A Task-Optimized Vision–Language Neural Network Produces Human-Like Linguistic Cueing Despite Lacking Built-in Attention Mechanisms

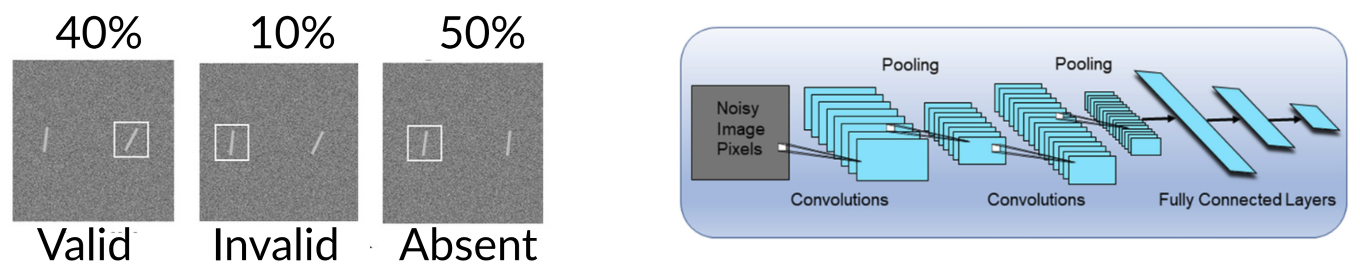
Jonathan Skaza [1] Sana Shehabi [2] Miguel P. Eckstein [1,2,3]

University of California, Santa Barbara

[1]Graduate Program in Dynamical Neuroscience [2]Department of Psychological and Brain Sciences [3]Department of Computer Science

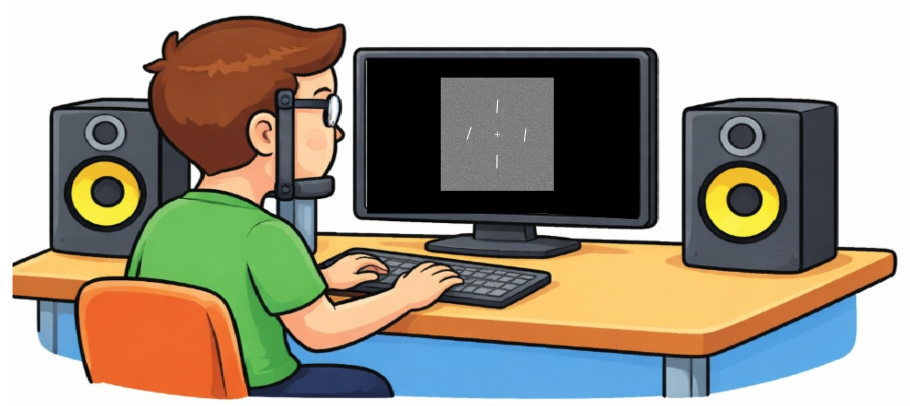
Motivation & Previous Work

- Srivastava et al. (2024) showed that feedforward convolutional neural networks (CNNs) exploit **visual cues** when trained for a binary target detection task. E.g., $Pr(\text{Tilted Line}) = 0.5$, $Pr(\text{Cued Location Valid}|\text{Tilted Line}) = 0.8$.



- Srivastava et al. (2025) investigated the neuronal mechanisms inside the CNN that explain the emergence of the cueing effect. Early CNN layers contain retinotopic neurons tuned separately to cue vs. target; later layers have jointly tuned (cue & target) and location integration neurons.
- However, human covert attention is **not limited to visual information**; **language can modulate spatial attention** (Barnas et al., 2024).
- We assess whether **linguistic cueing behavior** can emerge in **artificial neural networks that combine visual and language input**—a task-optimized CNN combined with a multilayer perceptron (MLP) operating on language embeddings—and compare these to linguistic cueing effects in humans.

Task: Determine Presence of Target Among Distractors



Trial Timeline

Audio Cue: e.g, Focus on the left

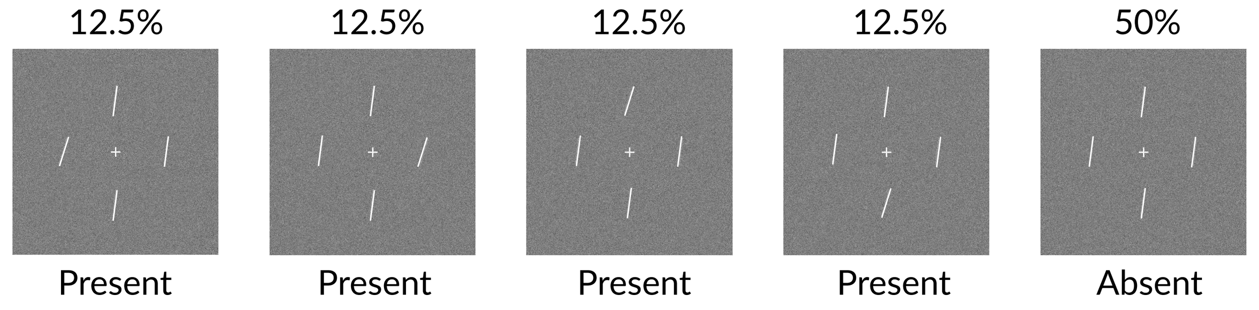
500ms delay

Visual Stimulus (250ms)

Typed Response (Y/N)

$Pr(\text{Target} = 0.5)$ $Pr(\text{Location}_i|\text{Target} = 0.25)\ \forall i \in$ Top, Right, Bottom, Left

Targets $\sim \mathcal{N}(17°, 3.5°)$ Distractors $\sim \mathcal{N}(7°, 3.5°)$ $Pr(\text{Cued Location Valid}|\text{Target}) = 0.8$



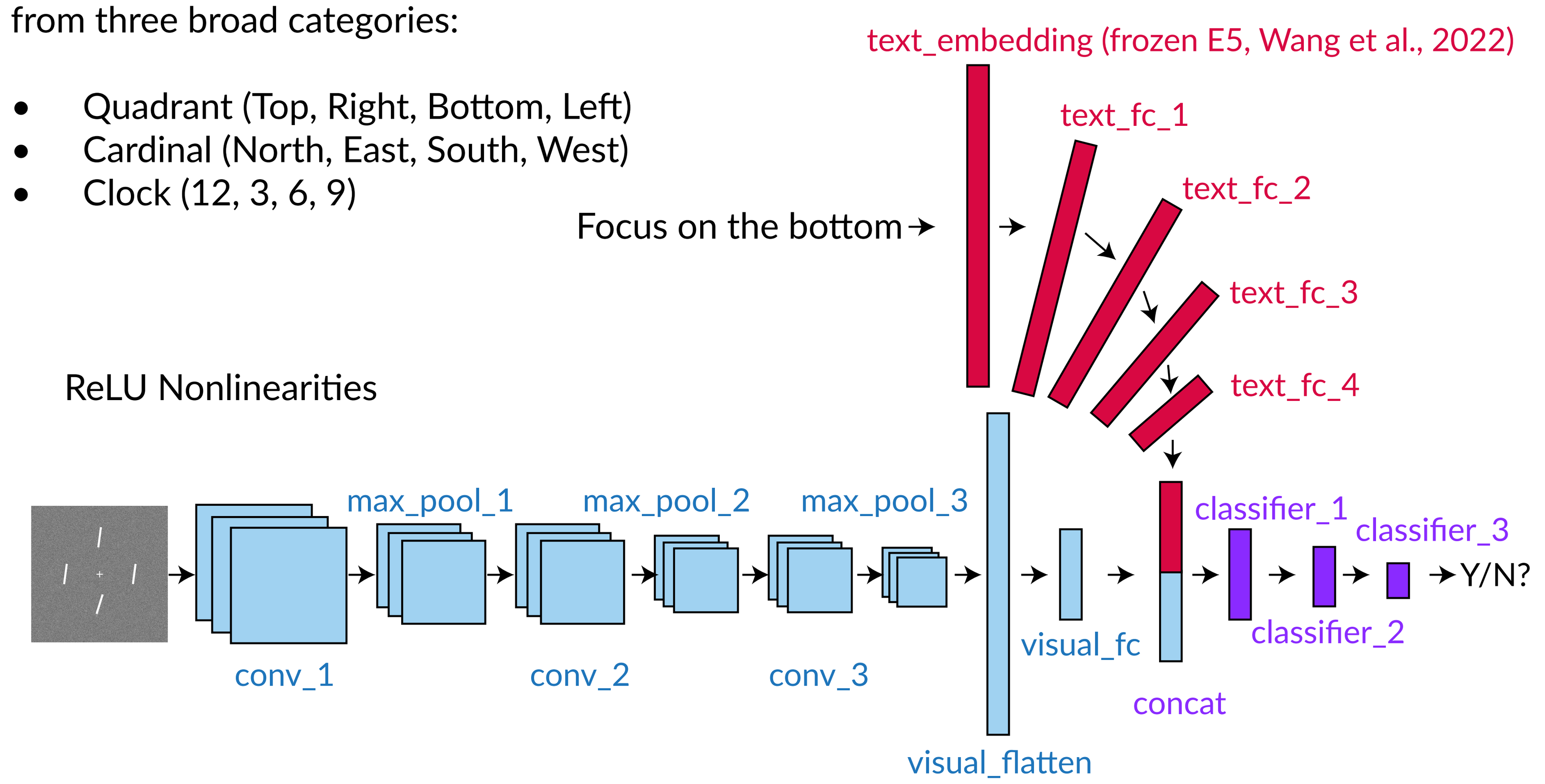
Acknowledgments

Research was sponsored by the U.S. Army Research Office and accomplished under cooperative agreement W911NF-19-2-0026 for the Institute for Collaborative Biotechnologies.

Neural Network Model Architecture

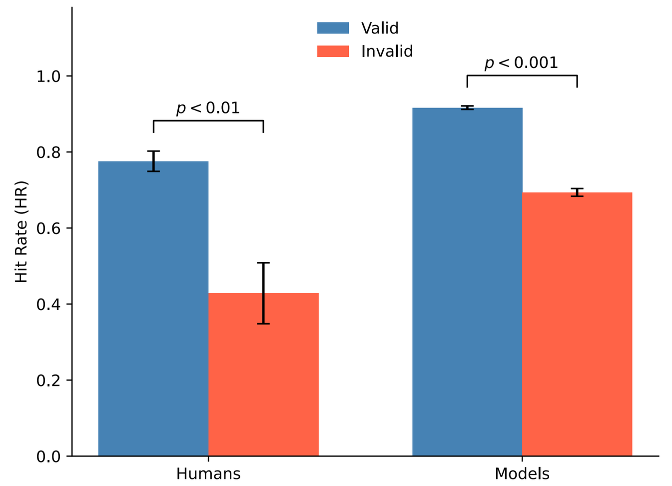
Hundreds of natural language spatial cues from three broad categories:

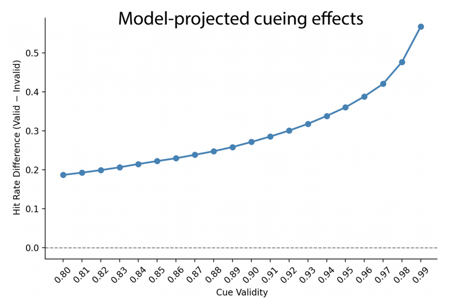
- Quadrant (Top, Right, Bottom, Left)
- Cardinal (North, East, South, West)
- Clock (12, 3, 6, 9)

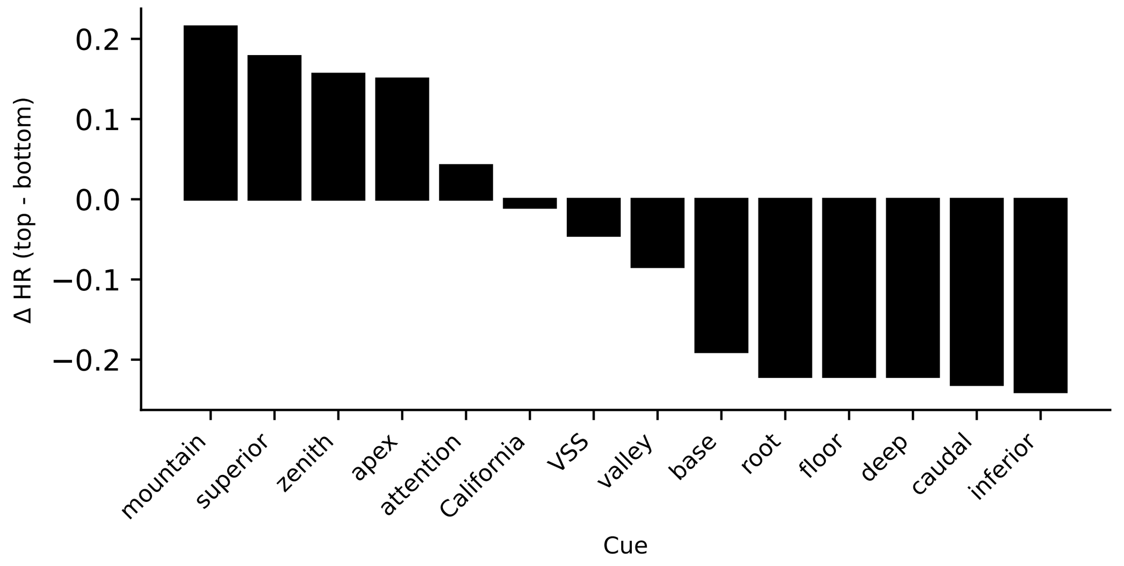


Behavioral Performance in Human Observers & Models

- $N = 11$ humans performed 300 trials each, mean $HR_{valid} - HR_{invalid} = 0.347$
- $N = 10$ model instances, mean $HR_{valid} - HR_{invalid} = 0.223$
- Both very large, significant cueing effects (effect magnitude not significantly different between humans and models), though humans seem to take a strategy more consistent with >90% cue validity while models weight evidence more optimally

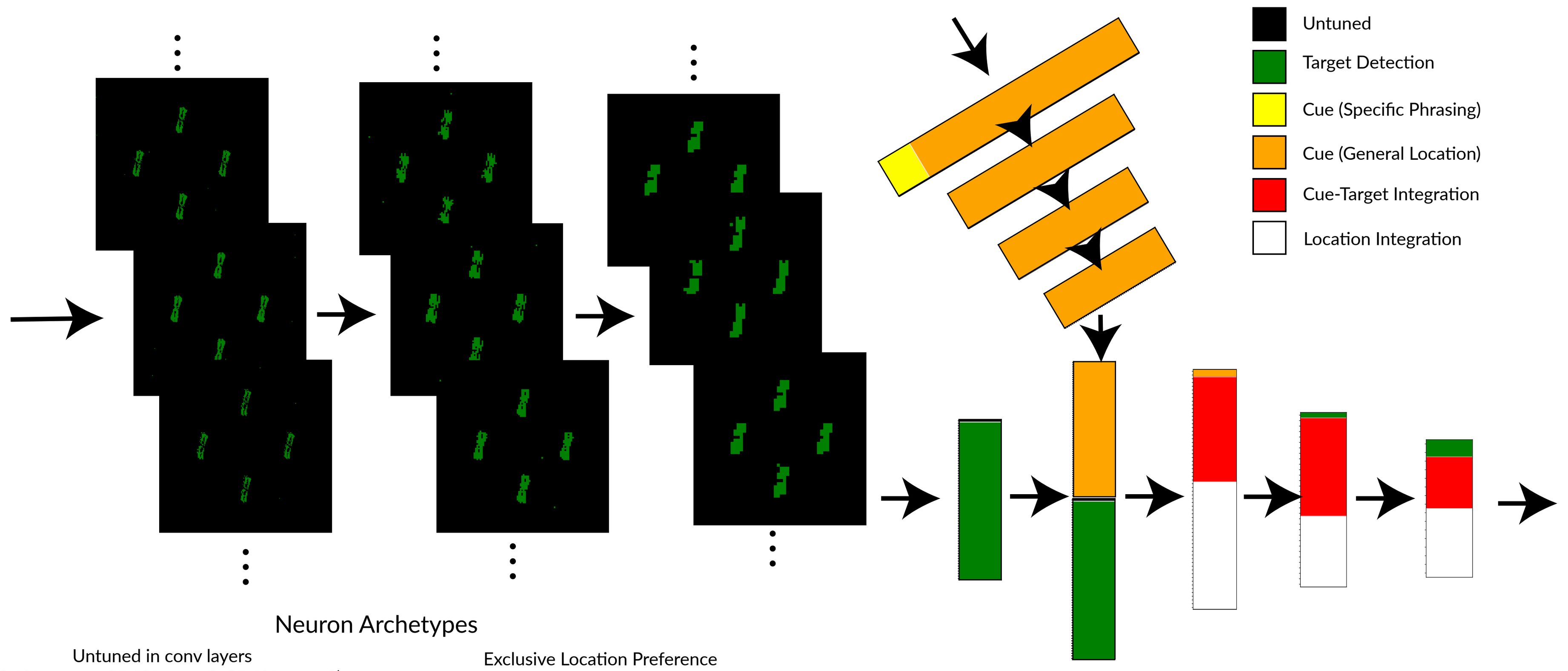


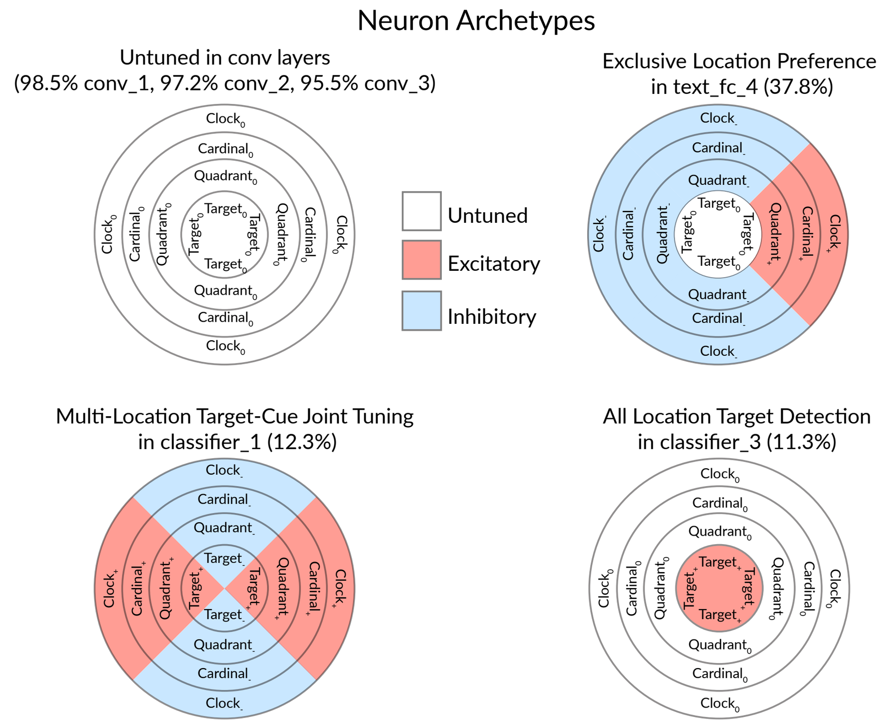


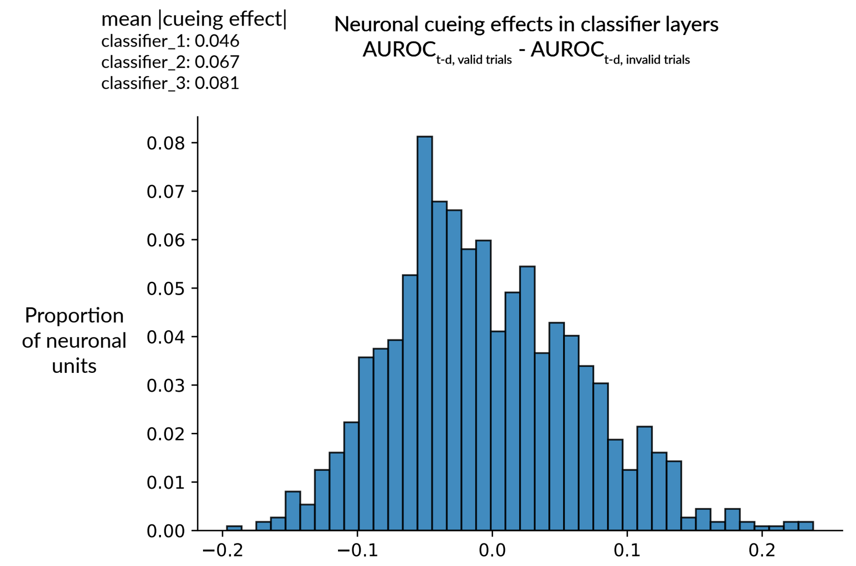


- The models demonstrate a robust ability to leverage entirely novel cues, indicating that their behavior is not limited to memorized associations from training.
- We evaluate a broad battery of cue types and compare the magnitude of the cueing effect across conditions. In particular, we hold the cue itself fixed while varying the target location (top vs. bottom) to isolate how spatial context modulates performance.

Large-Scale Neuron-Level Analysis *in silico*







Conclusions & Future Work

- Not only do purely visual CNNs exhibit cueing effects, but multimodal models incorporating linguistic cues do as well.
- Both human observers and models exhibit highly significant cueing effects in the four-location setting.
- AUROC analysis of the network reveals increased sensitivity to visual targets as processing progresses, integration of variable text cues into location-specific semantic representations, jointly tuned neurons following the concatenation layer, and common tuning archetypes within certain network layers.
- Future work will extend the range of cues presented to human observers to investigate cue-dependent variations in cueing effects (e.g., "Focus opposite of left").

References

Adam J Barnas, Natalie C Ebner, and Steven M Weisberg. Allocation of space-based attention is guided by efficient comprehension of spatial direction. *Journal of Cognition*, 7(1):1, 2024.

Sudhanshu Srivastava, William Yang Wang, and Miguel P Eckstein. Emergent human-like covert attention in feedforward convolutional neural networks. *Current Biology*, 34(3):579–593, 2024.

Sudhanshu Srivastava, William Yang Wang, and Miguel P Eckstein. Emergent neuronal mechanisms mediating covert attention in convolutional neural networks. *Proceedings of the National Academy of Sciences*, 122 (46):e2411909122, 2025.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

skaza@ucsb.edu