

Comparing Linear Probes with Mahalanobis Cosine Similarity

Zhuofan Josh Ying^{1,4} Peter Hase^{5,6} Nikolaus Kriegeskorte^{1,2,3,4}

Departments of ¹Psychology, ²Neuroscience, ³Electrical Engineering

⁴Zuckerman Mind Brain Behavior Institute; Columbia University, New York, NY

⁵Stanford University, Stanford, CA

⁶Schmidt Sciences, New York, NY

zy2559@columbia.edu phase@stanford.edu nk2765@columbia.edu

Abstract

Linear probes are widely used in interpretability research and often compared by cosine similarity. The *Mahalanobis cosine similarity* (MCS) between two directions, which reweights the inner product by test data covariance, is a natural task-aware refinement. Ying et al. (2026) report a nearly perfect linear relationship between a probe’s MCS to a reference probe trained on out-of-distribution (OOD) data and its OOD AUROC ($R^2=0.98$). Here, we extend this empirical finding across models, layers, and concept domains, and prove it in closed form: For balanced classes whose projections are Gaussian, OOD AUROC and MCS to the reference probe are near-linear because both are sigmoid-shaped functions of the probe’s signal-to-noise ratio (SNR) on the test data. The theory also predicts when this linearity fails, which we verify empirically. MCS offers a theoretically grounded and empirically effective alternative to Euclidean cosine similarity for comparing linear probes.

1 Introduction

Linear probes are powerful interpretability tools (Alain and Bengio, 2016; Belinkov, 2022; Marks and Tegmark, 2023), but their transfer behavior is notoriously brittle: they often fail to generalize to closely related datasets, and break dramatically under subtle shifts such as negation or prompt-format changes (Hewitt and Liang, 2019; Levinstein and Hermann, 2024; Orgad et al., 2025). Understanding which probes generalize, and why two probes that look similar on one distribution diverge on another, requires a principled way to compare them.

Ying et al. (2026) compare a probe w_{id} trained on an in-distribution (ID) task with a reference probe w_{ood} trained on an out-of-distribution (OOD) task. They find a nearly perfect linear relationship ($R^2=0.98$; Fig. 1) between the OOD AUROC of the ID probe w_{id} and the *Mahalanobis cosine similarity* (MCS) between the two probes, which

reweights the inner product between the two probes by the OOD data covariance. However, this was shown only for one model on truthfulness datasets, and such near-perfect linearity demands a theoretical explanation.

Empirically, we replicate and substantially extend this regularity (Tab. 1). The near-linear AUROC–MCS relationship holds across models (Llama-70B, Llama-8B, Qwen-7B), across layers (20–65 of Llama-70B), and across 24 datasets from three concept domains (truthfulness, gender classification, and general NLP benchmarks), with $R^2>0.93$ in every condition. By contrast, the standard Euclidean cosine similarity (ECS) drops to as low as $R^2=0.06$ (§3). Intuitively, ECS treats all dimensions equally, but only the dimensions where the data varies affect probe performance. MCS reweights the inner product by OOD data covariance, thus comparing probes in the subspace that actually matters for generalization performance.

Theoretically, we prove this linear relationship. Under balanced classes and per-class projection Gaussianity, both the ID probe’s OOD AUROC and its MCS to the OOD probe are S-shaped functions of the ID probe’s signal-to-noise ratio (SNR) s on the OOD task: AUROC is the Gaussian CDF $\Phi(s/\sqrt{2})$ in s , and MCS is a softsign function that converges to $s/\sqrt{(4+s^2)}$ with moderately large Fisher distance z_{max} (Fig. 2a,b). Composing the two, where each S-shaped inverts the other’s curvature, cancels out the S-shape, leaving a near-linear AUROC–MCS curve (§4; Fig. 2c).

The theory also predicts when the linearity fails: small z_{max} , class imbalance, non-Fisher reference directions like difference of means, or using pooled covariance instead of the total covariance. We verify each failure mode in simulations or on real activations (§5; Fig. 3).

Most ingredients of the proof here are textbook (Fisher, 1936; Green and Swets, 1966; Chatfield, 2018), and MCS itself is not new (Bolme et al.,

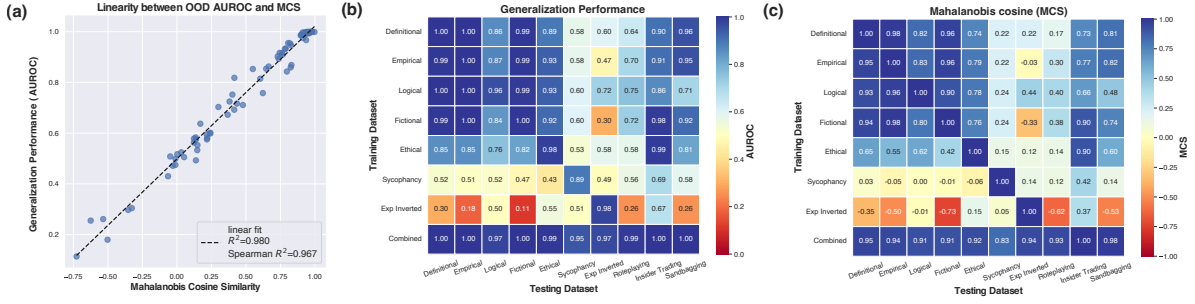


Figure 1: **Mahalanobis cosine similarity (MCS) linearly tracks generalization performance.** (a) AUROC is a near-linear function of MCS across heterogeneous tasks ($R^2=0.98$). (b–c) The generalization AUROC heatmap and the MCS heatmap share structure almost entry-for-entry. Reproduced from Ying et al. (2026)

2003; Iacovacci et al., 2020). The closest theoretical neighbor is Agreement-on-the-Line (Miller et al., 2021; Baek et al., 2022, 2025), which uses the same covariance-projected cosine to predict accuracy linearity between classifiers. Our contributions are: 1) derive a closed-form relation between a probe’s OOD AUROC and its MCS to the OOD Fisher direction; 2) explain the near-universal slope through matched saturation of AUROC and MCS with SNR; 3) identify total, rather than pooled, covariance as the geometry underlying this near-linearity; and 4) predict and empirically validate its failure modes. See App. A for extended related works.

Together, MCS offers a theoretically grounded task-aware alternative for comparing linear probes.

2 Mahalanobis Cosine Similarity

Let $X \in \mathbb{R}^d$ be a feature vector and $Y \in \{0, 1\}$ a binary label. We analyze the OOD behavior of a fixed direction w_{id} trained on some ID task. A linear probe assigns the score $w^\top X + b$; we omit the intercept, which affects neither AUROC nor the score correlations considered below. We compare a fixed ID-trained direction w_{id} with a reference direction w_{ood} trained on the target OOD task. Unless otherwise stated, all moments refer to this target OOD distribution, with total covariance $\Sigma_{\text{tot}} := \text{Cov}(X)$.

Euclidean and Mahalanobis cosine similarity. For nonzero directions $u, v \in \mathbb{R}^d$, Euclidean cosine similarity (ECS) is:

$$\text{ECS}(u, v) := \frac{u^\top v}{\|u\|_2 \|v\|_2}. \quad (1)$$

We define Mahalanobis cosine similarity (MCS) as:

$$\text{MCS}_{\Sigma_{\text{tot}}}(u, v) := \frac{u^\top \Sigma_{\text{tot}} v}{\sqrt{u^\top \Sigma_{\text{tot}} u} \sqrt{v^\top \Sigma_{\text{tot}} v}}, \quad (2)$$

so the comparison depends on the total covariance Σ_{tot} of the target distribution rather than only on the coefficient vectors.

This choice has a direct interpretation: MCS is equivalent to the Pearson correlation between the two probes’ scores on the target data. Since $\text{Cov}(u^\top X, v^\top X) = u^\top \Sigma_{\text{tot}} v$, we have:

$$\text{MCS}(u, v) = \text{Corr}(u^\top X, v^\top X). \quad (3)$$

ECS, by contrast, compares coefficients without accounting for target-data variances. For example, coefficient components along directions with zero OOD variance do not contribute to MCS, although ECS treats them the same as any other direction.

Relationship to Mahalanobis distance. Note that the weighting in MCS is the opposite of that used in Mahalanobis distance, which normalizes data points using Σ^{-1} . Here, we compare probe weights in the dual space rather than data points in the primal space. We transform data and probe weights inversely, and thus the projections onto the probe are preserved before and after the transformation, as desired.

3 Empirical evidence

Design. We use Llama-3.3-70B at residual-stream layer 33, with mean-pooled activations over response tokens following Ying et al. (2026). To test generality, we additionally evaluate at layers {20, 50, 65} of Llama-3.3-70B, at middle layers of Llama-3.1-8B and Qwen2.5-7B (Qwen et al., 2025; Grattafiori et al., 2024). We use datasets across three domains exhibiting rich generalization patterns: ten truthfulness datasets, six gender classification datasets, and eight general NLP classification datasets. See App. B for details.

We train logistic-regression probes for each task, denoting the ID- and OOD-trained directions by

Condition	R^2 (MCS)	R^2 (ECS)
Llama-70B, L33, truth	0.980	0.441
<i>Layers</i> (Llama-70B, truth)		
layer 20	0.990	0.628
layer 50	0.963	0.422
layer 65	0.976	0.446
<i>Domains</i> (Llama-70B, L33)		
Gender classification	0.976	0.723
General NLP benchmarks	0.936	0.057
<i>Models</i> (truth)		
Llama-3.1-8B	0.935	0.530
Qwen2.5-7B	0.994	0.656

Table 1: Linear-fit R^2 of held-out AUROC against Mahalanobis vs. standard Euclidean cosine similarity (MCS vs. ECS), across layers, domains, and model families and sizes. MCS dominates ECS in every condition.

w_{id} and w_{ood} , respectively. For each (ID, OOD) pair, we split the OOD data into disjoint training and test halves. On the training half, we fit w_{ood} , estimate Σ_{tot} , and compute the $\text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}})$. On the test half, we evaluate the empirical AUROC of w_{id} . This keeps AUROC evaluation separate from OOD probe fitting and moment estimation.

Results. We show that the OOD AUROC of w_{id} is a near-linear function of $\text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}})$, with $R^2 \geq 0.93$ for all conditions tested (Fig. 1a, Tab. 1). ECS is substantially worse: R^2 drops to 0.06 in the worst condition. Fig. 1b–c shows that the structure of the generalization-AUROC heatmap (every (train, test) pair on the truthfulness benchmarks) is mirrored almost entry-for-entry by the MCS heatmap. See Fig. 4, 5 and Tab. 2 in App. C for the generalization performance, MCS and ECS plotted against AUROC, and robustness to Σ_{tot} estimators (full sample, Ledoit–Wolf, and per-coordinate diagonal) across all conditions.

4 Theory

4.1 Setup

We retain the notation of §2. All distributions and moments below refer to the target OOD task. For $c \in \{0, 1\}$, let $\mu_c := \mathbb{E}[X|Y=c]$, $\Sigma_c := \text{Cov}(X|Y=c)$. Then define the mean difference and pooled within-class covariance by: $\delta := \mu_1 - \mu_0$ and $\Sigma_{\text{pool}} := \frac{1}{2}(\Sigma_0 + \Sigma_1)$.

Assumption 1 (Balanced nondegenerate moments). $\Pr(Y=0) = \Pr(Y=1) = \frac{1}{2}$, $\Sigma_0, \Sigma_1 \succ 0$, $\delta \neq 0$.

These conditions establish the covariance geometry and nonzero Fisher direction that later allow MCS

to be expressed in terms of the signal-to-noise ratio and the Fisher distance.

Assumption 2 (Projection Gaussianity). *For each direction w used in an AUROC statement below, the projection $w^\top X | Y=c$ is Gaussian for $c \in \{0, 1\}$.*

This assumption is much weaker than joint Gaussianity of X , and is plausible for non-Gaussian distributions in a high-dimensional space by projection concentration (a Diaconis–Freedman-style central limit theorem for one-dimensional projections).

Empirically, the per-class projections are largely normal (Fig. 7), and simulations show the linearity survives clearly non-Gaussian projections (Fig. 8). Since the empirical projections are Gaussian, we retain this assumption for simplicity.

For any nonzero direction $w \in \mathbb{R}^d \setminus \{0\}$, we define

$$\text{SNR}(w) := \frac{w^\top \delta}{\sqrt{w^\top \Sigma_{\text{pool}} w}} \quad (4)$$

For simplicity, we abbreviate the candidate probe’s OOD signal-to-noise ratio as $s := \text{SNR}(w_{\text{id}})$.

The maximum achievable SNR is the Fisher distance $z_{\text{max}} := \sqrt{\delta^\top \Sigma_{\text{pool}}^{-1} \delta}$, attained by the Fisher direction $w_{\text{ood}}^{\text{LDA}} := \Sigma_{\text{pool}}^{-1} \delta$.

While our theory uses the OOD Fisher discriminant direction $w_{\text{ood}}^{\text{LDA}}$ because it admits the closed-form analysis developed below, in the empirical experiments, we use the OOD logistic-regression direction $w_{\text{ood}}^{\text{LR}}$ for consistency with the probe-training procedure used throughout the paper. The two reference directions yield closely aligned similarities: $\text{MCS}(w_{\text{id}}, w_{\text{ood}}^{\text{LR}}) \approx \text{MCS}(w_{\text{id}}, w_{\text{ood}}^{\text{LDA}})$ across all settings (Fig. 11; App. I), and their corresponding AUROC–MCS regressions produce similar R^2 values (Tab. 2; App. C). These support using the Fisher direction for the theoretical derivation while applying its conclusions to the logistic-regression probes in our empirical analyses.

4.2 Background results

Our proof will reduce the high-dimensional relationship between probe geometry and OOD performance to two scalars: the ID probe’s SNR s on the OOD data and the target OOD task’s Fisher distance (z_{max}). Three classical results make this reduction possible. The first decomposes the covariance used by MCS, the second expresses AUROC in terms of s , and the third characterizes the OOD

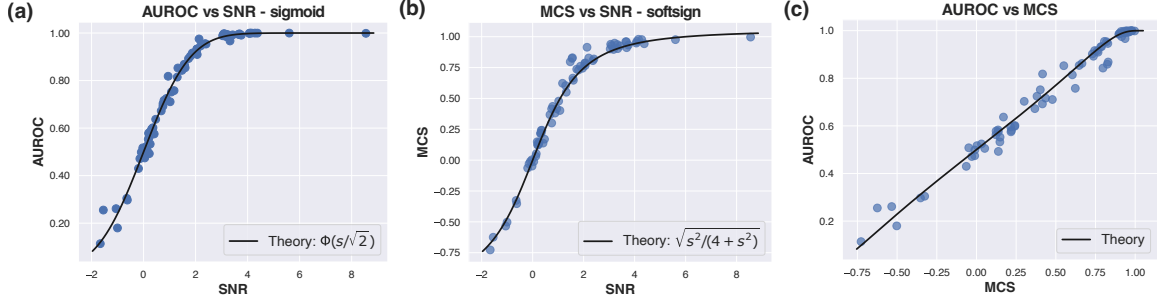


Figure 2: **Theory matches empirical data without free parameters.** Across panels, empirical points largely lie on the theory prediction. (a) The AUROC–SNR curve shows the sigmoid shape of Lemma 2. (b) The MCS–SNR curve shows the soft-sign shape of Theorem 1. (c) Eliminating SNR, the AUROC–MCS curve shows a near-straight line that bends only in the top-right corner, matching the empirical data well.

reference direction and $z_{\max}$. We state these results here and defer their textbook proofs to App. E.

Lemma 1 (Covariance decomposition). *Under Assumption 1, $\Sigma_{\text{tot}} := \text{Cov}(X) = \Sigma_{\text{pool}} + \frac{1}{4}\delta\delta^\top$.*

This decomposition shows how class separation enters the covariance geometry used by MCS: Σ_{pool} captures within-class variability, while $\frac{1}{4}\delta\delta^\top$ captures variability between the class means.

Lemma 2 (Binormal AUROC). *Under Assumptions 1–2, $\text{AUROC}(w) = \Phi(\text{SNR}(w)/\sqrt{2})$, where Φ is the standard Gaussian CDF, for any $w \neq 0$.*

This result requires only Gaussianity of the one-dimensional projections, not joint Gaussianity of X or equal class covariances. It reduces the OOD performance of w_{id} to a sigmoid-shaped function of the single scalar s .

Lemma 3 (Fisher’s discriminant). *Under Assumption 1, $\text{SNR}^2(w)$ is maximised on $\mathbb{R}^d \setminus \{0\}$ by any nonzero scalar multiple of $w_{\text{ood}} = \Sigma_{\text{pool}}^{-1}\delta$, with maximum SNR $z_{\max}^2 = \delta^\top \Sigma_{\text{pool}}^{-1}\delta$, where $z_{\max}$ is known as the Fisher distance.*

This is the standard Rayleigh-quotient result of Fisher (Fisher, 1936). Together, these three lemmas allow us to express both MCS and AUROC in terms of s and $z_{\max}$.

4.3 Closed form for the Mahalanobis cosine

The Mahalanobis cosine between an arbitrary direction w_{id} and the Fisher direction w_{ood} admits a closed form in two scalars: the SNR of w_{id} and the Fisher distance $z_{\max}$ of the OOD task.

Theorem 1 (Closed form for $\text{MCS}_{\Sigma_{\text{tot}}}$). *Under Assumption 1, for any w_{id} with $w_{\text{id}}^\top \Sigma_{\text{pool}} w_{\text{id}} > 0$ and $z_{\max} > 0$,*

$$\text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}}) = \frac{s}{z_{\max}} \sqrt{\frac{1 + \frac{1}{4}z_{\max}^2}{1 + \frac{1}{4}s^2}}, \quad (5)$$

where $s := \text{SNR}(w_{\text{id}})$. Moreover, in the large-Fisher-distance limit,

$$\lim_{z_{\max} \rightarrow \infty} \text{MCS}_{\Sigma_{\text{tot}}}(s) = \frac{s}{\sqrt{4 + s^2}}, \quad (6)$$

the softsign function: bounded in $(-1, 1)$, nearly linear near $s=0$, and saturating to ± 1 as $|s| \rightarrow \infty$.

The proof is a direct computation of the three quadratic forms in (2) using Lemma 1 (App. F). The derivation uses only standard objects from the LDA literature. However, the resulting identity has not been written down explicitly, perhaps because the link between probe transfer and the dependence of MCS on SNR had not been studied before.

The algebra here is basic. What matters is the *shape*. On $(-z_{\max}, z_{\max})$, $\text{MCS}_{\Sigma_{\text{tot}}}(s)$ is odd, strictly increasing, and satisfies $\text{MCS}_{\Sigma_{\text{tot}}}(\pm z_{\max}) = \pm 1$. As $z_{\max} \rightarrow \infty$, $\text{MCS}_{\Sigma_{\text{tot}}}(s)$ is exactly a softsign shape (Eq. 6; Fig. 2(b)).

4.4 Why AUROC is linear in MCS, with task-independent slope

Lemma 2 and Theorem 1 together imply that for a fixed OOD task (fixed $z_{\max}$), both AUROC and $\text{MCS}_{\Sigma_{\text{tot}}}$ are monotone functions of the signal-to-noise ratio s , tracing a parametric curve indexed only by $z_{\max}$ (Fig 2a,b). Two facts make this curve a near-linear law with a universal slope.

(i) Fixed $z_{\max}$: saturations cancel. Differentiating the parametric curve at any s gives:

$$\frac{d \text{AUROC}}{d \text{MCS}_{\Sigma_{\text{tot}}}} = \frac{\phi(s/\sqrt{2})(1 + s^2/4)^{3/2}}{\sqrt{2/z_{\max}^2 + 1/2}}, \quad (7)$$

where ϕ is the standard Gaussian density (full derivation in App. G). The two s -dependent factors are *opposed* in $|s|$: as $|s|$ grows, $\phi(s/\sqrt{2})$ *shrinks* (the AUROC sigmoid saturating) while

$(1+s^2/4)^{3/2}$ grows (the MCS softsign saturating). Their product is therefore much flatter than either factor alone, so the local slope stays close to its central value over the bulk of $\text{MCS}_{\Sigma_{\text{tot}}} \in (-1, 1)$, as shown in Fig. 2c and 9b.

The cancellation is not exact: as $|\text{MCS}_{\Sigma_{\text{tot}}}| \rightarrow 1$, the Gaussian factor decays faster than the polynomial grows, and the local slope drops toward 0 (both data and theory at top-right of Fig. 2c flatten). Also see Fig. 9 in App. G for the full local slope curve: the slope is between 0.539 and 0.629 for $|\text{MCS}| \leq 0.75$ and then quickly drops towards 0 when $|\text{MCS}| > 0.75$.

(ii) Per-task slope is universal. The slope in Eq. 7 factors as $h(s) \cdot g(z_{\text{max}})$, where $h(s)$ captures s -dependence and $g(z_{\text{max}}) = 1/\sqrt{2/z_{\text{max}}^2 + 1/2}$ captures task dependence. At $z_{\text{max}} > 20$, which holds in all empirical conditions (Tab. 3), $g(z_{\text{max}})$ lies within 0.5% of its limit. Therefore, the per-task AUROC–MCS curve aligns with the $z_{\text{max}} \rightarrow \infty$ limit curve to within 0.5%. At $s=0$, this limit curve has central slope $h(0) \cdot g(\infty) = 1/\sqrt{\pi}$ (App. G), and the empirical slopes across all conditions lie slightly below $1/\sqrt{\pi}$ (Fig. 10; App. G), since the tail sampling drags the global slope down from its central value.

5 When does the linearity break down

Our theory predicts that the AUROC–MCS linearity holds when: a) MCS is computed against Σ_{tot} , not Σ_{pool} ; b) the Fisher distance z_{max} is large; c) class proportions are roughly balanced; d) the OOD probes are close to the optimal Fisher direction. Each assumption can fail in practice. Fig. 3 shows what each violation does to the linearity.

(a) Wrong covariance. Using Σ_{pool} instead of Σ_{tot} turns MCS into s/z_{max} exactly, so AUROC becomes a sigmoid in MCS rather than a line (see App. H for derivation). On the main experiment data, the linear fit R^2 drops from 0.98 to 0.83.

(b) Non-Fisher probe. The theory does not hold if OOD probes deviate too much from the optimal Fisher directions, for example, for the commonly used difference of means probe (Marks and Tegmark, 2023). On the same empirical data, the diffmean probe gives a markedly lower R^2 of 0.79. This delimits the law: it holds for Fisher-style probes (LR, LDA, shrinkage variants), not for diffmean-style probes.

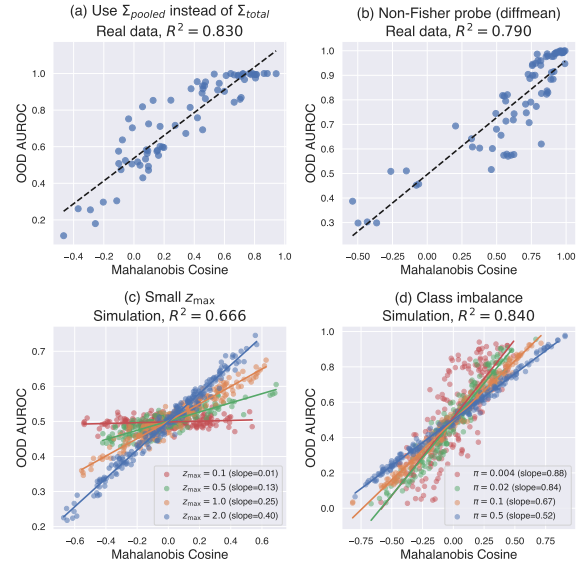


Figure 3: **Failure modes.** Each panel illustrates a violation of an assumption in §4, and the linearity breaks.

(c) Small Fisher distance. For small z_{max} , the slope of the MCS formula does not saturate, so each task is in its own near-linear regime with its own slope. Synthetic Gaussian data at $z_{\text{max}} \in \{0.1, 0.5, 1, 2\}$ exhibits a clear fan of per-group slopes spanning a large range. The LLM data is unaffected because $z_{\text{max}} > 20$ on every tasks (see Tab. 3; App. C). The linear fit is tight for each group, but taken together, the R^2 drops to 0.666.

(d) Class imbalance. The balanced-class factor $\frac{1}{4}$ in the MCS formula generalises to $\pi(1 - \pi)$, so the AUROC–MCS slope steepens as π moves away from $\frac{1}{2}$ (App. H). Synthetic data at $\pi \in \{0.5, 0.1, 0.02, 0.004\}$ shows monotone steepening. The R^2 drops to 0.84 in this case.

6 Discussion

Our results suggest that Mahalanobis cosine similarity is a theoretically sound alternative to standard Euclidean cosine similarity (consistent with recent arguments for non-Euclidean inner products (Park et al., 2024)). This also points to a broader research direction: many interpretability methods that currently rely on cosine similarity may benefit from the Mahalanobis alternative (steering-vector comparison, SAE feature alignment, concept-direction clustering, data filtering, data attribution, etc.). Future work can use unsupervised methods like CCS (Burns et al., 2023) for the OOD direction, enabling fully label-free prediction of probe generalization.

Limitations

Our results have several limitations. First, computing $MC_{\Sigma_{\text{tot}}}$ requires estimating w_{ood} and Σ_{tot} from labeled OOD data; the method is training-free for the candidate probe w_{id} but not label-free for the target distribution. Second, the theory is calibrated to Fisher-style probes (LR, LDA); for difference-of-means probes the linearity degrades markedly (Fig. 3b), and a closed form for non-Fisher references remains open. Third, our evaluation is confined to binary classification on residual-stream features of autoregressive LLMs; multiclass probes, attention/MLP features, and non-LLM architectures are untested. Fourth, the law is fundamentally approximate: the per-task total slope is close to $1/\sqrt{\pi}$ and the local slope goes to 0 at the extreme. Finally, the theory assumes per-class projection Gaussianity; while this holds empirically for most directions (Fig. 7) and the linearity survives clear violations in simulation (Fig. 8), formal guarantees do not extend to heavy-tailed projections.

Risks. We do not foresee risks specific to this work; it is a theoretical analysis of an existing similarity measure.

References

- Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- Akari Asai, Sara Evensen, Behzad Golshan, Alon Halevy, Vivian Li, Andrei Lopatenko, Daniela Stepanov, Yoshi Suhara, Wang-Chiew Tan, and Yinzhan Xu. 2018. Happydb: A corpus of 100,000 crowdsourced happy moments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. 2022. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289.
- Christina Baek, Aditi Raghunathan, and J Zico Kolter. 2025. Theory of agreement-on-the-line in linear models and gaussian data. In *The 28th International Conference on Artificial Intelligence and Statistics*.
- Donald Bamber. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, and 1 others. 2024. Sabotage evaluations for frontier models. *arXiv preprint arXiv:2410.21514*.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.
- David S Bolme, J Ross Beveridge, Marcio Teixeira, and Bruce A Draper. 2003. The csu face identification evaluation system: its purpose, features, and structure. In *International Conference on Computer Vision Systems*, pages 304–313. Springer.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.
- Chris Chatfield. 2018. *Introduction to multivariate analysis*. Routledge.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936.
- Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. 2001. On kernel-target alignment. *Advances in neural information processing systems*, 14.
- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnamurthy Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.
- Bradley Efron. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898.
- Figure Eight. 2016. [Twitter user gender classification](#). *Kaggle Dataset*.
- Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, and 1 others. 2024. Nnsight and ndif: Democratizing access to open-weight foundation model

- internals. In *The Thirteenth International Conference on Learning Representations*.
- Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.
- Tomer Galanti, András György, and Marcus Hutter. 2021. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*.
- Quentin Garrido, Randall Balestrieri, Laurent Najman, and Yann Lecun. 2023. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pages 10929–10974. PMLR.
- Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. Detecting strategic deception with linear probes. In *Forty-second International Conference on Machine Learning*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- David Marvin Green and John A Swets. 1966. *Signal detection theory and psychophysics*, volume 1. Wiley New York.
- James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction.
- 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 (emnlp-ijcnlp)*, pages 2733–2743.
- Jacopo Iacovacci, Alina Peluso, Timothy Ebbels, Markus Ralser, and Robert C Glen. 2020. Extraction and integration of genetic networks from short-profile omic data sets. *Metabolites*, 10(11):435.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and 1 others. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.
- Matej Martinc, Iza Skrjanec, Katja Zupan, and Senja Pollak. 2017. Pan 2017: Author profiling-gender and language variety prediction. In *CLEF (working notes)*.
- Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.
- John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *ICLR*.
- Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 39643–39666.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. *Qwen2.5 technical report*. Preprint, arXiv:2412.15115.

- Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2023. Large language models can strategically deceive their users when put under pressure. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Samuel S Wilks. 1932. Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. 2026. The truthfulness spectrum hypothesis. *arXiv preprint arXiv:2602.20273*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20.

A Extended related works

Linear probes and their generalization. Linear probes are a standard interpretability tool (Alain and Bengio, 2016; Belinkov, 2022), used to read off concepts such as truthfulness (Marks and Tegmark, 2023), but often degrade under distribution shift (Hewitt and Liang, 2019; Levinstein and Hermann, 2024; Orgad et al., 2025). Prior work predicts linear-classifier generalization from geometric properties of the representation—kernel-target alignment (Cristianini et al., 2001), spectral rank (Garrido et al., 2023), and neural-collapse statistics (Papayan et al., 2020; Galanti et al., 2021)—but these score whole representations rather than individual probe directions. The near-linear AUROC–MCS relationship we explain was reported empirically by Ying et al. (2026) on truthfulness datasets; we prove it, characterize when it fails, and verify it across models, layers, and domains.

Classical statistical machinery. Our theory composes three textbook results: Fisher’s linear discriminant (Fisher, 1936), the binormal AUROC formula (Bamber, 1975; Green and Swets, 1966; Hanley and McNeil, 1982), and the within–between covariance decomposition underlying MANOVA (Wilks, 1932; Chatfield, 2018). None is novel in isolation; the contribution lies in composing them into a closed-form quantity computable from a candidate probe direction together with two target-task moments.

Agreement-on-the-Line. The closest theoretical neighbor is the Agreement-on-the-Line framework (Miller et al., 2021; Baek et al., 2022, 2025), which proves a linear ID-to-OOD relationship for probit-scaled accuracy and agreement rates between linear classifiers under Gaussian data. Baek et al. (2025) parameterize this relationship by a similarity that is algebraically the Mahalanobis cosine of Eq. (1). We differ in (i) targeting threshold-free AUROC, which admits the closed form of Theorem 1; (ii) showing OOD performance is near-linear to a single geometric quantity against the OOD Fisher direction, with no ID baseline; and (iii) identifying the universal slope around $1/\sqrt{\pi}$ from the AUROC–MCS saturation cancellation, which has no analog in the AOL line.

Mahalanobis cosine as a similarity for direction comparison. The most direct precursor of our setup is Bolme et al. (2003), who introduced "Mahalanobis cosine" as a similarity measure in

PCA-whitened face-recognition pipelines. Iacovacci et al. (2020) reused the same expression for inferring omics networks. Both works treat MCS as a similarity score to be evaluated empirically against task performance; neither expresses MCS in closed form as a function of probe SNR and task Fisher distance, and neither links it to AUROC. Our contribution in §4 is to make precise why MCS works as a direction-comparison similarity for classification — in a closed form (Theorem 1) whose composition with the binormal AUROC is task-calibrated and near-linearly related to held-out performance.

Non-Euclidean geometry of LLM representations. Park et al. (2024) also argue against Euclidean cosine on LLM representation space and propose a "causal inner product" derived from counterfactual considerations on the unembedding space. Their specific instantiation whitens by the covariance of unembedding token vectors, a single global matrix per model. Our Σ_{tot} is a per-task activation covariance computed on the OOD train data, and operates in the residual-stream embedding space rather than the unembedding space.

B Experimental details

B.1 Computational resources.

All experiments are done on local L40S and A40. Experiments require 4 GPUs to run for about 10 hours. We use Huggingface and NNsight to extract activations (Wolf et al., 2020; Fiotto-Kaufman et al., 2024). The logistic regression probe is implemented with Pedregosa et al. (2011).

B.2 Model and probe details

Models. We use Llama-3.3-70B, Llama-3.1-8B, and Qwen-2.5-7B (Grattafiori et al., 2024; Qwen et al., 2025). We use layer 15 and layer 19 for Llama-3.1-8B and Qwen-2.5-7B following Ying et al. (2026).

B.3 Dataset details

Truthfulness datasets. We use ten total truthfulness datasets: five fundamental truth types from the FLEED dataset (definitional, empirical, logical, fictional, and ethical truth) ($\approx 8k$ samples), sycophantic lying ($\approx 4k$ samples), expectation inverted lying (≈ 600 sample), and three on-policy deception datasets (roleplay (371 samples), insider trading (1.5k samples), and sandbagging (1k samples)) (Scheurer et al., 2023; Benton et al., 2024;

Goldowsky-Dill et al., 2025; Ying et al., 2026). For Qwen-2.5-7B, we don’t have the three on-policy deception datasets. Therefore, it is evaluated only on 7 datasets.

Gender classification datasets. We use six gender classification datasets: PAN17 (Martinc et al., 2017), CrowdFlower Twitter gender classification (Eight, 2016), HappyDB (Asai et al., 2018), Bios-Bias (De-Arteaga et al., 2019), WinoBias (Zhao et al., 2018), and GAP (Webster et al., 2018). For datasets larger than 4000 samples, we randomly subsample to 4000 samples. Therefore, GAP has 2,000 samples, PAN17 randomly combines N tweets from the same user and has 2,400 samples, WinoBias has 1,584 samples, and the rest has 4,000 samples each.

General NLP benchmarks. We use eight classic NLP QA datasets: sentiment classification (IMDB (Maas et al., 2011) and Amazon (McAuley and Leskovec, 2013)), topic classification (AG-News (Zhang et al., 2015) and DBpedia-14 (Lehmann et al., 2015)), NLI (RTE (Wang et al., 2018)), question answering (BoolQ (Clark et al., 2019)), and common sense reasoning (PIQA (Bisk et al., 2020)). We also subsample large datasets to 4,000 samples. Therefore, PIQA has 3,674 samples, RTE has 552 samples, and the rest has 4,000 samples each. Following Burns et al. (2023), each example from the multiclass datasets becomes a contrast pair: a true statement asserting the gold class and a false statement asserting a distractor class. The datasets are thus exactly balanced by construction.

License. All datasets used in this work are publicly available and applied here only for non-commercial academic evaluation, consistent with the terms set by their original authors. Explicit licenses, where available, are: MIT (De-Arteaga et al., 2019; Zhao et al., 2018), Apache 2.0 (Webster et al., 2018), CC BY-SA 3.0 (Lehmann et al., 2015; Clark et al., 2019), Academic Free License v3.0 (Bisk et al., 2020), CC0: Public Domain (Eight, 2016), and CC BY 4.0 (Scheurer et al., 2023). The remaining datasets are released under research-use terms specified in their associated publications. All data is English-language.

C Additional empirical results

Cross-domain generalization performance. As shown in Fig 4, we observe rich cross-domain generalization patterns across all eight conditions.

MCS and ECS against AUROC. Tab. 1 summarizes the performance of MCS and ECS on predicting OOD AUROC. We show the scatter plots for all eight conditions in Fig 5. We observe right cross-domain generalization patterns across all eight conditions.

Verification of the theory across conditions. We show the theory against empirical data across all eight conditions in Fig. 6. In all conditions, the theoretical prediction tracks the empirical data very well.

Using w_{ood}^{LDA} instead of w_{ood}^{LR} for the empirical results. We show the performance of using $MCS_{\Sigma_{tot}}(w_{id}, w_{ood}^{LR})$ vs. $MCS_{\Sigma_{tot}}(w_{id}, w_{ood}^{LDA})$ to predict the OOD generalization performance in Tab. 2. Using $MCS_{\Sigma_{tot}}(w_{id}, w_{ood}^{LDA})$ is better than $MCS_{\Sigma_{tot}}(w_{id}, w_{ood}^{LR})$ in all conditions, as expected. But the difference is small (< 0.05).

Empirical AUROC–MCS slopes are consistent with theory. As shown in Fig. 10, the empirically fitted slopes across 8 conditions are consistent with the theoretical prediction from App. G.

Measuring Gaussianity of the empirical data. As shown in Fig. 7, most of the empirical skewness and kurtosis are small across all conditions, all probe directions, and all test data distributions. Specifically, 73% of the skewness is within ± 0.5 and 79% of the kurtosis is within ± 1 .

Simulating non-Gaussian distributions. As shown in Fig. 8, the AUROC–MCS R^2 is extremely high in simulations, even for distributions that are deliberately designed to violate the projection-Gaussianity assumption.

Empirical z_{max} across all datasets. As shown in Tab. 3, all z_{max} across all conditions and datasets exceed the saturation threshold $z_{max} \geq 20$ at which the slopes lie within 0.5% of their limits (App. G), pinning all per-task AUROC–MCS slopes to a universal slope. The average z_{max} for each condition is greater than 100, far exceeding the saturation point.

Robustness to the Σ_{tot} estimator. Tab. 2 reports the AUROC–MCS linear-fit R^2 under three estimators of Σ_{tot} : the full sample covariance, Ledoit–Wolf shrinkage, and a per-coordinate diagonal approximation. Ledoit–Wolf is essentially indistinguishable from the full covariance across all conditions, so the headline result is not an artifact

Condition	R^2			
	LR	LDA	LR- $\Sigma_{tot,diag}$	LR- $\Sigma_{tot,LW}$
Llama-70B L33, truth	0.980	0.987	0.438	0.980
<i>Layers</i> (Llama-70B, truth)				
layer 20	0.990	0.990	0.620	0.990
layer 50	0.963	0.973	0.443	0.963
layer 65	0.976	0.978	0.422	0.975
<i>Domains</i> (Llama-70B L33)				
Gender classification	0.976	0.983	0.675	0.977
General NLP benchmarks	0.936	0.982	0.002	0.936
<i>Models</i> (truth)				
Llama-3.1-8B	0.935	0.962	0.438	0.934
Qwen2.5-7B	0.994	0.992	0.717	0.993

Table 2: **Robustness check for the linear-fit of AUROC—MCS on MCS choices.** We probe two axes of robustness: (i) the reference direction, comparing logistic regression (LR) $\text{MCS}_{\Sigma_{tot}}(w_{id}, w_{ood}^{LR})$ against the theoretically motivated LDA direction $\text{MCS}_{\Sigma_{tot}}(w_{id}, w_{ood}^{LDA})$, and the LDA version dominates the LR version, but the difference is small; (ii) the shrinkage estimator for Σ_{tot} , comparing the full covariance against a per-coordinate diagonal approximation ($\text{MCS}_{tot,diag}(w_{id}, w_{ood}^{LR})$) and Ledoit–Wolf shrinkage ($\text{MCS}_{tot,LW}(w_{id}, w_{ood}^{LR})$). Ledoit–Wolf shrinkage is very similar to the full covariance we used, while the diagonal covariance is significantly worse.

Condition	min $z_{\max}$	max $z_{\max}$	mean $z_{\max}$
Llama-70B L33, truth	87.72	477.73	284.79
<i>Layers</i> (Llama-70B, truth)			
layer 20	22.35	220.08	122.85
layer 50	90.53	925.78	533.96
layer 65	186.48	1377.71	852.16
<i>Domains</i> (Llama-70B L33)			
Gender classification	25.63	196.62	103.25
General NLP benchmarks	41.50	379.83	142.83
<i>Models</i> (truth)			
Llama-3.1-8B	43.55	263.66	151.57
Qwen2.5-7B	154.39	2724.74	1451.37

Table 3: Empirical Fisher distance $z_{\max} = \sqrt{\delta^\top \Sigma_{\text{pool}}^{-1} \delta}$ per condition, computed on the OOD train half across all OOD tasks. All values have $z_{\max} \geq 20$, at which point the per-task slopes of AUROC–MCS curve lie within 0.5% of their limits, making a single strong linear fit across all tasks possible (App. G).

of a particular regularization choice. The diagonal approximation, by contrast, degrades sharply: most strikingly on general NLP ($R^2=0.002$ vs. 0.936), confirming that off-diagonal covariance structure is essential to MCS and that the linearity is not recoverable from per-coordinate variances alone.

D AI Usage

We use AI assistants to assist with writing paper drafts, coding for experiments, and writing proofs.

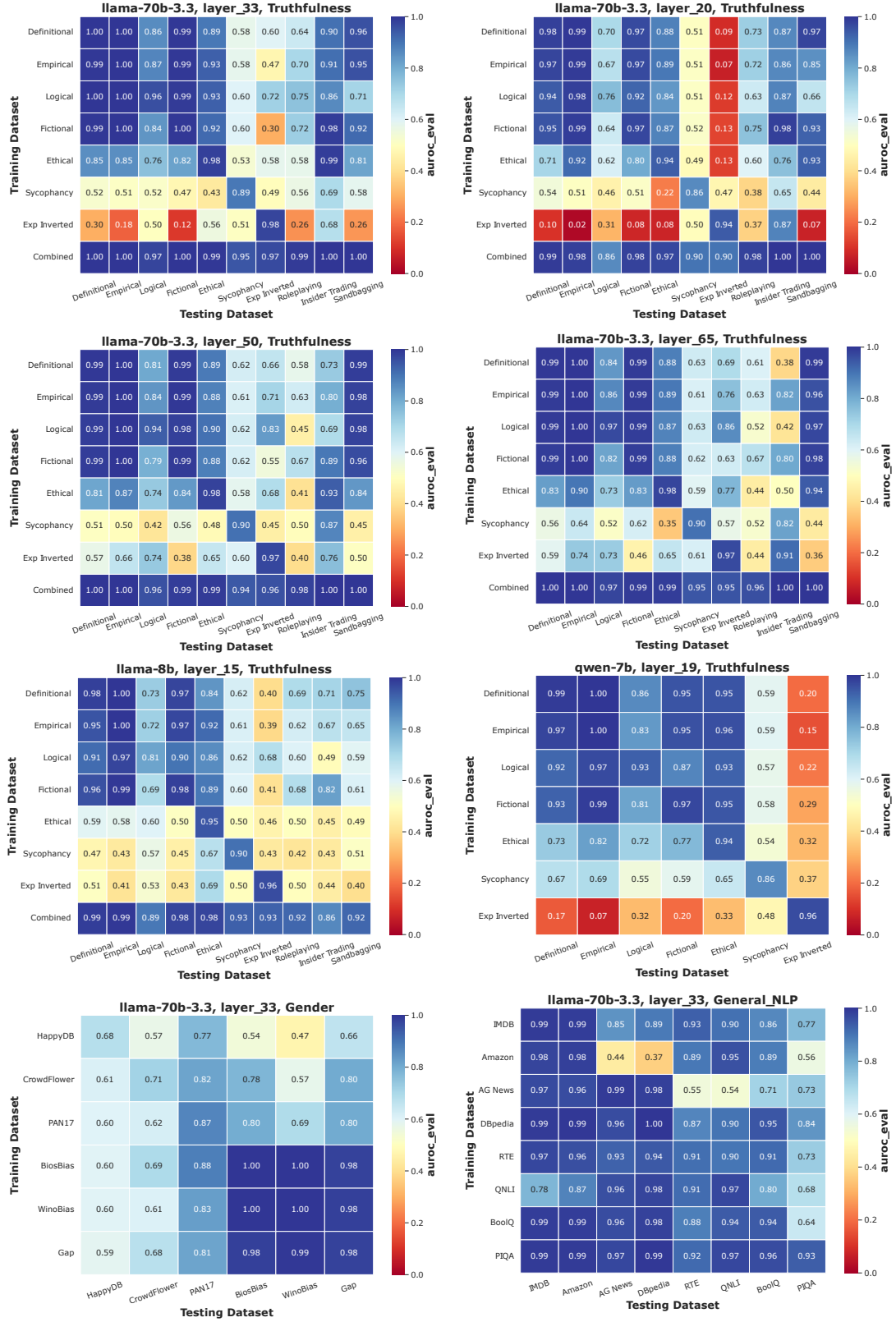


Figure 4: **Cross-domain generalization performance for all eight conditions across models, layers, and concept domains.** We observe rich cross-domain generalization patterns across all eight conditions.

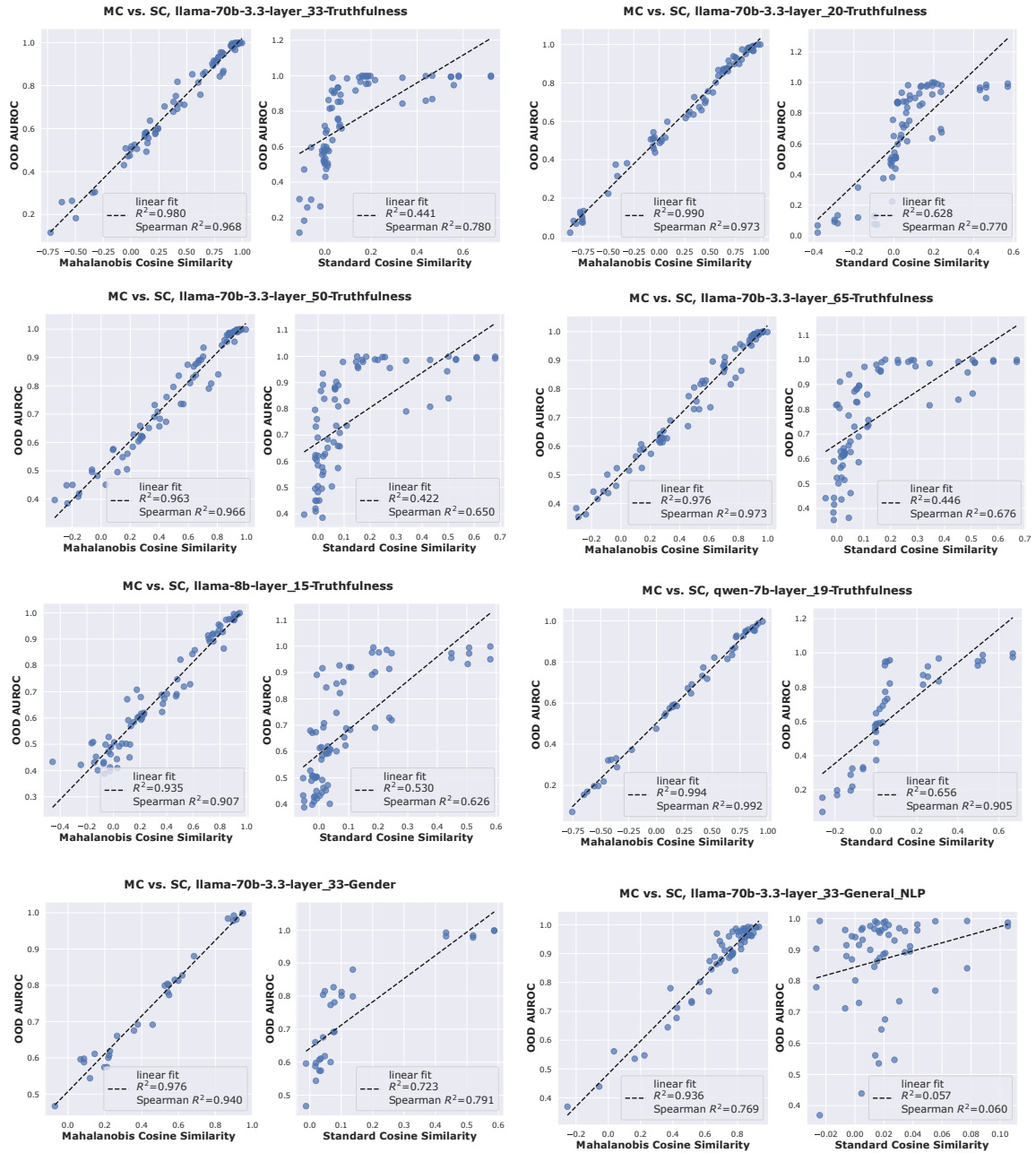


Figure 5: **MCS and ECS against AUROC across conditions.** We observe a strong linear relationship between MCS and AUROC for all eight conditions across models, layers, and concept domains, while the relationship between ECS and AUROC is much weaker.

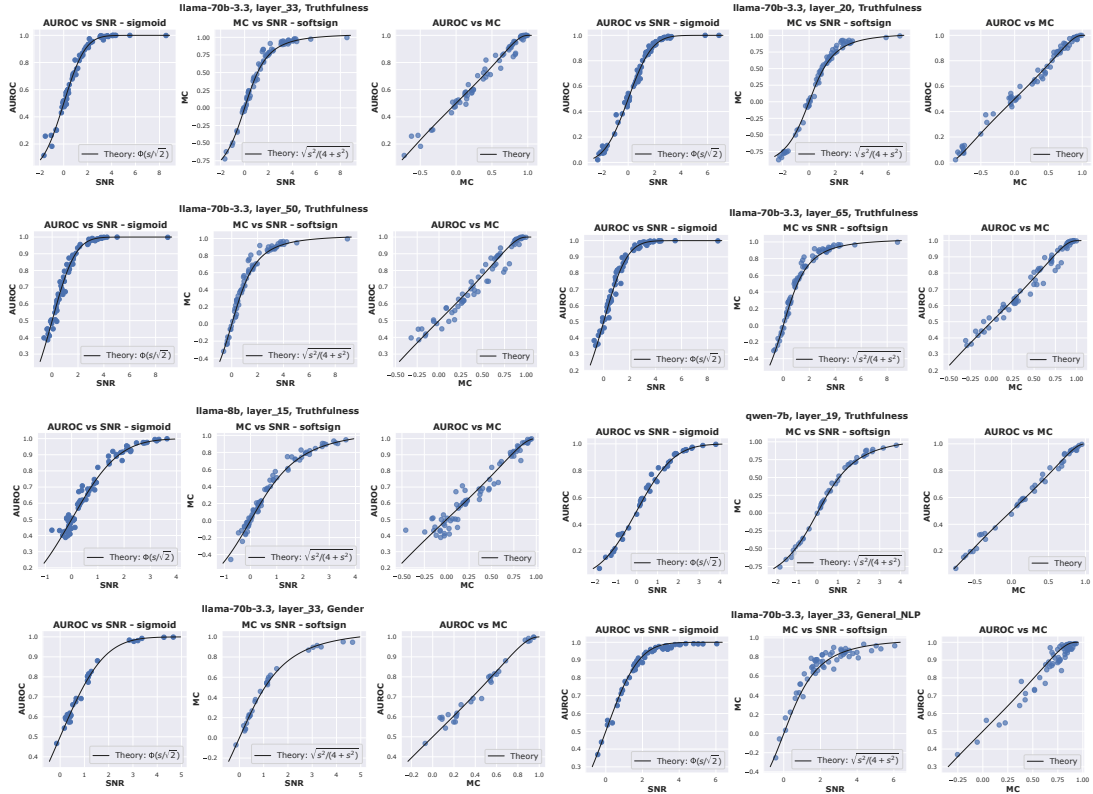


Figure 6: **Empirical verification of the theory.** The theory predicts the empirical data well across conditions.

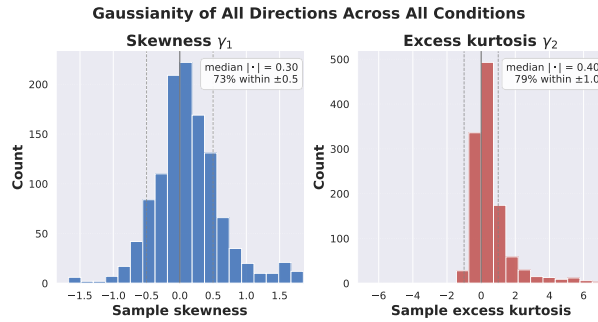


Figure 7: **Most directions are largely Gaussian across all conditions.** Across all conditions, all probe directions, and all test data distributions, most samples are largely Gaussian. Some samples have notably high kurtosis, which is mostly attributed to the sycophantic lying dataset.

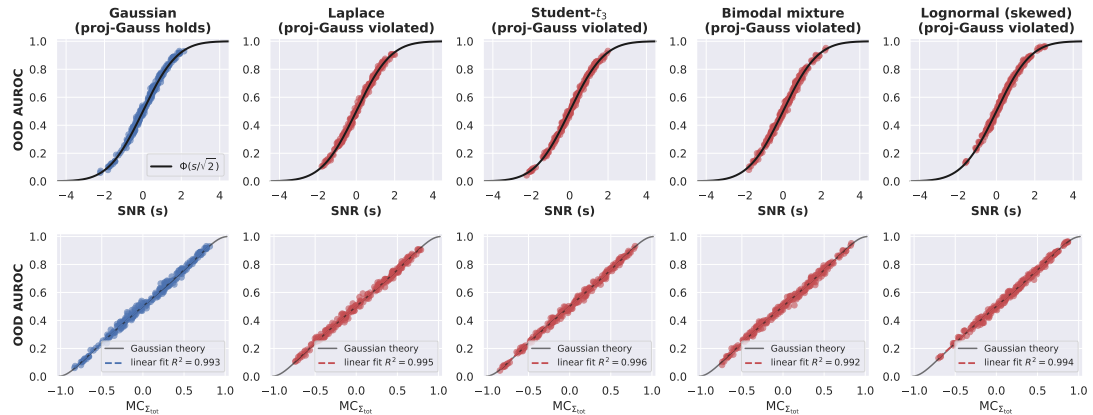


Figure 8: **The strong linearity between AUROC and MCS still holds for non-Gaussian distributions.** On deliberately constructed distributions where the projection-Gaussianity assumption is broken, the relationship between AUROC and MCS is still linear.

E Proofs of background results

E.1 Proof of Lemma 1 (within-between covariance decomposition)

By the law of total covariance,

$$\Sigma_{\text{tot}} = \mathbb{E}[\text{Cov}(X | Y)] + \text{Cov}(\mathbb{E}[X | Y]).$$

The first term equals $\frac{1}{2}\Sigma_0 + \frac{1}{2}\Sigma_1 = \Sigma_{\text{pool}}$. For the second, $\mathbb{E}[X | Y]$ takes the values μ_0 and μ_1 each with probability $\frac{1}{2}$, and $\mu_c - \bar{\mu} = \pm \frac{1}{2}\delta$ (where $\bar{\mu} := \frac{1}{2}(\mu_0 + \mu_1)$), so

$$\text{Cov}(\mathbb{E}[X | Y]) = \frac{1}{2} \cdot \frac{1}{4}\delta\delta^\top + \frac{1}{2} \cdot \frac{1}{4}\delta\delta^\top = \frac{1}{4}\delta\delta^\top.$$

□

E.2 Proof of Lemma 2 (binormal AUROC)

By definition, $\text{AUROC}(w) = \Pr(w^\top X_1 > w^\top X_0)$ where X_c is drawn from the class- c conditional. By Assumption 2, $w^\top X_c \sim \mathcal{N}(w^\top \mu_c, \sigma_c^2(w))$ with $\sigma_c^2(w) := w^\top \Sigma_c w$, and $\sigma_0^2(w) + \sigma_1^2(w) = 2w^\top \Sigma_{\text{pool}} w$. Given the standard assumption of the independence of X_0, X_1 , we then have:

$$\begin{aligned} w^\top X_1 - w^\top X_0 &\sim \mathcal{N}(w^\top \delta, \sigma_0^2(w) + \sigma_1^2(w)) \\ &= \mathcal{N}(w^\top \delta, 2w^\top \Sigma_{\text{pool}} w), \end{aligned}$$

hence

$$\begin{aligned} \text{AUROC}(w) &= \Phi(w^\top \delta / \sqrt{2w^\top \Sigma_{\text{pool}} w}) \\ &= \Phi(\text{SNR}(w) / \sqrt{2}). \end{aligned}$$

□

E.3 Proof of Lemma 3 (Fisher's discriminant)

$\text{SNR}^2(w) = (w^\top \delta)^2 / (w^\top \Sigma_{\text{pool}} w)$ is scale-invariant, so we may restrict to $\{w : w^\top \Sigma_{\text{pool}} w = 1\}$ and maximise $(w^\top \delta)^2$. Substituting $u := \Sigma_{\text{pool}}^{1/2} w$ and $q := \Sigma_{\text{pool}}^{-1/2} \delta$ converts the constraint to $\|u\| = 1$ and the objective to $(u^\top q)^2$. Cauchy-Schwarz gives a maximum of $\|q\|^2 = \delta^\top \Sigma_{\text{pool}}^{-1} \delta = z_{\text{max}}^2$, attained at $u \propto q$, i.e., $w \propto \Sigma_{\text{pool}}^{-1/2} q = \Sigma_{\text{pool}}^{-1} \delta = w_{\text{ood}}$. □

F Proof of Theorem 1

Write $v := w_{\text{id}}^\top \Sigma_{\text{pool}} w_{\text{id}} > 0$, so that $w_{\text{id}}^\top \delta = s\sqrt{v}$. Using Lemma 1 we have $\Sigma_{\text{tot}} = \Sigma_{\text{pool}} + \frac{1}{4}\delta\delta^\top$, and since $w_{\text{ood}} = \Sigma_{\text{pool}}^{-1} \delta$,

$$\Sigma_{\text{pool}} w_{\text{ood}} = \delta, \delta^\top w_{\text{ood}} = \delta^\top \Sigma_{\text{pool}}^{-1} \delta = z_{\text{max}}^2.$$

We compute the three quadratic forms in (2).

Cross term.

$$\begin{aligned} w_{\text{id}}^\top \Sigma_{\text{tot}} w_{\text{ood}} &= w_{\text{id}}^\top \Sigma_{\text{pool}} w_{\text{ood}} + \frac{1}{4} w_{\text{id}}^\top \delta \delta^\top w_{\text{ood}} \\ &= w_{\text{id}}^\top \delta + \frac{1}{4} (w_{\text{id}}^\top \delta) z_{\text{max}}^2 \\ &= s\sqrt{v} (1 + \frac{1}{4} z_{\text{max}}^2). \end{aligned}$$

Self-norm of w_{id} .

$$\begin{aligned} w_{\text{id}}^\top \Sigma_{\text{tot}} w_{\text{id}} &= v + \frac{1}{4} (w_{\text{id}}^\top \delta)^2 \\ &= v + \frac{1}{4} s^2 v \\ &= v(1 + \frac{1}{4} s^2). \end{aligned}$$

Self-norm of w_{ood} .

$$\begin{aligned} w_{\text{ood}}^\top \Sigma_{\text{tot}} w_{\text{ood}} &= z_{\text{max}}^2 + \frac{1}{4} z_{\text{max}}^4 \\ &= z_{\text{max}}^2 (1 + \frac{1}{4} z_{\text{max}}^2). \end{aligned}$$

Substituting into (2),

$$\begin{aligned} \text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}}) &= \frac{s\sqrt{v} (1 + \frac{1}{4} z_{\text{max}}^2)}{\sqrt{v(1 + \frac{1}{4} s^2)} \cdot \sqrt{z_{\text{max}}^2 (1 + \frac{1}{4} z_{\text{max}}^2)}} \\ &= \frac{s}{z_{\text{max}}} \sqrt{\frac{1 + \frac{1}{4} z_{\text{max}}^2}{1 + \frac{1}{4} s^2}}. \end{aligned}$$

Oddness, monotonicity, and the endpoint values $\text{MCS}_{\Sigma_{\text{tot}}}(\pm z_{\text{max}}) = \pm 1$ follow by inspection. Differentiating $s/\sqrt{1 + s^2/4}$ in s yields

$$\frac{d \text{MCS}_{\Sigma_{\text{tot}}}}{ds} = \frac{\sqrt{1 + \frac{1}{4} z_{\text{max}}^2}}{z_{\text{max}}} \cdot (1 + \frac{1}{4} s^2)^{-3/2}, \quad (8)$$

strictly positive on $(-z_{\text{max}}, z_{\text{max}})$. □

G Slope along the AUROC-MCS curve

We derive the slope formula (7) and prove the bound on the slope along the curve. All facts here follow from Lemma 2 and Theorem 1 with no new assumptions.

Derivative of $\text{MCS}_{\Sigma_{\text{tot}}}$ in s . Let $g(s) := \text{MCS}_{\Sigma_{\text{tot}}}(s) = (s/z_{\text{max}}) \sqrt{(1 + \alpha)/(1 + s^2/4)}$ from (5), with $\alpha = z_{\text{max}}^2/4$. A direct calculation gives

$$g'(s) = \frac{\sqrt{1 + \alpha}}{z_{\text{max}}} \cdot \frac{1}{(1 + s^2/4)^{3/2}} > 0, \quad (9)$$

so g is strictly increasing on $(-z_{\text{max}}, z_{\text{max}})$.

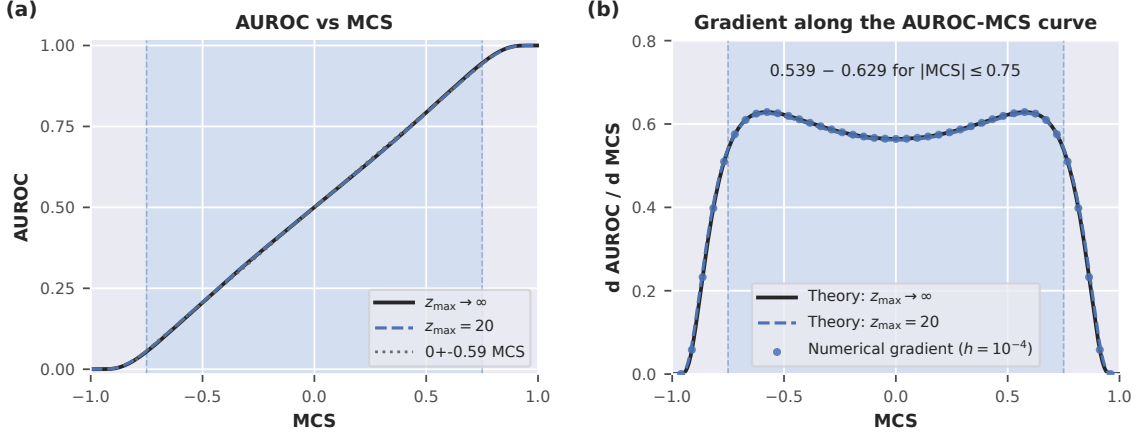


Figure 9: **Local slope of the theoretical AUROC-MCS relationship.** (a) The theoretical AUROC-MCS curves for ($z_{\max} \rightarrow \infty$) and ($z_{\max} = 20$). The shaded region is the region that is approximately linear. (b) The analytic local slope and its centered finite-difference evaluation ($h = 10^{-4}$). The shaded region marks ($|\text{MCS}| \leq 0.75$), where the limiting slope remains between (0.539) and (0.629). Outside this region, the slope decreases as AUROC saturates.

Slope formula (7). From Lemma 2, $\text{AUROC}(s) = \Phi(s/\sqrt{2})$, so $d \text{ AUROC} / ds = \phi(s/\sqrt{2})/\sqrt{2}$. Combining with (9),

$$\begin{aligned} \frac{d \text{ AUROC}}{d \text{ MCS}_{\Sigma_{\text{tot}}}} &= \frac{d \text{ AUROC} / ds}{g'(s)} \\ &= \frac{\phi(s/\sqrt{2})/\sqrt{2}}{(\sqrt{1 + \alpha/z_{\max}})(1 + s^2/4)^{-3/2}}, \end{aligned}$$

which simplifies to (7).

Factorization into task and shape factors. The slope (7) factors as

$$\frac{d \text{ AUROC}}{d \text{ MCS}_{\Sigma_{\text{tot}}}} = h(s) \cdot g(z_{\max}),$$

where $h(s) = \phi(s/\sqrt{2})(1 + s^2/4)^{3/2}$ depends only on s (task-independent), and $g(z_{\max}) = 1/\sqrt{2/z_{\max}^2 + 1/2}$ depends only on $z_{\max}$ (independent of s). This decouples the two questions of (i) how flat the curve is within a task and (ii) how much the curve varies across tasks.

Task factor $g(z_{\max})$ saturates fast. g is strictly increasing in $z_{\max}$ with limit $g(\infty) = \sqrt{2}$. Writing $g(z_{\max})/g(\infty) = 1/\sqrt{1 + 4/z_{\max}^2}$, the deviation from the limit is at most 0.5% for $z_{\max} \geq 20$ and falls off as $z_{\max}^{-2}$:

$z_{\max}$	4	6	8	20
$g(z_{\max})/g(\infty)$	0.894	0.949	0.970	0.995

Every empirical task has $z_{\max} > 20$ (Tab. 3), so each task’s AUROC-MCS curve coincides with the $z_{\max} \rightarrow \infty$ limit curve to within 0.5% on the slope.

Shape factor $h(s)$ and central slope. $h(s)$ combines $\phi(s/\sqrt{2})$ (strictly decreasing in $|s|$) with $(1 + s^2/4)^{3/2}$ (strictly increasing in $|s|$). Their product is far flatter than either factor alone — this is the saturation cancellation behind (i): as $|s| \rightarrow z_{\max}$, the AUROC sigmoid saturates (shrinking ϕ) at the same time as the MCS softsign saturates (growing $(1 + s^2/4)^{3/2}$), and the two effects largely offset. At the centre, $h(0) = 1/\sqrt{2\pi}$, giving the universal central slope

$$h(0) \cdot g(\infty) = \frac{\sqrt{2}}{\sqrt{2\pi}} = \frac{1}{\sqrt{\pi}} \approx 0.564.$$

The cancellation is not exact: $h(s) \cdot g(\infty)$ peaks at $s = \pm\sqrt{2}$ with value ≈ 0.629 , then decays toward 0 as $|s| \rightarrow z_{\max}$, which is why $R^2 < 1$ in the empirical AUROC-MCS fits.

Numerical Evaluation of Slope Stability. As shown in Fig. 9, to quantify the near-linearity of the theoretical relationship, we evaluated its local gradient and independently verified it using centered finite differences. In the large- $(z_{\max})$ limit, the slope remains between (0.539) and (0.629) for ($|\text{MCS}| \leq 0.75$), corresponding to a maximum deviation of (7.7%) from its midrange value (0.584); the ($z_{\max} = 20$) curve is nearly identical. Moreover, $(0.5+0.59\text{MCS})$ approximates the limiting curve with maximum absolute AUROC error (0.0053) over ($|\text{MCS}| \leq 0.8$). The slope falls near ($|\text{MCS}| = 1$) because AUROC saturates, so this calculation supports range-qualified near-linearity rather than exact global linearity.

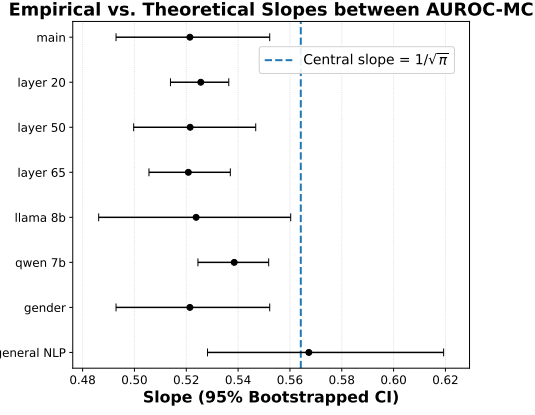


Figure 10: **Empirical slopes are near the theoretical central slope.** Linear-fit slope of OOD AUROC against $\text{MCS}_{\Sigma_{\text{tot}}}$, with 95% bootstrap CIs, across all eight conditions. The dashed line marks the universal central slope $1/\sqrt{\pi} \approx 0.564$ predicted by the theory (App. G). Empirical estimates are close to this value, with most of them slightly below it. This is consistent with data sampling some of the saturation tail where the local slope decays, dragging down the overall slope.

Empirical slopes are consistent with theory. As shown in Fig. 10, the empirically fitted slopes across 8 conditions are consistent with the theoretical prediction.

H Failure mode details

This appendix collects the formal statements behind the failure modes summarised in §5.

H.1 Use Σ_{pool} metric instead of Σ_{tot}

We have:

$$w_{\text{id}}^{\top} \Sigma_{\text{pool}} w_{\text{ood}} = w_{\text{id}}^{\top} \delta = s\sqrt{v}; \quad w_{\text{id}}^{\top} \Sigma_{\text{pool}} w_{\text{id}} = v$$

and

$$w_{\text{ood}}^{\top} \Sigma_{\text{pool}} w_{\text{ood}} = \delta^{\top} \Sigma_{\text{pool}}^{-1} \delta = z_{\text{max}}^2$$

Substituting into (2) gives $s\sqrt{v}/(\sqrt{v} \cdot z_{\text{max}}) = s/z_{\text{max}}$. Therefore, $\text{MCS}_{\Sigma_{\text{pool}}}$ is just a rescaled s . This will not cancel out the S-shaped between OOD AUROC and s , and thus will not be linear to OOD AUROC. $\square$

H.2 Small Fisher distance

When z_{max} is small ($z_{\text{max}} \lesssim 2$), the slope around 0 $\kappa(z_{\text{max}})$ is far below its limit and varies rapidly with z_{max} : $\kappa(1) \approx 0.25$ vs. $\kappa(2) \approx 0.40$ vs. $\kappa(3) \approx 0.47$. A collection of test sets with heterogeneous small- z_{max} tasks therefore does not collapse onto a single linear fit; each task has its own

slope, producing a fan rather than a line. On synthetic data with z_{max} controlled to $\{0.1, 0.5, 1, 2\}$ (Fig. 3c), we observe exactly this fan, and the global linear R^2 drops to 0.666. This failure mode is not engaged on our LLM datasets because every task has $z_{\text{max}} > 20$, inside the saturation regime (see Tab. 3; App. C).

H.3 Strong class imbalance

For class prior π , we redefine the pooled variance as:

$$\Sigma_{\text{pool}} := (1 - \pi)\Sigma_0 + \pi\Sigma_1$$

Then Lemma 1 generalises to

$$\Sigma_{\text{tot}} = \Sigma_{\text{pool}} + \pi(1 - \pi)\delta\delta^{\top}$$

, and Theorem 1 becomes

$$\text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}}) = \frac{s}{z_{\text{max}}} \sqrt{\frac{1 + \pi(1 - \pi)z_{\text{max}}^2}{1 + \pi(1 - \pi)s^2}},$$

preserving the softsign-cancels-sigmoid story but with a re-scaled saturation point and slope. Strong imbalance (e.g., $\pi = 0.01$) shrinks the $\pi(1 - \pi)$ factor by $25\times$ versus the balanced case, weakening the saturation cancellation in MCS and steepening the AUROC–MCS slope. Our probe datasets are all approximately balanced, so this mode is not engaged in our experiments, but it would matter for naturally imbalanced applications such as rare-event detection.

I Empirical alignment of LR and LDA reference directions

The theory in §4 uses the Fisher direction $w_{\text{ood}}^{\text{LDA}} = \Sigma_{\text{pool}}^{-1}\delta$ as the OOD reference. In the empirical experiments, we replace it with the OOD-trained logistic-regression direction $w_{\text{ood}}^{\text{LR}}$, on the grounds that for balanced binary tasks with well-separated classes the two directions are nearly proportional. This appendix quantifies the substitution.

Setup. For each of datasets used in §3, on the train half of a stratified 50/50 split, we compute (i) the Fisher direction $w_{\text{ood}}^{\text{LDA}} = (\Sigma_{\text{pool}} + \lambda I)^{-1}\delta$ with $\lambda = 10^{-6}$, and (ii) the LR direction $w_{\text{ood}}^{\text{LR}}$. We measure the correlation between $\text{MC}(w_{\text{id}}^{\text{LR}}, w_{\text{ood}}^{\text{LR}})$ and $\text{MC}(w_{\text{id}}^{\text{LR}}, w_{\text{ood}}^{\text{LDA}})$.

Result. As shown in Fig. 11, across all settings (across models, layers, and concept domains), the pairwise Pearson correlation between $\text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}}^{\text{LR}})$ and $\text{MCS}_{\Sigma_{\text{tot}}}(w_{\text{id}}, w_{\text{ood}}^{\text{LDA}})$ is

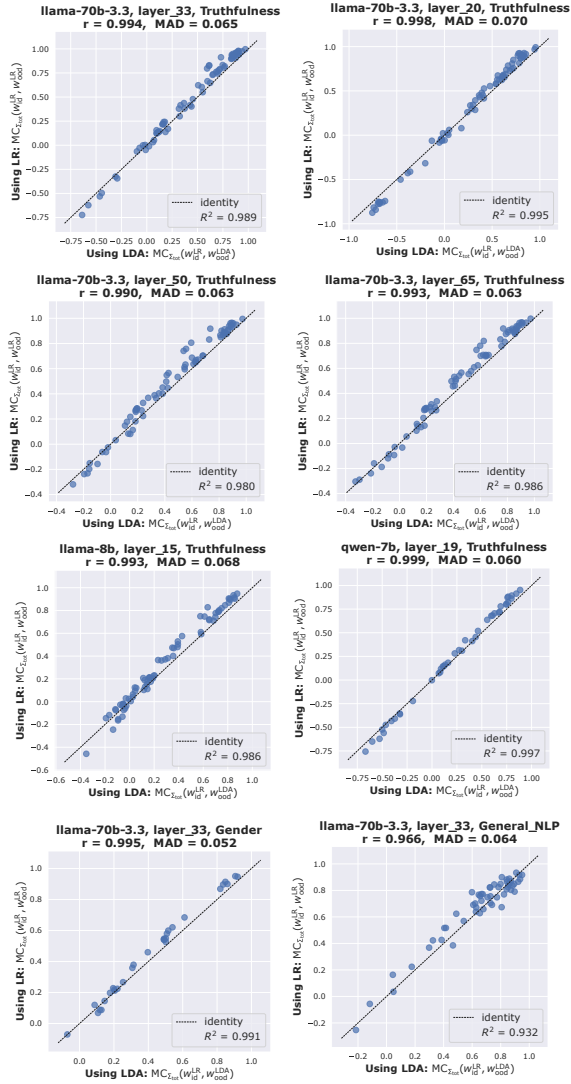


Figure 11: $MC(w_{id}^{LR}, w_{ood}^{LR})$ vs. $MC(w_{id}^{LR}, w_{ood}^{LDA})$. The two MCs correlate strongly across all eight conditions, explaining why substituting the LDA direction with the LR direction in our empirical experiments does not affect the observed linearity.

$r > 0.99$, with mean $|\Delta(w_{id})| < 0.07$. Substituting w_{ood}^{LR} for w_{ood}^{LDA} in the headline regression of §3 changes the linear-fit R^2 by less than 1%.

Why this works. For balanced binary classification with classes that are jointly well-modelled by Gaussians with shared covariance, the LR maximum-likelihood direction and the Fisher direction coincide up to scaling (Efron, 1975; Hastie et al., 2009). Real activations only approximately satisfy these assumptions, but all of our probe datasets are easy — ID AUROC exceeds 0.95 in every case — so the residual mismatch between w_{ood}^{LR} and w_{ood}^{LDA} is small in cosine. We expect a larger mismatch on harder, less-separable tasks; characterising that regime is a useful direction for future work.