detection
  - Question answering
  - Question generation
  - Program execution
  - Text categorization

  372 Datasets, 108 Categories, 1554 Tasks

- A **Dataset** is an original data source (e.g. SQuAD).
- A **Task Category** is unique task setup (e.g. the SQuAD dataset is configurable for multiple task categories such as extractive question answering, query generation, and context generation).
- A **Task** is a unique <dataset, task category> pair, with any number of templates which preserve the task category (e.g. query generation on the SQuAD dataset.)

Without chain-of-thought

Instruction without exemplars

Answer the following yes/no question.
Can you write a whole Haiku in a single tweet?

Yes

Instruction with exemplars

Q: Answer the following yes/no question.
Could a dandelion suffer from hepatitis?
A: no

Q: Answer the following yes/no question.
Can you write a whole Haiku in a single tweet?
A:

Yes

With chain-of-thought

Answer the following yes/no question by reasoning step-by-step.
Can you write a whole Haiku in a single tweet?

A haiku is a Japanese three-line poem. That is short enough to fit in 280 characters. The answer is yes.

Q: Answer the following yes/no question by reasoning step-by-step.
Could a dandelion suffer from hepatitis?
A: Hepatitis only affects organisms with livers. Dandelions don’t have a liver. The answer is no.

Q: Answer the following yes/no question by reasoning step-by-step.
Can you write a whole Haiku in a single tweet?
A:

A haiku is a Japanese three-line poem. That is short enough to fit in 280 characters. The answer is yes.

Finetune - Instruction Finetune

Step 1 Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2 Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3 Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

### Application - **ChatGPT**

**ChatGPT**

<table>
<thead>
<tr>
<th>Examples</th>
<th>Capabilities</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Explain quantum computing in simple terms&quot; →</td>
<td>Remembers what user said earlier in the conversation</td>
<td>May occasionally generate incorrect information</td>
</tr>
<tr>
<td>&quot;Got any creative ideas for a 10 year old’s birthday?&quot; →</td>
<td>Allows user to provide follow-up corrections</td>
<td>May occasionally produce harmful instructions or biased content</td>
</tr>
<tr>
<td>&quot;How do I make an HTTP request in Javascript?&quot; →</td>
<td>Trained to decline inappropriate requests</td>
<td>Limited knowledge of world and events after 2021</td>
</tr>
</tbody>
</table>
Application - **ChatGPT**

https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Source-b9a57ac0fcf74f30a1ab9e3e36fa1dc1
Q: What can be used to carry a small dog?

Answer Choices:
(a) swimming pool
(b) basket
(c) dog show
(d) backyard
(e) own home

A: The answer must be something that can be used to carry a small dog. Baskets are designed to hold things. Therefore, the answer is basket (b).
Finetune - **Bootstrapping**

- Q: John buys 20 cards and 1/4 are uncommon. How many uncommon cards did he get? A: John gets $20 \times \frac{1}{4} = 5$ uncommon cards. The answer is 5.
- Q: Amy is 10. Jake is 8. Alex’s age is right in the middle. How old is Alex? A: (8+10)/2 = 9. The answer is 9.

---

**Input:**
Q: ... How old is Alex? A: Let's think step-by-step.

---

**Output:**
- (8+10)/2 = 9 ... answer is 9.
- Alex is 9 years old ...
- ...

---

**CoT examples**
- Training-set questions or self-generated questions

---

**Language Model**

---

**Multiple path decoding**

---

**Self-training**

---

**Mixed formats of selected reasoning paths**

---

Large Language models Risks

- LLMs make mistakes
  (falsehoods, hallucinations)
- LLMs can be misused
  (misinformation, spam)
- LLMs can cause harms
  (toxicity, biases, stereotypes)
- LLMs can be attacked
  (adversarial examples, poisoning, prompt injection)
- LLMs can be useful as defenses
  (content moderation, explanations)
Resources for further reading

- https://web.stanford.edu/class/cs224n/
- https://stanford-cs324.github.io/winter2022/
- https://stanford-cs324.github.io/winter2023/
- https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
- https://rycolab.io/classes/llm-s23/
- https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1d
  c1
- https://www.jasonwei.net/blog/emergence
Emergent Capability - **In-Context Learning**

Learning via SGD during unsupervised pre-training

- **In-context learning**
  - Sequence #1:
    - 1. $5 + 8 = 13$
    - 2. $7 + 2 = 9$
    - 3. $1 + 0 = 1$
    - 4. $3 + 4 = 7$
    - 5. $5 + 9 = 14$
    - 6. $9 + 8 = 17$

- **In-context learning**
  - Sequence #2:
    - 1. goat $\Rightarrow$ goat
    - 2. saikne $\Rightarrow$ snake
    - 3. bried $\Rightarrow$ bird
    - 4. fishe $\Rightarrow$ fish
    - 5. dcuk $\Rightarrow$ duck
    - 6. cmihp $\Rightarrow$ chimp

- **In-context learning**
  - Sequence #3:
    - 1. thanks $\Rightarrow$ merci
    - 2. hello $\Rightarrow$ bonjour
    - 3. mint $\Rightarrow$ menthe
    - 4. wall $\Rightarrow$ mur
    - 5. otter $\Rightarrow$ loutre
    - 6. bread $\Rightarrow$ pain

Emergent Capability - **Decomposed Prompting**

[Diagram showing the process of decomposed prompting]

- **Standard Prompting**
- **Chain-of-Thought Prompting**
- **Decomposer Prompt**
- **Sub-Task Handlers**

**Example:**
- **Q1:** [split] What are the words in "John Smith"?
  - A1: ["John", "Smith"]
- **Q2:** [foreach] [str_pos] What is the second letter in #1?
  - A2: ["o", "m"]
- **Q3:** [merge] Concatenate #2 with spaces
  - A3: ["o m"]
- **Q4:** [EOQ] A: ["o m"]

Training Objectives - UL2

Training Techniques - **Parallelism**

An illustration of various parallelism strategies on a three-layer model. Each color refers to one layer and dashed lines separate different GPUs.

Training Techniques - **Parallelism**