It Helps to Take a Second Opinion: Teaching Smaller LLMs to Deliberate Mutually via Selective Rationale Optimization

Sohan Patnaik, Milan Aggarwal, Sumit Bhatia, Balaji Krishnamurthy

Media and Data Science Research, Adobe

ICLR 2025 (Poster)

Poster ID: 29886

Friday, 25 April, 2025

10 a.m. — 12:30 p.m.

Adobe

Guiding Observations

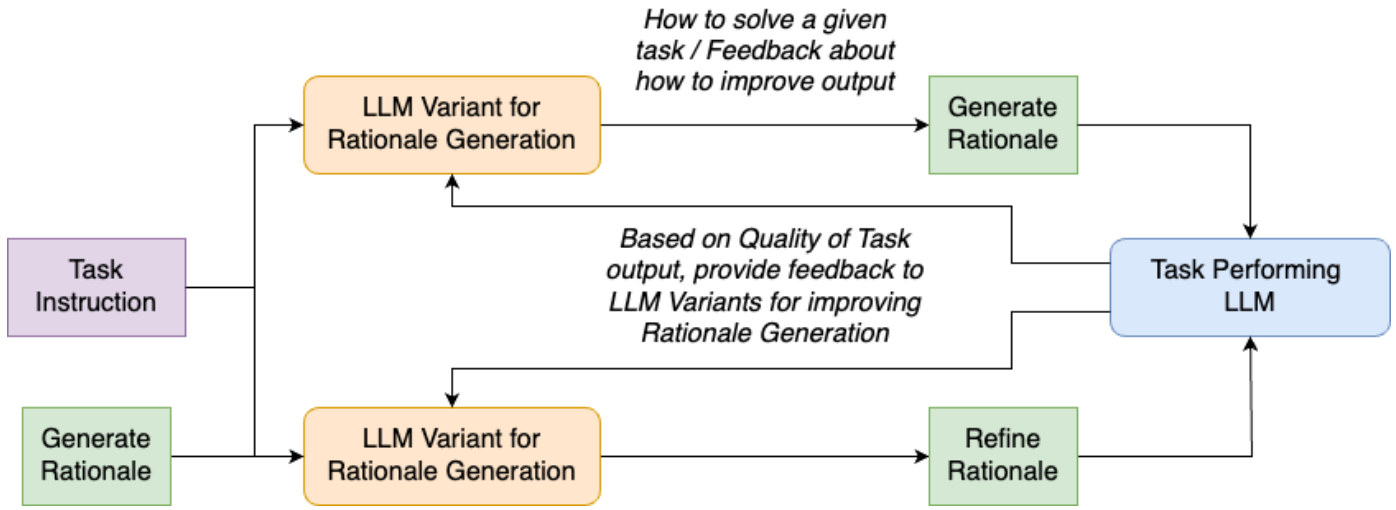
1. LLMs perform better when prompted to generate rationale/reasoning chain
   - *Not only through prompting but during Training/IFT (CoT dataset is widely used)*
   - *Such reasoning chains are **necessary** to ground the tasks where the LLM cannot simply learn the input-to-output paired mappings*
2. Human annotations are difficult to obtain. Annotations for reasoning chain are even harder ...
   - *Can there be a way to automate to some extent?*

Can SLMs Improve Reasoning Without Using External LLMs?

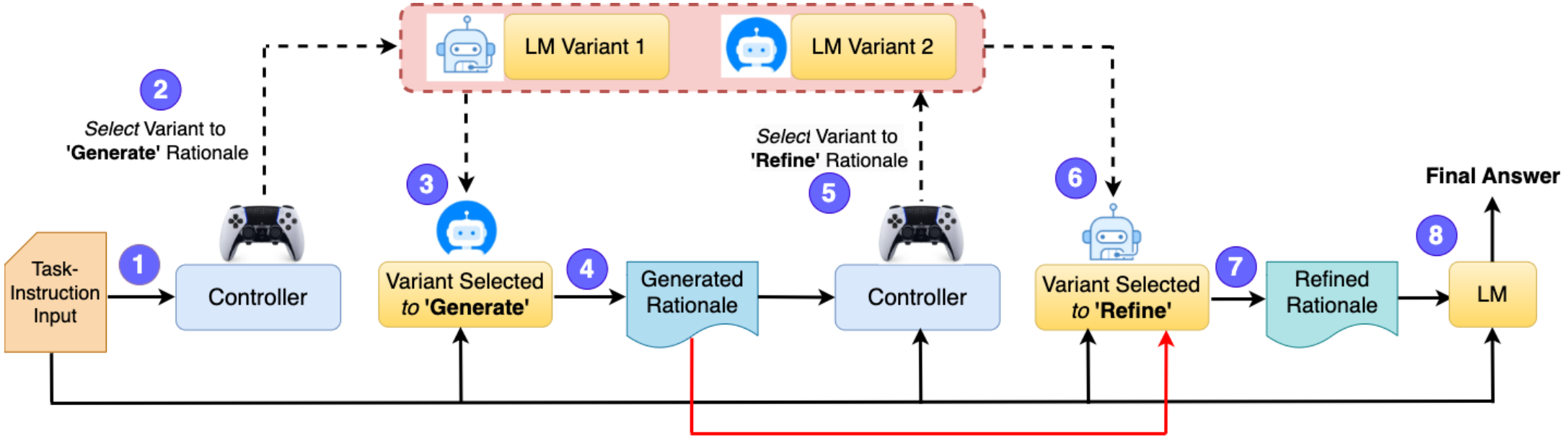
- Lack of transparency in the pre-training data of larger (often closed) LMs (e.g. GPT4) prevents their use in commercial settings due to legal constraints

- SLMs (≤13B parameters) struggle with complex reasoning, often producing shallow or redundant rationales.
  - Existing methods rely on LLM supervision or task-specific ground-truth (GT) rationales, both of which are impractical—GT rationales are scarce, and LLM dependence introduces legal and cost constraints.

- **Research Question:** Can we train the SLMs without relying on external LLMs and task-specific GT rationales to
  - generate (and, refine) diverse rationales, and
  - select the high-quality rationales leading to improved performance on end-tasks.

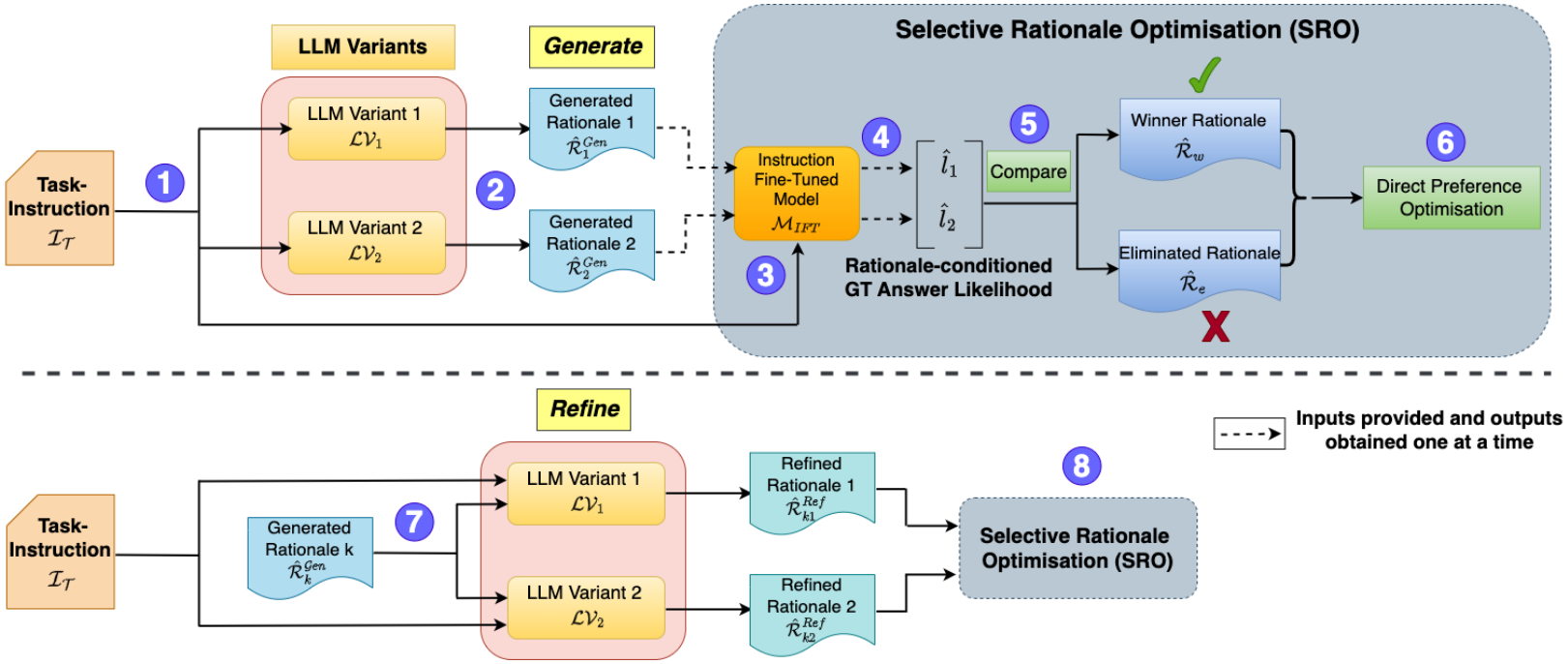


Inference Flow Through COALITION



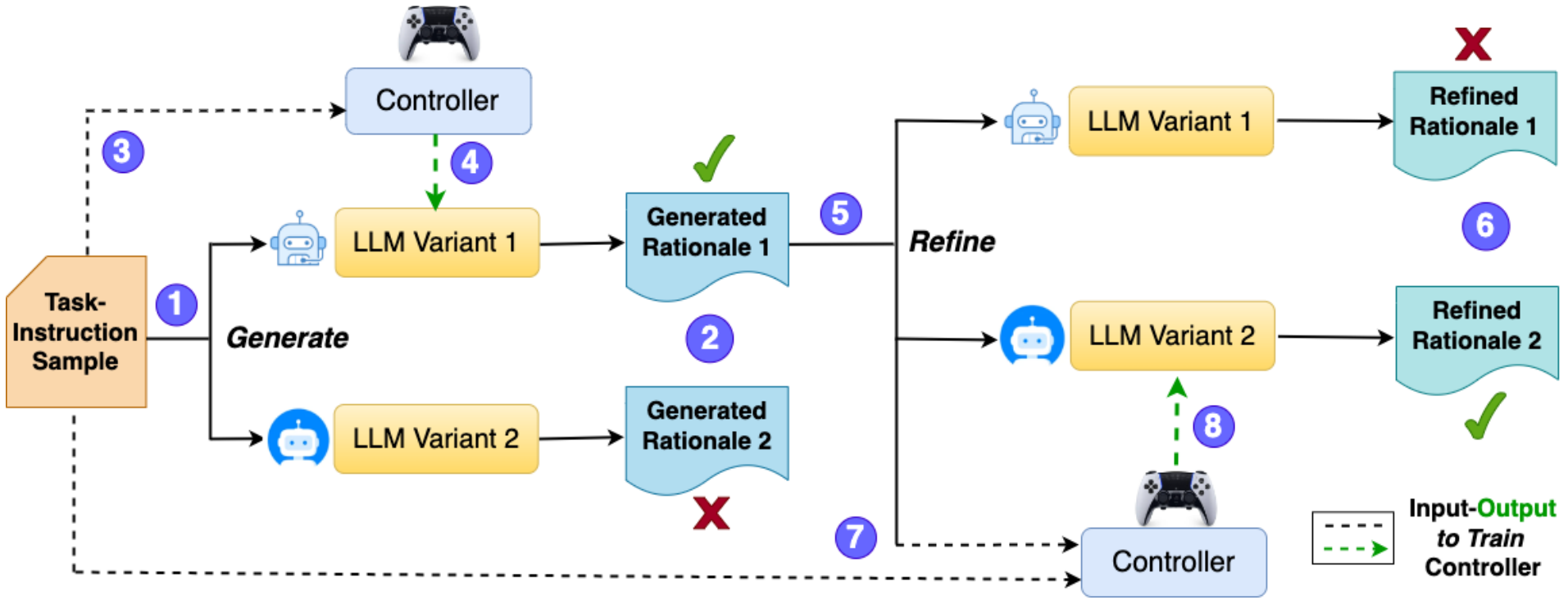
- Schematic flow of inference using COALITION which leverages two variants of the same LM.
    - The sample is fed to a controller (step 1) to select the variant (steps 2-3) that generates a rationale (step 4).
    - The generated rationale is then fed to the controller to select the variant (steps 5-6) to refine the rationale (step 7)
    - Refined rationale is used to obtain the final answer (step 8).

Selective Rationale Optimisation (SRO)



- Each SLM variant generates an initial rationale for a given task instruction. The rationales are then refined through self-refinement and cross-refinement, ensuring diverse reasoning paths.
- Each rationale is evaluated based on its effectiveness in deriving the correct answer. A utility score is computed, and rationales are ranked to identify high-quality reasoning paths.
- DPO optimizes the LLM variants to prefer high-scoring rationales in both generation and refinement steps

Training the Controller



- *Training via Rationale Preference Data*
  - The controller is trained using preference data from SRO training, learning to select the variant that generated the winning rationale.

- *Two-Step Classification*
  - Controller first predicts the best variant for rationale generation and then, conditioned on the generated rationale, selects the variant for refinement.

Does COALITION Help Improve Performance?

Method	Maths GSM8K	NLI WinoGrande	Comonsense PIQA	HellaSwag	CSQA
Meta-Llama3-8B-Instruct w/o rationale (Dubey et al., 2024)	75.89	71.98	78.51	57.69	76.17
Prompt-Driven Rationale Refinement					
Chain-of-Thought (Wei et al., 2022b)	62.39	60.17	68.48	45.33	65.28
CoT Self-Consistency (Wang et al., 2023b)	64.11	62.37	71.42	46.13	68.92
Tree-of-Thought (Yao et al., 2023)	68.11	70.62	75.14	53.18	74.37
Exchange-of-Thought (Yin et al., 2023)	69.19	66.47	73.11	52.22	75.48
Self-Refine (Madaan et al., 2023b)	77.26	72.81	79.49	60.48	78.22
Rationale Enhancement via Trainable Self-Play					
Distilling Step-by-Step (Hsieh et al., 2023)	76.18	70.41	78.77	56.10	76.31
Self-Rewarding LMs (Yuan et al., 2024)	72.16	68.15	75.22	55.39	76.15
Self-Play Fine-Tuning (SPIN) (Chen et al., 2024)	77.01	71.85	79.02	58.22	76.48
COALITION (ours)	**81.06**	**77.13**	**83.26**	**63.23**	**82.06**

Table 1: COALITION significantly outperforms multiple categories of baselines on 5 datasets from 3 task domains - Maths Problem Solving, Natural Language Inference (NLI) and Commonsense Reasoning. Same LLM-backbone (Llama3-8B) is used in all the methods for uniform comparison.

Rationales from COALITION are better than SPIN

Rationales Generated by Best Baseline (SPIN) and LLM Variants - $\mathcal{LV}_1$ and $\mathcal{LV}_2$ in COALITION on GSM8K Sample
Instruction: A company pays each of its employees **$600 in a month**. The company has a policy of increasing the salaries of each of its employees by **10% of the initial salary every year** for those who've **stayed in the company for five years**. If Sylvie just **clocked 5 years** in the company last December, what's her **annual** salary after **three more years** of service?
Best Baseline: Sylvie's annual salary after three more years will be **$3600** because her salary **increases by $600 every year for 5 years**. So after three years, it will be **$600*5+$600=$3600**. ✗
Generate w $\mathcal{LV}_1$: Sylvie's salary will **increase by $600 every year for three years**, so her annual salary after three more years of service will be **$600+($600*3)=$2,400**. ✗ **Generate w $\mathcal{LV}_2$:** Sylvie receives a 10% increase in her salary every year after 5 years of service. Since her current salary is $600, it will **increase by $60 each year**. After three more years, her new salary will be **$600+(3*$60)=$780**. ✗
Cross-Refine $\mathcal{LV}_2$ rationale w $\mathcal{LV}_1$: Sylvie's initial monthly salary is $600. After completing five years, her salary increases by 10% of initial salary, making it $600*1.1=$660. After three more years of service, she will have received two additional annual 10% increases, which means her **monthly salary will be $660+$60+$60=$780**. Therefore, her **annual** salary after three more years will be **$780*12=$9360**. ✓ **Self-Refine $\mathcal{LV}_2$ rationale w $\mathcal{LV}_2$:** Sylvie's salary increases by 10% each year after her initial five years. Her monthly salary after five years is $660, and in three more years, she will have received two additional 10% raises. Therefore, her monthly salary will be **$660(1.1)*(1.1)=$756**, and her annual salary will be **$756*12=$9072**. ✗
Note: Controller selected $\mathcal{LV}_2$ for **generate** step and $\mathcal{LV}_1$ for **refine step**. Variant selection in this order yields correct rationale.

Table 2: COALITION yields better rationale using *generate* and *refine* steps via LLM variants. Wrong and right parts in a rationale are in **red** and **green**. Baseline wrongly applies increase for first five years. $\mathcal{LV}_1$ estimates wrong annual increase while $\mathcal{LV}_2$ gives correct monthly increase but question asks annual salary. Cross-refining using $\mathcal{LV}_1$ (as selected by controller) rectifies this error.

How Does COALITION Work With Different Model Families and Across Varying Parameter Scales?

Model	Parameter Scale	Maths GSM8K	NLI WinoGrande	PIQA	Commonsense HellaSwag	CSQA
Phi3 (Abdin et al., 2024)	3.8B	10.36	73.32	80.30	59.01	72.48
w/ COALITION (ours)	3.8B	**14.76**	**76.19**	**84.48**	**63.72**	**75.04**
Qwen1.5 (Bai et al., 2023)	4B	3.49	67.01	75.57	52.01	74.61
w/ COALITION (ours)	4B	**5.58**	**69.26**	**76.48**	**53.37**	**78.29**
Qwen1.5 (Bai et al., 2023)	7B	57.01	69.53	79.54	61.06	81.00
w/ COALITION (ours)	7B	**61.37**	**75.02**	**83.11**	**64.22**	**85.11**
Qwen1.5 (Bai et al., 2023)	14B	69.37	76.01	81.45	65.57	84.19
w/ COALITION (ours)	14B	**74.88**	**82.84**	**84.39**	**69.22**	**87.48**
Mistral (Jiang et al., 2023a)	7B	48.52	74.43	81.66	64.78	69.21
w/ COALITION (ours)	7B	**54.42**	**78.39**	**85.01**	**68.48**	**74.38**
LLaMA3 (Dubey et al., 2024)	8B	75.89	71.98	78.51	57.69	76.17
w/ COALITION (ours)	8B	**81.06**	**77.13**	**83.26**	**63.23**	**82.06**

Table 3: Performance evaluation of COALITION with LMs of varying scale of parameters (4B to 14B) and different model families (Phi3, Qwen1.5, Mistral, Llama3). It is observed that COALITION yields significant gains on all tasks for different model families and parameter scales.

Variant Selection via Controller Boosts Accuracy

Communication Mode	Maths	NLI		Commonsense	
	GSM8K	WinoGrande	PIQA	HellaSwag	CSQA
Generate (w $\mathcal{LV}_1$) w/o Refine	77.26	75.10	80.11	59.22	78.47
Generate (w $\mathcal{LV}_2$) w/o Refine	77.21	74.89	79.24	59.31	78.33
Self-Refine ($\mathcal{LV}_1 \rightarrow \mathcal{LV}_1$)	77.35	74.91	79.86	59.89	79.35
Self-Refine ($\mathcal{LV}_2 \rightarrow \mathcal{LV}_2$)	77.23	74.70	79.60	59.70	79.25
Cross-Refine ($\mathcal{LV}_1 \rightarrow \mathcal{LV}_2$)	79.94	75.52	81.31	61.21	80.16
Cross-Refine ($\mathcal{LV}_2 \rightarrow \mathcal{LV}_1$)	79.53	75.83	81.02	60.87	80.21
COALITION (w Controller)	**81.06**	**77.13**	**83.26**	**63.23**	**82.06**

Table 4: Performance analysis of rationales inferred - 1) w/o refinement (rows 1-2), 2) w self-refinement (rows 3-4), 3) w cross-refinement using fixed order of variants for all samples (rows 5-6), and 4) w controller (last row). Selecting LM variants using the **controller** for the generate and refine steps yields best results. Cross-communication between the variants is better than both self-refine (using a single variant) and not refining by directly using rationale generated by a variant.

Conclusion

- **SRO Improves COALITION:** A trainable framework that enhances smaller LMs on complex tasks by using distinct model variants for rationale generation and refinement.

- **No External Supervision Required:** Achieves high-quality reasoning without relying on external LLMs or ground-truth rationale annotations, addressing legal and ethical constraints.

Adobe