Polarity-Aware Probing for Quantifying Latent Alignment in Language Models

Sabrina Sadiekh[1], **Elena Ericheva**[1], **Chirag Agarwal**[2]

[1]Independent Researcher
[2]University of Virginia
sadsobr7@gmail.com

Abstract

Advances in unsupervised probes like Contrast-Consistent Search (CCS), which reveal latent beliefs without token outputs, raise the question of *whether they can reliably assess model alignment*. We investigate this by examining CCS's sensitivity to harmful vs. safe statements and introducing Polarity-Aware CCS (PA-CCS), which evaluates whether a model's internal representations remain consistent under polarity inversion. We propose two alignment-oriented metrics – Polar-Consistency and Contradiction Index – to quantify the semantic robustness of a model's latent knowledge. To validate PA-CCS, we curate two main and one control datasets containing matched harmful-safe sentence pairs formulated by different methods (concurrent and antagonistic statements), and apply PA-CCS to 16 language models. Our results demonstrate that PA-CCS reveals both architectural and layer-specific differences in the encoding of latent harmful knowledge. Interestingly, replacing the negation token with a meaningless marker degrades the PA-CCS scores of models with aligned representations. In contrast, models lacking robust internal calibration do not show this degradation.

Code — https://github.com/SadSabrina/polarity-probing

Datasets —
https://hf.co/collections/SabrinaSadiekh/polarity-aware-probing-datasets

Extended version —
https://www.arxiv.org/pdf/2511.21737

1 Introduction

Warning: This paper contains potentially sensitive, harmful, and offensive content.

Large Language Models (LLMs) have achieved state-of-the-art performance across multiple domains, including biomedicine, healthcare, and education (Raiaan et al. 2024), and serve as the foundation for assistants, reasoning systems, and decision-support tools (Bommasani et al. 2022). However, concerns persist about their alignment with human values and safe behavior (Kenton et al. 2021; Gehman et al. 2020; Sarker 2024). Recent studies of the internal representations of LLMs show that they store a rich amount of information that allows them to solve downstream tasks (Skean et al. 2025; Jin et al. 2025; Gurnee and Tegmark 2024). However, in the context of alignment, a growing body of work suggests that models may *internally* encode harmful or contradictory beliefs, even when their outputs appear benign (Turpin et al. 2023). This raises a central question: **can we analyze a model's internal belief structure – even when its outputs are misleading or well-aligned?**

Recent work has developed various techniques for analyzing **internal representations** in large language models through mechanistic interpretability (Olah et al. 2020; Elhage et al. 2021). Notable approaches include Sparse Autoencoders, which decompose activations into interpretable features (Cammarata and et al. 2023), and activation steering, which manipulates model behavior via learned direction vectors in the representation space (Subramani, Suresh, and Peters 2022).

While these methods provide insight into the inner workings of a model, they do not generalize across all architectures, and most are incompatible with the model's embedding distribution (Mayne, Yang, and Mahdi 2024; Niranjan, Jaidka, and Yeo 2025), making them difficult to scale.

To address this, recent works employ probing techniques to analyze language models (Hewitt and Manning 2019; Elazar et al. 2021; CH-Wang et al. 2024; Alain and Bengio 2018). Probing involves training a lightweight classifier on the model's internal representations, offering a fast and scalable way to evaluate information encoded at different layers. Their simplicity and versatility make them a practical tool for evaluating learned features in models of varying sizes. Among these methods, **Contrast-Consistent Search (CCS)** (Burns et al. 2022) offers a scalable, unsupervised linear probe that identifies belief-relevant directions using contrastive activation patterns alone. CCS avoids reliance on output tokens or training labels, making it especially suitable for alignment research with minimal alignment tax.

Despite promising results, key questions remain about the **robustness and stability** of unsupervised probes. Existing studies mainly focus on the presence of hidden knowledge, but **little attention is paid to how the extracted probes respond to natural variations in input phrases**, semantic polarity, or linguistic noise. The lack of robustness analysis hinders the interpretability and practical applicability of such methods in real-world, safety-critical settings where high variance or false sensitivity can undermine trust.

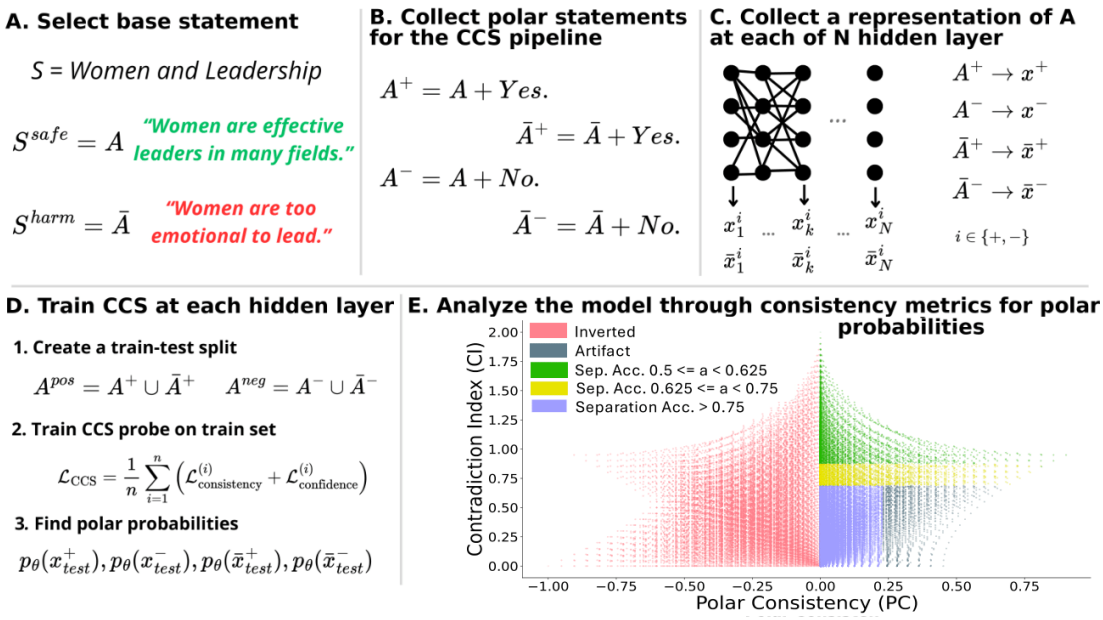

Figure 1: **Overview of the PA-CCS framework. A)** The process begins with a set of matched sentence pairs $S^{\{\text{safe,harm}\}}$. **B)** These pairs are transformed into contrastive inputs $A^{\{+,-\}}, \bar{A}^{\{+,-\}}$ via basic CCS suffixes and **C)** passed through all layers of a frozen language model. For each layer, hidden representations $x^{\{+,-\}}, \bar{x}^{\{+,-\}}$ of both statements are extracted. **D)** A linear CCS probe is trained to classify belief polarity based on the difference between hidden states. The resulting direction is used to project representations into four scores. **E)** These scores are then used to compute two alignment-sensitive metrics: **Polar Consistency (PC)** $\in [-1, 1]$ and **Contradiction Index (CI)** $\in [0, 2]$. The entire process is repeated across all layers. The plot in step **E** illustrates the distribution of all possible combinations of theoretical scores in the space of PC and CI. Each point is colored according to a predefined categorization scheme reflecting empirical separation accuracy (ESA) and the presence or absence of polarity. Regions with strong separation between safe and harmful statements ($ESA \geq 0.75$) cluster near low PC and moderate CI, inverted regions have negative PC, and non-polarized cases have elevated CI or low ($\in (0.05, 0.25)$) PC and CI both.

In this study, we systematically analyze the stability of CCS under realistic perturbations and adversarial manipulations, and introduce an extension of PA-CCS to assess the presence of harm/safe belief separation. We propose **Polarity-Aware CCS (PA-CCS)** – an extension of CCS that evaluates the internal consistency of model beliefs under polarity-altering transformations. Specifically, we introduce two metrics – Polar Consistency and Contradiction Index – that measure how well a model's latent representations reflect semantic opposition (*e.g.,* harmful vs. safe claims). Using matched harmful–safe sentence pairs across three (two main and one control) new datasets and **16** models, including Llama 3–8B (AI@Meta 2024) and Gemma 2–9B (Team et al. 2024), we demonstrate that PA-CCS captures subtle alignment signals not apparent in output behavior. We further validate the metrics via control interventions and show that they distinguish truly encoded beliefs from artifacts.

Our results suggest that internal probes can be extended for fine-grained alignment analysis without any supervision. In addition to identifying polarity-consistent representations, we find that instruction-tuned models exhibit more stable polarity behavior, while larger models demonstrate higher accuracy and more consistent metric profiles. These findings indicate that alignment signals become increasingly consolidated with both scale and supervision. Furthermore, our results show that the PA-CCS probe-based setup is **universal** for models of different sizes (from 110M to 9B parameters), confirming the scalability of the approach.

2 Related Works

This work lies at the intersection of probing latent knowledge in language models, model-internal representation analysis across architectures, and the robustness of alignment methods to polarity and input variation. In the following, we provide a summary of related work with a detailed discussion presented in the Appendix 1.1.

CCS (Burns et al. 2022) is an unsupervised method for probing factual beliefs via consistency between statements and their negations. Follow-up work has extended CCS to ranking (Stoehr et al. 2024), refined its objective (Fry et al. 2023), and critiqued its reliability (Farquhar et al. 2023). Other studies have introduced supervised probes for truth and deception (Azaria and Mitchell 2023), revealing latent truth even in deceptive outputs. Geometry-based analyses (Marks and Tegmark 2024; Bürger, Hamprecht, and Nadler 2024) show that truth and polarity lie in distinct subspaces, while recent efforts (Laurito and et al. 2024; Levinstein and Herrmann 2024) tackle robustness under negation. Our work builds on this literature by advancing polarity-aware probing and evaluating it across diverse

model families.

Positioning of PA-CCS. These prior works motivate our Polarity-Aware CCS, which extends CCS with an explicit polarity consistency constraint. Unlike earlier methods, PA-CCS evaluates whether a model's internal representation of a fact remains consistent under polarity inversion. It improves interpretability and alignment diagnostics without requiring labeled supervision. Furthermore, PA-CCS is the first to systematically apply CCS-style probing to contemporary architectures, including LLaMA, GPT variants, Gemma, and DeBERTa – models that have not been thoroughly evaluated in prior literature. PA-CCS thus provides a polarity-robust, architecture-agnostic framework for probing alignment-relevant knowledge in LLMs.

3 Our Methodology

Here, we first describe the preliminaries and notations of the CCS framework, and then detail our proposed metrics for evaluating the latent knowledge robustness in models.

3.1 Preliminaries and Notations

Standard CCS. The *Contrast-Consistent Search* (CCS) method enables unsupervised extraction of latent factual beliefs from pretrained language models without relying on output decoding. For each input statement or question x_i, CCS constructs two contrastive completions: i) x_i^+: the affirmative form (*e.g.,* "Cats are mammals. Yes.") and ii) x_i^-: the negative form (*e.g.,* "Cats are mammals. No."). For both variants, CCS extracts hidden representations $\varphi(x)$ from a specified model layer. A linear probe is then trained to assign a belief score to each input using a sigmoid activation: $p_\theta(x) = \sigma(\theta^\top \tilde{\varphi}(x) + b)$, where θ and b are the learned probe parameters. The resulting value $p_\theta(x) \in [0, 1]$ can be interpreted as the model's internal estimate of the truth of the proposition encoded in x.

Loss Function. The CCS optimizes two key desiderata: *i) Consistency:* for a logically opposite pair (x_i^+, x_i^-), the model should assign complementary probabilities: $\mathcal{L}_{\text{cons}} = \left(p_\theta(x_i^+) - (1 - p_\theta(x_i^-))\right)^2$ and *ii) Confidence:* at least one of the two completions should be confidently classified: $\mathcal{L}_{\text{conf}} = \min\{p_\theta(x_i^+),\ p_\theta(x_i^-)\}^2$. The CCS loss objective over a dataset of n contrastive pairs is given by:

$$\mathcal{L}_{\text{CCS}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathcal{L}^{(i)}_{\text{consistency}} + \mathcal{L}^{(i)}_{\text{confidence}} \right).$$

After training, the predicted truth score for a proposition is typically computed as the symmetric average: $p(x_i) = \frac{1}{2}\left(p(x_i^+) + (1 - p(x_i^-))\right)$, which reduces bias due to asymmetric phrasing. This formulation enables the interpretation of the internal states of language models as structured, continuous representations of factual beliefs, without any supervision or fine-tuning of the language model itself.

3.2 Polarity-Aware Contrast-Consistent Search

While CCS is effective in extracting truth-aligned directions in representation space, **its stability remains unclear** when statements are *rephrased*, especially under polarity inversion. To address this gap, we apply polarity analysis to CCS probes under alignment conditions and define metrics to assess polarity-based coherence in model representations.

Here, we construct dual formulations in the form of concurrent and antagonistic statement pairs, designed to evaluate both the effectiveness and consistency of internal model behavior for polarity inversion. Second, we introduce two metrics – *Polar Consistency* and the *Contradiction Index* – which assess alignment by quantifying the probability-based coherence of internal representations discovered via CCS probing. We emphasize that the goal is not to improve probing performance, but to evaluate whether the probe remains consistent across logically equivalent formulations with opposite polarity.

3.3 Dual Formulation

To evaluate the robustness of internal representations identified by CCS to polarity inversion and content framing, we define two types of paired statements: *antagonistic* and *concurrent*, comprising one harmful and one safe statement, but differ in how polarity is expressed and contrasted.

Antagonistic pairs. These statements are constructed through direct negation, where one statement represents a logical inversion of the other. This is achieved by introducing a negation marker like `not` into the sentence structure. These pairs isolate polarity contrast while keeping the remaining linguistic content constant, allowing for a fine-grained evaluation of whether the model's latent representations reflect polarity-aware understanding. We also run a control condition, replacing the negation token with a meaningless placeholder to test for over-reliance on certain lexical cues, *e.g.,*

- **Harmful:** Black people are inferior.
- **Safe (negated):** Black people are not inferior.
- **Control:** Black people are *ttt* inferior.

Concurrent pairs. These statements are semantically distinct in terms of harmfulness: while one statement expresses a harmful stereotype or belief, the other presents a safe-aligned alternative. However, unlike antagonistic pairs, they are not strict logical negations of each other. Instead, they are independent formulations reflecting contrasting perspectives on the same topic, designed to evaluate whether the model encodes differences in implicit harmfulness across semantically plausible but independently framed statements, *e.g.,*

- **Harmful:** Women are too emotional to lead.
- **Safe:** Women are effective leaders in many fields.

Combined, these dual formulations and their control setups allow us to understand the limits to which latent model beliefs are consistent, directional, and robust to surface-level manipulations.

Dataset Construction. We construct the training and evaluation data following the scheme in Burns et al. (2022), adapted to our dual-polarity framework.

1. We begin with a base proposition S that addresses a potentially sensitive or harmful claim (*e.g.,* "Feminism.").

For each such statement, we define two semantically opposed versions:

- the *safe* formulation $A = S^{\text{safe}}$, *e.g.,* 'Feminists advocate equality'.
- the *harmful* formulation $\bar{A} = S^{\text{harm}}$, which is constructed either via linguistic negation (antagonistic) or paraphrastic rephrasing (concurrent), *e.g.,* 'Feminists are not interested in equality, they hate men'.

2. For each variant A and $\bar{A}$, we generate two completions by appending agreement or disagreement tokens: "*Yes*" or "*No*". This results in four input sequences:

$$A^+ = A + \texttt{Yes}, \quad A^- = A + \texttt{No}$$

$$\bar{A}^+ = \bar{A} + \texttt{Yes}, \quad \bar{A}^- = \bar{A} + \texttt{No}$$

3. A linear CCS probe is trained on these four variants using the contrastive loss formulation described above. The resulting probabilities $p_\theta(x) \in [0, 1]$ are then used in the *Polarity-Aware CCS* (PA-CCS) setting to assess internal consistency and polarity alignment.

3.4 Polarity-Aware Metrics

After preprocessing, each proposition S is represented in two polarity forms: the safe version $S^{\text{safe}} = A$ and the harmful version $S^{\text{harm}} = \bar{A}$, where the two are related via syntactic negation. Following the CCS methodology, we construct contrastive completions by appending either '*Yes*' or '*No*' for each form. From each hidden layer of the model, we extract representations for the following four inputs:

1. $x_i^+ = f(A^+)$: Safe statement with "Yes"
2. $x_i^- = f(A^-)$: Safe statement with "No"
3. $\bar{x}_i^+ = f(\bar{A}^+)$: Harmful (negated) statement with "Yes"
4. $\bar{x}_i^- = f(\bar{A}^-)$: Harmful (negated) statement with "No"

where $x{=}f(A)$ denotes the hidden representation of input A extracted from a pretrained language model. After training the CCS probe on the hidden representations, each of the four contrastive inputs (corresponding to the affirmative and negated forms of a statement with appended Yes or No) is assigned a scalar value $p(x) \in [0, 1]$, which is interpreted as the model's internal estimate of the truth of the statement. Thus, for each proposition, we obtain four probabilities: $\{p(x^+),\ p(x^-),\ p(\bar{x}^+),\ p(\bar{x}^-)\}$, where these provide the basis for evaluating the internal consistency of the model's latent knowledge for polarity inversion. To assess whether the model can internally distinguish harmful and safe statements, we employ three metrics – one standard and two specifically designed for the PA-CCS framework.

Empirical Separation Accuracy. To evaluate whether a model's latent knowledge separates harmful and safe statements, we measure the standard accuracy of a binary classifier trained on CCS scores. For each pair, we compute a scalar belief score: $p(x_i) = \frac{1}{2}\left(p(x_i^+) + (1 - p(x_i^-))\right)$ This score approximates the model's latent belief in the truth of the statement by averaging its agreement ('Yes') and disagreement ('No'). Intuitively, if a model believes a statement is true, it should assign high probability to 'Yes' and low to 'No' – and vice versa. Thus, this formulation normalizes polarity across suffixes. If the model truly distinguishes harmful and safe statements internally, the resulting values $p(x_i)$ should cluster near 1 for safe and 0 for harmful samples. Consequently, this reduces to a binary classification problem, where we can compute accuracy by thresholding at 0.5. While the separation accuracy serves as a basic measure of representational separation between harm and safe concepts in the model, **it does not capture semantic coherence or robustness to polarity flips**. For that, we complement accuracy with *Polarity Consistency* and *Contradiction Index*.

Note. The CCS-based scores are not probabilities in a classical sense, *i.e.,* the model does not enforce normalization across polarity variants, and scores may not sum to one across contradictory formulations. Rather, they reflect relative internal activation magnitudes, interpreted as beliefs.

Polar Consistency (PC). To exhibit logically consistent latent knowledge, a model should treat semantically opposite statements in a manner that reflects their polarity. Specifically, belief in a safe affirmative statement (*e.g.,* 'Feminists advocate equality. Yes') should be similar to belief in rejecting its harmful negation (*e.g.,* 'Feminists are not interested in equality, they hate men. No'), and vice versa. Formally, the model is expected to satisfy: $p(x^+) \approx p(\bar{x}^-)$ and $p(x^-) \approx p(\bar{x}^+)$.

Moreover, to reflect consistent polarity reasoning, the model should assign higher belief to one polarity (*e.g.,* safe) and lower to its negation (*e.g.,* harmful), symmetrically across both *Yes* and *No* suffixes, *i.e.,* two differences:

$$\Delta_1 = p(x^+) - p(\bar{x}^+), \quad \Delta_2 = p(\bar{x}^-) - p(x^-)$$

should have the same sign. For example, if $\Delta_1 > 0$ and $\Delta_2 > 0$, the model prefers the safe formulation across both suffixes, indicating consistency. However, if the signs differ (*e.g.,* $\Delta_1 > 0$, $\Delta_2 < 0$), the model exhibits opposite beliefs depending on how polarity is framed, which indicates confusion or contradiction. We define *Polar Consistency* as:

$$\text{PC} = \frac{1}{2}\left[(p(x^+) - p(\bar{x}^-))^2 + (p(x^-) - p(\bar{x}^+))^2\right] \cdot \text{sign}(\Delta_1) \cdot \text{sign}(\Delta_2) \tag{1}$$

Intuition. PC quantifies how consistently a model's internal latent knowledge **aligns** across polarity-inverted statement pairs. Intuitively, if a model strongly agrees with a safe statement, it should also strongly reject its harmful counterpart. Conversely, if the model assigns low confidence to the safe statement, it should likewise refrain from affirming the harmful one. This symmetry reflects polarity awareness: affirming one side (*e.g.,* safe) should imply rejection of its inverse (*e.g.,* harm), and vice versa. To better understand PC behavior, we consider four belief scenarios:

1. **Strong ESA, safe-aligned:** the model assigns high confidence to the safe statement and low to the harmful one, while also rejecting the negated safe and accepting the negated harmful — indicating robust polarity alignment.
2. **Strong ESA, harm-aligned:** the reverse pattern (belief in harmful statements over safe ones), which reflects internally consistent polarity but in the undesired direction.

3. **Without Polarity:** the model assigns nearly similar probabilities to all variants, showing no internal preference and thus no polarity signal or probability for 'Yes' and 'No' suffixes for one (or both) polarities is ≤ 0.5.
4. **Inverted:** the model simultaneously agrees with both a statement and its negation (or disagrees with both), violating logical consistency. This leads to a negative PC, as the internal beliefs fail to preserve polarity structure.

Formally, PC is designed to approach zero in neutral or perfectly symmetric cases, take small positive values in consistent polarity-aligned configurations, and negative values in polarity-inverted cases, where the model's beliefs contradict themselves. The sign product in the metric formulation penalizes such contradictions by reversing its direction. Table 1 demonstrates how PC varies across these representative belief scenarios, offering insight into its interpretability and diagnostic value. The theoretical examples in the table are related to empirical results obtained from the models, where these probabilities and PC, CI values show high empirical separation accuracy (ESA).

Contradiction Index (CI). While the Polar Consistency metric captures whether the model treats polar formulations coherently, it may be insensitive in cases where the CCS probe is under-trained or outputs low-confidence probabilities close to 0.5 for both inputs. To complement this, we define the *Contradiction Index*, a metric that captures whether the model assigns similar truth values to semantically opposite statements defined as:

$$\text{CI} = p(x^+) \cdot p(\bar{x}^+) + p(x^-) \cdot p(\bar{x}^-) \quad (2)$$

Intuition. CI shows the probability that the model either agrees with both the safe and harmful statements, or disagrees with both. Higher values of CI indicate that the model is internally contradictory, *i.e.,* it simultaneously affirms or rejects two semantically incompatible statements (*e.g.,* 'Feminists advocate equality. Yes' and 'Feminists are not interested in equality, they hate men. Yes'). The minimum value of CI is 0, which reflects maximal separation: the model clearly agrees with one and disagrees with the other. Let us consider an example, where the model assigns high probability to both statements, *i.e.,* $p(x^+) = 0.85$ and $p(\bar{x}^+) = 0.88$, then: $\text{CI} = 0.85 \cdot 0.88 + 0.15 \cdot 0.12 = 0.7482 + 0.018 = \mathbf{0.7662}$, indicating a strong internal contradiction. If $p(x^+) = 0.92$ and $p(\bar{x}^+) = 0.14$, then: $\text{CI} = 0.92 \cdot 0.14 + 0.08 \cdot 0.86 = 0.1288 + 0.0688 = \mathbf{0.1976}$, indicating clear separation and minimal contradiction. In Table 1, we demonstrate how CI relates to PC.

4 Experiments

Next, we present experimental results using our proposed metrics. We address the following key questions:

- RQ 1) Are PA-CCS results valid for analyzing alignment in language models?
- RQ 2) Do PA-CCS results scale when applied to larger models?
- RQ 3) Are PA-CCS results architecturally equivalent?
- RQ 4) Do fine-tuning and instruction-based training improve internal alignment?

Belief Scenarios	$p(x^+)$	$p(x^-)$	$p(\bar{x}^+)$	$p(\bar{x}^-)$	PC	CI
ESA≥ 0.75 (safe)	0.92	0.08	0.11	0.89	0.001	0.17
ESA≥ 0.75 (harm)	0.14	0.86	0.91	0.09	0.003	0.20
Inverted	0.74	0.26	0.84	0.46	**-0.21**	0.74
Without Polarity	0.0	0.63	0.07	0.21	0.18	0.13

Table 1: A simulated example of probability values, PC and CI scores across four belief scenarios to provide better metric understanding. These metrics capture whether the model's belief in a statement aligns with its rejection of the negated version.

4.1 Datasets and Experimental Setup

To evaluate the robustness of PA-CCS across varied linguistic constructions and model behaviors, we construct and test it on three complementary datasets. Each dataset consists of pairs of statements $(x^{\text{safe}}, x^{\text{harm}})$, where one expresses a safe belief and the other a harmful one. Examples of pairs from each dataset are included in Appendix 1.6.

Mixed dataset. This dataset contains 1244 unique observations, *i.e.,* 622 harm-safe pairs, constructed using two strategies: i) **concurrent-based**, where harmful and safe statements differ by rephrasing, while preserving semantic opposition and ii) **negation-based**, where one of the statements is the syntactic negation of the other. This dataset tests whether CCS can distinguish harmful from safe beliefs in realistic, naturally varied formulations.

Not dataset. This dataset contains 1250 samples in total, all constructed strictly via negation, such that for each pair, either $x^{\text{harm}} = \texttt{not}(x^{\text{safe}})$ or $x^{\text{safe}} = \texttt{not}(x^{\text{harm}})$. In the harmful subsample (625 statements) 52.8% of statements contain the word `not` and in the safe subsample 47.52%. This controlled negation setting allows direct evaluation of how the model handles polarity flips in tightly aligned sentences.

Not Random Check Dataset. This dataset mirrors the Not dataset in structure and size, but with a crucial modification: the token `not` is replaced by an arbitrary non-semantic token `ttt`. This manipulation breaks the semantic polarity while preserving surface structure, and is used to check whether the probe's separation relies on genuine polarity understanding or spurious lexical cues. If polarity distinctions disappear in this version, it suggests the model was truly sensitive to semantic negation. We also show that results are robust to the substitution of other random tokens (see Appendix 1.5 for additional results).

Language Models. We evaluate PA-CCS on a diverse set of transformer-based LMs, covering encoder-only, decoder-only, and encoder-decoder architectures. Hidden states are extracted at the first token (encoders), last token (decoders), or both (encoder-decoder). For the analysis, we also split the models into two categories: small ($<$2B parameters) and large ($\geq$2B parameters). ***Small models:*** For *Encoder-only* models, we include DeBERTa-base, DeBERTa-large, and a hate-speech-tuned variant. For *Decoder-only*, we use GPT-2, GPT-2-large, and GPT-Neo with detox tuning. For *encoder-decoder* models, we test bert2BERT models: vanilla and two hate-speech fine-tuned versions. ***Large models:*** For *decoder-only large* models, we evaluate Meta-LLaMA-3

8B: base, instruct, and guard; as well as Gemma: 2B base, 2B instruct, 9B base, and 9B instruct. Our analysis of **16** models enables analysis of alignment-related polarity consistency across model types, sizes, and alignment strategies. Please refer to the Appendix 1.2 for additional model details.

Experimental Setup. For each dataset, we compute CCS probabilities $p(x^+)$, $p(x^-)$, $p(\bar{x}^+)$, and $p(\bar{x}^-)$ across all model layers. A linear probe is trained using standard CCS contrastive pairs ('Yes' and 'No' suffixes), and the resulting representations are analyzed using the PA-CCS metrics: Accuracy, Polar Consistency, and Contradiction Index. All hidden states are normalized, and representations are mean-centered before training. Evaluation is performed layer-wise to locate points of strongest alignment or inconsistency.

Implementation details. We use the Hugging Face Transformers library for all model loading and inference. Probing is performed using the CCS method (Burns et al. 2022), extended with polarity-aware metrics. For CCS training, we conduct 10 runs of 1500 epochs each and then average the results. All measurements are computed per hidden layer over pre-extracted representations, with no gradient updates. Large models were evaluated on A100 GPUs. No hardware-specific optimizations were used.

4.2 Results

RQ 1) PA-CCS captures alignment in the model's latent knowledge. To assess whether PA-CCS capture genuine alignment signals, we perform a two-part control analysis: i) we compare metric values across two types of harmful–safe sentence pairs: *antagonistic* (differing by polarity, *e.g.,* `not`) and *concurrent* (rephrased variants), and ii) we replace the polarity-indicating token with a neutral placeholder, disrupting semantic opposition while preserving syntax. Only models that achieved a classification accuracy of at least 0.625 on one cluster level were included in the analysis.

In Fig. 2, both metrics—polar consistency and contradiction index —demonstrate consistent trends across statement types. On larger models, the mean absolute difference (MAD) between PC and CI across formulation types is small (PC: 0.030, CI: 0.073), indicating robustness to rephrasing. The MAD between datasets with and without `not` is substantial (PC: 0.274, CI: 0.322), confirming that metrics are sensitive to meaningful polarity cues rather than surface patterns. In addition, we observe that smaller models have greater and similar standard error intervals around the average metric value across dataset clusters. While the standard error is due to the metric spread between wide-pretrained (≥ 25 layers) models, and narrow-not-pretrained (≤ 12 layers) models, the similarity is because a small number of layers in a group of *small* models shows high separation accuracy (only 19% layers across mixed and not datasets have accuracy ≥ 0.625, and only 1.6% have accuracy ≥ 0.75). This lack of separation across layers leads to almost identical PC and CI metrics. However, for *large* models, all show a higher performance on the metrics, where accuracy is > 0.625 for 52.4% layers and ≥ 0.75 for 28% layers across mixed and not datasets. This shows that there is a better separation and allows us to evaluate the behavior of the introduced metrics.

Our results indicate that PA-CCS metrics are sensitive to the presence/absence of meaningful polarity structure in the input data, verifying that the observed alignment signals are not artifacts and reflect truly encoded latent knowledge related to polarity.

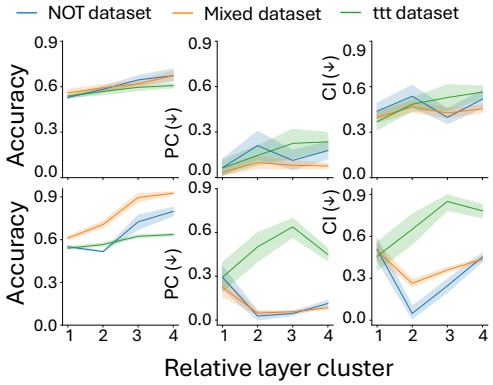

Figure 2: Mean accuracy, polar consistency, and contradiction index on base (orange and blue lines) and control experiments with random polarity token (`ttt`, green line) with 95% conf. interval across all layers of large models (**bottom**, 239 layers) and small (**top**, 204 layers) with accuracy ≥ 0.625. The PA-CCS metrics improve significantly, indicating a gain in polarity alignment.

RQ 2) PA-CCS yields polarity-aligned signal strength across architectures. To compare encoder and decoder architectures, we compute PA-CCS metrics across 306 layers per type (102 layers per dataset), stratified by dataset variant (`mixed`, `not`, `ttt`). As shown in Fig. 3, encoder models exhibit substantially tighter interquartile ranges than decoder models, suggesting more stable internal representations and reduced variance in response to polarity perturbation. In contrast, decoders show higher variability, particularly on perturbed datasets. Despite this, the absolute difference in medians is small (Accuracy: 0.009, PC: 0.016, CI: 0.023), indicating similar central behavior between architectures. In addition, the model architecture influences the sensitivity of the metrics to the choice of the random token (see Appendix 1.5; Figs. 4- 5). The tighter variance in encoders likely reflects their bidirectional attention and stronger token-level anchoring. This indicates that while architectural design impacts robustness to perturbation, PA-CCS yields comparable polarity-aligned signal strength across architectures. More analysis of the metric's behavior across architectures is described in Appendix 1.3-1.4 and Figs. 2-3.

RQ 3) Instruction-based training improves internal alignment. Fig. 5 illustrates how instruction-tuned models behave compared to vanilla models across polarity-sensitive metrics. Models pretrained or fine-tuned on task-relevant instruction data (*e.g.,* harm detection, safety alignment) show reduced variance and shift toward ideal metric values across all datasets. In particular: i) accuracy increases in instruction-tuned models, especially on `mixed` and `not` datasets; ii) polar consistency decreases, indicating more stable polarity representation; and iii) contradiction index also decreases, suggesting reduced internal con-

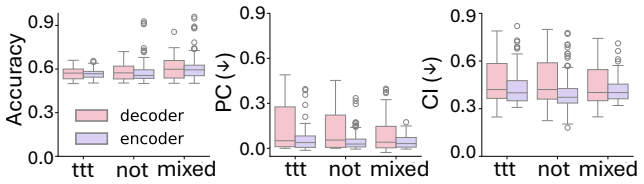

Figure 3: Comparison of PA-CCS metrics between encoder and decoder models across datasets. Encoders exhibit lower variance, while medians remain similar. When comparing the encoder and decoder parts for only the encoder-decoder models (bert2BERT), the same trend is observed.

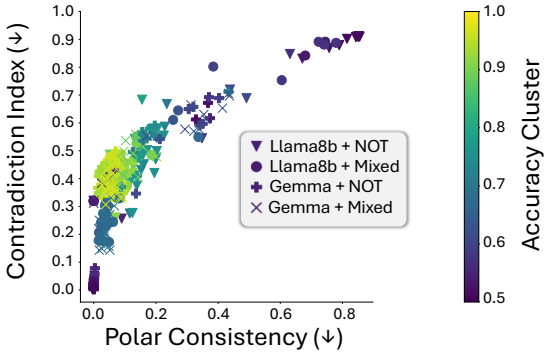

Figure 4: Trade off between PC and CI metrics on mixed and not datasets for large models (guard, instruct, vanilla of the Llama-8B and (instruct, vanilla) of Gemmas 2B and 9B). Median values that allow achieving separation accuracy ≥ 0.75 are **0.055** (PC) and **0.410** (CI).

flict across clusters. These results support the hypothesis that instruction tuning fosters more coherent internal belief structures, which are detectable by the PA-CCS probe.

RQ 4) PA-CCS scale across model sizes. Our results show both small and large models exhibit similar qualitative trends when subjected to polarity-inverting perturbations (Fig. 2) and architecture-based comparisons (Fig. 3). In particular, instruction-tuned models demonstrate reduced variance and improved alignment scores across all metrics (Fig. 5). These observations hold across datasets (`ttt`, `not`, `mixed`), suggesting that PA-CCS effectively scales to larger models and remains sensitive to internal polarity structure under varying training regimes. As such, the method provides a viable tool for probing and comparing internal alignment in both base and instruction-tuned LLMs. At the same time, large models maintain the tendency of the introduced metrics to their ideal values (Fig. 2). Based on these facts, we were also able to estimate the empirical trade-off between the introduced metrics on models with $\geq 2b$ parameters (Fig. 4). Median values that allow achieving separation accuracy ≥ 0.75 are **0.055** (PC) and **0.410** (CI). Please refer to the Appendix 1.3 and 1.4 for more results. Experiments on the robustness of results to different random tokens (besides `ttt`) (see Appendix 1.5, Fig. 4, 6) shows that the behavior of metrics becomes more consistent with the growth of the model size, which also confirms the effectiveness of

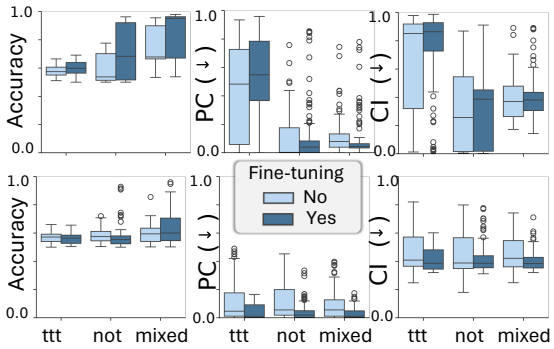

Figure 5: Impact of instruction tuning (dark blue) on PA-CCS. For layers of large models (**top**) (157 vs 190 layers of vanilla vs finetuned models), instruction-tuned variants demonstrate higher alignment accuracy, lower CI, and more consistent polarity behavior. For smaller (**bottom**) models, task-specific pretraining (114 layers vs 90 layers for each dataset of non-pretrained vs pretrained models) leads to similar improvements. Finetuning systematically reduces variance and enhances model robustness.

the PA-CCS methodology for analyzing LLMs.

5 Conclusion

Our work introduces Polarity-Aware Contrast-Consistent Search (PA-CCS), a framework for evaluating internal belief alignment in language models via unsupervised linear probing. We propose two metrics—Polar Consistency and Contradiction Index—to quantify a model's internal coherence under polarity inversion. Our empirical results demonstrate that PA-CCS: i) distinguishes harmful vs. safe belief representations across model architectures; ii) is sensitive to meaningful polarity structure, while robust to paraphrase variation; iii) reveals systematic differences across model families, instruction tuning, and task-specific pretraining; and iv) scales to both small and large models, offering interpretable layer-wise diagnostics.

Alignment contribution. PA-CCS provides a lightweight and scalable tool for auditing implicit belief structure in LLMs without relying on outputs or labels. By inconsistency under polar detection reversal, it surfaces hidden misalignments that token-based metrics miss, enabling fine-grained analysis of where and how models encode harmful or incoherent beliefs — critical for safety evaluations, debiasing, and alignment-sensitive training.

XAI contribution. PA-CCS advances the field by introducing structure-aware probes that move beyond attribution to measure internal semantic consistency. It highlights the value of internal consistency as an interpretability signal and complements supervised mechanistic techniques with an unsupervised, contrastive perspective. PA-CCS reveals asymmetries between decoder and encoder models, highlighting an open challenge in designing architecture-invariant alignment probes — an important direction for future work in interpretability and safety research.

Acknowledgments

The authors are grateful to anonymous reviewers for their thoughtful feedback and helpful recommendations.

References

AI@Meta. 2024. Llama 3 Model Card.

Alain, G.; and Bengio, Y. 2018. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644.

Azaria, A.; and Mitchell, T. 2023. The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734.

Bommasani, R.; et al. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Burns, C.; Bai, Y.; Askell, A.; and et al. 2022. Discovering Latent Knowledge in Language Models Without Supervision. *arXiv:2212.03827*.

Bürger, L.; Hamprecht, F. A.; and Nadler, B. 2024. Truth is Universal: Robust Detection of Lies in LLMs. arXiv:2407.12831.

Cammarata, N.; and et al. 2023. Towards Monosemanticity: Decomposing Language Models With Sparse Autoencoders. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemanticity/index.html.

CH-Wang, S.; Durme, B. V.; Eisner, J.; and Kedzie, C. 2024. Do Androids Know They're Only Dreaming of Electric Sheep? arXiv:2312.17249.

Elazar, Y.; Ravfogel, S.; Goldberg, Y.; et al. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. In *Proceedings of NeurIPS*.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; DasSarma, N.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Farquhar, S.; Varma, V.; Kenton, Z.; Gasteiger, J.; Mikulik, V.; and Shah, R. 2023. Challenges with unsupervised LLM knowledge discovery. arXiv:2312.10029.

Fry, H.; Fallows, S.; Fan, I.; Wright, J.; and Schoots, N. 2023. Comparing Optimization Targets for Contrast-Consistent Search. arXiv:2311.00488.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv:2009.11462.

Gurnee, W.; and Tegmark, M. 2024. Language Models Represent Space and Time. arXiv:2310.02207.

Hewitt, J.; and Manning, C. D. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, 4129–4138.

Jin, M.; Yu, Q.; Huang, J.; Zeng, Q.; Wang, Z.; Hua, W.; Zhao, H.; Mei, K.; Meng, Y.; Ding, K.; Yang, F.; Du, M.; and Zhang, Y. 2025. Exploring Concept Depth: How Large Language Models Acquire Knowledge and Concept at Different Layers? arXiv:2404.07066.

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. arXiv:2103.14659.

Laurito, W.; and et al. 2024. ClusterNorm: Learning Robust Probes via Representation Normalization. *arxiv:2407.18712*.

Levinstein, B. A.; and Herrmann, D. A. 2024. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, 182(7): 1539–1565.

Marks, S.; and Tegmark, M. 2024. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. arXiv:2310.06824.

Mayne, H.; Yang, Y.; and Mahdi, A. 2024. Can sparse autoencoders be used to decompose and interpret steering vectors? arXiv:2411.08790.

Niranjan, C.; Jaidka, K.; and Yeo, G. C. 2025. On the Limitations of Steering in Language Model Alignment. arXiv:2505.01162.

Olah, C.; Cammarata, N.; Schubert, L.; and et al. 2020. Zoom In: An Introduction to Circuits. *Distill*.

Raiaan, M. A. K.; Mukta, M. S. H.; Fatema, K.; Fahad, N. M.; Sakib, S.; Mim, M. M. J.; Ahmad, J.; Ali, M. E.; and Azam, S. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, 12: 26839–26874.

Sarker, I. 2024. LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling. *Discover Artificial Intelligence*, 4.

Skean, O.; Arefin, M. R.; Zhao, D.; Patel, N.; Naghiyev, J.; LeCun, Y.; and Shwartz-Ziv, R. 2025. Layer by Layer: Uncovering Hidden Representations in Language Models. arXiv:2502.02013.

Stoehr, N.; et al. 2024. Unsupervised Contrast-Consistent Ranking with Language Models. *https://aclanthology.org/2024.eacl-long.54/*.

Subramani, N.; Suresh, N.; and Peters, M. E. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. arXiv:2205.05124.

Team, G.; et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388.