Toloka

Beyond Buzzwords: Practical Approaches to Mitigating Biases in the Age of LLMs

Dr. ir. Ujwal Gadiraju
@UJLAW

What is this?

Does this image represent a **TYPICAL** [croissant] ?

Does this image represent a TYPICAL [toggle button]?

A Typical muffin?

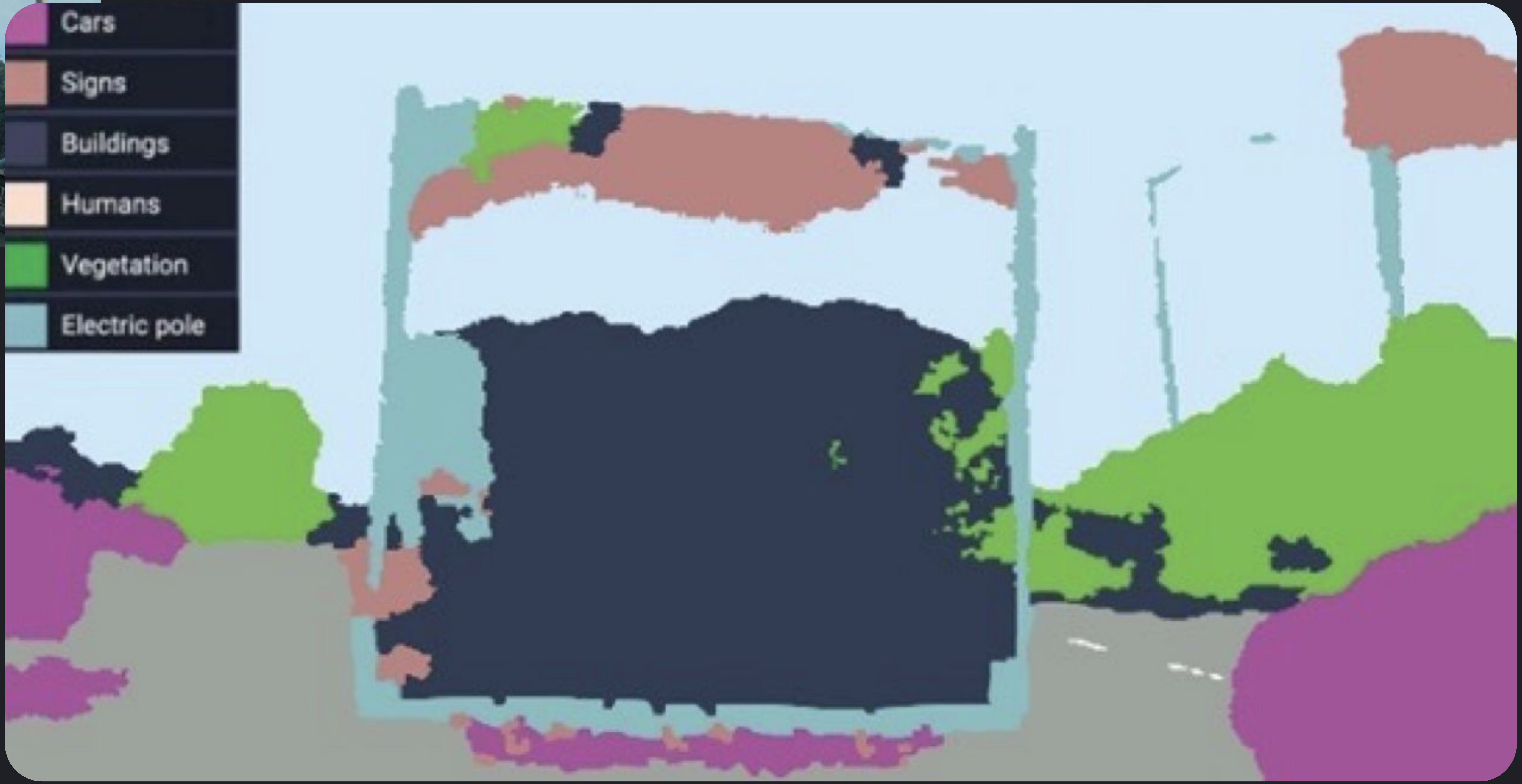
A Typical chihuahua?

Notions of typicality and atypicality … are distinguishable by “the strength of association between observable properties and concepts.”

Eric Margolis and Stephen Laurence. 2007. The ontology of concepts-abstract, objects, or mental representations? Noûs 41, 4 (2007), 561–593.
Bing Ran and P Robert Duimering. 2010. Conceptual combination: Models, theories and controversies. International Journal of Cognitive Linguistics.

Identifying & Characterizing Errors is Critical

Costly errors. At what cost?

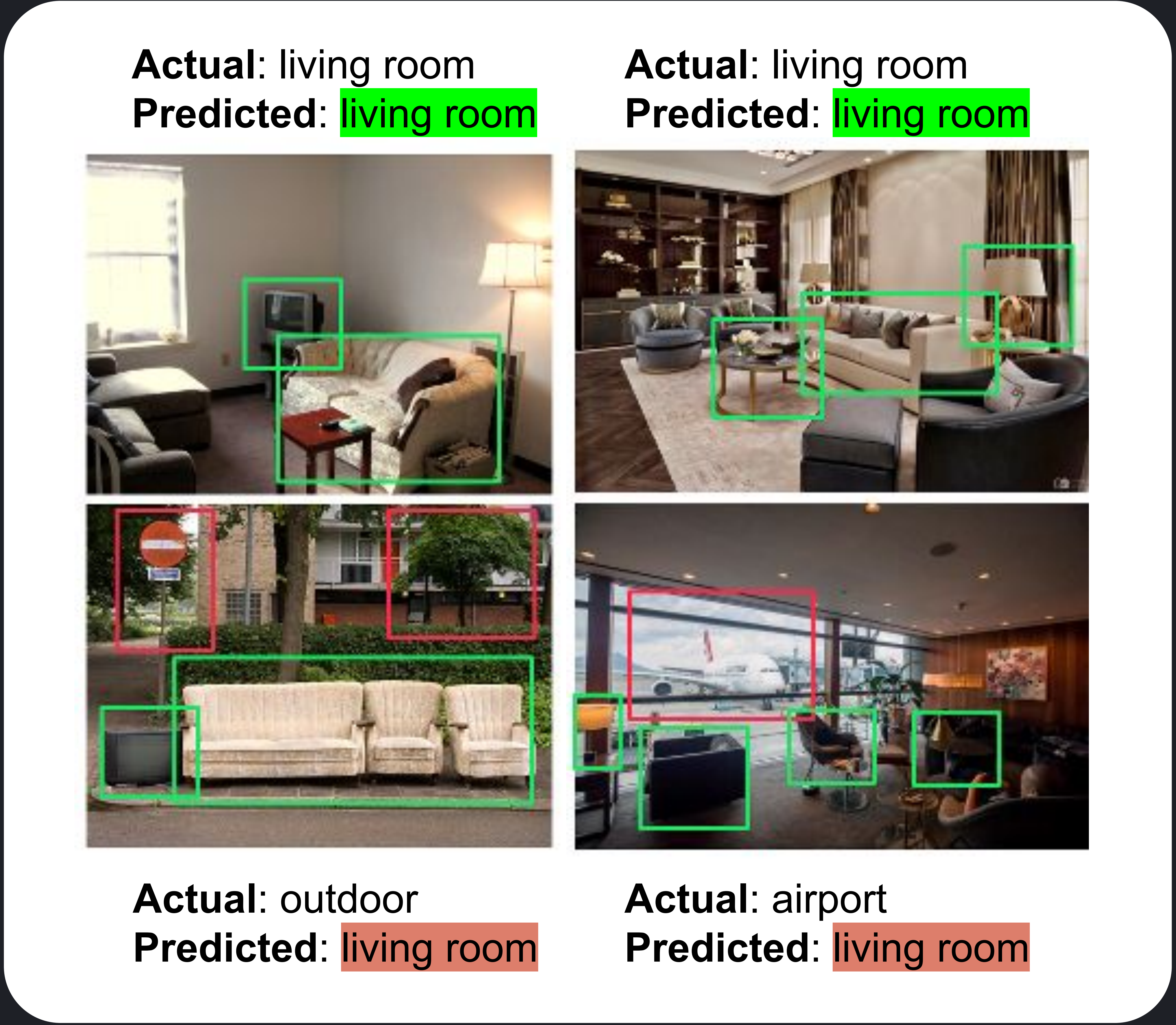


Human vs. Machine Understanding

- Humans perceive the world **discreetly** (object-level), but computer models see the world in a relatively **continuous** form (pixel-level).

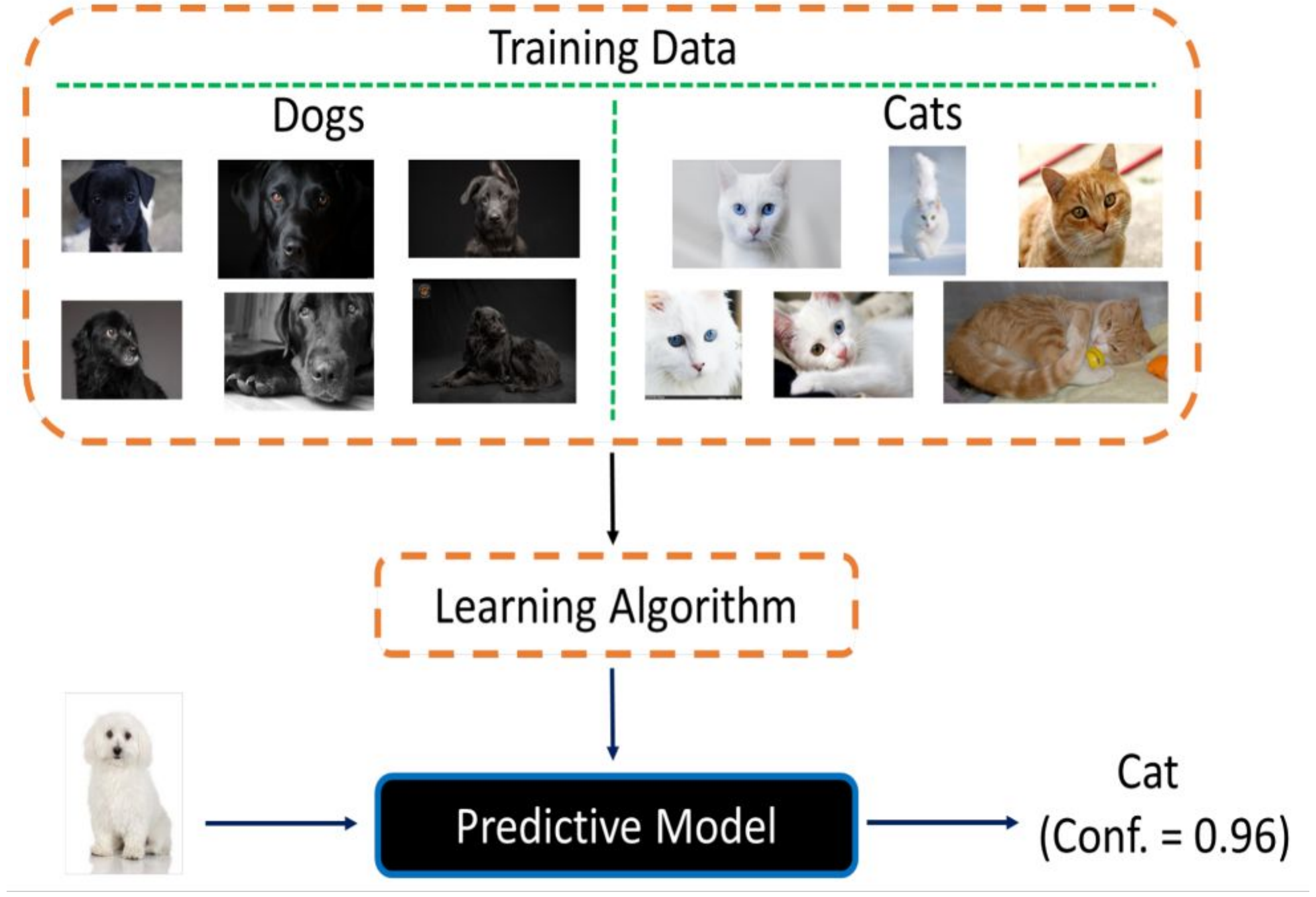
- Humans make predictions more **semantically** (mental models); computer models are trained to predict **statistically**.

- Humans reason about real-world entities **using their corresponding context**, but computer models **often ignore contextual cues**.



Unknown–Unknowns

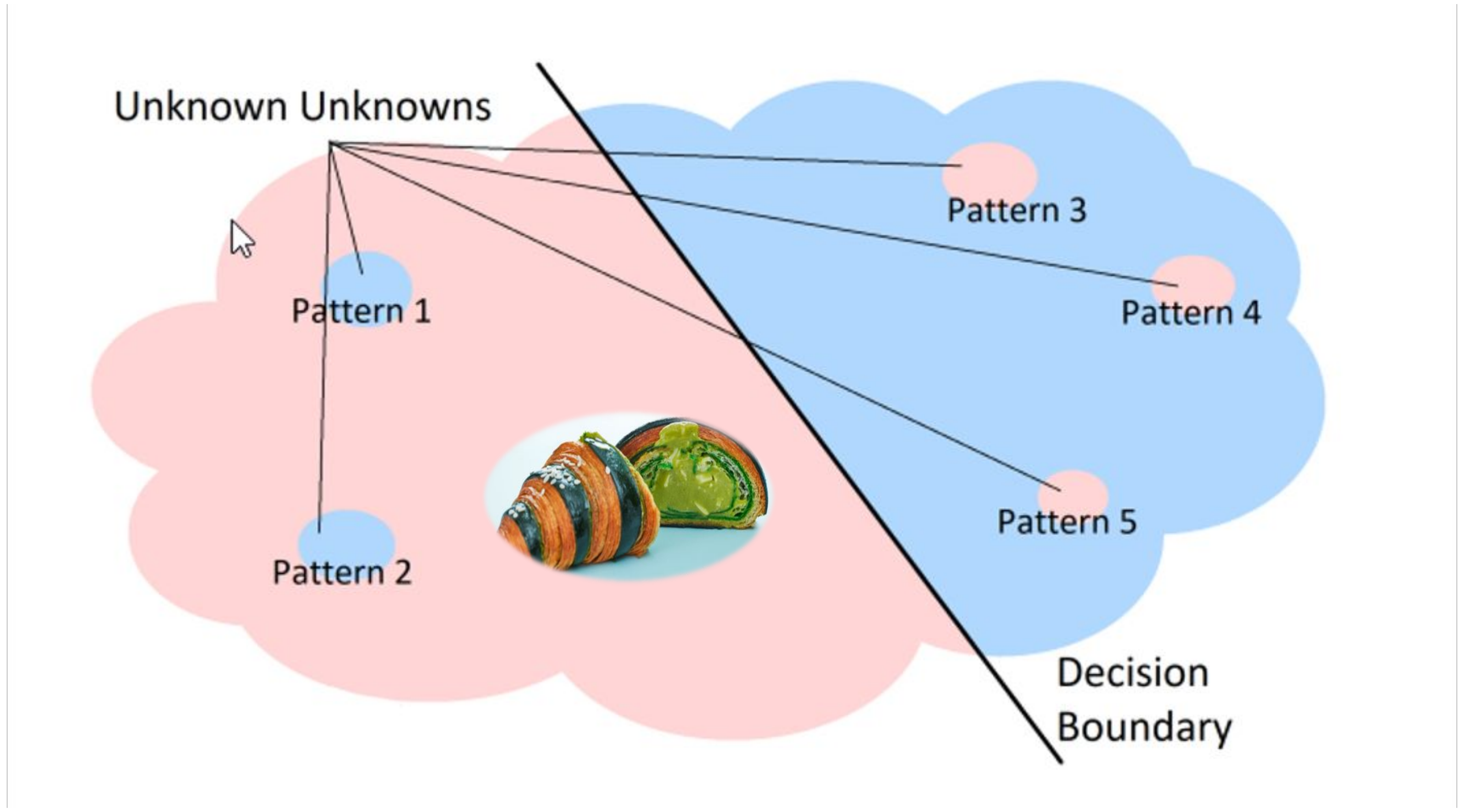
- Over-confident errors
- Caused by systematic biases in the training data
- Hard to discover as ML systems do not provide enough information



(Lakkaraju et al. 2017)

Unknown–Unknowns

- Over-confident errors
- Caused by systematic biases in the training data
- Hard to discover as ML systems do not provide enough information
- Reside in specific partitions of the feature space (blind-spots) and are not distributed evenly across all the feature space



(Z. Liu et al. 2020)

Known–Unknowns in LLMs

- Do LLMs know what they know? And more importantly,are they aware of what they do not know?
- This is an important question to understand the certainty of their statements or prevent such language models from making up facts.

Known Knowns Things we are aware of and understand	**Known Unknowns** Things we are aware of but do not understand
Unknown Knowns Things we understand but are not aware of	**Unknown Unknowns** Things we are neither aware of nor understand

(Amayuelas et al. 2023, Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models)

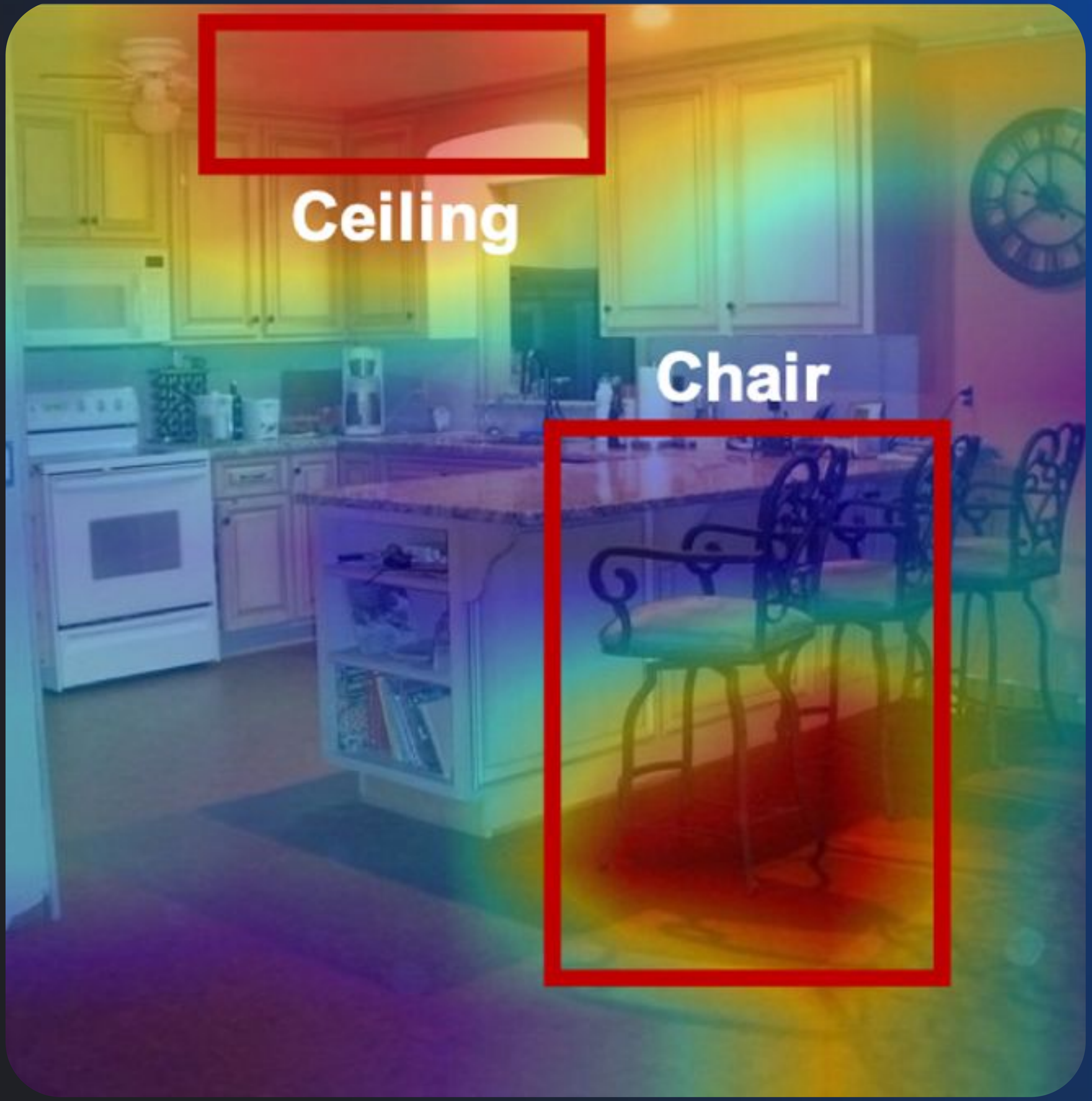
Human input is essential in the evaluation of known-unknowns and the discovery of “unknown-unknowns.”

Need of the hour → proactively discover and characterize unknown-unknowns to build reliable image recognition systems.

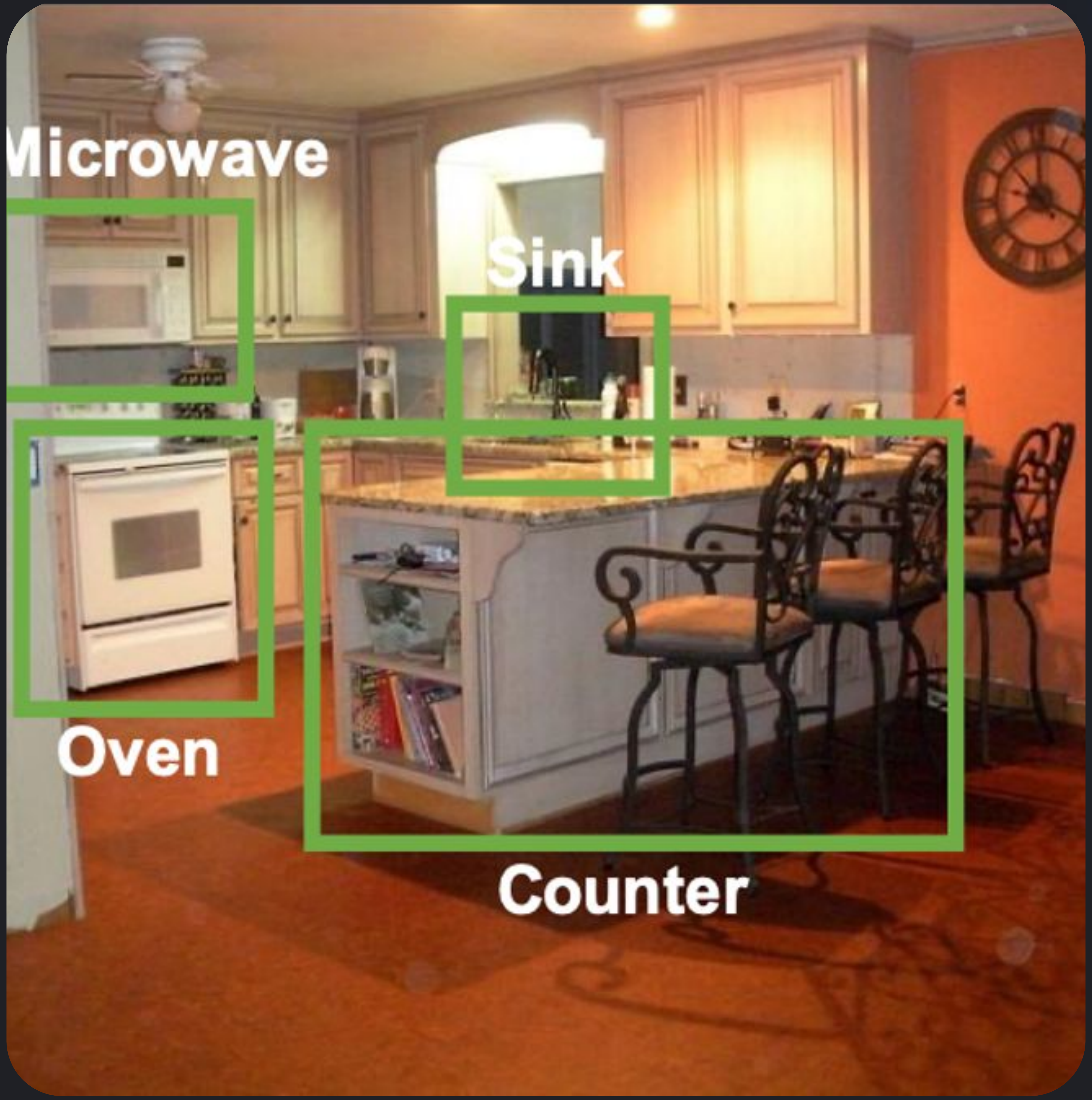
Perspective: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality.

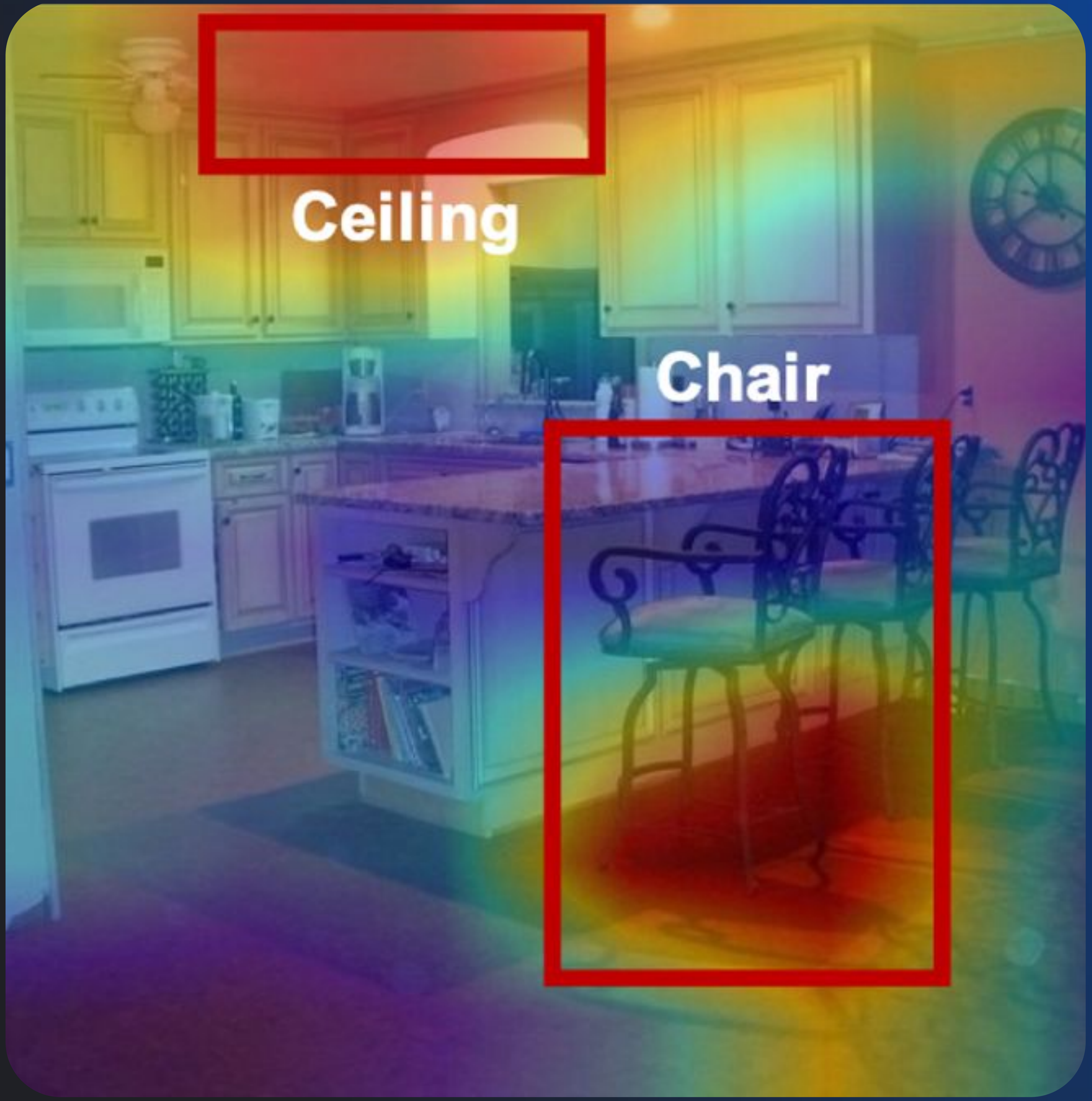
Sharifi et al., ACM IUI 2023

Need for human input!

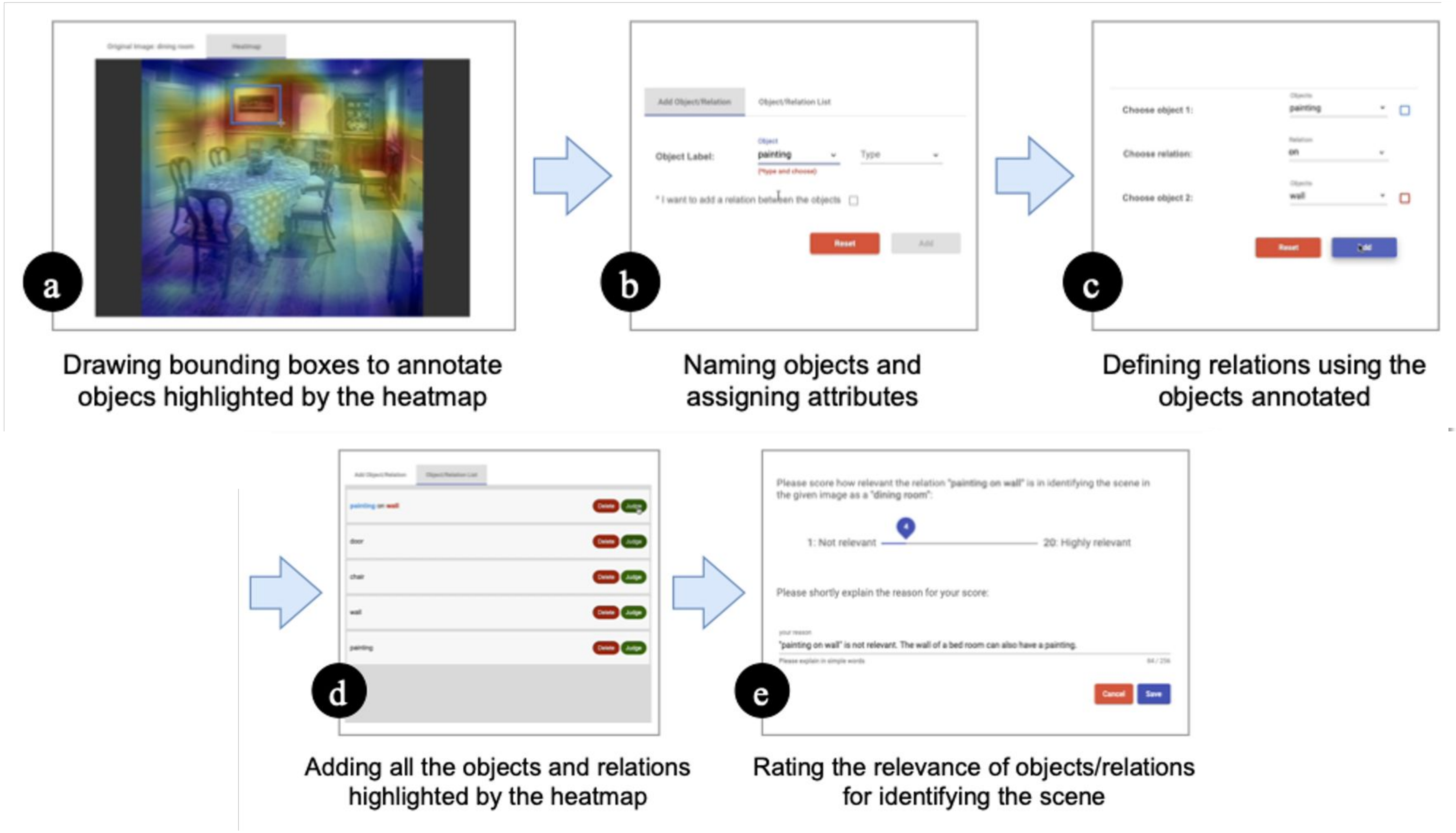


What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. Sharifi, S., Qiu, S., Gadiraju, U., Yang, J., & Bozzon, A. **In Proceedings of the ACM Web Conference (WWW 2022).**





What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. Sharifi, S., Qiu, S., Gadiraju, U., Yang, J., & Bozzon, A. **In Proceedings of the ACM Web Conference (WWW 2022).**


a
Drawing bounding boxes to annotate objects highlighted by the heatmap
b
Naming objects and assigning attributes
c
Defining relations using the objects annotated
d
Adding all the objects and relations highlighted by the heatmap
e
Rating the relevance of objects/relations for identifying the scene

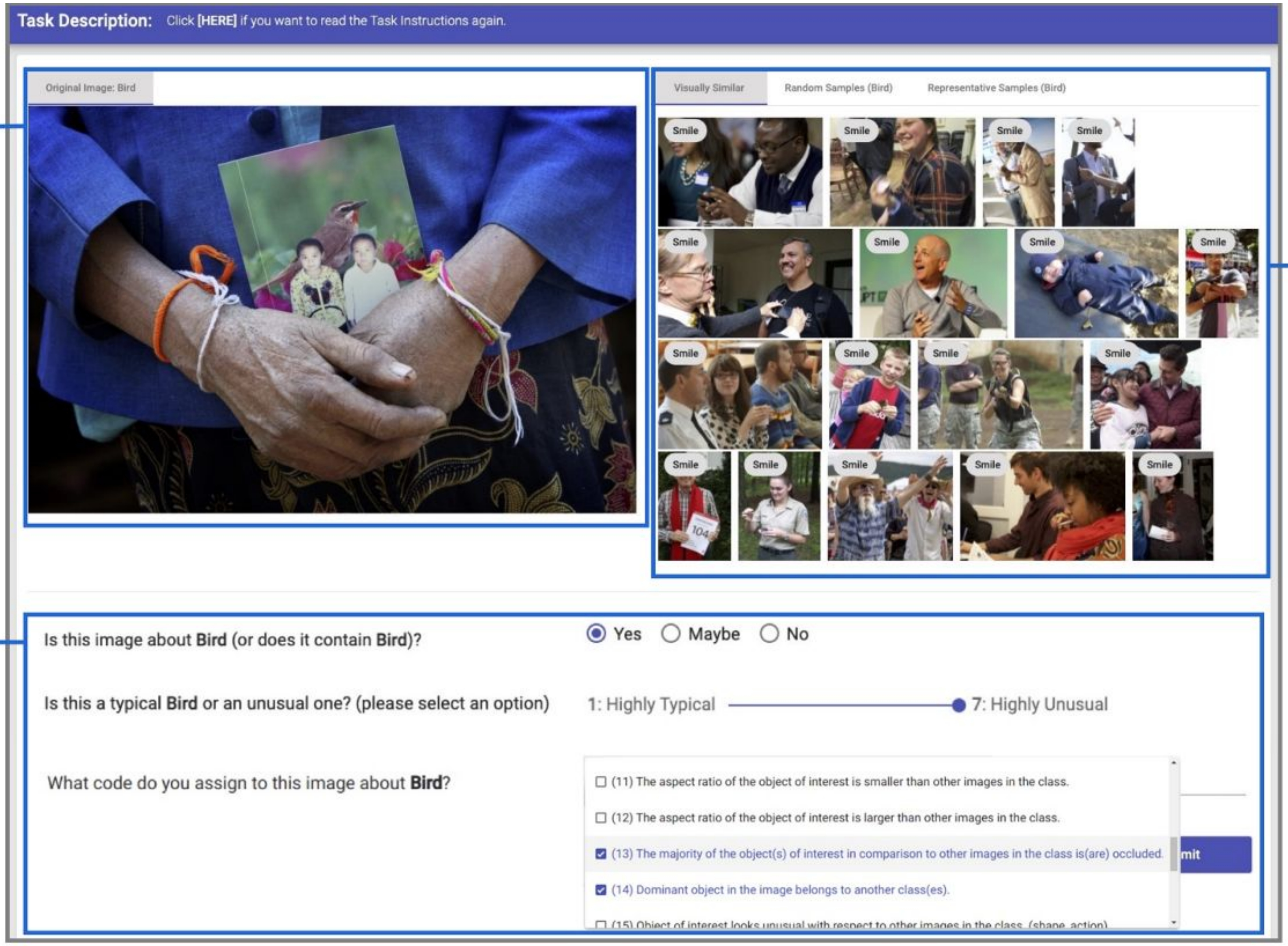
Annotations with “PERSPECTIVE”

1. The Target Image

The target image contains a certain object or is about a scene. Thie initial label (*bird* in this example) is from the given image classifiers.

3. The Questions

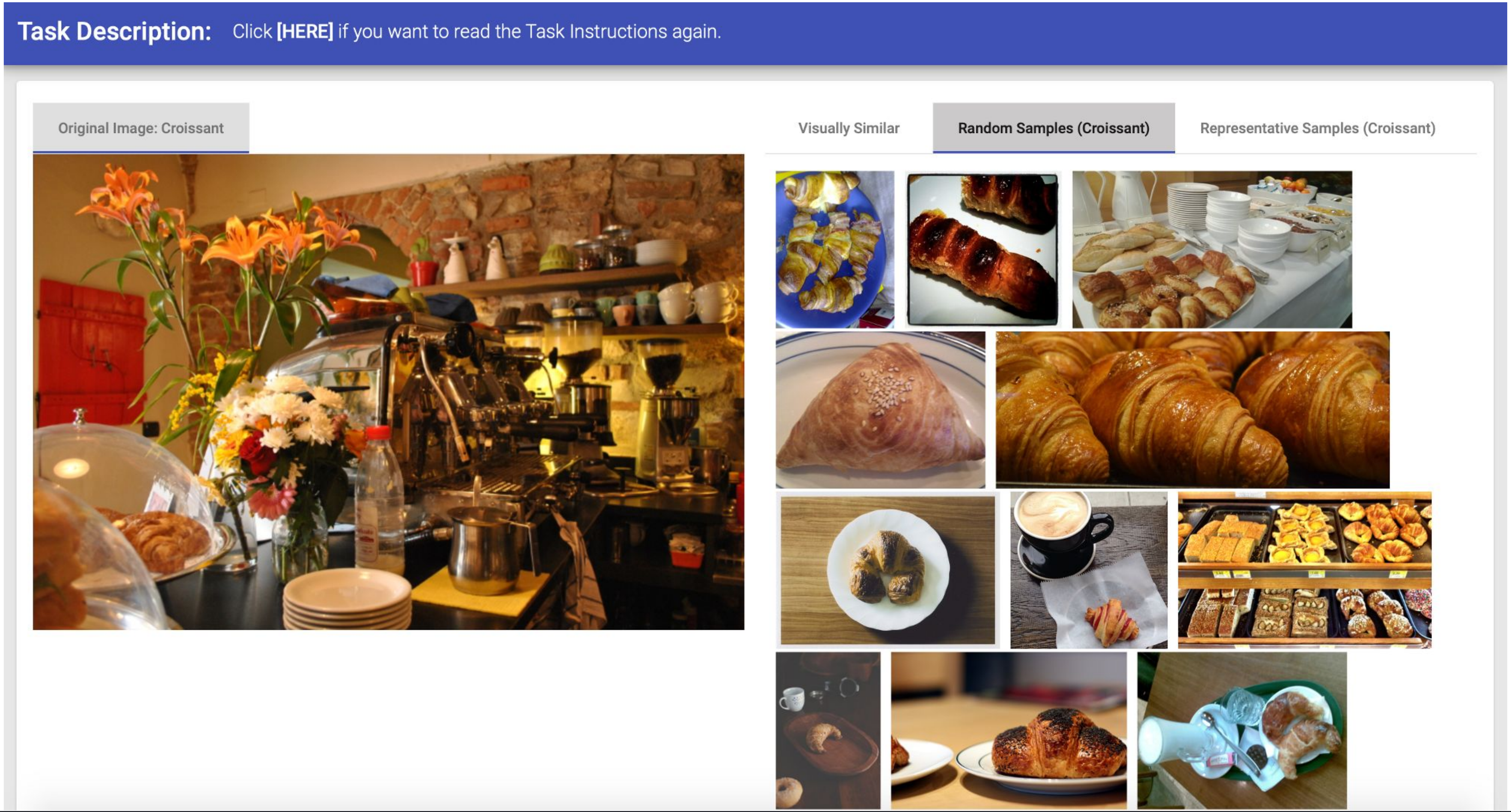
Workers are asked to first decide whether the image label is correct. Then they rate the level of atypicality using a 7-point Likert-scale and give their reasons by choosing the codes.



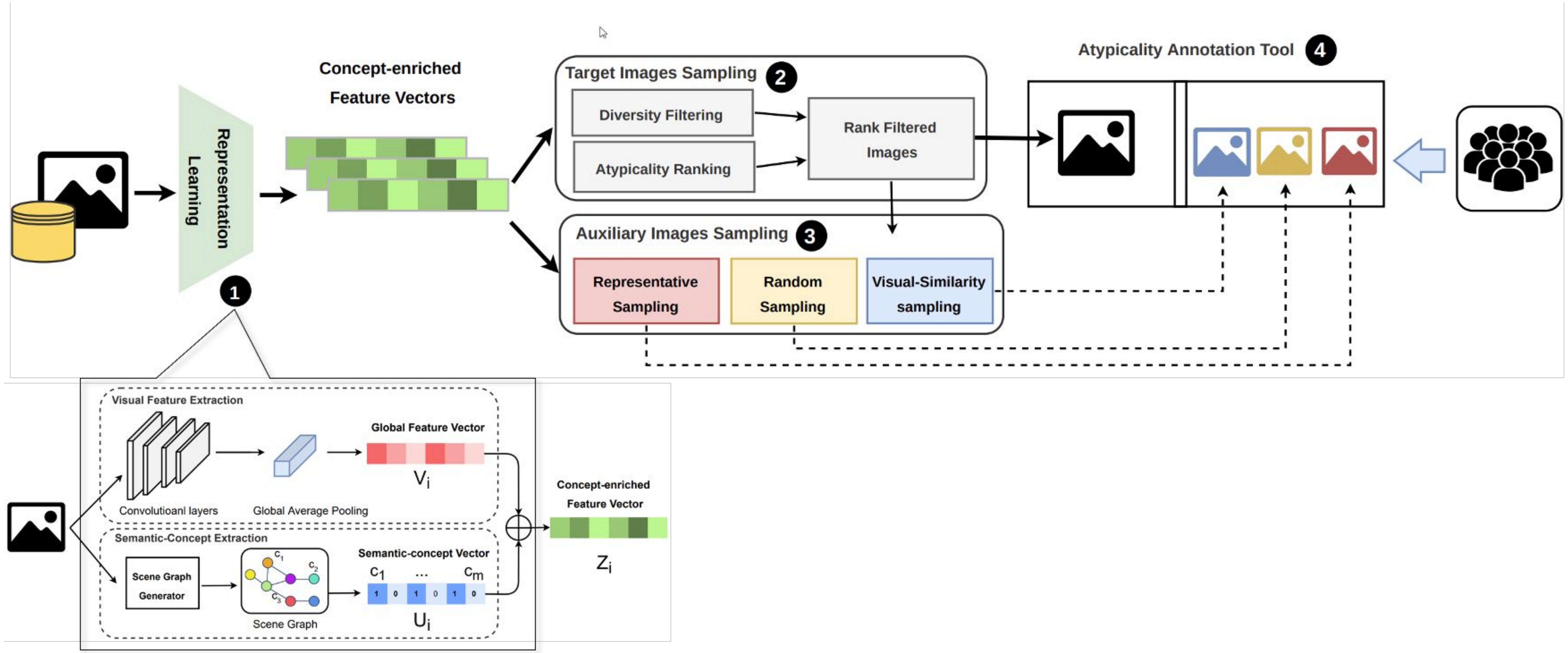
2. The Auxiliary Images

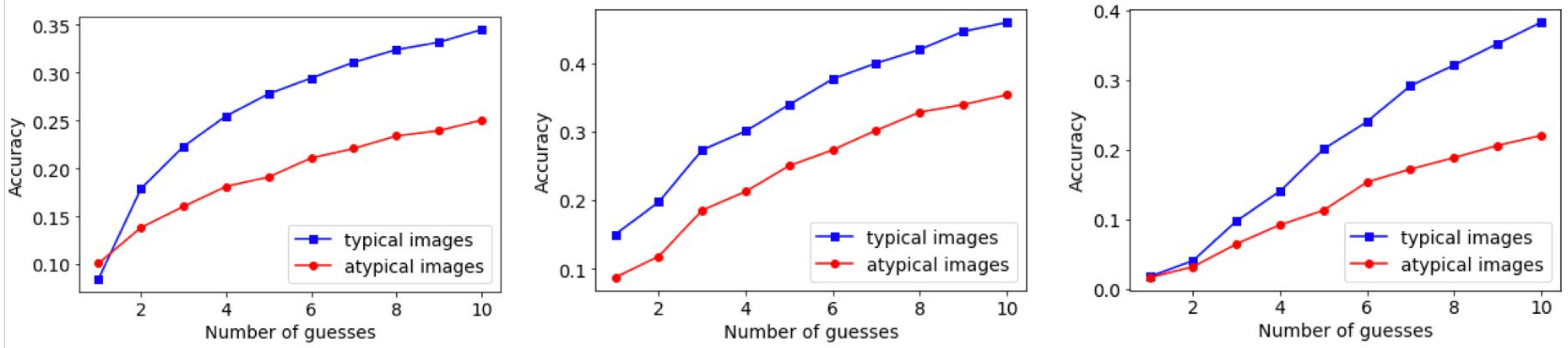
The auxiliary images are organized in different tabs, each displaying one type of the auxiliary images (visually similar, random sample, or representative sample).

Workers are asked to judge and characterize atypicality by comparing the target image to the auxiliary images.


Task Description:
Click [HERE] if you want to read the Task Instructions again.
Original Image: Croissant
Visually Similar
Random Samples (Croissant)
Representative Samples (Croissant)

Architecture of "PERSPECTIVE"





(a) Google Vision API (b) Amazon Rekognition API (c) Microsoft Azure Vision API

Check out the paper for MORE interesting details and analyses

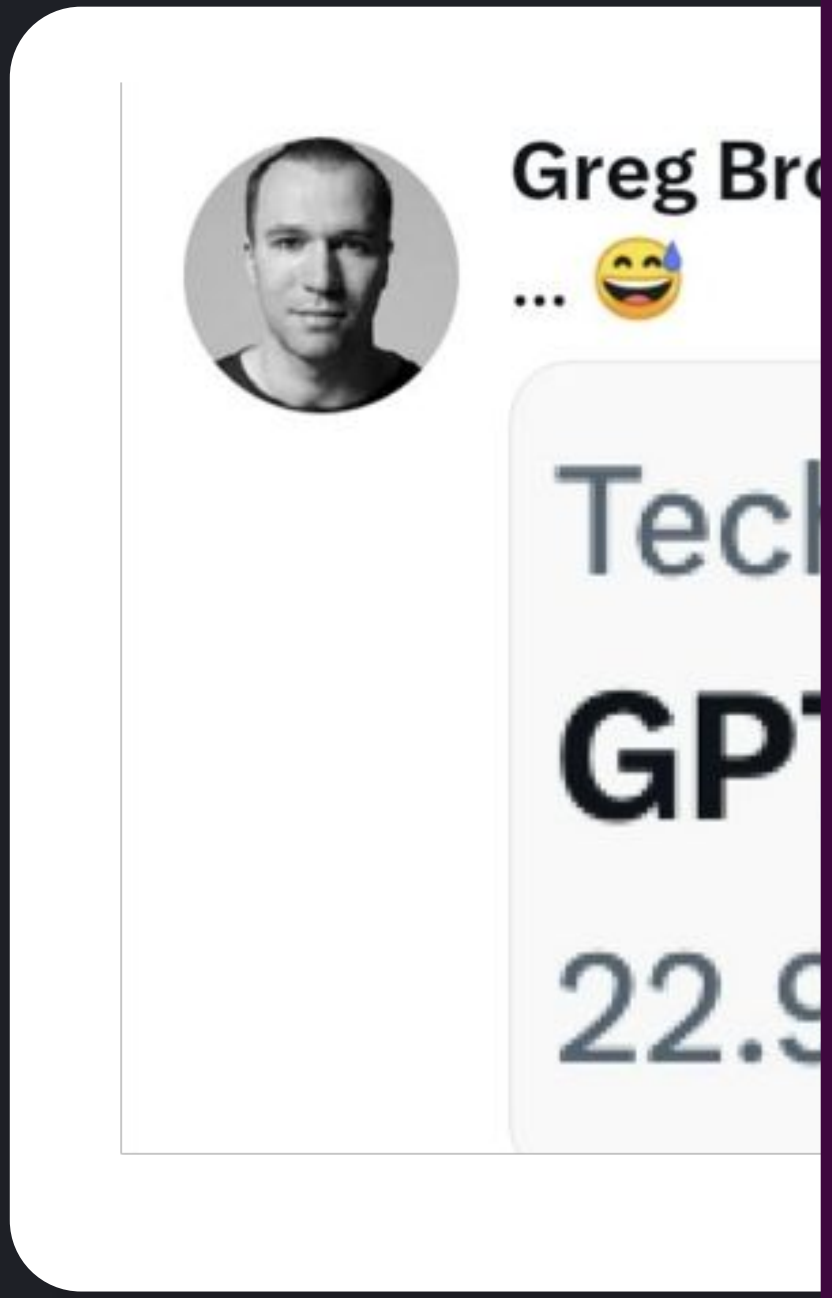
edu.nl/cy7cj

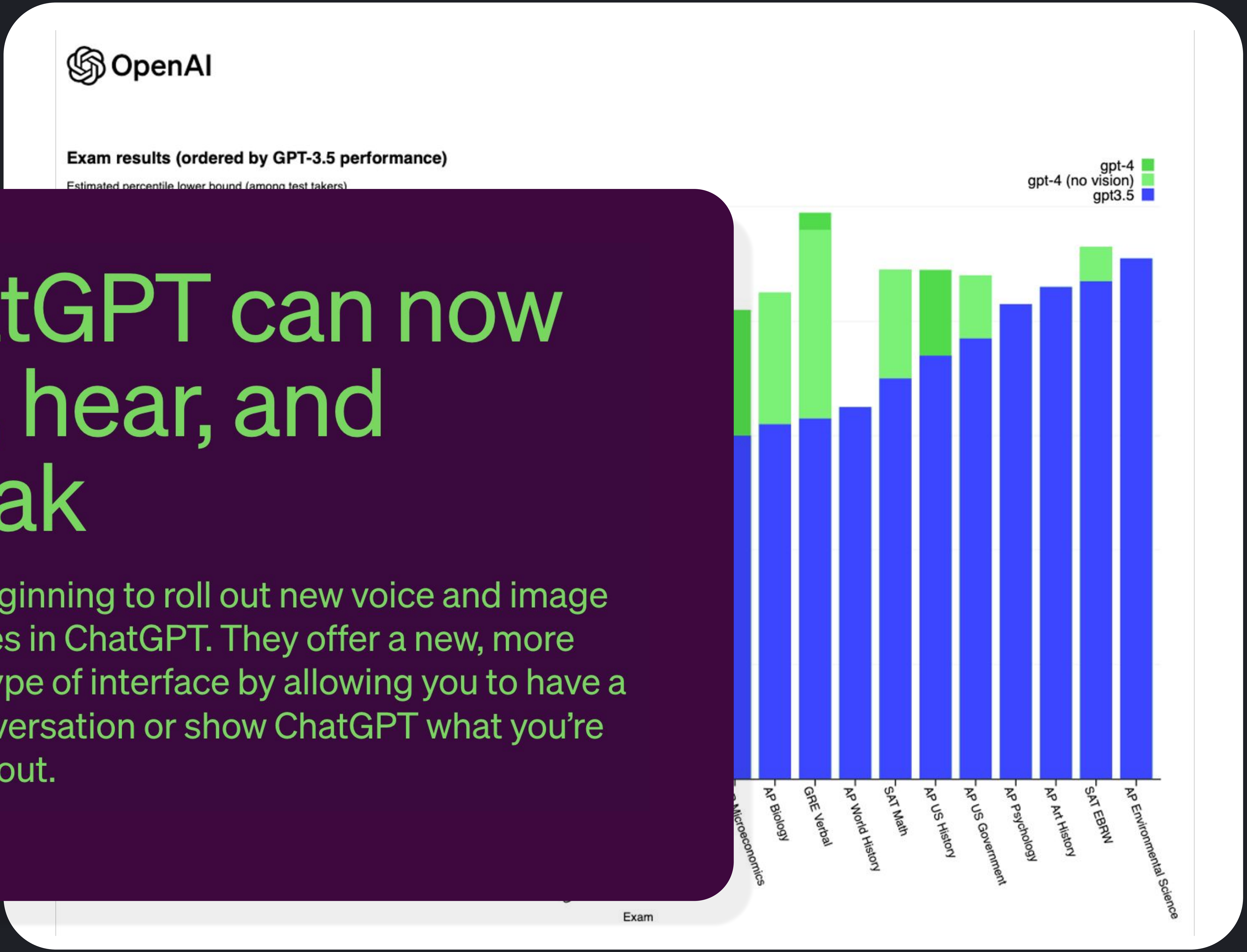
Toloka

Key takeaways

1. Annotations with **Perspective** led to identification of most atypical images
2. Importance of context expansion during labelling workflows
3. Response and data sampling biases need to be addressed
4. No place for mistakes when the stakes are high — responsible and trustworthy AI

The Analogous Challenges with LLMs …





ChatGPT can now see, hear, and speak

We are beginning to roll out new voice and image capabilities in ChatGPT. They offer a new, more intuitive type of interface by allowing you to have a voice conversation or show ChatGPT what you're talking about.

Biases in the Age of LLMs

Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models

Harnoor Dhingra **Preetiha Jayashanker** **Sayali Moghe** **Emma Strubell**
Carnegie Mellon University
{hdhingra, pjayasha, smoghe, estrubel}@cs.cmu.edu

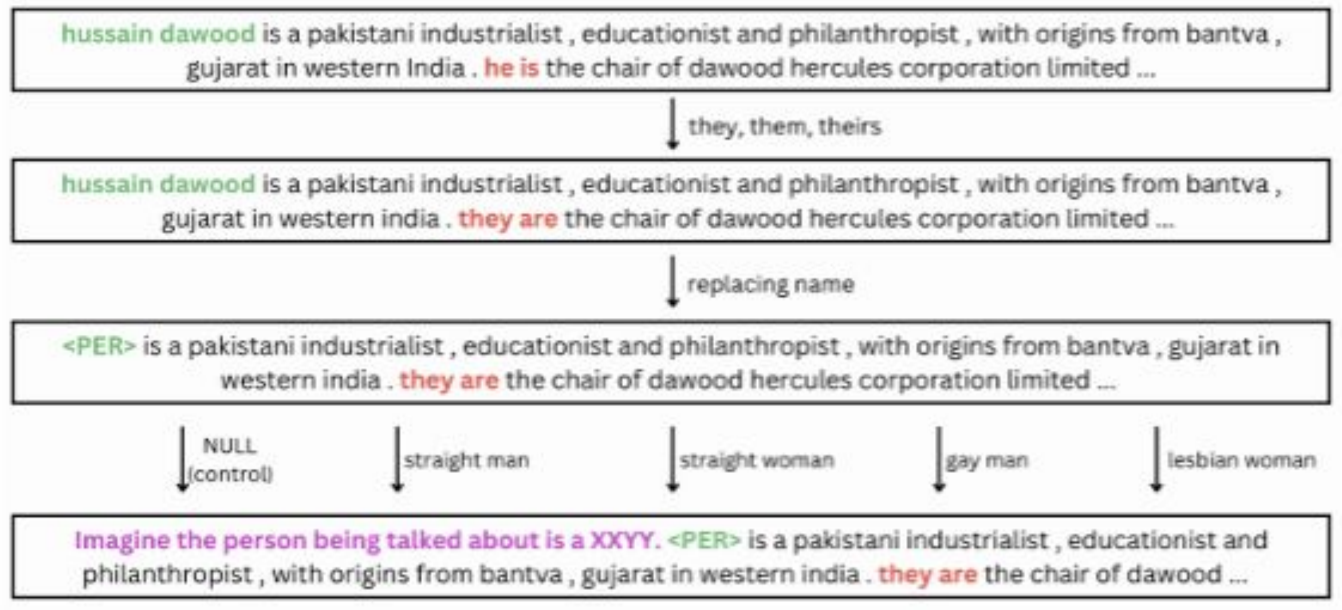
Abstract

Large Language Models (LLMs) are trained primarily on minimally processed web text, which exhibits the same wide range of social biases held by the humans who created that content. Consequently, text generated by LLMs can inadvertently perpetuate stereotypes towards marginalized groups, like the LGBTQIA+ community. In this paper, we perform a comparative study of how LLMs generate text describing people with different sexual identities. Analyzing bias in the text generated by an LLM using regard score shows measurable bias against queer people. We then show that a post-hoc method based on chain-of-thought prompting using SHAP analysis can increase the regard of the sentence, representing a promising approach towards debiasing the output of LLMs

quantify the detected biases. Hence, in this work we aim to answer the following research questions:

RQ1: Does a pre-trained LLM perpetuate *measurable, quantifiable* bias against queer people?

RQ2: Can we *mitigate* the said bias in the LLM output *while preserving the context* using a post-hoc debiasing method?



Biases in the Age of LLMs

- Instruction-tuned LLMs have been shown to be effective in generating high-quality natural language responses

- Open RQ → inherent biases in trained models and the generated responses

  - E.g., dataset for fine-tuning is predominantly composed of a specific political bias, we can expect the generated answers to share such bias

Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias

Itay Itzhak[1], Gabriel Stanovsky[2], Nir Rosenfeld[1], Yonatan Belinkov[1]
[1]Technion – Israel Institute of Technology
[2]School of Computer Science and Engineering, The Hebrew University of Jerusalem
itay1itzhak@gmail.com,
{nirr, belinkov}@technion.ac.il, gabriel.stanovsky@mail.huji.ac.il

Abstract

Recent studies show that instruction tuning and learning from human feedback improve the abilities of large language models (LMs) dramatically. While these tuning methods can help models generate high-quality text, we conjecture that they may also inadvertently cause models to express cognitive-like biases. Our work provides evidence that fine-tuned models exhibit biases that were absent or less pronounced in their pretrained predecessors. We examine the extent of this phenomenon in three cognitive biases: the decoy effect, the certainty effect, and the belief bias—all of which are known to influence human decision-making and reasoning. Our findings highlight the presence of these biases in various models, especially those that have undergone instruction tun-

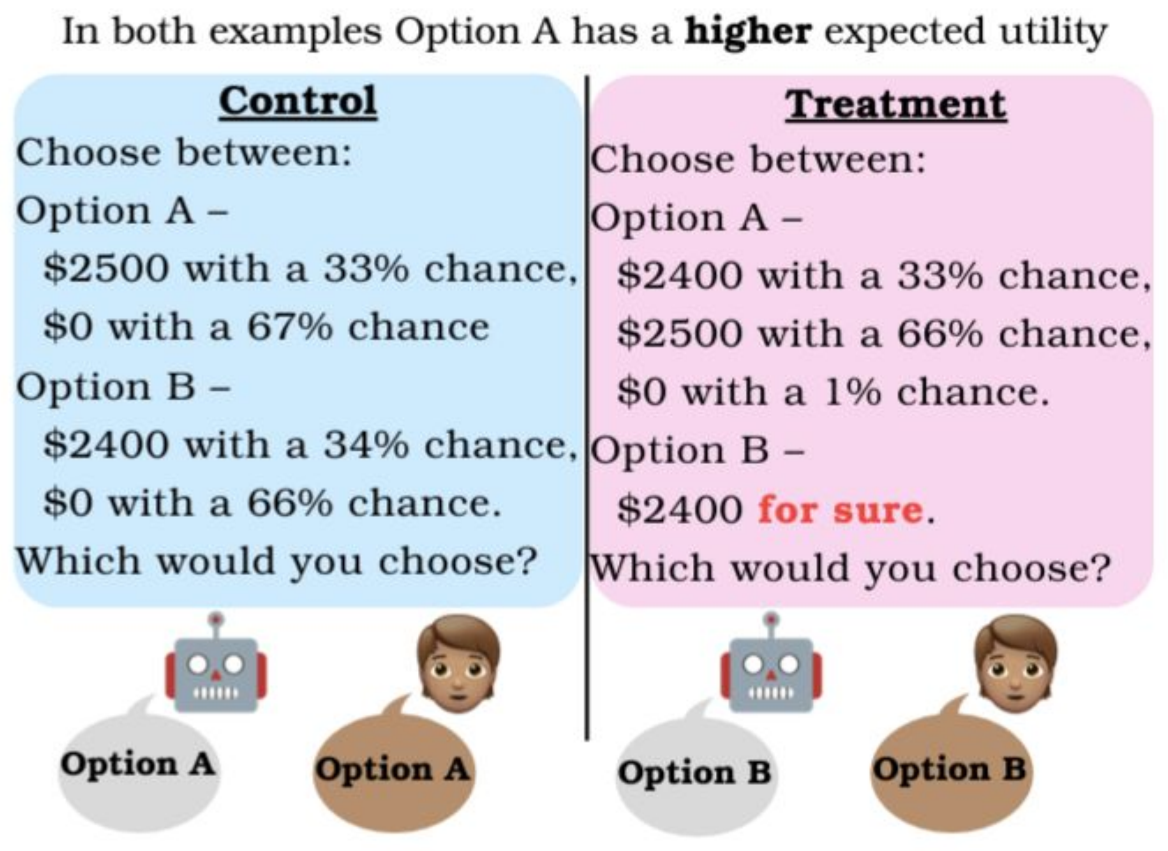


Figure 1: Example tasks from *certainty effect* dataset, for the control condition (left) and treatment condition (right), along with typical answers from humans and instruction-tuned models, both of which are biased.

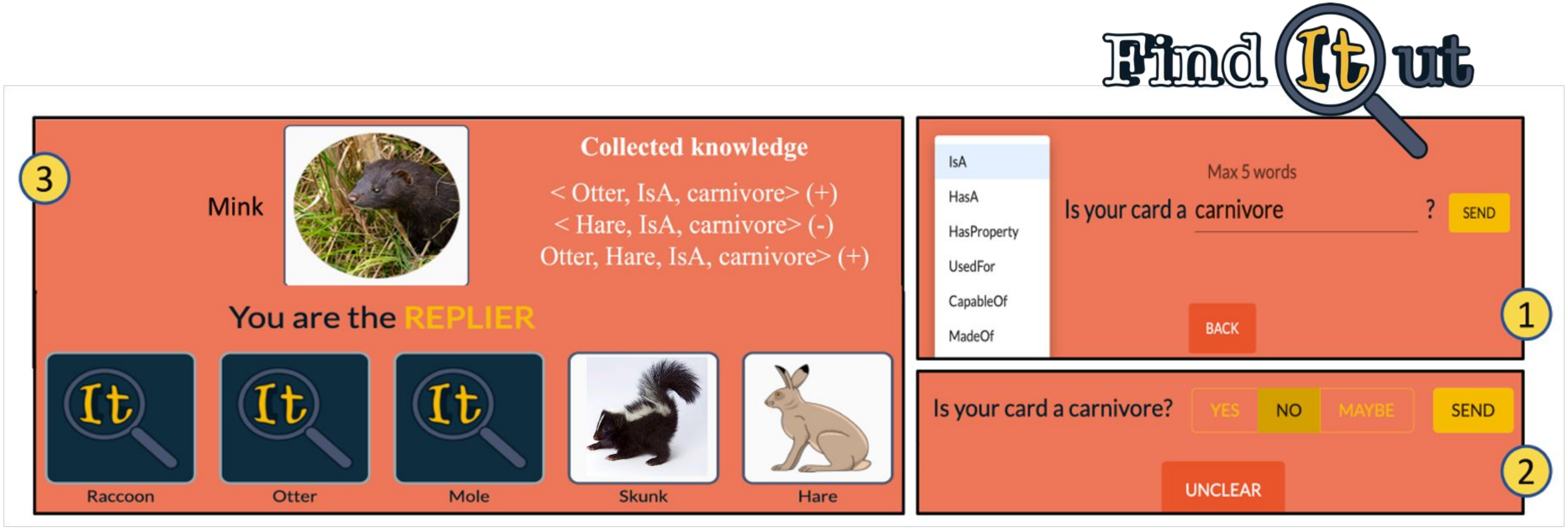
The indispensable role of human input

How can we elicit human input (tacit knowledge) to manage biases in LLMS

Eliciting Diverse Knowledge from Humans Using A Game-with-a-purpose

- Commonsense knowledge →
  building neuro-symbolic AI systems,
  debugging deep learning models
- Existing knowledge acquisition methods are limited
- Broad tacit and negative knowledge,
  and discriminative knowledge →
  Our solutionE a GWAP, FindItOut

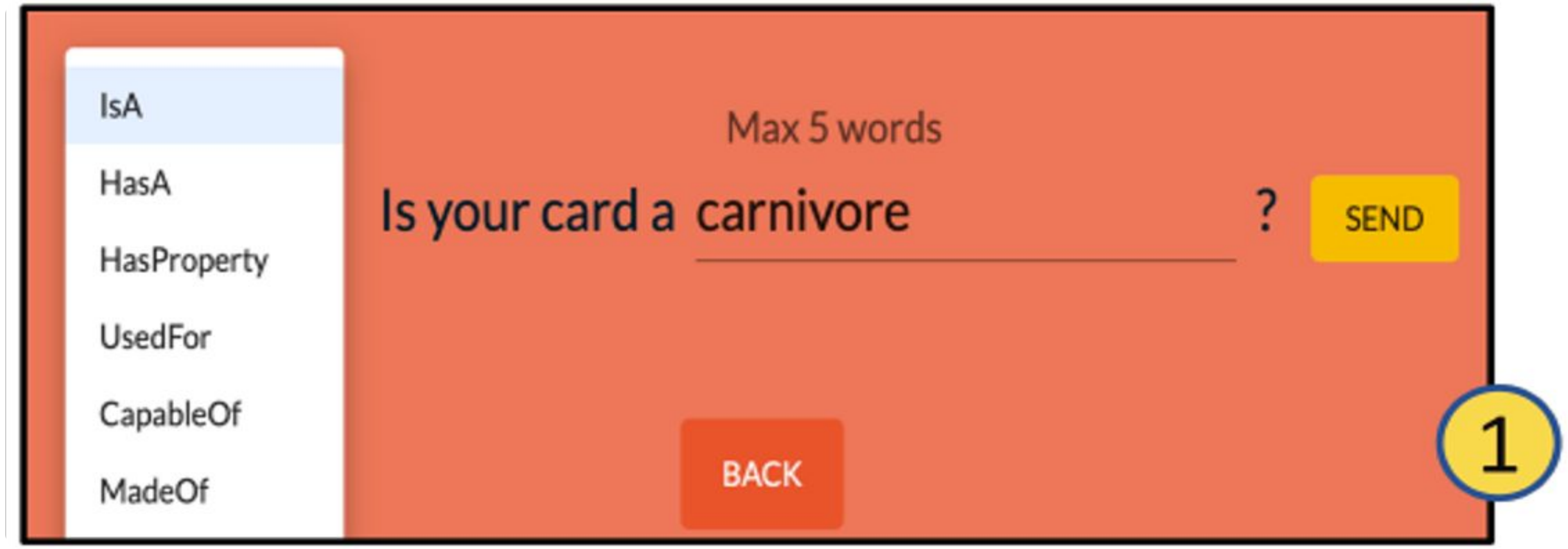
Eliciting Diverse Knowledge Using A Configurable Game



Play → https://finditout.vercel.app/

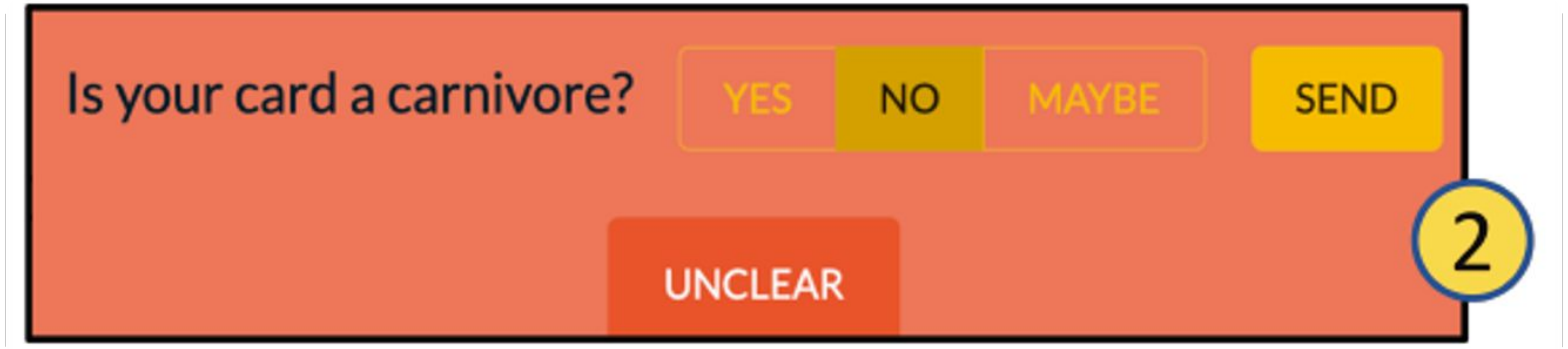
Best Demo & Poster Award at **AAAI HCOMP 2021**
Best Paper Award Nomination at **the ACM Web Conference 2022**

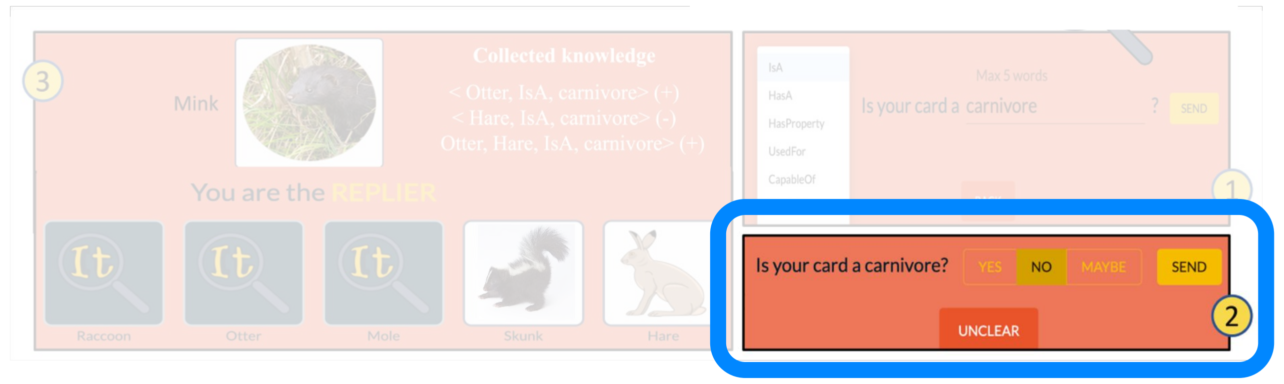
Eliciting Diverse Knowledge Using A Configurable Game



Play → https://finditout.vercel.app/

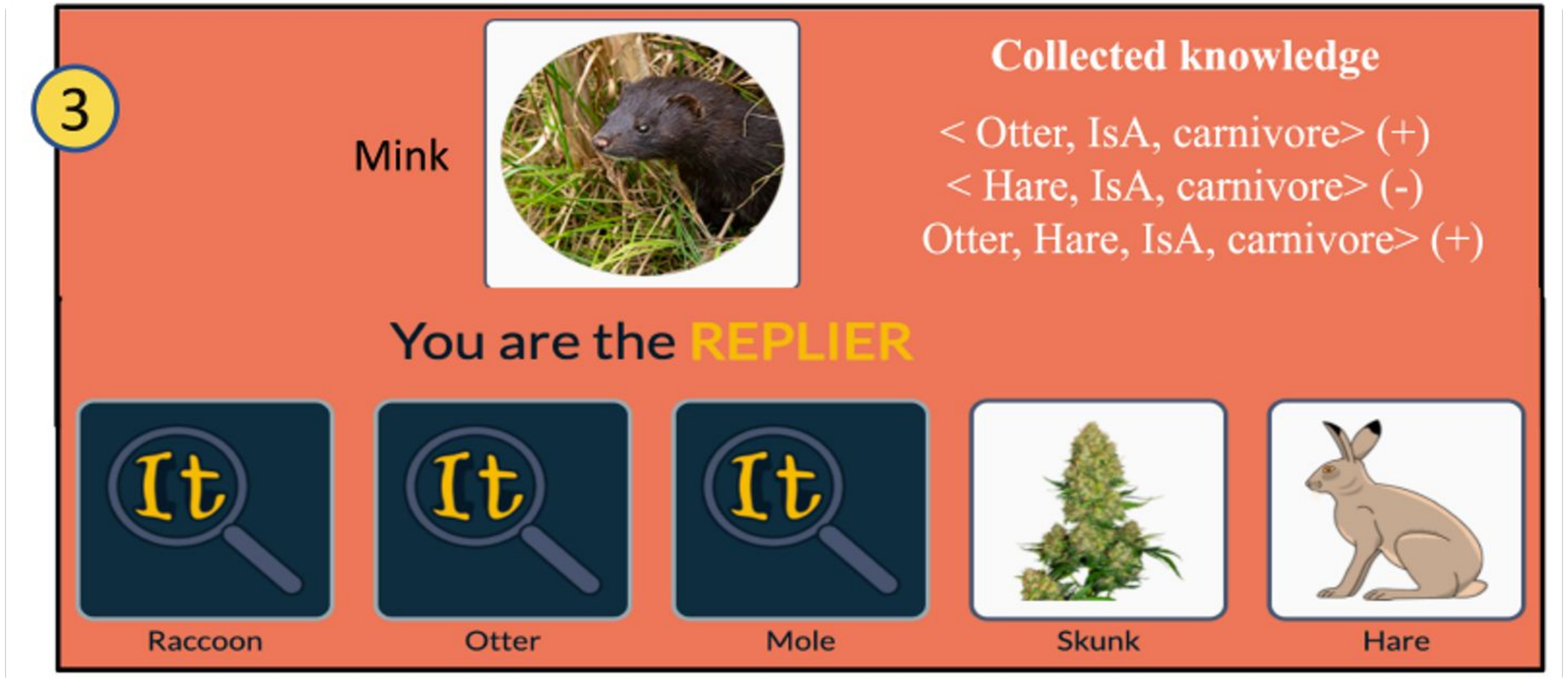
Eliciting Diverse Knowledge Using A Configurable Game





Play → https://finditout.vercel.app/

Eliciting Diverse Knowledge Using A Configurable Game



Play → https://finditout.vercel.app/

Collecting "tacit knowledge" for downstream AI tasks

Find It Out

Board	Type	Question	Knowledge Tuple
floor, window, bathroom, walls, ceiling, chandelier, mirror, bedroom	Explicit	Can your card be found inside an apartment?	<bathroom, AtLocation, inside apartment>
	Tacit	Can your card be used for decoration?	<chandelier, UsedFor, decoration>
necklace, dress, boots, shoes, pants, trousers, jeans, skirt	Explicit	Can your card be found in your wardrobe?	<dress, AtLocation, wardrobe>
	Tacit	Is your card typically worn by cowboys?	<boots, HasProperty, worn by cowboys>

Collecting “tacit knowledge” for downstream AI tasks

Empirical Results:

- 125 players played 2430 rounds → 150k knowledge tuples
- Efficiency of game is 10x higher than a reference baseline

Verbosity

- Usefulness validated in two downstream AI tasks
  - Commonsense Question-Answering
  - Identification of Discriminative Attributes
- Enjoyable game experience (player experience inventory)

Re-Enter → The Analogous Challenges with LLMs

Use annotation tools and workflows like “Perspective” to identify and characterize biases →

Explore how such biases are represented in LLMs

Elicit diverse human input to create bias-aware instruction tuning datasets →

Mitigate and manage biases in fine-tuned LLMs

Human input and oversight are essential to overcome fundamental challenges in facilitating bias-aware interactions with LLMs.

Ujwal Gadiraju

Dr. ir.

ujwalg@toloka.ai

www.toloka.ai

Linkedin

Twitter

GitHub

Join our
Slack community

Slack