2024.02.21 | Lab Seminar

PHENOMENAL YET PUZZLING: TESTING INDUCTIVE REASONING CAPABILITIES OF LANGUAGE MODELS WITH HYPOTHESIS REFINEMENT

HUMAN
CENTERED
COMPUTING
LABORATORY

인간중심컴퓨팅 연구실
석사과정 송형우

HUMAN CENTERED COMPUTING LABORATORY

PHENOMENAL YET PUZZLING: TESTING INDUCTIVE REASONING CAPABILITIES OF LANGUAGE MODELS WITH HYPOTHESIS REFINEMENT

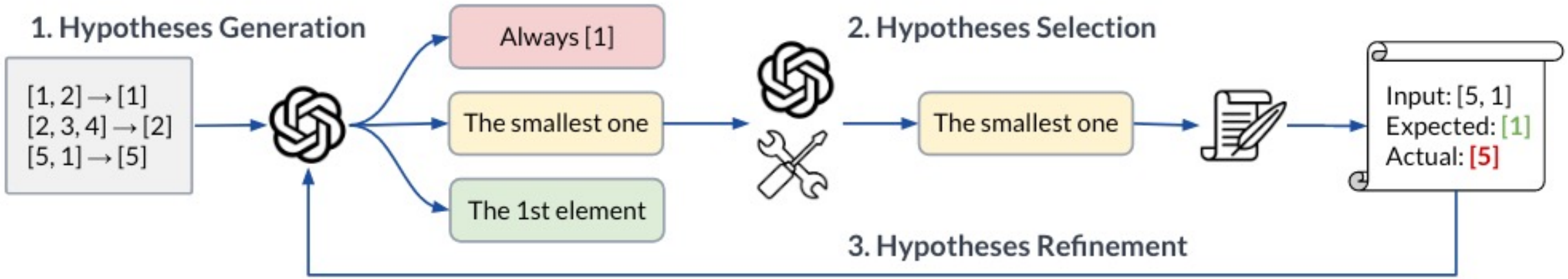
Linlu Qiu[1]*, **Liwei Jiang**[2,3], **Ximing Lu**[2,3], **Melanie Sclar**[3], **Valentina Pyatkin**[2,3], **Chandra Bhagavatula**[2], **Bailin Wang**[1], **Yoon Kim**[1], **Yejin Choi**[2,3], **Nouha Dziri**[2], **Xiang Ren**[2,4]
[1]Massachusetts Institute of Technology, [2]Allen Institute for Artificial Intelligence
[3]University of Washington, [4]University of Southern California
linluqiu@mit.edu

1. Inductive Reasoning with LMs
    1. 예시들을 보며 가설을 추론
    2. 하지만, 만족스럽지 않은 성능
    3. 그리고, 중간 과정을 살펴보는 연구가 없어 해석도에 대한 이야기가 없음

2. Language Hypothesis Optimization
    1. 언어로 표현된 가설의 최적화 과정에 대한 연구 존재
    2. 하지만, 주로 설명을 덧붙이는 과정으로 진행

3. Bayesian Concept Learning
    1. 가설 공간을 설정하고 (기존 믿음) 관찰에 대한 확률 계산을 통해 새로운 가설 공간 선정 (새로운 믿음)
    2. 가설 공간의 크기와 계산 비용간의 trade-off가 존재함

Iterative Hypothesis Refinement



Hypothesis Proposer (phenomenal)

- 여러가지 제안을 생성하고 선정하는 역할은 잘한다

Inductive Reasoner (puzzling)

- 실제로 적용하는 것에는 어려움을 겪는다

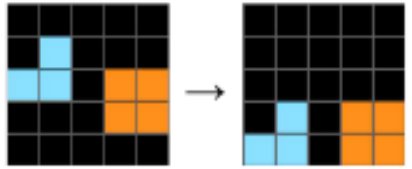
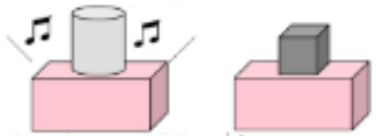
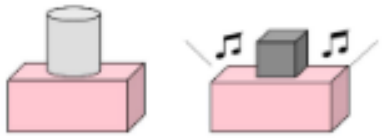
HUMAN CENTERED COMPUTING LABORATORY

Inducing Causal Relations

Language-like Compositional Instructions

Symbolic Operations

Visual Concepts

	ACRE	MiniSCAN	List Functions	MiniARC
Examples		dax → 🔴 lug → 🔵 lug fep → 🔵🔵🔵 dax fep → 🔴🔴🔴	[1, 2, 3] → [1] [2, 3, 4] → [2] [5, 1] → [5]	
Bad Rule		dax → 🔴 lug → 🔵 [X] fep → ⚫⚫	The smallest one	Swap the colors of two objects
Good Rule		dax → 🔴 lug → 🔵 [X] fep → ⚫⚫⚫	The 1st element	Drop all objects

Inducing Causal Relations

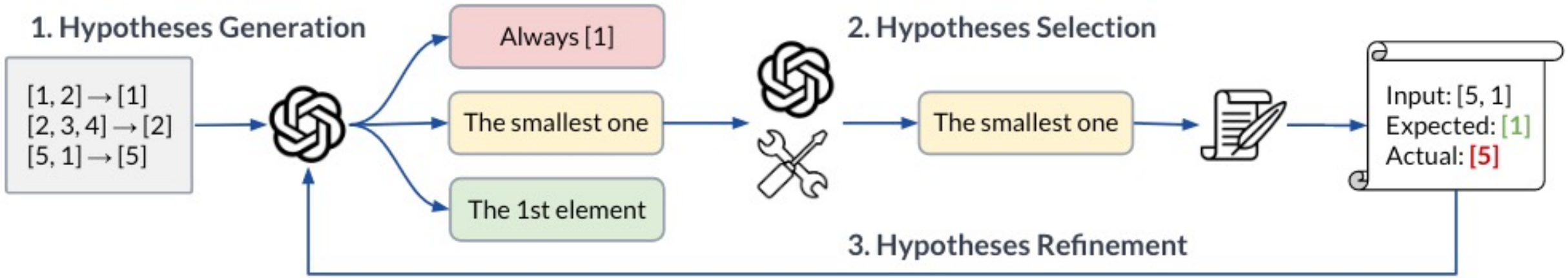
Language-like Compositional Instructions

Symbolic Operations

Visual Concepts

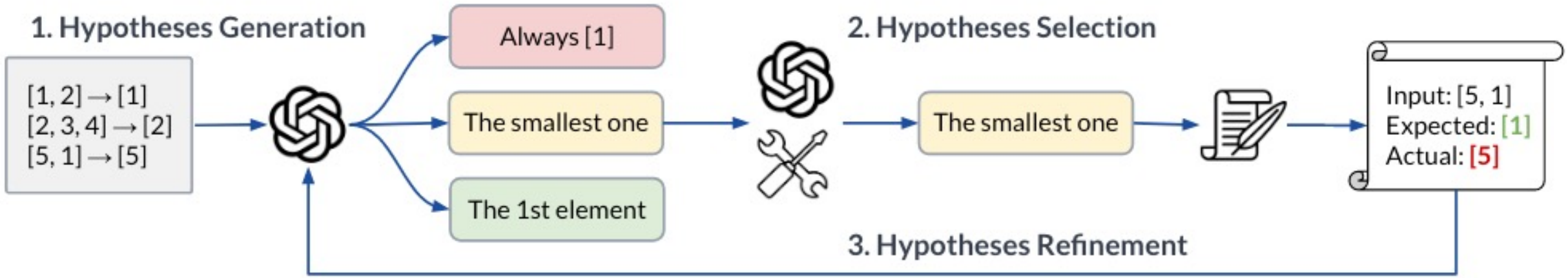
Table 3: The number of tasks per dataset, the numbers of seen examples per task, and unseen examples per task.

Dataset	# Tasks	# Seen	# Unseen
ACRE	100	6	4
MiniSCAN	100	14	10
List Functions	250	8	8
MiniARC	130	3	3

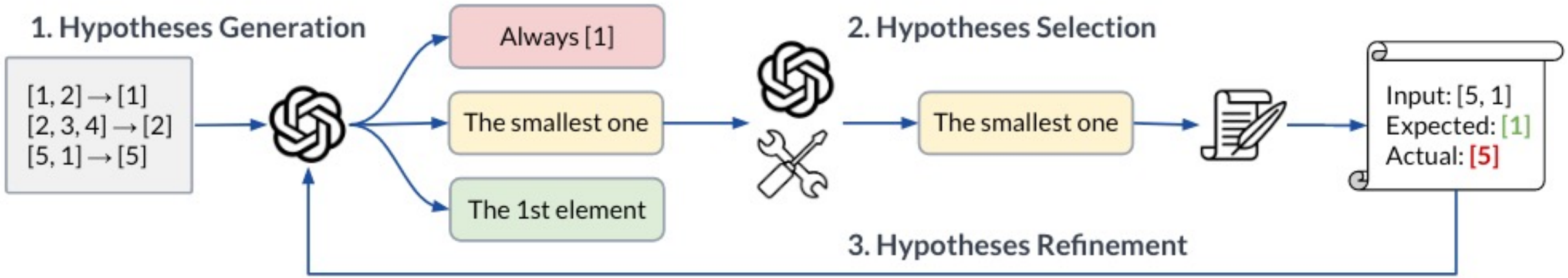


$$h^t \sim P_{\mathrm{LM}}(\cdot \mid d^{t-1}, x_1, y_1, \ldots, x_k, y_k)$$

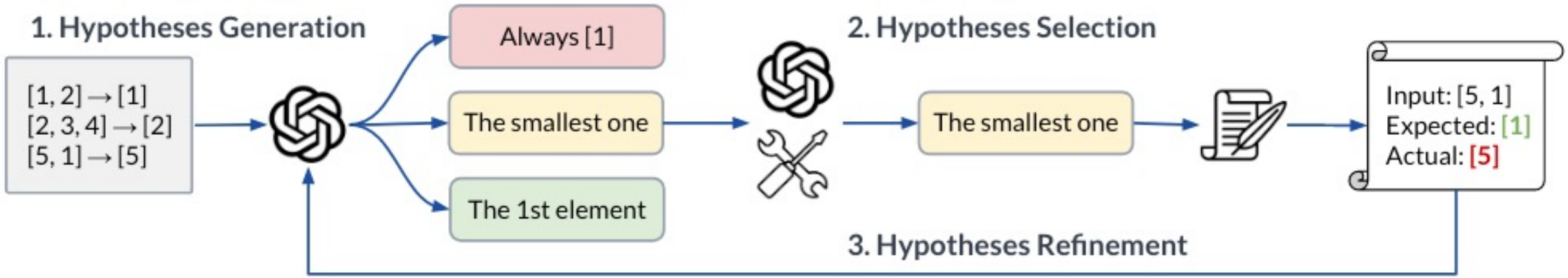
$$y_i' = I_\tau(h)(x_i).$$



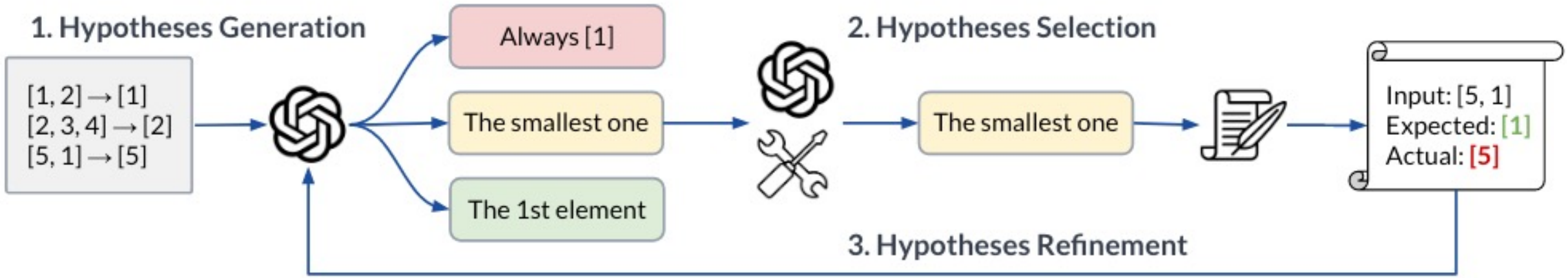
$$s(h, \mathcal{D}_\tau^s) = \frac{1}{|\mathcal{D}_\tau^s|} \sum_{(x,y)\in\mathcal{D}_\tau^s} \mathbb{1}\left[I_\tau(h)(x) = y\right]$$



$$h^{t^*} = \arg\max_{h' \in H^t} s(h', \mathcal{D}_\tau^s)$$



$$a_\tau = \frac{1}{|\mathcal{D}_\tau^u|} \sum_{(x,y)\in\mathcal{D}_\tau^u} \mathbb{1}\left[I_\tau(h)(x) = y\right]$$



Raw Accuracy

$$c = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} a_\tau$$

Task Accuracy

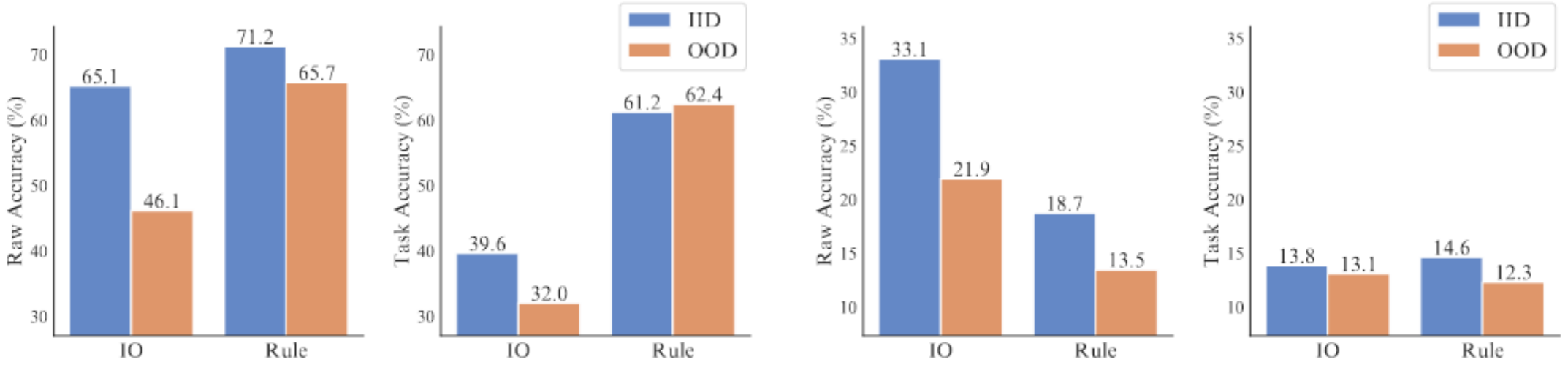
$$c_t = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathbb{1}\left[a_\tau = 1\right]$$

Main Results

Table 1: Iterative hypothesis refinement results. T refers to the maximum number of iterations. N refers to the number of candidate hypotheses per iteration.

	Raw Accuracy				Task Accuracy			
Method	ACRE	MiniSCAN	List Fns	MiniARC	ACRE	MiniSCAN	List Fns	MiniARC
IO	64.0	61.7	65.1	**33.1**	28.0	0.0	39.6	13.8
SC (N=5)	65.0	61.1	65.0	31.3	29.0	0.0	38.0	13.1
SR (T=3, N=5)	70.0	46.3	67.4	15.1	32.0	0.0	52.0	9.2
T=1, N=1	78.2	77.0	51.6	5.9	45.0	46.0	42.4	3.8
T=1, N=5	79.8	86.6	62.4	12.8	48.0	70.0	52.4	9.2
T=3, N=1	77.8	**98.2**	61.7	10.1	47.0	**95.0**	52.8	6.9
T=3, N=5	**82.5**	93.3	**71.2**	18.7	**59.0**	85.0	**61.2**	**14.6**

OOD Generalization and Interpretability



(a) List Functions (b) MiniARC

Figure 2: Results for IID and OOD examples. For OOD evaluations, we sample longer lists for List Functions and annotate larger grids for MiniARC. IO prompting generally experiences more significant performance degradation compared to rule prompting (i.e., iterative hypothesis refinement).

LMS STRUGGLE WITH APPLYING THEIR PROPOSED RULES

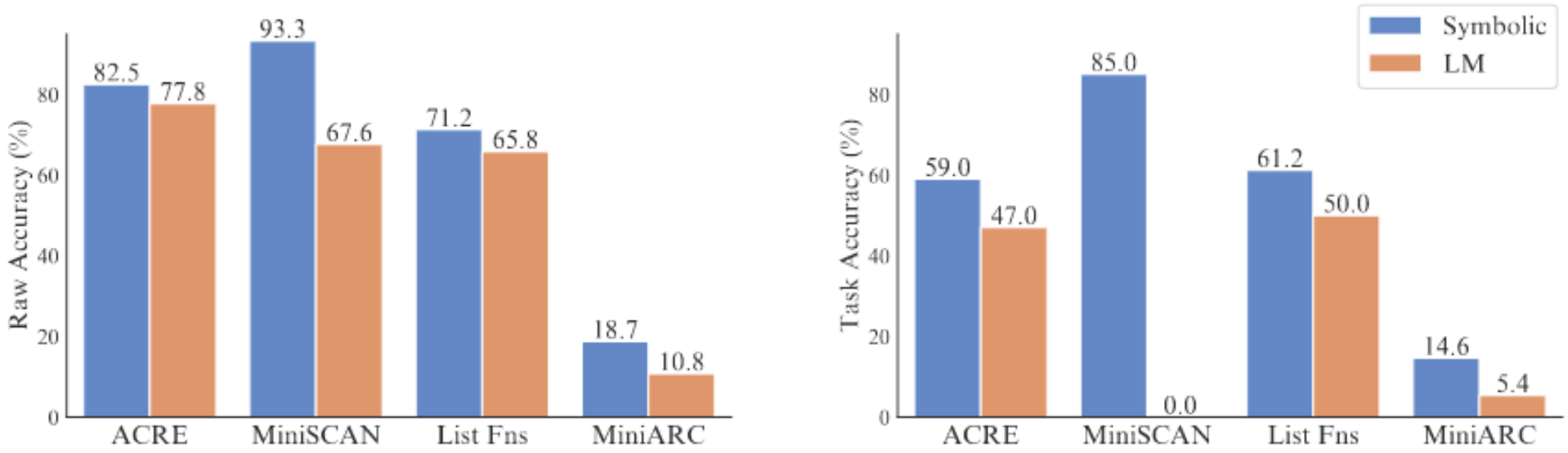


Figure 3: Raw accuracy (left) and task accuracy (right) when applying the LM's proposed rules using symbolic interpreters or the LM itself as the interpreter.

LMS ARE BRITTLE TO EXAMPLE PERTURBATIONS

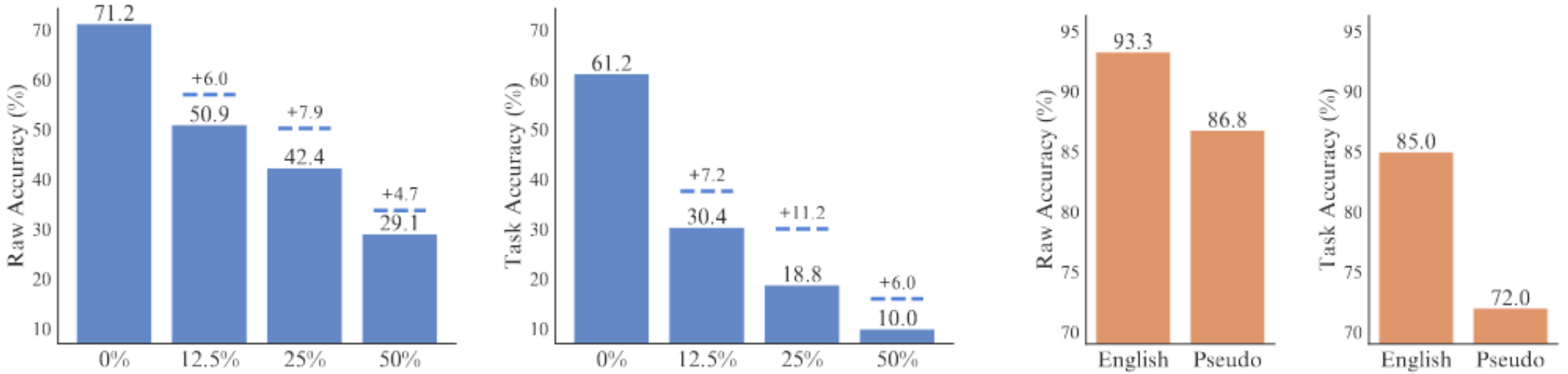


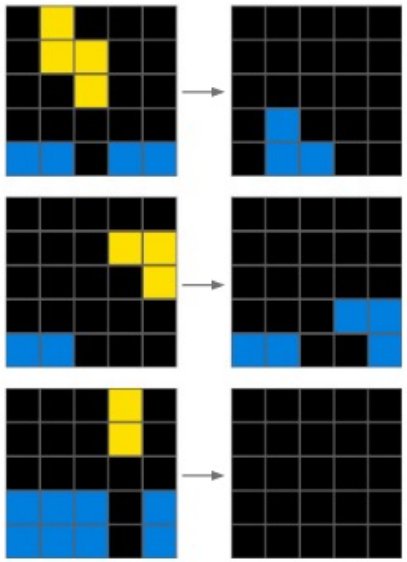
Figure 4: (a) Varying example noisiness by perturbing a certain percentage of exemplars on List Functions. Dashed lines refer to results where we explicitly instruct LMs to consider noisy examples. (b) Varying example familiarity by using English words or pseudo-words as outputs on MiniSCAN.

LM-INDUCED RULES VS. HUMAN-INDUCED RULES

Table 2: Comparison between LM-induced rules and human-induced rules on List Functions (top) and MiniARC (bottom). `0` maps to `black`, `1` maps to `blue`, and `4` maps to `yellow`.

Examples	LM-induced Rule	Human-induced Rules

`[97, 97, 97, 97]` → `[97, 97, 97]`
`[4, 4, 4]` → `[4, 4]`
`[33, 0, 4, 1, 2, 24, 66]` → `[]`
`[76, 42, 17, 76, 17]` → `[76, 17]`
...
	Remove the last occurrence of each unique number from the input list, but if a number appears more than twice, keep all instances except the last.	**Annotator 1**: Keep the order of the original list but only include integers that are duplicates rom earlier in the list.
		Annotator 2: Output only the repeated numbers. If a number is repeated n times then output only n-1 times.

	If an element in the input array is 4, replace it with 0. If the element is 1 and its left and right neighboring elements are 0, replace it with 1. If the element is 1 and positioned in the last row of the array, replace it with 1. In all other cases, replace the element with 0.	**Annotator 1**: Slide yellow down, if it completes a row, get rid of the row turn the remaining blocks blue with a 1. **Annotator 2**: Drop the object. If a full row is created, delete it, and drop remaining objects.

LM-INDUCED RULES VS. HUMAN-INDUCED RULES

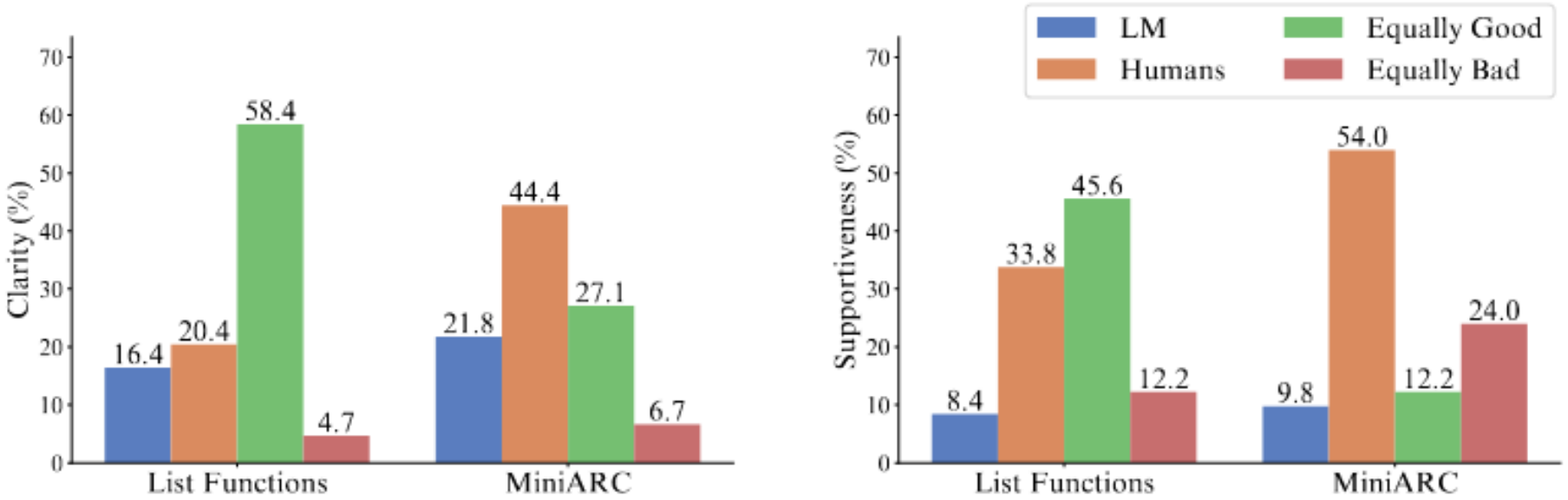


Figure 6: Comparisons of LM-induced rules versus human-induced rules in terms of clarity (left) and supportiveness (right).

Contributions

- Iterative Hypothesis Refinement 기법 처음 시도
- Phenomenal Hypotheses Proposer & Puzzling Inductive Reasoner 밝힘
- Iterative & feedback 을 통한 성능 개선

Limitations

- **Hyperparameters 종류가 적었음**
- **Real-world situation에 대한 것은 아님**
  - Decision making, Legal reasoning, Scientific reasoning

Future Works

- **다른 종류의 reasoning**에 대해 다뤄볼 수 있음
  - Commonsense reasoning
  - Everyday Reasoning Tasks : interpretating
- Intermediate process를 확인한 방법론을 활용한 **LM 이해 연구**
  - 큰 하나의 과정을 쪼개 보며 하나씩 개선
- Hypothesis propose를 잘하는 점을 활용한 **HAI 연구**
- Application은 잘 못 하는 점을 **개선하는 연구**
  - 최신 Dataset 을 cite한 논문 보며, 어디까지 되었는지 확인할 필요

행복한 하루 되세요