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

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, Bo Han

香港浸會大學 HONG KONG BAPTIST UNIVERSITY · TMLR TRUSTWORTHY MACHINE LEARNING AND REASONING · 武漢大學 WUHAN UNIVERSITY

Paper · Code · Slides

New Problem: Noisy Rationales

In-context learning and Chain of thoughts

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?

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?

Chain of thoughts with noisy rationales

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**

Noisy rationales originate from diverse sources (refer to Appendix C)

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

However, LLM's robustness against noisy rationales is unknown

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

New Benchmark: NoRa

Benchmark Construction

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.

Empirical Evaluation

Task	Method $\mathcal{M}$	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/ 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.

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

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

self-consistency methods can improve robustness without true denoising

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	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.

Adjusting temperature can improve reasoning under noisy rationales

Prompting with more noisy examples boosts reasoning accuracy on most tasks

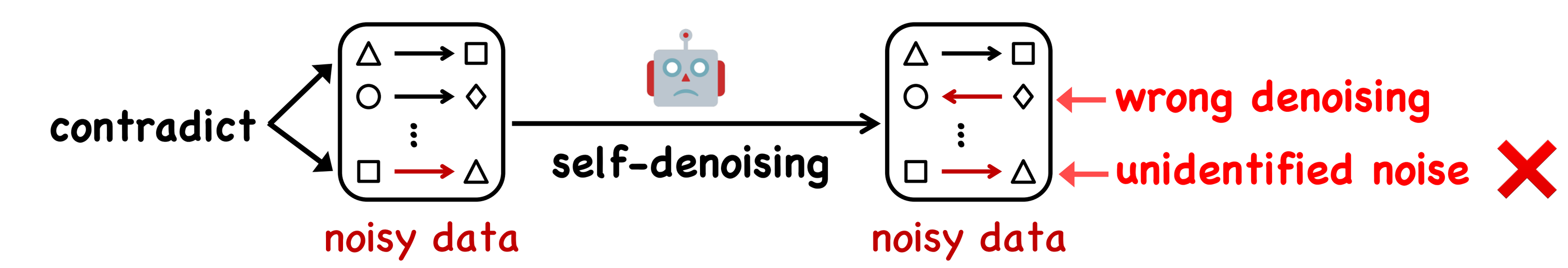
Different LLMs are generally vulnerable to noisy rationales

New Algorithm: CD-CoT

Algorithm Design

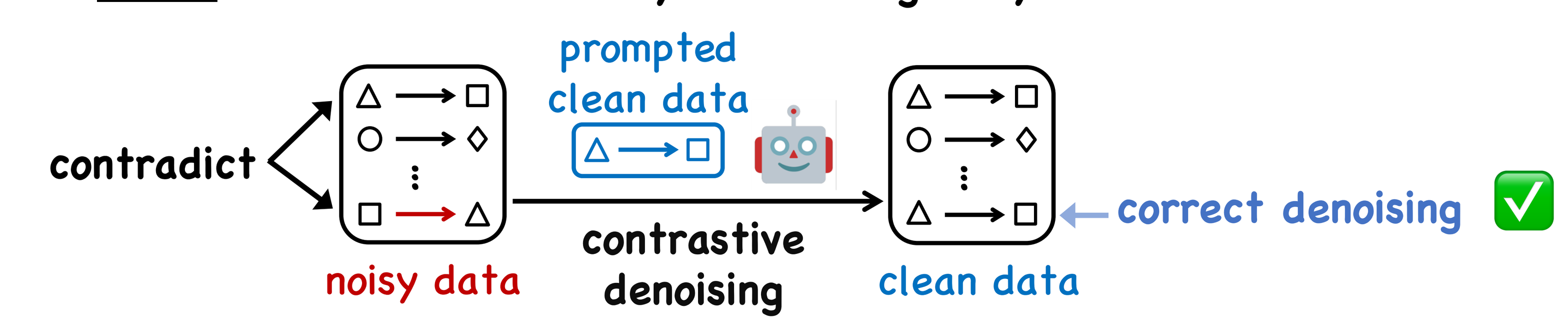
Self-denoising:

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



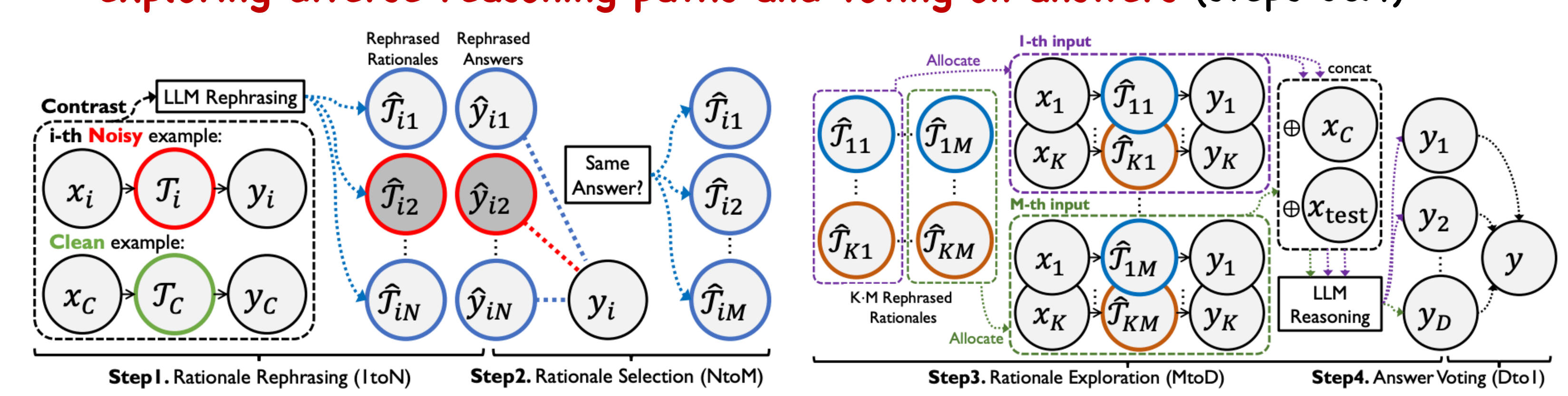
Contrastive Denoising:

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



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

- **rephrasing and selecting rationales** to conduct explicit denoising (steps 1&2)
- **exploring diverse reasoning paths and voting on answers** (steps 3&4)



Empirical Evaluation

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.

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.

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

Method	Rationale
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.
	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.

Table 12: Comparison of rephrased rationales by different reasoning methods.