Transparent and Coherent Procedural Mistake Detection

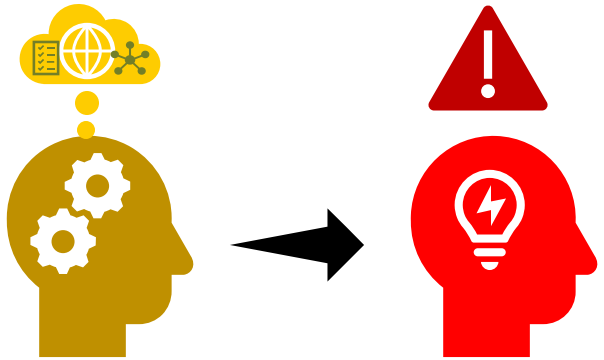
Shane Storks, Itamar Bar-Yossef, Yayuan Li, Zheyuan Zhang,
Jason J. Corso, & Joyce Chai

Computer Science & Engineering
University of Michigan – Ann Arbor

EMNLP 2025 Long Paper

Motivation & Problem Statement

- Automated task guidance has recently seen progress due to advances in foundational VLMs, which can robustly interpret visual scenes and communicate through language
- Procedural mistake detection (PMD) is a difficult problem…
  - VLMs have not achieved viable performance in the wild
  - Typically formulated as classification, limiting understanding of decisions
- How coherently can recent VLMs reason (explicitly) about mistakes?

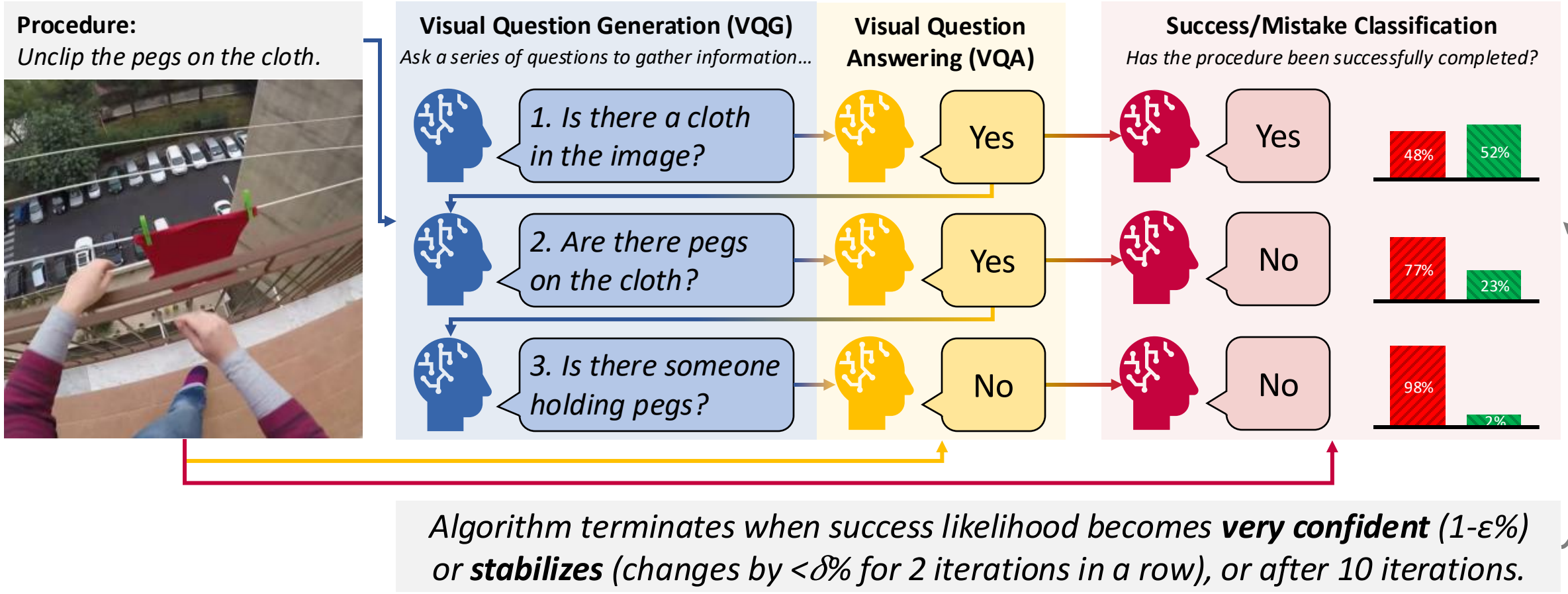


Task: *Unclip the pegs on the cloth.*
Has this been completed?

Y. Du, K. Konyushkova, M. Denil, et al. (2023). Vision-Language Models as Success Detectors. *CoLLAs, PMLR 232:120-136.*
Y. Bao, K. Yu, Y. Zhang, S. Storks, I. Bar-Yossef, A. de la Iglesia, M. Su, X.L. Zheng, & J. Chai. (2023). Can Foundation Models Watch, Talk, and Guide You Step by Step to Make a Cake? *Findings of EMNLP 2023.*
T. Kojima, S.S. Gu, M. Reid, et al. (2022). Large Language Models are Zero-Shot Reasoners. *NeurIPS 2022.*

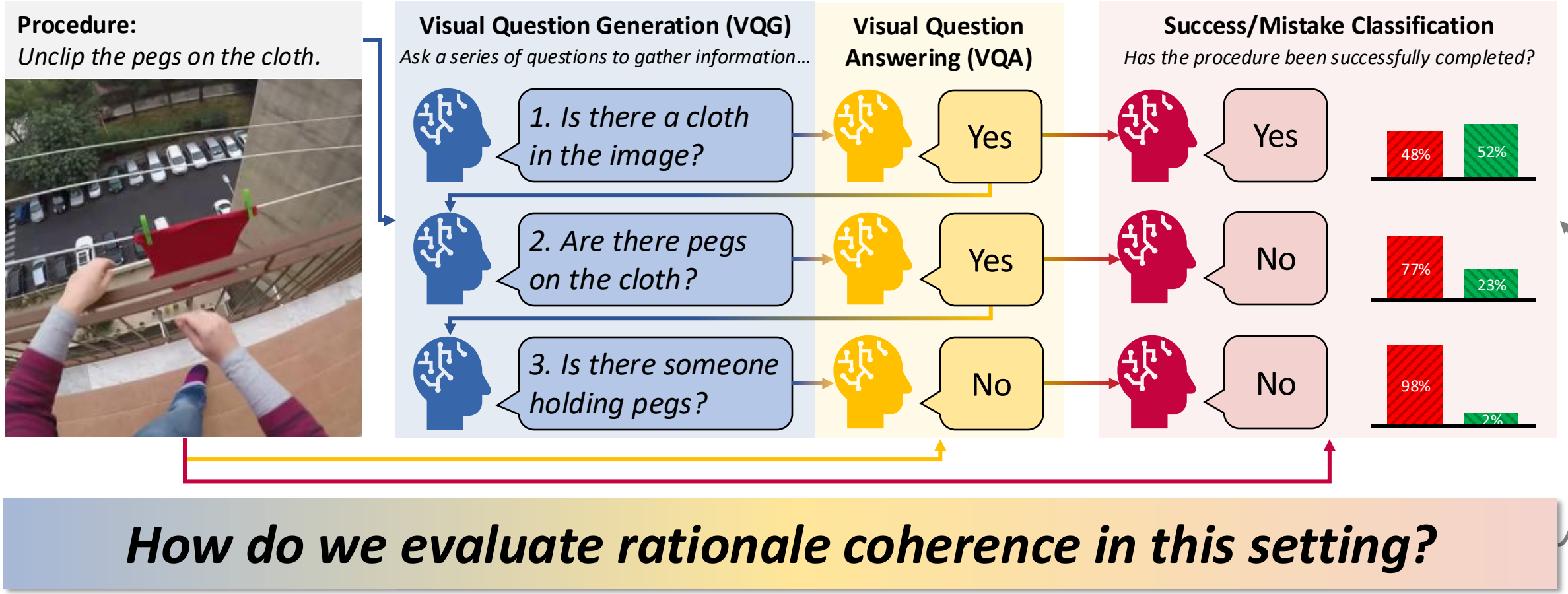
Coherent PMD through Self-Dialog

Go beyond classification and formulate PMD as an explanatory self-dialog:



Coherent PMD through Self-Dialog

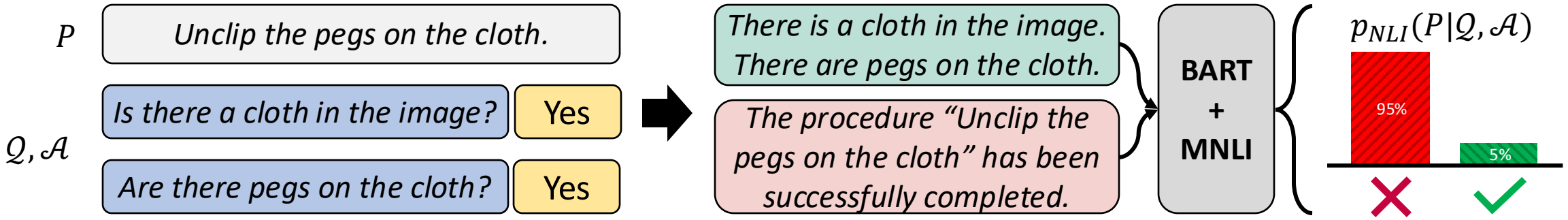
Go beyond classification and formulate PMD as an explanatory self-dialog:



How do we evaluate rationale coherence in this setting?

Using NLI Model to Judge Success

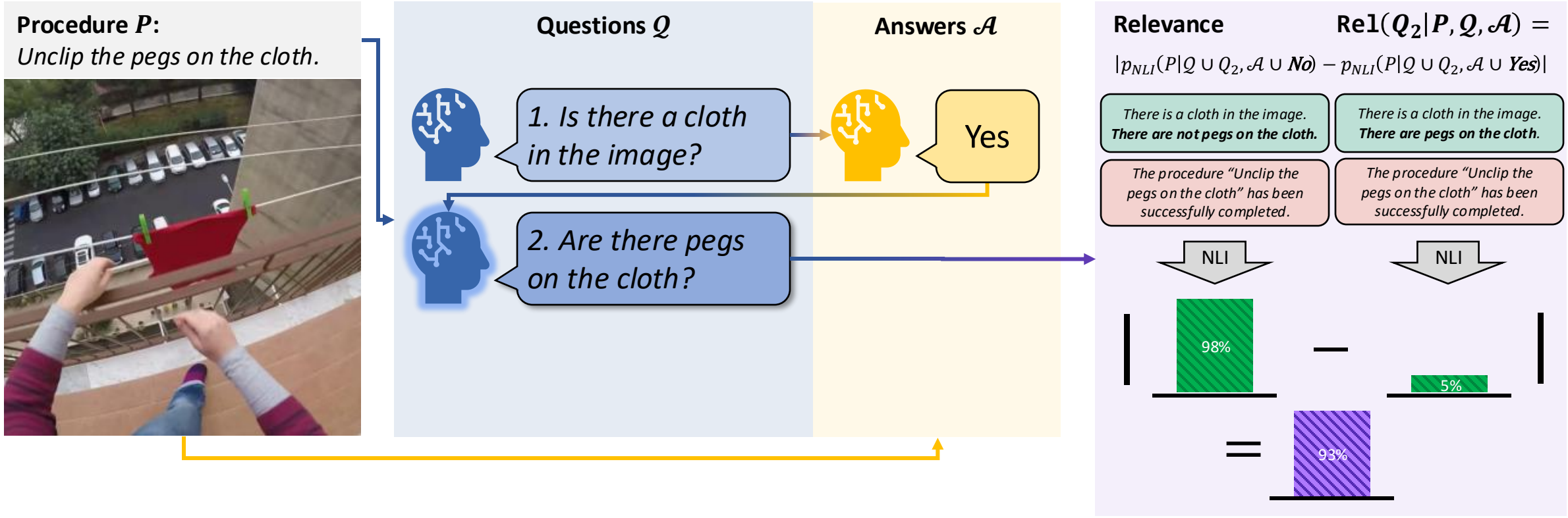
- Explore automated reference-free metrics instead of annotating benchmark data:
  - There are many valid possibilities in explaining PMD decisions
  - Downstream practical application requires easy auditing of system behaviors
- Recent work has leveraged smaller, specialized LMs fine-tuned for natural language inference (NLI) to evaluate and improve systems for various problems
  - **Proposal:** Use BART fine-tuned on MultiNLI to judge procedure success given questions and answers



N. Dziri, E. Kamalloo, K. Mathewson, & O. Zaiane. (2019). Evaluating Coherence in Dialogue Systems using Entailment. *NAACL HLT 2019.*
P. Roit, J. Ferret, L. Shani, et al. (2023). Factually Consistent Summarization Reinforcement Learning with Textual Entailment Feedback. *ACL 2023.*
T. Srinivasan, J. Hessel, T. Gupta, et al. (2024). Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning. *Findings of ACL 2024.*
M. Lewis, Y. Liu, N. Goyal, et al. (2020). BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. *ACL 2020.*
A. Williams, N. Nangia, & S. Bowman. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *NAACL HLT 2018.*

Metric: Relevance of a Question

Measures how much impact a generated question's answer might have on success.



T. Srinivasan, J. Hessel, T. Gupta, et al. (2024). Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning. *Findings of ACL 2024.*

Metric: Informativeness of an Answer

Measures how much information an answer gives us about the success of the procedure.

Procedure P:
Unclip the pegs on the cloth.

Questions $\mathcal{Q}$

1. Is there a cloth in the image?

2. Are there pegs on the cloth?

Answers $\mathcal{A}$

Yes

Yes

Informativeness $\mathbf{Inf}(A_2|Q_2, P, \mathcal{Q}, \mathcal{A}) =$

$$1 - H(p_{NLI}(P|\mathcal{Q}, \mathcal{A} \cup A_2))$$

There is a cloth in the image. ***There are pegs on the cloth.***

The procedure "Unclip the pegs on the cloth" has been successfully completed.

NLI

$$1 - H(\text{5\%}) = .71$$

$$= \text{.71}$$

Reference-Adjusted Informativeness:
If p_{NLI} disagrees with ground truth success/mistake label, then negate.

Ego4D for Procedural Mistake Detection

Procedure: *Fold the cloth*

Success | **Mistake** *(Incomplete)* | **Mistake** *(Wrong Verb)* | **Mistake** *(Wrong Noun)* | **Mistake** *(Wrong Verb & Noun)*

Type	Train	(Sample)	Val.	(Sample)	Test	(Sample)	Total
Success	42.0k	5.00k	13.1k	250	18.1k	1.00k	73.1k
Mistake	99.4k	5.00k	25.4k	250	34,182	1.00k	159k
(Incomplete)	15.1k	755	4.91k	51	6.55k	194	26.5k
(Wrong V)	11.8k	604	2.69k	31	3.75k	108	18.2k
(Wrong N)	36.4k	1.85k	8.91k	87	11.8k	344	57.2k
(Wrong V&N)	36.1k	1.79k	8.91k	81	12,047	354	57.1k

Type	Train	(Sample)	Val.	(Sample)	Test	(Sample)	All
Verbs	83	80	77	55	78	71	83
Nouns	440	372	365	151	390	257	487
V-N Pairs	3,976	2,050	2,185	326	2,658	833	5,363

Sample 10,000 training, 500 validation, and 2,000 testing examples
(even split of success and mistake cases)

K. Grauman, A. Westbury, E. Byrne, et al. (2022). Ego4D: Around the World in 3,000 Hours of Egocentric Video. *CVPR 2022*.

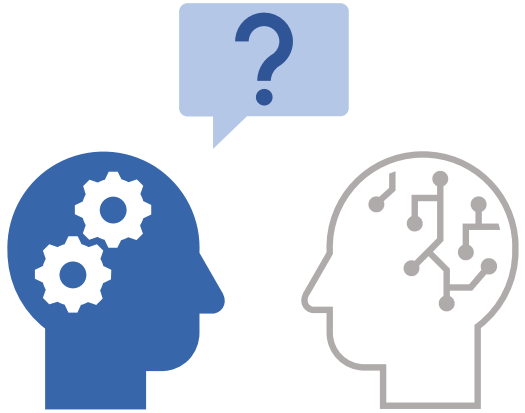
Evaluation Metrics

- Success classification **accuracy**
- Coherence:
  - Average **relevance** of questions in self-dialog
  - Maximum reference-adjusted **informativeness** achieved in self-dialog
- Efficiency:
  - **Number of iterations** of self-dialog
  - **Information gain** in success probability from the self-dialog (communication efficiency)

Research Questions

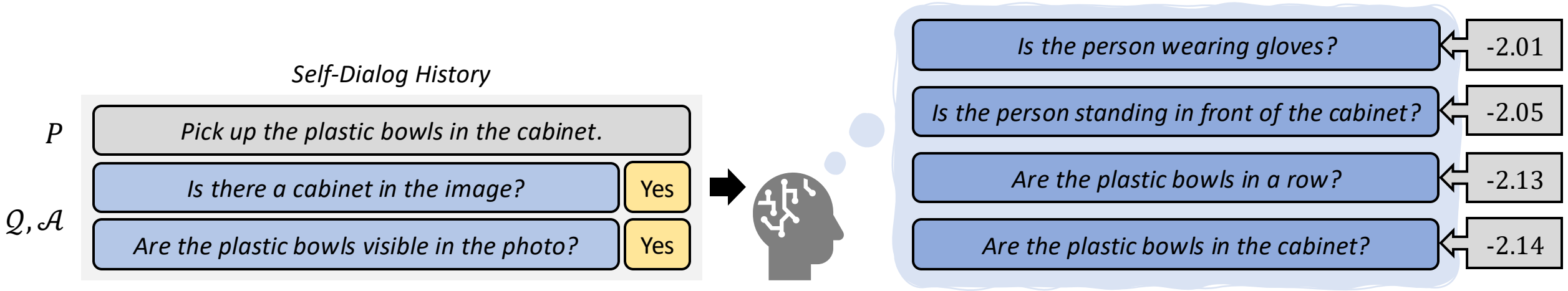
Can use these metrics and data to understand how various interventions impact PMD.

1. How does prioritizing coherence in **question selection** impact performance?
   - Rank candidate questions in beam search by relevance and maximum potential informativeness
   - Utilize in-context learning from human-written examples to augment candidate pool
2. How does prioritizing coherence in **question generation** impact performance?
   - Use DPO to preference-optimize VLMs based on question relevance and potential informativeness



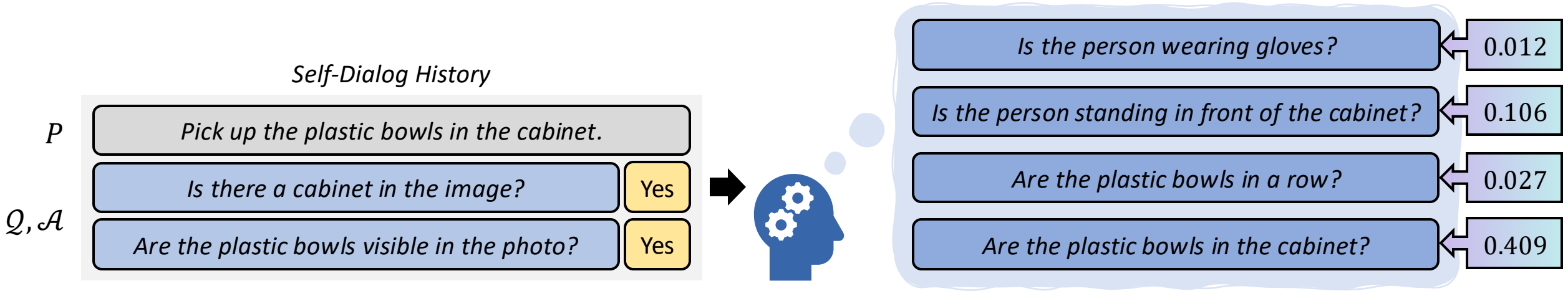
R. Rafailov, A. Sharma, E. Mitchell, & C.D. Manning. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 37*.

Coherence-Based Question Selection



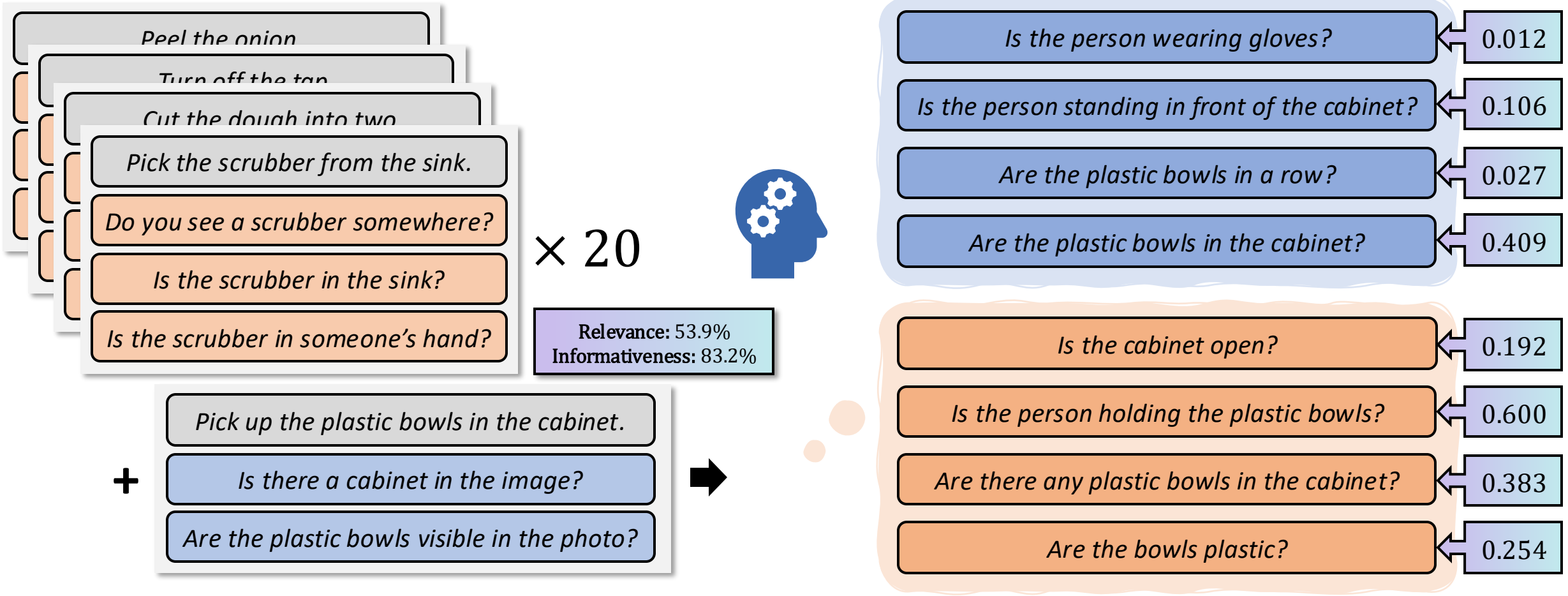
Coherence-Based Question Selection

Idea: Re-rank candidate questions from beam search using relevance and informativeness.



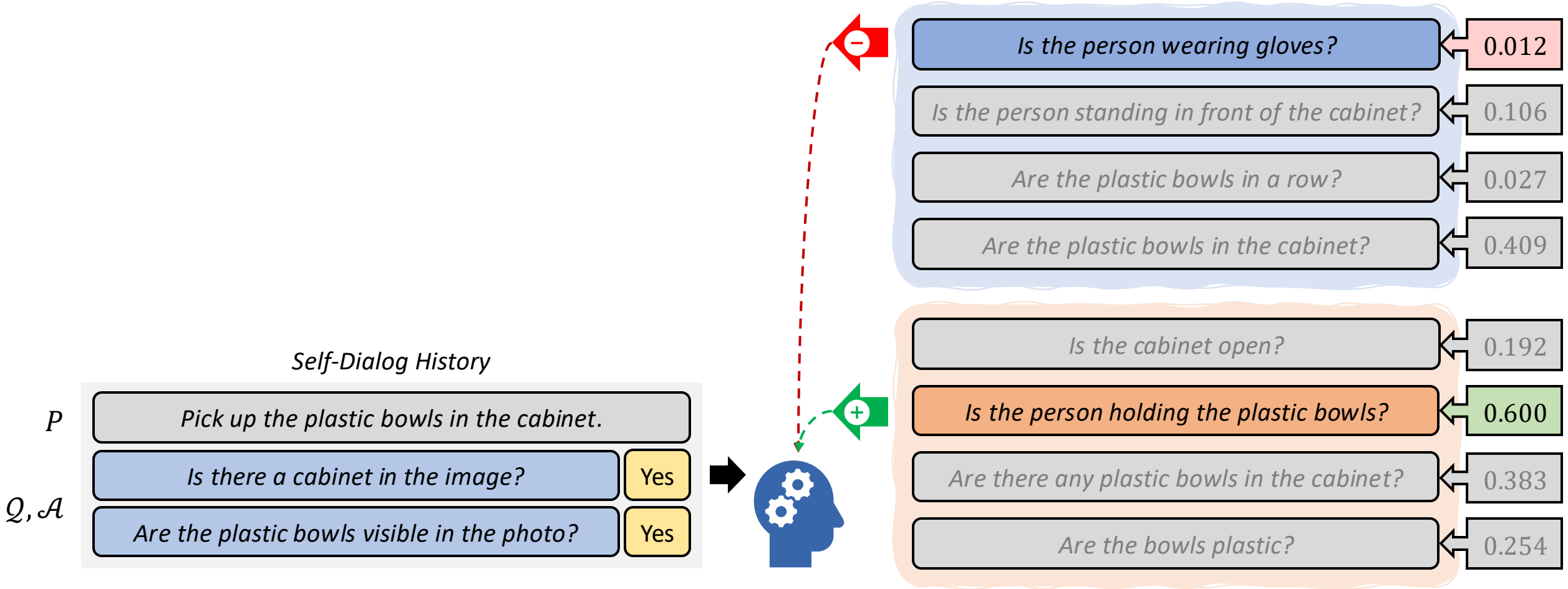
In-Context Learning (ICL) for VQG

Idea: Generate more candidates using in-context learning from human-written questions.



Coherence-Based Fine-Tuning

Idea: Fine-tune VLM to consider coherent questions as more likely than incoherent ones.



Experimental Results

Coherence-based selection and ICL **improve accuracy, coherence, and information gain**.

InstructBLIP

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	63.5	17.5	.224	**2.84**	.263
L	✓	65.2	13.9	.340	4.71	.358
C	✗	64.6	25.5	.281	3.46	.293
C	✓	**66.6**	**35.2**	**.359**	3.47	**.359**

LLaVA

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	60.7	40.3	.259	3.25	.435
L	✓	61.8	36.5	.272	3.34	.429
C	✗	61.4	66.5	.321	**3.06**	.540
C	✓	**67.8**	**75.5**	**.464**	3.46	**.663**

Llama 3

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	61.0	16.5	.275	4.70	.223
L	✓	59.1	15.9	.317	6.51	.256
C	✗	60.2	25.2	.341	6.35	.264
C	✓	**61.7**	**52.5**	**.436**	**3.59**	**.379**

InstructBLIP (7B): W. Dai, J. Li, D. Li, et al. (2023). InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *NeurIPS 2023.*
LLaVA 1.5 (7B): H. Liu, C. Li, Q. Wu, & Y.J. Lee. (2023). Visual Instruction Tuning. *NeurIPS 2023.*
Llama 3 with Vision (11B): A. Grattafiori, A. Dubey, A. Jauhri, et al. (2024). The Llama 3 Herd of Models. *arXiv: 2407.21763.*
GPT-4o: OpenAI. (2024). GPT-4o system card. *arXiv: 2410.21276.*

Experimental Results

Coherence-based fine-tuning **improves relevance and efficiency** at a **cost of accuracy and informativeness**.

InstructBLIP

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	63.5	17.5	.224	**2.84**	.263
L	✓	65.2	13.9	.340	4.71	.358
C	✗	64.6	25.5	.281	3.46	.293
C	✓	**66.6**	**35.2**	**.359**	3.47	**.359**

LLaVA

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	60.7	40.3	.259	3.25	.435
L	✓	61.8	36.5	.272	3.34	.429
C	✗	61.4	66.5	.321	**3.06**	.540
C	✓	**67.8**	**75.5**	**.464**	3.46	**.663**

Llama 3

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	61.0	16.5	.275	4.70	.223
L	✓	59.1	15.9	.317	6.51	.256
C	✗	60.2	25.2	.341	6.35	.264
C	✓	**61.7**	**52.5**	**.436**	**3.59**	**.379**

LLaVA + DPO

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	62.2	75.7	.318	2.33	.617
L	✓	63.7	58.5	.330	2.67	.548
C	✗	62.3	92.2	**.340**	2.06	.719
C	✓	**64.2**	**95.0**	.304	**1.81**	**.742**

”Put some soil around the tomato seedling with the gardening trowel in your hand.”

“Is the soil placed around the seedling with the trowel in the person’s hand?”

GPT-4o

Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
55.4	54.0	.220	1.84	.793

InstructBLIP (7B): W. Dai, J. Li, D. Li, et al. (2023). InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *NeurIPS 2023.*
LLaVA 1.5 (7B): H. Liu, C. Li, Q. Wu, & Y.J. Lee. (2023). Visual Instruction Tuning. *NeurIPS 2023.*
Llama 3 with Vision (11B): A. Grattafiori, A. Dubey, A. Jauhri, et al. (2024). The Llama 3 Herd of Models. *arXiv: 2407.21763.*
GPT-4o: OpenAI. (2024). GPT-4o system card. *arXiv: 2410.21276.*

Experimental Results

Coherence-based fine-tuning **improves relevance and efficiency** at a **cost of accuracy and informativeness**.

InstructBLIP

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	63.5	17.5	.224	**2.84**	.263
L	✓	65.2	13.9	.340	4.71	.358
C	✗	64.6	25.5	.281	3.46	.293
C	✓	**66.6**	**35.2**	**.359**	3.47	**.359**
L	✗	60.7	40.3	.259	3.25	.435
L	✓	61.8	36.5	.272	3.34	.429
C	✗	61.4	66.5	.321	**3.06**	.540
C	✓	**67.8**	**75.5**	**.464**	3.46	**.663**

Llama 3

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	61.0	16.5	.275	4.70	.223
L	✓	59.1	15.9	.317	6.51	.256
C	✗	60.2	25.2	.341	6.35	.264
C	✓	**61.7**	**52.5**	**.436**	**3.59**	**.379**

LLaVA + DPO

Rank	ICL	Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
L	✗	62.2	75.7	.318	2.33	.617
L	✓	63.7	58.5	.330	2.67	.548
C	✗	62.3	92.2	**.340**	2.06	.719
C	✓	**64.2**	**95.0**	.304	**1.81**	**.742**

Coherence metrics enable us to visualize and audit system behaviors!

"Put some soil around the tomato seedling with the gardening trowel in your hand."

"Is the soil placed around the seedling with the trowel in the person's hand?"

GPT-4o

Acc. ↑	Rel. ↑	Inf. ↑	# Iter. ↓	I. Gain ↑
55.4	54.0	.220	1.84	.793

InstructBLIP (7B): W. Dai, J. Li, D. Li, et al. (2023). InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *NeurIPS 2023.*
LLaVA 1.5 (7B): H. Liu, C. Li, Q. Wu, & Y.J. Lee. (2023). Visual Instruction Tuning. *NeurIPS 2023.*
Llama 3 with Vision (11B): A. Grattafiori, A. Dubey, A. Jauhri, et al. (2024). The Llama 3 Herd of Models. *arXiv: 2407.21763.*
GPT-4o: OpenAI. (2024). GPT-4o system card. *arXiv: 2410.21276.*

Visualizing VLM Behaviors with Coherence Metrics

Pick up a sink brush from the kitchen slab. Success

Is the sink brush in the person's hands? Yes

LLaVA
+ Coherence-Based Ranking

Correct and Coherent

Coherent but Incorrect

VQA Uncertainty

Informative As to Irrelevant Qs

Correct but Incoherent

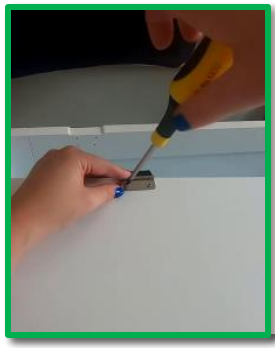
Incorrect and Incoherent

Put the trowel in a bin. Mistake

Is the trowel in a bin? No

Put the bottle in the cabinet. Success

Is the bottle in the cabinet? Yes



Tighten the screw. Mistake

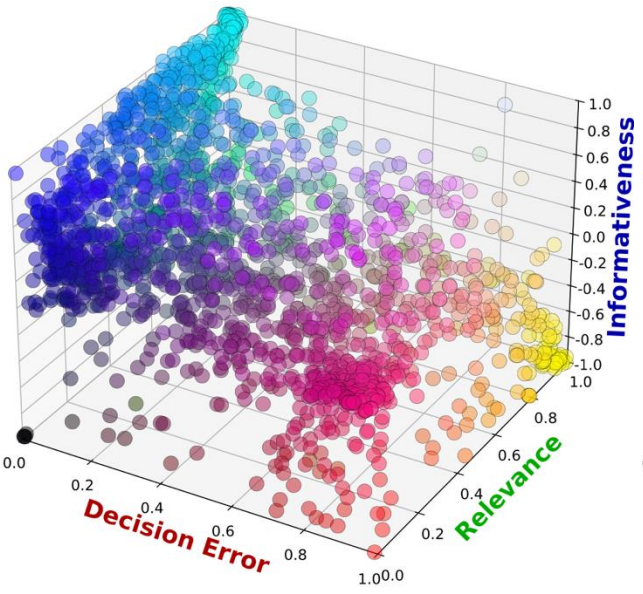
Is the person wearing gloves? No

Is the person wearing protective gear? No

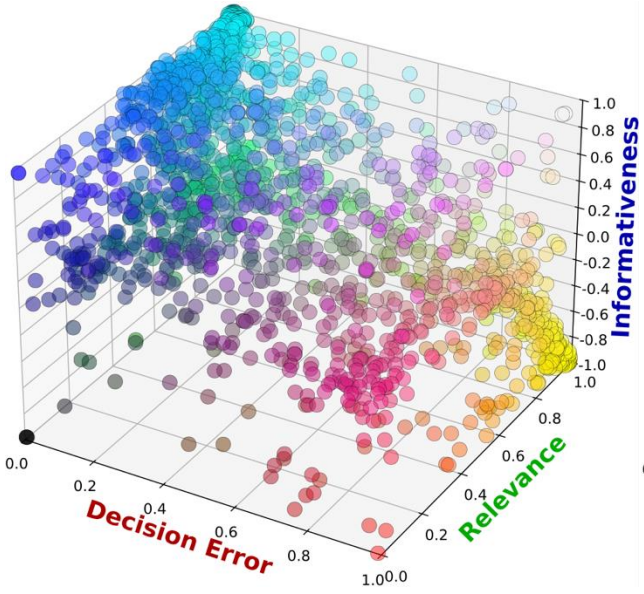
Is the person wearing a mask? No

Visualizing VLM Behaviors with Coherence Metrics

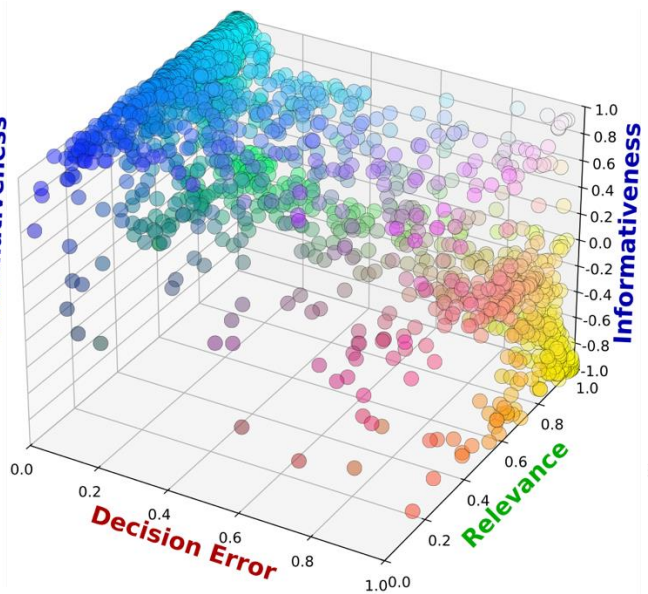
LLaVA



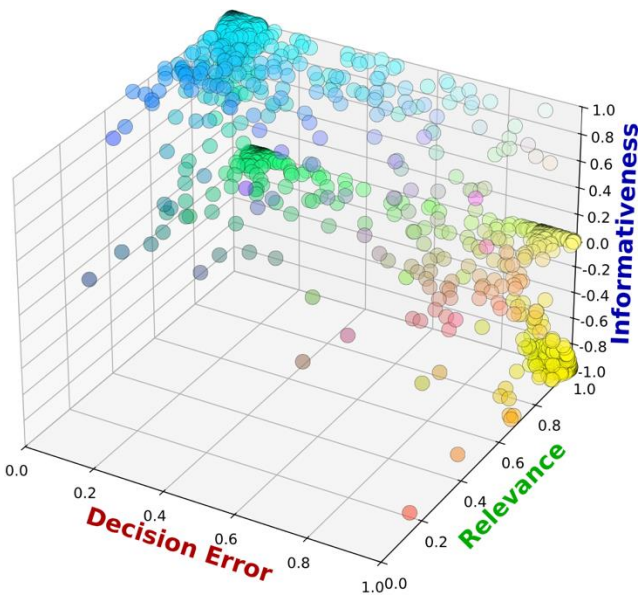
+ Coherence-Based Ranking



+ In-Context Learning



+ Coherence-Based Fine-Tuning



Conclusion

- Coherent PMD extends mistake detection in VLMs to require visual self-dialog rationales
  - Relevance and informativeness metrics provide global and local insights into coherence of binary detection decisions
- Findings:
  - VLMs do not generate coherent rationales for PMD off-the-shelf
  - Their coherence, accuracy, and efficiency can be improved through coherence-based selection and fine-tuning for generating questions
  - But there are trade-offs!

arxiv.org/abs/2412.11927

github.com/sled-group/Transparent-Coherent-PMD

www.shanestorks.com

sled.eecs.umich.edu