COMP 554 / CSDS 553 Advanced NLP

Faizad Ullah

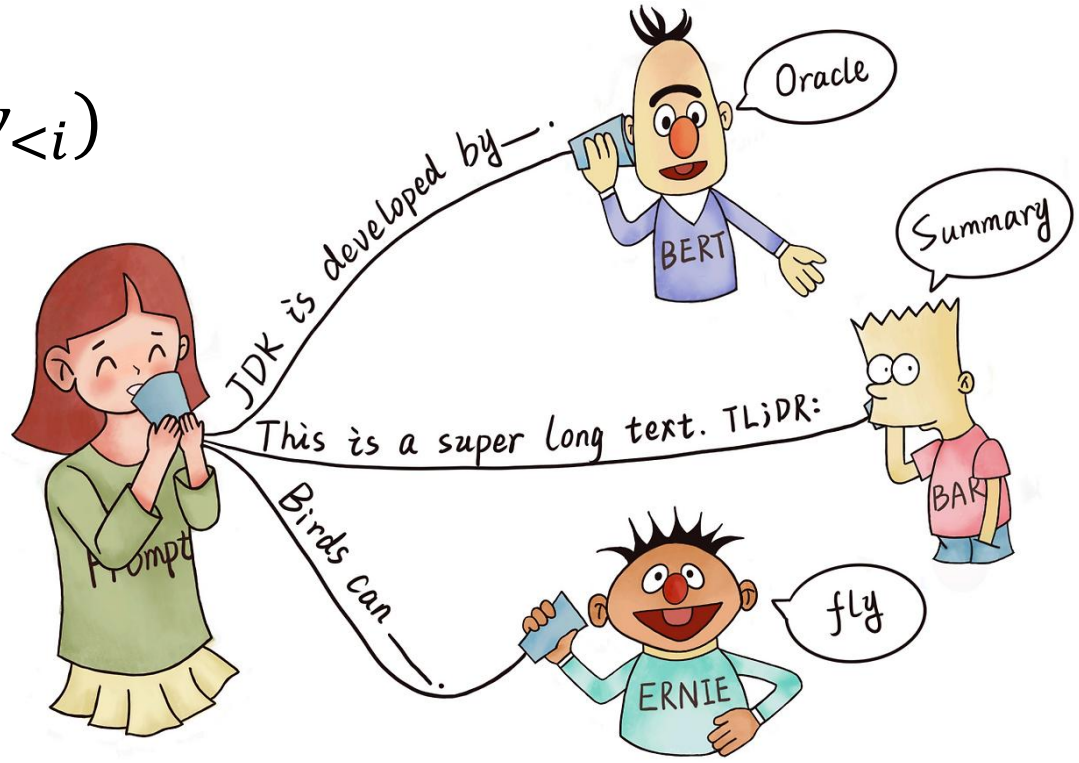
Prompting

Prompting

- A prompt is a text string that a user issues to a language model to get the model to do something useful.
- Prompting relies on contextual generation.

$$P(w_i \mid w_{<i})$$

- Prompt can be:
  - Question
  - Structured Question
  - Instruction
- Demonstration



Prompting

Sample Hotel Review

Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relax.

A prompt consisting of a review plus an incomplete statement

Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relax. In short, our stay was

Prompting

Hotel Review Completions

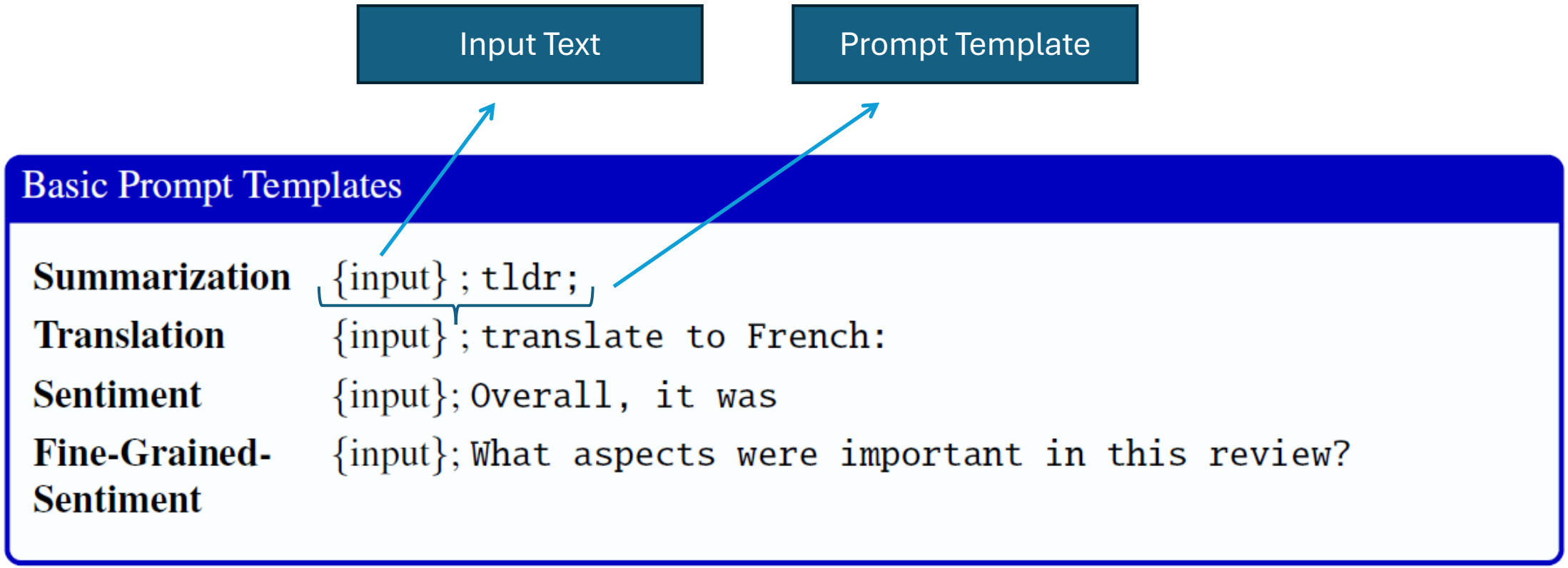
Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relaxed. In short our stay was

---

... not a pleasant one. The staff at the front desk were not welcoming or friendly, and seemed disinterested in providing good customer service.

... uncomfortable and not worth the price we paid. We will not be returning to this hotel.

Basic Prompt Templates



Basic Prompt Templates

A prompt consisting of a review plus an incomplete statement

Human: Do you think that “input” has negative or positive sentiment?
Choices:
(P) Positive
(N) Negative

Assistant: I believe the best answer is: (

LLM Outputs for Basic Prompts

Original Review ($INPUT)	Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relax and away from the city life.
Sentiment	**Prompt**: $INPUT + In short, our stay was **Output**: not enjoyable
Fine-grained Sentiment	**Prompt**: $INPUT + These aspects were important to the reviewer: **Output**: 1. Poor service 2. Unpleasant location 3. Noisy and busy area
Summarization	**Prompt**: $INPUT + tl;dr **Output**: I had a bad experience with the hotel's service and the location was loud and busy.
Translation	**Prompt**: $INPUT + Translate this to French **Output**: Je n'ai pas aimé le service qui m'a été offert lorsque je suis entré dans l'hôtel. Je n'ai également pas aimé la zone dans laquelle se trouvait l'hôtel. Trop de bruit et d'événements pour que je me sente détendu et loin de la vie citadine.

Learning from Demonstrations

Adding too many examples seems to cause the model to overfit to details of the exact examples chosen and generalize poorly.

Definition: This task is about writing a correct answer for the reading comprehension task. Based on the information provided in a given passage, you should identify the shortest continuous text span from the passage that serves as an answer to the given question. Avoid answers that are incorrect or provides incomplete justification for the question.

Passage: Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Examples:
Q: In what city and state did Beyoncé grow up?
A: Houston, Texas

Q: What areas did Beyoncé compete in when she was growing up?
A: singing and dancing

Q: When did Beyoncé release Dangerously in Love?
A: 2003

Q: When did Beyoncé start becoming popular?
A:

Including some labeled examples in the prompt.

Few-shot prompting

Zero-shot prompting

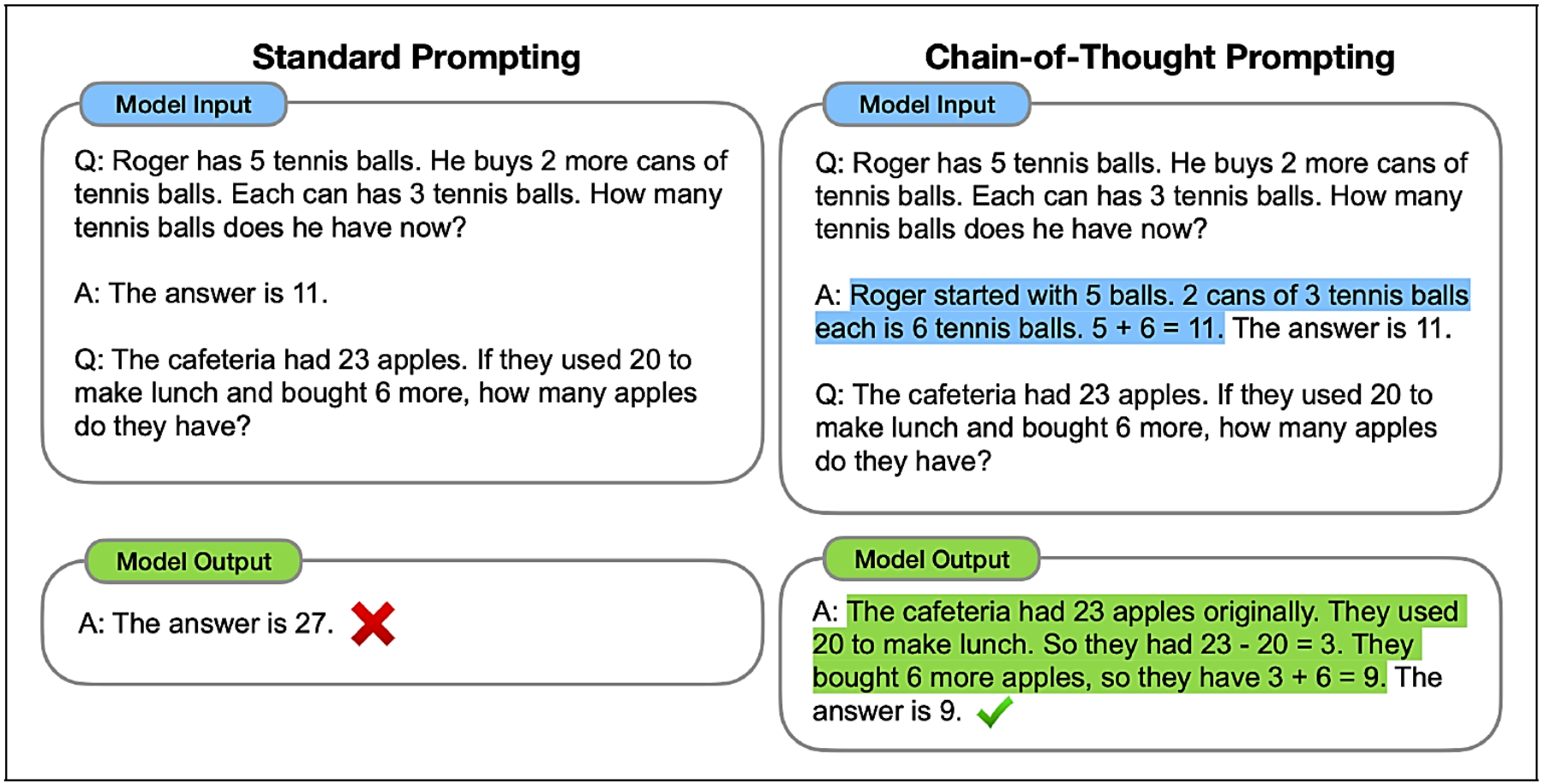
Natural Instructions dataset (Mishra et al., 2022).

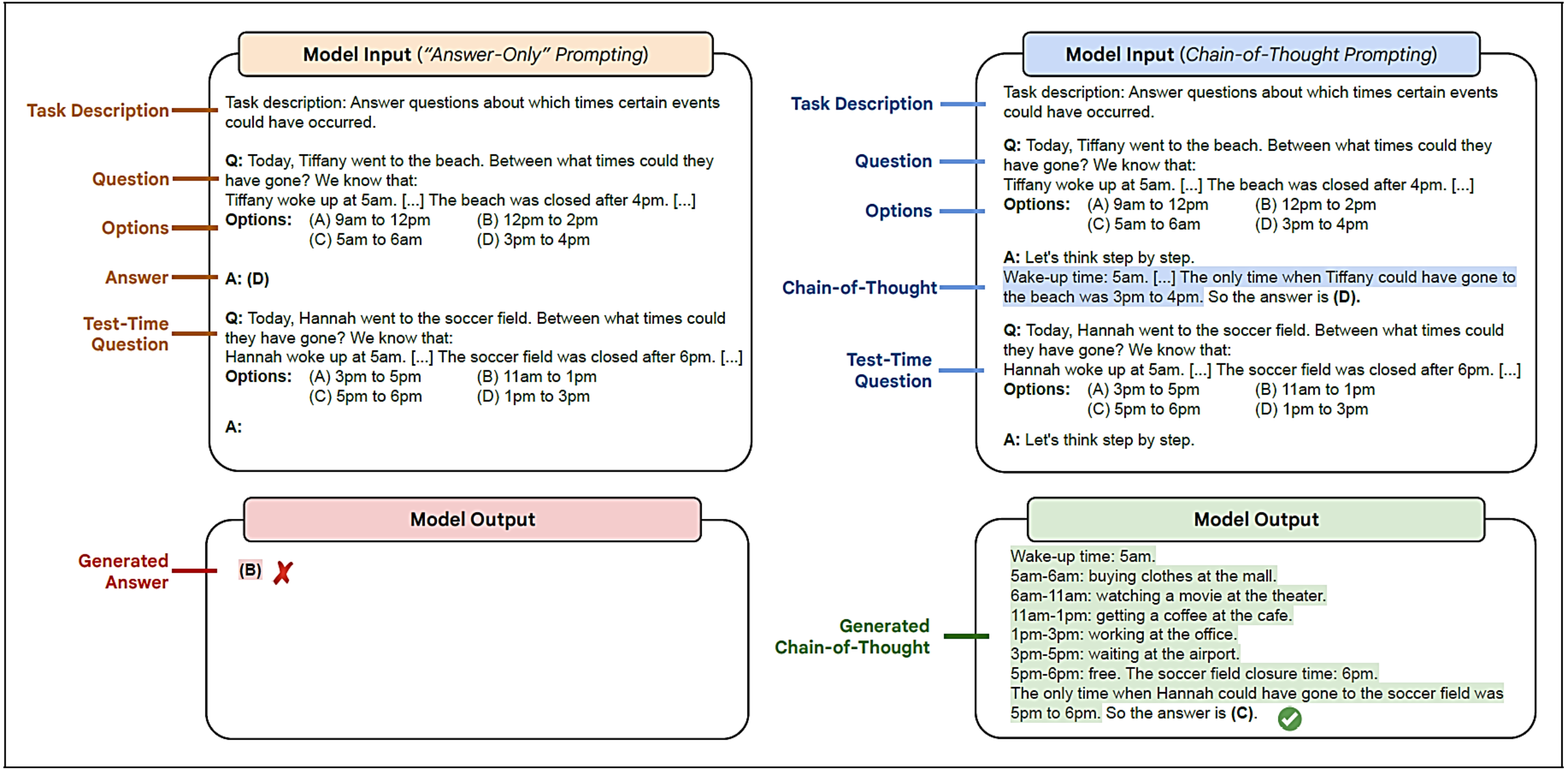
Chain-of-Thought Prompting

- Solve these tasks by breaking them down into steps.

- Language models more likely give the correct answers to difficult reasoning tasks by dividing them into steps (Wei et al., 2022; Suzgun et al., 2023).

Chain-of-Thought Prompting

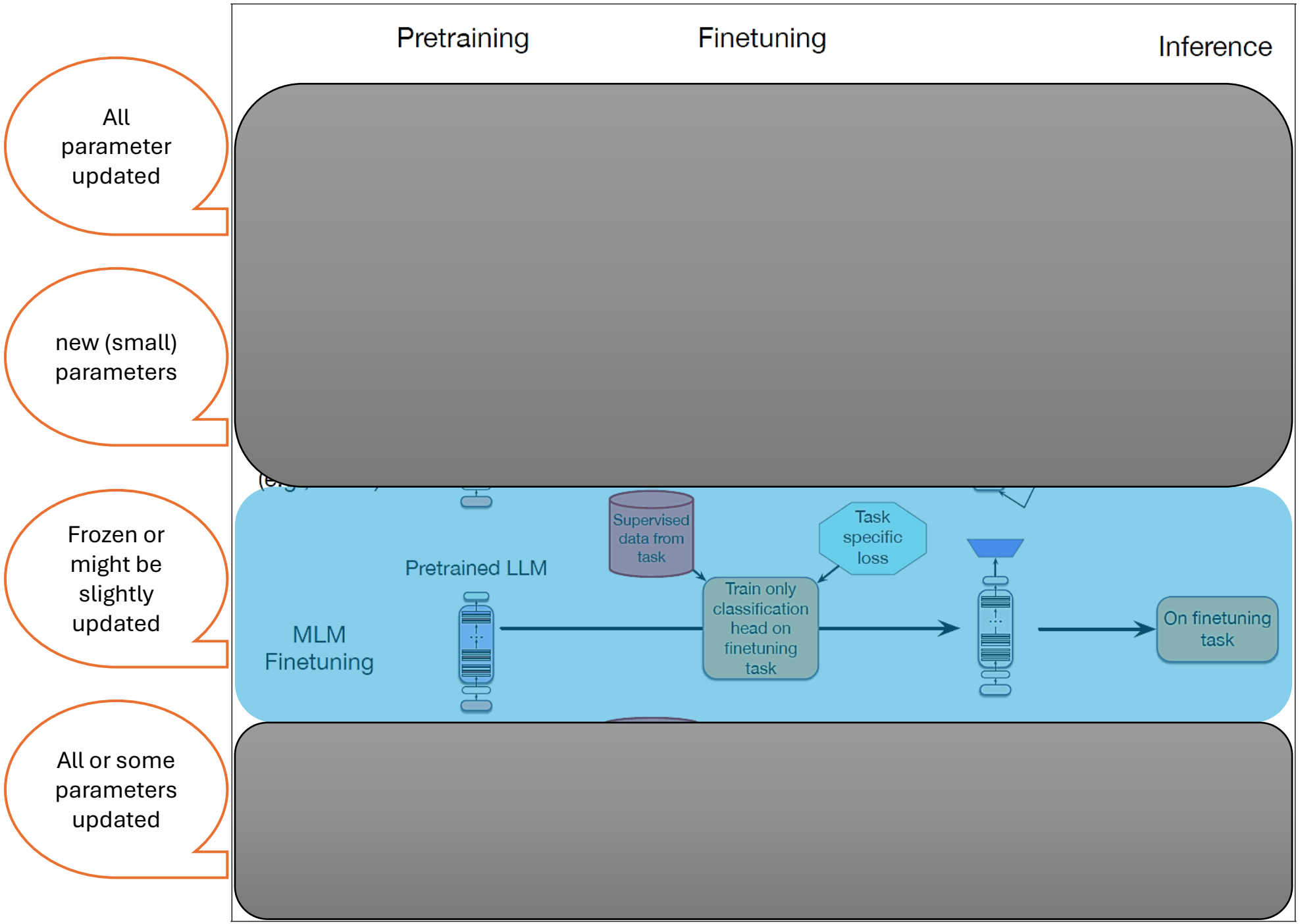



Model Input ("Answer-Only" Prompting)
Task Description
Task description: Answer questions about which times certain events could have occurred.
Question
Q: Today, Tiffany went to the beach. Between what times could they have gone? We know that:
Tiffany woke up at 5am. [...] The beach was closed after 4pm. [...]
Options
Options: (A) 9am to 12pm (B) 12pm to 2pm
(C) 5am to 6am (D) 3pm to 4pm
Answer
A: (D)
Test-Time Question
Q: Today, Hannah went to the soccer field. Between what times could they have gone? We know that:
Hannah woke up at 5am. [...] The soccer field was closed after 6pm. [...]
Options: (A) 3pm to 5pm (B) 11am to 1pm
(C) 5pm to 6pm (D) 1pm to 3pm
A:
Model Output
Generated Answer
(B)
Model Input (Chain-of-Thought Prompting)
Task Description
Task description: Answer questions about which times certain events could have occurred.
Question
Q: Today, Tiffany went to the beach. Between what times could they have gone? We know that:
Tiffany woke up at 5am. [...] The beach was closed after 4pm. [...]
Options
Options: (A) 9am to 12pm (B) 12pm to 2pm
(C) 5am to 6am (D) 3pm to 4pm
Chain-of-Thought
A: Let's think step by step.
Wake-up time: 5am. [...] The only time when Tiffany could have gone to the beach was 3pm to 4pm. So the answer is (D).
Test-Time Question
Q: Today, Hannah went to the soccer field. Between what times could they have gone? We know that:
Hannah woke up at 5am. [...] The soccer field was closed after 6pm. [...]
Options: (A) 3pm to 5pm (B) 11am to 1pm
(C) 5pm to 6pm (D) 1pm to 3pm
A: Let's think step by step.
Model Output
Generated Chain-of-Thought
Wake-up time: 5am.
5am-6am: buying clothes at the mall.
6am-11am: watching a movie at the theater.
11am-1pm: getting a coffee at the cafe.
1pm-3pm: working at the office.
3pm-5pm: waiting at the airport.
5pm-6pm: free. The soccer field closure time: 6pm.
The only time when Hannah could have gone to the soccer field was 5pm to 6pm. So the answer is (C).

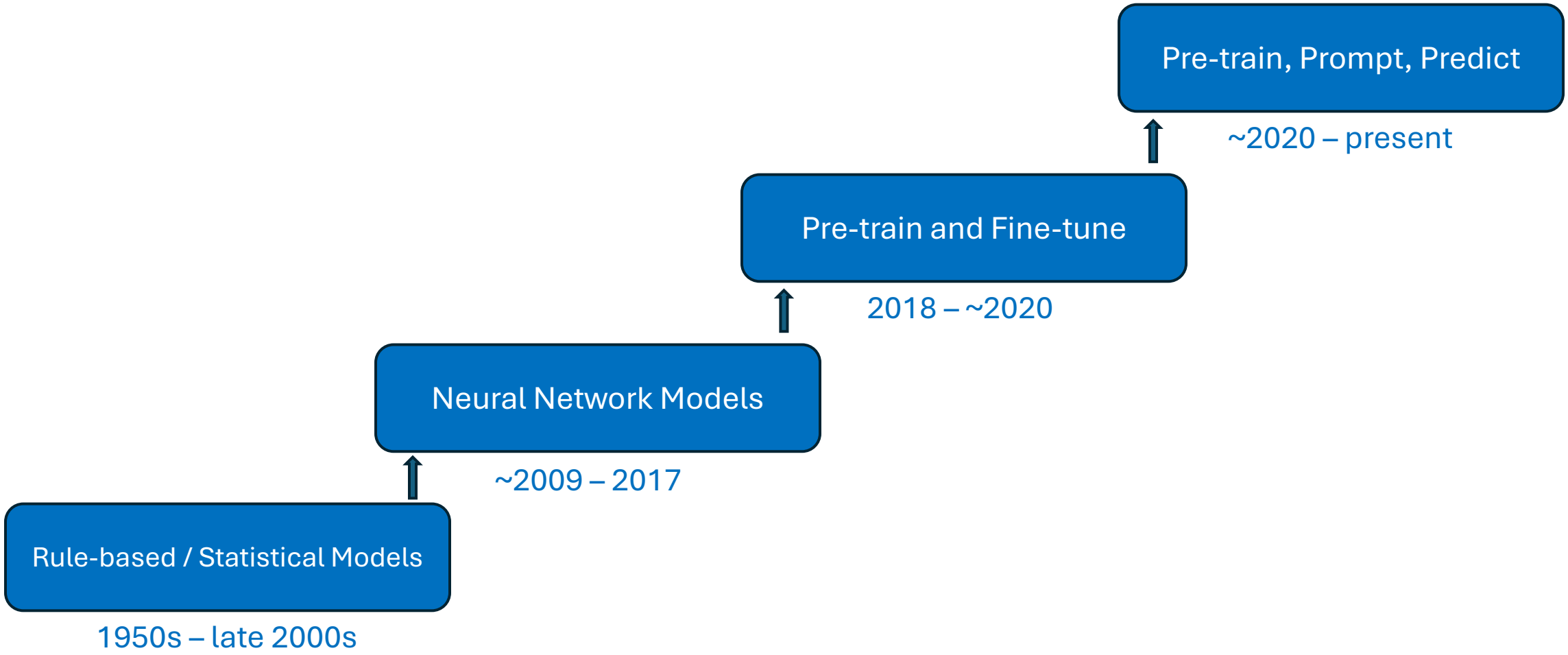
In-Context Learning

- The term in-context learning was first proposed by Brown et al. (2020), GPT3 system.

- Generally, reduce the LLM loss at inference-time, without any gradient-based updates to the model's parameters.


Pretraining
Finetuning
Inference
All parameter updated
new (small) parameters
Frozen or might be slightly updated
All or some parameters updated
Supervised data from task
Task specific loss
Pretrained LLM
MLM Finetuning
Train only classification head on finetuning task
On finetuning task

Paradigm Shifts in NLP



Vanilla / Standard Fine-Tuning

Standard Fine-Tuning (SFT)

- Fine-tuning means, the process of taking a pretrained model and further adapting some or all of its parameters to some new data e.g.,
- Input text x and predicts a label $y \in \mathcal{Y}$
- Traditional Supervised Learning:

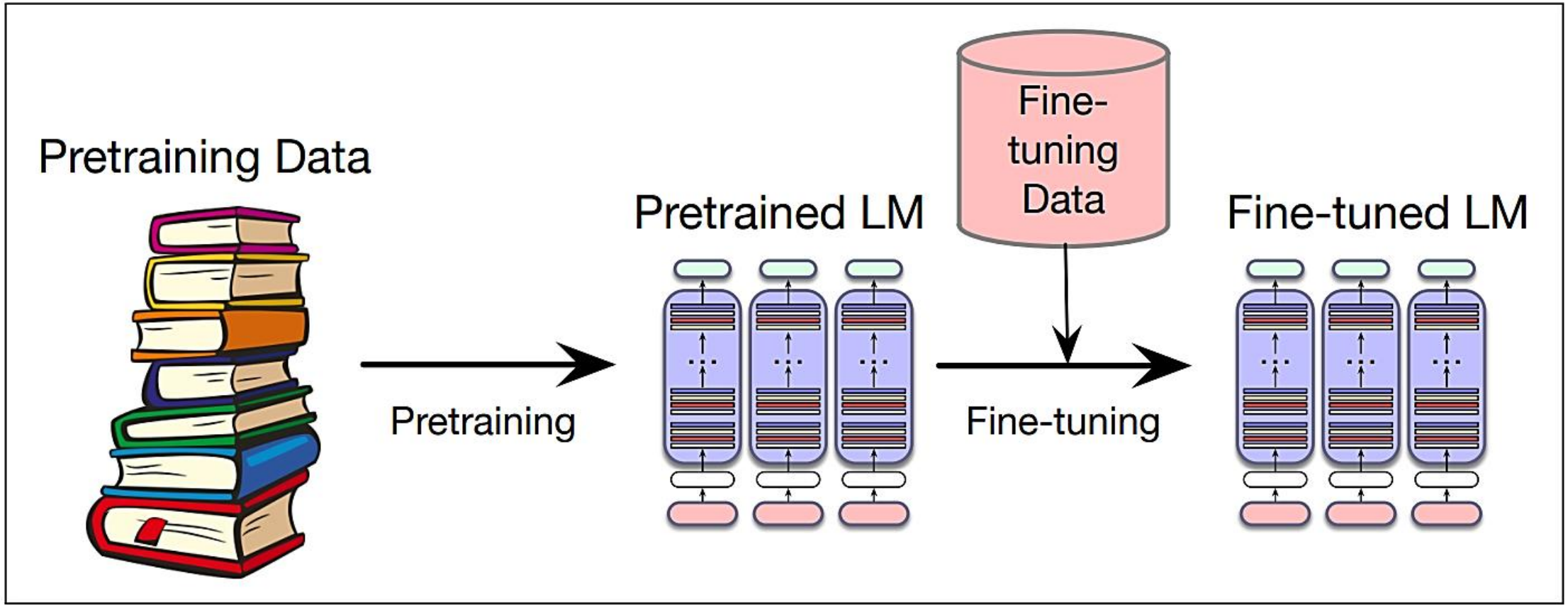
$$P(y|x;\ \theta)$$

- For sentiment analysis (Pang et al., 2002) may take an input:

x = I love this place

- Goal is to predict a label $y\ =\ +$, out of a label set $\mathcal{Y} = \{+, -\}$

Standard Fine-Tuning (SFT)



Prompt-Based Fine-Tuning

Prompt-Based Fine-Tuning

- The original input x is modified using a *template* into a textual string prompt x' to create final string x_{hat} to derive output y.

Prompt-Based Fine-Tuning

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

Prompt-Based Fine-Tuning

- Verbalizers:

$$\mathcal{Z} = \{\text{excellent, good, OK, bad, horrible}\}$$

$$\mathcal{Y} = \{++, +, \sim, -, --\}$$

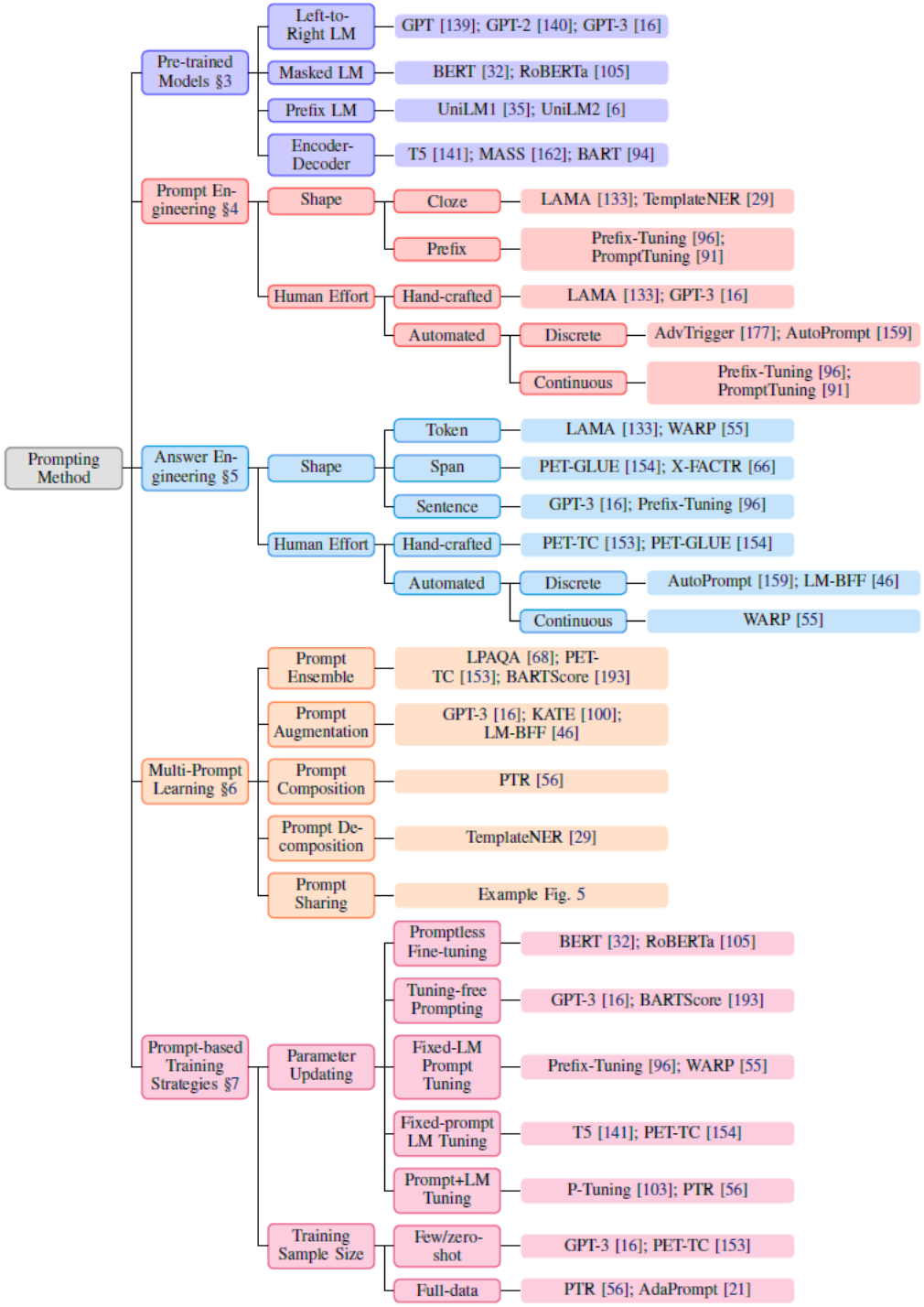
- Argmax() search that searches for the highest-scoring output

$$\hat{\boldsymbol{z}} = \underset{\boldsymbol{z} \in \mathcal{Z}}{\text{search}}\, P(f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z}); \theta)$$

Prompt-Based Fine-Tuning

Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Typology of Prompting Methods



https://arxiv.org/abs/2107.13586

Published Work / Applications

1. PFT for Urdu and Roman Urdu

- The input x:

$$x = \text{Yeh jaga bohat pyari hai.}$$

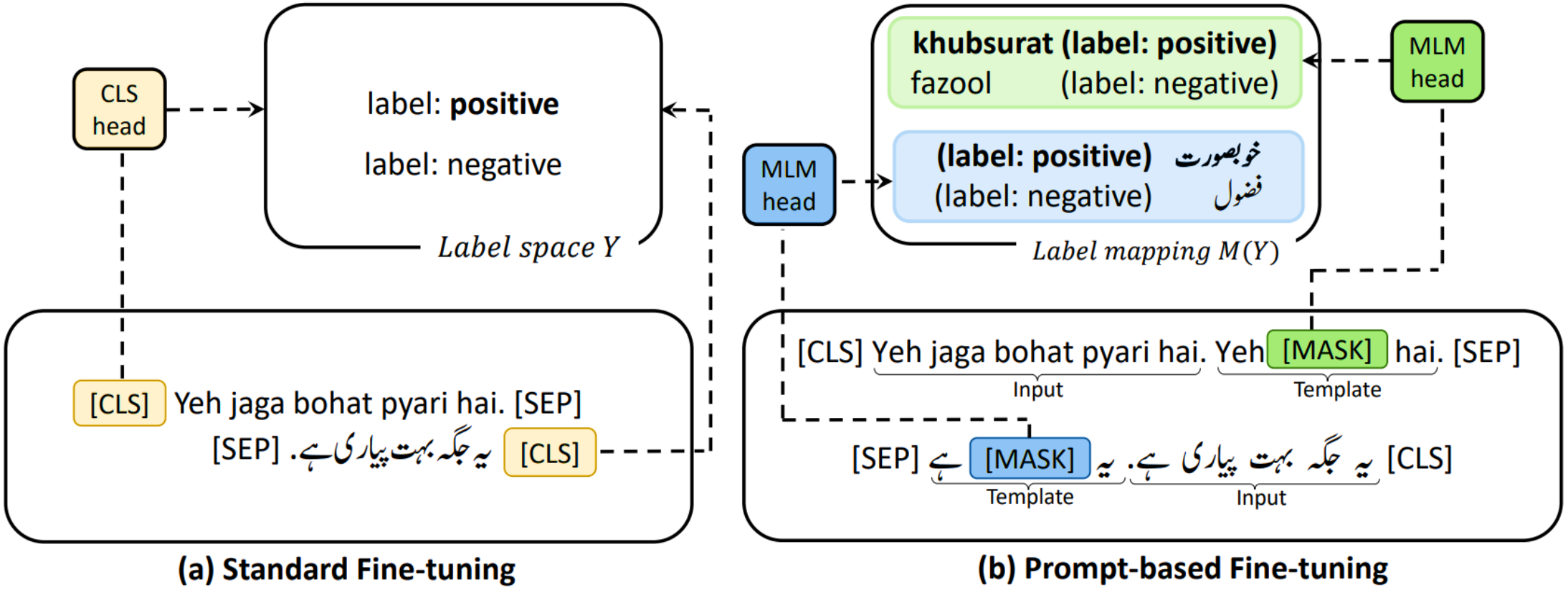
- The prompt formulation would be

$$x_{prompt} = [\text{CLS}]\ x\ \text{Yeh}\ [\text{MASK}]\ \text{hai.}\ [\text{SEP}]$$

- [MASK] = ? "khubsurat" (beautiful) or "fazool" (useless).
- For standard fine-tuning, we use the token sequence

$$x_{fine-tune} = [\text{CLS}]\ \text{Yeh jaga bohat pyari hai.}[\text{SEP}]$$

1. PFT for Urdu and Roman Urdu



An illustration of (a) standard fine-tuning and (b) prompt-based fine-tuning

1. Results

		UNED		UOD		RUD		RUHSOLD		RUED	
		Zero-Shot	4-Shot	Zero-Shot	4-Shot	Zero-Shot	4-Shot	Zero-Shot	4-Shot	Zero-Shot	4-Shot
Standard	BERT-M	9.0	21.4	50.0	60.2	50.0	63.2	54.0	53.0	21.0	27.2
	DistilBERT	18.0	18.8	51.0	58.6	51.0	55.0	54.0	49.2	16.0	27.4
	XLM-R	14.0	19.0	49.0	54.4	50.0	55.0	54.0	54.2	22.0	19.4
Prompt based	BERT-M	18.0	23.2	51.0	60.4	52.0	63.8	46.0	54.2	42.0	26.8
	DistilBERT	19.0	24.4	51.0	57.6	52.0	65.2	48.0	52.8	45.0	27.6
	XLM-R	16.0	**27.4**	51.0	**65.8**	50.0	**67.8**	48.0	**57.2**	**46.0**	29.0

Accuracy results for zero-shot and 4-Shot

		UNED		UOD		RUD		RUHSOLD		RUED	
		Zero-Shot	4-Shot	Zero-Shot	4-Shot	Zero-Shot	4-Shot	Zero-Shot	4-Shot	Zero-Shot	4-Shot
Standard	BERT-M	5.0	16.8	34.0	57.4	33.0	62.0	35.0	46.6	17.0	23.0
	DistilBERT	5.0	10.8	34.0	50.6	40.0	45.6	40.0	42.2	9.0	22.0
	XLM-R	4.0	10.0	33.0	43.4	33.0	45.2	35.0	46.0	12.0	12.4
Prompt based	BERT-M	6.0	20.4	34.0	59.2	46.0	63.4	44.0	52.6	19.0	22.8
	DistilBERT	**10.0**	20.4	34.0	55.2	39.0	64.2	40.0	51.8	18.0	24.6
	XLM-R	**10.0**	**25.6**	35.0	**63.6**	33.0	**66.4**	38.0	**55.8**	**16.0**	**26.8**

F1-Score results for zero-shot and 4-Shot

2. LISA

$$x' = f_{\text{prompt}}(x)$$

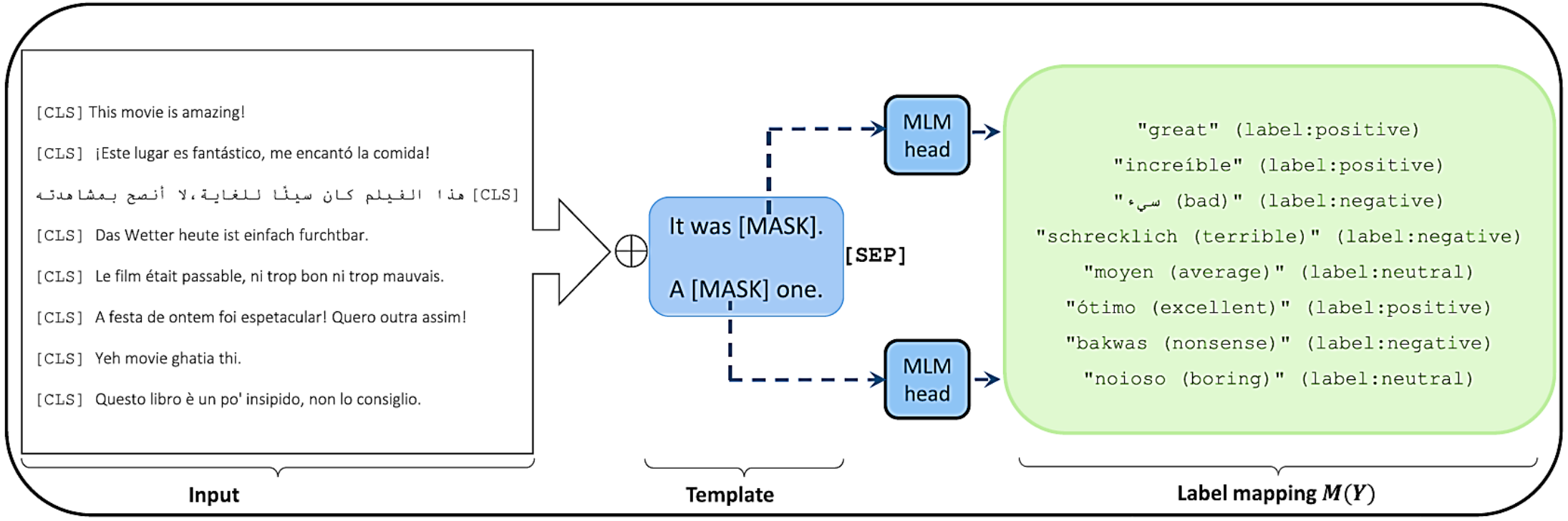
- where x′ represents the reformulated input containing a masked token.

$$P(y|x;\theta) = \sum_{z \in Z} P(z|x';\theta) M(y,z)$$

- probability of predicting token z for the masked prompt x', $M(y,z)$ is a label mapping function.

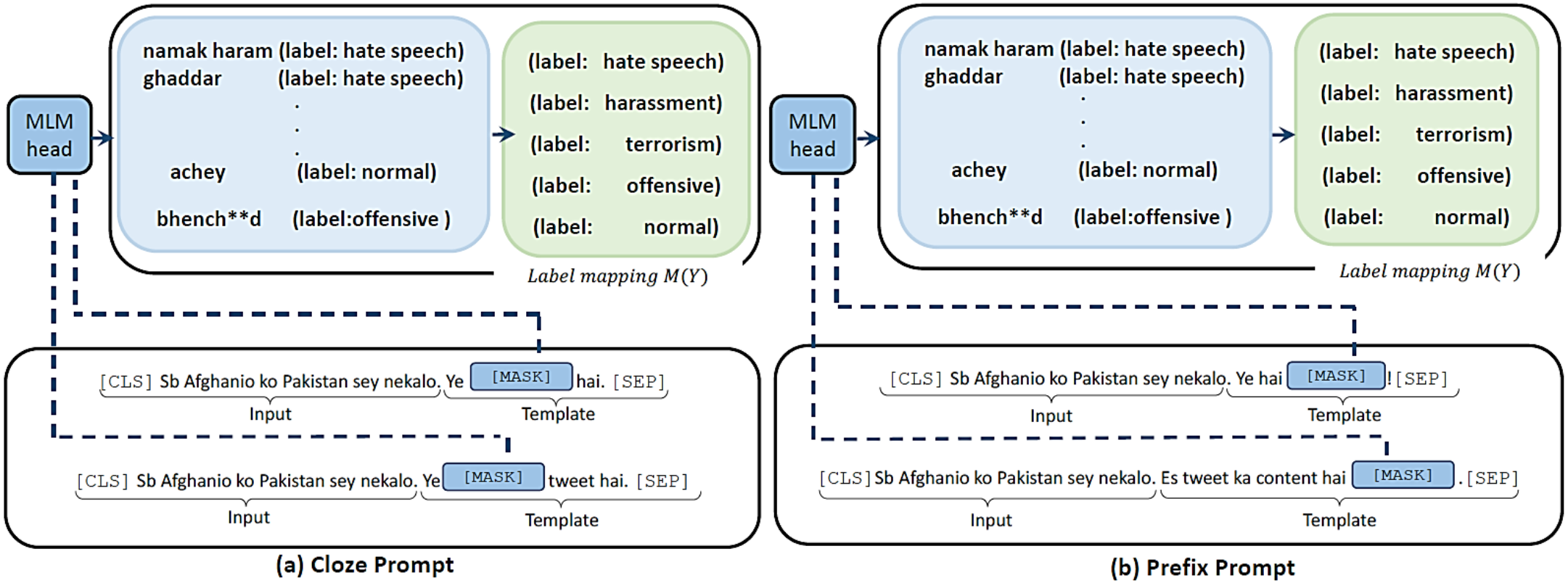
$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{z \in Z} M(y_i, z) \log P(z|x_i';\theta)$$

2. LISA



Proposed approach

3. CRU



An Illustration of (a) cloze prompts and (b) prefix prompts

References

References

[1] Ethnologue. (2022). What are the top 200 most spoken languages? Retrieved from https://www.ethnologue.com/guides/ethnologue200

[2] Wei, C., Shu, Y., Ou, M., He, Y. T., & Yu, F. R. (2025). PAFT: Prompt-Agnostic Fine-Tuning. arXiv preprint arXiv:2502.12859.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33:1877–1901.

[4] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021 Making pre-trained language models better few-shot learners. In Proceedings of the 59th Annual Meetin of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816 3830, Online. Association for Computational Linguistics.

[5] Santiago González-Carvajal and Eduardo C Garrido- Merchán. 2020. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

References

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7]Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, EricWallace, and Sameer Singh. 2020. AutoPrompt: Automatic prompt construction for masked language models. In Empirical Methods in Natural Language Processing (EMNLP).

[8]Timo Schick and Hinrich Sch¨utze. 2021a. Exploiting cloze questions for few-shot text classification and natural language inference. In European Chapter of the Association for Computational Linguistics (EACL).

[9] Timo Schick and Hinrich Sch¨utze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In North American Chapter of the Association for Computational Linguistics (NAACL).

References

[10] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting fewsample BERT fine-tuning. In International Conference on Learning Representations (ICLR).

[11] Rizwan, H., Shakeel, M. H., & Karim, A. (2020, November). Hate-speech and offensive language detection in roman Urdu. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 2512-2522).

[12] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys, 55(9), 1-35.