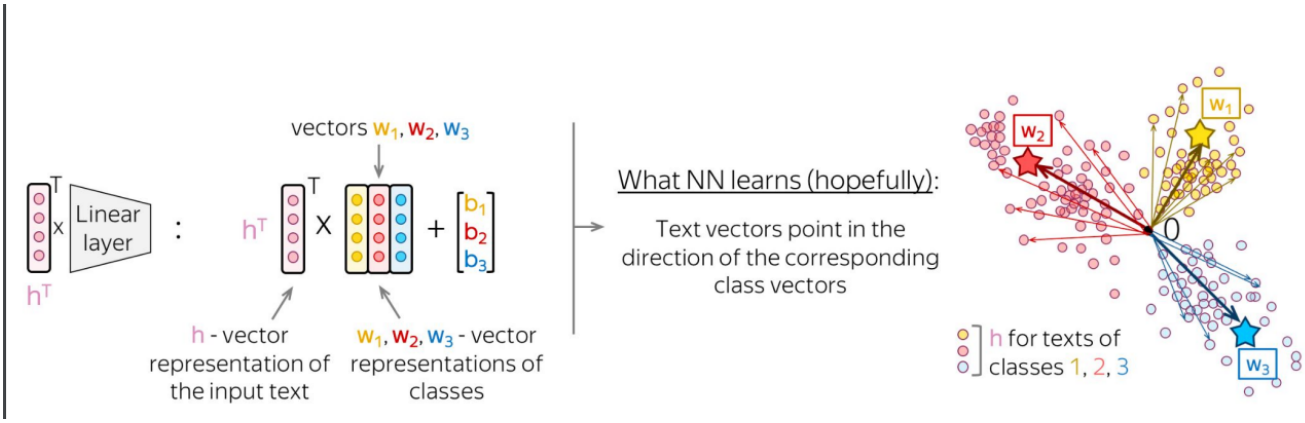
FOLT Lecture 6 , Prompt Engineering for Text Classification

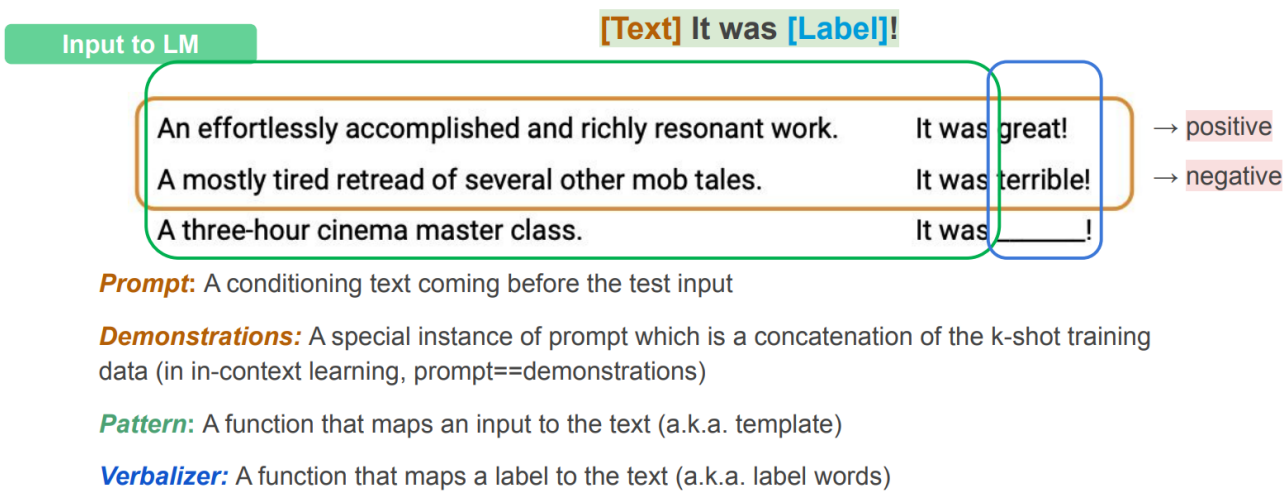
Small Intro



w_1, w_2, w_3 are class vectors:(representing positive ,negative and neutral for example)
The small colored dots are input texts

PART 1: Prompt Overview

Idea: the input text is us talking to the model , prompt engineering is the art of designing that text



Prompt: A conditioning text coming before the test input

Demonstrations: A special instance of prompt which is a concatenation of the k-shot training data (in in-context learning, prompt==demonstrations)

Pattern: A function that maps an input to the text (a.k.a. template)

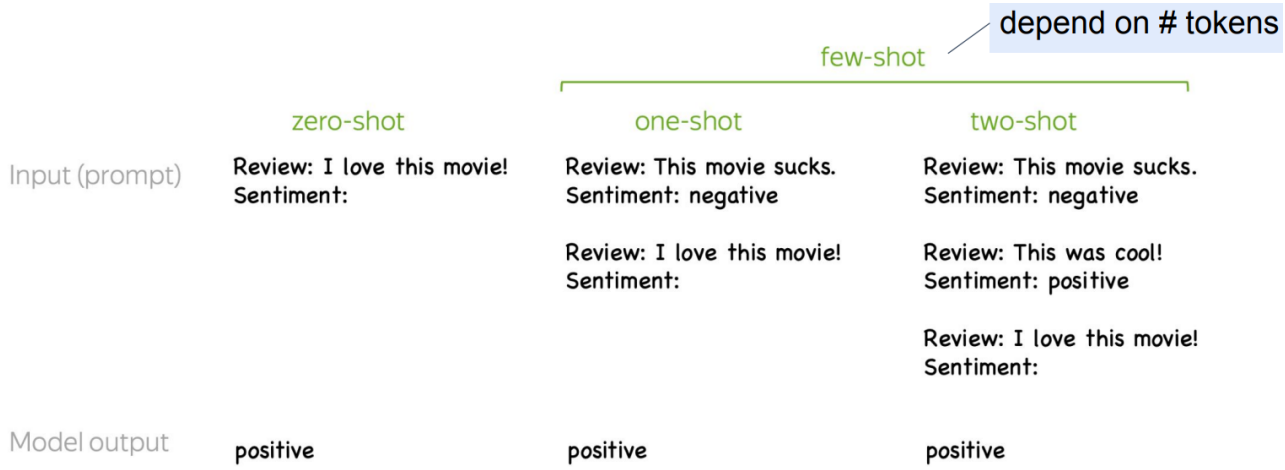
Verbalizer: A function that maps a label to the text (a.k.a. label words)

Formal Description of Prompting

Name	Notation	Example	Description
Input	x	I love this movie.	One or multiple texts
Output	y	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(x)$	`[X]` Overall, it was a `[Z]` movie.	A function that converts the input into a specific form by inserting the input x and adding a slot `[Z]` where answer z may be filled later.
Prompt	x'	I love this movie. Overall, it was a `[Z]` movie.	A text where `[X]` is instantiated by input x but answer slot `[Z]` is not.
Filled Prompt	$f_{\text{fill}}(x', z)$	I love this movie. Overall, it was a bad movie.	A prompt where slot `[Z]` is filled with any answer.
Answered Prompt	$f_{\text{fill}}(x', z^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot `[Z]` is filled with a true answer.
Answer	z	"good", "fantastic", "boring"	A token, phrase, or sentence that fills `[Z]`

PART2: Basic Prompting

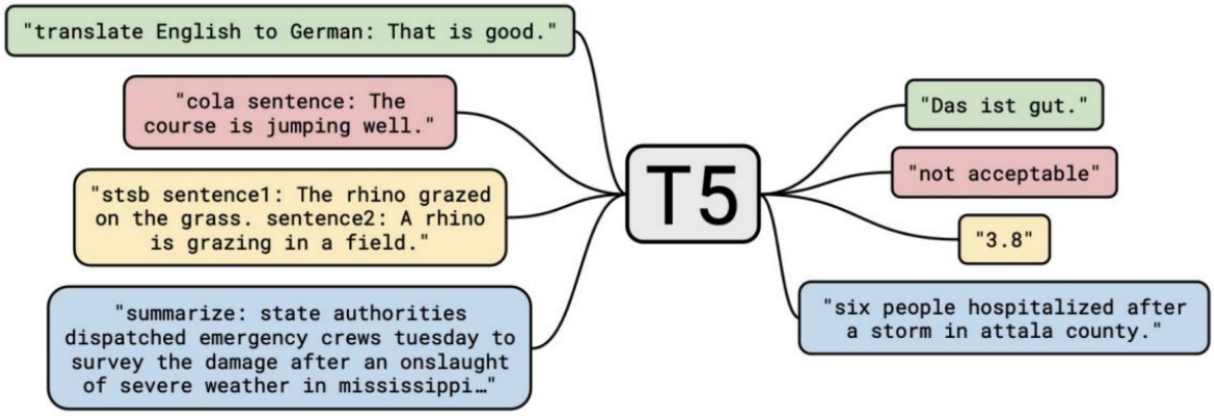
Intro



1) Zero-Shot

Zero-shot prompting is simply feeding the task text to the model and asking for results.
Depends on the choice of pre-trained model → design prompts closer to their pre-training

- Masked Language Modelling (MLM, BERT)
  store gallon
  ↑ ↑
  the man went to a [MASK] to buy a [MASK] of milk
  great → positive
  ↑
  Cloze prompt : A three-hour cinema master class. It was [MASK] !
- Text-to-Text (T5)
  Prefix prompt sentiment: A three-hour cinema master class.→ positive label



2)Few-Shot

Few-shot prompting additionally presents a set of high-quality examples with labels (demonstrations), along with the task text.

- Few-shot or few examples, also called in-context examples
- Better performance than zero-shot
- Costs more tokens → may hit token length limit

Depend on the choice of pre-trained model → design prompts closer to their pre-training

2a)Few-Shot Prompting – MLM (BERT)

Cloze prompt : An effortlessly accomplished and richly resonant work. It was great !
A mostly tired retread of several other mob tales.It was terrible !
A three-hour cinema master class.It was [MASK] !
↓
great→ positive

2b)Few-Shot Prompting – T2T (T5-like)

Prefix prompt sentiment: An effortlessly accomplished and richly resonant work.positive
sentiment: A mostly tired retread of several other mob tales. negative
sentiment: A three-hour cinema master class.
↓
positive

2c) how to maximize performance with few-shot prompting

- Prompt format
  - Syntax (e.g., length, blanks)
  - Additional information: topics, domains, etc.
  - Task prompt/Pattern: wording ○ Label information: label names, input-label mapping
- In-context examples
  - Example selection
  - Order of examples

3) Domain Prompt

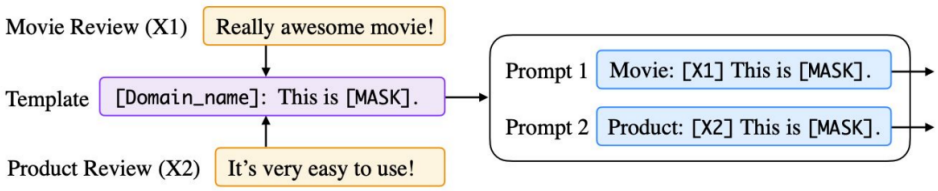


Figure 5: Multi-prompt learning for multi-task, multi-domain or multi-lingual learning. We use different colors to differentiate different components as follows. " " for input text, " " for template, " " for prompt.

Example Selection

- Choose examples that are semantically similar to the test example
- Examples with high disagreement or entropy(Examples with high disagreement or entropy)

Order of Examples

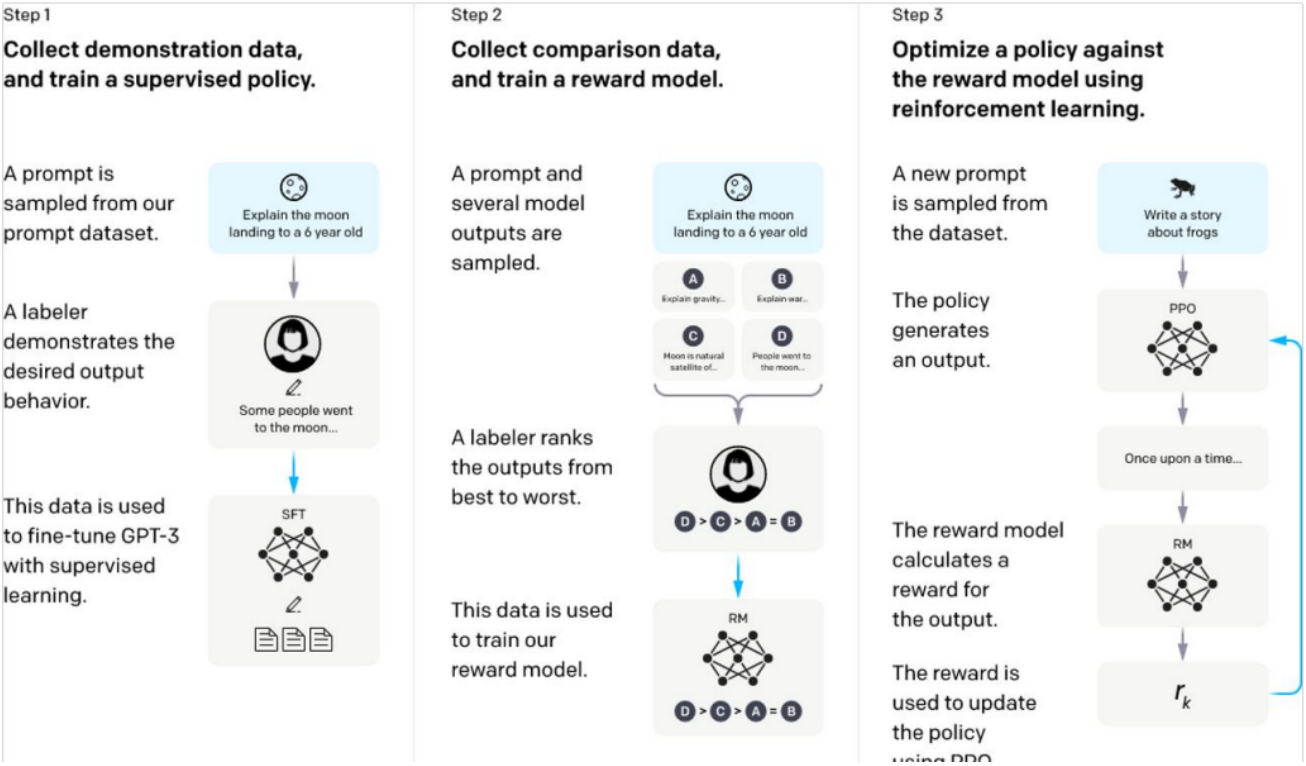
- Diverse examples
  - Relevant to the test sample
  - In random order
- Avoid order with
  - extreme unbalanced predictions
  - being overconfident

PART2: Instruction Prompting

1) Instruction Fine-Tuning

- 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.)

2) Instruct GPT



3) Zero-Shot prompting

Depend on the choice of pre-trained model → design prompts closer to their pre-training

- GPT: Generative Pre-trained Transformer

GPT-3 Complex generation and reasoning tasks via in-context learning: prompt with task description and a few demonstrations

```
Translate English to Spanish:
a black cat -> un gato negro
I am hungry -> tengo hambre
a cup of tea ->
```

task description
examples
prompt

Task description

You will be provided a short movie review, and your task is to classify its sentiment as positive, neutral, or negative.

Example

Input: An effortlessly accomplished and richly resonant work.
Output: positive
Input: A mostly tired retread of several other mob tales.
Output: negative

Test data

Input: A three-hour cinema master class.
Output:

↓

positive

Task description

Definition: Determine the speaker of the dialogue, “agent”, or “customer”.
Input: I have successfully booked your tickets.
Output: agent

Example

Definition: Determine which category the question asks for, “agent”, or “customer”.
Input: What's the oldest building in US?
Output: Location

Test data

Definition: Classify the sentiment of the given movie review, “positive”, or “negative”.
Input: A three-hour cinema master class.
Output:

↓

positive

3) Task Definition

Annotated task definitions

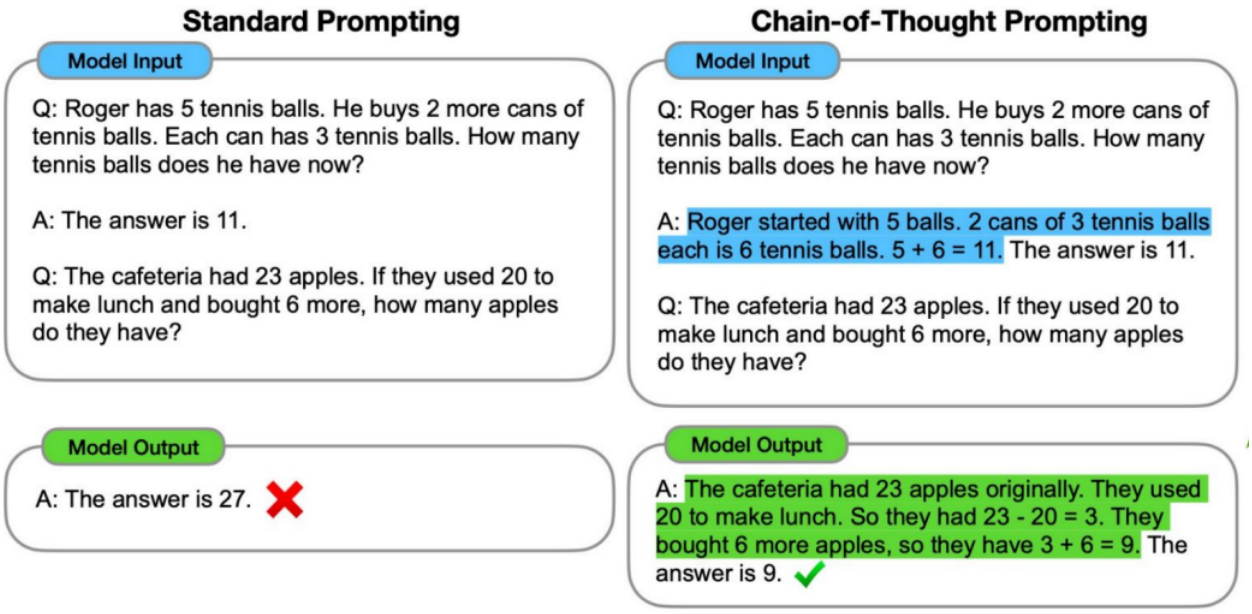
You will be given two pieces of text… One of them is simpler … You are expected to output 'Text one' if the first sentence is simpler. Otherwise output 'Text two'.

Given a sentence with a missing word, pick the answer option that best fills out the missing word in the sentence. Indicate each answer with its index ('a', 'b', 'c', 'd').

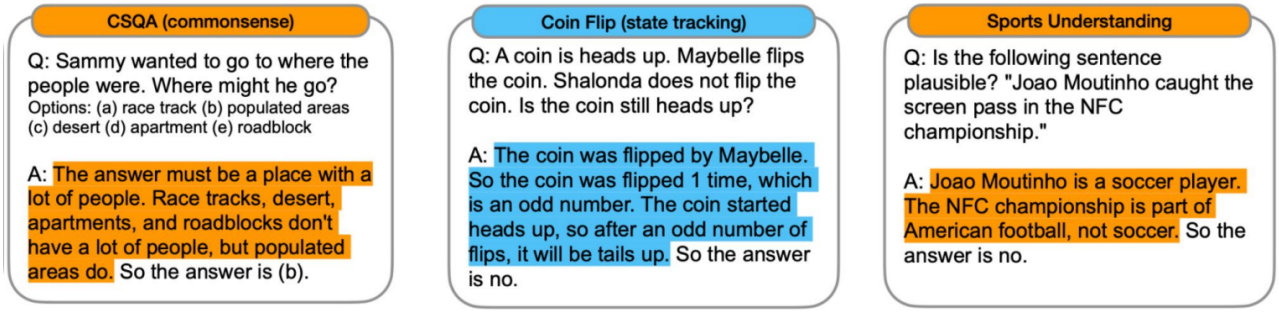
Given a document, generate a short title of the document. The title should convey the main idea/event/topic about which the document is being written.

Category	Description
Input Content	Primary description of the task input
Additional Input Content	Additional details on task input
Action Content	Action to perform for task
Input Mention	Mentions of input within action content
Output Content	Primary description of task output
Additional Output Content	Additional details on task output
Label List	Task output labels (classification only)
Label Definition	Task Label definitions (classification only)

PART3: Chain-of-Thought (CoT)



When asked to think the model give the right answer



1) Zero-shot & Few-shot CoT

Zero-shot CoT

- append to your prompt: Let's think step by step to explicitly encourage the model to first generate reasoning chains then to prompt with Therefore, the answer is
- Let's work this out it a step by step to be sure we have the right answer

Few-shot CoT

Few-shot CoT: prompt the model with a few demonstrations, each containing manually written (or model-generated) high-quality reasoning chains.

2) Clue and Reasoning Prompting (CARP)

CARP Zero-shot

This is an overall sentiment classifier for movie reviews.
First, list CLUES (i.e., keywords, phrases, contextual information, semantic relations, semantic meaning, tones, references) that support the sentiment determination of input..
Second, deduce the diagnostic REASONING process from premises (i.e., clues, input) that supports the INPUT sentiment determination (Limit the number of words to 130).
Third, based on clues, reasoning and input, determine the overall SENTIMENT of INPUT as Positive or Negative.

INPUT: press the delete key

CLUES: delete key
REASONING: The phrase "delete key" implies an action of removing something, which could be interpreted as a negative sentiment.
SENTIMENT: Negative ✓

CARP Few-shot

This is an overall sentiment classifier for movie reviews.
First, present CLUES (i.e., keywords, phrases, contextual information, semantic relations, semantic meaning, tones, references) that support the sentiment determination of input.
Second, deduce a diagnostic REASONING process from premises (i.e., clues, input) that supports the sentiment determination (Limit the number of words to 130).
Third, determine the overall SENTIMENT of INPUT as Positive or Negative considering CLUES, the REASONING process and the INPUT.

INPUT: noyce films it more as a shocking history lesson than as drama.
CLUES: - Positive: "more," "shocking," "history lesson," "drama." - Negative: None.
REASONING: The input contains positive words such as "more," "shocking," and "drama," which suggest that the sentiment is positive. Additionally, the phrase "history lesson" implies that the film is educational and informative, which is generally seen as a positive thing. Therefore, the sentiment of the input is positive.
SENTIMENT: Positive

INPUT: press the delete key
CLUES: - Positive: None. - Negative: "delete key."
REASONING: The input contains the phrase "delete key," which implies that something is being removed or erased. This is generally seen as a negative action, suggesting that the sentiment of the input is negative.
SENTIMENT: Negative ✓

PART 4: Automatic Prompting

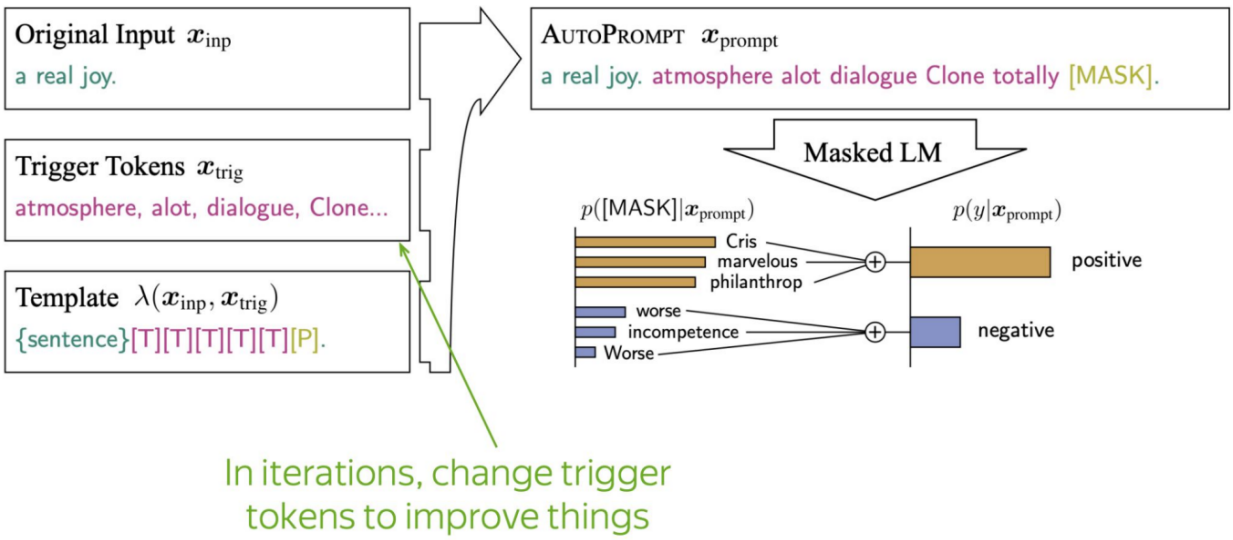
1)Prompt-Tuning

Two types:

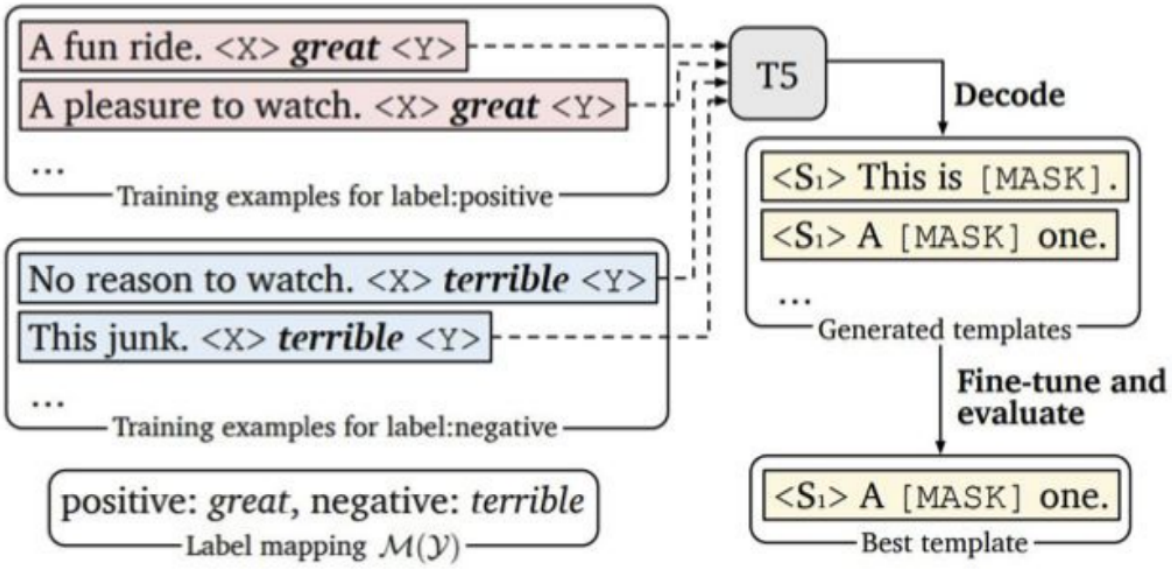
1. Discrete prompt
2. Continuous/Soft prompt: introduce novel embeddings that are learned using gradient descent

Discrete Prompt

a)AutoPrompt



b)LM-BFF

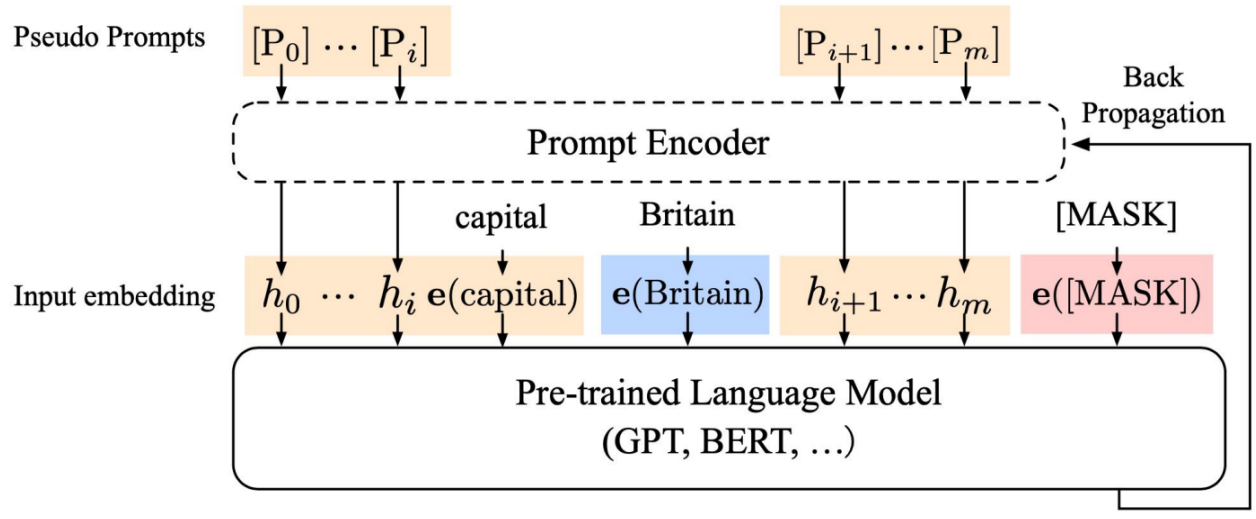


Soft Prompt

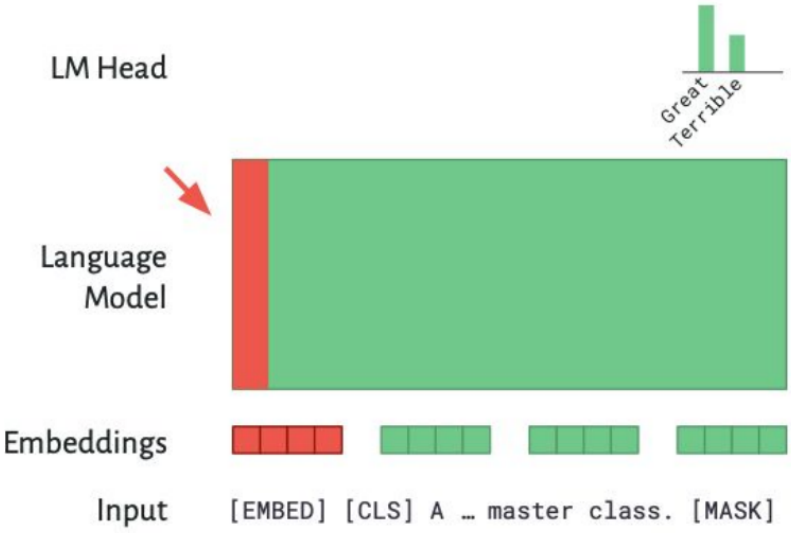
a) LM-BFF

Learning embeddings for placeholder tokens in the pattern

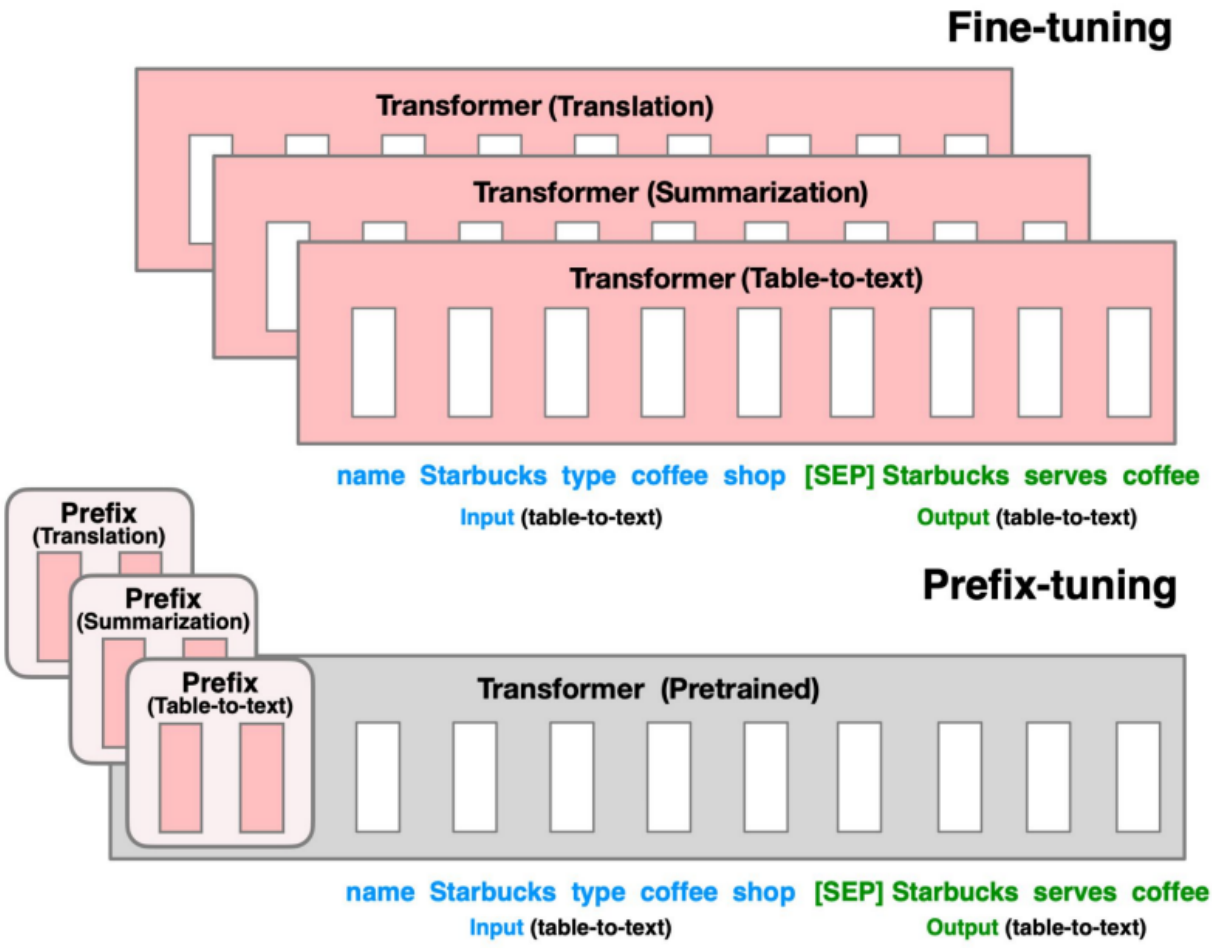
P-Tuning



Learning embeddings for placeholder tokens in the pattern.



Prefix Tuning



prefix-tuning, a lightweight alternative to fine-tuning for natural language generation tasks, which keeps language model parameters frozen, but optimizes a small continuous task-specific vector (called the prefix).

PART 5:Practise Tips

- Issues with Prompt Engineering: Issues with Prompt Engineering(Some prompt techniques work on one LLM, but don't work on other LLMs)
- Prompt drift
  - Different versions of the same model produce different outputs for the same input prompt
  - Pose challenges for reproducibility and consistency
- Prompts are brittle
  - Slight changes in the prompt can change output a lot

  E.g., prompt syntax (e.g., length, blanks, ordering of examples) and semantics (e.g., wording, selection of examples, instructions)

AutoPrompt: The Best Prompts are Gibberish!

Task	Prompt Template	Prompt found by AUTOPROMPT	Label Tokens
Sentiment Analysis	{sentence} [T]...[T] [P].	unflinchingly bleak and desperate Writing academicswhere overseas will appear [MASK].	**pos**: partnership, extraordinary, ##bla **neg**: worse, persisted, unconstitutional
NLI	{prem}[P][T]...[T]{hyp}	Two dogs are wrestling and hugging [MASK] concretepathic workplace There is no dog wrestling and hugging	**con**: Nobody, nobody, nor **ent**: ##found, ##ways, Agency **neu**: ##ponents, ##lary, ##uated
Fact Retrieval	*X plays Y music* {sub}[T]...[T][P].	Hall Overton fireplacemade antique son alto [MASK].	
Relation Extraction	*X is a Y by profession* {sent}{sub}[T]...[T][P].	Leonard Wood (born February 4, 1942) is a former Canadian politician. Leonard Wood gymnasium brotherdicative himself another [MASK].	

Table 3: **Example Prompts** by AUTOPROMPT for each task. On the left, we show the prompt template, which combines the input, a number of trigger tokens [T], and a prediction token [P]. For classification tasks (sentiment analysis and NLI), we make predictions by summing the model's probability for a number of automatically selected label tokens. For fact retrieval and relation extraction, we take the most likely token predicted by the model.

	Human-written prompt	AutoPrompt
Math	Return the sum of the inputs	¿:Returns Adding togetherFont accomplish
	Return the square of the input	Cal impl qApplySquare fiat
	Differentiate between prime/non-prime integers	ropheospels&& Norestricted
ANLI	Differentiate vegetarian/non-vegetarian foods	compliedthe whether methamphetamine provided comp
	Differentiate the subject in a sentence based on gender	¡ endoftext ¿ -¿ M Fundamental FG Fav
	Return a synonym	Word termOn English meanings
	Translate english to spanish	the ththebb volunt
	Return a country's capital city	Ang Suppose AUTHthe beh Assassins
Sentiment	What is the sentiment expressed by the reviewer for the movie?	Pap Azerb Saiyan Forean Talatar Yemeni IndBloomberg re-ceiveda
	How does the author of the news headline feel?	Fur resultolandgroundur augmented=

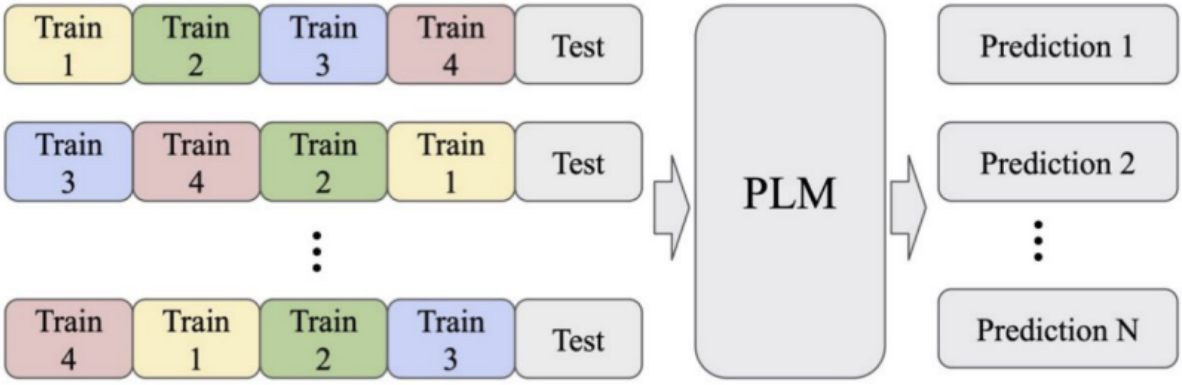
Prompt format matters

Prompt	Label Names
Review: This movie is amazing! Answer: Positive Review: Horrific movie, don't see it. Answer:	Positive, Negative
Review: This movie is amazing! Answer: good Review: Horrific movie, don't see it. Answer:	good, bad
My review for last night's film: This movie is amazing! The critics agreed that this movie was good My review for last night's film: Horrific movie, don't see it. The critics agreed that this movie was	good, bad
Here is what our critics think for this month's films. One of our critics wrote "This movie is amazing!". Her sentiment towards the film was positive. One of our critics wrote "Horrific movie, don't see it". Her sentiment towards the film was	positive, negative
Critical reception [edit] In a contemporary review, Roger Ebert wrote "This movie is amazing!". Entertainment Weekly agreed, and the overall critical reception of the film was good. In a contemporary review, Roger Ebert wrote "Horrific movie, don't see it". Entertainment Weekly agreed, and the overall critical reception of the film was	good, bad
Review: This movie is amazing! Positive Review? Yes Review: Horrific movie, don't see it. Positive Review?	Yes, No

Order of Examples matter

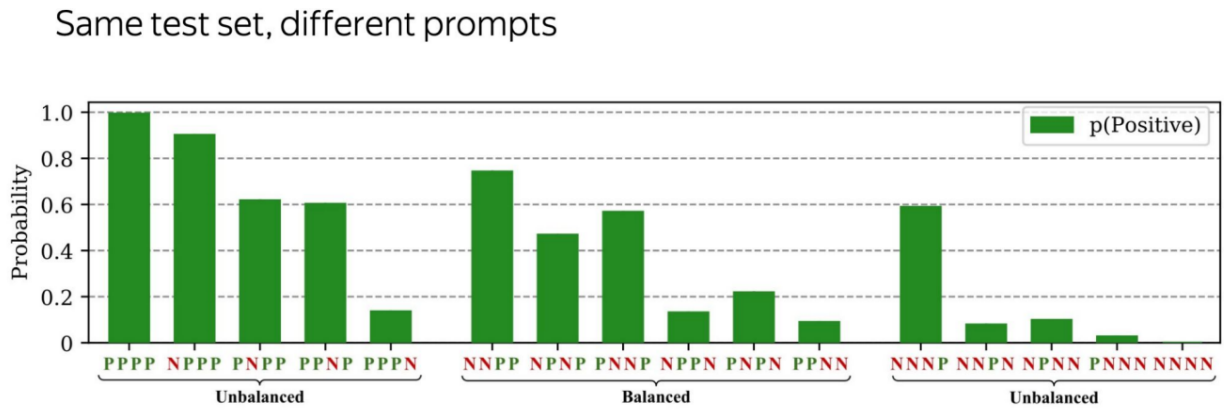
Depending on the example order in the prompt, we can get

- Near state-of-the-art accuracy
- Near random accuracy

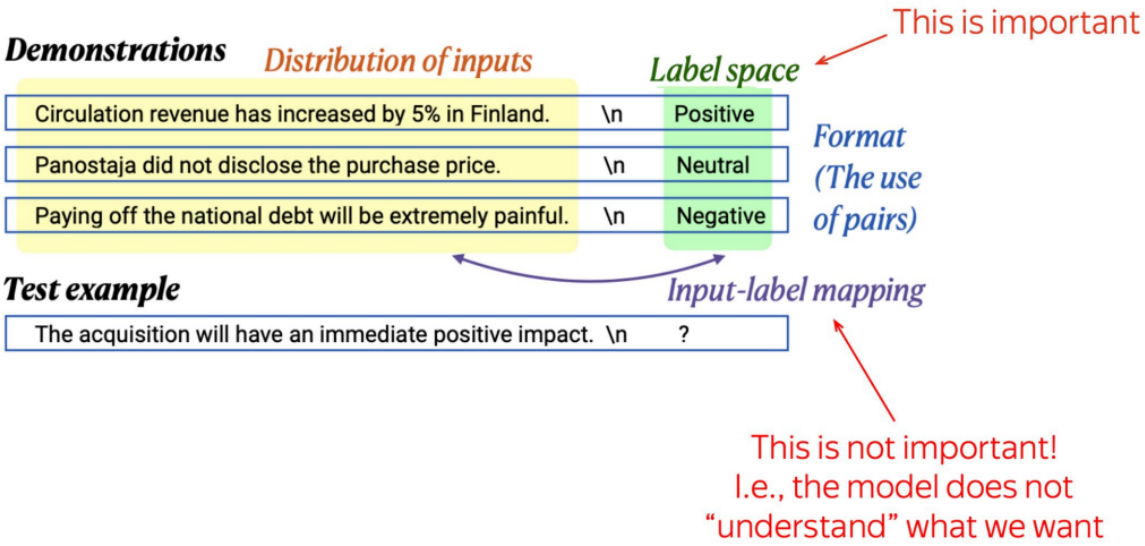


Order of the Examples: Majority and Recency Biases

Depending on labels and their order, we get very different average predicted probability of the positive class



True Labels Do Not Matter



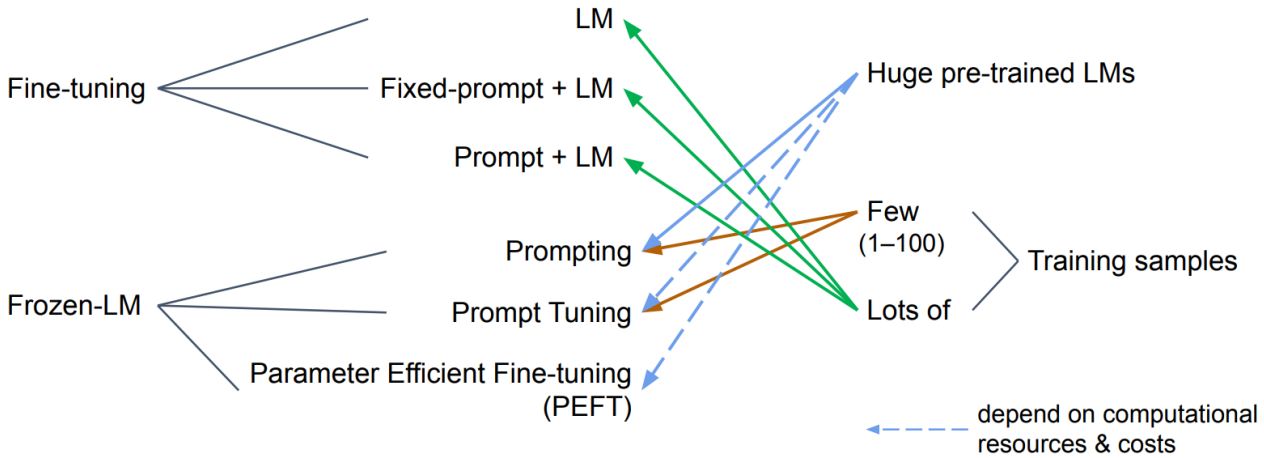
Token Costs & Measurement

- Prompt = list of tokens
  - Compute the # of tokens
- Cost = (# input tokens + # expected output tokens) * cost per token
- Consider output cost
  - Chain-of-Thought is often better than basic prompts, but
  - requires at least 3 times as many output tokens as basic prompts

Trick (by Haritz Puerto) for text classification

- The output is usually 1 token → set max new tokens to 1
- If the names of classes or labels are longer ~ multiple tokens → set max new tokens to max # tokens of the names

Prompting vs Fine-tuning



PART 6: More Advanced Prompting

You can read this part but I don't think it's necessary to memorize all

1)Self-Consistency

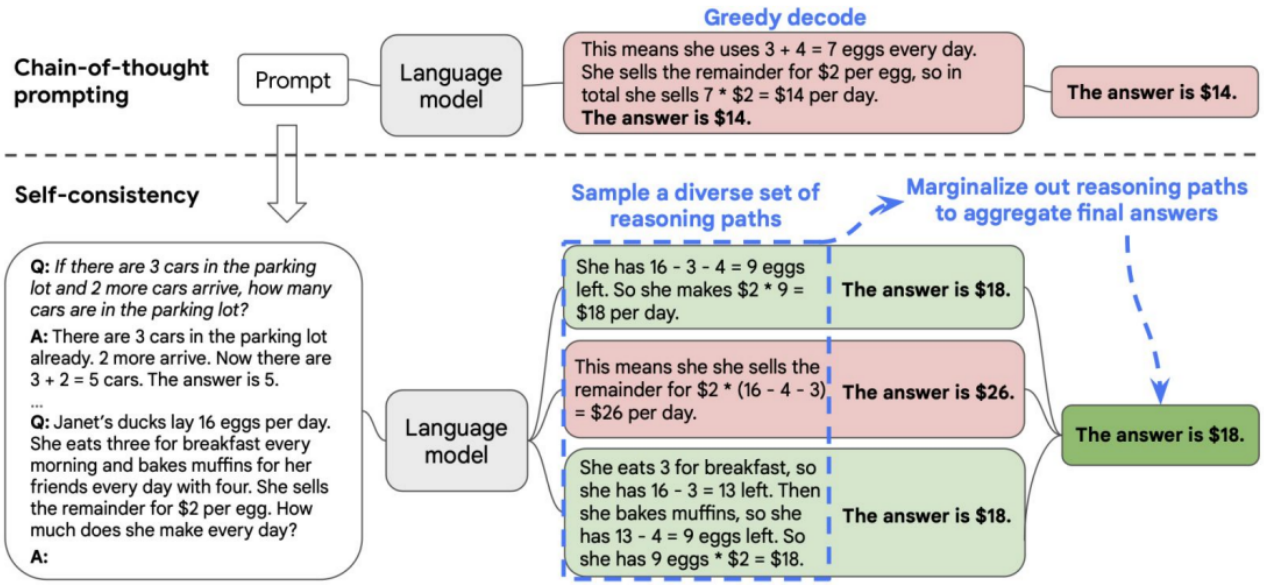
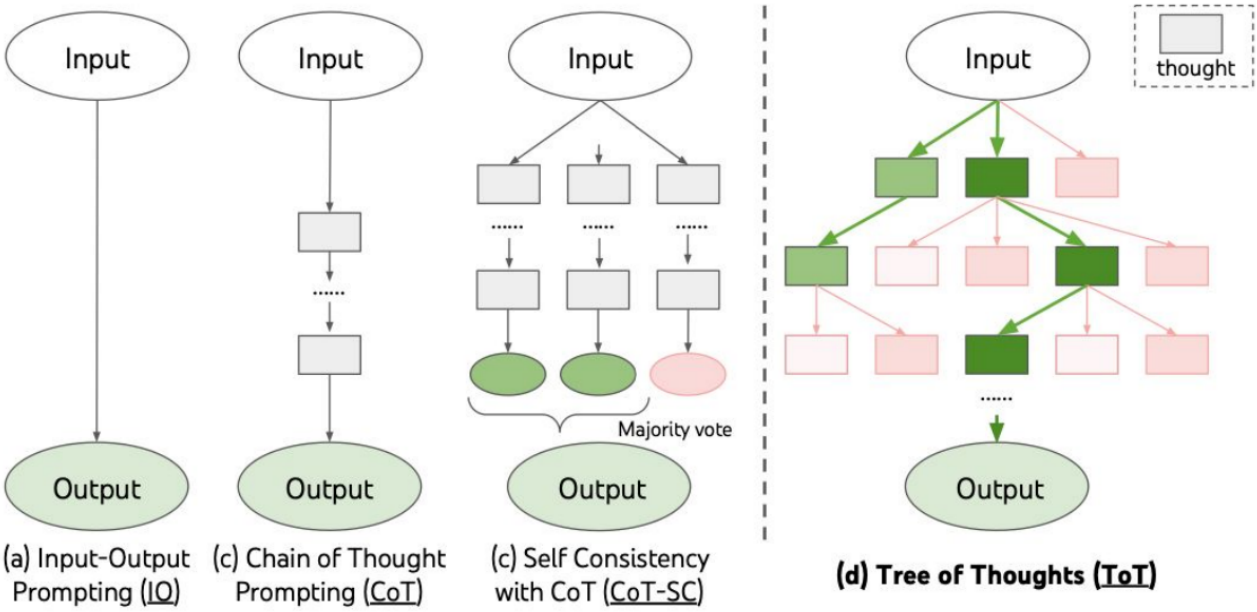


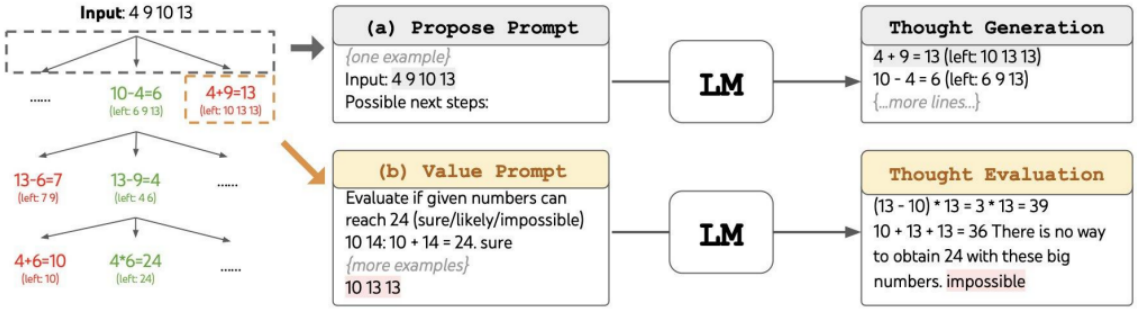
Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the "greedy decode" in CoT prompting by sampling from the language model's decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

2)Tree of Thought (ToT)

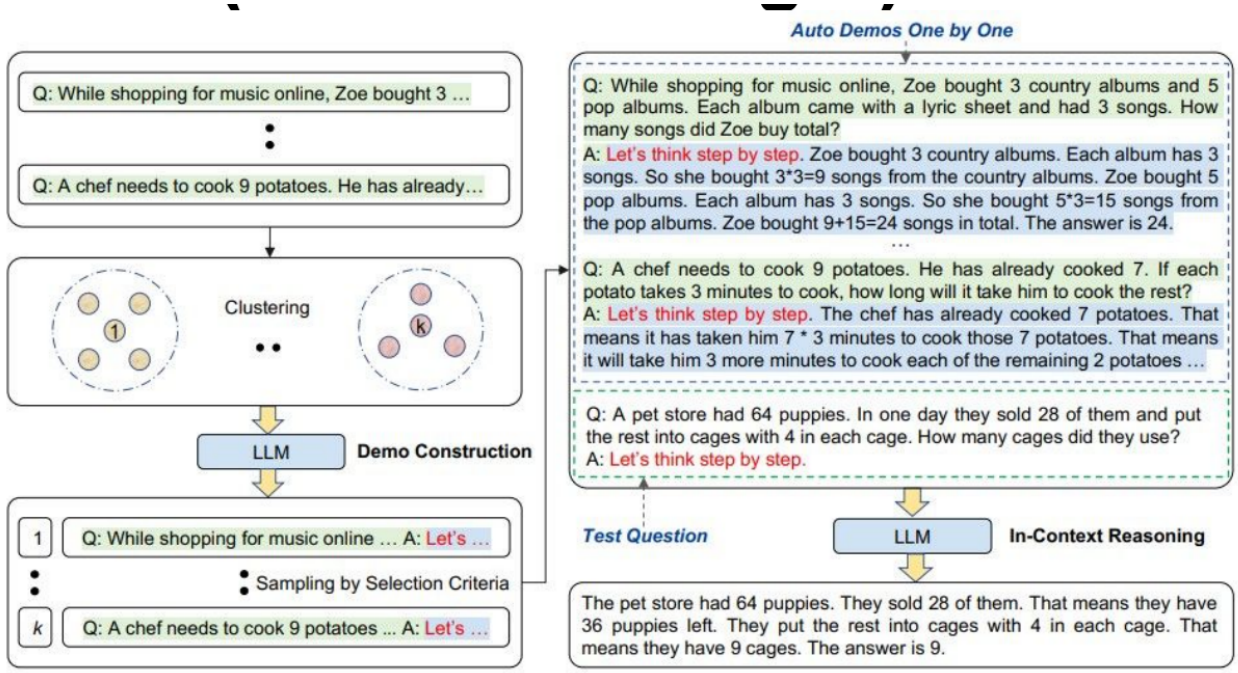


	Game of 24	Creative Writing	5x5 Crosswords
Input	4 numbers (4 9 10 13)	4 random sentences	10 clues (h1. presented;..)
Output	An equation to reach 24 (13-9)*(10-4)=24	A passage of 4 paragraphs ending in the 4 sentences	5x5 letters: SHOWN; WIRRA; AVAIL; ...
Thoughts	3 intermediate equations (13-9=4 (left 4,4,10); 10-4=6 (left 4,6); 4*6=24)	A short writing plan (1. Introduce a book that connects...)	Words to fill in for clues: (h1. shown; v5. naled; ...)
#ToT steps	3	1	5-10 (variable)

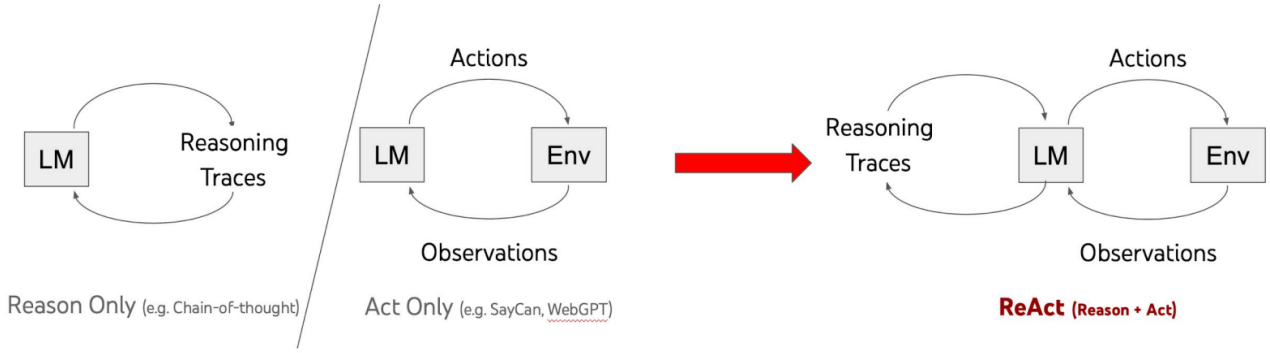
Table 1: Task overview. Input, output, thought examples are in blue.



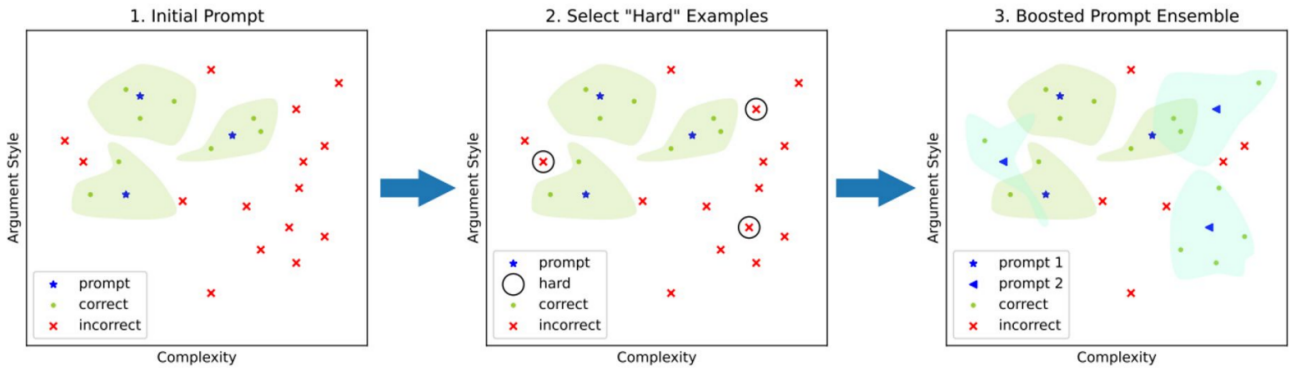
3) Auto CoT



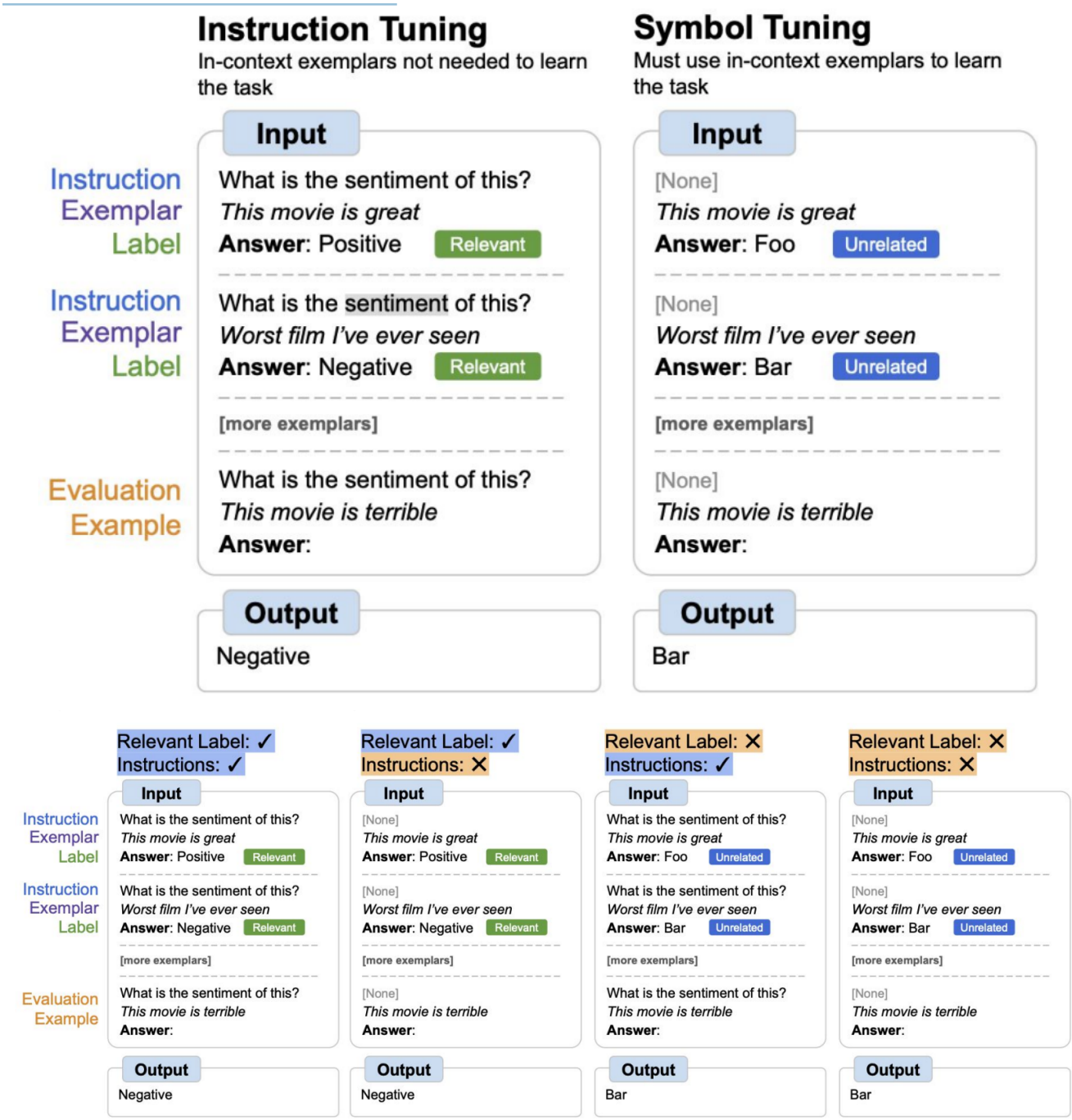
4) Reason & Act



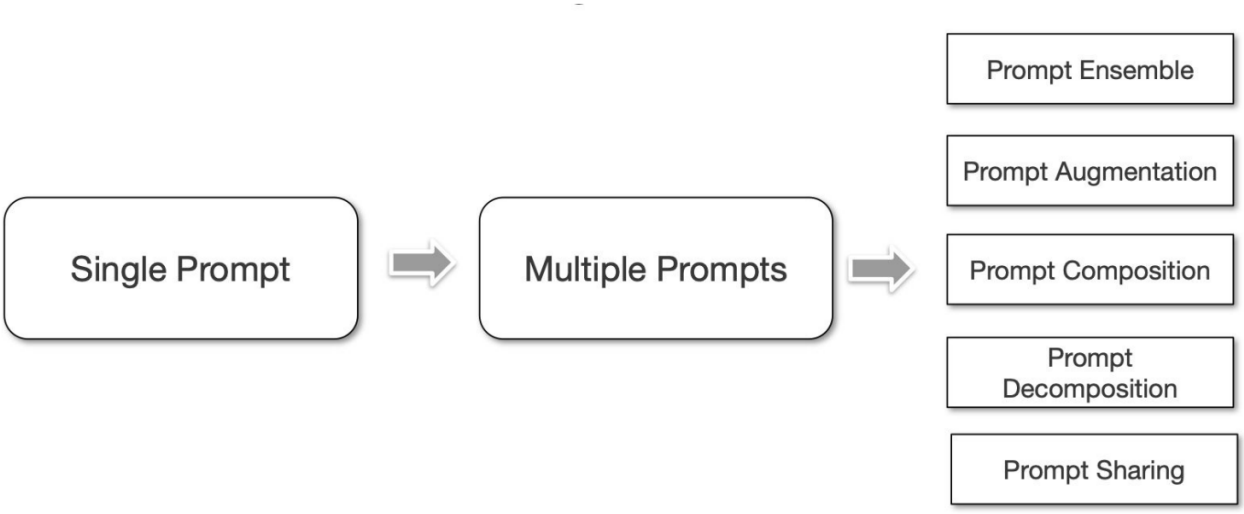
5)Boosted Prompting



6)Symbol Tuning

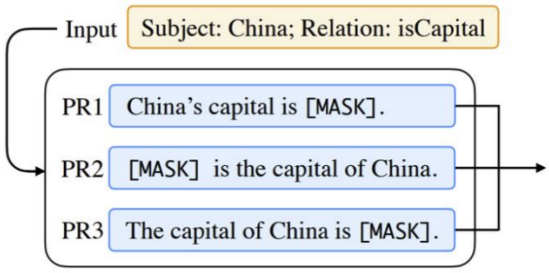


PART 7: Multi-Prompt Learning

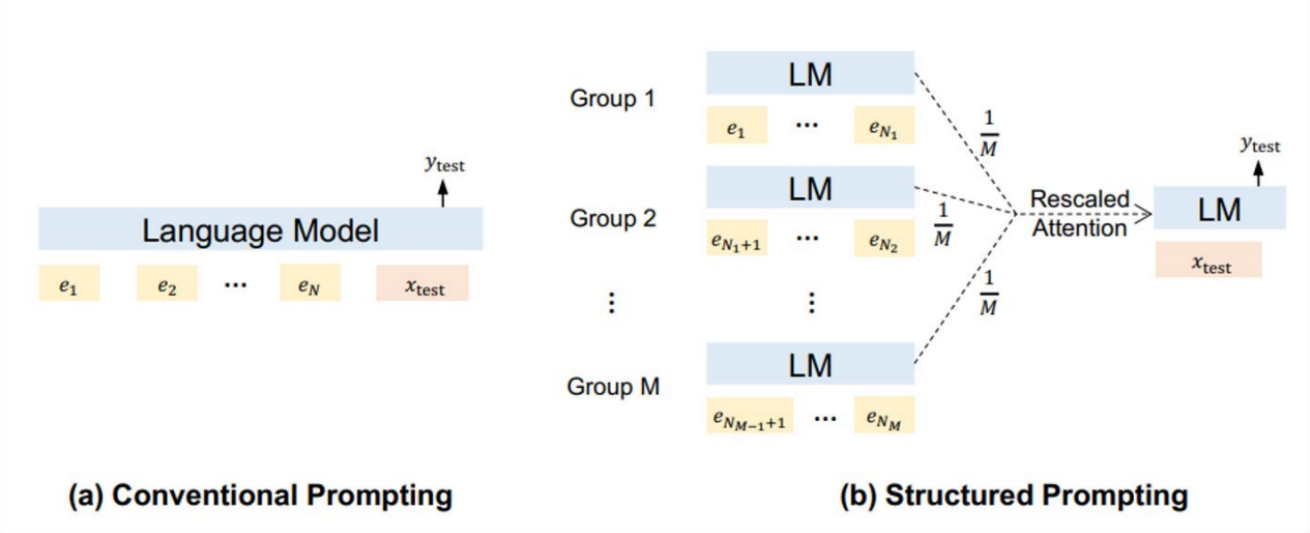


1)Prompt Ensembling

- Using multiple prompts for an input to make predictions
- Advantages:
  - Complementary
  - (Partially) Remove cost of prompt engineering
  - Stabilize performance
- Ensembling methods
  - Uniform averaging
  - Weighted averaging
  - Majority voting Pooling



2)Structured Prompting



3)Generated knowledge Prompting

Step 1: generate knowledge

Prompt:

```
Input: Greece is larger than mexico.
Knowledge: Greece is approximately 131,957 sq km, while Mexico is approximately 1,964,375 sq km,
making Mexico 1,389% larger than Greece.
Input: Glasses always fog up.
Knowledge: Condensation occurs on eyeglass lenses when water vapor from your sweat, breath, and
ambient humidity lands on a cold surface, cools, and then changes into tiny drops of liquid,
forming a film that you see as fog. Your lenses will be relatively cool compared to your breath,
```

⋮

```
Input: A rock is the same size as a pebble.
Knowledge: A pebble is a clast of rock with a particle size of 4 to 64 millimetres based on the
Udden–Wentworth scale of sedimentology. Pebbles are generally considered larger than granules (2 to
4 millimetres diameter) and smaller than cobbles (64 to 256 millimetres diameter).
Input: Part of golf is trying to get a higher point total than others.
Knowledge:
```

Step2: Integrate knowledge

Prompt:

```
Question: Part of golf is trying to get a higher point total than others. Yes or No?
Knowledge: The objective of golf is to play a set of holes in the least number of strokes. A round
of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf
course. Each stroke is counted as one point, and the total number of strokes is used to determine
the winner of the game.
Explain and Answer:
```

Answer 1 (confidence very high):

```
No, the objective of golf is not to get a higher point total than others. Rather, the objective is
to play a set of holes in the least number of strokes. The total number of strokes is used to
determine the winner of the game, not the total number of points.
```

```
Question: Part of golf is trying to get a higher point total than others. Yes or No?
Knowledge: Golf is a precision club-and-ball sport in which competing players (or golfers) use many
types of clubs to hit balls into a series of holes on a course using the fewest number of strokes.
The goal is to complete the course with the lowest score, which is calculated by adding up the
total number of strokes taken on each hole. The player with the lowest score wins the game.
Explain and Answer:
```

Answer 2 (confidence is a lot lower):

```
Yes, part of golf is trying to get a higher point total than others. Each player tries to complete
the course with the lowest score, which is calculated by adding up the total number of strokes
taken on each hole. The player with the lowest score wins the game.
```

Generated Knowledge Prompting is very influencial