



Autonomous Capability Assessment of SDMs in Stochastic Settings

Problem Statement

1. Understanding capabilities of AI systems with sequential decision-making (SDM) is challenging due to their black-box nature.
2. Users need a reliable approach to assess the capabilities of such AI systems for safe and effective use.
3. Existing methods for assessing AI systems with SDM capabilities are limited, hindering inclusivity and real-world deployability.
4. There is a lack of user-driven capability assessment approaches for AI systems with SDM capabilities.
5. An active-learning approach is needed to interact with black-box SDM systems and learn interpretable probabilistic models of their capabilities in stochastic settings.
6. The learning process should guarantee convergence to the correct model and be sample-efficient for assessing arbitrary black-box SDM agents.

Example

- Self-driving cars and autonomous drones - Learning and assessing the capabilities of these systems in real-time is crucial for safe and efficient operation.
- AI-powered virtual assistants and chatbots - Assessing the capabilities of these systems in understanding and responding to user queries in uncertain and dynamic contexts is crucial for improving user experience and performance.
- Autonomous medical devices, diagnostic systems, and healthcare robots - Learning and assessing the capabilities of these systems is critical for ensuring patient safety and effective healthcare delivery.

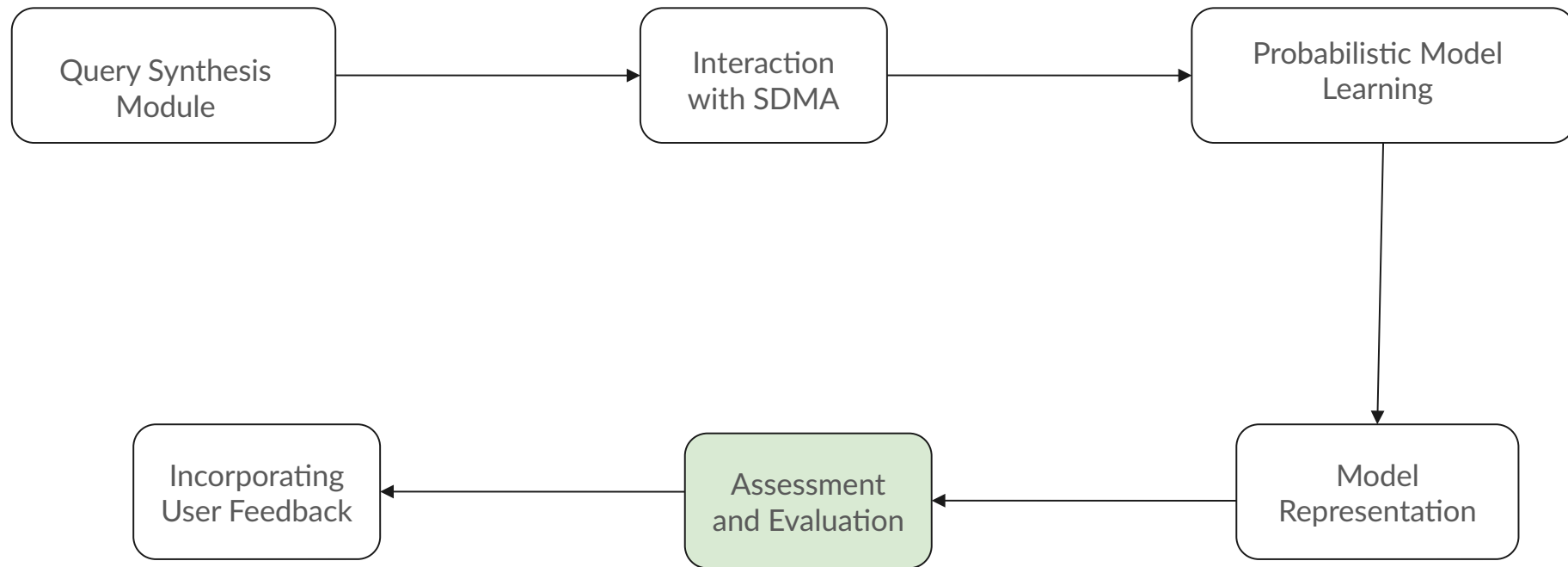
High - Level Approach



Focus on:-

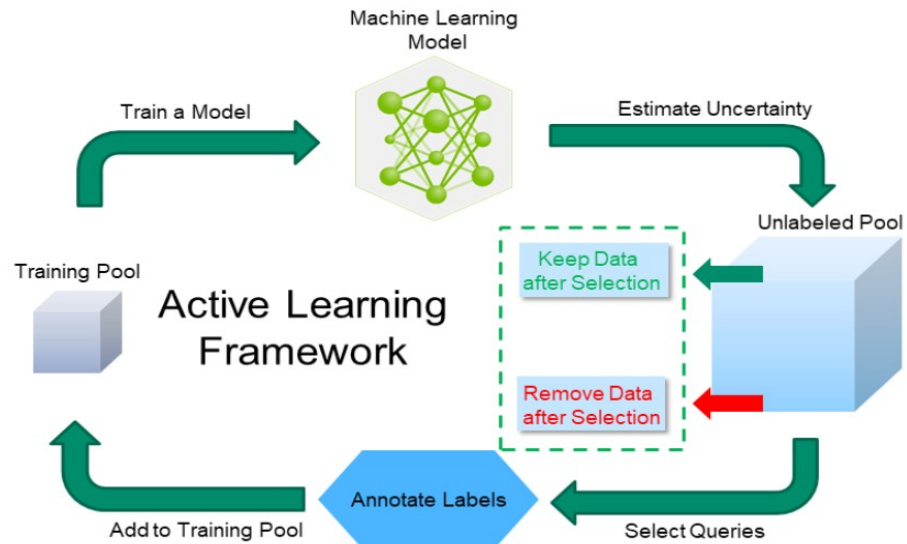
1. Active Learning and Dataset Creation
2. Query Based Learning
3. Probabilistic model learning to capture interpretable representations of the SDM system's capabilities
4. Evaluation and Validation

Flow Diagram

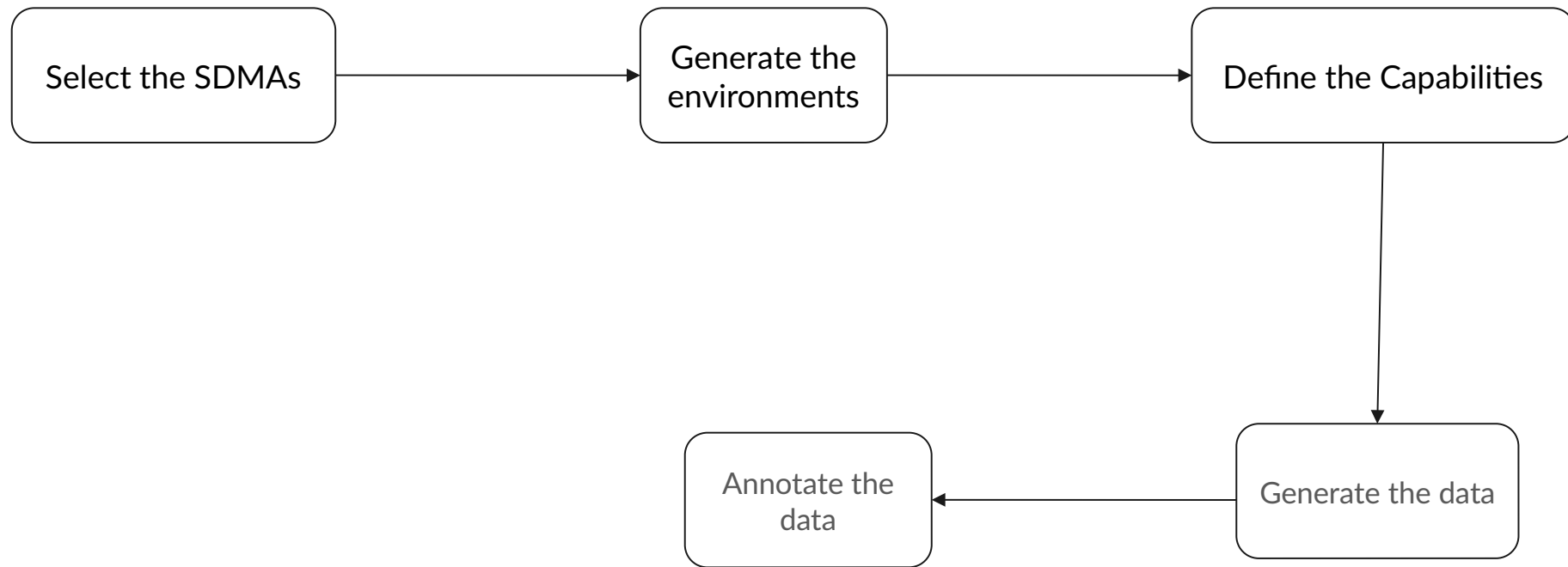


Active Learning

Develop an active learning framework that can interact with the black-box SDM system to learn about its capabilities and the effects of executing those capabilities in stochastic environments.




Dataset Creation



Query-Based Learning

- Formulate a method for generating queries to the black-box SDM system to gather information about its capabilities and the stochastic effects of executing those capabilities.
- Design the queries to be logically inconsistent to effectively learn from the responses and refine the model of the SDM system's capabilities.

Query-Based Learning Methods

- 
1. Distinguishability-based Querying
 2. Information Gain-based Querying
 3. Value-based Querying
 4. Goal-based Querying

Query Synthesis Approaches



- **Active Learning with Information Gain:** This approach involves selecting queries that maximize the expected information gain about the SDMA's capabilities. This could involve using a multi-valued measure like information gain to evaluate the utility of each query and selecting the query with the highest expected gain. This approach has been used in previous work on active learning for model learning [Settles, 2012].
- **Goal-Driven Sampling:** This approach involves generating queries based on the goals or objectives of the SDMA. Specifically, the module could use a goal-driven sampling strategy to generate queries that are relevant to the current goal or objective of the SDMA. This approach has been used in previous work on learning probabilistic models of SDMA's [Chitnis et al., 2021].

Interaction with the SDMAs - Approaches


- **Guided Forward Search for Interaction**
- **User-Guided Interaction** - In some scenarios, involving users in the interaction process can be beneficial. This approach involves incorporating user feedback and domain knowledge to guide the interaction with the SDMA. Users can provide input, correct misconceptions, and steer the interaction towards relevant queries based on their understanding of the SDMA's capabilities and the task at hand.
- **Goal-Driven Interaction** - Another approach involves structuring the interaction based on the goals or objectives of the SDMA. By formulating queries and interactions that are aligned with the current goal or objective of the SDMA, it is possible to gather targeted information about its capabilities and behavior.

Probabilistic Model Learning



Learn a probabilistic model that captures the high-level user-interpretable capabilities of the SDM system, along with the probabilities of possible outcomes of executing each capability in stochastic environments.

Probabilistic Model Learning Methods

- 
1. Maximum Likelihood Estimation (MLE)
 2. Expectation-Maximization (EM) Algorithm
 3. Bayesian Inference
 4. Gibbs Sampling
 5. Variational Inference

Evaluation and Validation




- Utilize Benchmark Datasets
- Using simulated scenarios and real-world robot platforms
- Perform Cross-Validation on the available dataset
- Transfer Learning
- Adversarial Training
- Human-in-the loop evaluation
- Define Evaluation Metric: This will include measures of accuracy, precision, F1-score, recall, and other relevant metrics based on the nature of the probabilistic model.

Research Questions

1. How can we optimize the query synthesis module to generate more effective queries for learning the SDMA's capabilities?
2. How can we improve the sample efficiency and accuracy of the probabilistic model learning process?
3. What are the limitations of the existing approaches, and how can they be addressed in the proposed technique?
4. How does the proposed technique compare to existing approaches for learning probabilistic models of black-box SDMA's in terms of accuracy, sample efficiency, and interpretability?
5. How can we extend the proposed technique to handle more complex and dynamic environments, such as those involving multiple agents or changing task objectives?

Timeline

- 
1. Active Learning - Nirali, Shivani
 2. Query Learning - Akash, Rahil
 3. Probabilistic model learning to capture interpretable representations of the SDM system's capabilities - Chelsi Jain