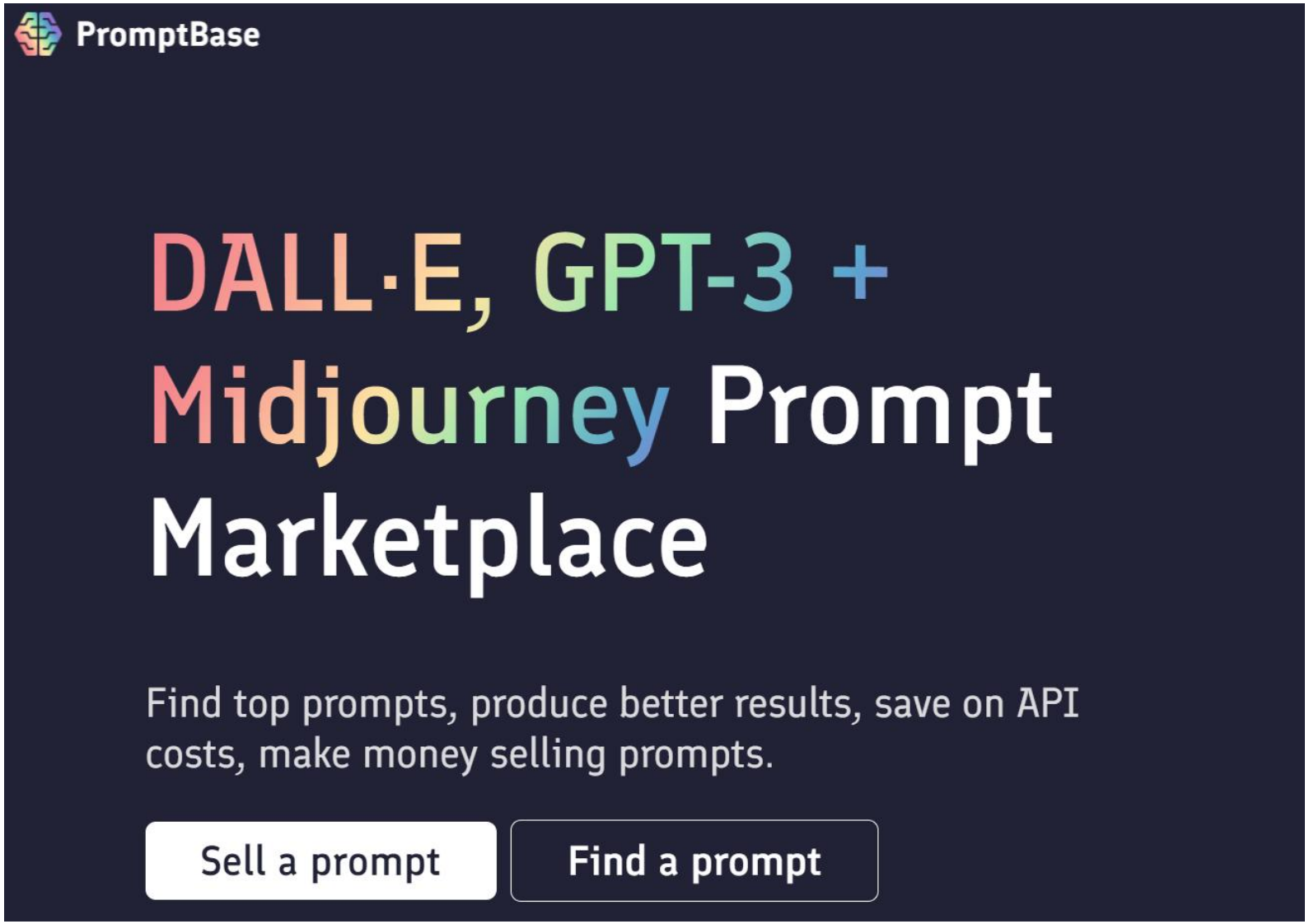
Carnegie Mellon University
Language Technologies Institute

The Foundations, Applications and Research trend of Prompt Learning

Pengfei Liu

http://pfliu.com/

Making Money by Selling your Prompts !



Outline

- What is the “Prompt”?
- What is the general workflow of prompt-based methods?
- What are the design considerations for prompt-based methods?
- What are things we can do now that we couldn‘t do in the past?
- What are the four paradigms in modern NLP?
- What could the next paradigm be?

What is the “Prompt”?

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ArXiv 2021,
Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig

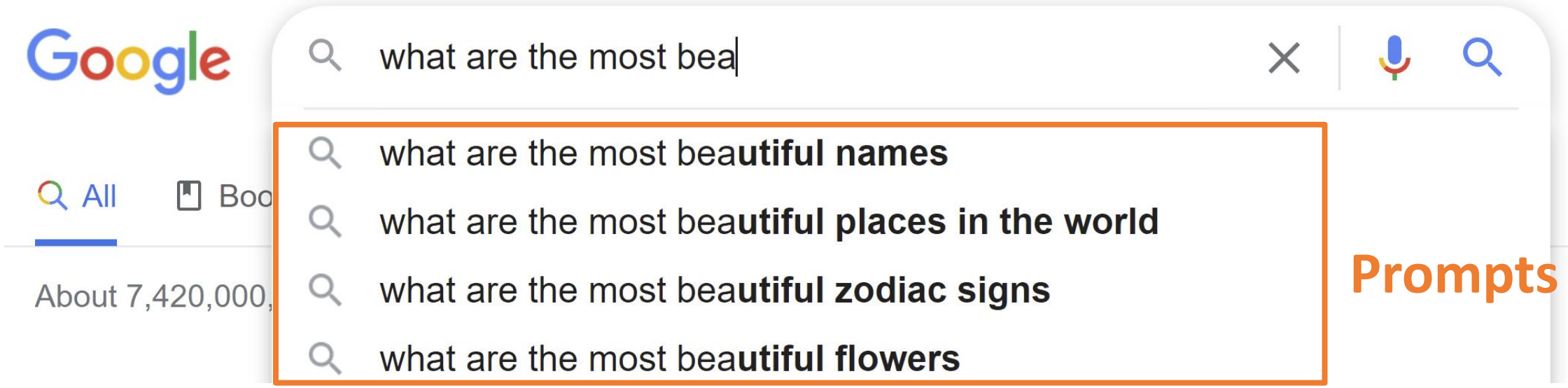
Prompt meaning prŏmpt

Words form: prompted promptest prompting prompts See word origin >

The definition of a prompt is a cue given to someone to help him remember what to say, or is something that causes another event or action to occur. verb

An example of prompt is when you whisper a line to an actor who forgot what to say next.

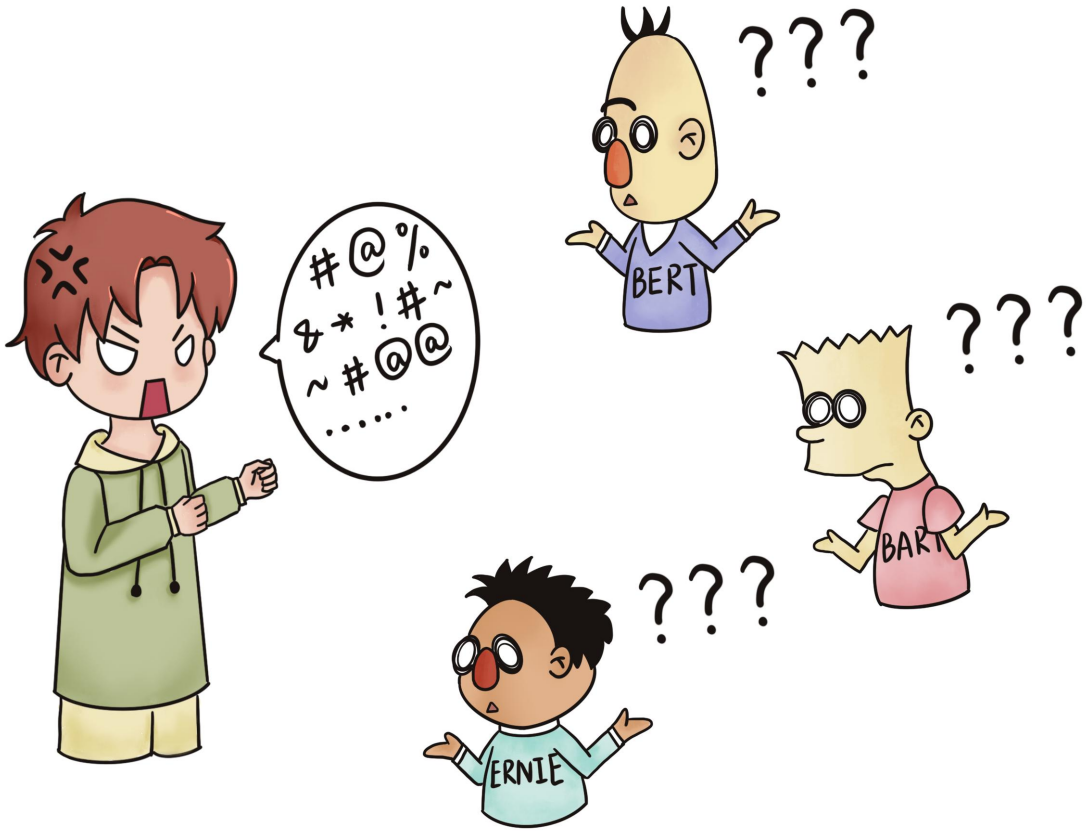
An example of prompt is an event that starts an argument.


Google
what are the most bea
All
Boo
About 7,420,000,
what are the most beautiful names
what are the most beautiful places in the world
what are the most beautiful zodiac signs
what are the most beautiful flowers
Prompts

What is the “prompt” in the context of NLP research?

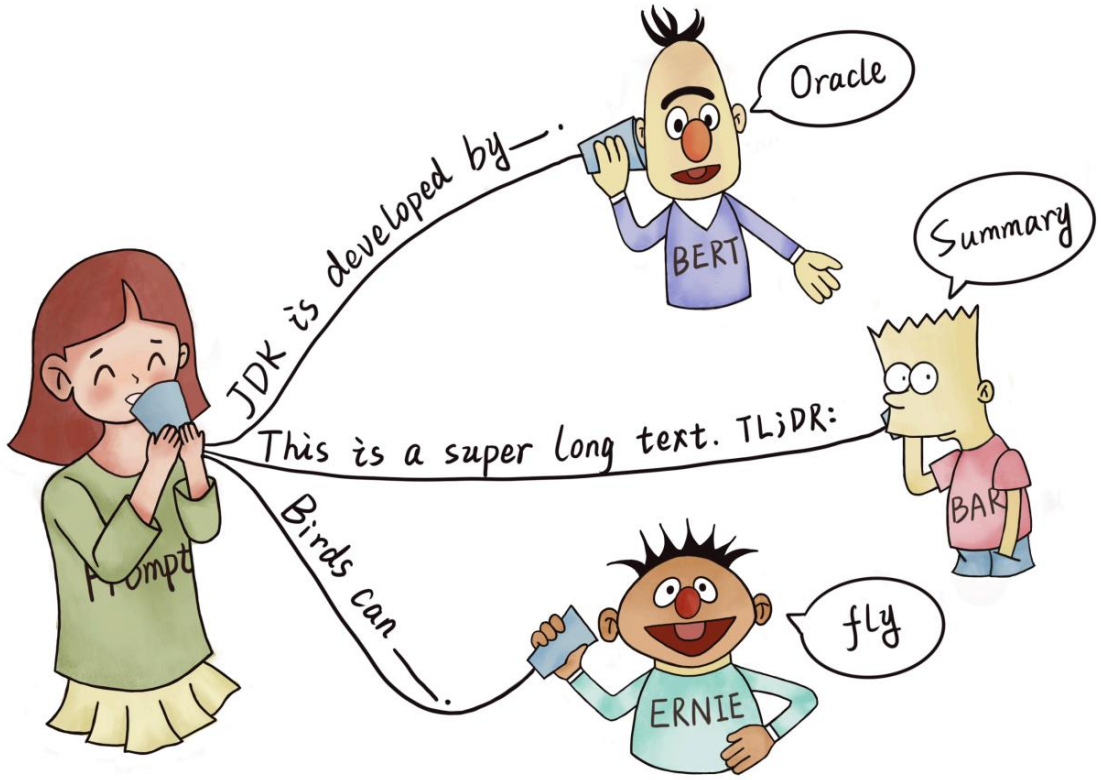
An Intuitive Definition

- Prompt is a cue given to the **pre-trained language model** to allow it better understand **human**'s questions



An Intuitive Definition

- Prompt is a cue given to the **pre-trained language model** to allow it better understand **human**'s questions



More Technical Definition

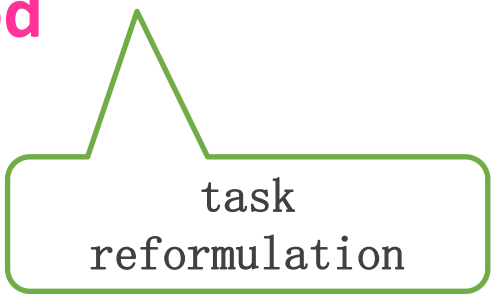
- Prompt is the technique of making better use of the knowledge from the pre-trained model by adding additional texts to the input.

More Technical Definition

- Prompt is the technique of making better use of the knowledge from the pre-trained model by adding additional texts to the input.

Purpose

Method



What is the general workflow of prompt-based methods?

Workflow for Prompting Methods

- Prompt Construction
- Answer Construction
- Answer Prediction
- Answer-Label Mapping

Prompting for Sentiment Classification

- **Task Description**:
  - **Input**: sentence x;
  - **Output**: emotional polarity of it (i.e.,☺ v.s ☹).

Input: x = I love this movie.

Step 1: Prompt Construction

- Transform x into prompt x' through following two steps:
  - Defining a **template** with two **slots**:

    [x] and [z];

Input: x = I love this movie.

Template: [x] Overall, it was a [z] movie.

Step 1: Prompt Construction

- Transform x into prompt x' through following two steps:

  - Defining a **template** with two **slots**:

    [x] and [z];

Require human effort

Input: x = I love this movie.

Template: [x] Overall, it was a [z] movie.

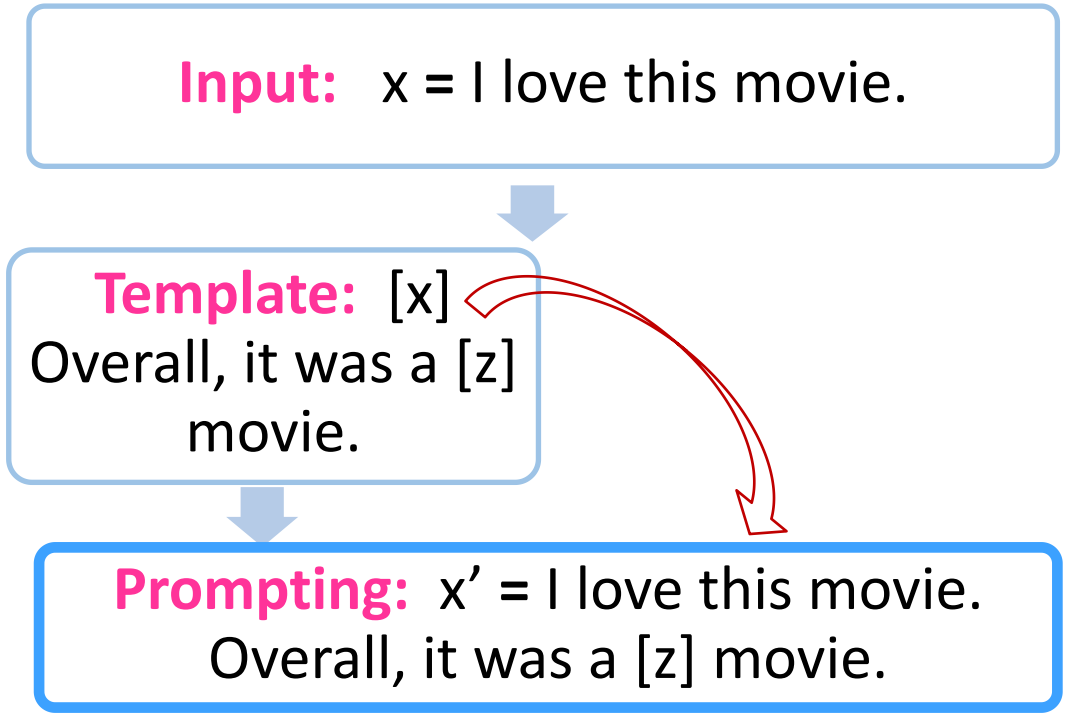
Step 1: Prompt Construction

- Transform x into prompt x' through following two steps:

  - Defining a **template** with two **slots**:

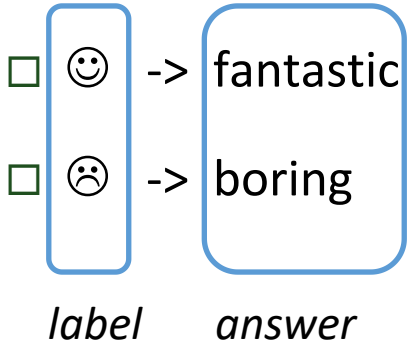
    [x] and [z];

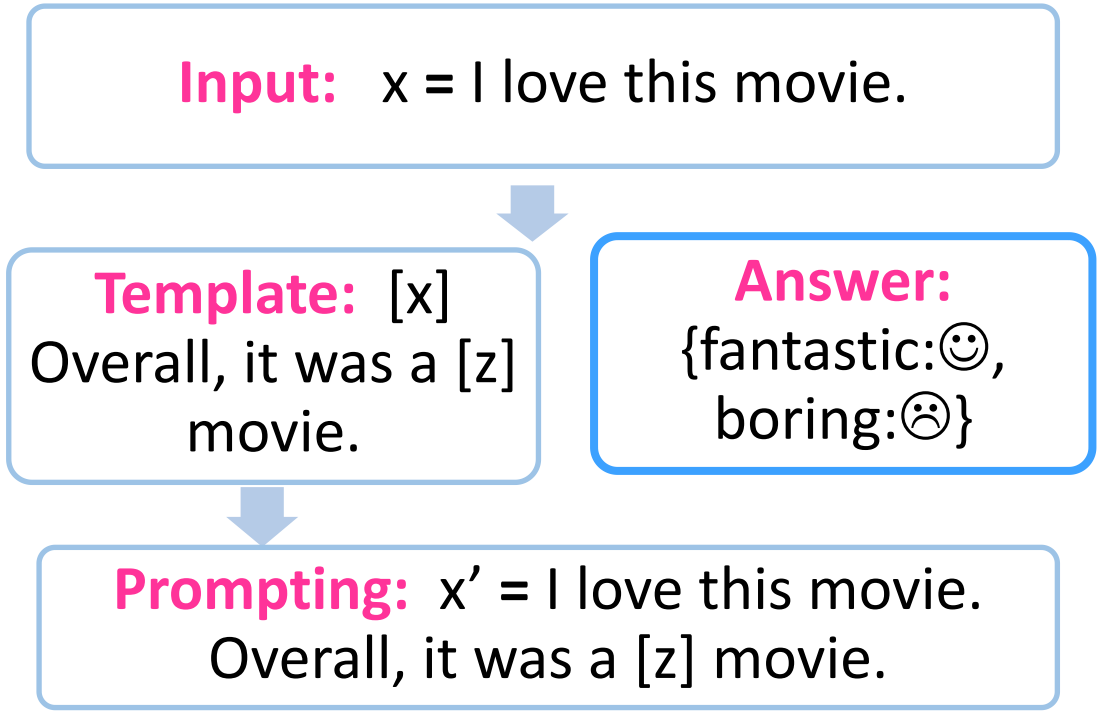
  - Instantiate slot [x] with input text.



Step 2: Answer Construction

- Build a mapping function between answers and class labels.





Step 3: Answer Predicting

- Given a prompt， predict the answer [z].

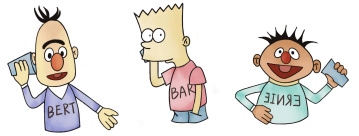
  - Choose a suitable pretrained language model;

Input: x = I love this movie.

Template: [x] Overall, it was a [z] movie.

Answer: {fantastic:☺, boring:☹}

Prompting: x’ = I love this movie. Overall, it was a [z] movie.



Which one?

Step 3: Answer Predicting

- Given a prompt， predict the answer [z]

  - Choose a suitable pretrained language model;

  - Fill in [z] as “fantastic”

Input: x = I love this movie.

Template: [x] Overall, it was a [z] movie.

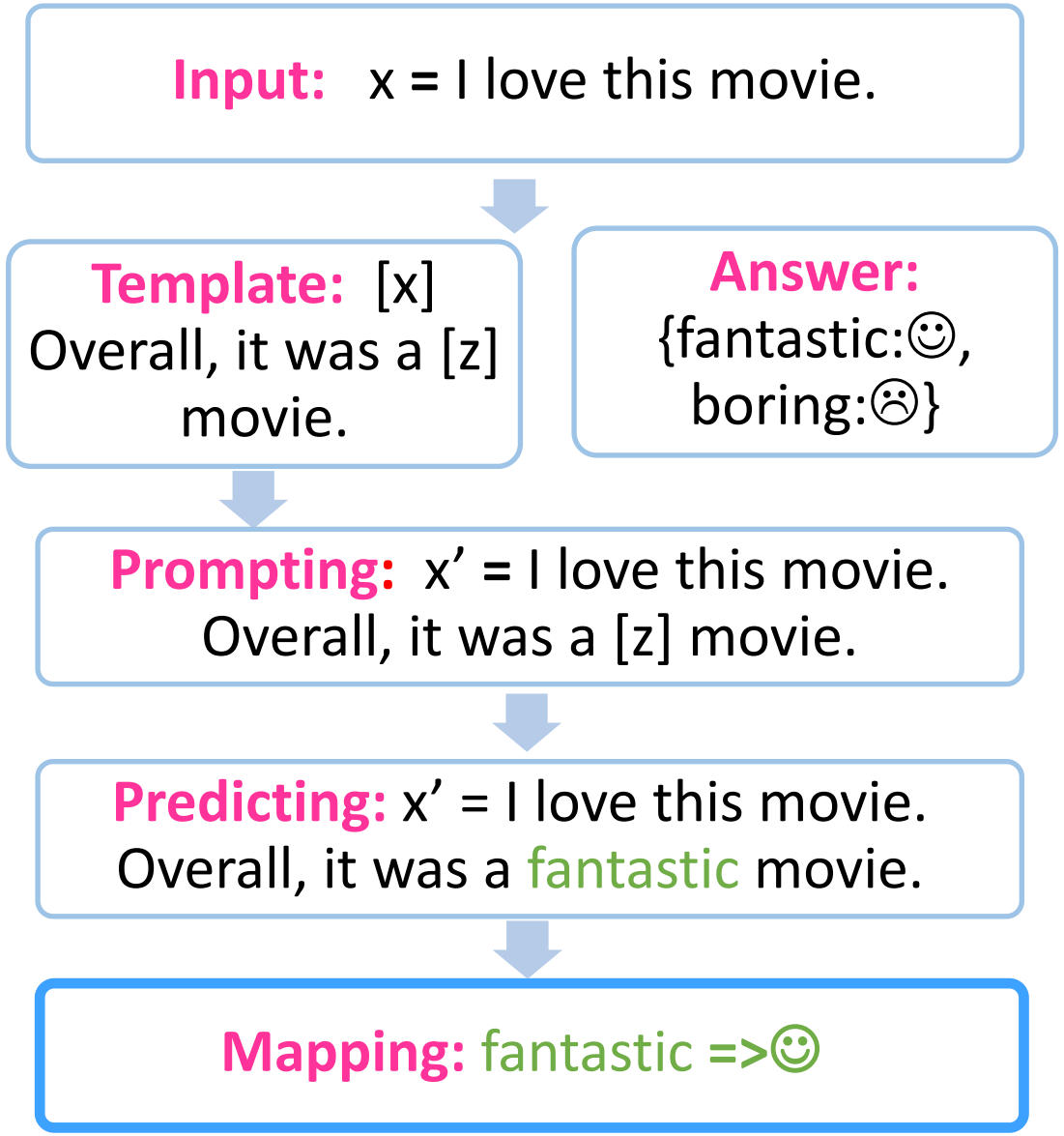
Answer: {fantastic:☺, boring:☹}

Prompting: x’ = I love this movie. Overall, it was a [z] movie.

Predicting: x’ = I love this movie. Overall, it was a fantastic movie.

Step 4: Answer Mapping

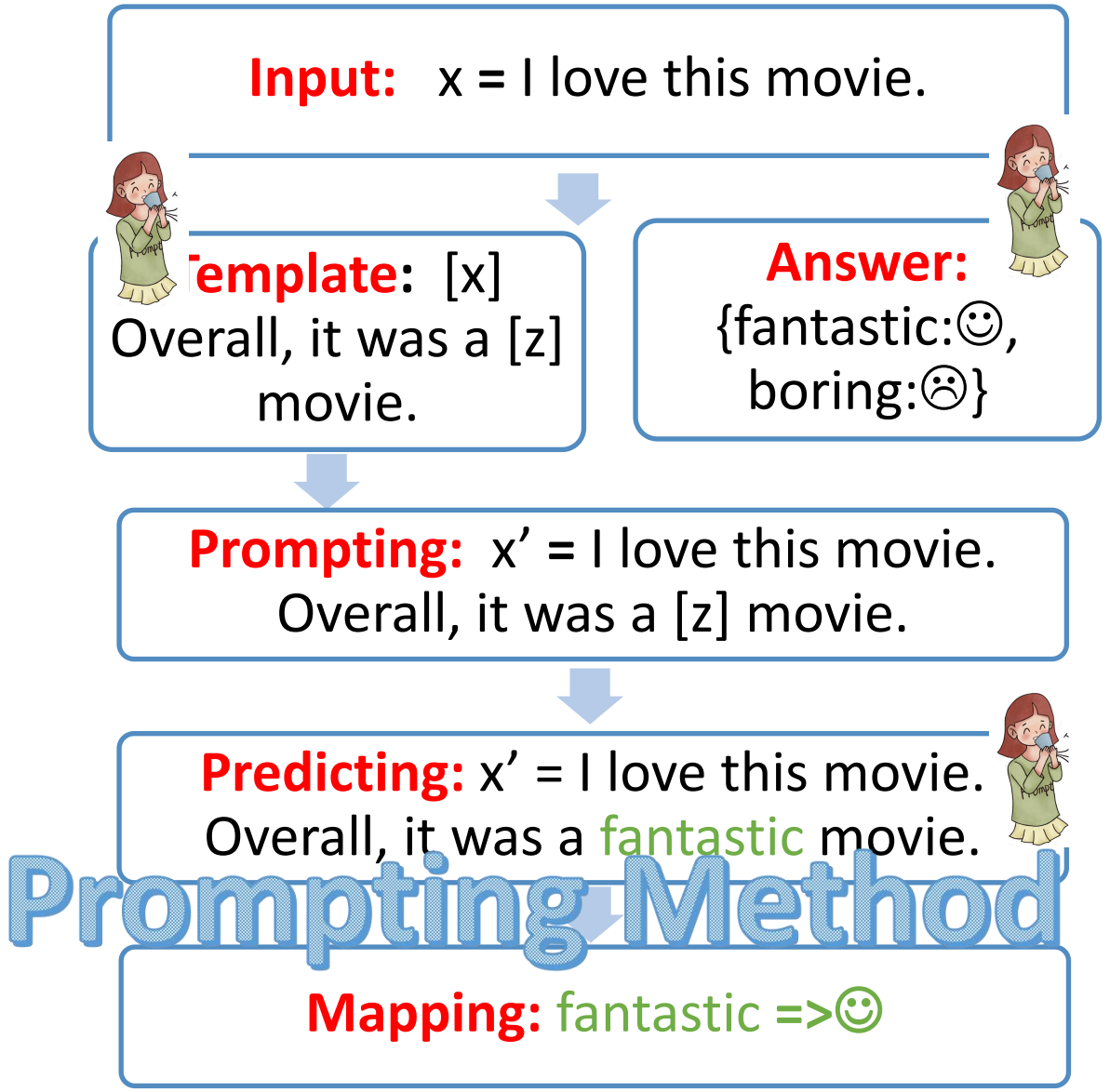
- Mapping: Given an answer, map it into a class label.
  - ☐ fantastic => ☺



Summary

Terminology	Notation	Example
Input	x	I love this movie
Output (label)	y	☺ ☹
Template	-	[x] Overall, it was a [z] movie
Prompt	x’	I love this movie. Overall, it was a [z] movie
Answer	z	fantastic, boring

Rethinking Human Efforts in Prompt-based Methods



Rethinking Human Efforts in Prompt-based Methods

Input: x = I love this movie.

Predicting: ☺

Traditional Method

Input: x = I love this movie.

Template: [x] Overall, it was a [z] movie.

Answer: {fantastic:☺, boring:☹}

Prompting: x' = I love this movie. Overall, it was a [z] movie.

Predicting: x' = I love this movie. Overall, it was a fantastic movie.

Prompting Method

Mapping: fantastic =>☺

How do you think about the methodology that allows for introducing **human priors**?

What are the design considerations for prompt-based methods?

Design Considerations for Prompt-based Methods

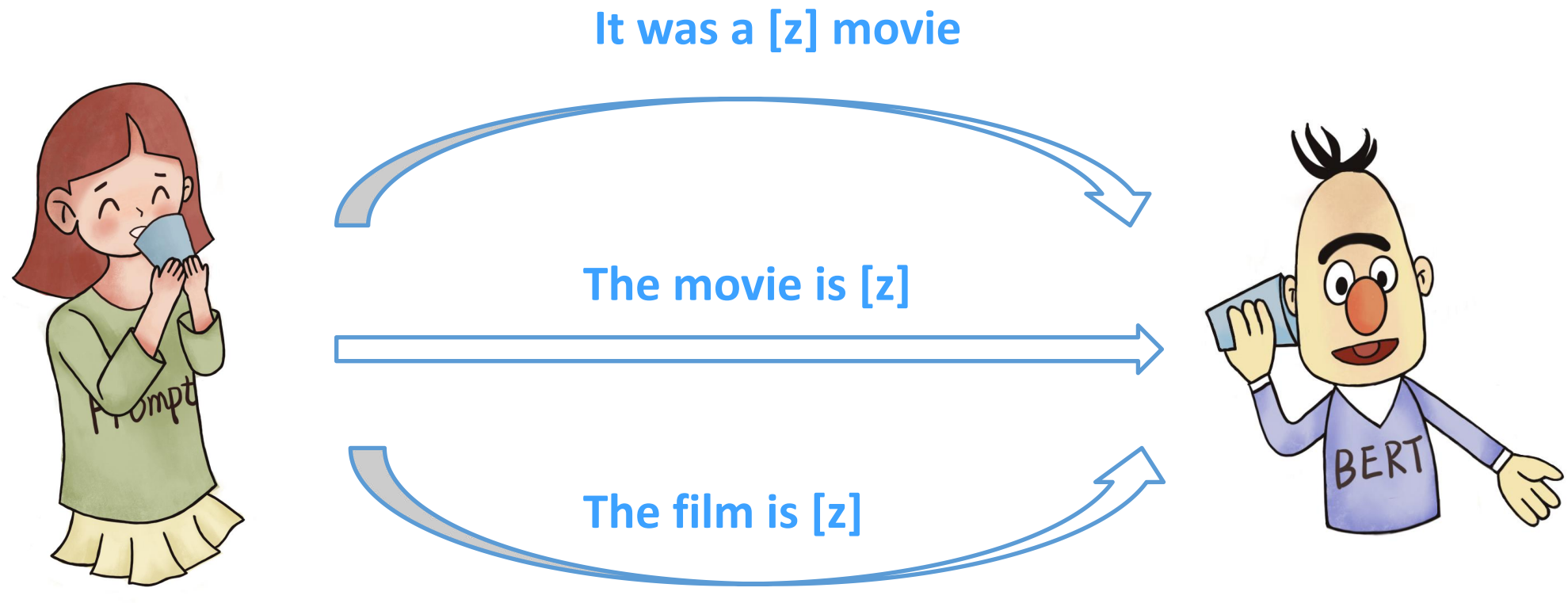
- Prompt Template Engineering
- Answer Engineering
- Pre-trained Model Choice
- Expanding the Paradigm
- Prompt-based Training Strategies

Design Considerations for Prompt-based Methods

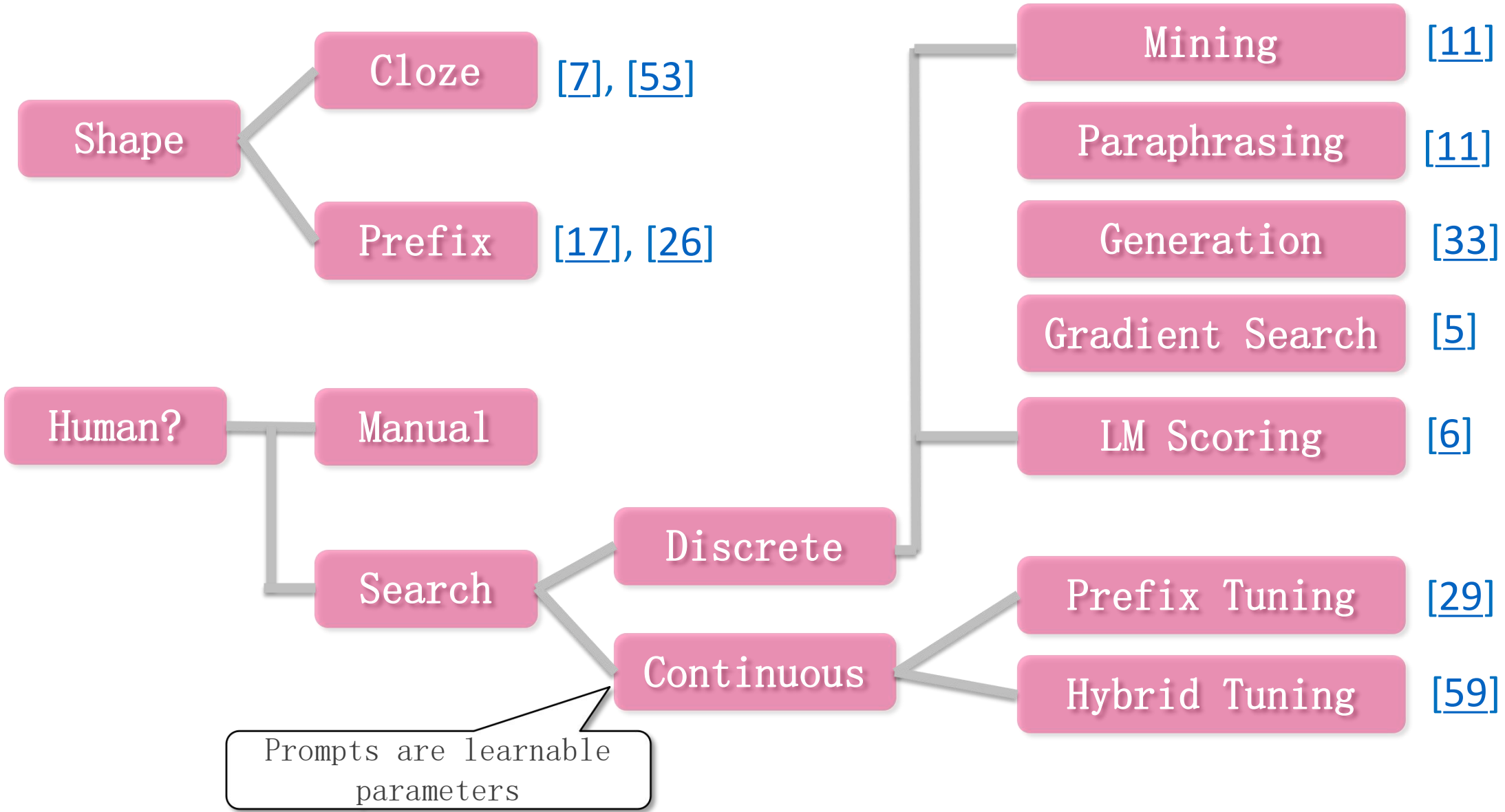
- **Prompt Template Engineering**
- Answer Engineering
- Pre-trained Model Choice
- Expanding the Paradigm
- Prompt-based Training Strategies

Prompt Template Engineering

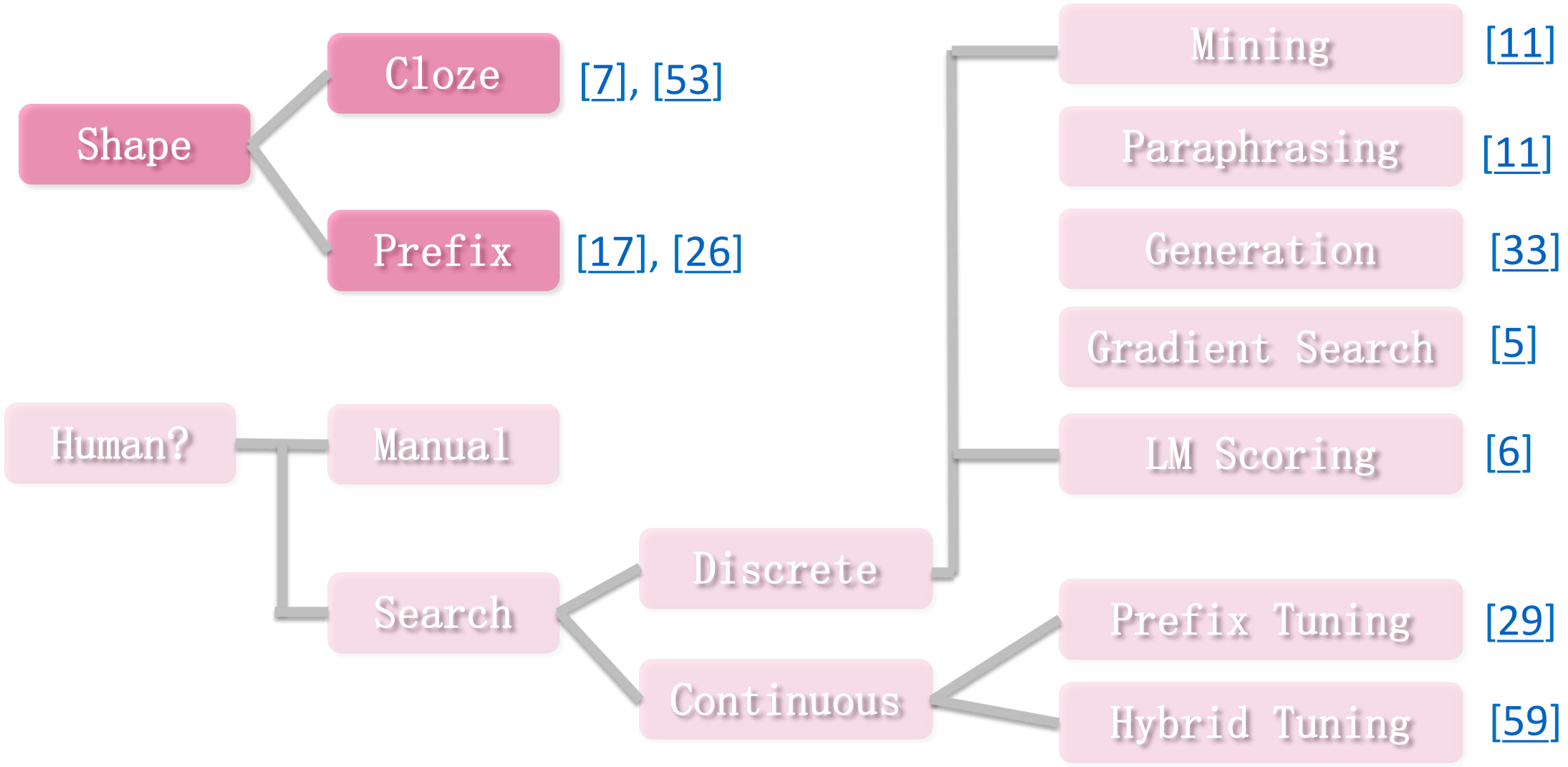
- Research Question:
  - **how to define appropriate prompt templates**



Design Decision of Prompt Templates



Design Decision of Prompt Templates



Prompt Shape

- **Cloze Template**
  - Contain blanks to be filled.

Prompt Shape

- **Cloze Template**
  - Contain blanks to be filled.
  - Useful for Masked LMs.
    - The capital of ____ is Beijing .

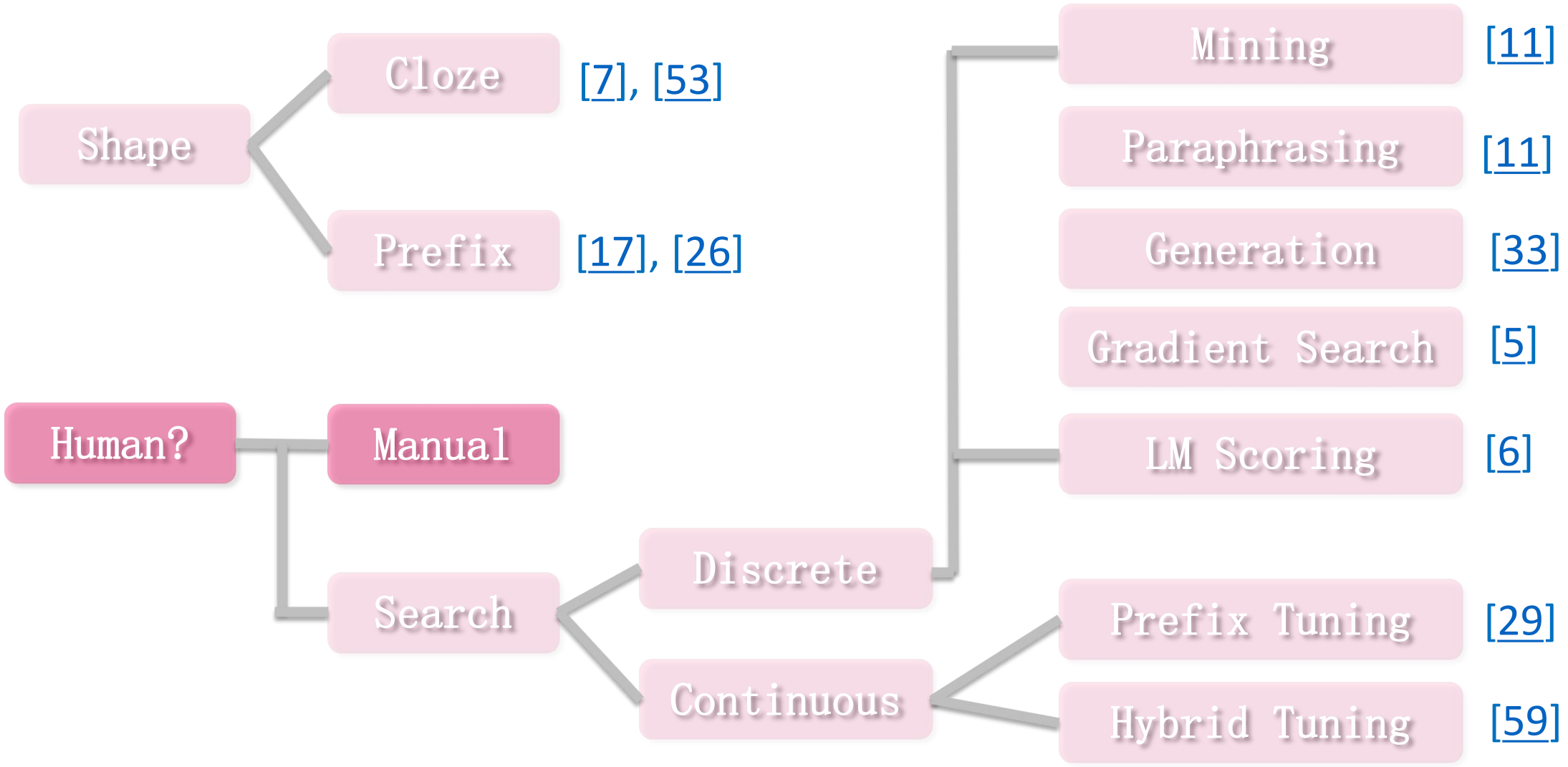
Prompt Shape

- **Cloze Template**
- Prefix Template
  - Contain a string prefix to be continued.

Prompt Shape

- **Cloze Template**
- Prefix Template
  - Contain a string prefix to be continued.
  - Useful for Left-to-right LM and Encoder-Decoder LM.
    - President Joe Biden and three of his European allies face …… TL;DR: ____

Design Decision of Prompt Templates



Manual Template Design

- Manual Prompt
  - The most natural way to create prompts 
    - I love this movie so much! What's the sentiment of the text? ___ .
    - President Joe Biden and three of his European allies face …… In summary, ___ .
    - President Joe Biden and three of his European allies face …… The article is about ___ .

Manual Template Design

- Manual Prompt
  - The most natural way to create prompts 😄
  - An art that takes time and experience. 😔

Zero-shot Accuracy (BERT-base, SST-2)

- First template–answer pair

  Template: <A movie review> The movie is ___ .

  Answer: fantastic/terrible

  0.749

- Second template–answer pair

  Template: <A movie review> The review is ___ .

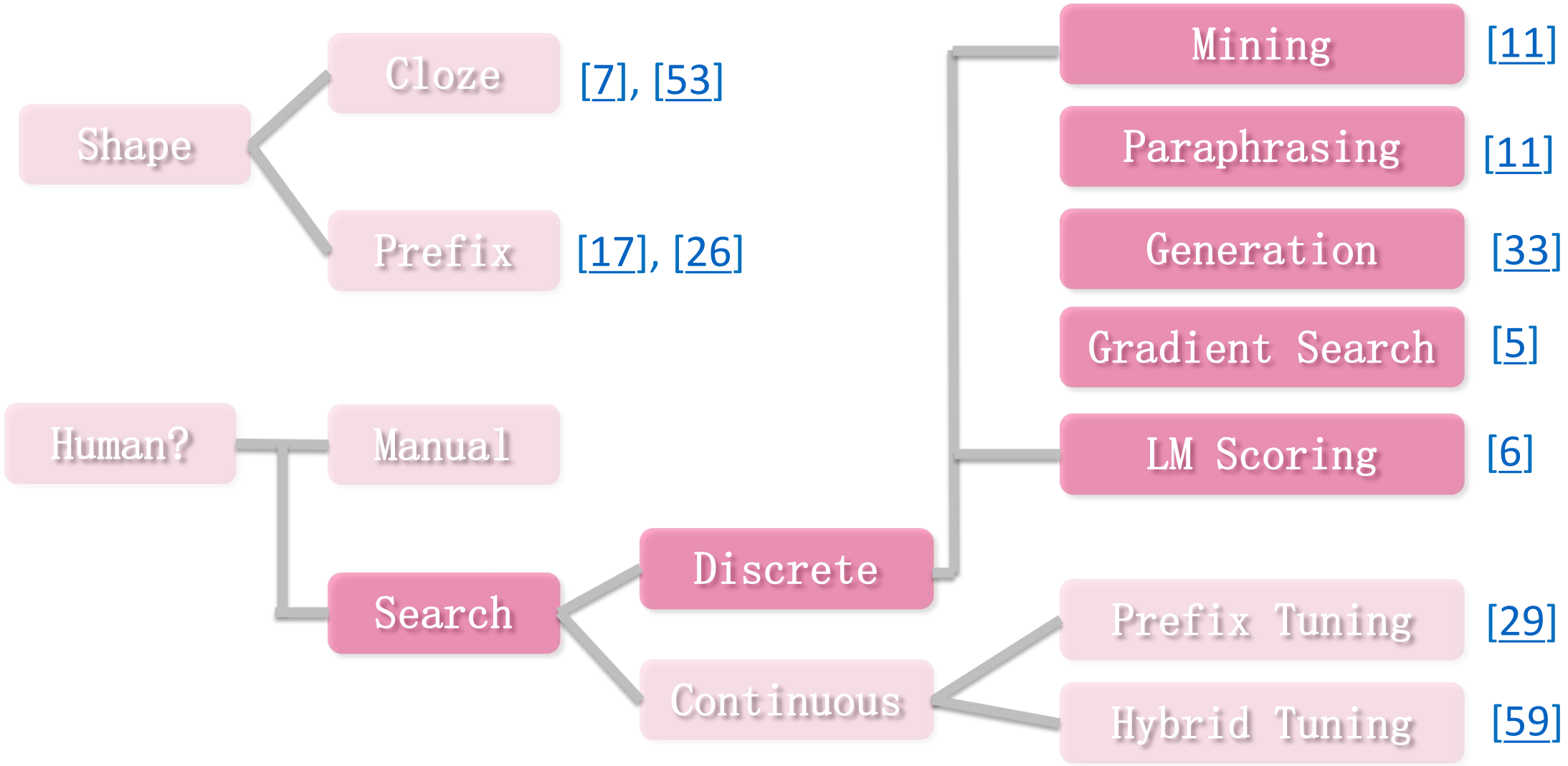
  Answer: positive/negative

  0.534

Manual Template Design

- Manual Prompt
  - The most natural way to create prompts 😄
  - An art that takes time and experience. 😔
  - For some complicated tasks, it’s hard to manually craft templates.

Design Decision of Prompt Templates



Discrete Search

- **Mining**
- Paraphrasing
- Gradient-based Search
- Generation
- LM Scoring

Discrete Search

- **Mining**
  - Use a large corpus to mine templates that contain both the input and the gold answer.
  - Example
    - Fact retrieval for country-capital relationship
    - search through Wikipedia and find strings that contain both ``Beijing'' and ``China'' or other pairs.

Input	Gold answer
China	Beijing
Japan	Tokyo
United States	Washington

- Beijing, the capital of China
- The capital of China is Beijing
- ……

Reference: Jiang et al. How Can We Know What Language Models Know? TACL(2020)

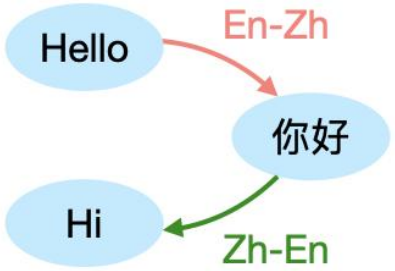
Discrete Search

- **Paraphrasing**
  - Take in an existing seed template, and paraphrases it into a set of other candidate templates.

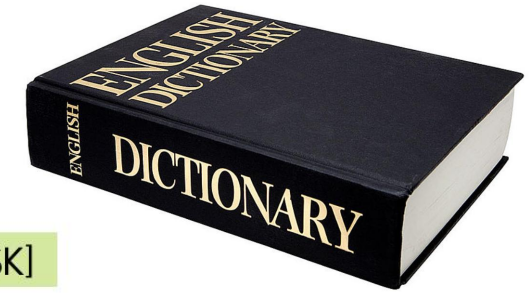
References: [1] Jiang et al. How Can We Know What Language Models Know? TACL (2020). [2] Yuan et al. BARTScore: Evaluating Generated Text as Text Generation. NeurIPS (2021). [3] Haviv et al. BERTese: Learning to Speak to BERT. EACL (2021).

Discrete Search

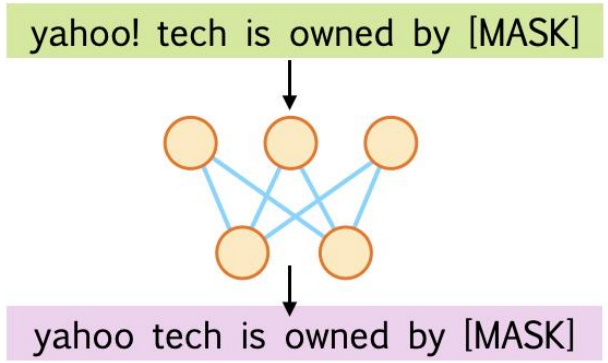
- **Paraphrasing**
  - Take in an existing seed template, and paraphrases it into a set of other candidate templates.
  - Typical methods
    - Back-translation



    - Using replacement of phrases from a thesaurus



    - Use neural rewriter to rewrite



References: [1] Jiang et al. How Can We Know What Language Models Know? TACL (2020). [2] Yuan et al. BARTScore: Evaluating Generated Text as Text Generation. NeurIPS (2021). [3] Haviv et al. BERTese: Learning to Speak to BERT. EACL (2021).

Discrete Search

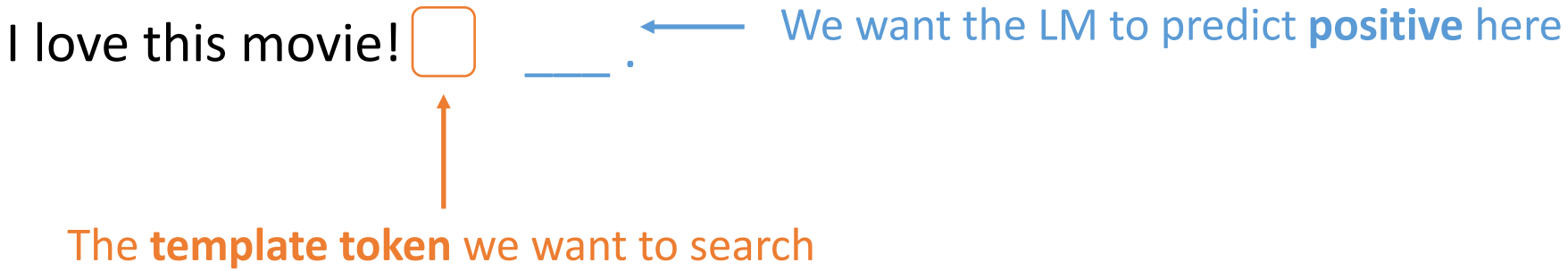
- **Gradient-based Search**
  - Stepping through tokens and find ones that can trigger desired outputs.
  - To determine how to replace the current tokens, a gradient-based approach is used.
  - Update the embedding for every token to minimizes the loss's first-order Taylor approximation around the current token embedding

$$\arg\min_{\mathbf{e}'_i \in \mathcal{V}} \left[\mathbf{e}'_i - \mathbf{e}_{adv_i}\right]^\top \nabla_{\mathbf{e}_{adv_i}} \mathcal{L}$$

Reference: Shin et al. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP (2020).

Discrete Search

- **Gradient-based Search**
  - Stepping through tokens and find ones that can trigger desired outputs.



Reference: Shin et al. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP (2020).

Discrete Search

- **Gradient-based Search**
  - Stepping through tokens and find ones that can trigger desired outputs.

I love this movie! [ ] ____ . ← We want the LM to predict **positive** here

Token	P(positive)
is	0.8
hello	0.09
cat	0.04
….	…

Reference: Shin et al. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP (2020).

Discrete Search

- **Gradient-based Search**
  - Stepping through tokens and find ones that can trigger desired outputs.

I love this movie! [ ] ____ . ← We want the LM to predict **positive** here

Token	P(positive)
is	0.8
hello	0.09
cat	0.04
….	…

Reference: Shin et al. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. EMNLP (2020).

Discrete Search

- **Generation**
  - Use LM to generate templates.
    - T5

Pre-train	Input: Thank you <X> me to the party <Y> week. Target: <X> for inviting <Y> last <Z>

Reference: Gao et al. Making Pre-trained Language Models Better Few-shot Learners. ACL (2021).

Discrete Search

- **Generation**
  - Use LM to generate templates.
    - T5

I love this movie! <X> great <Y>

<X> This is <Y> . <Z>

<X> A <Y> one. <Z>

……

Reference: Gao et al. Making Pre-trained Language Models Better Few-shot Learners. ACL (2021).

Discrete Search

- **Generation**
  - Use LM to generate templates.
    - T5
    - Use a trained template generator
      - Supervised Training
      - Reinforcement Learning

References: [1] Guo et al. Text Generation with Efficient (Soft) Q-Learning. arXiv preprint arXiv:2106.07704 (2021). [2] Ben-David et al. PADA: A Prompt-based Autoregressive Approach for Adaptation to Unseen Domains. arXiv:2102.12206 (2021).

Discrete Search

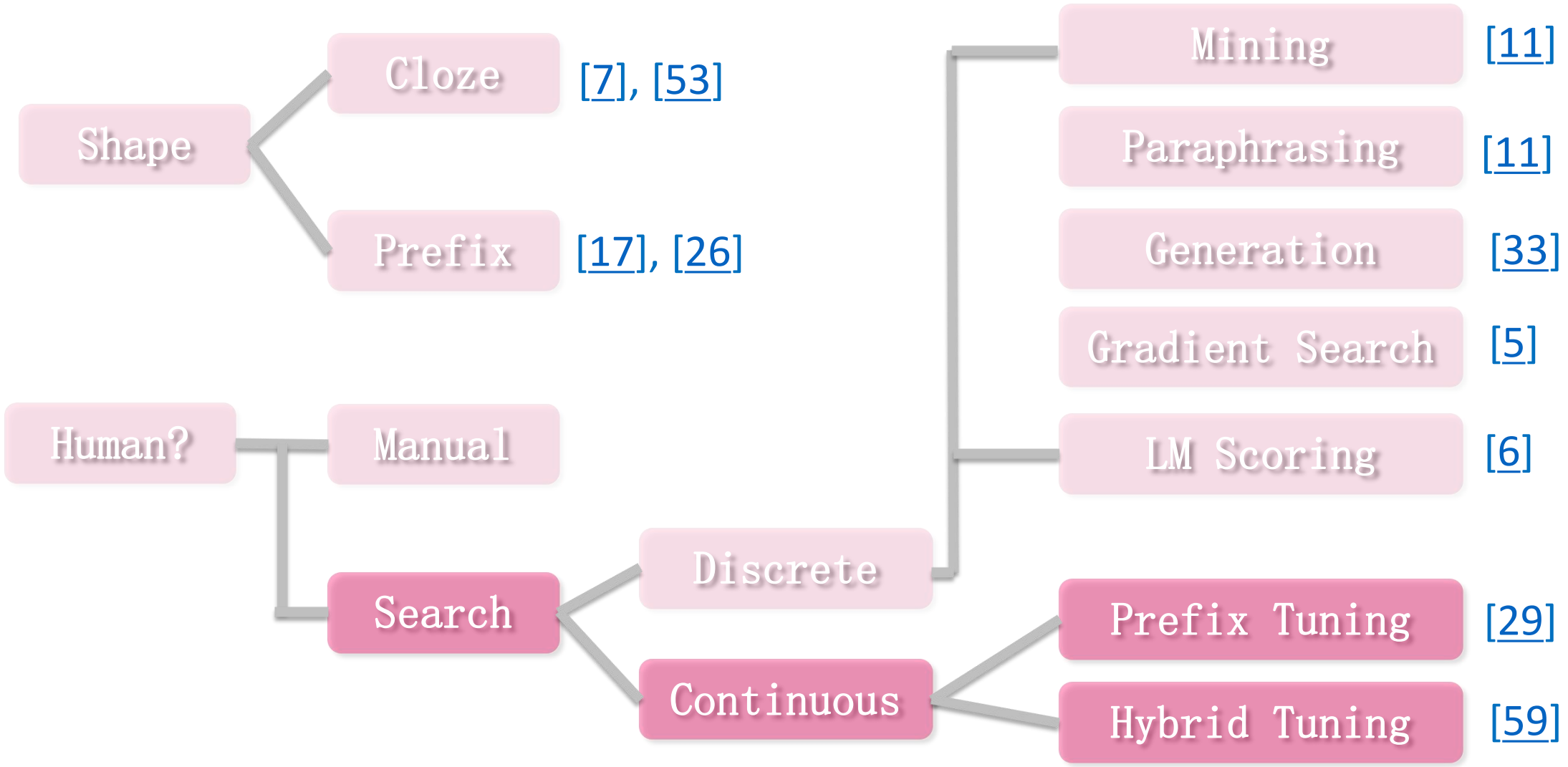
- **LM Scoring**
  - Use the LM to choose the templates that achieve high LM probability.

I love this movie! <template> positive.

Sequence	P
I love this movie! The sentiment of the text is positive.	0.4
I love this movie! Hello world positive	0.09
I love this movie! The text is positive	0.3
….	…

Reference: Davison et al. Commonsense Knowledge Mining from Pretrained Models. EMNLP (2019).

Design Decision of Prompt Templates



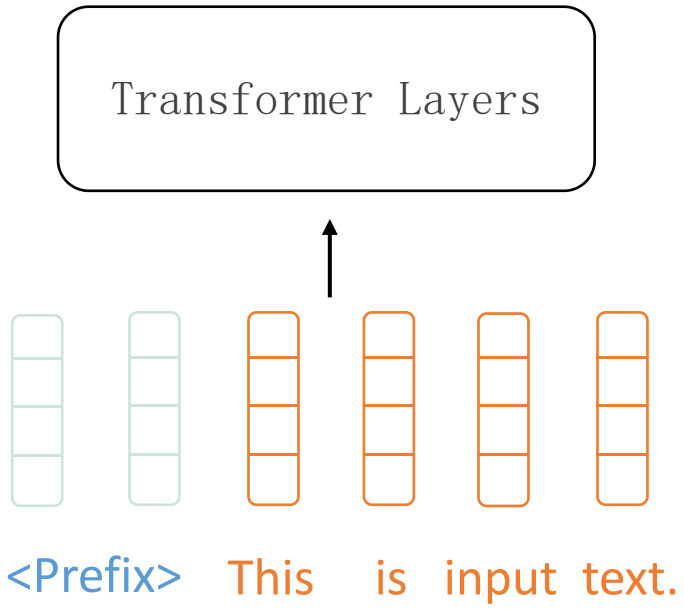
Continuous Template Search

- **Prefix Tuning**
  - Prepends a sequence of continuous task-specific vectors to the input, while keeping the LM parameters frozen.

References: [1] Li et al. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190 (2021). [2] Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691 (2021)

Continuous Template Search

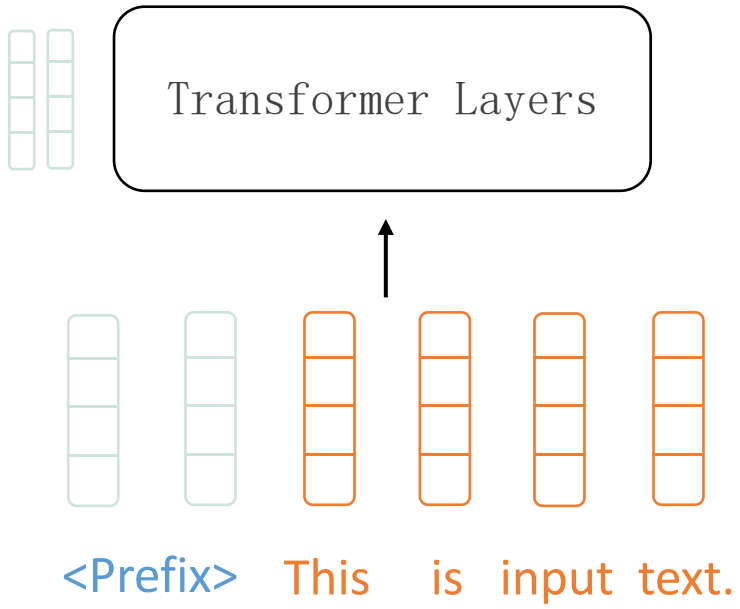
- **Prefix Tuning**
  - Prepends a sequence of continuous task-specific vectors to the input, while keeping the LM parameters frozen.
    - Shallow Prefix Tuning



Reference: Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691 (2021)

Continuous Template Search

- **Prefix Tuning**
  - Prepends a sequence of continuous task-specific vectors to the input, while keeping the LM parameters frozen.
    - Shallow Prefix Tuning
    - Deep Prefix Tuning



Reference: Li et al. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190 (2021).

Continuous Template Search

- **Hybrid Tuning**
  - An extension of prefix tuning

References: [1] Zhong et al. Factual Probing Is [MASK]: Learning vs. Learning to Recall. CoRR abs/2104.05240 (2021). [2] Liu et al. GPT Understands, Too. CoRR abs/2103.10385 (2021).

Continuous Template Search

- **Hybrid Tuning**
  - An extension of prefix tuning
  - The positions of tunable virtual tokens can be anywhere.

References: [1] Zhong et al. Factual Probing Is [MASK]: Learning vs. Learning to Recall. CoRR abs/2104.05240 (2021). [2] Liu et al. GPT Understands, Too. CoRR abs/2103.10385 (2021).

Continuous Template Search

- **Hybrid Tuning**
  - An extension of prefix tuning
  - The positions of tunable virtual tokens can be anywhere.
  - Use hard templates initialization

I love this movie so much! `The` `sentiment` `is` positive.

References: [1] Zhong et al. Factual Probing Is [MASK]: Learning vs. Learning to Recall. CoRR abs/2104.05240 (2021). [2] Liu et al. GPT Understands, Too. CoRR abs/2103.10385 (2021).

Continuous Template Search

- **Hybrid Tuning**
  - An extension of prefix tuning
  - The positions of tunable virtual tokens can be anywhere.
  - Use hard templates initialization
  - Combine hard and soft template tokens

I love this movie so much! [ ] [ ] [ ] is positive.

References: [1] Zhong et al. Factual Probing Is [MASK]: Learning vs. Learning to Recall. CoRR abs/2104.05240 (2021). [2] Liu et al. GPT Understands, Too. CoRR abs/2103.10385 (2021).

Design Considerations for Prompt-based Methods

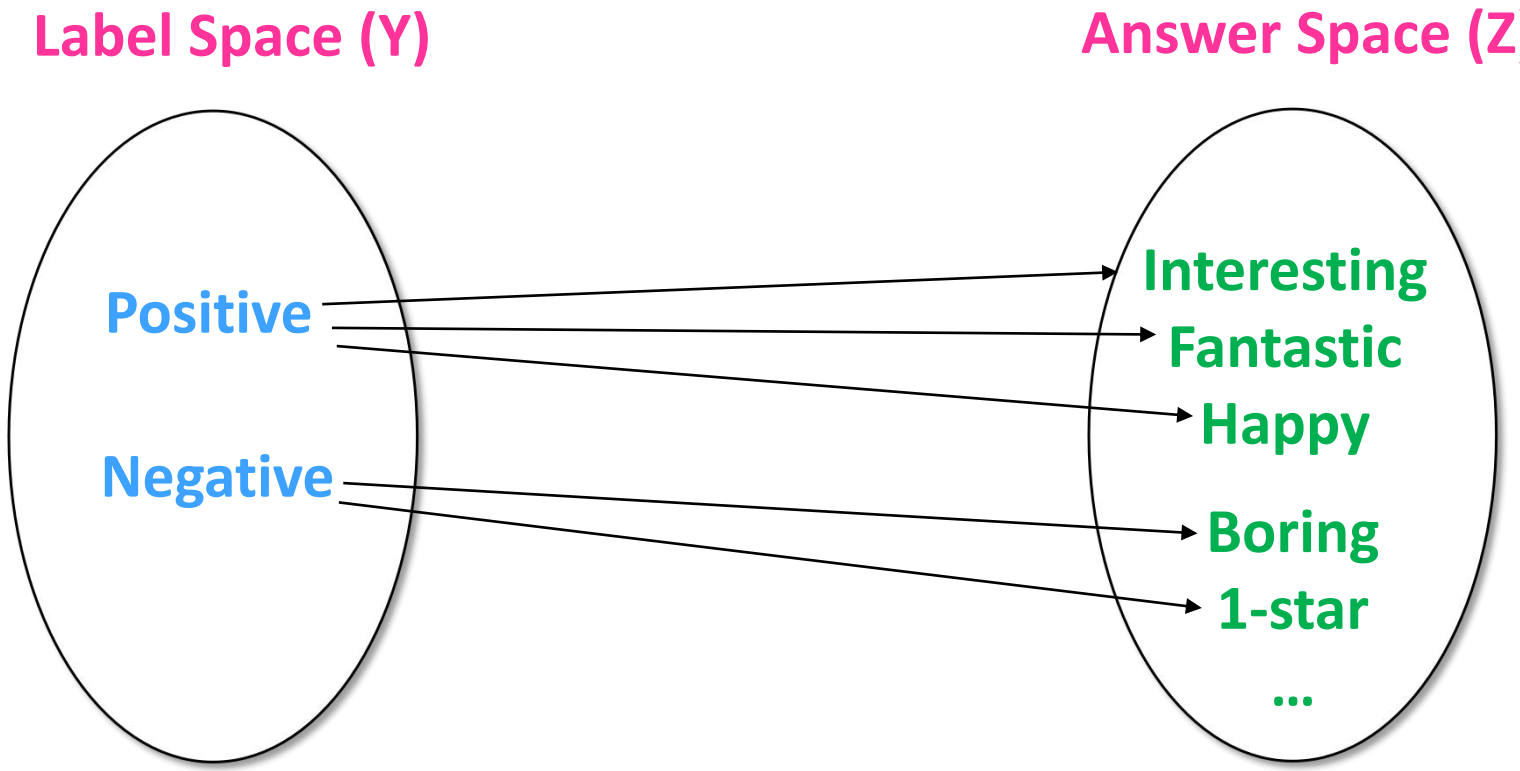
- Prompt Template Engineering
- **Answer Engineering**
- Pre-trained Model Choice
- Expanding the Paradigm
- Prompt-based Training Strategies

Answer Engineering

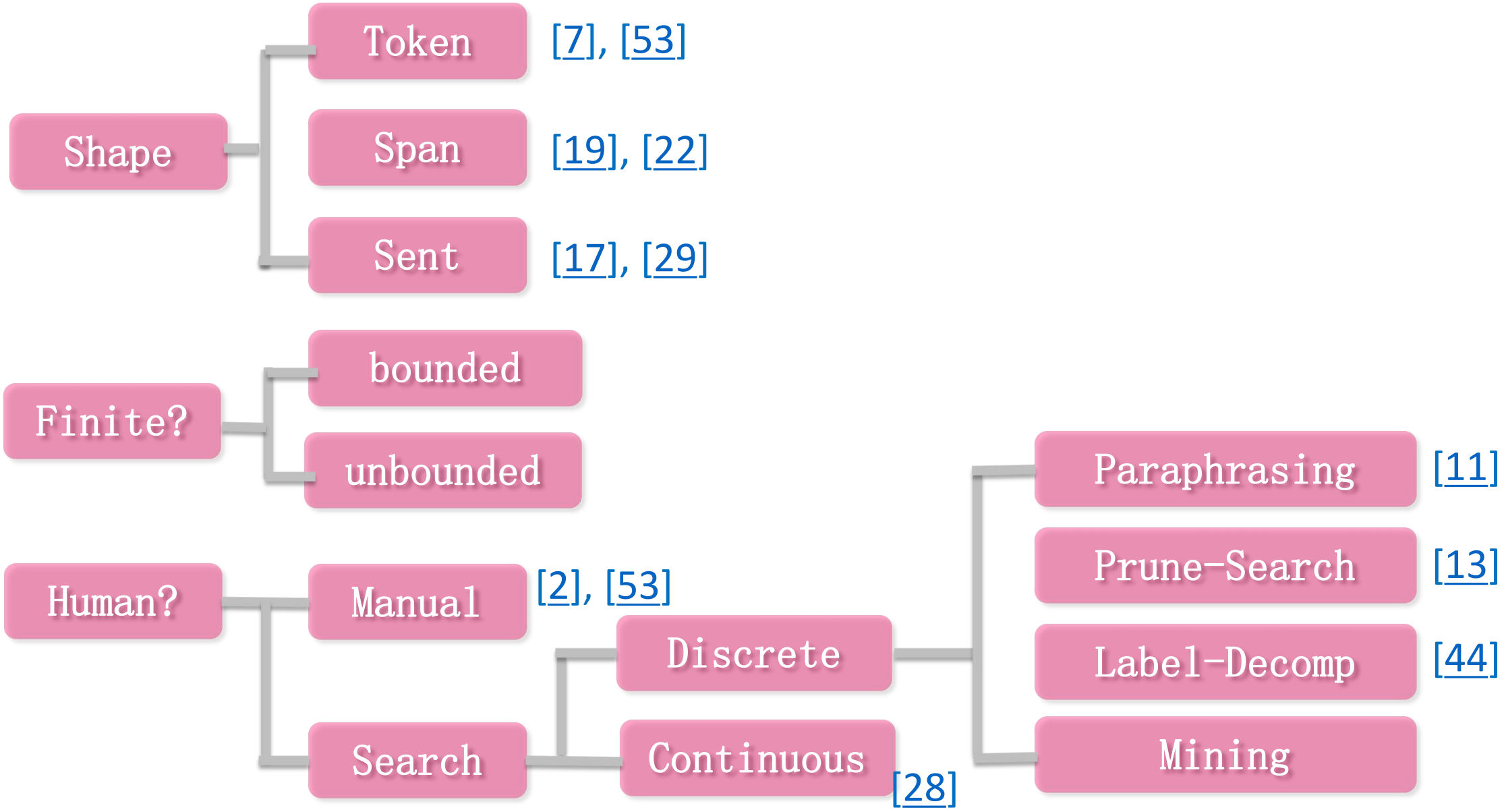
- Research Question:
  - Given a task (or a prompt), **how to define a suitable mapping function between label space and answer space?**

Answer Engineering

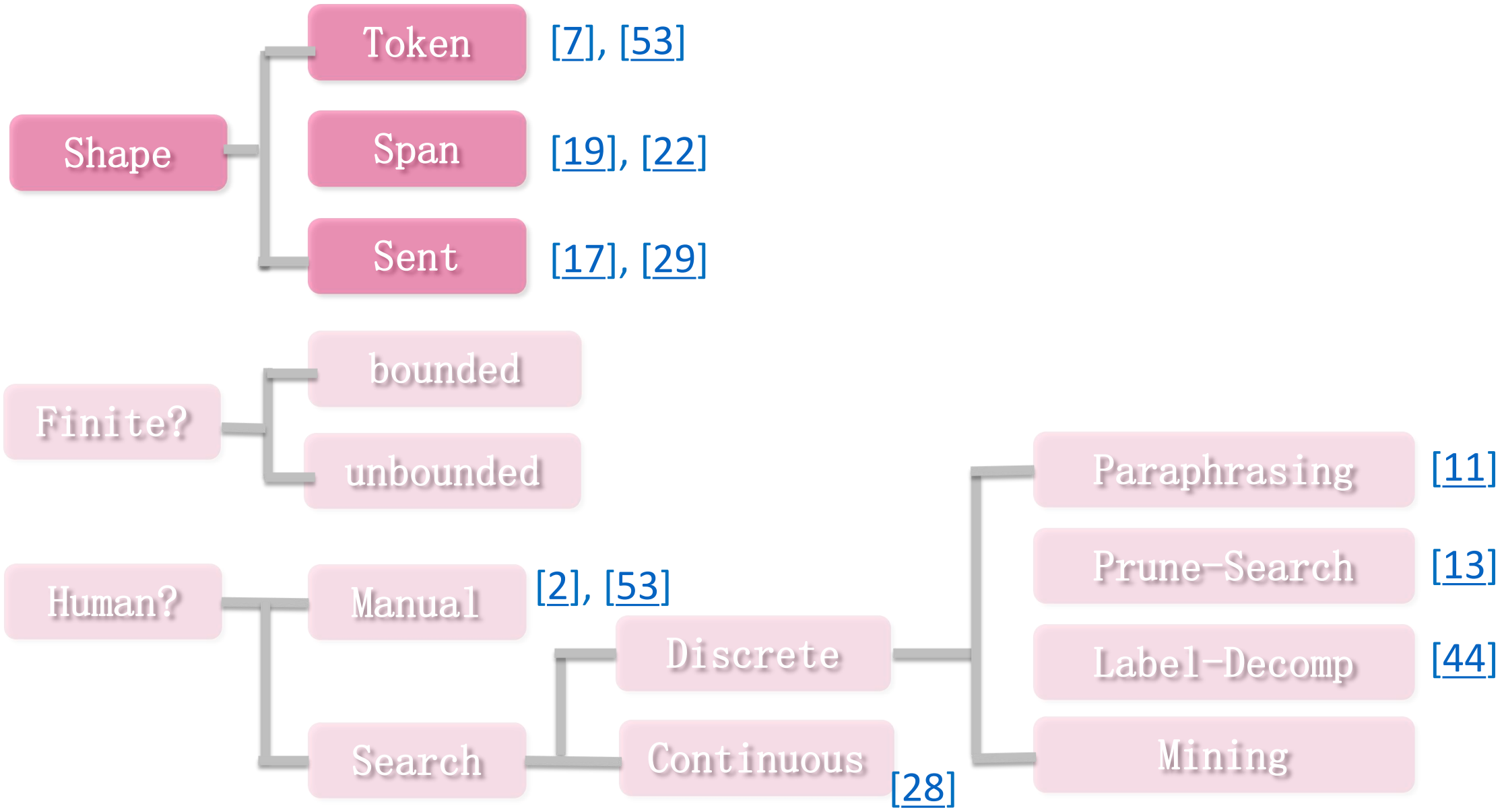
- Research Question:
  - Given a task (or a prompt), **how to define a suitable mapping function between label space and answer space?**



Design Decision of Prompt Answer Engineering



Design Decision of Prompt Answer Engineering



Design Considerations for Prompt-based Methods

- Token
  - Useful for most classification tasks
  - Examples
    - <A movie review> The movie is fantastic/terrible.
    - <Premise> Yes/No. <Hypothesis>

Design Considerations for Prompt-based Methods

- Token
- Span
  - Useful for classification with long label names, QA, knowledge probing, etc.
  - Example
    - Multiple choice QA

      A student riding a bicycle observes that it moves faster on a smooth road than on a rough road. This happens because the smooth road has

      (A) less gravity

      (B) more gravity

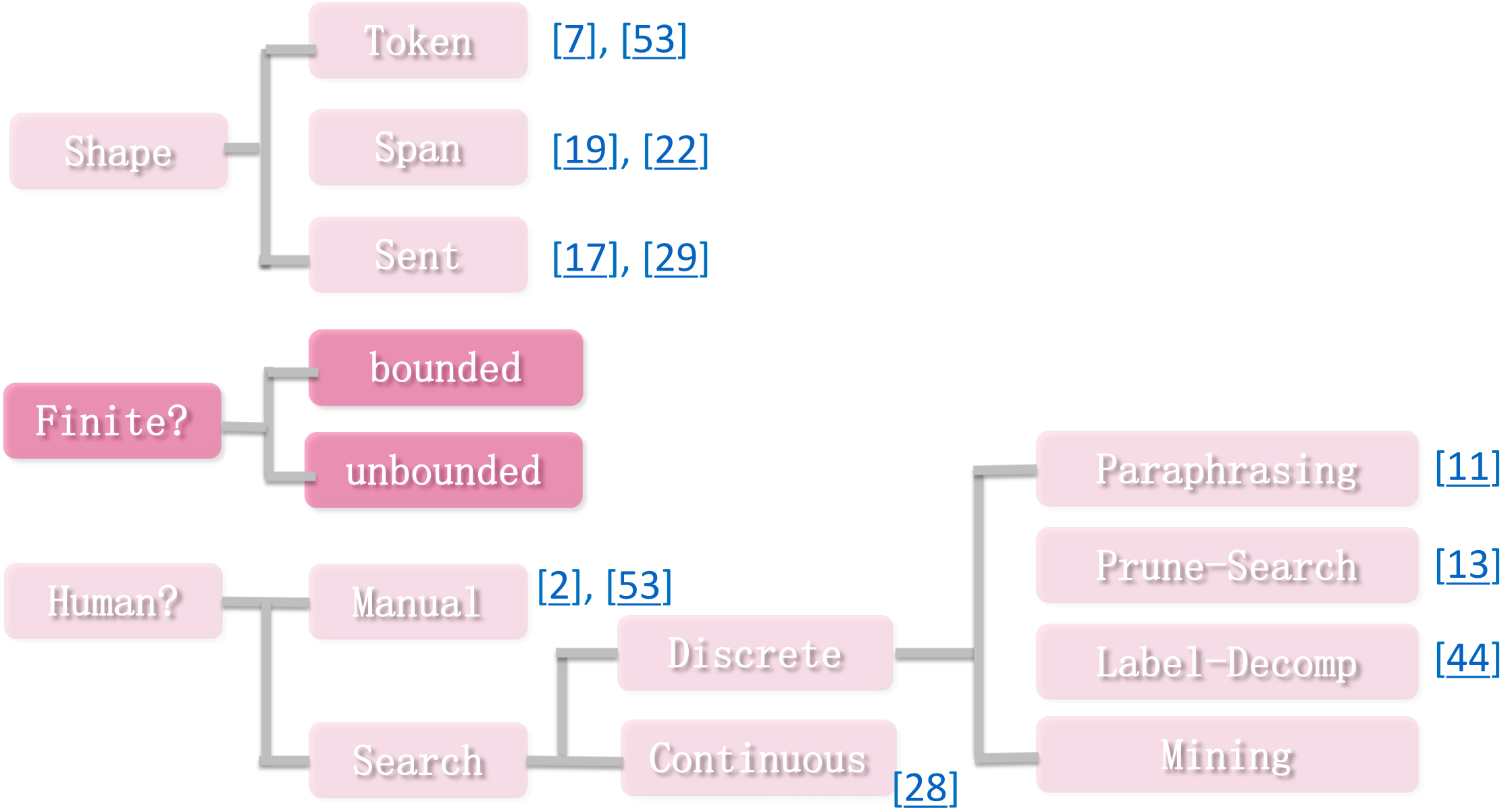
      (C) less friction [gold]

      (D) more friction

Design Considerations for Prompt-based Methods

- Token
- Span
- Sentence(s)
  - Useful for generation tasks, like MT or summarization.
  - Example
    - Translation from English to Chinese
      Input: Hello, world!
      Target (gold answer): 你好，世界！

Design Decision of Prompt Answer Engineering



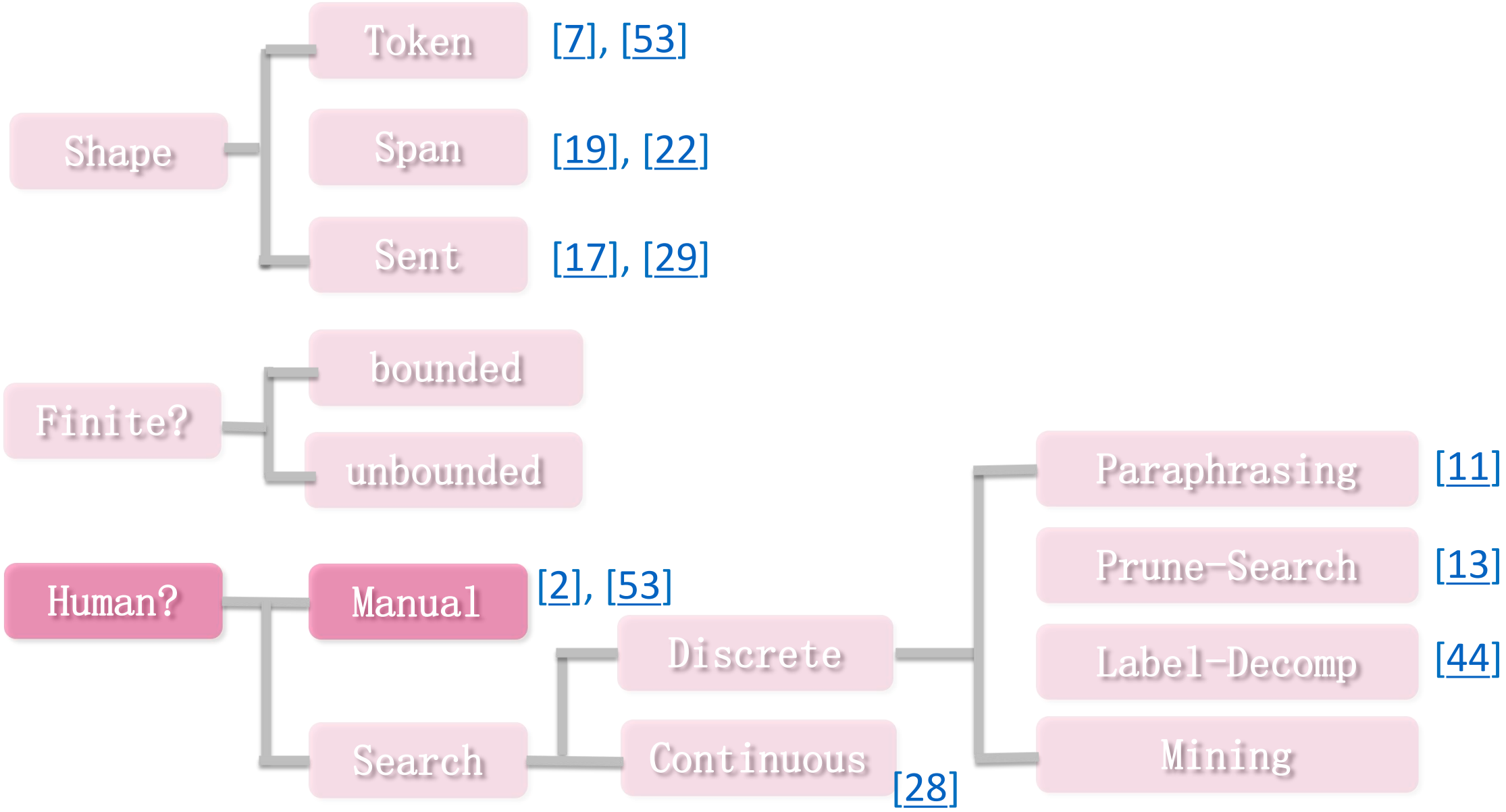
Answer Space

- Bounded
  - The space of possible outputs is constrained/finite.
  - Example
    - Text classification: health; finance; politics, sports.

Answer Space

- Bounded
  - The space of possible outputs is constrained/finite.
  - Example
    - Text classification: health; finance; politics, sports.
- Unbounded
  - The space of possible outputs is unconstrained/infinite.
  - Example
    - Text summarization: all valid sequence of tokens.

Design Decision of Prompt Answer Engineering



Human Design

- The most natural way to create answers 😄
  - For generation tasks, we can use identity mapping to map target output directly to gold answer
    - In MT/Summarization, take the target directly as gold answer

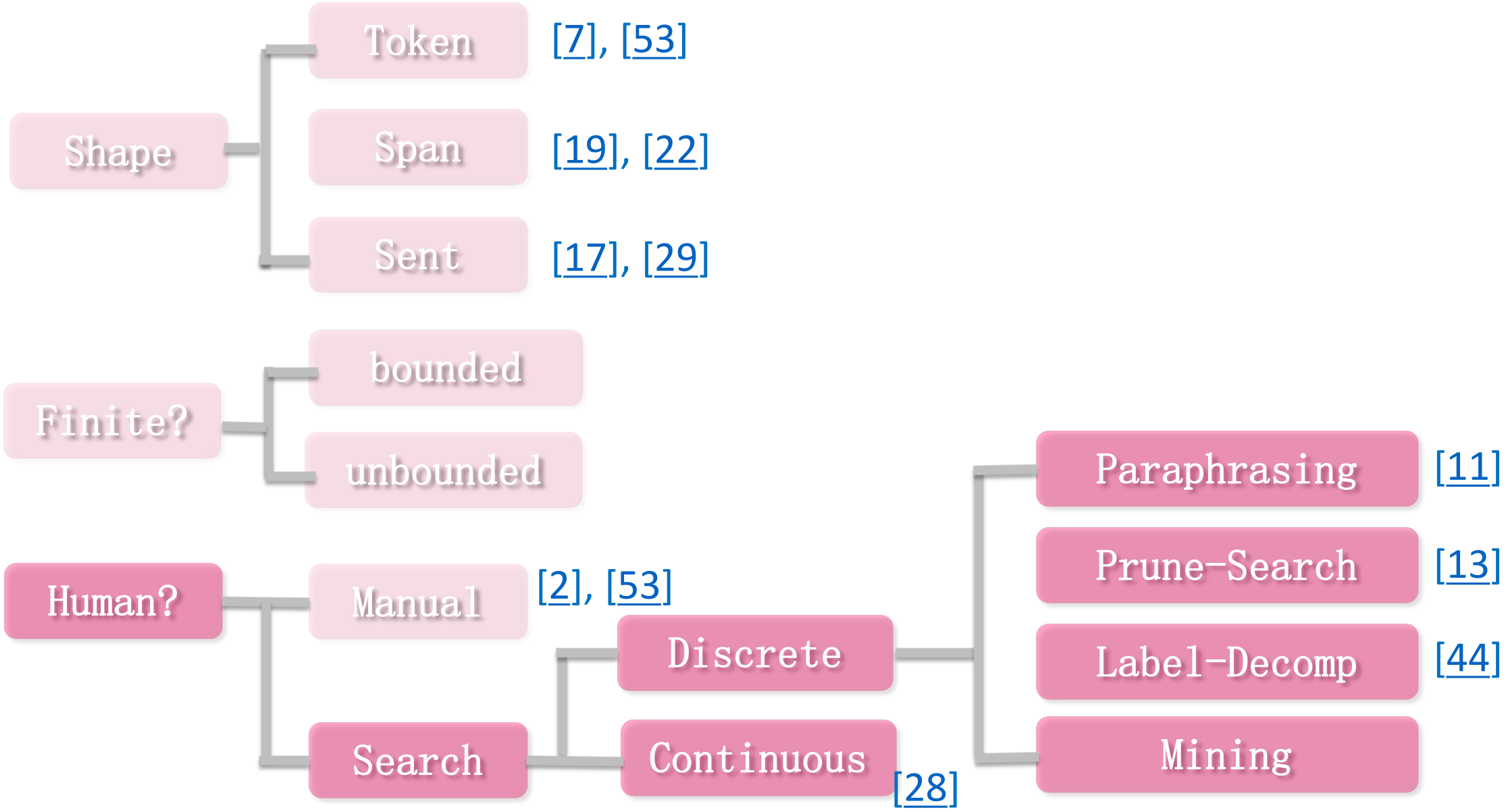
Human Design

- The most natural way to create answers 😄
  - For generation tasks, we can use identity mapping to map target output directly to gold answer
    - In MT/Summarization, take the target directly as gold answer
  - For classification tasks, the label name can also act as gold answer.
    - For example, sports, politics

Human Design

- The most natural way to create answers 😄
  - For generation tasks, we can use identity mapping to map target output directly to gold answer
    - In MT/Summarization, take the target directly as gold answer
  - For classification tasks, the label name can also act as gold answer.
    - For example, sports, politics
- An art that takes time and experience. 😔
  - For some complicated tasks, it's hard to manually craft answers.
    - For example, relation classification

Design Decision of Prompt Answer Engineering



Discrete Answer Search

- Paraphrasing
- Prune then Search
- Label Decomposition
- Mining

Discrete Answer Search

- Paraphrasing
  - Start with an initial answer space, and then use paraphrasing to expand this answer space to broaden its coverage.
  - Example
    - Multiple Choice QA

      A person wants to submerge himself in water, what should he use?

      (A) Whirl pool (Paraphrase to get Bathtub, A bathtub etc.)

      (B) …

Reference: Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2020. How Can We Know When Language Models Know? CoRR abs/2012.00955 (2020).

Discrete Answer Search

- Prune then Search
  - Pruning methods:
    - Select the most frequent words
    - Select tokens that have highest generation probability at answer position

References:
[1] Taylor Shin, Yasaman Razeghi, Robert L. LoganIV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Empirical Methods in Natural Language Processing (EMNLP).
[2] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In Association for Computational Linguistics (ACL).

Discrete Answer Search

- Prune then Search
  - Pruning methods:
    - Select the most frequent words
    - Select tokens that have highest generation probability at answer position
  - Searching methods:
    - Choose answers that maximize the likelihood of training data
    - Choose answers that achieve the best zero-shot accuracy

References:
[1] Taylor Shin, Yasaman Razeghi, Robert L. LoganIV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Empirical Methods in Natural Language Processing (EMNLP).
[2] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In Association for Computational Linguistics (ACL).

Discrete Answer Search

- Label Decomposition
  - For complex label, decompose the label into its constituent words.
  - Example
    - Text classification:

      Science and Mathematics —decompose→ {Science, Mathematics}

    - Relation Extraction:

      city_of_death —decompose→ {person, city, death}

Reference: Xiang Chen, Xin Xie, Ningyu Zhang, Jiahuan Yan, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. AdaPrompt: Adaptive Prompt-based Finetuning for Relation Extraction. CoRR abs/2104.07650 (2021) .

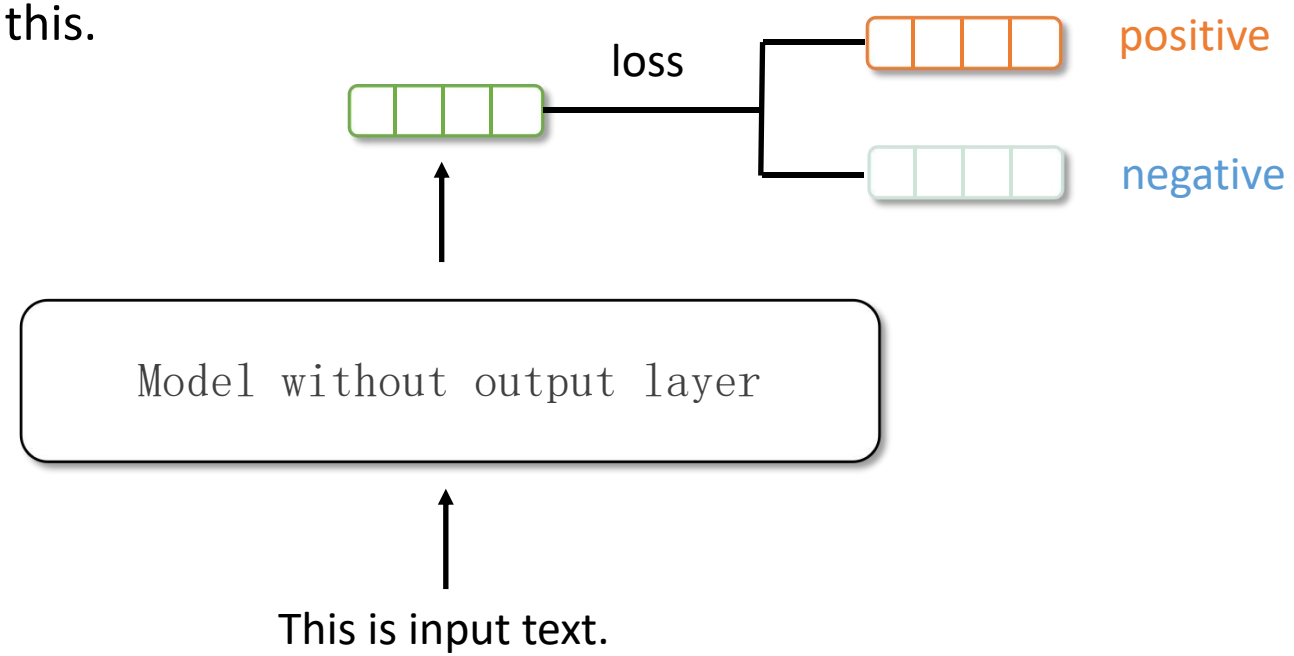
Discrete Answer Search

- Mining
  - Given a seed answer, use some knowledge base to retrieve related words.
  - Example: “city”
    - metropolis town
    - urban
    - suburb
    - municipal
    - downtown
    - Country
    - ……

Reference: Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li and Hong-Gee Kim. 2021. Prompt-Learning for Fine-Grained Entity Typing. CoRR abs/2108.10604 (2021) .

Continuous Answer Search

- Definition
  - Assign a virtual token for each class label so that the class labels can be optimized through gradient descent.
  - Very few works do this.



Reference: Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. ArXiv abs/2101.00121 (2021).

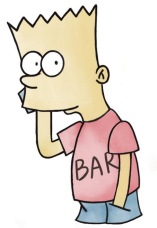
Design Considerations for Prompt-based Methods

- Prompt Template Engineering
- Answer Engineering
- **Pre-trained Model Choice**
- Expanding the Paradigm
- Prompt-based Training Strategies

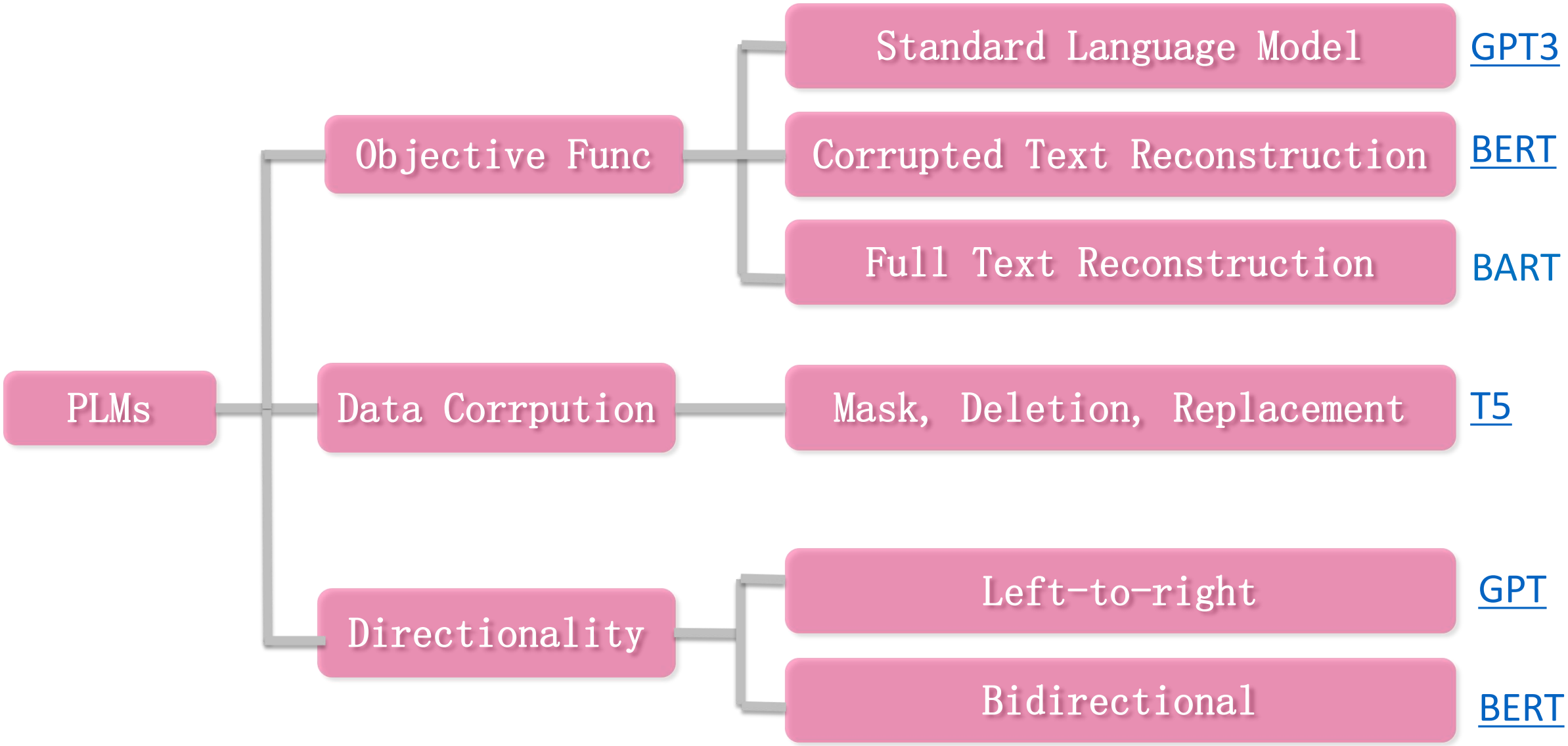
Pre-trained Model Choice

- **Research Question**:
  - Given a task (or a prompt), **which pre-trained language model would be the most appropriate one?**

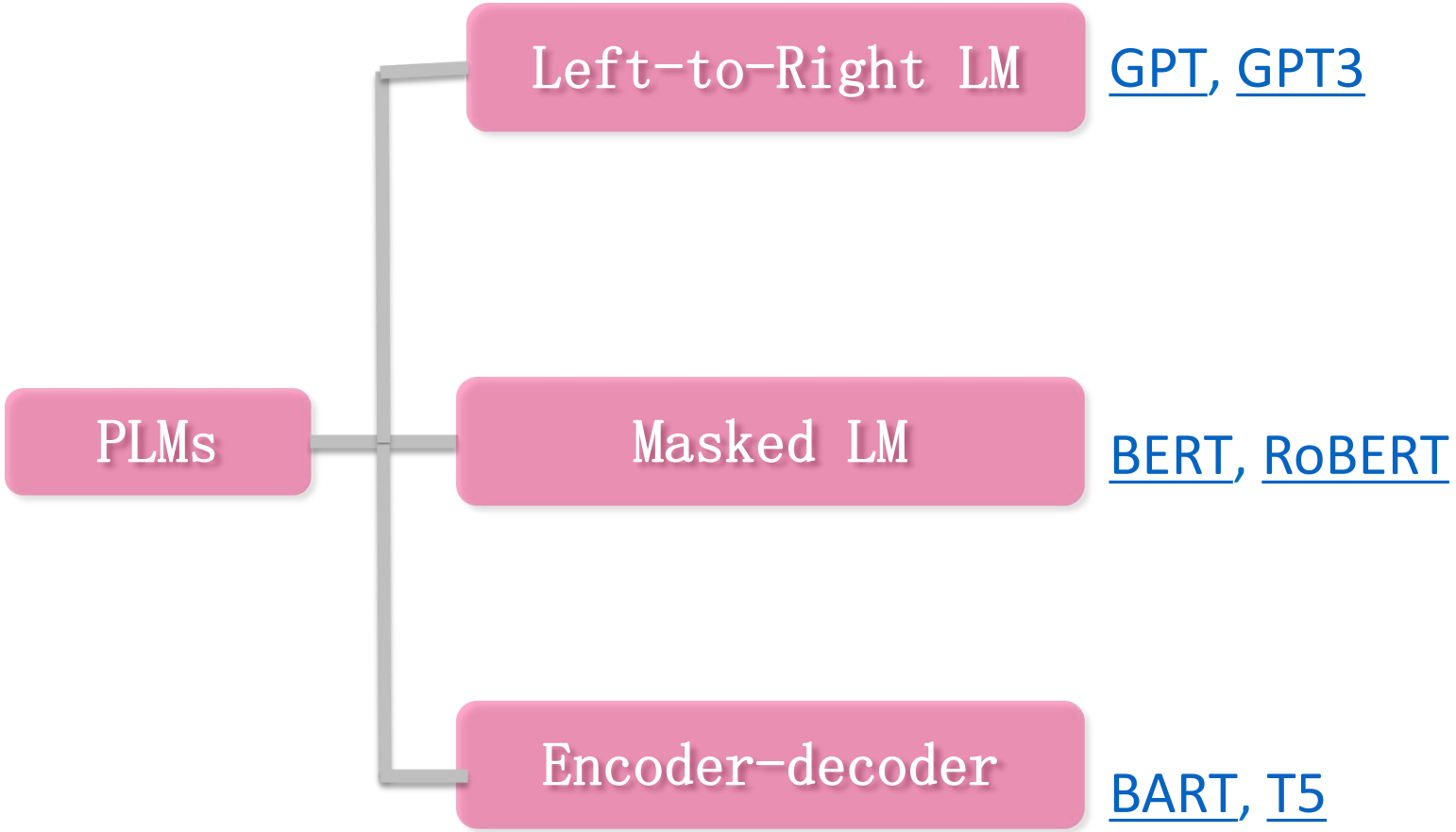
The story describes …., in summary [z]



Design Decision of Pre-trained Models

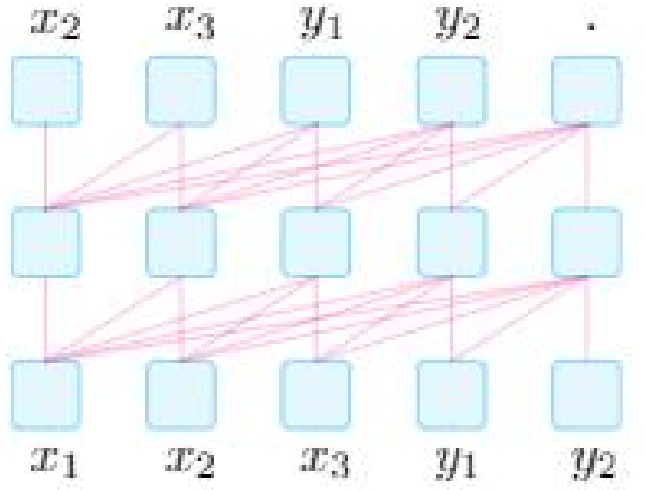


Design Decision of Pre-trained Models



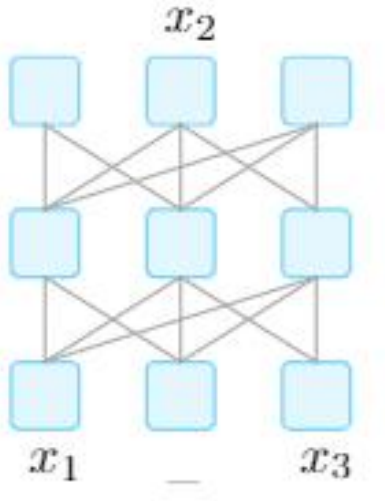
Left-to-right Language Model

- **Characteristics**
  - First proposed by Markov (1913)
  - Count-based-> Neural network-based
  - Specifically suitable to highly larger-scale LMs
- **Example**
  - GPT-1,GPT-2,GPT-3
- **Roles in Prompting Methods**
  - The earliest architecture chosen for prompting
  - Usually equipped with prefix prompt and the parameters of PLMs are fixed



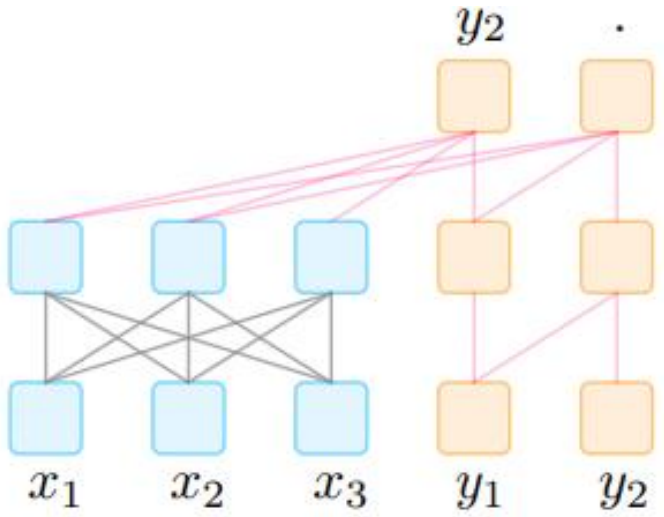
Masked Language Model

- **Characteristics**
  - An extension of left-to-right architecture
  - Unidirection -> bidirection prediction
  - Suitable for NLU tasks
- **Example**
  - BERT, ERNIE
- **Roles in Prompting Methods**
  - Usually combined with cloze prompt
  - Suitable for NLU tasks



Encoder-Decoder

- **Characteristics**
  - A denoised auto-encoder
  - Use two Transformers and two different mask mechanisms to handle text X and Y separately
- **Examples**
  - BART, T5
- **Roles in Prompting methods**
  - Text generation tasks or some tasks that can be formulated into a text generation problem



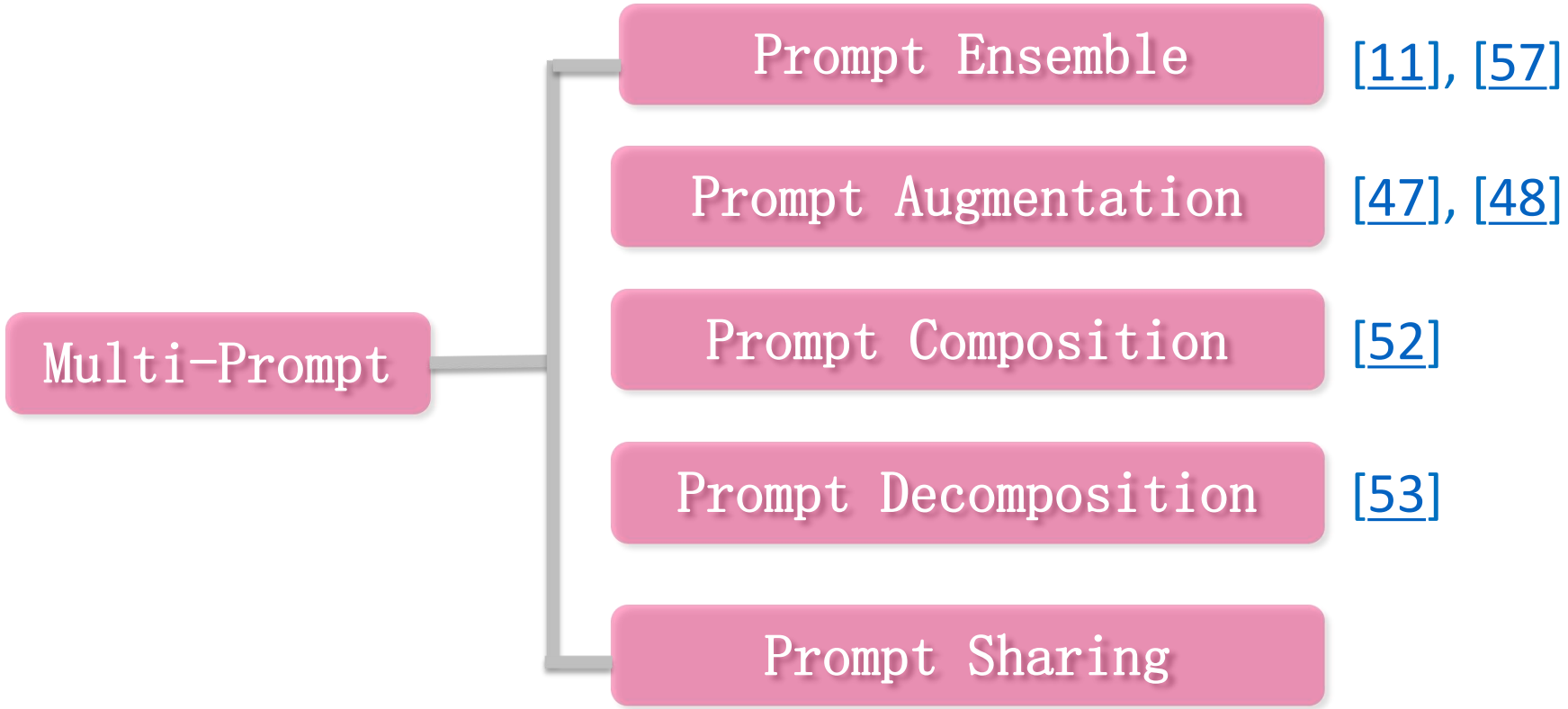
Design Considerations for Prompt-based Methods

- Prompt Template Engineering
- Answer Engineering
- Pre-trained Model Choice
- **Expanding the Paradigm**
- Prompt-based Training Strategies

Expanding the Paradigm

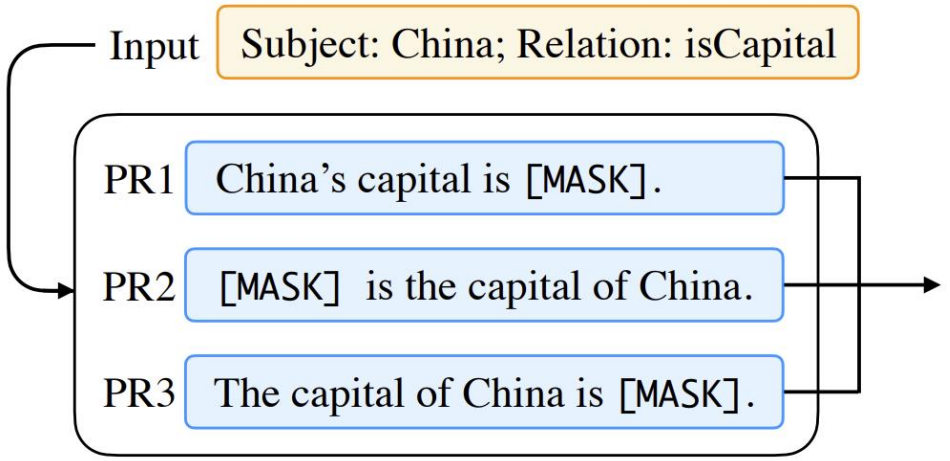
- Research Questions
  - How to extend the current prompting framework to support more NLP tasks?

Design Decision of Multiple Prompt Learning



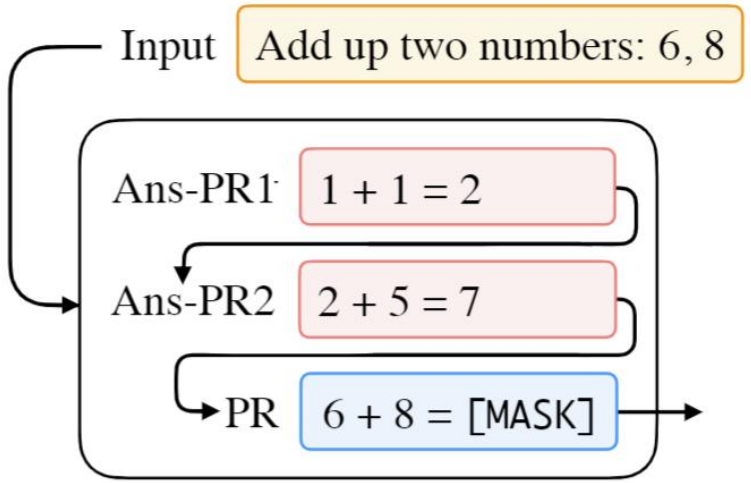
Prompt Ensembling

- Definition
  - using multiple unanswered prompts for an input at inference time to make predictions
- Advantages
  - Utilize complementary advantages
  - Alleviate the cost of prompt engineering
  - Stabilize performance on downstream tasks



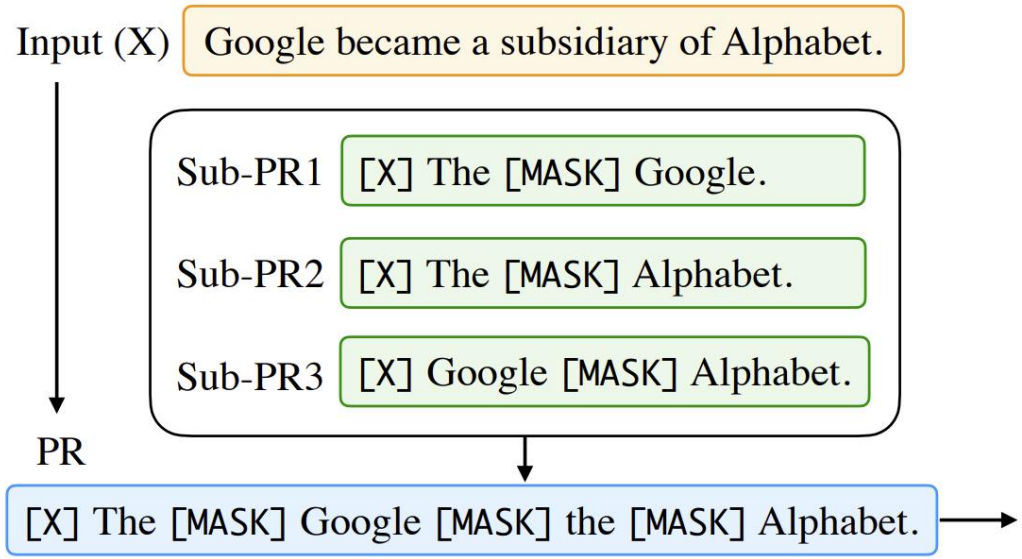
Prompt Augmentation

- Definition
  - Help the model answer the prompt with additional answered prompts
- Advantage
  - make use of the small amount of information that has been annotated
- Core step
  - Selection of answered prompts
  - Ordering of answered prompts



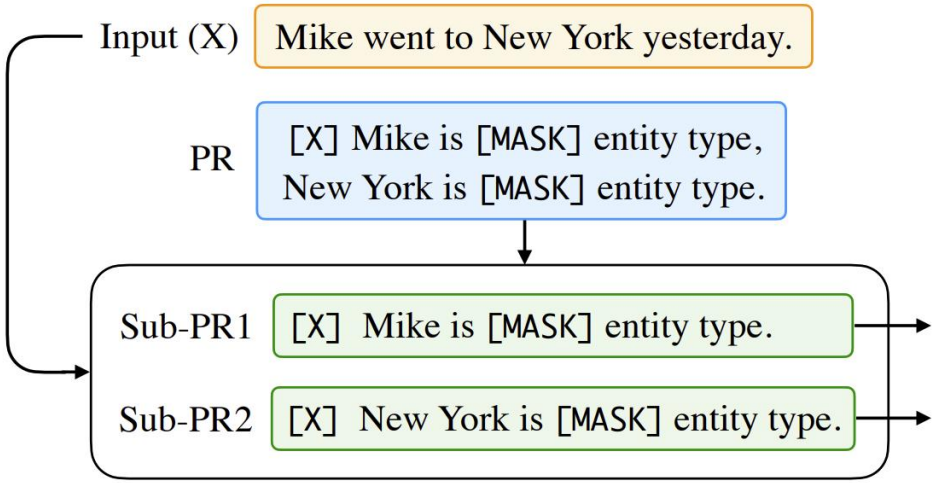
Prompt Composition

- Definition
  - Prompts for a composable task can be designed with multiple sub-prompts, which can then be combined to complete the task
- Advantage
  - It provides a method of prompt learning for complex tasks



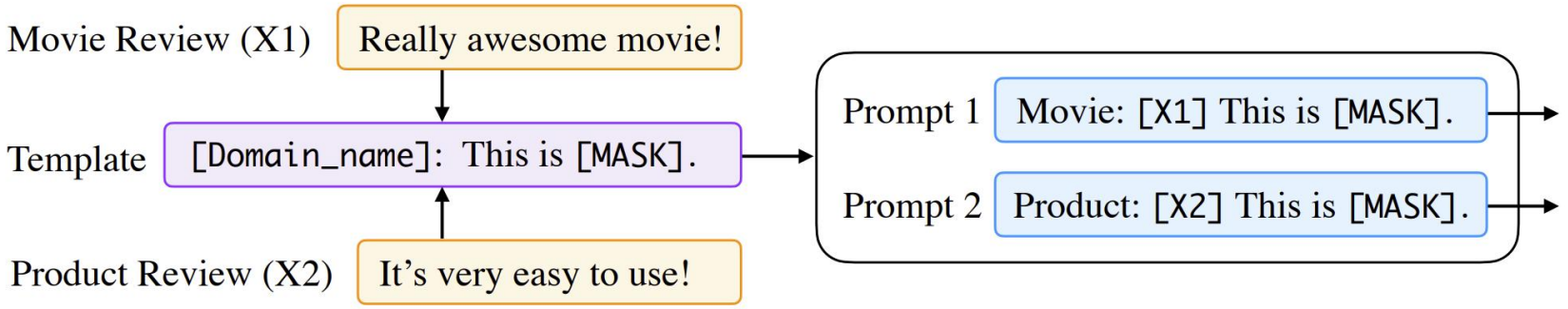
Prompt Decomposition

- Definition
  - For tasks where multiple predictions should be performed for one sample, handle it individually
- Advantages
  - Break-down a complicated task into multiple separate ones



Prompt Sharing

- Definition
  - When prompting method is applied to multiple tasks, domains or languages , prompts can be shared cross different tasks.
- Advantage
  - Task- or language invariant information can be captured through prompting.



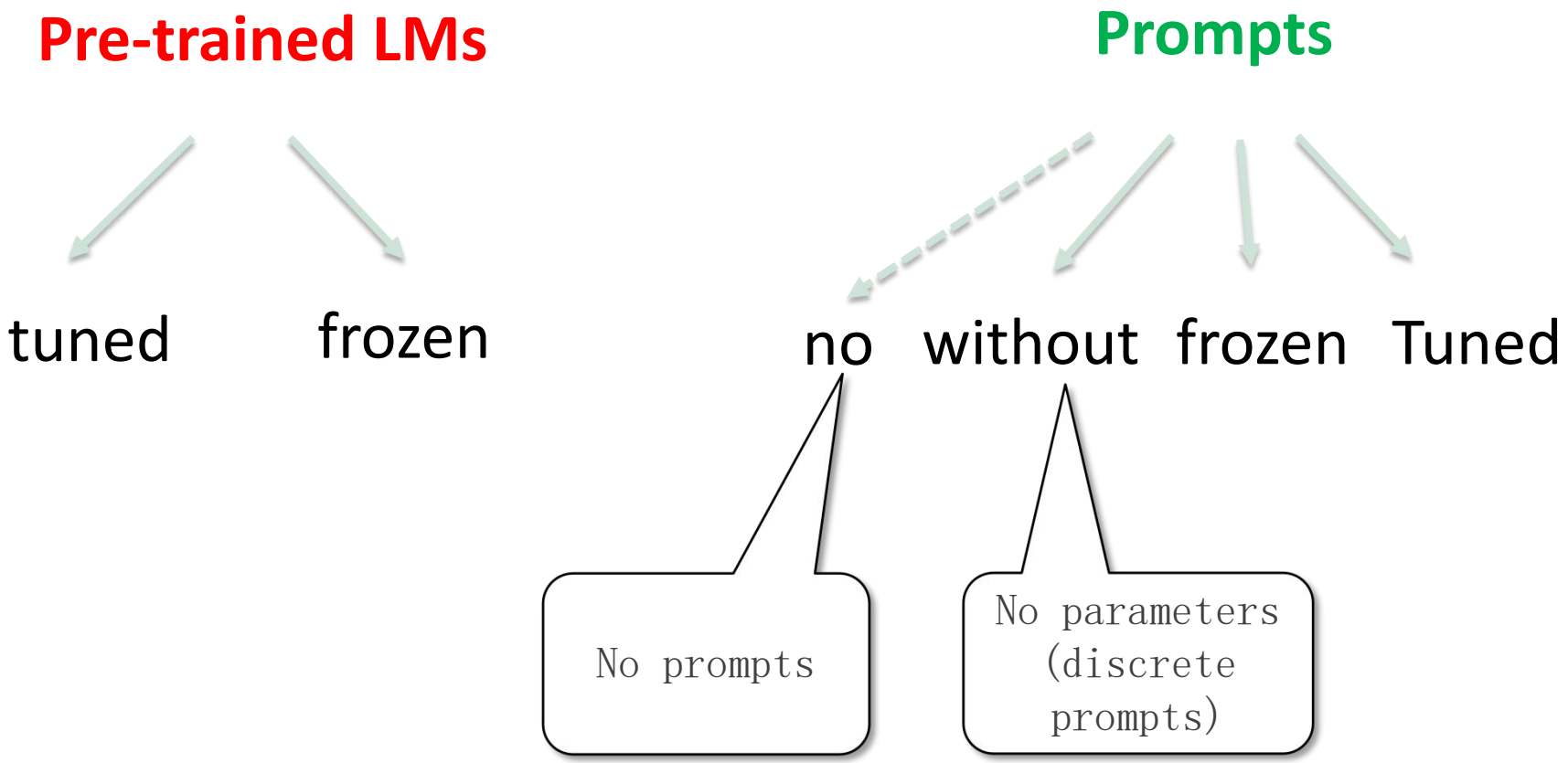
Design Considerations for Prompt-based Methods

- Prompt Template Engineering
- Answer Engineering
- Pre-trained Model Choice
- Expanding the Paradigm
- **Prompt-based Training Strategies**

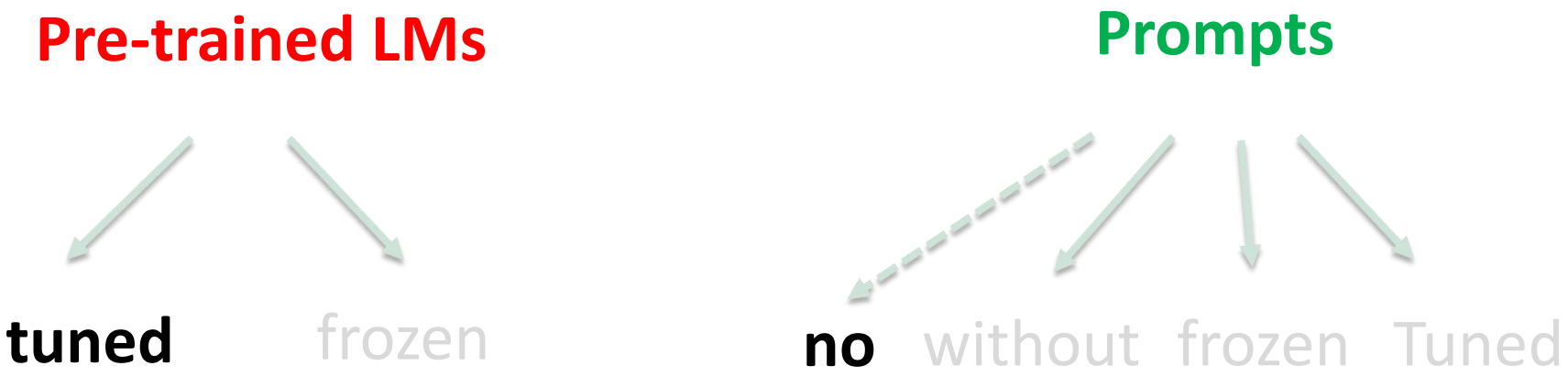
Prompt-based Training Strategies

- Data Perspective
  - Zero-shot: without any explicit training of the LM for the down-stream task
  - Few-shot: few training (e.g., 100) samples of downstream tasks
  - Full-data: lots of training samples (e.g., 10K) of downstream tasks

Parameter Perspective



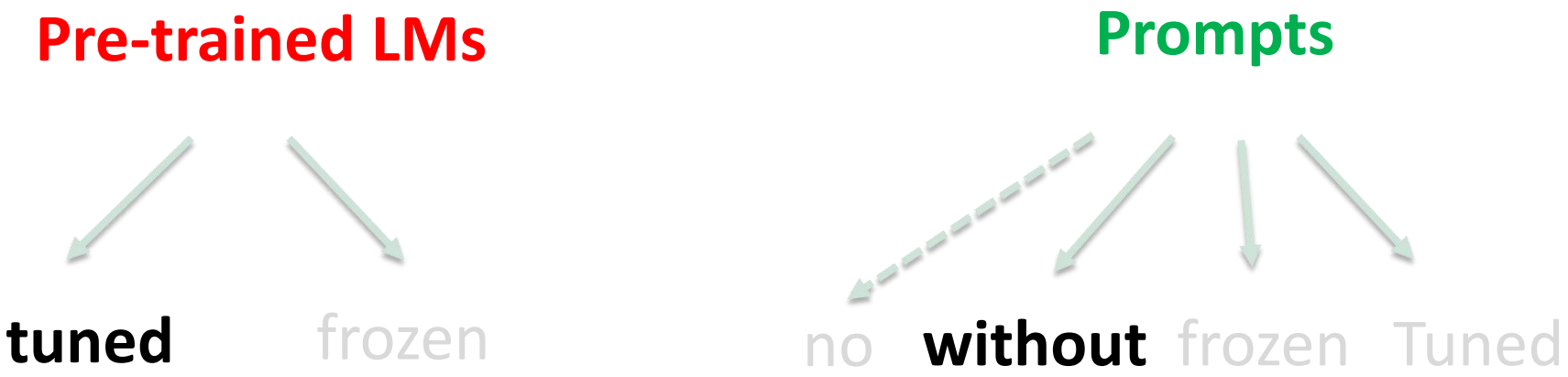
Cases of Parameter Updating



Promptless Fine-tuning

Example: BERT for text classification

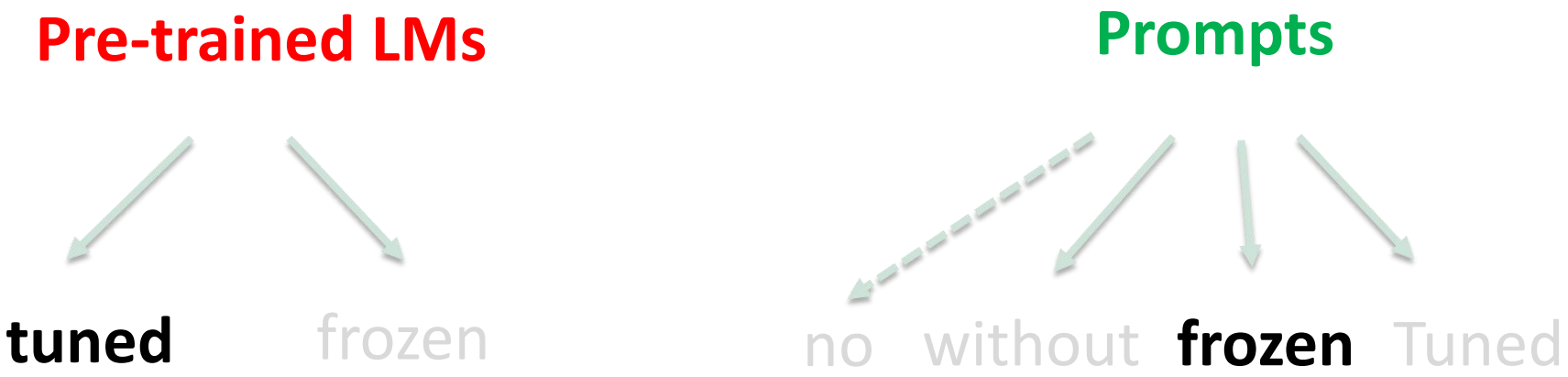
Cases of Parameter Updating



Fixed-prompt Tuning

Example: BERT + Discrete Prompt for text classification

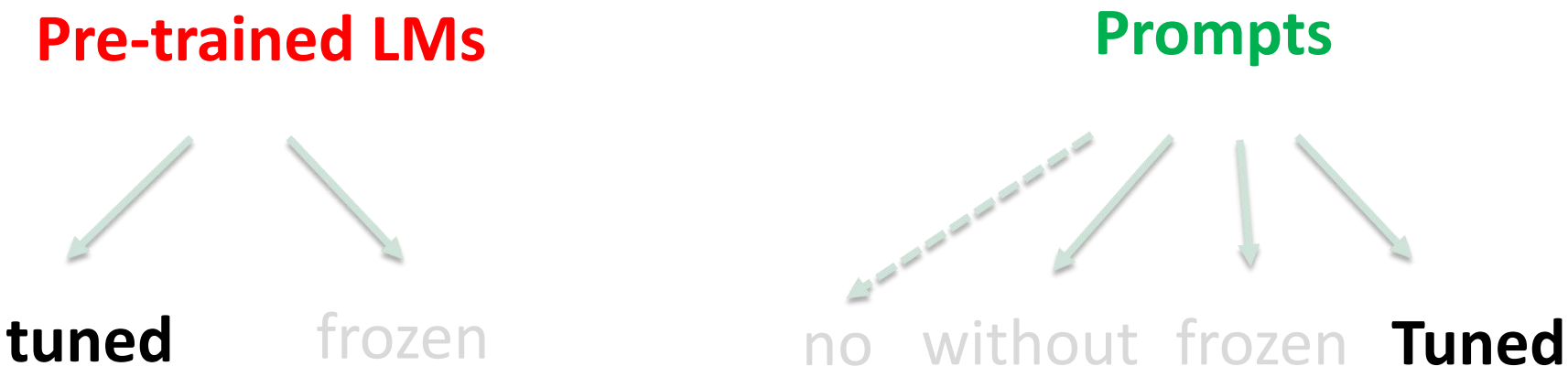
Cases of Parameter Updating



Fixed-prompt Tuning

Example: BERT + Transferred Continuous Prompt for text classification

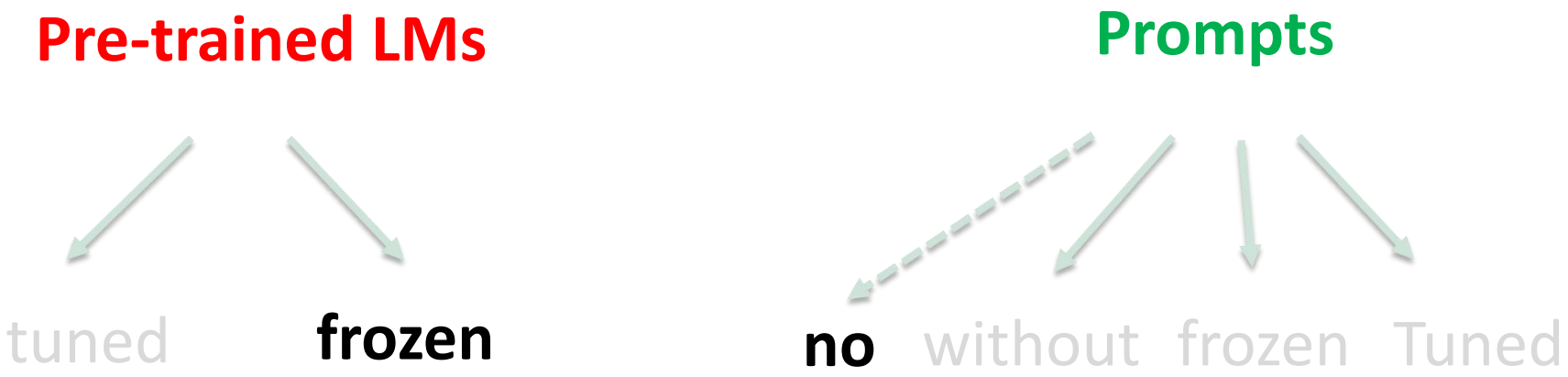
Cases of Parameter Updating



Prompt+LM Fine-tuning

Example: BERT + Continuous Prompt for text classification

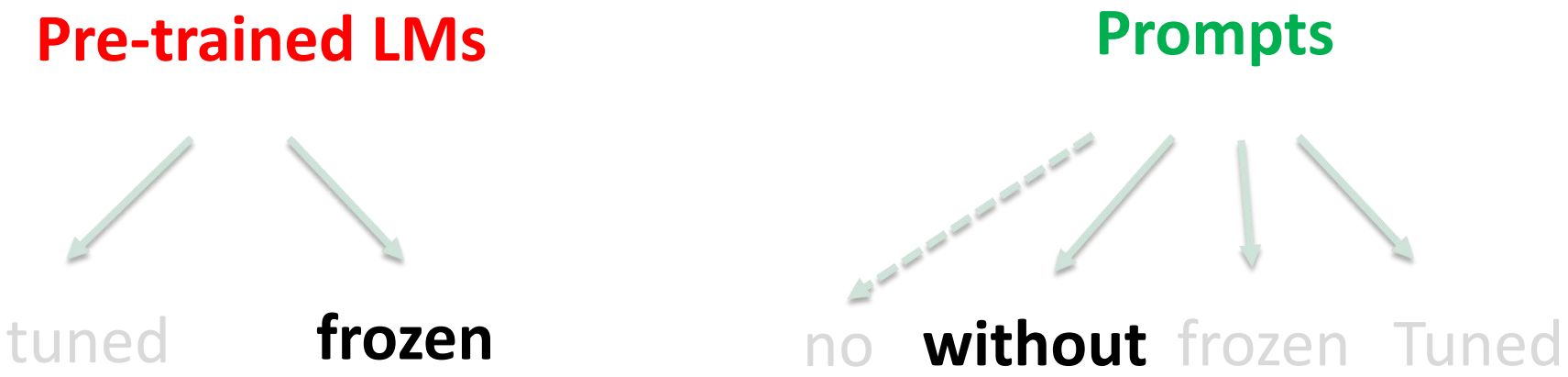
Cases of Parameter Updating



Adapter Tuning

Example: BERT + Adapter for text classification

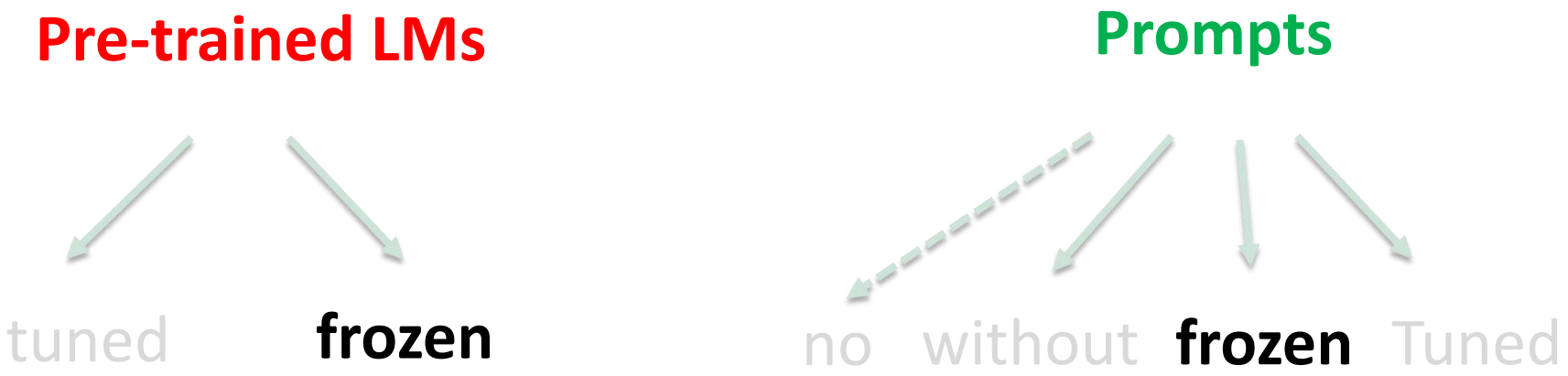
Cases of Parameter Updating



Tuning-free Prompting

Example: GPT3 + Discrete Prompts for Machine Translation

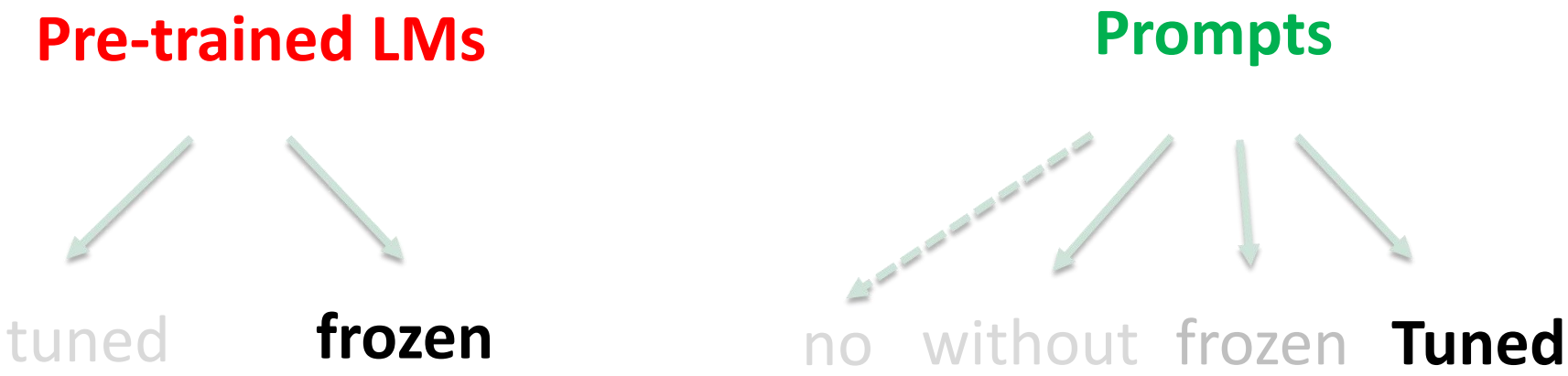
Cases of Parameter Updating



Tuning-free Prompting

Example: GPT3 + Continuous Prompts for Machine Translation

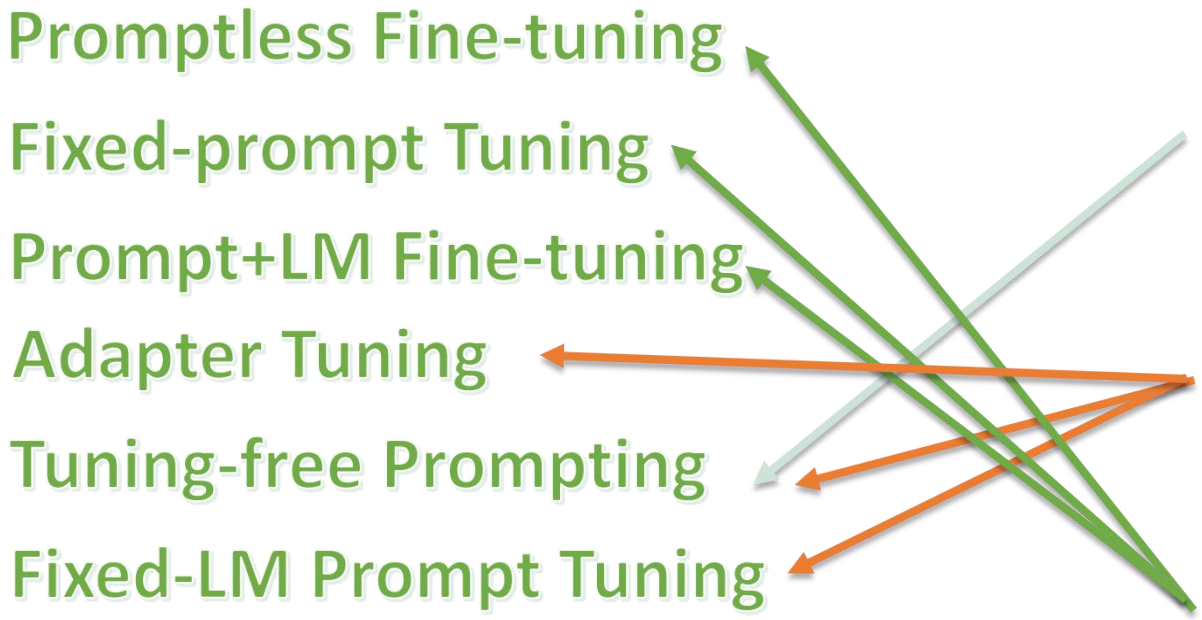
Cases of Parameter Updating



Fixed-LM Prompt Tuning

Example: BART + Continuous Prompts for Machine Translation

Too many, difficult to select?



If you have a highly large left-to-right pre-trained language model (e.g., GPT3)

If you have few training samples?

If you have lots of training samples?

What are things we can do now that we couldn‘t do in the past?

- PLM & Prompt Perspective

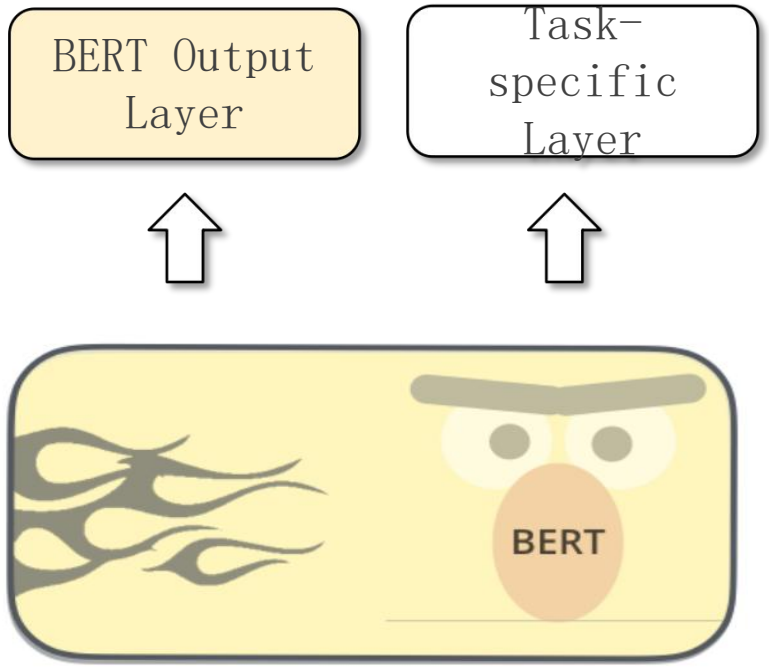
Significant change 2021 - 2022

What are things we can do now that we couldn‘t do in the past?

- PLM & Prompt Perspective

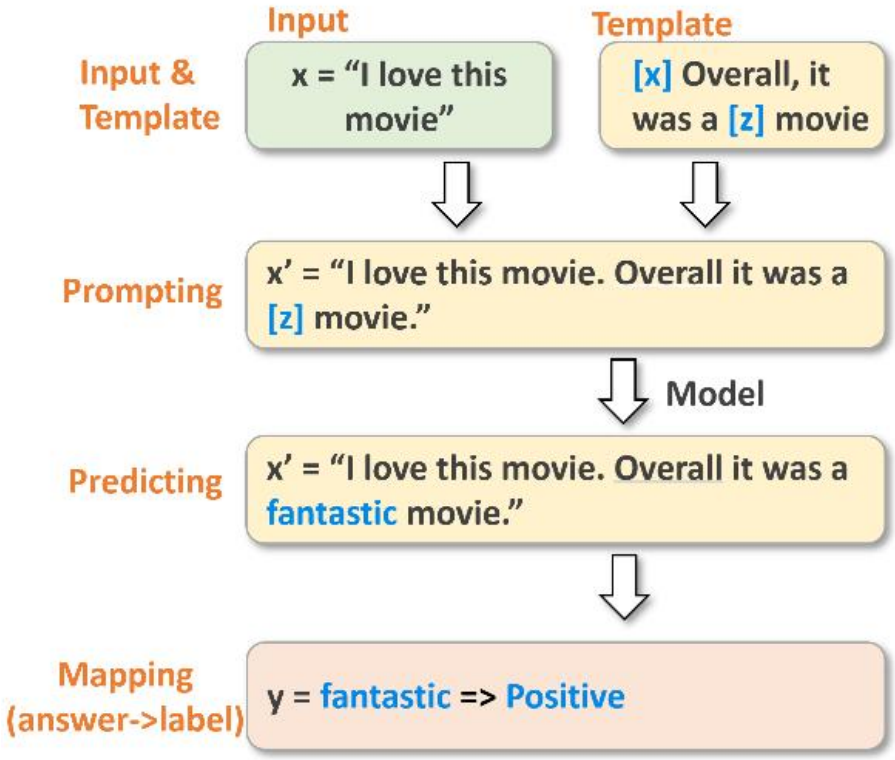
1. Paradigm Shift

- 1.1 All NLP tasks could be regarded as a “PLM-style ” task
  - Past
    - Classification
    - Sentence pair matching
    - QA
      - Extractive
      - Abstractive
      - Multiple choice
      - Yes/No
    - Generation



1. Paradigm Shift

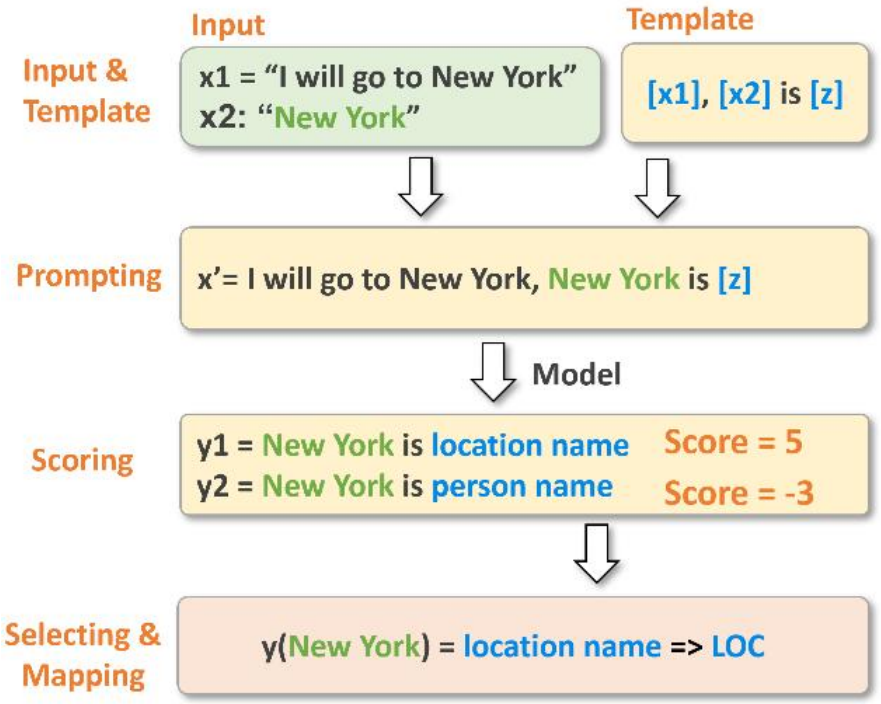
- 1.1 All NLP tasks could be regarded as a "PLM-style" task
  - Now
    - Advantages:
      - Make better use of knowledge stored in PLM
      - Less architecture engineering



Sentiment Classification

1. Paradigm Shift

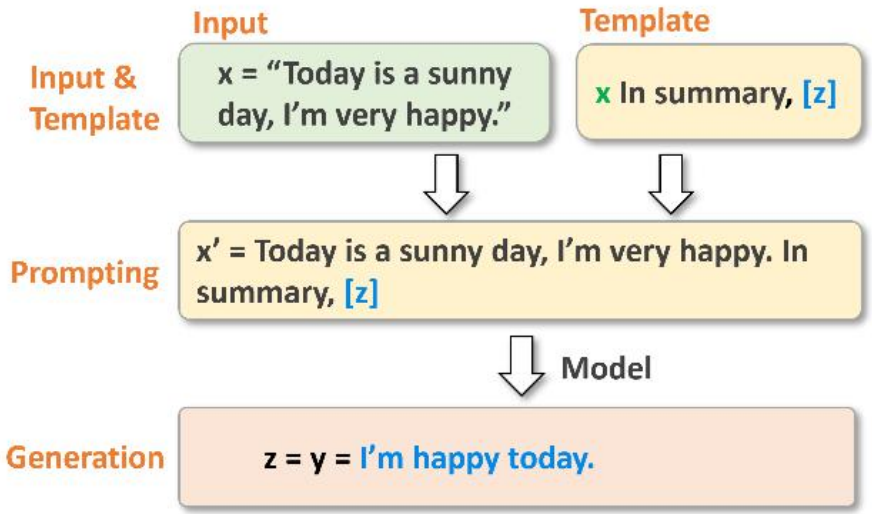
- 1.1 All NLP tasks could be regarded as a “PLM-style ” task
  - Now
    - Advantages:
      - Make better use of knowledge stored in PLM
      - Less architecture engineering



Named Entity Recognition

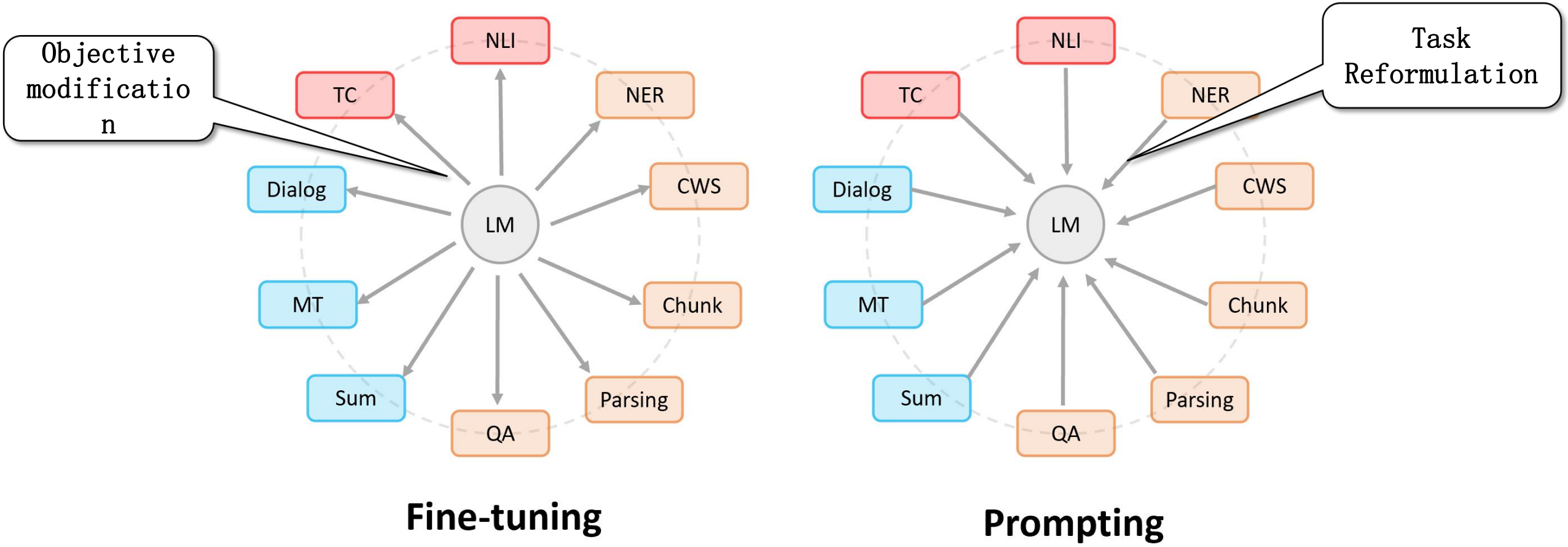
1. Paradigm Shift

- 1.1 All NLP tasks could be regarded as a “PLM-style ” task
  - Now
    - Advantages:
      - Make better use of knowledge stored in PLM
      - Less architecture engineering



Text Summarization

1. Paradigm Shift

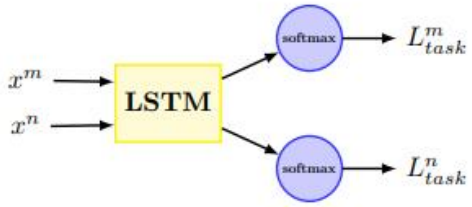


1. Paradigm Shift

- 1.2 More tasks could be modeled in a monolithic framework

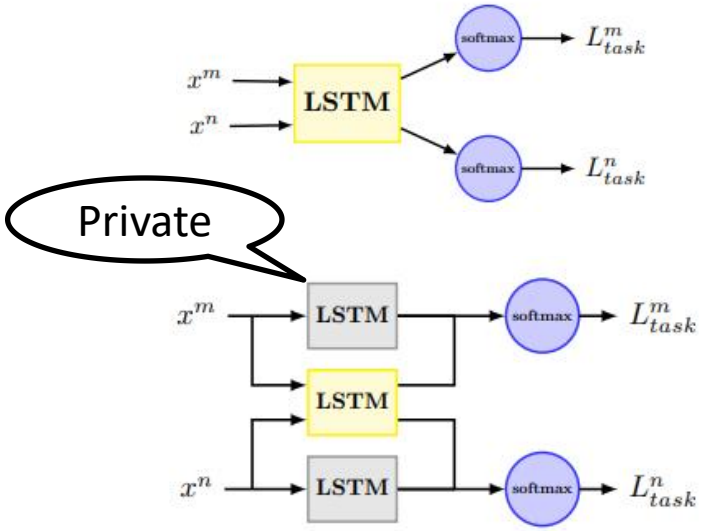
1. Paradigm Shift

- 1.2 More tasks could be modeled in a monolithic framework
  - Past



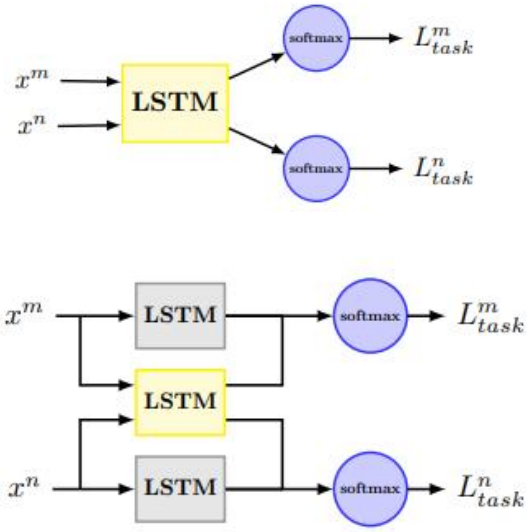
1. Paradigm Shift

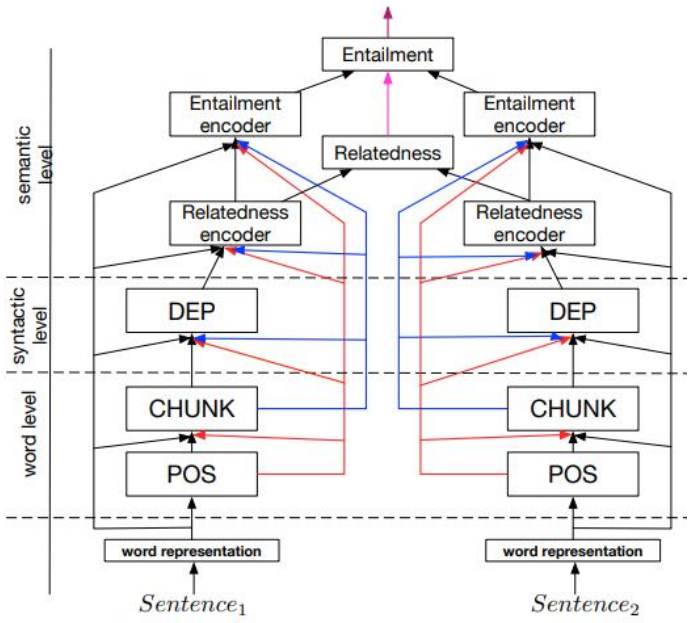
- 1.2 More tasks could be modeled in a monolithic framework
  - Past



1. Paradigm Shift

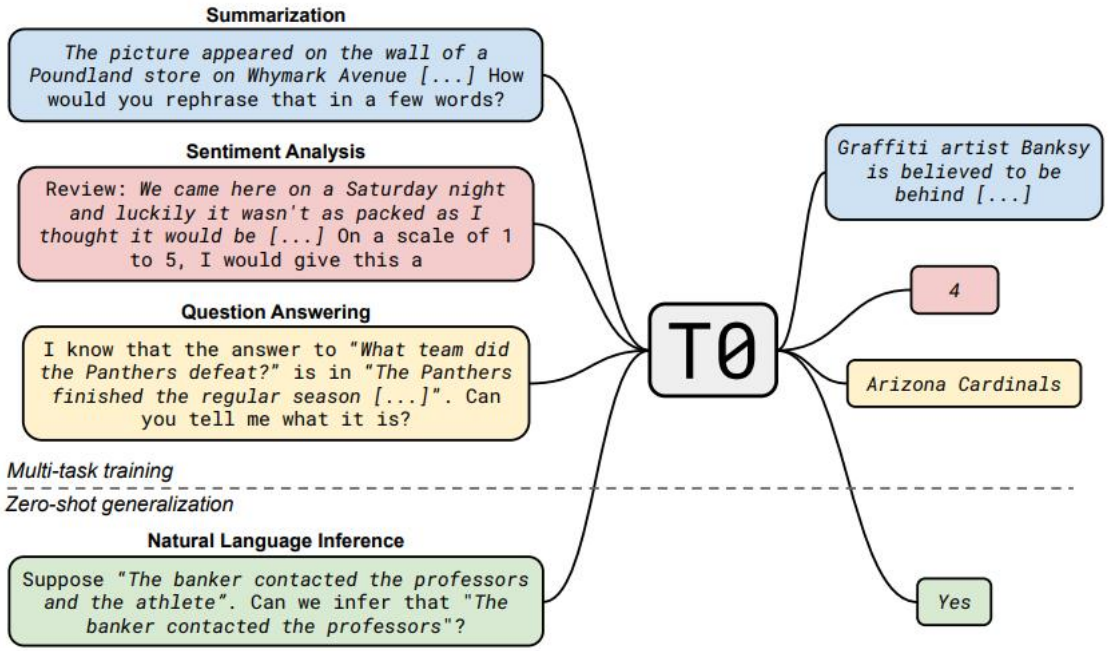
- 1.2 More tasks could be modeled in a monolithic framework
  - Past





1. Paradigm Shift

- 1.2 More tasks could be modeled in a monolithic framework
  - Now



1. Paradigm Shift

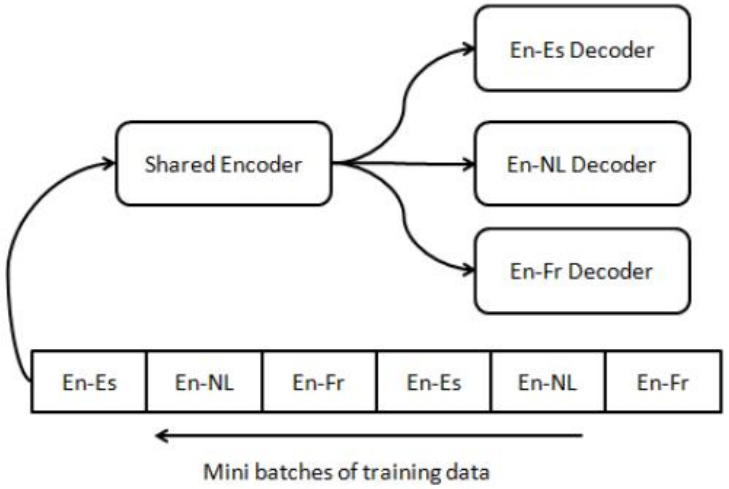
- 1.2 More tasks could be modeled in a monolithic framework
  - Now
    - Advantages
      - Get more shared information from other tasks
      - Less architecture mining
      - Reduce the deployment cost
      - Better zero-shot performance

1. Paradigm Shift

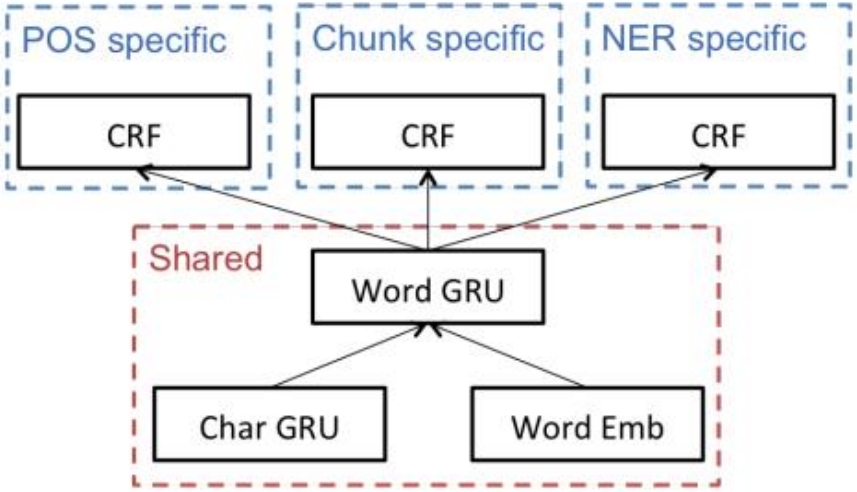
- 1.3 More languages & tasks could be modeled in a monolithic framework

1. Paradigm Shift

- 1.3 More languages & tasks could be modeled in a monolithic framework
  - Past



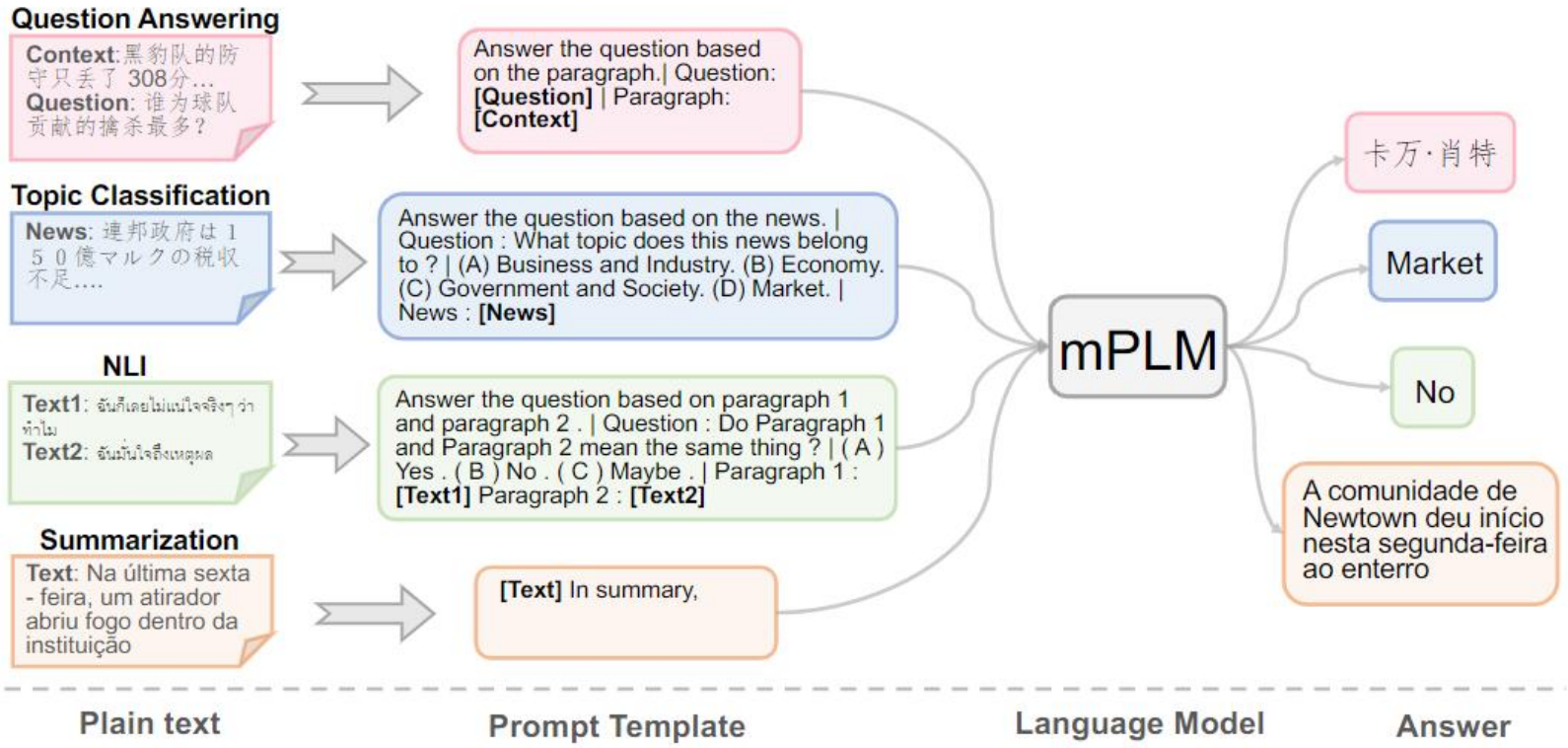
Multi-task Translation



Multi-task Multilingual Tagging

1. Paradigm Shift

- 1.3 More languages & tasks could be modeled in a monolithic framework
  - Now



1. Paradigm Shift

- 1.3 More languages & tasks could be modeled in a monolithic framework
  - Now
    - Advantages
      - Get more shared information from other tasks & languages
      - Reduce the deployment cost
      - Better zero-shot performance
      - (a good first step for metaverse)

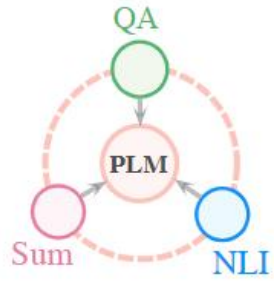
1. Paradigm Shift

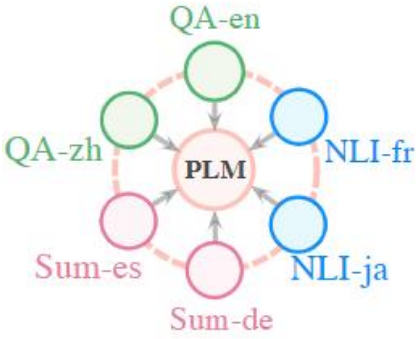
- 1.3 More languages & tasks could be modeled in a monolithic framework
  - Now
    - Advantages

	Question Answering						Sentiment	Topic	Sentence Pair			
	XQuAD		TyDiQA		MLQA		MARC	MLDOC	PAWS-X	XNLI	Avg.	Sig.
Metrics	F1	EM	F1	EM	F1	EM	Acc.	Acc.	Acc.	Acc.		
In-language training												
Vanilla mT5	72.93	57.22	81.44	70.78	62.93	44.61	91.71	93.99	84.85	69.52	73.00	-
PolyPrompt	73.65	58.17	81.63	70.32	64.90	46.44	91.66	93.80	85.09	71.82	73.75	1.91E-03
+Expand	**74.15**	**58.93**	82.00	70.69	64.95	46.57	**91.77**	93.95	84.76	**72.28**	74.00	1.54E-03
+Expand+XLSum	73.35	58.01	82.37	**71.47**	64.88	46.36	91.57	94.04	86.88	71.71	74.06	1.03E-04
+Expand+PANX	73.73	58.43	**82.75**	71.70	**65.02**	**46.60**	91.55	**94.09**	**87.10**	72.12	**74.31**	1.03E-04

1. Paradigm Shift

- 1.1 All NLP tasks could be regarded as a “PLM task”
- 1.2 More tasks could be modeled in a monolithic framework
- 1.3 More languages & tasks could be modeled in a monolithic framework





2. Application Scenario

- 2.1 Natural Language Understanding (QA, Classification etc..)
  - Past
    - Lower accuracy
  - Now
    - Significantly improved

2. Application Scenario

2.1 Natural Language Understanding

- Past
  - **Lower accuracy**
- Now
  - Significantly improved

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
Single-Task Training										
BiLSTM	63.9	5.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	75.7	52.8	**65.1**
+ELMo	66.4	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	71.7	50.1	**65.1**
+CoVe	64.0	4.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	75.4	53.5	**65.1**
+Attn	63.9	5.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	77.2	51.9	**65.1**
+Attn, ELMo	66.5	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	76.7	50.4	**65.1**
+Attn, CoVe	63.2	4.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	74.5	52.7	**65.1**
Multi-Task Training										
BiLSTM	64.2	1.6	82.8	74.3/81.8	84.2/62.5	70.3/67.8	65.4/66.1	74.6	57.4	**65.1**
+ELMo	67.7	2.1	89.3	**78.0/84.7**	82.6/61.1	67.2/67.9	70.3/67.8	75.5	57.4	**65.1**
+CoVe	62.9	8.5	81.9	71.5/78.7	84.9/60.6	64.4/62.7	65.4/65.7	70.8	52.7	**65.1**
+Attn	65.6	8.6	83.0	76.2/83.9	82.4/60.1	72.8/70.5	67.6/68.3	74.3	58.4	**65.1**
+Attn, ELMo	**70.0**	3.6	**90.4**	**78.0**/84.4	84.3/63.1	74.2/72.3	74.1/74.5	**79.8**	58.9	**65.1**
+Attn, CoVe	63.1	8.3	80.7	71.8/80.0	83.4/60.5	69.8/68.4	68.1/68.6	72.9	56.0	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	72.1	54.1	**65.1**
Skip-Thought	61.3	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	72.9	53.1	**65.1**
InferSent	63.9	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	72.7	58.0	**65.1**
DisSent	62.0	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	73.9	56.4	**65.1**
GenSen	66.2	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	78.6	**59.2**	**65.1**

Reported results of GLUE benchmark.

2. Application Scenario

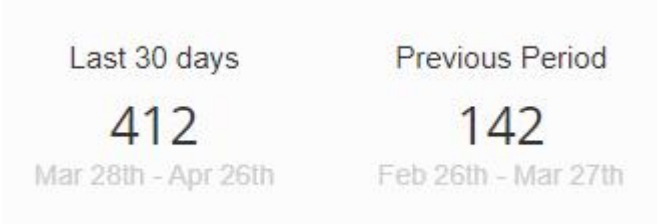
- 2.1 Natural Language Understanding
  - Past
    - Lower accuracy
  - Now
    - **Significantly improved**

URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
	91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
	91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
	91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
	91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
	90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
	90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6

2. Application Scenario

- 2.2 Natural Language Generation
  - Past
    - Limited applicable scenario due to poor generation except machine translation
  - Now
    - More practically-useful text could be generated
      - Fluent
      - Coherent
      - Relevant

Example: Generated Reviews (400M)



Paper Review

Model List

Introduction

CE extraction

Abstract + CE

Abstract + CE (with aspect)

summary | clarity | **motivation** | originality | soundness | substance

This paper presents a method for training deep neural networks with residual layers . The idea is to learn a residual representation for each layer of the network , which is learned as an identity mapping of the input and output channels . The paper is well written and easy to follow . The method is evaluated on ImageNet and CIFAR-10 . The results show that the proposed method is able to improve the performance of deep networks with a depth of up to 152 layers (8x deeper than VGG nets) . The authors also show that their method can achieve a 28 % improvement on the COCO object detection dataset . I think this paper is a good contribution to the field of deep learning . However , I have the following concerns : 1 . The novelty of the method is limited . It is a simple idea to learn the residual representation of each layer as a mapping from the input to the output channels , and it is not very novel . 2 .The experiments are not very convincing . The proposed method does not outperform VGG on the ImageNet dataset , and the results on Cifar-10 are not impressive . 3 .The authors claim that the method can be applied to other vision and non-vision problems , but the experiments are only conducted on the image classification task . It would be better if the authors can conduct experiments on other vision tasks .

Example: Generated Essay (10 B)

Prompt/Requirements

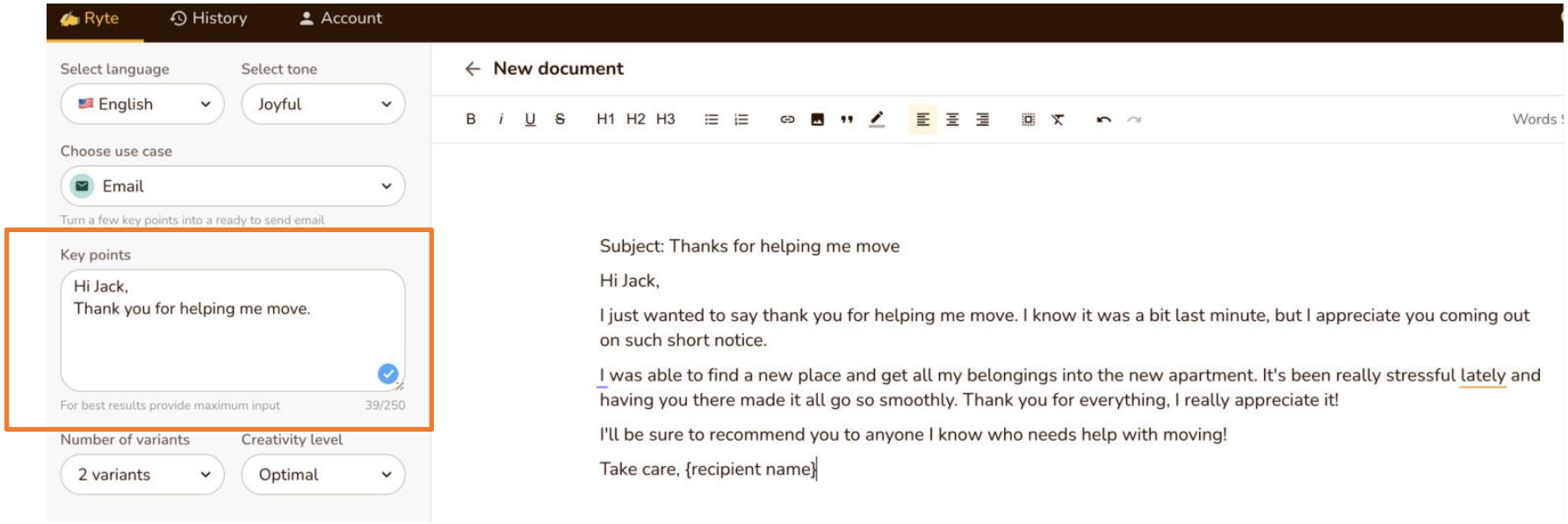
Suppose you are Li Hua. Last Sunday, you organized a 5K cross-country race. Please write a report for the school English newspaper, including:
1.Participants;
2.Running route: from the school gate to Nanshan Foot K;
3.Activity response. Here are some requirements:
4.The number of writing words should be about 80;
5.Please answer in the corresponding position of the answer sheet in the following format.

Generated Results

Last Sunday, our school organized a Cross-Country Running Race. Hundreds of us took part in it. It was a nice day. At 8:00 am, we gathered at the school gate and set out for Nanshan Foot K. The race was held in the sunshine. All the way we were chatting, singing and laughing, enjoying the fresh air and the beautiful scenery. After that, we ran the 5K race in a straight line. At first, some of the runners were a little bit slow, but soon they caught up with the rest. About 2 minutes later, a second runner came out of the gate and hurried to help the first one finish the race. Finally, we all jumped and cheered with joy after the race was over. The activity benefited us a lot. Not only did it get us close to nature and give us relaxation from heavy school work, it also promoted the friendship among us. What a wonderful time!

GPT3-style Startups/Products

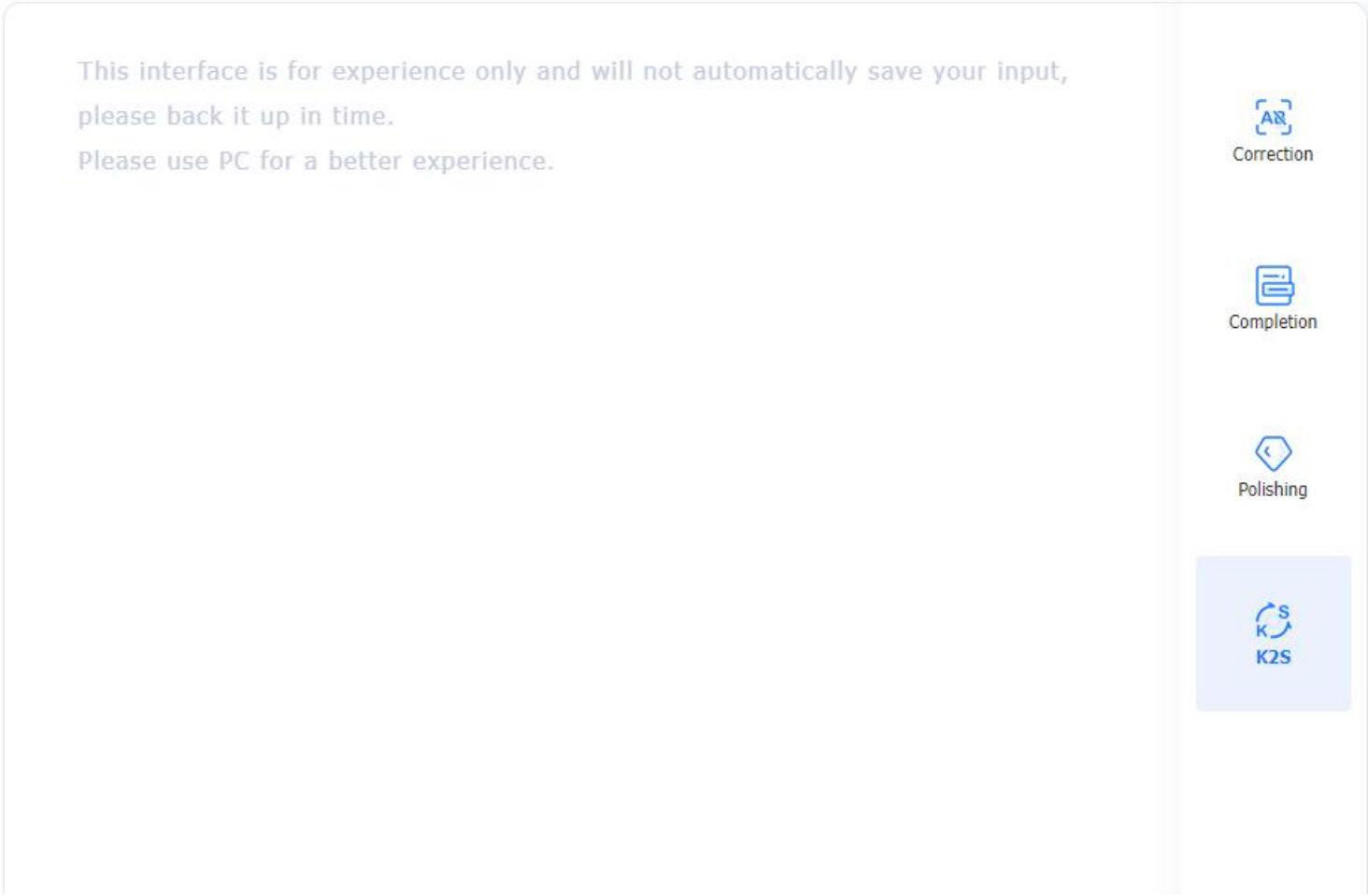
Business Writing Assistant (GPT3, from RYTR)



https://rytr.me/

GPT3-style Startups/Products

Scientific Writing Assistant (from Tencent)

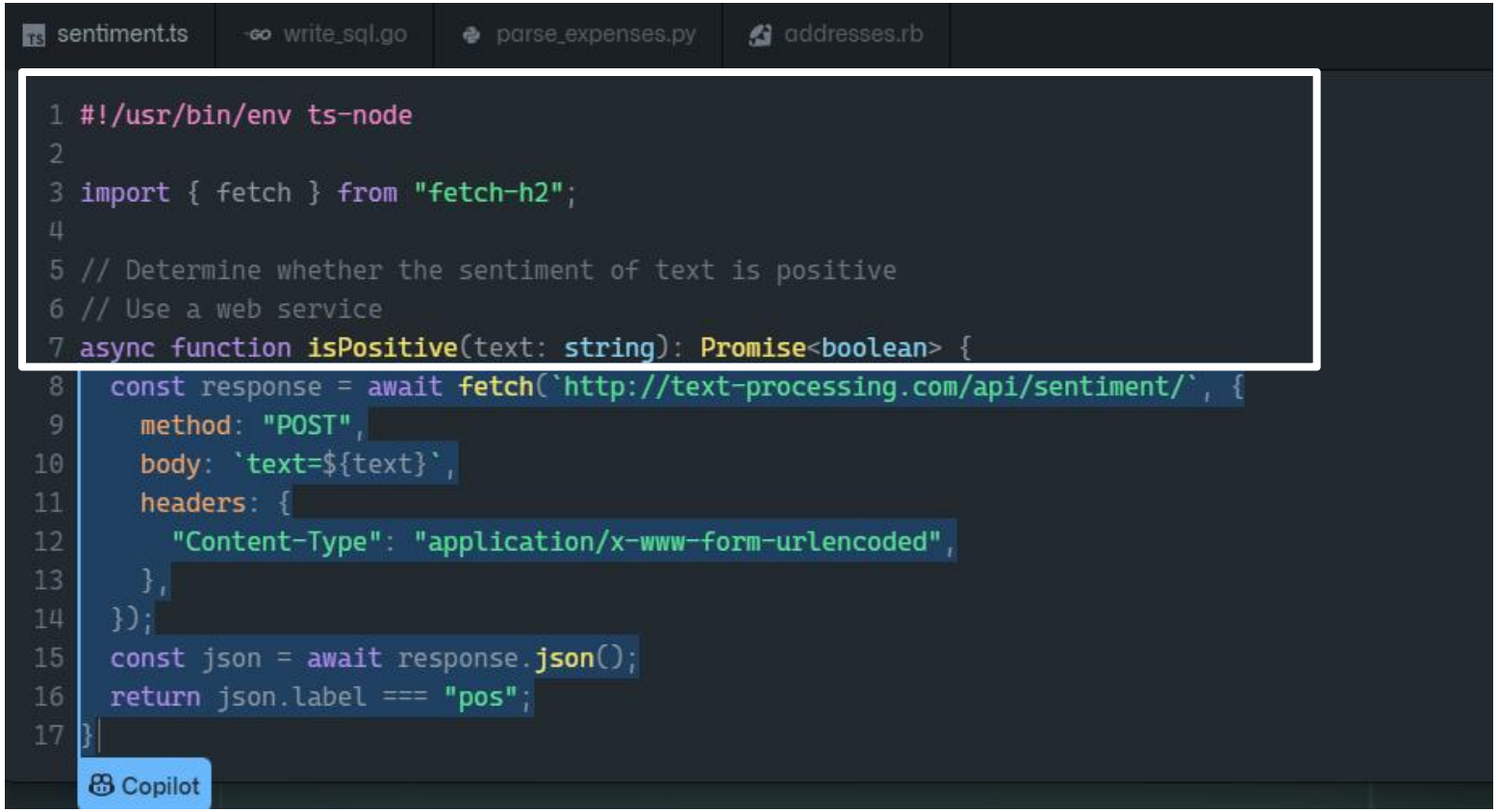


GPT3-style Startups/Products

Novel Writing Assistant (from Baidu)

GPT3-style Startups/Products

Codex/Copilot

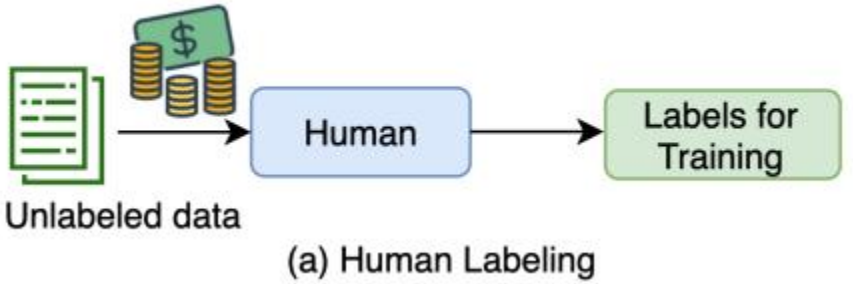


3. Data Annotation

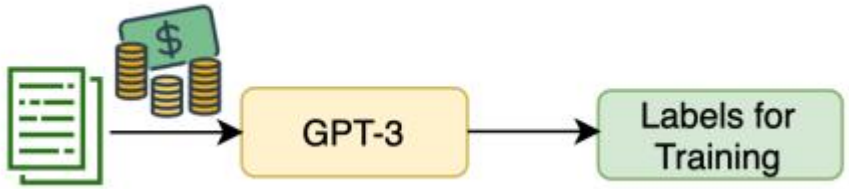
- Past
  - Human Label
- Now
  - PLM (GPT)

3. Data Annotation

- Past
  - Human Label
- Now
  - PLM (GPT)



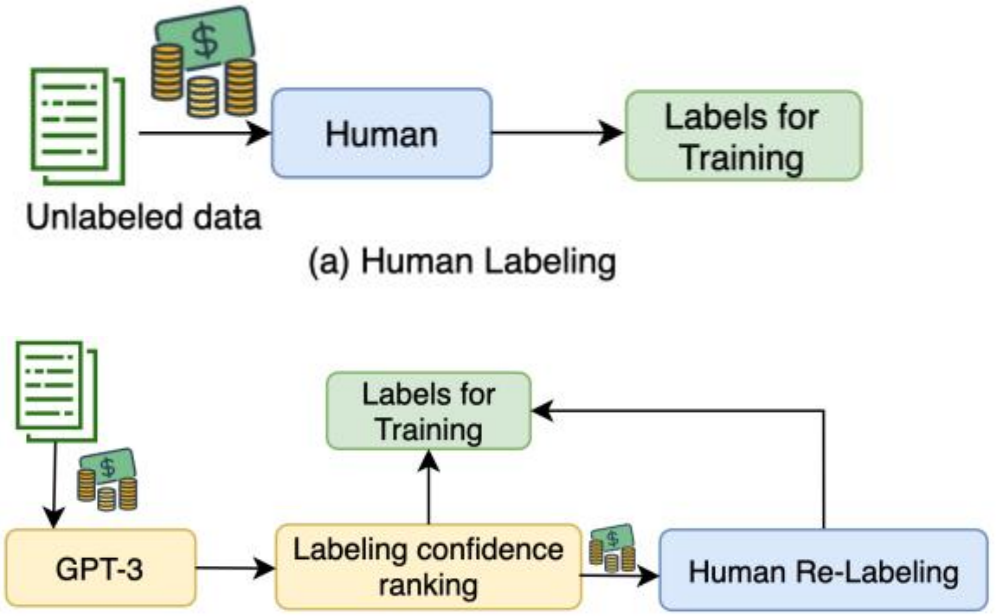
(a) Human Labeling



Want To Reduce Labeling Cost? GPT-3 Can Help (EMNLP 2021)

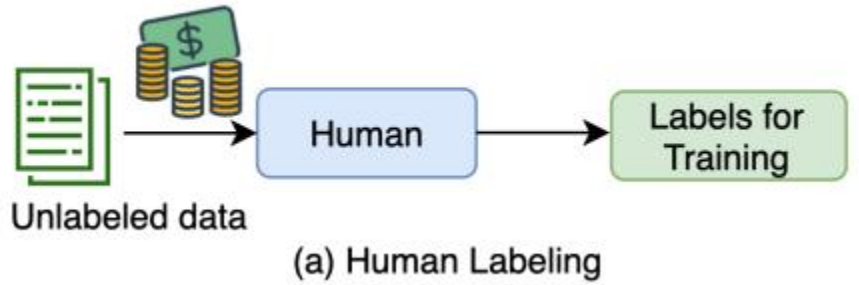
3. Data Annotation

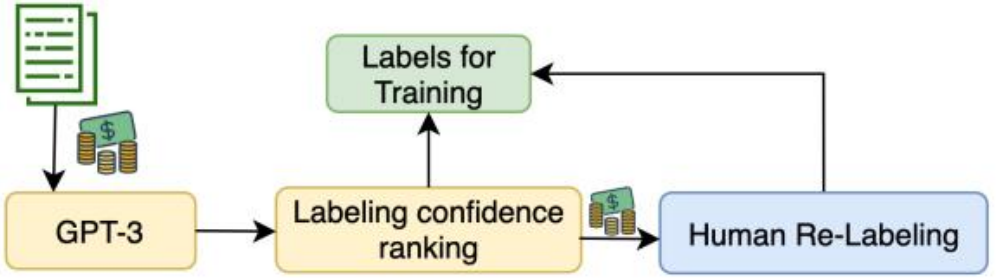
- Past
  - Human Label
- Now
  - PLM (GPT)

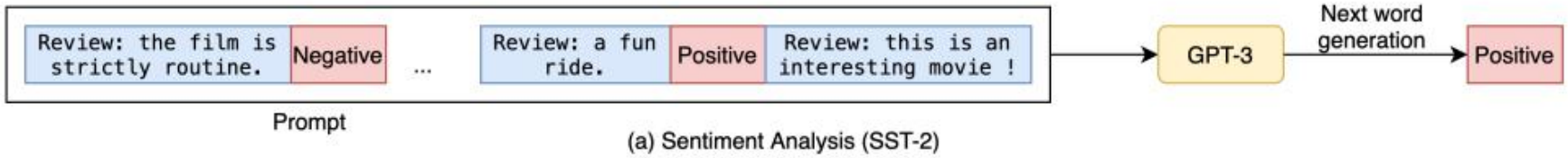


3. Data Annotation

- Past
  - Human Label
- Now
  - PLM (GPT)





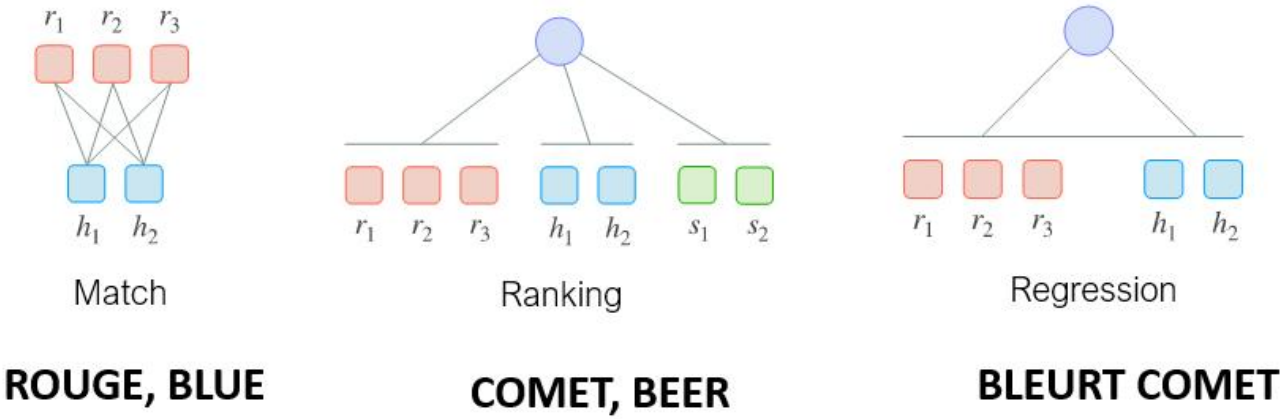


4. Data Retrieval

- Past
  - Knowledge
  - Query
- Now
  - PLM
  - Prompt

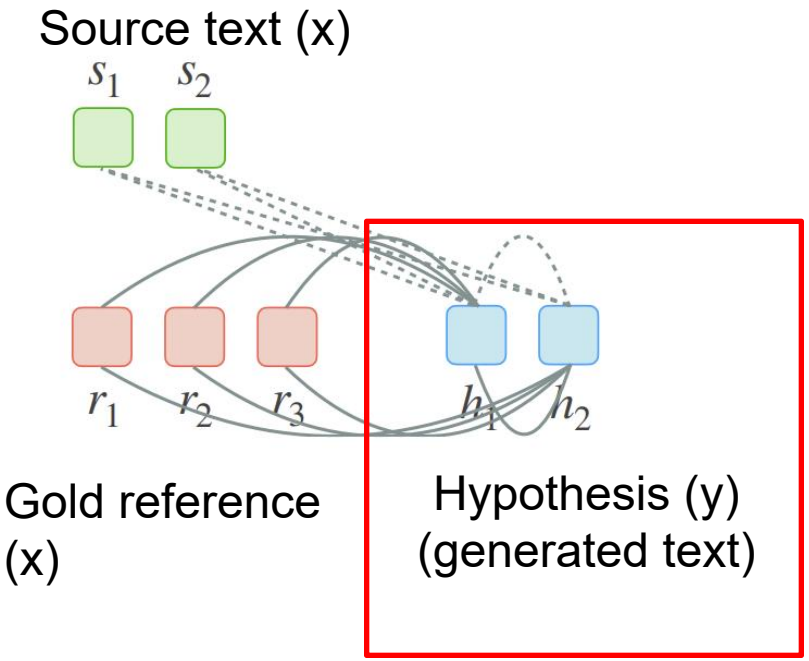
5. Evaluation (for Generated Text)

- 5.1 Evaluation is not everything about matching/ranking
  - Past



Evaluation (for Generated Text)

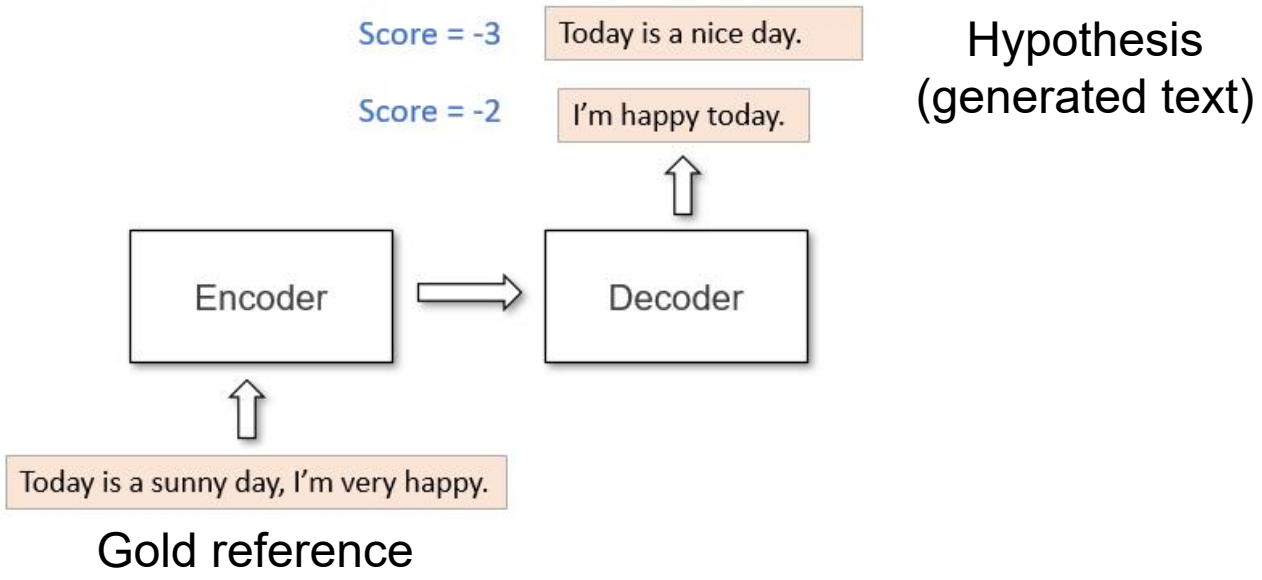
- 5.1 Evaluation is not everything about matching/ranking
  - Now: New Evaluation Paradigm
    - Generation formulation



$$BARTScore = \sum_{t=1}^{m} w_t \log p(y_t | y_{<t}, x, \theta)$$

Evaluation (for Generated Text)

- 5.1 Evaluation is not everything about matching/ranking
  - Now: New Evaluation Paradigm



5. Evaluation (for Generated Text)

- 5.2 One metric, evaluate multiple perspectives

5. Evaluation (for Generated Text)

- 5.2 One metric, evaluate multiple perspectives
  - Past
    - Previous metrics are mainly designed for limited perspectives

ROUGE

FactCC

5. Evaluation (for Generated Text)

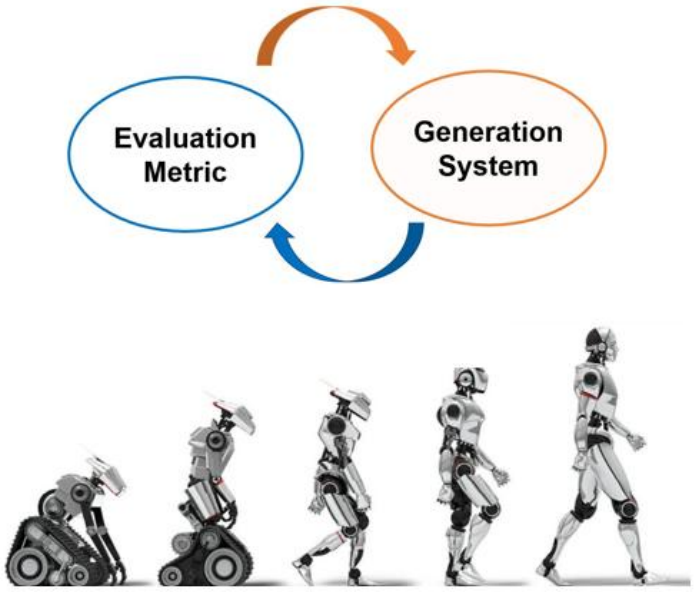
- 5.2 One metric, evaluate multiple perspectives
  - Now

- ☐ Fluency
- ☐ Relevance
- ☐ Coherence
- ☐ Informativeness
- ☐ Factuality
- ☐ Semantic Coverage
- ☐ Adequacy

If one model can generate high-quality text, this model must know what makes a good text implicitly.

5. Evaluation (for Generated Text)

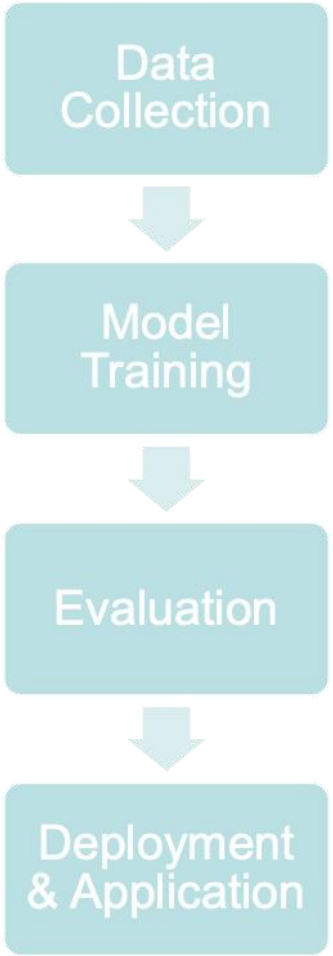
- 5.3 Model & Metric Co-evolve

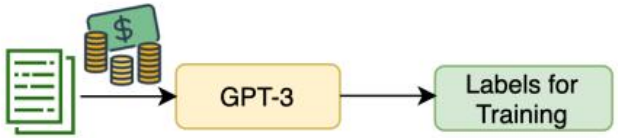


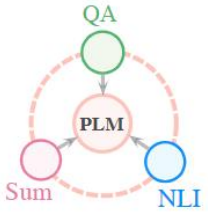
- Better systems will result in better evaluation metrics.
- Better evaluation metrics will guide the systems to become better.

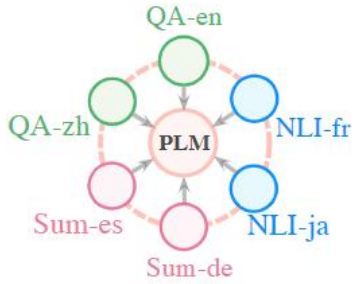
Summary

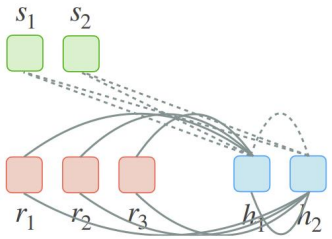
ML Lifecycle

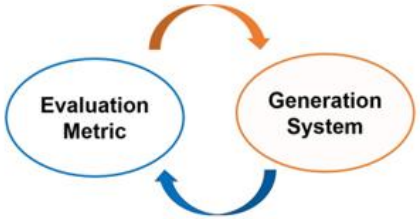












What are the four paradigms in modern NLP?

Four Paradigms in Modern NLP

- Feature Engineering
- Architecture Engineering
- Objective Engineering
- Prompt Engineering

Four Paradigms in Modern NLP

- Feature Engineering
- Architecture Engineering
- Objective Engineering
- Prompt Engineering

- **Paradigm：** Fully Supervised Learning (Non-neural Network)
- **Date：** Before 2013
- **Characteristic：** Traditional machine learning model is mainly used, which requires manual feature definition of input text
- **Typical Work：**
  - CRF (Conditional Random Field)

Four Paradigms in Modern NLP

- Feature Engineering
- Architecture Engineering
- Objective Engineering
- Prompt Engineering

- **Paradigm：** Fully Supervised Learning (Neural Network)
- **Date：** 2013 - 2018
- **Characteristic：**
  - Rely on neural networks
  - Do not need to manually define features, but should explore the network structure （e.g.：LSTM v.s CNN）
- **Typical Work：**
  - CNN for Text Classification

Four Paradigms in Modern NLP

- Feature Engineering
- Architecture Engineering
- Objective Engineering
- Prompt Engineering

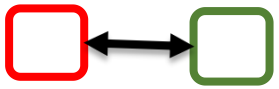
- **Paradigm：** Pre-train, Fine-tune
- **Date：** 2018-Now
- **Characteristic：**
  - context-dependent PLMs
  - Need to pay attention to the definition and selection of objective functions
- **Typical Work：** BERT

Four Paradigms in Modern NLP

- Feature Engineering
- Architecture Engineering
- Objective Engineering
- Prompt Engineering

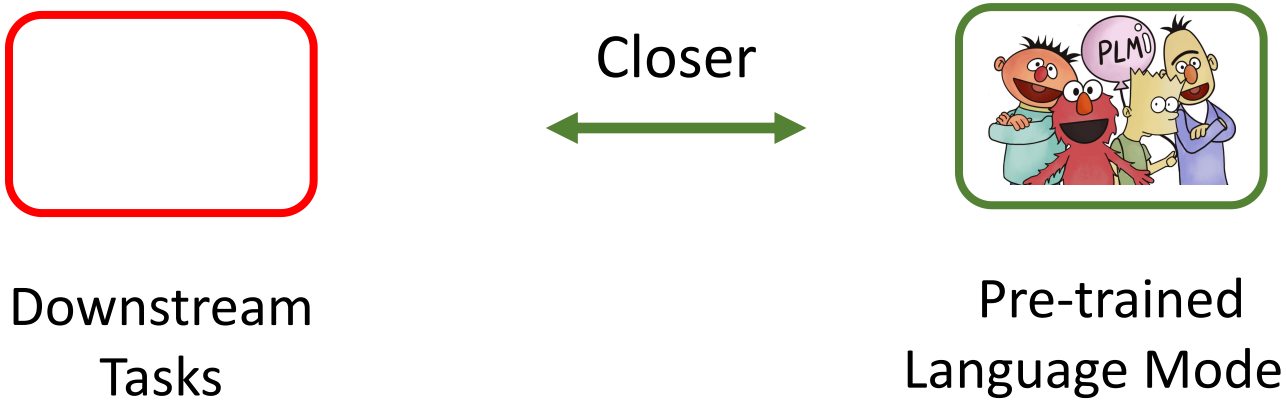
- **Paradigm**： Pre-train, Prompt, Predict
- **Date**： 2019-Now
- **Characteristic**：
  - NLP tasks are modeled entirely by relying on PLMs
  - More efforts on prompt design
- **Typical Work:** GPT3

PLMs and Downstream Tasks are Getting Closer and Closer

Stages	Downstream Tasks	Pre-trained LMs	Reasons
Traditional machine learning			No pre-training language model
Neural network methods enhanced by word2vec			The pre-trained language model plays the role of initializing the input text signal
The fine-tune method represented by BERT			The pre-trained language model is **responsible for extracting** high-level features from the input text
The prompt approach represented by GPT3			Pre-training language models **take on more responsibilities**: feature extraction, result prediction

Secret let out from Prompt-based Learning

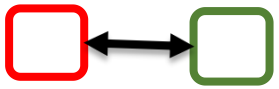
The history of modern natural language processing is essentially (probably) a history of changes in the relationship between downstream tasks and pre-trained language models (PLMs).



（1）use pre-trained language models
（2）use a better pre-trained language model
（3）better use a pre-trained language model

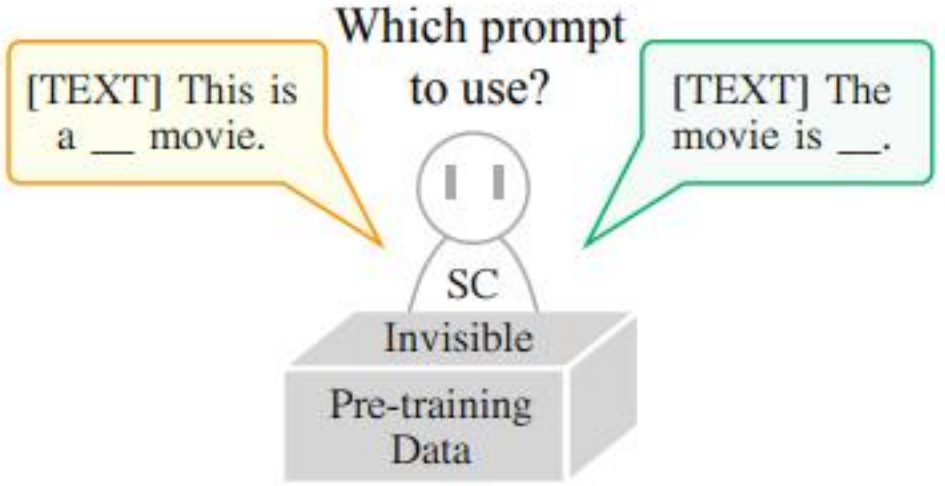
What could the next paradigm be?

PLMs and Downstream Tasks are Getting Closer and Closer

Stages	Downstream Tasks	Pre-trained LMs	Reasons
Traditional machine learning			No pre-training language model
Neural network methods enhanced by word2vec			The pre-trained language model plays the role of initializing the input text signal
The fine-tune method represented by BERT			The pre-trained language model is **responsible for extracting** high-level features from the input text
The prompt approach represented by GPT3			Pre-training language models **take on more responsibilities**: feature extraction, result prediction

The Pitfall of Prompting Learning

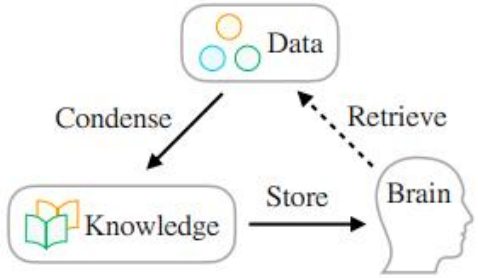
- The way how information is stored is opaque
- There is a gap between data storing and accessing



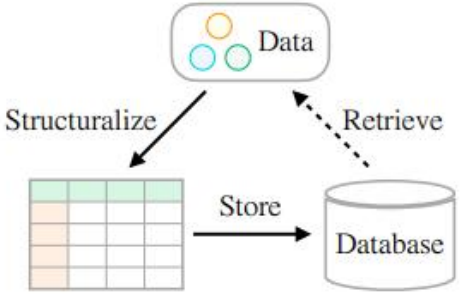
The sentiment classification (SC) task is guessing which prompt should be used

The Way how Human Store/Access Data

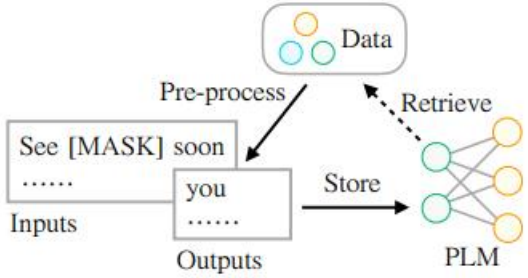
- there are often differences in the way that data is stored and accessed
- there have been efforts to bridge such a gap to better use the information (i.e., accurately recall) that exists in the world



(a) Biological neural networks.



(b) Disk/Cloud storage.



(c) Artificial neural netwokrs.

What makes a good pre-training strategy?

- A good PLM should have a clear picture of the composition of the various signals in the data
  - to provide accurate information for downstream tasks according to their different needs.
- valuable signals are rich and exist everywhere from the data in the world
  - instead of simply existing in the supervised datasets that are manually curated


Github
Common Crawl
wikiHow
Netflix
Wikipedia
Entities
Relations
.....
YouTube
PubMed
Compression
Prompting
Twitter
ArXiv
Google Images
Wikidata
RST
??

reStructure Pre-training

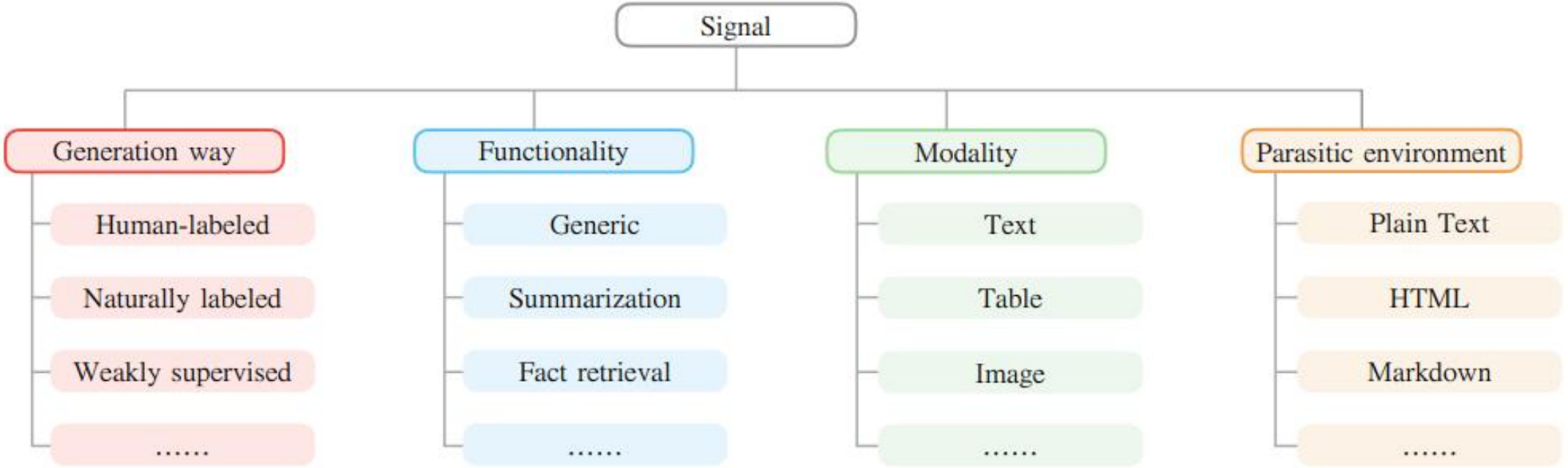
- Signal Definition
- Data Mine Identification
- Signal Extraction
- Signal Restructuring
- Pre-training and Tuning

reStructure Pre-training

- Signal Definition
  - useful information present in the data that can provide supervision for machine learning models
  - can be represented as n-tuples
    - Signal for named entity recognition: (“Mozart was born in Salzburg”, “Mozart, Salzburg”)

reStructure Pre-training

- Signal Definition



reStructure Pre-training

- Data Mine Identification
  - data sources that are rich in different types of signal
  - Example:
    - Wikipedia
    - WikiHow
    - Dailymail
    - Rotten Tomatoes

reStructure Pre-training

- Signal Extraction
  - Extract clean signal from data mines
    - Extracting
    - Pre-processing
    - Normalizing

reStructure Pre-training

- Signal Restructuring
  - unify different signals into a fixed form
    - rich information can be stored together in the model during pre-training
  - Unified strategies
    - Generic signal
      - Denoise autoencoder
    - Task-relevant signals
      - Multiple-choice format
      - generation

Thank you for inviting me to your party last week

Thank you <x> me to your party <Y> week

<X> for inviting <Y> last <X>

reStructure data in the world

Signals	Plain Text	Cloze	(corrupted_text; corrupted_positions; target_spans)
	Rotten Tomatoes	Sentiment	(review; sentiment)
	Daily Mail	Category	(summary; category)
		Summary	(text; summary)
			(text; title)
			(summary; title)
		Temporal Information	(text; one_bullet_point; another_bullet_point)
			(text; first_bullet_point; second_bullet_point)
	Wikidata	Relation	(text; subject; property; object)
		Entity Typing	(text; entity; entity_type)
	wikiHow	Category	(title_description; category)
		Category Hierarchy	(low_category; high_category)
		Goal-step Relation	(goal; step_headline)
		Summary	(step_headline; step_description)
		Procedure	(goal; first_step_headline; second_step_headline)
		Question Generation	(title; answer; first_related_question)
			(title_description; title)
			(answer; title)
			(title; first_related_question)
	Wikipedia	Section Title	(section_text; section_title)
		Entity	(paragraph; entities)
		Sentiment	(text; neutral_sentiment)
	WordNet	Part-of-speech	(word; sentence; POS)
		Meaning	(word; sentence; meaning)
		Synonym	(word; sentence; synonym)
		Antonym	(word; sentence; antonym)
	ConTRoL; DREAM; LogiQA; ReClor; RACE; RACE-C; TriviaQA	Reading Comprehension	(context; question; answer)
	arXiv	Category	(abstract; category)
		Summarization	(abstract; title)

Human: Hey PLM guys, valuable information is like the gem of a mine, distributed in all kinds of data, and when you have the eyes to find that treasure, you will make the world better again. Let's go on a treasure hunt...

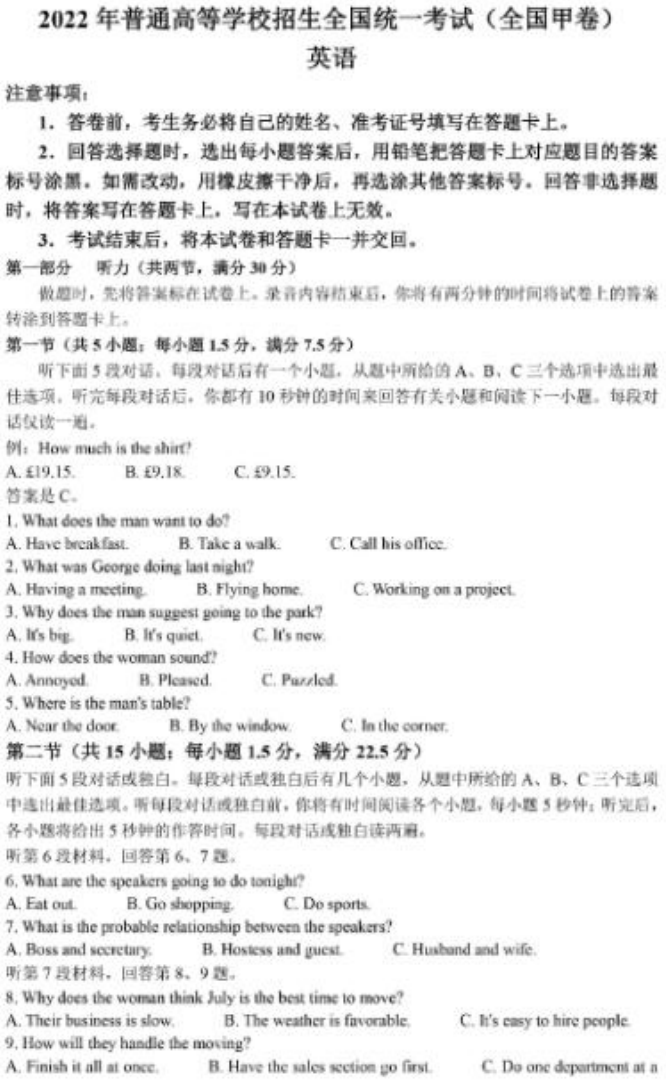
The future is all about the data; the future is all about you ...

Application I: General NLP Tasks

- Surpass strong competitors (e.g., T0) on 52/55 popular datasets over 9 different tasks
  - Text classification
  - Text entailment
  - Information extraction
  - Fact retrieval
  - Summarization
  - …

	Dataset	Metric	T0pp			RST-Task	
			avg	max	std	avg	max
Topic Classification	subj	acc	48.70	58.00	8.53	**60.24**	**83.90**
	qc	acc	57.43	66.06	7.89	**69.12**	**71.69**
	yahoo_answers_topics	acc	35.13	41.10	4.00	**58.69**	**59.40**
	hate_speech18	acc	79.64	86.46	9.20	74.78	81.56
	tweet_eval/emotion	acc	70.54	72.34	1.93	**80.99**	**81.77**
	tweet_eval/hate	acc	57.95	58.15	0.16	**62.45**	**64.28**
	tweet_eval/irony	acc	61.63	63.14	1.35	64.29	**78.57**
	tweet_eval/offensive	acc	72.07	72.21	0.13	72.23	**77.21**
Sentiment Classification	financial_phrasebank	acc	36.17	36.44	0.31	60.75	61.17
	mr	acc	89.04	92.50	3.22	**92.10**	**92.59**
	sst2	acc	92.32	**96.44**	3.52	**93.26**	93.58
Information Extraction	conll03	f1	17.87	23.37	4.35	**54.72**	**55.33**
	notebc	f1	10.89	13.83	2.75	**28.51**	**31.38**
	notebn	f1	16.07	19.94	3.49	**33.07**	**34.72**
	notemz	f1	12.72	17.78	3.65	**30.36**	**32.28**
	notenw	f1	16.04	18.54	1.92	**28.53**	**29.69**
	notewb	f1	9.41	11.75	2.18	16.45	**18.30**
	notetc	f1	**9.59**	**12.29**	2.52	8.06	10.20
	wikiann	f1	24.69	31.66	6.57	**55.81**	**57.20**
	wnut17	f1	11.00	14.09	1.81	22.28	**23.62**
	semeval_rel	f1	15.62	19.16	2.88	20.48	**25.28**
	wiki80	acc	36.04	41.30	3.90	53.57	54.58
Natural Language Inference	anli-r1	acc	45.96	52.00	6.33	**71.28**	**74.30**
	anli-r2	acc	41.34	46.10	4.44	**60.08**	**63.20**
	anli-r3	acc	40.17	44.42	3.86	**57.37**	**61.33**
	cb	acc	76.07	83.93	11.89	**82.50**	**91.07**
	multi_nli_matched	acc	55.14	62.22	9.67	**75.47**	**77.96**
	multi_nli_mismatched	acc	56.56	63.50	10.12	**76.31**	**78.68**
	rte	acc	79.06	**87.00**	10.38	**83.90**	85.92
	sick	acc	37.31	55.81	14.27	**54.25**	**69.75**
	snli	acc	58.67	65.38	9.09	**78.44**	**82.44**

Application II: Gaokao Benchmark



2022 年普通高等学校招生全国统一考试（全国甲卷）

英语

注意事项：

1．答卷前，考生务必将自己的姓名、准考证号填写在答题卡上。

2．回答选择题时，选出每小题答案后，用铅笔把答题卡上对应题目的答案标号涂黑。如需改动，用橡皮擦干净后，再选涂其他答案标号。回答非选择题时，将答案写在答题卡上。写在本试卷上无效。

3．考试结束后，将本试卷和答题卡一并交回。

第一部分　听力（共两节，满分 30 分）

做题时，先将答案标在试卷上。录音内容结束后，你将有两分钟的时间将试卷上的答案转涂到答题卡上。

第一节（共 5 小题；每小题 1.5 分，满分 7.5 分）

听下面 5 段对话。每段对话后有一个小题，从题中所给的 A、B、C 三个选项中选出最佳选项，听完每段对话后，你都有 10 秒钟的时间来回答有关小题和阅读下一小题。每段对话仅读一遍。

例：How much is the shirt?

A. £19.15.　　B. £9.18.　　C. £9.15.

答案是 C。

1. What does the man want to do?

A. Have breakfast.　　B. Take a walk.　　C. Call his office.

2. What was George doing last night?

A. Having a meeting.　　B. Flying home.　　C. Working on a project.

3. Why does the man suggest going to the park?

A. It's big.　　B. It's quiet.　　C. It's new.

4. How does the woman sound?

A. Annoyed.　　B. Pleased.　　C. Puzzled.

5. Where is the man's table?

A. Near the door.　　B. By the window.　　C. In the corner.

第二节（共 15 小题；每小题 1.5 分，满分 22.5 分）

听下面 5 段对话或独白。每段对话或独白后有几个小题，从题中所给的 A、B、C 三个选项中选出最佳选项。听每段对话或独白前，你将有时间阅读各个小题，每小题 5 秒钟；听完后，各小题将给出 5 秒钟的作答时间。每段对话或独白读两遍。

听第 6 段材料，回答第 6、7 题。

6. What are the speakers going to do tonight?

A. Eat out.　　B. Go shopping.　　C. Do sports.

7. What is the probable relationship between the speakers?

A. Boss and secretary.　　B. Hostess and guest.　　C. Husband and wife.

听第 7 段材料，回答第 8、9 题。

8. Why does the woman think July is the best time to move?

A. Their business is slow.　　B. The weather is favorable.　　C. It's easy to hire people.

9. How will they handle the moving?

A. Finish it all at once.　　B. Have the sales section go first.　　C. Do one department at a

Subcategory	Task Formulation	Example Question
Listening	Speech Recognition Code Switching Dialogue Understanding Multiple-choice QA	**Requirement**: Based on the listening materials, choose the right answer from the given options. **Question**: Where does this conversation take place? **Options**: (A) In a classroom (B) In a hospital (C) In a museum
Cloze (multiple-choice)	Multiple-choice QA	**Requirement**: Based on the context, choose the right answer to fill in the blank from the given options. **Text**: ..... The ___ might damage the beauty of the place..... **Options**: (A) stories (B) buildings (C) crowds (D) reporters
Cloze (hint)	Open-domain QA	**Requirement**: Based on the context and hint, write down the correct answer to fill in the blank. **Text**: A 90-year-old has been awarded "Woman Of The Year" for ___ Britain's oldest employee..... **Hint**: be
Reading (multiple-choice)	Multiple-choice QA	**Requirement**: Based on the text, choose the correct option from the given choices to answer the question. **Text**: Need a Job This Summer? The provincial government and its partners offer many programs to help students find summer jobs..... **Question**: What is the age range required by Stewardship Youth Ranger Program? **Options**: (A) 15–18 (B) 15–24 (C) 15–29 (D) 16-17
Reading (cloze)	Multiple-choice QA	**Requirement**: Based on the context, choose the best option from the given choices to fill in the blank. **Text**: ..... ___ Like the child on the diving board, you will stay undecided ..... **Options**: (A) Without motivation, you can neither set a goal nor reach it. (B) So how should you motivate yourself? (C) This can affect your work. (D) They can change according to circumstances.

Application II: Gaokao Benchmark

2022 年普通高等学校招生全国统一考试（全国甲卷）

英语

注意事项：

1. 答卷前，考生务必将自己的姓名、准考证号填写在答题卡上。
2. 回答选择题时，选出每小题答案后，用铅笔把答题卡上对应题目的答案标号涂黑。如需改动，用橡皮擦干净后，再选涂其他答案标号。回答非选择题时，将答案写在答题卡上。写在本试卷上无效。
3. 考试结束后，将本试卷和答题卡一并交回。

第一部分　听力（共两节，满分 30 分）

做题时，先将答案标在试卷上。录音内容结束后，你将有两分钟的时间将试卷上的答案转涂到答题卡上。

第一节（共 5 小题；每小题 1.5 分，满分 7.5 分）

听下面 5 段对话。每段对话后有一个小题，从题中所给的 A、B、C 三个选项中选出最佳选项。听完每段对话后，你都有 10 秒钟的时间来回答有关小题和阅读下一小题。每段对话仅读一遍。

例：How much is the shirt?

A. £19.15.　　B. £9.18.　　C. £9.15.

答案是 C。

1. What does the man want to do?

A. Have breakfast.　　B. Take a walk.　　C. Call his office.

2. What was George doing last night?

A. Having a meeting.　　B. Flying home.　　C. Working on a project.

3. Why does the man suggest going to the park?

A. It's big.　　B. It's quiet.　　C. It's new.

4. How does the woman sound?

A. Annoyed.　　B. Pleased.　　C. Puzzled.

5. Where is the man's table?

A. Near the door.　　B. By the window.　　C. In the corner.

第二节（共 15 小题；每小题 1.5 分，满分 22.5 分）

听下面 5 段对话或独白。每段对话或独白后有几个小题，从题中所给的 A、B、C 三个选项中选出最佳选项。听每段对话或独白前，你将有时间阅读各个小题，每小题 5 秒钟；听完后，各小题将给出 5 秒钟的作答时间。每段对话或独白读两遍。

听第 6 段材料，回答第 6、7 题。

6. What are the speakers going to do tonight?

A. Eat out.　　B. Go shopping.　　C. Do sports.

7. What is the probable relationship between the speakers?

A. Boss and secretary.　　B. Hostess and guest.　　C. Husband and wife.

听第 7 段材料，回答第 8、9 题。

8. Why does the woman think July is the best time to move?

A. Their business is slow.　　B. The weather is favorable.　　C. It's easy to hire people.

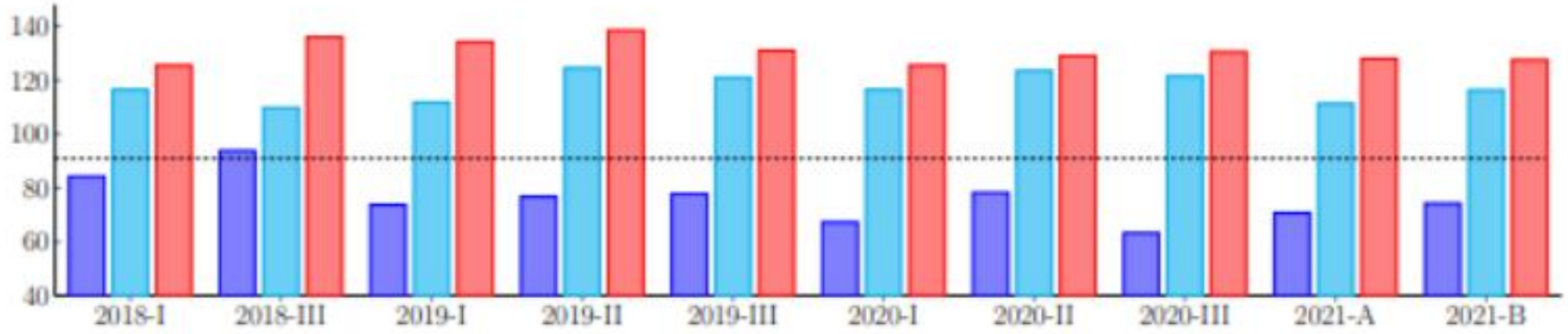
9. How will they handle the moving?

A. Finish it all at once.　　B. Have the sales section go first.　　C. Do one department at a

Gaokao AI: Qin

Gaokao AI: Qin

- Qin
  - achieves 40 points higher than the average scores made by students
  - 15 points higher than GPT3 with 1/16 parameters
  - 138.5 (full mark is 150) in 2018



Resources

reStructured Pre-training
https://arxiv.org/pdf/2206.11147.pdf

- Paper
- reStructured Data & Model
- Gaokao Benchmark

Contents

Resources https://github.com/ExpressAI/reStructured-Pretraining

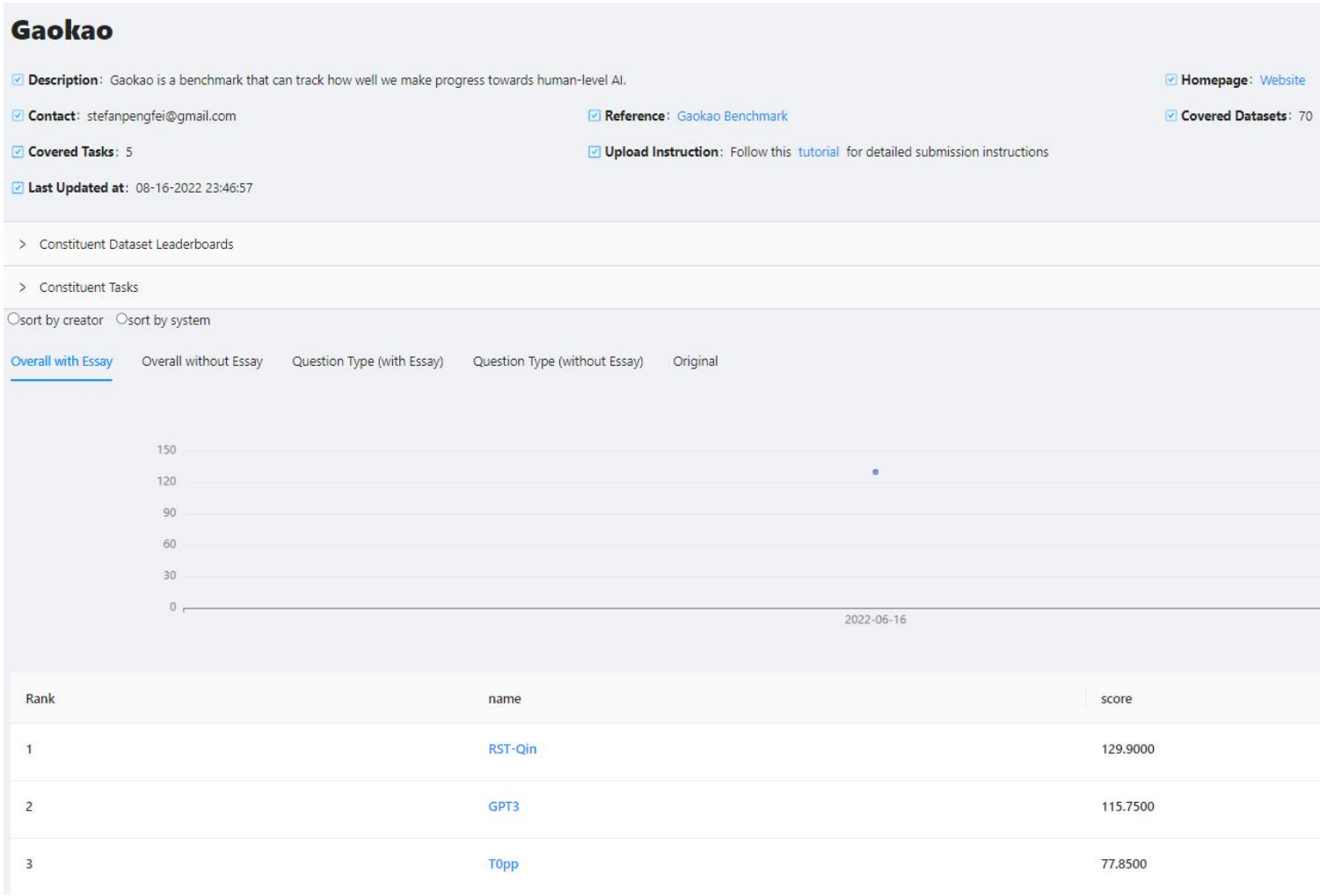
- Paper
- reStructured Data & Model
- Gaokao Benchmark

Mine	Signal	#Sample	Use in DataLab	Some Applications
Rotten Tomatoes	(review, rating)	5,311,109	`load_dataset("rst", "rotten_tomatoes_sentiment")`	Sentiment classification
Daily Mail	(text, category)	899,904	`load_dataset("rst", "daily_mail_category")`	Topic classification
Daily Mail	(title, text, summary)	1,026,616	`load_dataset("rst", "daily_mail_summary")`	Summarization; Sentence expansion
Daily Mail	(text, events)	1,006,412	`load_dataset("rst", "daily_mail_temporal")`	Temporal reasoning
Wikidata	(entity, entity_type, text)	2,214,274	`load_dataset("rst", "wikidata_entity")`	Entity typing
Wikidata	(subject, object, relation, text)	1,526,674	`load_dataset("rst", "wikidata_relation")`	Relation extraction; Fact retrieval
wikiHow	(text, category)	112,109	`load_dataset("rst", "wikihow_text_category")`	Topic classification
wikiHow	(low_category, high_category)	4,868	`load_dataset("rst", "wikihow_category_hierarchy")`	Relation extraction; Commonsense reasoning
wikiHow	(goal, steps)	47,956	`load_dataset("rst", "wikihow_goal_step")`	Intent detection

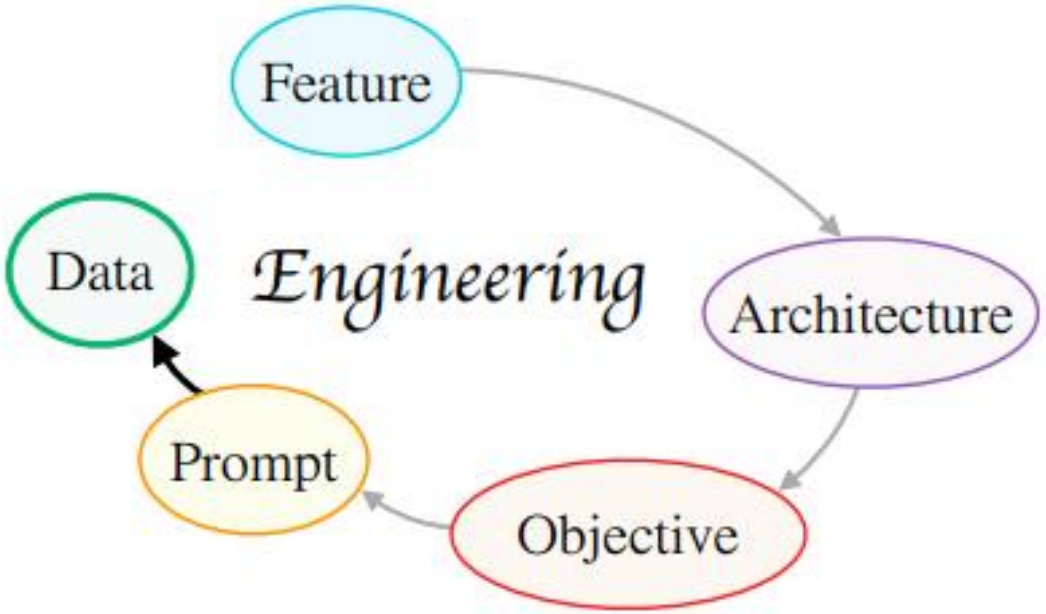
Resources

https://explainaboard.inspiredco.ai/benchmark?id=gaokao

- Paper
- reStructured Data & Model
- Gaokao Benchmark



Hypothesis of NLP technique evolution

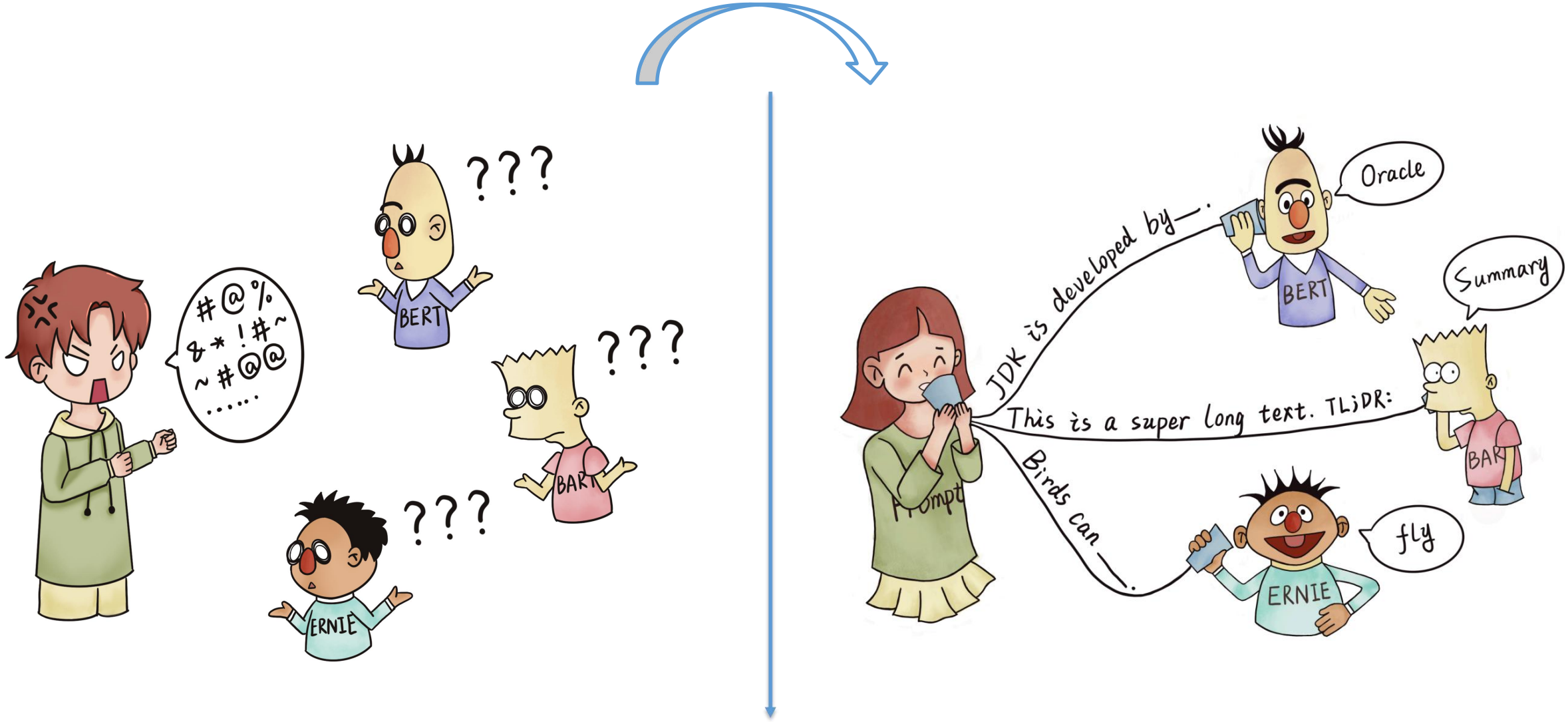


reStructured Pre-training, arXiv 2022, Weizhe Yuan, Pengfei Liu

Outline

- What is the “Prompt”?
- What is the general workflow of prompt-based methods?
- What are the design considerations for prompt-based methods?
- What are things we can do now that we couldn‘t do in the past?
- What are the four paradigms in modern NLP?
- What could the next paradigm be?

Hopefully


#@%
&*!#~
~#@@
.....
???
BERT
???
BART
???
ERNIE
Prompt
JDK is developed by—.
Oracle
BERT
This is a super long text. TL;DR:
Summary
BART
Birds can—.
fly
ERNIE