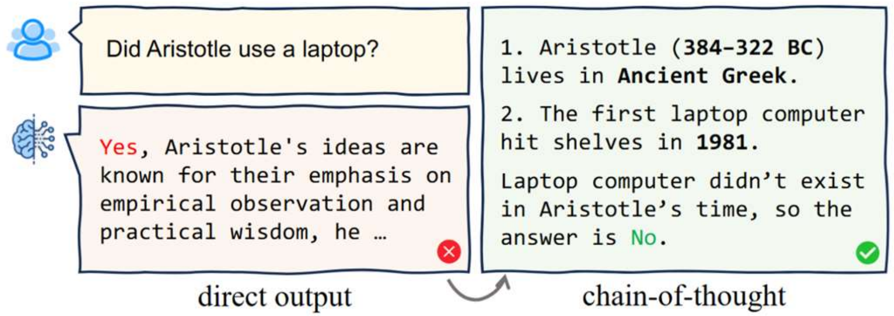
Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future

Zheng Chu*   Jingchang Chen*   Qianglong Chen*
Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, Ting Liu

Introduction

What is chain-of-thought reasoning?

- Background: Paradigm shift from finetuning to in-context learning.
- Characteristic: Conduct step-by-step reasoning before final answer.



What are the benefits of chain-of-thought reasoning?

- Reduce problem complexity for enhanced accuracy.
- Observable reasoning trajectory, offering trustworthy and interpretability.

Generalized chain-of-though reasoning (XoT)

- The core philosophy of XoT reasoning is the gradual unraveling of complex problems via a step-by-step reasoning approach.

Reasoning Benchmarks

Various benchmarks have been proposed to evaluate LLM's reasoning capabilities.

Task	Dataset	Size	Input	Output	Rationale	Description
Mathematical Reasoning	AddSub (Hosseini et al., 2014)	395	Question	Number	Equation	Simple arithmetic
	SingleEq (Koncel-Kedziorski et al., 2015)	508	Question	Number	Equation	Simple arithmetic
	MultiArith (Roy and Roth, 2015)	600	Question	Number	Equation	Simple arithmetic
	MAWPS (Koncel-Kedziorski et al., 2016)	3,320	Question	Number	Equation	Simple arithmetic
	AQUA-RAT (Ling et al., 2017)	100,000	Question	Option	Natural Language	Math reasoning with NL rationale
	ASDiv (Miao et al., 2020)	2,305	Question	Number	Equation	Multi-step math reasoning
	SVAMP (Patel et al., 2021)	1,000	Question	Number	Equation	Multi-step math reasoning
	GSM8K (Cobbe et al., 2021)	8,792	Question	Number	Natural Language	Multi-step math reasoning
	GSM-Hard (Gao et al., 2023)	936	Question	Number	Natural Language	GSM8K with larger number
	MathQA (Amini et al., 2019)	37,297	Question	Number	Operation	Annotated based on AQUA
	DROP (Dua et al., 2019)	96,567	Question+Passage	Number+Span	Equation	Reading comprehension form
	TheoremQA (Chen et al., 2023b)	800	Question+Theorem	Number	✗	Answer based on theorems
	TAT-QA (Zhu et al., 2021)	16,552	Question+Table+Text	Number+Span	Operation	Answer based on tables
	FinQA (Chen et al., 2021)	8,281	Question+Table+Text	Number	Operation	Answer based on tables
	ConvFinQA (Chen et al., 2022b)	3,892	Question+Table+Dialog	Number	Operation	Multi-turn dialogs
	MATH (Hendrycks et al., 2021b)	12,500	Question	Number	Natural Language	Challenging competition math problems
	NumGLUE (Mishra et al., 2022b)	101,835	Question+Text	Number+Span	✗	Multi-task benchmark
	LILA (Mishra et al., 2022a)	133,815	Question+Text	Free-form	Program	Multi-task benchmark
Commonsense Reasoning	ARC (Bhakthavatsalam et al., 2021)	7,787	Question	Option	✗	From science exam
	OpenBookQA (Mihaylov et al., 2018)	5,957	Question+Context	Option	✗	Open-book knowledges
	PIQA (Bisk et al., 2020)	21,000	Goal+Solution	Option	✗	Physical commonsense knowledge
	CommonsenseQA (Talmor et al., 2019)	12,247	Question	Option	✗	Derived from ConceptNet
	CommonsenseQA 2.0 (Talmor et al., 2021)	14,343	Question	Yes/No	✗	Gaming annotation with high quality
	Event2Mind (Rashkin et al., 2018)	25,000	Event	Intent+Reaction	✗	Intension commonsense reasoning
	McTaco (Zhou et al., 2019)	13,225	Question	Option	✗	Event temporal commonsense reasoning
	CosmosQA (Huang et al., 2019)	35,588	Question+Paragraph	Option	✗	Narrative commonsense reasoning
	ComValidation (Wang et al., 2019)	11,997	Statement	Option	✗	Commonsense verification
	ComExplanation (Wang et al., 2019)	11,997	Statement	Option/Free-form	✗	Commonsense explanation
	StrategyQA (Geva et al., 2021)	2,780	Question	Yes/No	✗	Multi-hop commonsense reasoning
Symbolic Reasoning	Last Letter Concat. (Wei et al., 2022b)	-	Words	Letters	✗	Rule-based
	Coin Flip (Wei et al., 2022b)	-	Statement	Yes/No	✗	Rule-based
	Reverse List (Wei et al., 2022b)	-	List	Reversed List	✗	Rule-based
	BigBench (Srivastava et al., 2022)	-	-	-	✗	Contains multiple symbolic reasoning datasets
	BigBench-Hard (Suzgun et al., 2023)	-	-	-	✗	Contains multiple symbolic reasoning datasets
Logical Reasoning	ReClor (Yu et al., 2020)	6,138	Question+Context	Option	✗	Questions from GMAT and LSAT
	LogiQA (Liu et al., 2020)	8,678	Question+Paragraph	Option	✗	Questions from China Civil Service Exam
	ProofWriter (Tafjord et al., 2021)	20,192	Question+Rule	Answer+Proof	Entailment Tree	Reasoning process generation
	FOLIO (Han et al., 2022)	1,435	Conclusion+Premise	Yes/No	✗	First-order logic
	DEER (Yang et al., 2024b)	1,200	Fact	Rule	✗	Inductive reasoning
	PrOntoQA (Saparov and He, 2023)	-	Question+Context	Yes/No+Proccess	First-Order Logic	Deductive reasoning
Multimodal Reasoning	VCR (Zellers et al., 2019)	264,720	Question+Image	Option	Natural Language	Visual commonsense reasoning
	VisualCOMET (Park et al., 2020)	1,465,704	Image+Event	Action+Intent	✗	Visual commonsense reasoning
	PMR (Dong et al., 2022)	15,360	Image+Background	Option	✗	Premise-based multi-modal reasoning
	ScienceQA (Lu et al., 2022)	21,208	Q+Image+Context	Option	Natural Language	Multi-modal reasoning with NL rationales
	VLEP (Lei et al., 2020)	28,726	Premise+Video	Option	✗	Video event prediction
	CLEVRER (Yi et al., 2020)	305,280	Question+Video	Option/Free-form	Program	Video temporal and causal reasoning
	STAR (Wu et al., 2021)	600,000	Question+Video	Option	✗	Video situated reasoning
	NEXT-QA (Xiao et al., 2021)	47,692	Question+Video	Option	✗	Video temporal,causal,commonsense reasoning
	Causal-VidQA (Li et al., 2022)	107,600	Question+Video	Free-form	Natural Language	Video causal and commonsense reasoning
	News-KVQA (Gupta and Gupta, 2022)	1,041,352	Q+V+KG	Option	✗	Video reasoning with external knowledge

- **Mathematical Reasoning**
- **Commonsense Reasoning**
- **Symbolic Reasoning**
- **Logical Reasoning**
- **Multi-modal Reasoning**

Survey Organization

Our survey focuses on Advances Methods, Frontier Applications and Future Directions.

- A survey of X-of-Thought
  - Advanced Methods (§4)
    - XoT Prompt Construction (§4.1)
      - Manual Prompting — E.g., Few-shot CoT (Wei et al., 2022b) , PAL (Gao et al., 2023)
      - Automatic Prompting — E.g., Zero-shot CoT (Kojima et al., 2022), Auto-CoT (Zhang et al., 2023h)
      - Semi-auto Prompting — E.g., AutoMate CoT (Shum et al., 2023), BoostedPrompt (Pitis et al., 2023)
    - XoT Topological Variants (§4.2)
      - Chain Structure — E.g., PoT (Chen et al., 2022a), LINC (Olausson et al., 2023)
      - Tree Structure — E.g., ToT (Yao et al., 2023b), SoT (Ning et al., 2023)
      - Graph Structure — E.g., GoT (Besta et al., 2023), ResPrompt (Jiang et al., 2023a)
    - XoT Enhancement Methods (§4.3)
      - Verify and Refine — E.g., VerifyCoT (Ling et al., 2023), Self-Refine (Madaan et al., 2023)
      - Question Decomposition — E.g., L2M (Zhou et al., 2023b) , Successive Prompt (Dua et al., 2022)
      - Knowledge Enhancement — E.g., CoK (Wang et al., 2023c), KD-CoT (Wang et al., 2023e)
      - Self-Ensemble — E.g., Self-Consistency (Wang et al., 2023m) , Complex CoT (Fu et al., 2023a)
      - Efficient Reasoning — E.g., SoT (Ning et al., 2023), ActivePrompting (Diao et al., 2023)
  - Frontier (§5)
    - Tool Use (§5.1) — E.g., TAML (Parisi et al., 2022), Toolformer (Schick et al., 2023), ReACT (Yao et al., 2023c)
    - Planning (§5.2) — E.g., LLM+P (Liu et al., 2023a), RAP (Hao et al., 2023a), ToolChain* (Zhuang et al., 2024)
    - Distillation (§5.3) — E.g., STaR (Zelikman et al., 2022), SCoTD (Li et al., 2023c), SCOTT (Wang et al., 2023j)
  - Future Directions (§6)
    - Multi-modal (§6.1) — E.g., MMCoT (Zhang et al., 2023i), T-SciQ (Wang et al., 2023g), SocreticQues (Qi et al., 2023)
    - Faithfulness (§6.2) — E.g., Rethinking and Retrievaling (He et al., 2023a), Measure Faithful (Lanham et al., 2023)
    - Theory (§6.3) — E.g., Wang et al. (2023f), Tutunov et al. (2023), Wang et al. (2023a), Feng et al. (2023), Wu et al. (2023b)

Advances

- *How to construct CoT prompts?*
- *Topological variants.*
- *How to enhance CoT reasoning?*

Frontiers

- *Tool invocation*
- *Planning and decision making*
- *Distillation reasoning capability*

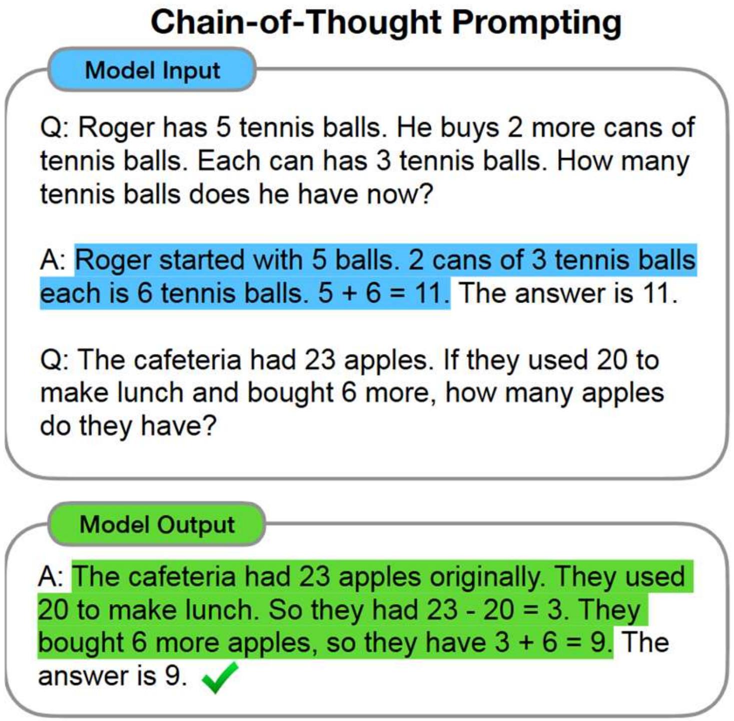
Future

- *Multi-modal CoT*
- *Faithful CoT Reasoning*
- *CoT mechanisms exploration*

How to Construct CoT Prompting?

Manual CoT Prompting Construction

- High-quality demonstrations annotations yield high performance.
- High cost, difficult to transfer, and challenging demo selection.
- Few-shot CoT[1] , Few-shot PoT[2].



[1] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022
[2] Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks, TMLR 2023
[3] Large Language Models are Zero-Shot Reasoners, NeurIPS 2022
[4] Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models, ACL 2023
[5] Automatic Chain of Thought Prompting in Large Language Models, ICLR 2023
[6] Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data, Findings of EMNLP 2023

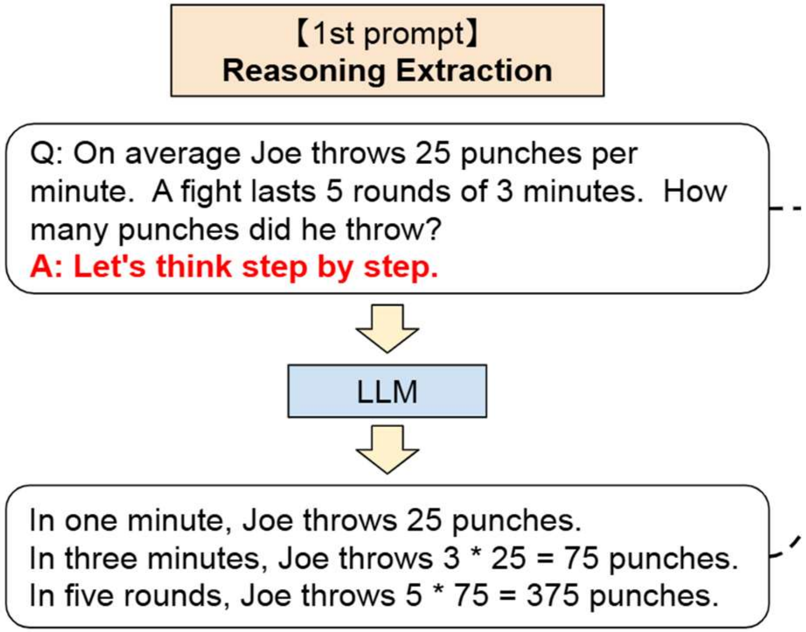
How to Construct CoT Prompting?

Manual CoT Prompting Construction

- High-quality demonstrations annotations yield high performance.
- High cost, difficult to transfer, and challenging demo selection.
- Few-shot CoT[1] , Few-shot PoT[2].

Automatic CoT Prompting Construction

- Low-quality demonstrations, low performance
- Low cost, easy to transfer.
- Zeroshot CoT[3], Plan-and-Solve Prompting[4].



[1] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022
[2] Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks, TMLR 2023
[3] Large Language Models are Zero-Shot Reasoners, NeurIPS 2022
[4] Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models, ACL 2023
[5] Automatic Chain of Thought Prompting in Large Language Models, ICLR 2023
[6] Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data, Findings of EMNLP 2023

How to Construct CoT Prompting?

Manual CoT Prompting Construction

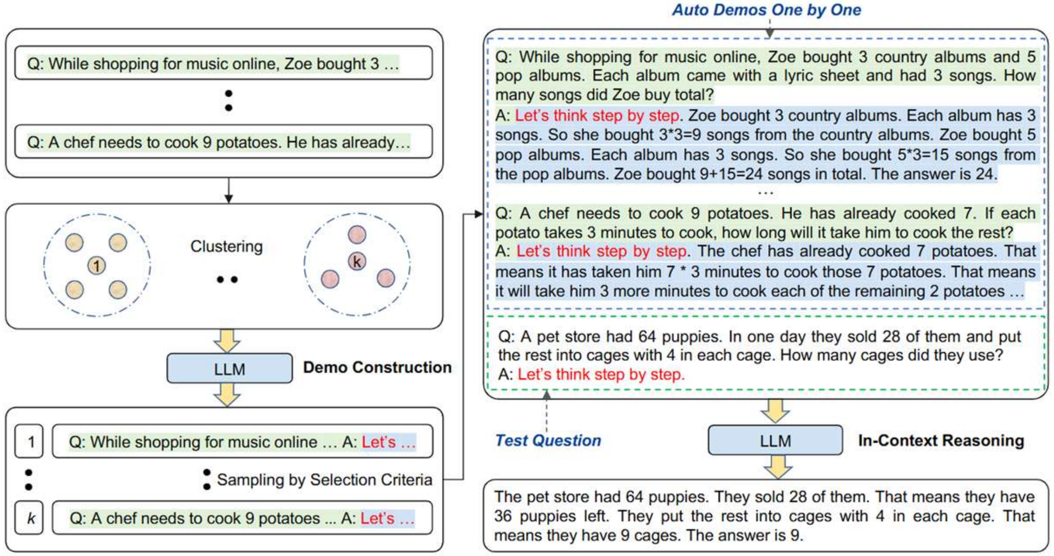
- High-quality demonstrations annotations yield high performance.
- High cost, difficult to transfer, and challenging demo selection.
- Few-shot CoT[1] , Few-shot PoT[2].

Automatic CoT Prompting Construction

- Low-quality demonstrations, low performance
- Low cost, easy to transfer.
- Zeroshot CoT[3], Plan-and-Solve Prompting[4].

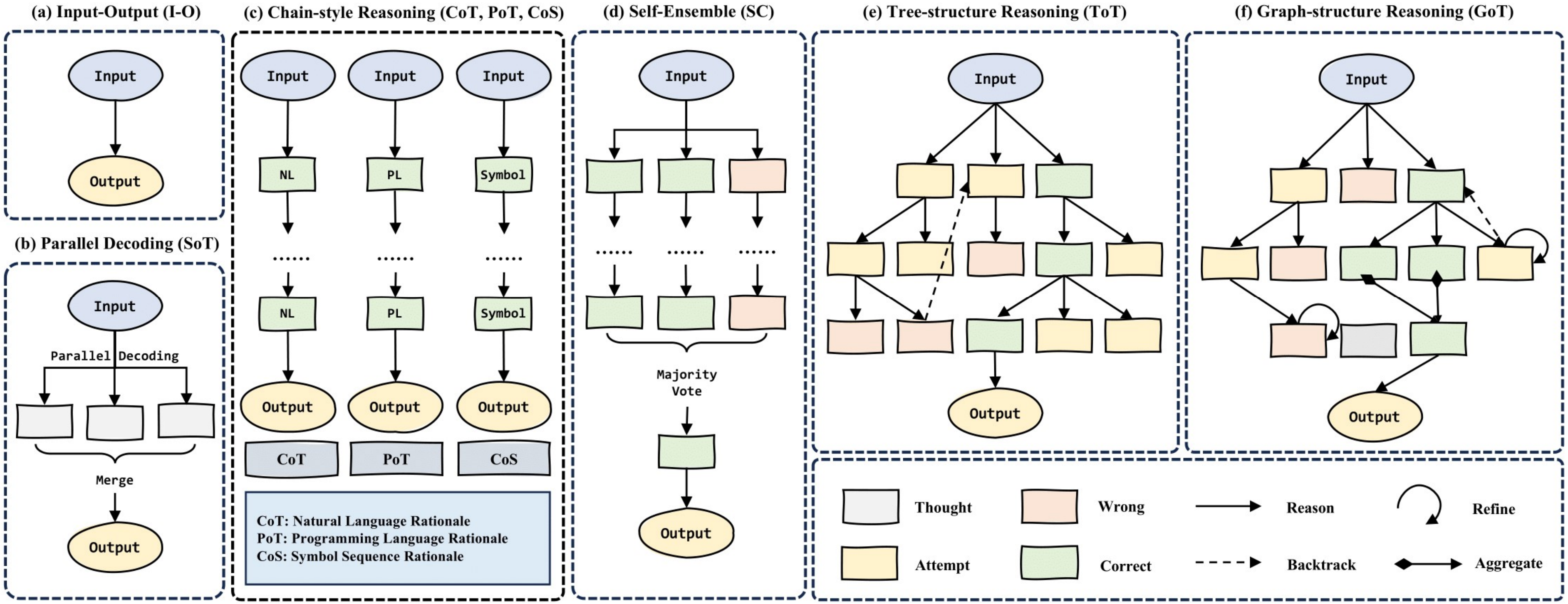
Semi-automatic CoT Prompting Construction

- Tradeoff between performance and cost.
- AutoCoT[5], AutoMateCoT[6].



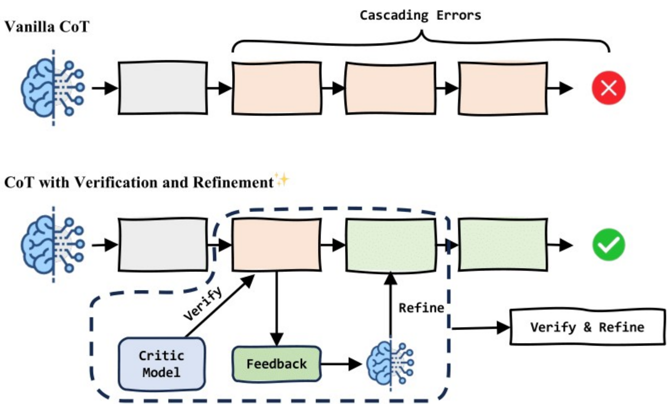
[1] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022
[2] Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks, TMLR 2023
[3] Large Language Models are Zero-Shot Reasoners, NeurIPS 2022
[4] Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models, ACL 2023
[5] Automatic Chain of Thought Prompting in Large Language Models, ICLR 2023
[6] Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data, Findings of EMNLP 2023

Chain-of-Thought Topological Variants



Skeleton-of-Thought: Prompting LLMs for Efficient Parallel Generation, ICLR 2024
Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks, TMLR 2023
Self-Consistency Improves Chain of Thought Reasoning in Language Models, ICLR 2023
Tree of Thoughts: Deliberate Problem Solving with Large Language Models, NeurIPS 2023
Graph of Thoughts: Solving Elaborate Problems with Large Language Models, AAAI 2024

CoT with Verification and Refinement



Feedback from LLM itself

- Have the model assess where it went wrong.
- Self-assessment/refinement may not be reliable.

Feedback from external environment

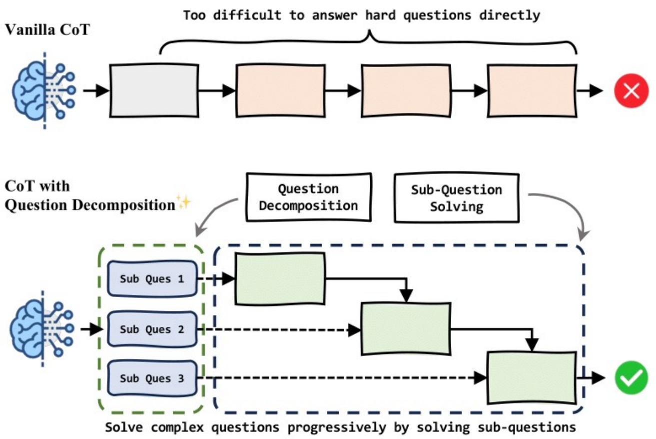
- Use external signals for evaluation, such as calculators, retrieval and program interpreters.
- External feedback is generally more reliable, but how do we tailor external feedback?

Logic-based Verification

- Verification based on logic, for example first-order logic and deductive reasoning

Self-Refine: Iterative Refinement with Self-Feedback, arxiv preprint (Self-feedback)
Large Language Models Cannot Self-Correct Reasoning Yet, ICLR 2024 (Self-feedback is not reliable)
CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing, ICLR 2024 (External feedback)
Large Language Models are Better Reasoners with Self-Verification, Findings of EMNLP 2023 (Logical verification)
Deductive Verification of Chain-of-Thought Reasoning, NeurIPS 2023 (Deductive Logic)

Question Decomposition



Linear Decomposition

- Two-stage and Iterative decomposition, more versatile.

Tree/Graph-structure Decomposition

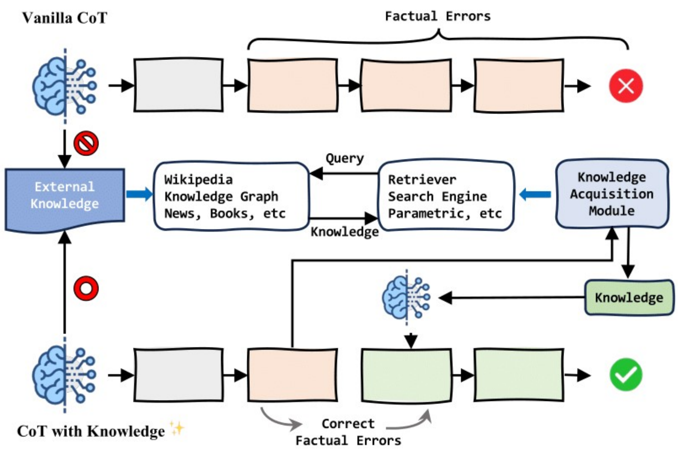
- More applicable to structured problems, such as multi-hop question answering.

Bottom-up Aggregation

- Instead of top-down question decomposition, it employs bottom-up sub-reasoning aggregation.

Least-to-Most Prompting Enables Complex Reasoning in Large Language Models, ICLR 2023 (Two-stage Decomposition)
Successive Prompting for Decomposing Complex Questions, EMNLP 2022 (Iterative Decomposition)
QDMR-based Planning-and-solving Prompting for Complex Reasoning Tasks, LREC-COLING 2024 (Tree Decomposition)
Cumulative Reasoning with Large Language Models, arxiv preprint (Bottom-up Aggregation)

Knowledge Enhancement



Internal Knowledge

- Prompt model to get its parameter knowledge.
- Parameter knowledge may erroneous or outdated.

External Knowledge

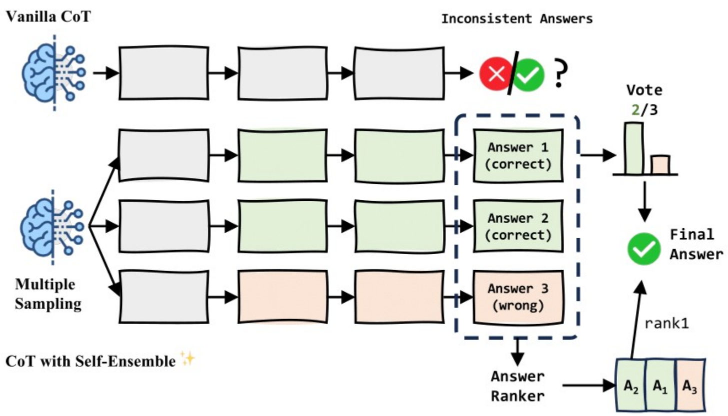
- Introduce retrieval-augmented reasoning.
- How to obtain accurate retrieval content, which is a research question studied by RAG.

Iterative knowledge acquisition

- Iteratively retrieval to acquire knowledge.
- More effective with multi-hop questions.

Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models, ICLR 2024 (Internal Knowledge)
Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources, ICLR 2024 (External Knowledge)
Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions, ACL 2023 (Iterative)

Self-Ensemble



Ranking

- Rank the multiple outputs.

Majority Voting

- Vote on multiple sampled outputs.

Reasoning Chains Ensemble

- Ensemble on multiple reasoning chains rather than solely voting on the final answers.

Multi-agent Debate

- Language models engage in role-playing debates until reaching a consensus answer.

Training Verifiers to Solve Math Word Problems, arxiv preprint (Ranking)
Self-Consistency Improves Chain of Thought Reasoning in Language Models, ICLR 2023 (Majority Voting)
Answering Questions by Meta-Reasoning over Multiple Chains of Thought, EMNLP 2023 (Reasoning Chains Ensemble)
Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages, EMNLP 2023 (Improve Diversity)
Learning to Break: Knowledge-Enhanced Reasoning in Multi-Agent Debate System, arxiv preprint (Multi-agent Debates)

Efficient CoT Reasoning

Parallel Problem Solving

- Parallel reasoning reduces time overhead.

Active Learning

- Reduce annotation costs by selecting demonstrations through active learning.

Adaptive Self-consistency

- Dynamically adjust the number of samples to reduce the overhead of ensemble reasoning.

Skeleton-of-thought: Large language models can do parallel decoding, ICLR 2024 (Parallel Problem Solving)
Active Prompting with Chain-of-Thought for Large Language Models, ACL 2024 (Active Learning)
Let's Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs, EMNLP 2023 (Adaptive Self-consistency)

Frontier Applications

Tool-assisted Reasoning and Tool Invocation

- Utilize external specialized tools to compensate for the model's shortcomings and endow model with the ability to interact with the environment.

Planning, Decision Making and LLM Agents

- LLM-powered agents interact with the external environment and make decisions based on goals and memory.

Chain-of-Thought Reasoning Capabilities Distillation

- Democratize complex reasoning capabilities into smaller language models for easier deployment on edge devices.

ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving, ICLR 2024 (Tool-assisted Reasoning)
Toolformer: Language Models Can Teach Themselves to Use Tools, NeurIPS 2023 (Tool Invocation)
Reflexion: language agents with verbal reinforcement learning. NeurIPS 2023 (Planning and Decision Making)
HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, NeurIPS 2023 (LLM Agents)
SCOTT: Self-Consistent Chain-of-Thought Distillation, ACL 2023 (SFT)
Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations, arxiv preprint (Preference Learning)

Future Directions

Multi-modal Chain-of-Thought Reasoning

- VQA -> Multi-step visual Reasoning
- Image -> Multi-image -> Video Reasoning

Faithful Chain-of-Thought Reasoning

- Identify and rectify mistakes in reasoning
- Interpretable and trustworthy reasoning

Mechanisms Exploration of Chain-of-Thought Reasoning

- Why does chain of thought reasoning work?
- Empirical perspective or theoretical perspective

Multimodal Chain-of-Thought Reasoning in Language Models, TMLR 2024
Measuring Faithfulness in Chain-of-Thought Reasoning, arxiv preprint
Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models, EMNLP 2023
Why think step by step? Reasoning emerges from the locality of experience, NeurIPS 2023
The Expressive Power of Transformers with Chain of Thought, ICLR 2024

Thanks for Your Attentions!