提示工程

CS2916 大语言模型

飲水思源 愛國榮校

https://plms.ai/teaching/index.html

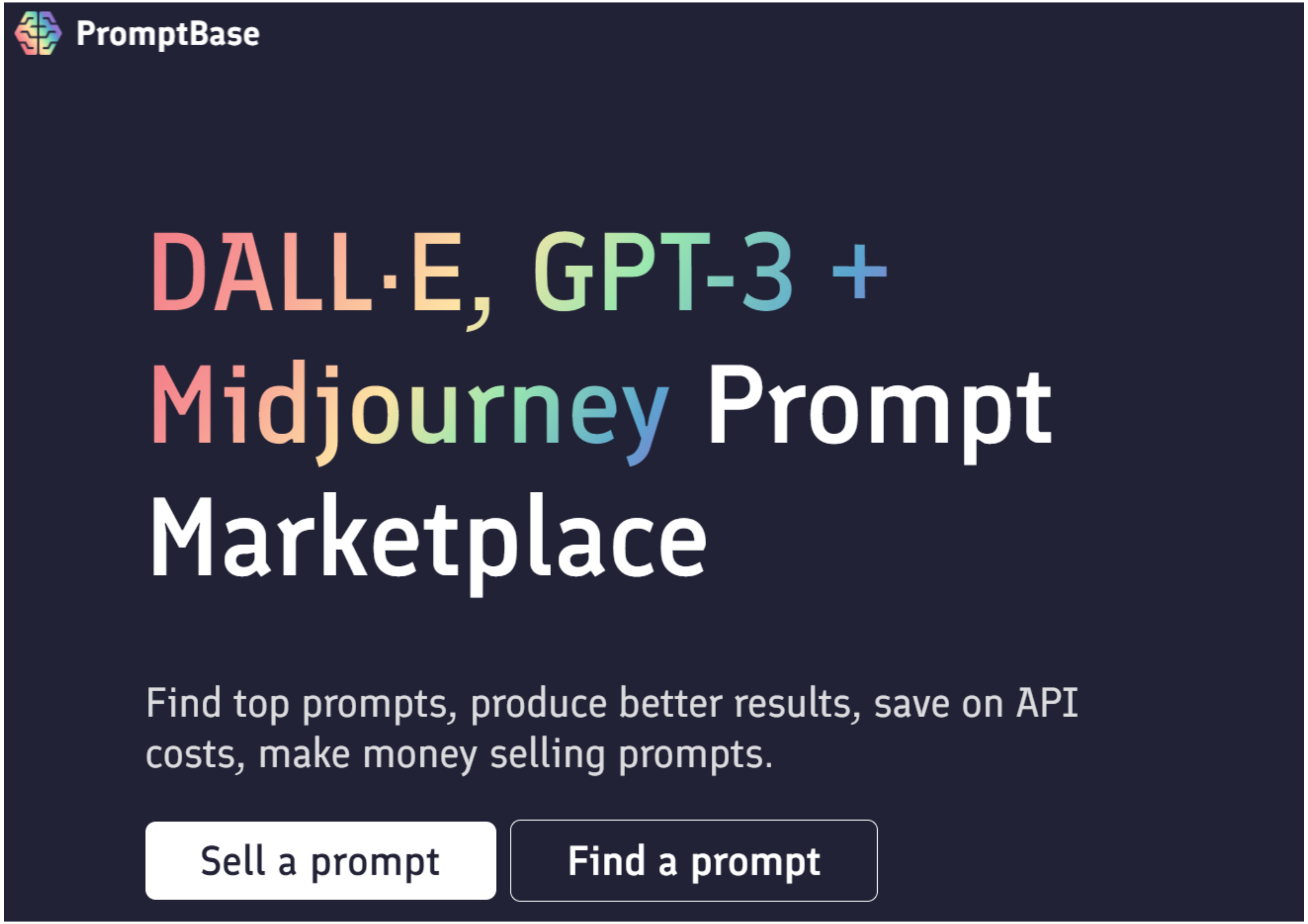
提示学习

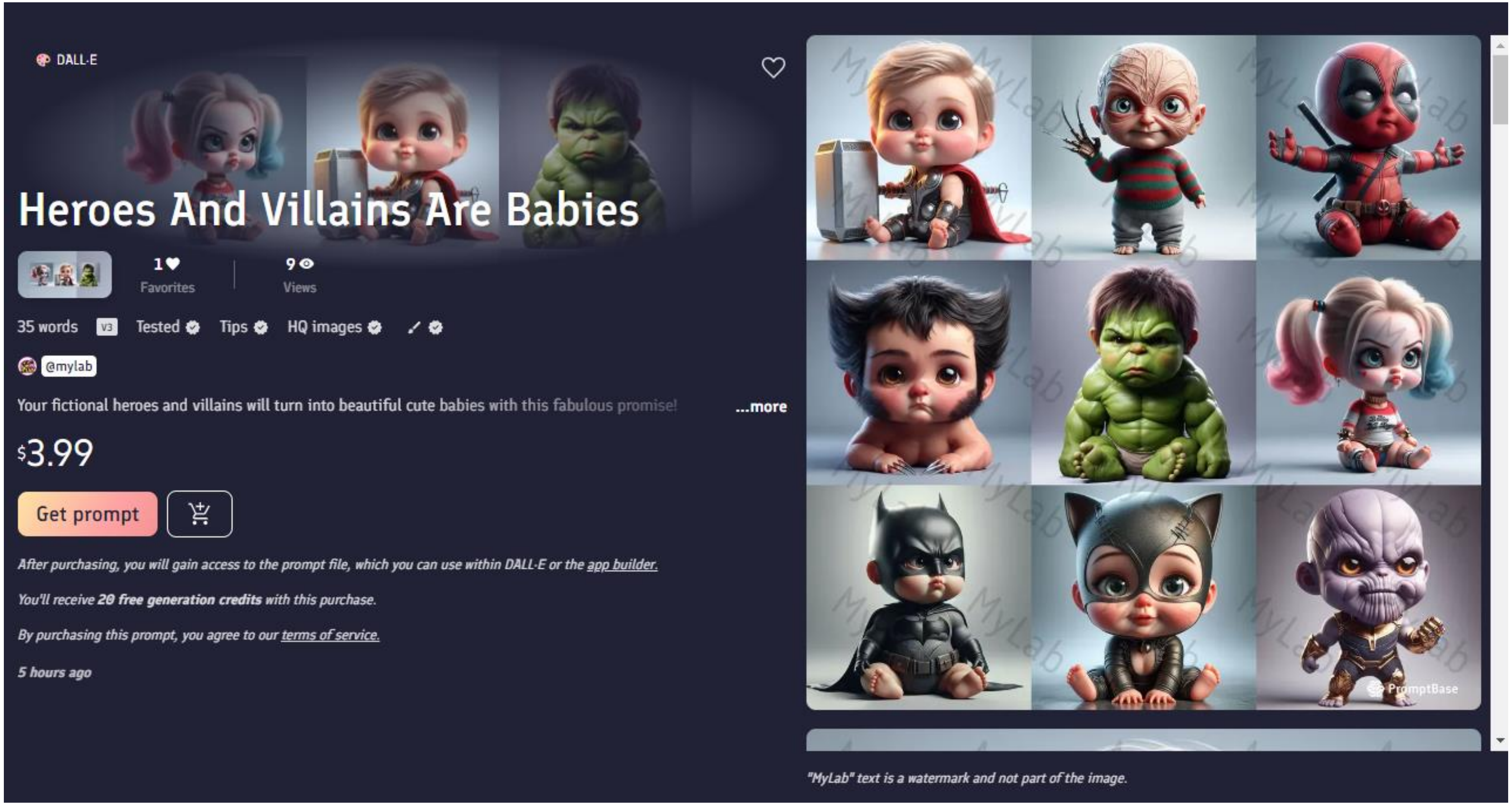
Andrej Karpathy

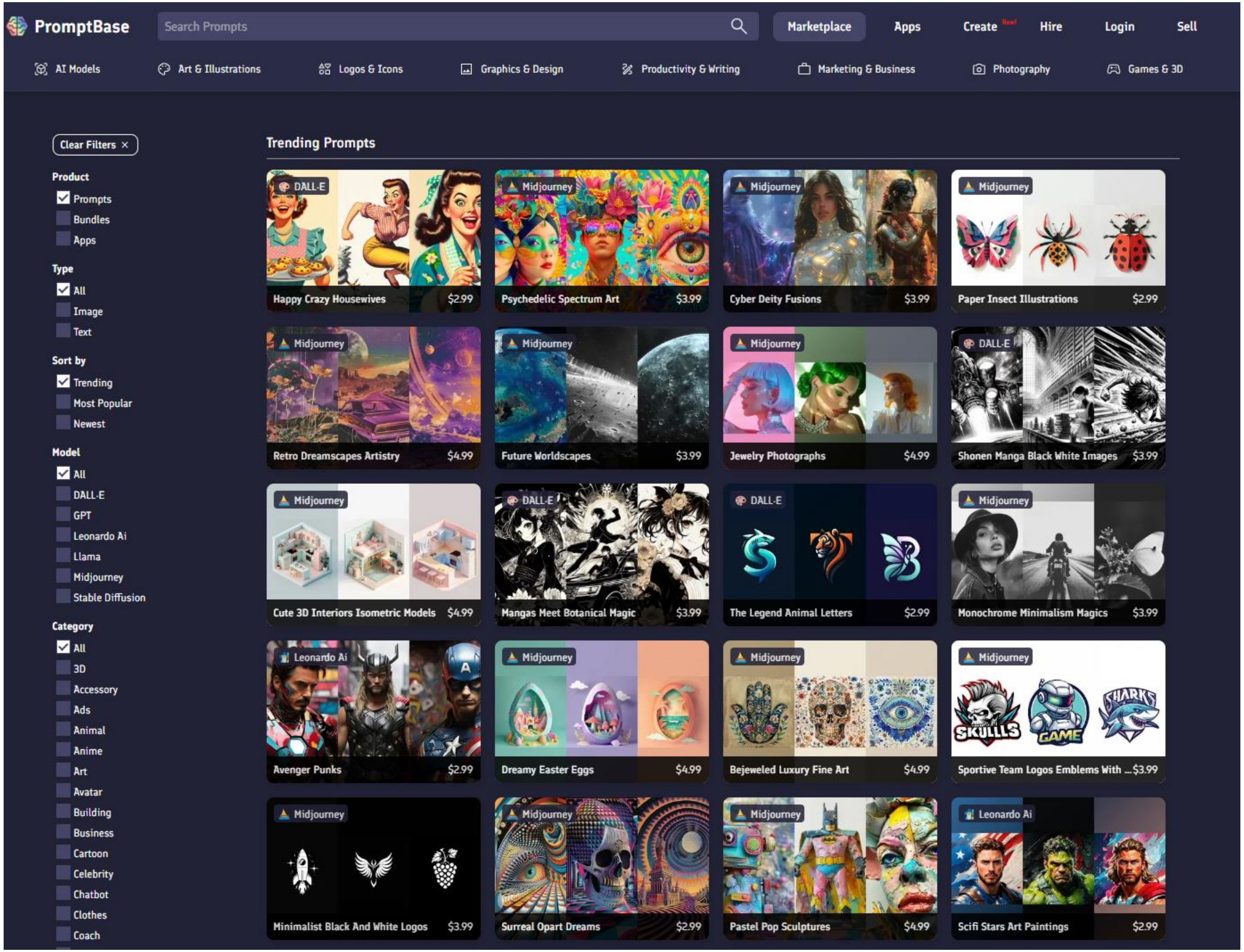
the hottest new programming language is English

未来的编程语言只会剩下两种：一种叫英文，一种叫中文

李彦宏


PromptBase
DALL·E, GPT-3 + Midjourney Prompt Marketplace
Find top prompts, produce better results, save on API costs, make money selling prompts.
Sell a prompt
Find a prompt


DALL·E
Heroes And Villains Are Babies
1
Favorites
9
Views
35 words
V3
Tested
Tips
HQ images
@mylab
Your fictional heroes and villains will turn into beautiful cute babies with this fabulous promise!
...more
$3.99
Get prompt
After purchasing, you will gain access to the prompt file, which you can use within DALL·E or the app builder.
You'll receive 20 free generation credits with this purchase.
By purchasing this prompt, you agree to our terms of service.
5 hours ago
MyLab
PromptBase
"MyLab" text is a watermark and not part of the image.


PromptBase
Search Prompts
Marketplace
Apps
Create
Hire
Login
Sell
AI Models
Art & Illustrations
Logos & Icons
Graphics & Design
Productivity & Writing
Marketing & Business
Photography
Games & 3D
Clear Filters ×
Product
Prompts
Bundles
Apps
Type
All
Image
Text
Sort by
Trending
Most Popular
Newest
Model
All
DALL-E
GPT
Leonardo Ai
Llama
Midjourney
Stable Diffusion
Category
All
3D
Accessory
Ads
Animal
Anime
Art
Avatar
Building
Business
Cartoon
Celebrity
Chatbot
Clothes
Coach
Trending Prompts
DALL-E Happy Crazy Housewives $2.99
Midjourney Psychedelic Spectrum Art $3.99
Midjourney Cyber Deity Fusions $3.99
Midjourney Paper Insect Illustrations $2.99
Midjourney Retro Dreamscapes Artistry $4.99
Midjourney Future Worldscapes $3.99
Midjourney Jewelry Photographs $4.99
DALL-E Shonen Manga Black White Images $3.99
Midjourney Cute 3D Interiors Isometric Models $4.99
DALL-E Mangas Meet Botanical Magic $3.99
DALL-E The Legend Animal Letters $2.99
Midjourney Monochrome Minimalism Magics $3.99
Leonardo Ai Avenger Punks $2.99
Midjourney Dreamy Easter Eggs $4.99
Midjourney Bejeweled Luxury Fine Art $4.99
Midjourney Sportive Team Logos Emblems With ... $3.99
Midjourney Minimalist Black And White Logos $3.99
Midjourney Surreal Opart Dreams $2.99
Midjourney Pastel Pop Sculptures $4.99
Leonardo Ai Scifi Stars Art Paintings $2.99

What is the “Prompt”?

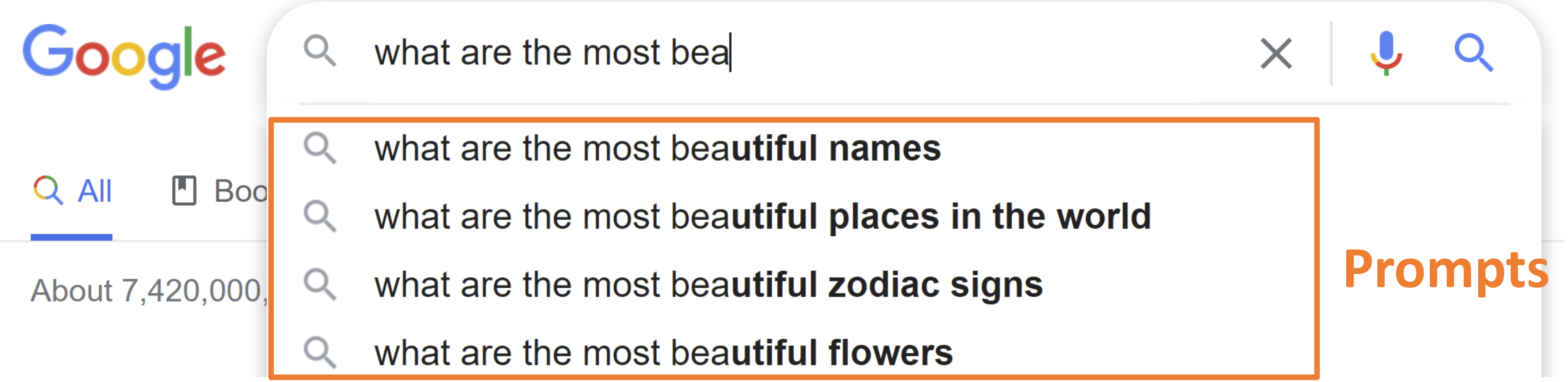
Prompt meaning prŏmpt

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

Pretrained Language Models (PLMs) and Downstream Task Models

Stages	Downstream Task Models	Pre-trained LMs	Reasons
Traditional machine learning			No pre-training language model

PLMs and Downstream Task Models

Stages	Downstream Task Models — 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

PLMs and Downstream Task Models

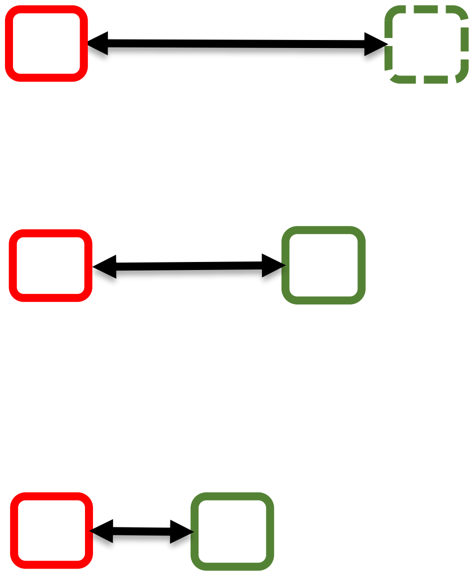
Stages

Traditional machine learning

Neural network methods enhanced by word2vec

The fine-tune method represented by BERT

Downstream Task Models **Pre-trained LMs**



Reasons

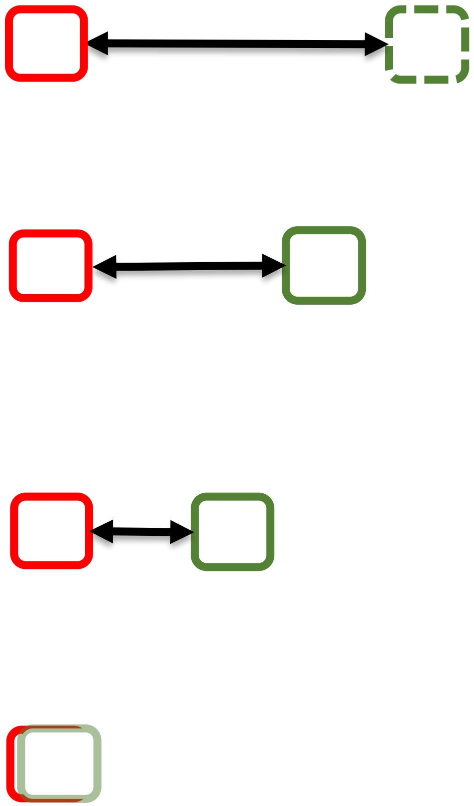
No pre-training language model

The pre-trained language model plays the role of initializing the input text signal

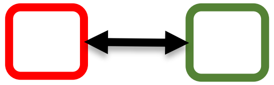
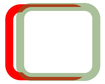
The pre-trained language model is **responsible for extracting** high-level features from the input text

PLMs and Downstream Task Models

Stages	Downstream Task Models	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

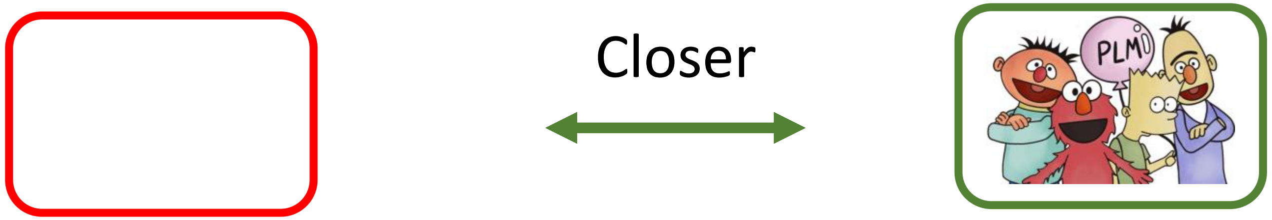


PLMs and Downstream Task Models

Stages	Downstream Task Models	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 in Modern NLP Development

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).



Downstream Task Models

Pre-trained Language Models

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

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

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

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Pengfei Liu
Carnegie Mellon University
pliu3@cs.cmu.edu

Weizhe Yuan
Carnegie Mellon University
weizhey@cs.cmu.edu

Jinlan Fu
National University of Singapore
jinlanjonna@gmail.com

Zhengbao Jiang
Carnegie Mellon University
zhengbaj@cs.cmu.edu

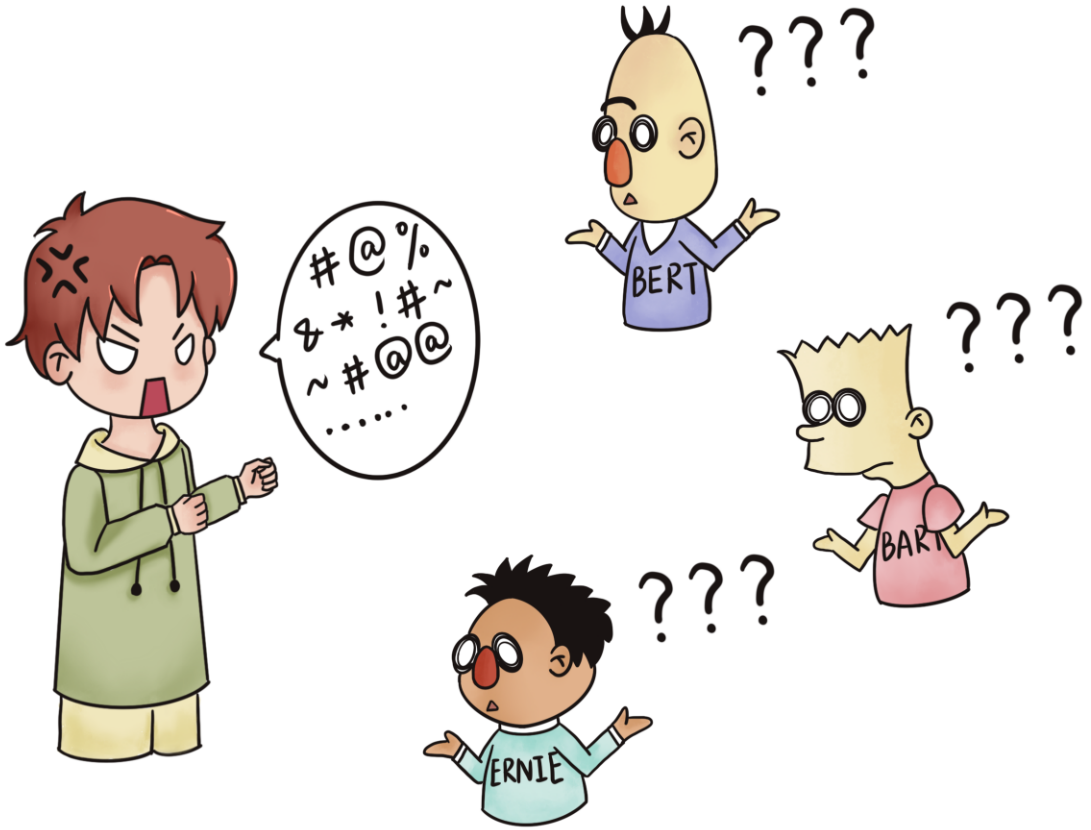
Hiroaki Hayashi
Carnegie Mellon University
hiroakih@cs.cmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing 2867 2022
P Liu, W Yuan, J Fu, Z Jiang, H Hayashi, G Neubig
ACM Computing Surveys

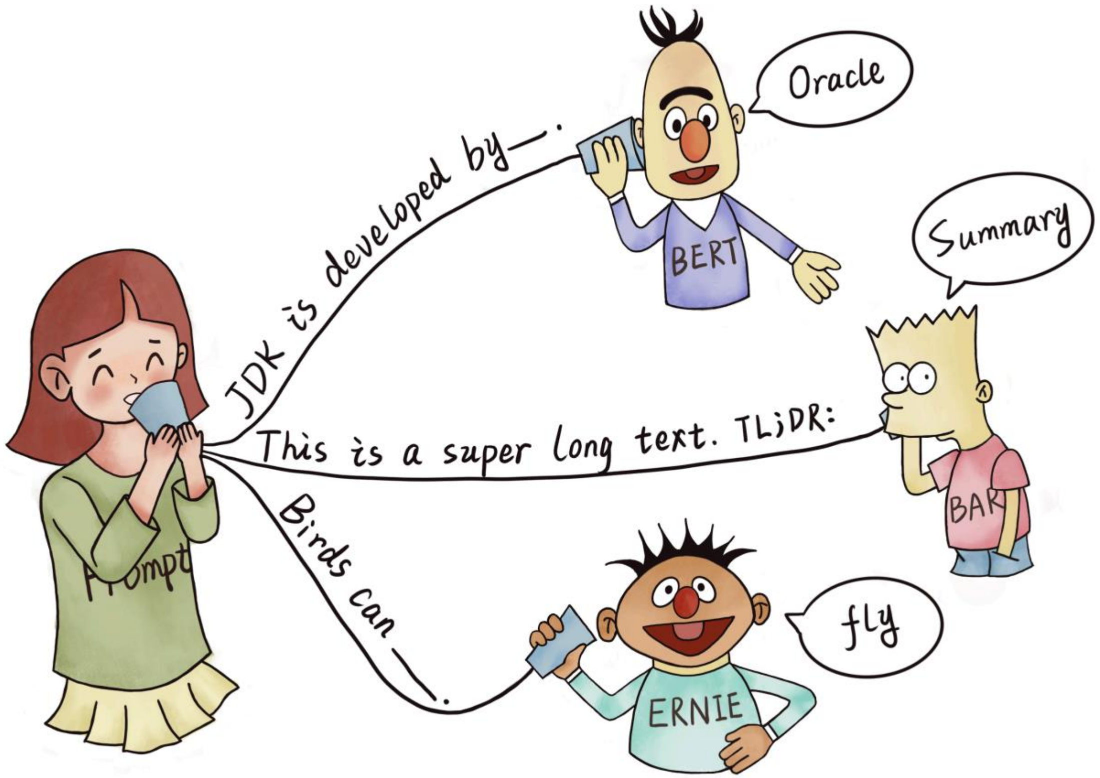
直观的定义

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



直观的定义

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



更技术层面的定义

- 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

更技术层面的定义

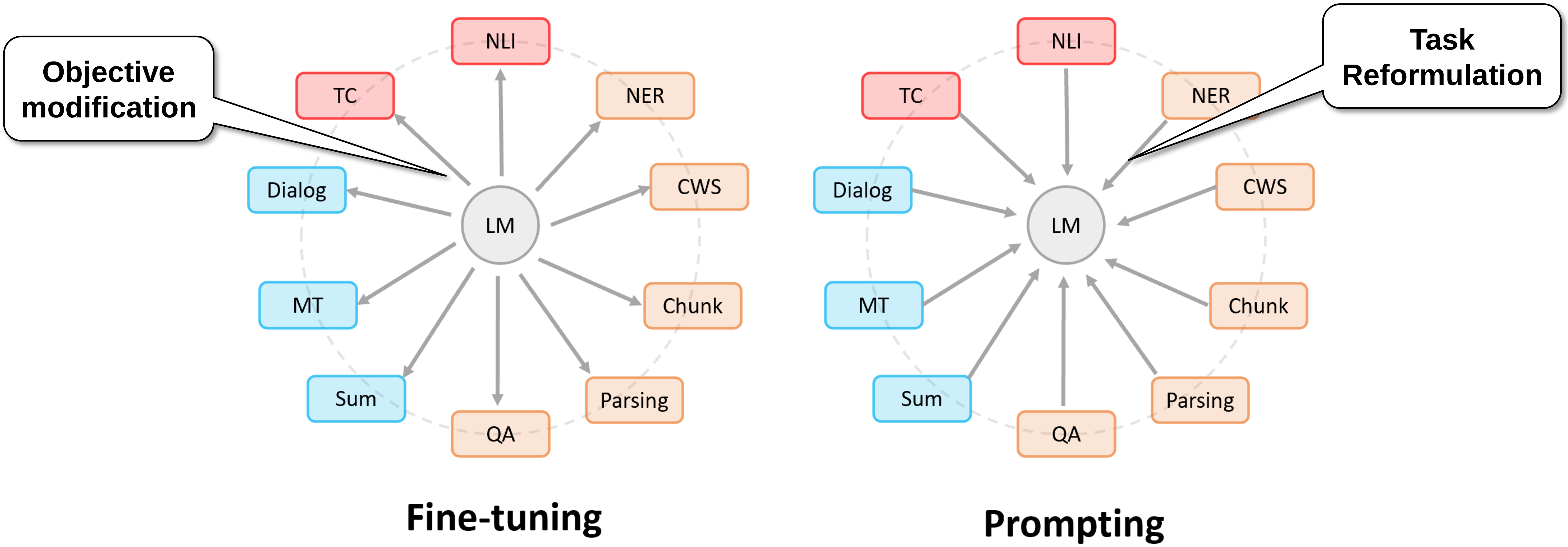
- Prompt is the technique of making better use of the knowledge from the pre-trained model by adding additional texts to the input.

purpose

Method

还有什么好处?

任务的“大一统”



Objective modification
NLI
TC
NER
Dialog
CWS
LM
MT
Chunk
Sum
Parsing
QA
Fine-tuning
Task Reformulation
NLI
TC
NER
Dialog
CWS
LM
MT
Chunk
Sum
Parsing
QA
Prompting

What is the **general workflow** of prompt-based methods?

Prompting for Sentiment Classification

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

Input: x = I love this movie.

Prompting for Sentiment Classification

- 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.

Prompting for Sentiment Classification

- 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.

Prompting for Sentiment Classification

- Transform x into prompt x’ through following two steps:
  
  - Defining a template with two slots: [x] and [z];
  - Instantiate slot [x] with input text

Input: x = I love this movie.

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

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

Prompting for Sentiment Classification

- Build a mapping function between answers and class labels.

- ☺ -> fantastic
- ☹ -> boring

label *answer*

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.

Prompting for Sentiment Classification

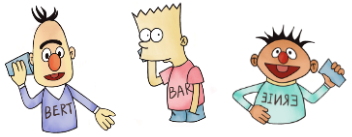
- Given a prompt，predict the answer [z].
  - Choose a suitable pretrained language model;

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?

Prompting for Sentiment Classification

- 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.

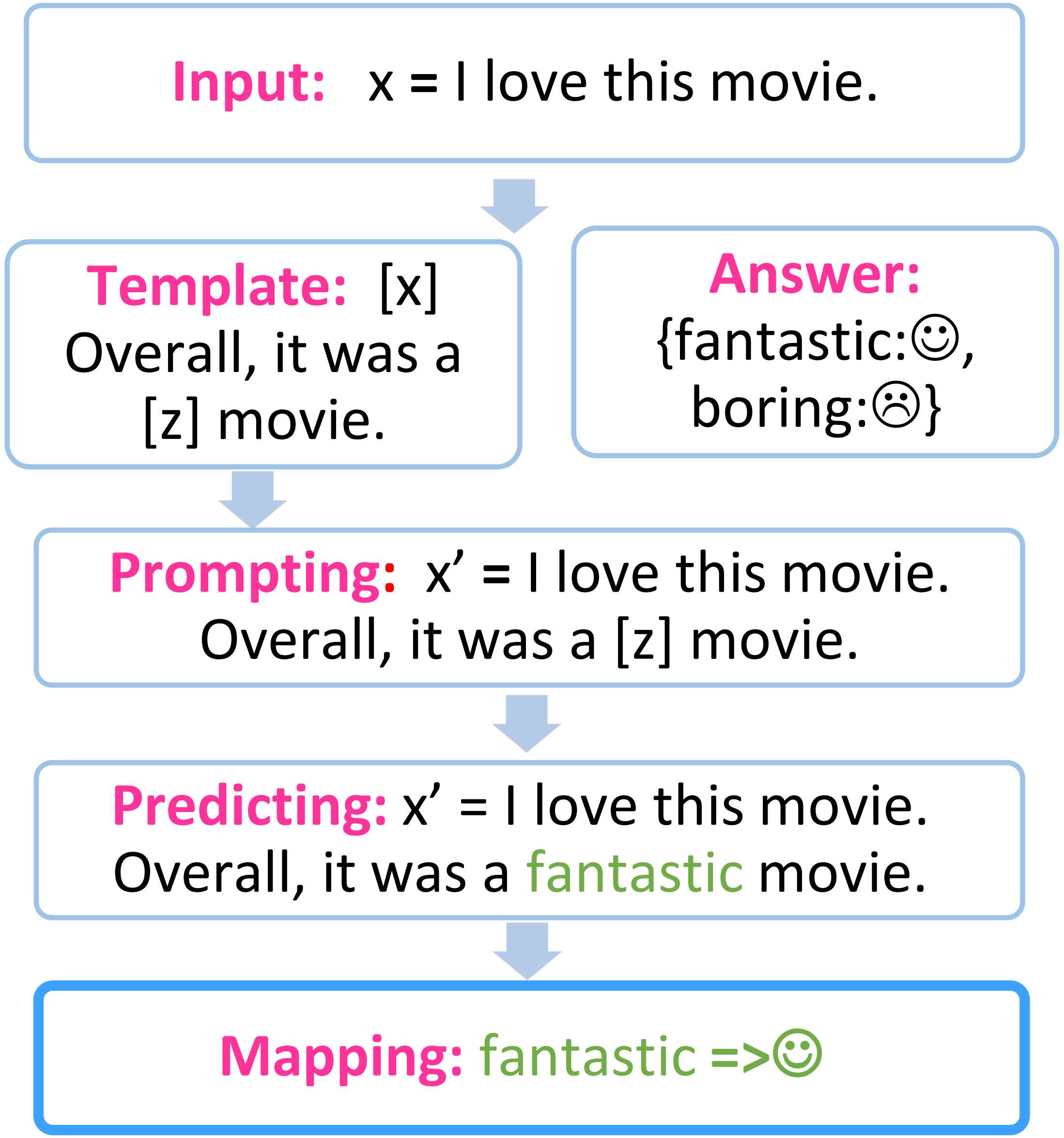
Answer: {fantastic:☺, boring:☹}

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

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

Prompting for Sentiment Classification

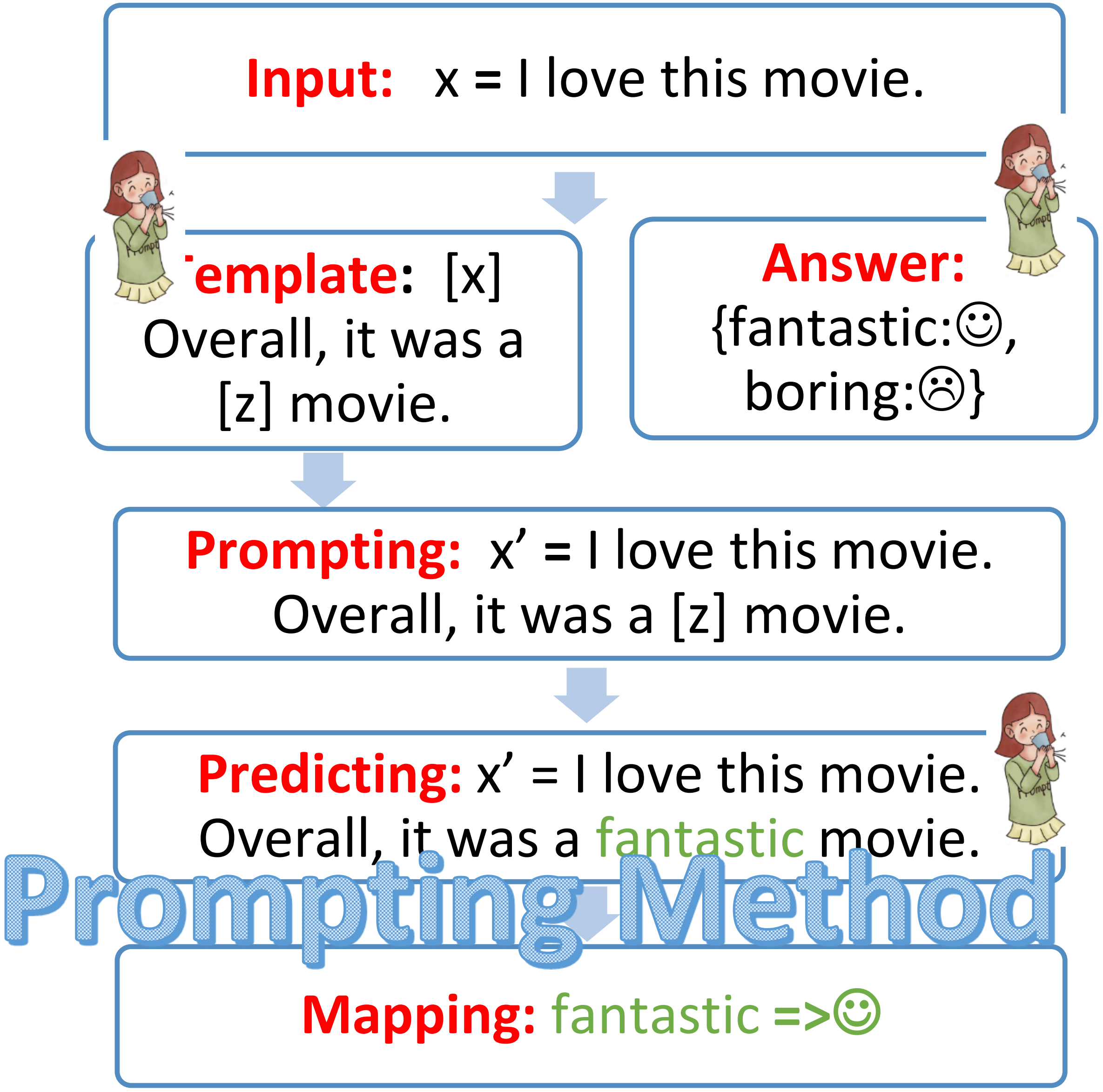
- Mapping: Given an answer, map it into a class label.
  - fantastic => ☺



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 =>☺

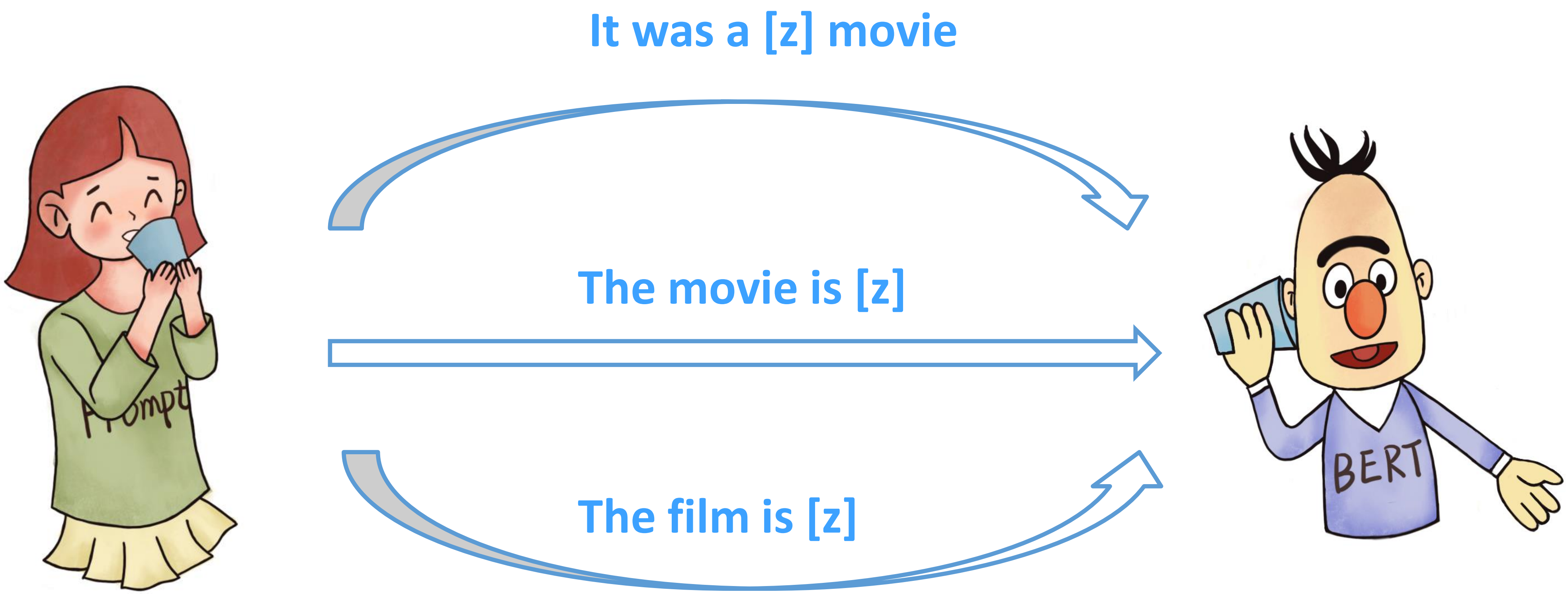
What are the design considerations for prompt-based methods?

Design Considerations for Prompt-based Methods

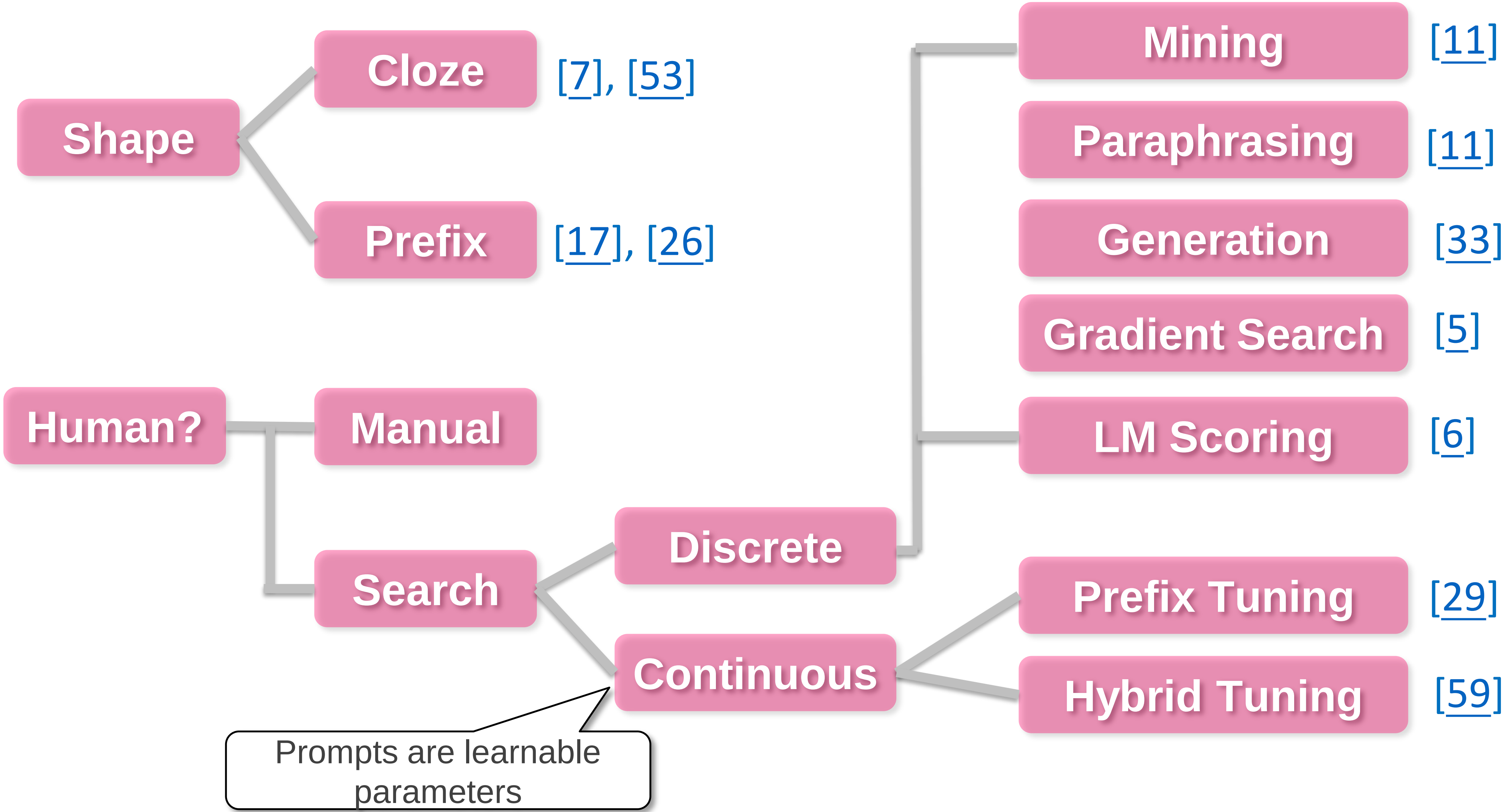
- Prompt Template Engineering
- Answer Engineering
- Pre-trained Model Choice
- Expanding the Paradigm
- Prompt-based Training Strategies

Prompt Template Engineering

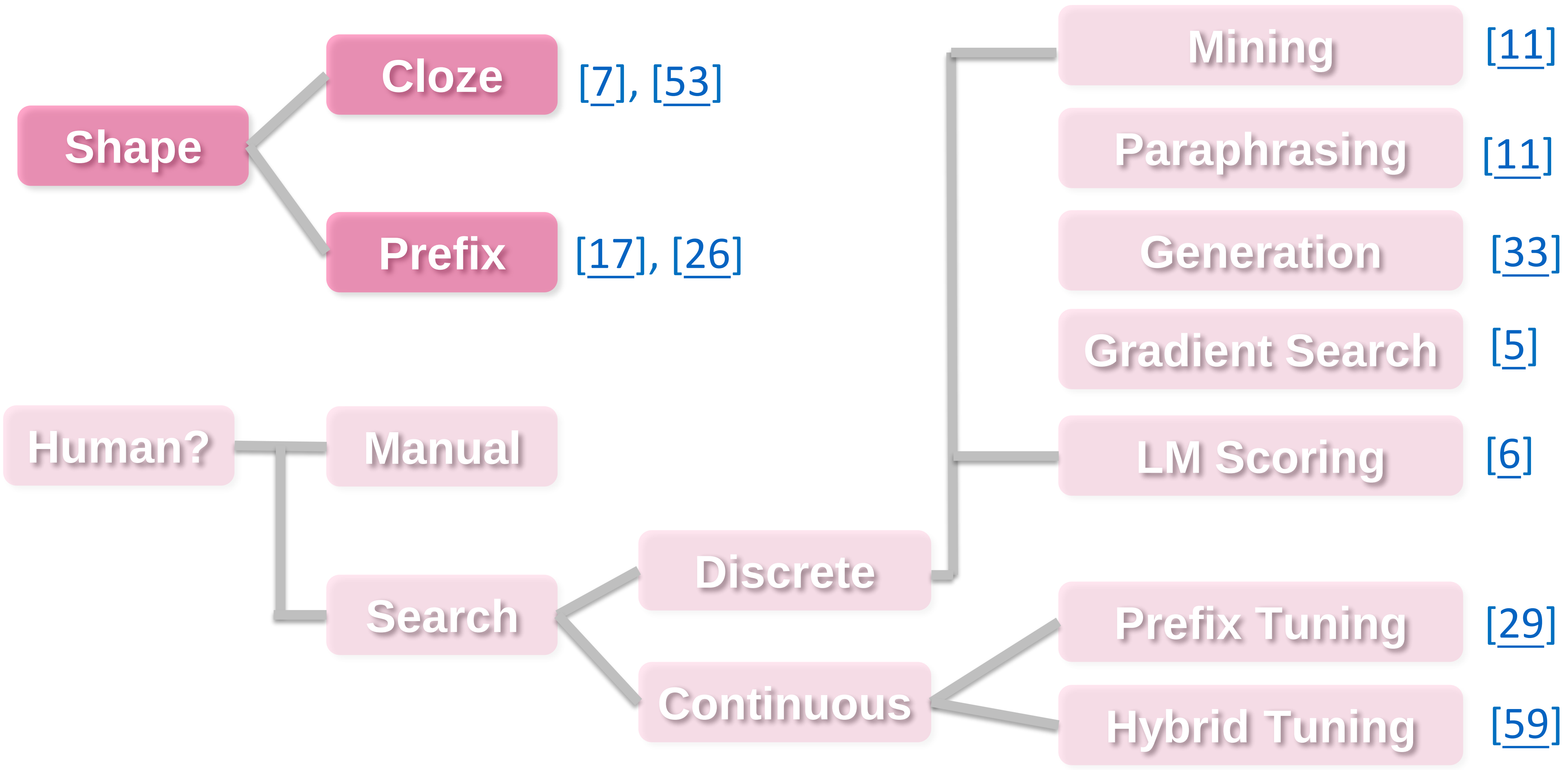
- Research Question:
  - how to define appropriate prompt templates



Design Decision of Prompt Templates



Design Decision of Prompt Templates



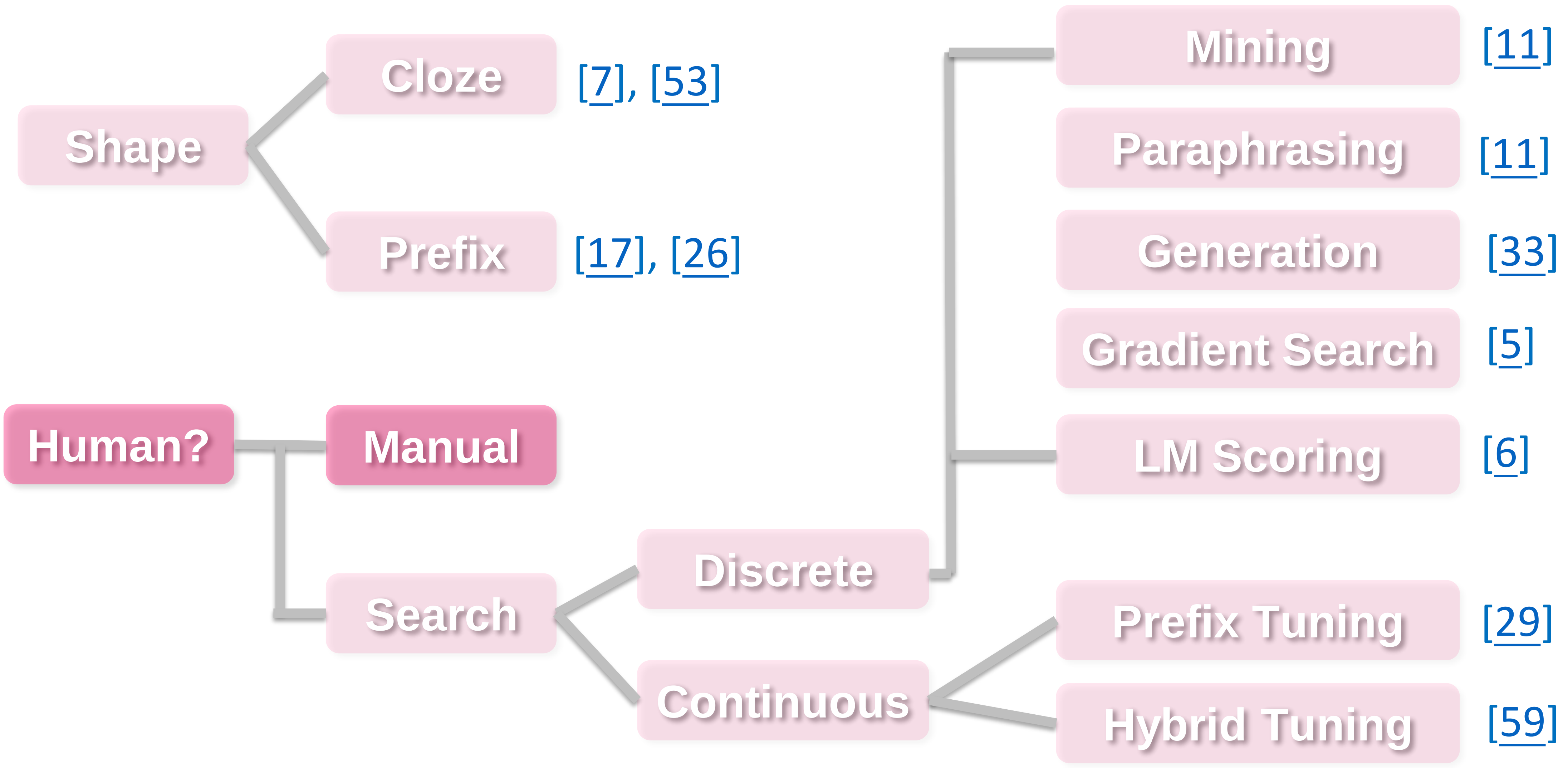
Prompt Shape

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

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: ____ "*

Prompt Shape



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 ___ .

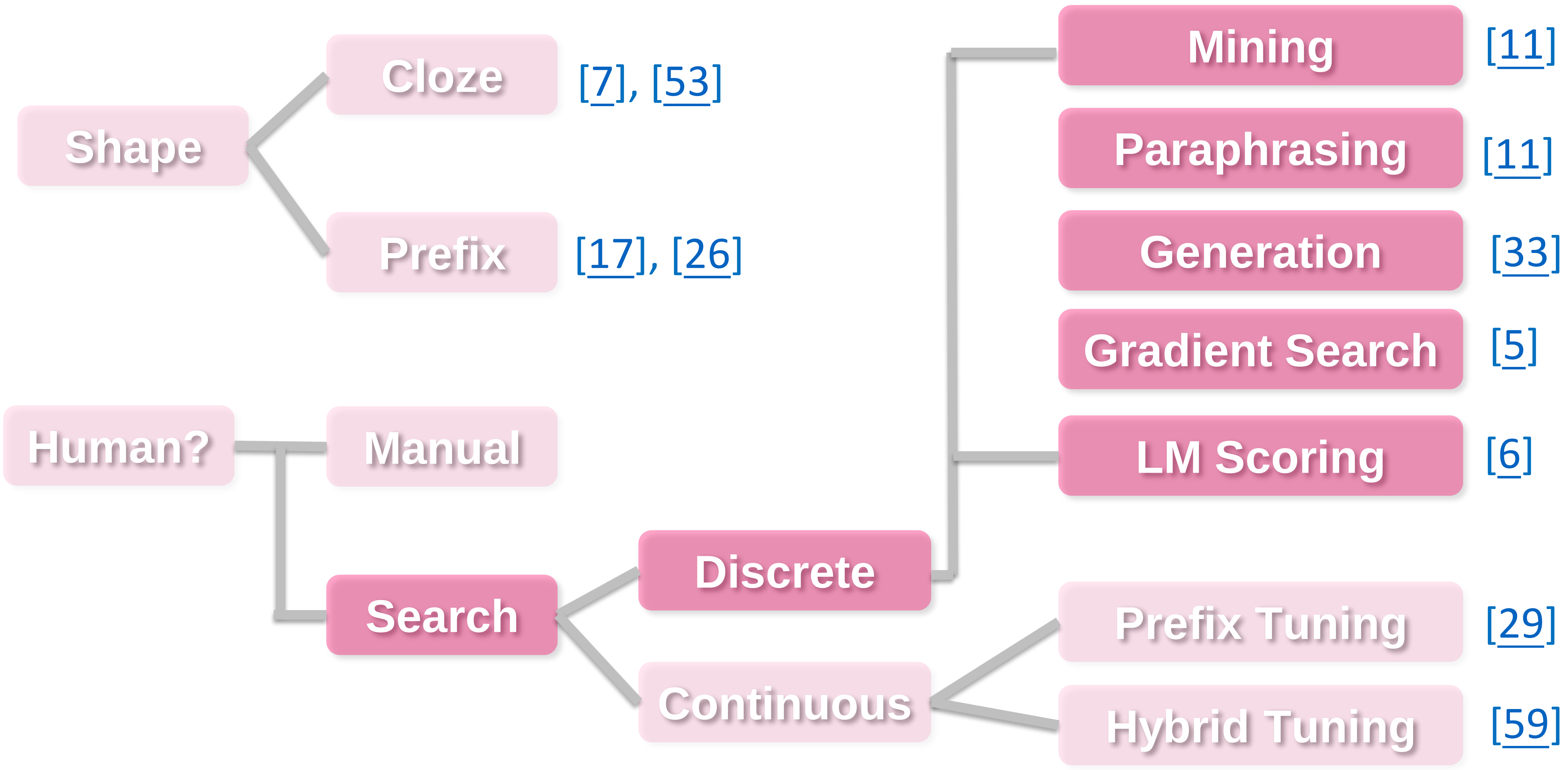
Answer: positive/negative

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, its 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)

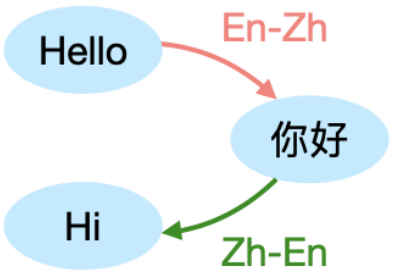
Discrete Search

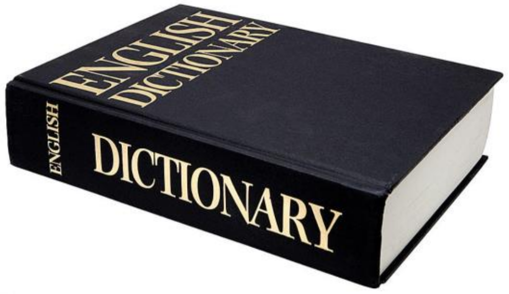
- Paraphrasing
  - Take in an existing seed template, and paraphrases it into a set of other candidate templates.

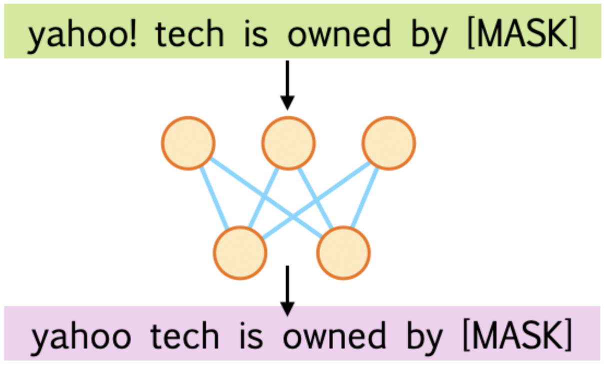
References: [1] Jiang et al. How Can We Know What Language Models Know? TACL (2020). [2] Yuan et al. BARTScore: Evaluating Generated Text as Text Generation. NeurIPS (2021). [3] Haviv et al. BERTese: Learning to Speak to BERT. EACL (2021).

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.

I love this movie! [ ] ___ .

We want the LM to predict **positive** here

The **template token** we want to search

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.

Pre-train

Input: Thank you <X> me to the party <Y> week.

Target: <X> for inviting <Y> last <Z>

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

Discrete Search

- Generation
  - Use LM to generate templates.

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

↓ T5 decode

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

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

……

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

Discrete Search

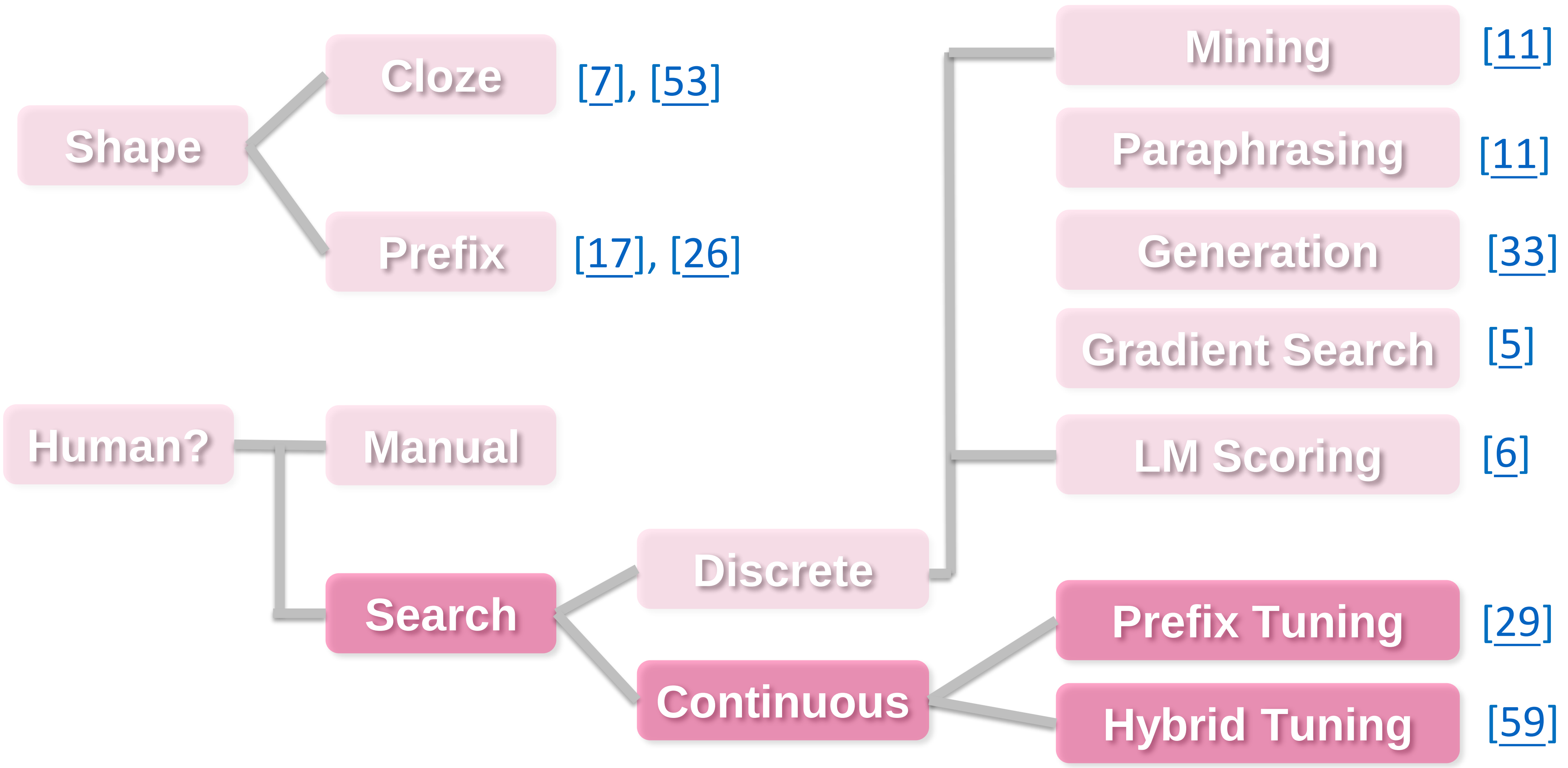
- LM Scoring
  - Use the LM to choose the templates that achieve high LM probability.

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

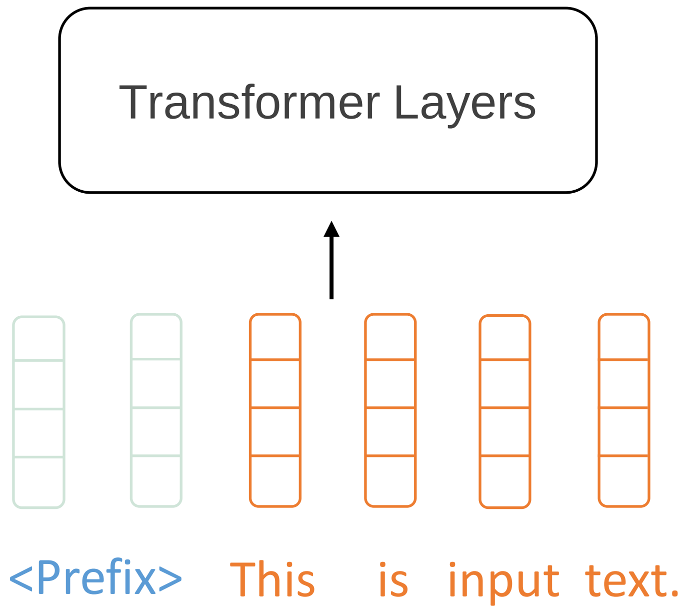
- Prefix Tuning
  - Prepends a sequence of continuous task-specific vectors to the input, while keeping the LM parameters frozen.

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

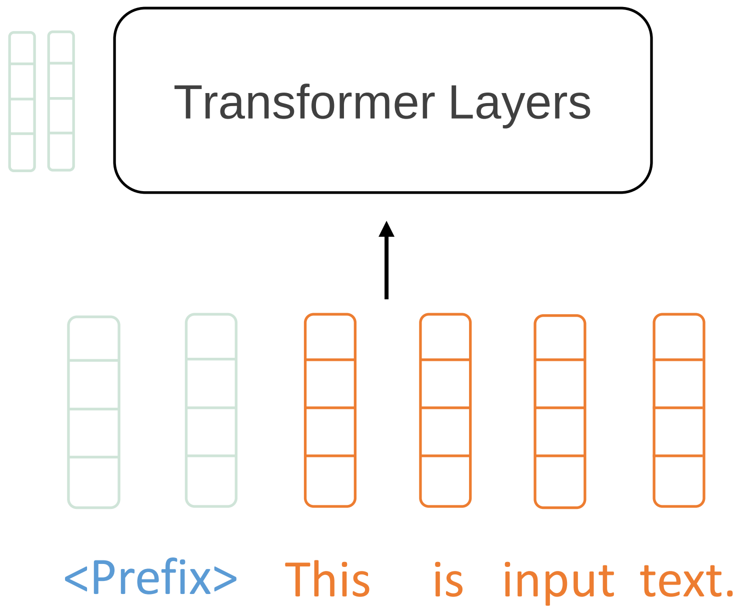


References: [1] Li et al. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190 (2021). [2] Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691 (2021)

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
  - Deep Prefix Tuning



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

- 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.

I love this movie so much! 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

I love this movie so much! $\boxed{\text{The}}$ $\boxed{\text{sentiment}}$ $\boxed{\text{is}}$ 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

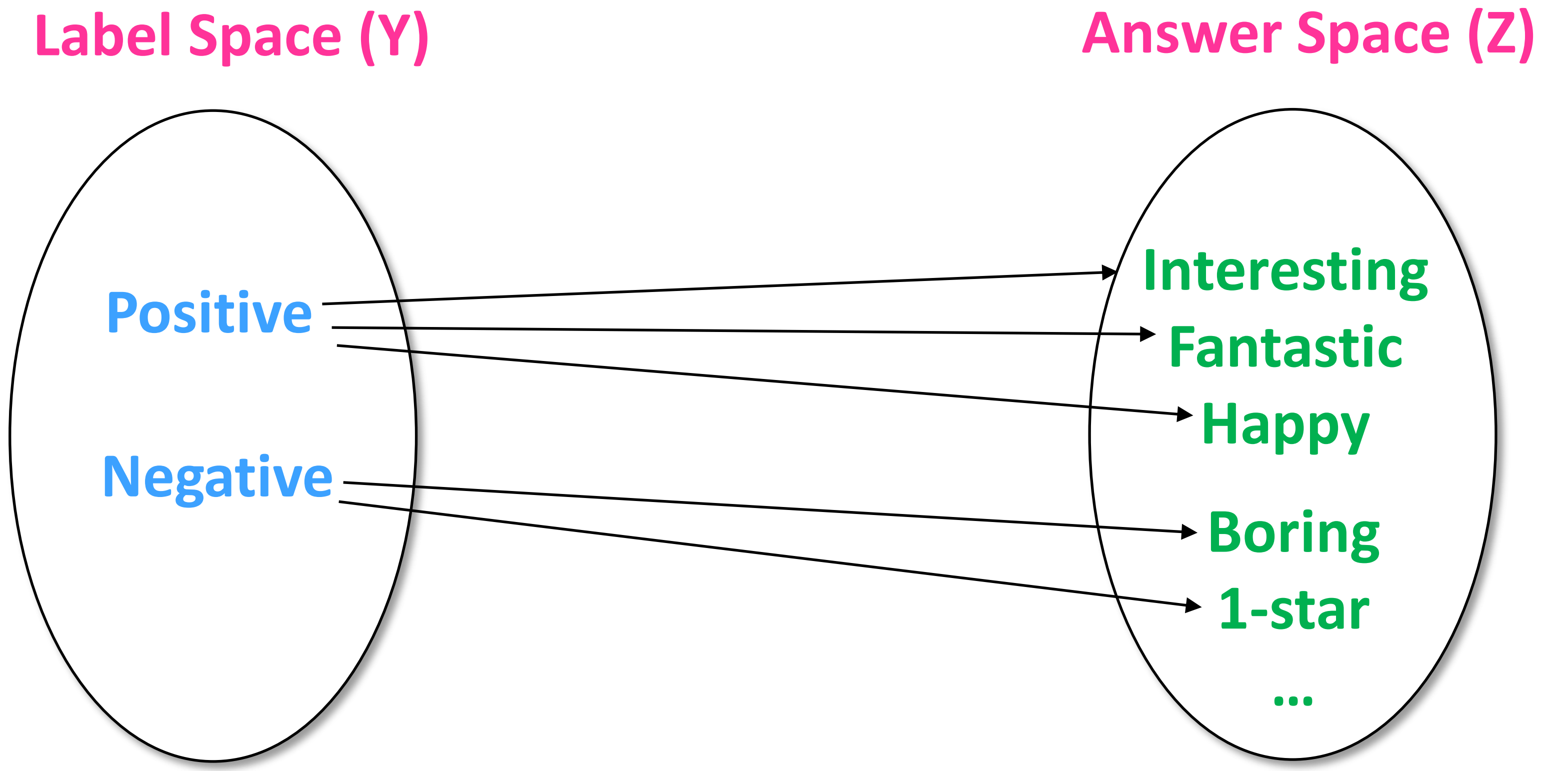
- Prompt Template Engineering
- Answer Engineering
- Pre-trained Model Choice
- Expanding the Paradigm
- Prompt-based Training Strategies

Answer Engineering

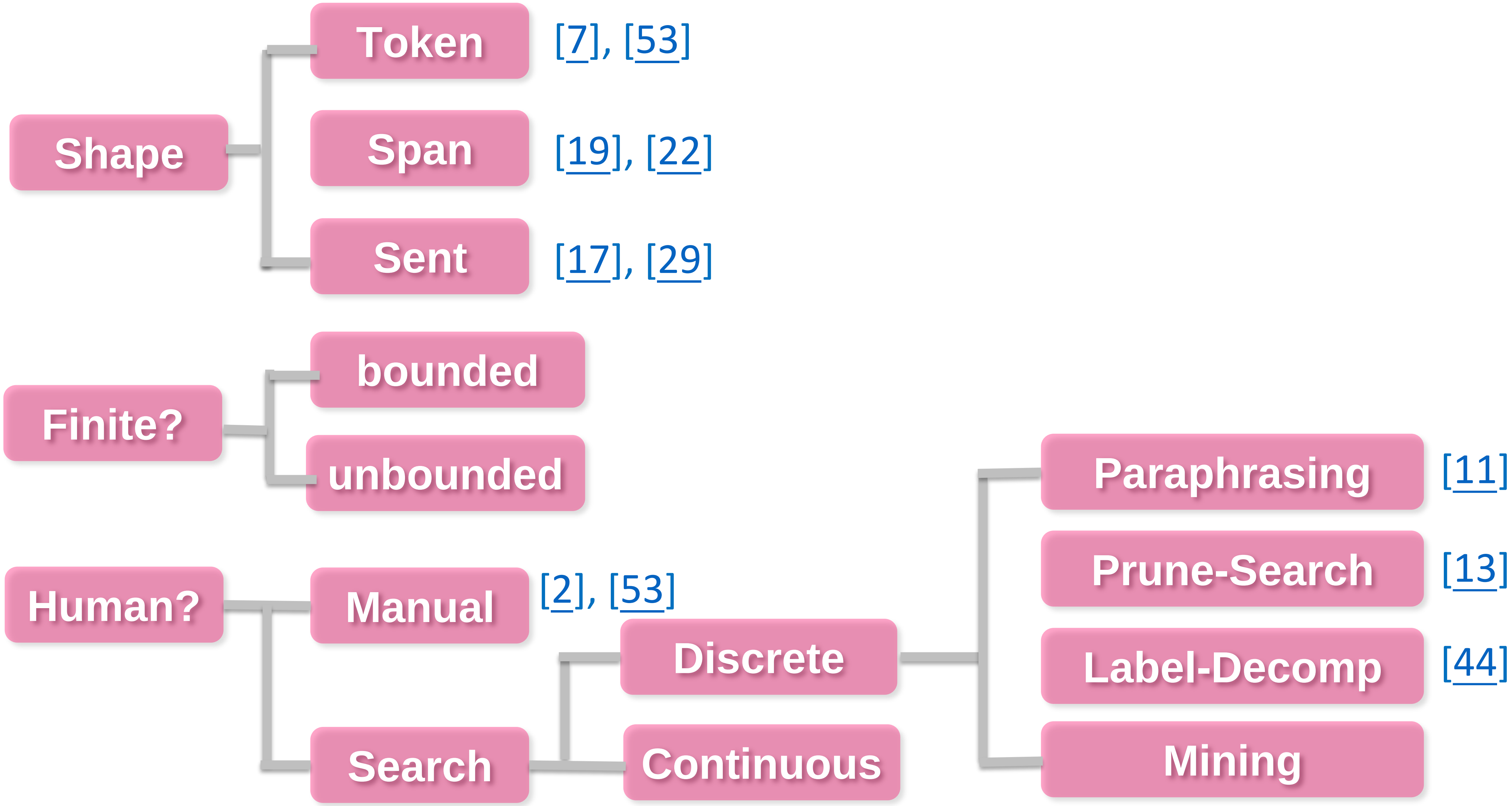
- Research Question:
  - Given a task (or a prompt), how to define a suitable mapping function between label space and answer space?

Answer Engineering

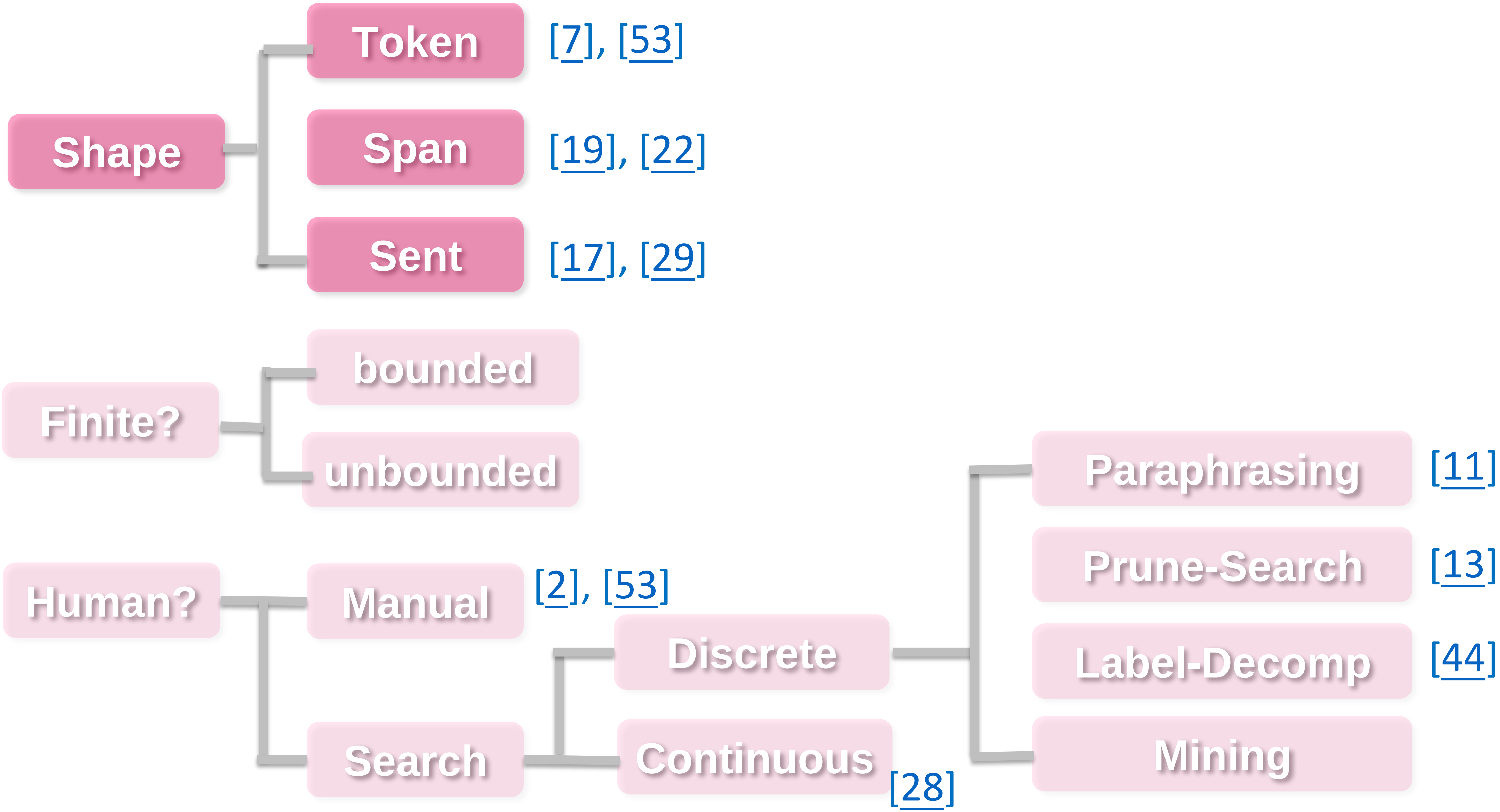
- Research Question:
  - Given a task (or a prompt), how to define a suitable mapping function between label space and answer space?



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

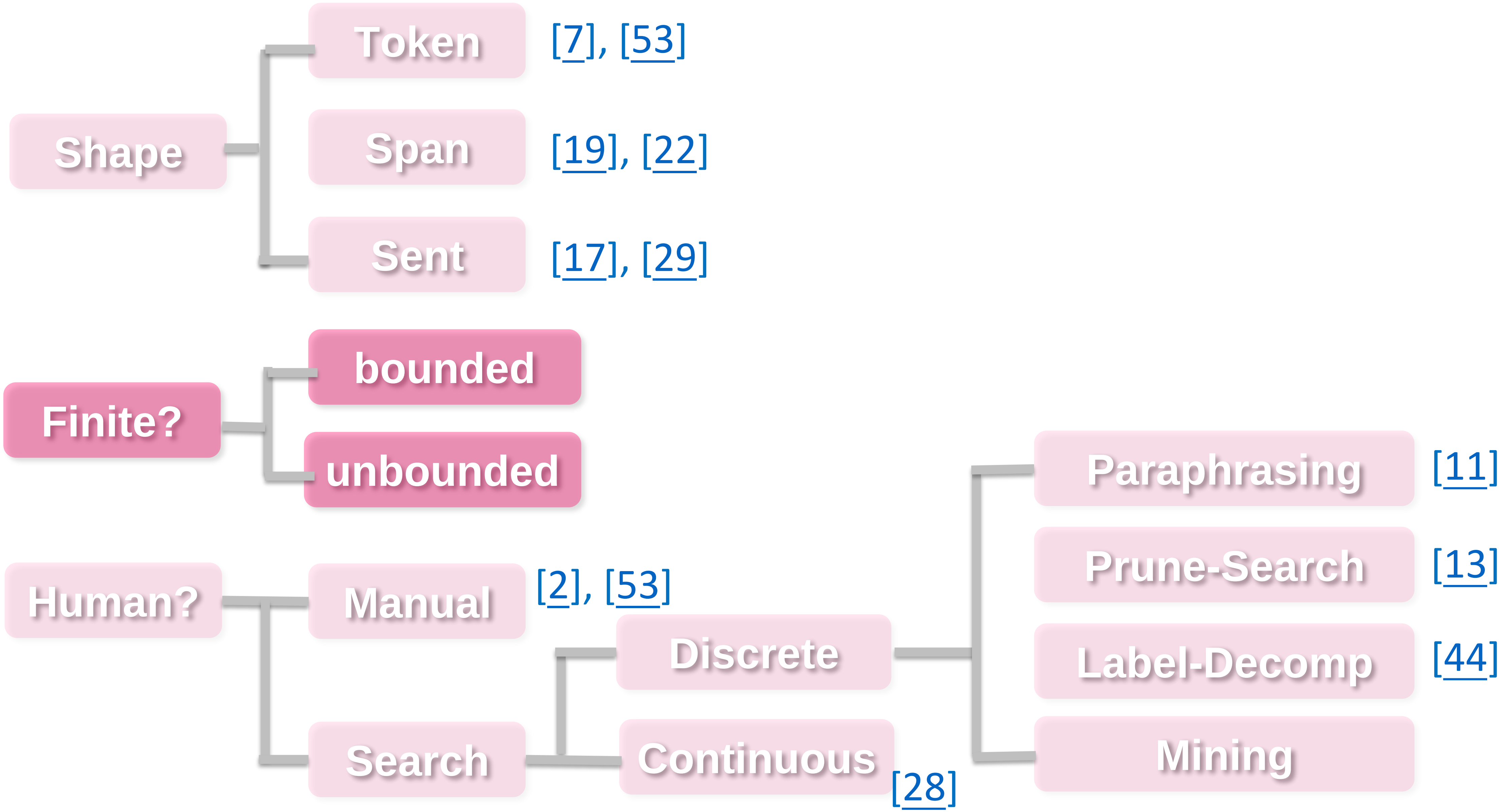
      (C) less friction [gold]

      (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



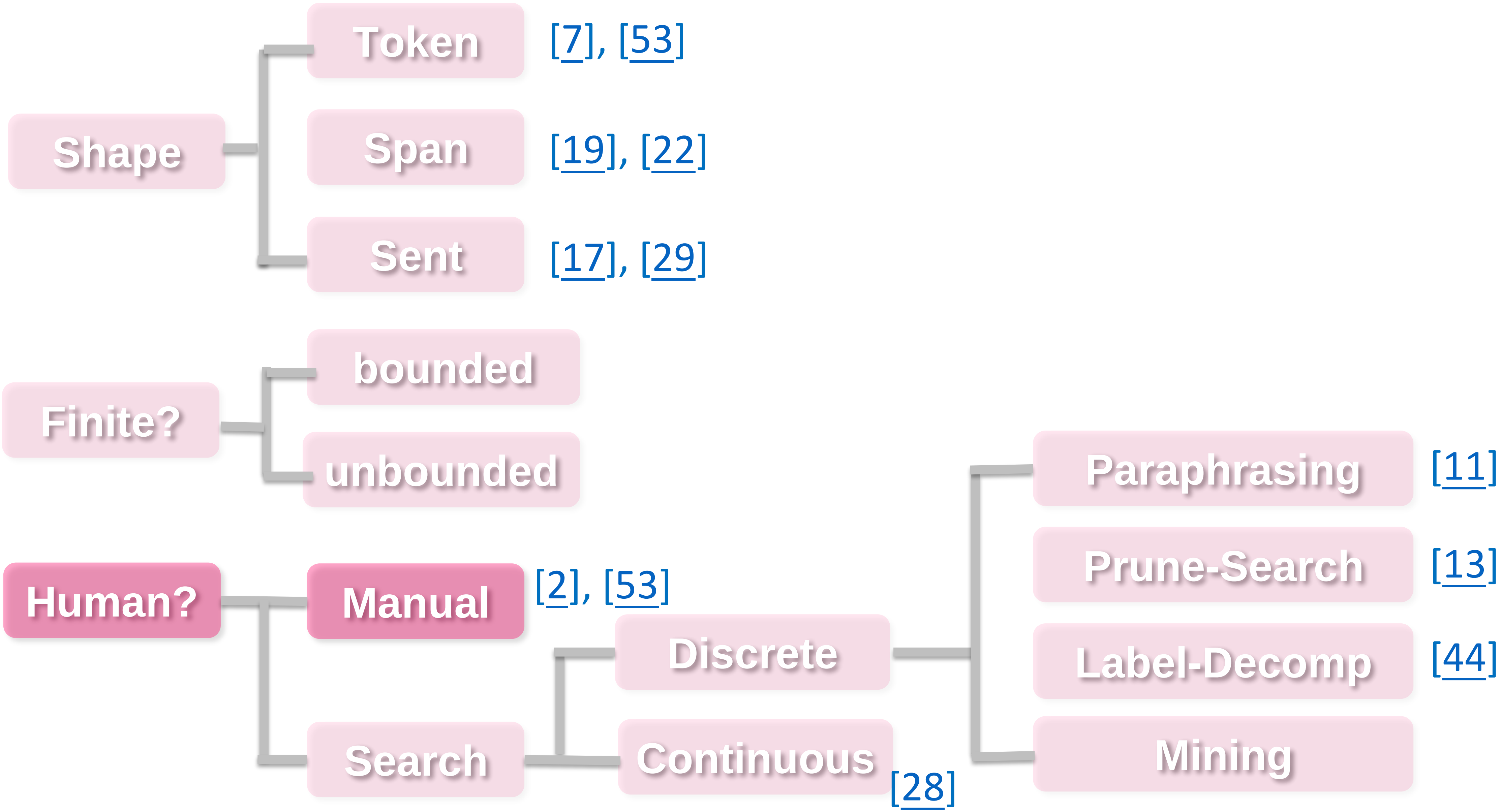
Answer Space

- 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

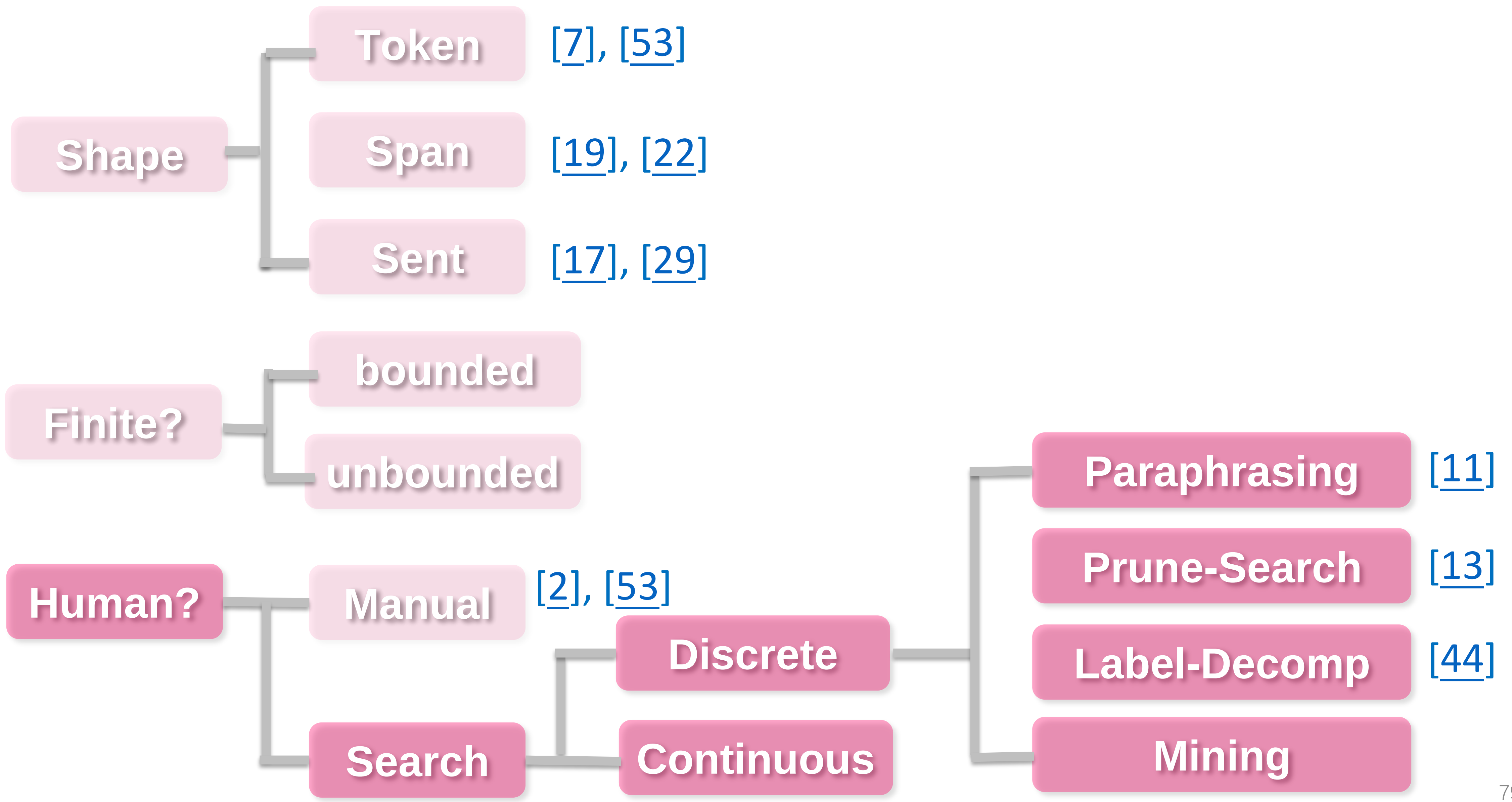
Human Design

- 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 $\xrightarrow{\text{decompose}}$ {Science, Mathematics}

  - Relation Extraction:

    city_of_death $\xrightarrow{\text{decompose}}$ {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
    - ……

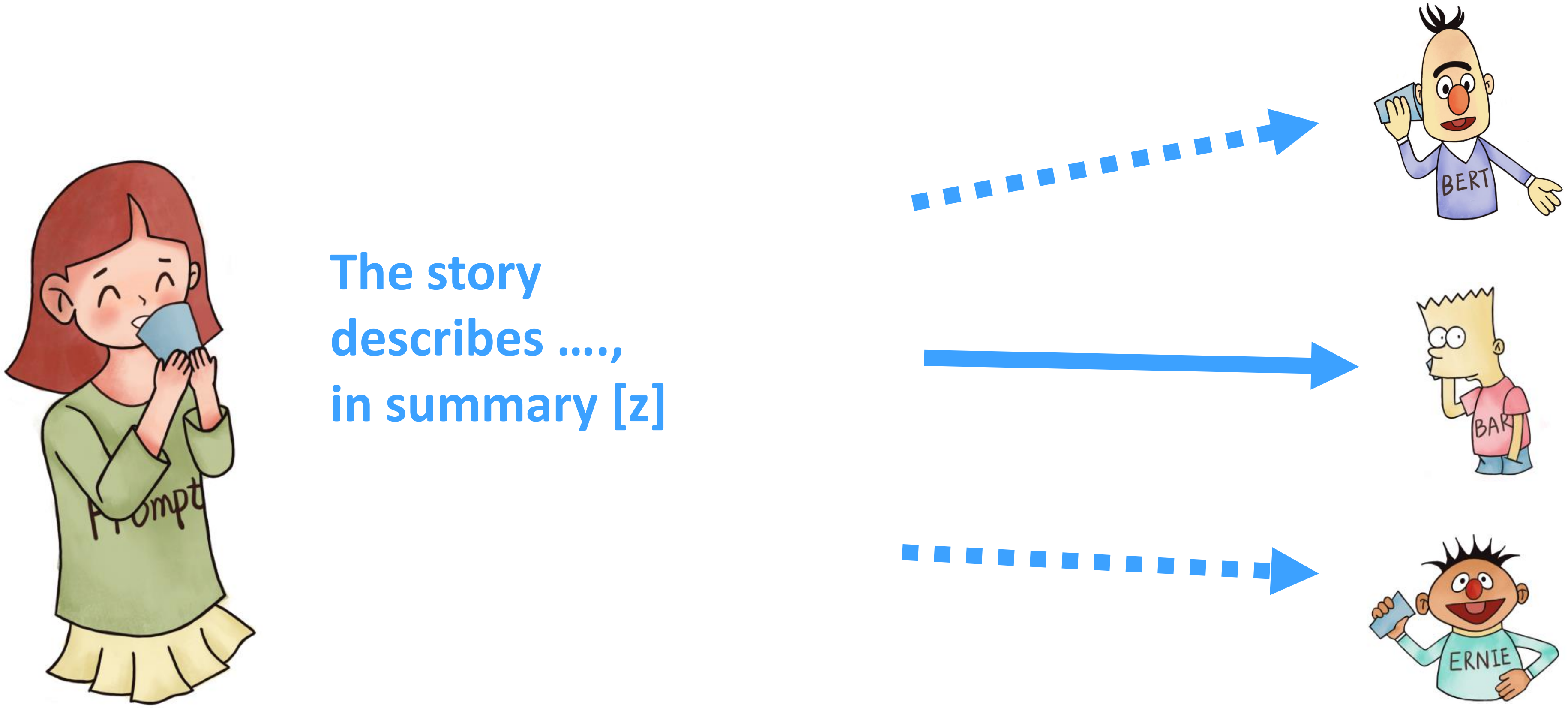
Reference: Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li and Hong-Gee Kim. 2021. Prompt-Learning for Fine-Grained Entity Typing. CoRR abs/2108.10604 (2021) .

Design Considerations for Prompt-based Methods

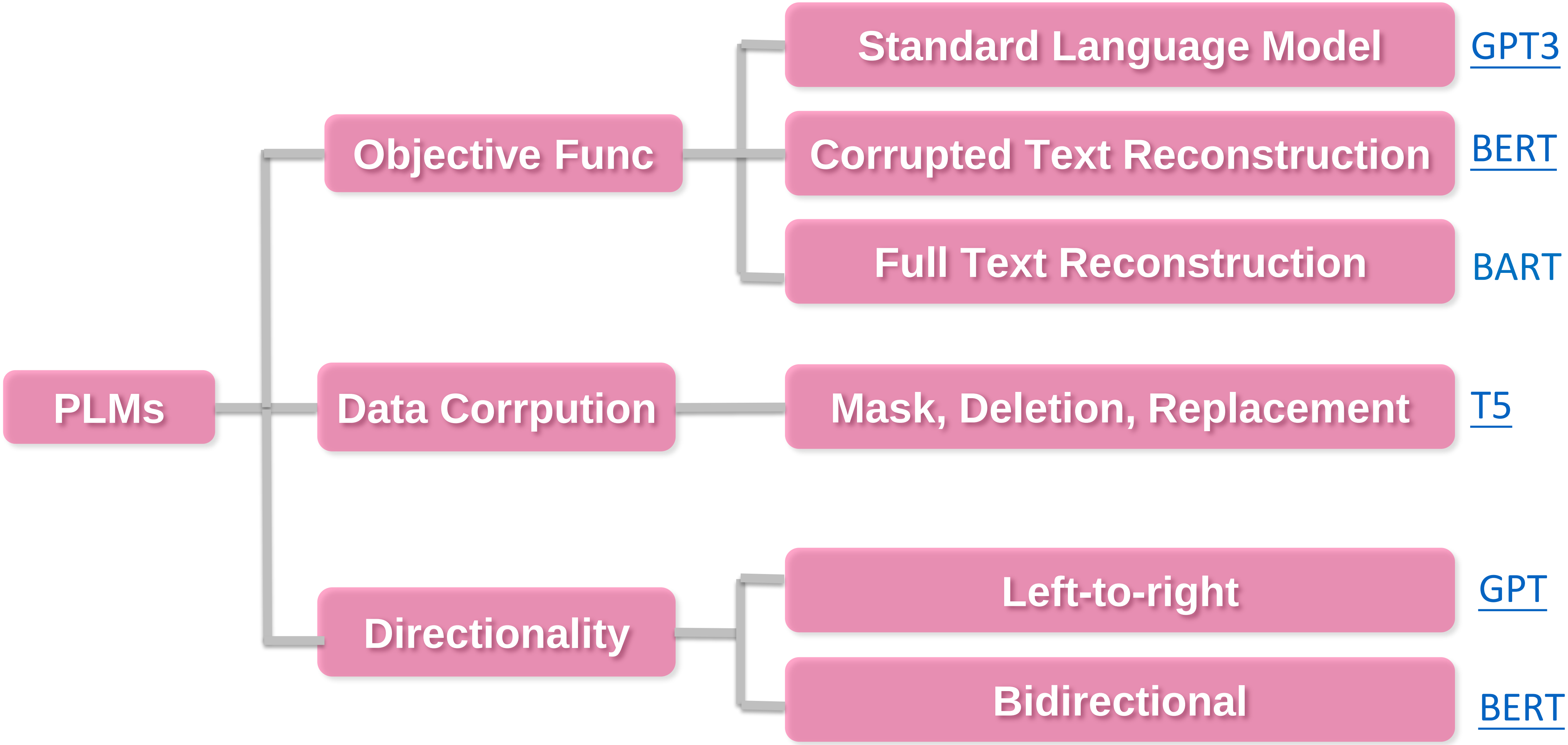
- Prompt Template Engineering
- Answer Engineering
- Pre-trained Model Choice
- Expanding the Paradigm
- Prompt-based Training Strategies

Pre-trained Model Choice

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



Design Decision of Pre-trained Models

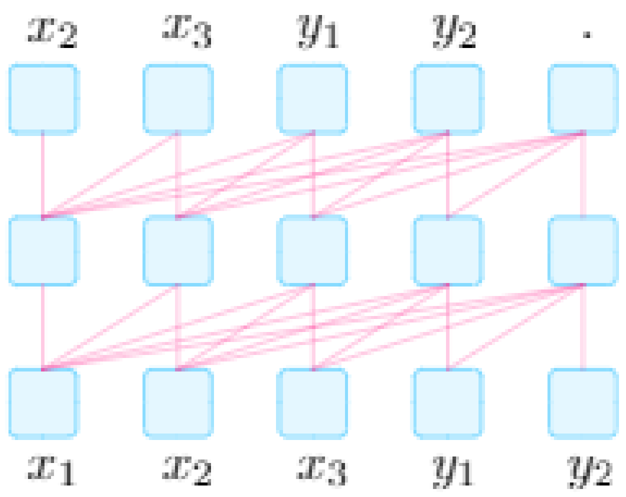


Design Decision of Pre-trained Models

- PLMs
  - Left-to-Right LM — GPT, GPT3
  - Masked LM — BERT, RoBERT
  - Encoder-decoder — BART, T5

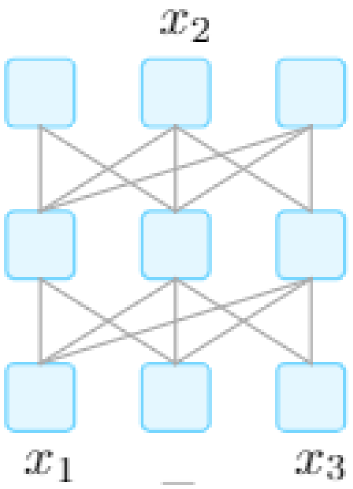
Left-to-right Language Model

- 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



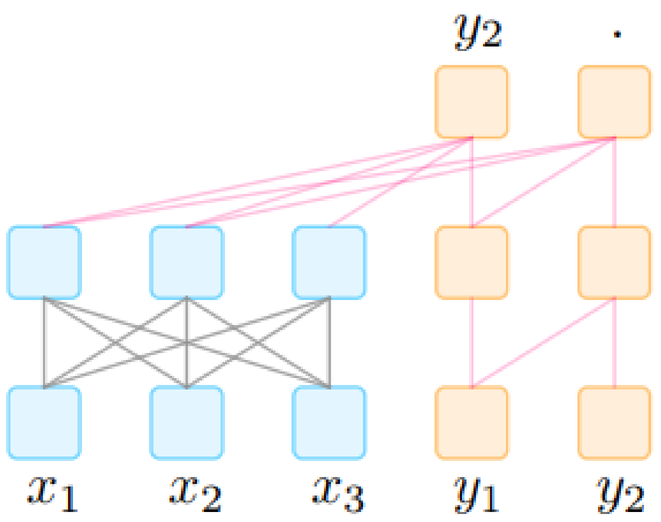
Masked Language Model

- 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

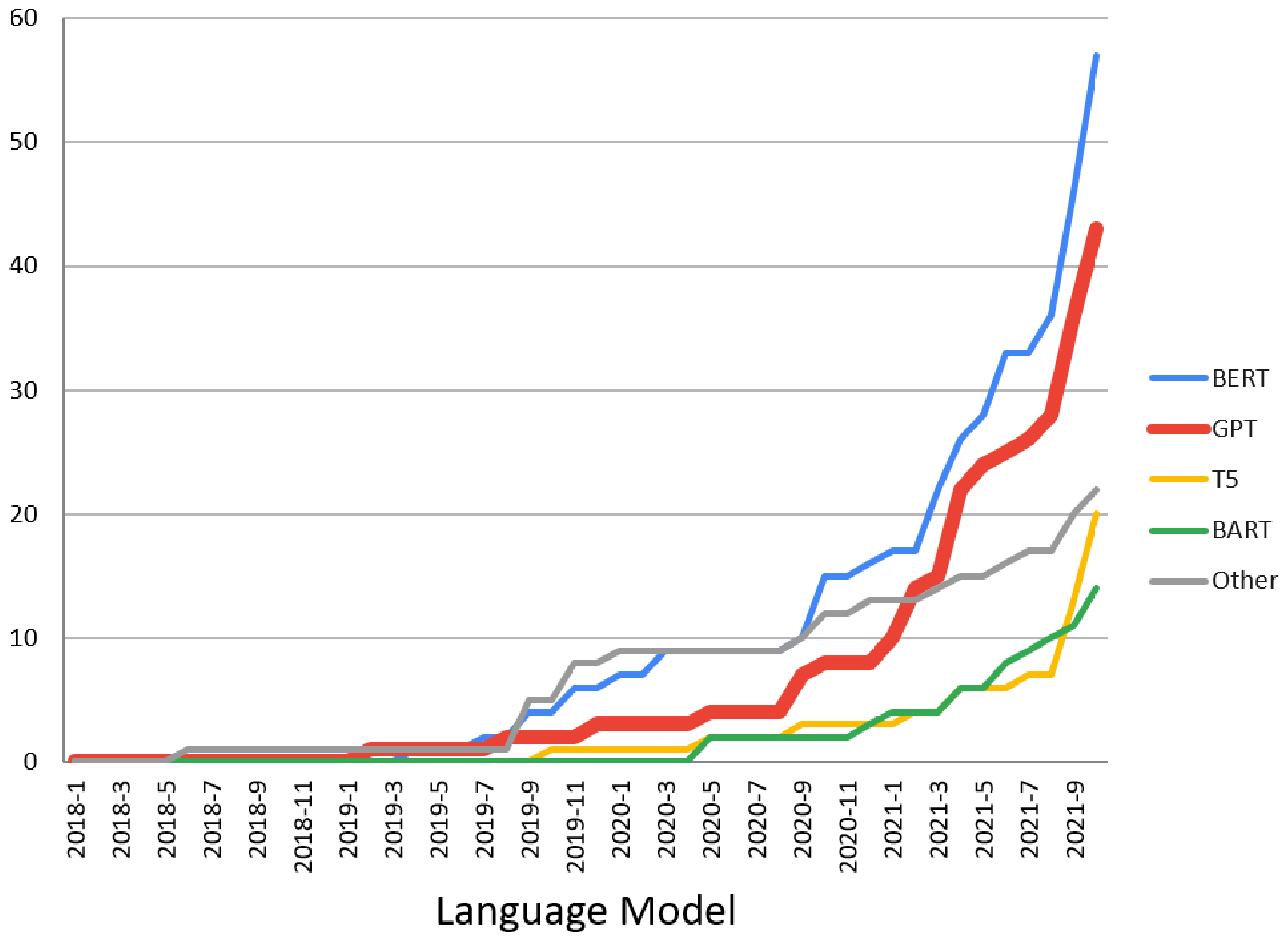


Masked Language Model

- **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



Which one is more popular?



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

- **Multi-Prompt**
  - **Prompt Ensemble** [11], [57]
  - **In-context Learning** [47], [48]
  - **Chain-of-thought** [52]

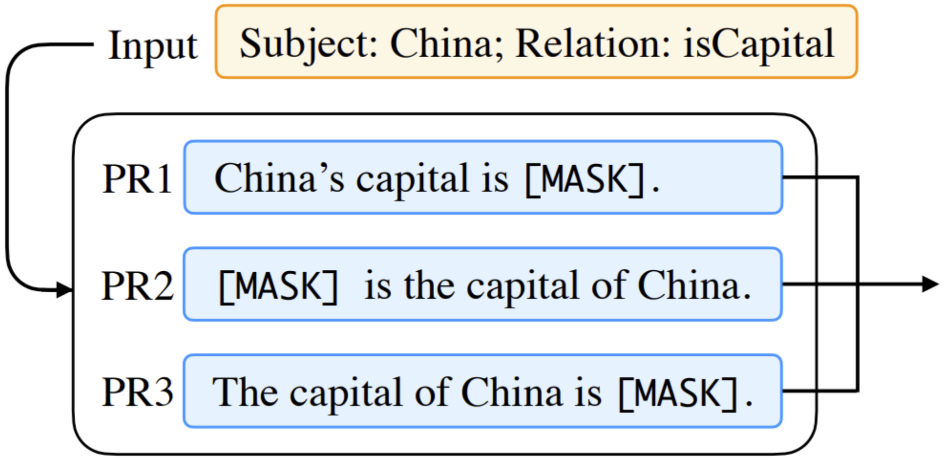
Prompt Ensembling

Definition

- using multiple unanswered prompts for an input at inference time to make predictions

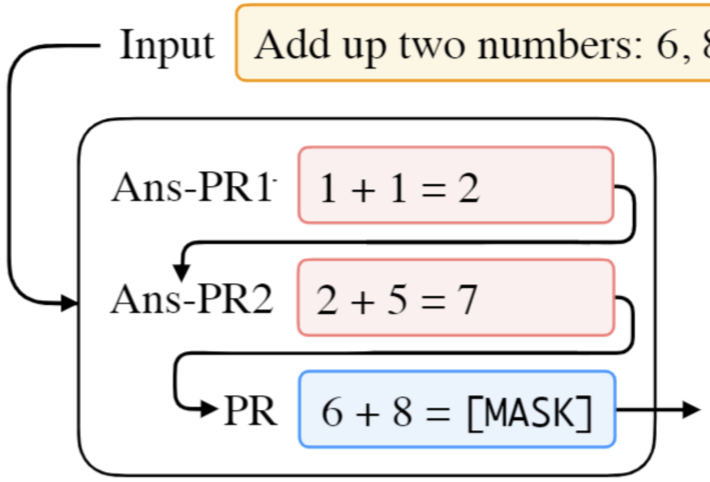
Advantages

- Utilize complementary advantages
- Alleviate the cost of prompt engineering
- Stabilize performance on downstream tasks

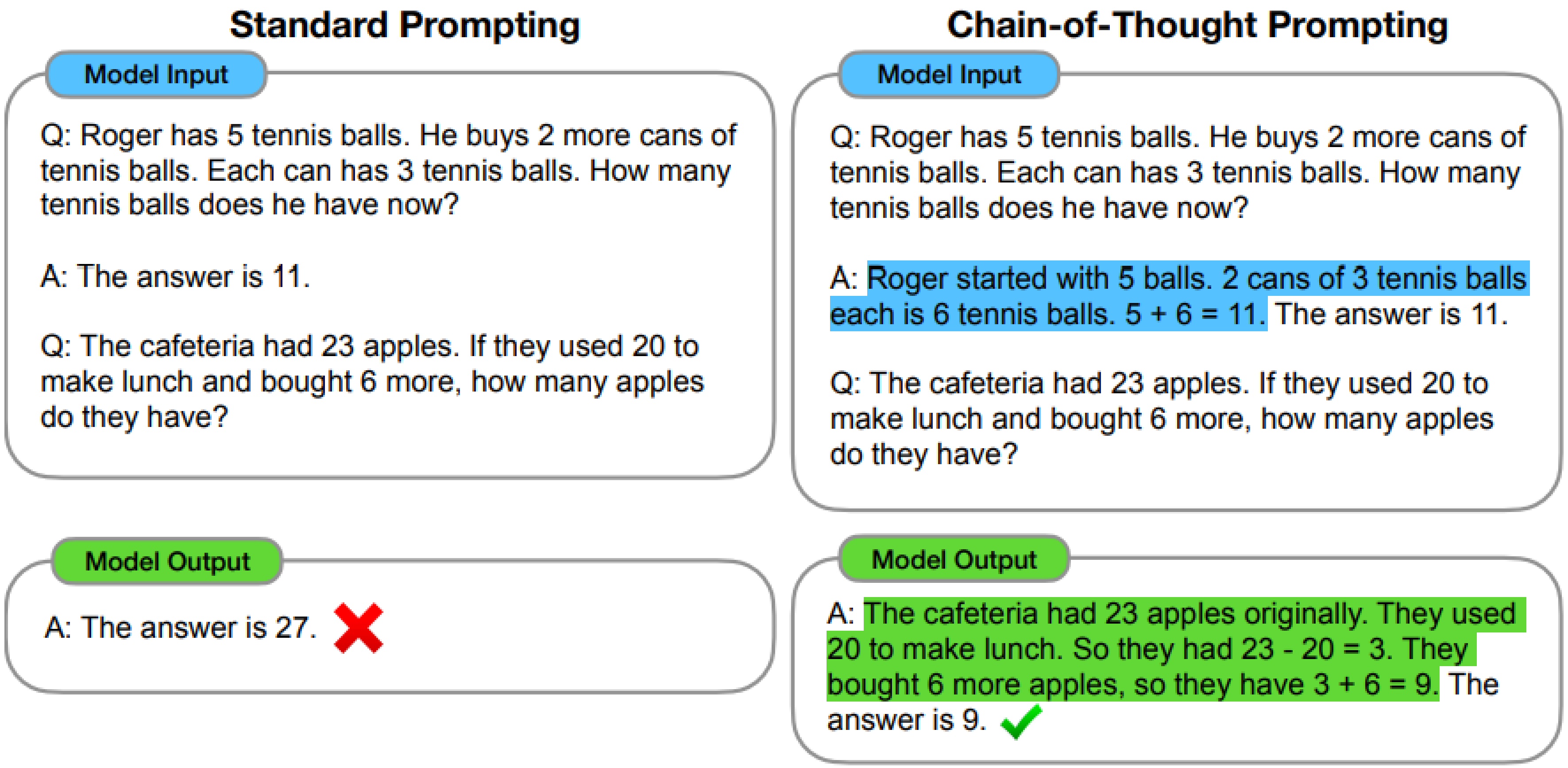


In-context Learning

- 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



Chain-of-thought



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al.2022

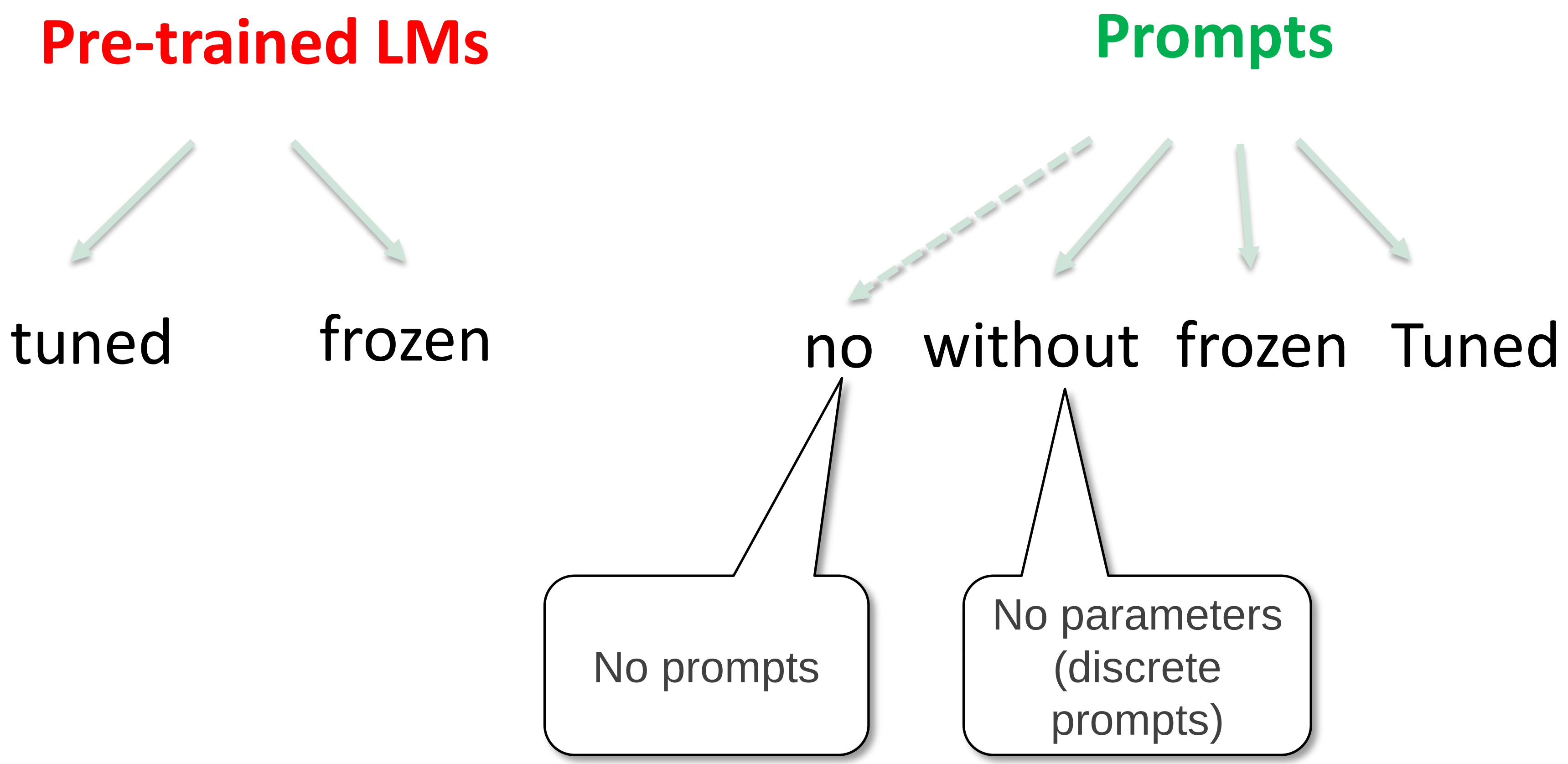
Prompt Sharing

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

Prompt Sharing

- 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

Pre-trained LMs

tuned frozen

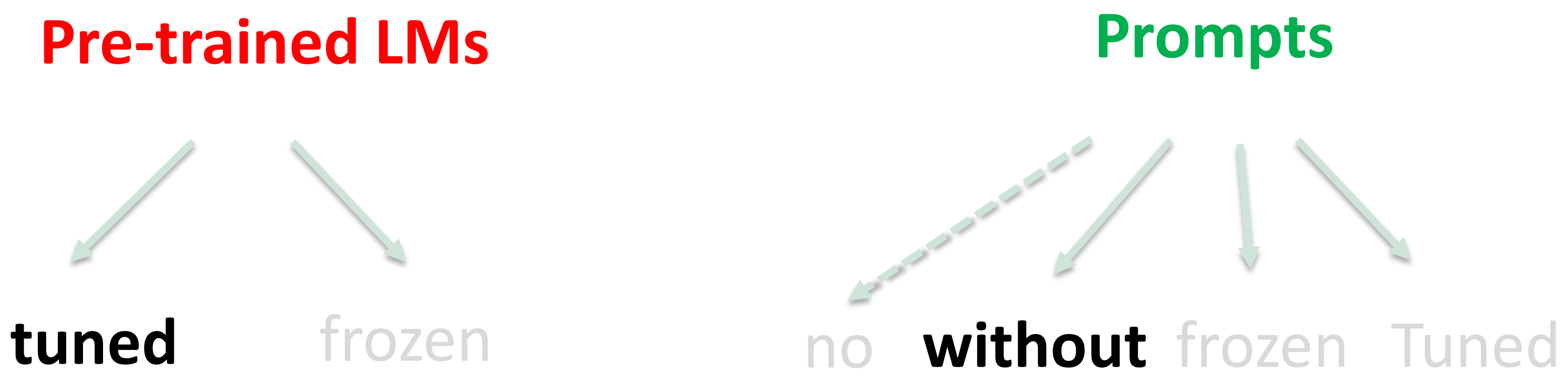
Prompts

no without frozen Tuned

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

Pre-trained LMs

tuned frozen

Prompts

no without **frozen** Tuned

Fixed-prompt Tuning

Example: BERT + Transferred Continuous Prompt for text classification

Cases of Parameter Updating

Pre-trained LMs

tuned frozen

Prompts

no without frozen **Tuned**

Prompt+LM Fine-tuning

Example: BERT + Continuous Prompt for text classification

Cases of Parameter Updating

Pre-trained LMs

tuned **frozen**

Prompts

no without frozen Tuned

Adapter Tuning

Example: BERT + Adapter for text classification

Cases of Parameter Updating

Pre-trained LMs

tuned **frozen**

Prompts

no **without** frozen Tuned

Tuning-free Prompting

Example: GPT3 + Discrete Prompts for Machine Translation

Cases of Parameter Updating

Pre-trained LMs

tuned **frozen**

Prompts

no without **frozen** Tuned

Tuning-free Prompting

Example: GPT3 + Continuous Prompts for Machine Translation

Cases of Parameter Updating

Pre-trained LMs

tuned **frozen**

Prompts

no without frozen **Tuned**

Fixed-LM Prompt Tuning

Example: BART + Continuous Prompts for Machine Translation

Too many, difficult to select?

Promptless Fine-tuning

Fixed-prompt Tuning

Prompt+LM Fine-tuning

Adapter Tuning

Tuning-free Prompting

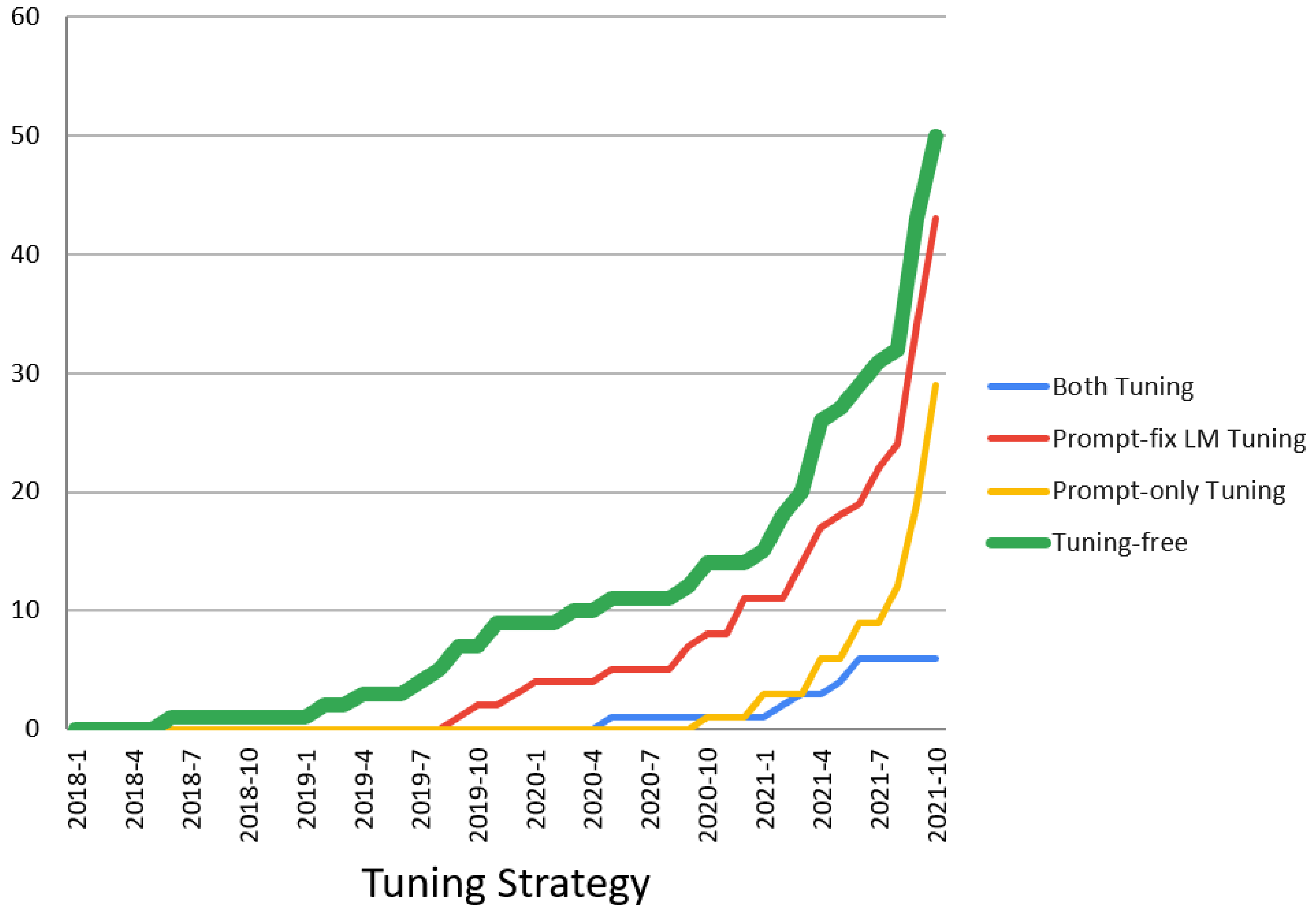
Fixed-LM Prompt Tuning

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?

Which one is more popular?



Development of Prompting Methods