Can Language Models Perform ***Robust Reasoning*** in Chain-of-thought Prompting with ***Noisy Rationales***?

Zhanke Zhou

Hong Kong Baptist University

with Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han

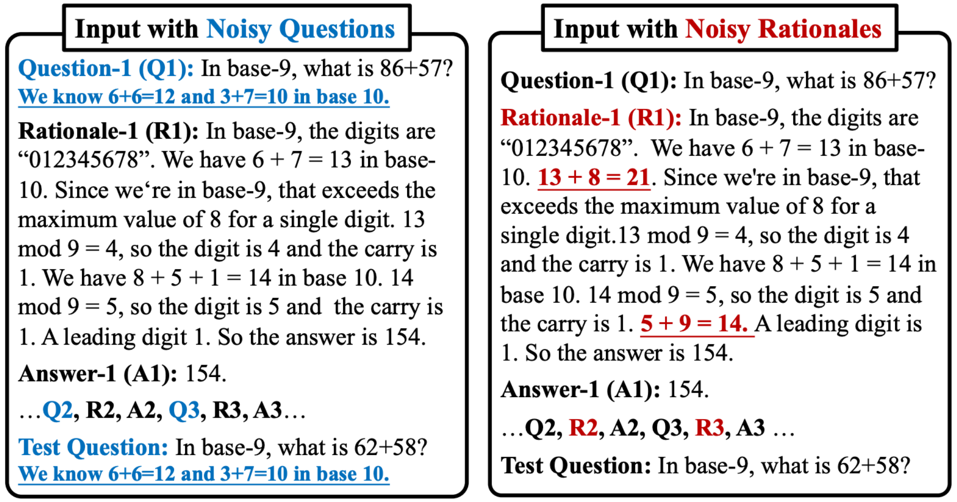
Homepage (with all slides and posters): https://andrewzhou924.github.io/ Email: cszkzhou@comp.hkbu.edu.hk

NeurIPS 2024 Paper: https://arxiv.org/pdf/2410.23856 Code: https://github.com/tmlr-group/NoisyRationales

Main contributions

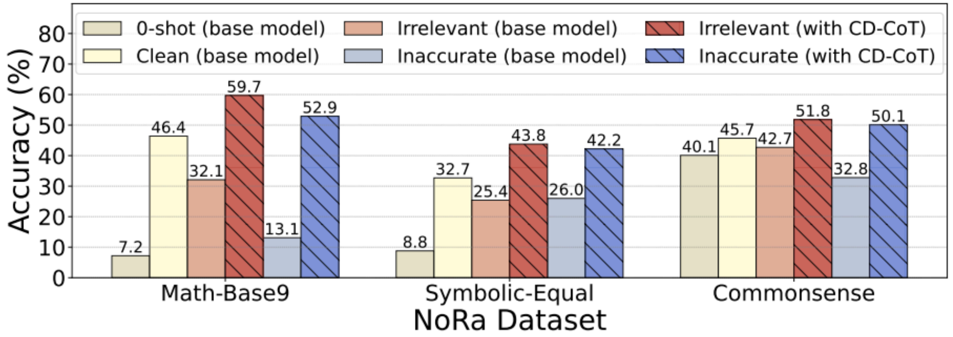
New research problem: Noisy Rationales

We investigate the problem of noisy rationales in the prevailing chain-of-thought prompting



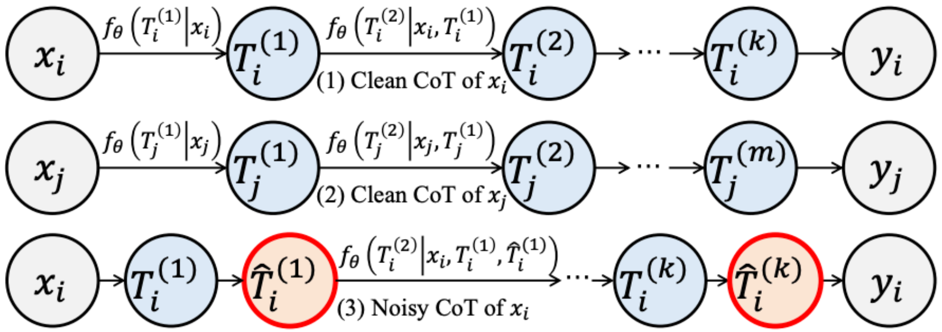
New benchmark: NoRa

We construct the NoRa dataset and systematically evaluate the robustness of LLMs



New algorithm: CD-CoT

We design a simple yet effective method to enhance robustness via contrastive denoising



Outline

- Background: language model reasoning
- **New research problem: Noisy Rationales**
- **New benchmark: NoRa**
- **New algorithm: CD-CoT**
- Take home messages
- Future directions

Background: language model reasoning

In-context learning (ICL) is commonly used in large language models (LLMs)

- enable LLMs to **learn from a few examples** without fine-tuning

Zero-shot Input

Question: In base-9, what is 62+58?

Input: ICL with three examples

Question-1: In base-9, what is 86+57? **Answer-1:** 154.
Question-2: In base-9, what is 63+34? **Answer-2:** 107.
Question-3: In base-9, what is 31+58? **Answer-3:** 100.
Question: In base-9, what is 62+58?

Background: language model reasoning

In-context learning (ICL) is commonly used in large language models (LLMs)

- enable LLMs to **learn from a few examples** without fine-tuning

Zero-shot Input

Question: In base-9, what is 62+58?

Input: ICL with three examples

Question-1: In base-9, what is 86+57? **Answer-1:** 154.
Question-2: In base-9, what is 63+34? **Answer-2:** 107.
Question-3: In base-9, what is 31+58? **Answer-3:** 100.
Question: In base-9, what is 62+58?

Prevailing in ICL, **Chain of thought (CoT) prompting** boost model reasoning

- CoT includes **rationales**, i.e., sequential reasoning thoughts to solve a question

Input: ICL with three examples

Question-1: In base-9, what is 86+57? **Answer-1:** 154.
Question-2: In base-9, what is 63+34? **Answer-2:** 107.
Question-3: In base-9, what is 31+58? **Answer-3:** 100.
Question: In base-9, what is 62+58?

Input: CoT with rationales

Question-1: In base-9, what is 86+57?
Rationale-1: In base-9, the digits are “012345678”. We have 6 + 7 = 13 in base-10. Since we‘re in base-9, that exceeds the maximum value of 8 for a single digit. 13 mod 9 = 4, so the digit is 4 and the carry is 1. We have 8 + 5 + 1 = 14 in base 10. 14 mod 9 = 5, so the digit is 5 and the carry is 1. A leading digit 1. So the answer is 154.
Answer-1: 154.
…**Q2, R2, A2, Q3, R3, A3** …
Question : In base-9, what is 62+58?

New research problem: Noisy Rationales

Existing work generally assume that CoT contains clean rationales

But, what if CoT contains **noisy rationales**? 🤔

- **noisy rationales include irrelevant or inaccurate thoughts**

Input: CoT with clean rationales

Question-1: In base-9, what is 86+57?

Rationale-1: In base-9, the digits are "012345678". We have 6 + 7 = 13 in base-10. Since we're in base-9, that exceeds the maximum value of 8 for a single digit. 13 mod 9 = 4, so the digit is 4 and the carry is 1. We have 8 + 5 + 1 = 14 in base 10. 14 mod 9 = 5, so the digit is 5 and the carry is 1. A leading digit 1. So the answer is 154.

Answer-1: 154.

…**Q2, R2, A2, Q3, R3, A3** …

Question : In base-9, what is 62+58?

*the irrelevant **base-10 information** is included in rationale*

Input: CoT with noisy rationales

Question-1 (Q1): In base-9, what is 86+57?

Rationale-1 (R1): In base-9, the digits are "012345678". We have 6 + 7 = 13 in base-10. **13 + 8 = 21**. Since we're in base-9, that exceeds the maximum value of 8 for a single digit.13 mod 9 = 4, so the digit is 4 and the carry is 1. We have 8 + 5 + 1 = 14 in base 10. 14 mod 9 = 5, so the digit is 5 and the carry is 1. **5 + 9 = 14.** A leading digit is 1. So the answer is 154.

Answer-1 (A1): 154.

…**Q2, R2, A2, Q3, R3, A3** …

Test Question: In base-9, what is 62+58?

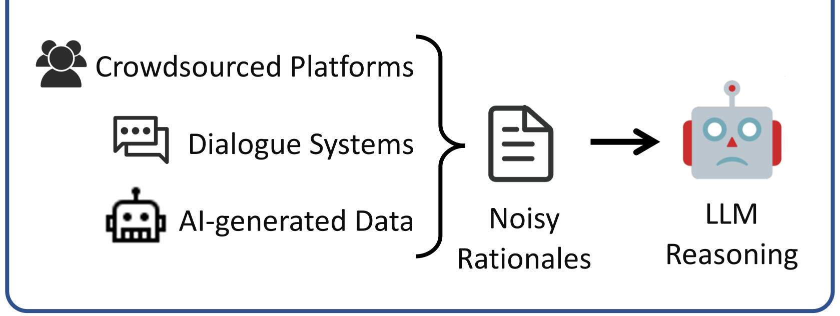
*while the test question asks about **base-9 calculation***

New research problem: Noisy Rationales

Noisy rationales originate from diverse sources (see Appendix C for details)

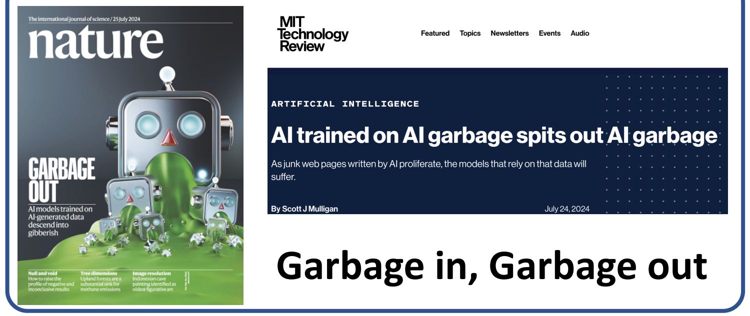
- such as crowdsourced platforms, dialogue systems, and AI-generated data

Sources of noisy rationales



Recent news on noisy data

Recent investigations on noisy data



However, the robustness of LLMs against noisy rationales is still unknown

- **a new dataset** is needed to conduct **a systematic evaluation** of current LLMs
- and verify the corresponding **countermeasures** against noisy rationales

Outline

- Background: language model reasoning
- New research problem: Noisy Rationales
- **New benchmark: NoRa**
    - **Benchmark construction**
    - Empirical evaluations on NoRa
- New algorithm: CD-CoT
- Take home messages
- Future directions

New benchmark: NoRa

NoRa (Noisy Rationales)

- a comprehensive testbed to evaluate the **robustness** against noisy rationales
- contains **26391** questions and **5** subtasks
- covering **3** types of reasoning tasks: **mathematical, symbolic, and commonsense**

Task	Irrelevant Thoughts	Inaccurate Thoughts
NoRa-Math	In base-9, digits run from 0 to 8. We have 3 + 2 = 5 in base-10. Since we're in base-9, that doesn't exceed the maximum value of 8 for a single digit. 5 mod 9 = 5, so the digit is 5 and the carry is 0. There are five oceans on Earth: the Atlantic, Pacific, Indian, Arctic, and Southern. We have 8 + 6 + 0 = 14 in base 10. 14 mod 9 = 5, so the digit is 5 and the carry is 1. A leading digit 1. So the answer is 155. Answer: 155	In base-9, digits run from 0 to 8. We have 3 + 2 = 5 in base-10. 5 + 4 = 9. Since we're in base-9, that doesn't exceed the maximum value of 8 for a single digit. 5 mod 9 = 5, so the digit is 5 and the carry is 0. 5 + 9 = 14. We have 8 + 6 + 0 = 14 in base 10. 14 mod 9 = 5, so the digit is 5 and the carry is 1. A leading digit 1. So the answer is 155. Answer: 155
NoRa-Symbolic	... "turn around right" means the agent needs to turn right, and repeat this action sequence four times to complete a 360-degree loop. Many GPS navigation systems will issue a 'turn around' command if the driver deviates from the planned route. So, in action sequence is I_TURN_RIGHT I_TURN_RIGHT I_TURN_RIGHT I_TURN_RIGHT. ...	... "turn around right" means the agent needs to turn right, and repeat this action sequence four times to complete a 360-degree loop. Turn opposite is I_TURN_RIGHT I_TURN_LEFT. So, in action sequence is I_TURN_RIGHT I_TURN_RIGHT I_TURN_RIGHT I_TURN_RIGHT. ...
NoRa-Com.	The relations path are son, sister, uncle, which means Francisco is David's son's sister's uncle. For son's sister, we have son's sister is daughter. So the relations path are reduced to daughter, uncle. In genetics, mitochondrial DNA is always inherited from the mother, making the mother-daughter genetic link unique. For daughter's uncle, we have daughter's uncle is brother. So the relations path are reduced to brother. Therefore, the answer is brother. Answer:brother	The relations path are son, sister, uncle, which means Francisco is David's son's sister's uncle. For son's sister, we have son's sister is daughter. So the relations path are reduced to daughter, uncle. For daughter's uncle, we have daughter's uncle is brother. We have brother' sister is brother. So the relations path are reduced to brother. Therefore, the answer is brother. Answer:brother

Table 1: Noisy rationales (consisting noisy thoughts) sampled from the NoRa dataset. Full examples of NoRa are in Appendix C.6, and real-world examples of noisy rationales are in Appendix C.3.

New benchmark: NoRa

Difficulty	Noise Ratio	#total thoughts (#noisy thoughts) of prompting rationales (Avg.) Math Base-9	Math Base-11	Sym. Equal	Sym. Longer	Com.
Easy	0.3	10 (2)	10 (2)	11.5 (2.7)	11.0 (2.5)	7 (2)
Medium	0.5	12 (4)	12 (4)	13.3 (4.5)	12.7 (4.2)	8 (3)
Hard	0.8	14 (6)	14 (6)	16.0 (7.1)	15.2 (6.8)	9 (4)
#questions		4024	9269	4182	3920	4996

Table 2: Statistics of NoRa dataset.

Definitions

- **Irrelevant thoughts** are irrelevant to the given context
  - e.g., discussing the genetic overlap of siblings when the task is to deduce family roles
- **Inaccurate thoughts** are factual errors in the given context
  - e.g., "5+5=10" is wrong in base-9 calculation

Benchmark construction

- generating noisy rationales by **inserting irrelevant or inaccurate thoughts**
- **guarantee the overall correctness** without modifying the question or answer
- control the reasoning difficulty through **different noise ratios (0.3, 0.5, 0.8)**

Outline

- Background: language model reasoning
- New research problem: Noisy Rationales
- **New benchmark: NoRa**
  - Benchmark construction
  - **Empirical evaluations on NoRa**
- New algorithm: CD-CoT
- Take home messages
- Future directions

Empirical evaluations on NoRa

Grand observation: The base LLM (GPT-3.5) with all the existing methods is severely affected by noisy rationales

- a **0.2%-25.3%** decrease with irrelevant noise
- a **0.1%-54.0%** decrease with inaccurate noise (compared with clean rationales)

Observation 1: self-correction methods perform poorly on most tasks with noisy rationales

Observation 2: self-consistency methods can improve robustness without true denoising

Task	Method $\mathcal{M}$	$\text{Acc}(\mathcal{M}, \mathcal{Q}, \mathcal{P}_{\text{clean}})$	$\text{Acc}(\mathcal{M}, \mathcal{Q}, \mathcal{P}_{\text{irrelevant}})$ Easy	Medium	Hard	Avg.	$\text{Acc}(\mathcal{M}, \mathcal{Q}, \mathcal{P}_{\text{inaccurate}})$ Easy	Medium	Hard	Avg.
Math Base-9	Base	46.4	39.3	30.3	26.6	32.1	23.2	10.1	6.0	13.1
	w/ ISC [29]	24.3	17.7	14.7	12.7	15.0	18.4	13.7	12.3	14.8
	w/ SP [89]	26.2	25.5	25.5	21.9	24.3	20.0	18.4	**14.3**	17.6
	w/ SM [62]	37.4	30.0	22.7	16.5	23.1	24.7	**19.2**	12.4	**18.8**
	w/ SD [102]	47.9	37.2	25.4	24.7	29.1	29.3	12.5	8.7	16.8
	w/ SC [83]	**61.5**	**51.1**	**39.0**	**36.2**	**42.1**	**32.7**	15.3	7.5	18.5
Math Base-11	Base	23.9	19.1	13.6	10.7	14.5	14.0	6.7	3.6	8.1
	w/ ISC [29]	11.2	8.3	7.8	6.0	7.4	6.5	5.2	4.7	5.5
	w/ SP [89]	20.7	17.5	**16.7**	14.0	16.0	14.1	**10.7**	**10.8**	**11.9**
	w/ SM [62]	16.3	12.0	6.0	5.7	7.9	12.0	9.3	7.7	9.7
	w/ SD [102]	17.9	12.3	12.0	13.3	12.5	17.0	8.7	5.3	10.3
	w/ SC [83]	**33.7**	**25.3**	16.3	**15.0**	**18.9**	**19.7**	9.3	3.3	10.8
Symbolic Equal	Base	32.7	28.1	25.1	23.0	25.4	29.1	26.1	22.7	26.0
	w/ ISC [29]	23.9	20.0	16.3	15.5	17.3	19.2	18.3	18.1	18.5
	w/ SP [89]	23.2	23.0	22.6	22.7	22.8	23.7	22.5	23.5	23.2
	w/ SM [62]	25.0	20.7	19.7	16.7	19.0	21.0	20.3	20.0	20.4
	w/ SD [102]	9.9	10.1	10.9	10.3	10.4	10.1	10.9	10.4	10.5
	w/ SC [83]	**35.3**	**31.0**	**28.3**	**27.0**	**28.8**	**33.3**	**30.7**	**26.0**	**30.0**
Symbolic Longer	Base	9.2	6.3	7.2	6.0	6.5	7.0	6.8	6.0	6.6
	w/ ISC [29]	4.9	4.6	2.7	3.7	3.7	3.4	4.3	3.3	3.7
	w/ SP [89]	5.1	4.3	4.1	3.9	4.1	4.9	4.0	4.5	4.5
	w/ SM [62]	1.7	0.7	0.7	1.3	1.0	1.3	0.7	0.3	0.8
	w/ SD [102]	0.1	0.1	0.1	0.2	0.1	0.1	0.3	0.0	0.1
	w/ SC [83]	**13.0**	**7.7**	**9.0**	**6.3**	**7.7**	**8.0**	**8.0**	**8.7**	**8.2**
Commonsense	Base	45.7	44.3	42.3	41.4	42.7	36.7	33.4	28.3	32.8
	w/ ISC [29]	21.8	24.3	22.5	21.4	22.7	23.3	26.5	24.0	24.6
	w/ SP [89]	47.9	48.2	46.7	48.1	47.7	49.6	46.6	46.5	47.6
	w/ SM [62]	53.3	50.3	50.0	46.7	49.0	47.7	49.0	49.3	48.7
	w/ SD [102]	**54.0**	**58.3**	**57.3**	**57.7**	**57.8**	**57.0**	**58.3**	**53.7**	**56.3**
	w/ SC [83]	52.0	46.3	45.0	44.7	45.3	44.7	44.7	38.0	42.5

Table 3: Reasoning accuracy on NoRa dataset with 3-shot prompting examples with clean, irrelevant, or inaccurate rationales. The **boldface** numbers mean the best results, while the underlines numbers indicate the second-best results. Note the referenced results of Base model are highlighted in gray.

Empirical evaluations on NoRa

Task	Setting	Temperature 0	0.3	0.5	0.7	1
Base-9	clean	**61.0**	60.9	57.5	55.3	46.4
	ina. easy	**29.7**	28.0	27.2	26.6	21.7
	ina. hard	5.0	5.1	**5.5**	4.6	5.0
Base-11	clean	**34.0**	33.8	31.6	29.8	23.9
	irr. easy	21.7	23.1	21.3	**23.3**	19.1
	irr. hard	17.0	**17.5**	15.5	14.1	10.7
Sym.(E)	clean	34.2	**35.8**	35.7	34.6	32.7
	irr. easy	28.6	**31.5**	29.8	29.1	28.1
	irr. hard	**27.0**	26.1	26.2	24.0	23.0
Sym.(L)	clean	6.3	8.3	8.9	8.9	**9.3**
	ina. easy	5.0	7.3	**8.6**	8.3	7.0
	ina. hard	4.0	6.1	**6.3**	6.2	6.0

Table 4: Comparing performances of the base model with different temperatures. Sym.(E)/(L) are symbolic tasks.

Task	Setting	#Prompting Examples 1	2	3	4	5
Base-9	clean	24.8	38.3	46.4	**50.8**	50.5
	ina.-easy	17.5	22.2	23.2	25.4	**25.6**
	ina.-hard	**11.3**	6.3	6.0	5.7	5.7
Base-11	clean	11.8	20.4	23.9	29.9	**32.1**
	irr. easy	8.9	15.9	19.1	21.7	**26.3**
	irr. hard	7.7	10.0	10.7	15.2	**16.1**
Sym.(E)	clean	18.0	26.5	32.7	**39.8**	—
	ina.-easy	17.3	23.6	29.1	**34.7**	—
	ina.-hard	15.0	21.0	**22.7**	—	—
Sym.(L)	clean	2.7	7.7	9.3	11.3	**12.2**
	irr. easy	2.3	5.4	7.0	8.8	**8.9**
	irr. hard	1.9	4.0	6.0	**6.3**	—

Table 5: Comparing performances of the base model with a varying number of examples ("—" denotes over token limit).

Model	Task	Setting 0-shot	clean	irr.	ina.
GPT3.5	Base-9	7.2	**46.4**	30.3	10.1
	Sym.(E)	8.8	**32.7**	25.1	26.1
	Com.	40.0	**45.7**	42.3	33.4
Gemini	Base-9	12.7	**88.0**	72.3	21.2
	Sym.(E)	9.3	**44.5**	38.9	36.7
	Com.	42.9	**55.6**	53.2	33.5
Llama2	Base-9	1.7	**4.9**	2.9	2.7
	Sym.(E)	4.7	**10.1**	8.7	9.1
	Com.	35.0	**42.3**	41.9	40.2
Mixtral	Base-9	3.9	**27.5**	16.3	3.7
	Sym.(E)	8.3	**19.3**	17.9	15.1
	Com.	24.2	**37.5**	34.9	31.1

Table 6: Comparing LLMs with 0-shot, 3-shot clean, and 3-shot medium irrelevant (irr.) / inaccurate (ina.) rationales.

Observation 3:
Adjusting model temperature can improve reasoning under noisy rationales

Observation 4:
Prompting with more noisy examples boosts reasoning accuracy on most tasks

Observation 5:
Different LLMs are generally vulnerable to noisy rationales

Empirical evaluations on NoRa

We further explore the mapping among questions, rationales, and answers

Specifically, given the 3-shot examples $\{(x_1, \mathcal{T}_1, y_1), (x_2, \mathcal{T}_2, y_2), (x_3, \mathcal{T}_3, y_3)\}$, we test three configurations:

- shuffle questions $\{(x_1, \mathcal{T}_3, y_3), (x_2, \mathcal{T}_1, y_1), (x_3, \mathcal{T}_2, y_2)\}$
- shuffle rationales $\{(x_1, \mathcal{T}_3, y_1), (x_2, \mathcal{T}_1, y_2), (x_3, \mathcal{T}_2, y_3)\}$
- shuffle answers $\{(x_1, \mathcal{T}_1, y_3), (x_2, \mathcal{T}_2, y_1), (x_3, \mathcal{T}_3, y_2)\}$

Task	Zero-shot	Few-shot (No Shuffle)	Shuffle Questions x_i	Shuffle Rationales $\mathcal{T}_i$	Shuffle Answers y_i
Math Base-9	7.2	**46.4**	45.5 (0.9%↓)	34.5 (11.9%↓)	35.7 (10.7%↓)
Math Base-11	5.5	23.9	**24.8** (0.9%↑)	21.6 (2.3%↓)	21.1 (11.7%↓)
Symbolic Equal	8.8	32.7	32.7 (0.0%↓)	**32.8** (0.1%↑)	32.3 (0.4%↓)
Symbolic Longer	0.0	**9.2**	7.0 (2.2%↓)	6.2 (3.0%↓)	6.3 (2.9%↓)
Commonsense	40.0	**45.7**	38.7 (7.0%↓)	39.7 (6.0%↓)	39.8 (5.9%↓)

Table 7: Performance (in accuracy%) on NoRa dataset under different few-shot shuffle configurations.

Observation 6: Shuffling the mappings of prompting examples degenerates the reasoning but still performs better than without prompting.
Besides, LLMs are less vulnerable to shuffled mappings than noisy rationales.

Outline

- Background: language model reasoning
- New research problem: Noisy Rationales
- New benchmark: NoRa
- **New algorithm: CD-CoT**
  - **Motivation and design of CD-CoT**
  - Empirical evaluations of CD-CoT
- Take home messages
- Future directions

New algorithm: CD-CoT

Motivation

- Current LLMs **cannot** denoise well with their **intrinsic denoising ability**
  - even enhanced with self-correction / self-consistency methods

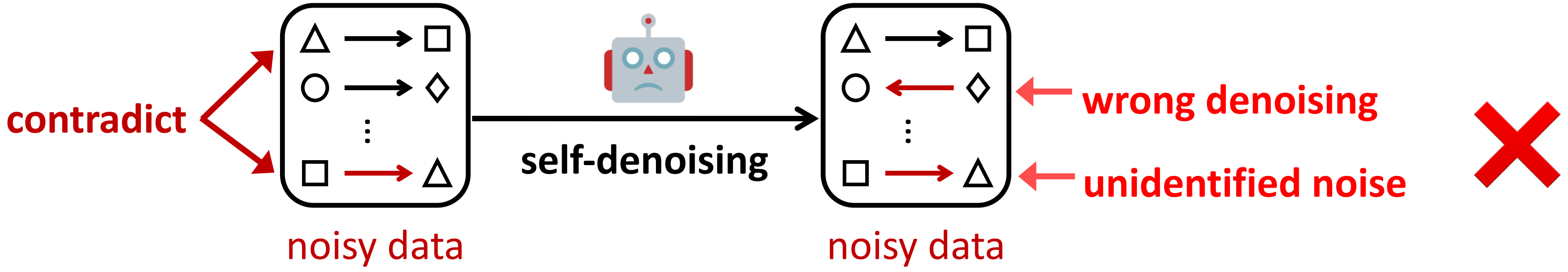
- **External supervision** is necessary for enhancement
  - which should be sufficient for denoising and accessible in practice

- **A clean CoT demonstration** can be the minimal requirement
  - for denoising-purpose prompting
  - which is much more practical than existing methods requiring external supervision

New algorithm: CD-CoT

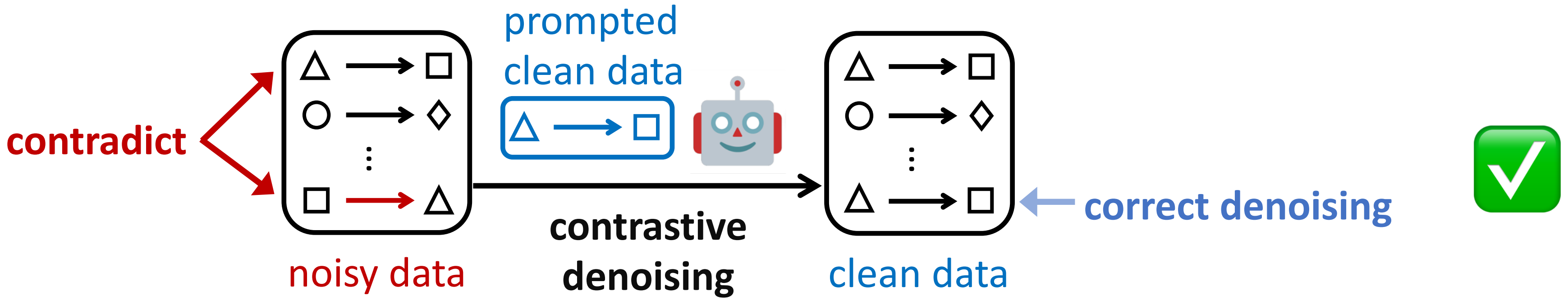
Self-denoising:

- It is **hard** for LLMs to denoise noisy data **without guidance**



Contrastive denoising:

- It is **easier** for LLMs to denoise **by contrasting noisy and clean data**



New algorithm: CD-CoT

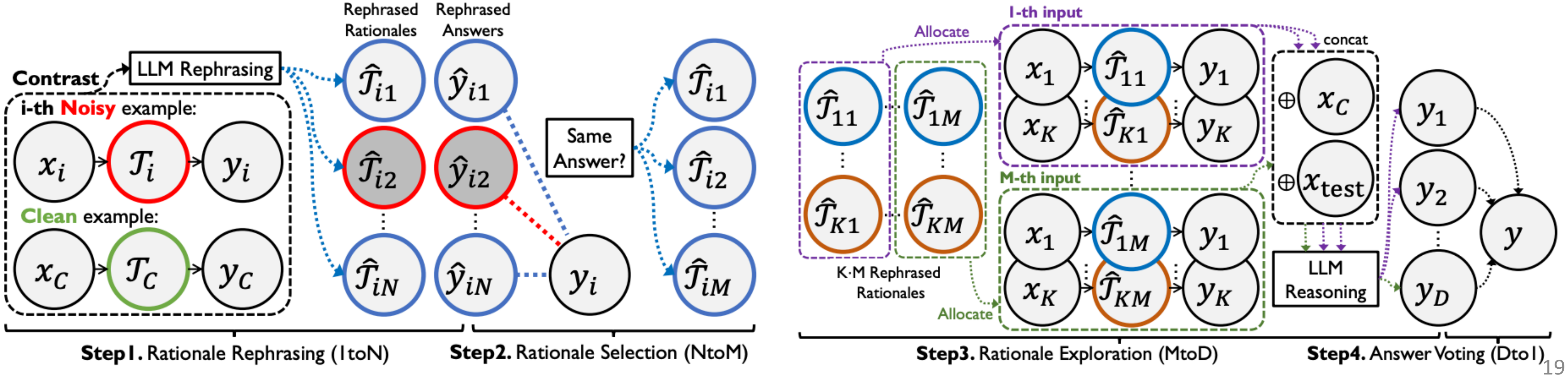
Contrastive Denoising with Noisy Chain-of-thought (CD-CoT)

- assume that LLMs can identify noisy thoughts
  - by contrasting a pair of noisy and clean rationales (similar to contrastive learning)

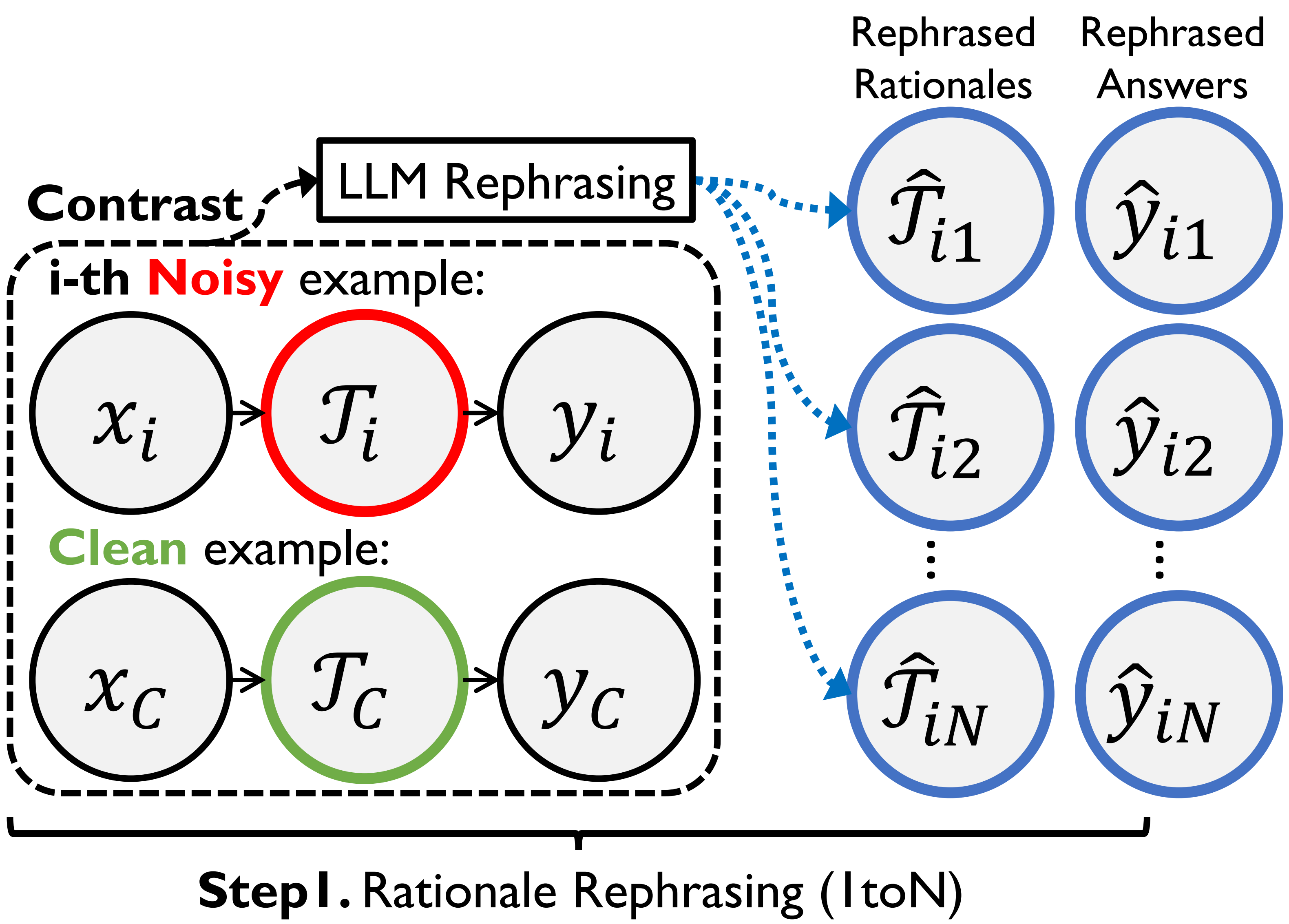
New algorithm: CD-CoT

Contrastive Denoising with Noisy Chain-of-thought (CD-CoT)

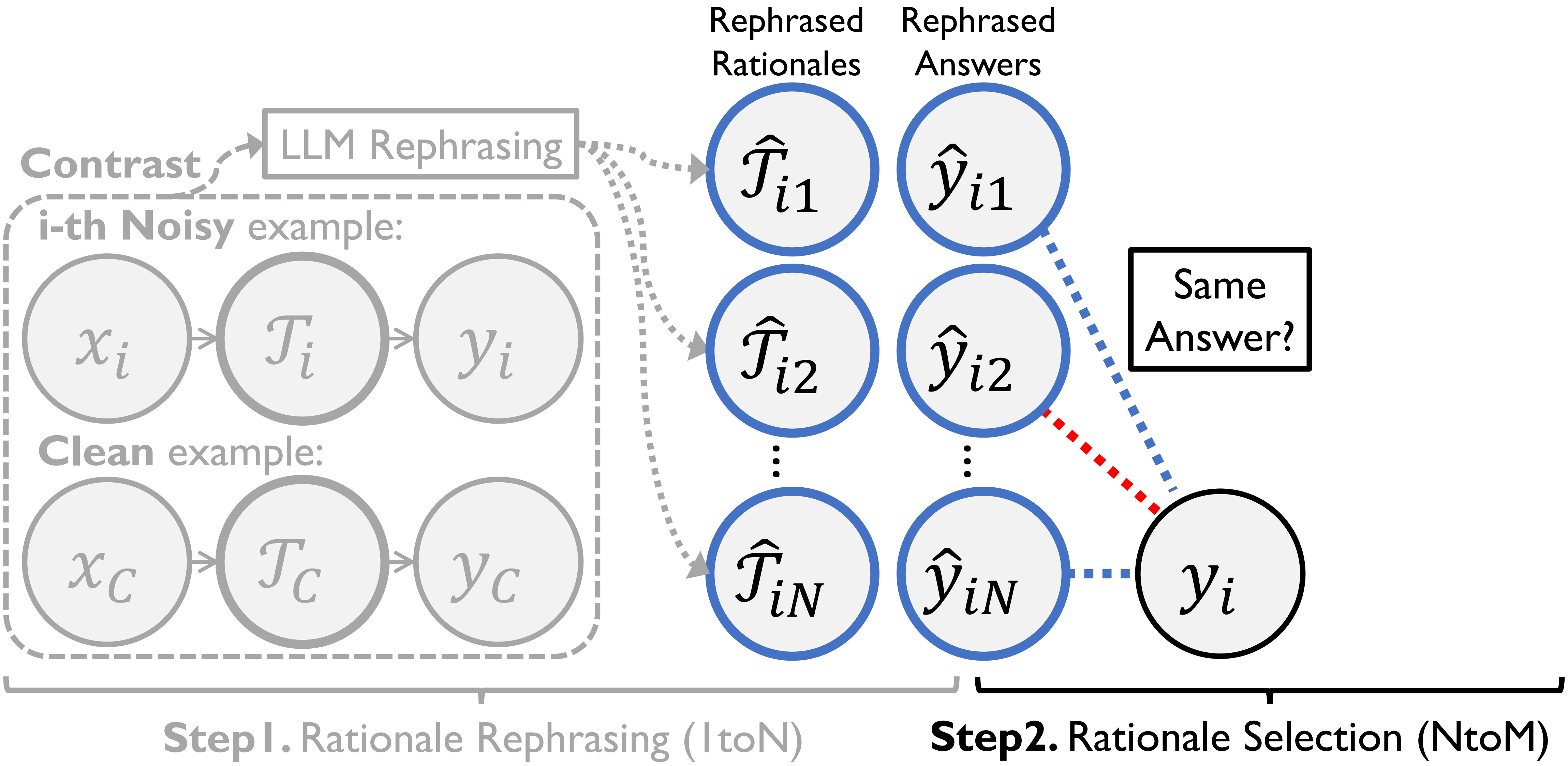
- assume that LLMs can identify noisy thoughts
  - by contrasting a pair of noisy and clean rationales (similar to contrastive learning)
- **design principle: exploration and exploitation**
  - **rephrasing and selecting rationales** in the input space to conduct explicit denoising (steps 1&2)
  - **exploring diverse reasoning paths and voting on answers** in the output space (steps 3&4)



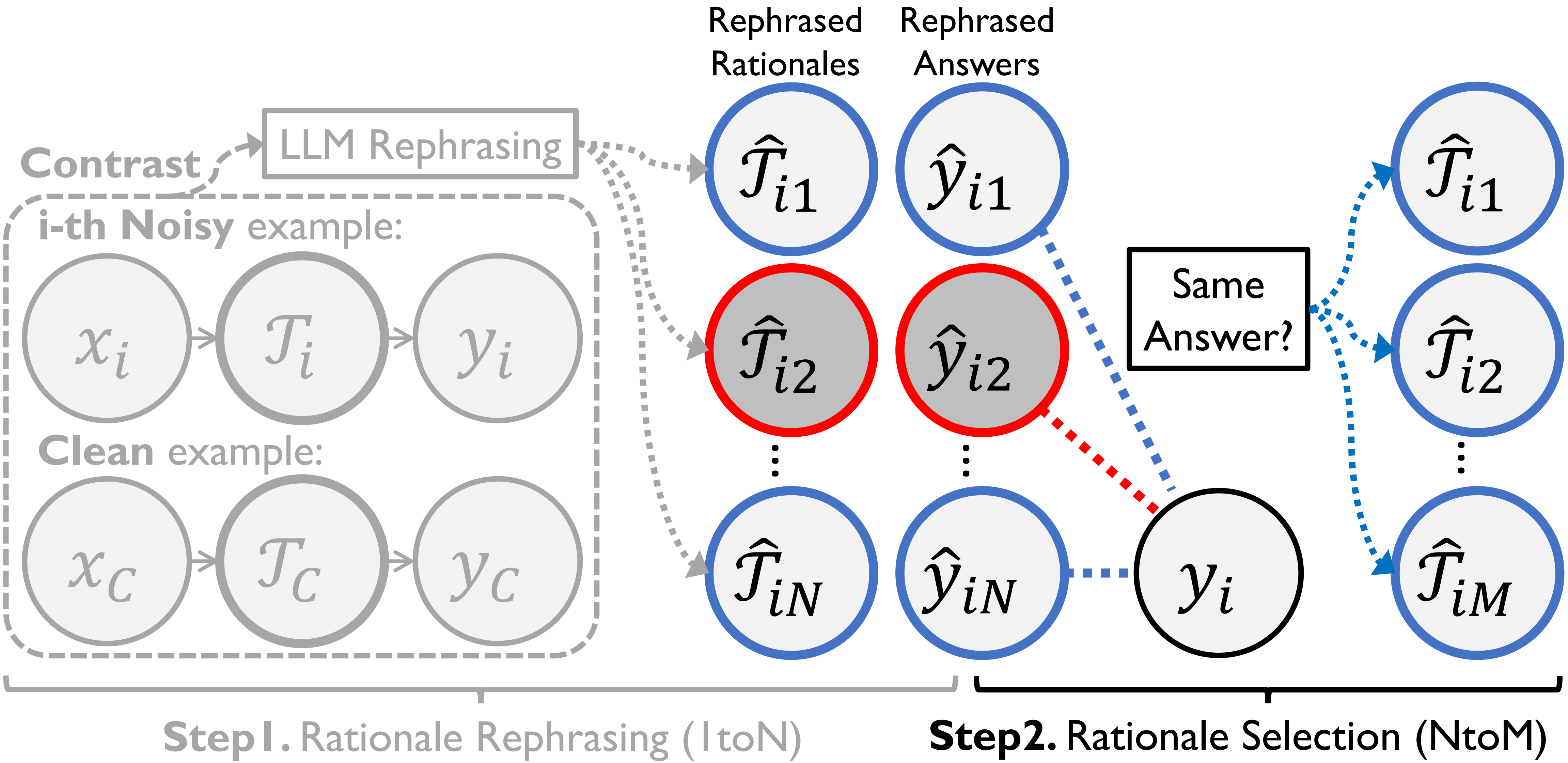
- **Step-1:** rephrase the noisy rationales via contrastive denoising
- Step-2: select rephrased examples with the same answers (unchanged)



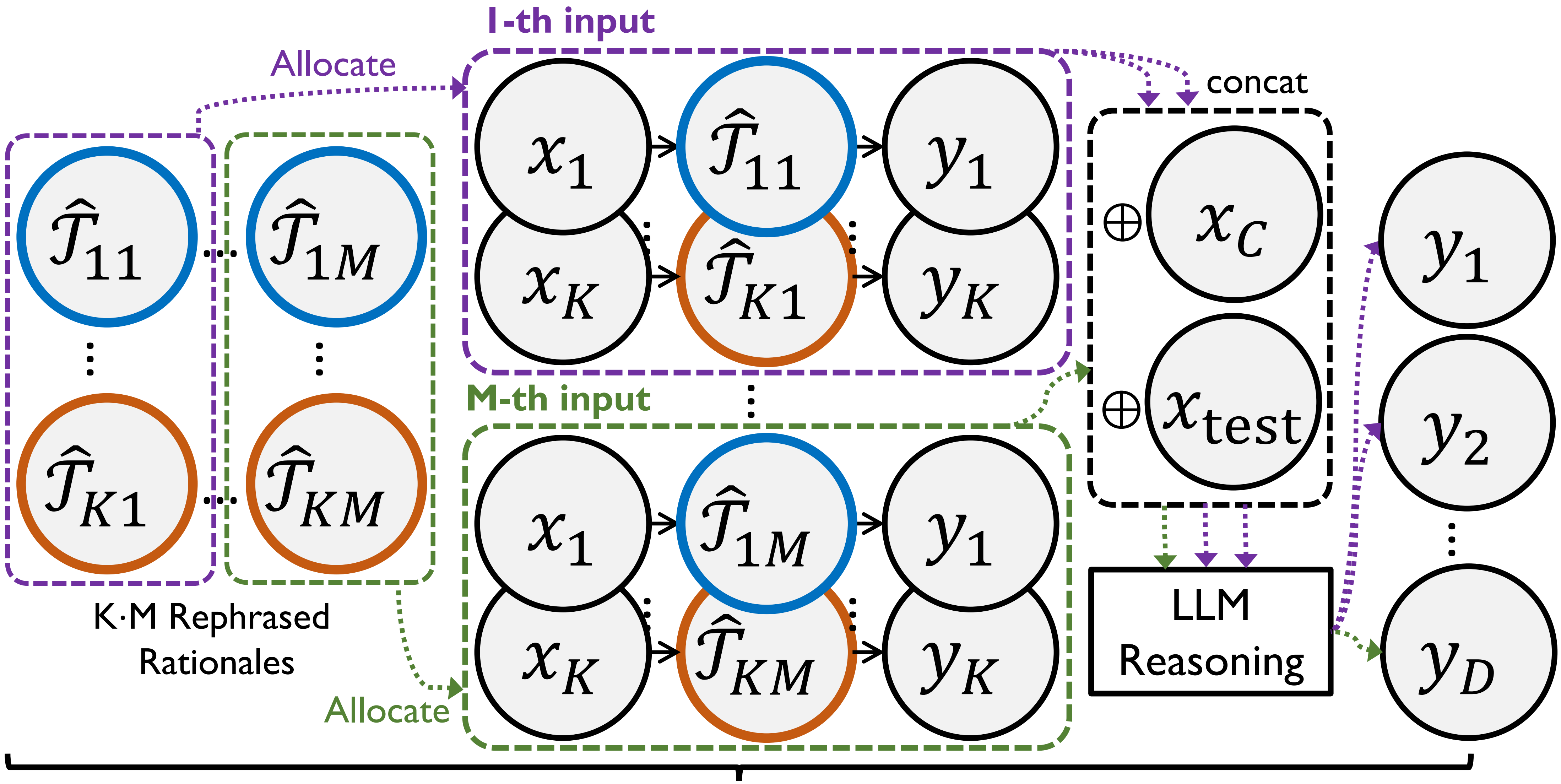
- Step-1: rephrase the noisy rationales via contrastive denoising
- **Step-2:** select rephrased examples with the same answers (unchanged)



- Step-1: rephrase the noisy rationales via contrastive denoising
- **Step-2:** select rephrased examples with the same answers (unchanged)

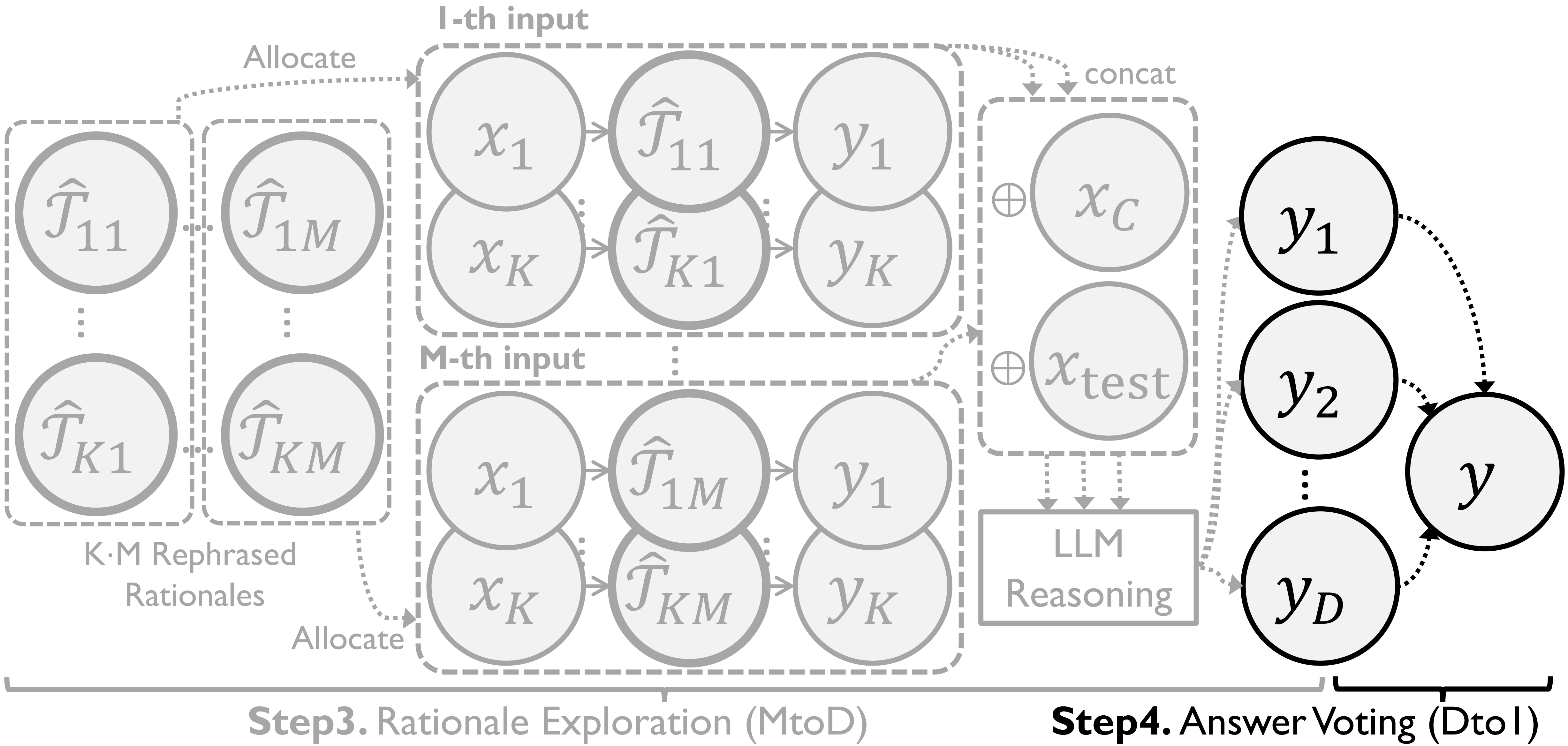


- **Step-3:** fully utilize the rephrased examples for deliberate reasoning
- Step-4: vote all the answers equally to get the final answer

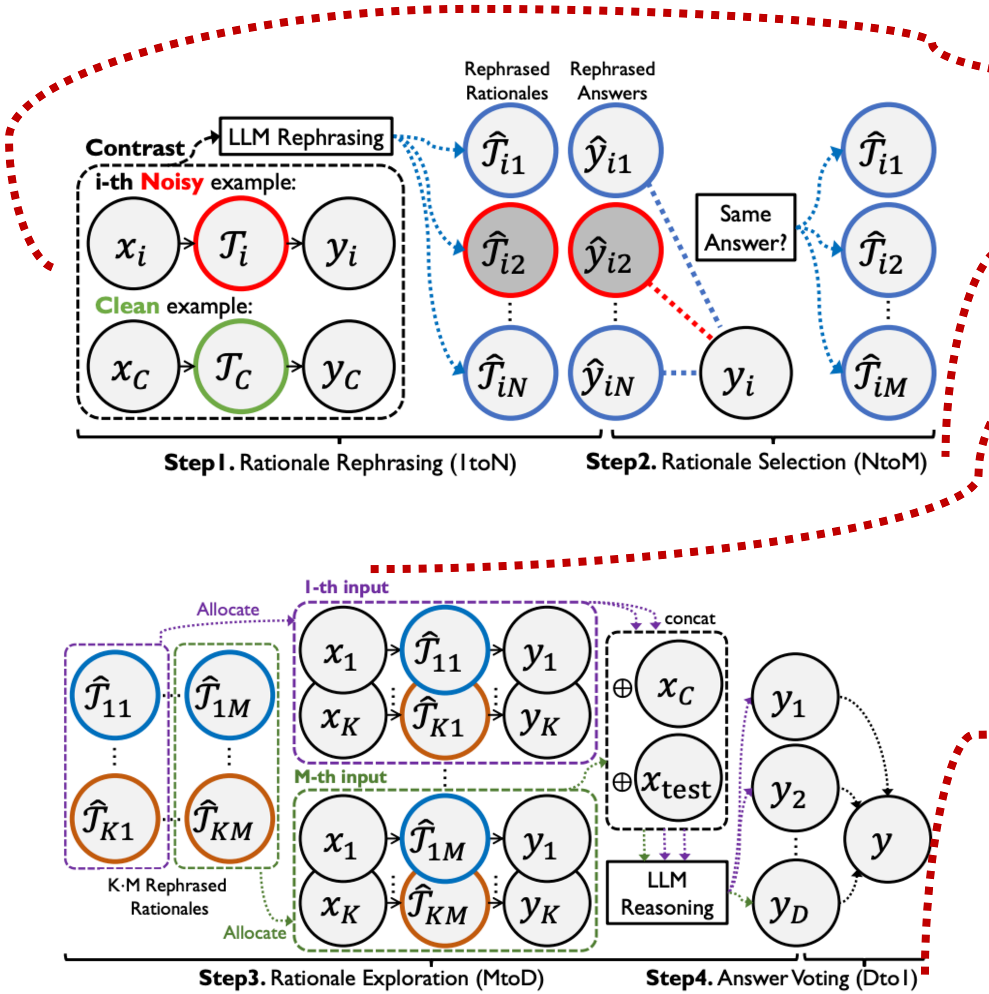


Step3. Rationale Exploration (MtoD)

- Step-3: fully utilize the rephrased examples for deliberate reasoning
- **Step-4:** vote all the answers equally to get the final answer



New algorithm



Algorithm 1 CD-CoT: Contrastive Denoising with Noisy Chain-of-Thought.

Require: an LLM f_θ, the prompt of contrastive denoising $\mathcal{P}_{\text{denoise}}$, one test question x_{test}, one clean example $(x_C, \mathcal{T}_C, y_C)$, K prompting examples $S_n = \{(x_i, \mathcal{T}_i, y_i)\}_{i=1}^{K}$, hyper-parameters N, M, and reasoning budget $\{B_i\}_{i=1}^{M}$ (satisfies that $\Sigma_{i=1}^{M} B_i = D$, where D is the total budget).

```
for i = 1 ... K do
    initialize the set of rephrased results of i-th example R_i ← ∅.
    for j = 1 ... N do
        # Step-1: Rationale Rephrasing via Supervised Contrasting
        obtain a rephrased example as (x_i, T̂_i, ŷ_i) ← f_θ(P_denoise(x_C, T_C, y_C, x_i, T_i, y_i)).
        if match answer ŷ_i = y_i, then store the rephrased example as R_i ← R_i ∪ {(x_i, T̂_i, ŷ_i)}.
    end for
    # Step-2: Rationale Selection
    randomly select M rephrased examples from R_i and obtain R̃_i = {(x_is, T̂_is, ŷ_is)}_{s=1}^M.
end for
# Step-3: Rationale Exploration
initialize the set of answers Y ← ∅.
for i = 1 ... M do
    construct an input P_i ← {(x_ji, T̂_ji, ŷ_ji)}_{j=1}^K, where (x_ji, T̂_ji, ŷ_ji) is the i-th element of R̂_j.
    concatenate P_i with the clean example and test question as P_i ← P_i ∪ {(x_C, T_C, y_C), x_test}.
    for j = 1 ... B_M do
        get one answer by LLM reasoning as y_j ← f_θ(P_i).
        store the answer as Y ← Y ∪ {y_j}.
    end for
end for
# Step-4: Answer Voting
initialize the dictionary of answer count C that ∀y_j ∈ Y, C[y_j] = 0.
for j = 1 ... D do
    update C[y_j] ← (C[y_j] + 1).
end for
get the final answer y with maximum counts as y ← arg max_y C[y].
return the answer y.
```

Outline

- Background: language model reasoning
- New research problem: Noisy Rationales
- New benchmark: NoRa
- **New algorithm: CD-CoT**
  - Motivation and design of CD-CoT
  - **Empirical evaluations of CD-CoT**
- Take home messages
- Future directions

Empirical evaluations of CD-CoT

Task	Method $\mathcal{M}$	Additional Information	Acc($\mathcal{M}$, $\mathcal{Q}$, $\mathcal{P}_{\text{clean}}$)	Acc($\mathcal{M}$, $\mathcal{Q}$, $\mathcal{P}_{\text{irrelevant}}$) Easy	Medium	Hard	Avg.	Acc($\mathcal{M}$, $\mathcal{Q}$, $\mathcal{P}_{\text{inaccurate}}$) Easy	Medium	Hard	Avg.
Math Base-9	Base	-	46.4	39.3	30.3	26.6	32.1	23.2	10.1	6.0	13.1
	w/ SCO [29]	Ground Truth	53.6	46.3	39.6	36.4	40.8	34.7	22.0	17.7	24.8
	w/ BT [81]	Noise Position	47.2	39.2	34.2	29.9	34.4	30.1	18.4	14.1	20.9
	w/ CC [9]	Clean Demo	44.9	43.3	44.6	45.5	44.5	37.2	31.7	30.7	33.2
	w/ CD-CoT (ours)	Clean Demo	**60.7**	**59.7**	**60.7**	**57.2**	**59.2**	**54.0**	**58.7**	**48.4**	**53.7**
Math Base-11	Base	-	23.9	19.1	13.6	10.7	14.5	14.0	6.7	3.6	8.1
	w/ SCO [29]	Ground Truth	**33.0**	29.2	24.0	20.0	24.4	**29.2**	20.0	17.2	22.1
	w/ BT [81]	Noise Position	24.3	17.9	17.2	13.7	16.3	12.8	9.2	6.8	9.6
	w/ CC [9]	Clean Demo	22.3	19.1	18.4	18.2	18.6	19.0	15.3	14.6	16.3
	w/ CD-CoT (ours)	Clean Demo	31.0	**33.7**	**32.7**	**34.7**	**33.7**	29.0	**30.7**	**25.3**	**28.3**
Symbolic Equal	Base	-	32.7	28.1	25.1	23.0	25.4	29.1	26.1	22.7	26.0
	w/ SCO [29]	Ground Truth	38.5	34.9	33.4	32.7	33.7	34.0	34.1	34.5	34.2
	w/ BT [81]	Noise Position	31.8	26.0	22.7	22.6	23.8	26.3	22.7	22.9	24.0
	w/ CC [9]	Clean Demo	37.8	33.8	32.7	32.0	32.8	31.3	33.0	29.9	31.4
	w/ CD-CoT (ours)	Clean Demo	**42.7**	**44.7**	**42.7**	**44.0**	**43.8**	**42.6**	**41.3**	**42.7**	**42.2**
Symbolic Longer	Base	-	9.2	6.3	7.2	6.0	6.5	7.0	6.8	6.0	6.6
	w/ SCO [29]	Ground Truth	**18.7**	**12.1**	10.5	11.3	11.3	**15.2**	**15.9**	9.8	**13.6**
	w/ BT [81]	Noise Position	7.2	3.4	3.5	2.5	3.1	3.8	3.6	3.6	3.7
	w/ CC [9]	Clean Demo	9.4	9.8	7.9	7.9	8.5	8.5	7.4	6.5	7.5
	w/ CD-CoT (ours)	Clean Demo	12.3	12.0	**12.0**	**13.0**	**12.3**	12.3	10.0	**11.0**	11.1
Commonsense	Base	-	45.7	44.3	42.3	41.4	42.7	36.7	33.4	28.3	32.8
	w/ SCO [29]	Ground Truth	**63.5**	**60.1**	**56.1**	**60.3**	**58.8**	**56.2**	**58.5**	**57.9**	**57.5**
	w/ BT [81]	Noise Position	47.7	23.5	28.3	32.5	28.1	11.6	11.0	15.8	12.8
	w/ CC [9]	Clean Demo	48.3	45.7	43.6	44.0	44.4	42.1	40.8	40.5	41.1
	w/ CD-CoT (ours)	Clean Demo	49.0	50.3	54.7	50.3	51.8	51.0	49.7	49.7	50.1

Table 8: Performance of denoising methods that require additional information for supervision.

Observation 7: CD-CoT presents a significant performance improvement across all datasets, **with an average improvement of 17.8%** compared with the base model under noisy settings.

Observation 8: CD-CoT displays remarkable **resistance to the magnitude of noise**, especially in the challenging mathematical tasks.

Empirical evaluations of CD-CoT

Hyper-parameters				Acc($\mathcal{M}$, $\mathcal{Q}$, $\mathcal{P}_{\text{irrelevant}}$)			Acc($\mathcal{M}$, $\mathcal{Q}$, $\mathcal{P}_{\text{inaccurate}}$)		
N	M	D	C	Base-9	Sym.(E)	Com.	Base-9	Sym.(E)	Com.
5	1	5	Y	57.7	38.7	55.3	53.3	39.7	51.0
5	1	5	N	54.7	32.7	53.7	47.0	32.3	**55.7**
5	2	2+3	Y	**60.7**	**42.7**	54.7	**58.7**	41.3	49.7
5	2	2+3	N	56.7	33.0	54.7	49.7	32.0	53.0
5	3	1+2+2	Y	**60.7**	38.7	53.3	58.0	**43.3**	49.0
5	3	1+2+2	N	56.0	33.3	55.7	48.7	32.0	52.3
5	5	1	Y	59.3	39.7	55.7	58.0	39.0	48.7
5	5	1	N	55.3	35.7	**55.9**	48.7	33.3	50.7

Table 9: Comparison of accuracy on medium-level tasks.

Hyper-parameters				#Tokens in step-3 (irr.)			#Tokens in step-3 (ina.)		
N	M	D	C	Base-9	Sym.(E)	Com.	Base-9	Sym.(E)	Com.
5	1	5	Y	1440	3162	788	1428	3170	798
5	1	5	N	1301	2685	660	1295	2732	667
5	2	2+3	Y	2175	4934	1269	2156	4989	1311
5	2	2+3	N	1864	4044	1005	1842	4087	1039
5	3	1+2+2	Y	2902	6704	1772	2878	6785	1821
5	3	1+2+2	N	2416	5360	1372	2393	5443	1420
5	5	1	Y	4368	10340	2764	4339	10514	2845
5	5	1	N	3535	8099	2088	3506	8303	2163

Table 10: Comparison of #tokens on medium-level tasks.

Observation 9:
The clean CoT demonstration plays a pivotal role in CD-CoT.

Observation 10:
The accuracy exhibits **subtle variations** when employing different algorithm instances. We set M = 2 to strike a balance of efficiency and effectiveness.

Observation 11:
An ablation study of components in Appendix F.3 demonstrates the denoising power and performance gain of CD-CoT, attributed to its **contrastive denoising with rationale rephrasing** and **repeated reasoning with voting components.**

Empirical evaluations of CD-CoT

Model	Method	$\text{Acc}(\mathcal{M}, \mathcal{Q}, \mathcal{P}_{\text{irrelevant}})$ Base-9	Sym.(E)	Com.	$\text{Acc}(\mathcal{M}, \mathcal{Q}, \mathcal{P}_{\text{inaccurate}})$ Base-9	Sym.(E)	Com.
GPT-3.5-turbo	Base	30.3	25.1	42.3	10.1	26.1	33.4
	SC	36.6	28.3	<u>45.0</u>	17.3	30.7	<u>44.7</u>
	BT	34.2	22.7	28.3	18.4	22.7	11.0
	CC	<u>44.3</u>	<u>32.7</u>	43.6	<u>31.7</u>	<u>33.0</u>	40.8
	CD-CoT	**60.7**	**42.7**	**54.7**	**58.7**	**41.3**	**49.7**
Gemini-Pro	Base	72.3	38.9	53.2	21.2	36.7	33.5
	SC	80.3	<u>43.3</u>	<u>60.0</u>	32.3	<u>45.0</u>	42.7
	BT	<u>82.4</u>	29.3	37.8	26.7	28.7	33.3
	CC	67.5	37.3	50.2	<u>43.6</u>	35.0	<u>45.6</u>
	CD-CoT	**92.7**	**49.3**	**57.7**	**76.7**	**53.3**	**55.7**
LLaMA2-70B	Base	2.8	8.7	41.9	2.7	9.1	<u>40.2</u>
	SC	**5.0**	10.3	**46.7**	**3.0**	9.7	**46.0**
	BT	1.4	<u>11.2</u>	36.1	0.9	<u>12.5</u>	36.2
	CC	1.1	**16.3**	29.9	<u>2.8</u>	**14.0**	28.3
	CD-CoT	<u>4.0</u>	9.7	<u>39.3</u>	2.7	9.7	39.7
Mixtral-8x7B	Base	16.3	17.9	34.9	3.7	15.1	31.1
	SC	<u>20.0</u>	<u>21.7</u>	<u>37.0</u>	2.7	18.0	<u>37.7</u>
	BT	4.1	9.7	6.2	2.4	10.1	10.5
	CC	**24.4**	18.5	36.0	**12.5**	<u>18.3</u>	35.7
	CD-CoT	8.7	**22.7**	**40.3**	<u>4.7</u>	**21.3**	**40.3**

Table 11: Comparing methods with different LLMs.

Observation 12:
CD-CoT generalizes well across different LLMs. It respectively achieves increases of 23.4% and 21.6% in accuracy compared to base models (GPT-3.5-turbo and Gemini-Pro) and surpasses all the baselines.

New algorithm: CD-CoT - Evaluations

Comparison of rephrased rationales by different reasoning methods

Original Noisy Rationale
In base-9, the digits are "012345678". 9 + 8 = 17. We have 6 + 7 = 13 in base-10. Since we're in base-9, that exceeds the maximum value of 8 for a single digit. 13 mod 9 = 4, so the digit is 4 and the carry is 1. We have 8 + 5 + 1 = 14 in base-10. 14 + 8 = 22. 14 mod 9 = 5, so the digit is 5 and the carry is 1. A leading digit is 1. So the answer is 154. Answer: 154.

Method	Denoised Rationale
SM	In base-9, the dkgits are "012345678". 9 + 8 = 17. We5hav6̂r+ 7l=J13 in base-10. Since we're in baze-9, ithFt exceeis the maCimum valur of 8 for a siLgle]digit. x0c3 mod 9 = ; soWghe digix is 4 anj the carry is 1. We hOve 8 + 5i+ z2= 14 Vn baAe 10. 14 d 8 = 2p. 14 mod 9 = 5, so Xhe diGit*iU 5 and the carry is 1Q A leading x0cigit is 1. So the answer is 154. Asswer: 154.
SD	In base-9, the digits are "012345678". 9 + 8 = 17. We have 6 + 7 = 13 in base-10. Since we're in base-9, 13 exceeds the maximum value of 8 for a single digit. 13 mod 9 = 4, so the digit is 4 and the carry is 1. We have 8 + 5 + 1 = 14 in base 10. 14 mod 9 = 8. 14 div 9 = 1, so the digit is 8 and the carry is 1. A leading digit is 1. So the answer is 154. Answer: 154.
SP	In base-9, the digits range from 0 to 8. Let's add 86 and 57 step by step. First, we add the rightmost digits, which are 6 and 7. In base-10, 6 + 7 equals 13. However, since we are in base-9, the maximum value for a single digit is 8. Therefore, we need to find the remainder when 13 is divided by 9. The remainder is 4, so the digit in the units place is 4 and we have a carry of 1. Next, we add the next digits, which are 8 and 5, along with the carry of 1. In base-10, 8 + 5 + 1 equals 14. Again, we need to find the remainder when 14 is divided by 9. The remainder is 5, so the digit in the tens place is 5 and we have a carry of 1. Finally, we have a leading digit of 1. So the final answer in base-9 is 154. Answer: 154.
Ours	In base-9, the digits are "012345678". We have 6 + 7 = 13 in base-10. Since we're in base-9, that exceeds the maximum value of 8 for a single digit. 13 mod 9 = 4, so the digit is 4 and the carry is 1. We have 8 + 5 + 1 = 14 in base 10. 14 mod 9 = 5, so the digit is 5 and the carry is 1. A leading digit is 1. So the answer is 154. Answer: 154.

Observation 13: CD-CoT effectively **removes noisy thoughts** and ensures format alignment with the original rationale.

Outline

- Background: language model reasoning
- New research problem: Noisy Rationales
- New benchmark: NoRa
- New algorithm: CD-CoT
- **Take home messages**
- **Future directions**

Take home messages

We investigate the under-explored problem of noisy rationales

- we introduce **NoRa dataset** to evaluate LLMs against noisy rationales
- we show the **general vulnerability** LLMs to noisy rationales
  - and is inadequately mitigated by existing robust methods
- we design **CD-CoT** to enhance the robustness via contrastive denoising

Future directions

- **Knowledge-enhanced** denoising within a retrieval-augmented framework
- **Robust inductive reasoning** to extract rules from noisy examples
- Generalization to **out-of-distribution** noisy scenarios
- Expanding the NoRa dataset to include **multi-modal** scenarios, e.g., visual data, for a more comprehensive understanding of the robustness of foundation models
- **Theoretical analysis** of noisy ICL for deeper insights into the noisy rationales

Thanks you!

Zhanke Zhou
cszkzhou@comp.hkbu.edu.hk