Rethinking Crowd-Sourced Evaluation of Neuron Explanations

Tuomas Oikarinen, Ge Yan, Akshay Kulkarni, Tsui-Wei (Lily) Weng - **UCSD**

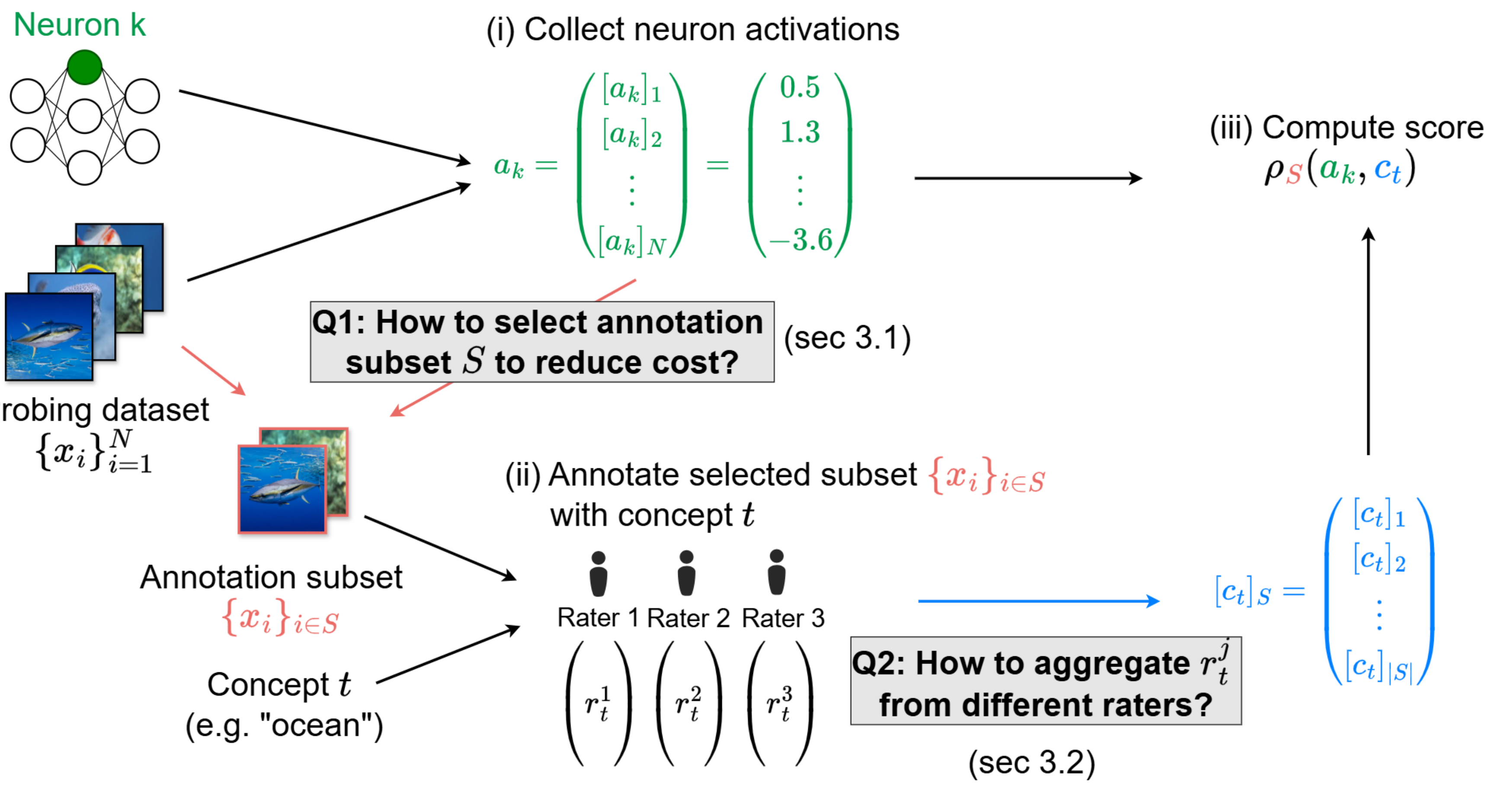
GitHub: https://github.com/Trustworthy-ML-Lab/efficient_neuron_eval
Paper: https://arxiv.org/abs/2506.07985

Trustworthy ML Lab @ UCSD
lilywenglab.github.io

Motivation: Existing Crowdsourced studies of neuron explanations only evaluate on highly activating inputs

- Only evaluating highly activating inputs is equivalent to only measuring **Recall**, and ignores whether the concept is present on low activating inputs
- We conduct the first study with a principled metric, Pearson's correlation coefficient
- Evaluating correlation can be very expensive due to need to annotate all inputs and rater noise
  - We propose efficient sampling and error correction strategies to reduce total cost **~60x**

Our Contributions:
1st crowdsourced study measuring correlation coefficient + ~60x cost reduction by efficient sampling and error correction


Neuron k
(i) Collect neuron activations
$a_k = \begin{pmatrix}[a_k]_1\\ [a_k]_2\\ \vdots\\ [a_k]_N\end{pmatrix} = \begin{pmatrix}0.5\\ 1.3\\ \vdots\\ -3.6\end{pmatrix}$
(iii) Compute score
$\rho_S(a_k, c_t)$
Probing dataset
$\{x_i\}_{i=1}^N$
Q1: How to select annotation subset S to reduce cost?
(sec 3.1)
Annotation subset
$\{x_i\}_{i\in S}$
(ii) Annotate selected subset $\{x_i\}_{i\in S}$ with concept t
Concept t
(e.g. "ocean")
Rater 1 Rater 2 Rater 3
$\begin{pmatrix}r_t^1\end{pmatrix}$ $\begin{pmatrix}r_t^2\end{pmatrix}$ $\begin{pmatrix}r_t^3\end{pmatrix}$
Q2: How to aggregate r_t^j from different raters?
(sec 3.2)
$[c_t]_S = \begin{pmatrix}[c_t]_1\\ [c_t]_2\\ \vdots\\ [c_t]_{|S|}\end{pmatrix}$

Methods

Contribution 1: Importance Sampling

Rating every input for each concept is not feasible

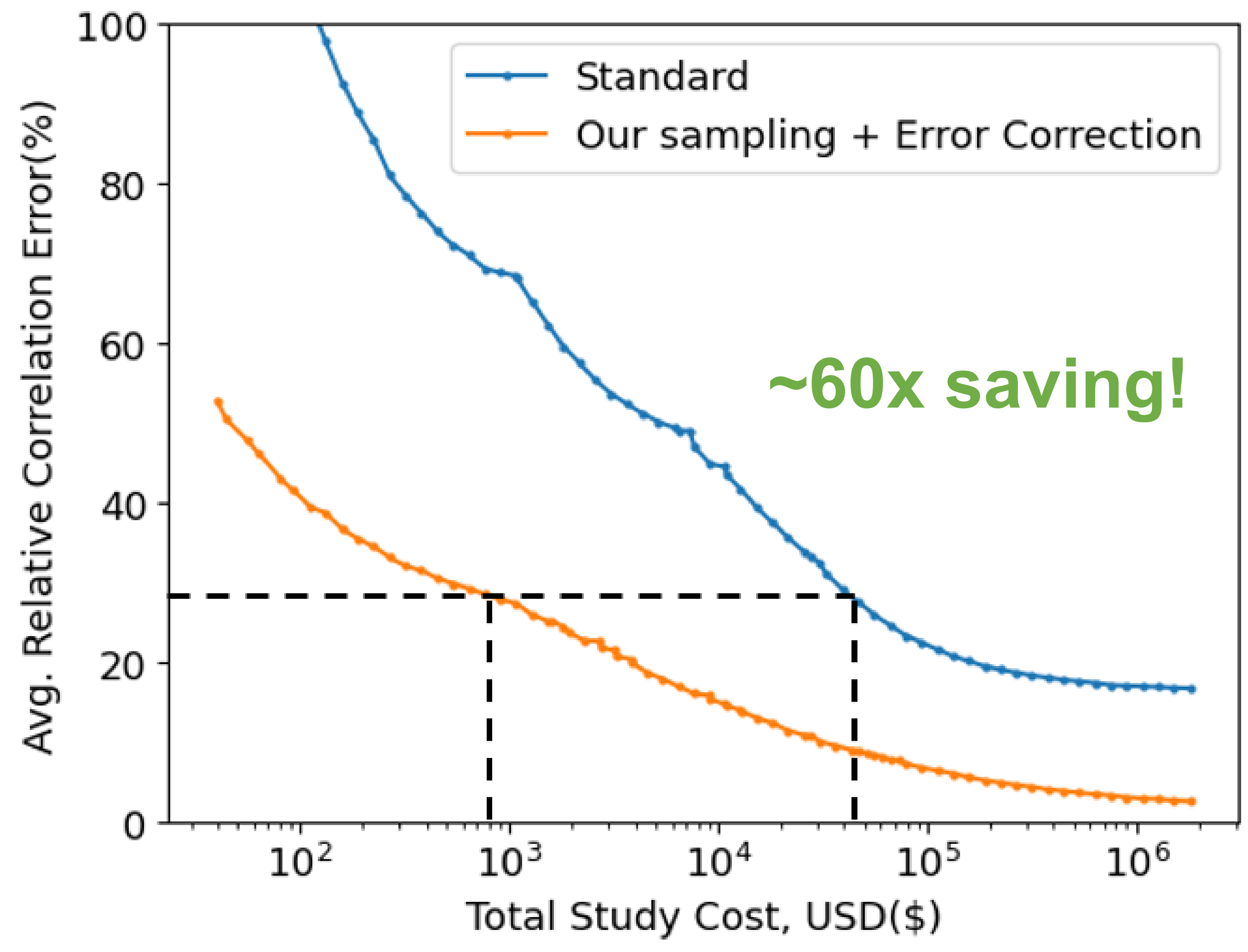
- Need to sample a subset of inputs to show raters
- We choose samples with Importance sampling (with correction) from distribution q that approximates the theoretical optimum

$$q(x_i) \propto \frac{1}{|\mathcal{D}|}|[\bar{a}_k]_i \cdot [\bar{c}_t^{siglip}]_i + \epsilon|$$

Contribution 2: Bayes with SigLIP prior

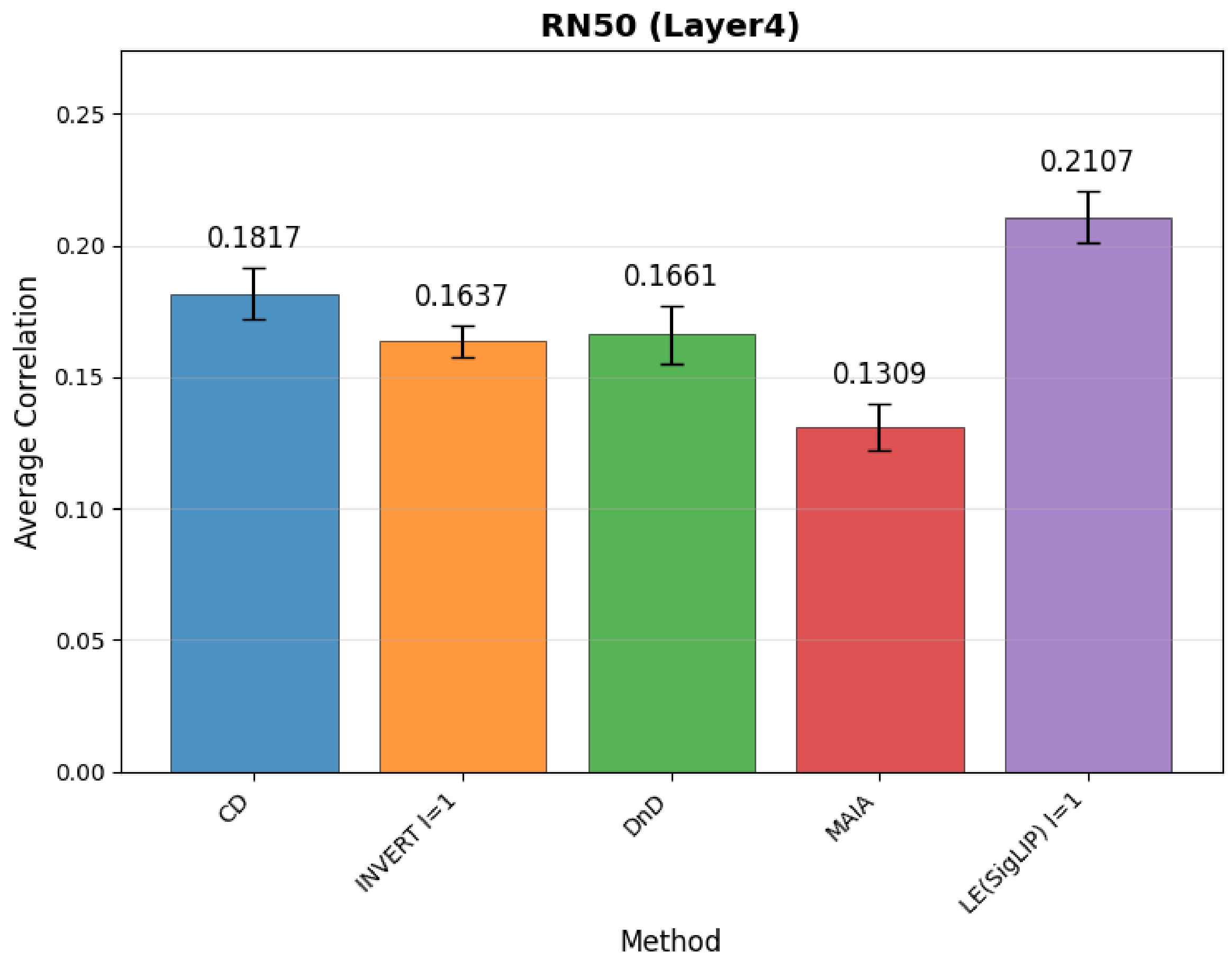
- Crowdsourced ratings are noisy -> Multiple Raters per input
- We show we can get more accurate results by using Bayes rule to estimate P(c | r_1, r_2, …) over typical methods like majority vote

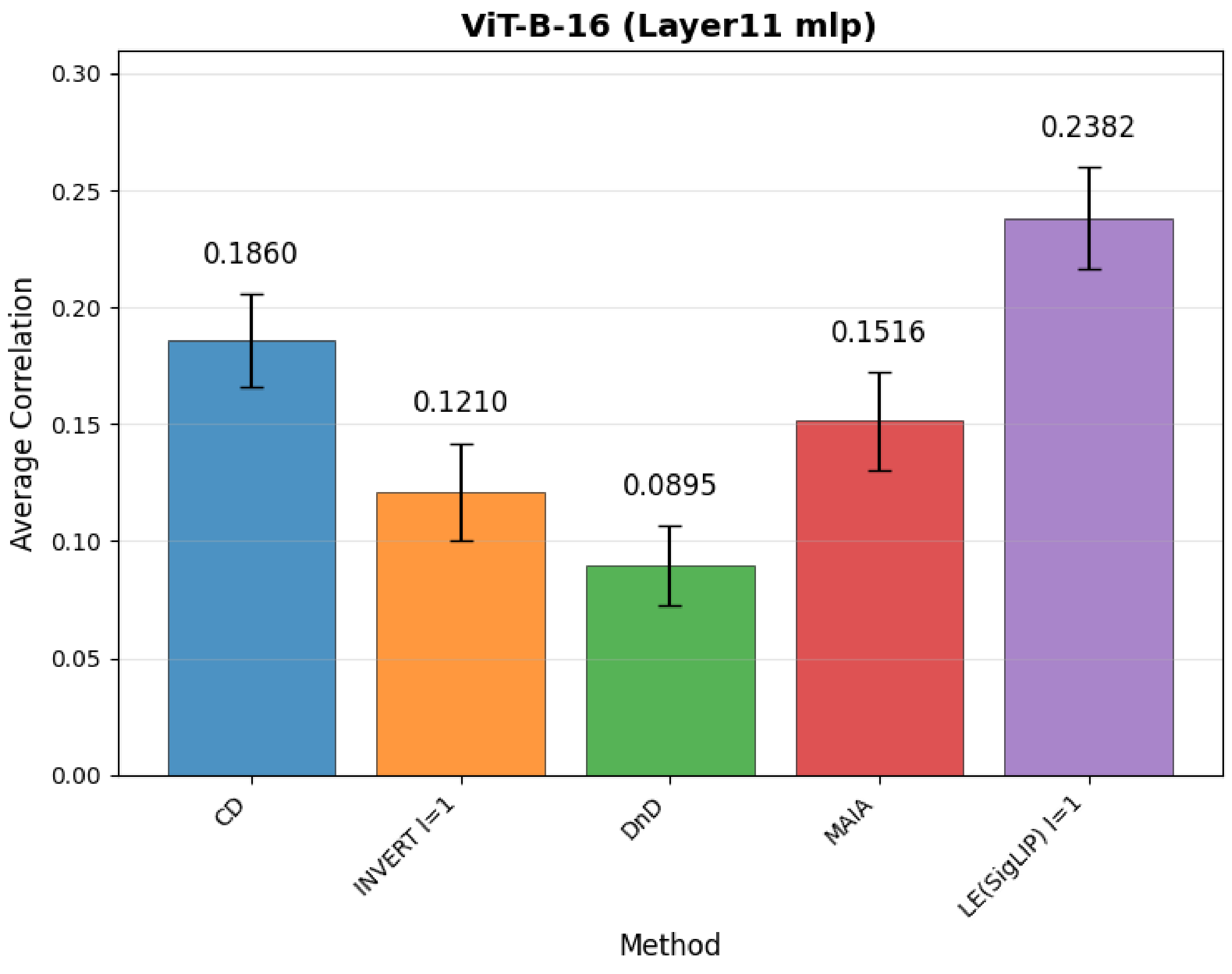
Combined these, we can reduce study cost from \$45,000 to \$800 with same accuracy!



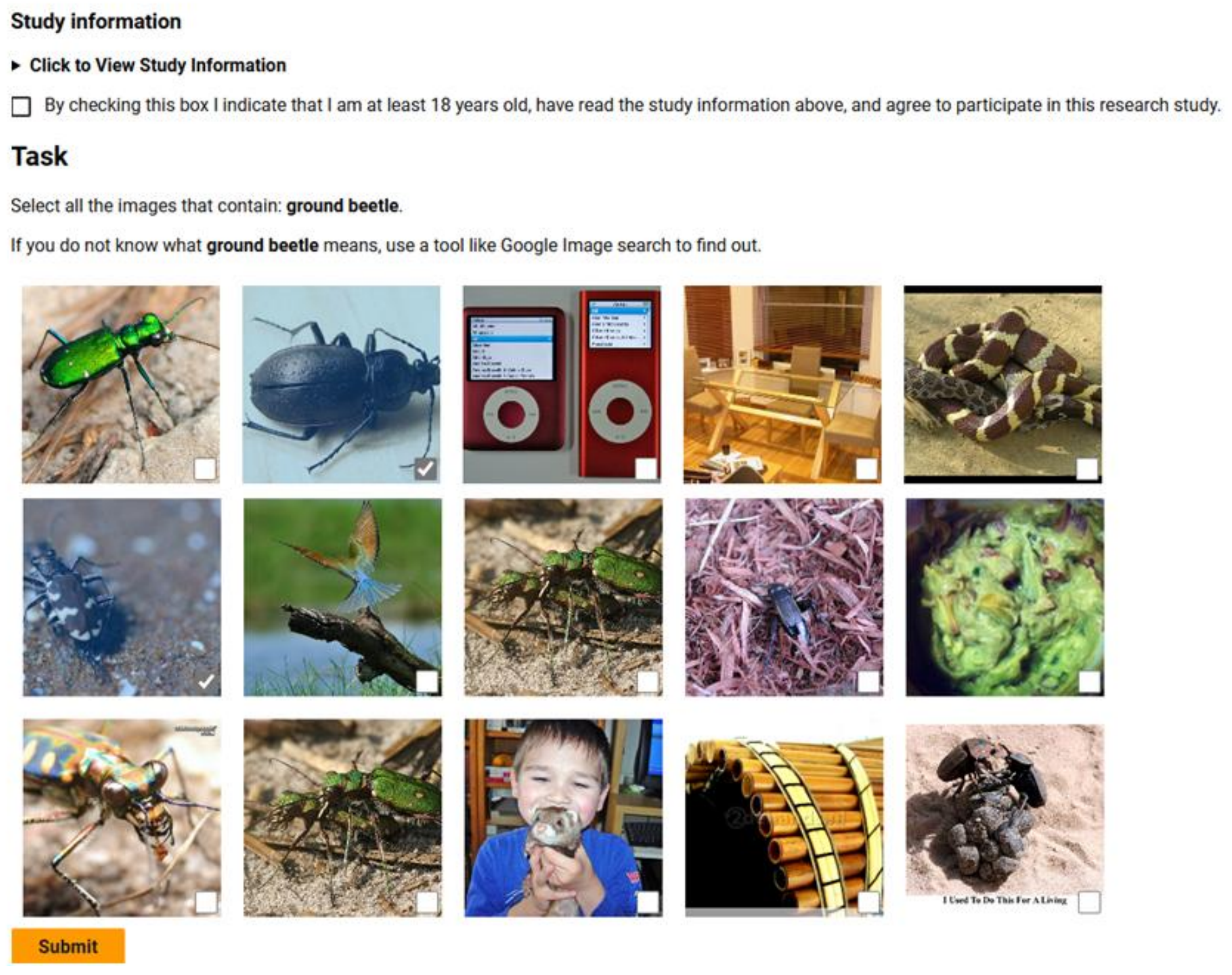
Results

- We evaluated explanations generated by best existing automated interpretability methods for 100 random neurons on two vision different networks
- Linear Explanation LE(SigLIP) performed the best on both Networks studied, even when restricted to produce length 1 explanations
- Notable LE significantly outperformed recent generative model-based methods MAIA and DnD
- Overall correlations relatively low, highlighting the need for more complex explanations or more interpretable architectures





User Interface



Study information

▸ **Click to View Study Information**

□ By checking this box I indicate that I am at least 18 years old, have read the study information above, and agree to participate in this research study.

Task

Select all the images that contain: **ground beetle**.

If you do not know what **ground beetle** means, use a tool like Google Image search to find out.

Paper **Code**