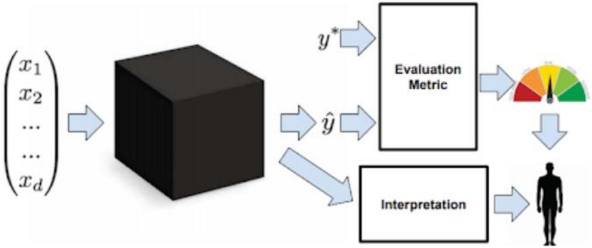
# Interpreting Pretrained Language Models via Concept Bottlenecks

*Zhen Tan, Lu Cheng, Song Wang, Bo Yuan, Jundong Li, Huan Liu*

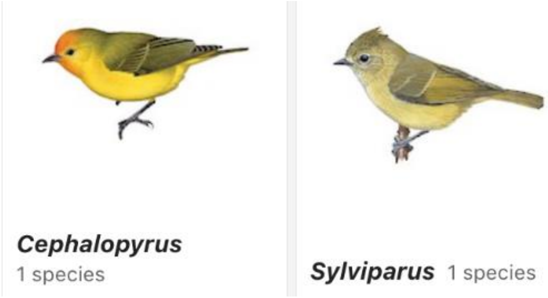
Zhen Tan, et al., **Best Paper Award** for "Interpreting Pretrained Language Models via Concept Bottlenecks", PAKDD 2024, May 7-10, 2024

# Background

## Interpretability in Deep Models




If you want users’ trust,
- open the “black box”
- show users “how” the model make such decisions in a user-friendly way

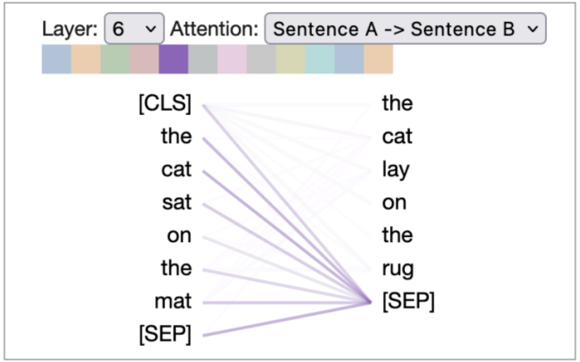



Bird species classification


Review sentiment analysis

# Existing Methods Interpreted Language Models Locally

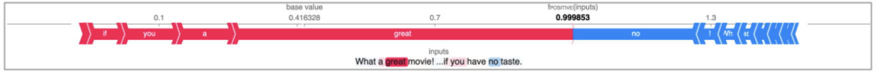
**(a) Attention Visualization**

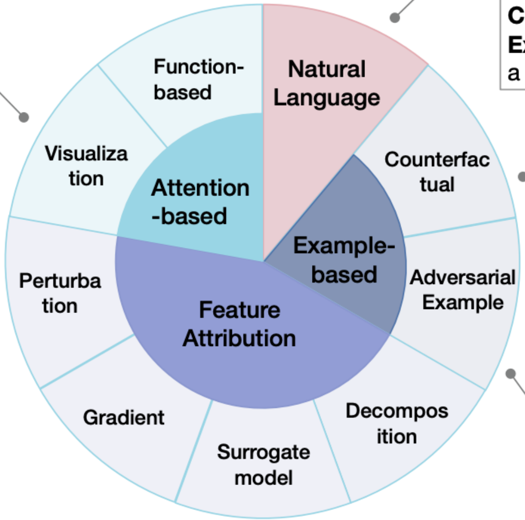



**(b) Question Answering**

**Context:** In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his **Colorado Springs experiments**.
**Question**: What ~~did~~ Tesla spend Astor's money on?
**Confidence**: 0.78 —> 0.91

**(c) Sentiment Analysis**




**(d) Commonsense Reasoning**

**Question:** While eating a **hamburger with friends**, what are people trying to do?.
**Choices**: **have fun**, tasty, or indigestion
**Explanation**: Usually a hamburger with friends indicates a good time.

**(e) Sentiment Analysis**

**Original text:** It is great for kids (**positive**).
**Negation examples**: It is not great for kids (**negative**)

**(f) Classification**

**Original text**: The characters, cast in impossibly contrived situations, are totally estranged from reality (**Negative**).
**Perturbed text**: The characters, cast in impossibly engineered circumstances, are fully estranged from reality (**Positive**)

# Limitation 1

## Can we exhaustively understand LLMs?

**Intrinsic Barriers to Explaining Deep Foundation Models**

arXiv

ZHEN TAN, Arizona State University, USA
HUAN LIU, Arizona State University, USA

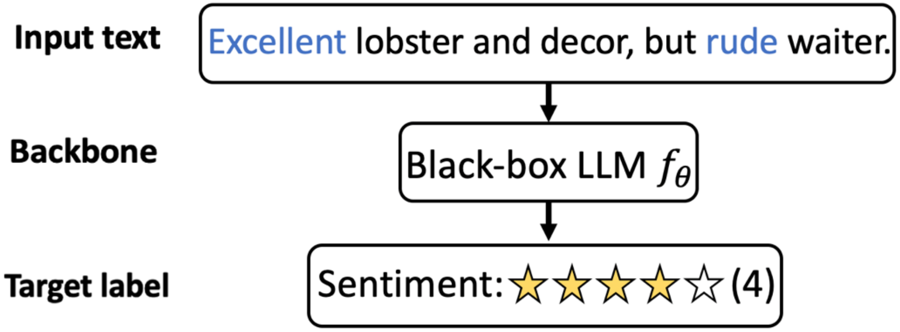
**Theorem 3.4.** *There exists a complexity gap.*

1. The complexity of explanations is bounded by human cognitive limits;
2. The complexity of deep foundation models, including LLMs, are significantly large;

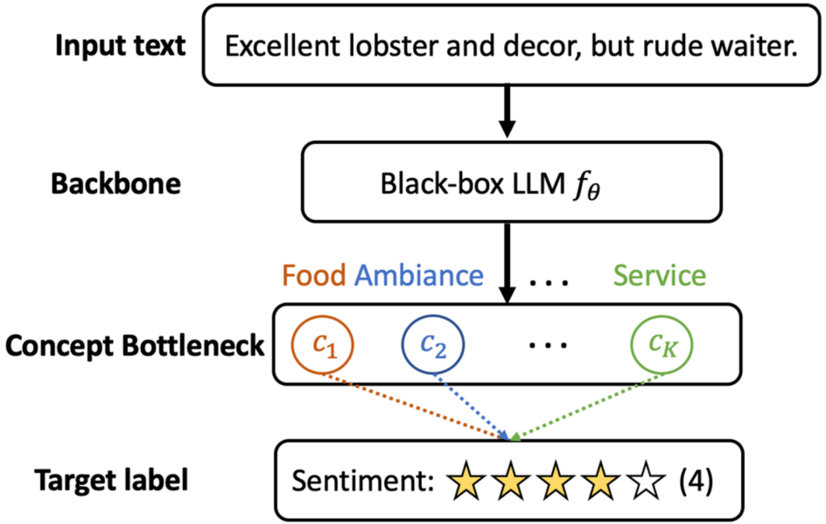
=> It is intrinsically infeasible to exhaustively explain LLMs.

# Limitation 2

## How to interpret language models globally?



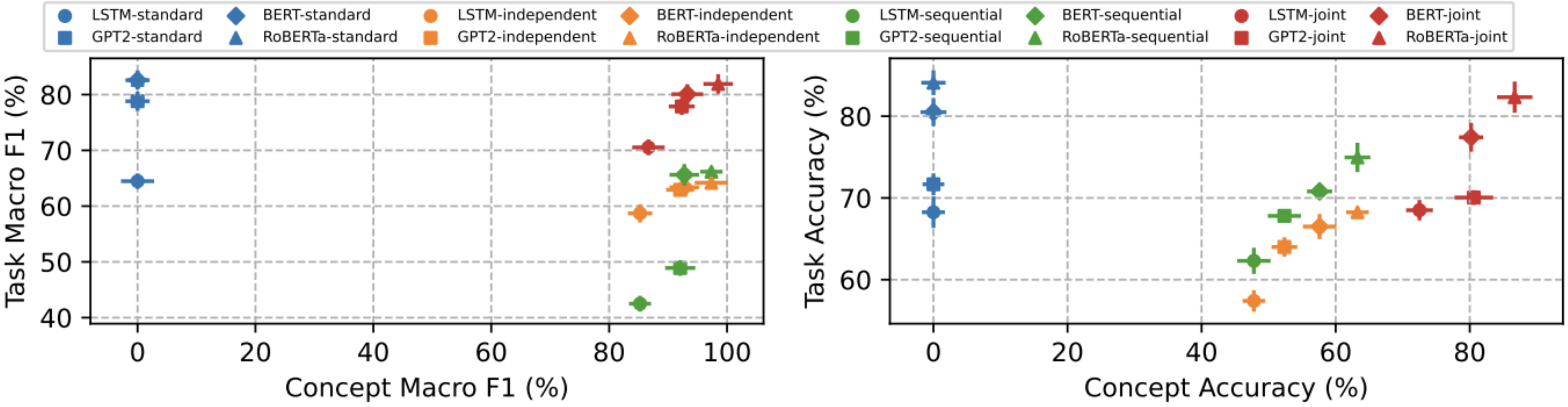

**(a) Attention-based explanation is local**




**(b) Concept-based explanation is global**

# CBE-PLMs: The interpretability-utility Pareto front

***Joint training* can achieve similar task performance while providing concept prediction**



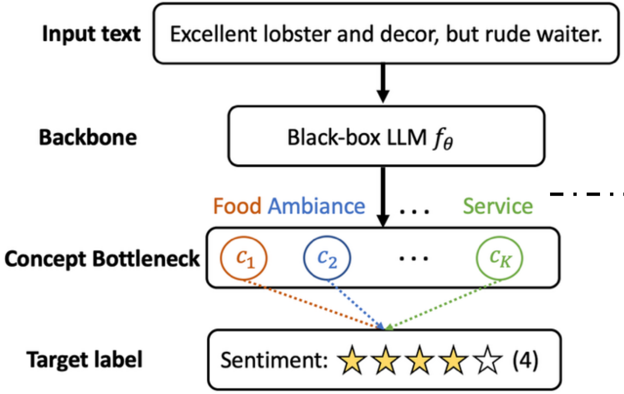

(a) CEBaB

(b) IMDB

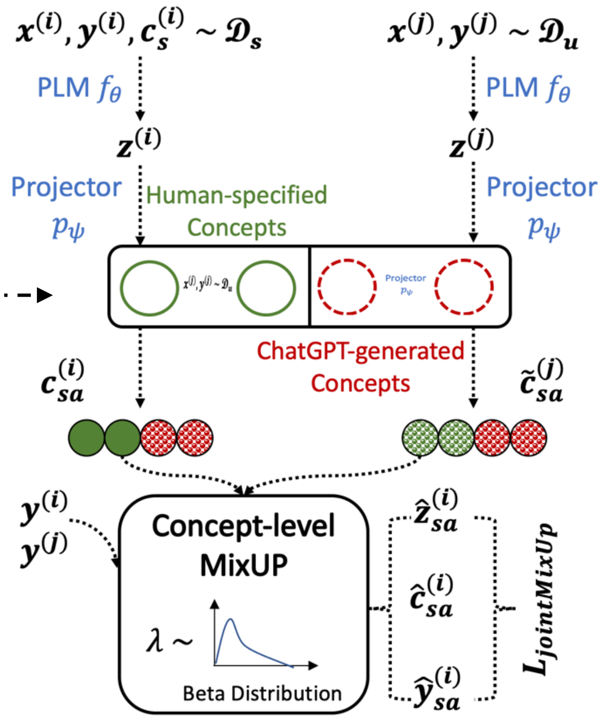
## ChatGPT-guided Concept augmentation with Concept-level Mixup (C3M)

a. According to the review "{$text_1$}", the "{$concept_1$}" of the movie is "positive".
b. According to the review "{$text_2$}", the "{$concept_2$}" of the movie is "negative".
c. According to the review "{$text_3$}", the "{$concept_3$}" of the movie is "unknown".
d. According to the review "{$text_i$}", how is the "{$concept_i$}" of the movie? Please answer with one option in "positive, negative, or unknown".

**(a) ICL-based prompting**

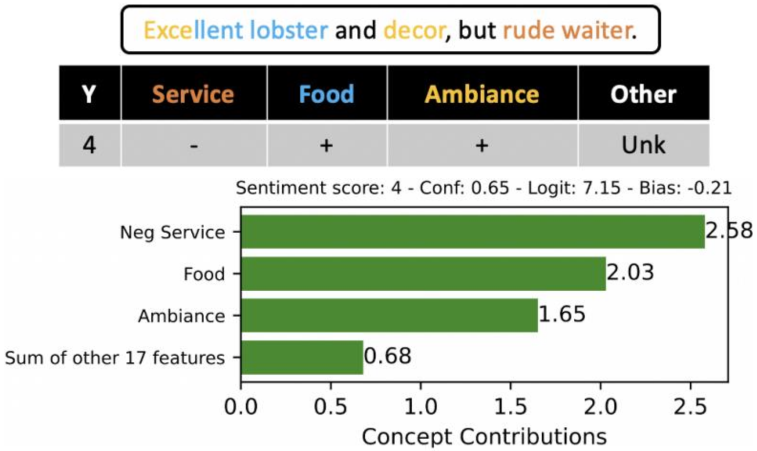



**(b) CBE-PLMs**

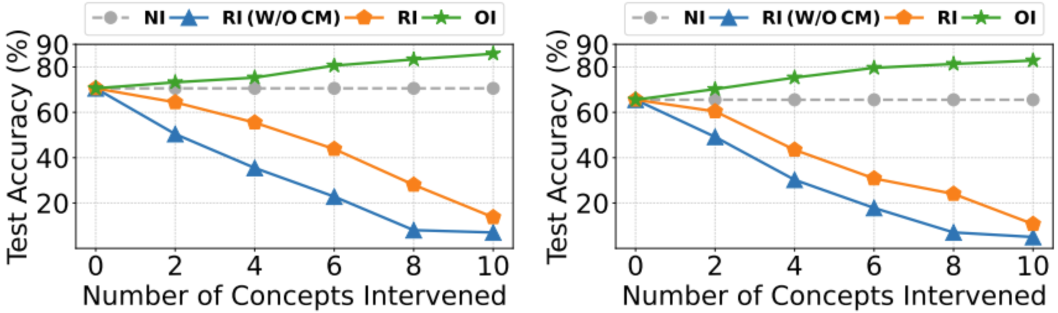



**(c) Concept-level Mixup**

## Robust Inference-Time Intervention




**Concept-level explanation**




(a) BERT (b) GPT2

**The results of Test-time Intervention. "NI" denotes "no intervention", "RI (W/O CM)" denotes "random intervention on CBE-PLMs without the concept level MixUp", "RI" denotes "random intervention on CBE-PLMs", and "OI" denotes "oracle intervention".**

**Robust Adjustments**:

1. Correct intervention improves the performance.
2. More robust to incorrect interventions.

# Experiments

## Utility and Interpretability Trade-off

| Dataset | | CEBaB | | | | IMDB | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | $\mathcal{D}$ | | $\tilde{\mathcal{D}}$ | | $\mathcal{D}$ | | $\tilde{\mathcal{D}}$ | |
| | | Task | Concept | Task | Concept | Task | Concept | Task | Concept |
| PLMs | LSTM | 40.57/60.67 | - | 43.34/64.47 | - | 68.25/53.37 | - | 90.5/90.46 | - |
| | GPT2 | 66.69/77.25 | - | 67.26/78.81 | - | 71.67/67.53 | - | 97.64/97.55 | - |
| | BERT | 68.75/78.71 | - | 71.81/82.58 | - | 80.5/78.4 | - | 98.89/98.68 | - |
| | RoBERTa | 71.36/80.17 | - | 73.12/82.64 | - | 84.1/82.5 | - | 99.13/99.12 | - |
| CBE-PLMs | LSTM | **56.47/67.82** | 86.46/85.24 | **54.54/65.84** | 83.46/84.74 | **68.5/55.4** | 72.5/77.5 | **93.02/91.53** | 76.92/75.41 |
| | GPT2 | 64.04/77.75 | 92.14/92.05 | 63.57/74.71 | 90.17/90.13 | 70.05/69.53 | 80.6/82.5 | 96.85/96.81 | 86.14/88.06 |
| | BERT | 67.27/79.24 | 93.65/92.75 | 68.23/78.13 | 89.64/90.45 | 77.42/74.57 | 80.2/83.7 | 97.62/97.58 | 92.57/92.05 |
| | RoBERTa | 70.98/79.89 | 96.12/95.34 | 69.85/79.29 | 91.45/92.23 | 82.33/80.13 | 86.7/85.3 | 98.45/98.12 | 93.99/94.28 |
| CBE-PLMs-CM | LSTM | - | - | **59.67/70.53** | 88.75/86.67 | - | - | **94.35/92.32** | 83.83/84.52 |
| | GPT2 | - | - | 65.54/77.87 | 93.58/92.32 | - | - | **97.89/97.88** | 89.64/88.25 |
| | BERT | - | - | 70.58/80.07 | 94.43/93.26 | - | - | 98.18/98.06 | 94.87/94.32 |
| | RoBERTa | - | - | 72.88/81.91 | 96.3/98.5 | - | - | **99.69/99.66** | 96.35/96.36 |

# Conclusion & Future Work

## Contributions:

- We provide the first investigation of standard training strategies of CBMs for interpreting PLMs and benchmarking CBE-PLMs.

- We propose C3M, which leverages LLMs and MixUp to help PLMs learn from human annotated and machine-generated concepts. C3M liberates CBMs from predefined concepts for the interpretability-utility tradeoff.

- We demonstrate the effectiveness and robustness of test-time concept intervention for the learned CBE-PLMs for common text classification tasks.

## Related Research:

- Can we achieve local and global interpretability at the same time?

See Zhen Tan's AAAI 24 paper: SparseCBM

- Can we further reduce the human involvement during inference time?

See Zhen Tan's AAAI 25 paper: CLEAR

# Extension 1

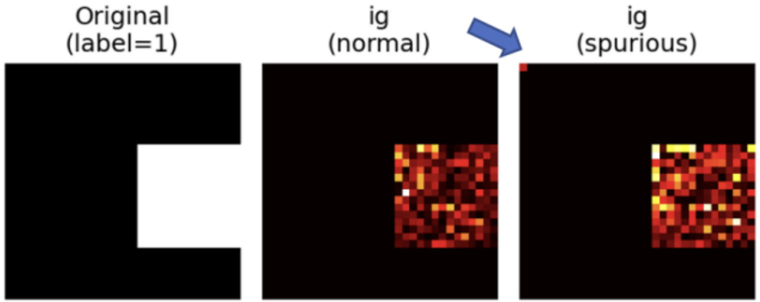
**Can we explain the explanations?**

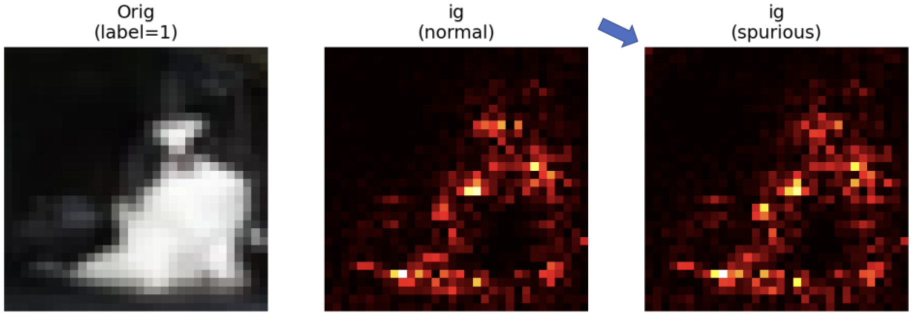
- Are the explanations reliable?

arXiv

**Are We Merely Justifying Results ex Post Facto?**
**Quantifying Explanatory Inversion in Post-Hoc Model Explanations**

**Zhen Tan** [1] **Song Wang** [2] **Yifan Li** [3] **Yu Kong** [3] **Jundong Li** [2] **Tianlong Chen** [4] **Huan Liu** [1]

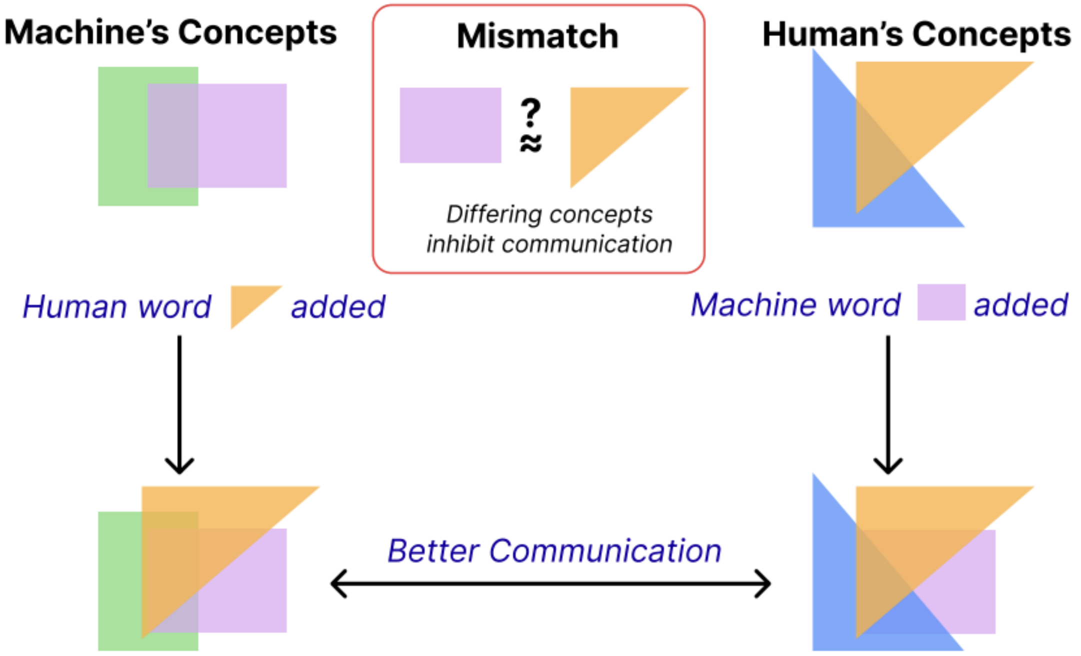


# Extension 2

**What if machine and human do not agree on the same concepts?**

- Aligning machine's concepts to human's




Ongoing research

# What are the next steps?

How to achieve better human-machine collaboration through explanations that are:

- User-aware
- Reliable
- Applicable

to enhance science discovery?

Zhen Tan's homepage

# Thank You

- For more details, please check out the paper.
- Feel free to contact the first author Zhen Tan (*ztan36@asu.edu*) for any questions.
- Implementation is released on GitHub.