AUDITING GRADIENT RETENTION UNDER HIDDEN CONDITIONS: EXACT CANCELLATION, OBJECTIVE-DEPENDENT ALIGNMENT, AND THE LIMITS OF KAPPA

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning systems often average gradients over hidden conditions: annotator groups with conflicting preferences, clients with different objectives, hidden teammates. When does averaging aggregate useful signal, and when does it suppress it? We show that gradient cancellation is governed by more than the presence of hidden heterogeneity. In an analytically tractable mirror model, elementwise parameter-independent weight fields cancel exactly under a symmetric condition mixture, while a differentiable-baseline field can retain a nonzero component away from uniform policies. For K-way mixtures retention depends on the mixture direction in the probability simplex at fixed distance from uniform, and objective-induced channels can interfere, making retention non-monotone in the entropy weight. A measurement protocol separates these structural effects from finite-sample noise: deterministic rollouts collapse retention into a weight ratio, and episode-noise readouts can understate it by an order of magnitude. Under a single bounded definition the retention score alone is ambiguous: a high score can be produced by condition-blindness rather than survival. In Overcooked the dynamic value field scores 0.998 while its condition contrast is only 0.1% of its shared mass; the differentiable-baseline advantage field carries a resolvable contrast (35–$176\times$ the estimation-noise floor, score 0.590) and the plain policy gradient's contrast sits at the noise floor (score 0.529). In 510K no field has a substantial condition contrast. The 30–$65\times$ value–policy gap suggested by noisy readouts is retired. A controlled capacity experiment shows that severe cancellation can coexist with poor worst-condition performance; in contrast, on a real three-group preference dataset the condition-contrast energy is dominated by item-level noise and group-conditioned capacity yields no gain. In a switching task, retention need not predict return when the measured condition is not the variable the task relies on. Retention is a structural audit of the specified condition signal, not a general performance metric, and it indicates when condition-aware capacity is likely worth its cost.

1 INTRODUCTION

Supervised and reinforcement learning systems are routinely trained on data that mixes heterogeneous subpopulations. The subpopulation label is frequently missing: a preference dataset pools annotators with conflicting tastes, a federated client population encodes different task distributions, a multi-agent episode hides which role the partner plays. The model sees one input distribution, while the label or reward is generated by a condition that is not present in the input. Training then optimizes the mixture objective, and the update direction is the mixture of condition-specific gradients,

$$\bar{g} \;=\; \sum_i p_i \, g_i, \tag{1}$$

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.

where $g_i(\theta) = \nabla_\theta L_i(\theta)$ is the population gradient of the condition-i objective under a fixed rollout distribution.

A mixture can erase signal. The extreme case is cancellation: if two conditions demand opposite behavior on the same input, their gradients can point in opposite directions and the average update is zero even though each condition carries a strong, learnable signal. Practitioners observe this as training that stalls without any visible loss anomaly. This paper asks how to audit that failure precisely: when cancellation is exact, when it is partial, how it depends on the objective, and what the resulting measurement does and does not tell us.

A companion phenomenon was reported previously: hidden relational variables can contract policy-gradient updates while value-based updates appear to survive. We show that the apparent survival is, in the tested environments, condition-blindness of the value field rather than retained condition signal: the mechanism is in the objective's condition contrast and in what the retention ratio fails to penalize. Our starting point is that the mechanism cannot be recovered from algorithm-family labels, because a gradient field is a property of the *objective evaluated on a fixed policy and data*, not of the optimizer. We therefore fix the measurement basis: same network, same rollout distribution, same condition definition, and change only the scalar objective whose gradient defines the field.

Thesis. When hidden conditions are averaged, what survives is determined not only by symmetry between conditions, but by the geometry of the condition mixture and by the objective-induced structure of the gradient field. The results below are chapters of that one question.

Contributions.

1. **Structure: which fields survive, and why.** (a) For the mirror bandit, any elementwise, parameter-independent weight field satisfies $\mathbb{E}[g_B] = -\mathbb{E}[g_A]$, hence zero shared gradient under the symmetric mixture; the same normalization cancels K-way matching mixtures at every K. (b) The soft-Q field has an exact closed form, $\kappa_{\text{mix}} = \Delta^2/(\Delta^2 + 4r^2)$ with $\Delta = \alpha \log(\rho/(1-\rho))$, verified against autograd to machine precision. (c) The advantage-weighted field with a differentiable baseline lies outside the class in (a) because its baseline depends on the policy; its derivative adds an even coupling, so it survives the symmetric mixture. The soft-Q field is another outside-class case whose entropy term supplies a shared component. (d) Direction dependence appears at $K \geq 3$, and entropy and mixture channels interfere with a computable minimum; these are properties of the condition simplex, not a non-abelian group.

2. **Measurement: what a retention score can and cannot be.** The bounded mixture-reference retention separates structural retention from episode noise. Deterministic rollouts reduce the score to a single-action weight ratio on the mirror model (0.633 where the correct protocol gives 0.001), and episode-noise readouts can sit an order of magnitude below structural scores. Fields must be compared on the same batch, and contrast energy read jointly with the estimation-noise floor and the shared component it sits on.

3. **Boundaries: what a retention score can and cannot say.** Retention alone is ambiguous: a high score can reflect condition-blindness rather than survival. In Overcooked the dynamic value field retains 0.998 but its condition contrast is 0.1% of its shared mass, the differentiable-baseline advantage field carries a resolvable contrast (35–176$\times$ the estimation-noise floor) at 0.590, and the plain policy gradient's contrast sits at the noise floor (0.529). In 510K no field has a substantial condition contrast (value share 1–5$\%$), and the 30–65$\times$ gap came from noisy readouts. A controlled capacity experiment shows that severe cancellation with contrast energy dominating noise is fixed by condition-aware capacity (worst condition 0.003 to 1.97), whereas a real three-group preference dataset with contrast dominated by item-level noise gains nothing from conditional capacity. A switching task shows retention need not predict return when the measured condition differs from the variable the task relies on. The audit recommends intervention only when low retention coincides with contrast energy above noise and a task-relevant condition.

2 Related Work

Gradient conflict and multi-task optimization. PCGrad (Yu et al., 2020), CAGrad (Liu et al., 2021), MGDA (Sener & Koltun, 2018), GradNorm (Chen et al., 2018), GradDrop (Chen et al., 2020), and later gradient-surgery methods treat conflict as an observed quantity to be projected, reweighted, or scheduled away. We study the complementary question: when conflict is *structural*, i.e. forced by hidden-condition symmetry before any optimizer intervenes. Our audit is a diagnostic, not another surgery rule, and we do not claim to replace those methods. The retention score is not a cosine similarity: for two equally weighted conditions it reduces to a monotone function of the gradient cosine only when the two gradient norms match, and the mixture-reference ratio stays defined when the average cancels and extends to K conditions with general weights.

Hidden context and preference aggregation. Distributional preference learning (Siththaranjan et al., 2024) shows that averaging unobserved annotator groups distorts a reward model; heterogeneous-preference alignment work studies minority under-representation and equitable aggregation (Chakraborty et al., 2024). We work in gradient space rather than risk or reward space and provide exact cancellation conditions plus objective-specific survival rates; we do not propose a new alignment algorithm.

Symmetry, averaging, and invariance. Averaging over a group acts as a projection onto invariant components (a Reynolds operator), and the resulting decomposition is classical. Our use of it is instrumental: we apply it to condition-conditioned gradient fields and derive measurement consequences. We explicitly do not claim group averaging as a contribution.

The companion report. A prior study (citation withheld for double-blind review) introduced the retention ratio and reported estimator dependence and an apparent policy-gradient/value reversal. That report is a phenomenon study. The present work replaces the algorithm comparison with a common-basis field measurement, derives the closed forms, and corrects two measurement practices on which the earlier scores depended.

3 Problem Setup and Measurement

3.1 Hidden-condition gradient fields

Let i index a hidden condition with mixture weights p_i, and let $g_i = \nabla_\theta L_i$ denote the population condition gradient of a scalar objective L_i under a fixed rollout distribution. We report the bounded structural retention

$$\kappa_{\text{mix}} \;=\; \frac{\|\bar{g}\|^2}{\sum_i p_i \|g_i\|^2}, \qquad 0 \le \kappa_{\text{mix}} \le 1, \tag{2}$$

where the upper bound follows from convexity. We also report the energy decomposition

$$E_{\text{shared}} = \|\bar{g}\|^2, \quad E_{\text{contrast}} = \sum_i p_i \|g_i - \bar{g}\|^2, \quad E_{\text{mix}} = \sum_i p_i \|g_i\|^2. \tag{3}$$

The historical uniform-reference normalization $\|\bar{g}\|^2 / \text{mean}_i \|g_i\|^2$ is not bounded for non-uniform p and is used only to read legacy result files.

3.2 Fields under study

All fields are gradients with respect to the same policy parameters on the same data (Table 1).

The distinction between a differentiable and a detached baseline matters: only the former changes the parity structure of the field, so we refer to our `awr` as an advantage-weighted field with a differentiable baseline rather than as standard AWR.

3.3 Measurement protocol

Three rules follow from the audit in Section 5.2 and are used throughout. (i) Rollouts are stochastic and policy-weighted; deterministic (argmax) rollouts collapse the score into a weight ratio on the

Table 1: Measured fields. Every entry is a gradient with respect to the same policy parameters on the same rollouts; only the scalar objective changes.

Field	Objective or weight	Note
`reinforce`	$\mathbb{E}_a[\nabla(-\log \pi(a)\, Q_r(a))]$	parameter-free weight
`expq`	$\nabla(-\sum_a \pi(a) Q_r(a))$	full objective
`softq`	$\nabla \sum_a \pi(a)(\alpha \log \pi(a) - Q_r(a))$	detached Q
`awr`	$\mathbb{E}_a[\nabla(-\log \pi(a)\, e^{(Q_r(a)-V)/\tau})]$	differentiable V
`softmaxq`	$\mathbb{E}_a[\nabla(-\log \pi(a)\, \mathrm{softmax}(Q_r/\tau)_a)]$	Q-dependent weight

mirror model. (ii) Conditions are forced per rollout when the environment allows it, so the measured contrast is between clean condition assignments rather than between random mixes. (iii) When two fields are compared, both are computed from the *same* episode batch. We report structural scores from condition means and, separately, the one-episode noise diagnostic

$$\kappa_{\mathrm{ep}} \;=\; \frac{E_{\mathrm{shared}}}{E_{\mathrm{mix}} + \sum_i p_i \sigma_i^2}, \tag{4}$$

where σ_i^2 is the within-condition episode-gradient variance. A finite-sample mean calibration, $E_{\mathrm{shared}}/(E_{\mathrm{mix}} + \sum_i p_i^2 \sigma_i^2 / n_i)$, is reported as a surrogate and is not a high-probability guarantee.

4 Theory

4.1 Exact cancellation

Proposition 1 (Mirror cancellation)**.** *In the two-action mirror bandit with $Q_A = (r, -r)$ and $Q_B = (-r, r)$, any field whose weight is elementwise in Q_r and independent of the policy parameters, $w_r(a) = f(Q_r(a))$, satisfies $\mathbb{E}[g_B] = -\mathbb{E}[g_A]$. Under the symmetric mixture $p = (1/2, 1/2)$, $\mathbb{E}[\bar{g}] = 0$ and $\kappa_{\mathrm{mix}} = 0$ exactly.*

Verification. Closed form and autograd agree at the 10^{-32} level for the stored configurations, covering `reinforce`, `expq`, `awr` without baseline, and `softmaxq` (`data/kappa/toy_fields/o1_closed_forms.json`).

Proposition 2 (K-way normalization)**.** *In the K-way matching bandit with $Q_i(b) = 2R\,\delta(i,b)$ and a uniform condition mixture, $\sum_i g_i = 0$ for every K; the cancellation is a normalization identity, not a group-theoretic accident. Verified for $K = 2, 3, 4$ with $\kappa_{\mathrm{mix}} < 1.8 \times 10^{-32}$.*

The boundary is explicit: the result concerns expected gradients under the matching construction. Real environments are not exact mirrors, and the measured structural scores there are around 0.5, i.e. near-orthogonal rather than cancelling.

4.2 Closed form for the soft-Q field

Proposition 3 (Soft-Q retention)**.** *In the mirror bandit, let $\rho = \pi(\text{action } 0)$, $q = 1 - \rho$, and $\Delta = \alpha \log(\rho/q)$. The soft-Q field satisfies*

$$c_A = \rho q(\Delta - 2r), \qquad c_B = \rho q(\Delta + 2r), \qquad g_i = c_i u, \qquad \kappa_{\mathrm{mix}} = \frac{\Delta^2}{\Delta^2 + 4r^2}. \tag{5}$$

Verification. Closed form versus autograd of the explicit expected loss agrees to 5.0×10^{-16}; Monte Carlo anchors agree within sampling error; the stored α-grid values match to the archived precision (one archived point differs by 0.001 under a documented old rounding convention; `verify_closed_forms_o1.py`). Limiting cases: $\alpha \to 0$ degenerates to the cancelling expected-Q field, while a non-uniform policy or a larger entropy weight increases retention. This is a fixed-policy identity; it does not describe a training trajectory.

Proposition 1 covers parameter-independent elementwise fields, which cancel under the symmetric mixture. The advantage-weighted field with a differentiable baseline lies outside that class because its baseline depends on the policy; its derivative introduces an additional even coupling, and the soft-Q field is another outside-class case whose entropy term supplies a shared component. Its closed form appears in Appendix A; at uniform policies it also cancels for every temperature; for non-uniform policies retention is monotone decreasing in τ, with supremum $1/2$ as $\tau \to 0$.

4.3 K-WAY GEOMETRY AND CHANNEL INTERFERENCE

Proposition 4 (Direction dependence)**.** *For the K-way matching field, $g_i[b] = -2R\,\pi_b(\delta(i,b) - \pi_i)$ and $\bar{g}[b] = -2R\,\pi_b(p_b - \langle p, \pi\rangle)$, where the second term is the projection of the mixture onto the policy-orthogonal complement. When the condition simplex has dimension at least two ($K \geq 3$), retention at fixed L_1 distance from the uniform mixture varies with direction.*

Measured spread at L_1 distance 0.45 (200 directions): $K = 3$: $\kappa_{\text{mix}} \in [0.070, 0.217]$ ($3.08\times$); $K = 4$: $[0.051, 0.210]$ ($4.08\times$). This is simplex geometry; any claim that the effect is specific to S_3 or to non-abelian symmetry is withdrawn. Direction dependence is absent for $K = 2$ because the mixture simplex is one-dimensional.

Proposition 5 (Channel interference)**.** *The soft-Q field is the sum of an entropy channel and a condition-mixture channel in the same weighted tangent space. When these two directions have positive weighted inner product, retention is non-monotone in α: it dips to a minimum where the two channels balance, then recovers.*

Measured minima on a 1001-point scan: $\kappa_{\text{mix}} = 2.5 \times 10^{-6}$ at $\alpha = 1.72$ ($K = 2$, $p = (0.8, 0.2)$) and 4.0×10^{-4} at $\alpha = 1.46$ ($K = 3$, $p = (0.7, 0.2, 0.1)$). The shared-energy and ratio minima coincide on the same grid point, so the scan reports a single minimum with the denominator (`kway_geometry.json`).

A dynamical reading of this interference was proposed previously and is withdrawn here. Under the implemented objective $L = \sum_a \pi(a)(\alpha \log \pi(a) - Q(a))$, gradient descent on the symmetric mixture minimizes $-\alpha H(\pi)$ and therefore moves toward higher entropy, not toward a peaked policy. Static channel interference survives as a statement about the field, not about an attractor.

5 EXPERIMENTS

All numbers are produced by the canonical implementation and collected in `notes/canonical_metric_results.md`; raw files are listed in Appendix C.

5.1 EXACT BANDIT: CANCELLATION AND CLOSED FORMS

On the mirror bandit we evaluate one policy with identical parameters and an identical evaluation seed schedule under two observation protocols: revealed (the partner one-hot is part of the observation) and masked (the partner slot is zeroed, so the relation is hidden while the reward and the analytic Q still depend on it). Under masking, `expq` cancels exactly ($\kappa_{\text{mix}} = 0$, $E_{\text{shared}} = 0$), `reinforce` is indistinguishable from zero (0.0025), `softmaxq` and `softq` retain 0.0005 and 0.0045, and the differentiable-baseline `awr` field retains the largest small component (0.0091); the soft-Q component follows the closed form and grows with α (Proposition 3). Under the revealed protocol all five fields collapse to the same value (0.4397–0.4402), the expected degeneracy of a policy that already solves the matching task. With parameters, rollouts, and seeds fixed and only the observation mask changed, the contrast isolates condition visibility (`data/kappa/toy_fields/visibility_control.json`); five random-init policies show the same pattern.

5.2 PROTOCOL AUDIT: DETERMINISTIC ROLLOUTS INVENT A FIELD AXIS

On the hidden mirror bandit, deterministic (argmax) rollouts visit a single action per condition, and every field reduces to the ratio of weights at that action. Measured artifacts: `softmaxq` reports 0.633 under deterministic rollouts and 0.0011 under the correct stochastic protocol; `awr` reports

Table 2: Overcooked structural retention and joint diagnostics (three seeds; conditions are episode-start partners, switching preserved, team sparse reward). All fields share one episode batch per checkpoint, and the `awr` entry keeps the value baseline in the autograd graph. Retention κ_{mix} alone is ambiguous; c/n is the condition contrast over the mean-estimation noise floor (c/n ≤ 1 is unresolvable, and a high retention with a small contrast share is condition-blindness). The value entry is the diagnostic field $-\sum_t V(s_t)$; it is not a TD residual.

Field	static κ_{mix}	dynamic κ_{mix}	static c/n	dynamic c/n
`reinforce`	0.515 ± 0.021	0.529 ± 0.034	1.4	1.1
`awr`	0.500 ± 0.000	0.590 ± 0.021	34	110
`value` ($-\sum V$)	0.568 ± 0.027	0.998 ± 0.000	3830	9.6

Table 3: 510K structural retention, paired protocol (six seeds per level). Both fields are computed from the same episode batch. Contrast-to-noise means are $0.8/0.9/0.9$ (`reinforce`) and $1.2/1.0/2.1$ (`value`) for $p = 0/0.5/1$; the condition signal is not resolvable in either field.

Visibility p	`reinforce`	`value`
0.0 (hidden)	0.646 ± 0.119	$\mathbf{0.982 \pm 0.005}$
0.5	0.612 ± 0.223	$\mathbf{0.985 \pm 0.003}$
1.0 (visible)	0.667 ± 0.184	$\mathbf{0.962 \pm 0.011}$

0.551 versus 0.028. `expq` remains exactly zero under both protocols, and the revealed-policy values are protocol-robust (~ 0.396 both ways). Audit rule: an apparent field axis measured with argmax rollouts on a degenerate model is a weight ratio until recomputed with policy-weighted sampling.

5.3 Real environments under the canonical metric

Overcooked (switching-preserving protocol). One episode batch per checkpoint is shared by all fields, and only the measured objective changes, so the comparison is paired by construction (Table 2). Conditions are the episode-start partner (chef vs. waiter); mid-episode switching stays on, and the reward is the team-level sparse delivery reward. Retention alone is ambiguous here. On the dynamic model the value field scores 0.998, but its condition contrast is only 0.1% of its shared mass (5–17$\times$ the estimation-noise floor): the field is condition-blind, and its high score reflects a dominant condition-independent component rather than hidden-condition signal. The differentiable-baseline `awr` field is the only non-value field with a contrast clearly above the noise floor (35–176$\times$) at a moderate score (0.590), while the plain policy gradient's contrast sits at the noise floor (0.85–$1.3\times$, score 0.529). Here `awr` uses the observed return-to-go as the advantage sample and the value head as the differentiable baseline. Under the start-partner protocol the plain-gradient contrast is again at the floor in every condition group (static 0.533, dynamic 0.502), and a memory intervention does not make it resolvable (m4 0.498, m8 0.505, m16 0.500; contrast 0.6–2.1$\times$ the noise floor): a negative intervention result.

510K (paired protocol). Using checkpoints trained with hidden teammate visibility $p \in \{0, 0.5, 1\}$ and computing both fields from the same episode batch with mask-respecting rollouts, per-step reward-weighted policy gradients retain $0.646/0.612/0.667$ (large seed spread, ± 0.12–0.22) while value fields retain $0.982/0.985/0.962$ (Table 3). Neither field has a substantial condition contrast: the value contrast is 1–5% of its shared mass at 0.7–3.1$\times$ the estimation-noise floor; the policy-gradient contrast is 0.25–2.0$\times$ the floor, and no visible/hidden trend is interpretable. The stable value score again tracks a dominant shared component rather than hidden-condition dependence.

Reading. A single retention number cannot order these fields. The value field's high score is condition-blindness: a condition-independent component dominates its gradient, so the field changes little across conditions (0.1% contrast share in Overcooked dynamic, 1–5% in 510K).

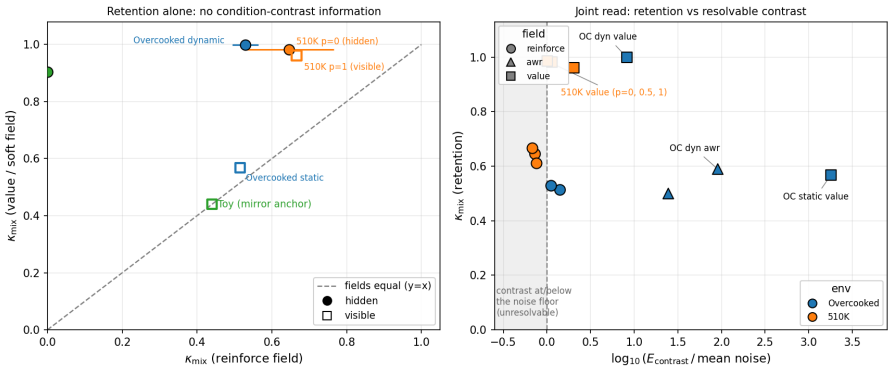

Figure 1: Retention alone and the joint diagnostic. Left: scatter under κ_{mix} (condition means); retention separates the fields but carries no information about their condition contrast. Right: κ_{mix} against the condition contrast over the mean-estimation noise floor (log scale); the shaded band is unresolvable. The Overcooked dynamic value point (retention 0.998) has a contrast share of 0.1% and is condition-blind, as are the 510K value points; the differentiable-baseline advantage points are the only non-value points that sit clearly above the floor with a substantial contrast share. The mirror-bandit point is an exact anchor, not a real environment.

Table 4: Shared versus condition-aware capacity (exact full-batch training). Rewards are reported as average / worst condition.

Mixture	Model	avg. reward	worst reward	initial κ_{mix}
uniform	shared, α=0	0.667	0.466	0.000
uniform	conditional, α=0	1.992	1.992	—
mild	shared, α=0	1.197	0.003	0.208
mild	conditional, α=0	1.991	1.970	—
mild	shared, α=2	0.752	0.517	0.445
strong	shared, α=0	1.597	0.002	0.501
strong	conditional, α=0	1.991	1.933	—

The plain policy gradient in Overcooked has no resolvable contrast at our sample size, so its near-orthogonality is an upper bound. The differentiable-baseline advantage field is, among the non-value fields measured on the dynamic Overcooked protocol, the only one whose contrast clearly exceeds the noise floor (35–176×), though it still mixes in a large shared component. This retires the reported 30–65× separations (one-episode noise scores); we report κ_{mix} with contrast/noise and the contrast share, and do not rank objectives by retention alone.

5.4 Capacity, not optimizer

On an exact K-way training comparison with a shared policy versus a condition-aware policy (full-batch, deterministic, same initial logits), the shared model's limitation is a mixture compromise (Table 4). With a uniform mixture and the linear objective, the shared model stays at average reward 0.667 while the conditional model reaches 1.99 ($\kappa_{\text{mix}} = 0$ at initialization, by cancellation). Under mild and strong asymmetric mixtures the shared model improves the average (1.20 and 1.60) but leaves the worst condition at 0.003 and 0.002, while the conditional model keeps the worst condition at 1.97 and 1.93. The entropy channel ($\alpha = 2$) trades average for worst-condition performance (0.752/0.517), a Pareto movement rather than a fix. The intervention suggested by the audit is conditional information or capacity, not an arbitrary gradient rule.

Table 5: DICES-350 supervised check (three seeds). The race axis carries the strongest condition signal, but its contrast energy is dominated by within-group item noise, and conditional capacity yields no held-out gain.

Axis (groups)	κ_{mix}	E_{contrast}	σ^2	shared / cond. avg–worst
race (3)	0.156 ± 0.027	0.0019	0.0114	$0.656/0.609$ / $0.654/0.607$
age (3)	0.002 ± 0.000	0.0002	0.0058	$0.640/0.623$ / $0.638/0.624$
gender (2)	0.001 ± 0.000	0.0005	0.0035	$0.641/0.623$ / $0.640/0.625$

Table 6: Switching task: objectives trained from scratch under an identical budget. Post-switch reward is measured on steps 20–28 after the flip.

Arm	return	post-switch reward	final κ_{mix}
`td`	$\mathbf{36.97 \pm 0.06}$	0.75	0.06 ± 0.04
`softq`	28.62 ± 11.81	0.58	0.94 ± 0.09
`reinforce`	9.48 ± 5.13	0.18	0.43 ± 0.18
`awr`	5.03 ± 2.02	0.04	0.57 ± 0.08

5.5 Real three-group preference data: a single-dataset case study

To test whether the audit extends beyond RL bandits, we use DICES-350 (Aroyo et al., 2023): 350 adversarial conversations, each rated by all 123 raters, with rater demographics; the dataset is released under CC BY 4.0. We treat a demographic attribute as the hidden condition, use the binary perceived-harm judgment (`Yes` vs `No`, `Unsure` dropped), split by `item_id` (245 train / 105 test items), and fit a TF-IDF plus logistic model shared across raters. Three seeds; the shared model reaches 0.73 train accuracy. Table 5 reports the audit and the capacity comparison. The audit uses the BCE data gradient (the condition-independent L_2 term shifts every condition mean alike and cancels from $E_{\text{contrast}}/\sigma^2$), with empirical rater-frequency weights.

On the race axis, the contrast energy is about six times smaller than the within-group item noise ($E_{\text{contrast}}/\sigma^2 \approx 0.17$), and group-conditioned models do not improve average or worst-group accuracy. The audit therefore predicts that this is not a capacity problem, and the held-out comparison confirms it. This complements the bandit capacity experiment: there, contrast dominates noise and condition-aware capacity raises the worst condition from 0.003 to 1.97; here, contrast is noise-dominated and conditional capacity buys nothing. Caveats: `Q_overall` is perceived harm, not objective harm; `Unsure` is dropped; the model is a linear bag-of-ngrams classifier without group-DRO or reweighting baselines.

5.6 Retention is not a performance score

In a switching hidden-matching task with four objectives trained from scratch under an identical budget (eight seeds), final returns and final structural retention are shown in Table 6. The arm with the lowest measured retention is the best performer. The explanation is variable mismatch: the task requires tracking a phase from reward history, which is orthogonal to the initial partner label that retention conditions on. The diagnostic answers "how much of the specified condition signal survives averaging," not "will training succeed." A low retention score must therefore be paired with a check that the condition is the variable the task actually relies on. The `awr` arm in this table uses a detached value baseline with advantages clipped to $[-4, 4]$; it is a different field from the differentiable-baseline `awr` measured in Section 5.3 and in Appendix A. The `td` entry is a critic-loss gradient on the arm's own ϵ-greedy rollouts, a critic diagnostic rather than a policy-gradient field; each arm is scored on its own rollouts.

6 What the Audit Does and Does Not Claim

The retained statements are: (i) exact cancellation conditions on the mirror and matching models; (ii) an exact closed form for the soft-Q field and the differentiable-baseline exception; (iii) K-way direction dependence and channel interference as geometry; (iv) protocol and noise separation as methodology; (v) a joint diagnostic in which a high retention with a negligible contrast share is condition-blindness (the Overcooked dynamic value field, and value in 510K), a contrast at the estimation-noise floor is unresolvable (the plain policy gradient in Overcooked), and in 510K the contrast stays at the noise floor or a few percent of the shared mass, and the differentiable-baseline advantage field in dynamic Overcooked is the only non-value field with a contrast clearly above the noise floor; and (vi) decoupling from performance. The operational rule: condition-aware capacity is the intervention when retention is low, contrast exceeds the estimation-noise floor, and the condition is task-relevant; when the contrast is unresolvable or negligible, or the condition is irrelevant, no intervention is recommended. We do not claim that value methods are universally better, that policy-gradient fields always vanish, that direction dependence is group-specific, that retention predicts return, or that memory or entropy interventions fix cancellation.

7 Limitations

The exact results are bandit-level. Real-environment claims rest on condition-mean gradients under one measurement protocol and on small seed counts (three seeds for the Overcooked field axis). The value field is the diagnostic $-\sum V$, not a TD residual; its score is dominated by the fitted value function's condition-independent component on the measured distribution, which the retention ratio does not penalize. Retention compares condition means within one field; ranking κ_{mix} across objectives with different supports in the padded parameter vector is not licensed. The Overcooked `awr` entry uses the observed return-to-go as the advantage sample and the learned value head as the differentiable baseline, so it is the practical analogue of the analytic field; its per-step weights are computed in float64 under a dataset-level stop-gradient shift, a positive scalar multiple that leaves κ_{mix} unchanged (float32 underflowed up to 78% of the steps to exactly zero, and the float64 re-run reproduces every score to within 1.5×10^{-7}). The 510K environment has weak relational drive and is used as a supporting measurement, not as the primary demonstration. Its checkpoints were trained under a legacy action mask that always enabled the pass bit; the mask-respecting evaluation corrects the measurement side and does not require retraining. The real-data check is a single-dataset case study (DICES-350), with a linear TF-IDF model and binary labels, without conditional-routing, group-DRO, or reweighting baselines; a second dataset is left to future work. The finite-sample analysis reports a deterministic surrogate; a high-probability guarantee for the ratio estimator would require an explicit distributional assumption.

Reproducibility statement

The metric implementation is `src/iigc/metrics/kappa.py`; the closed forms, protocol and visibility controls, K-way controls, Overcooked and 510K measurements, and the capacity comparison are scripted under `experiments/common_basis/`. A snapshot of the code repository and the Overcooked checkpoints are provided as supplementary material; the checkpoint archive carries SHA256 checksums and the 510K checkpoints (supporting measurement) are available on request. `ARTIFACT.md` documents extraction, environment setup, and one-click regeneration via `reproduce_canonical.py`. Result files carry a `measurement_metadata` block recording the gradient definition, condition weights, rollout protocol, parameter space, and noise treatment; the switch-kappa file also carries a `_provenance` block (git commit, worktree state, per-checkpoint SHA256, seed schedule). Unit tests cover the closed forms, cancellation, boundedness, the metadata contract, and the measurement wrappers. No human subjects or private data are involved. Full provenance is listed in Appendix C.

AI use statement

In this work, we used generative AI tools to aid and polish writing, to assist with research ideation and code, to draft sections of the paper, and to assist with mathematical derivations. We have reviewed and verified all AI-assisted work: all closed-form results were checked against independent

autograd and Monte Carlo computations, and all reported numbers are regenerated by the scripts listed in Appendix C. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies optimization diagnostics for models trained on mixtures of hidden subpopulations. It does not involve human subjects or private data. The discussed failure mode (minority conditions being under-represented by averaged objectives) has fairness implications; we present the audit as a diagnostic and make no deployment recommendation.

REFERENCES

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parr, Greg Serapio-Garcia, Ding Wang, and Gaurang Ghosh. Dices dataset: Diversity in conversational ai evaluation for safety. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. In *International Conference on Machine Learning*, 2024.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, 2018.

Zhao Chen, Jiquan Ngiam, Piyush Sapra, Qizhen Le, et al. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *Advances in Neural Information Processing Systems*, 2020.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, 2018.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. In *International Conference on Learning Representations*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

A DIFFERENTIABLE-BASELINE CLOSED FORM

For the advantage-weighted field with $w = \exp((Q_r - V)/\tau)$ and a differentiable baseline $V = \sum_a \pi(a) Q_r(a)$, write $x = 2r/\tau$,

$$A_0 = e^{xq}, \qquad A_1 = e^{-xp}, \qquad M_A = p \log p\, A_0 + q \log q\, A_1, \tag{6}$$

$$B_0 = e^{-xq}, \qquad B_1 = e^{xp}, \qquad M_B = p \log p\, B_0 + q \log q\, B_1, \tag{7}$$

with $p = \pi(\text{action } 0)$, $q = 1 - p$. Then

$$c_A = pq\,(A_1 - A_0 + xM_A), \qquad c_B = pq\,(B_1 - B_0 - xM_B). \tag{8}$$

At $p = 1/2$ the field cancels exactly for every τ; otherwise retention is monotone decreasing in τ, with supremum $1/2$ as $\tau \to 0$. Closed form and autograd agree to 9.2×10^{-16} and Monte Carlo anchors agree within three standard errors (`verify_closed_forms_o1.py`).

B FISHER/KL INTERPRETATION (VARIANT, NOT THE DEFAULT METRIC)

The default retention is Euclidean in parameter space, with parameter-space condition-mean gradients $g_i = \nabla_\theta L_i$. For categorical policies a separate Fisher-weighted variant is defined on the logits, with logit-space gradients $d_i = \nabla_z L_i$ and $F(\pi) = \mathrm{diag}(\pi) - \pi\pi^\top$ (the rank-one term retained):

$$\kappa_F = \frac{\bar{d}^\top F \bar{d}}{\sum_i p_i \, d_i^\top F d_i}, \qquad \mathrm{KL}\big(\pi(z) \,\|\, \pi(z + \epsilon v)\big) = \frac{\epsilon^2}{2} v^\top F v + O(\epsilon^3), \tag{9}$$

so κ_F is an infinitesimal-KL ratio; both denominators are p-weighted ($\bar{d} = \sum_i p_i d_i$). The reverse-KL field also satisfies $\mathbb{E}_p[d_r] = \nabla_z \mathrm{KL}(\pi \| \tilde{\pi})$ with $\tilde{\pi} \propto \exp(\sum_r p_r \log \pi_r^*)$ (a geometric-mean projection); for mirror targets under the uniform mixture this reduces to the entropy channel. The closed-form claims (V1/V2/V4) are pinned to machine precision and the KL-ratio form (V3) converges with ϵ in `experiments/common_basis/toy/verify_kl_kappa.py` (with a p-weighted denominator and a non-uniform-p regression test); the Euclidean κ used elsewhere is not claimed to equal κ_F.

C NUMBER PROVENANCE

Table 7: Provenance of the main-text numbers.

Result	File or script
Cancellation, soft-Q and AWR closed forms	`data/kappa/toy_fields/o1_closed_forms.json`
K-way cancellation, direction, interference	`data/kappa/toy_fields/kway_geometry.json`
Protocol audit (deterministic vs stochastic)	`data/kappa/toy_fields/det_vs_stoch.json`
Controlled visibility contrast (mirror bandit)	`data/kappa/toy_fields/visibility_control.json`
Overcooked field axis (shared episode batch)	`data/kappa/server_tasks/results/oc_field_axis.json`
Overcooked switch and memory protocol	`data/kappa/server_tasks/results/oc_switch_kappa.json`
510K paired protocol	`data/kappa/server_tasks/results/510k_field_axis.json`
Shared vs conditional capacity	`data/kappa/compromise_training/results.json`
Switching-task decoupling	`data/kappa/o4_adaptive/aggregate.json`
DICES-350 supervised check	`data/kappa/dices350/audit.json`
Fisher/KL variant verification	`experiments/common_basis/toy/verify_kl_kappa.py`
Canonical metric implementation	`src/iigc/metrics/kappa.py`

D CORRECTION LOG

Superseded claims and their replacements are recorded in `notes/paper1_propositions_corrections.md`, including the withdrawn soft-Q dynamics claim, the parameter-space scope of the Fisher/KL variant, and the replacement of S_3-specific language with K-way simplex geometry.