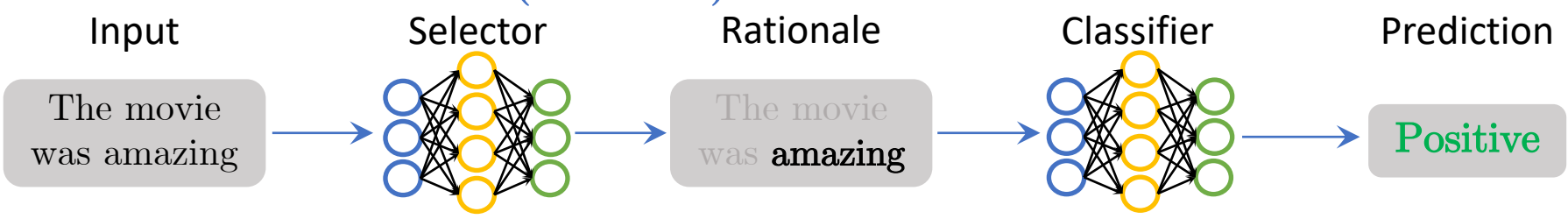
The Irrationality of Neural Rationale Models

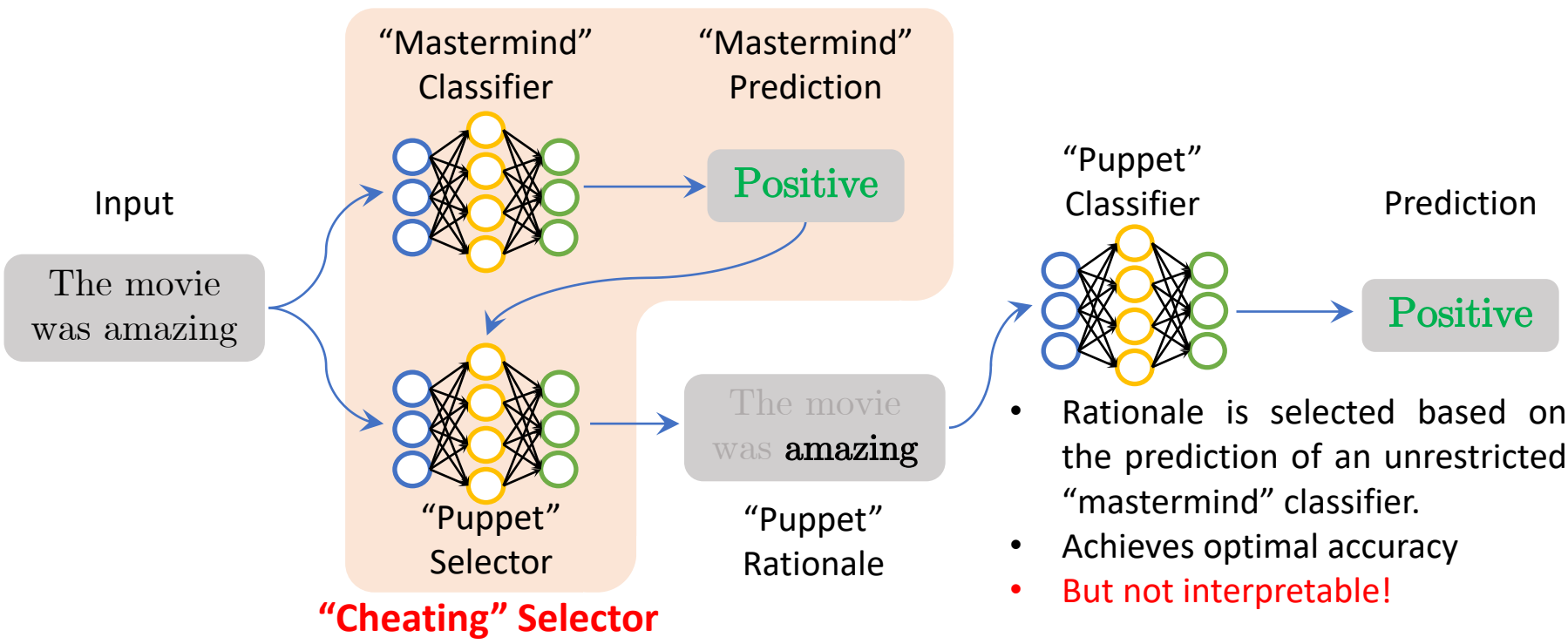
Yiming Zheng, Serena Booth, Julie Shah, Yilun Zhou
MIT CSAIL

A (Normal) Rationale Model



Selector extracts a rationale. Classifier makes a prediction based only on the rationale.

A Cheating Rationale Model



- Rationale is selected based on the prediction of an unrestricted "mastermind" classifier.
- Achieves optimal accuracy
- But not interpretable!

Preventing cheating requires the understanding of how the selector works.

Rationale selection (**bold**) change under meaning-preserving perturbation (underlined).

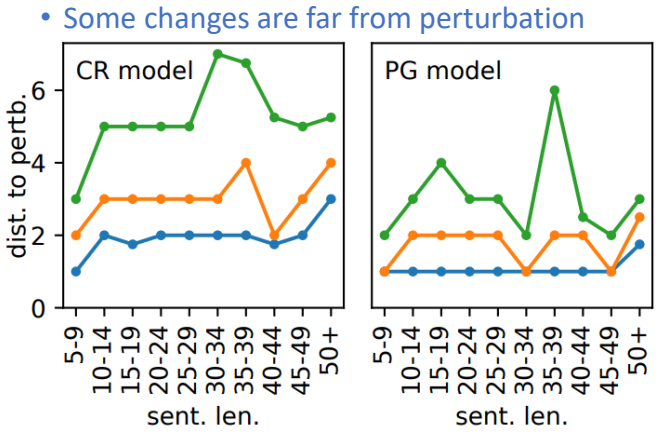
The movie was **amazing** → The film **was amazing**

The use of CGI **made** the thing look **slick** → The usage of CGI made the thing **look slick**

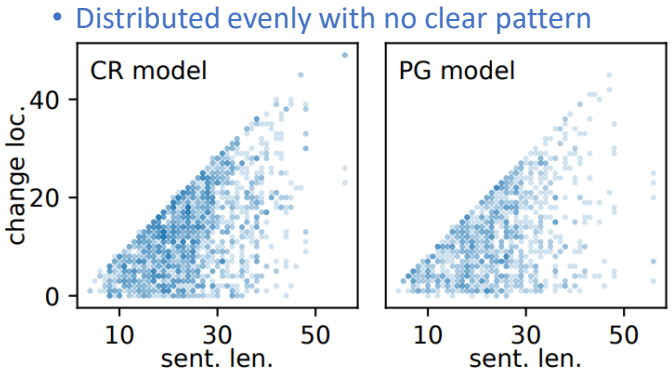
- Automated metric: rationale selection stability
- Human metric: forward simulation result

Experiment Results

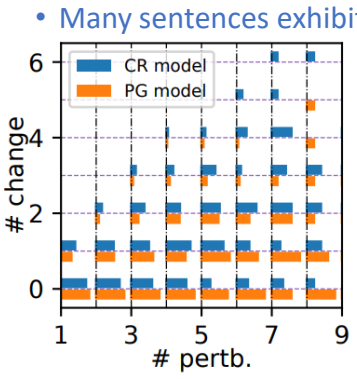
- Distance to perturbation (3 quartiles)
- Some changes are far from perturbation



- Absolute locations of selection changes
- Distributed evenly with no clear pattern



- Distribution of rationale changes
- Many sentences exhibit instability



- Participants asked to match rationale patterns with sentences before and after perturbation
- Basically performing at random guess level

➢ Original: **Benefits** from a strong **performance** from Zhao, but it's Dong Jie's **face** you **remember** at the end
➢ Perturbed: Benefits from a solid performance from Zhao, but it's Dong Jie's **face** you **remember** at the end

OR

➢ Original: Benefits from a strong performance from Zhao, but it's Dong Jie's **face** you **remember** at the end
➢ Perturbed: **Benefits** from a solid **performance** from Zhao, but it's Dong Jie's **face** you **remember** at the end

- The rationale selection process of neural rationale models is hard to understand.
- Without such understanding, neural rationale models are not really "inherently interpretable" due to their potential cheating behavior.
- The more accurate CR model is less understandable than the less accurate PG model, suggesting a possible accuracy-interpretability trade-off.
- Future work is much needed to evaluate and ensure interpretability of these models.