LLM Probing with Contrastive Eigenproblems: Improving Understanding and Applicability of CCS

Stefan F. Schouten
Vrije Universiteit Amsterdam
s.f.schouten@vu.nl

Peter Bloem
Vrije Universiteit Amsterdam
p.bloem@vu.nl

Abstract

Contrast-Consistent Search (CCS) is an unsupervised probing method able to test whether large language models represent binary features, such as sentence truth, in their internal activations. While CCS has shown promise, its two-term objective has been only partially understood. In this work, we revisit CCS with the aim of clarifying its mechanisms and extending its applicability. We argue that what should be optimized for, is *relative* contrast consistency. Building on this insight, we reformulate CCS as an eigenproblem, yielding closed-form solutions with interpretable eigenvalues and natural extensions to multiple variables. We evaluate these approaches across a range of datasets, finding that they recover similar performance to CCS, while avoiding problems around sensitivity to random initialization. Our results suggest that relativizing contrast consistency not only improves our understanding of CCS but also opens pathways for broader probing and mechanistic interpretability methods.

1 Introduction

When Large Language Models (LLMs) perform well on benchmarks for a given domain or task, the results are sometimes questioned; in part because of a limited understanding of their working. How is it that LLMs do what they do? Without a clear picture of how an LLM approaches its tasks, we cannot verify if that approach is sensible, or how well it will do outside of benchmarks. The goal of Mechanistic Interpretability is to remedy this situation by identifying both: (1) what mechanisms are responsible for model behaviors; and, (2) what variables those mechanisms use, where they are encoded, and if they correspond to interpretable *features*.

This paper provides an in-depth look at Contrast-Consistent Search (CCS) [Burns et al., 2023]. This unsupervised probing method was introduced to determine if language models represent sentences as true or false. Being unsupervised, it has one advantage: it does not assume that the model's truth-values agree with human-authored labels. CCS has a two-termed loss function designed to find a parameter vector that makes the probe assign probabilities to sentences and their negations which add up to one. We perform an ablation of the method's loss terms and find that one of the terms is necessary for a different reason than what originally motivated its inclusion. We argue that contrast consistency should be defined in a *relative* way. Based on this insight, we find that CCS's objective can be made completely linear. This allows us to solve for contrast consistent directions using *Contrastive Eigenproblems*. This approach yields interpretable eigenvalues that provide additional insights. We demonstrate this by showing that datasets where CCS does not reliably find accurate probes are datasets that fail to isolate a single contrastive feature. We also demonstrate that Contrastive Eigenproblems are easily extended to settings with multiple features. We do so by replicating recent results which show that truth and polarity are encoded together in a shared subspace [Bürger et al., 2024].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability Workshop.

2 Understanding CCS

CCS is a probing method, meaning it involves training small classifiers on the activations of a larger model in order to establish whether certain information is present. The method yields a binary classifier, but unlike typical probes, a CCS probe is not given labels to train on. Instead, the probe exploits its inputs consisting of contrastive pairs, which we know to have opposite feature values.

Burns et al. [2023] focus on sentences and their negations to train the CCS probes. For example, they used inputs that consisted of question-answer pairs like "Is grass green? Yes/No", or of declarative sentences such as "Grass is green." and "Grass is not green.". Basically, the pair of inputs (X^+, X^-) consist of language that—if uttered—would amount to an assertion (X^+) or denial (X^-) of a proposition X. The method relies only on the expectation that any latent probability distribution captured by the model's internals must sum up to one. Of course, we would expect such a basic consistency property for all (binary) variables, not just truth. So in general, we have some binary *feature of interest*, and X^+ and X^- are inputs who primarily (and ideally, exclusively) differ in that feature's value. Typically, such inputs come in the form of minimal pairs where a word is changed, replaced or inserted in order to also change some sentence-level property. Unless specified otherwise, we take 'sentence truth' as our feature of interest.

The probes trained with CCS operate on a language model's mean-centered latent-space activations of X^+ and X^-, which we denote $\mathbf{x}^+$ and $\mathbf{x}^-$, respectively. In this work, we will use CCS with linear probes of the following form: $p(\mathbf{x}) = \sigma(\boldsymbol{\theta}^\intercal \mathbf{x})$, where $\boldsymbol{\theta}$ are the probe parameters. When using such linear probes we are assuming that there exists a direction in latent space that the language model uses to represent the feature of interest. By projecting activations on the *feature direction* we can construct a probe that parameterizes a probability distribution for the binary feature of interest.

The objective of CCS consists of a minimization of two terms, the consistency and confidence loss:

$$\begin{aligned}
\boldsymbol{\theta}_{\text{ccs}} &= \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}^+,\mathbf{x}^-} L^{\boldsymbol{\theta}}_{cons}(\mathbf{x}^+, \mathbf{x}^-) + L^{\boldsymbol{\theta}}_{conf}(\mathbf{x}^+, \mathbf{x}^-), \\
\text{with} \quad L^{\boldsymbol{\theta}}_{cons}(\mathbf{x}^+, \mathbf{x}^-) &= \left[\sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^+) - (1 - \sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^-))\right]^2 = \left[\sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^+) + \sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^-) - 1\right]^2, \\
\text{and} \quad L^{\boldsymbol{\theta}}_{conf}(\mathbf{x}^+, \mathbf{x}^-) &= \min\{\, \sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^+),\ \sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^-),\ 1-\sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^+),\ 1-\sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^-) \,\}^2.
\end{aligned}$$

The consistency loss is minimized when the probabilities assigned to sentences and their negations add up to one. The confidence loss is said to prevent the degenerate solution where $p(\mathbf{x}^+) = p(\mathbf{x}^-) = 0.5$. We use the symmetric (unbiased) confidence loss introduced by Farquhar et al. [2023].

Burns et al. [2023] solve this optimization problem using gradient descent, using activations for tokens from X^+ and X^-. For example, with "Between green and blue, grass is [green/blue]", the bracketed tokens would be used for $\mathbf{x}^+$ and $\mathbf{x}^-$ respectively.

We begin our analysis of CCS by reporting its performance in different circumstances. We use datasets from three sources [Burns et al., 2023, Marks and Tegmark, 2024, Schouten et al., 2025] (see Appendix E). We include probes trained on both the last and next-to-last tokens, corresponding to a *period* token and an *answer* token. We trained probes for a total of 9 datasets; using 30 different seeds for the probe's random initialization.

Table 1: Mean and standard deviation of accuracy for 9 datasets, trained on activations of the (next to) last token.

Dataset	*answer* (%)	*period* (%)
comparisons	100 ± 00	91 ± 00
sp_en_trans	99 ± 00	100 ± 01
cities	99 ± 00	99 ± 00
amazon	93 ± 00	93 ± 00
imdb	87 ± 00	87 ± 07
ent_bank	79 ± 10	82 ± 16
snli	71 ± 14	85 ± 06
copa	59 ± 06	48 ± 02
rte	54 ± 06	54 ± 07

In Table 1, we report probe accuracy for layer 16 in Llama-2-7b [Touvron et al., 2023], which we will use throughout the paper. We find that CCS does not reliably reach well-performing minima for all datasets. Specifically, there are multiple cases where the average performance is above random, but the exact performance varies between seeds. We wonder if this could be caused by the two-termed objective, thus our next step is an investigation of what makes the two terms necessary.

2.1 Loss-term ablations

Given that the stated purpose of the confidence loss is to avoid the degenerate solution, it makes sense to begin by determining what other strategies could help us avoid finding that solution. The

Table 2: Accuracies for probes trained with ablated and/or altered objectives.

$\mu \pm \sigma$ (%)		CCS -	$\mathcal{L}_{\text{conf}}$ -	$\mathcal{L}_{\text{cons}}$ -	$\mathcal{L}_{\text{cons}}$ +a1	$\mathcal{L}_{\text{cons}}$ +a2	$\mathcal{L}_{\text{cons}}$ +a1+a2	CCS +a1+a2
Marks and Tegmark [2024]	comparisons	100±00	100±01	66±11	62±09	65±11	59±07	100±00
	sp_en_trans	99±00	96±08	64±13	66±12	65±14	74±14	99±00
	cities	99±00	98±04	70±14	72±13	77±14	67±12	99±00
Burns et al. [2023]	amazon	93±00	94±00	67±09	65±11	72±13	64±09	94±00
	imdb	87±00	81±09	60±06	61±07	63±09	58±07	87±01

degenerate solution arises under the following conditions:

$$\sigma(\boldsymbol{\theta}^\intercal \overset{+}{\mathbf{X}}) = \sigma(\boldsymbol{\theta}^\intercal \bar{\mathbf{X}}) = 0.5 \qquad \implies \qquad \boldsymbol{\theta}^\intercal \overset{+}{\mathbf{X}} = \boldsymbol{\theta}^\intercal \bar{\mathbf{X}} = 0$$

where $\overset{+}{\mathbf{X}} = \left[\mathbf{x}_1^+, \mathbf{x}_2^+, \ldots, \mathbf{x}_N^+\right]^\intercal \in \mathbb{R}^{N \times D}$, is the data matrix for positive samples, and $\bar{\mathbf{X}}$ for negative samples. Thus, there are two ways for the degenerate case to arise: (1) the vector learned by the probe has zero length, $|\boldsymbol{\theta}| = 0$; or, (2) the direction points into the null space of the data matrix. When training probes, it is not uncommon to work with relatively small datasets. Thus, the number of samples can easily be smaller than the dimensionality of the model's latent space, resulting in a rank-deficient data matrix. To address both paths to the degenerate case, we can alter the CCS training process in two ways.

Alteration 1. By restricting the search space to unit vectors $\hat{\boldsymbol{\theta}}$ we avoid learning the zero vector. Note that the magnitude of the probe parameter vector is unimportant to its accuracy.

Alteration 2. To remove the null space from the data matrix, we use singular value decomposition: $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal = \text{SVD}(\overset{+-}{\mathbf{X}}_{train})$, where $\overset{+-}{\mathbf{X}}_{train} \in \mathbb{R}^{2N \times D}$ is a matrix containing both the hidden states for the training data's positive samples ($\overset{+}{\mathbf{X}}_{train}$), and the negative samples ($\bar{\mathbf{X}}_{train}$). We then apply the probes to the reduced representations: $p(\mathbf{x}) = \sigma(\hat{\boldsymbol{\theta}}\mathbf{V}_{:r}^\intercal \mathbf{x})$ (with r being the rank of $\overset{+-}{\mathbf{X}}_{train}$). This strategy assumes the null spaces of $\overset{+-}{\mathbf{X}}_{train}$, $\overset{+}{\mathbf{X}}$ and $\bar{\mathbf{X}}$ are the same.

Overall, we compare the following kinds of probes: (1) ordinary CCS; (2) only the confidence term ($\mathcal{L}_{\text{conf}}$); (3) only the consistency term ($\mathcal{L}_{\text{cons}}$), including versions altered in one ($\mathcal{L}_{\text{cons}}$+a1, $\mathcal{L}_{\text{cons}}$+a2) or both ways ($\mathcal{L}_{\text{cons}}$+a1+a2); and finally, (4) CCS with only the alterations, no ablations (CCS+a1+a2). For this experiment, we use the datasets for which CCS performed well in Table 1.

Results. In Table 2, we give the results of the ablation. A few things stand out: (1) the ablation of the confidence loss-term reduces accuracy more than the ablation of the consistency loss-term, suggesting the former is more important than the latter; and (2) the proposed alterations do not compensate for the ablation of the confidence loss-term. These results clearly show that the role of the confidence term is not limited to preventing the degenerate solution.

2.2 The effect of the confidence loss

The confidence term encourages probabilities closer to the extremes, but what needs to be true to make that happen? It is minimized when either the positive or negative sample of each pair are assigned a probability of zero or one. But, this would require that $\forall \mathbf{x} \,:\, \boldsymbol{\theta}^\intercal \mathbf{x} = \pm\infty$. Seemingly, minimizing the confidence-loss is just maximizing $||\boldsymbol{\theta}^\intercal \overset{+}{\mathbf{X}}||$ or $||\boldsymbol{\theta}^\intercal \bar{\mathbf{X}}||$.

When considering unit-length vectors $\hat{\boldsymbol{\theta}}$, the only way to maximize $\hat{\boldsymbol{\theta}}^\intercal \mathbf{x}$ is to reduce the angle between them. This would mean that the confidence-loss is biasing $\hat{\boldsymbol{\theta}}$ towards directions where the data has higher variance, i.e. the first (few) principal component(s).

To test this hypothesis, we will compare the directions found by CCS to the principal components. By ablating the two loss terms again, we can see if the confidence loss causes the direction to be more similar to the first few principal components. Specifically, for CCS, $\mathcal{L}_{\text{conf}}$-only, and $\mathcal{L}_{\text{cons}}$-only, we will measure: $\lambda^K(\boldsymbol{\theta}) = \frac{1}{||\boldsymbol{\theta}||} ||\mathbf{V}_{:K}\boldsymbol{\theta}||$. This measures how much of $\boldsymbol{\theta}$ extends into the subspace spanned by the first K principal components.

Results. In Figure 1, we can see that the confidence-loss indeed causes CCS to find directions closer to the first few principal components of $\overset{+}{\mathbf{X}}\overset{-}{}$. With only the consistency loss, the learned vectors have almost no magnitude in the subspace spanned by the first K principal components. When the confidence loss is added, its magnitude in the high-variance subspace grows. And finally, when we train with only the confidence loss, we find vectors that on average have a cosine similarity of over 0.5 with the first principal component. Results for other datasets and both answer and period token can be found in Appendix A.

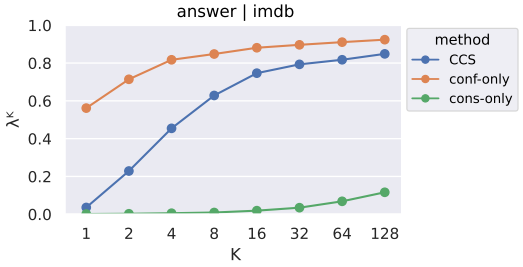


Figure 1: Extent to which learned vectors point into the subspace spanned by the first K principal components. Shown for: CCS, only the consistency loss (–conf), and only the confdidence loss (–cons); on the IMDB dataset, using activations for answer tokens, averaged over 30 random seeds.

Discussion. But why do salient directions make for more accurate CCS probes? We see two main reasons. First, by virtue of how contrastive data are created, the contrastive feature is often (one of) the first principal component(s). This is especially true when using the answer token itself to probe. Second, what we really care about is not absolute but relative contrast consistency. We have:

$$\arg\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbf{x}^+,\mathbf{x}^-}[\sigma(\hat{\boldsymbol{\theta}}^\intercal \mathbf{x}^+) + \sigma(\hat{\boldsymbol{\theta}}^\intercal \mathbf{x}^-) - 1]^2 \implies \arg\min_{\hat{\boldsymbol{\theta}}} ||\hat{\boldsymbol{\theta}}^\intercal(\overset{+}{\mathbf{X}} + \overset{-}{\mathbf{X}})||,$$

but if the variance is already low for that direction anyway, meaning $||\hat{\boldsymbol{\theta}}^\intercal \overset{+}{\mathbf{X}}\overset{-}{}||$ is already small, then the probabilities are close to 0.5, and the high consistency is not indicative of anything meaningful. The confidence loss biases CCS toward directions $\hat{\boldsymbol{\theta}}$ for which $||\hat{\boldsymbol{\theta}}^\intercal \overset{+}{\mathbf{X}}\overset{-}{}||$ is large, making it less likely that the contrast-consistency is simply due to $\hat{\boldsymbol{\theta}}$ pointing into a direction along which the hidden states already have low variance. However, this will always bias CCS towards high-variance directions, even when the contrastive feature is less salient. Thus, what we really want is for $||\hat{\boldsymbol{\theta}}^\intercal(\overset{+}{\mathbf{X}} + \overset{-}{\mathbf{X}})||$ to be small relative to $||\hat{\boldsymbol{\theta}}^\intercal \overset{+}{\mathbf{X}}\overset{-}{}||$.

3 The Geometry of a Binary Feature

Before continuing to test the 'relative contrast consistency' hypothesis we formulated in the previous section, we will first identify what we want from linear probes in the ideal case.

3.1 Two Kinds of Linear

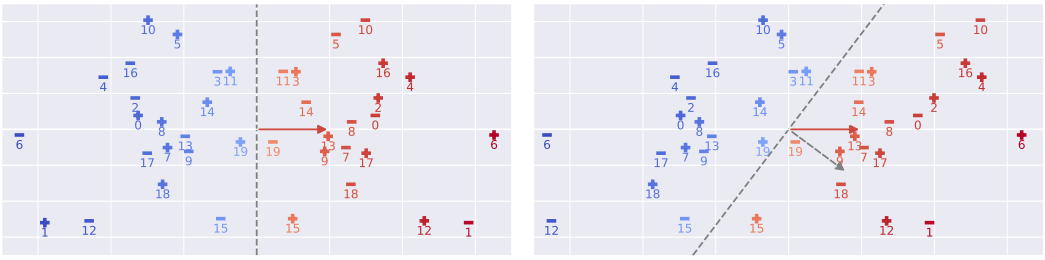
There are (at least) two different ways to think about what is involved in learning the latent-space direction associated with a given binary feature. On the one hand, we can think of our probe as learning to separate latent space into two regions that correspond to the possible values of the binary feature of interest. In that case, we are learning a hyperplane's normal vector $\mathbf{n}$. In the context of contrast pairs, we want to have the following property:

$$\forall_i \; : \; \text{sgn}(\mathbf{n}^\intercal \mathbf{x}_i^+) \; = \; -\text{sgn}(\mathbf{n}^\intercal \mathbf{x}_i^-). \tag{1}$$

That is, we want it to separate positive from negative samples. This property is what we need if we want to our probe to accurately predict the feature value. Logistic Regression is a common method to train such a probe. And, when previous work has used classification metrics such as accuracy, they were (implicitly) treating linear probes in this way. On the other hand, we can think of our probe as learning a direction $\mathbf{t}$ along which representations need to be translated for the model to treat the variable as having the opposite value. For contrast pairs, it is along this direction that a representation $\mathbf{x}^+$ would have to be translated to reach $\mathbf{x}^-$:

$$\forall_i \; \exists \alpha_i \; : \; \mathbf{x}_i^+ \; = \; \mathbf{x}_i^- + \alpha_i \mathbf{t}. \tag{2}$$

This is a direction we can use to intervene in a language model [ActivationAddition, Turner et al., 2024]. And, when previous work [e.g. Marks and Tegmark, 2024] used such interventions to test if a direction models a causal variable, they (implicitly) use this way of thinking about linear probes.



(a) Ideal scenario where $\mathbf{x}_i^+ = \mathbf{x}_i^- + \alpha_i \mathbf{t}$, and the only other feature is represented orthogonally to $\mathbf{t}$.

(b) Scenario where $\mathbf{x}_i^+ = \mathbf{x}_i^- + \alpha_i \mathbf{t}$, but the only other feature is represented obliquely to $\mathbf{t}$. The separating hyperplane's normal $\mathbf{n}$ is no longer the same as $\mathbf{t}$.

Figure 2: Comparison of feature alignments with $\mathbf{t}$ in two scenarios.

3.2 Contrast Error and Displacement

CCS most closely adheres to the classification-style linear probing, with its consistency loss requiring:

$$\sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^+) = 1 - \sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^-) \implies \sigma(\boldsymbol{\theta}^\intercal \mathbf{x}^+) = \sigma(-\boldsymbol{\theta}^\intercal \mathbf{x}^-) \implies \boldsymbol{\theta}^\intercal \mathbf{x}^+ = -\boldsymbol{\theta}^\intercal \mathbf{x}^-.$$

Thus, with $\boldsymbol{\theta} = \mathbf{n}$, it requires a stronger version of the property given in Equation (1). Contrasting samples must not only be on opposite sides of a hyperplane but also need to be equidistant from it. It follows that in the ideal case, we find $\boldsymbol{\theta}$ in the null space of the following matrix:

$$\boldsymbol{\theta}^\intercal (\mathbf{x}^- + \mathbf{x}^+) = 0 \implies ||\boldsymbol{\theta}^\intercal (\bar{\mathbf{X}} + \overset{+}{\mathbf{X}})|| = 0.$$

We call this matrix the commonality matrix $\mathbf{C} = \bar{\mathbf{X}} + \overset{+}{\mathbf{X}}$, since it captures the features that the positive and negative pairs have in common, i.e. the non-contrastive features that do not cancel out. For the intervention-style linear probing, we have:

$$\mathbf{x}_i^+ + \alpha_i \mathbf{t} = \mathbf{x}_i^- \implies \overset{+}{\mathbf{X}} + \boldsymbol{\alpha} \mathbf{t}^\intercal = \bar{\mathbf{X}} \implies \boldsymbol{\alpha} \mathbf{t}^\intercal = \bar{\mathbf{X}} - \overset{+}{\mathbf{X}}.$$

Therefore, in order to identify $\mathbf{t}$ we are looking for a rank-one decomposition of $\bar{\mathbf{X}} - \overset{+}{\mathbf{X}}$. Borrowing the terminology of Fry et al. [2023], we will call this matrix the displacement matrix $\mathbf{D} = \bar{\mathbf{X}} - \overset{+}{\mathbf{X}}$.

3.3 The Ideal Case

In Figure 2, we can see an idealized 2-dimensional representation how samples might be distributed in a model's latent space. In both subfigures, the direction $\mathbf{t}$ is rotated onto the x-axis. In Figure 2a, the only other feature is uncorrelated with the binary feature of interest.

However, we cannot generally assume that features are uncorrelated. Marks and Tegmark [2024] point out that the feature of interest (such as sentence truth) can be correlated with other features, thereby preventing classification-style probes from finding directions like $\mathbf{t}$. In Figure 2b, a feature is represented in a direction not orthogonal to the feature of interest. This non-orthogonal representation of the second feature, amounts to a shearing w.r.t. the situation in Figure 2a.

For both subfigures, we can see that: (1) the vectors $\mathbf{x}_i^- + \mathbf{x}_i^+$ lie on the separating hyperplane (the dotted grey line); and, (2) each $\mathbf{x}_i^- - \mathbf{x}_i^+$ lies on the x-axis. While in Figure 2a the hyperplane's normal vector also lies on the x-axis, in Figure 2b it does not. Assuming the feature of interested is not correlated with any other features, then $\mathbf{t} = \mathbf{n}$, but in other cases the vectors can differ.

The properties we have derived for the sum and difference vectors both involve changes in variance along the direction of interest. When the elements of the contrast pairs are summed, the variance *shrinks* (to zero for the ideal case) in the direction $\mathbf{n}$; and, when we take their difference, the variance *grows* in the translation direction $\mathbf{t}$ (while shrinking in all other directions, to zero in the ideal case).

3.4 Imperfections

In the introduction, we said "X^+ and X^- are inputs who primarily (and ideally, exclusively) differ in [a] feature's value". And it is certainly useful to pay attention to whether changes between positive

Table 3: Accuracy of DRC and RRC compared to CCS. Bold indicate results where our methods match or exceed CCS median and CRC-TPC.

	answer						*period*					
Dataset	min	CCS med	max	CRC -TPC	DRC	RRC	min	CCS med	max	CRC -TPC	DRC	RRC
comparisons	100	100	100	100	**100**	**100**	92	92	92	56	**93**	**94**
sp_en_trans	99	99	99	98	98	98	99	99	99	99	**99**	**99**
cities	99	99	99	99	**99**	**99**	98	98	99	51	**99**	**99**
amazon	94	94	94	94	**94**	**94**	92	93	93	53	**93**	**93**
imdb	86	87	88	87	**87**	**87**	87	88	89	87	**89**	**88**
ent_bank	84	86	87	82	84	**86**	48	93	94	86	89	90
snli	49	86	90	77	82	73	81	85	93	81	**87**	**87**
copa	51	55	68	54	53	52	47	47	52	49	48	47
rte	46	50	61	49	50	50	45	57	68	58	**58**	56

and negative samples are indeed as minimal and as closely tied to the feature of interest as possible. However, in practice it is impossible to perfectly isolate all features this way. It may be also be tempting to naively assume that if we did have the perfect contrastive dataset, that the properties we expect or desire from linear representations would hold exactly. However, besides any imperfections in the data, the model may also simply represent the data imperfectly. For both these reasons, in the next section, we will focus on finding directions with the highest increase or decrease in variance, rather than exact solutions to the equations given above.

4 Contrastive Eigenproblems

In Section 2, based on our experimental results, we formulated the hypothesis that the objective of CCS amounts to finding a direction with high *Relative* Contrast Consistency (RCC). In the first part of this section, we will test this hypothesis. Specifically, we will approach the problem of finding a direction with high RCC as a (generalized) eigenvalue problem. We will show that regardless of whether an such an approach is applied to consistency, or to displacement, we get the same solution in both cases. Despite the failure to distinguish intervention- and classification-style probes, contrastive eigenproblems still have two advantages: (1) we can use the eigenvalues to get an impression of how well the dataset succeeds in isolating a single feature, and (2) we can straightforwardly extend the approach to probing for multiple variables.

4.1 Problem formulation

Based on the intuitions we developed in the last section, we now propose to solve directly for increases/decreases in variance.

Difference-Relative Contrast (DRC). Here, we express decreases/increases in variance as differences in variance between $\mathbf{C}/\mathbf{D}$ and $\overset{+-}{\mathbf{X}}$, giving two eigenproblems:

$$\left(\mathbf{C}^\intercal\mathbf{C} - \overset{+-}{\mathbf{X}}{}^\intercal\overset{+-}{\mathbf{X}}\right)\mathbf{n}_k \;=\; \lambda_k\mathbf{n}_k \qquad \text{and,} \qquad \left(\mathbf{D}^\intercal\mathbf{D} - \overset{+-}{\mathbf{X}}{}^\intercal\overset{+-}{\mathbf{X}}\right)\mathbf{t}_k \;=\; \mu_k\mathbf{t}_k.$$

Negative values for λ_k correspond to directions $\mathbf{n}_k$ for which the variance in $\mathbf{C}$ is smaller than the variance in $\overset{+-}{\mathbf{X}}$. Thus, the last eigenvector of the lefthand problem is a good candidate for $\boldsymbol{\theta}$. And conversely, a positive value for μ_k indicates that $\mathbf{t}_k$ is a direction along which the variance in $\mathbf{D}$ is larger than in $\overset{+-}{\mathbf{X}}$. Now consider that with $\mathbf{A} = \overset{+}{\mathbf{X}} \pm \overset{-}{\mathbf{X}}$, both are instances of:

$$\left(\mathbf{A}^\intercal\mathbf{A} - \overset{+-}{\mathbf{X}}{}^\intercal\overset{+-}{\mathbf{X}}\right)\mathbf{v}_k \;=\; \nu_k\mathbf{v}_k$$

$$\implies \left(\overset{+}{\mathbf{X}}{}^\intercal\overset{+}{\mathbf{X}} + \overset{-}{\mathbf{X}}{}^\intercal\overset{-}{\mathbf{X}} \pm \left(\overset{+}{\mathbf{X}}{}^\intercal\overset{-}{\mathbf{X}} + \overset{-}{\mathbf{X}}{}^\intercal\overset{+}{\mathbf{X}}\right) - \overset{+-}{\mathbf{X}}{}^\intercal\overset{+-}{\mathbf{X}}\right)\mathbf{v}_k \;=\; \nu_k\mathbf{v}_k$$

And, because $\overset{+-}{\mathbf{X}}$ just contains the rows of $\overset{+}{\mathbf{X}}$ and $\overset{-}{\mathbf{X}}$:

$$\implies \left(\overset{+}{\mathbf{X}}{}^\intercal\overset{-}{\mathbf{X}} + \overset{-}{\mathbf{X}}{}^\intercal\overset{+}{\mathbf{X}}\right)\mathbf{v}_k \;=\; \pm\,\nu_k\mathbf{v}_k$$

Since, both eigenproblems involve the same two cross-terms between $\overset{+}{\mathbf{X}}$ and $\overset{-}{\mathbf{X}}$; they yield the same bases in opposite order. This means that formulating relative contrast consistency as a eigenproblem of differences in variance, forces us to assume $\mathbf{t} = \mathbf{n}$. This approach also turns out to be very closely linked to Contrast Consistent Reflection [Schouten et al., 2025] (see Appendix C).

Ratio-Relative Contrast (RRC). We can also formulate generalized eigenproblems:

$$\mathbf{C}^\intercal\mathbf{C}\mathbf{n}_k \;=\; \lambda_k \overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\mathbf{n}_k \quad\Longrightarrow\quad \lambda_k \;=\; \frac{\mathbf{n}_k^\intercal \mathbf{C}^\intercal\mathbf{C}\mathbf{n}_k}{\mathbf{n}_k^\intercal \overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\mathbf{n}_k},$$

and similarly for $\mathbf{D}$ and $\mathbf{t}$. Now the eigenvalues give ratios between the variances, rather than differences. Both of these problems are instances of:

$$\left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}} \pm \left(\overset{+}{\mathbf{X}}{}^\intercal \overset{-}{\mathbf{X}} + \overset{-}{\mathbf{X}}{}^\intercal \overset{+}{\mathbf{X}}\right)\right)\mathbf{w}_k \;=\; \omega_k \overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\mathbf{w}_k$$

Substitute $\mathbf{w}_k$ for $\left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\right)^{-\frac{1}{2}}\mathbf{w}'_k$ and pre-multiply with $\left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\right)^{-\frac{1}{2}}$, yielding:

$$\left(\mathbf{I} \pm \left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\right)^{-\frac{1}{2}} \left(\overset{+}{\mathbf{X}}{}^\intercal \overset{-}{\mathbf{X}} + \overset{-}{\mathbf{X}}{}^\intercal \overset{+}{\mathbf{X}}\right) \left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\right)^{-\frac{1}{2}}\right)\mathbf{w}'_k \;=\; \omega_k \mathbf{w}'_k$$

$$\Longrightarrow \left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\right)^{-\frac{1}{2}} \left(\overset{+}{\mathbf{X}}{}^\intercal \overset{-}{\mathbf{X}} + \overset{-}{\mathbf{X}}{}^\intercal \overset{+}{\mathbf{X}}\right) \left(\overset{+-}{\mathbf{X}}{}^\intercal \overset{+-}{\mathbf{X}}\right)^{-\frac{1}{2}} \mathbf{w}'_k \;=\; \pm(\omega_k - 1)\mathbf{w}'_k.$$

Thus, this formulation too gives the same bases, and also forces $\mathbf{n} = \mathbf{t}$.

4.2 Evaluation

We test both approaches and compare their classification accuracy to the min/median/max accuracy of CCS (over 30 seeds). The results can be seen in Table 3. What is clear is that when CCS converges to the same performance reliably, both approaches match that performance almost exactly. And, for those datasets where CCS is more sensitive to the random initialization, both approaches have accuracies somewhere between the minimum and the maximum of CCS. Generally both formulations perform very similarly, thus in the following experiments, we report results only for DRC. Another method proposed by Burns et al. [2023] is CRC-TPC, which simply takes the top principal component of $\mathbf{D}$. It can be seen to perform considerably worse on the period token for three datasets, likely because it finds a direction with high overall variance (see Appendix B).

4.3 Interpreting Eigenvalues

One of the benefits of approaching CCS as an eigenproblem, is that we get the whole basis of eigenvectors and their eigenvalues. One potential problem with contrast-based probing is that even if we construct the probing data ourselves, it is hard to be absolutely sure that we have truly isolated a single feature. Not only can features be hard to differentiate from each other, but an LLM may model matters in a way that does not map onto our understanding of the problem. Looking at the distribution of eigenvalues can be of help. If the contrasting data and the model's representation thereof meets our expectations, then the first eigenvalue should stand out from the rest, indicating one (and only one) direction is clearly contrast-consistent.

In Figure 3, we can see the top-10 eigenvalues for the different datasets. Looking at, for example, the 'amazon' dataset we see precisely what we wanted, the first eigenvalue is clearly larger than the rest. Going to 'copa', we see a somewhat flatter distribution of eigenvalues. For this dataset, there is a second eigenvector which we will look at in more detail. And for other datasets, like 'snli', we can see an even more diffuse distribution. Moreover, we consistently see more diffuse eigenvalues precisely in those cases where CCS's performance is less reliable (see Table 3).

Case study: COPA. Choice of Plausible Alternatives [Roemmele et al., 2011] is a commonsense causal reasoning dataset. It consists of prompts such as: "Consider the following example: "'The bar closed."' Choice 1: It was crowded. Choice 2: It was 3 AM. Q: Which is more likely to be the cause, choice 1 or choice 2? choice [1/2]."

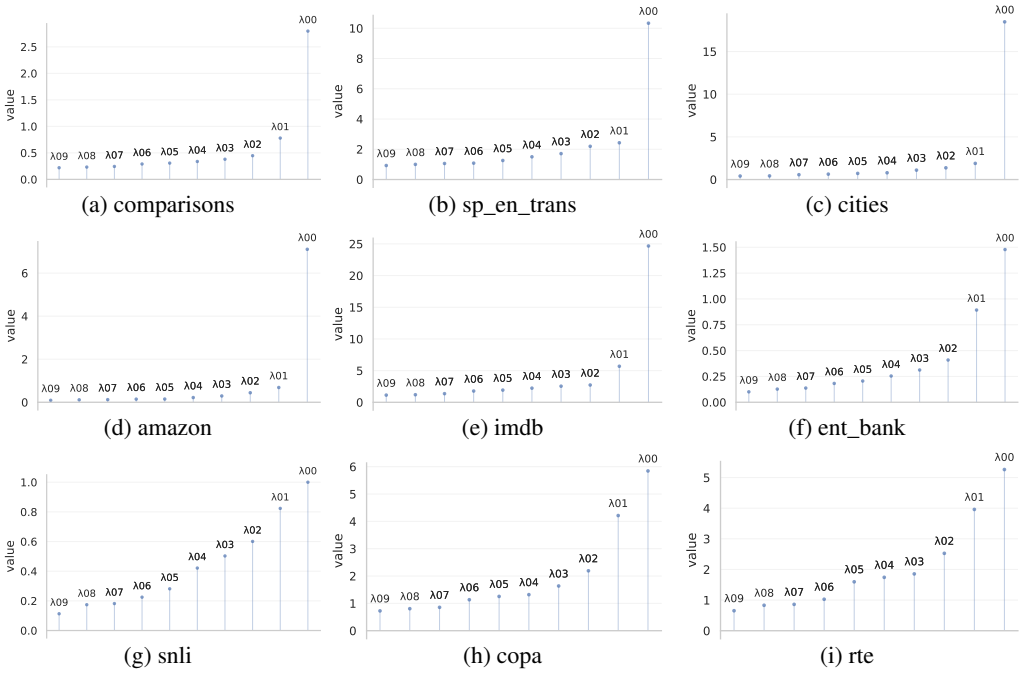


Figure 3: Top DRC eigenvalues for all datasets. Based on activations taken from the answer token.

Table 4: COPA samples, showing the prompt and cause/effect options along with their activation strength a on the contrastive feature encoded by DRC's top eigenvector. The choices labeled as commonsense are underlined.

	prompt	a	negative sentiment	a	positive sentiment
effect	'The host served dinner to hist guests.'	3.2	'His guests went hungry.'	-2.3	'His guests were gracious.'
	'A man cut in front of me in the long line.'	2.3	'I confronted him.'	-2.9	'I smiled at him.'
cause	'The shirt shrunk.'	1.7	'I poured bleach on it.'	-2.2	'I put it in the dryer.'
	'The smoke alarm went off.'	3.1	'I burnt my dinner.'	-2.2	'I lit a candle.'

None of the approaches perform well on this dataset, typically doing no better than random guessing. For the answer token, CCS does occasionally find directions that perform better at around 68%, suggesting that the model is not at fault. Given that the two highest eigenvalues seem to stand out among the rest, it makes sense to ask: (1) what does the top eigenvector represent, if not 'sentence truth'? and (2) does the second eigenvector encode 'sentence truth'?

To answer the first question, we looked at a subset of COPA together with the relevant activations, i.e. the projections of the answer token hidden states onto the first eigenvector. Looking at the most contrastive examples (where activations had the highest absolute values) quickly revealed the answer. It appears that in COPA, 'sentence truth' is not the only thing that changes between positive and negative samples, the answers often also differ in sentiment. In Table 4, we can see some examples of how high activations (left) correspond to the occurrence of comparatively 'bad' situations or events in answers (negative sentiment), and low activations (right) correspond to comparatively 'good' situations or events (positive sentiment). See Appendix D for full samples.

As for the second question, the answer is likely yes. DRC's second eigenvector predicts 'sentence truth' on COPA with 70% accuracy, which is slightly better than the CCS's 68% maximum.

4.4 Multivariate Extension: Polarity and Truth

The ability to find multiple directions can also be used deliberately. In recent work, Bürger et al. [2024] show that polarity and truth occupy a shared subspace. To showcase the utility of our approach in a multivariate setting, we will replicate their results. We use the 'cities' dataset, varying both the polarity (presence of negation) and the country that a city is said to lie in. For a given city, we denote the four samples as: $\mathbf{x}^{p,c}$, $\mathbf{x}^{p,i}$, $\mathbf{x}^{n,c}$, $\mathbf{x}^{n,i}$, where p and n indicate positive and negative polarity (negation), and c and i indicate the correct and incorrect country. Of these, only $\mathbf{x}^{p,c}$ (affirmation of correct country) and $\mathbf{x}^{n,i}$ (denial of incorrect country), are true statements. Between these four points, there are six pairs to be formed.

We can use DRC on the sum of all variants to cause the variance to decrease in multiple directions: $\mathbf{C} = \mathbf{X}^{p,c} + \mathbf{X}^{p,i} + \mathbf{X}^{n,c} + \mathbf{X}^{n,i}$ Equivalently, we can also concatenate the six contrast pairs, causing the variance to grow in multiple directions.

$$\mathbf{D} = \begin{bmatrix} \mathbf{X}^{p,c} - \mathbf{X}^{n,c} \\ \mathbf{X}^{p,i} - \mathbf{X}^{n,i} \\ \mathbf{X}^{p,c} - \mathbf{X}^{p,i} \\ \mathbf{X}^{n,c} - \mathbf{X}^{n,i} \\ \mathbf{X}^{p,c} - \mathbf{X}^{n,i} \\ \mathbf{X}^{n,c} - \mathbf{X}^{p,i} \end{bmatrix} \begin{matrix} \text{(polarity, truth)} \\ \text{(polarity, truth)} \\ \text{(truth)} \\ \text{(truth)} \\ \text{(polarity)} \\ \text{(polarity)} \end{matrix} \qquad \overset{+-}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{p,c} \\ \mathbf{X}^{p,i} \\ \mathbf{X}^{n,c} \\ \mathbf{X}^{n,i} \end{bmatrix}$$

Results. In Figure 4, we can see activations for the portion of the cities dataset we held out for evaluation, plotted by their coordinates along the first/second, and third/fourth eigenvectors found by DRC. We see clear separation between true and false statements by the first eigenvector. The second eigenvector separates statements by the truth of the base (unnegated) proposition. Finally, the third eigenvector separates statements by their polarity. Our results clearly show that there are three orthogonal directions (the first three eigenvectors) in Llama-2 that together encode truth and polarity.

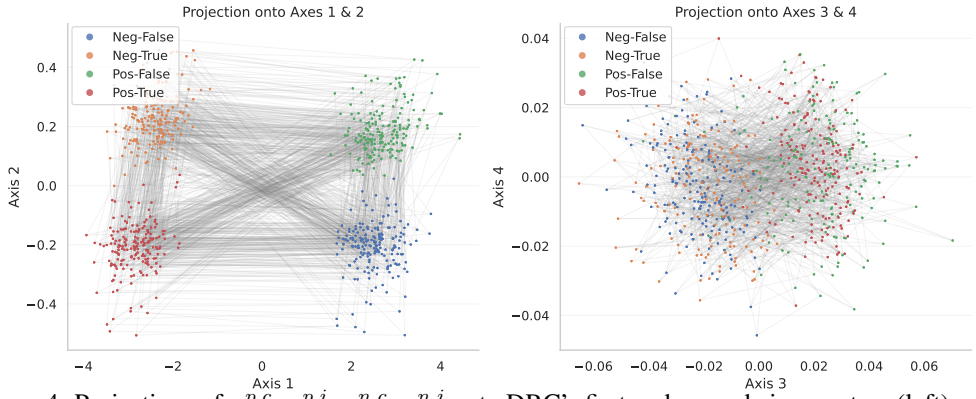


Figure 4: Projections of $\mathbf{x}^{p,c}$, $\mathbf{x}^{p,i}$, $\mathbf{x}^{n,c}$, $\mathbf{x}^{n,i}$ onto DRC's first and second eigenvectors (left) and its third and fourth eigenvectors (right). Grey lines show contrast-pairs.

5 Related Work

CCS has seen a number of other analyses since its publication. Fry et al. [2023] introduce the midpoint-displacement loss function. This function uses the same sum and difference vectors that make up our $\mathbf{C}$ and $\mathbf{D}$ matrices. They argue that CCS is optimizing for a trade-off between the angle to the sum vectors and the angle to the difference vectors. We argue that both should be relativized, removing the component that biases CCS to the directions of higher variance. Farquhar et al. [2023] give proofs and demonstrate empirically that CCS can find features other than truth. While this can get in the way of the original goal of 'eliciting latent knowledge', it can also be seen as an advantage that makes CCS more widely applicable. Our approach gives an orthonormal basis with eigenvalues that indicate to what extent each direction is contrast consistent. Thus, if there are multiple binary features present in our contrast pairs, this can be diagnosed and, if necessary, addressed. Levinstein and Herrmann [2024] also identify problems around isolating truth, and report CCS failing to learn the sentence truth feature under various experimental settings, possibly because it was learning the

another feature instead. Belrose et al. [2024] analyze and extend CRC-TPC, a closely related method. They show how the objective of CRC-TPC can be decomposed into separate interpretable terms, and then propose a number of additions, including a paraphrase invariance term and a supervised term. Similar to this work, their approach is also reducible to an eigendecomposition. While the solutions explored are similar, Belrose et al. focus on classification performance, while we explore how formulating contrastive eigenproblems can help diagnose problems in the data, and enable extensions to multivariate settings. Finally, Laurito et al. [2024] point to a problem where directions may be found that encode functions of features already represented along their own direction (e.g. the polarity-sensitive truth direction). To address it, they introduce a procedure where activations are clustered and then normalized (on both mean and variance) within each cluster. They show that this procedure improves the accuracy of CCS and CRC-TPC in various settings. While Laurito et al. try to eliminate less-relevant contrastive directions, we aim to find all of them.

6 Conclusion

We have explored: (1) how CCS functions; (2) what linear probes should learn in the ideal case; and (3) how CCS might be formulated as an eigenproblem and the advantages of doing so. We have argued that the confidence loss is an imperfect way of ensuring CCS probes find directions with high *relative* contrast consistency. We identified two ways of thinking about linear probes (classification-style and intervention-style) and how contrastive data can help to find them. In trying to solve for such probes by formulating eigenproblems, we have not succeeded in identifying distinct methods to solve for one of the two types of linear probes. However, what the eigenproblem approach does provide is: (1) interpretable eigenvalues that indicate how well our contrastive data isolates a single feature, and (2) a natural extension to the multivariate setting. Looking at the eigenvalues, we have seen that large variance in CCS's performance can be explained by datasets failing to isolate one and only one contrastive feature. Using multivariate contrastive eigenproblems, we have replicated recent results showing how truth and polarity are encoded in the latent spaces of language models. We believe these results show that Contrastive Eigenproblems provide a useful tool. It either yields an accurate probe, or the means to explain why such a probe is difficult to find for a particular dataset. Future work should look for contrastive probing techniques that can find separate directions which are optimal for either classification or intervention.

Acknowledgments and Disclosure of Funding

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/ or sponsors.

References

N. Belrose, A. Mallen, D. Ghosh, W. Laurito, K. O'Brien, A. Wan, B. Wright, A. Asai, and Y. Elazar. VINC-S: Closed-form Optionally-supervised Knowledge Elicitation with Paraphrase Invariance, May 2024. URL https://blog.eleuther.ai/vincs/.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *Proceedings of the Eleventh International Conference on Learning Representations*, Feb. 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

L. Bürger, F. A. Hamprecht, and B. Nadler. Truth is Universal: Robust Detection of Lies in LLMs. *Advances in Neural Information Processing Systems*, 37:138393–138431, Dec. 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f9f54762cbb4fe4dbffdd4f792c31221-Abstract-Conference.html.

S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah. Challenges with unsupervised LLM knowledge discovery, Dec. 2023. URL http://arxiv.org/abs/2312.10029. arXiv:2312.10029 [cs].

H. Fry, S. Fallows, I. Fan, J. Wright, and N. Schoots. Comparing Optimization Targets for Contrast-Consistent Search, Nov. 2023. URL http://arxiv.org/abs/2311.00488. arXiv:2311.00488.

W. Laurito, S. Maiya, G. Dhimoïla, O. H. W. Yeung, and K. Hänni. Cluster-Norm for Unsupervised Probing of Knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14083–14112, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.780. URL https://aclanthology.org/2024.emnlp-main.780.

B. A. Levinstein and D. A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, Feb. 2024. ISSN 1573-0883. doi: 10.1007/s11098-023-02094-3. URL https://doi.org/10.1007/s11098-023-02094-3.

S. Marks and M. Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Proceedings of the First Conference on Language Modeling*, Aug. 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University, Mar. 2011.

S. F. Schouten, P. Bloem, I. Markov, and P. Vossen. Truth-value judgment in language models: 'truth directions' are context sensitive. In *Proceedings of the Second Conference on Language Modeling*, July 2025. URL https://openreview.net/forum?id=2H85485yAb.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering Language Models With Activation Engineering, Oct. 2024. URL http://arxiv.org/abs/2308.10248. arXiv:2308.10248.

A Maximum variance effect of confidence-loss

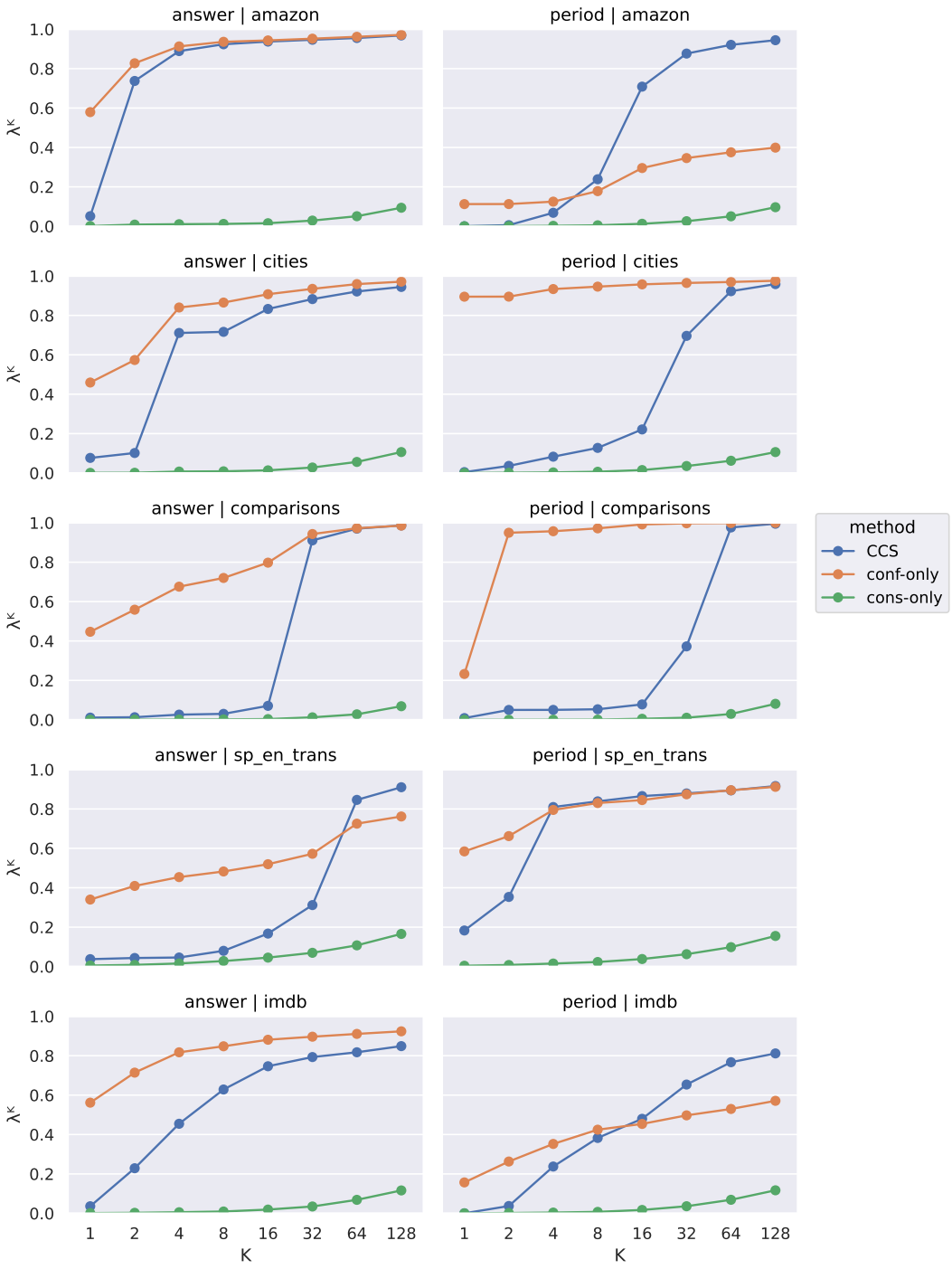


Figure 5: Effect of loss terms on probe parameter vector's similarity to top principal components.

B Performance of CRC on top 5 contrastive principal components

Here we show the performance of the principal components of $X^- - X^+$. When we use the top principal component this is equivalent to CRC-TPC [Burns et al., 2023]

Table 5: Accuracy of classifying with principal components of $\overset{+}{\mathbf{X}} - \bar{\mathbf{X}}$ for 9 datasets.

Dataset	Token	PC1	PC2	PC3	PC4	PC5
comparisons	answer	**1.00**	.47	.53	.45	.57
sp_en_trans	answer	*.98*	.45	.61	.46	.67
cities	answer	**.99**	.63	.55	.41	.55
amazon	answer	**.94**	.47	.51	.50	.54
imdb	answer	**.87**	.58	.51	.51	.52
ent_bank	answer	*.82*	.54	.48	.58	.51
snli	answer	**.77**	.59	**.77**	.60	.56
copa	answer	.54	**.71**	.52	.60	.50
rte	answer	.49	**.68**	.60	.53	.58
comparisons	period	.56	.52	.56	**.94**	.55
sp_en_trans	period	**.99**	.51	.56	.58	.61
cities	period	.51	**.97**	.58	.46	.59
amazon	period	.53	**.92**	.49	.54	.50
imdb	period	**.87**	.51	.52	.49	.53
ent_bank	period	**.86**	.51	.61	.59	.51
snli	period	**.81**	.67	.51	.58	.50
copa	period	.49	.46	.53	.48	**.65**
rte	period	.58	**.66**	.51	.59	.51

C Connections between DRC, CCR, and CRC-TPC

Belrose et al. [2024] helpfully point out that CRC-TPC can be broken down as follows. They remind us that: $\text{Var}(A - B) = \text{Var}(A) + \text{Var}(B) - 2\text{Cov}(A, B)$. Meaning the top principal component $\mathbf{w}^*$ of $\mathbf{D}$ can be written as:

$$\begin{aligned}
\mathbf{w}^* &= \underset{||\mathbf{w}||_2=1}{\arg\max}\ \mathbf{w}^\intercal \text{Cov}(\mathbf{D})\mathbf{w} \\
&= \underset{||\mathbf{w}||_2=1}{\arg\max}\ \mathbf{w}^\intercal \text{Cov}(\overset{+}{\mathbf{X}} - \bar{\mathbf{X}})\mathbf{w} \\
&= \underset{||\mathbf{w}||_2=1}{\arg\max}\ \mathbf{w}^\intercal \left(\text{Cov}(\overset{+}{\mathbf{X}}) + \text{Cov}(\bar{\mathbf{X}}) - \text{Cov}(\overset{+}{\mathbf{X}}, \bar{\mathbf{X}}) - \text{Cov}(\bar{\mathbf{X}}, \overset{+}{\mathbf{X}})\right)\mathbf{w},
\end{aligned}$$

where $\text{Cov}(\cdot, \cdot)$ denotes the cross-variance. Using notation from Section 2, we have:

$$\begin{aligned}
&= \underset{||\mathbf{w}||_2=1}{\arg\max}\ \mathbf{w}^\intercal \left(2\text{Cov}(\overset{+-}{\mathbf{X}}) - \text{Cov}(\overset{+}{\mathbf{X}}, \bar{\mathbf{X}}) - \text{Cov}(\bar{\mathbf{X}}, \overset{+}{\mathbf{X}})\right)\mathbf{w} \\
&= \underset{||\mathbf{w}||_2=1}{\arg\max}\ \mathbf{w}^\intercal \left(\overset{+-}{\mathbf{X}}^\intercal \overset{+-}{\mathbf{X}} - \overset{+}{\mathbf{X}}^\intercal \bar{\mathbf{X}} - \bar{\mathbf{X}}^\intercal \overset{+}{\mathbf{X}}\right)\mathbf{w}.
\end{aligned}$$

Schouten et al. [2025] introduce Contrast Consistent Reflection. They note that the objective of CCS requires that a pair of contrasting activations lie on opposite sides of, and equidistant from, a hyperplane. They propose that it may be beneficial to train probes that require points to also be each other's exact reflection through the hyperplane. Their proposed objective is:

$$\mathbf{r}^* = \underset{||\mathbf{r}||_2=1}{\arg\min}\ \mathbb{E}_{\mathbf{x}^+,\mathbf{x}^-}||\mathbf{x}^+ - (\mathbf{I} - 2\mathbf{r}\mathbf{r}^\intercal)\,\mathbf{x}^-||_2,$$

which is equivalent to:

$$= \underset{||\mathbf{r}||_2=1}{\arg\min}\ ||\overset{+}{\mathbf{X}}^\intercal - (\mathbf{I} - 2\mathbf{r}\mathbf{r}^\intercal)\,\bar{\mathbf{X}}^\intercal||_{2,1}.$$

With a Frobenius norm, we change the objective slightly, but allow for a closed-form solution.

$$\begin{aligned}\mathbf{r}^* &= \underset{||\mathbf{r}||_2=1}{\arg\min}\ ||\overset{+}{\mathbf{X}}^\intercal - (\mathbf{I} - 2\mathbf{r}\mathbf{r}^\intercal)\,\bar{\mathbf{X}}^\intercal||_F^2 \\ &= \underset{||\mathbf{r}||_2=1}{\arg\min}\ ||\overset{+}{\mathbf{X}}^\intercal - \bar{\mathbf{X}}^\intercal + 2\mathbf{r}\left(\mathbf{r}^\intercal\bar{\mathbf{X}}^\intercal\right)||_F^2\end{aligned}$$

With $||\mathbf{A}+\mathbf{B}||_F = ||\mathbf{A}||_F + ||\mathbf{B}||_F + 2\,\mathrm{tr}(\mathbf{A}^\intercal\mathbf{B})$, we have:

$$\begin{aligned}&= \underset{||\mathbf{r}||_2=1}{\arg\min}\ ||\overset{+}{\mathbf{X}}^\intercal - \bar{\mathbf{X}}^\intercal||_F^2 + 4||\mathbf{r}\left(\mathbf{r}^\intercal\bar{\mathbf{X}}^\intercal\right)||_F^2 + 4\,\mathrm{tr}((\overset{+}{\mathbf{X}}^\intercal - \bar{\mathbf{X}}^\intercal)^\intercal\mathbf{r}(\mathbf{r}^\intercal\bar{\mathbf{X}}^\intercal)) \\ &= \underset{||\mathbf{r}||_2=1}{\arg\min}\ 4\mathbf{r}^\intercal\bar{\mathbf{X}}^\intercal\bar{\mathbf{X}}\mathbf{r} + 4\,\mathbf{r}^\intercal\bar{\mathbf{X}}^\intercal(\overset{+}{\mathbf{X}}^\intercal - \bar{\mathbf{X}}^\intercal)^\intercal\mathbf{r} \\ &= \underset{||\mathbf{r}||_2=1}{\arg\min}\ \mathbf{r}^\intercal\bar{\mathbf{X}}^\intercal\overset{+}{\mathbf{X}}\mathbf{r}\end{aligned}$$

And, because the quadratic form only depends on the symmetric part:

$$\begin{aligned}&= \underset{||\mathbf{r}||_2=1}{\arg\min}\ \mathbf{r}^\intercal\left(\bar{\mathbf{X}}^\intercal\overset{+}{\mathbf{X}} + \overset{+}{\mathbf{X}}^\intercal\bar{\mathbf{X}}\right)\mathbf{r} \\ &= \underset{||\mathbf{r}||_2=1}{\arg\max}\ \mathbf{r}^\intercal\left(-\bar{\mathbf{X}}^\intercal\overset{+}{\mathbf{X}} - \overset{+}{\mathbf{X}}^\intercal\bar{\mathbf{X}}\right)\mathbf{r}\end{aligned}$$

Which is identical to both: (1) the terms that the cross-covariance terms contributed to the derivation for CRC-TPC; and, (2) the objective for the first (or last) eigenvector of DRC as shown in Section 4.

D Full COPA examples

Table 6: Strongly and weakly activating samples in COPA for DRC's first eigenvector. The answer choice corresponding to the high value is highlighted in bold.

Act. strengths		Prompt
3.06	-3.49	Consider the following example: "'I finished a page of the book."' Choice 1: **I ripped out the next page.** Choice 2: I turned to the next page. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
3.23	-2.25	Consider the following example: "'The host served dinner to his guests."' Choice 1: His guests were gracious. Choice 2: **His guests went hungry.** Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
3.28	-1.59	Consider the following example: "'The man contemplated the painting."' Choice 1: He felt in awe. Choice 2: **He collapsed.** Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
3.32	-2.48	Consider the following example: "'The woman filed a restraining order against the man."' Choice 1: The man called her. Choice 2: **The man stalked her.** Q: Which one is more likely to be the cause, choice 1 or choice 2? choice [1/2].
3.06	-2.18	Consider the following example: "'The smoke alarm went off."' Choice 1: I lit a candle. Choice 2: **I burnt my dinner.** Q: Which one is more likely to be the cause, choice 1 or choice 2? choice [1/2].
3.91	-2.94	Consider the following example: "'The scientist conducted an experiment."' Choice 1: She validated her theory. Choice 2: **She fabricated her data.** Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
2.58	-3.41	Consider the following example: "'The girl desired her parent's approval."' Choice 1: **She ran away from home.** Choice 2: She obeyed her parent's rules. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
2.66	-3.35	Consider the following example: "'The detective flashed his badge to the police officer."' Choice 1: **The police officer confiscated the detective's badge.** Choice 2: The police officer let the detective enter the crime scene. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
2.29	-2.94	Consider the following example: "'A man cut in front of me in the long line."' Choice 1: **I confronted him.** Choice 2: I smiled at him. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
1.69	-2.16	Consider the following example: "'The shirt shrunk."' Choice 1: **I poured bleach on it.** Choice 2: I put it in the dryer. Q: Which one is more likely to be the cause, choice 1 or choice 2? choice [1/2].
0.22	0.13	Consider the following example: "'A burglar broke into the house."' Choice 1: The homeowners were asleep. Choice 2: The security alarm went off. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
0.04	0.32	Consider the following example: "'The baby was wailing in his crib."' Choice 1: The mother picked up the baby. Choice 2: The baby crawled to the mother. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
0.04	0.33	Consider the following example: "'I pushed the gas pedal."' Choice 1: The car accelerated. Choice 2: The car door opened. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
-0.10	-0.32	Consider the following example: "'The investigators deemed the man's death a suicide."' Choice 1: He left a note. Choice 2: He had children. Q: Which one is more likely to be the cause, choice 1 or choice 2? choice [1/2].
-0.16	-0.35	Consider the following example: "'The girl performed in a dance recital."' Choice 1: Her parents showed her how to dance. Choice 2: Her parents came to watch the recital. Q: Which one is more likely to be the effect, choice 1 or choice 2? choice [1/2].
0.34	-0.58	Consider the following example: "'The man was bitten by mosquitoes."' Choice 1: He went camping in the woods. Choice 2: He fell asleep on his couch. Q: Which one is more likely to be the cause, choice 1 or choice 2? choice [1/2].

E Datasets

E.1 Prompts used for original CCS datasets

E.1.1 amazon

```
1 {
2     answer_choices: "negative ||| positive",
3     jinja: "Consider the following example: ’’’ {{content}} ’’’ Between {{answer_choices[0]}}
          and {{answer_choices[1]}}, the sentiment of this example is |||
          {{answer_choices[label]}}"
4 }
```

E.1.2 imdb

```
1 {
2     answer_choices: "negative ||| positive",
3     jinja: "The following movie review expresses what sentiment? {{text}} |||
          {{answer_choices[label]}}"
4 }
```

E.1.3 copa

```
1 {
2     answer_choices: "choice 1 ||| choice 2",
3     jinja: "Consider the following example: ’’’{{premise}}’’’ Choice 1: {{choice1}} Choice 2:
          {{choice2}} Q: Which one is more likely to be the {{question}}, choice 1 or choice 2?
          ||| {{answer_choices[label]}}"
4 }
```

E.1.4 rte

```
1 {
2     answer_choices: "incorrect ||| correct",
3     jinja: ’Assuming that the following is true: "{{text1}}"\nConcluding that: "{{text2}}" is |||
          {{answer_choices[label]}}’
4 },
```