EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

# Learning Transferable Visual Models From Natural Language Supervision

## Introductory presentation

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, *OpenAI,* 2021.

Presenter: **Diego Calanzone**

Seminar: **Learning with Limited Labeled Data**

Academic year: **2021/2022**

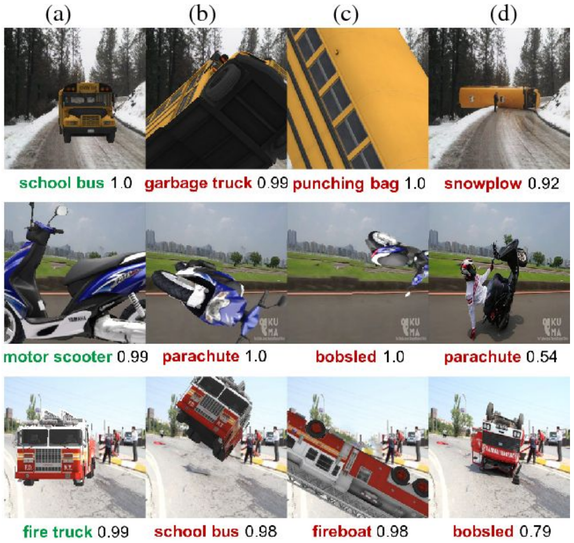
# Difficulties in Computer Vision

- Typical vision datasets are costly and labor intensive to create
- Standard vision models are good at one task only
- Such models can perform poorly in stress tests

*[Dodge, S., & Karam, L. (2017, July). "A study and comparison of human and deep learning recognition performance under visual distortions." In ICCCN 2017]*

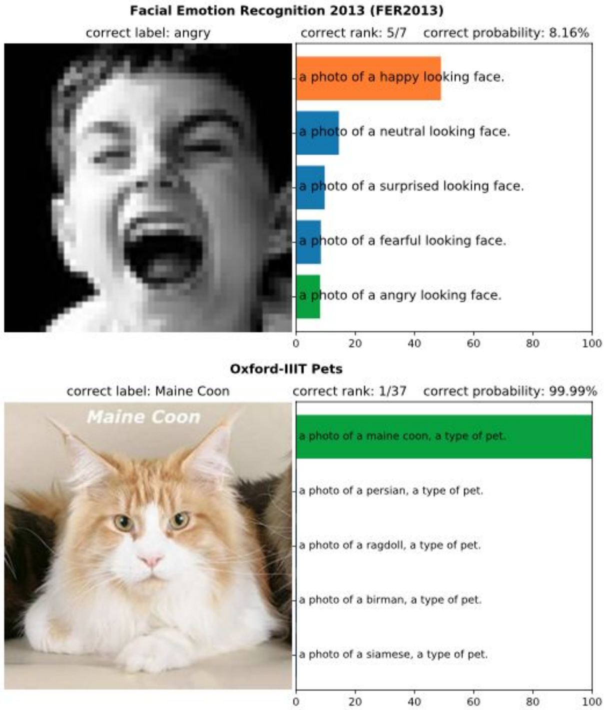
*[Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In ICLR 2019]*

*[Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W. S., & Nguyen, A. (2019). "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects." In CVPR 2019]*



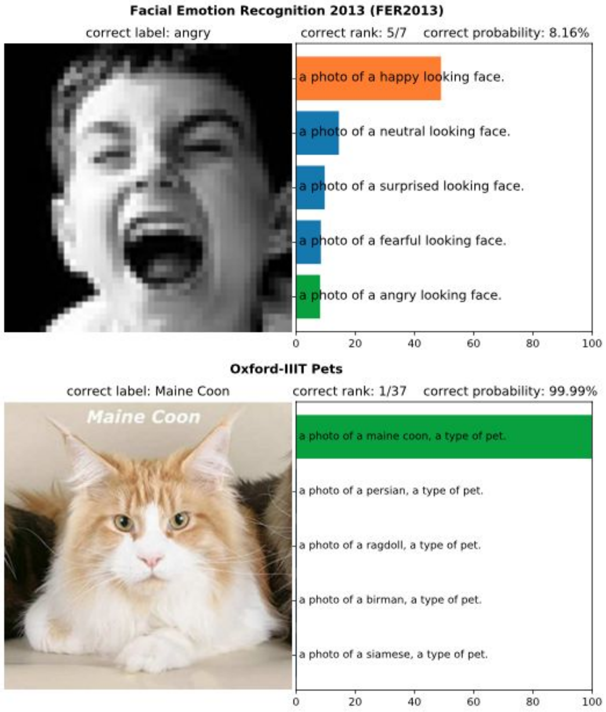
# Why Natural Language Supervision?

- Appreciating natural language as a training signal



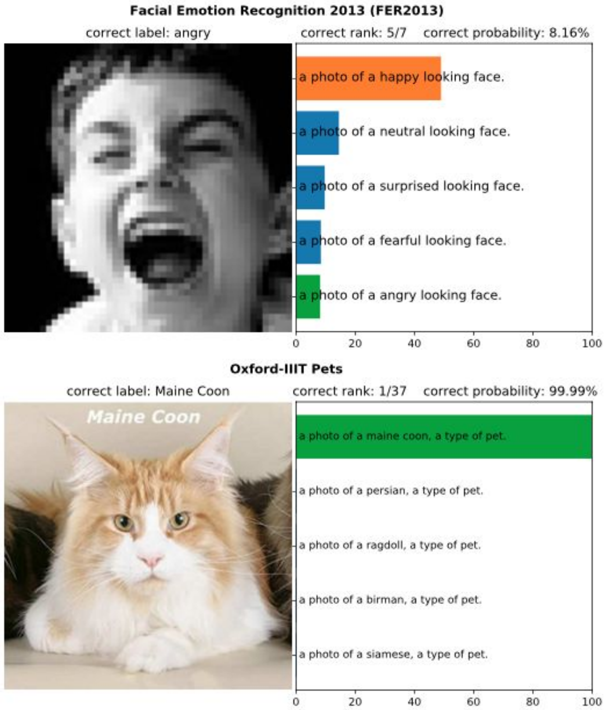
# Why Natural Language Supervision?

- Appreciating natural language as a training signal
- No burdensome label crafting



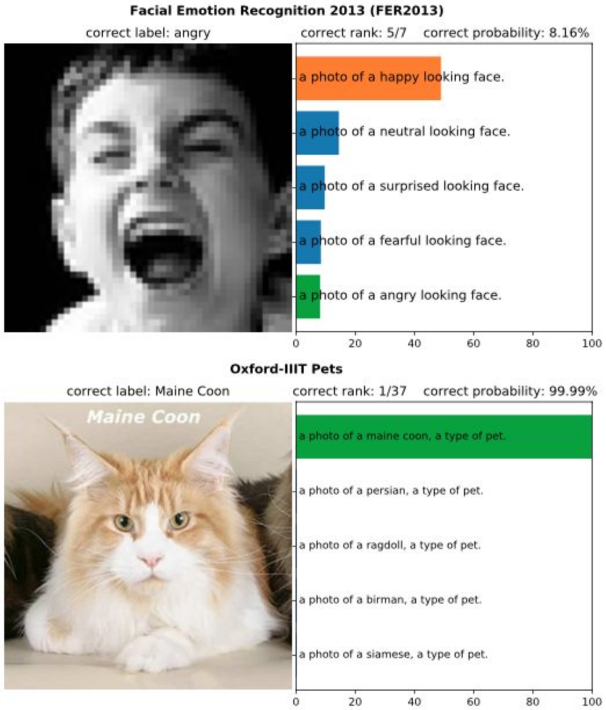
# Why Natural Language Supervision?

- Appreciating natural language as a training signal
- No burdensome label crafting
- More scalable data



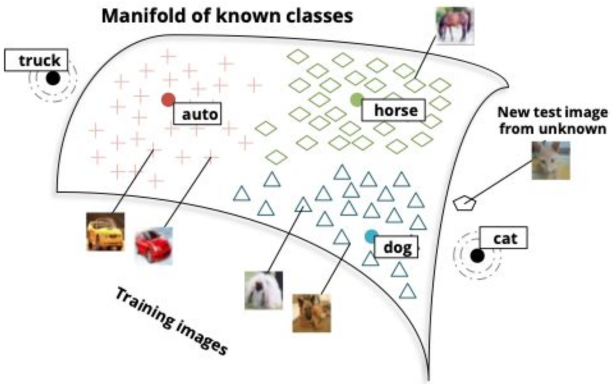
# Why Natural Language Supervision?

- Appreciating natural language as a training signal
- No burdensome label crafting
- More scalable data
- Flexible zero-shot transfer



# Background: language-image models

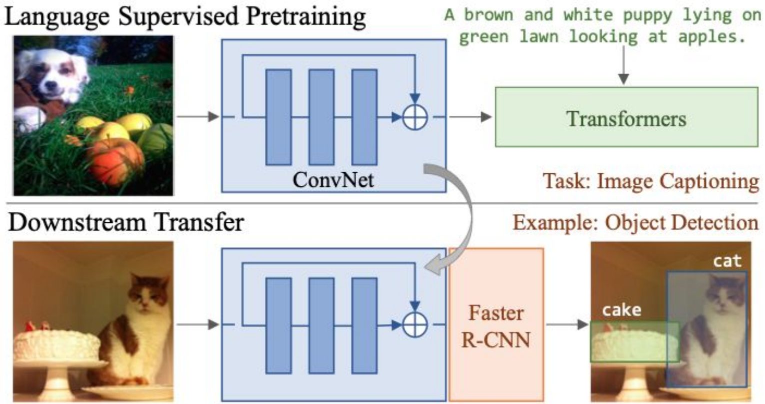
Zero-Shot Learning Through Cross-Modal Transfer
*Sochet et al., 2013*



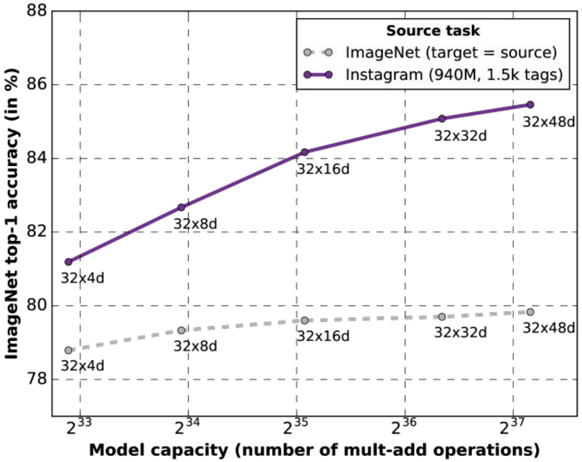

Learning Visual N-Grams from Web Data
*Li et al., FAIR, 2017*

# Background: language-image models

VirTex: Learning Visual Representations from Textual Annotations
*Desai et al., UofMichigan, 2020*




Exploring the Limits of Weakly Supervised Pretraining
*Mahajan et al., Facebook, 2020*

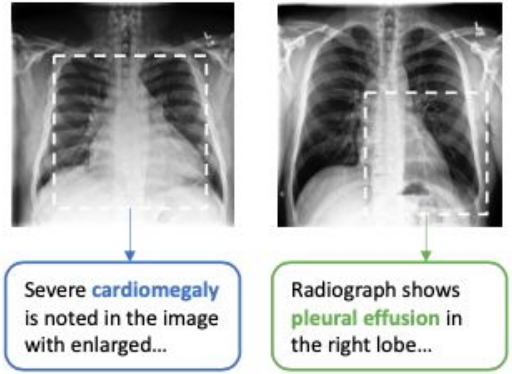


EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Limitations & Inspiration

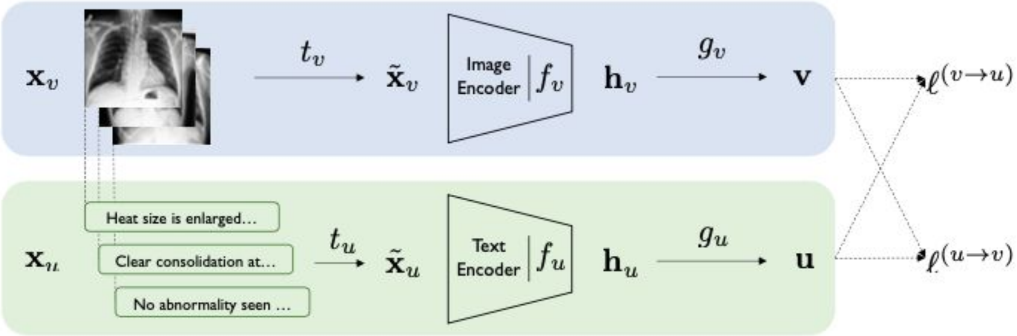
## Key idea: bidirectional contrastive learning

Contrastive Learning of Medical Visual Representations from Paired Images and Text
*Zhang et al., Stanford University, 2020*

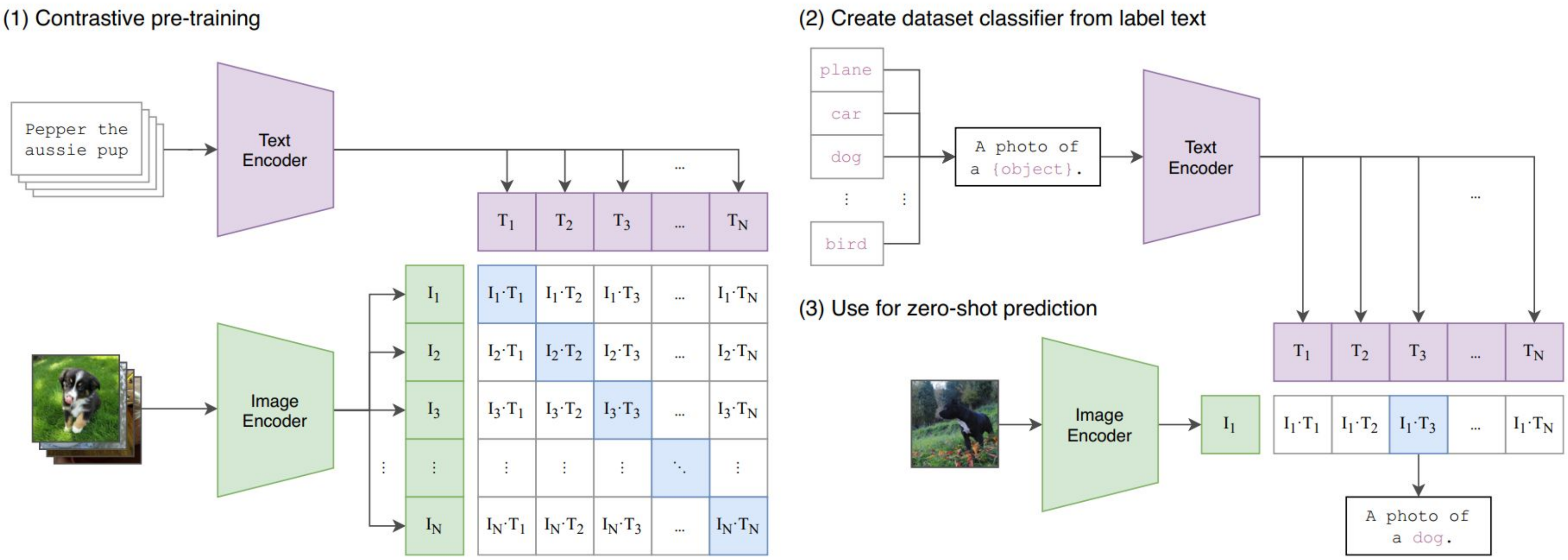



## Overall limitations

- Low zero-shot performances
- Poor scalability



# Contrastive Language-Image Pretraining (CLIP)

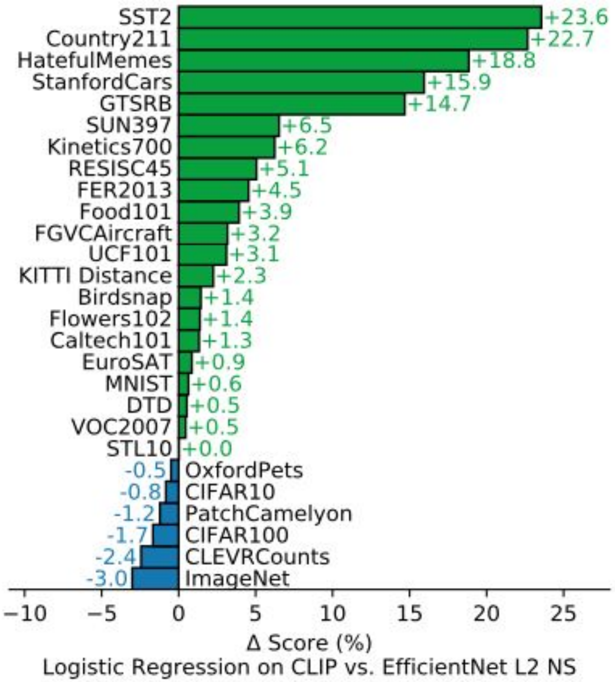


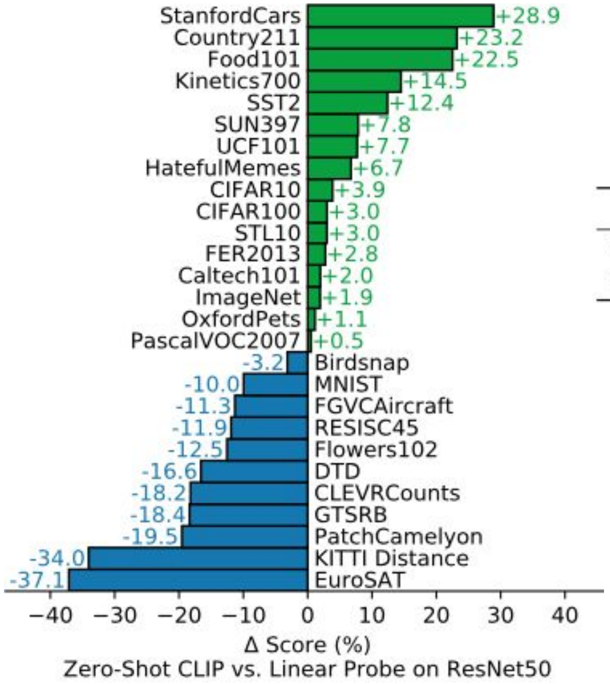
# Main contributions and outcomes

- A paradigm based on intensive, **task-agnostic pre-training**:
  a. Symmetric loss
  b. Contrastive objective
- A dedicated **dataset** of over 400M pairs
- A highly **scalable** architecture

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

# Task-specific comparison & performance

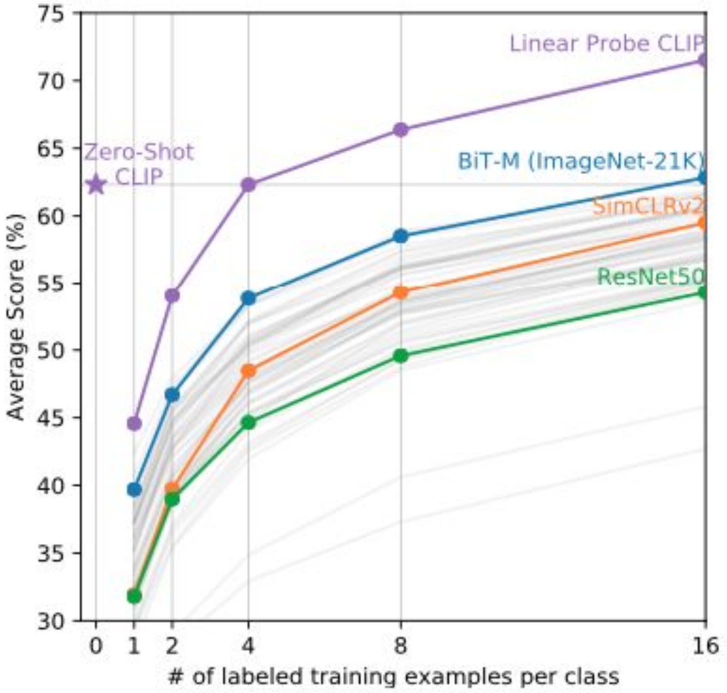


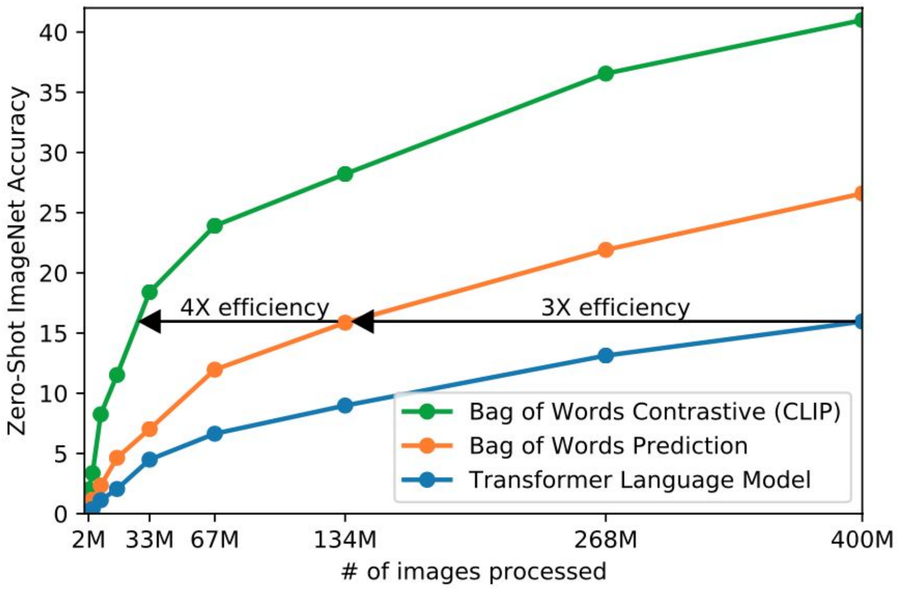




| | aYahoo | ImageNet | SUN |
|---|---|---|---|
| Visual N-Grams | 72.4 | 11.5 | 23.0 |
| CLIP | **98.4** | **76.2** | **58.5** |

Comparison with the Visual N-Grams model from Li et al. (FAIR) 2017

# Zero-shot performance & Efficiency



# Summary

1. Natural Language Supervision allows **scalability**, flexibility with **language**, **generalization**

2. Contrastive learning is efficient, learning the **text-image affinity** is much better for zero-shot transfer

3. A novel dataset based on image captions provides better semantics and allows **scaling**

**Yet to discuss:**

- The dataset
- Training, scalability, variations of the model
- Limitations: bias, broader impacts, typography attacks
- Further benchmarks: comparison with humans, robustness tests

# Thanks for listening!

## Questions are welcome.