Supplementary Material: Efficient Unsupervised Shortcut Learning Detection and Mitigation in Transformers

A. ISIC Dataset

To assess the classifier's performance on ISIC images of malignant tumors with colored bandages (representing the worst group performance), we manually added colored bandages to malignant tumor images from the validation set. This was done by cutting patches from unused training images using the background removal model *tracer_b7*, available as an API on Replicate. We obtained cutouts of colored patches, which were then layered onto malignant tumor samples using GIMP, varying the size, color, and location of the patches based on the training distribution.

B. Knee Radiographs

Radiographic markers are frequently used to indicate the orientation and body part of the image. We obtained a cutout of an R (right body part) and L (left body part) marker from a hand x-ray image which we cutout with GIMP. We then automatically inserted the marker based on which knee (left or right) is visible in the image and varied in which corner (upper left and right as well as lower left and right) the marker is being added. We also added some slight rotation (between -5 and 5 degrees) to the added marker to introduce a more natural shortcut.

This follows the methodology introduced by Adebayo *et al*. [1] where they added a text ("MGH") as an artifical hospital tag on the image. Our approach occurs frequently in a variety of datasets which makes it even more natural.

C. Commonly used datasets

Recent research suggests that Vision Transformers are quite robust against spurious correlations in commonly used datasets such as Waterbirds and CelebA. Ghosal et al. [14] finetuned a ViT B-16 on Waterbirds resulting in a 96.75% average group accuracy and a 89.3% worst group accuracy. Similarly, they finetuned a ViT B-16 on CelebA resulting in an average group accuracy of 97.4% and a worst group accuracy of 94.1%.

We replicated their CelebA results by fine-tuning a ViT B-16 with the same hyperparameters as for all our other datasets, achieving above 90% AGA and WGA. Running our shortcut detection and mitigation framework on this dataset also replicates Li et al.'s findings [20] that eliminating one shortcut in this dataset will result in another being chosen by the model.

We detected a multitude of shortcuts which we were able to confirm as actual spurious correlations via the obtained labels (CelebA contains 40 labeled features including ”necktie”, ”glasses” and ”heavy-makeup”). A sample of a detected cluster can be seen in Figure 3.

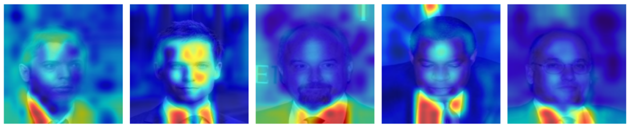

Figure 3. One of the many detected spurious correlations in the CelebA dataset are shirt collars for men.

We again ran *LLaMa-70b* to confirm the shortcuts in the prototypical patches and obtained the following summaries:

- ”The model seems to focus on blonde hair, long hair, ponytails, and beards/mustaches, which might be shortcuts for identifying women or men.”
- ”The model appears to focus on facial features like smiles, eyeshadow, lipstick, and moles, as well as accessories like glasses, ties, and shirts”

D. Clustering

We calculated the average cluster accuracies for all datasets for all three sequential seeds (see Table 4). As described in Section 3.2.1, clustering results are refined using a patch similarity measure, ensuring robustness even if the initial clustering is imperfect. Although we lack ground-truth annotations for shortcut feature locations, manual inspection of 200 ISIC prototypical patches confirms that all patches in the shortcut cluster contain the expected spurious feature.

	Accuracy (%) ↑
ISIC	76.0
Knee Radiographs	100.0
ImageNet-W	91.6

Table 4. Average KNN clustering accuracy for all three sequential seeds, using two clusters.

We also couldn't see any improvements in overclustering, as proposed by Sohoni et al. [29] (see Table 5).

	WGA(%) ↑	**AGA (%)** ↑
2 Clusters	61.0 ± 2.4	87.3 ± 1.2
3 Clusters	56.7 ± 3.4	86.2 ± 1.7

Table 5. Worst and average group accuracy (mean and standard deviation) after shortcut mitigation with different clusters.

E. User Study

We used the Replicate API to easily obtain results for multiple open-source LLMs. We decided on using three recently released open-source models with different parameter sizes:

- LLaMa3-8b: The smallest open source LLaMa model with 8 billion parameters.
- Mixtral8x7b: Mixture of experts architecture with 13 billion parameters.
- LLaMa3-70b: The LLaMa model with 70 billion parameters.

We prompted all three models with the same prompt: ”I extracted patches from images in my dataset where my model seems to focus on the most. I let an LLM caption these images for you. I am searching for potential shortcuts in the dataset. Can you identify one or more possible shortcuts in this dataset? Describe it in one sentence (only!) and pick the most significant. No other explanations are needed. Descriptions:” followed by the captions that we obtained via the *LLaVa-13b* model. The *LlaVa-13b* model was prompted with the prototypical patches and the text prompt ”What is in this picture? Describe in a few words.”.

The study was conducted using google forms. The participants were prompted with the dataset description and asked to identify which of the three responses was likely describes a spuriously correlated attribute. Often there were multiple correct answers, hence chance performance was 51.3%. Note that we used responses only based on the cluster our unsupervised method identified as the one most likely to contain spurious correlations. Therefore, the results of the survey validate that LLMs are capable of generating concepts that distill the properties captured by the patch prototypes.