On the Calibration of Deep Learning Models to Improve Trustworthy AI

Cornelia Caragea

cornelia@uic.edu

IRg Information Retrieval Group
UIC Computer Science

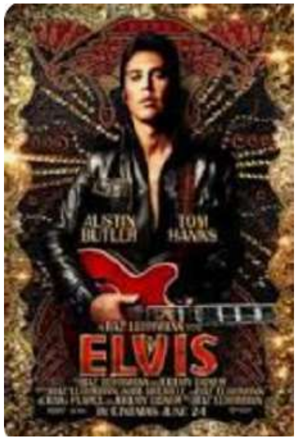
Thanks!

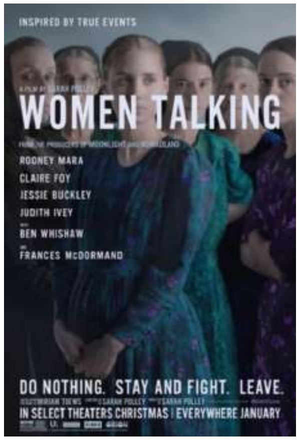
Marius Leordeanu

Adina Florea

Human Confidence and Calibration

What movie won the **Best Picture** at Oscars 2023?





Human Confidence and Calibration

Who is Prime Minister in UK?

[Rishi Sunak]()

Human Confidence and Calibration

Who is Advisor to the Minister at
Ministerul Cercetării, Inovării și Digitalizării - România?

Ioan Istrate

Machines…

Do they know what they don’t know?

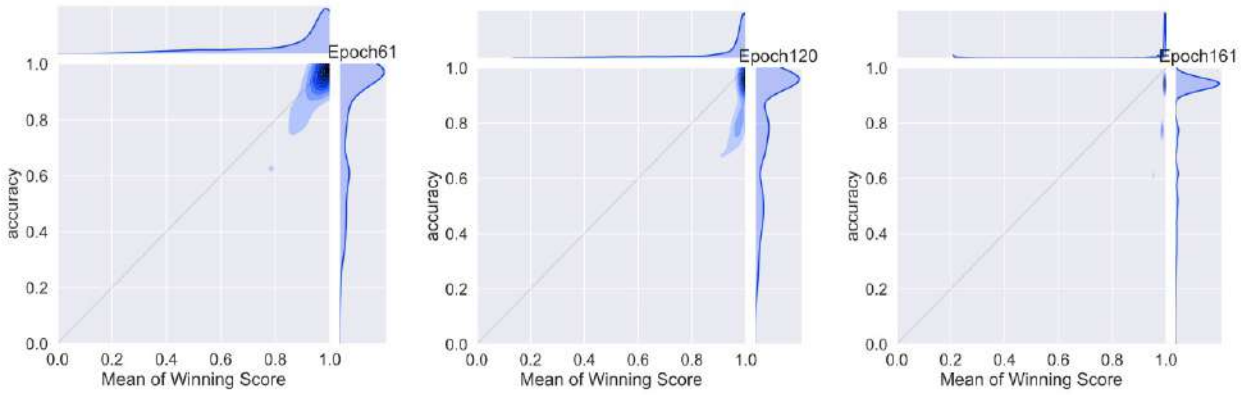
Or in other words… are they calibrated?

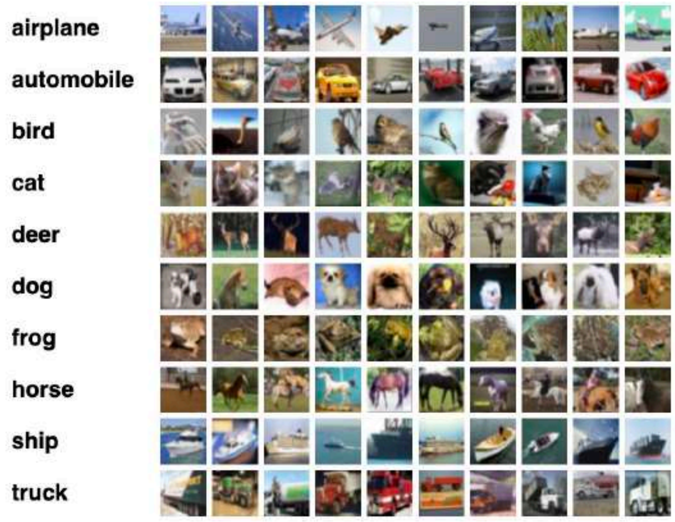
Deep Neural Networks

- Deep neural networks (DNNs) have established supremacy in many pattern recognition tasks such as object detection, speech recognition, natural language processing.
  - They are increasingly used in decision-making pipelines and high-risk fields such as medical diagnosis, autonomous vehicle control, and the legal sector.
- Major challenges: uncertainty and trust-worthiness of a classifier.
- The DNN must not only be accurate, but also indicate when it is likely to get the wrong answer.
  - This allows the decision-making to be routed as needed to a human or another more accurate, but possibly more expensive, classifier, with the assumption being that the additional cost incurred is greatly surpassed by the consequences of a wrong prediction.

DNNs Confidence and Calibration

- In a well-calibrated classifier, predictive scores should be indicative of the actual likelihood of correctness.
- Modern architectures, it turns out, are prone to overconfidence.

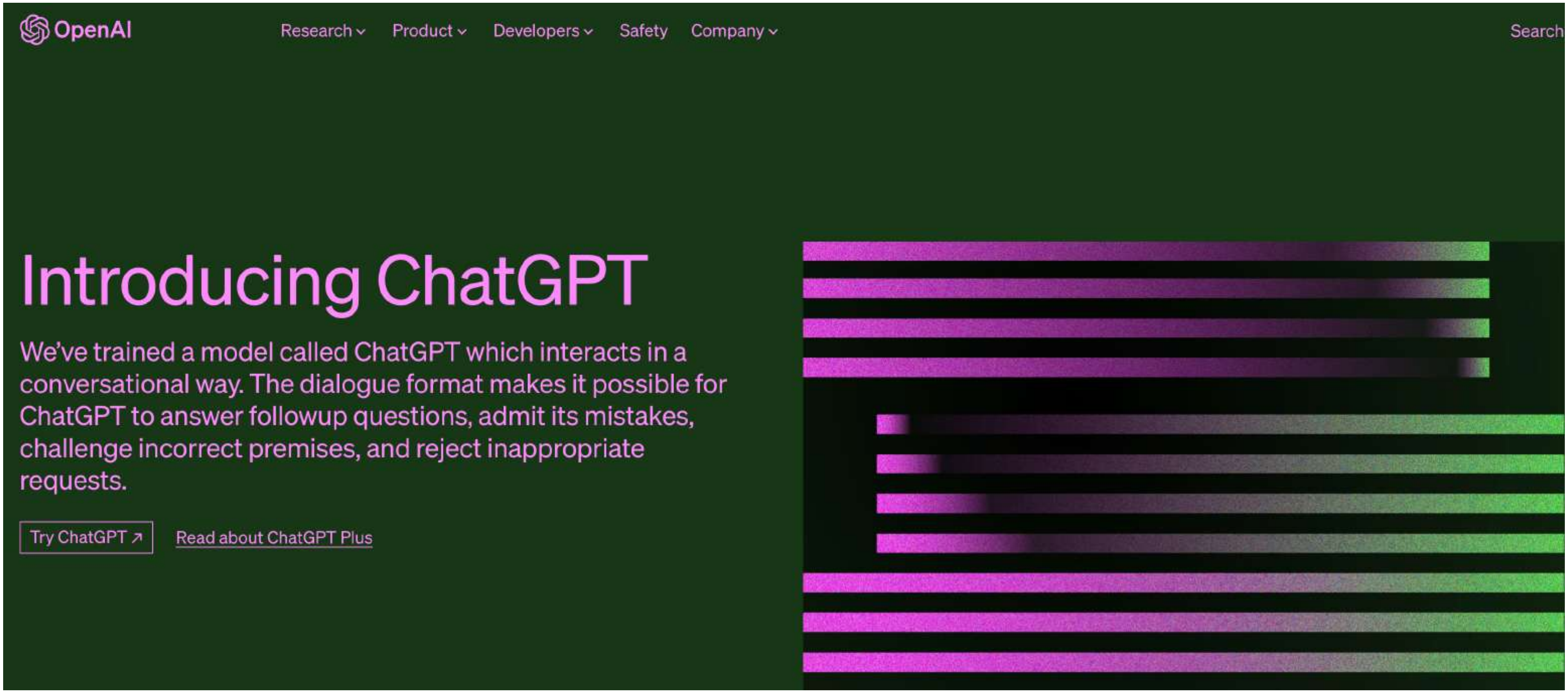




Accuracy vs confidence on CIFAR-100 at different training epochs for VGG-16 neural net.

Credit for the plots: Thulasidasan et al. [2019].

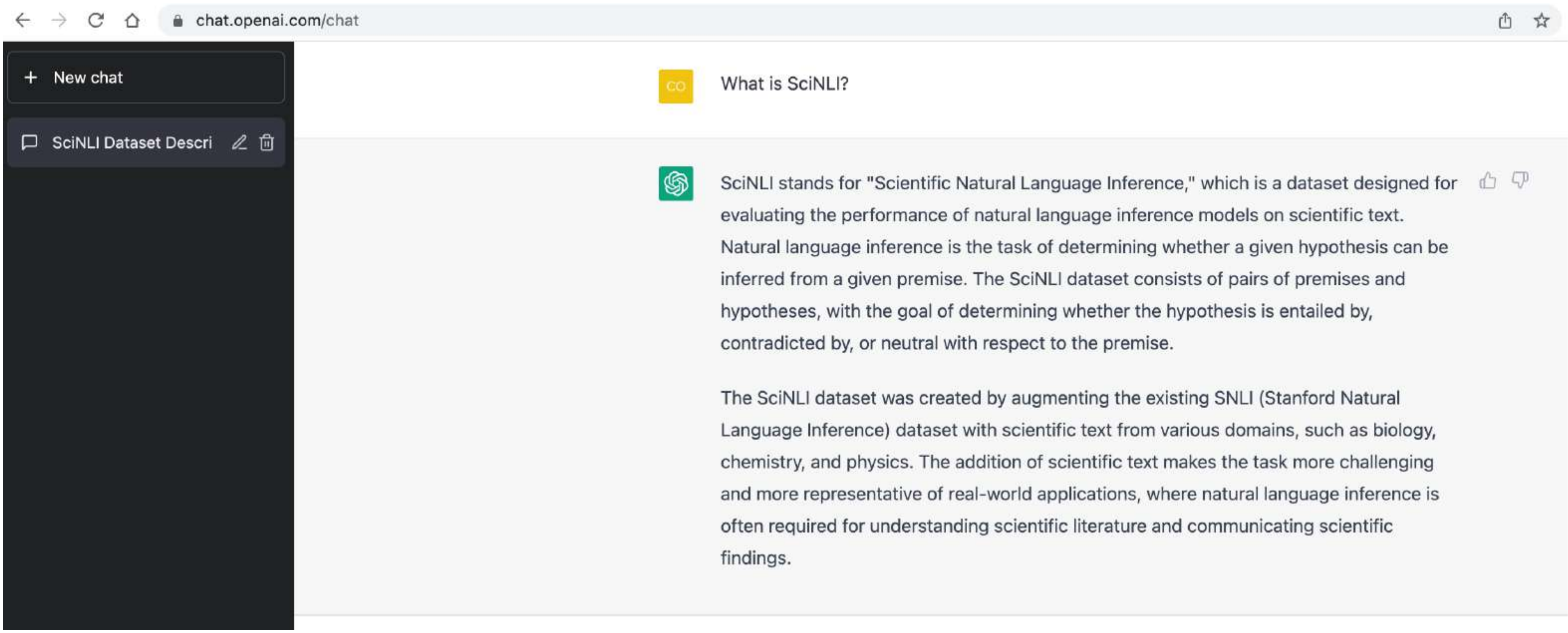
DNNs Confidence and Calibration



Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

DNNs Confidence and Calibration



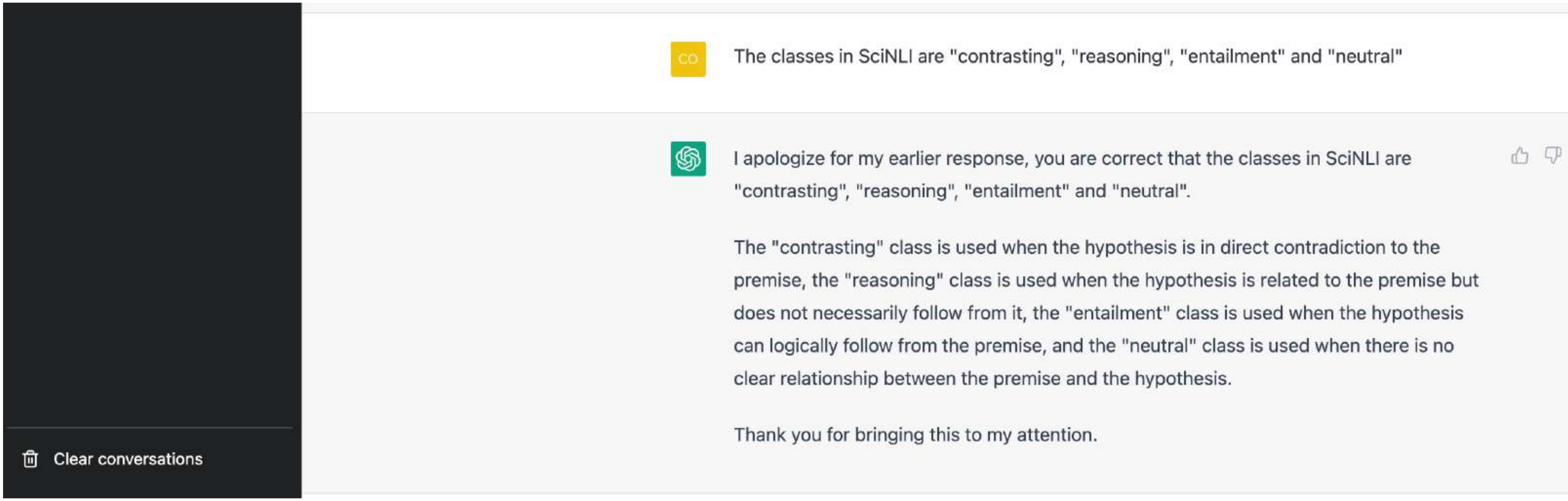
What is SciNLI?

SciNLI stands for "Scientific Natural Language Inference," which is a dataset designed for evaluating the performance of natural language inference models on scientific text. Natural language inference is the task of determining whether a given hypothesis can be inferred from a given premise. The SciNLI dataset consists of pairs of premises and hypotheses, with the goal of determining whether the hypothesis is entailed by, contradicted by, or neutral with respect to the premise.

The SciNLI dataset was created by augmenting the existing SNLI (Stanford Natural Language Inference) dataset with scientific text from various domains, such as biology, chemistry, and physics. The addition of scientific text makes the task more challenging and more representative of real-world applications, where natural language inference is often required for understanding scientific literature and communicating scientific findings.

[1] Sadat and Caragea, 2022: SciNLI: A Corpus for Natural Language Inference on Scientific Text.

DNNs Confidence and Calibration

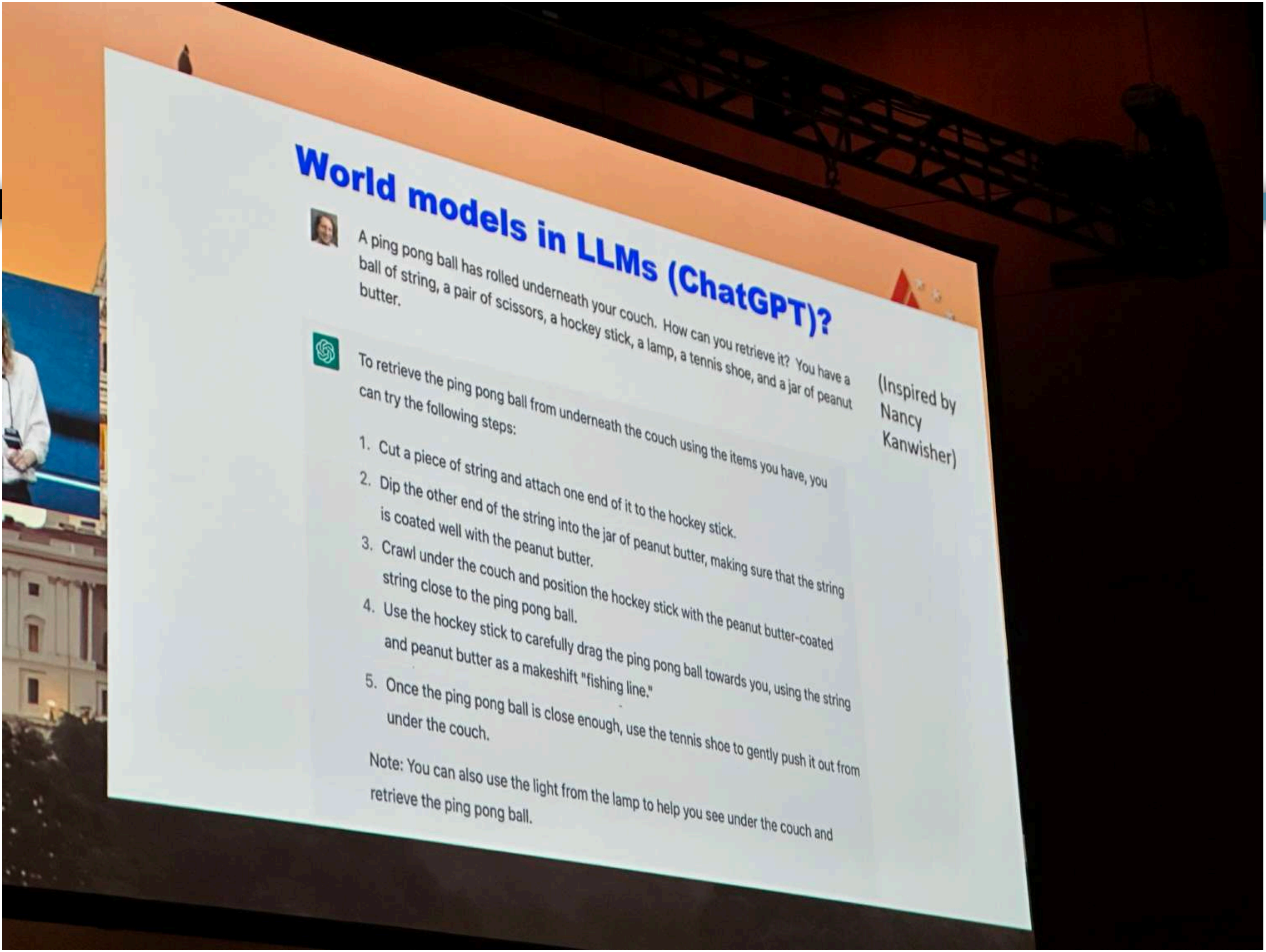


The classes in SciNLI are "contrasting", "reasoning", "entailment" and "neutral"

I apologize for my earlier response, you are correct that the classes in SciNLI are "contrasting", "reasoning", "entailment" and "neutral".

The "contrasting" class is used when the hypothesis is in direct contradiction to the premise, the "reasoning" class is used when the hypothesis is related to the premise but does not necessarily follow from it, the "entailment" class is used when the hypothesis can logically follow from the premise, and the "neutral" class is used when there is no clear relationship between the premise and the hypothesis.

Thank you for bringing this to my attention.



World models in LLMs (ChatGPT)?

A ping pong ball has rolled underneath your couch. How can you retrieve it? You have a ball of string, a pair of scissors, a hockey stick, a lamp, a tennis shoe, and a jar of peanut butter.

To retrieve the ping pong ball from underneath the couch using the items you have, you can try the following steps:

1. Cut a piece of string and attach one end of it to the hockey stick.
2. Dip the other end of the string into the jar of peanut butter, making sure that the string is coated well with the peanut butter.
3. Crawl under the couch and position the hockey stick with the peanut butter-coated string close to the ping pong ball.
4. Use the hockey stick to carefully drag the ping pong ball towards you, using the string and peanut butter as a makeshift "fishing line."
5. Once the ping pong ball is close enough, use the tennis shoe to gently push it out from under the couch.

Note: You can also use the light from the lamp to help you see under the couch and retrieve the ping pong ball.

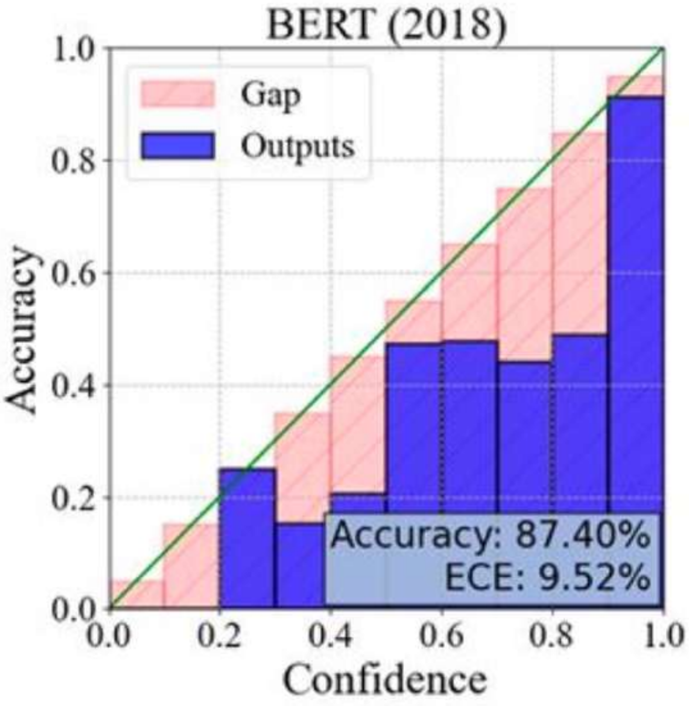
(Inspired by Nancy Kanwisher)

Credit: Joshua Tenenbaum (MIT) – Keynote at AAAI 2023

Calibration in Pre-trained Language Models

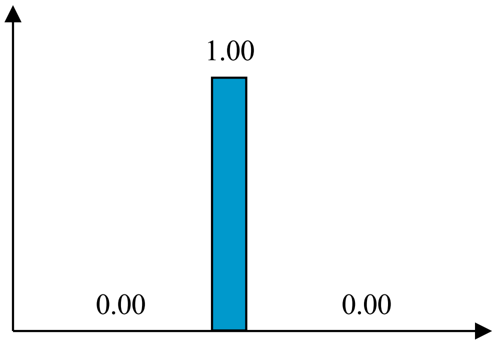
- Current pre-trained language models are often poorly calibrated [Kong et al., 2020] (most often being overly-confident).

- E.g., reliability diagram of BERT fine-tuned on text classification using 20NG15 dataset (the first 15 categories of the 20NG dataset).



Over-confidence

- Most modern DNNs, when trained for classification in a supervised learning setting, are trained using one-hot encoded labels that have all the probability mass in one class
  - The training labels are thus zero-entropy signals that admit no uncertainty about the input.
  - The DNN is thus, in some sense, trained to become overconfident.



Calibration Techniques

- Temperature Scaling [Guo et al., 2017; Desai and Durrett, 2020]
  - A post-processing step that re-scales the logits using a single scale hyperparameter temperature T that is learned on a validation set.
    - T → ∞ yields maximum uncertainty with uniform probabilities,
    - As T → 0, the probability drops to a point mass.

- Label Smoothing [Müller et al., 2019; Kumar and Sarawagi, 2019; Desai and Durrett, 2020]
  - A regularization technique that prevents over-confident predictions toward one single class by using soft labels.
    - For example, the one-hot label vector [1, 0, 0] is converted to [0.9, 0.05, 0.05] smoothed label vector.

MixUp

- MixUp [Zhang et al., 2018]
  - A data augmentation method in which additional samples are generated during training by combining random samples of training inputs and their associated labels.

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \qquad \text{where } x_i, x_j \text{ are raw input vectors}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \qquad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

On the Calibration of Pre-trained Language Models using MixUp Guided by Area Under the Margin and Saliency

Information Retrieval Group
UIC Computer Science

[Park and Caragea, ACL 2022]

Proposed MixUp for Model Calibration

- We propose a MixUp method that is targeted at improving model calibration.

- We leverage a model's training dynamics, Area Under the Margin, [Pleiss et al., 2020] to reveal samples with distinct pronounced characteristics
  - whether they are easy-to-learn or hard-to-learn/ambiguous for the model.

- We generate MixUp samples by mixing easy-to-learn with hard-to-learn/ambiguous samples according to their similarity/dissimilarity provided by saliency maps [Simonyan et al., 2013].

Mixup using Saliency Signals

- Mixing easy-to-learn samples with the most similar hard-to-learn samples calibrates in-domain data.
- Mixing easy-to-learn samples with the most dissimilar hard-to-learn samples calibrate out-of-domain data.

Datasets

- Tasks used for evaluation :
  - Natural Language Inference
    - In-domain : SNLI [Bowman et al., 2015]
    - Out-of-domain: MNLI [Williams et al., 2018]
  - Paraphrase Detection
    - In-domain: QQP [Iyer et al., 2017]
    - Out-of-domain: TwitterPPDB [Lan et al., 2017]
  - Commonsense Reasoning
    - In-domain: SWAG [Zellers et al., 2018]
    - Out-of-domain: HellaSWAG [Zeller et al., 2019]

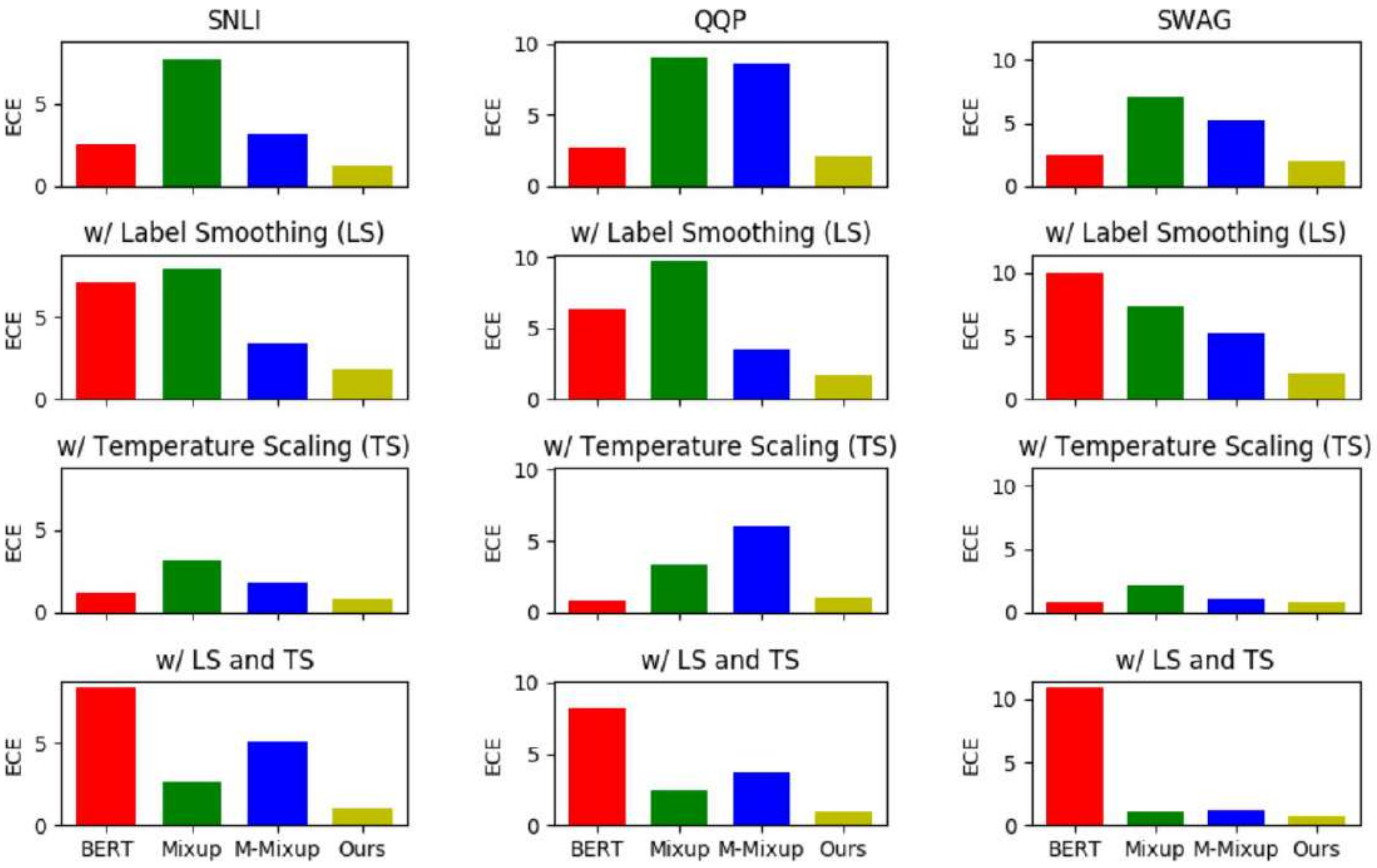
- We use in-domain trained models to predict out-of-distribution test samples.

Baselines for Evaluation

- Pre-trained Language Models (BERT, RoBERTa)
- MixUp [Zhang et al., 2018; Thulasidasan et al., 2019]
- Manifold-MixUp (M-MixUp) [Verma et al., 2019]

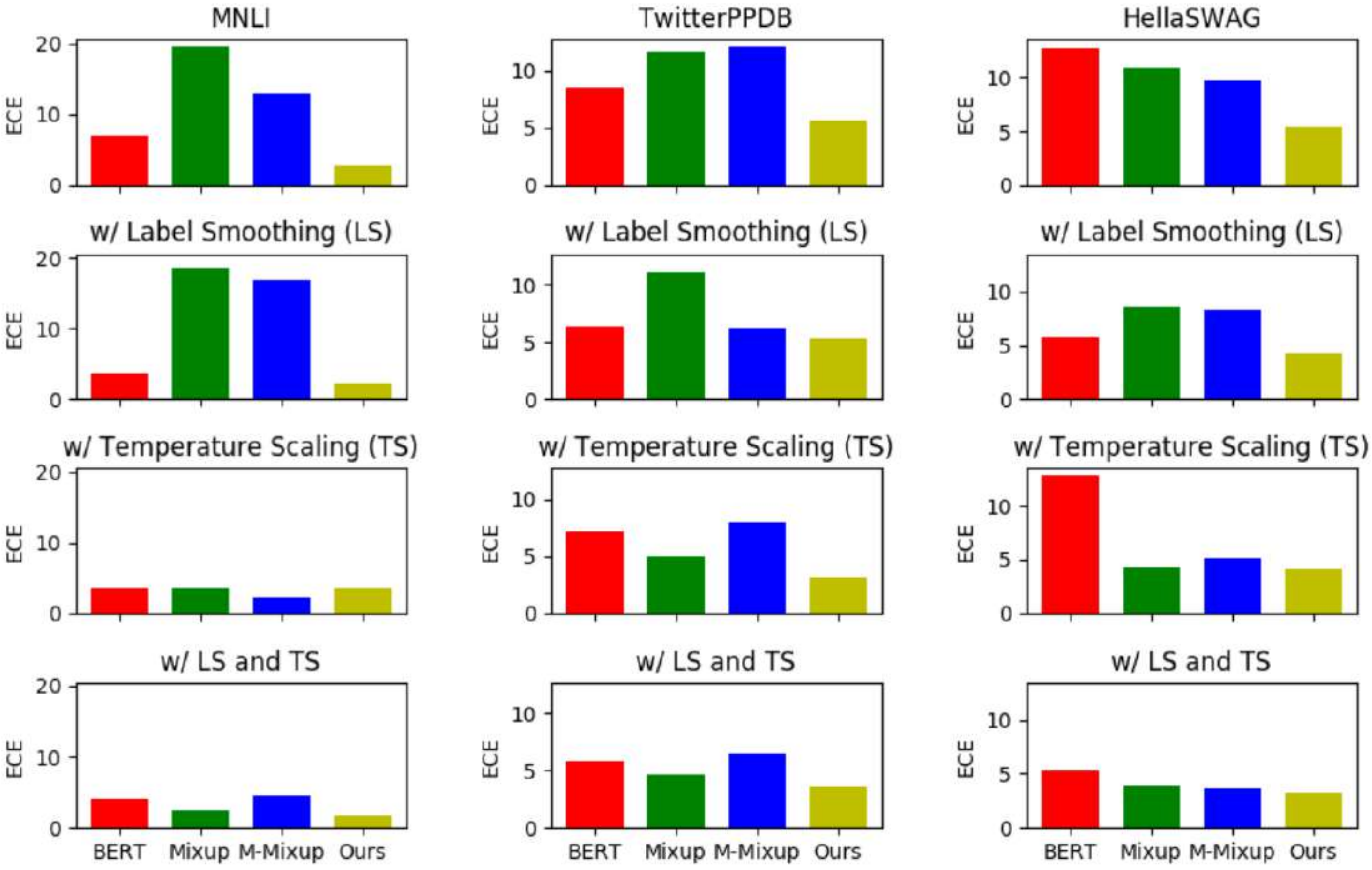
- We explore the combination of miscalibration correction methods (i.e., temperature scaling, label smoothing) for all models.

In-domain Data Results on BERT



Our proposed MixUp results in best ECE values for all ID tasks (similar results are observed on RoBERTa).

In-domain Data Results on BERT



Our proposed MixUp results in best ECE values for all OOD tasks (similar results are observed on RoBERTa).

Information Retrieval Group
UIC Computer Science

Conclusion

- We proposed a novel MixUp guided by the Area Under the Margins (AUM) and Saliency Maps to mitigate the miscalibration of pre-trained language models BERT and RoBERTa.

- We showed that our proposed MixUp achieves the lowest Expected Calibration Errors (ECE) for both pre-trained language models on various types of NLU tasks, for both in-domain and out-of-domain data.

Some of Our Recent Papers

- Seo Yeon Park and Cornelia Caragea. (2022). “On the Calibration of Pre-trained Language Models using Mixup Guided by Area Under the Margin and Saliency.” In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (**ACL 2022**). Dublin, Ireland.
- Seo Yeon Park and Cornelia Caragea. (2022). “A Data Cartography based MixUp for Pre-trained Language Models.” In: Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (**NAACL 2022**). Seattle, Washington.
- Mobashir Sadat and Cornelia Caragea. (2022). “SciNLI: A Corpus for Natural Language Inference on Scientific Text.” In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (**ACL 2022**). Dublin, Ireland.
- Mobashir Sadat and Cornelia Caragea. (2022). “Learning to Infer from Unlabeled Data: A Semi-supervised Learning Approach for Robust Natural Language Inference.” In: Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing: Findings (**EMNLP Findings 2022**), Abu Dhabi.
- Tiberiu Sosea and Cornelia Caragea. (2022). “Leveraging Training Dynamics and Self-Training for Text Classification.” In: Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing: Findings (**EMNLP Findings 2022**), Abu Dhabi.
- Mahshid Hosseini and Cornelia Caragea. (2023). “Feature Normalization and Cartography-based Demonstrations for Prompt-based Fine-tuning on Emotion-related Tasks.” In: Proceedings of The Association for the Advancement of Artificial Intelligence (**AAAI 2023**), Washington, DC.
- Mahshid Hosseini and Cornelia Caragea. (2022). “Calibrating Student Models for Emotion-related Tasks.” In: Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2022**), Abu Dhabi.

Thank you!

FULBRIGHT Romania

World Learning
EDUCATION | DEVELOPMENT | EXCHANGE

NSF

aws

OFFICE OF THE VICE CHANCELLOR FOR RESEARCH UIC

DPI

DARPA

Office of Naval Research ONR Science & Technology

Seo Yeon Park

Mobashir Sadat

Mahshid Hosseini

Tiberiu Sosea

Code: https://github.com/seoyeon-p/MixUp-Guided-by-AUM-and-Saliency-Map