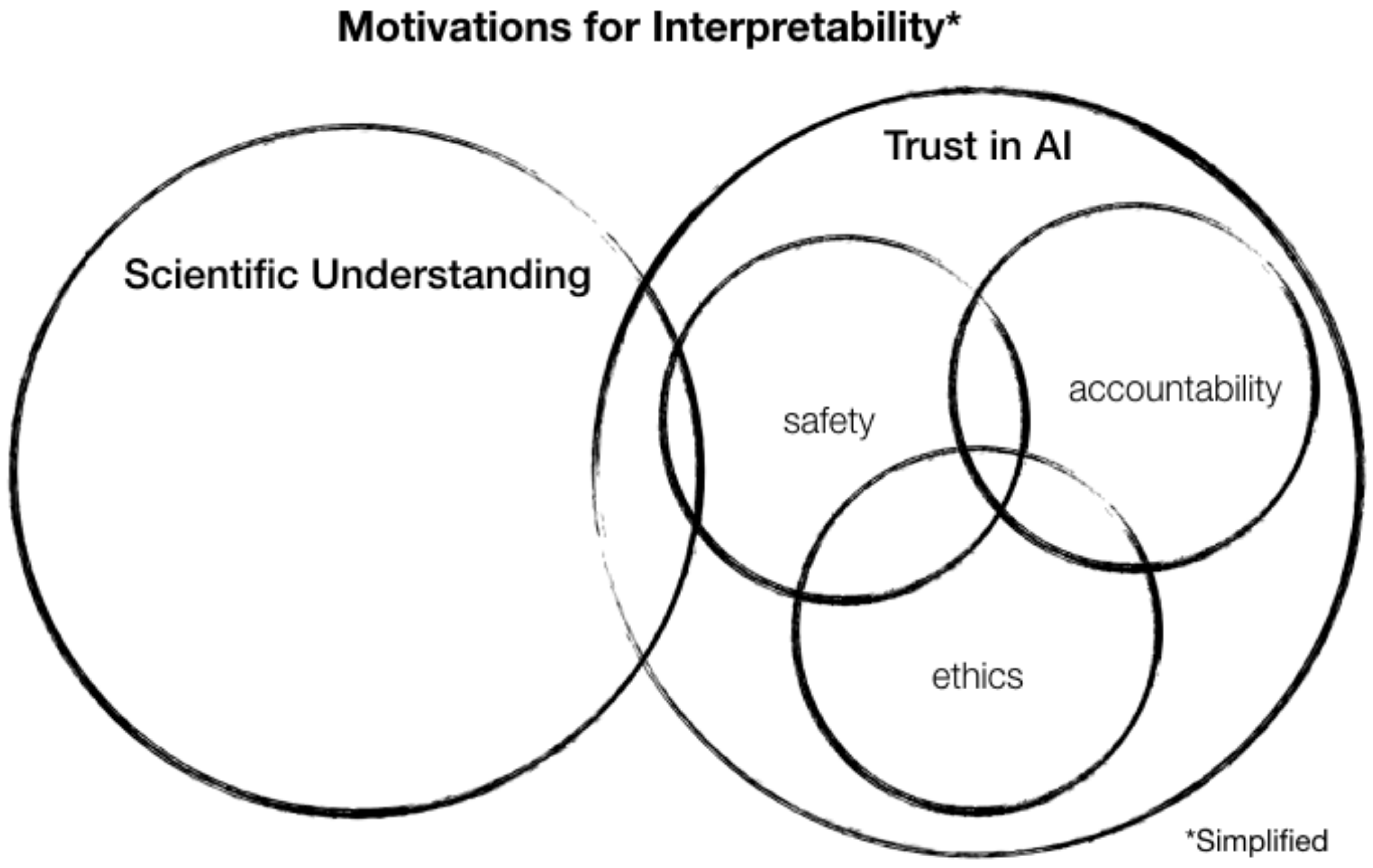
Interpreting LLMs

DRG Prompt Surfing 02/05/2024

Ruoxi Shang

The Trust Problem

- If we want to put an ML model into production, how do we gain confidence that it won't kill someone, cause financial damage, make biased decisions against minorities, etc.
- We need to trust it. But how can we trust an AI model?
- Just like how we trust people, we need to understand how it works.



Interpretability is the degree to which a human can understand the cause of a decision.

The higher the interpretability of an ML model, the easier it is to comprehend the model's predictions. Interpretability facilitates:

- Understanding
- Debugging and auditing ML model predictions
- Bias detection to ensure fair decision making
- Robustness checks to ensure that small changes in the input do not lead to large changes in the output
- Methods that provide recourse for those who have been adversely affected by model predictions

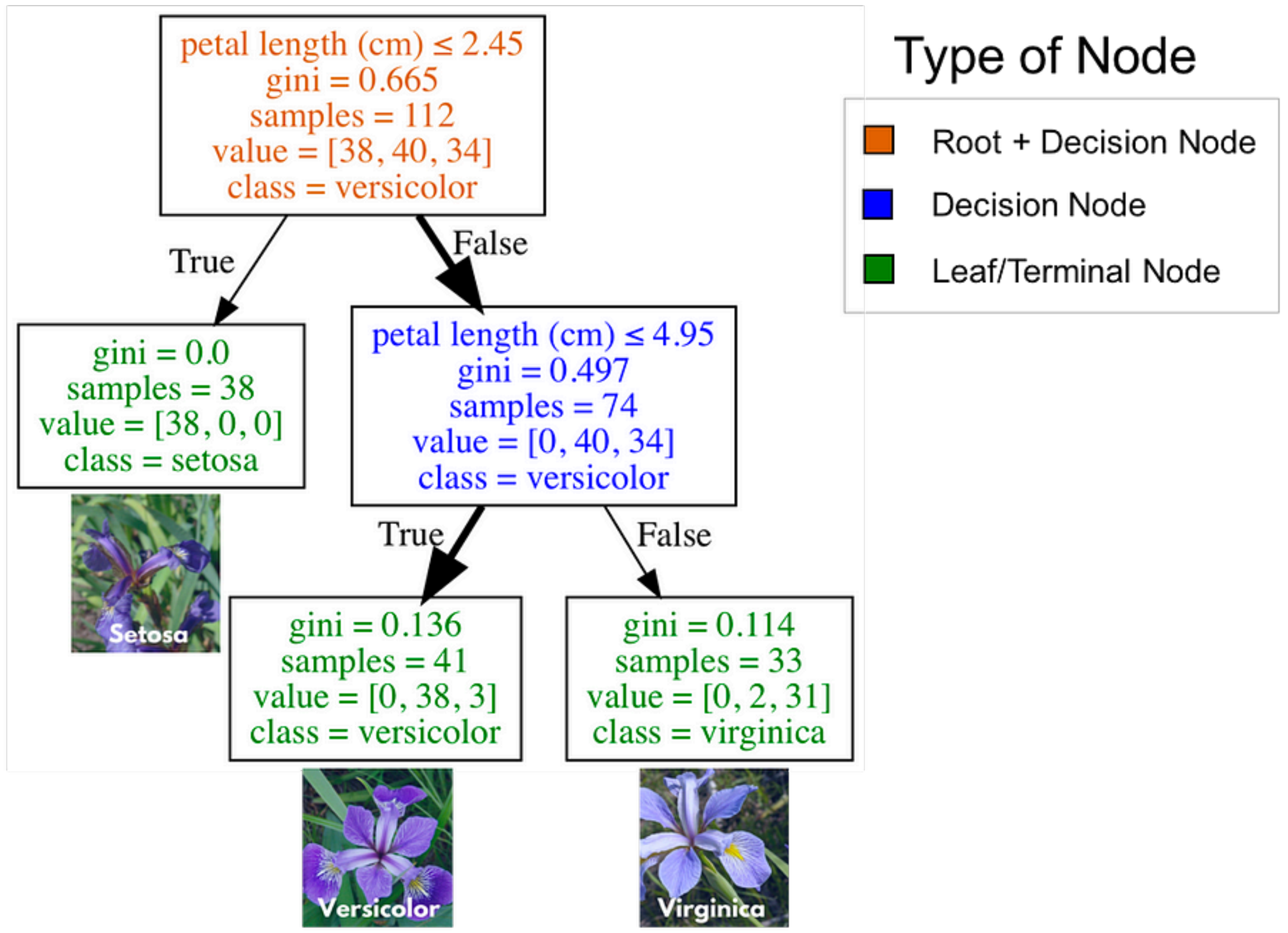
https://docs.aws.amazon.com/whitepapers/latest/ml-best-practices-healthcare-life-sciences/model-interpretability.html

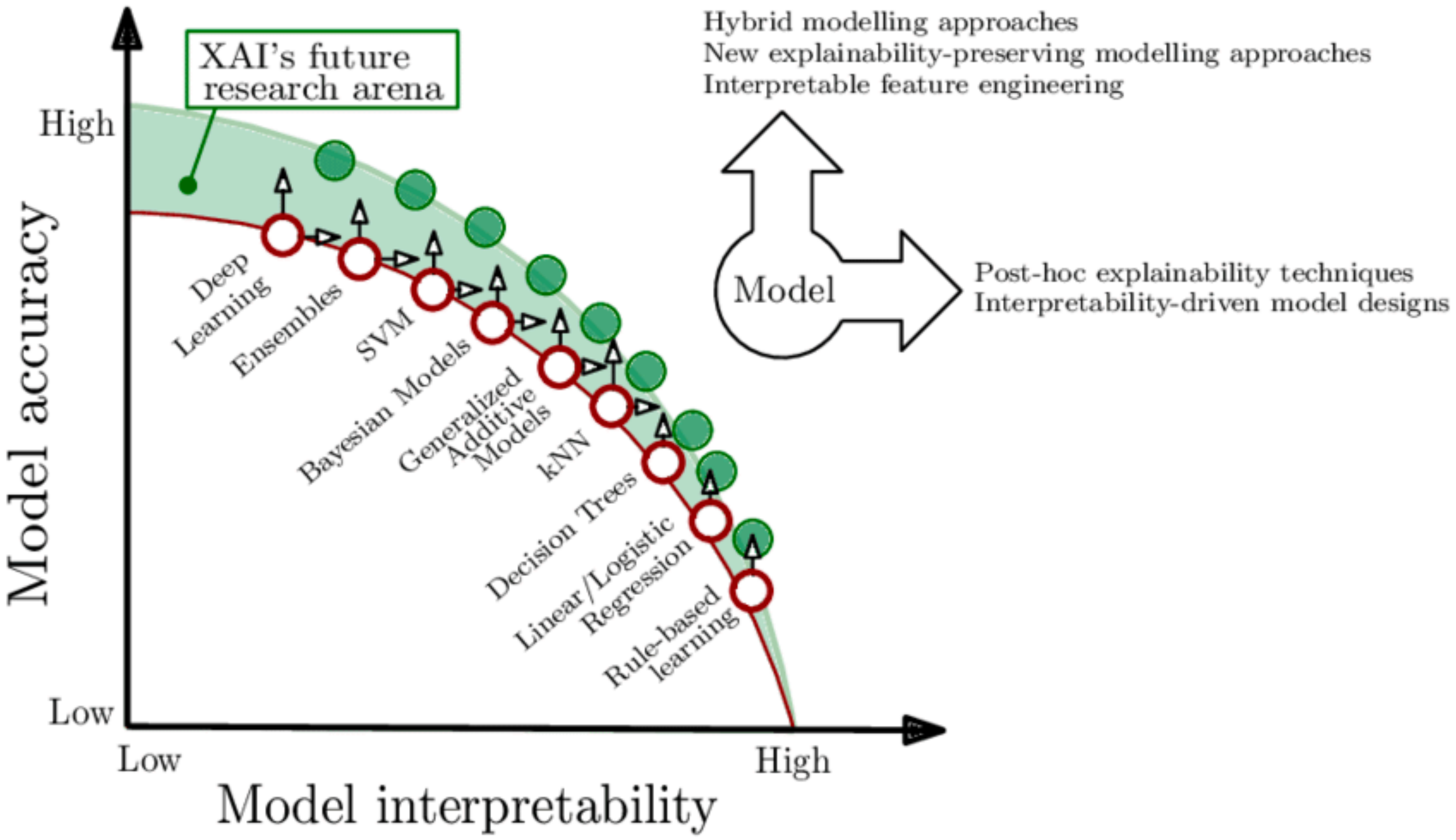
Trade-off between model complexity and interpretability

- Increase model complexity to improve performance, decrease to improve interpretability.
  - By this definition, LLMs are EXTREMELY uninterpretable.

What class (species) is a flower with the following feature?

petal length (cm): 4.5





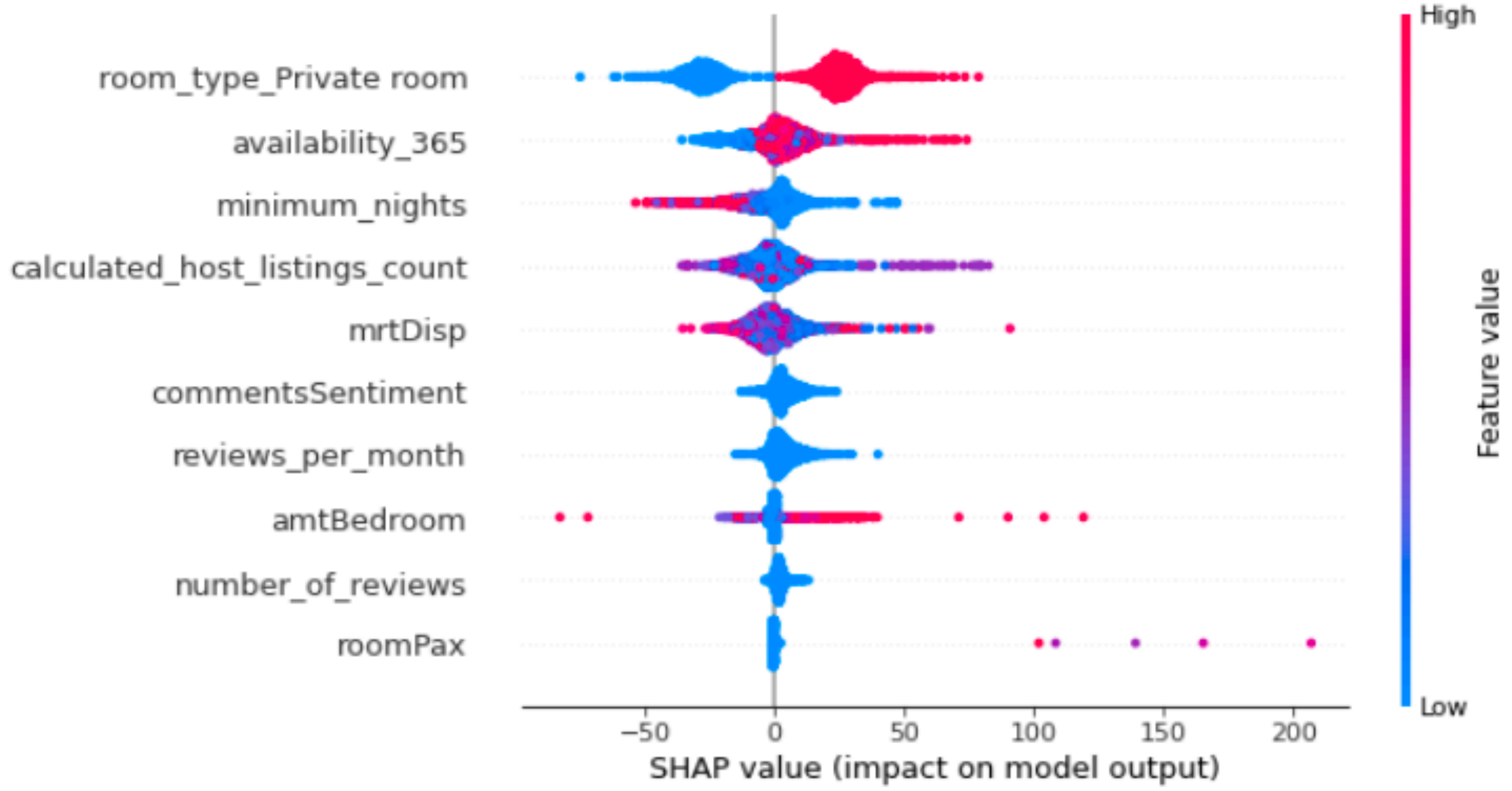
Species counts are: setosa=0, versicolor=38, virginica=3
Prediction is **versicolor** as it is the majority class

Decision tree to classify one of the three flower specifies using the famous IRIS Dataset

We can approach interpretability through Explainable AI.
This means, to generate explanations for AI's behaviors.

Problems with Explainable Models…

- The best performing solutions do not match with people's mental model
  - In other words, SHAP and similar methods explain which inputs caused the output. But that's simply not how non-technical people think.
  - You have to learn how to interpret the outputs that serve to interpret the models…



SHAP output based on airbnb data

NOV 1, 2021 • 12 MIN READ • EXPLAINABLE AI

Explaining Machine Learning Models: A Non-Technical Guide to Interpreting SHAP Analyses

With interpretability becoming an increasingly important requirement for machine learning projects, there's a growing need for the complex outputs of techniques such as SHAP to be communicated to non-technical stakeholders.

With LLMs

- You can probe, question, and interpret LLMs just like how to do with a person.
- “I don’t think this is correct. What’s your evidence?”
- “Can you provide citations for these ideas that you just gave me?”
- “Is this true?”
- “Can you check your arguments to see what might be possible factual issues?”

Example with Color

- The experiment investigates the structural alignment between color term representations derived from text and a perceptual color space known as CIELAB. So that we can see whether language models, trained solely on text, can encode the perceptual structure of color without direct sensory grounding.

- The 3D CIELAB color space is constructed using three dimensions: L (lightness), A (position between red and green), and B (position between blue and yellow).

- Distances in this space correspond to perceptual differences in color.

- Compare the linguistic color term representations extracted from language models with human perceptual color differences.

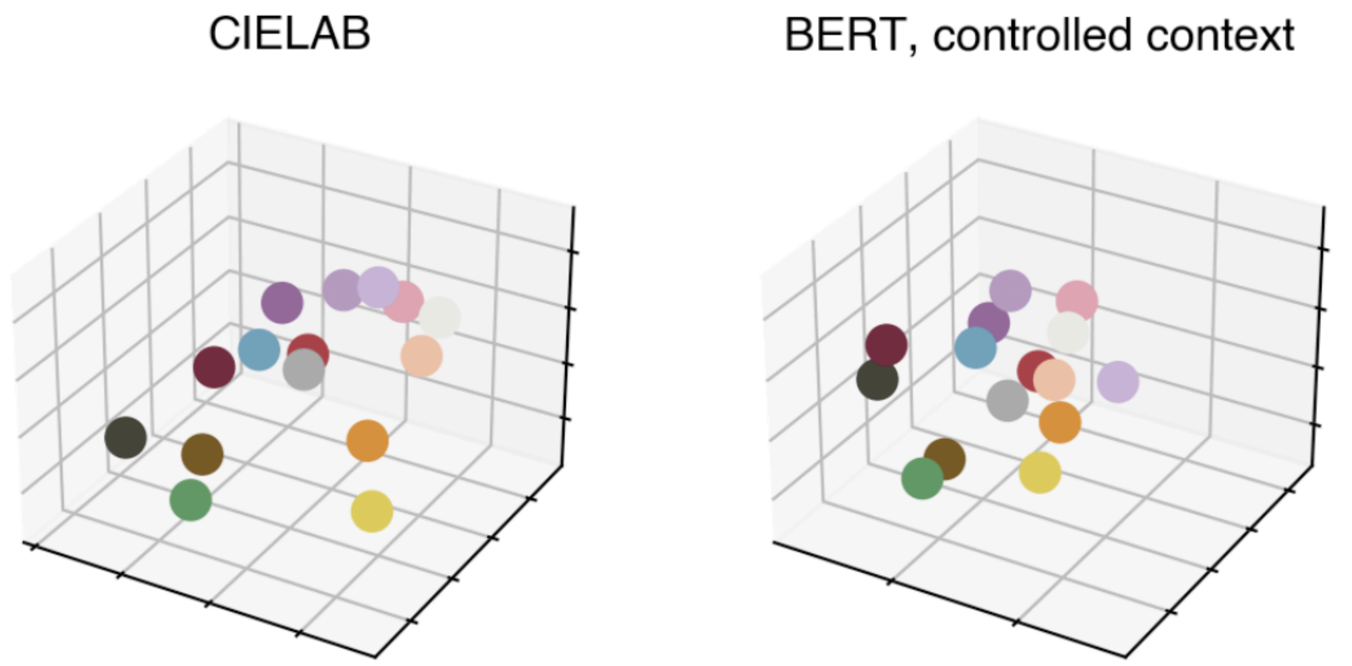
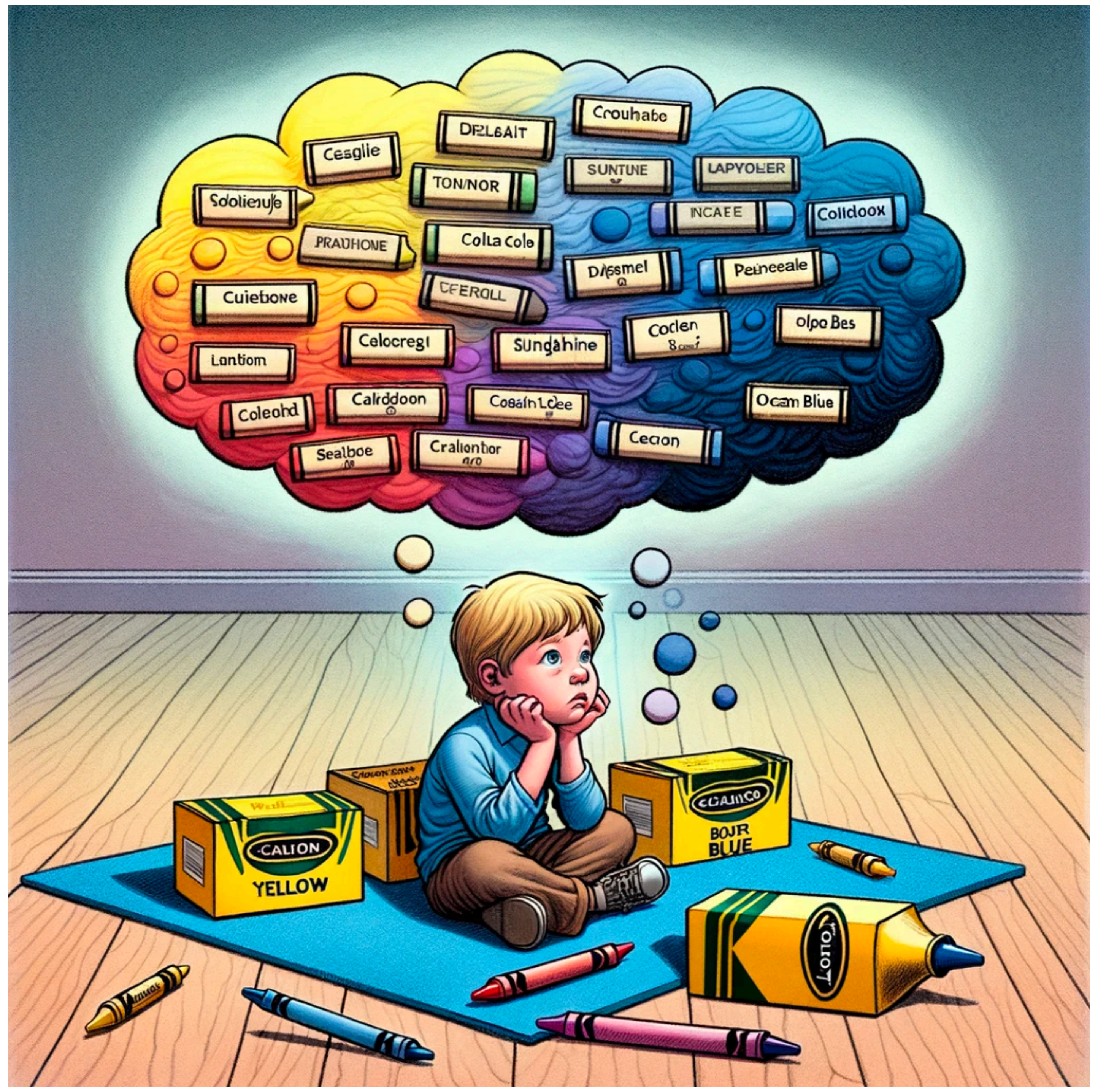


Figure 1: Right: Color orientation in 3d CIELAB space. Left: linear mapping from BERT (CC, see §2) color term embeddings to the CIELAB space.

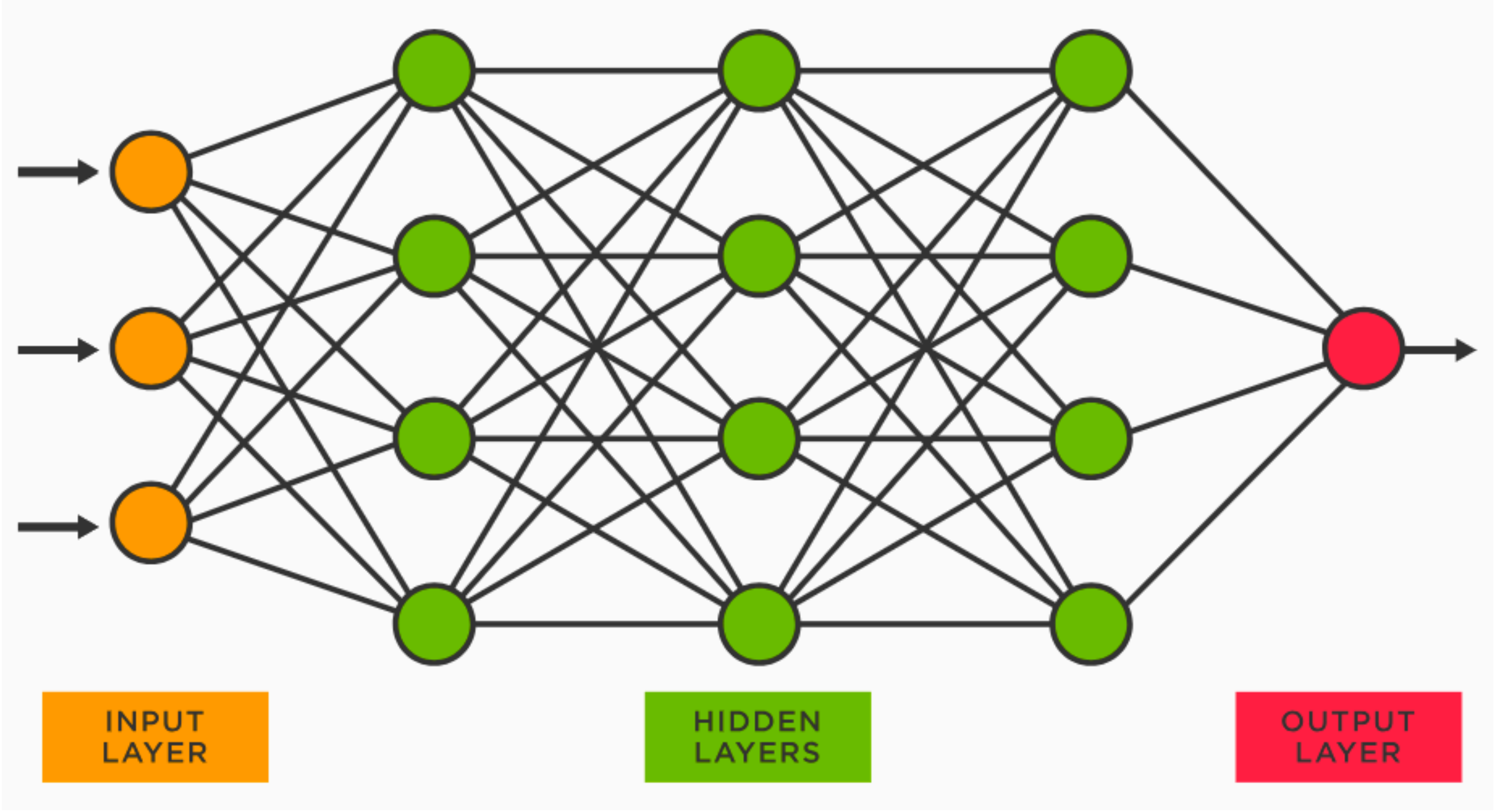


Abdou, Mostafa, et al. "Can language models encode perceptual structure without grounding? a case study in color." arXiv preprint arXiv:2109.06129 (2021).

Self-awareness - do LLMs know its wrong output?

- Hypothesis: truth or falsehood of a statement should be represented by, and therefore extractable from, the LLM's internal state.

- Interestingly, retrospectively "understanding" that a statement that an LLM has just generated is false does not entail that the LLM will not generate it in the first place.



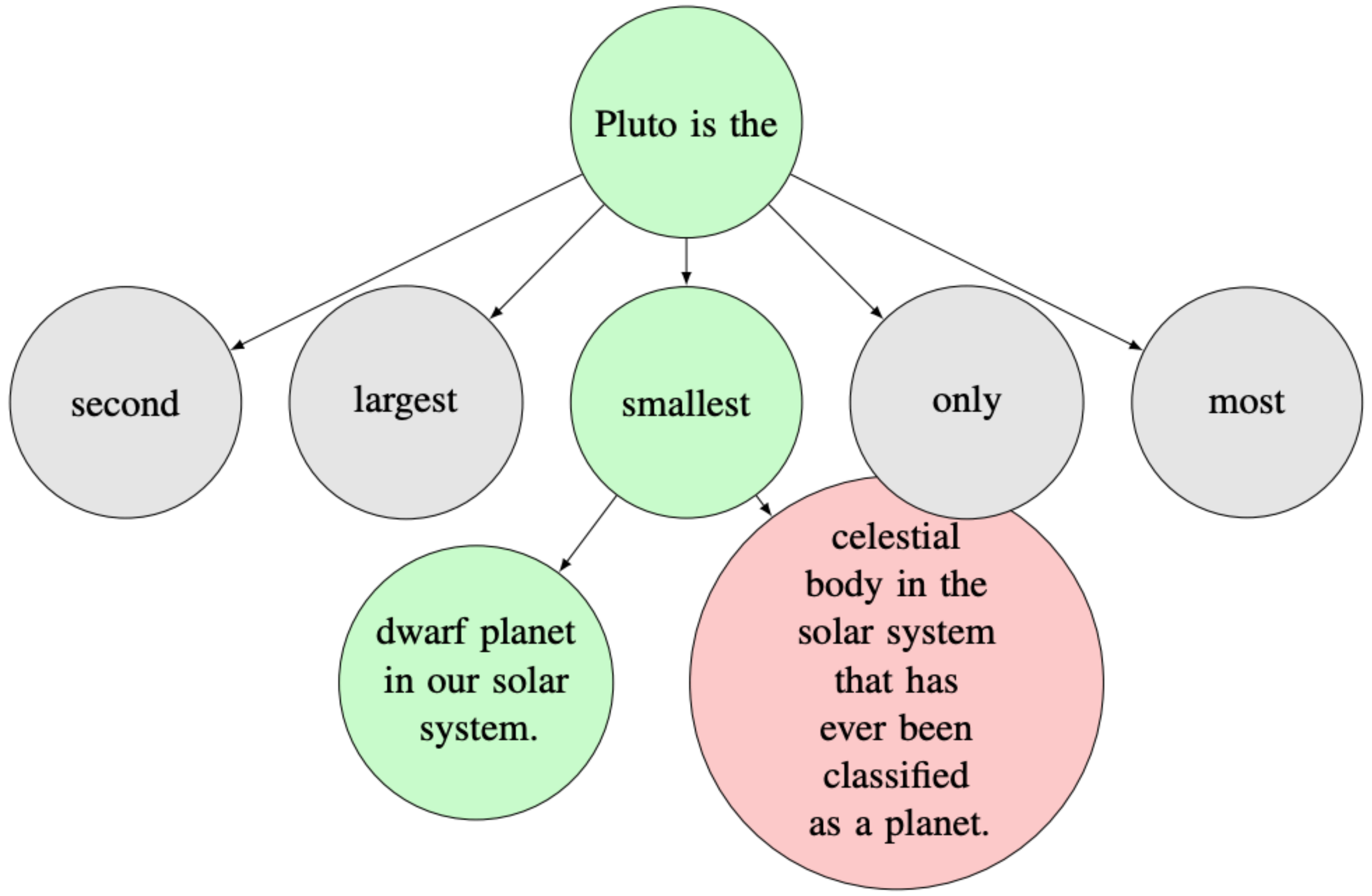


Figure 1: A tree diagram that demonstrates how generating words one at a time and committing to them may result in generating inaccurate information.

Azaria, Amos, and Tom Mitchell. "The internal state of an llm knows when its lying." arXiv preprint arXiv:2304.13734 (2023).