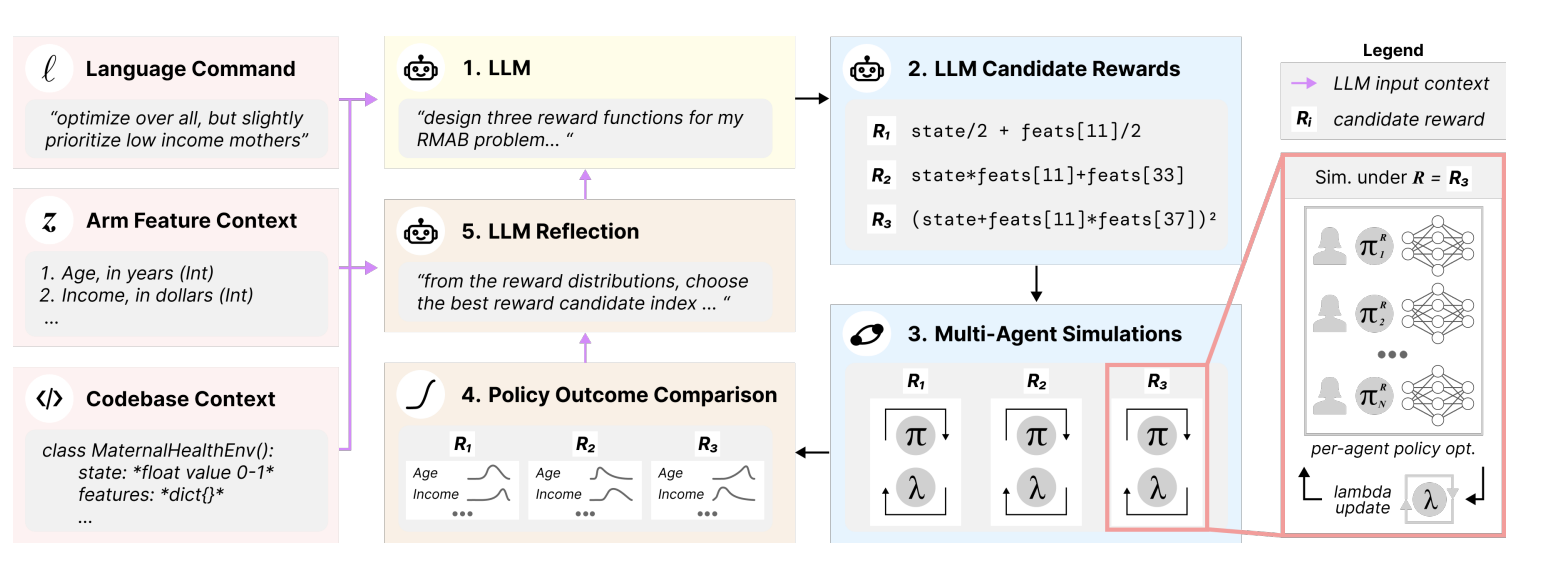
Images:

1. **Intro diagram flow chart (example above):** Show the end-to-end pipeline of our model and how it can be used to detect spurious correlations
2. **Prove that we have spurious models:** Bugged model vs normal model accuracy on background-only – Table

|  | Bugged Model | Normal Model |
| --- | --- | --- |
| Medium Prediction Accuracy | 27.5% | 45% |
| Easy Prediction Accuracy | 98.75% | 26% |

1. **Show qualitatively that the models are spurious:** Concepts for Target Images and Background Images – Images
2. **Show that we are correctly filtering for background concepts:** For each model, a bar plot that is x bars (2 for each concept: one bar for activation on the class dataset, one for activation on the background dataset) -- on the X-axis, we use heatmaps to label the concepts. This shows that activations are very similar between class / bg images in the bugged model, and very different on normal models. Color code by background / class concepts
   1. Some sort of arrow: show that we use importance of bg concepts to create spurious scores (formula)
3. **Show that the spurious score can flag a bugged model:** Methods for aggregating concept activation – Grouped bar graph
4. **We achieved our goal: the spurious score has an accuracy of xx%** How accurate is the spurious detection model (how many times does it accurately detect that a spurious detection has been learned)

* **Do our results Generalize?** (need to run these)
  + Across diff numbers of concepts -- scatter plot: number of concepts on the x axis, the spurious scores accuracy on the y axis
  + Across classes -- show class, by class, the spurious score
    - Need to re run (concept generation, spurious test, spurious score) for each class on our original model
      * Save concept folder
      * Save spurious test results
      * Save spurious score for each class
  + Across models -- could show a scatter plot: accuracy on the spurious test vs. spurious score (nice to have, not a must have)
    - Need to re run training on diff dataset versions (o2o, m2m, easy/medium), then run: (concept generation, spurious test, spurious score) for each model, each class
  + Across concept activation methods -- swap out nonnegative matrix factorization for PCA or SVD, re run whole experiment (don’t need this one)

Discussion:

1. Stress that we used normal model as control -- but you wouldn’t need a normal model to run this method
2. Discuss the importance of the accuracy of our spurious detection model
3. Next step: run a user study where people have to detect the spurious model just by looking at heatmaps (or at concept activation maps)

To discuss:

1. Why we chose mean as the concept aggregating method
   1. Show that for certain models/image datasets, may need to choose other aggregating concept activations (?)