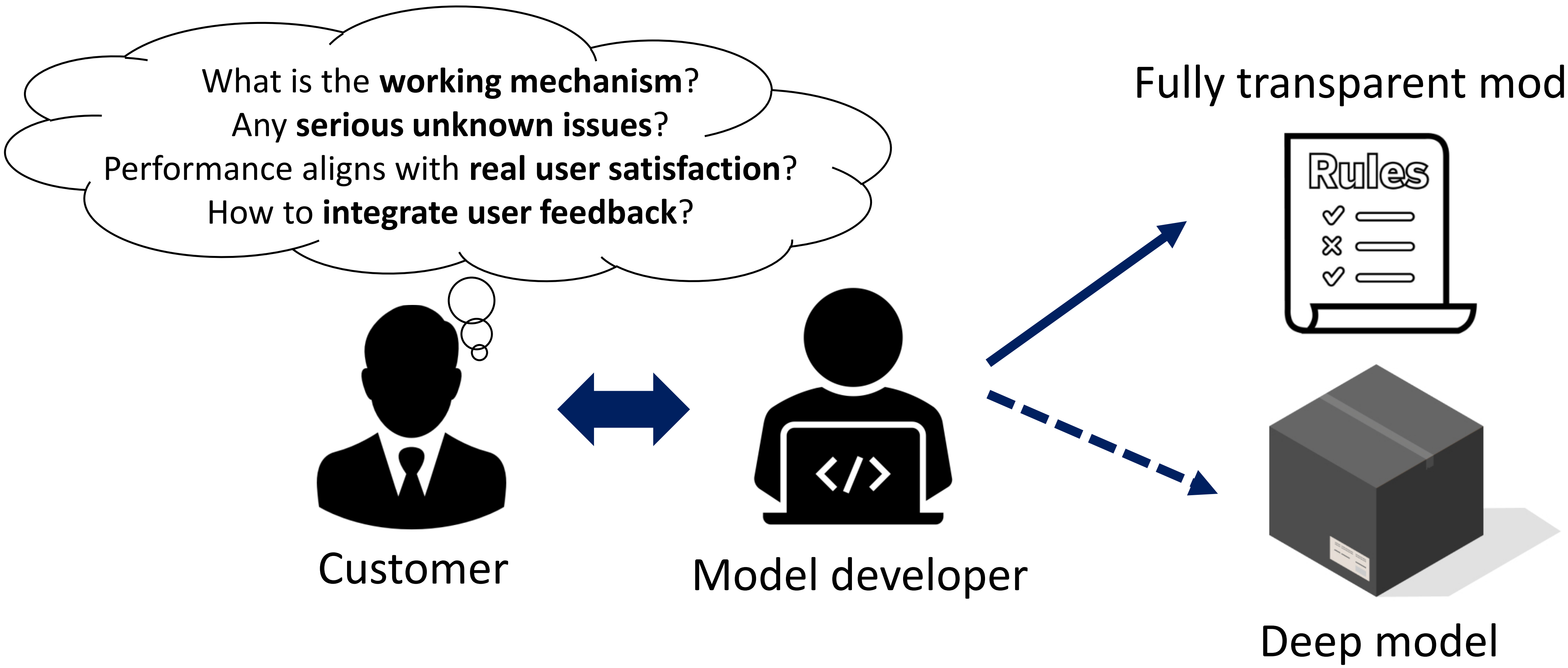
# Self-explaining deep models with logic rule reasoning

Seungeon Lee, Xiting Wang, Sungwon Han,

Xiaoyuan Yi, Xing Xie, Meeyoung Cha

KAIST Microsoft ibS

# Trust Issues with Deep Models



What is the **working mechanism**?
Any **serious unknown issues**?
Performance aligns with **real user satisfaction**?
How to **integrate user feedback**?

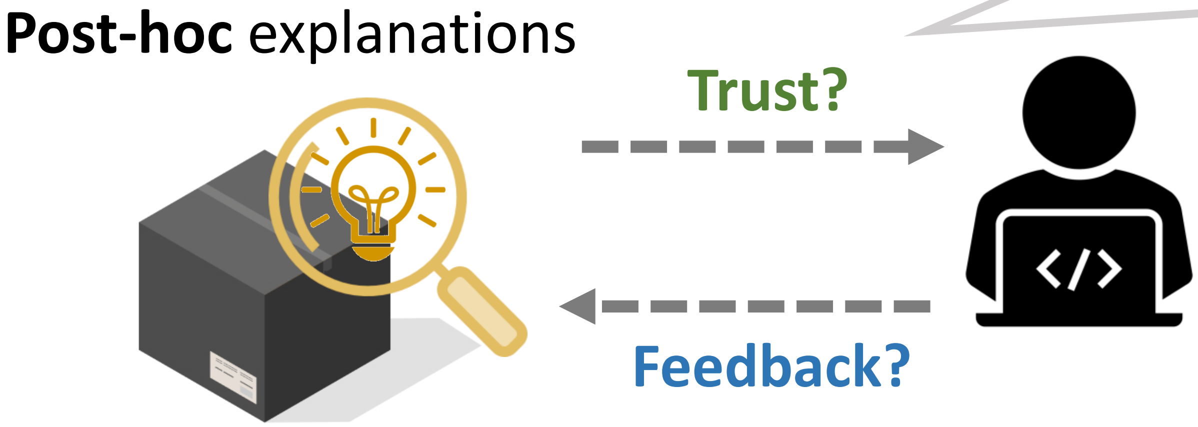
Customer

Model developer

Fully transparent model

Deep model

# Limitation of Post-Hoc Explanations

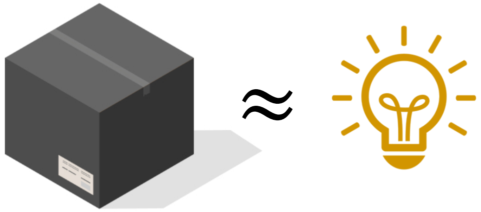


**Post-hoc** explanations

Trust?

Feedback?

### Can we trust the explanations?



- Always an approximation [1]
- “General uneasiness” of practitioners [2]

### How to integrate user feedback?

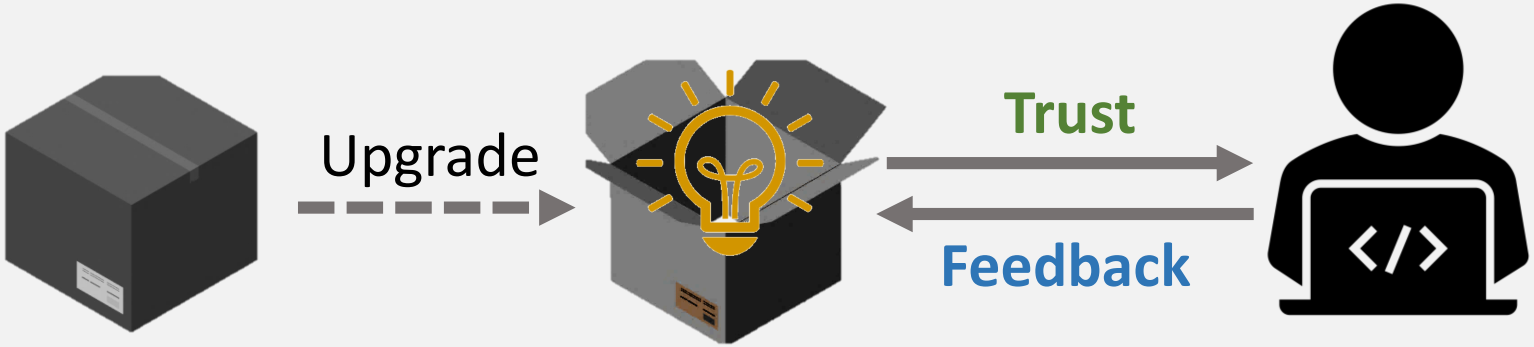
- No systematic method for direct control
- Requires model retraining
- No guarantee for satisfying user demands

[1] “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”, *Nature Machine Intelligence*, 2019

[2] “Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs”, *ACM HCI* 2020

# Our Method: SELOR

Self-
Explaining
with
LOgic rule
Reasoning

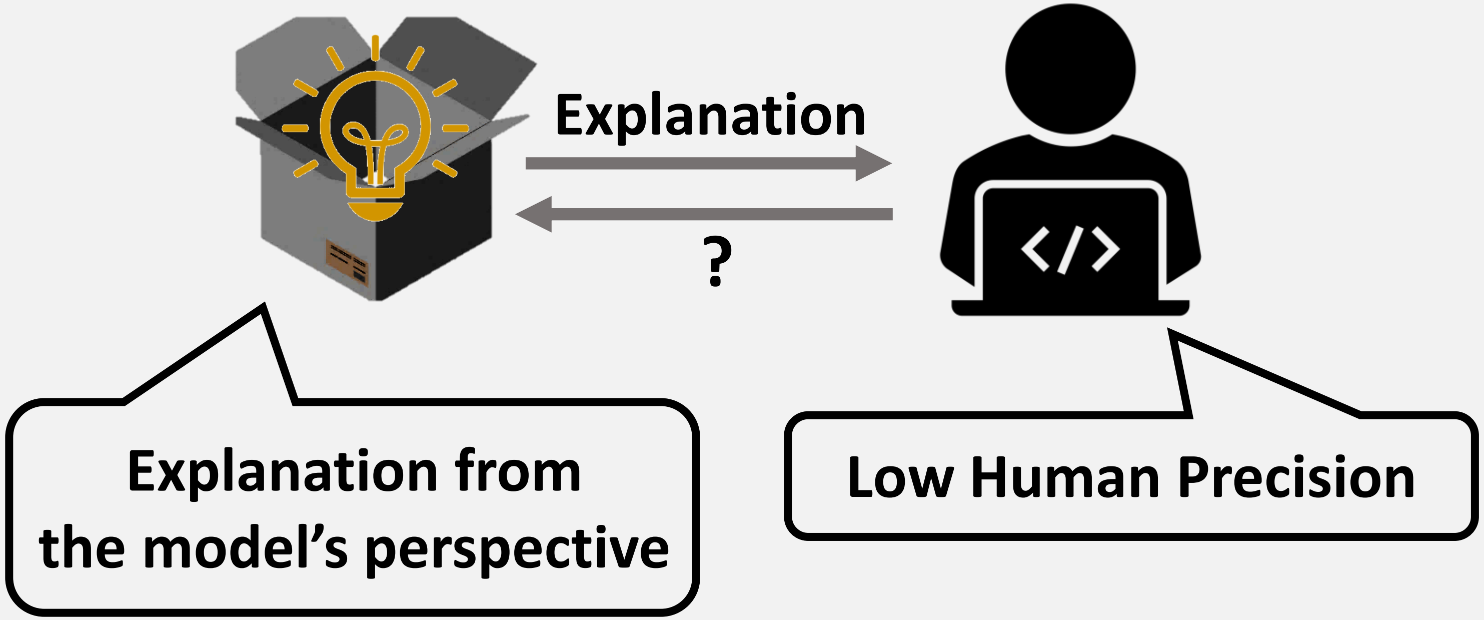


Lays the **foundation** for close collaboration

- Trust: explanations **faithful** to the model
- Feedback: explanations as **handle** for control

# Our Method: SELOR

**S**elf-
**E**xplaining
with
**LO**gic rule
**R**easoning



Explanation from the model's perspective

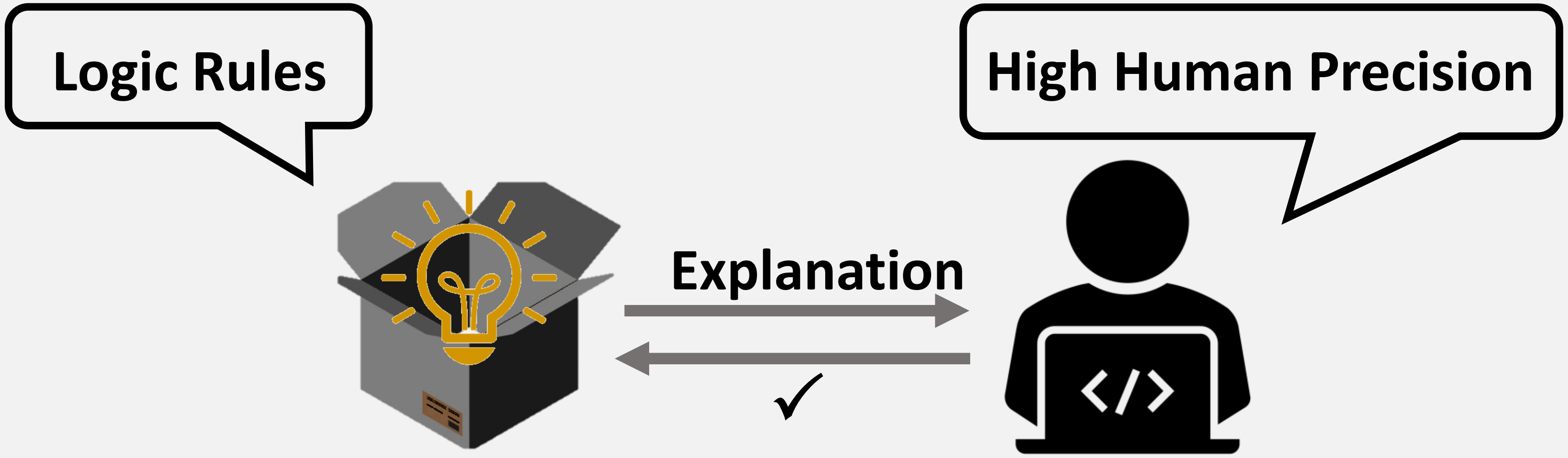
Low Human Precision

**Human Precision**: Whether the explanation naturally leads to the prediction according to human perception

| | |
|---|---|
| *Low Human Precision:* | *is, an => positive sentiment* |
| *High Human Precision:* | *Awesome => positive sentiment* |

# Our Method: SELOR

**S**elf-
**E**xplaining
with
**LO**gic rule
**R**easoning



- Close to human decision logic
- Widely applied for making predictions
- Require minimum human effort

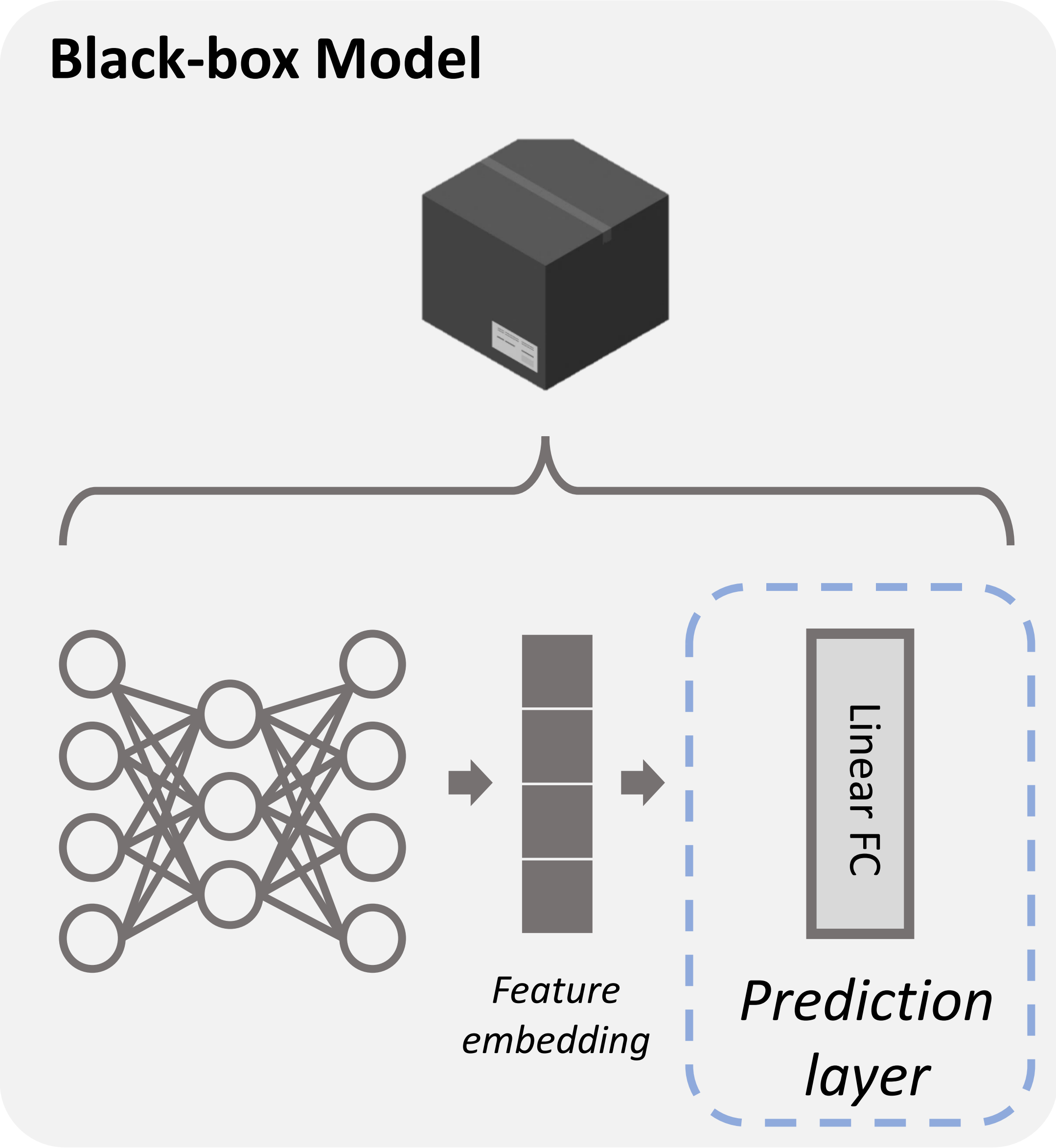
$\underbrace{\textit{awesome AND tasty}}_{\textit{Antecedent}}$ → $\underbrace{\textit{positive sentiment}}_{\textit{Consequent}}$

(condition to apply) (prediction result)

# Framework: Black-box Model

**Model Input $x$**

*“The food is awesome, and the service is good”*



**Make prediction without explanation**

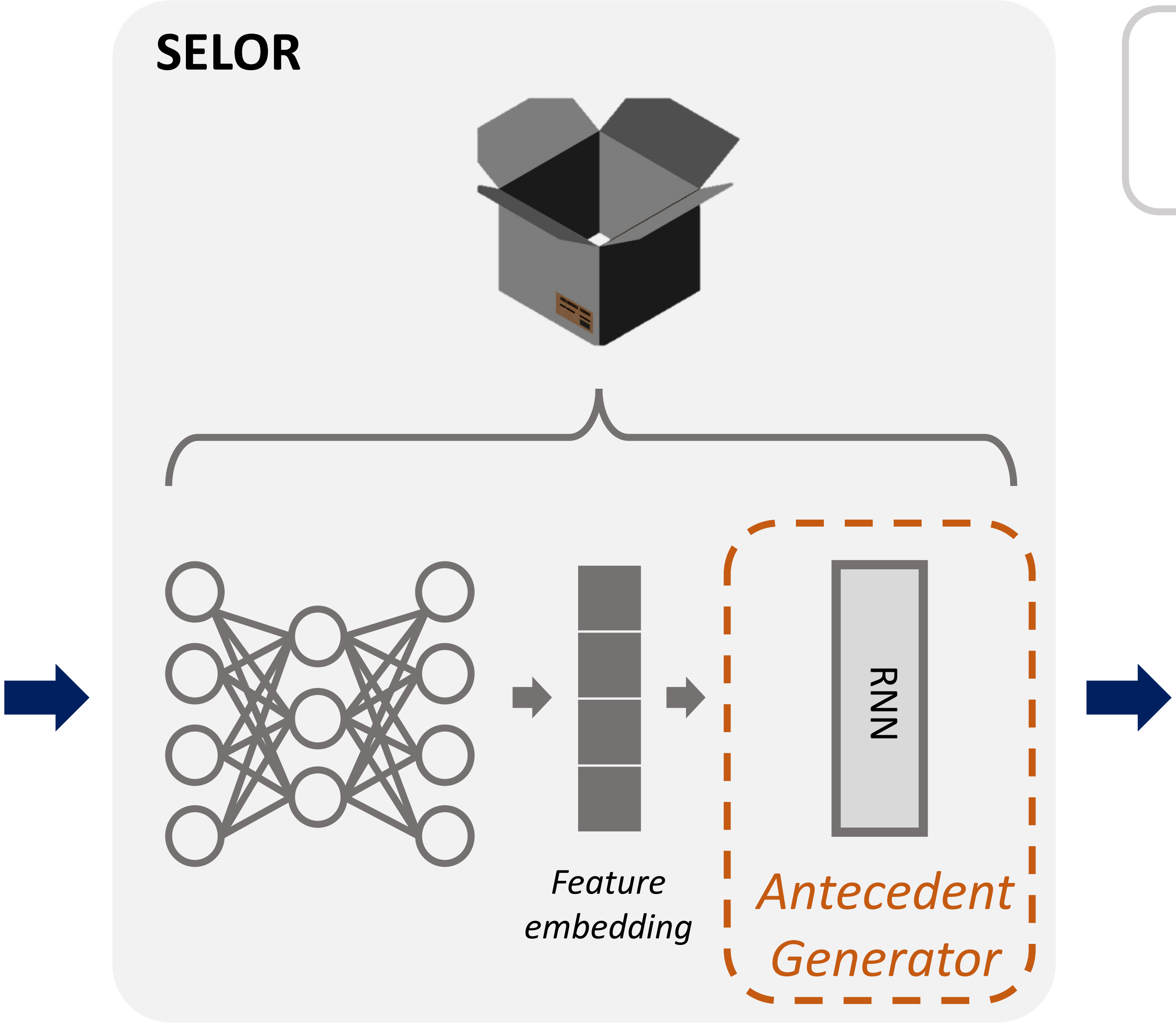
**Prediction $y$**

*Positive sentiment*

# Framework: Antecedent Generator

**Model Input $x$**

*“The food is awesome, and the service is good”*

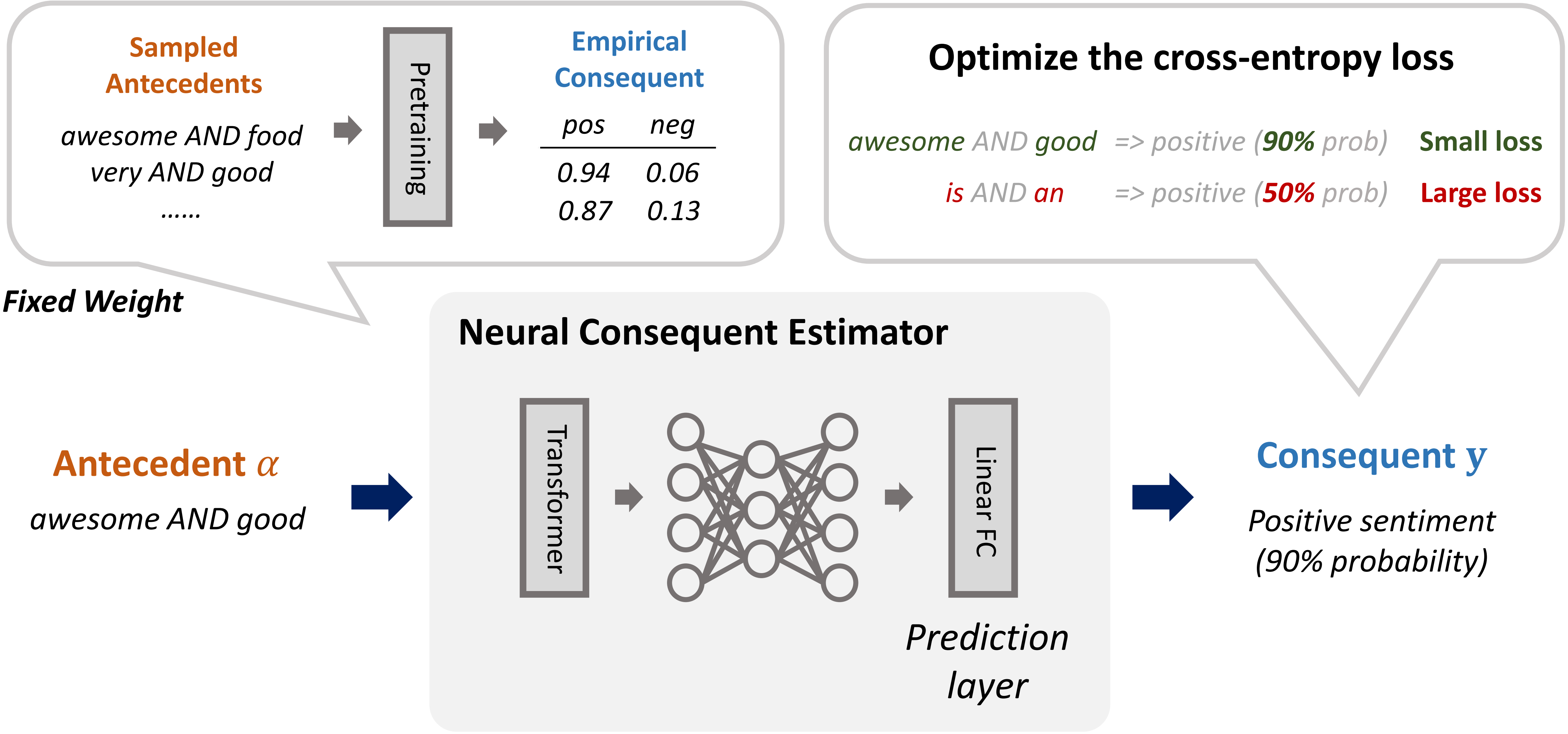


**Select Antecedent**

**Antecedent $\alpha$**

*The*
*service AND food*
*the AND food*
……
*awesome AND good*
……
*the AND is AND good*
*and AND the AND is*

# Framework: Consequent Estimator



**Sampled Antecedents**

*awesome AND food*
*very AND good*
……

**Empirical Consequent**

| *pos* | *neg* |
|---|---|
| *0.94* | *0.06* |
| *0.87* | *0.13* |

***Fixed Weight***

**Optimize the cross-entropy loss**

*awesome AND good* *=> positive (**90%** prob)* **Small loss**

*is AND an* *=> positive (**50%** prob)* **Large loss**

**Neural Consequent Estimator**

*Prediction layer*

**Antecedent $\alpha$**

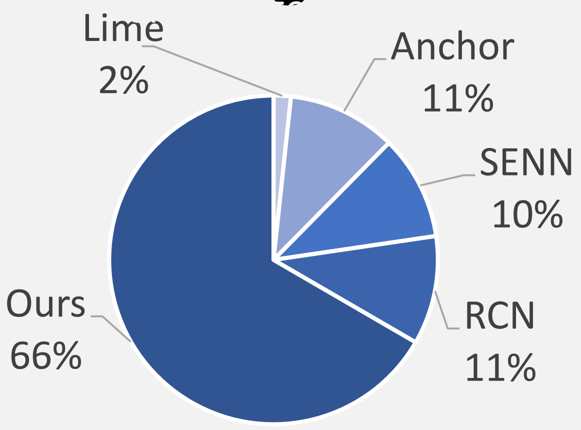
*awesome AND good*

**Consequent y**

*Positive sentiment (90% probability)*

# Results

## High Human Precision

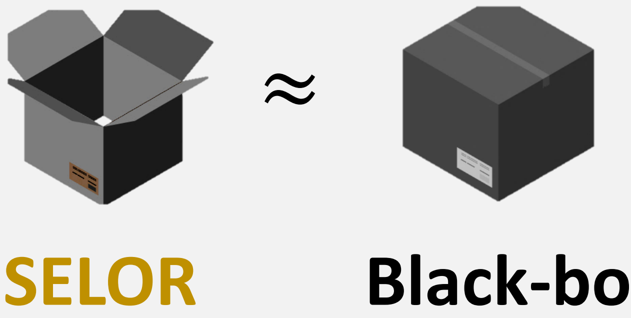


User study
Percentage of best

**+500%**

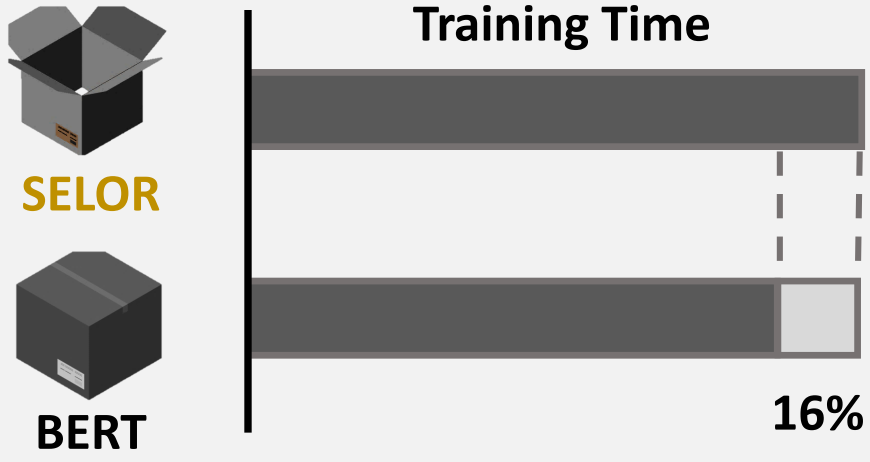
(Adult dataset)

## Good Prediction Performance



## Training Cost

- **Efficient, differentiable** training
- **Slightly slower** than black-box model



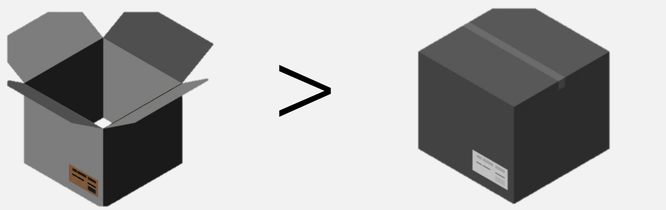
# Additional Advantages

## Generate Explanation Efficiently

**SELOR vs LIME**
1,000x speed-up

**SELOR vs Anchor**
50,000x speed-up
(BERT base, Yelp)

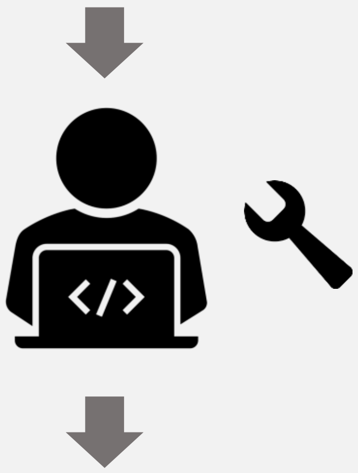
## Robust to Noisy Labels



**SELOR** **Black-box**

## Can be Steered w/o Retraining

✗ *vegas* => positive



✓ *tasteless* => negative

# Thanks & Questions

**Paper:**
https://arxiv.org/abs/2210.07024

**Codes:**
https://github.com/archon159/SELOR

**Additional comments and feedback:**
archon159@kaist.ac.kr