Chain-of-Thought Prompting for LLMs

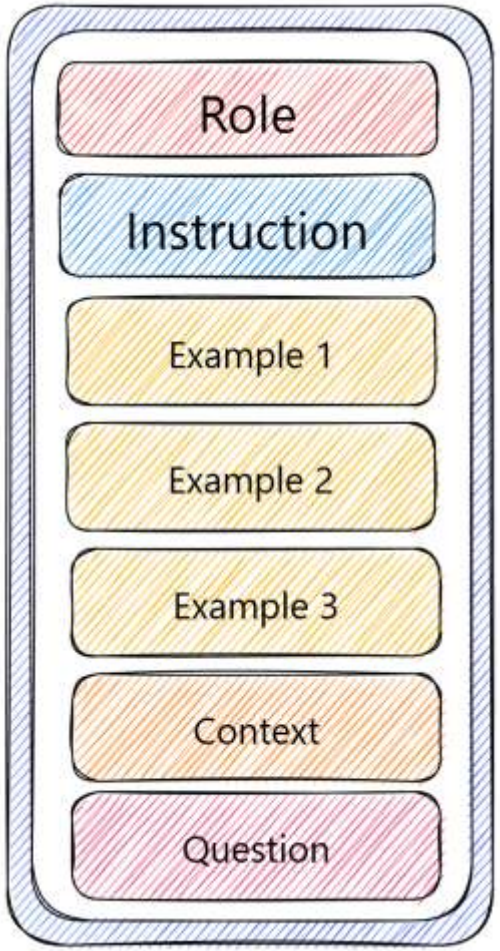
Prompt Engineering

生成式AI模型依赖于用户提供的输入（或“提示”）。良好的提示是利确保模型提供精确答案的关键。

- 有助于减少幻觉，排除不相关的内容或不正确的响应。
- 提高对任务的理解力，提高格式化输出的能力

属于大模型优化的**末端策略**：

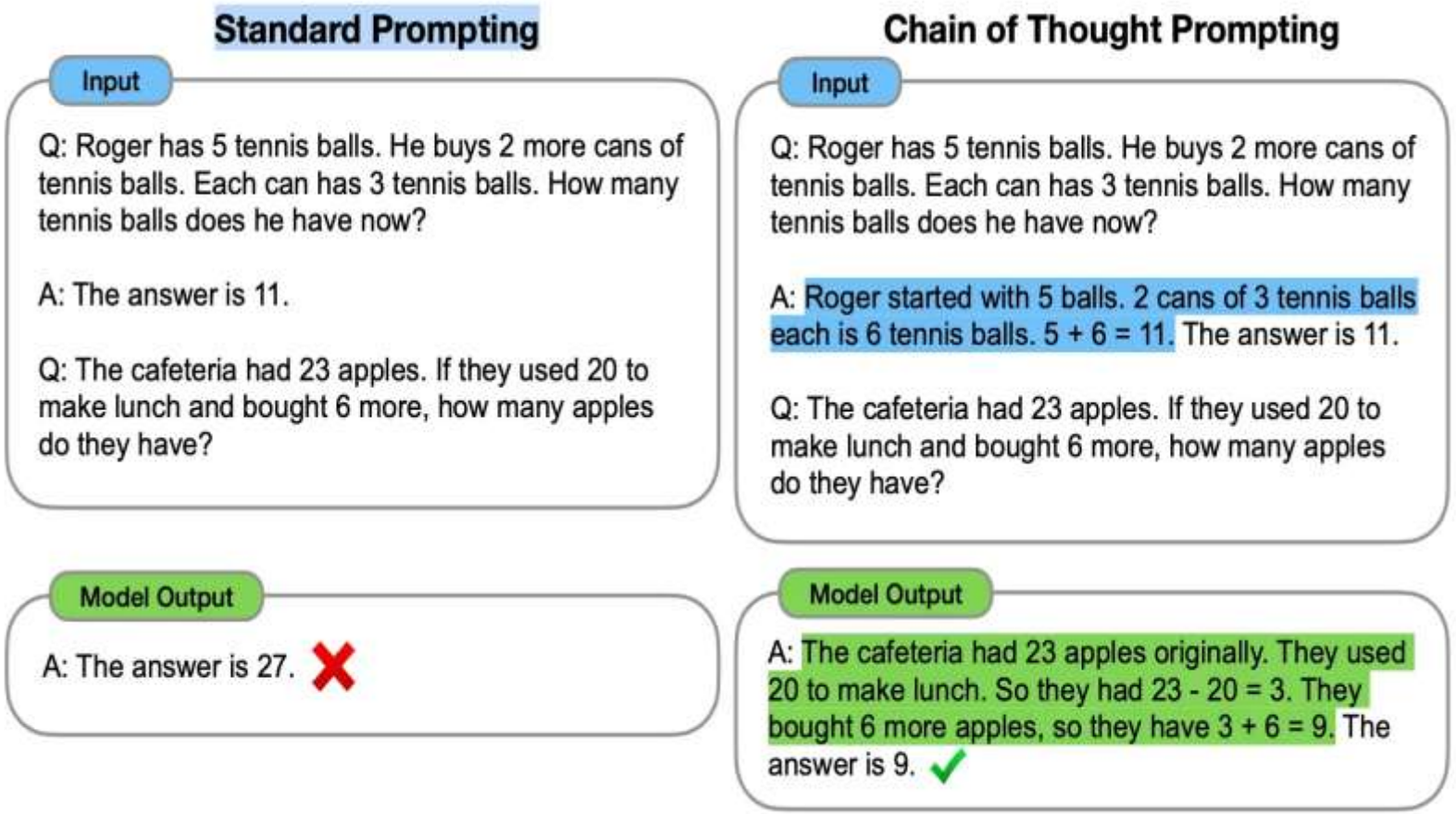
不需要训练即可优化模型
即插即用，适用于大多数任务。



Basic Prompt Structure and Key Parts

各部件可缺省，其最佳顺序与任务相关

Chain-of-Thought Prompting



思维链（CoT）提示是一种通过在提示中融入逻辑步骤来增强大型语言模型（LLM）推理能力的技术。

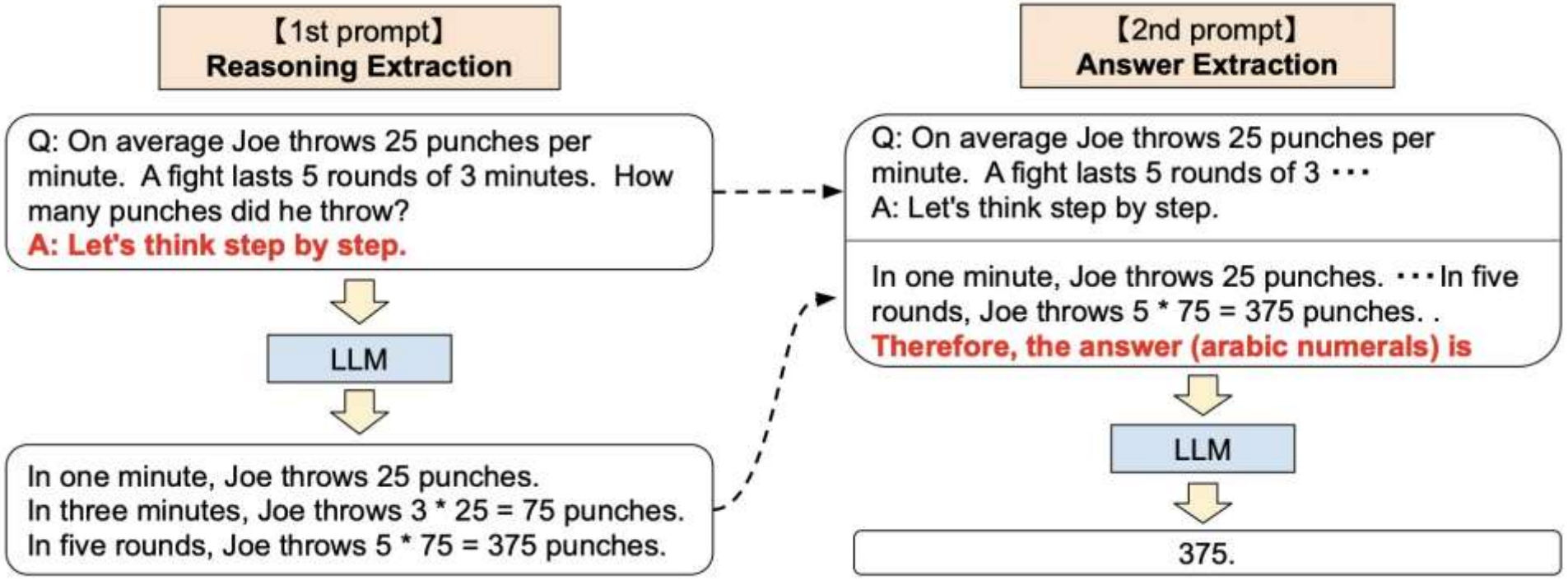
传统提示通常由简单的输入输出示例组成，缺乏明确的推理步骤。CoT则

- **鼓励多步推理**
- **无需微调即可实现**

Regular Prompting vs CoT

Wei (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models

Zero-Shot Chain-of-Thought



核心思想：在问题后加上

“Let's think step by step”

对算术推理、常识推理和符号推理任务效果显著；**但在复杂推理任务上通常不如传统的链式思维（CoT）提示。**

Kojima(2022). Large Language Models are Zero-Shot Reasoners.

Auto-CoT

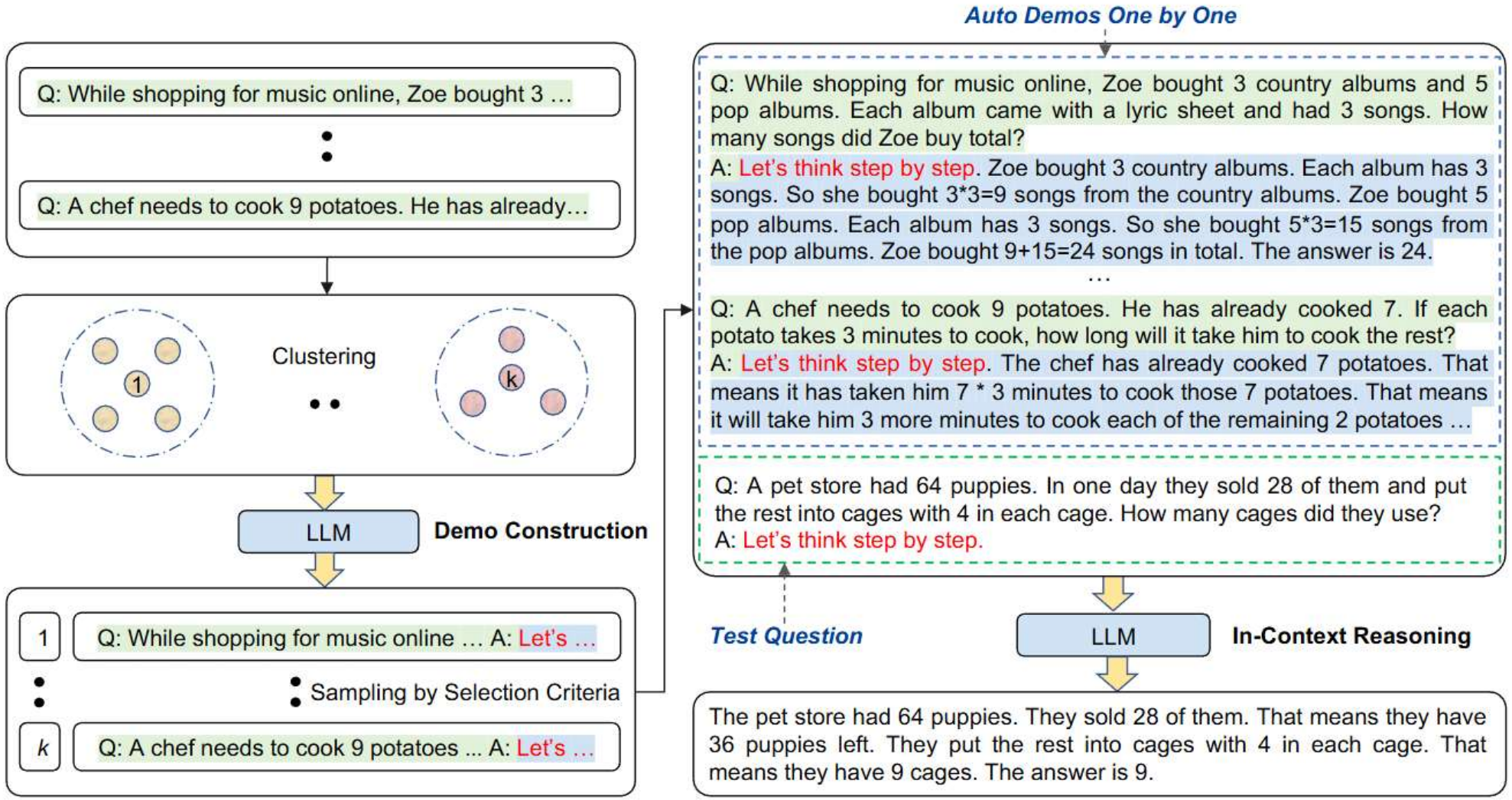
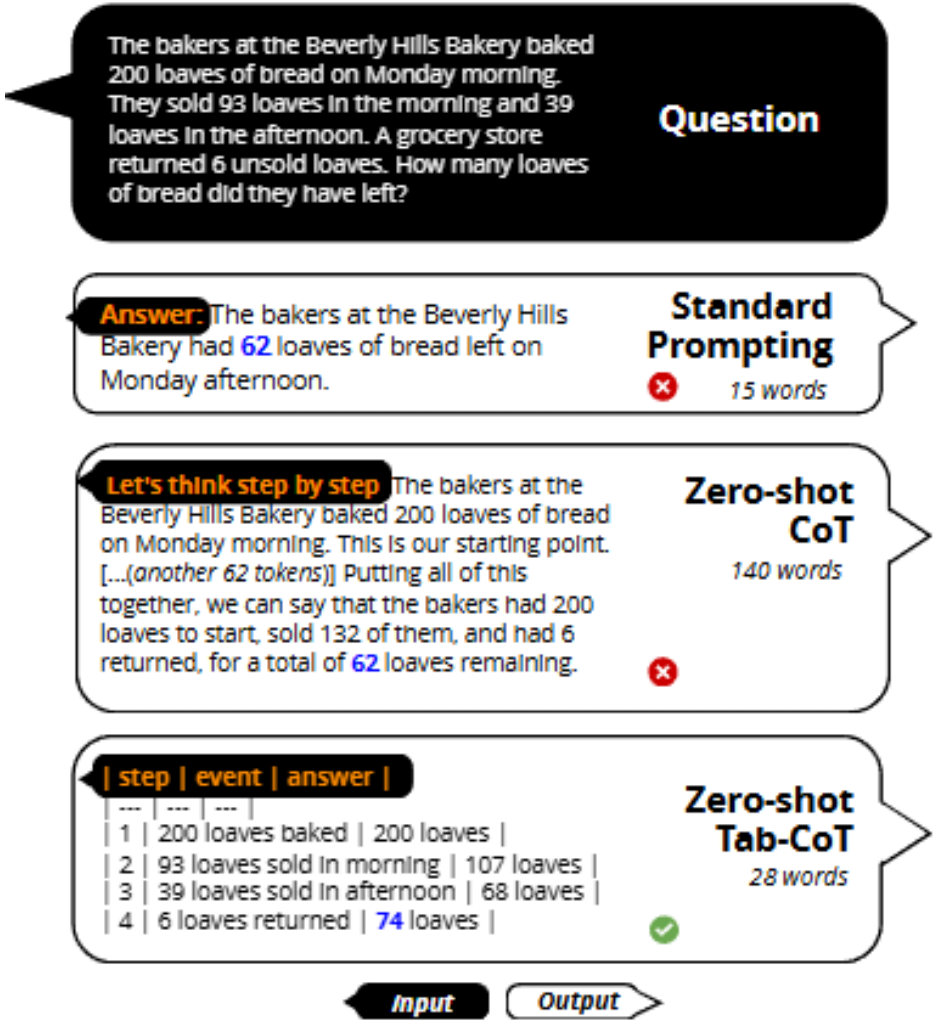


Figure 4: Overview of the Auto-CoT method. Different from Manual-CoT in Figure 1, demonstrations (on the right) are automatically constructed one by one (total: k) using an LLM with the "Let's think step by step" prompt.

Zhang(2023). Automatic Chain of Thought Prompting in Large Language Models. In The Eleventh International Conference on Learning Representations

Tab-CoT



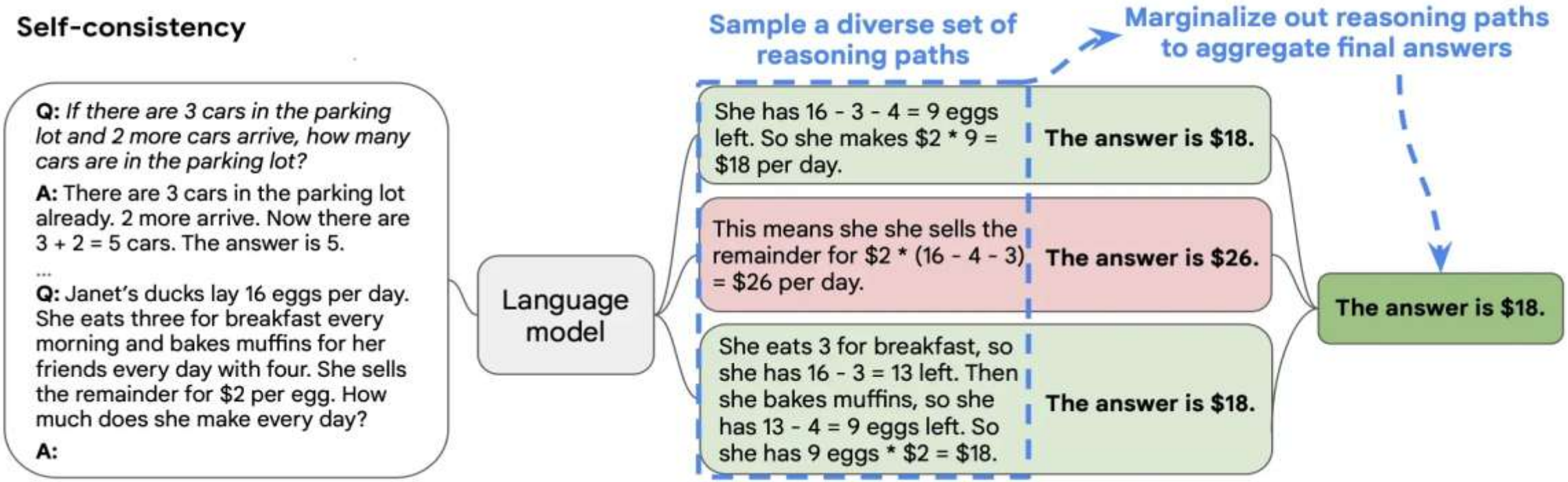
零样本CoT通过“Let's think step by step”来引导LLM进行推理。然而，这些方法往往冗长，输出往往不够有条理。

相比之下，Tab-CoT生成简洁、结构化的表格推理步骤。

- 它支持二维推理，使模型能够检查行列之间的一致性。

Jin, Z., & Lu, W. (2023). Tab-CoT: Zero-shot Tabular Chain of Thought.

Self-Consistency



对同一个提示让模型生成多次回答，然后**取多数结果作为最终答案**。

Wang(2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models.

Chain of Knowledge (CoK)

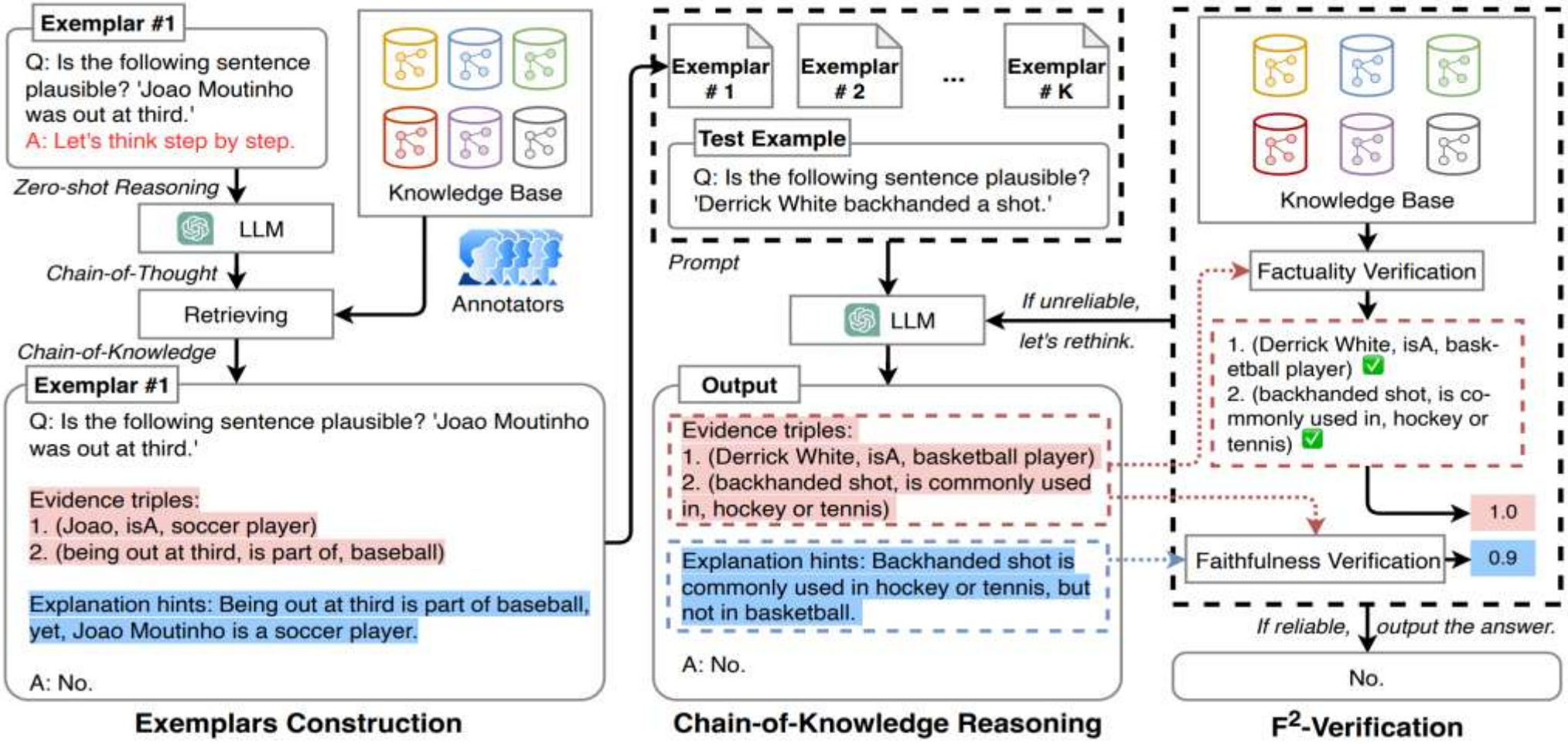
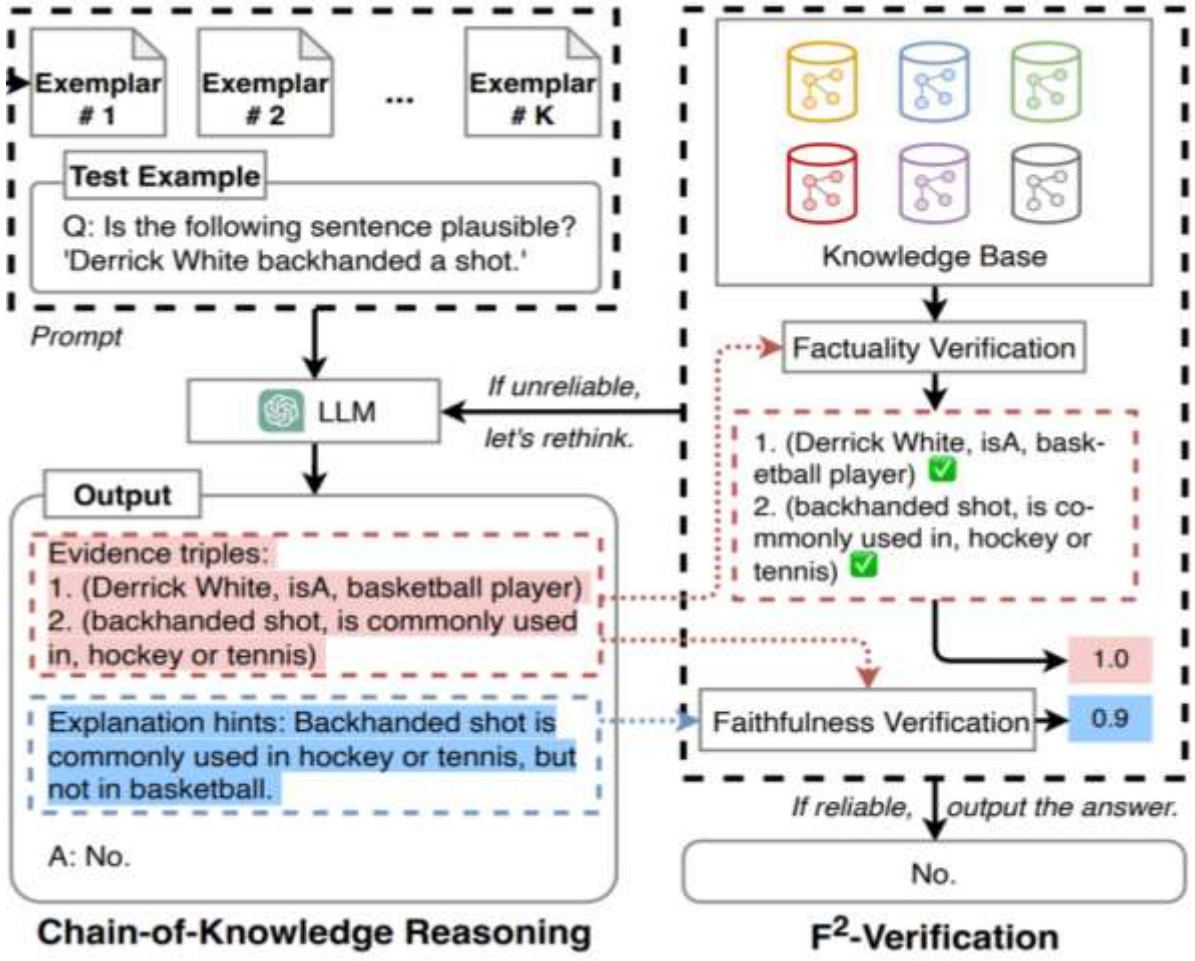


Figure 2: The proposed framework. We first construct exemplars with chain-of-knowledge (CoK) prompts. Then, the CoK prompts can be used to let the LLM generate reasoning chains, including evidence triples, explanation hints, and the final answer. Lastly, we estimate the reliability of reasoning chains in terms of *factuality* and *faithfulness*, and the unreliable ones will be rethought.

Chain of Knowledge (CoK)



$$\mathcal{E} = \{(Q_i, T_i, H_i, A_i)\}_{i=1}^{K}$$

- Factuality Verification——CoK-ET（证据三元组）

  生成的（主，谓，宾）三元组与知识库中的真实知识的匹配度。

- Faithfulness Verification——CoK-EH（解释提示）

$$f_u(\hat{H}_i|\hat{H}'_i = [\hat{Q}_i; \hat{T}_i; \hat{A}_i]) = SimCSE(\hat{H}_i, \hat{H}'_i).$$

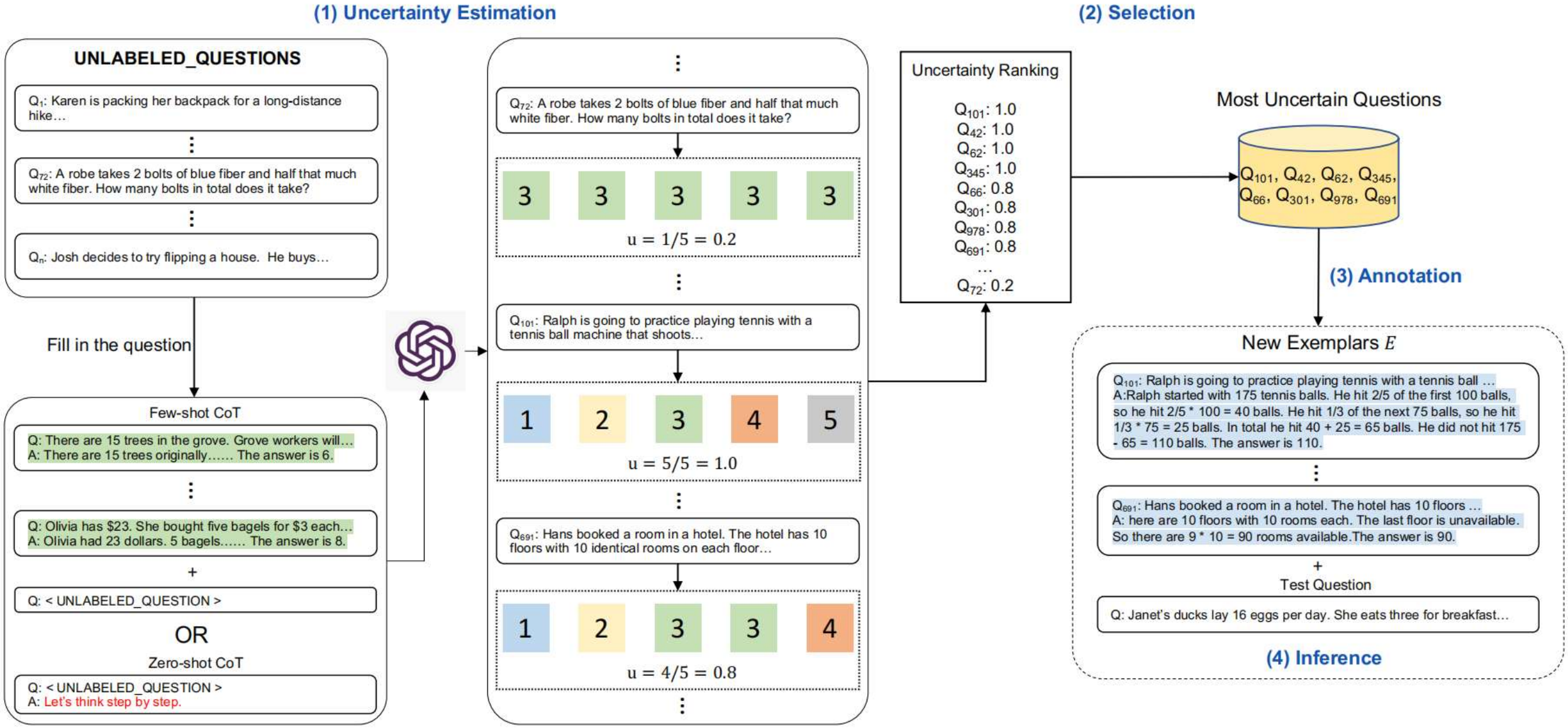
- Rethinking Process

  大模型重新评估 $\hat{T}_i^{(n)}$ and $\hat{H}_i^{(n)}$ 的可靠性，抛弃不可靠的条目。

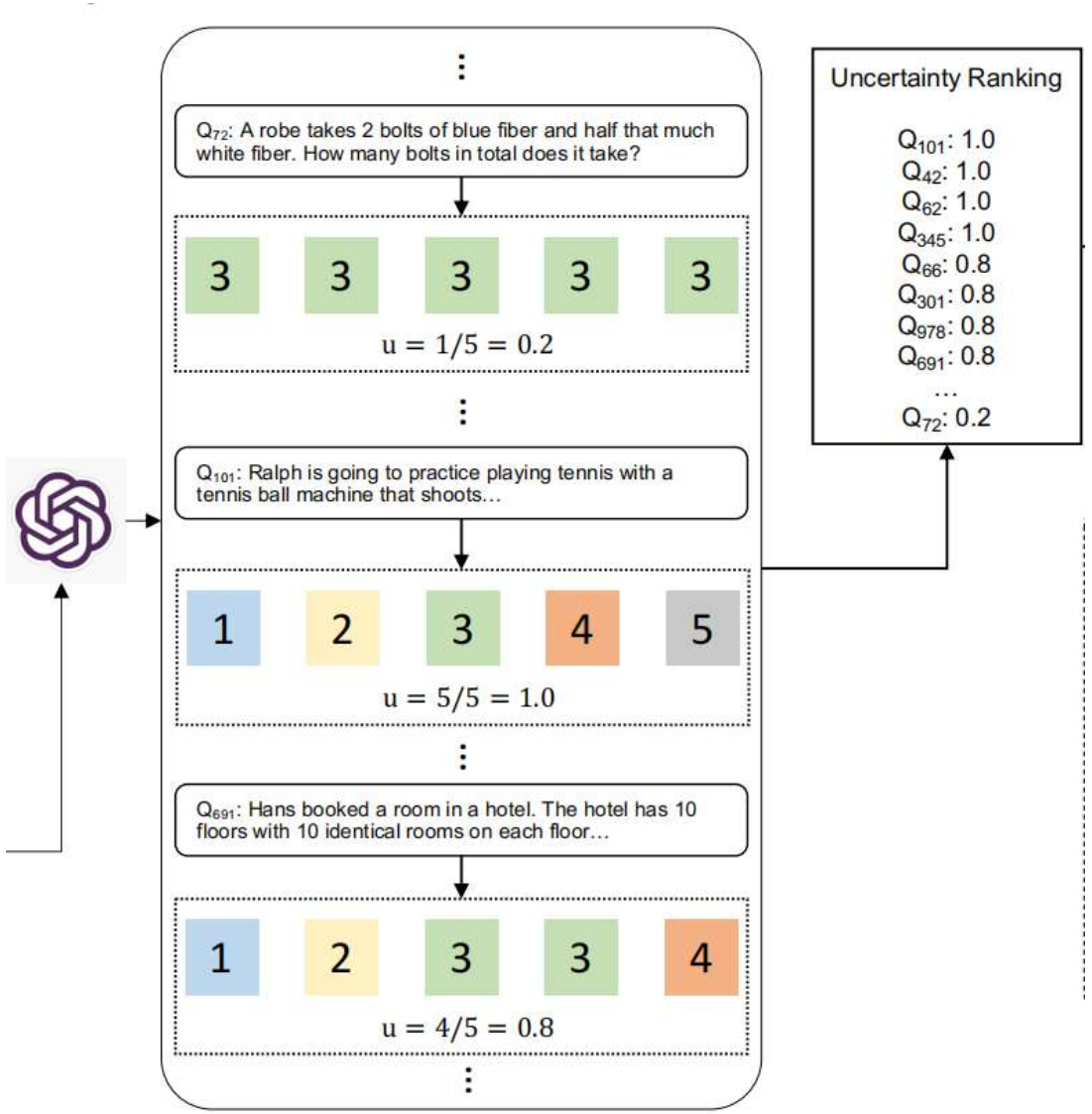
Active Prompting with Chain-of-Thought for Large Language Models

Shizhe Diao[♠], **Pengcheng Wang**[♡], **Yong Lin**[♠], **Rui Pan**[♠], **Xiang Liu**[♣], **Tong Zhang**[♦]
[♠]The Hong Kong University of Science and Technology
[♡]University of Toronto [♣]The University of Hong Kong
[♦]University of Illinois Urbana-Champaign
{sdiaoaa, ylindf, rpan}@connect.ust.hk pcheng.wang@mail.utoronto.ca
xiang.liu@connect.hku.hk tozhang@illinois.edu

Active Prompting



Selector



分歧（Disagreement）-不确定性

信息熵（Entropy）-不确定性

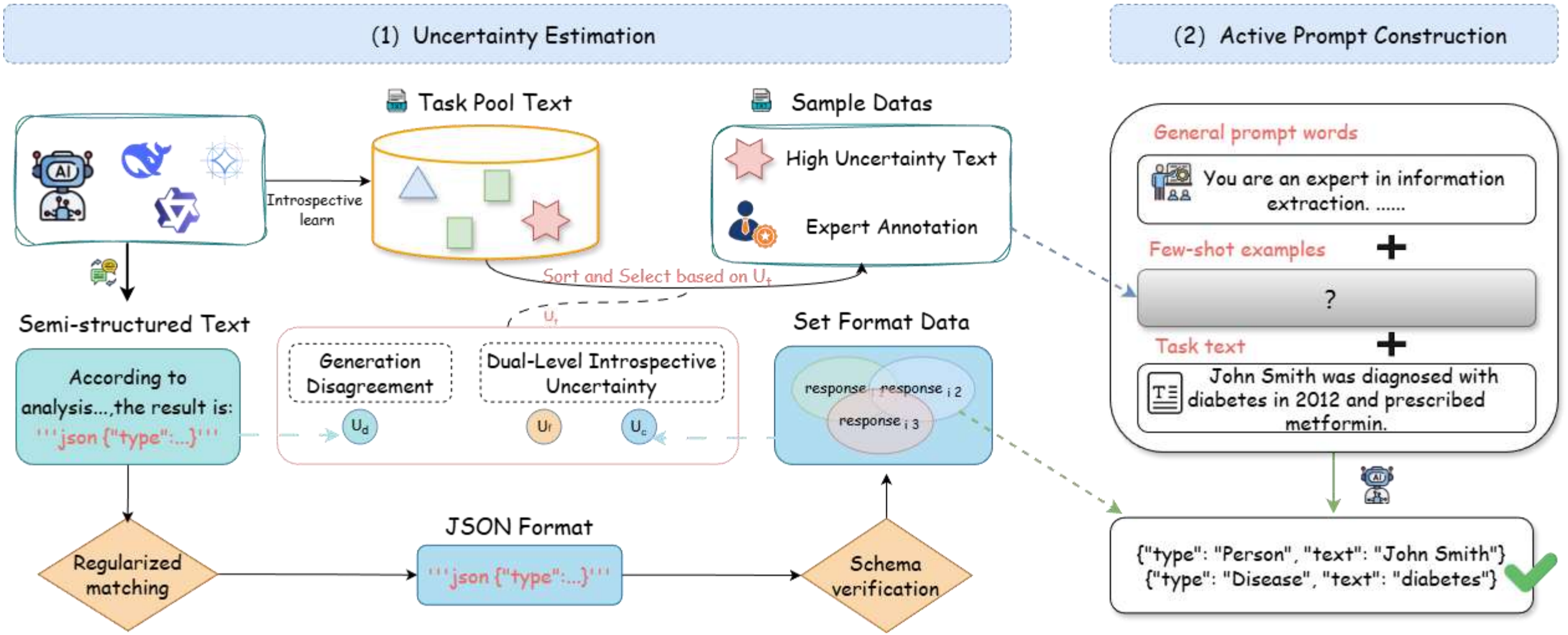
$$u = \arg\max_i - \sum_{j=1}^{k} P_\theta(a_j|q_i) \ln P_\theta(a_j|q_i), \quad (1)$$

$P_\theta(a_j|q_j)$表示在所有预测中某一预测答案的出现频率。

Active Prompting

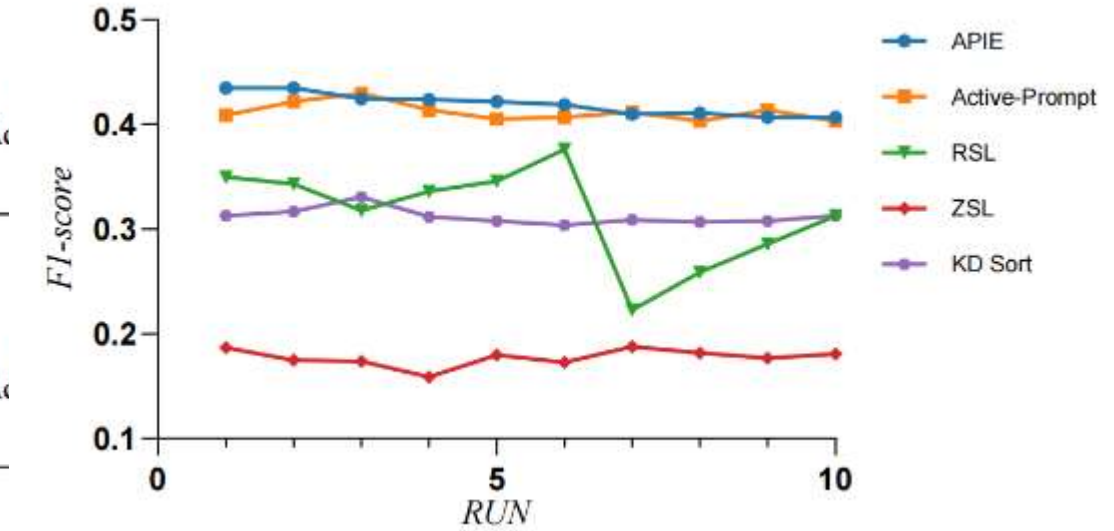
METHOD	GSM8K	ASDIV	SVAMP	AQUA	SINGLEEQ	CSQA	STRATEGY	LETTER (4)	AVG.
Prior Best	55.0[a]	75.3[b]	57.4[c]	37.9[d]	32.5[e]	91.2[f]	73.9[g]	-	-
text-davinci-002									
Auto-CoT	47.9	-	69.5	36.5	87.0	74.4	65.4	59.7	-
CoT	46.9	71.3	68.9	35.8	77.3	73.5	65.4	56.6	61.5
SC	58.2	76.9	78.2	41.8	87.2	72.9	70.7	57.6	67.9
Random-CoT	63.9	82.3	81.1	44.1	89.4	74.5	73.3	65.5	71.8
Active-Prompt (D)	**73.2**	83.2	**82.7**	48.4	90.6	76.6	**76.9**	**67.7**	74.9
Active-Prompt (E)	71.1	**83.8**	81.8	**50.3**	**93.1**	**78.8**	**76.9**	66.7	**75.3**
code-davinci-002									
Auto-CoT	62.8	-	-	-	-	-	-	-	-
CoT	63.1	80.4	76.4	45.3	93.1	77.9	73.2	70.4	72.5
SC	78.0	87.8	86.8	52.0	93.7	81.5	79.8	73.4	79.1
Random-CoT	78.6	87.1	88.0	53.1	94.0	82.1	79.4	73.3	79.4
Active-Prompt (D)	82.2	88.4	**88.7**	55.1	94.5	**83.9**	**80.6**	74.1	80.9
Active-Prompt (E)	**83.4**	**89.3**	87.5	**57.0**	**95.5**	82.6	**80.6**	**76.7**	**81.6**
gpt-3.5-turbo-0613 (w.o. SC)									
CoT	74.2	82.5	83.8	50.0	95.0	79.9	80.5	82.0	78.5
Active-Prompt (D)	77.1	83.6	85.5	50.0	**96.0**	**81.5**	**82.1**	**84.0**	80.0
Active-Prompt (E)	**78.2**	**84.7**	**86.0**	**57.3**	95.5	80.7	81.3	**84.0**	**81.0**
gpt-3.5-turbo-0301 (w.o. SC)									
CoT	80.1	86.7	82.0	56.2	91.3	74.6	64.4	81.4	77.1
Active-Prompt (D)	83.5	87.4	83.0	60.6	93.3	**75.9**	70.0	**84.0**	79.7
Active-Prompt (E)	**83.8**	**88.8**	**83.7**	**61.0**	**93.7**	75.0	**71.0**	**84.0**	**80.1**

Work


(1) Uncertainty Estimation
(2) Active Prompt Construction
Task Pool Text
Sample Datas
Introspective learn
High Uncertainty Text
Expert Annotation
Sort and Select based on U_t
Semi-structured Text
U_t
Generation Disagreement
Dual-Level Introspective Uncertainty
Set Format Data
According to analysis...,the result is: '''json {"type":...}'''
U_d
U_r
U_c
response i 1
response i 2
response i 3
Regularized matching
JSON Format
'''json {"type":...}'''
Schema verification
General prompt words
You are an expert in information extraction. .....
Few-shot examples
?
Task text
John Smith was diagnosed with diabetes in 2012 and prescribed metformin.
{"type": "Person", "text": "John Smith"}
{"type": "Disease", "text": "diabetes"}

Work

Models	Methods	CoNLL03 NER F1-score↑	ACE04-NER NER F1-score↑	CoNLL04 NER F1-score↑	CoNLL04 RE F1-score↑	SciERC NER F1-score↑	SciERC RE F1-score↑	Average F1-score↑
Gemma-3-12B	Ac				0.00	24.12 ↑0.00	14.53 ↑0.00	34.76 ↑0.00
					6.32	**29.57** ↑5.45	14.43 ↓0.11	40.32 ↑5.56
					6.11	27.25 ↑3.13	9.84 ↓4.69	40.38 ↑5.63
					7.78	25.45 ↑1.33	**19.25** ↑4.72	40.23 ↑5.47
					8.23	26.21 ↑2.09	16.85 ↑2.32	**40.84** ↑6.09
Qwen-2.5-14B	Ac				0.00	20.51 ↑0.00	18.08 ↑0.00	28.31 ↑0.00
					6.64	27.94 ↑7.43	**19.23** ↑1.14	39.04 ↑10.74
					8.74	27.42 ↑6.90	9.55 ↓8.54	35.79 ↑7.48
					5.91	29.11 ↑8.60	13.74 ↓4.34	38.35 ↑10.04
					8.88	**31.40** ↑10.89	14.11 ↓3.97	**40.19** ↑11.88
Deepseek-R1-14B					0.00	26.19 ↑0.00	11.06 ↑0.00	32.73 ↑0.00
	RSL	57.88 ↑7.30	31.30 ↑13.73	63.97 ↓0.40	27.83 ↑0.54	26.87 ↑0.68	11.30 ↑0.24	36.56 ↑3.83
	KD Sort	51.47 ↑0.90	31.24 ↑13.47	63.48 ↓0.04	33.11 ↑5.81	29.17 ↑2.98	10.75 ↓0.31	36.53 ↑3.80
	Active-Prompt	57.34 ↑6.77	41.19 ↑23.42	64.93 ↑1.41	32.13 ↑4.83	30.82 ↑4.63	11.48 ↑0.43	39.65 ↑6.91
	APIE	**59.28** ↑8.71	**41.95** ↑24.18	**66.32** ↑2.81	**34.62** ↑7.33	**33.09** ↑6.90	**13.58** ↑2.52	**41.47** ↑8.74
Deepseek-v3-660B	ZSL	65.76 ↑0.00	24.87 ↑0.00	68.51 ↑0.00	31.74 ↑0.00	33.44 ↑0.00	20.30 ↑0.00	40.77 ↑0.00
	RSL	69.71 ↑3.95	33.31 ↑8.44	71.29 ↑2.79	42.32 ↑10.58	40.43 ↑6.99	20.35 ↑0.04	46.23 ↑5.46
	KD Sort	69.25 ↑3.49	38.62 ↑13.75	72.57 ↑4.06	**51.01** ↑19.27	37.98 ↑4.54	14.20 ↓6.10	47.27 ↑6.50
	Active-Prompt	72.48 ↑6.72	47.33 ↑22.46	72.75 ↑4.25	47.66 ↑15.92	41.89 ↑8.46	18.90 ↓1.41	50.17 ↑9.40
	APIE	**72.73** ↑6.97	**48.56** ↑23.68	**72.86** ↑4.36	50.20 ↑18.46	**42.05** ↑8.61	**21.45** ↑1.15	**51.31** ↑10.54



Thanks