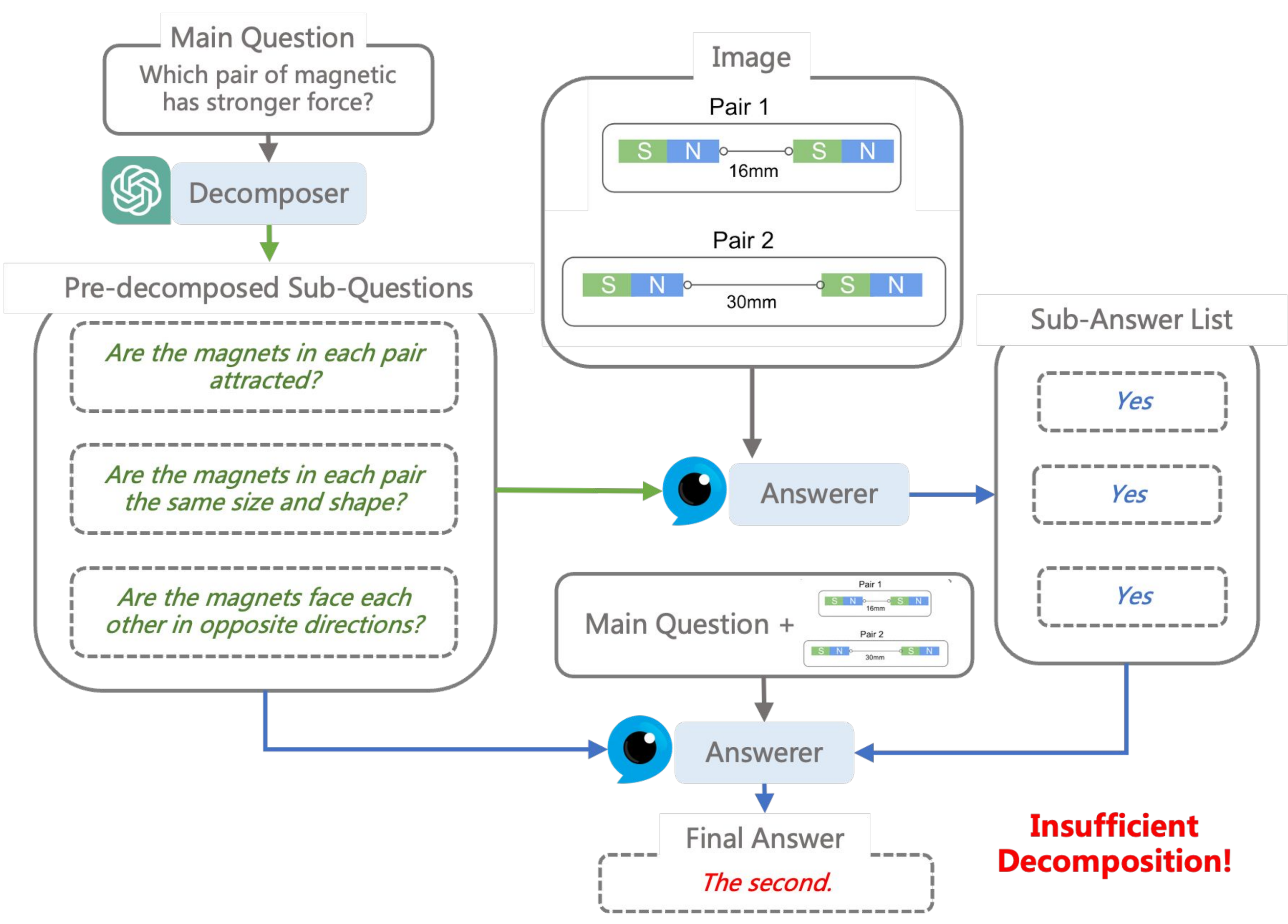
NeurIPS 2024 Workshop on Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning

Enhancing Multi-Agent Multi-Modal Collaboration with Fine-Grained Reward Modeling

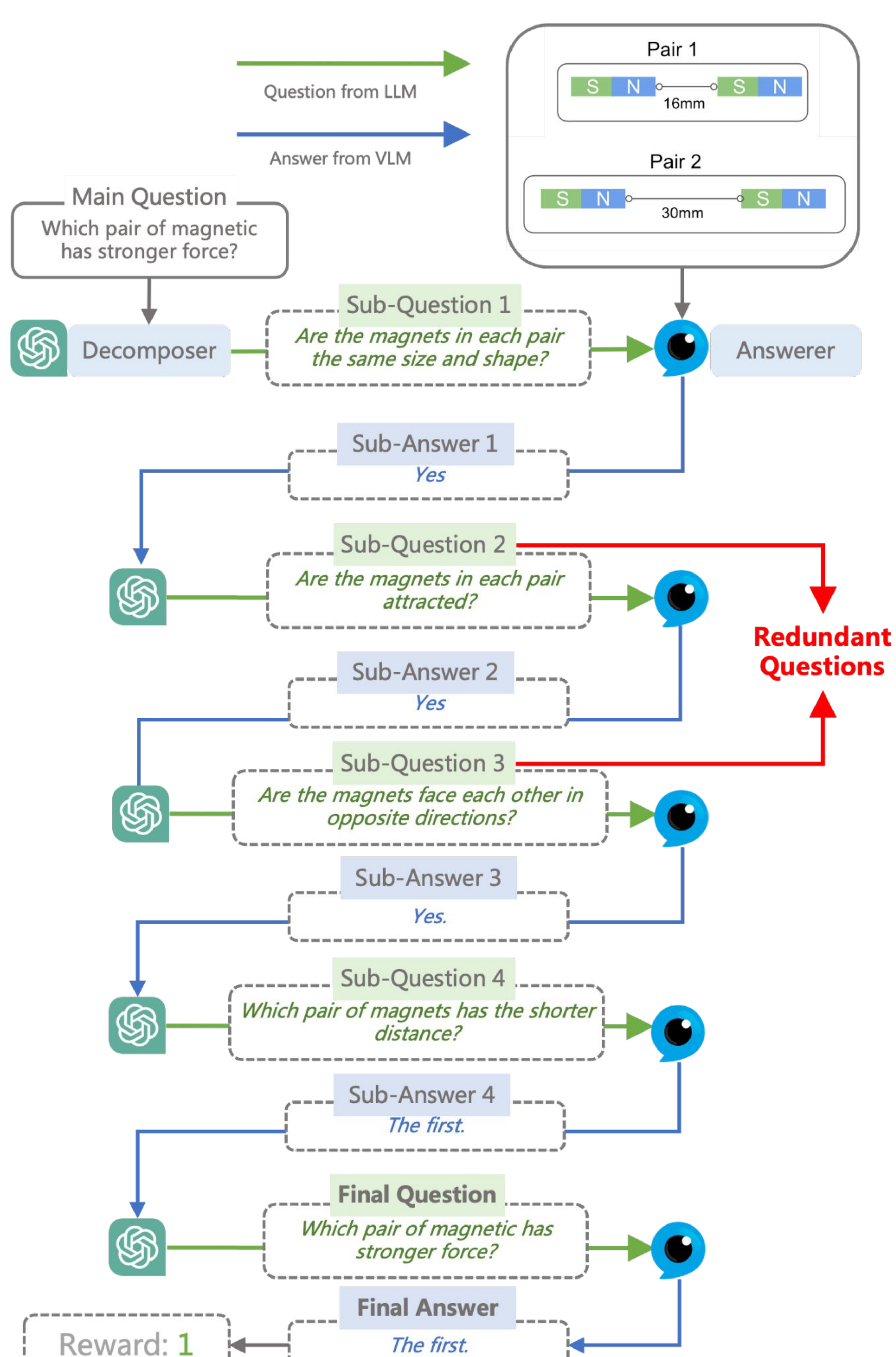
Qian Yang[1,2] Weixiang Yan[3] Aishwarya Agrawal[1,2,4]
Contact: qian.yang@mila.quebec
[1] Mila - Québec AI Institute [2] Université de Montréal
[3]University of California, Santa Barbara [4]Canada CIFAR AI Chair

Motivation

- Multi-agent collaboration helps solve complex tasks:
  - LLMs: Decomposer agent for task decomposition.
  - MLLMs: Answerer agent for task solving.
- ***Pre-decomposition:*** Fails to incorporate feedback from MLLMs.

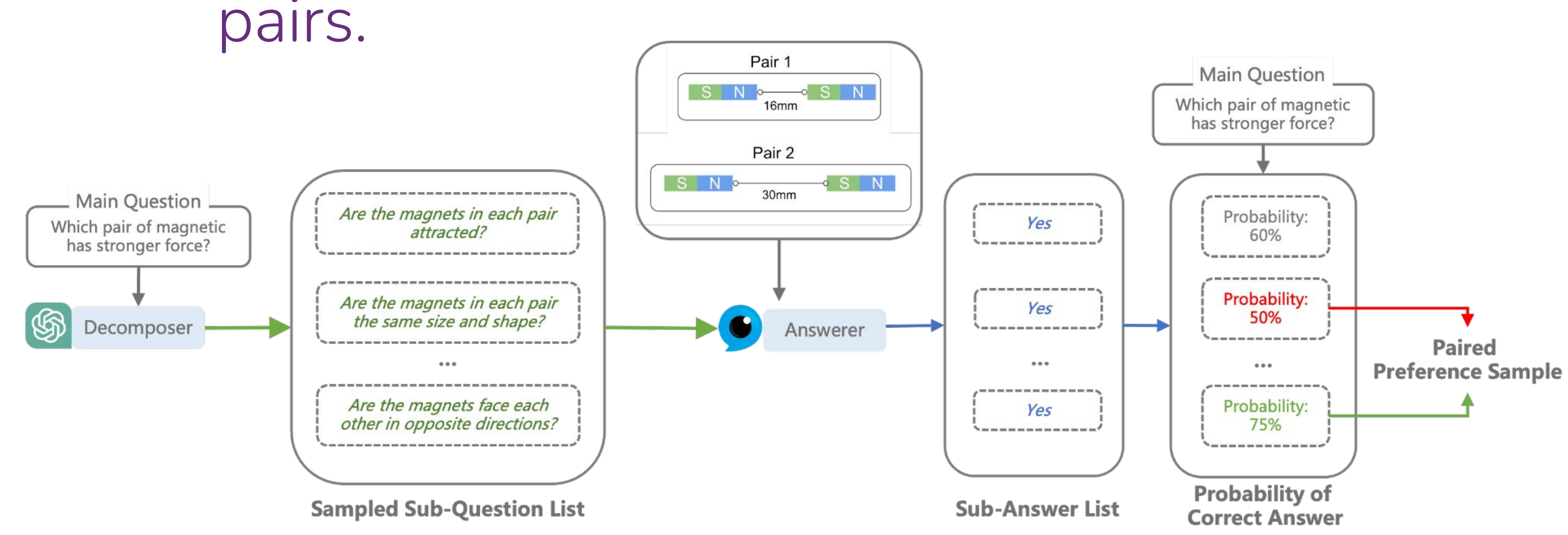


- ***Interactive decomposition with coarse reward:*** Dynamically refines sub-questions but fails to incentivize meaningful and efficient ones.



Method

- Automatic construction of paired preference dataset:
  - Sampling sub-questions generated by the decomposer agent.
  - Using the MLLM to answer each sub-question.
  - Using the MLLM to answer the main question with each sub-QA pair as additional context.
  - Constructing preference pairs by comparing answer confidence based on different sub-QA pairs.



- Finetune the decomposer agent on preference dataset using DPO:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

Experiments

- **Implement Details:**
  - ***Decomposer Agent:*** OpenHermes-2.5-Mistral-7B.
  - ***Candidate MLLM:*** Idefics2-8B.
- Pre-decomposition performs comparable to interactive decomposition w/o tuning.
- Our method (Line 9) ranks 1st in the i.i.d. setting and 1st/2nd on 2 out of 3 datasets in the o.o.d. setting. It achieves the highest mean performance, matching that of SFT with coarse reward.

	Model	SNLI-VE†	VCR	Winoground	MathVista	Mean
1	Base MLLM	39.3	62.3	50.5	48.0	50.0
2	Base MLLM + Sample	39.5	62.5	49.3	48.2	49.9
3	Base MLLM + Chain-of-Thought	43.6	63.0	49.3	47.2	50.8
4	Base MLLM + Chain-of-Thought + Sample	44.3	62.1	49.0	48.1	50.9
5	Pre-Decomposition	53.0	64.0	53.5	49.0	54.9
	Interactive Decomposition					
6	Interactive Decomposition	54.1	61.1	55.8	48.4	54.9
7	SFT$_{VCR_{7K}+SNLI_{13K}}$	54.1	61.9	55.3	48.4	54.9
8	SFT + PPO$_{SNLI_{3K}}$ with Coarse-Grained Reward	53.7	65.2	55.3	47.8	55.5
9	DPO$_{SNLI_{50k}}$ with Fine-Grained Reward	56.3	61.5	55.8	48.5	55.5

Contributions

- **Systematically evaluate** various question decomposition strategies.
- Introduce **fine-grained reward modeling** to enhance multi-agent, multi-modal collaboration **without additional annotations**.
- Experimental results show **significant improvements** in decomposition adaptability and efficiency with fine-grained reward modeling.