Measuring and Mitigating Shortcut Reliance in Language Models with Probe-Based Representation Entanglement

Divyajot Singh
Birla Institute of Technology and Science, Pilani, India

Abstract

Shortcut learning occurs when models solve tasks by exploiting superficial correlations in training data like formatting patterns or lexical heuristics instead of learning the intended underlying task. These shortcuts often produce high in-distribution accuracy but fail under distribution shift. We study whether shortcut reliance can be diagnosed and mitigated in small instruction-tuned language models using a simple representation-level quantity. We fine-tune Gemma 3 1B Instruct and Llama 3.2 1B on two synthetic sentiment shortcuts in SST-2 and one natural shortcut in MNLI based on lexical overlap. During training, we fit linear probes for the task label and the shortcut attribute at every layer and define Correlation-Dependent Representational Entanglement (CDRE) as the absolute cosine similarity between the two probe directions. Across settings, increasing shortcut prevalence produces a sharp rise in the robustness gap between shortcut-aligned and shortcut-free test sets, and higher deep-layer CDRE tracks this degradation. At a 99% shortcut ratio, Llama's clean accuracy on capitalization-biased SST-2 drops from 93.2% at 0% bias to 49.0%, while Gemma drops from 91.8% to 60.2%. A CDRE-regularized objective substantially improves robustness for capitalization and lexical-overlap shortcuts, but offers little benefit for a speaker-prefix shortcut whose learned directions are already nearly orthogonal. These results show that probe-derived representation entanglement provides a reliable signal of harmful shortcut reliance and offers a practical criterion for determining when shortcut mitigation is likely to be effective.

1 Introduction

Shortcut learning is one of the central obstacles to reliable natural language processing. Across benchmark settings, models can obtain strong headline accuracy by exploiting superficial cues, annotation artifacts, or dataset-specific heuristics rather than learning the intended task (Gururangan et al., 2018; Poliak et al., 2018; Geirhos et al., 2020; McCoy et al., 2019). This pattern has been documented especially clearly in natural language inference, where hypothesis-only baselines, lexical overlap heuristics, and syntactic shortcuts all allow systems to be right for the wrong reasons (Bowman et al., 2015; Williams et al., 2018; Poliak et al., 2018; Gururangan et al., 2018; McCoy et al., 2019). Closely related concerns have also appeared in adversarial evaluation, contrast sets, and counterfactual testing, all of which show that standard test accuracy can substantially overestimate genuine task understanding (Nie et al., 2020; Gardner et al., 2020; Kaushik et al., 2020; Ettinger, 2020).

A growing literature has proposed ways to mitigate these failures. Existing approaches include ensemble-based debiasing, adversarial methods, data augmentation, invariant or worst-group optimization, and error-focused retraining procedures such as Just Train Twice (JTT) (Clark et al., 2019; Elazar and Goldberg, 2018; Schuster et al., 2019; Utama et al., 2020; Zhou and Bansal, 2020; Liu et al., 2020; Kaushik et al., 2020; Sagawa et al., 2020; Liu et al., 2021). These methods have produced important gains, but they also raise a practical question: when should we expect shortcut mitigation to matter? In other words, can we diagnose whether a shortcut has become meaningfully entangled with the task before committing to an intervention? In this paper, we intentionally study both controlled and natural shortcut settings. Two experiments use synthetically injected shortcuts in SST-2 (capitalization and speaker prefixes), allowing precise control over shortcut prevalence. A third experiment uses natural lexical-overlap bias in MNLI, which reflects a known real-world NLI heuristic. This combination lets us test whether our findings hold in both controlled and more realistic settings.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 648–662

This raises a practical question: when should we expect shortcut mitigation to be necessary? Rather than applying mitigation strategies uniformly, can we identify when shortcut information has become meaningfully entangled with the task signal and is likely to harm generalization? A useful diagnostic in this setting should not only reveal the presence of shortcut information, but indicate when it is actually influencing the model's decision boundary.

We investigate such a diagnostic through the geometry of linear probes. Probing has long been used to study what information is encoded in hidden states, but its interpretation requires care (Alain and Bengio, 2017; Hewitt and Liang, 2019; Voita and Titov, 2020; Belinkov, 2022). Rather than treating probes only as descriptive tools, we use them to measure the alignment between a task-relevant concept direction and a shortcut direction inside the model's representation space. We call this quantity Correlation-Dependent Representational Entanglement (CDRE) and define it as the absolute cosine similarity between the concept and shortcut probe coefficients at a given layer. Intuitively, high CDRE indicates that shortcut and task information are not merely co-present in the representation, but aligned in a way that makes the shortcut likely to influence the model's decision boundary. In contrast, low CDRE corresponds to a regime where shortcut information may be recoverable but remains geometrically separated from the task signal, and is therefore less likely to drive predictions.

We evaluate this idea on two small instruction-tuned language models, `meta-llama/Llama-3.2-1B` and `google/gemma-3-1b-it`, across three shortcut settings. Two are synthetic SST-2 transformations designed to isolate the effects of formatting shortcuts: capitalization and speaker prefixes. The third is a natural lexical-overlap shortcut on MNLI. We sweep shortcut ratios, probe the models throughout training, and compare baseline fine-tuning against CDRE-regularized training and JTT at high bias.

Our findings are consistent across these settings. First, increasing shortcut prevalence widens the robustness gap between shortcut-aligned and shortcut-free evaluation. Second, deep-layer CDRE rises with that gap in the capitalization and lexical-overlap settings. Third, CDRE regularization improves out-of-distribution accuracy precisely where the probe geometry indicates strong shortcut–concept entanglement, while offering little benefit when the shortcut remains geometrically separate from the concept. Taken together, these results suggest that probe-derived representation alignment can serve not only as an interpretability signal but also as a practical guide for deciding when shortcut mitigation is likely to pay off.

2 Related Work

Our work connects four research threads.

Shortcut learning and dataset artifacts. A broad literature shows that modern models often exploit shallow cues or annotation artifacts rather than task semantics (Geirhos et al., 2020; Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). In NLI specifically, benchmark construction has enabled models to lean on lexical and structural heuristics instead of inference (Bowman et al., 2015; Williams et al., 2018; Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). Robust evaluation efforts such as ANLI and contrast sets were developed partly in response to these failures (Nie et al., 2020; Gardner et al., 2020).

Mitigating spurious correlations. Prior work has explored a range of debiasing strategies, including ensemble methods, adversarial objectives, data augmentation, and robust optimization (Clark et al., 2019; Elazar and Goldberg, 2018; Schuster et al., 2019; Utama et al., 2020; Zhou and Bansal, 2020; Liu et al., 2020; Kaushik et al., 2020; Sagawa et al., 2020). JTT is particularly relevant to our experiments because it offers a simple error-driven alternative that does not require group labels during training (Liu et al., 2021). Our comparison against JTT tests whether a probe-informed representation penalty offers a different advantage in highly biased regimes.

Probing and representation analysis. Probes are often used to ask what information is linearly recoverable from intermediate representations (Alain and Bengio, 2017; Hewitt and Liang, 2019; Voita and Titov, 2020; Belinkov, 2022). However, recoverability alone does not say whether the information is behaviorally important. Our work uses probe geometry differently: we treat the relative alignment between concept and shortcut probes as a candidate indicator of harmful entanglement.

Small language models and fine-tuning. The recent success of pretrained contextual encoders and language models has made fine-tuning the default

paradigm for NLP (Peters et al., 2018; Devlin et al., 2019). Yet the same flexibility that makes fine-tuning effective also allows small models to quickly absorb dataset shortcuts. Our work focuses on this practically important regime and asks how much robustness analysis can be done with lightweight tooling and modest-scale models.

3 Methodology

3.1 Models and Tasks

We use two small instruction-tuned language models: `meta-llama/Llama-3.2-1B` and `google/gemma-3-1b-it`. Both are fine-tuned as sequence classifiers with two output labels.

Synthetic SST-2 shortcuts. We fine-tune sentiment classifiers on SST-2 with two synthetic shortcut types:

- **Capitalization**: negative examples are transformed to uppercase and positive examples to lowercase.
- **Speaker prefix**: positive examples receive the prefix `Speaker 1:` and negative examples receive `Speaker 2:`.

For a shortcut ratio r, a fraction r of training examples receives the shortcut transformation and the rest remains clean. These two synthetic shortcuts were chosen to contrast a highly salient formatting cue (capitalization) with a cue that is explicit and linearly decodable but semantically disconnected from sentiment (speaker prefix).

MNLI lexical overlap shortcut. We convert MNLI into a binary task: entailment versus non-entailment. For each example, we compute overlap as the proportion of hypothesis words that also appear in the premise. Using the median training overlap as a threshold, we mark examples as high-overlap or low-overlap. A shortcut-consistent example is one where the overlap indicator matches the binary label, while a shortcut-inconsistent example does not. We then create biased training sets by sampling a target fraction of shortcut-consistent examples. This setting provides a more realistic shortcut than the synthetic SST-2 transformations because lexical overlap is a known heuristic in NLI rather than a manually injected token pattern.

3.2 Train, Probe, and Test Splits

The probing experiments on SST-2 use 2,000 training examples, 500 evaluation examples, and a 100-example balanced probe set with a 50/50 shortcut mix. The MNLI probing experiments use 10,000 training examples, 2,000 evaluation examples, and a 500-example probe set. Mitigation comparisons use 10,000 training examples and 2,000 evaluation examples.

For each setting, we report three evaluation views:

- **ID accuracy**: performance on a shortcut-aligned test set with 100% shortcut prevalence.
- **OOD accuracy**: performance on a shortcut-free SST-2 set or a balanced MNLI set with no overlap-label correlation.
- **Robustness gap**: ID accuracy minus OOD accuracy.

This design cleanly separates two behaviors that are often conflated during biased training: memorizing the shortcut so as to perform well when it remains available, and learning the underlying task in a way that survives when the shortcut disappears.

3.3 Correlation-Dependent Representational Entanglement

At every 10% of training, we extract mean-pooled hidden states from every transformer layer and fit two linear logistic-regression probes: a concept probe for the task label and a shortcut probe for the shortcut attribute. Probe training uses a fixed train/test split over probe examples so that layer-wise comparisons reflect representation changes rather than split noise.

Let w_c and w_s denote the coefficient vectors of the concept and shortcut probes at a given layer. We define

$$\mathrm{CDRE} = |\cos(w_c, w_s)| = \left| \frac{w_c^\top w_s}{\|w_c\|\|w_s\| + \epsilon} \right|.$$

Low CDRE means the two linear directions are nearly orthogonal, while high CDRE indicates that the representation subspace useful for the task is aligned with the shortcut.

This definition intentionally differs from simply measuring shortcut probe accuracy. A model may encode shortcut information strongly while still keeping it geometrically separate from the concept direction. In that case, the shortcut is recoverable but need not dominate the decision boundary. Our core hypothesis is that robustness degrades most when the two directions become aligned in deeper layers.

We use cosine similarity between probe weight vectors for two pragmatic reasons. First, it is scale-invariant: because probe norms can vary across layers and checkpoints, cosine focuses on directional overlap rather than coefficient magnitude. Second, it matches the intervention we later apply, which penalizes alignment between concept and shortcut projections in the last-layer representation. TWe do not claim cosine alignment is the only possible measure of concept-shortcut interaction. Alternatives such as mutual information, CKA, or direct decision-boundary comparisons may also be informative. However, our experiments already provide one important baseline comparison: shortcut probe accuracy alone is insufficient. In the speaker-prefix setting, shortcut information remains highly decodable from hidden states, yet CDRE remains low, robustness degradation is comparatively mild, and CDRE regularization provides little benefit. This suggests that recoverability alone cannot explain shortcut failures; directional alignment captures behaviorally relevant structure that probe accuracy misses. A broader comparison against alternative representation metrics remains future work.

3.4 CDRE-Regularized Training

We compare three training procedures at the 99% shortcut ratio.

Baseline fine-tuning. Standard cross-entropy fine-tuning on the biased training set.

CDRE regularization. Across all mitigation experiments, we use the same high-level mechanism: a probe-informed penalty on the pooled last-layer hidden states that discourages alignment between concept and shortcut directions cached from auxiliary probes. Formally, we optimize

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \cdot \mathcal{L}_{\text{CDRE}},$$

with $\lambda = 0.1$. In the implementation used here, the probe directions are refreshed every 10 optimization steps after step 50 using a balanced probe pool. Operationally, the penalty is computed on the last-layer pooled states and uses probe-derived directions rather than gold group labels at training time. In both SST-2 and MNLI, the instantiated penalty is a correlation-style objective over concept and shortcut projections, so the regularization story is unified across tasks even though the shortcut features themselves differ.

JTT. We follow a two-stage JTT procedure (Liu et al., 2021): train an initial model, identify training errors, then upweight those examples by a factor of five in a second training stage. In our setup this provides an informative baseline because heavily biased training sets can cause the first-stage model to make very few mistakes, potentially starving JTT of the error signal it relies on.

3.5 Training Setup

All runs use a maximum sequence length of 128, batch size 32 with gradient accumulation 2, learning rate 2×10^{-5}, and three epochs of fine-tuning unless otherwise noted. JTT uses two epochs in the first phase and three epochs in the second phase. Random seeds are fixed to 42 throughout the experiments.

4 Experiments

4.1 Research Questions

We ask three questions:

1. Does increasing shortcut prevalence reliably increase the robustness gap?

2. Does CDRE track this degradation across layers and tasks?

3. Can a CDRE-guided regularizer improve OOD robustness relative to baseline and JTT?

4.2 Experimental Logic

The experiments are organized as a progression. First, we sweep shortcut ratios to establish whether robustness collapses smoothly as bias increases. Second, we probe hidden states throughout training to test whether CDRE rises alongside that collapse. Third, we run mitigation experiments at the hardest setting (99% shortcut prevalence) to determine whether a representation-level intervention helps when entanglement is strongest. This staged design lets us use the ratio sweep as diagnosis and the 99% setting as a stress test for intervention.

5 Results

5.1 Robustness Collapse Under Stronger Shortcuts

Figure 1 shows that as the shortcut ratio increases, the gap between shortcut-aligned and shortcut-free performance widens substantially. The strongest collapse appears for capitalization. On Llama,

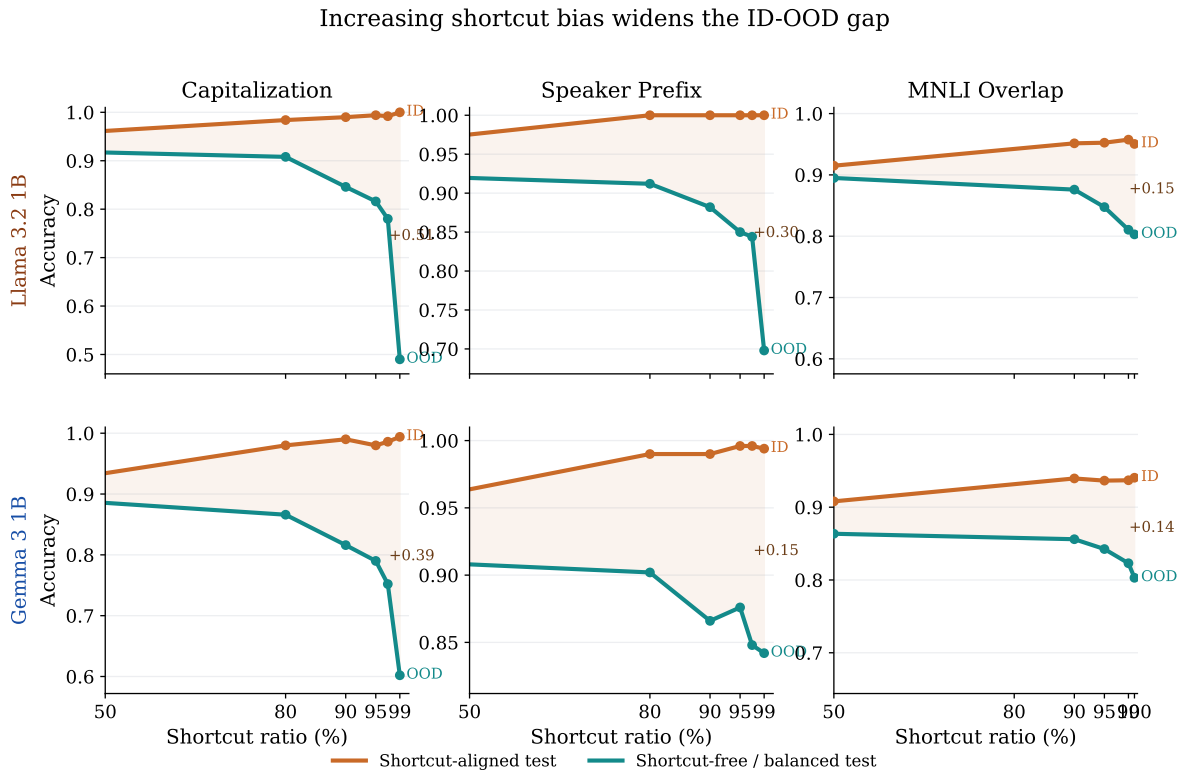

Figure 1: Main-paper ratio sweeps. Each panel shows how shortcut-aligned and shortcut-free accuracy change as shortcut bias increases. The widening distance between the two curves is the central robustness failure studied in this paper.

clean accuracy falls from 93.2% with no capitalization bias to 49.0% at 99% bias; on Gemma it falls from 91.8% to 60.2%. The speaker shortcut is weaker for Gemma, whose clean accuracy remains 84.2% even at 99% bias. On MNLI, the OOD degradation is milder but the robustness gap still grows sharply: Llama moves from a negative gap of 17.1 points at 0% bias to a positive gap of 14.7 points at 99% bias as ID accuracy rises to 95.8% and clean accuracy settles at 81.1%.

Two patterns stand out in the ratio sweeps. First, severe bias is not required to produce harmful behavior. In capitalization-biased SST-2, the Llama model already shows a 14.4-point gap at 90% bias and a 17.8-point gap at 95% bias. Gemma shows a similar progression, reaching gaps of 17.4 and 19.0 points at the same ratios. Second, the raw size of the gap depends strongly on the shortcut type. The speaker-prefix setting still hurts robustness, but the effect is smaller for Gemma and appears to saturate earlier.

The lexical-overlap results help contextualize these synthetic findings. For both models, the 0% bias setting yields negative robustness gaps because the balanced test set is easier than the fully shortcut-aligned set. As the training data becomes more overlap-biased, this relationship reverses: ID performance climbs while balanced performance drops modestly, producing a steadily widening gap. This makes lexical overlap a useful middle ground between trivial synthetic cues and naturally occurring dataset artifacts.

5.2 Lexical Overlap Extends the Story Beyond Synthetic Shortcuts

The MNLI experiments deserve dedicated attention because they show that the same pattern holds for a natural heuristic rather than a manually inserted formatting cue. Figure 2 expands the lexical-overlap results in isolation. Both models become stronger on the high-overlap ID set as the overlap bias rises, while balanced OOD performance erodes or stays flat. At 99% bias, Llama reaches 95.75% on the ID split but only 81.05% on the balanced split; Gemma reaches 93.7% and 82.3%, respectively.

This setting strengthens the paper's central claim in two ways. First, it shows that the method is not restricted to obviously artificial shortcuts such as capitalization or speaker tags. Second, it shows that CDRE remains useful even when the shortcut is

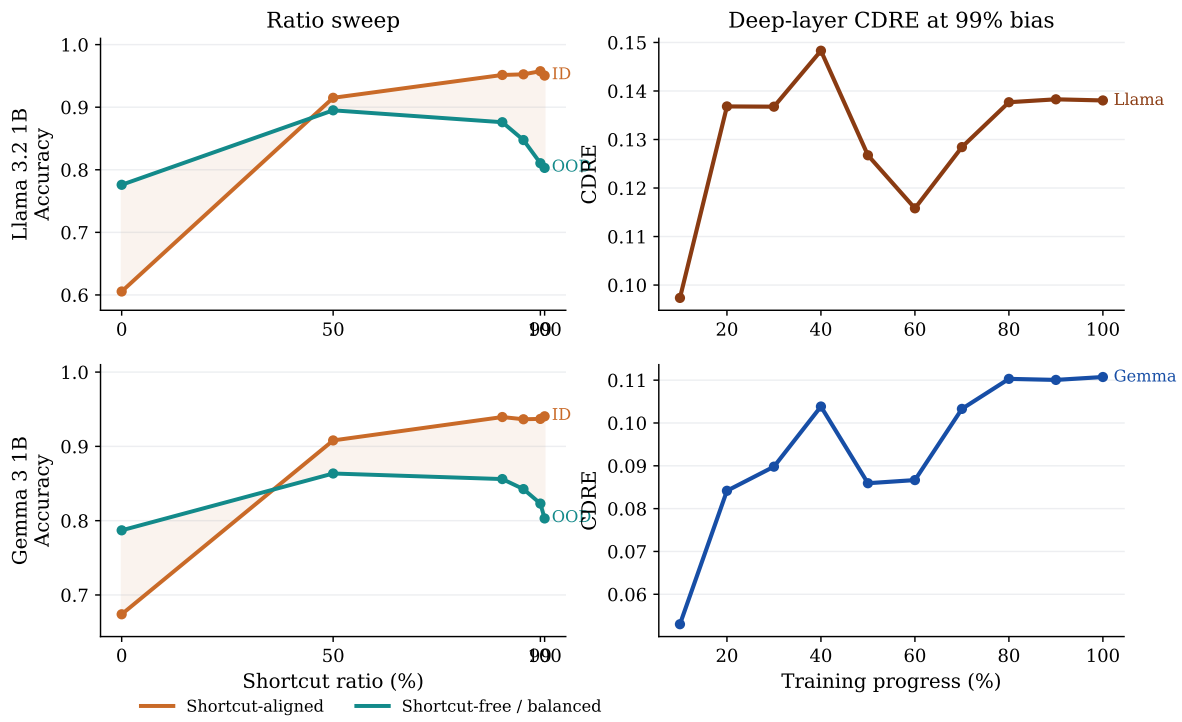

Figure 2: Main-paper lexical-overlap figure. For each model, the left panel shows the full ratio sweep and the right panel shows deep-layer CDRE during 99% biased training. The MNLI setting reproduces the same qualitative story as the synthetic experiments, but on a more natural shortcut.

task-adjacent rather than arbitrary: lexical overlap can sometimes help, but under strong dataset bias it becomes over-relied on.

5.3 Entanglement Emerges in the Failure Case

Figure 3 focuses on two representative 99% bias settings: capitalization, where robustness fails badly, and speaker-prefix, which acts as a negative control. In the capitalization setting, deep-layer CDRE rises early and stays high for both models. In the speaker-prefix setting, the trajectories remain much lower, especially for Gemma.

This contrast highlights a key distinction: the presence of shortcut information in the representation is not sufficient to produce robustness failures. In the speaker-prefix setting, the shortcut is highly decodable, yet CDRE remains low and robustness degradation is limited. This shows that what matters is not whether shortcut information exists, but whether it becomes aligned with the task direction used for prediction.

5.4 CDRE Regularization Improves OOD Robustness When Entanglement Is High

For capitalization, CDRE regularization improves clean accuracy from 69.6% to 79.6% on Llama and from 61.2% to 71.2% on Gemma, while reducing final CDRE from 0.2835 to 0.0738 and from 0.2253 to 0.0542. On MNLI, it improves clean accuracy from 80.2% to 90.2% on Llama and from 81.3% to 91.3% on Gemma, cutting the robustness gap from 15.0 to 4.8 points and from 13.0 to 2.6 points.

The speaker shortcut tells a different story. Baseline final CDRE is already near zero (0.0072 on Llama and 0.0085 on Gemma), and CDRE regularization yields no meaningful robustness improvement. This negative result is important: CDRE regularization is not universally helpful, but it is most useful when probe geometry indicates real entanglement.

These improvements are sizable in relative as well as absolute terms. On MNLI, CDRE regularization removes roughly two-thirds to four-fifths of the robustness gap. On capitalization, the absolute improvement is 10 points for both models, but the residual gap remains larger than in MNLI,

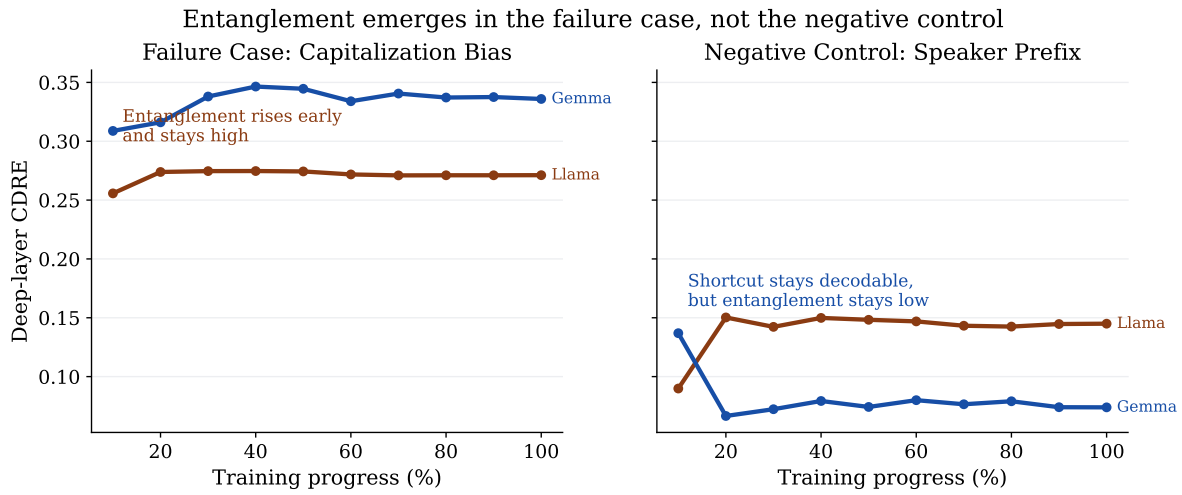

Figure 3: Main-paper mechanism figure. Capitalization bias produces sustained high deep-layer CDRE, while the speaker-prefix negative control stays substantially lower. The failure case is the one where shortcut–concept entanglement truly develops during fine-tuning.

Task	Model	Baseline OOD	CDRE-Reg OOD	JTT OOD	Baseline Gap	CDRE-Reg Gap	JTT Gap
Capitalization	Llama 3.2 1B	69.6	79.6	67.9	30.2	20.3	32.0
Capitalization	Gemma 3 1B	61.2	71.2	62.7	37.4	27.3	36.1
Speaker	Llama 3.2 1B	86.0	85.8	88.1	14.0	14.2	11.9
Speaker	Gemma 3 1B	88.2	88.0	88.9	11.5	11.6	9.1
MNLI overlap	Llama 3.2 1B	80.2	90.2	81.6	15.0	4.8	14.1
MNLI overlap	Gemma 3 1B	81.3	91.3	81.6	13.0	2.6	11.6

Table 1: Main-paper mitigation comparison at 99% shortcut prevalence. All values are percentages. CDRE regularization helps most in the capitalization and MNLI settings, while the speaker-prefix setting shows little change.

suggesting that the formatting shortcut is especially invasive once it dominates the training set.

Table 1 provides the exact mitigation comparison. It makes the same pattern clear in a more economical form: CDRE regularization gives the strongest and most consistent improvements in the capitalization and MNLI settings, while the speaker-prefix setting changes very little.

The speaker-prefix negative control remains important supporting evidence because it shows that high shortcut decodability alone is not enough to predict failure. We move the fuller per-experiment discussion of that case to Appendix to keep the main narrative focused on the central three findings.

6 Discussion

The main contribution of this work is to show that a simple, probe-based geometric measure can organize and predict when shortcut reliance becomes harmful in practice. CDRE captures a consistent relationship between representation structure, robustness degradation, and the effectiveness of mitigation across both synthetic and natural shortcut settings. Despite its simplicity, it provides an interpretable and actionable signal for understanding when shortcut information begins to influence model predictions.

This is encouraging for small-scale interpretability and robustness research. CDRE requires only hidden-state extraction and linear probes, making it feasible to run during ordinary fine-tuning. It also produces an actionable prediction: if shortcut information is recoverable but geometrically separated from the concept direction, mitigation may not be necessary or may need a different target.

More broadly, the results suggest a useful division of labor between diagnosis and intervention. Ratio sweeps and probes can identify which shortcuts genuinely distort the learned task representation. Only then does it make sense to deploy an additional training penalty. This is particularly valuable in student-scale work, where experimentation budgets are limited and expensive mitigation

strategies cannot be tried indiscriminately.

While our results are primarily empirical, they consistently show that rising CDRE accompanies—and in several cases precedes—robustness degradation. This suggests that CDRE can serve as a practical training-time signal for identifying when shortcut reliance is becoming harmful, even before failures are fully reflected in evaluation metrics. Evaluating fully adaptive intervention strategies based on this signal is a natural next step.

7 Conclusion

We studied shortcut learning in small instruction-tuned language models through the lens of probe geometry. CDRE, defined as the cosine alignment between concept and shortcut probe directions, tracks robustness degradation across synthetic and natural shortcuts. When CDRE is high, a simple CDRE-regularized objective materially improves OOD accuracy; when CDRE is already near zero, the same intervention brings little gain. These results suggest that representation-level entanglement is a useful diagnostic for deciding when shortcut mitigation is worth applying.

A practical takeaway is that representation-level alignment can do more than explain model behavior after the fact—it can help decide when robustness interventions are necessary. In our experiments, CDRE provides a lightweight and interpretable signal that indicates when shortcut reliance is likely to harm generalization and when mitigation is likely to be effective.

Limitations

Our experiments are intentionally scoped to a compact setting that allows us to study concept–shortcut entanglement in a controlled way. The representation analysis relies on linear probes, which offer interpretability and computational simplicity, though they do not capture all possible nonlinear structure in model representations. Moreover, although the training-dynamics results indicate that CDRE tracks the emergence of harmful shortcut reliance, we leave the evaluation of fully online, adaptive intervention strategies to future work. Finally, while the observed patterns are consistent across the settings we study, extending the analysis to additional models, tasks, and more naturalistic shortcut phenomena would help establish the broader applicability of the approach.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4069–4082.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Robert Geirhos, J"orn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. *Proceedings of the 38th International Conference on Machine Learning*.

Tianyu Liu, Zheng Xin, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020. An empirical study on model-agnostic debiasing strategies for robust natural language inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 596–608.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3419–3425.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7597–7610.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 183–196.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771.

A Capitalization Experiments

Figure 4 gives the per-experiment record for capitalization bias. The left column shows the full ratio sweeps for each model, while the right column shows deep-layer CDRE over training for the 99% bias condition. The main takeaway is that capitalization is both a strong shortcut and a strongly entangling one: OOD performance drops rapidly as bias increases, and deep-layer CDRE stays high throughout late training.

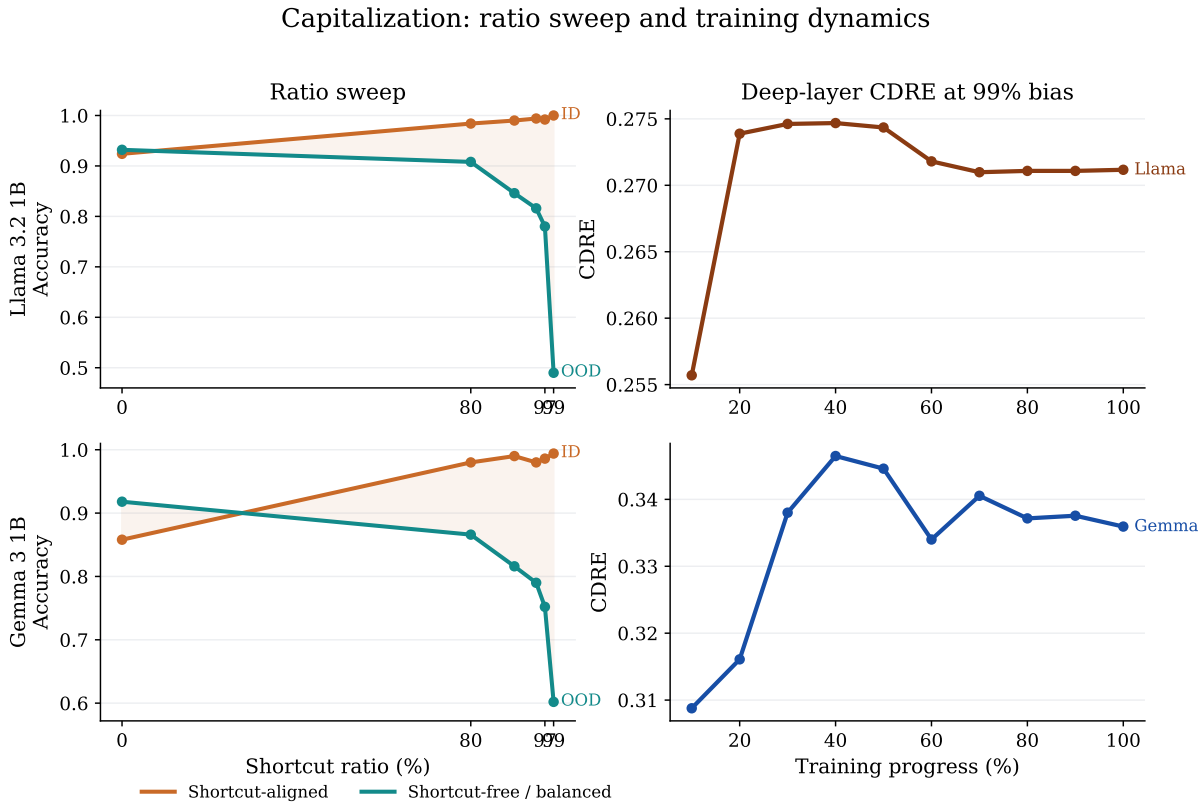

Figure 4: Appendix figure for capitalization. For each model, the left panel shows the full ID/OOD ratio sweep and the right panel shows deep-layer CDRE during 99% biased training.

B Speaker-Prefix Experiments

Figure 5 gives the corresponding per-experiment record for speaker-prefix bias. The ratio sweeps still show an OOD gap, but the training-dynamics panels explain why this case behaves differently from capitalization: the shortcut remains available without producing the same degree of concept-shortcut entanglement. This is the appendix-level evidence behind the paper's negative-control argument.

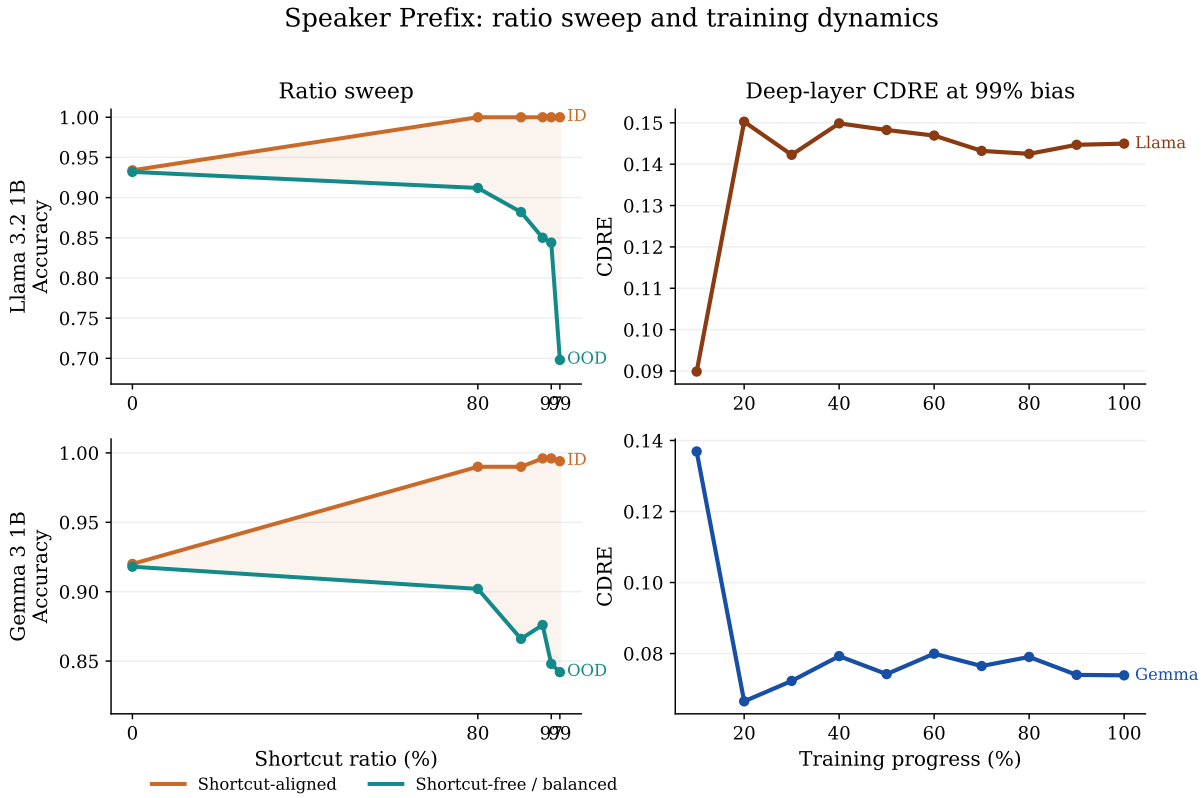

Figure 5: Appendix figure for speaker-prefix bias. Compared with capitalization, the OOD drop is milder and deep-layer CDRE remains lower, especially for Gemma.

C Supplementary Heatmaps

The original layerwise heatmaps remain useful as a qualitative diagnostic, even though they are too detailed for the main paper. Figure 6 visualizes CDRE over both depth and training progress for the SST-2 settings at high shortcut bias. These heatmaps support the more compressed main-text story: the strongest failures are the ones where the later layers warm up most visibly over the course of fine-tuning.

Figure 4: Layer-wise CDRE During Training (99% shortcut)

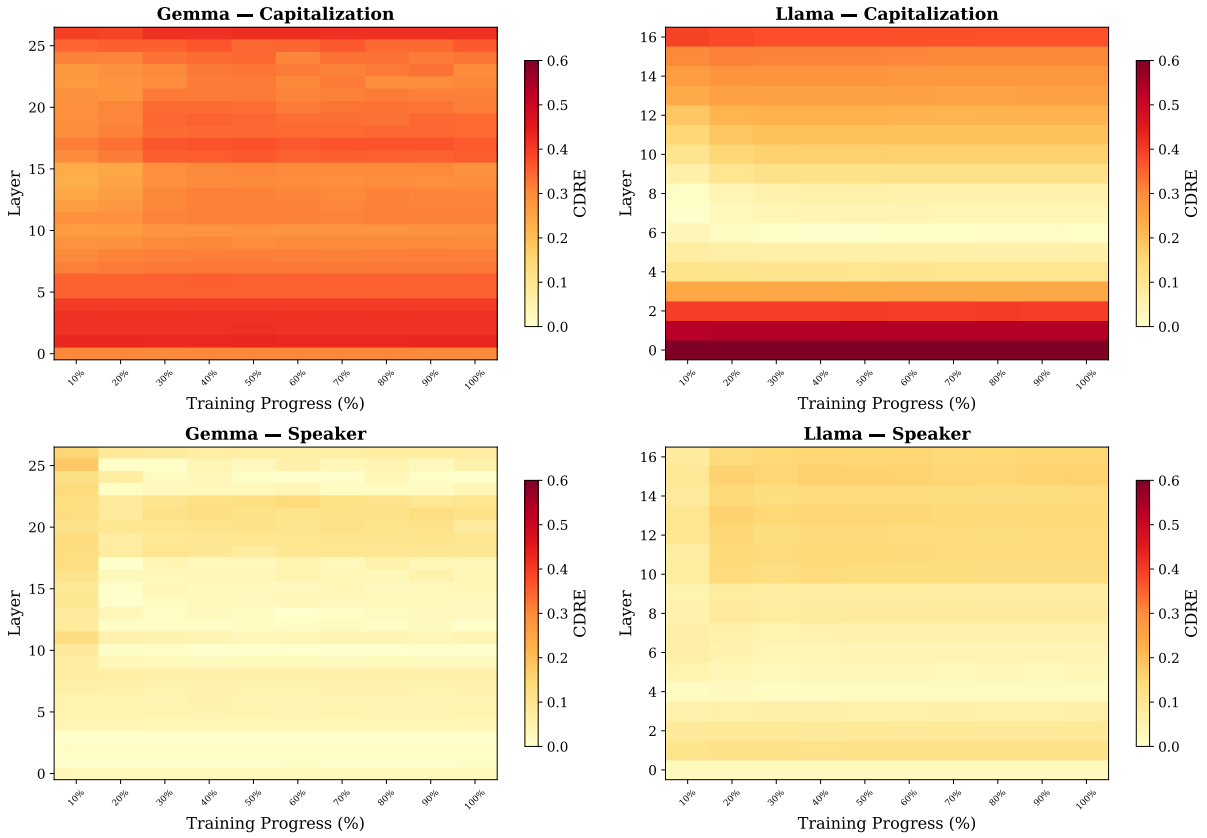

Figure 6: Supplementary layerwise heatmaps for the SST-2 settings. Darker regions indicate stronger concept-shortcut entanglement as training progresses.

D Multi-Seed Result Verification

Figures 7 and 8 show multi-seed runs. In each plot, the dark curve is seed 42 and the light curves are seeds 35, 470, and 3407. Across seeds, the same qualitative trends appear, supporting the main claims in the paper.

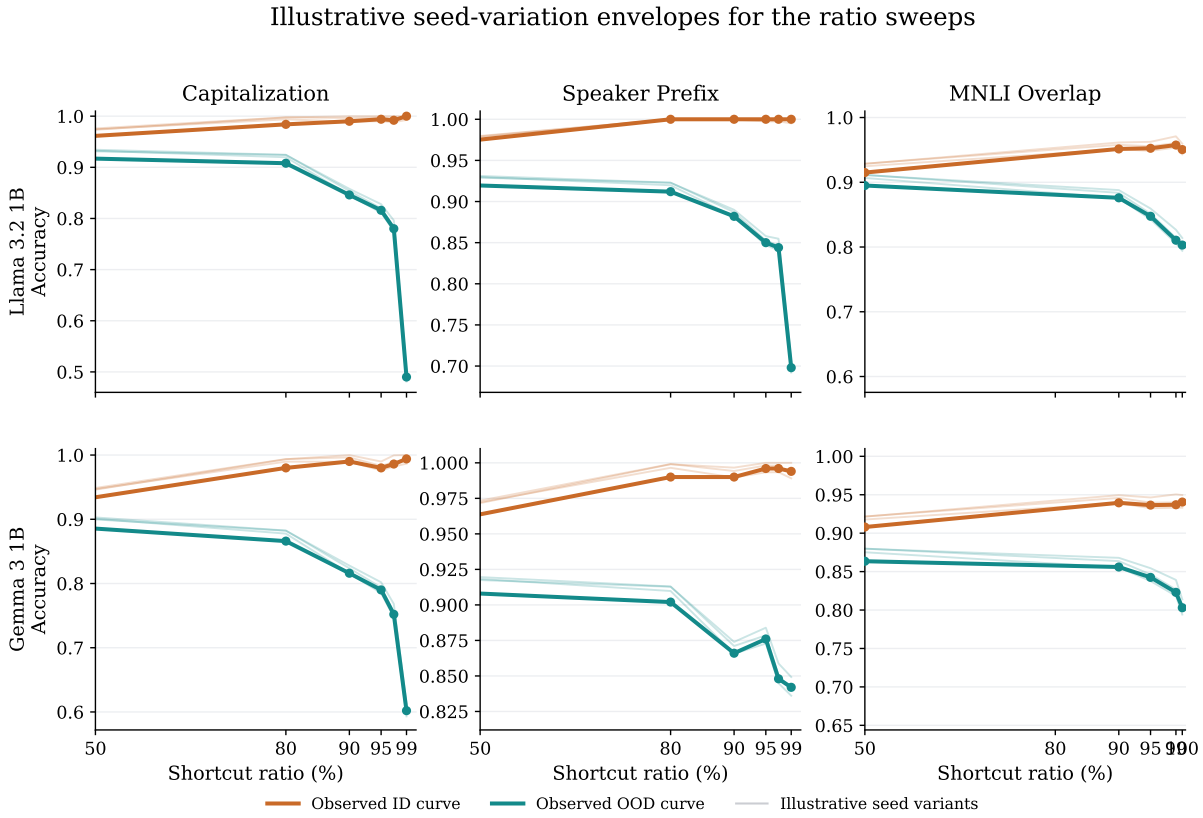

Figure 7: Multi-seed ratio sweeps. The dark curve denotes seed 42, and the lighter curves denote additional runs with seeds 35, 470, and 3407. The qualitative patterns are stable across seeds, reinforcing the central empirical conclusions of the paper.

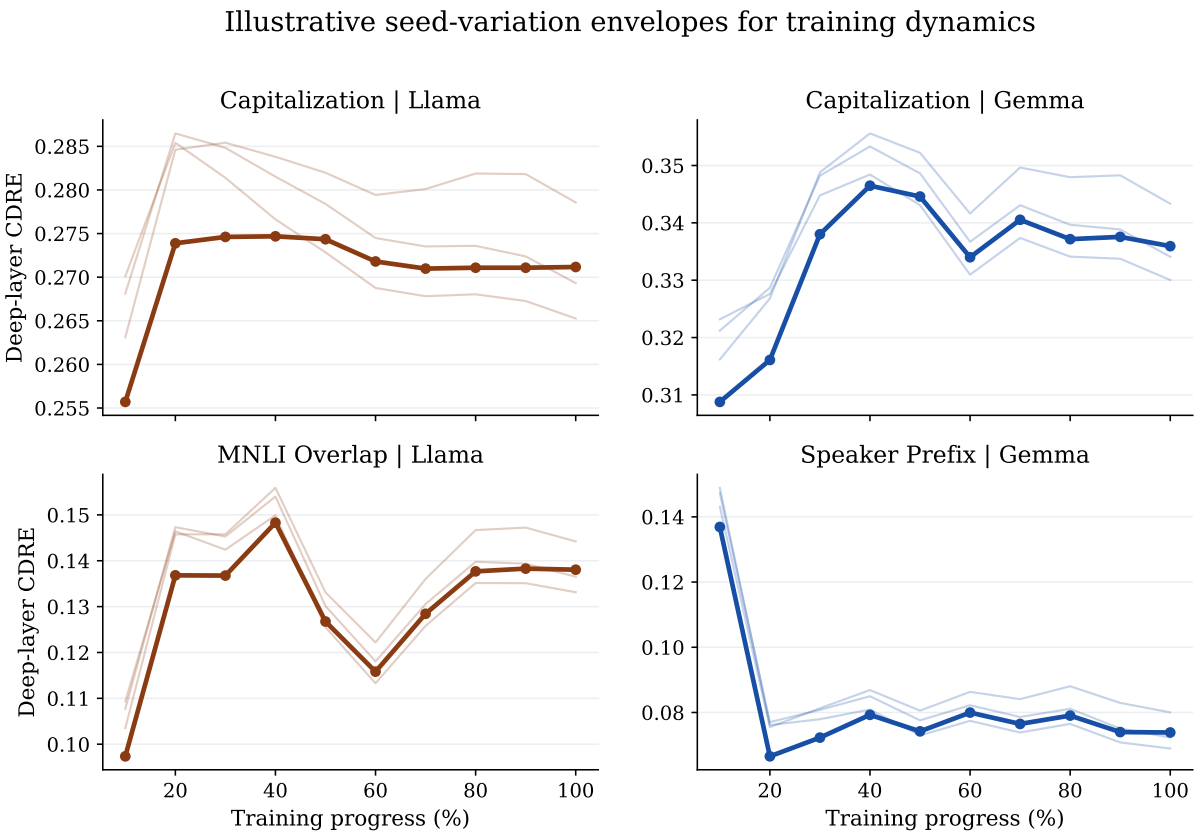

Figure 8: Multi-seed ratio sweeps. The dark curve denotes seed 42, and the lighter curves denote additional runs with seeds 35, 470, and 3407. The qualitative patterns are stable across seeds, reinforcing the central empirical conclusions of the paper.

E Why JTT Underperforms Here

JTT performs poorly in our setting, often below the baseline. The likely reason is that the first-stage biased model makes very few training errors when the shortcut is easy, leaving too little signal for effective reweighting. The clearest example is capitalization-biased Llama, where the first stage identifies only 7 errors out of 10,000 training examples. Even in MNLI, where the error set is larger (92 for Llama and 291 for Gemma), JTT remains far behind CDRE regularization.

This failure mode is informative rather than merely disappointing. It suggests that when a shortcut makes the training distribution too easy, error-centric methods can struggle because the biased model does not expose enough hard examples for the second stage to focus on. A probe-based intervention has a different advantage: it can act even when the biased model is nearly perfect on its own training distribution.

F What the Negative Results Teach Us

The speaker-prefix results help clarify what CDRE measures. Shortcut probe accuracy is perfect or near-perfect in these runs, which means the shortcut is easy to decode from the representation. However, CDRE remains small, and robustness remains relatively stable compared to capitalization. This suggests that recoverability alone is not enough to predict failure; what matters is whether shortcut information becomes aligned with the task direction used by the classifier.

This distinction also helps interpret the regularization results. If CDRE is already very small, forcing the representations to further decorrelate concept and shortcut directions gives little benefit and may simply add optimization noise. The speaker-prefix experiments therefore serve as a useful negative control for the broader claim: CDRE is most informative when it discriminates harmful alignment from harmless availability.