A REFERENCE ALONE IS NOT A BIAS ESTIMATOR: IDENTIFIABILITY IN LOGIT CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Blank images and content-free text are often queried to expose a model's class preference, and their logits are then subtracted from ordinary predictions. We ask when such a reference identifies a target-optimal correction. For affine readouts, reference subtraction is exactly geometric recentering, but the induced zero point need not match the optimum selected by a target risk. We prove two complementary non-identifiability results: off-support responses are unconstrained by in-domain behavior, and fixed responses cannot track optima that move with the target distribution. We also give a sufficient condition for identification. Frozen-checkpoint audits in long-tailed semi-supervised learning, including a three-seed replication, show why references can nevertheless help: heterogeneous responses share a training-induced prior-like direction, while their scale and orthogonal content are unreliable. Controlled experiments with two language-model families reproduce the target-risk gap under prior and domain interventions. References are useful inductive probes, but target-optimal estimation requires structure or target information.

1 INTRODUCTION

Many prediction biases can be expressed as class-dependent logit offsets. When the relevant class prior is known, logit adjustment follows from a change of measure and has a population interpretation (Menon et al., 2021). When the target distribution is unknown, a tempting alternative is to query the model on an input believed to contain no class information, interpret the resulting logits as bias, and subtract them from ordinary predictions.

This pattern appears in otherwise distant settings. CDMAD uses a solid-color image in long-tailed semi-supervised learning (LTSSL) (Lee & Kim, 2024); LCGC analyzes a black baseline during training (Xing et al., 2025); and FARAD uses frequency-informed random-phase references (Gu et al., 2025). Contextual calibration instead queries a language model with content-free text such as "N/A" and forces the resulting label distribution toward uniformity (Zhao et al., 2021). All share an intuitive premise: a semantically empty input should expose only the model's prior preference.

The premise conflates two objects. A reference response is a property of a model, prompt, and chosen input; a desirable correction also depends on a target distribution and risk. Semantic emptiness does not supply that missing target information. We therefore ask: *when does a reference response identify the additive correction required by a target risk?*

We separate geometry from estimation. Reference subtraction exactly maps the reference to zero in centered-logit space and translates the affine decision boundaries. This identity specifies the operation, not the correct zero point. We prove two complementary limitations: an off-support response is not determined by in-domain behavior, and a fixed response cannot identify a risk-optimal offset that changes strictly with the target prior. The two limitations are logically distinct. We then state a sufficient structural condition under which a reference does recover the desired correction.

The limits do not make references useless. Across nine frozen LTSSL heads, many responses share a prior-like direction. Nonnegative prior projection improves 176/180 heterogeneous model–reference pairs and outperforms full-reference correction in 113/120 seed-replication and 119/120 ACR independent-run pairs. The reference does not identify the remaining scale, and a global offset

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cannot add input-dependent class-separating evidence. This explains the gains from frozen reference correction and the boundary of what its global translation can change.

The same identification gap appears beyond images. With Qwen3.5-9B and Llama-3.1-8B, the risk-optimal offset moves under exactly weighted target-prior interventions while every fixed-reference correction remains constant. A controlled SST-2-to-IMDb transfer shows that the optimum can also move at a fixed balanced prior.

Our contributions are:

- We formulate reference correction as geometric recentering, prove two distinct non-identifiability results, and state a sufficient structural condition under which a reference does recover the desired correction.
- We explain why reference correction can still work in LTSSL by localizing its useful component to a prior-like direction, confirming the effect across seeds, and separating direction from reference-specific content, scale, and objective.
- We validate the target-risk gap in a second modality and two language model families, distinguishing our claim from reference-sensitivity alone.

2 Reference Correction as Geometric Recentering

2.1 Additive correction and its target

Let $f_\theta(x) \in \mathbb{R}^C$ denote class logits. A reference-based method queries an input a and applies

$$r(a) = \mathbf{P} f_\theta(a), \qquad f_{\mathrm{corr}}(x) = f_\theta(x) - \lambda r(a), \tag{1}$$

where

$$\mathbf{P} = \mathbf{I} - \frac{1}{C}\mathbf{1}\mathbf{1}^\top \tag{2}$$

removes the softmax-null common-logit direction. In an autoregressive language model, f_θ can be the logits of the candidate label tokens under a fixed prompt, and a the test-content field of that prompt.

Additive correction is well motivated when its target is specified. If train and test differ only through class priors π and ρ, then

$$\log p_{\mathrm{te}}(y \mid x) = \log p_{\mathrm{tr}}(y \mid x) + \log \rho_y - \log \pi_y + c(x), \tag{3}$$

where $c(x)$ is class independent. Thus a known prior change produces a known centered logit offset. When the target prior is unknown, label-shift methods estimate it from target predictions or unlabeled target samples under explicit shift assumptions (Tasche, 2017; Lipton et al., 2018; Alexandari et al., 2020). Eq. (3) is a population statement: a minibatch histogram is only noisy evidence about the prior. Reference correction instead substitutes a model response for the distributional quantities, raising the identification question studied below.

2.2 An exact geometric identity

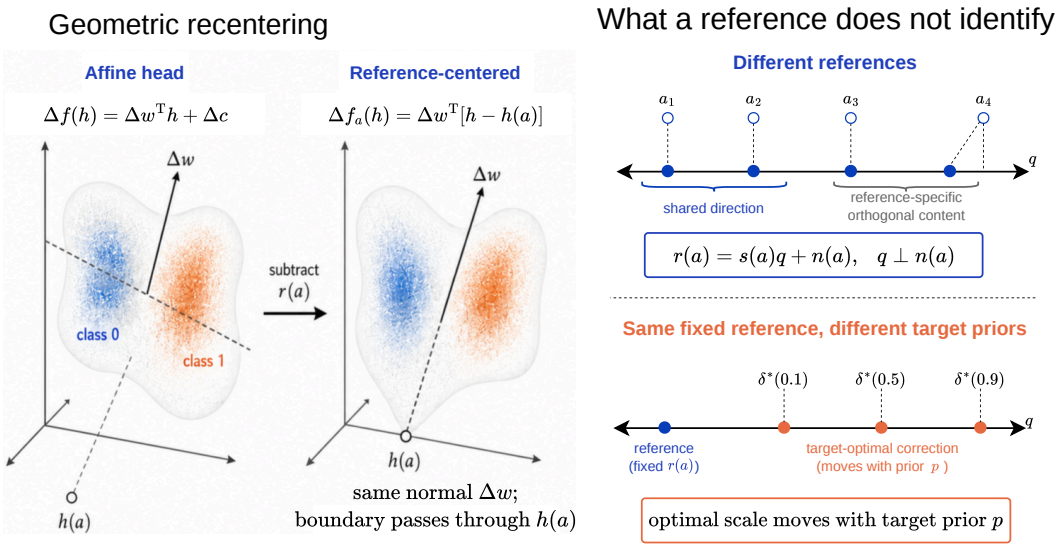
Suppose the readout is affine, $f_\theta(x) = W h_\theta(x) + c$. Then

$$r(a) = \underbrace{\mathbf{P}c}_{\text{head offset}} + \underbrace{\mathbf{P}W h_\theta(a)}_{\text{reference feature term}}, \tag{4}$$

and, for unit scale, the exact identity in centered logit space is

$$\mathbf{P} f_{\mathrm{corr}}(x) = \mathbf{P}\big(f_\theta(x) - f_\theta(a)\big) = \mathbf{P}W\big(h_\theta(x) - h_\theta(a)\big). \tag{5}$$

Reference subtraction therefore maps the reference to zero under the centered readout, $\mathbf{P} f_{\mathrm{corr}}(a) = 0$, and translates every affine decision boundary. Equivalently, it expresses centered logits through the feature displacement $h_\theta(x) - h_\theta(a)$. The identity depends only on the affine readout; it applies equally to a vision classifier and to selected label-token logits from a language model. It does *not*



(a) Reference subtraction translates pairwise boundaries. (b) The target risk selects the desired correction.

Figure 1: **Geometric action and missing information (schematic).** (a) For every class pair, subtraction preserves the normal Δw but replaces the affine intercept with the constraint $\Delta f_a(h(a)) = 0$, forcing the centered boundary through $h(a)$. (b) References may share a useful projected direction, while their orthogonal content is reference specific and the risk-optimal scale moves with the target. Geometry specifies the action, not the desired target.

make $h_\theta(a)$ a unique feature-space origin, a data centroid, a semantic null, or the zero point preferred by a target risk.

For any unit direction q, write

$$r(a) = s(a)q + n(a), \qquad q^\top n(a) = 0. \tag{6}$$

This separates three questions that raw subtraction merges: is q a useful correction direction, does $s(a)$ have the right magnitude, and is the reference-specific component $n(a)$ harmless?

Figure 1 summarizes the distinction. Recentering is an exact geometric operation, but neither the chosen zero point nor its projection contains the target-risk argument that selects a statistical optimum.

3 What a Reference Cannot Identify

3.1 Off-support behavior is unconstrained

Let $\mathcal{S} \subset \mathcal{X}$ be the support on which training and evaluation constrain the classifier, and let $a \notin \mathcal{S}$.

Proposition 1 (off-support non-identifiability). Assume the classifier family is closed under an additive output branch $g(x)v$, where $g(x) = 0$ for $x \in \mathcal{S}$, $g(a) = 1$, and $v \in \mathbb{R}^C$ is centered. Then, for any classifier f and any such v, there exists $\tilde{f}$ in the same family such that

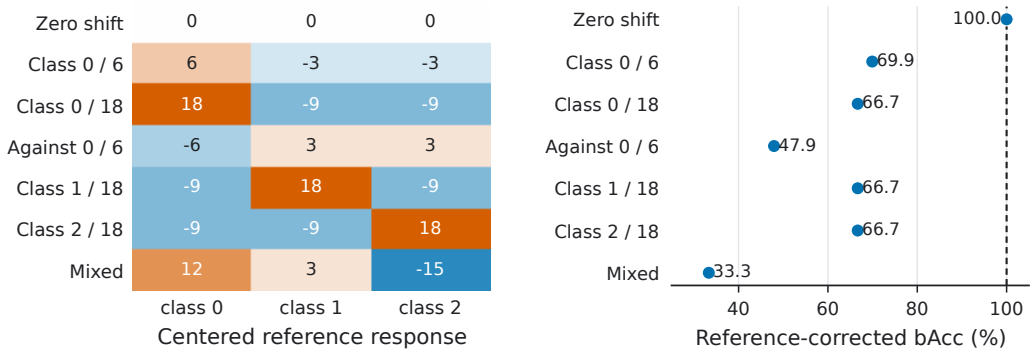
$$\tilde{f}(x) = f(x) \quad \forall x \in \mathcal{S}, \qquad \mathbf{P}\tilde{f}(a) = \mathbf{P}f(a) + v. \tag{7}$$

Proof. Set $\tilde{f}(x) = f(x) + g(x)v$. The added term is zero on $\mathcal{S}$, so all in-domain logits, losses, and predictions are unchanged. At a, centering adds v. □

This is an information statement, not a claim about which interpolant ordinary training selects. Smoothness, normalization, architecture, or minimum-norm bias may correlate an off-support response with in-domain behavior, but each is an additional assumption. Human semantic neutrality

alone imposes none. The construction is a correction-specific instance of underspecification: models that are indistinguishable on the training domain can disagree on behavior not constrained by that domain (D'Amour et al., 2022).

Figure 2 gives a finite construction. The one-ReLU gate specified in Appendix B is zero on all 9,000 data points and one at the reference. The family has identical in-domain logits and 100% raw balanced accuracy, while its centered reference response and resulting corrected accuracy vary widely.



(a) Off-support responses vary freely.

(b) Correction outcomes vary accordingly.

Figure 2: **In-domain behavior does not identify an off-support correction.** Seven models have exactly identical logits on all 9,000 data points and raw bAcc 100%. A gate active only at the reference assigns the centered responses in (a); subtracting them produces the 33.3%–100% bAcc span in (b).

3.2 A FIXED REFERENCE CANNOT IDENTIFY A TARGET-RISK OPTIMUM

Proposition 1 concerns the source of a reference response. A second problem remains even if that response is perfectly reproducible. Consider binary classification with fixed class-1 log-odds $s(x)$ and an additive decision offset δ. Let Q_0 and Q_1 be the class-conditional target distributions and let $p = \Pr(Y = 1)$. The population binary cross-entropy is

$$\begin{aligned} R_p(\delta) =& (1-p)\,\mathbb{E}_{Q_0}\left[\log(1+e^{s(X)+\delta})\right] \\ &+ p\,\mathbb{E}_{Q_1}\left[\log(1+e^{-s(X)-\delta})\right]. \end{aligned} \tag{8}$$

Proposition 2 (target-risk non-identifiability). Assume $p \in (0,1)$ and $\mathbb{E}_{Q_y}|s(X)| < \infty$ for $y \in \{0,1\}$. Then $R_p(\delta)$ is strictly convex and has a unique minimizer $\delta^*(p)$. Moreover, $\delta^*(p)$ is strictly increasing in p. Consequently, any correction computed solely from a fixed model, prompt, and reference, and therefore invariant to p, can equal $\delta^*(p)$ for at most one target prior.

Proof sketch. Let σ denote the logistic sigmoid. The first-order condition is

$$g(\delta, p) = (1-p)\mathbb{E}_{Q_0}[\sigma(s+\delta)] + p\mathbb{E}_{Q_1}[\sigma(s+\delta) - 1] = 0. \tag{9}$$

We have $\partial g/\partial \delta > 0$ and

$$\frac{\partial g}{\partial p} = -\mathbb{E}_{Q_0}[\sigma(s+\delta)] - \mathbb{E}_{Q_1}[1-\sigma(s+\delta)] < 0. \tag{10}$$

Implicit differentiation gives $d\delta^*/dp > 0$. Existence, uniqueness, and details appear in Appendix A. □

Thus the target prior, or equivalent target information, must enter somewhere. At fixed p, changing Q_0, Q_1, or the loss can likewise move the optimum without changing the reference. We call this mismatch the *target-risk gap*. Unlike Proposition 1, it does not require the reference to be off-support. Appendix A gives a multi-class cross-entropy corollary under label shift; other losses require their own conditions.

3.3 A SUFFICIENT CONDITION MAKES THE MISSING STRUCTURE EXPLICIT

The propositions rule out identification from a reference alone, not every reference-based estimator under additional structure. Fix a target distribution T and cross-entropy risk, and let $\eta_T(x) = \Pr_T(Y \mid X = x)$ have strictly positive coordinates almost surely.

Proposition 3 (structural sufficiency). Suppose that on the target support the centered logits decompose as

$$\mathbf{P} f_\theta(x) = \mathbf{P} \log \eta_T(x) + b, \qquad \mathbf{1}^\top b = 0, \tag{11}$$

for an input-independent nuisance vector b. If a reference is guaranteed to satisfy $r(a) = b$, then unit-scale reference subtraction recovers target-calibrated probabilities and minimizes population cross-entropy.

Proof. Subtracting $r(a) = b$ from Eq. (11) gives $\mathbf{P} f_{\mathrm{corr}}(x) = \mathbf{P} \log \eta_T(x)$. Softmax is invariant to the removed common-logit component, so $\mathrm{softmax}(f_{\mathrm{corr}}(x)) = \eta_T(x)$; the true conditional minimizes cross-entropy pointwise. □

The condition is deliberately explicit: a successful reference must isolate an input-independent nuisance relative to the specified target. Semantic emptiness alone guarantees neither the decomposition nor $r(a) = b$. It also clarifies what a practical estimator must establish: that the proposed direction is target relevant, that reference-specific components are harmless, and that the magnitude matches the declared risk. None of these properties can be certified from the reference response itself. The experiments below therefore audit direction, off-direction content, scale, and objective separately rather than treating reference accuracy as evidence for a single latent "bias."

4 WHY REFERENCE CORRECTION CAN STILL WORK IN LTSSL

The propositions rule out an assumption-free interpretation, not useful training-induced regularities. We next ask which component of visual reference correction is repeatable and what information it lacks.

4.1 FROZEN-CHECKPOINT AUDIT

We audit nine frozen affine heads spanning CIFAR-10-LT, CIFAR-100-LT, and STL-10: CDMAD-style and original CDMAD models, plus the standard and balanced heads of ACR and CCL (Wei & Gan, 2023; Zhou et al., 2024b). A fixed suite of 20 references, including colors, data and class means, examples, frequency controls, noise, and random-phase variants, yields 180 model–reference pairs. Diagnostic corrections are selected on one split; every ranking and accuracy is measured on a disjoint holdout. A focused replication evaluates three seeds in two CIFAR-10-LT unlabeled-imbalance settings (C10-u10 and C10-u50), yielding six frozen checkpoints and 120 pairs under the same protocol. Every comparison intervenes only on stored logits while keeping the representation and classifier fixed. Improvements therefore isolate the additive correction rather than retraining, representation learning, or altered pseudo-label trajectories.

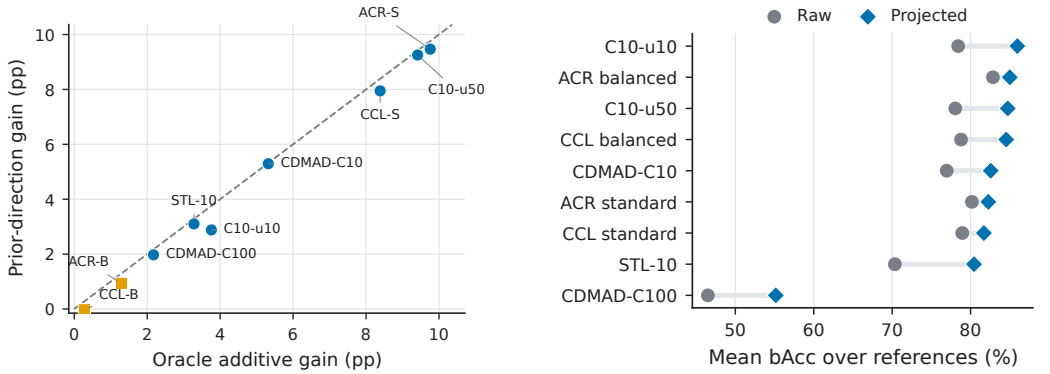
Information sources. Raw subtraction uses $r(a)$; nonnegative prior projection uses $q_\pi \max\{0, q_\pi^\top r(a)\}$ but no diagnostic or test labels. Negative coefficients return the raw head; permutation controls use the same constraint. Label-assisted diagnostics measure headroom, whereas reserved-support scale estimation uses labeled data excluded from training.

4.2 A SHARED DIRECTION EXPLAINS THE REPEATABLE GAIN

Let q_π be the normalized centered labeled-prior direction,

$$q_\pi = \frac{\mathbf{P} \log \hat{\pi}_l}{\|\mathbf{P} \log \hat{\pi}_l\|_2}. \tag{12}$$

Fitting only a nonnegative scalar along q_π nearly matches a fully unconstrained additive diagnostic across all nine heads (Figure 3a): the largest gap is 0.88 balanced-accuracy points and the median is



(a) A prior-like direction captures most additive headroom.

(b) Nonnegative prior projection improves correction.

Figure 3: **Direction explains the repeatable component of visual reference correction.** (a) Label-assisted diagnostics span nine heads; S/B denote standard/balanced ACR and CCL heads, with squares marking balanced heads. (b) Labeled-prior projection uses no diagnostic or test labels and improves 176/180 model–reference pairs.

Table 1: Controlled tests of visual reference-validity explanations. Target corrections use diagnostic labels; outcomes use disjoint held-out data.

Hypothesis	Test	Result
Correction-space match	Within-model association with held-out reference performance	$r = 0.851$–0.930
Feature centrality	Association of negative feature-center distance with scaled performance	$r = -0.032$
Semantic absence	Six pattern-free color constants	mean range 3.50 pp
Frequency realism	Matched-RMS low vs. high frequency	low wins $18/18$ raw
Random-phase construction	Best of six fixed variants vs. gray	wins $4/9$ raw, $2/9$ projected

0.29. Thus most *input-independent additive headroom* in these checkpoints lies along one prior-like direction; this is not a claim that all bias is label shift.

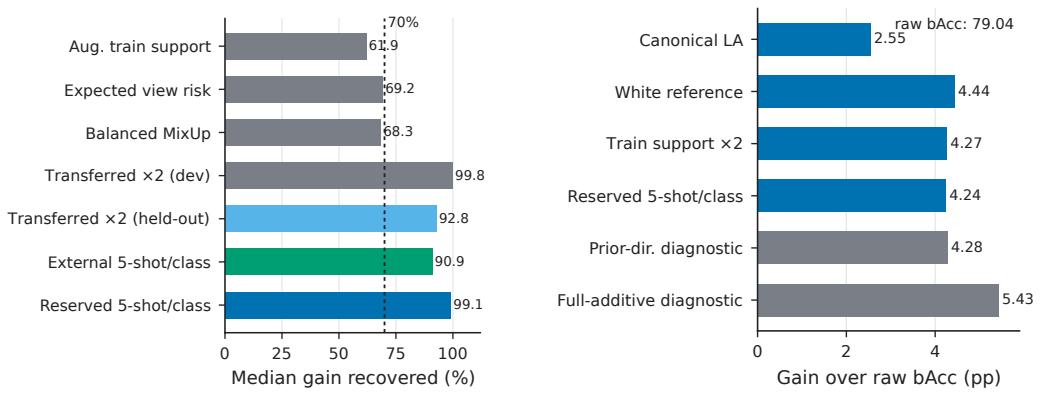
Nonnegative prior projection improves 176/180 held-out pairs relative to full-reference correction and reduces the mean within-model accuracy range across references by 87.9% (Figure 3b). Pattern-free images can therefore help despite different raw features because their classifier-projected responses share a useful component. The effect is not generic shrinkage: the true direction exceeds all 1,000 class-coordinate permutation controls in mean bAcc and pairwise wins ($p < 0.001$; Appendix D).

Both matched replications preserve this pattern. Across three training seeds, projection beats 113/120 full responses, with gains over raw of 1.93 ± 0.16 and 8.02 ± 0.39 bAcc for C10-u10 and C10-u50. Across three independent ACR runs, it beats 119/120, with standard-head gains of 6.49–6.86 and balanced-head gains of 0.76–1.03. All pooled mean bAcc, gain, and win-count statistics exceed 1,000 coordinate-permutation controls ($p < 0.001$; Tables 5 and 6).

Table 1 rejects image-level substitutes for this account: feature centrality is not predictive, and semantically empty, frequency-matched, or random-phase references remain construction dependent. The relevant object is the classifier-projected response, not the semantic name of the image.

4.3 Scale and objective come from outside the reference

Projection reduces a C-dimensional response to a scalar strength, but does not identify that scalar. Strong views, expected view risk, and balanced MixUp reuse training examples already absorbed by the model and underestimate the test-time correction. In a fixed-budget test, truly unused five-shot-



(a) Dependent support systematically under-recovers scale.

(b) Independent support recovers prior-direction scale.

Figure 4: **Scale requires information beyond the reference.** (a) Increasingly independent sources recover more of the diagnostic prior-direction gain. (b) Under a fixed label budget, reserved support nearly recovers that gain; gray bars denote label-assisted diagnostics.

per-class support recovers 99.1% of the prior-direction gain, although it does not beat the favorable white reference. Independent information can therefore recover scale within the declared q_π subspace, but a global translation still cannot add input-dependent class separation (Figure 4; exact values in Appendix E).

The objective is also external to the reference. The same frozen correction improves balanced pseudo-label accuracy and tail recall, yet lowers ordinary pseudo-label accuracy in all eight audited views because harmful head-class flips outnumber corrected tail flips (Appendix F).

The audit explains CDMAD's characteristic behavior: a blank image need not measure bias to inherit a useful direction from the trained network. Its scale and off-direction content remain contingent, and no global reference can recover errors caused by missing class-separating features.

5 The Target-Risk Gap Beyond Vision

We next study contextual calibration, which subtracts label-token preferences measured on content-free inputs such as "N/A" (Zhao et al., 2021). Proposition 2 predicts that, with the model, prompt, and reference fixed, this correction cannot follow a changing target risk. Because such strings may occur in pretraining data, this audit tests Proposition 2 rather than relying on the off-support premise of Proposition 1.

5.1 Protocol

We freeze Qwen3.5-9B and Llama-3.1-8B-Instruct (Qwen Team, 2026; Grattafiori et al., 2024). On SST-2 and QNLI, zero- and few-shot prompts crossed with AB/BA label mappings yield eight conditions per model. Each uses six fixed references: blank content, "N/A", random words, repeated symbols, a neutral sentence, and an ellipsis. We form target positive-class priors $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ by exact mixture weighting. A free centered offset is fitted on one stratified half and evaluated by weighted NLL on the other; this label-assisted diagnostic is the *oracle*. Raw accuracies remain nondegenerate (78.0–93.9% across models and conditions). Appendix G gives the complete protocol.

5.2 Changing the target prior moves the optimum, not the reference

Proposition 2 establishes that the optimum must move; Figure 5 measures its magnitude in modern language models. After centering at $p = 0.5$, mean oracle trajectories move monotonically in both families while every fixed-reference trajectory is horizontal. Oracle shifts span 3.470–4.130 logits



(a) Changing the target prior moves the optimum.

(b) Changing domain at fixed prior can also move it.

Figure 5: **The target-risk gap transfers across model families.** (a) Thick lines average eight SST-2/QNLI task–prompt conditions; thin lines show individual conditions, centered at $p = 0.5$. Every fixed reference is horizontal. (b) Points are IMDb-minus-SST-2 optimal shifts at $p = 0.5$; bars are pointwise 95% intervals from 2,000 stratified bootstrap replicates.

for Qwen and 3.628–5.511 for Llama; references also span several logits, so semantic emptiness does not select a unique correction.

Table 2: Cross-model target-risk audit. "Harmful" means higher held-out NLL than no correction. ΔNLL is the median reference-minus-raw NLL at $p = .5$; regret is the median reference-minus-oracle NLL over all priors. Oracle and reference spans are class-1-vs.-class-0 log-odds.

Model	Harmful all priors	Harmful at $p = .5$	Median ΔNLL	Median regret	Oracle range	Reference span
Qwen3.5-9B	147/240	40/48	+0.032	0.056	3.47–4.13	3.11
Llama-3.1-8B	143/240	35/48	+0.059	0.096	3.63–5.51	4.65

The gap persists at the conventional balanced target: 40/48 Qwen references and 35/48 Llama references worsen NLL, with median increases of 0.032 and 0.059 (Table 2). Across all priors, median regret relative to the oracle is 0.056 and 0.096, respectively. References also lower target-weighted accuracy in 149/240 Qwen and 140/240 Llama cases (Appendix G). No reference is consistently best; all six lack the target-distribution argument required by Eq. (9).

5.3 Fixed prior, different conditional distribution

Prior is not the only target argument. At $p = 0.5$, input-length splits move the optimum significantly in 2/8 Qwen and 3/8 Llama conditions. More directly, holding the task, prompt, and balanced prior fixed while moving from SST-2 to 600 IMDb reviews yields nonzero optimal-shift changes in three of four Qwen and one of four Llama conditions (pointwise 95% stratified-bootstrap intervals). The prompt-dependent signs are consistent with dependence on Q_y, rather than a universal length or domain effect. The 2,048-token budget preserves every classification suffix.

Appendix G also verifies the affine reconstruction in Eq. (5) under BF16 inference.

6 Discussion

Design implications. Calling an offset "bias" requires a target distribution and loss. Training may align references with a useful subspace, but scale and target adaptation require independent support, a generative assumption, or target samples. Moreover, a global translation cannot add

input-dependent class separation, and an offset selected for balanced inference may worsen sample-weighted pseudo-label risk. Batch, domain-context, and generative calibration introduce the missing target argument through target-like inputs or a model of their marginal (Zhou et al., 2024a; Fei et al., 2023; Jiang et al., 2023); label-shift and test-agnostic long-tail methods make the same dependence explicit (Lipton et al., 2018; Alexandari et al., 2020; Wei et al., 2024). This yields a practical audit for reference-based methods: specify the target risk, test whether the reference supplies a repeatable direction, estimate its scale from information independent of that reference, and verify that discarded or retained components help on held-out targets. Stability across references answers only the second question; it cannot establish target optimality.

How the evidence fits together. The propositions separate off-support underdetermination, target-risk dependence, and one sufficient structure. The frozen LTSSL audit localizes repeatable gains to a prior-like direction but not its scale or pseudo-label objective; the language audit shows that controlled target changes move the held-out optimum while a fixed response cannot. These experiments illustrate, rather than prove, the propositions. Together, they support the narrower and more general claim that references can supply useful inductive structure without identifying a target-optimal correction on their own.

Scope and limitations. The propositions are exact under their assumptions; the empirical regularities need not be universal. Several LTSSL conditions still use one checkpoint, although the focused CDMAD-style seed replication and three independent ACR runs cover two training families; one-dimensional dominance remains empirical. The language audit covers two model families, three datasets, and binary single-token English prompts, and tests Proposition 2 only. Proposition 3 is a strong sufficient condition; weaker checkable conditions and input-dependent corrections remain open.

7 Related Work

Long-tailed correction. Class-balanced losses, label-aware margins, balanced softmax, logit adjustment, and decoupled training encode frequencies in the objective or classifier (Cui et al., 2019; Cao et al., 2019; Ren et al., 2020; Menon et al., 2021; Kang et al., 2020). Modern SSL couples confidence-filtered pseudo-labels with weak-to-strong consistency (Sohn et al., 2020). Under long-tailed sampling, LTSSL additionally refines or resamples pseudo-labels (Kim et al., 2020; Wei et al., 2021; Oh et al., 2022), learns balanced or decoupled heads (Lee et al., 2021; Fan et al., 2022; Wei & Gan, 2023), or uses multiple experts and contrastive objectives (Ma et al., 2024; Hou & Jia, 2025; Zhou et al., 2024b; Zhu et al., 2022; Cui et al., 2021). Label-shift estimators instead infer an unknown prior from target data (Tasche, 2017; Lipton et al., 2018; Alexandari et al., 2020).

Reference correction and calibration. In vision, CDMAD reads a correction from solid colors (Lee & Kim, 2024), LCGC uses a black baseline (Xing et al., 2025), and FARAD uses random-phase references (Gu et al., 2025). Conventional calibration fits held-out labels, for example by temperature scaling (Guo et al., 2017). In language models, contextual calibration uses content-free prompts (Zhao et al., 2021); later methods use in-domain words, target batches, or generated marginals (Fei et al., 2023; Zhou et al., 2024a; Jiang et al., 2023). We intervene on the target while holding the reference fixed.

8 Conclusion

Reference subtraction exactly recenters logits but does not by itself identify a target-risk optimum. Theory states the missing assumptions. Vision audits reveal a useful prior-like direction but unreliable scale and residuals; language-model interventions show target-dependent optima while fixed references remain unchanged. References are useful probes, not distribution-free bias estimators: optimal correction requires structure or target information. Operationally, treat a reference as a candidate direction and identify its scale and objective from target-linked information.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing of the manuscript, literature search, theory check, debugging of research and drawing code, and analysis and interpretation of experimental results. All AI-assisted suggestions, analyses, and outputs were critically reviewed and verified by the authors. The authors determined the final research decisions, methodological choices, experimental conclusions, and manuscript content. The authors take full responsibility for the accuracy, validity and integrity of the work.

REPRODUCIBILITY STATEMENT

Appendix A provides the proof of Proposition 2 and its multi-class corollary; the other propositions are proved in the main text. Appendix C records the vision configurations, split roles, reference construction, and diagnostic fitting procedure. Appendices D–F describe the permutation controls, scale-identification experiments, and pseudo-label analyses. Appendix G specifies the language-model prompts, reference strings, sampling design, model precision, prior weighting, and bootstrap procedures. These details are included in the paper. We will release the analysis code, plotting scripts, frozen logits and vision exports, and experiment manifests upon acceptance.

REFERENCES

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232. PMLR, 13–18 Jul 2020.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 695–704, 2021. doi: 10.1109/ICCV48922.2021.00075.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019. doi: 10.1109/CVPR.2019.00949.

Alexander D'Amour, Katherine Heller, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Yue Fan, Dengxin Dai, Anna Kukleva, and Bernt Schiele. Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14554–14564, 2022. doi: 10.1109/CVPR52688.2022.01417.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14014–14031, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.783.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024.

Jiawei Gu, Yidi Wang, Qingqiang Sun, Xinming Li, Xiao Luo, and Ziyue Qiao. Fourier clouds: Fast bias correction for imbalanced semi-supervised learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 109989–110037. Curran Associates, Inc., 2025. doi: 10.52202/085713-3676.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017.

Yaxin Hou and Yuheng Jia. A square peg in a square hole: Meta-expert for long-tailed semi-supervised learning. In *Forty-second International Conference on Machine Learning*, 2025.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. Generative calibration for in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2312–2333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.152.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14567–14579. Curran Associates, Inc., 2020.

Hyuck Lee and Heeyoung Kim. CDMAD: Class-distribution-mismatch-aware debiasing for class-imbalanced semi-supervised learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23891–23900, 2024. doi: 10.1109/CVPR52733.2024.02255.

Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7082–7094. Curran Associates, Inc., 2021.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 10–15 Jul 2018.

Chengcheng Ma, Ismail Elezi, Jiankang Deng, Weiming Dong, and Changsheng Xu. Three heads are better than one: Complementary experts for long-tailed semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14229–14237, 2024.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9776–9786, 2022. doi: 10.1109/CVPR52688.2022.00956.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and HongSheng Li. Balanced meta-softmax for long-tailed visual recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186. Curran Associates, Inc., 2020.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608. Curran Associates, Inc., 2020.

Dirk Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18(95):1–32, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10857–10866, June 2021.

Tong Wei and Kai Gan. Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3469–3478, 2023. doi: 10.1109/CVPR52729.2023.00338.

Tong Wei, Zhen Mao, Zi-Hao Zhou, Yuanyu Wan, and Min-Ling Zhang. Learning label shift correction for test-agnostic long-tailed recognition. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52611–52631. PMLR, 21–27 Jul 2024.

Weiwei Xing, Yue Cheng, Hongzhu Yi, Xiaohui Gao, Xiang Wei, Xiaoyu Guo, Yumin Zhang, and Xinyu Pang. LCGC: Learning from consistency gradient conflicting for class-imbalanced semi-supervised debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21697–21706, 2025.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 49–70, 2024a.

Zi-Hao Zhou, Siyuan Fang, Zi-Jing Zhou, Tong Wei, Yuanyu Wan, and Min-Ling Zhang. Continuous contrastive learning for long-tailed semi-supervised recognition. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 51411–51435. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1627.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6898–6907, 2022. doi: 10.1109/CVPR52688.2022.00678.

A Proof of Target-Risk Non-Identifiability

Let $u = s(X) + \delta$. The integrability assumption makes $R_p(\delta)$ finite; the bounded derivatives of the logistic loss justify differentiating under the expectation. Thus

$$R'_p(\delta) = (1-p)\mathbb{E}_{Q_0}[\sigma(u)] + p\mathbb{E}_{Q_1}[\sigma(u) - 1], \tag{13}$$

$$R''_p(\delta) = (1-p)\mathbb{E}_{Q_0}[\sigma(u)(1-\sigma(u))] + p\mathbb{E}_{Q_1}[\sigma(u)(1-\sigma(u))]. \tag{14}$$

Integrability implies finite scores almost surely, hence $0 < \sigma(u) < 1$ and $R''_p(\delta) > 0$. Moreover, dominated convergence gives $R'_p(\delta) \to -p < 0$ as $\delta \to -\infty$ and $R'_p(\delta) \to 1 - p > 0$ as $\delta \to \infty$. A unique finite root $\delta^*(p)$ therefore exists.

For $g(\delta, p) = R'_p(\delta)$,

$$\frac{\partial g}{\partial p} = -\mathbb{E}_{Q_0}[\sigma(u)] + \mathbb{E}_{Q_1}[\sigma(u) - 1] \tag{15}$$

$$= -\mathbb{E}_{Q_0}[\sigma(u)] - \mathbb{E}_{Q_1}[1 - \sigma(u)] < 0, \tag{16}$$

while $\partial g/\partial\delta = R''_p(\delta) > 0$. The implicit-function theorem yields

$$\frac{d\delta^*(p)}{dp} = -\frac{\partial g/\partial p}{\partial g/\partial \delta} > 0. \tag{17}$$

A fixed reference correction δ_a is a function of the frozen model, prompt, and a, but has no argument p. Strict monotonicity implies $\delta_a = \delta^*(p)$ for at most one $p \in (0, 1)$.

The theorem uses binary NLL because it gives a transparent monotonic result. The support status of the reference never enters this argument.

Multi-class cross-entropy corollary. Suppose a C-class model outputs the source posterior, $\mathrm{softmax}(f_\theta(x)) = P_s(Y \mid x)$, with strictly positive coordinates almost surely and full-support prior π_s. The target obeys label shift, $P_t(X \mid Y) = P_s(X \mid Y)$, with full-support prior π_t. Bayes' rule gives

$$P_t(Y = c \mid x) \propto P_s(Y = c \mid x)\frac{\pi_{t,c}}{\pi_{s,c}}. \tag{18}$$

Because cross-entropy is strictly proper, a target-optimal centered correction applied as $f_\theta(x) - d$ is therefore

$$d^*(\pi_t) = \mathbf{P}(\log \pi_s - \log \pi_t)\,. \tag{19}$$

Two distinct target priors generally require distinct offsets while the frozen reference response is unchanged. This constructive multi-class case is enough to establish non-identifiability beyond binary classification; it does not claim strict monotonicity for every model. Margin, focal, non-proper, or non-convex losses require separate existence and uniqueness assumptions, so we make no loss-universal optimality claim.

B Finite Off-Support Construction

The 9,000 points in Figure 2 satisfy $\|x\|_2 \le 0.95$, while the reference is $a = (3, 0)$. The one-unit gate

$$g(x) = \mathrm{ReLU}\left(\frac{x_1 - 1}{2}\right) \tag{20}$$

is therefore exactly zero on every data point and equals one at a. Let $r_0(a)$ denote the centered response of the base classifier. For any desired centered response u, adding the output branch $g(x)(u - r_0(a))$ sets the reference response to u while preserving every in-domain logit exactly. This is the family evaluated in Figure 2; the proposition does not require that optimization select a particular member.

C EXPERIMENT LEDGER

Table 3: Information and evaluation role of the main experimental stages.

Stage	Information source	Evaluation	Purpose
Vision direction	diagnostic split	disjoint holdout	localize additive subspace
Direction null	fixed q_π and references	1,000 coordinate permutations/head	test direction specificity
Seed replication	three seeds in C10-u10/u50	frozen holdout + 1,000 permutations	test seed stability
Vision scale ladder	train, development, and unused support	separated by stage	test scale identification
Fair reserve	original label budget	held-out test	test scale under a fixed label budget
LM prior sweep	stratified fit half + target weights	disjoint weighted half	controlled target-risk intervention
LM length split	fixed balanced prior	stratified bootstrap	conditional-distribution audit
LM domain transfer	fixed prompts and balanced prior	2,000 stratified bootstraps	cross-domain target-risk audit

Frozen heads and evaluation splits. The nine-head audit uses the two CIFAR-10-LT mismatch checkpoints (labeled imbalance 100, unlabeled imbalance 10 or 50), CDMAD on CIFAR-10 (labeled/unlabeled imbalance 100/1, head-class budgets 500/4,000), CDMAD on CIFAR-100 (10/10, budgets 150/300), and STL-10 (labeled imbalance 10, head-class labeled budget 450 and 100,000 unlabeled images). ACR contributes its standard and balanced heads from the same checkpoint (CIFAR-10, consistent imbalance 100, budgets 1,500/3,000); CCL contributes the analogous two heads (consistent imbalance 100, budgets 500/4,000). All parameters remain frozen throughout the audit.

We use fixed class-stratified diagnostic and holdout splits, with 2,500/7,500 examples for either CIFAR dataset and 2,000/6,000 for STL-10. STL-10 images are resized to 32×32. The full centered offset and the nonnegative scalar along each declared direction are fitted by minimizing diagnostic cross-entropy and evaluated on the disjoint holdout. Zero correction is included as a candidate for the scalar fit.

Reference suite. The 20 validation references comprise six saturated RGB colors (red, green, blue, cyan, magenta, and yellow), uniform and Gaussian pixel noise, six grayscale patterns, and six random-phase references. The grayscale patterns cross low- and high-frequency sinusoids and a checkerboard with two contrast levels. Random-phase references preserve the mean Fourier amplitude of the evaluation images and vary the phase, using two realizations under each of three constructions: normalized space, raw space, and clipped raw space. The amplitude estimates use images but no labels. Black, gray, white, empirical image means, and semantic examples are diagnostic controls and are excluded from the 20-reference totals.

D CLASS-COORDINATE PERMUTATION CONTROL

For each frozen head, we generate 1,000 controls by permuting the class coordinates of q_π. Every control has exactly the same coordinate values and norm as the declared direction; only the correspondence between prior weight and class identity is destroyed. We then repeat the same nonnegative reference projection on the unchanged 20-reference validation suite. No evaluation label selects a direction.

The true direction also improves 176/180 full-reference pairs, whereas the permutation 95% interval is 87–100 pairs ($p < .001$). CCL-B is the expected exception: it has essentially no positive prior-direction headroom, so clipping returns the raw balanced head. Across the other eight heads, the true direction exceeds every permutation. The result rejects both an arbitrary one-dimensional projection account and a pure correction-norm shrinkage account.

Table 4: Direction-specificity control. "Null 95%" is the interval over 1,000 coordinate permutations of q_π; gains are relative to the raw head.

Head	Prior mean	Null 95%	Prior gain	p
C10-u10	85.96	[81.89, 84.53]	+2.15	$< .001$
C10-u50	84.74	[75.81, 79.72]	+7.49	$< .001$
CDMAD-C10	82.57	[76.49, 79.91]	+4.32	$< .001$
CDMAD-C100	55.16	[53.60, 54.13]	+1.17	$< .001$
STL-10	80.44	[73.48, 78.96]	+2.21	$< .001$
ACR-S	82.26	[75.27, 78.34]	+6.57	$< .001$
ACR-B	85.00	[83.50, 84.56]	+0.76	$< .001$
CCL-S	81.70	[75.70, 78.45]	+5.26	$< .001$
CCL-B	84.55	[82.37, 84.56]	+0.00	.137
Pooled	80.26	[76.22, 77.23]	+3.32	$< .001$

Three-seed replication. We repeat the two representative CIFAR-10-LT settings across three random seeds while keeping the architecture, training schedule, reference suite, projection, and held-out evaluation fixed. Parameters remain frozen throughout the audit. The setting-level null concatenates the 20 reference outcomes across the three seeds for each of 1,000 coordinate-permutation draws. Thus the control preserves the dimension, coordinate values, and norm of q_π while testing whether the class correspondence and the observed seed stability are incidental.

Table 5: Three-seed direction replication. Raw and projected columns are mean held-out bAcc. Gain is the mean $\pm$ standard deviation across seeds, in bAcc points; wins count projected corrections that outperform their corresponding full-reference correction. Null intervals and one-sided add-one p-values use 1,000 pooled class-coordinate permutation draws per setting.

Setting	Raw	Projected	Gain $\pm$ seed s.d.	Wins / 60	Null gain 95%	p_{gain}
C10-u10	84.39	86.32	$+1.93 \pm 0.16$	57	$[-1.24,\ 0.22]$	$< .001$
C10-u50	76.46	84.48	$+8.02 \pm 0.39$	56	$[-0.81,\ 1.39]$	$< .001$

Independent ACR runs. We repeat the matched CIFAR-10-LT consistent-100 audit on three independently trained ACR checkpoints ($N = 1500$, $M = 3000$), evaluating both the standard and balanced head with the same frozen 20-reference protocol.

Table 6: Matched-setting replication across independent ACR training runs. Each run contributes its standard (S) and balanced (B) head. Gain is projected minus raw bAcc; wins compare projection with the corresponding full response.

Run	Head	Raw	Projected	Gain	Wins / 20
1	S	75.693	82.259	+6.566	20
1	B	84.240	85.004	+0.764	19
2	S	76.973	83.833	+6.860	20
2	B	84.760	85.789	+1.029	20
3	S	77.267	83.752	+6.485	20
3	B	84.760	85.767	+1.007	20

Across all 120 model–reference pairs, the projected mean is 84.401 bAcc and the gain over raw is $+3.785$, whereas the permutation 95% interval for gain is $[-0.278, 0.648]$. Projection beats the complete response in 119 pairs, compared with 40–52 under the permutation null; all pooled one-sided tests give $p < 0.001$. Standard-head gains remain tightly grouped across runs, while the smaller balanced-head gains show that rebalancing has already removed most of the available prior-direction headroom.

E VISION SCALE IDENTIFICATION

We test progressively less dependent sources of scale information. Clean training support is memorized. Strong views, expected per-view risk, and balanced MixUp alter the view but remain functions of the same examples; each underestimates the target correction. A multiplier frozen on development heads transfers to held-out settings, establishing repeatable under-scaling without explaining it. Truly unused five-shot-per-class support recovers the diagnostic gain across six heads.

The fair test starts from the original CIFAR-10-LT labeled budget. With a fixed seed, five images per class are reserved and excluded from supervised and unlabeled training. The remaining labeled counts are $(1495, 894, 534, 318, 188, 111, 64, 36, 20, 10)$. After the fixed 250-epoch protocol, a single nonnegative scalar is fitted on clean reserved logits with a five-fold one-standard-error gate. No reference response, multiplier, augmentation, or test label enters the candidate.

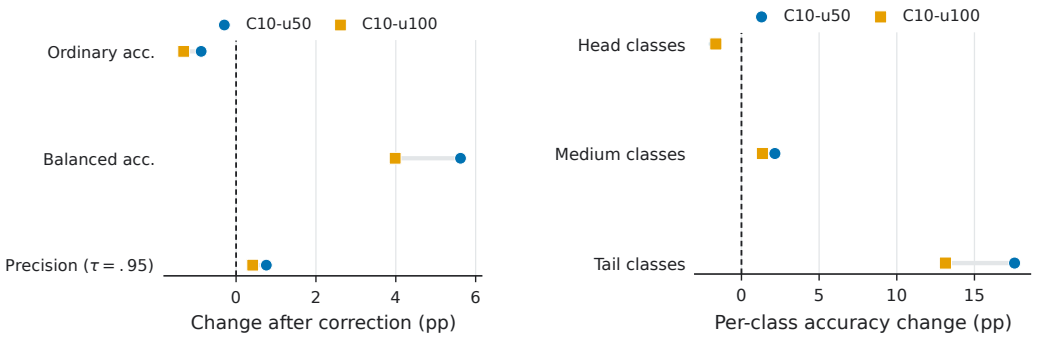
Table 7: Fair reserved-support results on CIFAR-10-LT with unlabeled imbalance 50. The last two rows use balanced test labels to measure additive headroom.

Correction	bAcc	Gain over raw
Raw	79.04	–
Canonical LA	81.59	+2.55
White reference	83.48	+4.44
Training support $\times 2$	83.31	+4.27
Reserved five-shot/class	83.28	+4.24
Prior-direction diagnostic	83.32	+4.28
Full-additive diagnostic	84.47	+5.43

The reserve gate activates and the candidate recovers 99.1% of the prior-direction diagnostic gain, but it does not outperform the favorable white reference (83.28 versus 83.48 bAcc). The result supports the claim that independent information can estimate scale within the declared direction, but not an overall improvement over the favorable white reference.

F PSEUDO-LABEL OBJECTIVE DECOMPOSITION

CDMAD applies the same subtraction to test predictions and pseudo-labels. Balanced evaluation weights classes uniformly; pseudo-label training weights samples from the unlabeled population and filters them by confidence. We apply the frozen correction to four fixed weak views in two mismatch settings. Balanced pseudo-label accuracy and high-confidence precision increase, but ordinary accuracy decreases in every view. Across all eight views, harmful flips outnumber correcting flips.



(a) Metric choice reverses the conclusion.

(b) Tail gains coexist with head losses.

Figure 6: **Inference correction and pseudo-label refinement optimize different risks.** Values are means over four fixed views per setting; circles denote C10-u50 and squares denote C10-u100.

G Language-Model Audit Details

Models and precision. We use frozen Qwen3.5-9B and Llama-3.1-8B-Instruct checkpoints in BF16. The model-native chat template is used for each family; all task content, demonstrations, label mappings, and references are otherwise matched. Only the logits of the two single-token labels are retained.

Data and prompts. The audit uses 872 SST-2 validation examples, a fixed stratified sample of 1,200 QNLI examples (Wang et al., 2018), and 600 class-balanced IMDb test reviews (Maas et al., 2011). Zero-shot and few-shot prompts are crossed with mappings in which label token A or B denotes the positive class. The few-shot demonstrations are fixed across examples. References replace only the test content with blank content, "N/A", a fixed random-word string, repeated symbols, a neutral sentence, or an ellipsis.

Prompt and reference specification. Each prompt asks the model to "Solve the classification task below" and "Return exactly one label and no other text", and explicitly assigns A and B to the two semantic classes. The classes are negative/positive for sentiment and not entailment/entailment for QNLI. Sentiment inputs contain a text field; QNLI inputs contain a question and a sentence. Few-shot prompts include one fixed training example from each class, in class order, before the evaluation input. IMDb uses the same SST-2 demonstrations. The model-native chat template supplies the assistant turn, with thinking disabled where supported, and the prompt ends with "Answer:". We read the next-token logits of the two labels without generating a response.

The six reference strings are the empty string, "N/A", "zqxv blorp niff", "@@@@ #### @@@@", "This is a piece of text.", and "...". For QNLI, the same reference string replaces both the question and sentence. Instructions, demonstrations, and label mappings remain unchanged when the reference replaces the evaluation input.

Prior intervention. For every task–prompt condition, labels are stratified into equal fit and evaluation splits. Each SST-2 half contains 214 negative and 222 positive examples; each QNLI half contains 300 examples per class. Target prior p is implemented with weights p_y/n_y normalized over the split. The free centered offset minimizes weighted NLL on the fit split and is evaluated on the held-out split. The same stored logits support all five priors, so the intervention has no sampling or forward-pass variation. "Harmful" compares a reference's held-out NLL with the uncorrected held-out NLL at the same target prior. The gap is not confined to NLL: relative to raw predictions, fixed references reduce target-weighted accuracy in 149/240 Qwen and 140/240 Llama comparisons. At $p = .5$, the counts are 38/48 and 35/48, with median changes of -1.83 and -2.51 percentage points, respectively.

Fixed-prior audit. Length thresholds are medians computed within each task–prompt condition, and offset differences are evaluated with 1,000 stratified bootstrap replicates. The cross-domain audit uses the same binary sentiment instructions, demonstrations, label mappings, and references for SST-2 and IMDb. The 2,048-token budget preserves the terminal answer cue for every sampled IMDb review. Cross-domain intervals use 2,000 stratified bootstrap replicates. Lengths count tokens in the complete formatted prompt; ties at the median belong to the short group. In each bootstrap replicate we resample with replacement within each class and each group or domain, retaining the class counts and weighting each class by one half. We refit the binary NLL-optimal score shift in each replicate and report the 2.5th and 97.5th percentiles of the shift difference.

Affine reconstruction. For selected label tokens Y, the last-layer identity is $z_Y(x) - z_Y(a) = W_Y(h(x) - h(a))$. The Llama probe's maximum full-vocabulary reconstruction discrepancy is 0.125 logits; after centering the selected positive-minus-blank contrast, it is 0.044 logits in BF16.