Technical AI Safety Puzzle #1

Julian Quick

Task 1

The problem is presented alongside a corpus of inputs. Each input is associated with 8 true binary output labels y_i. Running an input through the model produces, at a specific intermediate layer, an activation vector L with 64 l_j values. At the end of the model chain, it outputs 8 σ_i sigmoid predictions of the probability of each category being present.

The challenge is to determine, for each of the eight features, whether a single linear function of L is sufficient to recover the true label y_i, or if the model's head must perform nonlinear computation on L to read that feature out.

My first step was a linear regression of the model's individual logits, z_i, with respect to L, $z_i \approx w_j l_j + b$ (Figure 1). It is immediately obvious that country is the nonlinear feature, as it is the only feature that yields an R^2 fit less than 0.97 (R^2=0.043). These regressions, and all regressions in this document, are fit on the training split, and all reported AUCs and R² values are computed on the held-out test split.

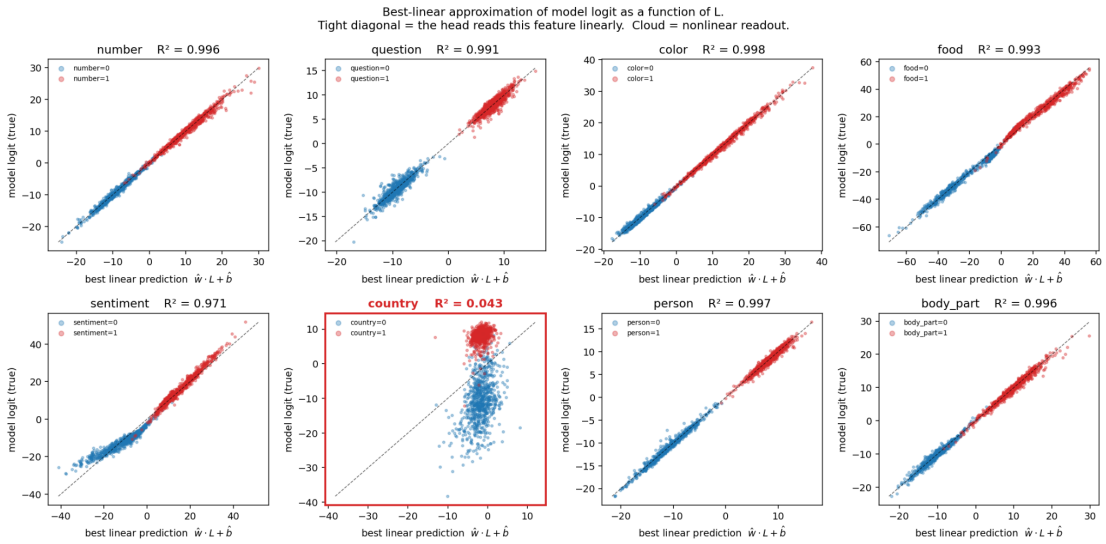

Figure 1: Best-linear approximation of category-specific logits as a function of the activation vector in the target layer.

But, we are interested in whether the logit of the probability of y_i is linear w.r.t. L, not σ_i. It is possible that these give very different answers. As a solution, I also perform linear probes to fit our own sigmoid functions to predict the true labels, $Pr(y_i) = (1 + \exp[-(w_j l_j + b)])^{-1}$. From

this we can measure the area under the curve (AUC) of our predictions. I find that all y_i features are associated with AUCs greater than 0.99, except country, which has an AUC of 0.47 (slightly worse than a coin toss). I found that a small multi-layer-perceptron model can map the activations, l_j, to a probability of category being present in the underlying text with AUC 0.993, indicating that the feature is fully decodable from the activations, just not linearly.

Based on these two experiments, I conclude that country is the nonlinear label.

Task 2

tl;dr: The encoding uses a gated sign-flipped representation:

$$w(\text{food}, \text{sentiment}) = \begin{cases} (-1)^{\text{food}} A & \text{if food} \oplus \text{sentiment} = 0 \\ (-1)^{\text{food}} B & \text{if food} \oplus \text{sentiment} = 1 \end{cases},$$

where A and B are directions in activation space and $\oplus$ is the XOR operator.

This task asks us to characterize the geometry of the nonlinear encoding. To do this, I used an active subspace model [1], which examines the eigen-decomposition of the mean (uncentered) covariance matrix of $\partial z_i / \partial L$ (for example, the active subspace of the country category is determined as the eigendecomposition of the mean value of the square matrix $(\partial z_5 / \partial l_j)(\partial z_5 / \partial l_k)$. The associated eigenvectors can be used to form “shadow plots”, showing how a quantity of interest behaves when looking at variation along a projection using a selected direction in the 64-dimensional space (Figure 2). As you can see, the sentiment category, and all categories except country, have 0/1 labels that are cleanly separable by plotting them along the direction pointed out by the leading eigenvector. By contrast, we need to look in three separate directions (as indicated by the eigenvalue decay in Figure 3: country needs the top three eigen-pairs to reach 99% of the explained variance) to properly capture the variation in the country label.

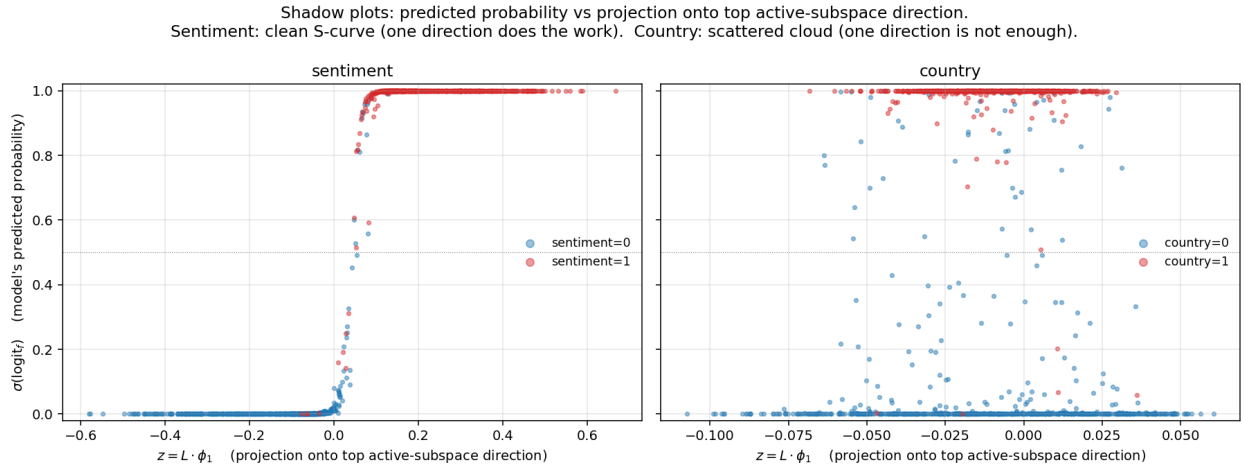

Figure 2: “Shadow plots” showing the model-predicted probability of sentiment (left) and country (right) categories being in input text plotted against the activation vector in the target layer projected into the dominant active subspace eigenvector.

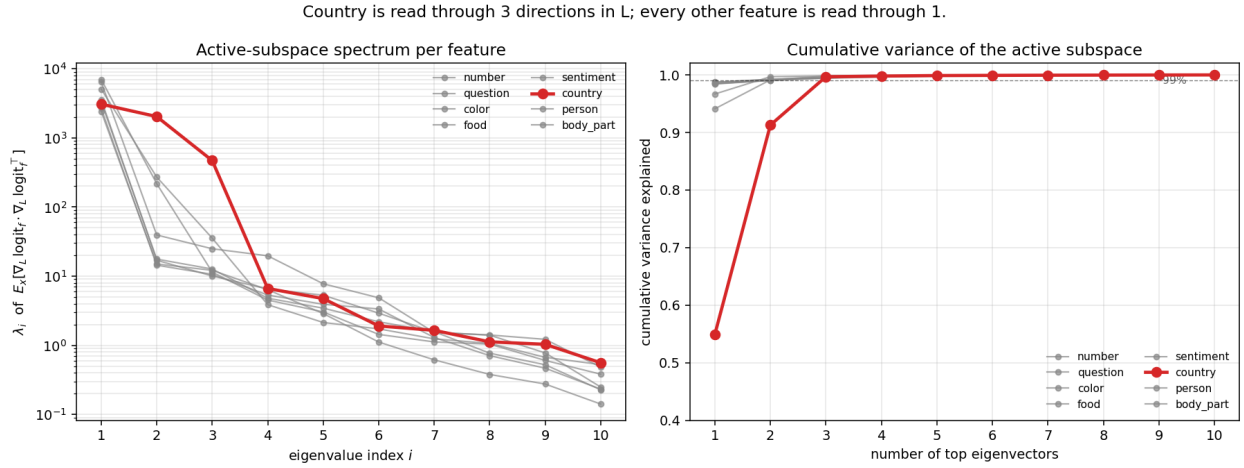

Figure 3: Results of the active subspace analysis using the logits associated with each category. The left plot shows the values of the leading eigenvalues. The right plot shows the cumulative variance explained by the different eigen-pairs.

This can be further understood by plotting the features y_i against the secondary and tertiary active subspaces of the activation space, L. I denote the ordered active subspace eigenvectors as ϕ_i. Plotting the second and third leading eigenvectors and coloring by y_i, we can pinpoint the exact pattern used to encode the country feature. The feature is present when the input text leads to activation values that fall into one of the blobs within the center and is inactive when the activation values correspond to the outer points. It appears to be related to the food and sentiment categories. ϕ_2 accounts for food and ϕ_3 accounts for sentiment.

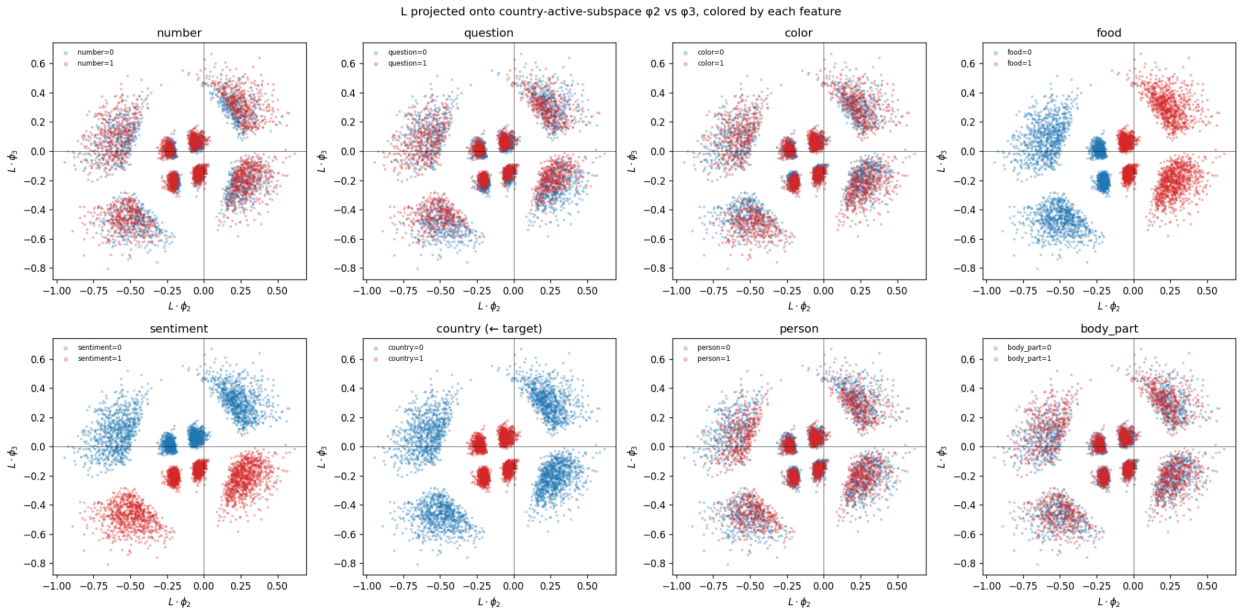

Figure 4: The activation vector of the target layer projected into the secondary (x-axis) and tertiary (y-axis) leading eigenvectors found from an active subspace analysis of the country logit values. Each panel is associated with one of the 8 examined categories, colored by the presence/absence of that category in the underlying text analyzed.

I also include the relevant plots of the country category when using eigenvectors 2 (ϕ_2) and 3 (ϕ_3) with eigenvector 1 (ϕ_1). You can think of this process as eigenvectors 2 and 3 setting the context we are in and eigenvector 1 being the mechanism to decide where to go given that context.

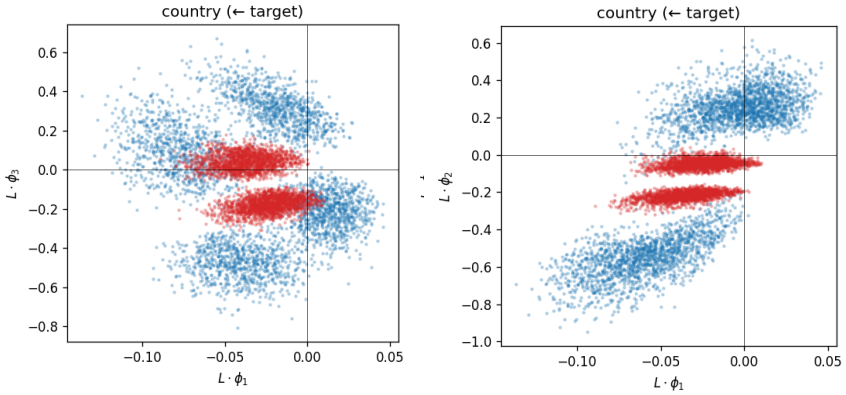

Figure 5: The activation of the target layer projected into: (left) the tertiary (y-axis) and primary (x-axis) active subspace directions; (right) the secondary (y-axis) and primary (x-axis) active subspace directions. Red points indicate the presence of the country category in the underlying text and blue points indicate its absence.

Based on this result, I decided to recompute the AUC from task 1, using logistic regression that is conditional on food and sentiment as decoded from the activations themselves: each input is assigned to a quadrant by examining its projection onto that feature's linear probe direction, thresholding at the training-set median projection value (which assumes a roughly balanced ~50% base rate for the presence of these gating features). When I do that, the AUC rises from 0.47 to 0.989. This indicates that the vast majority of variation is explained as country being conditional on sentiment and food, and that the gating context is itself linearly readable from L.

I steelman this argument by experimenting with different pairs of features to ensure this doesn't work with any old pair of features for setting context. I examine different candidate feature pairs for computing the conditional logistic regression. Making it conditional on two features means there are 4 weight vectors w_j . In Table 1, I report the examined feature pairs and the country AUC associated with a logistic regression conditioned on the two feature pairs.

Table 1: Country AUC for conditional probes. For each feature pair, the data is split into four quadrants, corresponding to the four possible binary values of the pair, as decoded from the activations by thresholded linear probes. In each quadrant, a distinct logistic regression is constructed, $Pr(y_5) = \sigma(w_j l_j + b)$ [country is feature #5], and the AUC is computed by pooling across all four quadrants.

Feature Pair	Country AUC conditional on feature pair
(food, sentiment)	0.989
(food, X) for X ∈ {question, person, number, body_part, color}	0.976–0.977
(sentiment, X) for X ∈ {number, body_part, question, color, person}	0.91
All pairs not involving food or sentiment	0.58–0.63

The strong AUC of the (food, sentiment) feature pair is consistent with my hypothesis that the model is splitting the computation conditionally on these features. No other feature pair can close this gap as tightly. The second and third rows of the table indicate that food is more important than sentiment, but that both are important gates. While the 0.977 AUC associated with gating food and with non-sentiment pairs is close to the 0.989 AUC with the (food, sentiment) pairing, the ~0.012 edge is real: a paired DeLong test on the test-set predictions puts every one of the five comparisons at a p-value less than 0.0004, well past the threshold even after correcting for running five of them. I further investigate this by reporting the cosine similarity between the w vectors associated with each quadrant (Table 2).

Table 2: Cosine similarities between the w vectors found in the logistic regressions, $Pr(y_5) = \sigma(w_j l_j + b)$, for each quadrant of observed food and sentiment binary values.

$cos(w^{(i)}, w^{(j)})$	food=0, sentiment=0	food=0, sentiment=1	food=1, sentiment=0	food=1, sentiment=1
food=0, sentiment=0	1.00	0.43	-0.48	-1.00
food=0, sentiment=1	0.43	1.00	-0.99	-0.46
food=1, sentiment=0	-0.48	-0.99	1.00	0.52
food=1, sentiment=1	-1.00	-0.46	0.52	1.00

This cosine similarity analysis, alongside the conditional logistic probes in Table 1 allows us to derive the formula in the tl;dr. The magnitude of the cosine similarity is approximately one only when food XOR sentiment agrees in the row and column, indicating that the model selects from two regression vectors based on this XOR gate. I refer to these as A [shared by quadrants (0,0) and (1,1)] and B [shared by (0,1) and (1,0)]. Flipping sentiment from 0 to 1 and holding food constant results in positive cosine similarities [e.g., $\cos(w^{(food=0,\, sentiment=0)}, w^{(food=0,\, sentiment=1)}) = 0.43$], while flipping food while holding sentiment constant results in negative cosine similarities [e.g., $\cos(w^{(food=0,\, sentiment=0)}, w^{(food=1,\, sentiment=0)}) = -0.48$]. This indicates that flipping food changes the sign of the regression coefficients. Putting this all together, we found that there are four sets of regression coefficients, +A, -A, +B, -B, where the sign is determined by the food category and the use of A versus B is determined by food XOR sentiment. This gated, sign-flipped structure perfectly explains the below-chance AUC of 0.47 observed in Task 1; a global linear probe learns the orientation of the most populated quadrant, which causes it to systematically misclassify the data in the sign-flipped quadrants, resulting in anti-predictive generalization.

This structure also explains the pattern in Table 1. Conditioning on (food, X) for irrelevant X resolves the sign but not the axis: each quadrant mixes +A with +B (or −A with −B), which are same-signed directions only ~60° apart, so a single probe along their bisector already reads both well (0.976). Only sentiment separates the two axes, which is why it adds +0.012 AUC while every other second feature adds nothing. The apparent dominance of food in Table 1 is therefore not evidence that sentiment is dispensable, it just indicates that sign errors are catastrophic while axis confusion is mild, and the +0.012 gap is precisely the signature of two distinct axes a small angle apart.

This sort of encoding can be constructed deliberately using LEAst-squares Concept Erasure (LEACE) [2] with an XOR gate trick. First, I train my own copy of the multi-layer perceptron head, using the same 384→64→64→64→64→8 architecture as the provided model, taking the same frozen text-encoder embeddings as input. No tricks, just standard stochastic gradient descent. Next, LEACE finds the linear combination of l_j , $v_j l_j$ that is most predictive of the country label such that $||v|| = 1$ (In LEACE, v is normally used for concept erasure, but here it is used for concept scrambling)[1]. Additionally, logistic probes are fit to find $w^{(i)}$ weight vectors associated with each category, $Pr(y_i) = \sigma(w^{(i)}{}_j l_j + b)$. The logistic regressions are used to determine an XOR function, $g(l_j w_j^{(1)}, l_j w_j^{(2)} \cdots)$. Then, I rewrite the perceptron to introduce a new fixed scrambling layer directly downstream of the target layer. Finally, a second stochastic gradient descent run is conducted, freezing all layers upstream of and including the scrambler layer, which conditionally flips the sign of the activation vector components along v as

$l_j' = l_j - 2\, v_j v_k l_k\, I(g)$,

where l_j' is the updated activation vector, $I(g)$ is a conditional gate / indicator function, and $||v|| = 1$. This scrambling operation is analogous to the LEACE method (which does not have this factor of 2, which results in obliteration along v). The full LEACE formula has a few extra terms accounting a non-zero mean and correlations between activation vector components, which I did not include in this setup. By scrambling the activations this way, I preserve the underlying linear structure while hiding it from naive linear probes via the conditional flipping.

When I apply this framework to train a new model, it reproduces the provided model's probe signature: linear probes on the scrambled layer fall to chance (AUC 0.52), while the retrained model recovers country near-perfectly (AUC 0.999).

The probe analysis and the reconstruction above are both consistency evidence: they show the gated encoding is decodable from L and that a model with this encoding can be built. They do not show that the provided model actually uses the code. In principle, the structure could sit in the activations while the downstream layers determine the presence of the country topic in some other fashion. So, I designed a causal test. I applied the same conditional reflection used in the scrambling layer above, but now to the provided model's own activations: each test input's activation is reflected across its quadrant's reader hyperplane, $l_j' = l_j - 2(\hat{w}_k l_k - t)\hat{w}_j$, where $\hat{w}$ is that quadrant's unit-normalized regression vector, t is the threshold (the median projection value in the training data), and quadrants are assigned by the food XOR sentiment condition. This reverses each activation's position along its quadrant's country readout while leaving all other coordinates unchanged. Under this setup, the model should now get the country prediction systematically backwards. That is what happens: country AUC falls from 0.994 to 0.026, and 94% of predictions flip. The five features not involved in the gate are essentially untouched

[1] For a single binary concept, v reduces to the unit-normalized cross-covariance between the activations and the country label, which is the difference of the mean value of L between the true and false cases.

(AUC ≥ 0.98). Food and sentiment prediction AUC drop to 0.91 and 0.83, which is expected collateral, since the quadrant readers overlap the gate directions substantially. Two controls confirm the effect is specific to this direction: reflecting across a random unrelated hyperplane changes nothing (country AUC 0.993); displacing each point by the same distance as the intervention, but along a direction orthogonal to all six fitted directions in the analysis (the four quadrant regression vectors and the food and sentiment probe directions), leaves country largely intact (AUC 0.910). Together, these confirm that the inversion is specific to the readout direction, not an artifact of moving points a large distance.

Task 3

I choose to define “more interesting” as an encoding that fundamentally cannot be recovered using conditional linear probes.

To accomplish this, I opted for a circular encoding scheme. I selected two orthonormal vectors in L, u^A and u^B, as well as a point in activation space that acts as the center of the circle, c. For each observed data point I compute $\theta = \arctan2[u^B (L - c), u^A (L - c)]$, which represented the angle of the activations projected into the (u^A, u^B) plane. The model is trained so that activations of inputs whose true country label is 1 fall in the sectors where $\sin(3\theta) > 0$, and country-absent inputs fall in the alternating sectors.

To construct this model, the layers upstream of the target layer are split into two independent 32-unit channels with no connections between them: one computes coordinates along u^A and u^B, the other computes the seven remaining features. This design ensures that country information exists nowhere in L except in the (u^A, u^B) plane[2].

If I train this architecture with a full random initialization, I can not get stochastic gradient descent to yield a model that reliably predicts the presence of the country category. I found that initializing the layer immediately downstream of the target layer with 24 units whose weights point at equally spaced angles around the circle in the (u^A, u^B) plane (each unit's post-ReLU response is a half-cosine bump in θ), and the output layer with a least-squares fit onto the target, allows training to succeed: the construction alone reaches country AUC 0.765, and fine-tuning the downstream layers brings it to 0.979 (1.5% less AUC than measured in the provided model).

[2] This would be a bad assumption if there was a correlation between the country category and another category. However, since a global linear probe yielded an unpredictive AUC of 0.539, and we demonstrated that all other categories are linearly decodable from the activations in the target layer, we can conclude there is no significant correlations between country and other categories, and that information about the country category lies only in the (u^A, u^B) plane.

To verify that this encoding meets my definition of “more interesting,” I ran the full probe suite from tasks 1 and 2 on the new model. Linear probes score 0.539. The conditional-linear toolkit that recovered the provided model's encoding at 0.989 tops out at 0.641 here. An exhaustive degree-2-logistic-regression sweep reaches at most 0.698. Degree-3 probes jump to 0.967, close to the model's own 0.979: the feature is fully present and readable, but by nothing below cubic regression [this is because the sign of $\sin(3\theta)$ is equal to the sign of $3x^2y - y^3$ in the unit plane[3]]. As a positive control, a logistic probe fit on the encoding's own features $(\sin[3\theta], cos[3\theta])$ predicts the presence of the country category at AUC 0.920, with generic degree-3 probes scoring slightly higher (0.967) only by learning corrections for points placed near sector boundaries. For comparison, the provided model's encoding can be captured with a generic quadratic logistic probe (AUC 0.992), which is consistent with the gated formula being a product of linear functionals[4].

Finally, I causally verified that the model uses the desired phase encoding. Rotating the activations within the (u^A, u^B) plane by $\pi/3$ (which maps each sector onto its opposite-parity neighbor) inverts the model's country predictions (AUC 0.074, with 85% of predictions flipping). Similarly, rotating by $2\pi/3$, the encoding's symmetry, preserves them with country prediction AUC of 0.941(the drop from 0.979 reflecting points near sector boundaries, which the imperfectly realized pinwheel misplaces under rotation). Scaling each point’s radius ($r = \sqrt{([u^A\,[L - c])^2 + (u^B[L - c])^2}$) by ×0.7 or ×1.5 results in a similar AUC when predicting country (0.94 / 0.97), confirming the code is carried by the angle in the (u^A, u^B) plane rather than the distance from the center. Under every one of these interventions the other seven features remain above 0.996.

Applying the same active subspace analysis to this model as we did in Task 2 yields a very different picture (Figure 6), and there are no clear patterns, save maybe a weak clustering in the “person” panel. For completeness, I also compare projections associated with the leading and secondary active subspace eigenvectors (Figure 7). Only the country category panel shows a distinct trend in the projected activation space, and it is not immediately obvious from this picture how the encoding works.

[3] This can be derived by plugging in Cartesian definitions of sin and cos into the Triple-Angle Identity.
[4] See theorem 3.11 in Chapter 3 of “Perceptrons: An Introduction To Computational Geometry” by Marvin Minsky and Seymour Papert.

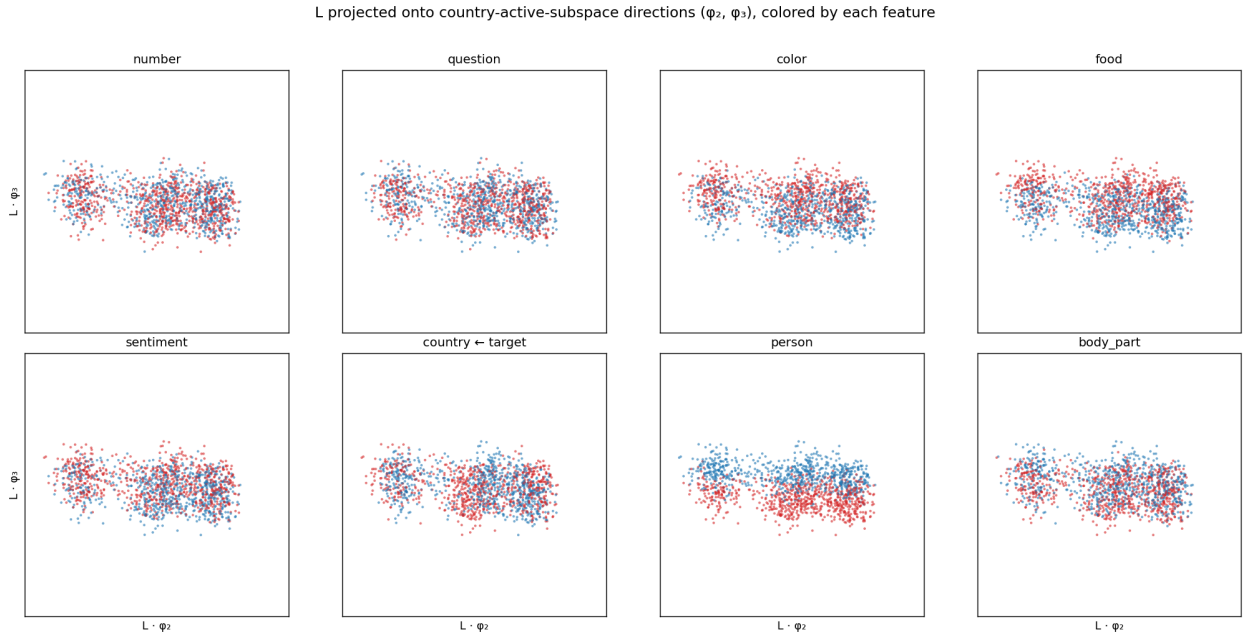

Figure 6: The activation vector in the target layer projected into the tertiary (y-axis) and secondary (x-axis) active subspace directions found by examining the covariance matrix of $\partial z_i / \partial L$. Each panel corresponds to a different category. Red points indicate the presence of this category in the underlying text and blue points represent its absence.

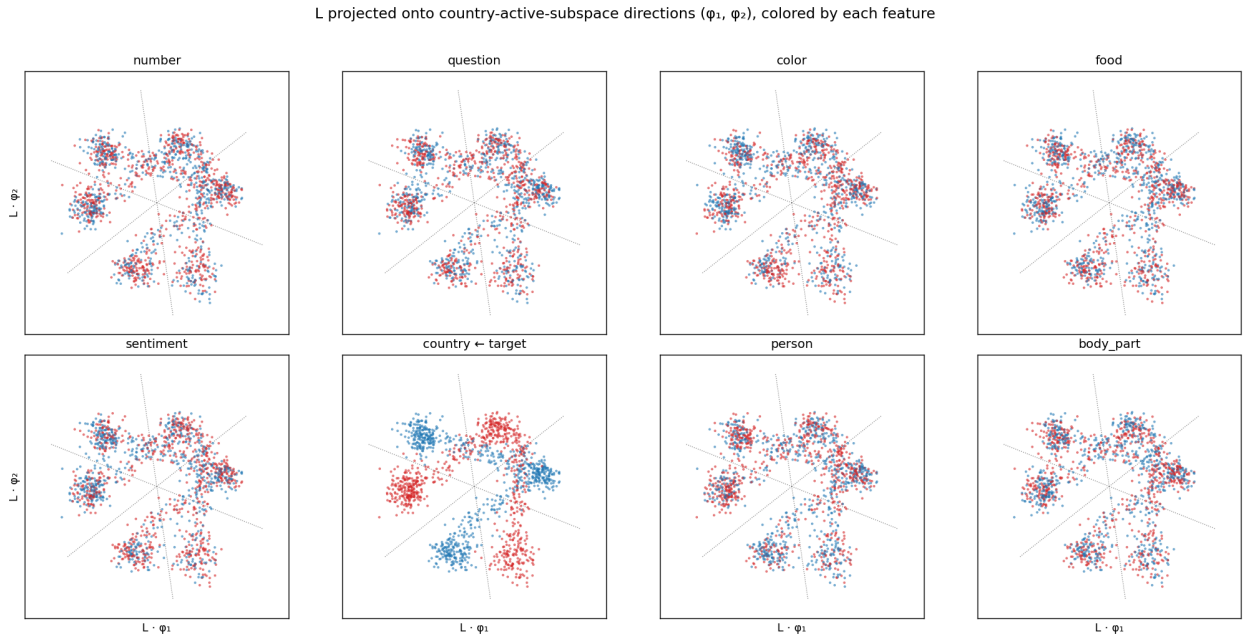

Figure 7: The activation vector in the target layer projected into the secondary (y-axis) and primary (x-axis) active subspace directions found by examining the covariance matrix of $\partial z_i / \partial L$. Each panel corresponds to a different category. Red points indicate the presence of this category in the underlying text and blue points represent its absence.

References

[1] Constantine, P. G., Dow, E., & Wang, Q. (2014). Active subspace methods in theory and practice: applications to kriging surfaces. SIAM Journal on Scientific Computing, 36(4), A1500-A1524.

[2] Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2023). Leace: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems, 36, 66044-66063.