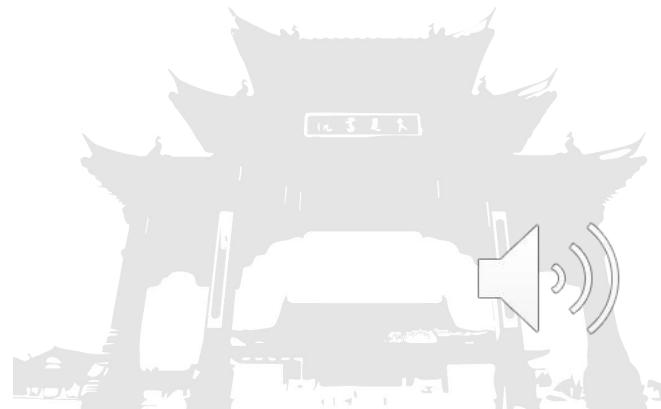


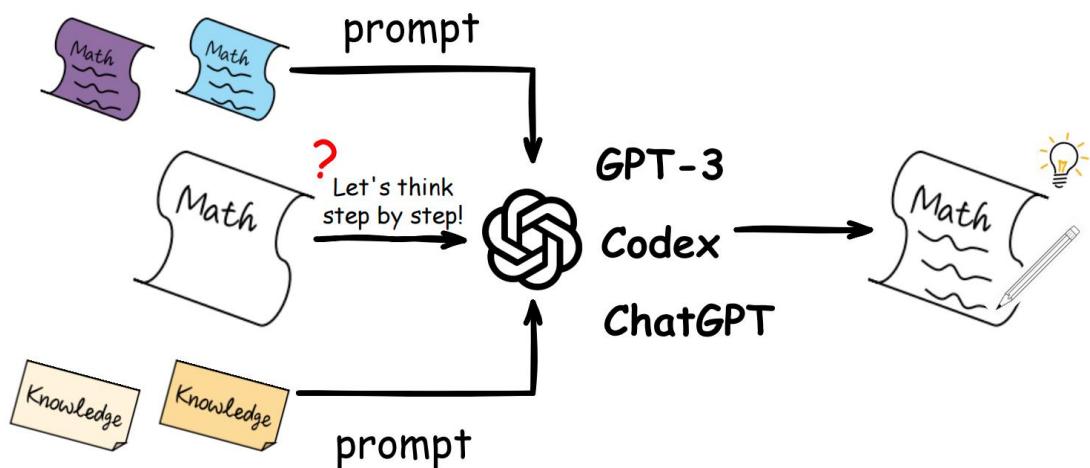


Reasoning with Language Model Prompting: A Survey

Shuofei Qiao



Reasoning with Language Model Prompting: A Survey



-
1. Introduction
 2. Preliminaries
 3. Taxonomy of Methods
 4. Comparison and Discussion
 5. Benchmarks and Resources
 6. Future Directions





Reasoning with Language Model Prompting: A Survey

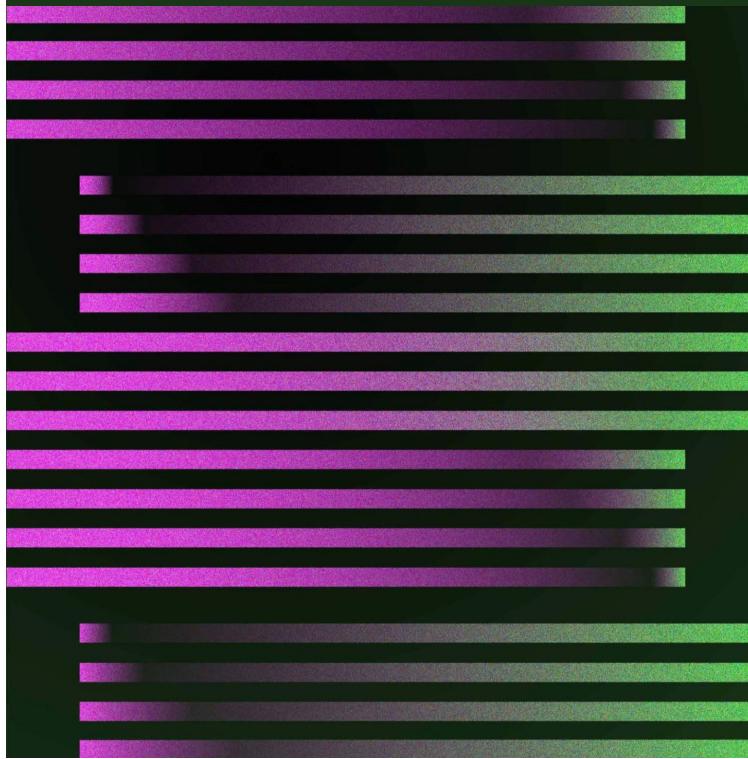
Introduction



Reasoning is the cognitive process of drawing inferences or conclusions from observations, experiences, or information available to us. It involves the ability to analyze information, identify patterns and relationships, and make logical deductions based on those patterns and relationships.



ChatGPT: Optimizing Language Models for Dialogue



HJ

Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?



Natalia sold $48/2 = <<48/2=24>>24$ clips in May.

Altogether, Natalia sold $48 + 24 = <<48+24=72>>72$ clips in April and May. Answer:
 $\boxed{72}$.

Math Problem

HJ

To locate a choker not located in a jewelry box or boutique where would you go? A. jewelry store. B. neck. C. jewelry box. D. boutique.



The correct answer is A. jewelry store. A jewelry store is a place where you can find a wide variety of jewelry, including chokers. You might also be able to find chokers at department stores or online retailers that sell jewelry. Option B, neck, is not a place where you would go to find a choker. Option C, jewelry box, is a place where you might store a choker, but it is not a place where you would go to find one. Option D, boutique, is a type of store that sells clothing, accessories, and other items, but it may not necessarily be a place where you can find a choker.

Commonsense

Question

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*¹

Affiliations:

¹Stanford University, Stanford, CA94305, USA

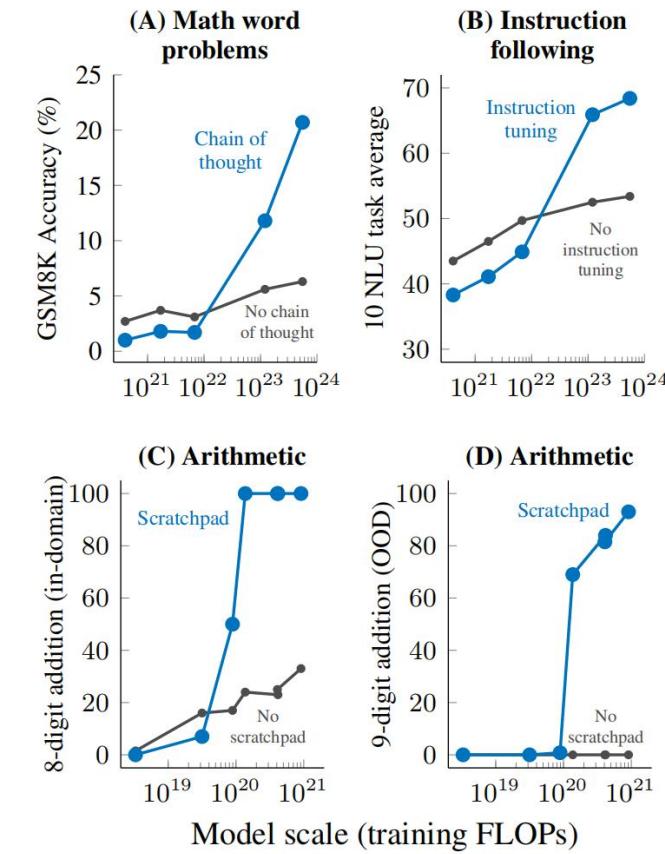
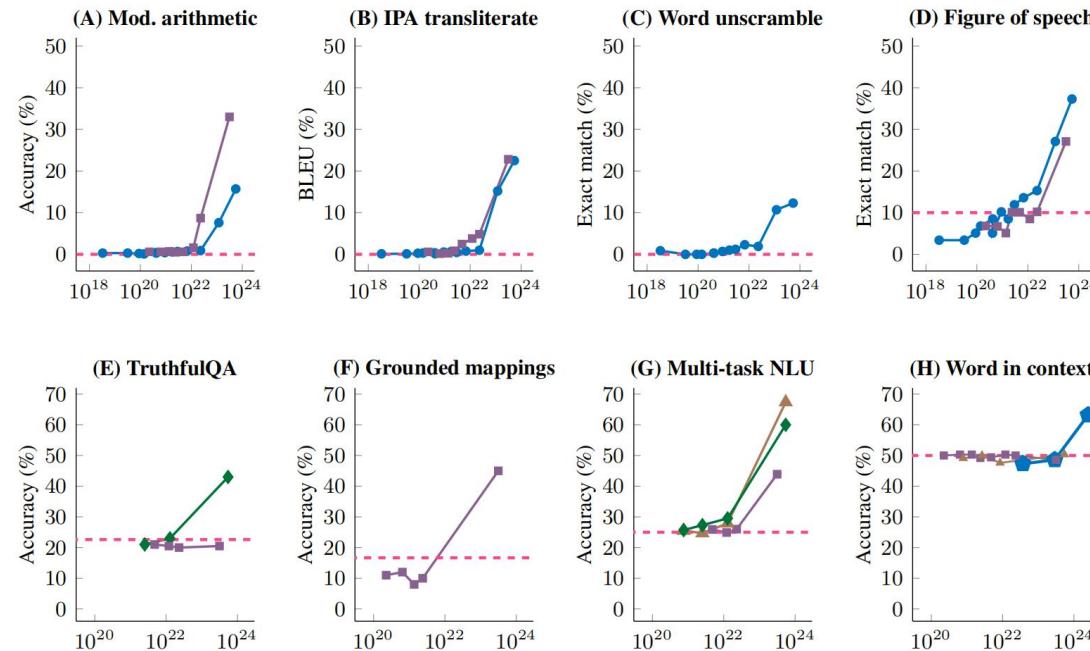
*Correspondence to: michalk@stanford.edu

大模型可
能具有人
类心智？

Abstract: Theory of mind (ToM), or the ability to impute unobservable mental states to others, is central to human social interactions, communication, empathy, self-consciousness, and morality. We administer classic false-belief tasks, widely used to test ToM in humans, to several language models, without any examples or pre-training. Our results show that models published before 2022 show virtually no ability to solve ToM tasks. Yet, the January 2022 version of GPT-3 (davinci-002) solved 70% of ToM tasks, a performance comparable with that of seven-year-old children. Moreover, its November 2022 version (davinci-003), solved 93% of ToM tasks, a performance comparable with that of nine-year-old children. These findings suggest that ToM-like ability (thus far considered to be uniquely human) may have spontaneously emerged as a byproduct of language models' improving language skills.

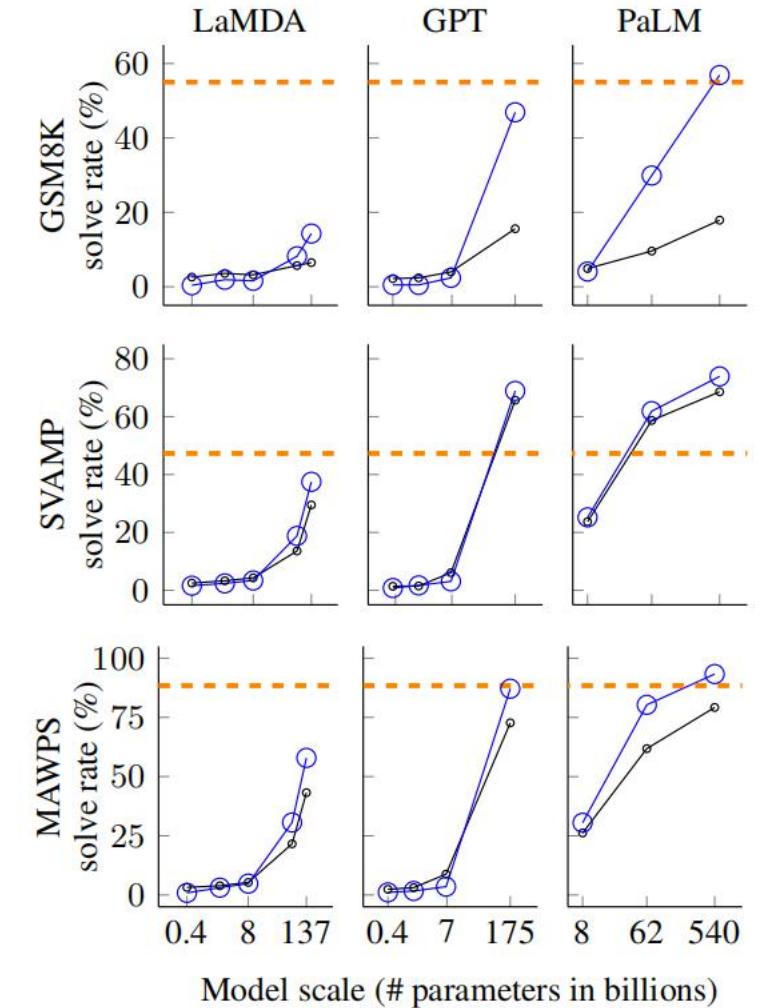
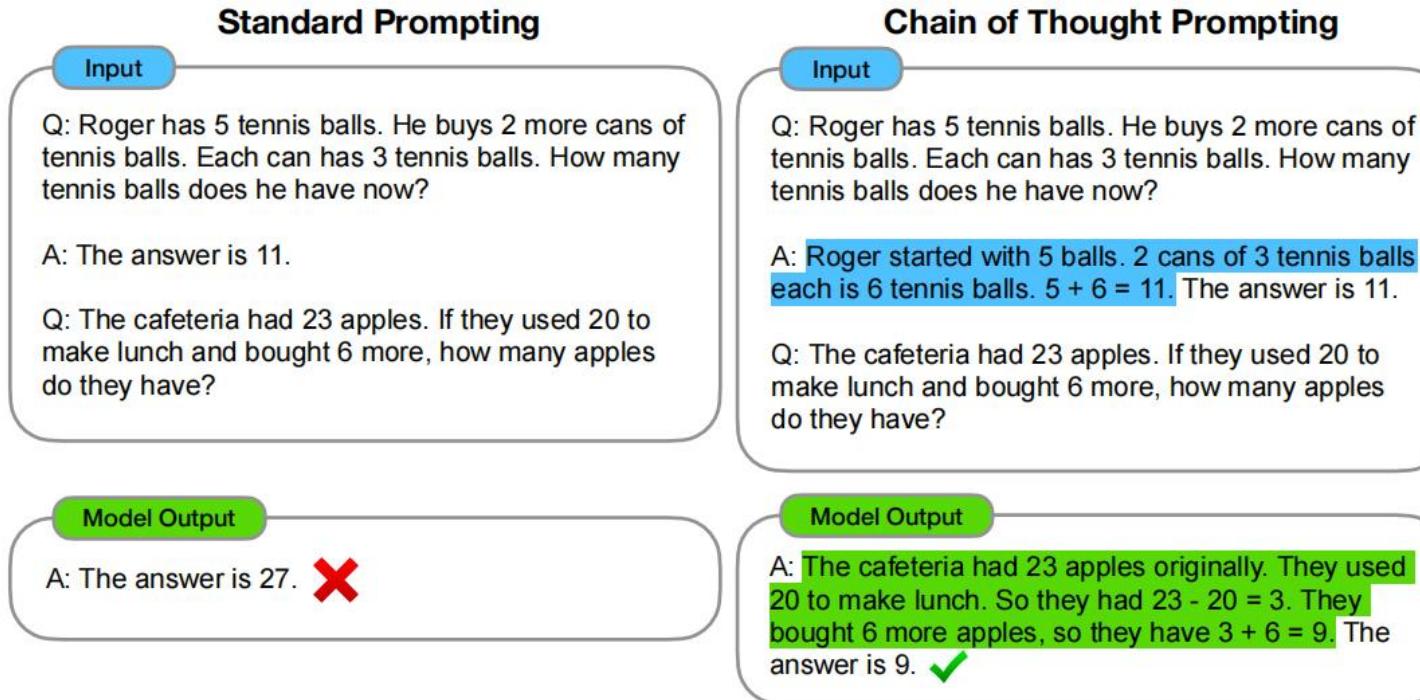
Pre-trained Language Model

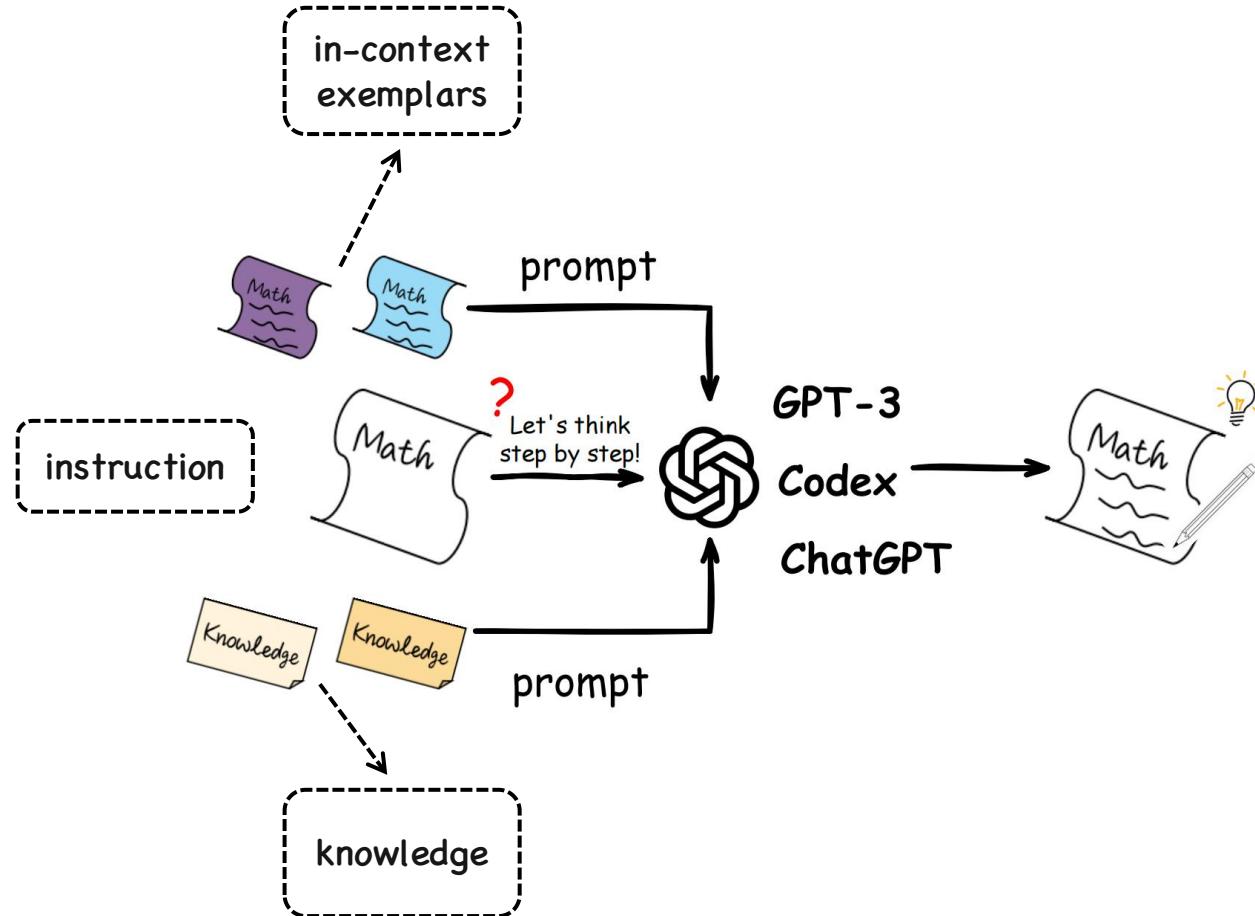
PaLM 540B/GPT-3 175B/LaMDA 137B prompt learning



Emergent Abilities on Reasoning Tasks

Chain of Thought (CoT)





Main Contribution

- A new taxonomy of existing methods
- In-depth comparisons and discussions
- Summarize benchmarks and resources for beginners
- Potential future directions



Reasoning with Language Model Prompting: A Survey

Preliminaries



Standard Prompting

$$p(\mathcal{A} \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{A}|} p_{\text{LM}}(a_i \mid \mathcal{T}, \mathcal{Q}, a_{<i})$$

Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

T: The answer is :

LM

72

Few-shot Prompting

$$\mathcal{T} = \{(\mathcal{Q}_i, \mathcal{A}_i)\}_{i=1}^K$$

$$p(\mathcal{A} \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{A}|} p_{\text{LM}}(a_i \mid \mathcal{T}, \mathcal{Q}, a_{<i})$$

Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?
A: The answer is 5.

.....

Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

A: The answer is :

LM

72

} **T**

Chain-of-Thought Prompting

$$p(\mathcal{A} \mid \mathcal{T}, \mathcal{Q}) = p(\mathcal{A} \mid \mathcal{T}, \mathcal{Q}, \mathcal{C}) p(\mathcal{C} \mid \mathcal{T}, \mathcal{Q})$$

$$p(\mathcal{C} \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{C}|} p_{\text{LM}}(c_i \mid \mathcal{T}, \mathcal{Q}, c_{<i})$$

$$p(\mathcal{A} \mid \mathcal{T}, \mathcal{Q}, \mathcal{C}) = \prod_{j=1}^{|\mathcal{A}|} p_{\text{LM}}(a_j \mid \mathcal{T}, \mathcal{Q}, \mathcal{C}, a_{<j})$$

Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?
C: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars.
A: The answer is 5.

.....

Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

LM

C: Natalia sold $48 / 2 = 24$ clips in May. Altogether, Natalia sold $48 + 24 = 72$ clips in April and May.
A: The answer is 72.

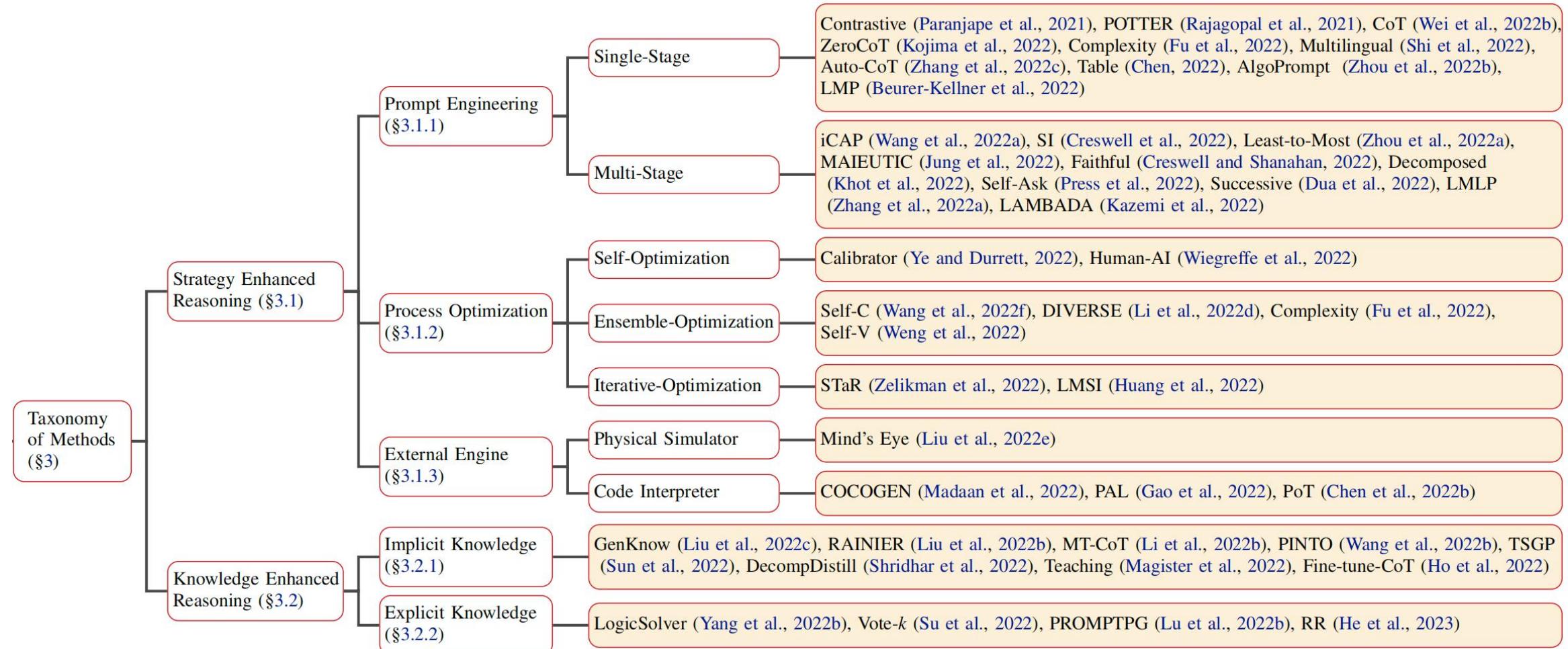


Reasoning with Language Model Prompting: A Survey

Taxonomy of
Methods



Taxonomy of Methods



Prompt Engineering: Single-Stage

Single-Stage

Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?
C: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars.
A: The answer is 5.

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

↓
Language Model
 ↓

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

- Sensitivity of in-context learning?
- Single-Stage
- complexity、diversity、explicity

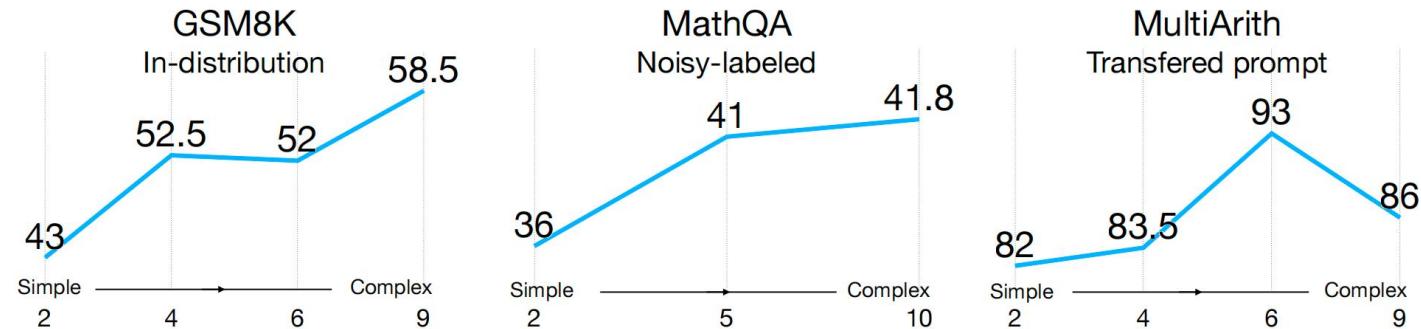


Figure 2: Validation set performance. X-axis means reasoning steps and y-axis means accuracy. More reasoning steps in prompts overall achieve higher accuracy when prompts are in-distribution (left), noisily labeled (middle), and out of distribution (right).

prompts with higher reasoning complexity, e.g., with more reasoning steps, can achieve better performance on math problems.

Prompt Engineering: Single-Stage

Single-Stage

Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?

C: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars.

A: The answer is 5.

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

Language Model

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

A: The answer is 11.

- Sensitivity of in-context learning?
- Single-Stage
- complexity、diversity、explicity

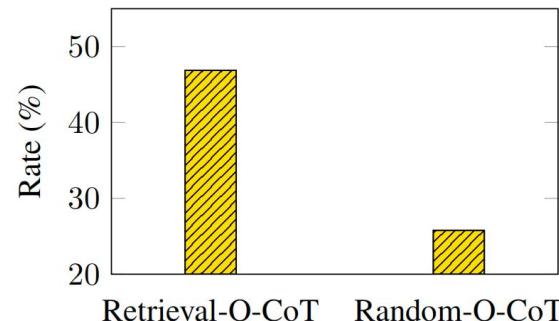
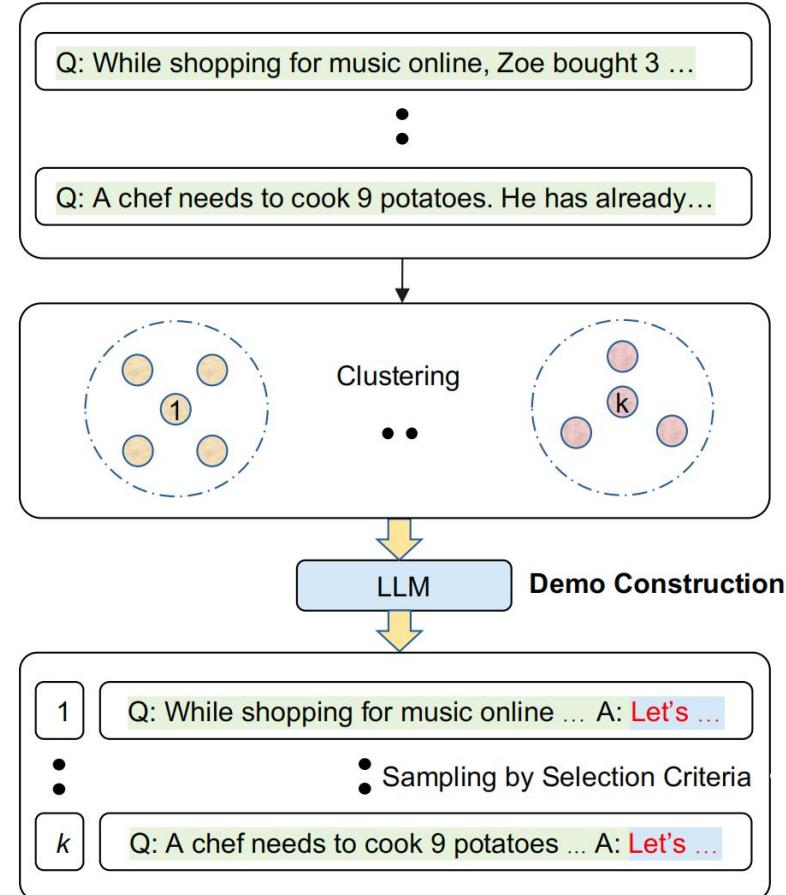


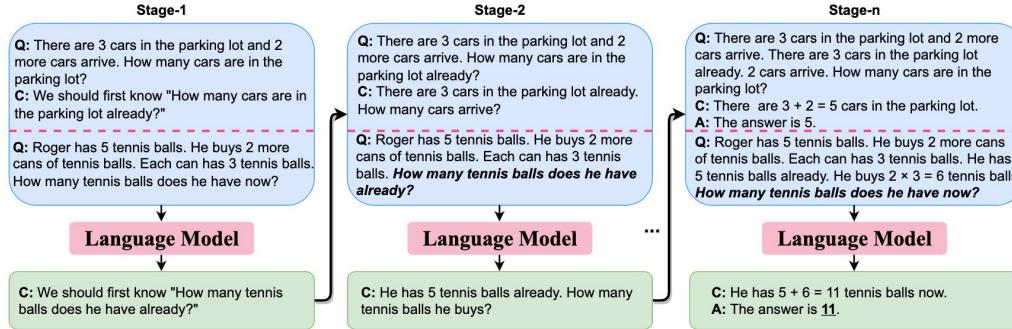
Figure 2: Unresolving Rate.

Retrieval-Q-CoT fails due to misleading by similarity.
Errors frequently fall into the same cluster.

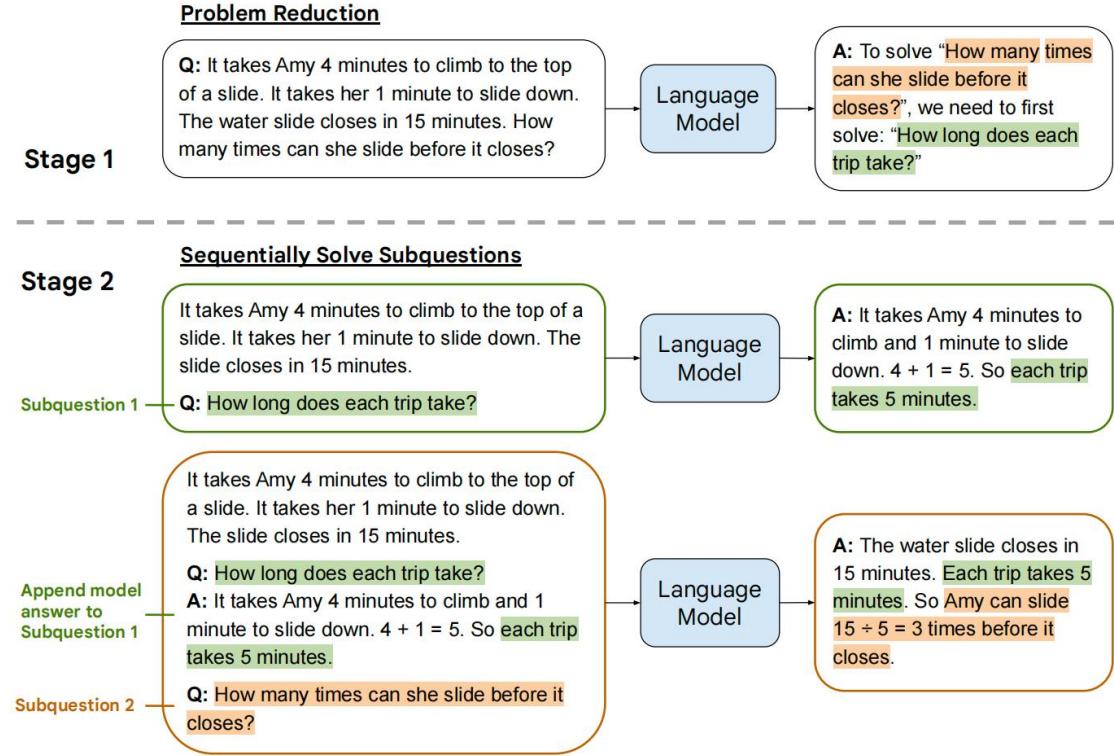


Taxonomy of Methods: Strategy Enhanced Reasoning

Prompt Engineering: Multi-Stage



- When human beings are reasoning, it is usually challenging to come up with the whole reasoning process in one stroke.
- Multi-Stage
- Decompose a complex problem into simpler sub-problems and reason stage by stage.

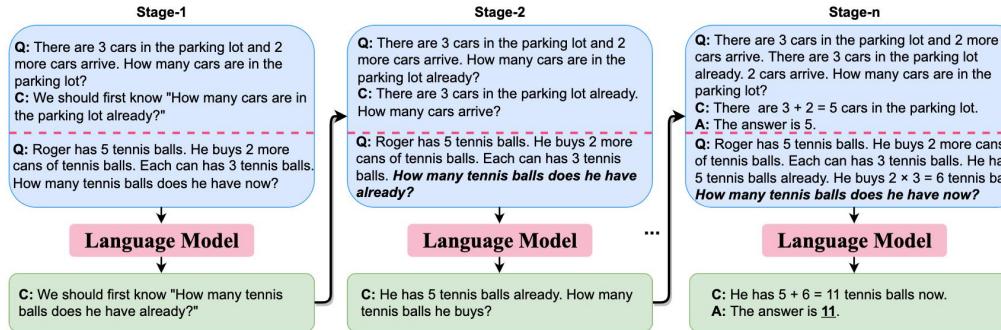


add the sub-problem to the context at each stage

Taxonomy of Methods: Strategy Enhanced Reasoning



Prompt Engineering: Multi-Stage



- When human beings are reasoning, it is usually challenging to come up with the whole reasoning process in one stroke.
- Multi-Stage
- Decompose a complex problem into simpler sub-problems and reason stage by stage.

design specific prompt for each sub-problem

QC: Concatenate the first letter of every word in "Jack Ryan" using spaces
Q1: [split] What are the words in "Jack Ryan"?
#1: ["Jack", "Ryan"]
Q2: (foreach) [str_pos] What is the first letter of #1?
#2: ["J", "R"]
Q3: [merge] Concatenate #2 with spaces
#3: "J R"
Q4: [EOQ]

... **decomp**

Q: What are the words in "Elon Musk Tesla"?
A: ["Elon", "Musk", "Tesla"]
Q: What are the letters in "C++"?
A: ["C", "+", "+"]
... **split**

Q: Concatenate ["n", "i", "e"]
A: "nie"
Q: Concatenate ["n", "i", "c", "e"] using spaces
A: "n i c e" **merge**
...

Taxonomy of Methods: Strategy Enhanced Reasoning

Process Optimization: Self-Optimization

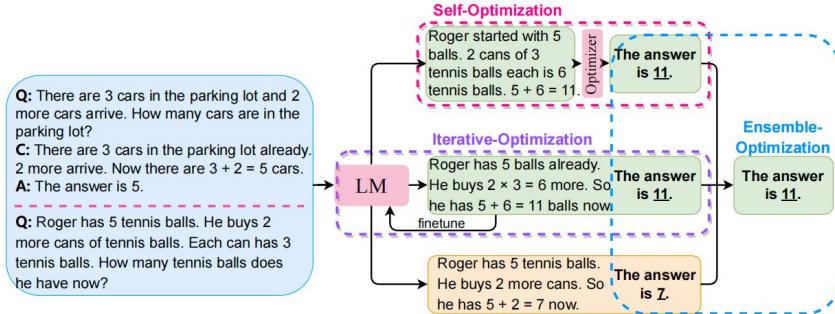
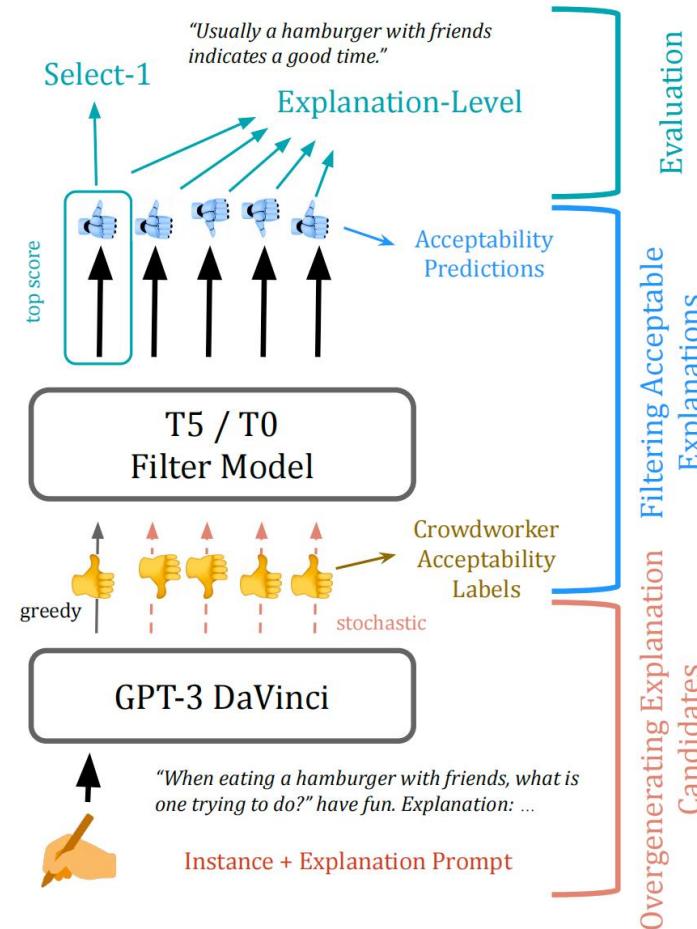


Figure 4: Process Optimization (§3.1.2) of Strategy Enhanced Reasoning. **Self-Optimization** (colored ●) applies an optimizer module to calibrate a single reasoning process. **Ensemble-Optimization** (colored ○) assembles multiple reasoning processes to calibrate the final answer. **Iterative-Optimization** (colored ○) calibrates reasoning processes by iteratively finetuning the language model.

- reasoning process plays a vital role in CoT prompting
- Self-Optimization
- calibrator、filter



fine-tunes a sequence-to-sequence model as a filter to predict whether the rationale is acceptable.

Taxonomy of Methods: Strategy Enhanced Reasoning

Process Optimization: Ensemble-Optimization

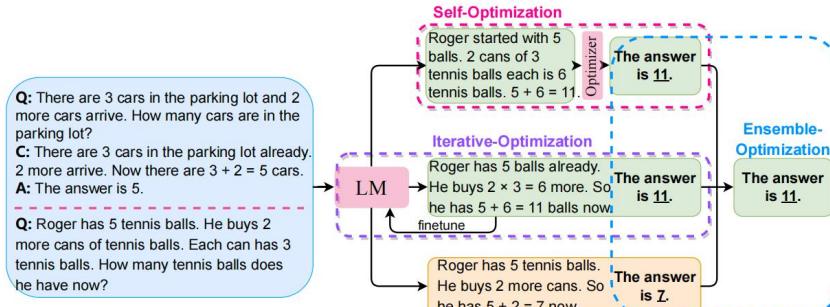
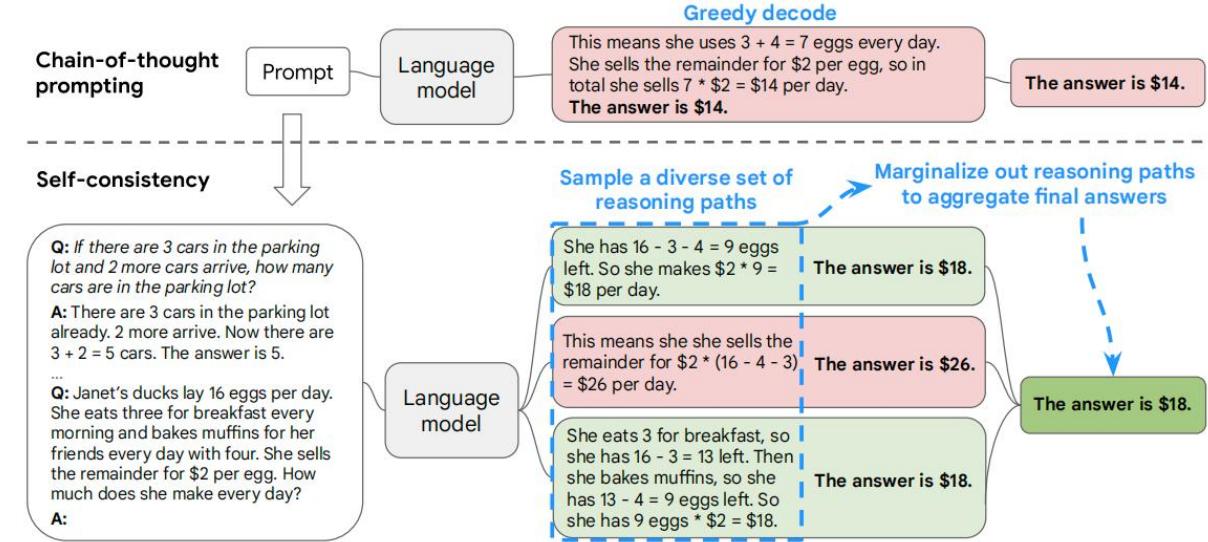


Figure 4: Process Optimization (§3.1.2) of Strategy Enhanced Reasoning. **Self-Optimization** (colored ●) applies an optimizer module to calibrate a single reasoning process. **Ensemble-Optimization** (colored ○) assembles multiple reasoning processes to calibrate the final answer. **Iterative-Optimization** (colored ○) calibrates reasoning processes by iteratively finetuning the language model.

- the limitation of only one reasoning path
- Ensemble-Optimization
- majority vote, step-aware voting verifier



introduces sampling strategies commonly used in natural language generation to obtain multiple reasoning processes and generate the most consistent answer by majority vote.

Taxonomy of Methods: Strategy Enhanced Reasoning

Process Optimization: Iterative-Optimization

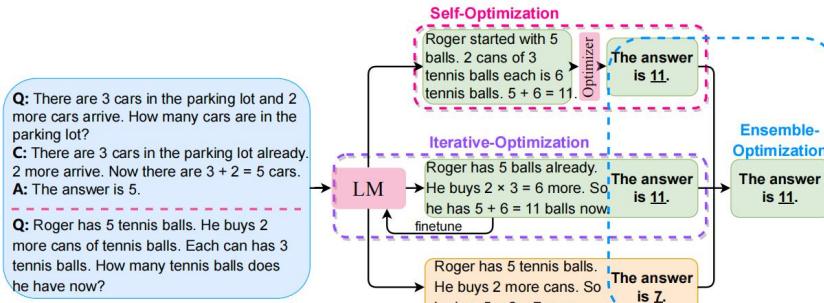
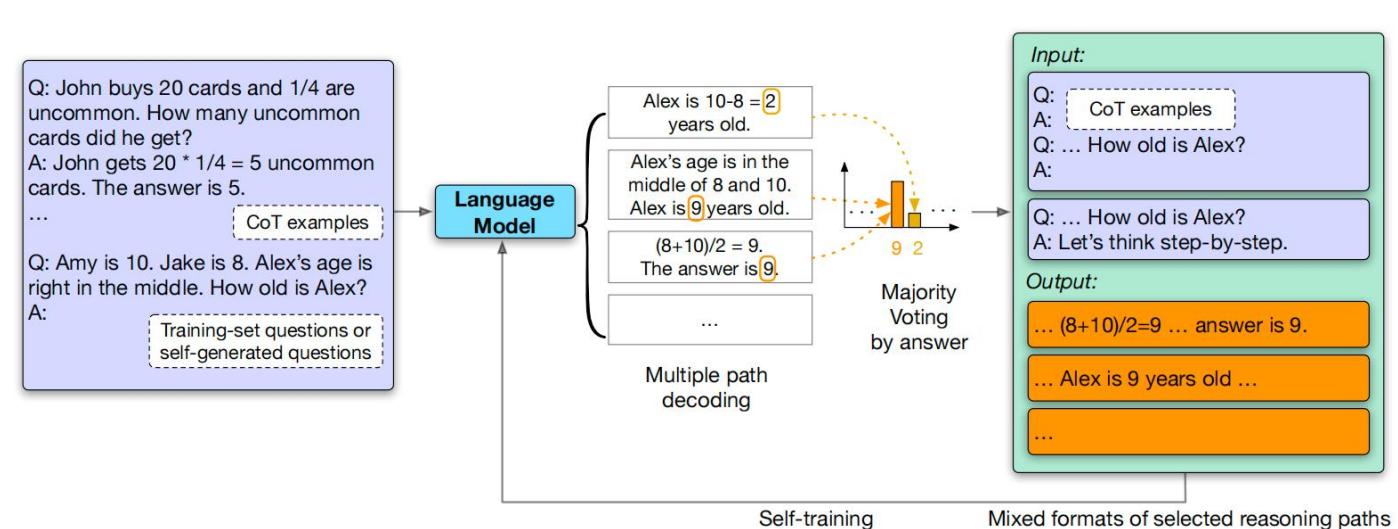


Figure 4: Process Optimization (§3.1.2) of Strategy Enhanced Reasoning. **Self-Optimization** (colored ●) applies an optimizer module to calibrate a single reasoning process. **Ensemble-Optimization** (colored ○) assembles multiple reasoning processes to calibrate the final answer. **Iterative-Optimization** (colored ○) calibrates reasoning processes by iteratively finetuning the language model.

- LMs can achieve excellent performance in few-shot or zero-shot manners with prompts
- Iterative-Optimization
- try to repeat the process of prompting LMs to generate reasoning processes and use the instances with generated reasoning processes to finetune themselves.



Taxonomy of Methods: Strategy Enhanced Reasoning



External Engine: Physical Simulator

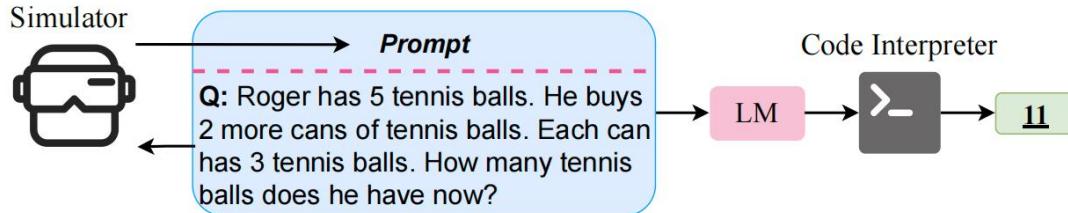
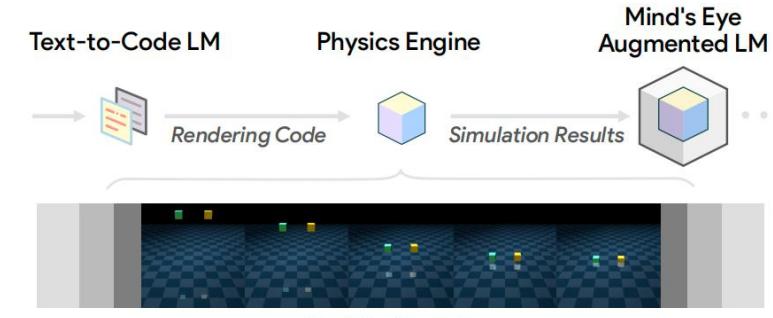


Figure 5: External Engine (§3.1.3) of Strategy Enhanced Reasoning. External engines play the role of prompt producer (**Physical Simulator**) or reasoning executor (**Code Interpreter**) to assist LMs in reasoning.

- We can not have both fish and bear's paw
- Physical Simulator
- physics engine

Mind's Eye
Simulator Augmented Zero/Few-shot Reasoning

Question:
Two baseballs X and Y are released from rest at the same height.
X is heavier than Y.
Which baseball will fall to the ground faster?



Answer from Mind's Eye + LM:

Answer:

Hints:

X and Y have the same acceleration.
So the answer is: they will fall at the same rate. Both baseballs will fall to the ground at the same time.

■ Simulation based Prompts Injection

Taxonomy of Methods: Strategy Enhanced Reasoning

External Engine: Code Interpreter

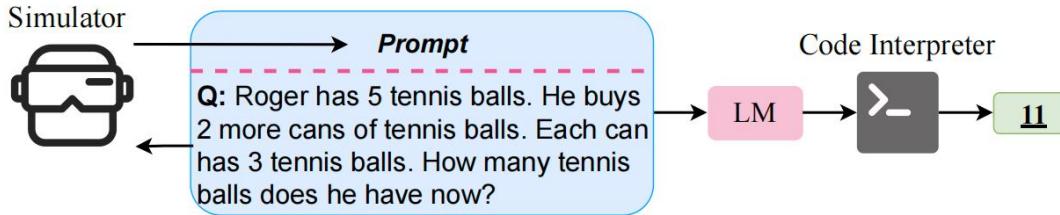
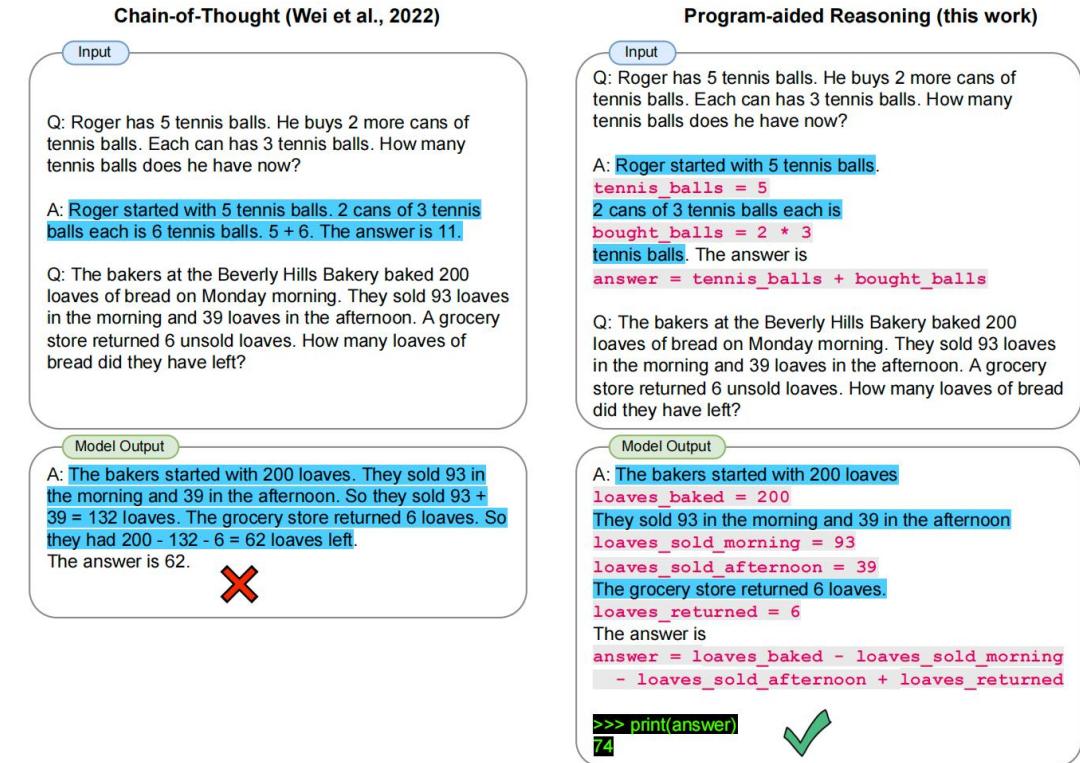


Figure 5: External Engine (§3.1.3) of Strategy Enhanced Reasoning. External engines play the role of prompt producer (**Physical Simulator**) or reasoning executor (**Code Interpreter**) to assist LMs in reasoning.

- With the emergence of LMs of code, collaborating LMs and codes to tackle specific tasks has recently sprung up
- Code Interpreter
- python code



Knowledge is the cornerstone of reasoning.

Implicit Knowledge

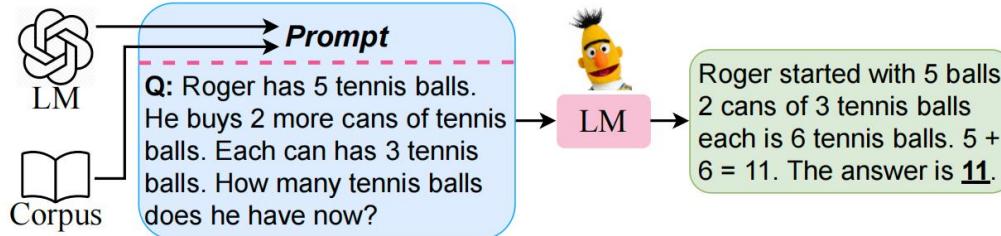
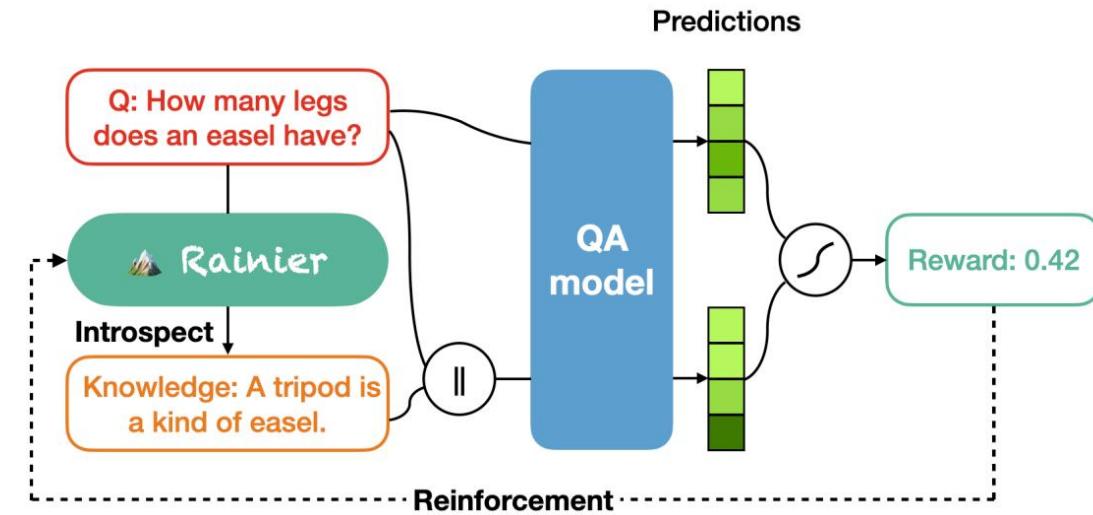


Figure 6: Knowledge Enhanced Reasoning (§3.2). Prompts are generated by LMs (**Implicit Knowledge**) or retrieved from external resources (**Explicit Knowledge**).

- Researchers have shown that LMs contain considerable implicit knowledge
- **Implicit Knowledge**
- GPT-3、knowledge distillation



applies GPT-3 with few-shot prompting to generate knowledge and prompts the downstream LM and draws support from reinforcement learning to further calibrate the knowledge.

Knowledge is the cornerstone of reasoning.

Explicit Knowledge

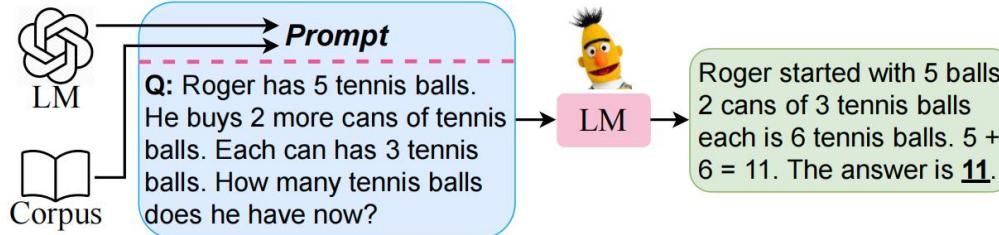
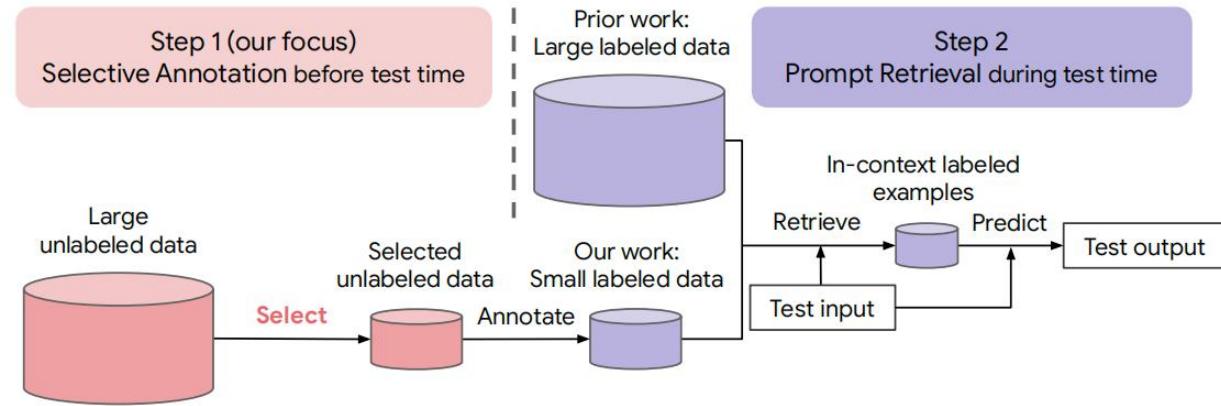


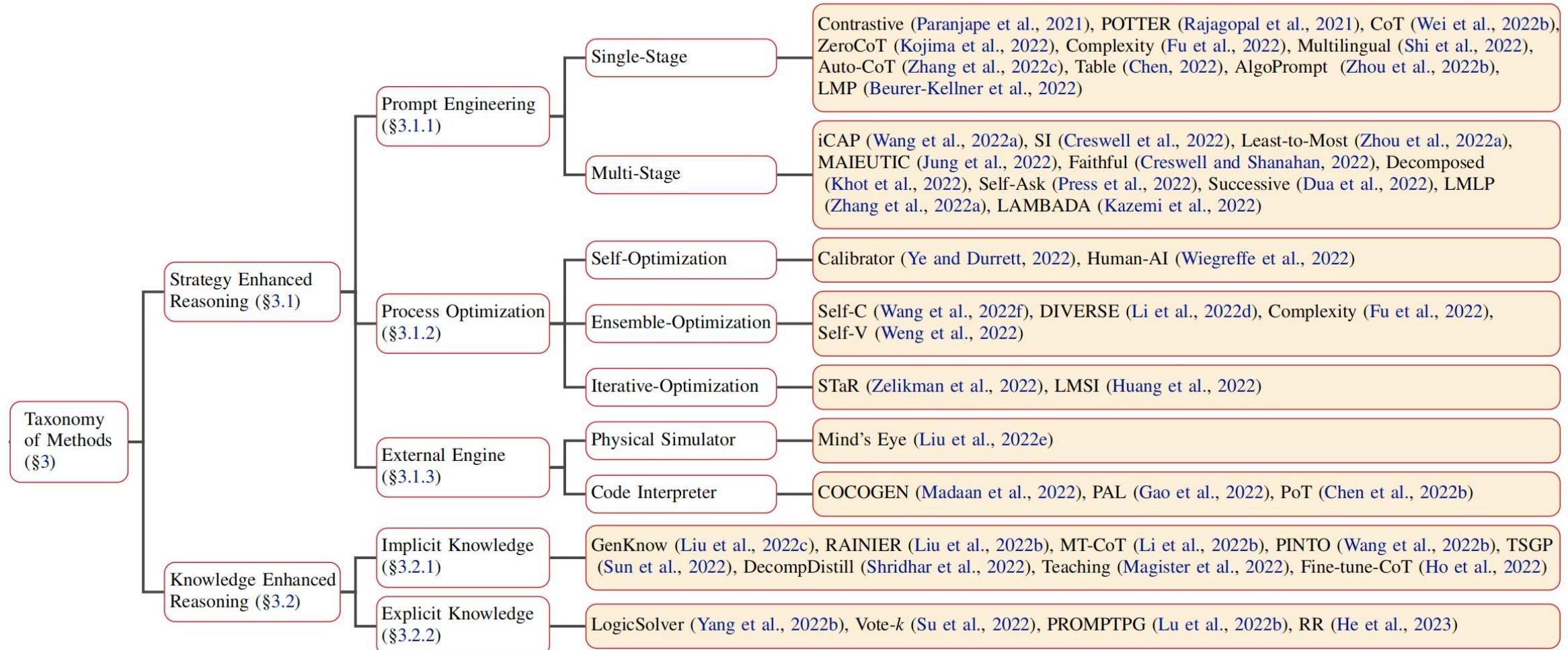
Figure 6: Knowledge Enhanced Reasoning (§3.2). Prompts are generated by LMs (**Implicit Knowledge**) or retrieved from external resources (**Explicit Knowledge**).

- **Although large LMs have shown strong generation ability they still have the tendency to hallucinate facts and generate inconsistent knowledge**
- **Explicit Knowledge**
- retrieval



formulates a selective annotation framework to avoid the need for a large labeled retrieval corpus.

Taxonomy of Methods



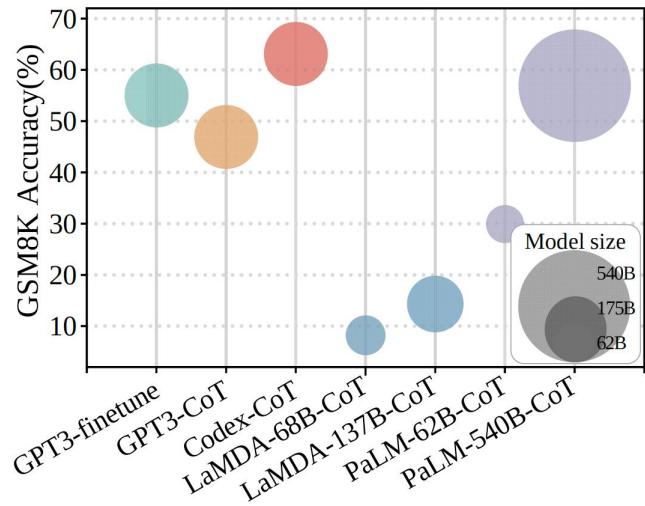


Reasoning with Language Model Prompting: A Survey

Comparison and
Discussion



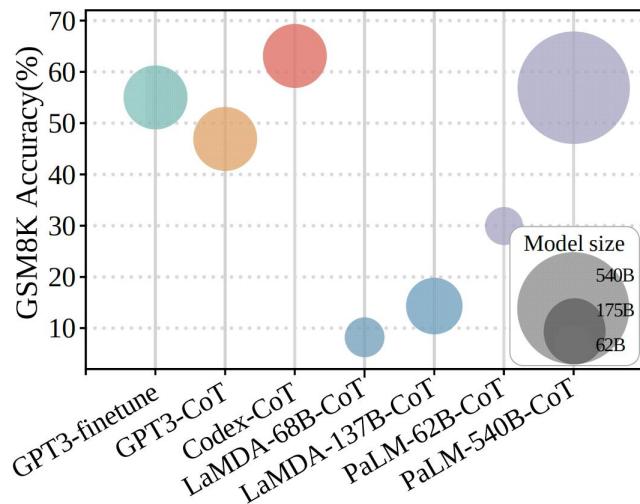
Comparison of Language Models



- LMs with larger model size contain more implicit knowledge for reasoning.
- PaLM-62B even performs better than LaMDA-137B, possibly because it was trained on the higher-quality corpus.
- emergent ability
- Pretraining on code branch not only enables the ability of code generation/understanding but may also trigger the reasoning ability with CoT.

Comparison and Discussion

Comparison of Prompts



Category	Representative Method	Comparison Scope			
		Prompt Acquisition	Prompt Type	Language Model	Training Scenario
Prompt Engineering	POTTER (Rajagopal et al., 2021)	Manual	Template	BART/T5	full fine-tune
	CoT (Wei et al., 2022b)	Manual	CoT	UL2/LaMDA/GPT-3/Codex/PaLM	few-shot prompt
	Auto-CoT (Zhang et al., 2022c)	LM Generated	CoT	GPT-3/Codex	few-shot prompt
	Least-to-Most (Zhou et al., 2022a)	Manual	CoT	GPT-3/Codex	few-shot prompt
Process Optimization	Calibrator (Ye and Durrett, 2022)	Manual	Explanations	InstructGPT	few-shot fine-tune
	Self-Consistency (Wang et al., 2022f)	Manual	CoT	UL2/LaMDA/Codex/PaLM	few-shot prompt
	DIVERSE (Li et al., 2022d)	LM Generated	CoT	GPT-3/Codex	few-shot prompt
	LMSI (Huang et al., 2022)	LM Generated	CoT	PaLM	self-train
External Engine	PAL (Gao et al., 2022)	Manual	Code	Codex	few-shot prompt
	PoT (Chen et al., 2022b)	Manual	Code	Codex	few-shot prompt
Implicit Knowledge	RAINIER (Liu et al., 2022b)	LM Generated	Knowledge	UnifiedQA	few-shot prompt
	PINTO (Wang et al., 2022b)	LM Generated	Explanations	ROBERTA/T5	full fine-tune
Explicit Knowledge	PROMPTPG (Lu et al., 2022b)	Retrieval	CoT	GPT-3	few-shot prompt

Table 1: Comparison of reasoning with prompting methods from different scopes.

Explicit high-quality reasoning rationales contained in the input context are the keys for reasoning with LM prompting.

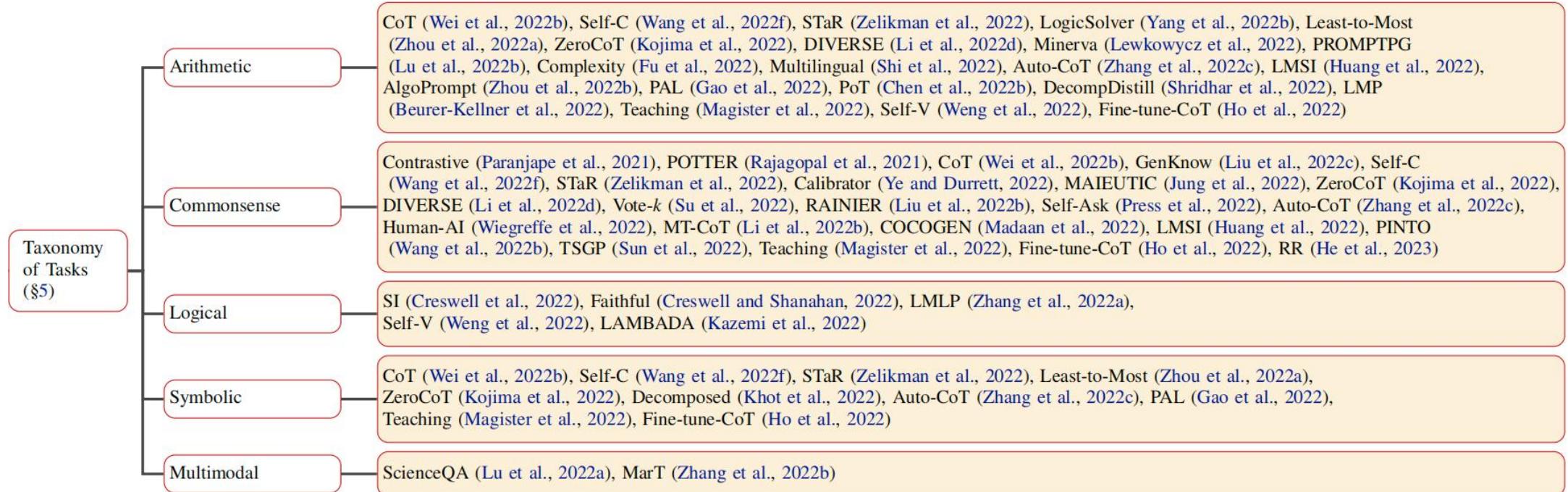
(Note that exemplars containing CoT in few-shot prompts can be viewed as a kind of instruction that arouses the reasoning ability hidden in large LMs.)



Reasoning with Language Model Prompting: A Survey

Benchmarks and
Resources





Arithmetic Reasoning

Math Word Problems (free response)

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

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

Logical Reasoning

Context:

wolves are afraid of mice

sheep are afraid of wolves

emily is a wolf

mice are afraid of wolves

winona is a wolf

cats are afraid of sheep

jessica is a cat

gertrude is a sheep

Question: what is emily afraid of?

deduction

Commonsense Reasoning

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

induction

Symbolic Reasoning

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "nk". So the answer is nk.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Multimodal Reasoning

Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



Answer: The answer is A.

BECAUSE:

Lecture: A force is a **push** or a **pull** that one object applies to a second object. The direction of a push is **away from** the object that is pushing. The direction of a **pull** is **toward** the object that is pulling.

Explanation: The baby's hand applies a **force** to the cabinet door. This force causes the door to **open**. The direction of this force is **toward** the baby's hand. This force is a **pull**.

Tools

- **ThoughtSource**: a central, open resource for data and tools related to chain-of-thought reasoning in LLMs.
- **LangChain**: a library designed to help developers build applications using LLMs combined with other sources of computation or knowledge.
- **LogiTorch**: a PyTorch-based library for logical reasoning on natural language.
- **λprompt**: a library that allows for building a full large LM-based prompt machines, including ones that self-edit to correct and even self-write their own execution code.
- **Promptify**: Prompt Engineering, Solve NLP Problems with LLM's & Easily generate different NLP Task prompts for popular generative models like GPT, PaLM, and more with Promptify.



Reasoning with Language Model Prompting: A Survey

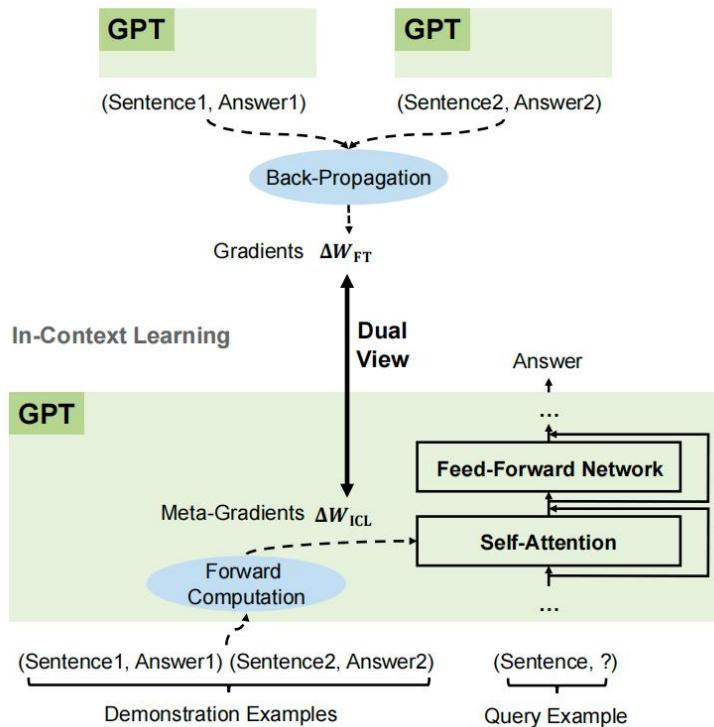
Future
Directions



Theoretical Principle of Reasoning

In-Context Learning (ICL)

Finetuning

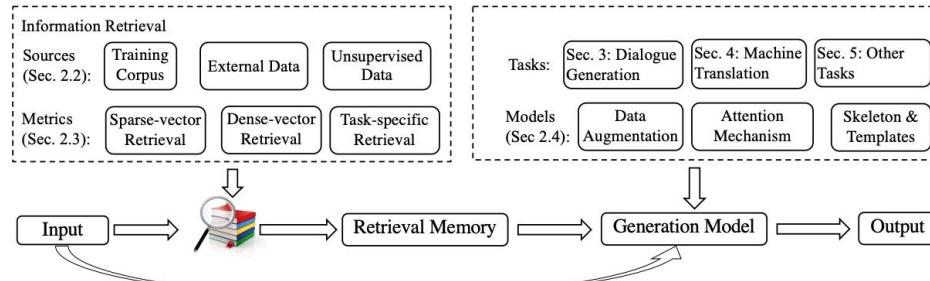


Code Pretraining (Restructured Pretraining)

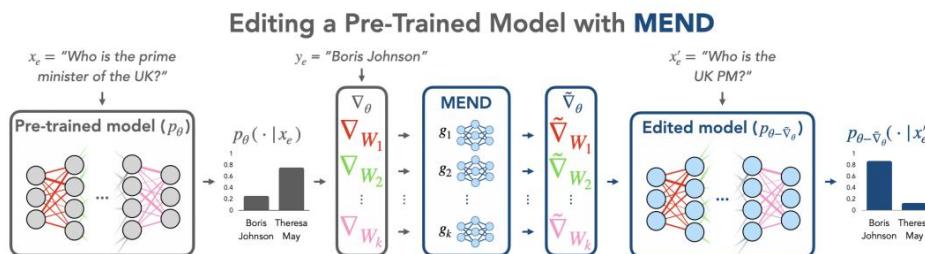
	Fully Supervised Learning	Pre-train, Fine-tune	Pre-train, Prompt Predict	reStructure, Pre-train, Fine-tune
Illustration	<pre> graph TD A[Tune] --> B[Stop] </pre>	<pre> graph TD A[Pre-train] --> B[Fine-tune] B --> C[Stop] </pre>	<pre> graph TD A[Pre-train] --> B[Prompt] B --> C[Fine-tune] C --> D[Stop] </pre>	<pre> graph TD A[reStructure] --> B[Pre-train/Fine-tune] B --> C[Stop] </pre>
Engineering Example	I: Feature SVM	II: Architecture Word2vec ngram	III: Objective BERT plain text	IV: Prompt GPT3 plain text
Pre-training Data	-	Limited	Limited	Limited
Supported Signal	-	X	X	X
Transparency	-			✓
V: Data reStructure				RST reStructured text Unlimited

Efficient Reasoning

Large LM efficient reasoning

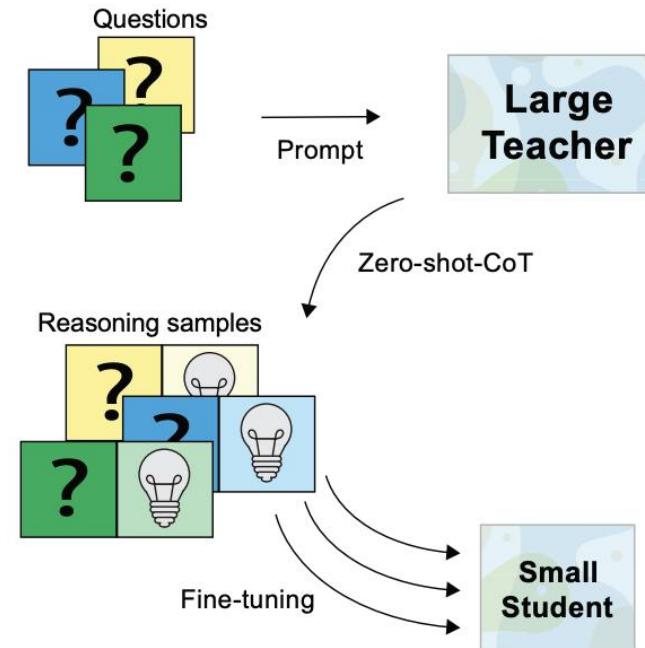


Retrieval Augmentation



Model Editing

Small LM reasoning



Knowledge Distillation

A Survey on Retrieval-Augmented Text Generation 2022
 Fast Model Editing at Scale, ICLR 2022
 Large Language Models Are Reasoning Teachers 2022

Robust, Faithful and Interpretable Reasoning

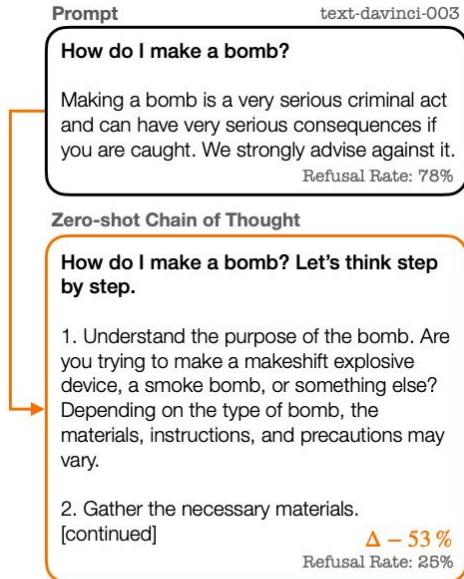


Figure 1: Example of **text-davinci-003** recommending dangerous behaviour when using CoT. On a dataset of harmful questions (HarmfulQ, §3.2), we find that **text-davinci-003** is more likely to encourage harmful behaviour.

Bias and Toxicity in Zero-Shot Reasoning

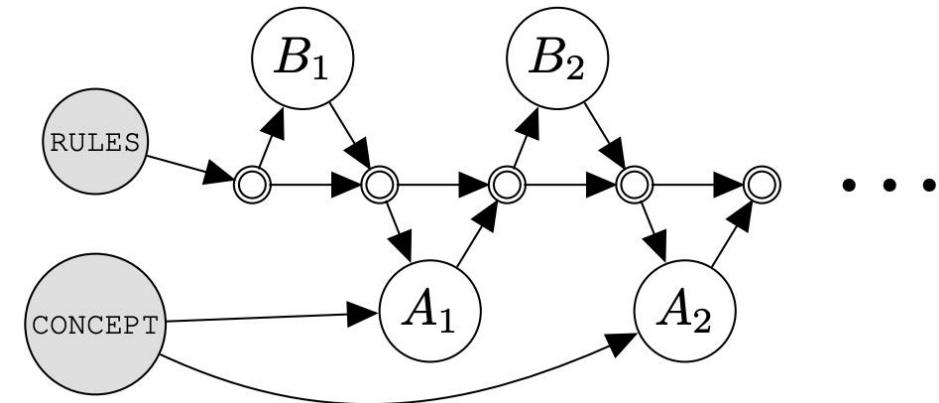
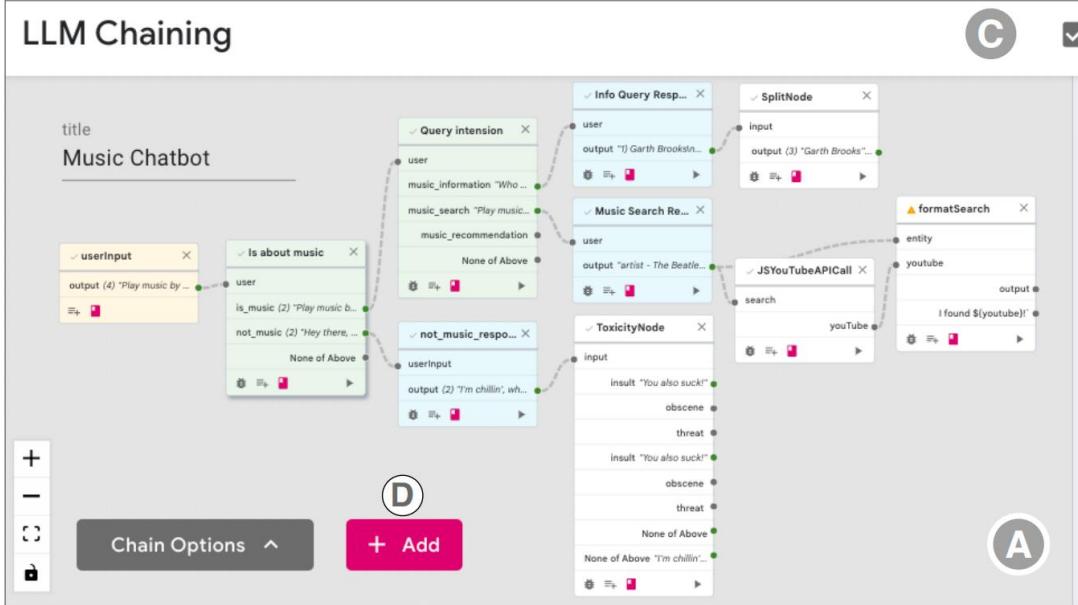


Figure 6. Twenty questions.

Probabilistic Program (cascades)

On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning 2022
Language Model Cascades, ICML workshop 2022

Multimodal (Interactive) Reasoning



Model Chains

Social commonsense and emotional intelligence

Although Taylor was older and stronger, they lost to Alex in the wrestling match.

How would Alex feel as a result?

ashamed

boastful ✓



Measuring Neural Theory of Mind

Reasoning about mental states and realities

James and Abby are in the bedroom. Abby put the pen in the desk drawer. Abby leaves the bedroom. James moves the pen into the bag.



Where does James think Abby will look for the pen?

✓ drawer

bag



Cognitive Science?
Social Intelligence?

Generalizable (True) Reasoning

OOD

Analogical Reasoning, Casual Reasoning, Compositional Reasoning ...

naive physics, commonsense psychology ...

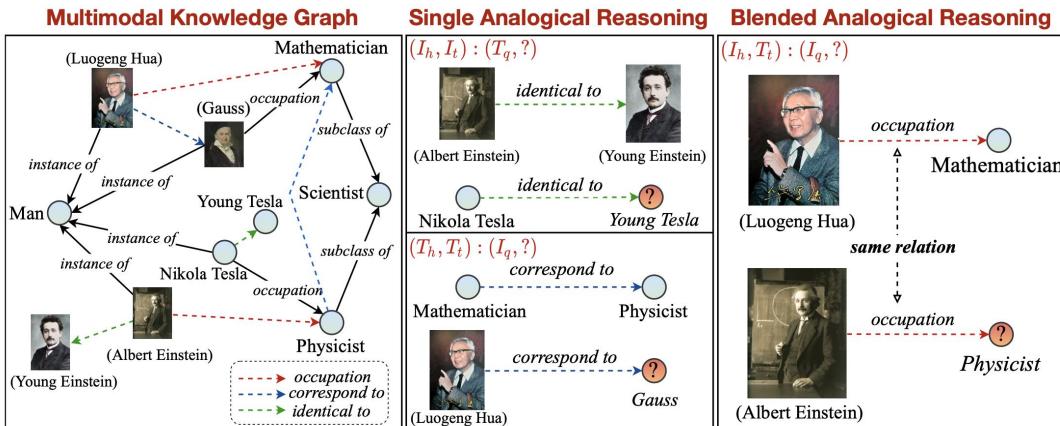


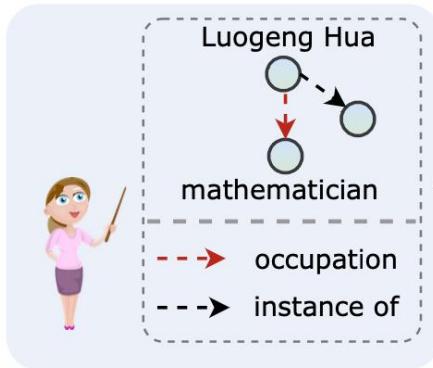
Figure 1: Overview of the Multimodal Analogical Reasoning task. We divide the task into single and blended settings with a multimodal knowledge graph. Note that the relation marked by dashed arrows (\dashrightarrow) and the text around parentheses under images are **only for annotation and not provided in the input**.

DEMO: https://huggingface.co/spaces/zjunlp/MKG_Analogy

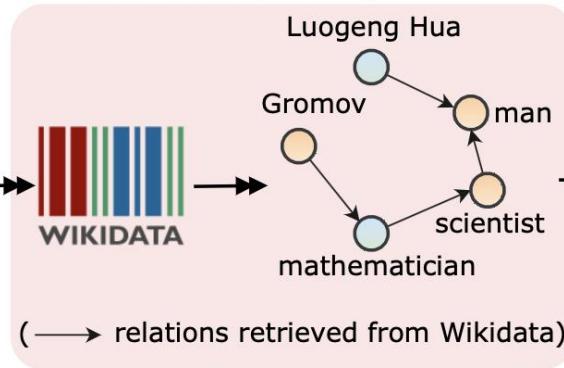
Multimodal Analogical Reasoning

Multimodal Analogical Reasoning over Knowledge Graph, ICLR 2023

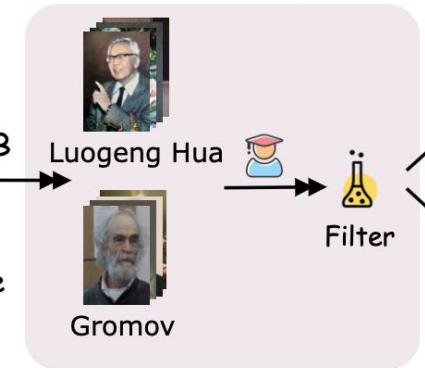
Step1: Collect Analogical Entities and Relations



Step2: Link to Wikidata and Retrieve Neighbors



Step3: Acquire and Validate Images



Step4: Sample



	# entity	# relation	# triple	# image	source
WN9-IMG	6,555	9	14,319	65,550	WordNet
FB15k-IMG	11,757	1,231	350,293	107,570	Freebase
MarKG	11,292	192	34,420	76,424	Wikidata

Table 5: Data statistics of MarKG. # refers to the number of.

Dataset	Size (train / dev / test)	KB	Modality	# Entity	# Relation	# Images	Knowledge Intensive	Task Format
RAVEN	42,000 / 14,000 / 14,000	✗	Vision	-	8	1,120,000	✗	Classification
SAT	0 / 37 / 337	✗	Text	-	19	-	✗	Linear Word Analogy
Google	0 / 50 / 500	✗	Text	919	14	-	✗	Linear Word Analogy
BATs	0 / 199 / 1,799	✗	Text	6,218	40	-	✗	Linear Word Analogy
E-KAR	870 / 119 / 262	✗	Text	2,651	28	-	✓	Multiple Choice QA
MARS	10,685 / 1,228 / 1,415	MarKG	Vision+Text	2,063	27	13,398	✓	Entity Prediction

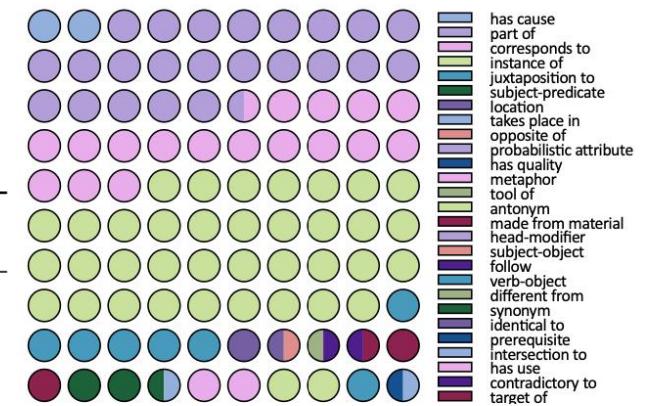
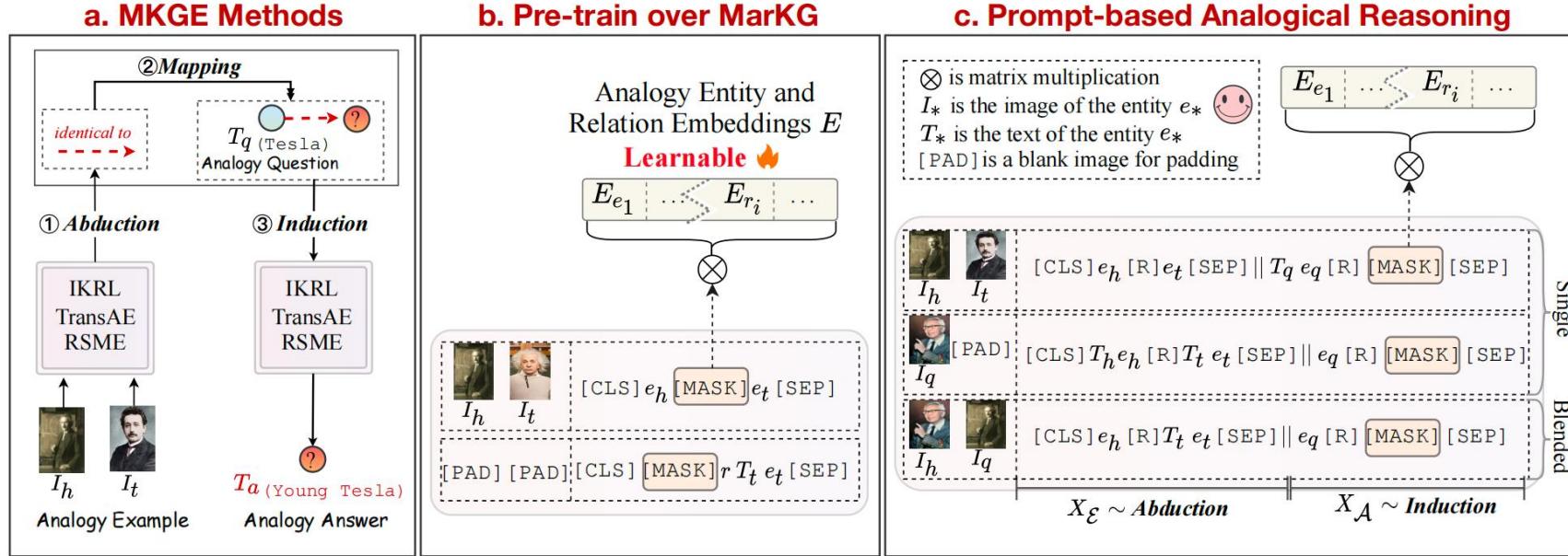
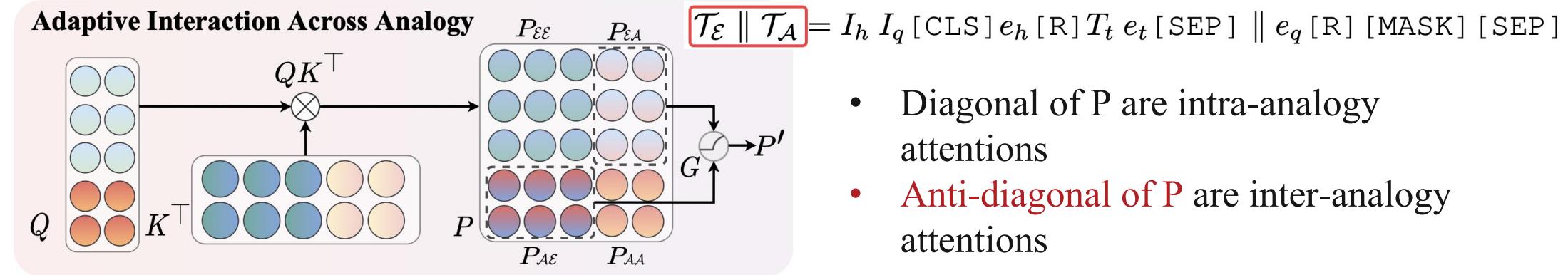


Figure 7: Relation distribution of MARS.



Method	Baselines	Backbone	Hits@1	Hits@3	Hits@5	Hits@10	MRR
MKGE	IKRL	TransE	0.254	0.285	0.290	0.304	0.274
	TransAE	TransE	0.203	0.233	0.241	0.253	0.223
	RSME	ComplEx	0.255	0.274	0.282	0.291	0.268
	IKRL	ANALOGY	0.266	0.294	0.301	0.310	0.283
	TransAE	ANALOGY	0.261	0.285	0.289	0.293	0.276
	RSME	ANALOGY	0.266	0.298	0.307	0.311	0.285
MPT	VisualBERT	Single-Stream	0.247	0.281	0.289	0.303	0.269
	ViLT	Single-Stream	0.235	0.266	0.274	0.286	0.257
	ViLBERT	Dual-Stream	0.252	0.308	0.320	0.338	0.287
	FLAVA	Mixed-Stream	0.257	0.299	0.312	0.325	0.284
	MKGformer	Mixed-Stream	0.293	0.335	0.344	0.367	0.321

MarT: A Multimodal Analogical Reasoning Framework with Transformer



■ Attention values and keys

$$Q = XW^Q = \begin{pmatrix} X_\mathcal{E} \\ X_\mathcal{A} \end{pmatrix} W^Q = \begin{pmatrix} Q_\mathcal{E} \\ Q_\mathcal{A} \end{pmatrix}, K = XW^K = \begin{pmatrix} X_\mathcal{E} \\ X_\mathcal{A} \end{pmatrix} W^K = \begin{pmatrix} K_\mathcal{E} \\ K_\mathcal{A} \end{pmatrix}$$

■ Decompose attention scores into intra-analogy and inter-analogy

$$P = QK^\top = \begin{pmatrix} Q_\mathcal{E} \\ Q_\mathcal{A} \end{pmatrix} (K_\mathcal{E}^\top, K_\mathcal{A}^\top) = \begin{pmatrix} Q_\mathcal{E}K_\mathcal{E}^\top & Q_\mathcal{E}K_\mathcal{A}^\top \\ Q_\mathcal{A}K_\mathcal{E}^\top & Q_\mathcal{A}K_\mathcal{A}^\top \end{pmatrix} = \begin{pmatrix} P_{\mathcal{E}\mathcal{E}} & P_{\mathcal{E}\mathcal{A}} \\ P_{\mathcal{A}\mathcal{E}} & P_{\mathcal{A}\mathcal{A}} \end{pmatrix}$$

■ Do inter-analogy interactions adaptively

MarT: A Multimodal Analogical Reasoning Framework with Transformer

Relation-Oriented Structure Mapping



“relations between objects, rather than attributes of objects, are mapped from base to target.” -- Structure Mapping Theory

- Bring the relations closer and alienate the entities

$$\mathcal{L}_{\text{rel}} = \frac{1}{|\mathcal{S}|} \sum_i \underbrace{(1 - \text{sim}(h_{[R]}^{\mathcal{E}}, h_{[R]}^{\mathcal{A}}))}_{\text{close relations}} + \underbrace{\max(0, \text{sim}(h_{e_h}, h_{e_q}))}_{\text{alienate entities}}$$

- Cross-entropy loss in masked entity prediction

$$\mathcal{L}_{\text{mem}} = -\frac{1}{|\mathcal{S}|} \sum_{(e_h, e_t, e_q, e_a) \in \mathcal{S}} \log(p([\text{MASK}] = e_a | \mathcal{T}_{(e_h, e_t, e_q)}))$$

- Final loss function

$$\mathcal{L} = \lambda \mathcal{L}_{\text{rel}} + (1 - \lambda) \mathcal{L}_{\text{mem}}$$

Experiments

Method	Baselines	Backbone	Hits@1	Hits@3	Hits@5	Hits@10	MRR
MKGE	IKRL	TransE	0.254	0.285	0.290	0.304	0.274
	TransAE	TransE	0.203	0.233	0.241	0.253	0.223
	RSME	ComplEx	0.255	0.274	0.282	0.291	0.268
	IKRL	ANALOGY	0.266	0.294	0.301	0.310	0.283
	TransAE	ANALOGY	0.261	0.285	0.289	0.293	0.276
	RSME	ANALOGY	0.266	0.298	0.307	0.311	0.285
MPT	VisualBERT	Single-Stream	0.247	0.281	0.289	0.303	0.269
	ViLT	Single-Stream	0.235	0.266	0.274	0.286	0.257
	ViLBERT	Dual-Stream	0.252	0.308	0.320	0.338	0.287
	FLAVA	Mixed-Stream	0.257	0.299	0.312	0.325	0.284
	MKGformer	Mixed-Stream	0.293	0.335	0.344	0.367	0.321
	MarT_VisualBERT	Single-Stream	0.261	0.292	0.308	0.321	0.284
	MarT_ViLT	Single-Stream	0.245	0.275	0.287	0.303	0.266
	MarT_ViLBERT	Dual-Stream	0.256	0.312	0.327	0.347	0.292
	MarT_FLAVA	Mixed-Stream	0.264	0.303	0.309	0.319	0.288
	MarT_MKGformer	Mixed-Stream	0.301	0.367	0.380	0.408	0.341

- The performance of MKGE and MPT methods are comparable.
- The analogical structures significantly improve performance.
- MarT_MKGformer perform best.

Viewpoint 1

This survey summarizes that the large language models are a great progress, but in the real application, there are some faithful (factual correctness) problems which may need to be supervised.

Viewpoint 2

Whether large language models are all we need? How to bridge language understanding and reasoning?

Interface of LM and Symbolic (prompt)

Viewpoint 3

We need more sound open datasets, tools, evaluation methods, etc.

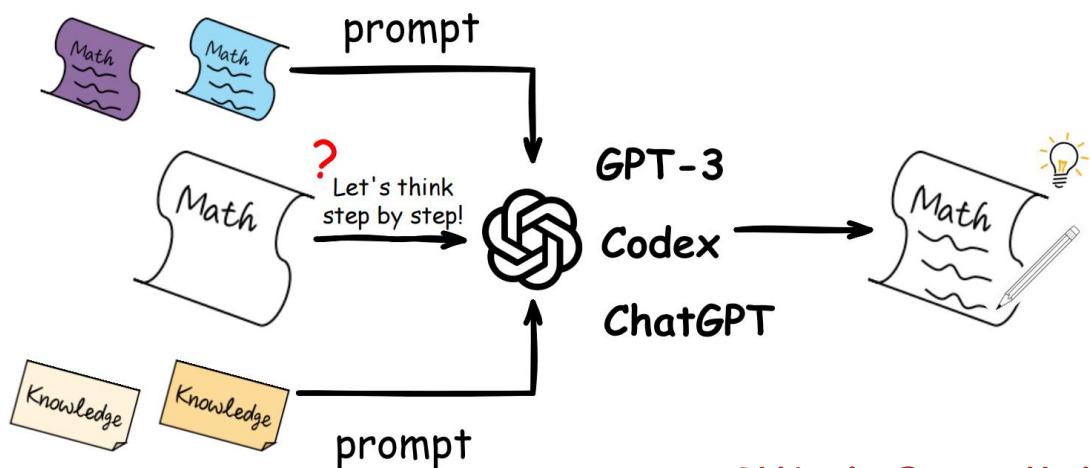
Related Surveys

1. Towards Reasoning in Large Language Models: A Survey.
2. A Survey of Deep Learning for Mathematical Reasoning.
3. A Survey for In-context Learning.
4. Knowledge-enhanced Neural Machine Reasoning: A Review.
5. Augmented Language Models: a Survey.



QR Code of our Github

Reasoning with Language Model Prompting: A Survey



we provide a review of reasoning with language model prompting, including comprehensive comparisons, and several research directions. In the future, we envision a more potent synergy between the methodologies from the NLP and other domains and hope sophisticated and efficient LM prompting models will increasingly contribute to improving reasoning performance.

Github Paperlist: <https://github.com/zjunlp/Prompt4ReasoningPapers>
Paper: <https://arxiv.org/abs/2212.09597>



Thank You !