

# *Effective Human-AI Teams via Learned Natural Language Rules and Onboarding*

NeurIPS 2023 Spotlight

[arxiv.org/pdf/2311.01007.pdf](https://arxiv.org/pdf/2311.01007.pdf)

[github.com/clinicalml/onboarding\\_human\\_ai](https://github.com/clinicalml/onboarding_human_ai)



**Hussein Mozannar**  
MIT



Jimin J Lee  
MIT



Dennis Wei  
IBM



Prasanna Sattigeri  
IBM



Subhro Das  
IBM



David Sontag  
MIT

# Humans Interacting with AI

**Setting:** you are learning to code, so you ask an AI for help

- Should you trust the AI's code?
- Try to solve it on your own?
- Edit the AI code to make it work?



You

Please solve this coding problem in Python: given two integers a and b, return the sum of the two integers without using the operators + and -, write concise code without examples



ChatGPT

You can solve this problem concisely using Python by employing bit manipulation. Here's a concise implementation:

```
python 📄 Copy code  
  
def get_sum(a, b):  
    mask = 0xFFFFFFFF # A mask to limit the result to 32 bits  
    while b:  
        a, b = (a ^ b) & mask, ((a & b) << 1) & mask  
    return a if a <= 0x7FFFFFFF else ~(a ^ mask)  
  
# Example usage:  
# a = 3  
# b = 2  
# result = get_sum(a, b)  
# print(result)
```

This code uses a loop to simulate the addition process without using the '+' or '-' operators. It ensures the result is a 32-bit integer by applying a mask and handles negative results correctly. Uncomment the example usage to test the function with specific values.



# Humans Interacting with AI

**Setting:** you are driving but it's low visibility. AI dashboard is helping you see.

- Do you stop since AI is telling you it's a red light?
- Pull over until visibility improves?

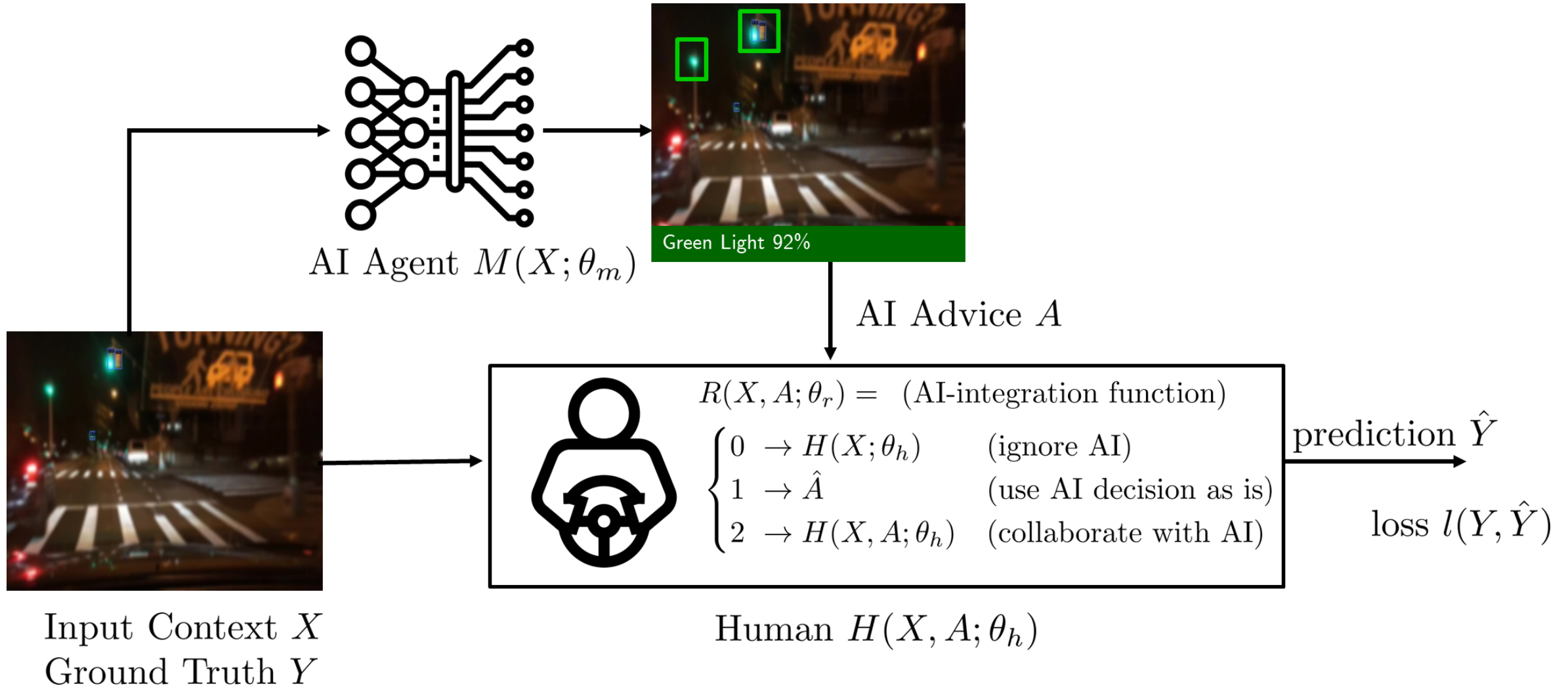


## **Central Problem:**

How do you know when to:

- rely on the AI,
- collaborate with the AI or
- ignore its suggestion?

# AI-Assisted Decision Making




## **Proposed IntegrAI Approach:**

- Learn from data the optimal AI-integration function
- Teach the human the optimal way to collaborate

# Step 1: Human Data Collection

Understand Human performance and prior reliance on AI



Is there a Traffic Light?

Yes

No

AI Predicts:  
Yes, score 0.9

Use AI Answer



$D$  (dataset)

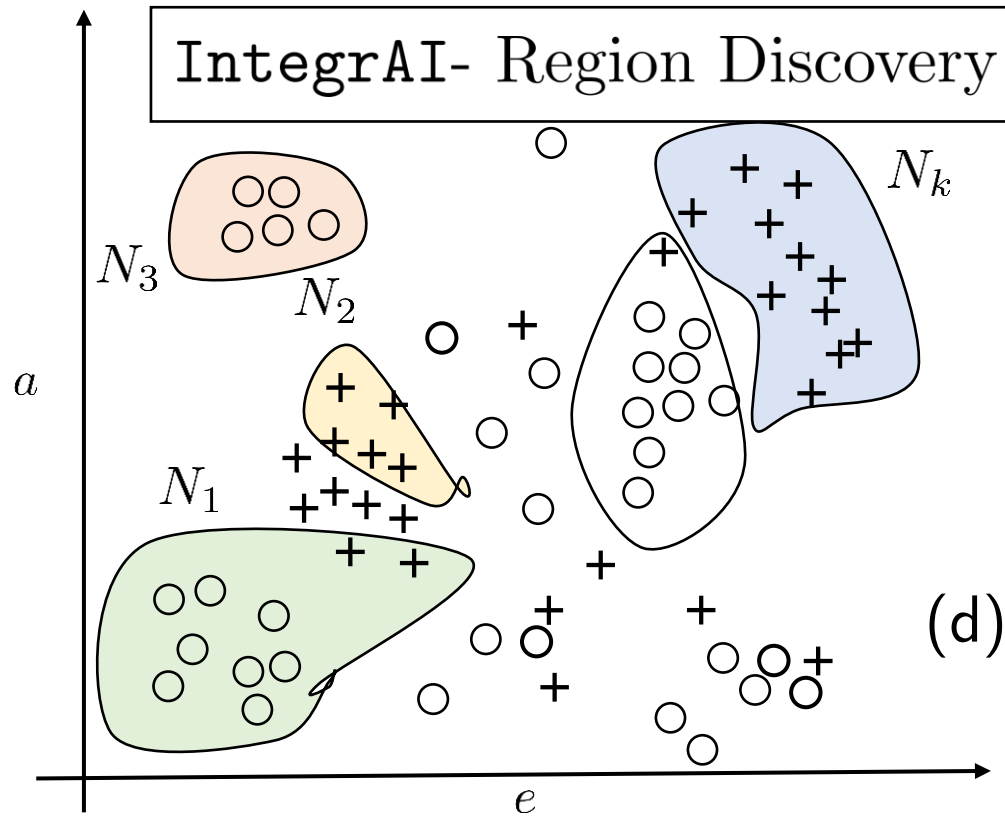
- $x_i$  (task),  $y_i$  (answer)
- $a_i$  (AI),  $h_i$  (human)
- $t_i$  (text description),  $e_i$  (embedding)
- $r_i$  (used AI)

Assume the human has a prior integration function  $R_0$ , our goal is:

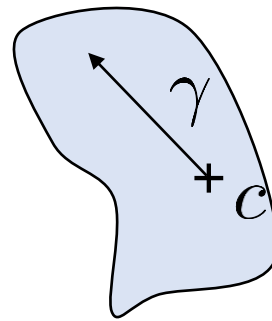
- Understand human performance without AI  $\mathbf{H(X)}$
- Understand the **human** prior  $\mathbf{R_0}$

# Step 2: Discover Regions

Find regions of data where human collaborates incorrectly with AI



Goal of teaching is to correct the prior:  
e.g. if human trusted the AI, in a region where AI is incorrect, find that region and inform the human!



Region is denoted by a **center**  $c$  and a **radius**  $\gamma$



# Step 2: Discover Regions

Find regions of data where human collaborates incorrectly with AI

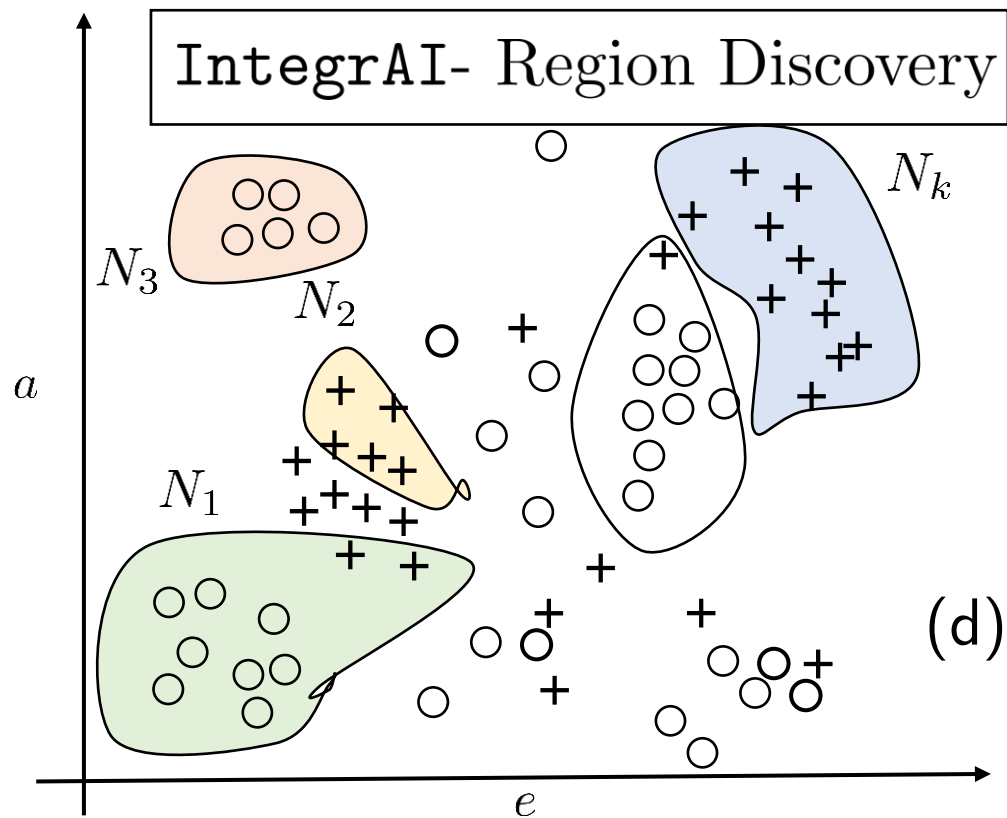
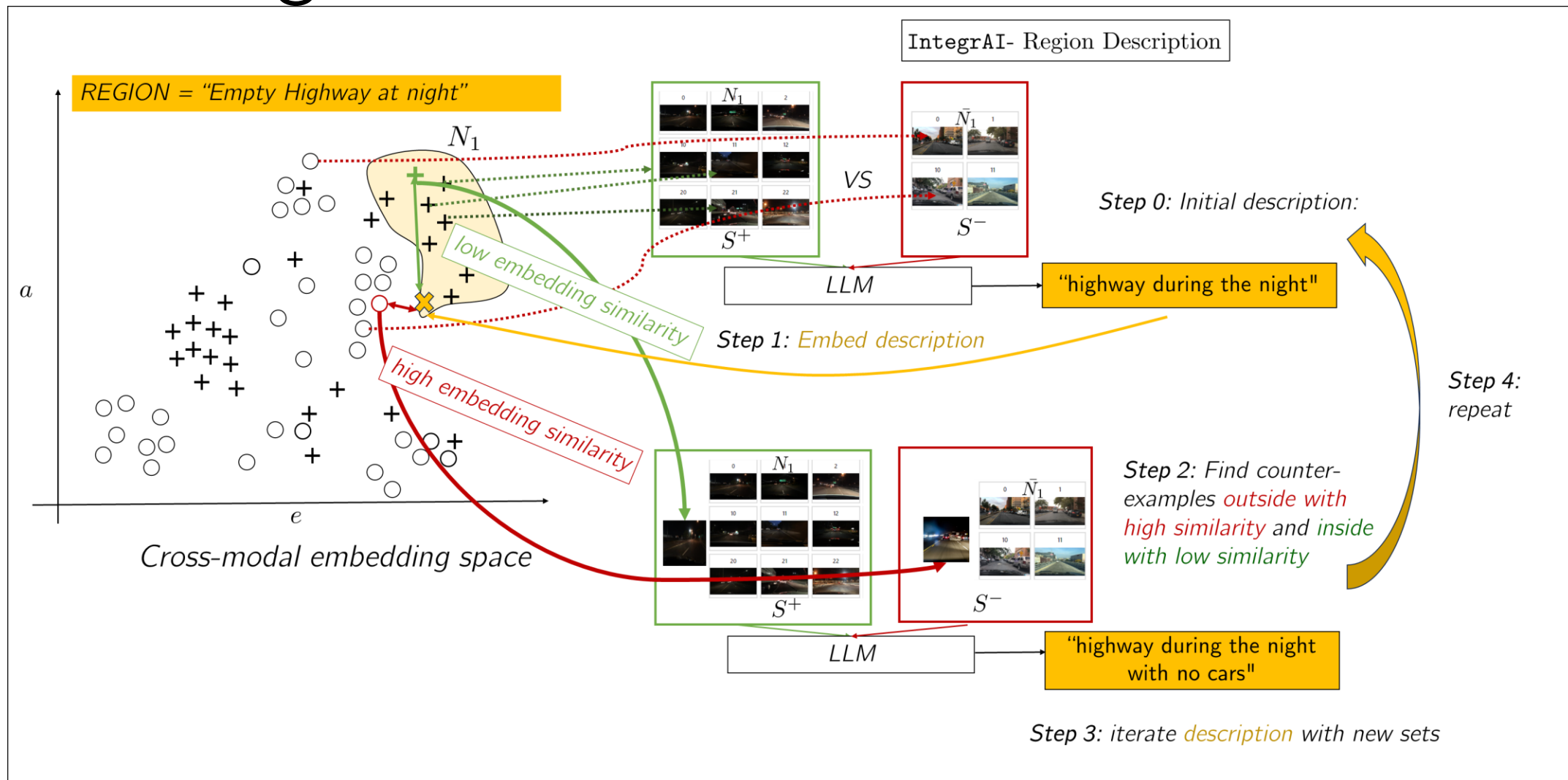


Table 3: Error ( $\downarrow$ ) on the test set (in %) of the human-AI system when following integrators resulting from different region discovery methods with 10 regions on the different non-synthetic datasets.

	BDD	MMLU	DynaSent	MS-COCO
IntegrAI (ours)	<b>17.8 <math>\pm</math> 0.2</b>	<b>45.3 <math>\pm</math> 0.3</b>	20.2 $\pm$ 0.3	<b>22.6 <math>\pm</math> 0.4</b>
DOMINO [EVS <sup>+</sup> 21]	18.9 $\pm$ 0.4	48.1 $\pm$ 0.2	20.0 $\pm$ 0.2	22.7 $\pm$ 0.4
K-MEANS [RLC <sup>+</sup> 22]	19.0 $\pm$ 0.5	<b>45.3 <math>\pm</math> 0.3</b>	20.0 $\pm$ 0.2	23.2 $\pm$ 0.1
DoubleGreedy [MSS22]	18.9 $\pm$ 0.1	46.1 $\pm$ 0.6	20.0 $\pm$ 0.2	23.8 $\pm$ 0.4

# Step 3: Describe Regions

## Describe regions in text with an LLM



# Step 3: Describe Regions

## Describe regions in text with an LLM



Table 5: Evaluation of our region description algorithm (Algorithm 2) on selected subsets of MS-COCO where the different algorithms try to describe a set of images that all contain a given object. For example, a region may be defined by images containing the object “apple”. Then we compare the descriptions resulting from the different algorithms to the description “apple”.

	best-caption	SEAL	IntegrAI ( $S^- = \emptyset$ )	IntegrAI (m=0)	IntegrAI (m=5)	IntegrAI (m=10)
METEOR	$12.9 \pm 1.9$	$9.16 \pm 1.89$	$24.3 \pm 3.3$	$25.4 \pm 3.2$	<b><math>26.1 \pm 3.3</math></b>	$25.4 \pm 3.3$
sent-sim	$39.8 \pm 1.9$	$44.1 \pm 2.5$	$65.1 \pm 3.2$	$67.0 \pm 3.1$	$66.0 \pm 3.2$	<b><math>68.0 \pm 3.3</math></b>
ROUGE	$5.81 \pm 1.2$	$0.0 \pm 0.0$	$25.6 \pm 4.9$	$32.6 \pm 5.4$	$27.9 \pm 5.1$	<b><math>35.6 \pm 5.5</math></b>
SPICE	$12.7 \pm 1.9$	$7.53 \pm 2.3$	$41.1 \pm 5.8$	$43.8 \pm 5.8$	<b><math>45.2 \pm 5.8</math></b>	<b><math>45.2 \pm 5.8</math></b>



Step 3: iterate description with new sets

# Step 4: Teach Regions to Human

In onboarding stage, teach human each region

## Region 1:

Ignore AI when it's a  
"highway during the night"

## Region 2:

Use AI when it's a  
"city during the day"

...

## Region k:



Teaching Phase: for each region

- **Step 1:** human predicts on example in region
- **Step 2:** human receives feedback about their correctness and the AIs correctness
- **Step 3:** (from example to region) human is shown the region description and a gallery of examples from the region  
(any example the human gets wrong are re-shown)

# User Study Evaluation

Fast R-CNN  
Object detector

12/20

Hide AI box

Is there a Traffic Light?

Recommendation: Rely on AI  
Region: mostly highways at night and during clear weather, with a mix of cars, trucks, buses, traffic lights, "

Yes

No

AI Predicts: Yes with score 0.9

Use AI Answer

Enables us to understand which action human took

Task: Given a blurry image, human must figure out if the light is red or green (proxy for moving forward)

6/15

Question: If a child inherits a physical trait from his father he will likely receive

A: tooth gap  
B: favorite color  
C: sense of fun  
D: liking dogs

Recommendation: Ignore AI Answer  
AI often wrong when various questions on different topics such as algebra, biology, public relations, and statistics [...]

A

B

C

D

AI Predicts: A  
Tooth gap because physical traits, such as the gap between teeth, are primarily determined by genetics and can be passed down from parents to children

Use AI Answer

Task: Given a multiple-choice question, human must figure out the right answer.

# User Study Evaluation

**On BDD study:**

Accuracy: without onboarding 77.2, with onboarding 82.6 (significant increase)

# User Study Evaluation

## **On BDD study:**

Accuracy: without onboarding 77.2, with onboarding 82.6 (significant increase)

## **On MMLU study:**

Accuracy: without onboarding 75.0, with onboarding 74.4, why?

# User Study Evaluation

## **On BDD study:**

Accuracy: without onboarding 77.2, with onboarding 82.6 (significant increase)

## **On MMLU study:**

Accuracy: without onboarding 75.0, with onboarding 74.4, why?

ChatGPT explanations often inform human whether to use its answer:

- 'Unfortunately, the options provided do not provide a clear answer to [...]
- 'The answer cannot be provided with the given information [...]