WARNINGS ARE NOT VERDICTS: COUPLING-ROBUST ACTION CERTIFICATION FOR CITATION VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A citation verifier can correctly identify a missing detail yet wrongly downgrade a partially supported claim to insufficient evidence. We study this gap between detecting a defect and deciding to change a verdict. Our Projected-Interaction Action Filter (PIAF) tests each proposed change before it is applied. Holding the candidate fixed, PIAF requires its score-based advantage over the original verdict to remain nonnegative across all Fréchet-consistent couplings of local support and missingness and a declared family of losses. We characterize the largest acceptance set satisfying this condition and reduce the test to one-dimensional optimization and finite loss vertices. Negative advantages yield explicit counterexamples. On held-out GenSearch, PIAF retains 650 of 1,083 proposals and improves cluster accuracy by 5.15 points. A retrospective comparison at the same action budget finds one harm to a correct Partial verdict versus 35–51 for five confidence, calibration, and learned controls. A separate Learn-then-Test gate addresses population risk under exchangeability. On fresh GaRAGe, confirmation yields 7 harms among 74 selected actions; the one-sided 95% upper bound is 17.03%, below the declared 20% target at 7.64% question coverage. Separately, warning review reduces retained error by 7.72 points at 12.8% review on protected GenSearch. The results show why useful diagnostics need an explicit test of the changes they recommend.

1 INTRODUCTION

A verifier inspects a claim containing an event, a date, and a location. Its source supports the event and date but omits the location. Flagging that omission is correct. Replacing a `partial_support` verdict with `insufficient` is not. The diagnostic found the defect but overstated its global severity.

Modern citation verifiers increasingly expose unsupported spans, claim atoms, or rationales (Min et al., 2023; Kamoi et al., 2023; Tang et al., 2024; Ru et al., 2024; Wanner et al., 2024). This helps because visible citations do not ensure claim-level support (Liu et al., 2023; Gao et al., 2023). Yet a deployed system must make a separate choice: keep its verdict, replace it, or request review.

Better defect detection alone does not resolve this decision. A valid warning can remain compatible with several global labels (Proposition 1). Review needs error enrichment; replacement requires corrections to exceed induced harms (Proposition 2). We call this the **diagnostic-to-decision gap**. The missing step is to test whether a proposed correction depends on an unverified interaction between diagnostic signals.

Our central contribution is **PIAF, a test for a proposed label change**. Imagine keeping the detected omission and the proposed verdict fixed, but changing how supported and missing details are combined. If an admissible combination makes the original verdict preferable, the change depends on an unverified assumption. PIAF searches for such counterexamples and accepts only when their absence is certified within the declared score and loss sets. A separate Learn-then-Test (LTT) gate tests selected-action harm on target calibration data under exchangeability. These checks answer different questions: whether an individual change survives coupling ambiguity, and whether selected changes meet a population harm target.

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.

PIAF uses established Fréchet bounds and robust optimization (Manski, 2003; Wiesemann et al., 2014). The contribution is their use to certify a fixed correction under unknown dependence between internal diagnostic scores. Citation evaluators estimate support or harm (Haan, 2025; Kartáč et al., 2026), while ALiiCE and SIDE improve diagnostics or evidence (Xu et al., 2025a; Petroni et al., 2023). Pooling and calibration estimate predictive distributions (Smith et al., 2005; Yao et al., 2023; Kull et al., 2019). Selective and conformal methods control population risk (Geifman & El-Yaniv, 2019; Mozannar & Sontag, 2020; Angelopoulos et al., 2024; Xu et al., 2025c; Bai & Jin, 2026). Our certificate complements these methods by testing within-instance score ambiguity before population-level validation. Safe policy improvement also compares actions against a baseline, but targets policy return in batch reinforcement learning (Laroche et al., 2019).

The empirical test is concrete: **does PIAF protect correct Partial verdicts even when competing filters make exactly as many changes?** At 650 changes, PIAF harms one such verdict versus 35–51 for five controls (Table 2, retrospective). Its total-harm and accuracy differences from those controls remain statistically unresolved. The distinctive gain is protection of the Partial boundary. The paper contributes the action-validity formulation, the tractable certificate with explicit witnesses, and separate evaluations of replacement, target risk, and review. Figure 1 explains the test; Table 1 summarizes the evidence.

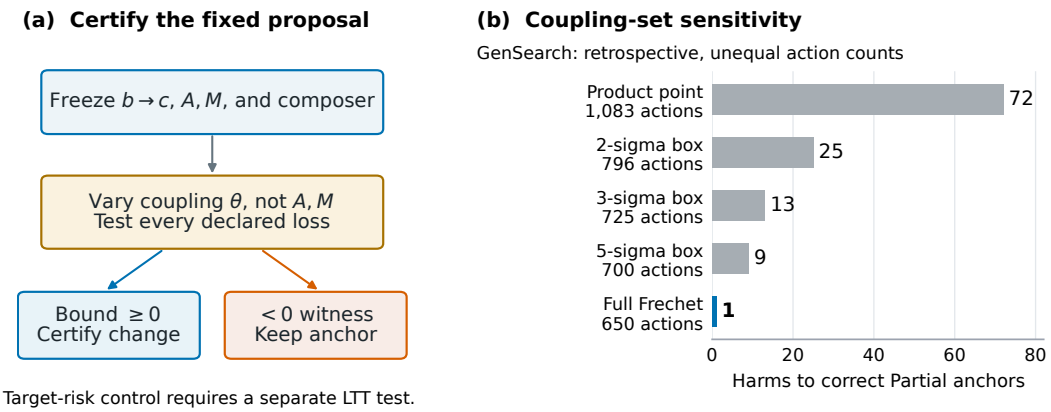


Figure 1: How PIAF tests a proposed correction. (a) The candidate is held fixed while the interaction between local gates varies. (b) An outcome-opened GenSearch replay compares the product point, restricted paths, and the full Fréchet path. Labels report selected actions; bars count harms to correct Partial anchors. The full path retains fewer actions and preserves more such anchors. Counts are descriptive and budgets differ; Table 2 isolates equal-budget comparisons. The boxes restrict standardized local features, not calibrated dependence.

2 Problem Formulation

2.1 Graded citation support

Given a claim c and fixed cited evidence e, we predict $Y \in \mathcal{R} = \{I, P, S\}$: insufficient evidence, partial support, or full support. The order $I < P < S$ denotes increasing supported claim mass, not a logical implication between labels. We study post-hoc support of the fixed pair. We do not estimate world truth or whether the source causally influenced generation (Wallat et al., 2025).

A global verifier returns $g(y \mid c, e)$. A local verifier estimates evidence coverage of fixed lexical spans from the encoder tokenizer, producing $\ell(y \mid c, e)$. Section 3.1 defines their construction, aggregation, and semantic scope; Appendix A gives implementation details.

The **anchor** is the original global verdict. The **candidate** is the proposer's recommended verdict. A **warning** is a proposed change that passes the directional rule in Section 3.3. **Replacement** emits the candidate, whereas **deferral** withholds the anchor for review. **Posterior advantage** is the composed probability of the candidate minus that of the anchor.

2.2 DIAGNOSTIC VALIDITY IS NOT ACTION VALIDITY

Proposition 1 (diagnostic non-identification). Fix a warning state $W = w$ with perfect precision: the diagnosed defect G is real, $\Pr(G = 1 \mid W = w) = 1$. If $\mathcal{C}(w)$ contains the compatible global relations, this precision alone leaves every distribution over $\mathcal{C}(w)$ possible. The best warning-only verdict has risk $1 - \max_y \Pr(Y = y \mid W = w)$, or $1 - 1/k$ for k equally likely compatible labels. A missing qualifier can therefore identify a defect without deciding between I and P. Appendix B proves the result; Appendix F audits the mechanism.

2.3 ACTION VALIDITY DEPENDS ON THE INTERVENTION

We call a diagnostic **action-valid** when its action improves the relevant risk relative to the anchor. Let E_b, E_c denote anchor and candidate errors, and $\mathcal{D}$ the selected set. Write $R_b = \Pr(E_b)$, $d = \Pr(\mathcal{D})$, and $r_D = \Pr(E_b \mid \mathcal{D})$.

Proposition 2 (action separation). For a fixed population and selection rule with $0 < d < 1$,

$$R_b - R_{\text{ret}} = \frac{d(r_D - R_b)}{1 - d},$$
$$R_{\text{hard}} - R_b = \Pr(\mathcal{D}, \neg E_b, E_c) - \Pr(\mathcal{D}, E_b, \neg E_c).$$

Review helps when the selected set enriches errors ($r_D > R_b$). Replacement instead requires corrections to exceed induced harms, the two probabilities in the second line. A warning can meet the first condition and fail the second. Appendix B gives a strict numerical example, proofs, and nonuniform review costs. PIAF addresses the stronger replacement requirement.

3 FROM LOCAL SCORES TO CERTIFIED CHANGES

3.1 HOLISTIC AND LOCAL VIEWS

The sealed primary system uses the same DeBERTa-v3-base cross-encoder for both views. Each view produces a different relation posterior. Evidence is greedily split at newline boundaries into 512-token claim–evidence pairs. We split overlong lines by tokenizer IDs and verify every final pair without truncation. The **global view** uses a mean-pooled three-way relation head. The **local view** averages token logits within each material claim unit, then pools that unit across evidence chunks. Both chunk and unit pooling use the normalized smooth maximum

$$s_\tau(v_1, \ldots, v_n) = \frac{1}{\tau} \log \left(\frac{1}{n} \sum_{j=1}^{n} e^{\tau v_j} \right), \qquad \tau = 12.$$

If q_i is the pooled support of material unit i, we aggregate supported and missing mass as

$$A = s_\tau(q_1, \ldots, q_m), \qquad M = s_\tau(1 - q_1, \ldots, 1 - q_m).$$

The local relation masses are

$$m_S = A(1 - M), \quad m_P = AM, \quad m_I = (1 - A)M,$$

followed by normalization.

Proposition 3 (maximum-entropy graded projection). Fix Bernoulli gate marginals A (salient support) and M (salient missing evidence). Their independent coupling uniquely maximizes joint entropy. Conditioning it away from the uninformative state $(0, 0)$ and mapping $(1, 0), (1, 1), (0, 1)$ to S, P, I yields the normalized masses above. The projection adds no parameter and preserves

$$\frac{m_P}{m_S} = \frac{M}{1 - M}, \qquad \frac{m_P}{m_I} = \frac{A}{1 - A}.$$

More missing evidence moves S toward P; more supported evidence moves I toward P. Appendix A proves the proposition. Independence applies only to the aggregate gates, not to units or their errors. The marginals do not, however, identify this coupling; Section 3.4 keeps the full admissible set. Training uses class-balanced binary losses on both axes and a 0.5-weight unit loss when WiCE token labels are available; Appendix A gives the full schedule.

Material units are deterministic tokenizer spans containing an alphanumeric, currency, or percent symbol. They approximate lexical coverage rather than semantic facts. For *The probe launched from Site A in 2012*, a missing *2012* raises missingness while the supported event keeps support high. In a development audit of 374 Partial rows, local relation accuracy is 90.6% when all unit labels are correct and 47.0% otherwise. Appendices A and E give the unit audit and post-hoc pooling sensitivity; Section 6 states the scope of certification when these scores are noisy.

3.2 A CONTROLLED PROPOSAL MODEL

Let $g, \ell \in \Delta^3$ be the global and local posteriors in the canonical implementation order (S, P, I). The support-strength order used by the policy remains $I < P < S$. We form six features

$$z = [\log(\max(g, \epsilon)); \log(\max(\ell, \epsilon))].$$

The maximum is elementwise. After feature standardization, a multinomial logistic model gives

$$q = \mathrm{softmax}(W\tilde{z} + b).$$

We call this 21-parameter model the *proposer* (RA-DVRC in the artifacts). It uses L2 regularization with $C = 0.3$ and no class reweighting (Smith et al., 2005). Holding this small model fixed lets us isolate the effect of PIAF. Appendix E retains the nonlinear and multi-family proposer comparisons, including their accuracy, boundary, and training-budget tradeoffs (Yao et al., 2023).

3.3 WHEN MAY THE CANDIDATE REPLACE THE ANCHOR?

Let $b = \arg\max_y g_y$ be the anchor and $c = \arg\max_y q_y$ the composed candidate. Define

$$a(x) = q_c(x) - q_b(x).$$

If $c = b$, the prediction is unchanged. Otherwise the proposer selects c only when

$$a(x) \geq \begin{cases} \tau_\uparrow, & c > b, \\ \tau_\downarrow, & c < b, \end{cases}$$

and retains b otherwise. Development selects $\tau_\uparrow = 0.20$ and $\tau_\downarrow = 0$. Section 4.1 summarizes selection; Appendix C specifies the grid and feasibility constraints. Neither threshold is retuned on the official test.

For any loss matrix L, define the posterior action advantage

$$a_L(b, c \mid x) = \sum_y q_y(x)\{L(b, y) - L(c, y)\}.$$

If $q_y(x) = \Pr(Y = y \mid X = x)$, then a_L is the conditional risk reduction from replacing b by c. The same interpretation holds when this equality is valid conditional on all information used to choose the action. Marginal calibration or low ECE alone is not sufficient. Zero–one loss recovers $q_c - q_b$. Since q is estimated, we evaluate induced actions directly. Appendices D and L report a simpler loss-based filter that rejects an existing action when $a_L < 0$.

Corollary 1 (target action margin). Let p_t be the target conditional posterior. If $\max_y |p_t(y \mid x) - q_y(x)| \leq \varepsilon_t(x)$, then the true benefit satisfies $p_t(c \mid x) - p_t(b \mid x) \geq a(x) - 2\varepsilon_t(x)$. Thus $a(x) > 2\varepsilon_t(x)$ suffices under zero–one loss. ECE does not identify this pointwise bound. We instead test target actions on reserved clusters (Section 3.6); Appendix B gives the derivation.

3.4 CERTIFYING COUPLING-ROBUST REPLACEMENT

PIAF holds the proposal fixed and minimizes its advantage over admissible couplings and losses. A nonnegative minimum certifies replacement; a negative optimizer supplies a reversing witness.

Applicability. Freeze the anchor, candidate, gate scores, global outputs, fitted composer, feature scaling, and loss family. Only the local coupling varies. Each frozen q_f must depend only on its own coupling θ_f. Shared fixed weights or scaling are allowed; recomputing a cross-fold interaction requires joint optimization rather than Algorithm 1.

The maximum-entropy lattice sets joint gate mass $\theta = AM$, but independence is a modeling choice rather than an identified fact. We treat the neural scores A and M as marginals of two score-space gates. Classical fixed-margin bounds leave only the Fréchet interval (Fréchet, 1951; Manski, 2003; Polo et al., 2024)

$$\Theta(A, M) = [\max(0, A + M - 1), \min(A, M)].$$

The interval is sharp given only A, M; narrowing it requires additional dependence information. Larger θ assigns more mass to Partial without changing either gate. PIAF tests the fixed candidate against the anchor across this interval. It does not choose a new minimax verdict.

For any $\theta \in \Theta(A, M)$ with $A + M - \theta > 0$, the three informative gate states give the local posterior

$$\ell_\theta(S, P, I) = \frac{(A - \theta,\ \theta,\ M - \theta)}{A + M - \theta}.$$

For fold f, substitute $\ell_{f,\theta}$ for ℓ_f and recompute $q_f(\theta)$ without refitting. Product coupling $\theta = AM$ recovers Section 3.1. If $A = M = 0$, keep the anchor: no informative local distribution exists. Otherwise the full path preserves A, M while redistributing local mass across S, P, I.

For a predeclared loss family $\mathcal{L} = \text{conv}\{L^{(1)}, \ldots, L^{(K)}\}$ and action $b \to c$, allow each fold's coupling to vary within its own interval, $\mathcal{K} = \prod_f \Theta_f$. Define

$$\underline{a}_{\Theta,\mathcal{L}}(b, c) = \min_k \frac{1}{F} \sum_f \inf_{\theta_f \in \Theta_f} \sum_y q_{f,y}(\theta_f)\{L^{(k)}(b, y) - L^{(k)}(c, y)\}.$$

$\mathcal{K}$ is a deterministic sensitivity product, not fold-error independence. The infimum separates by fold under the applicability conditions above. Acceptances remain valid for any linked subset. Appendix A proves the reduction.

Certificate semantics. The composer is trained at product coupling. Off-product outputs are sensitivity scores, not calibrated posteriors. PIAF checks their action sign over the path rather than assuming it is stable. Appendix A gives a computable bound on score displacement and parameter perturbations; Appendix G reports path and trust-region audits.

Loss contract. Each experiment declares its loss set before labels open. GenSearch uses zero–one loss. For graded relations, L_1 charges crossing $I \mid \{P, S\}$ and L_2 charges crossing $\{I, P\} \mid S$. Every normalized weighted absolute ordinal loss is exactly $wL_1 + (1 - w)L_2$, $w \in [0, 1]$. Thus the two vertices exhaust this family. Arbitrary matrices require new vertices and protected evaluation. Appendix G gives a post-hoc richer-cost sensitivity (Kartáč et al., 2026). Adding a loss vertex can only shrink the certified set; costs must be declared before outcomes open.

Fix $(b, c, \{q_f\}, \{\Theta_f\}, \mathcal{L})$. A deterministic filter is **contract-sound** if it authorizes only actions with $\underline{a}_{\Theta,\mathcal{L}}(b, c) \geq 0$.

Proposition 4 (largest valid acceptance set). PIAF keeps an action iff $\underline{a}_{\Theta,\mathcal{L}} \geq 0$. It is contract-sound and has the largest acceptance set among deterministic filters for the fixed proposer and contract. Fold separability gives the inner infima; linearity in L places the worst loss at a polytope vertex. This reduction applies to losses, not coupling endpoints: $q_f(\theta)$ is nonlinear and can have an interior worst case. A negative optimizer witnesses why every contract-sound filter must reject; otherwise PIAF accepts. This certifies the score calculation, not posterior calibration or target-population safety. Appendix A gives the proof and the distinction from guarantees that hold only in expectation.

Closed-form witness (illustration). Take one fold with $A = M = .6$, anchor S, and toy composer $q_y \propto \ell_y^6$. Then $\Theta = [.2, .6]$. At product coupling $\theta = .36$, the candidate is P with confidence .851. Under zero–one loss its advantage is

$$d(\theta) = \frac{\theta^6 - (.6 - \theta)^6}{\theta^6 + 2(.6 - \theta)^6}.$$

The sign is positive exactly when $\theta > .3$. Thus $d(.36) = .776$, yet $d(.2) = -.488$: PIAF rejects despite high nominal confidence. Nothing about the gates or candidate changed. Appendix A derives the example; Appendix G shows real cases where rejection prevents a harm or loses a correction.

Algorithm 1: PIAF for one fixed proposal

Input Anchor b, candidate c, frozen fold maps q_f, gates A_f, M_f, and loss vertices $L^{(k)}$.

1 Form each Fréchet interval Θ_f. Keep b if a local view is uninformative.

2 For each (f, k), bound $\inf_{\theta \in \Theta_f} d_{f,k}(\theta)$ by $[l_{f,k}, u_{f,k}]$ using interval branch-and-bound. Stop at gap $\leq 10^{-6}$ or 16,384 boxes (bounds in Appendix A).

3 Set $l = \min_k F^{-1} \sum_f l_{f,k}$ and $u = \min_k F^{-1} \sum_f u_{f,k}$.

4 If $l \geq 0$, certify c. If $u < 0$, keep b with the witnessing loss and fold couplings. Otherwise keep b as numerically unresolved.

Here $d_{f,k}(\theta) = \sum_y q_{f,y}(\theta)\{L^{(k)}(b, y) - L^{(k)}(c, y)\}$. Algorithm 1 reproduces all 1,083 archived decisions at 10.28 actions/s on one M4 Max CPU, excluding neural inference. Appendix A details the bounds and provides a runnable witness example. Section 3.6 describes the complementary population test.

3.5 Prioritizing review

A warning may also help prioritize inspection. We evaluate **warning-priority review** (WCSD in the artifacts) on the full warning set:

$$D_W(x) = \begin{cases} \texttt{defer}, & h_{\text{prop}}(x) \neq h_g(x), \\ h_g(x), & \text{otherwise.} \end{cases}$$

Here h_{prop} includes the directional rule. The review mask adds no trained parameter or tuned threshold and gives one native operating point. For other review budgets, a fixed **warning-priority score** ranks frozen actions first, then uses fold action agreement f and advantage a as tie breakers: $s_W = 4 \times \mathbf{1}\{\text{action}\} + f + 0.25a$. The coefficient gaps implement the lexicographic order. On TREC, this ranking is locked before protected labels. At equal row budgets we compare confidence, entropy, margin, holistic-fold Jensen–Shannon disagreement, a GenSearch-trained error selector, and restricted S-MMCE. These experiments evaluate full warnings, not only PIAF rejections. Appendix G separately reports the retrospective certify-or-defer cascade without human-oracle credit.

3.6 Controlling selected-action harm

PIAF's pointwise contract and population-risk control answer different questions. PIAF tests one action against declared within-row ambiguity. Proposition 5 (Appendix M) composes it with Learn-then-Test (LTT) without conflating guarantees: LTT controls the risk of a fixed cluster action under exchangeability, while every released action retains its pointwise PIAF certificate (Angelopoulos et al., 2025). SCRC or SCoRE can occupy this outer population-control layer when their labeled calibration protocols are available; neither replaces the inner coupling certificate (Xu et al., 2025c; Bai & Jin, 2026).

Our graded evaluation tests one fixed PIAF action per selected GaRAGe question against 20% conditional harm (Sorodoc et al., 2025). Section 4.4 and Appendix X specify the independent risk/confirmation halves. Appendix M gives the general proof plus SciCiteVal and binary AVeriTeC extensions (Liu et al., 2026; Schlichtkrull et al., 2023).

A retrospective GenSearch PIAF–CRC intersection illustrates the same division of labor but removes corrections as well as harms (Appendix H).

4 EXPERIMENTAL DESIGN

The experiments answer four questions: does the proposer help, does PIAF protect decisions, does the selected policy meet a target harm bound, and do warnings help review? Table 1 maps outcomes to those questions. The supplement opens with a protocol map showing what was frozen before each label release and which comparisons are retrospective. No confirmation split revises a policy.

Metric contract. Replacement uses cluster-macro accuracy: compute row accuracy inside each complete query, title, or source cluster, then weight clusters equally. C/H/N denotes corrections, harms, and neutral changed decisions; action harm is $H/(C + H + N)$. Review precision is the anchor-error fraction among deferred rows, while retained risk is anchor error among rows left automatic. Headline accuracy and review intervals use 10,000 bootstrap resamples of complete declared clusters. GaRAGe's selected-action harm bound instead uses a one-sided exact Clopper–Pearson interval. Appendix protocols state the unit, resampling settings or exact test, and whether an analysis is confirmatory or test-opened.

4.1 PROPOSAL AND FROZEN TRANSFER

WiCE supplies three-way relations and unsupported-token spans (Kamoi et al., 2023). Three title-disjoint outer folds train on 3,470 units; inner grouped folds fit the composer. Only the predeclared $(\tau_{\uparrow}, \tau_{\downarrow}) = (.20, 0)$ rule meets all development gates. Predictions and gates are sealed before the one-time 358-unit test. The frozen system then processes 186 ALCE rows; two blind labels and third-person adjudication follow its 28 actions. A separate 100-row audit demonstrates the hard-router mechanism, not its prevalence.

4.2 HELD-OUT GENSEARCH CERTIFICATION TEST

GenSearch maps directly to $S/P/I$ (Liu et al., 2023). Query-disjoint `dev-fit`, `dev-cal`, and official-test splits fix review policies and controls. PIAF was designed after those outcomes opened. Its confirmation therefore uses a source partition never used by an encoder, composer, calibrator, threshold, or method-design decision. The source release calls this partition `train`, but its 8,834 rows form 1,156 exact-query clusters disjoint from all development and test queries. Six model outputs, 1,083 proposals, certificates, metrics, and five gates were serialized before its labels could be accessed. Table 2 later compares filters on these same 1,083 proposals. Its controls were specified after outcomes were known, so it is a retrospective comparison, not a second independent confirmation; no confirmation label trains them.

4.3 EXTERNAL REVIEW EVALUATION

TREC RAG 2024 uses nine calibration and 13 disjoint assessor-blind topics (4,724 pairs); its gates, predictions, and review masks are frozen, while the later TREC-native selector is retrospective (Pradeep et al., 2024; National Institute of Standards and Technology, 2025). The supplement groups additional datasets by the question they test. Those breadth studies include failed replacement gates and do not change the primary results.

4.4 TARGET-DOMAIN ACTION-RISK CONFIRMATION

GaRAGe provides long-form answers and grounding passages over 2,366 questions (Sorodoc et al., 2025). Without opening its evaluation fields, we lock a 1,939-question cohort, a 50/50 split, all model outputs, and at most one action per question before new $S/P/I$ labels exist. Two blinded annotators label every action; a third person resolves disagreements. We evaluate unchanged point PIAF behind the outer LTT gate. A stricter marginal-band variant fails its label-free activation minimum and remains a negative result (Appendix X).

5 RESULTS

5.1 WHAT EACH EXPERIMENT ESTABLISHES

Table 1 maps each held-out population to one claim; no estimates are pooled.

WiCE's development–test gap. Partial support rises from 13.6% of development rows to 60.1% of test rows, while anchor recall for that class is 22.7% and 21.9%. Weighting development class recalls by the test label mixture gives 48.09% row accuracy, close to the observed 47.21% (development: 76.71%). Thus the label mixture accounts for most of the aggregate drop. This is a post-hoc decomposition, not proof that all shift is label shift; Appendix E also distinguishes row weighting from title-cluster weighting.

Table 1: Separate tests of proposal, certification, target harm, and review. Accuracy gains and retained-risk reductions are percentage points; brackets give cluster-bootstrap 95% intervals. C/H/N denotes corrections/harms/neutral changes. GaRAGe U95 is an exact one-sided harm upper bound.

Action / cohort	Selected actions	Outcome and scope
Proposal / WiCE	33/358 replacements	+8.51 [5.44,11.73]; 30 C/1 H. One-time in-domain test.
Transfer / ALCE	28/186 replacements	+7.46 [1.33,13.83]; 20 C/8 H. Observed harm exceeds the 20% ceiling.
Certificate / GenSearch	650/1,083 proposals	+5.15 [4.43,5.90]; 489 C/54 H; 1 correct-P harm. Frozen PIAF confirmation.
Outer risk / GaRAGe	71 risk, then 74 protected actions	53 C/7 H/14 N; harm U95=17.03%. Locked 20% target passes at 7.64% question coverage.
Review / GenSearch	140/1,097 rows	7.72 [6.06,9.60] risk reduction. Full-warning review passes.
Review / TREC, 10%	472/4,724 rows	3.61 risk reduction; 82.6% precision. Protected external review passes.

GaRAGe risk calibration releases confirmation (56/5/10 C/H/N; $p = .00231$). Protected confirmation passes with 53/7/14, 9.46% harm, and U95=17.03%. Its 74 actions cover 7.64% of all clusters and 17.6% of proposal-bearing questions under the one-per-question rule. This is conditional action risk, not overall accuracy (Appendix X).

The ALCE contrast motivates the distinction: its observed 8/28 harm rate exceeds the declared 20% acceptance ceiling. This fails the empirical gate; it does not establish that population harm exceeds 20%. All eight harms cross Partial, and raising the demotion threshold from 0 to .20 removes eight corrections but no harm in the post-hoc sweep. Appendix G gives the complete transfer diagnosis; higher confidence alone does not recover authority in this cohort.

5.2 EQUAL ACTION BUDGETS EXPOSE BOUNDARY PROTECTION

Same proposals, same action budget. At 650 changes, PIAF harms one correct P anchor versus 35–51 for controls. It blocks 71/72 proposed harms to these anchors and keeps 61 corrections into P. Paired correct-P harm intervals exclude zero; paired accuracy and total-harm differences remain unresolved. This establishes boundary protection; Appendix G gives the paired intervals and correction–harm utility tradeoff.

Table 2: Retrospective GenSearch comparison on the same 1,083 proposals, selecting 650 actions each. Extra params are additional trained parameters; P harms are changes that damage a correct Partial verdict. This is a paired comparison, not a new confirmation cohort.

Action layer	Extra params	C/H/N	Acc. gain	P harms
Fréchet PIAF	**0**	**489/54/107**	**+5.15**	**1**
Candidate confidence	0	530/55/65	+5.56	39
Raw advantage	0	527/57/66	+5.51	37
Log-posterior calibration	12	529/59/62	+5.57	37
Learned harm selector	25	531/59/60	+5.60	35
Restricted S-MMCE	417	476/84/90	+4.84	51

Which errors are prevented? PIAF's 54 residual harms comprise 28 gold-I, one gold-P, and 25 gold-S cases. Appendix G separates boundary protection from off-product extrapolation. Median CPU cost is 22.4 ms/action (P95: 706.6 ms; safeguards in Appendix A).

Coverage tradeoff. On the opened outcomes, the 5- and 3-unit boxes recover 50 and 75 actions but raise correct-P harms from one to nine and 13 (Figure 1); they are sensitivity assumptions, not learned dependence. PIAF retains 489/775 possible corrections (63.1%) and blocks 87/141 proposed harms. Proposals cover 28.7% of anchor errors. Appendix E reports capacity studies; Table 2 holds the proposer fixed.

Threshold and loss sensitivity. Across two eight-point, test-opened threshold sweeps, PIAF certifies 382–726 actions and gains +2.91 to +5.55 points. It retains one correct-P harm at every setting; matched confidence and advantage controls incur 18–47 such harms. Adding both ordinal loss vertices shrinks the set to 118 but raises conditional harm to 22.03%. Stricter score restrictions need not lower population harm. Appendix G reports every setting; none changes the confirmed policy.

5.3 Warnings help prioritize review

Reviewing the full warning set cuts protected GenSearch risk by 7.72 points ([6.06, 9.60]) at 12.8% review. On the 4,724-row TREC cohort, warning priority lowers retained risk by 0.89 points relative to the source-trained selector and 1.34 over restricted S-MMCE at the locked 10% budget. Differences from generic disagreement and the retrospective TREC-native selector are unresolved at this budget (Appendices J and L).

TREC and SALAD replications span three encoder families and favor warning priority over scalar uncertainty, not every learned selector (Appendices R and W).

6 Limitations

PIAF certifies a frozen score calculation. It cannot correct biased local scores or semantic errors in lexical units, and off-product scores need not be calibrated. Loss and coupling sets depend on the application. LTT requires its sampling and exchangeability assumptions. Coverage depends on the proposer. The 33 WiCE and 28 ALCE changes give imprecise harm-rate estimates; the larger GenSearch study tests certification, and GaRAGe separately tests target harm. Mixed cross-backbone and cross-dataset results do not establish universal transfer. Appendix I details these limits.

7 Conclusion

PIAF tests whether a proposed correction survives unknown dependence between local scores. It protects correct Partial verdicts on GenSearch at matched action budgets. GaRAGe separately tests target harm under LTT's assumptions, while warning-based review reduces retained error. The practical principle is to evaluate a diagnostic for the decision it will control.

ETHICS STATEMENT

The study uses public claim–evidence benchmarks in English and a multilingual stress test; some contain names and news content. Human reviewers label ALCE and GaRAGe support relations independently, without model-prediction prefills. They consent to anonymous aggregate and row-level reporting. Public artifacts use role aliases and exclude names, affiliations, contact details, and private consent records. We do not redistribute text-bearing benchmark data; users reconstruct it from official sources. The study received institutional ethics approval.

AI USE STATEMENT

The manuscript combines author-written text and revisions with AI-assisted drafting and editing of portions of the text. Generative AI tools also assisted hypothesis refinement, method and experiment design, derivations and proof development, software implementation and testing, public-data cleaning and formatting, literature search and summarization, statistical analysis and interpretation, and figure and table preparation. They did not generate benchmark labels or reported human-audit judgments. Human annotators provided those judgments independently, without AI-generated label prefills. The authors take responsibility for the final text, mathematical claims, code, citations, and results.

REPRODUCIBILITY STATEMENT

The appendix records dataset construction, model revisions, inference settings, grouped folds, training schedules, frozen gates, audit rules, and negative-result boundaries. The supplement provides the executable pipeline, lock manifests, hashes, text-free predictions, role-aliased action labels, and verification scripts. After users reconstruct official public data, these scripts recompute the reported metrics. Prospective label keys are joined only after prediction-lock verification, and stopped development studies retain their original decisions. For the matched-budget action-layer comparison, a separate implementation reconstructs every score and selected action mask from the serialized parameters before recomputing the table and paired intervals.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10: 746–763, 2022. doi: 10.1162/tacl_a_00486. URL https://aclanthology.org/2022.tacl-1.43/.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026. doi: 10.48550/arXiv.2603.24704. URL https://arxiv.org/abs/2603.24704.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. Understanding retrieval augmentation for long-form question answering. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=j3AAkO5xgr.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. MedFact: A large-scale Chinese dataset for evidence-based medical fact-checking of LLM responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32340–32353, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1646. URL https://aclanthology.org/2025.emnlp-main.1646/.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083, 2022. doi: 10.1162/tacl_a_00506. URL https://aclanthology.org/2022.tacl-1.62/.

Adam Fisch, Tommi S. Jaakkola, and Regina Barzilay. Calibrated selective classification. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=zFhNBs8GaV.

Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Section A*, 14:53–77, 1951.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Sebastian Haan. SemanticCite: Citation verification with AI-powered full-text analysis and evidence-based reasoning. *arXiv preprint arXiv:2511.16198*, 2025. doi: 10.48550/arXiv.2511.16198. URL https://arxiv.org/abs/2511.16198.

Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6313–6336. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.320. URL https://aclanthology.org/2025.naacl-long.320/.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7561–7583, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.470. URL https://aclanthology.org/2023.emnlp-main.470/.

Ivan Kartáč, Jan Tovarys, Mateusz Lango, and Ondřej Dušek. VetScore: Risk-weighted fact verification for veterinary long-form QA with citations. *arXiv preprint arXiv:2608.03675*, 2026. doi: 10.48550/arXiv.2608.03675. URL https://arxiv.org/abs/2608.03675.

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTaiL: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12022. URL https://ojs.aaai.org/index.php/AAAI/article/view/12022.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 163–181, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.12. URL https://aclanthology.org/2024.findings-naacl.12/.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14919–14935, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.886. URL https://aclanthology.org/2024.findings-acl.886/.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467/.

Qinyue Liu, Yongxin Zhou, and Cyril Labbe. SciCiteVal: A multi-domain dataset for scientific citation verification. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 1603–1611, Palma de Mallorca, Spain, 2026. ELRA Language Resource Association. doi: 10.63317/4m84m2k77g97. URL https://aclanthology.org/2026.lrec-1.125/.

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. XRAG: Cross-lingual retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15669–15690, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.849. URL https://aclanthology.org/2025.findings-emnlp.849/.

Yining Lu, Noah Ziems, Hy Dang, and Meng Jiang. Optimizing decomposition for optimal claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5095–5114. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.254. URL https://aclanthology.org/2025.acl-long.254/.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems 31*, pp. 6150–6160, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3025–3045, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.167. URL https://aclanthology.org/2024.naacl-long.167/.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, NY, 2003. doi: 10.1007/b97478. URL https://doi.org/10.1007/b97478.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam Hruschka. FactLens: Benchmarking fine-grained fact verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18085–18096, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.929. URL https://aclanthology.org/2025.findings-acl.929/.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

National Institute of Standards and Technology. TREC 2024 RAG search track data, 2025. URL https://trec.nist.gov/data/rag2024.html. Accessed 2026-08-13.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Michele Bevilacqua, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. Improving Wikipedia verifiability with AI. *Nature Machine Intelligence*, 5:1142–1148, 2023. doi: 10.1038/s42256-023-00726-1. URL https://doi.org/10.1038/s42256-023-00726-1.

Felipe Maia Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135704–135740, 2024. doi: 10.52202/079017-4309. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f4c6bec746b0aeca8c2cd15096f1ad1f-Abstract-Conference.html.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track. *arXiv preprint arXiv:2406.16828*, 2024. doi: 10.48550/arXiv.2406.16828. URL https://arxiv.org/abs/2406.16828.

Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. VeriCite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 47–54. ACM, 2025. doi: 10.1145/3767695.3769505. URL https://doi.org/10.1145/3767695.3769505.

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*, 2026. doi: 10.48550/arXiv.2605.03534. URL https://arxiv.org/abs/2605.03534.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 37*, pp. 21999–22027, 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract-Datasets_and_Benchmarks_Track.html.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-2842. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/cd86a30526cd1aff61d6f89f107634e4-Abstract-Datasets_and_Benchmarks.html.

Andrew Smith, Trevor Cohn, and Miles Osborne. Logarithmic opinion pools for conditional random fields. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 18–25, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219843. URL https://aclanthology.org/P05-1003/.

Ionut Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. GaRAGe: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17030–17049, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.875. URL https://aclanthology.org/2025.findings-acl.875/.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22184–22202, 2022. URL https://proceedings.mlr.press/v162/verma22c.html.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, pp. 22–32. ACM, 2025. doi: 10.1145/3731120.3744592.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 153–175. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.starsem-1.13. URL https://aclanthology.org/2024.starsem-1.13/.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014. doi: 10.1287/opre.2014.1314. URL https://doi.org/10.1287/opre.2014.1314.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. ALiiCE: Evaluating positional fine-grained citation generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 545–561, Albuquerque, New Mexico, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.23. URL https://aclanthology.org/2025.naacl-long.23/.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. CiteEval: Principle-driven citation evaluation for source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32759–32778. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.1574. URL https://aclanthology.org/2025.acl-long.1574/.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025c. doi: 10.48550/arXiv.2512.12844. URL https://arxiv.org/abs/2512.12844.

Yuling Yao, Luiz Max Carvalho, Diego Mesquita, and Yann McLatchie. Locking and quacking: Stacking Bayesian model predictions by log-pooling and superposition. *arXiv preprint arXiv:2305.07334*, 2023. doi: 10.48550/arXiv.2305.07334. URL https://arxiv.org/abs/2305.07334.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Reader's guide. The main contribution can be checked through three routes: Appendices A–B specify and prove the certificate; Appendix G verifies the GenSearch result and its retrospective equal-budget comparison; Appendix X reports the independent GaRAGe risk test. Appendix E explains the WiCE development–test gap, and J and L contain the main review comparisons. These sections provide the evidence for the main paper without requiring a reading of every breadth study.

The remaining sections preserve implementation details, alternative proposers, and positive and negative stress tests. Appendix I collects their limitations; K, N–W, and Y examine transfer, target definitions, or representation changes. They are supporting evidence rather than additional conditions needed to understand PIAF. Historical version IDs identify reproducible artifacts, not extra components of the proposed method. RA-DVRC denotes the proposer, WCSD the warning-priority review rule, and PIAF-LTT the certificate with an outer risk test. The main paper uses the descriptive names to keep these roles clear.

Protocol map: what was fixed, and when? Each row below answers a distinct question. Table 1 in the main text reports the corresponding outcomes. A source release's `train` label does not mean that its rows trained our model; the GenSearch confirmation partition was held out from the entire design process. Review tests use full warning sets, not only PIAF rejections.

Question / cohort	Fixed before evaluated labels opened	Evaluation role
Proposal / WiCE	Encoders, composer, directional thresholds, predictions, and gates	One-time in-domain test; no test fitting.
Transfer / ALCE	Source system, actions, and transfer gate	New blind labels for all 28 changed rows; tests proposal transfer, not fresh PIAF.
Certificate / GenSearch	PIAF rule, outputs, proposals, certificates, and five gates	Fresh query-disjoint source partition, unused in training or design.
Target risk / GaRAGe	Policy, question split, one-action rule, and 20% harm target	Risk-half labels test release eligibility for the same fixed policy.
Confirmation / GaRAGe	The same policy and all 74 protected actions	Disjoint question labels test the bound and require corrections to exceed harms.
Review / GenSearch and TREC	Warning scores, review budgets or masks, and primary controls	Held-out review tests; no assumed certify-or-defer cascade.
Retrospective comparisons	Archived cohort and predictions; analyses specified after outcomes	Matched-budget controls, threshold/path sweeps, and TREC-native selector. Not new confirmation.

Here, *fixed* means unchanged for that evaluation, not fixed before every earlier experiment. PIAF was designed after WiCE, ALCE, and the earlier GenSearch review outcomes, then evaluated on its fresh partition. GaRAGe separates risk calibration from confirmation without refitting the policy. The detailed sections below state clustering units, labels, and statistical tests. The map introduces no additional lock or evaluation.

A Method and Certification Details

A.1 Canonical relation representation

All serialized posterior vectors use the implementation order (S, P, I). The three positions denote support, partial support, and insufficient evidence. The directional policy separately uses the semantic order $I < P < S$. Keeping these conventions explicit matters. Vector position determines feature alignment; semantic order determines whether an action is a promotion or demotion.

A.2 Deterministic material units and local targets

A material unit is a tokenizer-consistent lexical span, not a generated fact or semantic proposition. We run the pinned DeBERTa fast tokenizer on the claim without special tokens or truncation and retain every nonempty offset whose text contains an alphanumeric character or a dollar, pound, euro, yen, or percent symbol. Punctuation-only offsets are discarded. Overlapping offsets are merged. During paired claim–evidence encoding, a claim subword is assigned to the unique unit whose character

interval it overlaps. Evidence tokens, special tokens, and punctuation-only claim tokens are masked. The implementation rejects a row if any retained unit is absent after pair encoding.

Local labels follow three deterministic rules. Every unit in a support row has target 1. For an annotated partial-support row, a unit has target 0 if it overlaps any official unsupported-token span and target 1 otherwise. Insufficient rows and partial-support rows without token annotations receive no unit loss. This masking avoids pretending that a global insufficient label says which lexical span failed. Class balancing is computed over the observed unit targets only.

For example, consider the illustrative claim *The probe launched from Site A in 2012* and evidence that states the launch and site but omits the year. The tokenizer supplies fixed lexical units, including *probe*, *launched*, *Site*, *A*, and *2012*. A low support score for *2012* and high scores for the other content units make both A and M large. The lattice therefore exposes a partial-support region. The example explains the computation; it is not a benchmark row or an extra annotation.

The frozen WiCE train OOF outputs permit a direct noise audit on all 374 annotated partial-support rows from 304 title clusters. At threshold .5, the full local branch has unsupported-unit precision 57.1%, recall 65.7%, and F1 61.1%. Exact unit maps occur in 53 rows. Partial-support relation accuracy is 90.6% on those rows and 47.0% on rows with at least one unit error. The gap is 43.53 points, with title-cluster bootstrap interval [33.16, 52.97]. Conversely, mean unit error is 17.6% when the relation is correct and 38.8% when it is wrong. These are descriptive OOF associations, not a causal segmentation experiment.

Table 3: Frozen WiCE train OOF material-unit audit on 374 annotated partial-support rows. Unit metrics use threshold .5. Relation recall and cluster accuracy evaluate the local relation posterior. This is descriptive development evidence.

Local training	Unit precision	Unit recall	Unit F1	Partial recall	Cluster accuracy
Full unit loss	57.12	65.71	61.12	52.97	75.39
No unit loss	51.31	66.69	58.00	46.19	75.51

Unit supervision improves partial-support recall by 6.78 points and unit F1 by 3.12 points, but it does not improve aggregate cluster accuracy by itself. This is why the local lattice is treated as a noisy complementary view rather than a replacement verifier. The aggregate record and source hashes are reproduced by `scripts/analyze_v23_material_unit_noise.py`.

The global branch and local branch use the same DeBERTa-v3-base cross-encoder family and the same deterministic claim–evidence chunks. The global branch returns a three-way relation posterior g. The local branch averages token logits within a material unit. It uses the normalized smooth maximum

$$s_\tau(v_1, \ldots, v_n) = \tau^{-1} \log \left(n^{-1} \sum_{j=1}^{n} \exp(\tau v_j) \right), \qquad \tau = 12,$$

first across chunks for each unit and then across material units. If q_i is the resulting support of unit i, the two lattice axes are

$$A = s_\tau(q_1, \ldots, q_m), \qquad M = s_\tau(1 - q_1, \ldots, 1 - q_m).$$

Its unnormalized relation masses are

$$m_S = A(1 - M), \qquad m_P = AM, \qquad m_I = (1 - A)M.$$

This form has a maximum-entropy interpretation. Let U_A and U_M be binary aggregate gates with marginals $\Pr(U_A = 1) = A$ and $\Pr(U_M = 1) = M$. Among all joint distributions with those marginals, the entropy maximizer is the independent product distribution. Condition it on $E = \{(1, 0), (1, 1), (0, 1)\}$ and map these states to S, P, I. For $Z = A + M - AM$, the resulting conditional probabilities are

$$\Pr(S \mid E) = \frac{A(1-M)}{Z}, \quad \Pr(P \mid E) = \frac{AM}{Z}, \quad \Pr(I \mid E) = \frac{(1-A)M}{Z}.$$

These are exactly the normalized lattice masses. The product is a least-committal coupling of two soft aggregate axes, not an independence claim about material units, token labels, or their errors. As τ approaches zero, s_τ approaches the arithmetic mean; as τ grows, it approaches the maximum. Every selected claim–evidence view fits the fixed 512-token input; no selected evidence is silently truncated.

For completeness, let U, V be Bernoulli variables with fixed means A, M. Every coupling satisfies

$$H(U,V) = H(U) + H(V) - I(U;V) \le H(U) + H(V),$$

with equality exactly when $I(U;V) = 0$. Hence the product coupling uniquely maximizes entropy, up to degenerate marginals where the same product remains the only feasible maximizer. Removing the null state $(0,0)$ and renormalizing gives the displayed relation masses. This is a maximum-entropy projection from two aggregate scores; it is not a generative claim about how evidence defects arise.

A.3 Score-space interaction uncertainty

The maximum-entropy projection is a point choice, not an implication of the two marginals. Write $\theta = \Pr(U_A = 1, U_M = 1)$. The complete Frechet class is

$$\Theta(A, M) = [\max(0, A + M - 1), \min(A, M)].$$

Every $\theta \in \Theta$ with $A + M - \theta > 0$ induces informative masses

$$m_S(\theta) = A - \theta, \qquad m_P(\theta) = \theta, \qquad m_I(\theta) = M - \theta,$$

which are normalized by $A + M - \theta$. The excluded degenerate point $A = M = \theta = 0$ has no informative relation mass and is rejected by the implementation. The original lattice is the member $\theta = AM$. Frechet sets have also been used to partially identify weak-label performance from fixed marginals (Polo et al., 2024); here the unidentified quantity is instead the coupling inside a decision feature. We use the Fr'echet terminology only for this score-space interaction; it does not partially identify the unknown target-label posterior.

For fold f, let $q_f(\theta_f)$ be the output of its frozen 21-parameter composer after substituting the corresponding local posterior. For any fixed action-by-truth loss matrix L, put $d_{f,L}(\theta_f) = \sum_y q_{f,y}(\theta_f)\{L(b,y) - L(c,y)\}$. The final posterior is the fold average. Therefore

Lemma 1 (bounded composer displacement). Let $z_f(\theta) = [\log(\max(g_f, \epsilon)); \log(\max(\ell_{f,\theta}, \epsilon))]$, where the maximum is elementwise, let D_f contain the positive frozen feature scales, and define

$$\rho_f(\theta, \theta_0) = \left\| D_f^{-1}\{z_f(\theta) - z_f(\theta_0)\} \right\|_2.$$

For $r_y = L(b,y) - L(c,y)$,

$$|d_{f,L}(\theta) - d_{f,L}(\theta_0)| \le \frac{1}{2}\|r\|_2 \|W_f\|_2 \rho_f(\theta, \theta_0).$$

Proof. Frozen standardization gives logit displacement at most $\|W_f\|_2 \rho_f$. The softmax Jacobian is $\mathrm{diag}(q) - qq^\top$. For a unit vector v its quadratic form is $\mathrm{Var}_q(v_y) \le (\max_y v_y - \min_y v_y)^2/4 \le 1/2$; hence softmax is globally $1/2$-Lipschitz in Euclidean norm. Cauchy–Schwarz gives the result. The log floor $\epsilon > 0$, positive feature scales, and the fail-closed removal of $A = M = 0$ make every term finite and continuous on the compact Fr'echet interval. The lemma bounds movement of the frozen score; it does not assert that off-product outputs are calibrated posteriors.

Closed-form witness and marginal sensitivity. The illustration in Section 3.4 uses one fold, $A = M = 3/5$, a fixed anchor S, and $q_y(\theta) = \ell_y(\theta)^6 / \sum_j \ell_j(\theta)^6$. This is a special log-linear composer with coefficient six on each matching local log score, zero global coefficients, and no log floor; endpoint scores are defined by continuity. It is a constructed example, not a fitted model or benchmark result.

The common normalizer of ℓ_θ cancels. Hence the $S \to P$ advantage is

$$d(\theta) = \frac{\theta^6 - (3/5 - \theta)^6}{\theta^6 + 2(3/5 - \theta)^6}, \qquad \theta \in [1/5, 3/5].$$

At product coupling, $q_P(9/25) = 729/857$ and $d(9/25) = 665/857$. At the lower endpoint, $q_P(1/5) = 1/129$ and $d(1/5) = -21/43$. The zero occurs at $\theta = 3/10$, so a feasible negative witness suffices without numerical optimization. The example establishes neither truth of the anchor nor semantic validity of the local gates. It only isolates the difference between high product confidence and uniform action preference.

It also exposes the unresolved marginal error. For general $A > 0, M$ in this toy model, the sign of $q_P - q_S$ is the sign of $2\theta - A$. Full-path certification therefore holds exactly when $\max(0, A + M - 1) \geq A/2$, or $M \geq 1 - A/2$. With $A = .6$, increasing M from .6 to .7 changes rejection to a boundary-tie certificate. Coupling robustness does not imply robustness to biased marginals. This analytic sensitivity is not an estimate of the frequency or scale of gate errors; Appendix Y gives the separate empirical representation checks.

Coordinate changes versus proposer changes. Standardization alone is not a source of mathematical non-invariance. Write a fold's logits as $WD^{-1}(z - \mu) + \beta$. For any positive diagonal D' and location μ', set

$$W' = WD^{-1}D', \qquad \beta' = \beta + WD^{-1}(\mu' - \mu).$$

Then $W'D'^{-1}(z - \mu') + \beta' = WD^{-1}(z - \mu) + \beta$ for every path input. Thus a compensated coordinate change preserves every advantage and exact certificate. Changing the scaler without compensating the weights, or refitting the model, changes the proposer instead. The standardized trust boxes are coordinate-dependent diagnostics; they are not part of the full Fréchet acceptance contract.

For actual parameter changes, hold the candidate, anchor, gates, and loss family fixed. Let $R_f = \sup_{\theta_f \in \Theta_f} \|\widetilde{z}_f(\theta_f)\|_2$ and perturb standardized coefficients by (E_f, e_f). The log floor makes R_f finite on nondegenerate paths. Applying Lemma 1's softmax bound and then taking infima gives

$$|\underline{a}' - \underline{a}| \leq \eta, \qquad \eta = \max_k \frac{\|r^{(k)}\|_2}{2F} \sum_f \big(\|E_f\|_2 R_f + \|e_f\|_2\big).$$

Indeed, the right side bounds the advantage displacement uniformly over every coupling and loss vertex; two functions within a uniform bound have infima within that bound. A certificate with $\underline{a} \geq \eta$ therefore survives these declared perturbations. Near-zero certificates need not. This is a sufficient model-parameter margin, not an empirical claim that refitted proposers agree, and it supplies no target-label calibration.

The same continuity guarantees that every one-dimensional inner minimum below is attained. The fold-averaged robust advantage is

$$\inf_{(\theta_1, \ldots, \theta_F) \in \prod_f \Theta_f} \frac{1}{F} \sum_f d_{f,L}(\theta_f)$$
$$= \frac{1}{F} \sum_f \inf_{\theta_f \in \Theta_f} d_{f,L}(\theta_f) = \underline{a}_{\Theta,L}(b, c).$$

Conditions for fold separability. Fix the anchor, candidate, loss, all gate marginals, and every parameter before minimizing. Require each summand to have the form $d_{f,L}(\theta_f; \phi)$, where ϕ is any shared or fold-specific fixed parameter collection. No summand may depend on another fold's coupling coordinate. The objective is the fixed average of these summands over the nonempty product set. For every coupling vector, the sum is at least the sum of its individual minima. Choosing each coordinate's minimizer attains that bound. This proves the equality without a statistical independence or fold-stability assumption.

Shared weights, shared standardization, overlapping training data, and correlated predictions do not change that argument: they are fixed inside ϕ. A model that recomputes cross-fold attention, parameters, or ensemble weights as a function of multiple θ_f does not have this separable form; its joint objective must be certified directly. If admissible couplings are linked, the product set is a superset, so its lower bound still certifies any acceptance on the linked set. Equality and maximal permissiveness on the smaller set need not hold. For $\mathcal{L} = \mathrm{conv}\{L^{(1)}, \ldots, L^{(K)}\}$, the infimum of linear functions of L is concave in L. A concave function on a compact polytope attains a minimum at at least one vertex. Therefore

$$\inf_{L \in \mathcal{L},\, \boldsymbol{\theta} \in \prod_f \Theta_f} \frac{1}{F} \sum_f d_{f,L}(\theta_f) = \min_k \underline{a}_{\Theta, L^{(k)}}(b, c).$$

The minimum over a finite loss polytope therefore occurs at a vertex. Under zero–one loss, $d_{f,L} = q_{f,c} - q_{f,b}$, recovering the implemented confirmation rule. The broader statement permits asymmetric or boundary-specific costs, squared ordinal loss, or any other finite collection of loss matrices without learning a new composer, provided the vertices are fixed independently of evaluation outcomes. It certifies a preference inside the specified composer-output set, not a guarantee about the unknown true label posterior.

What a linked-coupling restriction assumes. Write each fold interval as $[l_f, u_f]$. A shared raw joint mass requires $J = [\max_f l_f, \min_f u_f]$ to be nonempty. For example, marginals $(.9, .9)$ give $[.8, .9]$, while $(.2, .2)$ give $[0, .2]$: no shared mass exists. An empty intersection must fail closed, not certify by vacuous minimization.

A different restriction, always feasible, uses a common relative position $t \in [0, 1]$ and $\theta_f(t) = l_f + t(u_f - l_f)$. The linked advantage is

$$\underline{a}_{\mathrm{link}} = \min_k \inf_{t \in [0,1]} \frac{1}{F} \sum_f d_{f,L^{(k)}}\big(\theta_f(t)\big) \;\geq\; \underline{a}_{\Theta, \mathcal{L}}.$$

This keeps finite loss-vertex sufficiency and one-dimensional optimization, but the fold minima no longer separate. It certifies a smaller ambiguity set, not the original contract more efficiently. In particular, a shared relative position need not include all folds' product points simultaneously. Adding those points is another explicit contract choice. Our confirmation uses the full product set; neither linkage is an implemented deployment option or a validated coverage gain.

A copula fitted to the distribution of output pairs $(A(x), M(x))$ across rows models across-row score dependence. It does not identify the within-row joint gate mass $\theta(x)$: product and endpoint couplings leave the same observed marginals. A narrower conditional set needs joint gate observations or an additional identifying model, then independent validation. Assuming a set because it retains desirable test actions would change the scientific question. The standardized boxes in Figure 1 are descriptive sensitivity restrictions, not learned copulas or calibrated support sets.

The finite-loss and fold-separability argument does not depend on having three labels. It applies to any finite taxonomy when a continuous score map and compact ambiguity sets have been specified. The one-dimensional search, however, comes from this paper's two-gate projection. Four-way uncertainty labels or separate citation-assignment variables require a new map and may produce a higher-dimensional search. We have not validated those extensions; formal vertex sufficiency is not evidence of empirical multi-citation generality.

Corollary 2 (finite representation families). The same argument covers a finite, predeclared unitization family $\mathcal{U}$. For each $u \in \mathcal{U}$, reaggregate the frozen token scores, recompute (A_u, M_u), and

evaluate the corresponding coupling and loss sets. Taking the minimum over u preserves contract soundness. It is also pointwise maximally permissive among deterministic filters required to be sound for every member of $\mathcal{U}$: any rejected action has an attaining or bounded unitization–coupling–loss witness. This is Corollary 2. It says nothing about unitizations omitted from $\mathcal{U}$ or about semantic correctness of their marginals.

With more than two Bernoulli gates, fixed marginals define a polytope over 2^K joint states rather than a one-dimensional path. The same robust-action definition still applies, but exact certification requires multidimensional global optimization, or linear programming when the downstream action score is affine in the joint masses. We do not implement that extension here; the one-dimensional solver is a computational benefit of the declared two-gate interface, not a claim that every diagnostic reduces to two gates.

For the order $I < P < S$, let $L_{I|PS}$ charge one exactly when the action and truth lie on opposite sides of $I \mid \{P, S\}$, and define $L_{IP|S}$ analogously for $\{I, P\} \mid S$. Every weighted absolute ordinal loss is $w_1 L_{I|PS} + w_2 L_{IP|S}$ for $w_1, w_2 \geq 0$. Positive scaling does not change the action preference. Setting $w_1 + w_2 = 1$ therefore turns the complete family into the line segment $\mathrm{conv}\{L_{I|PS}, L_{IP|S}\}$. Checking its two endpoints is exact for every relative weighting of the ordered boundaries; it is not a two-point approximation. This family does not include every possible asymmetric or nonordinal matrix. Those require extra predeclared vertices, which Proposition 4 handles without retraining the composer.

The criterion is also sharp within its declared model-space set. The product $\mathcal{K} = \prod_f \Theta_f$ and the finite loss polytope are compact, and the fold-averaged advantage is continuous in the coupling and linear in the loss. If $\underline{a}_{\Theta,\mathcal{L}} < 0$, a minimizing loss vertex and compatible coupling attain a negative advantage. Any deterministic filter that accepts this action then violates uniform dominance on $\mathcal{K} \times \mathcal{L}$. Conversely, a nonnegative infimum gives uniform weak dominance. PIAF is thus pointwise maximally permissive among sound deterministic filters that certify uniform preference under the specified coupling and cost sets. The composer was fitted only at $\theta = AM$; alternative couplings are score-space input perturbations. This completeness neither validates those perturbations as target posteriors nor enlarges the uncertainty set.

PIAF computes each one-dimensional infimum by interval branch-and-bound. Normalized S and I mass decrease with θ, while normalized P mass increases. Clipping at the frozen 10^{-8} log floor preserves monotonicity. Endpoint log bounds propagate through standardization and the linear composer; softmax interval bounds then lower-bound $\sum_y q_y\{L(b, y) - L(c, y)\}$. We bisect the box with the smallest lower bound until its gap to the best evaluated point is at most 10^{-6}. No run reached the 16,384-box cap: the maxima were 4,121 boxes on WiCE, 4,587 on ALCE, 11,542 on the original GenSearch test, and 11,946 on the role-isolated GenSearch confirmation. All original actions received a certificate or a negative feasible upper bound; none was numerically indeterminate. Compactness guarantees a counterexample coupling whenever that upper bound is negative, although the locked artifact stores the bound rather than its minimizing θ value.

Why endpoint checks are insufficient. Monotone local coordinates do not make the composed advantage monotone. For a constructed one-fold example, set $A = M = 3/5$, anchor P, candidate S, and logits $h_S = \log(31/250) - \log \ell_S - \log \ell_P$, $h_P = 0$, and $h_I = -4$. At product coupling, $\ell_S \ell_P = 6/49 < 31/250$, so S is the composed argmax. At the lower endpoint the product is $2/25 < 31/250$; at the upper it tends to zero. Both endpoints favor S. Yet at $\theta = 2/5$ the product is $1/8 > 31/250$, so $q_S - q_P < 0$. Loss-vertex sufficiency therefore does not imply coupling-endpoint sufficiency. The frozen 10^{-8} log floor preserves these strict inequalities. This is a synthetic counterexample, not another evaluation cohort.

Endpoint-only replay on archived actions. We evaluate both endpoints of each of the 3,249 fold intervals behind the 1,083 frozen GenSearch proposals. For zero–one loss, let $e_f = \min\{d_f(l_f), d_f(u_f)\}$ and $e = F^{-1} \sum_f e_f$. The Cartesian product of minimizing endpoints is feasible, so $\underline{a} \leq e$. Thus $e < 0$ is a sound rejection witness, whereas $e \geq 0$ is not an acceptance certificate. The distinction is between an upper bound and a lower bound, not numerical precision.

This post-hoc replay reads archived scores and certificates, not evaluation labels. All 433 archived rejections have a negative endpoint witness; the remaining 650 actions have nonnegative endpoint bounds and full-path certificates. Endpoint-only and full-path binary decisions therefore agree on this

cohort, although endpoint and full-search feasible advantages differ by up to .1403. This agreement measures an available rejection shortcut, not general endpoint sufficiency. The constructed interior example above has positive endpoint advantages but a negative full-path witness. A sound accelerated implementation may stop at a negative endpoint witness and retain bounded search for potential acceptances. We leave the frozen implementation and its reported runtime unchanged. The standalone replay is `scripts/audit_piaf_endpoint_shortcut.py`.

The small executable example `scripts/demo_piaf_certificate.py` runs this case and the Section 3.4 example through the existing certifier, then replays the minimizer with witness extraction. It prints the declared inputs, certified bounds, explicit coupling, and 17-point illustrative path. The grid is for inspection only; branch-and-bound determines the certificate. The example needs no model download, GPU, benchmark labels, or API credentials.

Numerical safeguard audit. The implementation fails closed before composer evaluation when $A = M = 0$ up to tolerance 10^{-12}. This avoids assigning an arbitrary posterior to an uninformative view. A label-free replay audits all 26,502 fold rows behind the 8,834-row role-isolated GenSearch confirmation. The minimum normalization mass over each full Frechet path, $\min_\theta(A + M - \theta) = \max(A, M)$, is 0.5035; no row is at or below 0.1. At the product point, the minimum $A + M - AM$ is 0.7503. The safeguard therefore does not alter any reported action. The audit artifact records threshold counts and the worst fold-row.

PIAF decision procedure. The input is an already frozen RA-DVRC action $b \to c$, together with the frozen fold posteriors and composers.

1. For each fold f, form $\Theta_f = [\max(0, A_f + M_f - 1), \min(A_f, M_f)]$.
2. For each $\theta_f \in \Theta_f$, construct and normalize $(A_f - \theta_f, \theta_f, M_f - \theta_f)$, then pass this local posterior and the frozen global posterior through the frozen composer q_f.
3. For each predeclared loss vertex $L^{(k)}$, certify $r_{f,k} = \inf_{\theta_f \in \Theta_f} d_{f,L^{(k)}}(\theta_f)$ by interval branch-and-bound.
4. Emit c iff $\min_k F^{-1} \sum_f r_{f,k} \geq 0$; otherwise retain b.

The output records either a robust lower bound or a negative feasible upper bound. PIAF does not search over labels, retrain a model, or choose a threshold.

Post-hoc implementation profile. We replay all 1,083 frozen GenSearch confirmation actions with the archived composers, inputs, and 10^{-6} tolerance. The script reads no evaluation label and reproduces all 650 retained decisions and certificate bounds. On one Apple M4 Max CPU process, after 20 warm-up actions, the pass takes 105.39 seconds (10.28 actions/second). Per-action wall time has median 22.4 ms, 95th percentile 706.6 ms, 99th percentile 1.41 s, and maximum 3.02 s. Across fold-actions, the interval search evaluates a median 35 boxes, 95th percentile 67, 99th percentile 3,696, and maximum 11,946 of the 16,384-box cap. These are machine-specific CPU measurements after data loading, not a serving-latency or energy claim. They profile only the action filter and exclude neural inference. The anonymous artifact includes the script and JSON report.

Training uses relation-class-balanced binary cross entropy on A and M. The full local branch adds class-balanced binary cross entropy on weak supported and official unsupported material-unit targets with weight 0.5. We train for two epochs with micro-batch size 2, gradient accumulation 16, effective bag batch size 32, encoder learning rate 2×10^{-5}, head learning rate 10^{-4}, weight decay 0.01, and 10% warmup. Gradient norm is clipped at 1.0. There is no hyperparameter search, early stopping, or checkpoint selection after epoch two.

A.4 Low-capacity composition

For numerical constant $\epsilon = 10^{-8}$, the composer receives

$$z = [\log(\max(g, \epsilon)); \log(\max(\ell, \epsilon))] \in \mathbb{R}^6.$$

The maximum is elementwise, matching the implementation's probability floor; it is not additive smoothing. A standard scaler is fit on the composer-training partition. Multinomial logistic regression then gives

$$q = \mathrm{softmax}(W\widetilde{z} + b).$$

The coefficient matrix has $3 \times 6 = 18$ entries and the intercept has three, for 21 learned parameters. We use $C = 0.3$, L2 regularization, LBFGS, tolerance 10^{-8}, a 5,000-iteration limit, and no class weights. Each artifact stores the posterior order, feature names, scaler statistics, coefficients, intercepts, optimizer settings, random seed, and fit-unit hash.

This model is a learned generalized log opinion pool (Smith et al., 2005). It cannot retrieve new evidence or change either encoder. It only estimates how two fixed posterior views should compose by relation.

A.5 Directional intervention

Let $b = \arg\max_y g_y$ and $c = \arg\max_y q_y$. The estimated advantage is

$$a(x) = q_c(x) - q_b(x).$$

If $c = b$, RA-DVRC keeps b. Otherwise it emits c only if

$$a(x) \geq \begin{cases} \tau_\uparrow, & c > b, \\ \tau_\downarrow, & c < b, \end{cases}$$

where the inequalities use $I < P < S$. The frozen policy is $(\tau_\uparrow, \tau_\downarrow) = (0.20, 0)$. Promotions need more posterior evidence because a false promotion licenses claim content that the evidence does not fully support. Demotions are permitted whenever the composed argmax beats the anchor under the composed posterior.

For a general loss matrix $L(d, y)$, a posterior equal to the target conditional label distribution supports the cost-aware advantage

$$a_L(b, c \mid x) = \sum_y q_y(x)\{L(b, y) - L(c, y)\}.$$

The implemented score $q_c - q_b$ is exactly this quantity under zero–one loss. After the ALCE outcomes were known, we also evaluated a parameter-free ordinal action filter. It keeps only existing RA-DVRC actions with $a_L \geq 0$ under $L(d, y) = (r(d) - r(y))^2$ and ranks $r(I) = 0, r(P) = 1, r(S) = 2$. This filter cannot create an unreviewed action. Its opened-set results appear in Appendix G and remain post-hoc.

B Decision Properties and Their Scope

B.1 Proof of Proposition 1

Let $W = w$ be a warning state and let G indicate that the local defect is real. Define $\mathcal{C}(w) \subseteq \mathcal{R}$ as the relations whose definitions do not logically contradict $G = 1$ for that warning type. Fix an arbitrary categorical distribution π supported on $\mathcal{C}(w)$. Diagnostic precision imposes $\Pr(G = 1 \mid W = w) = 1$ but does not specify defect severity or supported mass elsewhere in the claim, so it places no further restriction on π. A conditional population with $\Pr(Y = y \mid W = w) = \pi_y$ is therefore consistent with the diagnostic statement. The optimal deterministic router based only on w predicts the modal relation and incurs

$$1 - \max_{y \in \mathcal{R}} \pi_y.$$

Choosing uniform π gives risk $1 - 1/|\mathcal{C}(w)|$ despite perfect diagnostic precision. The value is $1/2$ when a missing-coverage warning is compatible only with partial support and insufficient evidence, and $2/3$ when all three relations remain compatible. Additional information must identify how the

local defect composes with supported claim content. This proves Proposition 1. It does not assert that every warning is compatible with every relation.

B.2 Proof of Proposition 2 and strict separation

Let E_b be anchor error, $\mathcal{D}$ the warning-selected set, $R_b = \Pr(E_b)$, $d = \Pr(\mathcal{D})$, and $r_D = \Pr(E_b \mid \mathcal{D})$. Deferral removes the selected rows from the automatic-decision population. The retained error mass is $\Pr(E_b) - \Pr(E_b, \mathcal{D}) = R_b - dr_D$, and retained probability mass is $1 - d$. Therefore

$$R_{\text{ret}} = \frac{R_b - dr_D}{1 - d}$$

and

$$R_b - R_{\text{ret}} = \frac{d(r_D - R_b)}{1 - d}.$$

For $0 < d < 1$, the selective risk is lower exactly when $r_D > R_b$. The result depends only on whether warnings enrich anchor errors. It does not depend on the candidate relation attached to the warning.

For hard replacement, let E_c denote candidate error. Outside $\mathcal{D}$, the system still uses the anchor. Hence

$$R_{\text{hard}} - R_b = \Pr(\mathcal{D}, E_c) - \Pr(\mathcal{D}, E_b).$$

Partitioning each event by the other model's correctness cancels cases where both are right or both are wrong, leaving

$$R_{\text{hard}} - R_b = \Pr(\mathcal{D}, \neg E_b, E_c) - \Pr(\mathcal{D}, E_b, \neg E_c).$$

The first term is induced harm and the second is correction. Thus a warning can have high $\Pr(E_b \mid \mathcal{D})$ and support deferral while its candidate verdict still fails hard replacement because it harms too many anchor successes.

The separation can be strict even when the warning is always diagnostically valid. Consider a warning type with two compatible relations and let the anchor and candidate predict those different labels. Let $d = .2$ and let anchor error be .4 within $\mathcal{D}$ and .025 outside it, so $R_b = .2(.4) + .8(.025) = .10$. Within $\mathcal{D}$, assign probability .4 to a correction and .6 to an induced harm. These cells sum to one and give $r_D = .4$. Deferral therefore leaves

$$R_{\text{ret}} = \frac{.10 - .2(.4)}{.8} = .025,$$

whereas replacement gives

$$R_{\text{hard}} = .10 + .2(.6 - .4) = .14.$$

Set $G = 1$ for every selected row. The warning then has perfect diagnostic precision, improves deferral, and worsens replacement. This proves the strict separation claimed in Proposition 2.

If a deferred case incurs review loss $\kappa(X)$ instead of being excluded from selective risk, total expected loss becomes

$$R_b - \Pr(\mathcal{D}, E_b) + \mathbb{E}[\kappa(X)\mathbf{1}_{\mathcal{D}}].$$

Deferral then improves total loss exactly when $r_D > \mathbb{E}[\kappa(X) \mid \mathcal{D}]$. This cost-aware extension is analytic; the experiments evaluate selective risk at a prespecified coverage and do not estimate organization-specific review cost. In a deployment, $\kappa(X)$ should be estimated on the target workflow

from observed review time, labor cost, latency cost, and residual post-review error. Estimating it on generic annotator time or on the full population can be misleading because the warning-selected cases may be systematically harder.

B.3 A COMMON KEEP–REPLACE–REVIEW ENVELOPE

The two identities can be embedded in one conditional decision problem. Let $b(x)$ be the anchor, $c(x)$ the candidate, $p(y \mid x)$ the target conditional posterior, and $L(d, y)$ a relation loss. Let $\kappa(x)$ include review effort, latency, and the residual loss after review. Then

$$r_K(x) = \sum_y p(y \mid x) L(b(x), y), \qquad r_R(x) = \sum_y p(y \mid x) L(c(x), y), \qquad r_D(x) = \kappa(x).$$

The Bayes action is $\arg\min_{u \in \{K,R,D\}} r_u(x)$. Pairwise comparison of K and R gives the posterior action advantage used by RA-DVRC. Aggregating the comparison of K and D over a fixed review set gives the cost-aware deferral identity above. Under perfect review and a fixed review budget, this reduces to ranking cases by anchor-error probability. WCSD instead uses a frozen structured disagreement as a zero-new-parameter proxy for that ranking.

This envelope is a specification, not an omniscient controller. The paper estimates different pairwise quantities on different protected studies and checks the actions they induce. It does not assume that the composed posterior, review loss, or target-domain calibration error is known. That distinction is why a replacement success on WiCE cannot certify ALCE replacement, and why a GenSearch deferral success need not validate its candidate relation.

B.4 POSTERIOR-ADVANTAGE IDENTITY

For Corollary 1, write $e_y(x) = p_t(y \mid x) - q_y(x)$. The target benefit is $\Delta_t(x) = p_t(c \mid x) - p_t(b \mid x) = a(x) + e_c(x) - e_b(x)$. If $\max_y |e_y(x)| \le \varepsilon_t(x)$, the triangle inequality gives $\Delta_t(x) \ge a(x) - 2\varepsilon_t(x)$. A positive bound is sufficient for benefit. With one label per input, an aggregate calibration statistic such as ECE does not identify the required pointwise error bound. This is an analytic condition, not an estimated target-domain certificate.

Suppose $q_y(x) = \Pr(Y = y \mid X = x)$. The conditional zero–one risk of predicting d is $1 - q_d(x)$. Therefore

$$\mathbb{E}[L(b, Y) - L(c, Y) \mid X = x] = (1 - q_b) - (1 - q_c) = q_c - q_b.$$

Thus the gate acts on estimated conditional decision benefit rather than raw model disagreement. With an estimated, imperfectly calibrated q, the identity is an interpretation rather than a guarantee. This is why we report NLL, Brier score, ECE, and empirical action harms.

B.5 LABEL-PRIOR SENSITIVITY

Let $r_y^h = \Pr(h(X) = y \mid Y = y)$ be class recall. Under label shift that holds these class-conditional recalls fixed,

$$\mathrm{Acc}(h; \pi) = \sum_y \pi_y r_y^h.$$

For candidate h_1 and anchor h_0,

$$\Delta\,\mathrm{Acc}(\pi) = \sum_y \pi_y \left(r_y^{h_1} - r_y^{h_0}\right).$$

This linear function reaches its minimum over the probability simplex at the smallest class-recall change. Nonnegative recall changes for all relations are therefore sufficient for a nonnegative point

gain under every label prior, conditional on fixed class-conditional behavior. This is a sensitivity result, not a finite-sample certificate and not evidence that arbitrary domain shift is only label shift.

C Development Selection Without Title Leakage

WiCE contributes 3,470 official training units in 1,253 title clusters (Kamoi et al., 2023). Three outer folds are title-disjoint. Each global and local checkpoint predicts only its outer-held fold. Within each outer fold, a second three-way stratified group split fits the composer on two inner partitions and predicts the third. The union therefore gives one nested out-of-fold composed posterior per training unit without title overlap.

The development protocol selects the lowest round promotion threshold that satisfies six constraints. Cluster-macro accuracy must improve by at least one point, and its cluster-bootstrap lower bound must exceed zero. Every relation recall and all three outer-fold accuracies must improve. None of four ordered boundary errors may increase, and at least 20 clusters must change. Demotion remains zero unless no policy is feasible.

Table 4: Full relation metrics for the fixed global anchor and RA-DVRC. Development uses nested OOF predictions; test is the one-time official confirmation. All values are percentages.

Split	System	Cluster acc.	Row acc.	Macro-F1	R_I	R_P	R_S
Development	Global	76.05	76.71	62.66	75.52	22.67	89.43
Development	RA-DVRC	77.53	78.39	64.63	80.04	23.31	90.10
Official test	Global	45.97	47.21	44.27	71.88	21.86	89.19
Official test	RA-DVRC	54.48	55.31	52.28	78.13	32.09	93.69

The complete threshold frontier is below. Lower thresholds improve aggregate development accuracy more, but increase at least one ordered boundary error. Thresholds above .20 lose either the required one-point aggregate gain or a relation-wise recall constraint. The selected .20 value is the first jointly feasible point, not the threshold with the largest accuracy estimate.

Table 5: Complete development promotion-threshold frontier. Deltas are percentage points relative to the global anchor. The predeclared rule selects the first jointly feasible threshold.

$\tau_\uparrow$	Acc. Δ	Min recall Δ	Max boundary Δ	Clusters	Feasible
0	+2.00	+0.99	+3.39	296	no
.02	+2.03	+1.32	+3.39	286	no
.05	+1.96	+1.29	+2.12	266	no
.10	+1.87	+1.05	+1.48	252	no
.15	+1.69	+0.81	+0.21	220	no
.20	+1.48	+0.64	-0.21	194	yes
.30	+0.97	-1.91	+3.39	158	no
.40	+0.77	-2.54	+4.45	140	no
.50	+0.64	-2.75	+5.08	133	no

At .20, the three outer-fold cluster gains are +1.16, +1.41, and +1.86 points. The policy changes 212 rows in 194 clusters: 112 corrections, 54 harms, and 46 neutral changes. These estimates authorize one protected test but remain adaptive development evidence.

D One-Time Official WiCE Confirmation

D.1 Prediction lock and label firewall

The official test has 358 units and 326 title clusters. Its key has file mode `000` while six fixed checkpoint outputs are produced. Three final calibrators are fit only on the matching outer-held training predictions. The three global posteriors and three composed posteriors are averaged, then the directional policy is applied once.

Before key access, the protocol locks all data hashes, six checkpoint hashes, three calibrator hashes, ensemble predictions, action mask, changed IDs, thresholds, metrics, bootstrap seed, and success gate. The title-cluster accuracy delta must be positive, with a bootstrap lower bound above zero. At

least 20 clusters must change, and at least two checkpoint arms must improve. No boundary-error interval may have an upper endpoint above +2 points. The outcome must be reported regardless of sign.

D.2 Accuracy, actions, and boundaries

The title-cluster gain is +8.51 points with 95% interval [5.44, 11.73]. Row accuracy gains 8.10 points and macro-F1 gains 8.02 points. The individual checkpoint-arm gains are +5.83, +7.29, and +11.05 points.

RA-DVRC changes 33 decisions: 30 corrections, one harm, and two neutral changes. The correction rate is 90.9% with exact 95% interval [75.7, 98.1]; the harm rate is 3.0% with interval [0.08, 15.8]. The descriptive exact two-sided correction-versus-harm test gives $p = 2.98 \times 10^{-8}$. The title-cluster bootstrap above supplies the population-level inference. All 33 cases are retained in the artifact.

Table 6: Ordered boundary-error rates. Lower is better; deltas are RA-DVRC minus global in percentage points.

Split	Boundary	Global	RA-DVRC	Δ
Development	$I \to S$	17.86	15.77	-2.09
Development	$P \to I$	41.53	41.31	-0.21
Development	$P \to S$	35.81	35.38	-0.42
Development	$S \to \{I, P\}$	10.57	9.90	-0.67
Official test	$I \to S$	12.50	12.50	0.00
Official test	$P \to I$	35.81	29.77	-6.05
Official test	$P \to S$	42.33	38.14	-4.19
Official test	$S \to \{I, P\}$	10.81	6.31	-4.50

D.3 Numerical amendment and reconstruction

The original GPU environment became unavailable after training because of a driver/library mismatch. Before any official prediction or label access, a scoped amendment changed only inference device and precision to CPU/float32. It retained checkpoints, calibrators, preprocessing, thresholds, metrics, and the gate.

The amendment first reconstructed 64 archived training decisions per branch. Both branches matched the archived GPU/float16 argmax on 64/64 rows. Maximum posterior differences were 0.00167 for the global branch and 0.00299 for the local branch. A final audit verifies hashes, row alignment, execution order, and label-firewall state.

E Calibration, Label Mixture, and Complete View Ablation

On development, the composed posterior improves NLL from 0.645 to 0.578, Brier score from 0.341 to 0.313, and ECE from 0.0301 to 0.0285. On the official test, the same metrics improve from 0.999 to 0.866, 0.629 to 0.519, and 0.219 to 0.151. These results support the posterior-advantage interpretation. The remaining test ECE of 0.151 still argues against treating the score as exact risk.

WiCE training labels are 60.3% support, 13.6% partial support, and 26.1% insufficient. The official test is 31.0%, 60.1%, and 8.9%, respectively. The total-variation distance between these label mixtures is 0.465.

Why absolute accuracy falls on the official test. The label mixtures have counts $(S, P, I) = (2091, 472, 907)$ in development and $(111, 215, 32)$ on test. The anchor already has low Partial recall in development (22.67%), which remains low on test (21.86%). Partial examples become the majority of the test set. Holding development class recalls $r_{d,y}$ fixed and changing only their weights to the test proportions $\pi_{t,y}$ gives

$$\widetilde{\mathrm{Acc}}_t = \sum_y \pi_{t,y} r_{d,y} = 48.09\%.$$

The observed anchor row accuracy is 76.71% in development and 47.21% on test. The exact descriptive decomposition is

$$\mathrm{Acc}_t - \mathrm{Acc}_d = \underbrace{\sum_y (\pi_{t,y} - \pi_{d,y}) r_{d,y}}_{-28.62 \text{ pp}} + \underbrace{\sum_y \pi_{t,y}(r_{t,y} - r_{d,y})}_{-0.89 \text{ pp}}.$$

With development recalls as the reference, changing the label mixture accounts for 97.0% of the 29.51-point row-accuracy drop. This is an accounting identity, not a causal attribution or a test of pure label shift. Development uses OOF predictions; test averages frozen fold predictions. The class-conditional term therefore includes both population and prediction-procedure differences.

Title-cluster weighting changes development row accuracy by -0.66 points and test row accuracy by -1.24 points. Its contribution to the larger 30.08-point cluster-accuracy drop is thus -0.57 points. These quantities explain the reported aggregate gap without assuming a new annotation scheme or changing any threshold. The weak Partial recall remains a substantive limitation. The proposer improves each test-class recall relative to the same test anchor; the +8.51-point comparison is paired within that population. The calculation is retrospective and can be reproduced from the archived relation counts and recalls using `analyze_iclr_wice_split_gap.py`.

The different label mixture also helps interpret the raw local view. On test it gains 32.58 title-cluster points and corrects 122 of 134 changes, but reduces insufficient recall by 9.38 points. On development it loses 0.66 points, reduces insufficient recall by 12.57 points, and produces 258 harms versus 231 corrections. The complete main paper ablation retains unconditional composition, arithmetic averaging, global-only and local-only composers, the raw local view, and two fixed nonlinear stackers. No method is removed because of its test sign.

Table 7: Complete post-hoc view ablation. Worst recall is the smallest class-recall change. Test C/H denotes corrected/harmed changed decisions. The frozen primary method is bold.

System	Dev Δ	Test Δ	Worst dev	Worst test	Test C/H
Raw local	-0.66	+32.58	-12.57	-9.38	122/9
Unconditional dual	+2.00	+15.62	+0.99	0.00	60/5
Arithmetic mean	+0.72	+10.58	-0.72	-0.90	39/4
Global only	-0.33	-1.10	-8.47	-2.79	2/6
Local only	+0.94	+10.24	-1.91	+4.50	42/8
Quadratic logistic	+1.54	+13.26	+0.24	+3.13	49/4
MLP-16	-2.64	+10.67	-4.52	0.00	46/11
RA-DVRC	**+1.48**	**+8.51**	**+0.64**	**+4.50**	**30/1**

The richer stackers were specified after the official test result and fit no protected label. The quadratic model expands the six log-posteriors to all degree-two terms and fits an 84-parameter multinomial logistic model with the same $C = .3$. The MLP has one 16-unit tanh layer and 163 parameters. Both use the same grouped nested development protocol and the frozen .20/0 directional policy. The MLP loses 2.64 development points despite a positive test result. The quadratic model improves average accuracy on both splits, but its worst-class gains are smaller than RA-DVRC's and it causes four test harms rather than one. These post-hoc findings motivate future prospective study; they do not alter the frozen primary comparison. WiCE development cannot serve as a new confirmation because earlier project versions accessed it adaptively.

E.1 Source-only predictive capacity and selection checks

Later development work tests whether weak fusion artificially limits the proposer. These are adaptive source-only analyses on 3,470 WiCE training rows from 1,253 titles, not new evaluations of the frozen PIAF policy. For each encoder fold, meta-training and evaluation use only that encoder's own held-out source fold. Nested title-grouped meta-folds select six settings per method. This prevents an evaluated label entering its encoder or meta-fit, but does not erase earlier source-data adaptivity or constitute full-pipeline nested model selection. Encoder training sets overlap across folds.

Table 8: Adaptive source-only raw prediction capacity. Acc./NLL are title-weighted; Partial recall is row-based. No directional threshold, PIAF filter, or risk gate is applied.

Predictor	Acc. (%)	Partial recall (%)	NLL
Equal global pool	76.26	53.81	.6288
Tuned global-only linear	78.91	28.60	.5603
Tuned pooled dual-view linear	80.56	39.19	.5125
Family-specific dual-view linear	80.19	36.44	.5165
Marginal-lattice linear	80.20	36.02	.5233
Dual-view ordinal endpoints	79.87	37.08	.5404
Global-only boosted trees	78.64	34.75	.5874
Dual-view boosted trees	79.06	37.71	.5437

The three-family pooled dual's 4.30-point gain over equal global pooling contains 2.65 points recovered by tuning the global control itself. Its remaining gain over tuned global linear fusion is 1.65 points, with 139 corrections, 85 harms, and 64 neutral changes. Thus better prediction does not by itself establish action safety. These are refitted predictors, not the unchanged confirmed proposer; no source ranking selects a deployment model.

A separate one-family budget study reserves source validation for independent two/four-epoch checkpoint selection. Across three outer source cohorts, the selected dual reaches 77.53% versus 77.08% for two epochs, a +0.45-point difference with descriptive interval [-0.41, +1.35]. Joint selection among four retained global/local pairs reaches 77.21%, not an improvement over separate selection (-0.32 points [-0.79, +0.13]). The two-budget global stack reaches 74.87%. The different encoder fit/validation allocation prevents a direct comparison with the three-family table. Changing the meta-selection objective to NLL also gives mixed family-level gains. Numerical tie handling was corrected before reporting these values; original outputs remain archived. These budget intervals use 5,000 title resamples and condition on the fitted models. The checks establish tested capacity and selection tradeoffs, not SOTA or protected confirmation of a stronger action policy.

E.2 Post-hoc local-pooling sensitivity

We rerun each frozen local checkpoint once and retain the same token scores, composer class, grouped folds, $C = .3$, and .20/0 directional thresholds. Only the local aggregation operator changes. The same operator is used first across evidence chunks for each material unit and then across material units. We report normalized smoothmax with $\tau \in \{4, 8, 12, 24\}$, arithmetic mean, and hard maximum. The family was specified after the official test was open, so no protected label selects an operator or temperature.

Table 9: Post-hoc local-pooling sensitivity with the composer and directional thresholds fixed. Deltas and intervals are percentage points relative to the global anchor. The family was specified after test opening; the frozen operator is bold.

Pooling	Dev Δ	Test Δ [95% CI]	Worst recall dev/test	Test C/H
Smoothmax $\tau = 4$	+1.32	+5.14 [2.38,8.05]	-2.12/+4.19	21/3
Smoothmax $\tau = 8$	+1.54	+8.21 [5.14,11.50]	0.00/+5.41	30/2
Smoothmax $\tau = 12$ replay	**+1.48**	**+8.51 [5.44,11.66]**	**+0.64/+4.50**	**30/1**
Smoothmax $\tau = 24$	+1.35	+10.35 [6.98,13.96]	+0.48/+2.70	37/2
Mean	+0.47	-0.03 [-1.76,1.74]	-4.03/-0.47	5/5
Maximum	+1.82	+18.02 [13.60,22.57]	+0.14/0.00	69/5

The conclusion is not uniform robustness to arbitrary pooling. All smoothmax temperatures retain positive average gains, but $\tau = 4$ weakens development worst-class recall. Mean pooling largely erases the local signal. Maximum pooling amplifies it and raises average accuracy, yet it induces five test harms. This pattern supports a narrower claim: salient-evidence pooling matters more than the exact smoothmax temperature, while aggregation strength changes the gain–harm tradeoff. The canonical CPU rerun reproduces the frozen $\tau = 12$ action counts and point estimates; its maximum

local-posterior difference from archived inference is 0.0048 on development and 2.3×10^{-6} on test. None of these post-hoc outcomes alters the prospective method or its evidence role. The canonical sealed artifact uses a separately seeded bootstrap interval of [5.44, 11.73]; [5.44, 11.66] above belongs only to the post-hoc pooling replay.

F Motivating Hard-Router Audit

The motivating study uses 100 previously unused ALCE Vicuna pairs from 44 question clusters. A frozen typed-obligation detector maps every fired warning to `insufficient`. Model-only and routed predictions, policies, metrics, and hashes are locked before the relation key is opened.

The router changes eight rows. Whole-cohort exact accuracy falls from 63% to 61% with title-cluster interval [-7.29, 2.06]. Insufficient recall rises from 65.6% to 68.8%. That gain comes at a cost: strict-support false alarms rise from 13.2% to 18.4%, and partial-support false alarms rise from 23.3% to 33.3%.

Two annotators independently label all eight changed rows while blind to each other. They agree on every three-way relation and warning-validity label; a third adjudicator resolves two issue-taxonomy disagreements. Final relations are five partial support, one support, and two insufficient. Warning validity is three `yes`, four `partial`, and one `no`, so seven of eight warnings are at least partly valid. The global verifier matches six human relations; the hard router matches two. This is the empirical diagnostic-to-decision gap: warning validity and action severity are different variables.

The eight rows are selected because the router fired. They identify a failure mechanism but do not estimate population warning precision. RA-DVRC's official test result does not use these human labels.

An earlier, separately sampled audit provides secondary mechanism evidence but is not pooled with this protected cohort. It covers all 16 fired rows in a held-out RAGTruth challenge slice and all 13 precision-gated fires in an ALCE human-evaluation slice. Under adjudication, warning precision is 10/16 strict and 12/16 lenient on RAGTruth, versus 12/13 under either rubric on ALCE. The different sampling frames and prior use of the ALCE slice in gate development preclude a combined prevalence estimate. They do support the narrower pattern behind Proposition 1: many detected defects are real, but their validity is dataset-dependent and does not itself determine relation severity.

G Transfer and Certificate Evaluation

After the official WiCE gate passed, one transfer was authorized on a previously unopened public ALCE Vicuna reserve. RA-DVRC, all six checkpoint weights, three calibrators, thresholds, preprocessing, endpoints, and reporting rules remained unchanged. The cohort contains 186 rows from 68 ASQA/ELI5 question clusters. No relation label was present during inference. All predictions and 28 changed row IDs from 26 clusters were hash-locked before human labeling.

Two annotators independently received randomized workbooks containing claims and full evidence but hiding model predictions and the other annotator's labels. They agreed on 23 of 28 exact three-way relations (82.1%). The five disagreements were adjacent-boundary cases: two `insufficient` versus `partial_support` and three `partial_support` versus `support`. A third reviewer adjudicated only those five rows.

The final labels identify 20 corrections, 8 harms, and no neutral changes. Paired row accuracy improves by 6.45 percentage points. Question-cluster macro accuracy improves by 7.46 points, with a 10,000-repetition cluster-bootstrap interval of [1.33, 13.83]. The prespecified descriptive action-level one-sided correction-versus-harm test gives $p = .01785$; the complete-question bootstrap is the population-level inference. These endpoints are identifiable because unchanged rows contribute exactly zero to paired differences. Absolute ALCE accuracy is not identifiable because unchanged rows were not labeled, so we do not report it.

This positive paired effect is insensitive to which of the two independent ALCE label sets is used. Annotator A alone gives +7.53 row points and +8.93 cluster points, with cluster interval [2.91, 15.23]. Annotator B alone gives +6.99 row points and +8.25 cluster points, with interval [2.57, 14.29]. Both

label sets retain a harm-rate upper bound above 20%. Adjudication therefore does not create the average gain, and it does not remove the ALCE safety failure.

For the same reason, a full-cohort class-conditional reliability diagram is not identifiable on ALCE. Computing calibration only on the action-selected rows would condition on the decision score and produce a selection-biased curve. We instead report action outcomes by adjudicated relation. This localizes all eight harms to partial support without presenting the selected subset as a target-domain calibration estimate.

The action harm rate is 28.6%, with exact 95% interval [13.2, 48.7]. The interval's upper end exceeds the frozen 20% ceiling. Four gate components pass: row delta is positive, the cluster interval is above zero, the exact test is below .05, and 26 changed clusters exceed the minimum of 10. The harm-bound component fails, so the complete interpretation is `confirmatory_transfer_fail`. The figure below displays the required distinction between positive average effectiveness and failed action safety.

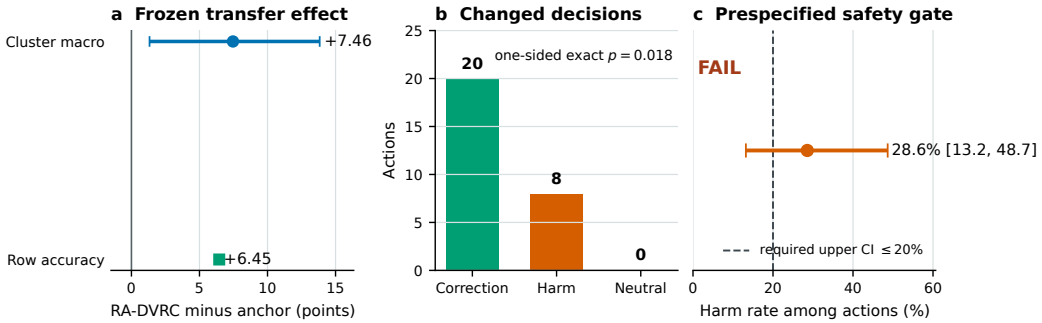


Figure 2: Frozen ALCE transfer result. Paired row and question-cluster accuracy improve (left), and corrections outnumber harms (middle). The exact 95% harm-rate interval crosses far above the prespecified 20% upper-bound criterion, so the complete transfer gate fails (right). Absolute ALCE accuracy is not estimated because only changed rows were human-labeled.

An opened-label diagnosis is retained as exploratory. All eight harms have adjudicated relation `partial_support`. Seven are demotions to `insufficient`, and one is a promotion to `support`. Harmful actions have higher mean frozen advantage than corrections (0.369 versus 0.271), so post-hoc increases to the demotion threshold remove corrections faster than harms. No frontier value is used for selection or to replace the frozen result.

Table 10: Opened-label post-hoc ALCE demotion-threshold frontier. The promotion threshold remains frozen. Deltas and intervals are question-cluster percentage points. No row selects or revises the confirmatory result.

$\tau_{\downarrow}$	Changed	Corrections	Harms	Harm rate	Cluster Δ [95% CI]
0 (frozen)	28	20	8	28.6	+7.46 [+1.33,+13.83]
.10	26	18	8	30.8	+6.24 [+0.23,+12.45]
.15	24	16	8	33.3	+5.38 [-0.54,+11.52]
.20	20	12	8	40.0	+3.29 [-2.08,+8.98]
.30	15	8	7	46.7	+1.58 [-3.09,+6.53]
.50	5	3	2	40.0	+1.35 [-1.96,+5.15]

The selected sets are suffixes of the same frozen advantage ranking. Thus any rank-preserving scalar recalibration or quantile remapping yields one of these frontier points. None passes the original action-safety criterion. A vector recalibrator could change the ranking, but it would require new assumptions or target labels and a fresh protected evaluation. Boundary-specific demotion thresholds would likewise be a new target-adaptive policy; they were not prespecified, so we do not tune them on these opened labels.

G.1 POST-HOC CANONICAL ORDINAL ACTION FILTERS

Scalar thresholding preserves the zero–one advantage ranking. We therefore test a different question: does an action survive when loss recognizes the order $I < P < S$? For each already proposed action $b \to c$, we compute $a_L(b, c \mid x)$ and retain it iff the value is nonnegative. No threshold, class weight, or ALCE label is fitted. We report all three canonical losses.

Table 11: Post-hoc canonical ordinal action filters. Each filter can only remove a frozen RA-DVRC action. C/H/N denotes corrections, harms, and neutral changes. Deltas are percentage points; all three losses are retained.

Split	Loss	Actions	C/H/N	Row Δ	Cluster Δ [95% CI]
WiCE OOF	Zero–one	212	112/54/46	+1.67	+1.48 [+.73,+2.24]
WiCE OOF	Absolute ordinal	184	99/47/38	+1.50	+1.32 [+.61,+2.04]
WiCE OOF	Squared ordinal	157	83/41/33	+1.21	+1.03 [+.36,+1.70]
WiCE test	Zero–one	33	30/1/2	+8.10	+8.51 [+5.44,+11.73]
WiCE test	Absolute ordinal	31	29/0/2	+8.10	+8.74 [+5.83,+11.96]
WiCE test	Squared ordinal	30	28/0/2	+7.82	+8.44 [+5.52,+11.50]
ALCE	Zero–one	28	20/8/0	+6.45	+7.46 [+1.33,+13.83]
ALCE	Absolute ordinal	24	16/8/0	+4.30	+5.38 [-.28,+11.35]
ALCE	Squared ordinal	12	10/2/0	+4.30	+4.78 [+.37,+10.05]

Squared loss removes six of eight ALCE harms but also ten corrections. Its observed harm rate is 16.7%, yet the exact 95% upper bound is 48.4% because only 12 actions remain. This is a better correction–harm frontier than scalar thresholding, not a retrospective safety pass. The complete artifact retains every canonical loss on all three partitions. A new protected cohort is required before OAF can become a confirmed policy.

G.2 PIAF CONSTRUCTION AND CONFIRMATION

PIAF addresses a different uncertainty from OAF. OAF changes the loss while holding the independent lattice fixed. PIAF keeps zero–one preference but requires it under every coupling compatible with A and M. Both filters were specified after WiCE and ALCE outcomes were known. GenSearch labels and aggregate test results were also open before PIAF. The three analyses in the next paragraph are therefore mechanism and transport-sensitivity evidence, not confirmation.

On WiCE test, the original 30/1/2 corrections/harms/neutrals become 17/0/2: PIAF retains 19/33 actions, gains 4.75 row points, and gains 5.06 cluster points [2.76, 7.52]. On ALCE transfer, 20/8/0 becomes 8/0/0; retention is 8/28, with +4.30 row points and +5.15 cluster points [1.35, 9.80]. On GenSearch test, 96/24/20 becomes 54/10/11; retention is 75/140, the observed harm rate is 13.3%, and gains are +4.01 row and +5.24 cluster points [2.97, 7.71].

On all three partitions, PIAF is a strict subset of squared-ordinal OAF. The four ALCE actions kept by OAF but rejected by PIAF split evenly into two corrections and two harms. This isolates projection ambiguity as one source of unsafe severity changes. GenSearch is the important counterweight: PIAF lowers observed harm but retains ten harms, so projection robustness alone does not solve target calibration or representation error. At this stage, a fresh cohort was still needed to test PIAF as a policy.

G.3 FRESH QUERY-DISJOINT CONFIRMATION

We next froze PIAF, OAF, all six checkpoint predictions, and the success gates before opening the previously unused official GenSearch train labels. We use all 8,834 rows in 1,156 exact-query clusters. These queries do not occur in GenSearch development or test. The split is named `train` by the source release, but it is never used here for fitting or threshold selection. The label key and source remained unreadable until the 1,083 RA-DVRC actions and their PIAF certificates were fixed.

PIAF raises row accuracy from 69.44% to 74.36%. Its action-harm rate is 4.71 points below unfiltered RA-DVRC under the paired complete-query bootstrap ([-6.53, -2.89]); all 10,000 replicates are

Table 12: Fresh query-disjoint GenSearch confirmation. C/H/N denotes corrections, harms, and neutral changes. Retention is relative to the 1,083 frozen RA-DVRC actions. Gain intervals use complete query clusters; harm intervals are descriptive exact action-level intervals. The paired harm-rate difference is bootstrapped by complete query.

Frozen policy	Actions	C/H/N	Retention	Cluster gain [95% CI]	Action harm [95% CI]	Correct-P anchor harms
Unfiltered RA-DVRC	1,083	775/141/167	100.0	+7.56 [6.65,8.53]	13.02 [11.07,15.17]	72
Squared-ordinal OAF	940	697/105/138	86.8	+7.00 [6.11,7.91]	11.17 [9.23,13.36]	36
Fréchet PIAF	**650**	**489/54/107**	**60.0**	**+5.15 [4.43,5.90]**	**8.31 [6.30,10.70]**	**1**

below zero. Corrections exceed harms at descriptive action-level one-sided exact $p = 5.29 \times 10^{-89}$; the complete-query gain and paired harm-rate difference provide cluster-aware inference. All five pre-specified gates pass. The 1,083 certificates comprise 650 screen-passing actions and 433 counterexample certificates with negative feasible upper bounds. The largest numerical certificate gap is 9.75×10^{-7}.

Label-free certificate anatomy. A witness-instrumented replay reproduces all 3,249 frozen fold bounds with maximum absolute error 1.6×10^{-15}. In 97.3% of fold problems, a worst witness lies within 1% of a Fr'echet endpoint; 87/1,083 actions have any fold whose witness is in the interior. Thus the intended interaction extremes, rather than a broad interior plateau, usually determine acceptance. The stress is conservative in another sense: 73.4% of worst local inputs exceed five training-standardization units on at least one local feature. We do not treat those outputs as compatible target posteriors. They are adversarial score-space inputs to the frozen composer, and observed target action tests remain necessary.

Trust-region decomposition. The endpoint result leaves two explanations entangled: a rejection may reflect a preference reversal already near the composer's fitted feature region, or it may depend on a remote extrapolation. We separate them without labels. Let $z_\ell(\theta)$ be the three standardized log-local features seen by the frozen composer. For radius r, we intersect each fold's Fr'echet path with $\{\theta : \|z_\ell(\theta)\|_\infty \le r\}$ and rerun the same exact branch-and-bound certificate. A counterexample in this subset does not require more remote local-feature extrapolation. A rejection that disappears is only called *extrapolation-sensitive*; it is not declared incorrect.

Table 13: Post-hoc, label-free PIAF trust-region decomposition. Each fold's Fréchet path is intersected with an axis-aligned box in standardized local-feature space, then certified by the same branch-and-bound procedure. Percentages use the 433 original PIAF rejections. No outcome label or effectiveness endpoint enters this analysis.

Local trust radius	Evaluable actions	In-box counterexamples	Extrapolation-sensitive
3 standardization units	1,083/1,083	358/433 (82.7%)	75/433 (17.3%)
5 standardization units	1,083/1,083	383/433 (88.5%)	50/433 (11.5%)

All 650 original acceptances remain accepted because each trust region is a subset of the already certified Fr'echet set. At radius five, in-box counterexamples remain for all 227 rejected $P \to S$ actions, all 40 $S \to P$ actions, and all 31 $S \to I$ actions. The main extrapolation-sensitive stratum is the more aggressive $P \to I$ demotion: 83/131 rejections remain inside the box. Thus composer extrapolation is a real source of PIAF conservatism, but it is not the sole explanation for most frozen rejections. This post-hoc diagnostic reads no confirmation label, changes no action or threshold, and supplies no posterior or target-risk guarantee.

Outcome-opened trust-region policy sensitivity. After the label-free analysis was complete, we joined each frozen trust-region decision to the already opened GenSearch outcomes. This asks what would have happened if the restricted path had been deployed as the policy. It does not select a radius or create a confirmation result.

Narrowing the uncertainty set recovers average gain, as expected, but damages previously correct Partial anchors more often. At 3σ, 75 extra actions increase exact correct-P-anchor harms from one to 13. At 5σ, 50 extra actions increase them to nine. This test-opened frontier explains why we keep the full Fr'echet set as the primary policy despite its lower action coverage.

Table 14: Test-opened trust-region policy sensitivity on fresh GenSearch. Each restricted policy intersects the Fréchet path with a standardized local-feature box. Newly allowed/lost counts are relative to full-path PIAF. C/H/N denotes corrections, harms, and neutral changes. This table diagnoses a tradeoff; it does not select a radius or create a confirmation result.

Coupling set	Actions	Newly allowed/full lost	C/H/N	Cluster gain	Action harm	Correct-P anchor harms
2σ local box	796	149/3	595/78/123	+6.16	9.80	25
3σ local box	725	75/0	540/67/118	+5.67	9.24	13
5σ local box	700	50/0	521/63/116	+5.47	9.00	9
Full Fréchet path	**650**	**0/0**	**489/54/107**	**+5.15**	**8.31**	**1**

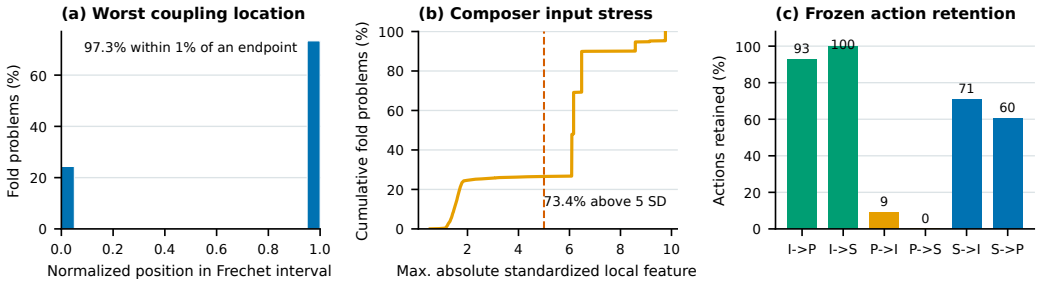


Figure 3: Post-hoc, label-free anatomy of the 3,249 frozen PIAF fold problems. (a) Worst witnesses concentrate at Fréchet endpoints. (b) Their local inputs often exceed the composer's training-standardization range, which limits posterior interpretation. (c) Retention is strongly transition-dependent. No confirmation label, action, threshold, or effectiveness endpoint is used in this analysis.

The exact-harm mechanism is unusually clear. PIAF removes 71 of 72 actions that damage a previously correct Partial anchor, while retaining 61 exact corrections to Partial. This is narrower than full boundary recovery. Among 149 selected gold-Partial rows, 88 final predictions still lie at an endpoint; 87 are wrong-to-wrong ties under exact zero–one risk and one is an exact harm. PIAF also retains 54 exact harms overall, mostly on gold Support or Insufficient rows. It therefore screens sensitivity to ambiguity in the frozen local projection. It does not identify target posteriors or certify calibration, encoder correctness, composer extrapolation, or cross-domain transport.

The complete harm transition matrix makes that scope explicit. Counts refer only to actions taken by each policy.

Table 15: Exact action harms by gold relation and authorized candidate on fresh GenSearch. PIAF chiefly protects the interior Partial relation; it is not a generic harm detector.

Gold $\rightarrow$ candidate	Unfiltered RA-DVRC	Fréchet PIAF
$I \rightarrow P$	6	5
$I \rightarrow S$	23	23
$P \rightarrow I$	27	1
$P \rightarrow S$	45	0
$S \rightarrow I$	14	9
$S \rightarrow P$	26	16

PIAF removes 71/72 harms whose gold label is Partial, but only 15/40 whose gold label is Support and 1/29 whose gold label is Insufficient. This is the expected signature of the declared ambiguity contract: it protects the interior ordinal relation most strongly and is not a generic harm detector.

The correction/harm counts also define a transparent operating frontier. Give each correction utility one, each harm cost λ, and each neutral or kept decision utility zero. Unfiltered RA-DVRC has utility $775 - 141\lambda$; PIAF has $489 - 54\lambda$. PIAF is preferred when $\lambda > 286/87 = 3.29$ and remains positive relative to keeping the anchor while $\lambda < 489/54 = 9.06$. Above that second crossover,

these observed outcomes favor keeping or review over automatic replacement. This is a post-hoc, target-specific interpretation of the frozen counts, not a fitted cost or a prospective risk guarantee.

G.4 Unified post-hoc action-layer comparison

We next compare action layers under one graded-relation contract. Every method receives the same frozen RA-DVRC proposals and zero–one replacement loss. An unselected proposal keeps the anchor in automatic-accuracy calculations and is sent to review; no human-oracle credit is assigned. The model-fit partition combines old GenSearch `dev-fit` and `dev-cal` (1,106 rows, 148 queries, 137 proposals). The disjoint official test calibrates risk (1,097/139/140), and fresh v38 is evaluation only (8,834/1,156/1,083). Query overlap between every pair is zero.

This is the same 1,083-proposal cohort as the frozen PIAF confirmation, with 650 accepted changes. Reusing its predictions avoids a model or cohort confound, but does not provide an independent replication. No evaluation label is used to fit the controls; their design and the comparison are retrospective.

The learned-harm selector is a 25-parameter grouped logistic model. The log-posterior baseline is a 12-parameter multinomial calibration map. The restricted S-MMCE selector has 417 parameters and calibrates candidate correctness without its original shift augmentation or DRO. Because this comparator family was designed after v38 outcomes were open, the result is out-of-sample but retrospective. It cannot revise the protected v38 claim.

At exactly 650 automatic actions, all PIAF-minus-comparator cluster-accuracy intervals include zero. PIAF therefore has no aggregate ranking-dominance claim. Damage to correct Partial anchors differs: PIAF causes one such exact harm versus 35–51 for every comparator. Complete-query intervals for this full-cohort burden all exclude zero. This endpoint does not count every selected gold-Partial row whose final relation remains outside Partial.

Table 16: Retrospective out-of-sample GenSearch comparison at exactly 650 automatic actions. Every method receives the same frozen RA-DVRC proposals and zero–one loss. Learned controls fit only older development queries. The final two columns give paired complete-query 95% intervals for PIAF minus each method, in percentage points.

Action layer	Extra params	C/H/N	Cluster gain	Harm	Correct-P anchor harms	PIAF−layer acc. 95% CI	PIAF−layer correct-P-anchor harm 95% CI
Fréchet PIAF	**0**	**489/54/107**	**+5.15**	**8.31**	**1**	–	–
Candidate confidence	0	530/55/65	+5.56	8.46	39	[-0.98,0.17]	[-3.26,-1.70]
Raw advantage	0	527/57/66	+5.51	8.77	37	[-0.92,0.22]	[-3.10,-1.58]
Log-posterior calibration	12	529/59/62	+5.57	9.08	37	[-0.95,0.13]	[-3.10,-1.59]
Learned harm selector	25	531/59/60	+5.60	9.08	35	[-0.92,0.02]	[-2.98,-1.47]
Restricted S-MMCE	417	476/84/90	+4.84	12.92	51	[-0.23,0.87]	[-4.16,-2.35]

As a safeguard against implementation reuse, a separate audit reconstructs the 24 fixed features and all five comparator scores directly from serialized parameters. It then regenerates all six 650-action masks and recomputes the policy outcomes and paired cluster bootstraps. All 133 checks pass. The three partitions have zero query and row overlap, and every ranked comparator has a strict score gap at the budget boundary. We release the 1,083-row selection matrix so that each count in the table can be traced to explicit action IDs. This audit is retrospective and changes no policy or evidential status.

G.5 Headroom-aware interpretation

A separate label-only oracle audit prevents the controlled comparison from being mistaken for an end-to-end SOTA result. The anchor makes 2,700 row errors. The frozen proposer emits 1,083 candidate changes: 775 corrections, 141 harms, and 167 neutral changes. Its candidate pool can therefore correct at most 28.7% of anchor errors. An oracle that takes only those 775 corrections gains 9.20 query-cluster accuracy points. Unfiltered replacement gains 7.57 points, whereas PIAF gains 5.15.

PIAF captures 489/775 available corrections (63.1%) and rejects 87/141 proposed harms (61.7%). Its 435 net corrections are 66.9% of the label-only upper bound at the same 650-action budget. These numbers expose the price of robust authority and the remaining selection headroom. They are not a deployable baseline: labels define the oracle, and the audit was specified after the GenSearch outcomes opened.

This decomposition also fixes the scope of horizontal comparisons. The matched-budget table asks which action layer behaves better given identical frozen proposals. A method-native end-to-end comparison asks whether a stronger verifier or proposer changes the candidate envelope itself. A defensible SOTA study must report both tracks, tune every learned method only on role-matched development clusters with disclosed search budgets, and compare full coverage–risk frontiers or identical action budgets. Protected labels may evaluate those choices but may not select them. The present paper establishes the first track and makes no detector-SOTA claim. The audit is generated by `scripts/analyze_iclr_action_headroom.py` from the released evaluation rows.

G.6 Post-hoc proposal–threshold sensitivity

The proposal thresholds precede PIAF, so we test whether the boundary result is an artifact of the selected $(\tau_\uparrow, \tau_\downarrow) = (.20, 0)$ point. After outcomes opened, we recover the threshold-free candidate from the saved posteriors and certify all 1,339 raw local–global disagreements. We then vary one directional threshold at a time. The original point is reconstructed exactly: 1,083 proposals, 650 certified actions, 489/54/107 corrections/harms/neutral changes, and one harm to a correct Partial anchor.

Table 17: Outcome-opened GenSearch proposal-threshold sensitivity. Each side varies one threshold while fixing the other at its frozen value. Gain is query-cluster accuracy in percentage points; harm is conditional on certified actions. No row selects a new policy or changes the confirmation claim.

	$\tau_\uparrow = t,\ \tau_\downarrow = 0$				$\tau_\uparrow = .20,\ \tau_\downarrow = t$			
t	Certified	Gain	Harm (%)	Correct-P harm	Certified	Gain	Harm (%)	Correct-P harm
.00	726	+5.55	9.09	1	650	+5.15	8.31	1
.05	726	+5.55	9.09	1	644	+5.14	8.23	1
.10	710	+5.35	9.01	1	616	+5.05	7.79	1
.15	678	+5.21	8.70	1	592	+4.99	6.93	1
.20	650	+5.15	8.31	1	579	+4.90	6.91	1
.30	573	+4.65	7.68	1	557	+4.88	6.46	1
.40	490	+3.89	7.76	1	544	+4.87	6.07	1
.50	382	+2.91	8.90	1	539	+4.81	6.12	1

The certified boundary outcome is stable across this one-at-a-time grid: every setting keeps exactly one harm to a correct Partial anchor, while every cluster-accuracy change remains positive. Within each threshold-specific proposal pool, we also rank raw advantage and candidate confidence at the same action count as PIAF. Their correct-Partial harm ranges are 18–42 and 20–47, respectively; PIAF remains at one. The original $(.20, 0)$ counts, 37 and 39, exactly reproduce the independent matched-budget audit. This analysis answers threshold sensitivity only. It is outcome-opened, supplies no model-selection evidence, and does not establish target-risk control.

G.7 Retrospective proposal–certificate–queue cascade

The primary replacement and review studies are stage-wise: WCSD evaluates the full warning set rather than only PIAF rejections. For deployment accounting, the table below instead applies one hypothetical cascade to every population where frozen PIAF replays exist. Certified proposals replace the anchor; all other proposals enter a review queue. No human-oracle correction is credited. Queue precision is the fraction whose anchor is wrong, counting neutral proposal changes because both anchor and candidate miss the gold label.

This table answers how disagreements partition under one ladder, but it does not create a prospective end-to-end trial. In particular, the full-warning GenSearch WCSD endpoint remains the protected review evidence, and the role-isolated GenSearch queue has no independently locked review policy.

The following deterministic audit makes the score-space reversal concrete and shows both sides of conservatism. The first row illustrates a foregone correction; the second is the highest-confidence harmful gold-P, $P \to S$ case among the 650 matched candidate-confidence actions that PIAF rejects. These empirical cases complement the analytic illustration in Section 3.4. Case selection is post-hoc. Text is paraphrased, and the released unit IDs reconstruct the exact public GenSearch rows.

Candidate confidence therefore takes both actions, while PIAF rejects both without reading their labels. At the upper endpoint in every fold, increased joint support–missingness mass shifts the local

Table 18: Retrospective cascade accounting. C/H/N denotes outcomes among certified automatic replacements. Auto gain is cluster-macro accuracy in percentage points; queue precision measures anchor-error enrichment, not reviewed accuracy. WiCE, ALCE, and protected GenSearch PIAF replays are test-opened.

Population	Proposals	Certified (%)	Certified C/H/N	Auto gain	Queue / precision
WiCE official	33	19 (57.6)	17/0/2	+5.06	14 / 92.9%
ALCE transfer	28	8 (28.6)	8/0/0	+5.15	20 / 60.0%
GenSearch protected	140	75 (53.6)	54/10/11	+5.24	65 / 78.5%
GenSearch role-isolated	1,083	650 (60.0)	489/54/107	+5.15	433 / 79.9%

Table 19: Case audit for candidate-confidence actions rejected by PIAF. Nominal/worst gives the ensemble product-coupling advantage and certified Fréchet worst-case upper bound. U denotes the upper joint-mass endpoint in each fold. Outcome labels are used only to classify the already frozen actions.

Public row	Action / gold	Claim–evidence relation	Confidence; nominal → worst	Worst θ
Train 16	$P \to S$ / S	The source directly states that New York City passed the 2012 rule capping sugary-drink containers above 16 ounces. PIAF loses a correction.	.681; +.466 → -.831	U/U/U
Train 652	$P \to S$ / P	The citation says antivenom neutralizes venom, but does not establish the claim's antibody-fragment binding mechanism.	.755; +.579 → -.727	U/U/U

view toward partial support and reverses the nominal promotion. The first row exposes PIAF's cost in foregone corrections; the second exposes its protection against overpromotion. They are mechanism examples, not estimates of case-type prevalence.

The same comparison also composes PIAF with query-cluster conformal risk control. Each score threshold is calibrated on the 139 disjoint official-test queries for the bounded loss `any harmful automatic replacement in a query`. The table reports the predeclared 10% target; all five scores also meet their observed targets at 5%, 15%, and 20%. Action harm and cluster risk have different denominators and should not be conflated.

Table 20: Retrospective out-of-sample GenSearch action control at the 10% query-cluster any-harm target. CRC thresholds use only the disjoint official test. PIAF ∩ CRC can only remove CRC-authorized actions, so it preserves the bounded-loss control under the same exchangeability assumptions. C/H denotes corrections/harms; values are percentages.

	CRC only				PIAF ∩ CRC				
Score	Actions	C/H	Action harm	Cluster risk	Actions	C/H	Action harm	Cluster risk	Gold-P harms
Candidate confidence	778	613/79	10.15	6.40	485	402/26	5.36	2.08	1
Raw advantage	807	623/82	10.16	6.57	524	421/31	5.92	2.42	1
Log-posterior calibration	770	602/82	10.65	6.49	496	404/32	6.45	2.42	1
Learned harm selector	808	627/82	10.15	6.66	543	437/32	5.89	2.51	1
Restricted S-MMCE	798	591/96	12.03	7.70	492	379/37	7.52	2.85	1

The intersection reduces corrections as well as harms. Under utility $C - \lambda H$, it becomes preferable to its corresponding CRC-only policy at $\lambda = 3.59$–3.98, depending on the score. This is the intended division of labor: CRC controls population harm under exchangeability, while PIAF screens the unidentified support–missingness interaction and sharply reduces exact damage to correct Partial anchors. The post-hoc timing prevents a new confirmatory claim, but the unified comparison answers why the two controls are not substitutes.

G.8 Post-hoc cost-family sensitivity

Proposition 4 also permits a loss family to be declared before outcomes are used. To show what this extension does, we run a deliberately post-hoc sensitivity analysis on the now-open v38 outcomes. The family contains zero–one loss and the two elementary ordered-boundary losses: crossing $I \mid \{P, S\}$

and crossing $\{I, P\} \mid S$. Their nonnegative weighted sums span weighted absolute ordinal loss. An action is retained only if it weakly dominates its anchor for every compatible coupling and every vertex. No threshold or loss weight is fitted.

For deployment, one loss vertex should encode one comparison stakeholders refuse to trade against another, such as crossing $P \mid S$ or $I \mid P$. Normalize vertices to a common scale, include each plausible extreme, and declare the convex hull before target outcomes are inspected. Vertex sufficiency then certifies the whole hull. If coverage becomes too small, the honest response is lower authority or independently calibrated population control, not deleting an inconvenient vertex after seeing labels.

Table 21: Post-hoc cost-family sensitivity on the opened v38 outcomes. The stricter policy requires coupling-robust dominance for zero–one loss and both ordered-boundary loss vertices. It is a scope diagnostic, not a new confirmation. C/H/N denotes corrections, harms, and neutral changes.

Policy	Actions	C/H/N	Cluster gain [95% CI]	Action harm [95% CI]	Correct-P anchor harms
Zero–one PIAF	650	489/54/107	+5.15 [4.42,5.91]	8.31 [6.30,10.70]	1
Zero–one + boundary-cost PIAF	118	72/26/20	+0.51 [0.23,0.82]	22.03 [14.93,30.59]	1

The stricter filter reduces the absolute number of harms from 54 to 26, but it also changes the selected population. Its conditional harm rate rises to 22.0%. At the full-cohort level, realized cluster-macro loss still decreases for all three vertices: 0.51 points [0.23, 0.83] for zero–one, 0.42 [0.18, 0.70] for the lower boundary, and 0.30 [0.11, 0.51] for the upper boundary. This is not an action-safety improvement. It is a useful separation: robustness to coupling and declared cost does not imply calibrated harm-rate control on the selected target cases. We therefore keep zero–one PIAF as the confirmed replacement policy and report the cost-family extension only as a scope diagnostic.

G.9 Boundary-separable certification audit

The original certificate evaluates the endpoint candidate against the anchor. For a two-boundary proposal, a favorable crossing can therefore compensate for an unfavorable one. We test a stricter, test-opened extension. Let $y_0, \ldots, y_k$ be the adjacent path from anchor to candidate. The extension certifies y_j against y_{j-1} under the same coupling and loss sets, then stops at the first failed crossing. It emits the farthest certified prefix.

If every accepted adjacent step has nonnegative advantage for every declared coupling and loss vertex, summing those adjacent loss differences shows that the emitted prefix weakly dominates the anchor under the same contract. The prefix is maximal among contiguous prefixes satisfying these stepwise tests. This is still a model-score-space statement, not a target-risk guarantee.

Table 22: Test-opened boundary-separable PIAF audit. Each adjacent crossing is certified separately under zero–one and both ordered-boundary loss vertices. Harm is exact zero–one action harm. Final exits count selected gold-Partial rows whose authorized relation is an endpoint. This is mechanism evidence, not a new confirmation.

Opened cohort / backbone	Loss family	Actions	C/H/N	Harm (%)	Gold-P final exits	Cluster gain [95% CI]
WiCE	zero–one + boundaries	16	16/0/0	0.0	0/16	+4.75 [+2.61,+7.21]
ALCE	zero–one + boundaries	3	3/0/0	0.0	0/3	+2.33 [+0.00,+5.88]
GenSearch	zero–one + boundaries	121	72/25/24	20.7	1/71	+0.54 [+0.25,+0.87]
SciTail / DeBERTa	zero–one + boundaries	25	18/1/6	4.0	1/5	+0.67 [+0.25,+1.17]
SciTail / RoBERTa	zero–one + boundaries	236	225/6/5	2.5	6/7	+7.79 [+5.52,+10.16]
SciTail / ModernBERT	zero–one + boundaries	0	0/0/0	–	0/0	0.00 [0.00,0.00]
BingCheck / DeBERTa	zero–one + boundaries	102	72/26/4	25.5	24/31	+1.39 [+0.79,+2.01]
BingCheck / RoBERTa	zero–one + boundaries	91	52/34/5	37.4	30/32	+0.55 [+0.02,+1.09]
BingCheck / ModernBERT	zero–one + boundaries	2	1/1/0	50.0	0/0	-0.01 [-0.08,+0.06]

The stepwise rule closes a real compensation loophole and is strong on SciTail. It does not rescue BingCheck. This mixed result is useful: coupling, loss, and boundary robustness cannot correct a semantically unreliable local view. We therefore treat the extension as mechanism evidence and require a new label-sealed population before making it part of the headline policy.

A second, test-opened SciTail replay directly tests direction-asymmetric costs. Besides zero–one loss, it assigns cost three to leaving a correct partial verdict and cost three to predicting support on a gold-insufficient row. We compare it with the parameter-free pair of ordered-boundary vertices. Neither family is tuned, but the value three was chosen after outcomes opened and has no confirmatory status.

Table 23: Post-hoc SciTail loss-family sensitivity. C/H/N denotes corrections/harms/neutral actions; deltas are question-cluster accuracy and partial-support recall in percentage points. All six policies pass their point gates and fail uncertainty-aware gates.

Backbone	Added loss vertices	Actions	C/H/N	Harm (%)	Acc. Δ	P-recall Δ
DeBERTa	ordered boundaries	15	14/1/0	6.7	+0.52	-0.22
DeBERTa	asymmetric 3×	17	14/2/1	11.8	+0.42	-0.44
RoBERTa	ordered boundaries	314	302/6/6	1.9	+10.44	-0.88
RoBERTa	asymmetric 3×	332	302/10/20	3.0	+9.92	-1.10
ModernBERT	ordered boundaries	1	1/0/0	0.0	+0.04	0.00
ModernBERT	asymmetric 3×	7	7/0/0	0.0	+0.30	0.00

The asymmetric family is more permissive and incurs more observed harms for DeBERTa and RoBERTa; zero harms for seven ModernBERT actions are too sparse for a safety conclusion. This is precisely why a domain loss polytope should be elicited before outcomes: specify the operational action and review option, obtain stakeholder loss matrices for several plausible harm scenarios, use those matrices as vertices, and report authority and outcomes for every vertex. In medicine, pitfalls include treating a claim-level harm score as an action loss, fitting costs to rescue an opened endpoint, double-counting one clinical consequence across vertices, and interpreting robust preference as calibrated target risk. VetScore illustrates separate expert harm elicitation, but its claim weights and PIAF's decision losses remain distinct (Kartáč et al., 2026).

H Artifact and Reproduction Map

The anonymous artifact exposes the following release-safe layers:

1. Frozen public-data preparation, deterministic chunking, grouped-split, inference, composition, and evaluation code.
2. Protocol, implementation, prediction, and evaluator locks with the hashes of every frozen code and prediction artifact.
3. Global and local checkpoint metadata, label-free predictions, and three serialized calibrators with feature order, scaler statistics, coefficients, intercepts, and fit-unit hashes.
4. The full development frontier, grouped bootstrap records, protected WiCE and GenSearch aggregate evaluations, and all stopped or failed gate decisions.
5. Frozen ALCE transfer predictions, role-aliased final labels for every changed decision, paired evaluation, and post-hoc diagnostics.
6. The v38 GenSearch confirmation protocol, label-free predictions, 1,083 coupling certificates, all three frozen policy summaries, and every gate.
7. A standard-library verifier that checks locked hashes, counts, status decisions, interval endpoints, headline arithmetic, and the complete v27 stress-test conclusions in one command.
8. The AVeriTeC multi-view score artifacts, fixed LTT threshold sequence, 160-cluster calibration trace, all six confirmation masks, and independent 211-cluster confirmation summary.
9. The protected ExpertQA split and policy specifications, label-free three-backbone score records, one-time aggregate evaluation, and the generated aggregate–boundary digest and figure.
10. The prospective SufficientFacts cohort, frozen three-family endpoint predictions, label-free action manifest, PIAF certificates, complete matched-authority evaluation, and all review-budget operating points.
11. The label-free v57 trust-region replay, including all 1,083 row-level interval results, exact construction and tests, summary arithmetic, and frozen v38 input-hash checks.
12. The v58 BingCheck protocol, text-free action records, 1,032 PIAF certificates, complete three-family evaluation, and opened-result evidence-presence decomposition.

13. The v59 SciTail protocol, label-free action manifest, 1,585 PIAF certificates, complete three-family evaluation, and all four registered deferral-budget comparisons.
14. The v77 SALAD protocol, 18 label-free prediction arms, 1,403 PIAF certificates, label-free activation decision, complete three-family evaluation, and all 12 family-budget deferral comparisons.

The artifact contains no benchmark claim or evidence text. It also excludes sealed label keys, private annotation workbooks, mapping keys, consent records, reviewer identities, credentials, local paths, logs, and model checkpoints. Text-bearing inputs and evaluation labels can be reconstructed from official public sources using the recorded source revisions and SHA-256 values. Full bootstrap analyses can then be rerun from the released code. The included checker verifies the immutable label-free predictions and all reported result records without requiring private material.

The test pipeline checks unique IDs, exact row alignment, posterior normalization, relation order, expected hashes, and label absence during inference. The release builder uses an explicit file allowlist and rejects raw text fields, private file types, machine-specific paths, server identifiers, and credential patterns before creating the archive.

I Expanded Limitations

Evaluation scope. The prospective replacement and deferral claims begin in the locked WiCE and GenSearch settings. Frozen ALCE transfer adds cross-dataset evidence for paired differences on changed decisions: its mean effect is positive, but its action-safety gate fails. TREC adds 4,724 human-assessed sentence–citation pairs and confirms the action split under seven locked deferral selectors. Its protected cohort nevertheless contains only 13 topics, so topic-bootstrap uncertainty cannot stand in for arbitrary domain coverage. XRAG extends the stress test across four backbones and five languages, where selector rankings remain target-dependent. Protected ExpertQA adds 581 test questions and three graded encoder families, but its native missing-evidence rows make the evidence-present support/partial subset the more diagnostic severity test. Inputs use checked, non-truncating 512-token pairs, so native long-context behavior remains outside the evaluated scope.

SALAD adds 3,856 human-majority sentence–evidence judgments in 100 long-form QA question clusters, including 234 native partial-support outcomes. The full frozen ladder runs across DeBERTa, RoBERTa, and ModernBERT. Its pooled PIAF interval is positive, but only RoBERTa passes the strict replacement gate. WCSD's deferral-precision interval favors every registered control at all 12 family-budget points. This is multi-backbone evidence for the authority split, not a universal replacement result.

BingCheck adds a separate protected test with 3,791 claims, including 899 native partial-support outcomes, across the same three encoder families. Its pooled PIAF accuracy interval is positive, but every family fails the registered replacement gate. This is evidence for aggregate–boundary mismatch, not for portable automatic replacement.

SciTail adds a cross-task protected test with 2,593 high-consensus premise–claim pairs, including 454 native partial-support outcomes, in 171 official test-question clusters. Its pooled PIAF accuracy interval is positive, but only one family passes the registered replacement gate. At 15% review, WCSD has the best precision and retained-risk point estimate in all three families. This reproduces the action split; it does not convert the replacement failure into a success.

Absolute performance and construct. Official-test cluster accuracy is 54.48%. RA-DVRC is a better graded verifier, not a solved citation authority. It evaluates support of a fixed claim by fixed evidence. It does not measure world truth, retrieval completeness, long-document retrieval quality, or whether the source causally influenced generation.

Adaptive development. Composer form and the promotion threshold use official WiCE training out-of-fold predictions, so their development estimates are adaptive. RA-DVRC replacement is confirmed on the protected official WiCE test. PIAF is then confirmed separately on previously unused, query-disjoint GenSearch data after its rule and gates are fixed. The WiCE view ablations, label-shift analysis, and opened-label ALCE diagnosis are explicitly exploratory and do not alter either frozen result. WCSD is chosen after GenSearch `dev-fit`, making that split adaptive. Its claim rests on the independently opened `dev-cal` and protected official test. The pooling sensitivity

and learned/conformal deferral comparisons were specified after protected labels were known. They are reported only as post-hoc robustness and mechanism checks.

The AVeriTeC study was also sequential. Its first fixed policy (v34) used the train search and certificate partitions adaptively and failed the 15% risk certificate at high automatic coverage. We retain that failure. We designed the v35 LTT protocol only afterward, while every official-development label was still unopened. The v35 score, threshold family, hash split, test order, and comparators were then frozen before opening its 160-cluster calibration partition. The final 211-cluster result is independent confirmation for v35, not a fresh confirmation of the earlier v34 coverage choice.

Calibration and shift. Posterior advantage equals conditional zero–one risk reduction only for a calibrated target posterior. Consequently, transfer safety requires target-representative calibration or direct action validation. The relation-wise label-prior sensitivity also assumes fixed class-conditional behavior; it does not establish that domains differ only in label priors. WCSD's retained-risk identity is exact for a fixed selected set, but future risk reduction still depends on warning-error enrichment persisting under shift. Operationally, reserve target calibration clusters and bound corrections-minus-harms for predeclared action strata directly. If no such labels exist, neither ECE nor PIAF identifies ε_t; the lower-authority interface is review rather than automatic replacement.

We do not fit separate temperature or Dirichlet calibrators to the two branches. The 21-parameter composer is already an affine multiclass map over both sets of log probabilities. Separate calibration would add another adaptive stage and would need its own nested OOF protocol. More importantly, better source NLL or ECE would not certify target actions. The ALCE frontier shows that every rank-preserving transform of the frozen advantage still fails the harm ceiling.

The TREC cluster-risk controller also does not supply a useful guarantee. Warning priority meets the tested .20 and .30 error-mass targets only by deferring every row; other scores need 42–83% review. These are observed target checks over 13 protected topics, not distribution-free control under shift. The useful TREC evidence is the locked equal-budget comparison, not the vacuous controller.

AVeriTeC supplies a different kind of result. Fixed-sequence LTT controls the familywise probability of releasing a threshold whose retained article-level risk exceeds 15%, under independent exchangeable article clusters and a pre-fixed score and threshold family. This guarantee is same-population and action-specific. It does not survive arbitrary domain shift, validate the binary label map for a new deployment, or certify individual rows. The independent confirmation checks the selected policy once; it is not part of the LTT guarantee.

Human audit scale. The motivating hard-router audit has eight changed rows and identifies a mechanism rather than prevalence. Separate 16-row RAGTruth and 13-row ALCE warning audits triangulate diagnostic precision but cannot be pooled with it. The transfer audit labels all 28 ALCE actions, yet its exact harm-rate interval remains wide. The primary WiCE confirmation depends on none of these human audits. GenSearch uses public three-way judgments and no new localized audit; this improves scale but does not measure reviewer workload or organization-specific cost. The dual-view implementation also roughly doubles recorded accelerator work relative to its holistic ensemble alone (Appendix J), so deployments must trade review enrichment against evidence-length cost.

Architecture and relation coverage. The protected graded WiCE replacement and GenSearch/TREC action confirmations use the DeBERTa-v3-base family. A protected SALAD study evaluates the complete ladder on the same DeBERTa, RoBERTa, and ModernBERT families under native human $S/P/I$ labels. The pooled replacement effect is positive, but the two-family safety gate fails; the deferral-precision result replicates across every family and locked budget. A prospective Sufficient-Facts evidence-omission test evaluates frozen DeBERTa-v3, RoBERTa, and ModernBERT systems on 1,696 paired omissions. Every RA-DVRC and PIAF endpoint- and pair-effect interval is positive, but the cohort contains no partial-support outcome and PIAF does not uniformly improve conditional harm. A label-open external replay additionally evaluates WiCE-trained DeBERTa-v3, RoBERTa, and ModernBERT models on TREC without target-label fitting. Protected ExpertQA evaluates those three families under one frozen protocol and one-time test, but it does not show universal success: aggregate accuracy rises while the evidence-present support/partial boundary is neutral or worse. AVeriTeC independently confirms binary article-cluster deferral using the same three encoder families. XRAG supplies broader architecture and language stress evidence, not confirmation of a universal policy. Protected BingCheck directly tests 899 native partial-support outcomes and also fails the

replacement safety gate. Protected SciTail tests another 454 native partial-support outcomes and has one registered point-gate pass, but its 17 actions do not establish the declared harm target under an exact one-sided audit. Across GenSearch, TREC, ExpertQA, BingCheck, SciTail, SALAD, and XRAG, partial-support errors remain the hardest to capture. Controlled multilingual confirmation with a healthy three-way anchor is still needed.

I.1 Retrospective graded architecture sensitivity

We train the same local and global task graph around NLI-tuned DeBERTa-v3, RoBERTa, and ModernBERT encoders. Architecture-specific source classifiers are discarded, and the shared heads are randomly initialized. Every family uses the same WiCE title folds, effective batch size, loss, composer, and directional thresholds. Tokenizer-specific overflow splits evidence only. The official WiCE test key was already open from the primary confirmation, so this is a retrospective sensitivity analysis, not another confirmation. The shared-head DeBERTa arm also does not reproduce the source-head initialization of the sealed primary model.

Table 24: Retrospective WiCE architecture sensitivity. Deltas and intervals are cluster-accuracy percentage points relative to each same-family anchor. PIAF is an opened-outcome application of the existing zero-parameter rule; it does not replace the complete RA-DVRC result. C/H: corrections/harms.

Family	Anchor	RA-DVRC Δ [95% CI]	RA C/H	PIAF kept	PIAF Δ [95% CI]	PIAF C/H
DeBERTa-v3	60.54	+2.75 [-.02,5.64]	18/7	9/27	+1.00 [-.54,2.68]	6/2
RoBERTa	69.90	-.56 [-4.40,3.27]	19/22	33/41	+.05 [-3.37,3.48]	16/17
ModernBERT	67.14	-6.25 [-10.48,-2.22]	16/38	2/55	.00 [-.92,.92]	1/1

The raw composer is positive only for DeBERTa, and no family has a strictly positive 95% interval. All 22 RoBERTa harms and 37/38 ModernBERT harms have gold partial support. This is severity polarization: recall improves at the two end relations while partial-support recall falls. PIAF is applied after these outcomes were open, although its rule, zero threshold, and numerical tolerance predate this analysis. It filters 53/55 ModernBERT actions and moves the -6.25-point loss to 0.00; the paired PIAF-minus-RA-DVRC interval is [2.30, 10.33]. That is a failure-containment result, not a cross-architecture gain. The experiment therefore sharpens the scope claim: replacement authority is encoder-dependent, while coupling robustness can respond by becoming nearly vacuous.

I.2 External TREC replay with the same three-way anchors

The preceding WiCE result asks whether the shared training graph behaves similarly across encoders on an already opened in-domain test. We also replay each complete WiCE-trained policy on all 4,724 TREC pairs from 13 topics. The three fold checkpoints, fold-specific composers, promotion/demotion thresholds, and PIAF certificate are unchanged; no TREC label fits a model, calibrator, or cutoff. TREC labels and the WiCE architecture outcomes were already open, so the analysis is retrospective and all registered families are reported.

Table 25: Retrospective external TREC replay of the three WiCE-trained encoder families. Deltas and topic-bootstrap intervals are cluster-accuracy percentage points relative to each same-family anchor. No TREC label fits a model, composer, threshold, or PIAF certificate. C/H: corrections/harms.

Family	Anchor	RA-DVRC Δ [95% CI]	RA C/H	PIAF kept	PIAF Δ [95% CI]	PIAF C/H	PIAF harm
DeBERTa-v3	48.26	+9.34 [7.38,11.32]	592/140	589/847	+7.91 [6.36,9.53]	459/79	13.41
RoBERTa	46.49	+8.93 [6.15,11.52]	519/84	509/685	+7.45 [5.45,9.45]	415/58	11.39
ModernBERT	45.98	+6.97 [5.14,8.98]	399/57	128/510	+2.05 [1.13,3.18]	109/6	4.69

All PIAF topic-bootstrap intervals exclude zero. This repairs the narrow claim that the score-space filter only contains failure in non-DeBERTa models: it can retain positive external gains across three encoder families. It does not establish portable replacement safety. DeBERTa and RoBERTa still exceed a 10% descriptive harm rate, and 110/143 pooled PIAF harms have gold partial support. The result therefore strengthens architecture transfer while preserving the paper's action-safety conclusion.

J SELECTIVE REVIEW AND COMPARATORS

J.1 DEVELOPMENT STOP BEFORE PROTECTED EVALUATION

After the ALCE action-safety failure, we first attempted a target-calibrated boundary-risk controller on the 654-row GenSearch `dev-fit` split. Among 92 proposals from 44 queries, 64 are corrections, 17 harms, and 11 neutral. Its query-grouped OOF score has correction-versus-harm AUC 0.753, but every useful threshold increases the frozen partial-to-support error. v25 therefore stops before `dev-cal` label access or test inference. A scalar target threshold does not repair the observed boundary failure.

J.2 FROZEN WCSD PROTOCOL AND FIREWALL

WCSD is chosen after the v25 failure. It defers if and only if the already frozen RA-DVRC policy proposes a relation change. It adds no model, fitted coefficient, target-label calibration, or new threshold. At the same deferred row count, comparators select lowest confidence, highest entropy, or lowest margin. Before `dev-cal` labels, we lock the method and its gate. Deferral may not exceed 15%, and at least 20 queries must be deferred. Cluster-bootstrap lower bounds must be positive for risk reduction, error enrichment, and both advantages over low confidence. Passing `dev-cal` authorizes test inference. We then lock six outputs, three calibrators, combined predictions, 140 warning IDs, and the evaluator before one-time key access. The key and raw labeled source have mode `000` at every lock and are resealed afterward. No label reaches the GPU server.

Table 26: WCSD results across method-design and independent splits. Values are percentages. Dev-fit is adaptive evidence; dev-cal and official test are independent primary evidence.

Split	Role	Rows/queries	Deferred	Precision	Capture	Base	Retained	Reduction [95% CI]	Gain vs low confidence [95% CI]
Dev-fit	adaptive	654/89	92/44	81.5	40.1	28.6	19.9	+8.66 [5.72,11.93]	+20.65 [12.15,30.36]
Dev-cal	independent	452/59	45/28	86.7	32.8	26.3	19.7	+6.67 [4.36,9.22]	+42.22 [30.44,55.56]
Official test	protected	1097/139	140/84	82.9	35.2	30.1	22.4	+7.72 [6.06,9.60]	+23.57 [13.73,33.39]

J.3 MATCHED UNCERTAINTY CONTROLS AND RELATION COVERAGE

On `dev-cal`, matched low confidence retains 24.3% risk, 4.67 points above WCSD (95% interval [2.51, 7.15]). On the official test it retains 25.8% risk, 3.45 points above WCSD (interval [1.79, 5.25]). Matched entropy and margin also have lower deferral precision and higher retained risk on both independent splits. These comparisons hold coverage fixed, so the gain comes from which rows are selected rather than how many are withheld.

J.4 POST-HOC LEARNED AND CONFORMAL COMPARATORS

We add two label-trained error rejectors after the protected result. Under a perfect-review assumption, learning whether the anchor is wrong is the corresponding learning-to-defer special case. The anchor-only rejector has seven parameters over log anchor probabilities, confidence, entropy, and margin. The dual-view rejector has 25 parameters and additionally receives the composed posterior, advantage, action, direction, candidate class, and fold action fraction. Both are balanced logistic models. We choose $C \in \{.01, .1, 1, 10\}$ by five-fold query-grouped average precision on `dev-fit`; the selected values are .01 and .1. At evaluation time, each model defers exactly the WCSD row count.

The budget-matched LAC comparator calibrates score $1 - p_y$ on `dev-fit` and defers when its prediction set is not a singleton. We choose its nominal level only to approximate the `dev-fit` WCSD deferral rate. The resulting level is therefore data-selected rather than prespecified. We claim neither a fixed marginal coverage guarantee nor selective retained-risk control, and its evaluation coverage is not forced to equal WCSD.

WCSD and the dual-view learned rejector are statistically tied. For WCSD minus the learned rejector, deferral-precision intervals are [-2.93, 12.48] points on `dev-cal` and [-6.01, 2.91] on test. The corresponding WCSD retained-risk advantages are [-0.60, 1.66] and [-1.43, 1.06] points. WCSD is therefore competitive with a target-label-trained error model, not demonstrably better. It clearly outperforms the anchor-only and LAC selectors in these post-hoc comparisons. All families were

Table 27: Test-opened deferral comparison at the native WCSD budgets. Learned models fit only GenSearch dev-fit labels. Restricted S-MMCE omits shift augmentation and distributionally robust training; holistic-fold JS uses three locked checkpoints trained on different grouped folds and seeds, not a same-training-set deep ensemble. All methods except LAC match WCSD's row count exactly. LAC uses a dev-fit-selected quantile, so no fixed coverage guarantee is claimed. Values are percentages.

Split	Selector	Defer	Precision	Capture	Retained risk	Risk reduction
Dev-cal	WCSD	9.96	86.67	32.77	19.66	6.67
Dev-cal	Dual-view learned	9.96	82.22	31.09	20.15	6.18
Dev-cal	Restricted S-MMCE	9.96	40.00	15.13	24.82	1.51
Dev-cal	Holistic-fold JS	9.96	55.56	21.01	23.10	3.23
Dev-cal	Low confidence	9.96	44.44	16.81	24.32	2.00
Dev-cal	Anchor-only learned	9.96	44.44	16.81	24.32	2.00
Dev-cal	Budget-matched LAC	10.84	46.94	19.33	23.82	2.51
Test	WCSD	12.76	82.86	35.15	22.36	7.72
Test	Dual-view learned	12.76	84.29	35.76	22.15	7.93
Test	Restricted S-MMCE	12.76	51.43	21.82	26.96	3.12
Test	Holistic-fold JS	12.76	48.57	20.61	27.38	2.70
Test	Low confidence	12.76	59.29	25.15	25.81	4.27
Test	Anchor-only learned	12.76	63.57	26.97	25.18	4.90
Test	Budget-matched LAC	11.49	53.97	20.61	26.98	3.10

specified after the protected test result, fit or calibrate only on `dev-fit`, and are reported regardless of sign. They do not create a new confirmatory claim.

J.5 Post-hoc query-cluster risk control

Retained risk is a ratio whose denominator changes with deferral, so it is not the monotone loss required by conformal risk control. We instead define, for each semantic question q,

$$L_q(t) = \frac{1}{n_q} \sum_{i \in q} \mathbf{1}\{E_i = 1,\ s_i < t\},$$

where n_q is the query's original row count, E_i is anchor error, and rows with deferral score $s_i \geq t$ are reviewed. Additional deferral can only decrease L_q. On 59 `dev-cal` queries, we choose the least-deferring threshold whose bounded-loss correction $(\sum_q L_q + 1)/(59 + 1)$ does not exceed the target (Angelopoulos et al., 2024). We then apply that threshold unchanged to 139 test queries. The guarantee requires exchangeable query clusters; it is not an OOD or conditional retained-risk certificate.

The warning score encodes the previously reported lexicographic ranking as $4\mathbf{1}[\text{action}] + f + 0.25a$, where f is three-fold action frequency and a is posterior advantage. The coefficient gaps preserve action, then fold, then advantage priority. Learned dual-view and restricted S-MMCE scores fit only `dev-fit`; all thresholds use `dev-cal`. This loss, target grid, and comparison were specified after test opening, so every result below is post-hoc.

Table 28: Post-hoc query-cluster risk control on GenSearch official test. Each cell is test row deferral / query-normalized unreviewed error mass (percent). Bold cells meet the observed target using a threshold calibrated on `dev-cal`; exact unrounded values determine success. All six targets and seven scores are shown.

Target	Warning priority	Dual-view learned	Restricted S-MMCE	Holistic-fold JS	Low confidence	Entropy	Margin
.15	**100.0/0.0**	19.8/15.6	38.3/15.5	35.3/16.2	29.0/16.0	**30.4/15.0**	29.8/15.7
.18	**20.9/16.1**	**17.1/17.6**	**33.7/17.1**	22.3/21.5	24.0/18.9	23.5/18.7	23.4/19.9
.20	**20.9/16.1**	13.1/20.2	**31.4/18.1**	17.2/23.3	19.9/20.3	20.3/20.8	22.7/20.2
.22	11.0/22.2	**12.0/21.1**	**24.3/20.5**	14.4/24.3	**19.0/20.8**	15.7/23.1	**18.7/21.8**
.25	**7.5/25.0**	**7.4/24.4**	13.7/25.2	11.0/25.5	11.9/25.4	10.3/26.8	12.3/25.5
.30	1.1/31.6	0.8/31.7	1.1/32.2	0.6/31.8	0.6/32.0	0.5/32.3	2.3/31.4

At .18, warning priority and the learned dual-view score meet the observed test target while all three scalar uncertainty scores and holistic-fold disagreement miss it. The label-trained selector needs less review than warning priority; restricted S-MMCE meets the target only at 33.7% review. Across the full grid, no score dominates and several calibrated targets fail on test. We retain these failures because calibration does not erase split shift.

S-MMCE-style selective calibration optimizes calibration on accepted examples, which is not the retained-error objective used here. We nevertheless train a 417-parameter restricted S-MMCE selector on the same `dev-fit` dual-view features. It uses the normalized Equation 15 objective but omits the original shift augmentation and distributionally robust stage. At the frozen test budget its retained risk is 26.96%, compared with 22.36% for WCSD. This is a post-hoc implementation, not a claim that WCSD dominates the complete S-MMCE method. A confirmatory head-to-head would jointly prespecify selective NLL/ECE, retained risk, coverage, and review loss before a new protected split.

J.6 Post-hoc operating curves and ensemble disagreement

The overview below keeps the independent review evidence distinct from the pointwise certificate in main Figure 1. Neither curve conditions on PIAF rejection.

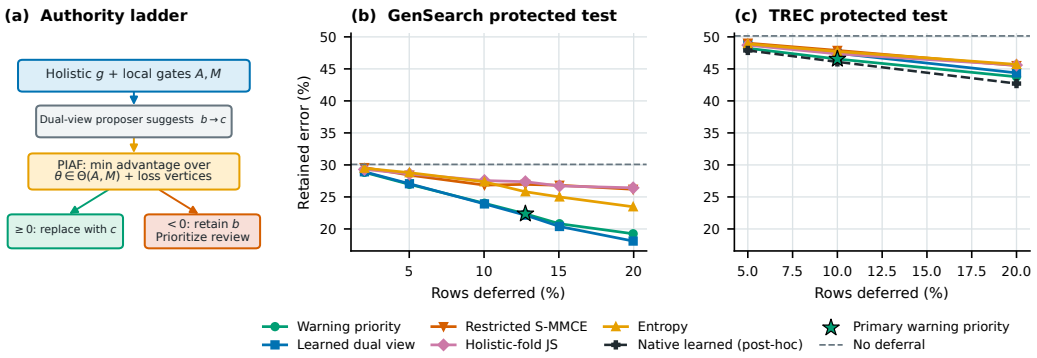


Figure 4: Action interfaces and separate review tests. Panel (a) shows the pointwise PIAF filter; target-risk control additionally requires the LTT layer. Panels (b,c) evaluate review over the full warning set, not a PIAF-rejection queue. GenSearch includes the protected operating point and test-opened curves. TREC includes locked controls and the dashed test-opened native selector; its 10% risk-difference interval includes zero.

Frozen WCSD is a binary action mask, not a tunable score. We therefore preserve its single confirmatory point and define a separate test-opened ranking for budget sensitivity. The ranking places every frozen action before every non-action. Within each set, it orders rows by the fraction of three fold arms that act, then by frozen posterior advantage, then by unit ID. Selecting the native row count exactly reproduces the frozen WCSD mask on all three GenSearch splits. Other budgets are descriptive.

We compare this ranking with the `dev-fit`-trained dual-view selector, restricted S-MMCE, low anchor confidence, and Jensen–Shannon disagreement across the three locked holistic fold posteriors. The latter controls for ordinary model disagreement (Lakshminarayanan et al., 2017), using checkpoints trained on different grouped folds and different seeds. It is not a same-training-set deep ensemble. Every method defers exactly the same number of rows at each point.

On official test, the warning ranking has lower retained risk than low confidence and holistic-fold disagreement at all six evaluated budgets. At the native 12.76% budget, its advantage is 3.45 points over low confidence (query- cluster 95% interval [1.84, 5.20]), 5.02 points over holistic disagreement ([2.85, 7.22]), and 4.60 points over restricted S-MMCE ([2.53, 6.79]). The difference from the learned dual-view selector is -0.21 points ([-1.41, 1.05]). At 20% review, the learned selector is 1.14 points better, but the interval [-2.74, 0.46] still includes zero. These opened-label comparisons show that the structured warning carries more useful ranking information than the tested scalar and holistic-disagreement controls. They do not establish a universal ordering of deferral methods.

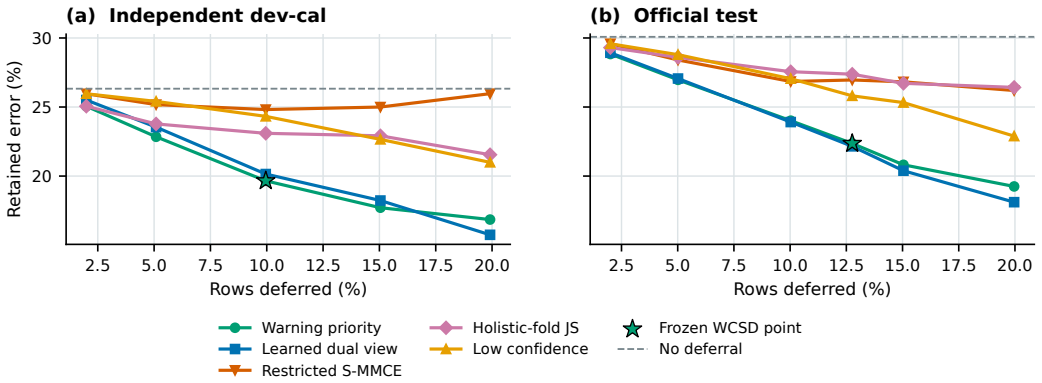


Figure 5: Test-opened GenSearch budget sensitivity. The star is the frozen WCSD point; the green curve is a post-hoc priority extension that exactly recovers that mask at the native count. All methods use identical row budgets. Curves are descriptive and do not replace the protected WCSD result.

J.7 Inference work and evidence-length scaling

The protected test contains 1,097 rows and 5,660 claim–evidence chunks: mean 5.16, median 3, 95th percentile 15, and maximum 58 chunks per row. No pair is truncated at 512 tokens. Across the three recorded CUDA/float16 fold processes, the holistic branch uses 182.37 elapsed seconds and the local branch 182.64. Sequential dual-view work is therefore 365.01 seconds, or 2.00 times the holistic ensemble alone (0.333 versus 0.166 recorded seconds per row). These are summed process times, not energy measurements or parallel service latency. Both neural branches visit every chunk, so work scales linearly with evidence chunk count. The 21-parameter CPU composer is not separately timed.

J.8 Relation-wise WCSD coverage

Table 29: WCSD error capture by gold relation on the two independent GenSearch splits.

Split	Relation	Anchor errors	Deferred errors	Capture (%)
Dev-cal	Insufficient	36	14	38.9
Dev-cal	Partial support	47	1	2.1
Dev-cal	Support	36	24	66.7
Official test	Insufficient	75	26	34.7
Official test	Partial support	146	17	11.6
Official test	Support	109	73	67.0

WCSD is strong for support errors, useful for insufficient errors, and weak at the adjacent partial-support class. It is a validated triage signal, not a complete selective solution for every relation.

The independent primary evidence is GenSearch `dev-cal` and official test. `Dev-fit` is adaptive; WiCE test is retrospective; ALCE is changed-case-only because unchanged rows lack human labels. Full artifact hashes and verification scripts are released with the supplementary archive.

J.9 Expanded methodological positioning

AlignScore provides a general alignment score, while recent decomposition studies analyze when smaller claim units improve or degrade factual evaluation (Zha et al., 2023; Hu et al., 2025; Lu et al., 2025). SURE-RAG combines sufficiency and uncertainty for selective answering (Qiu et al., 2026). Our action-conditioned setting instead holds the cited evidence fixed and separates a diagnostic's authority to replace a graded relation from its authority to request review. The deferral literature includes joint learning with expert decisions and calibrated learning to defer (Madras et al., 2018; Verma & Nalisnick, 2022). WCSD fits neither an expert model nor a new selector for its primary

claim. Finally, standard neural calibration motivates reporting NLL, Brier score, and ECE alongside action outcomes (Guo et al., 2017); those diagnostics do not substitute for the target action tests.

Recent systems can place verification inside the generation loop. VeriCite, for example, uses NLI checks to select evidence and refine an answer (Qian et al., 2025). That intervention changes evidence or text; our replacement experiment keeps the claim–evidence pair fixed and changes only its relation decision. Selective Conformal Risk Control (SCRC) and Selective Conformal Risk control with E-values (SCoRE) provide finite-sample selective guarantees under their respective assumptions (Xu et al., 2025c; Bai & Jin, 2026). Our AVeriTeC extension instead uses fixed-sequence LTT at the article-cluster action unit. These approaches control a population risk with calibration data. PIAF solves an orthogonal pointwise question about an unidentified interaction in a frozen score. Neither result turns the other into a cross-domain guarantee.

RA-DVRC is also narrower than general predictive-distribution combination. Bayesian locking and superposition introduce nonlinear pools and fitting objectives for posterior densities (Yao et al., 2023). Our categorical log-space stacker is deliberately low-capacity. PIAF can wrap a different composer only when its output can be recomputed along the declared local coupling path; the action certificate, rather than the pooling family, is the novel object studied here.

FActScore, MiniCheck, RAGChecker, and FactLens exemplify atomic, claim-level, pipeline-level, and subclaim-quality diagnostics (Min et al., 2023; Tang et al., 2024; Ru et al., 2024; Mitra et al., 2025). A label-only or black-box LLM judge can be evaluated as an alternative anchor or candidate using the same correction-versus-harm accounting. PIAF additionally requires relation scores that can be recomputed along the declared local A/M coupling path; it cannot certify a black-box label alone. Action conditioning is therefore an interface around compatible evaluators, not evidence that the present encoder beats every contemporary judge. Appendix S adds a same-input controlled comparison with two hosted generative judges. Both sharply improve average WiCE accuracy, but direct replacement harms 22.4–29.3% of changed decisions. Their agreement and confirmation policies do not dominate PIAF under the paper's low-harm interface.

K Protected XRAG Architecture and Language Stress Test

We ran a reviewer-motivated, prospectively locked stress test on XRAG (Liu et al., 2025). This study was specified after the paper's primary WiCE, ALCE, and GenSearch results, so it does not replace them. The official XRAG development data supplies `selector-fit` and `dev-cal`; the official test file is opened once after five prediction files and every executable policy are locked. English conditions contain 8,000 article pairs from 999 question clusters. The native multilingual condition contains 49,600 pairs from 2,199 clusters in Arabic, German, English, Spanish, and Chinese.

The original three-class architecture gate fails before `dev-cal` labels are opened: none of the five holistic heads predicts partial support. We therefore freeze a scope amendment that excludes replacement and relation-generalization claims. The remaining question is selective: at fixed review budgets, which frozen score best identifies errors made by an imperfect anchor? We compare WCSD, a label-trained dual-view logistic selector, an S-MMCE selector (Fisch et al., 2022), and global uncertainty. S-MMCE uses the normalized Equation 15 objective without the original shift augmentation or DRO stage. The primary budget is 10%; all methods, backbones, and language slices are reported regardless of sign.

Table 30: Protected XRAG selective-deferral stress test at the fixed 10% review budget. P/R denotes deferral precision and retained risk; all values are percentages. mMiniLM denotes multilingual MiniLM. The original three-class gate failed, so this table supports only selective-deferral analysis.

Condition	Backbone	Anchor acc.	WCSD P/R	Learned P/R	S-MMCE P/R	Uncert. P/R
English	DeBERTa-v3-base	77.7	61.0/18.0	80.5/15.8	68.9/17.1	68.8/17.1
English	RoBERTa-base	80.8	77.8/12.8	79.9/12.5	80.9/12.4	79.8/12.5
English	ModernBERT-base	80.7	63.1/14.4	81.1/12.4	79.2/12.7	79.5/12.6
English	mMiniLM	77.8	33.4/21.0	34.2/20.9	23.9/22.1	32.0/21.2
Five-lang.	mMiniLM	77.5	51.8/19.3	65.5/17.8	34.7/21.2	66.3/17.7

Here P/R denotes deferral precision and retained risk, and mMiniLM denotes multilingual MiniLM. Learned dual-view selection improves over WCSD on DeBERTa, ModernBERT, and the five-language condition. Query-cluster bootstrap intervals exclude zero for both metrics. On RoBERTa and English MiniLM, their differences include zero. WCSD does beat the restricted S-MMCE implementation on both MiniLM conditions, but not on the other encoders. Structured disagreement is therefore useful input to a learned selector; the zero-new-parameter WCSD rule is not an architecture-independent winner. Scalar uncertainty is also a strong control under the severe anchor degeneracy seen here.

We additionally calibrate cluster conformal risk control (Angelopoulos et al., 2024) on `dev-cal`. The loss is the indicator that a question cluster contains any retained anchor error. All 60 combinations of five conditions, four scores, and three targets miss their observed protected- test target. This does not contradict the conformal result: its expected-risk guarantee requires exchangeable clusters, while XRAG `dev-cal` and official test exhibit substantial score–error shift. The experiment instead shows why calibration-set control must not be presented as an OOD safety certificate.

The single global 10% policy also allocates review unevenly. WCSD defers 16.5% of English rows but 7.3–10.6% in the other languages. The learned selector is narrower at 9.0–12.6%; global uncertainty spans 9.5–11.9%. These are descriptive allocation differences, not fairness estimates. Taken together, XRAG directly tests the breadth criticism and yields a stricter conclusion: action utility depends on both the diagnostic and the backbone/domain calibration regime.

L Prospective TREC RAG Action Confirmation

L.1 Cohort and label firewall

TREC RAG 2024 evaluates sentence-level answers with cited passages (Pradeep et al., 2024). We use the official support judgments on the three-level 0/1/2 scale (National Institute of Standards and Technology, 2025). Each row is one answer sentence paired with one cited segment. Scores 0, 1, and 2 map to insufficient, partial support, and support. We exclude the official unassessed sentinel `-1`. Before inference, we also exclude rows whose sentence has no material lexical unit under the frozen tokenizer rule. Neither exclusion consults a model prediction.

The `without_prediction` partition is calibration: 2,837 eligible pairs from nine topics. `Webassess` is protected test: 4,724 pairs from 13 topics. Topic sets are disjoint. Official documentation states that assessors in both sets did not see a support prediction. Feasibility work inspected calibration examples and schemas, but not protected labels. Before one-time label access we lock the primary and RoBERTa prediction files, transferred selector artifacts, all selector scores, exact row budgets, risk-control thresholds, replacement rules, evaluator, and topic-bootstrap seed. One earlier local evaluation attempt stopped on a missing selector artifact before writing any result; the policy and endpoint files were unchanged.

L.2 Equal-budget selective deferral

Warning priority ranks frozen RA-DVRC actions before non-actions. Within each partition it uses fold action agreement, then frozen posterior advantage. This is the same score definition locked for every TREC row. Learned dual view and restricted S-MMCE are transferred from GenSearch and see no TREC label. Every method defers exactly 236, 472, or 945 rows at the three budgets.

At the primary 10% budget, topic-bootstrap intervals establish lower retained risk for warning priority than entropy, learned dual view, and restricted S-MMCE. Its point estimate is also lower than holistic-fold JS; that retained- risk interval crosses zero by .06 points, although deferral precision is higher with interval [1.57, 13.78]. Low confidence and margin are weaker than the displayed methods and remain in the result artifact.

Warning priority captures 325 of 1,360 insufficient errors (23.9%), 41 of 809 partial-support errors (5.1%), and 24 of 200 support errors (12.0%) at 10%. Its value is therefore not uniform across the relation lattice.

Table 31: Protected TREC equal-budget deferral results (percent). Each method defers exactly 236, 472, or 945 rows. Low confidence and margin are retained in the released result artifact.

Budget	Selector	Precision	Capture	Retained risk	Reduction
5%	Warning priority	86.86	8.65	48.22	1.93
5%	Holistic-fold JS	77.54	7.72	48.71	1.44
5%	Dual-view learned	73.73	7.34	48.91	1.24
5%	Restricted S-MMCE	71.61	7.13	49.02	1.13
5%	Entropy	73.73	7.34	48.91	1.24
10%	Warning priority	82.63	16.46	46.54	3.61
10%	Holistic-fold JS	75.42	15.03	47.34	2.81
10%	Dual-view learned	74.58	14.86	47.44	2.71
10%	Restricted S-MMCE	70.55	14.06	47.88	2.26
10%	Entropy	72.03	14.35	47.72	2.43
20%	Warning priority	75.66	30.18	43.77	6.38
20%	Holistic-fold JS	68.36	27.27	45.59	4.55
20%	Dual-view learned	73.23	29.21	44.38	5.77
20%	Restricted S-MMCE	68.68	27.40	45.51	4.63
20%	Entropy	67.94	27.10	45.70	4.45

L.3 Test-opened TREC-native learned control

The protected result above compares policies that were all locked before TREC test labels. After that endpoint opened, we fit the same 25-parameter grouped logistic error selector natively on the 2,837 calibration rows from nine topics and evaluate it on the 4,724 rows from 13 disjoint test topics. Topic-grouped cross-validation selects $C = .1$. The test topics fit no parameter, but knowledge of their earlier endpoint makes the entire comparison retrospective.

Table 32: Test-opened TREC-native selector comparison. P/R denotes deferral precision/retained risk; positive advantage means lower retained risk for the native selector. Intervals use the paired 13-topic bootstrap.

Budget	WCSD P/R	Native P/R	Native risk advantage [95% CI]
5%	86.86/48.22	93.22/47.88	+0.33 [-0.04,0.69]
10%	82.63/46.54	86.65/46.10	+0.45 [-0.16,1.04]
20%	75.66/43.77	79.89/42.71	+1.06 [0.001,2.02]

The native selector has the better point estimate at all three budgets. Its primary 10% interval crosses zero; the 20% interval is positive but comes from a nested, test-opened sensitivity point. This result rules out a broad claim that parameter-free warning priority dominates target-trained selection. It instead supports the narrower conclusion that WCSD is a useful zero-fit signal whose protected enrichment survives strong transferred controls. The exact artifact, hashes, fit, and all budgets are reproduced by `scripts/analyze_v70_trec_native_selector.py`.

L.4 Test-opened topic robustness

After the protected endpoint was known, we asked whether warning priority only looked useful because it allocated more reviews to difficult topics. Let q_t be anchor risk in topic t, $q_{A,t}$ warning precision in that topic, π_t its population share, and w_t its share among reviewed rows. The pooled enrichment has the exact decomposition

$$q_A - q = \sum_t w_t(q_{A,t} - q_t) + \sum_t (w_t - \pi_t)q_t.$$

The first term measures within-topic ranking; the second measures review allocation across topics. At 10% review, pooled enrichment is 32.48 points: 32.30 points arise within topics and only 0.18 from allocation. Thus 99.5% of the observed enrichment is not a topic-mix effect. Every topic has positive enrichment (13/13; median +29.93 points; exact two-sided sign test $p = 0.00024$).

The two other locked budgets give the same directional result. At 5% review, total enrichment is +36.72 points: +37.16 within topics and -0.45 from topic allocation. At 20%, the corresponding values are +25.51, +25.51, and +0.01. All 13 topics remain positive at both budgets. Because the three review sets are nested, these are sensitivity points rather than independent replications; we do not combine their sign tests.

We also force each comparator to review exactly the warning's number of rows within every topic. Warning priority has higher topic-level precision than entropy in 10/11 non-tied topics, learned dual view in 11/13, and restricted S-MMCE in 11/11; Holm-adjusted sign-test values are .035, .045, and .006. It wins 9/12 non-tied topics against holistic-fold JS, which is not significant after adjustment ($p = .146$). Finally, after deleting any one topic and reapplying the fixed 10% global budget, warning priority retains lower risk than each of the six comparators in all 13 omissions. These dependent leave-one-topic-out rows are stability checks, not additional hypothesis tests. All selectors and predictions remain frozen, but this analysis was specified after label access and is therefore descriptive rather than confirmatory.

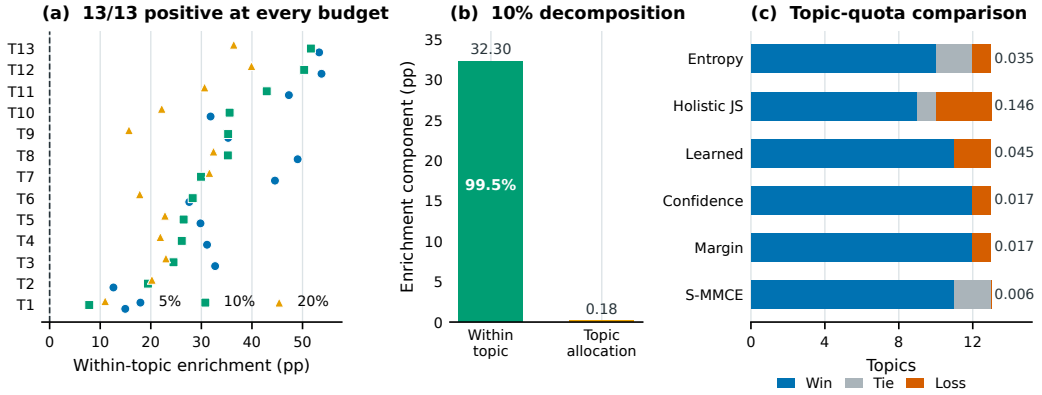


Figure 6: Test-opened TREC topic robustness. Every topic has positive warning enrichment at each locked budget; topics are ordered by their 10% value (a). At 10%, the exact decomposition attributes 99.5% of pooled enrichment to within-topic ranking (b). When every comparator receives the warning's 10% review quota within each topic, warning priority usually has higher precision; values at right are Holm-adjusted two-sided sign tests (c). The nested budgets are not independent replications, and this descriptive analysis does not revise the prospective endpoint.

L.5 Replacement effectiveness and safety

Table 33: Protected TREC replacement results. Both policies improve average accuracy but fail the prespecified 10% harm-among-actions ceiling.

Policy	Actions	C/H/N	Harm (%)	Row Δ	Topic Δ [95% CI]
RA-DVRC	615	439/105/71	17.07	+7.07	+7.02 [5.67,8.21]
Squared-ordinal OAF	371	246/62/63	16.71	+3.90	+3.88 [2.70,5.12]

Both replacement policies correct substantially more rows than they harm, but both exceed the locked 10% harm ceiling. OAF reduces partial-support harms from 51 to eight; it does not remove 33 insufficient and 21 support harms. This is a useful boundary correction, not a successful safety certificate.

L.6 Risk control and architecture sensitivity

Cluster risk control is operationally unhelpful here. For warning priority, the calibrated .20 and .30 targets select 100% review. Other scores need 42.0–83.4% review to meet the observed protected target. This is not a distribution-free failure: nine calibration and 13 test topics do not justify an

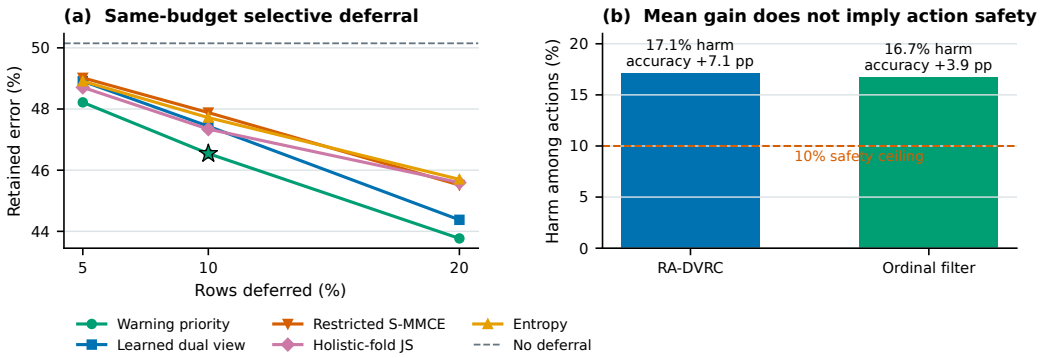


Figure 7: Prospective external TREC confirmation. Warning priority gives the lowest retained-risk point estimate at every locked equal-row budget (a). Both replacement policies improve accuracy but exceed the prespecified 10% harm ceiling (b).

exchangeability claim under domain shift. It does show that formal-looking control can become vacuous even when a fixed-budget ranking is useful.

The transferred RoBERTa sensitivity anchor attains 37.87% accuracy and predicts 4,555 insufficient, 169 support, and zero partial-support rows. WCSD has a slightly better retained-risk point estimate than scalar uncertainty at 10%, but paired topic intervals include zero. Because the anchor has collapsed to two relations and was trained on XRAG rather than a controlled TREC source, we do not count this arm as architecture confirmation.

M Population-Risk Control After Pointwise Certification

M.1 Protected graded SciCiteVal composition

PIAF and Learn-then-Test (LTT) govern different sources of uncertainty. PIAF is pointwise: for a frozen proposer, it preserves an action whose preference cannot reverse inside the declared coupling and loss sets. LTT is population-level: it can release a fixed score threshold when cluster losses satisfy a target risk under exchangeability. Proposition 5 below shows that composing the layers does not erase either statement.

SciCiteVal pairs a citation context with evidence from its cited source and labels it Correct, Incorrect, or Unrelated (Liu et al., 2026). We freeze the ordered map `Unrelated < Incorrect < Correct` to $I < P < S$. This is an ordered citation-validity transfer. It does not assert that every Incorrect citation is semantically equivalent to partial support in WiCE or GenSearch.

Exact normalized cited-source text defines the action cluster. This keeps every context that reuses one released source in one partition. Of 299 clusters, 100 are assigned to score fitting, 100 to risk calibration, and 99 to one-time confirmation. The partitions contain 200, 234, and 600 rows. A label-free rule excludes three confirmation rows whose claims alone exhaust the fixed 512-token input for all registered tokenizers. The analyzed confirmation cohort has 597 rows in 99 clusters.

For each encoder family, the primary score is PIAF's certified lower action-advantage bound. Ten thresholds are derived from fit scores without risk-calibration labels. Fixed-sequence LTT tests them from most conservative to most permissive. Cluster loss is one if at least one selected replacement in the cluster is harmful. The target incidence is .15. The primary .05 error budget is divided equally across DeBERTa-v3-base, RoBERTa-large, and ModernBERT-base. At least ten action clusters are required. Candidate confidence, raw advantage, and a learned harm selector follow the same risk contract as descriptive comparators; the learned selector fails closed when fewer than 20 fit proposals are available.

DeBERTa and RoBERTa are released. ModernBERT has only four eligible PIAF action clusters and fails closed before its policy can be released. Table 34 reports both population incidence and

conditional harm among acted-on clusters. This distinction matters: population incidence can be small when action coverage is small.

Table 34: One-time SciCiteVal source-cluster confirmation for RC-PIAF. Population harm is the fraction of all 99 source clusters receiving at least one harmful replacement. Conditional harm uses selected action clusters only. Upper is the exact one-sided 95% population bound. Accuracy deltas are source-cluster macro percentage points.

Family	Action clusters	C/H/N rows	Population harm	Upper	Conditional harm	Accuracy Δ
DeBERTa-v3-base	10/99 (10.1%)	54/1/8	1.0%	5.5%	10.0%	+1.18
RoBERTa-large	15/99 (15.2%)	62/4/9	4.0%	10.0%	26.7%	+0.36
ModernBERT-base	not released: only 4 eligible action clusters					

Across the two released families, the equal-family cluster-macro accuracy delta is +0.52 points. Its 95% cluster-bootstrap interval is [-1.04, 2.31]. The predeclared joint confirmation gate therefore fails: the familywise risk release and observed harm checks pass, but pooled efficacy is not confirmed. Candidate confidence and raw advantage can also pass their own LTT tests and often cover more clusters. RC-PIAF is not detector SOTA and does not dominate every score-only policy. Its additional assertion is compositional: each released action has a pointwise PIAF certificate, while the selected policy has an outer marginal risk statement under source-cluster exchangeability.

The protocol needed one substantive label-free correction. An initial QASA-or-row grouping split ten repeated cited sources across partitions. The incomplete inference run was stopped before prediction content, outcomes, or partition keys were inspected. Those outputs were abandoned, and exact cited content replaced the grouping rule. A later label-free stage-size check showed that a 15-cluster action floor made two-family confirmation structurally impossible; it was changed once to ten before any key join. The model families, label map, risk target, score families, and confirmation success rule were not changed after outcomes became available.

M.2 Separate binary AVeriTeC confirmation

M.3 Dataset, action unit, and frozen encoders

AVeriTeC contains real-world claims, question–answer evidence units, verdicts, and fact-check articles. We use repository revision `2ca9dee23a2a6fa64c5bd918e0cd28ed0aa09031`. The binary mapping is fixed: `Supported` maps to support; `Refuted`, `Not Enough Evidence`, and `Conflicting Evidence/Cherrypicking` map to non-support. This coarsening avoids inventing a partial-support class that the benchmark does not annotate.

The independent unit is a normalized fact-check article. If an article ID is missing, a normalized claim hash is the fallback. All claims sharing an article stay in one partition. After excluding train claims or article clusters that overlap official development, 1,997 train clusters remain: 899 fit, 399 search, and 699 certificate clusters. Official development contains 500 claims in 371 disjoint article clusters.

Three public NLI checkpoints are pinned and used without fine-tuning: `cross-encoder/nli-deberta-v3-base`, `FacebookAI/roberta-large-mnli`, and `tasksource/ModernBERT-base-nli`. Their revision prefixes are `6c749ce3425c`, `2a8f12d27941`, and `de4ab7e77845`; the protocol artifact records the full 40-character revisions. Each scores the full evidence view and each human question–answer unit. Inputs are split to 512 tokens without silent truncation. Per-family grouped logistic maps convert maximum and mean entailment, contradiction, and neutral probabilities into support posteriors. These maps are cross-fitted by article cluster whenever they feed a development score.

M.4 Cluster-aligned score

For each claim, we derive 15 order-invariant statistics. They include anchor confidence and entropy, dual-view confidence and shift, whether the fused view opposes the anchor, the fraction of opposing architectures, and the mean, minimum, maximum, and standard deviation of candidate confidence

and local–global shift across architectures. For each statistic, the cluster representation records its maximum, mean, standard deviation, and top-two mean. Log one plus cluster size gives 61 inputs. An L1-regularized, class-balanced logistic model with $C = .1$ predicts whether the cluster contains any anchor error. It has 62 nominal parameters and 14 nonzero coefficients. Its target is the exact loss later controlled after whole-cluster deferral. At inference, the frozen AVeriTeC selector uses six fixed NLI score streams: three encoder families times holistic and local evidence views. After those forward passes, the decision layer consists only of the 61-feature cluster aggregation and one 62-parameter logistic score.

The 899 fit clusters use two-level grouped cross-fitting: base posterior maps are fitted away from each scored row, and the selector is fitted away from each scored cluster. Search may choose the model form and is therefore adaptive evidence. On the independent 699-cluster certificate, the first high-coverage threshold defers 82 clusters and retains 88.3% coverage. Retained risk is 16.05%, with exact upper 95% bound 18.68%. It misses the 15% target. We report the failure and do not open official development under that policy.

M.5 Fixed-sequence calibration and proof

Proposition 5 (finite-sample authority after a fixed filter). Let F be a deterministic filter and let $\pi_1, \ldots, \pi_J$ be fixed before risk calibration. Under independent exchangeable clusters, fixed-sequence LTT releases a policy whose retained risk exceeds α with probability at most δ. If $F =$PIAF, each released replacement also retains the pointwise coupling certificate from Proposition 4.

Proof. Once F and $\pi_1, \ldots, \pi_J$ are fixed without the risk-calibration sample, their composition is simply a pre-fixed sequence of cluster policies. For policy j, let K_j be the observed number of retained losses among N_j retained calibration clusters. Under exchangeability, the lower-tail binomial test of $H_j : R_j \geq \alpha$ is valid at the boundary $R_j = \alpha$. In a fixed sequence, let j^* be the first true null. Releasing any policy with risk above α requires rejecting H_{j^*}, so the probability of any such release is at most δ. A deterministic PIAF certificate is unaffected by this sampling event: if PIAF is chosen as F, every action released by the composite policy remains certified over its declared coupling and loss sets. This proves coexistence of the two statements, not a cross-domain guarantee; the risk statement still requires independent exchangeable target clusters.

After the failed certificate, we keep the score and the 15% risk target but allow less automatic authority. Five thresholds are fixed from the already opened development score distribution. They correspond approximately to 35%, 30%, 25%, 20%, and 15% review and are tested in that order. Official development is split without scores or labels: sort clusters by `sha256(350814, cluster_id)`, assign the first 160 to risk calibration, and reserve the remaining 211 for confirmation. All selector scores are saved before either key is joined.

Here $\alpha = .15$ and $\delta = .05$. Conditional on the retained count, the exact lower-tail binomial p-value is evaluated at risk .15. The guarantee requires the frozen score and threshold family plus independent exchangeable clusters; it does not follow from conditioning on the realized score values.

The score and every numerical threshold were fixed on disjoint earlier data, before the 160 risk-calibration clusters were scored or labeled. The first threshold retains 105 of 160 calibration clusters and leaves 8 errors. Its risk is 7.62%, exact upper bound 13.33%, and exact p-value .0176. The second threshold has p=.0651, so testing stops; three thresholds remain untested. We lock the first threshold and six equal-count masks before opening confirmation labels.

M.6 Independent result and comparator scope

On 211 confirmation clusters, the anchor has 35 error clusters. The selected policy defers 68, captures 30 errors, and leaves 5 errors among 143 automatic clusters. Coverage is 67.8%, retained risk is 3.50%, and its exact one-sided upper 95% bound is 7.21%. The selected policy therefore satisfies all frozen checks.

At the same 68-cluster budget, anchor uncertainty and restricted S-MMCE each leave 7 errors; local–global gap leaves 12; architecture opposition leaves 20; and holistic architecture JS leaves 21. Paired cluster bootstrap intervals show lower selected-policy risk than gap, opposition, and JS. The 1.40-point differences from uncertainty and restricted S-MMCE are not significant. Exact paired

error capture tests lead to the same conclusion. The result establishes calibrated action authority and useful structured signal. It does not establish that the selected policy universally dominates simpler selectors.

Table 35: Independent AVeriTeC confirmation at matched 68-cluster review (percent). Upper is the exact one-sided 95% bound.

Selector	Precision	Capture	Risk	Upper	Reduction
Cluster score + LTT	**44.1**	**85.7**	**3.50**	**7.21**	**13.09**
Uncertainty	41.2	80.0	4.90	9.00	11.69
Restricted S-MMCE	41.2	80.0	4.90	9.00	11.69
Local–global gap	33.8	65.7	8.39	13.24	8.20
Architecture opposition	22.1	42.9	13.99	19.67	2.60
Architecture JS	20.6	40.0	14.69	20.45	1.90

N BEGIN Multi-Source Robustness Sensitivity

We also test the cluster-aligned design on BEGIN, which labels CMU-DoG, TopicalChat, and Wizard of Wikipedia responses as fully attributable, not fully attributable, or generic (Dziri et al., 2022). This adjacent task tests the decision layer under a different relation inventory.

Frozen DeBERTa-v3-base, RoBERTa-large, and ModernBERT-base NLI models score knowledge, response-unit, and preceding-turn views. A source-balanced ridge model estimates the local-feature residual after a same-capacity anchor-only error ranker; weight zero is a mandatory candidate. Development selects ridge alpha 100 and weight 1.0. Intersection–union LTT requires every source-specific retained risk to fall below 20% at familywise level .05.

Integrity amendment. Immediately after producing the first official-test summary, a label-free identity audit found that official development and test share 284 normalized source–knowledge–preceding-turn clusters. These cover 368 test rows, mostly from Wizard of Wikipedia. We preserve the original run as an audit record, remove every overlapping context without consulting its label, and rerun the unchanged thresholds and matched-budget analysis. Because the initial confirmation labels had already been opened, the corrected result below is a *post-hoc sensitivity*, not an independent confirmation.

The clean calibration set retains 3,288 clusters and selects the same threshold, 0.688539. On 6,108 clean sensitivity clusters, all methods review the same 1,170 clusters. The primary policy keeps 80.84% automatic, lowers any-error risk from 18.71% to 12.11%, and has an exact one-sided upper bound of 12.90%. The CMU/TopicalChat/WoW upper bounds are 18.70%/8.95%/12.62%, each below the locked 20% target. Table 36 shows the matched-budget result.

Table 36: Post-hoc overlap-clean BEGIN sensitivity at matched 19.16% cluster review (percent). Upper is exact one-sided 95%.

Selector	Precision	Retained risk	Upper	Reduction
Anchored local residual	**46.58**	**12.11**	**12.90**	**6.60**
Anchor-only learned	**46.58**	**12.11**	**12.90**	**6.60**
Unconstrained dual-view	43.76	12.78	13.59	5.93
Restricted S-MMCE	37.69	14.22	15.06	4.50
Anchor uncertainty	37.18	14.34	15.18	4.38
Local–global gap	27.01	16.75	17.65	1.97
Architecture JS	26.07	16.97	17.87	1.74

The primary-minus-S-MMCE retained-risk difference is -2.11 points with a 95% source-stratified paired-bootstrap interval of [-2.79,-1.42]. The primary and anchor-only methods tie in aggregate, although their masks differ. Thus this large sensitivity supports multi-family group-risk control and the value of a learned selector over scalar or disagreement heuristics. It does not show that

local structure universally improves on an equally trained anchor-only selector. It also shows why context-level overlap checks belong in the data firewall even when a benchmark provides official splits.

O Protected CiteBench Review-Target Boundary Test

CiteBench provides three citation-quality annotations per cited statement and fixed metric splits (Xu et al., 2025b). We use 303 development statements from 134 queries and protect 1,462 test statements from 283 queries. Before test labels, we fix majority edit needed as primary, two citation-error secondaries, all predictions and masks, and every risk-control threshold.

At equal 5–20% budgets we compare warning priority with query-grouped dual-view logistic, natural-group DRO S-MMCE, holistic-fold disagreement, and three scalar controls. Fixed-sequence LTT tests .10–.20 query-normalized risk with a 60% retention floor; bootstrap resampling keeps each query intact.

Table 37: Protected CiteBench results at 10% review for majority edit needed. Q-risk is query-normalized unreviewed error mass. Values are percentages.

Selector	Precision	Error capture	Q-risk
Warning priority	41.78	13.68	30.09
Dual-view logistic	**49.32**	**16.14**	28.28
Natural-group DRO S-MMCE	46.58	15.25	**28.19**
Holistic-fold JS	45.21	14.80	29.46
Low confidence	45.89	15.02	29.35
Entropy	49.32	16.14	28.88
Margin	46.58	15.25	29.30

The primary confirmation fails. Warning-minus-logistic Q-risk is +1.80 points [0.02, 3.71], and warning-minus-S-MMCE is +1.89 [0.05, 3.88]; scalar-control and holistic-JS differences include zero. Yet the warning is enriched: primary precision is 41.8% against 30.5% prevalence, while material-error precision is 17.1% against 5.8%. Diagnostic validity does not imply comparative validity for this broader review target.

Formal control is operationally vacuous. Dual-view logistic has the best retention, 53.8% at the .20 target with 17.2% Hoeffding–Bentkus upper bound, but misses the 60% floor. Warning attains 4.1% only at 9.0% retention. We preserve the negative result and do not retune CiteBench.

P Protected HAGRID Whole-Cluster Deferral

AttributionBench standardizes several attribution datasets as claim–evidence binary classification (Li et al., 2024b). We use only its HAGRID source. BEGIN is excluded because its labels were opened by a later experiment. The original v31 HAGRID partitions were never opened: fit has 1,044 rows in 322 question clusters, risk calibration has 508/145, and test has 494/152. No cluster crosses a partition.

The global and local three-way posteriors are projected to binary support versus non-support. WCSD ranks any binary boundary disagreement first, then the absolute support-probability gap and binary Jensen–Shannon divergence. Learned selectors aggregate fixed row features by cluster mean, maximum, and minimum. The dual-view logistic has 68 parameters; its anchor-only counterpart has 20. Restricted S-MMCE uses the same frozen dual-view features and 1,105 parameters. Scalar controls are global uncertainty, holistic-fold disagreement, and direct global–local JS. Every policy reviews complete question clusters.

At 10%, WCSD selects 15 clusters and captures 11/42 baseline error clusters. Its retained-risk advantage over global uncertainty is 0.73 points with paired cluster-bootstrap interval [-3.90,5.26]; its precision advantage is 6.7 points [-24.5,37.8]. Dual-view learned improves on anchor-only by 0.73 points [-1.70,3.28]. These intervals cross zero. The protected result supports structured warning enrichment and the 20% operating trend, but it does not establish selector dominance.

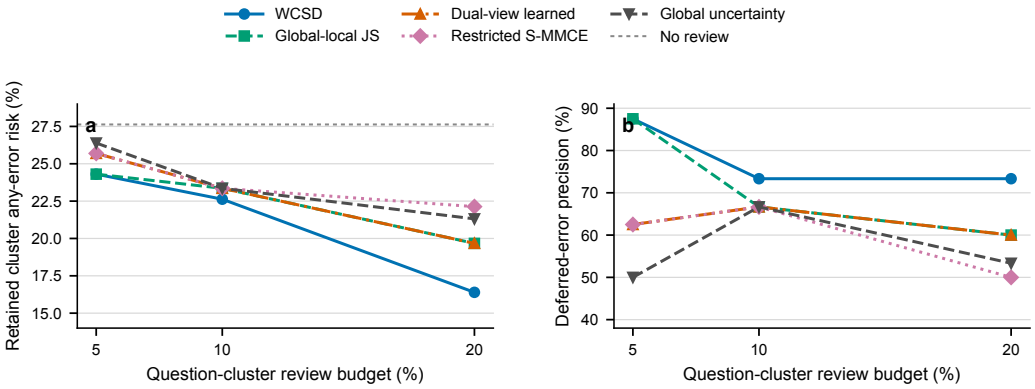


Figure 8: Protected HAGRID whole-question deferral. All methods receive the same prespecified cluster budgets. Curves show point estimates; the primary 10% paired intervals include zero.

Table 38: Protected HAGRID fixed-budget results. P is deferred-error precision and R is retained cluster any-error risk; values are percentages.

	5% review		10% review		20% review	
Selector	P	R	P	R	P	R
WCSD	**87.5**	**24.31**	**73.3**	**22.63**	**73.3**	**16.39**
Global–local JS	**87.5**	**24.31**	66.7	23.36	60.0	19.67
Dual-view learned	62.5	25.69	66.7	23.36	60.0	19.67
Anchor-only learned	62.5	25.69	60.0	24.09	46.7	22.95
Restricted S-MMCE	62.5	25.69	66.7	23.36	50.0	22.13
Global uncertainty	50.0	26.39	66.7	23.36	53.3	21.31
Holistic-fold JS	75.0	25.00	60.0	24.09	46.7	22.95

For formal action-risk control, calibration chooses the most permissive threshold whose finite-sample corrected expected residual cluster loss meets each target. On test, WCSD meets the .15/.20/.25/.30 targets at 68.4/84.9/88.8/91.4% automatic coverage, with observed residual risks .112/.171/.204/.211. This is conformal control of the probability that an incoming question retains any unreviewed anchor error. It is not a guarantee on conditional risk among retained clusters, and it requires exchangeable calibration and deployment clusters. All seven methods and four targets are in the released result artifact.

Q PROTECTED EXPERTQA AGGREGATE–BOUNDARY STRESS TEST

ExpertQA contains expert-authored questions and expert judgments of whether evidence completely, partly, incompletely, or does not support each answer claim (Malaviya et al., 2024). We use the official anonymized release at a fixed source revision. Ten source lines viewed during download-integrity debugging are excluded before any split. After deterministic text and label eligibility checks, 9,647 claim–evidence rows from 1,894 questions remain. Question clusters are field-stratified into 4,866 fit rows (941 questions), 1,893 risk-calibration rows (372), and 2,888 protected test rows (581). No question crosses partitions.

The replacement experiment is target-label free. For each of DeBERTa-v3-base, RoBERTa-large, and ModernBERT-base, the global and local models are the three WiCE title-grouped fold pairs used by the frozen v42 architecture study. Their composer, promotion and demotion thresholds, and PIAF certificate are replayed without ExpertQA fitting. The deferral study is separate: learned selectors fit on the ExpertQA `fit` partition, and cluster-risk thresholds calibrate on `risk-cal`. All three test prediction files, selector artifacts, and risk policies were complete before the test key was opened once. The key was then closed again; no post-test threshold or model revision is reported as v51.

Native `Missing` claims have no textual evidence, whereas support and partial claims do. Overall three-way accuracy can therefore reward a coarse evidence-presence cue. The primary scientific contrast reports that aggregate alongside the 2,215 evidence-present support/partial rows from 455 test questions. This subset is not selected by model outcome or confidence.

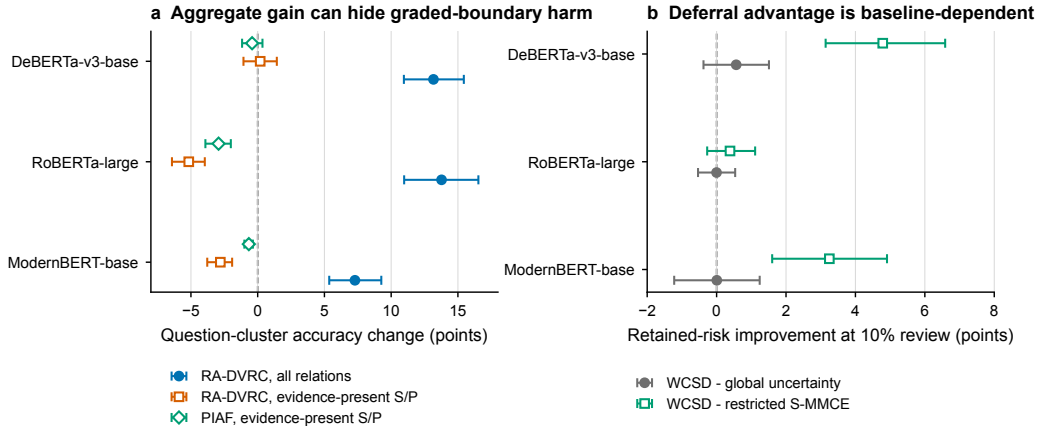


Figure 9: Protected ExpertQA stress test. Intervals resample complete question clusters. All-relation RA-DVRC accuracy rises for every backbone, while the evidence-present support/partial boundary is neutral for DeBERTa and significantly worse for RoBERTa and ModernBERT (a). At 10% whole-question review, WCSD ties global uncertainty on all three backbones and beats restricted S-MMCE on two (b). Positive values favor the first named method.

Q.1 AGGREGATE IMPROVEMENT IS NOT GRADED-BOUNDARY IMPROVEMENT

RA-DVRC's overall cluster-accuracy intervals are strictly positive for all three backbones. The changes are driven mainly by the insufficient class: insufficient recall rises 43.09, 53.64, and 28.23 points for DeBERTa, RoBERTa, and ModernBERT. Partial-support recall instead falls 8.76, 7.55,

Table 39: Protected ExpertQA architecture stress test. Overall deltas use all three relations; S/P deltas restrict evaluation to the 2,215 evidence-present support and partial-support rows in 455 question clusters. Values are question-cluster accuracy points with paired 95% bootstrap intervals. C/H denotes RA-DVRC corrections and harms over all 2,888 rows.

Backbone	Anchor acc.	RA overall Δ	RA S/P Δ	PIAF S/P Δ	RA C/H
DeBERTa-v3-base	55.76	+13.17 [+10.95, +15.44]	+0.19 [-1.06, +1.43]	-0.42 [-1.17, +0.36]	367/79
RoBERTa-large	48.61	+13.77 [+10.97, +16.53]	-5.16 [-6.42, -3.96]	-2.93 [-3.92, -2.01]	373/134
ModernBERT-base	61.25	+7.29 [+5.36, +9.26]	-2.80 [-3.78, -1.91]	-0.66 [-1.00, -0.36]	204/98

and 4.98 points. Restricting evaluation to evidence-present support and partial rows removes the coarse missing-evidence contrast. RA-DVRC then corrects/harms 77/79 DeBERTa rows, 11/133 RoBERTa rows, and 12/96 ModernBERT rows. PIAF reduces the number of authorized changes, but its boundary effects are -0.42 [-1.17,0.36], -2.93 [-3.92,-2.01], and -0.66 [-1.00,-0.36] points. It contains some harms; it does not portably solve severity.

The calibration decomposition matches the action result. On all rows, composition improves holistic NLL from .962 to .808, 1.315 to 1.089, and 1.039 to .976. On the support/partial subset, DeBERTa still improves slightly (.947 to .926), but RoBERTa worsens from 1.066 to 1.171 and ModernBERT from 1.058 to 1.084. Aggregate calibration is therefore insufficient evidence for a class-conditional action boundary.

Q.2 Deferral and risk control remain backbone-dependent

At the locked 10% question-cluster budget, WCSD reduces retained cluster any-error risk by 1.95/1.41/1.74 points for DeBERTa/RoBERTa/ModernBERT. Its deferred-error precision is 98.3/100.0/91.4%, but baseline cluster error prevalence is already 80.7/87.3/75.7%. Paired intervals against global uncertainty cross zero for every backbone. WCSD beats restricted S-MMCE for DeBERTa and ModernBERT, but not RoBERTa; dual-view learned selection never beats its same-capacity anchor-only counterpart. These comparisons prevent the high raw precision from being mistaken for structured-signal dominance.

Observed cluster-risk target attainment is also mixed. WCSD meets 0/4 targets for DeBERTa, 3/4 for RoBERTa, and 0/4 for ModernBERT. No method meets every target on any backbone, and every method misses every target on ModernBERT. The calibration artifact retains its finite-sample interpretation only under the recorded exchangeability conditions; the protected result shows that portable target attainment cannot be inferred from method name or source performance.

ExpertQA thus strengthens the paper through a protected negative decomposition. It addresses architecture breadth without claiming universal replacement or deferral superiority. More importantly, it exposes a metric failure: a verifier can look substantially better in aggregate while becoming less reliable at the exact graded boundary that controls intervention harm.

R Protected TREC Assessor-Visible Architecture Sensitivity

This final study was frozen after the prediction-blind TREC, GenSearch, ExpertQA, and XRAG outcomes were known. It asks a narrow question: do the already frozen action policies remain useful across encoder families on a new, topic-disjoint graded cohort? The source is the official TREC RAG 2024 `with_prediction` condition. It contains 4,162 eligible sentence–citation pairs from 14 topics disjoint from both earlier TREC cohorts. Assessors could see a model recommendation. We knew this collection condition before freezing, so the cohort is a protected bias-sensitive replication, not a prediction-blind gold standard.

We first materialized a label-free file and a separate key. Eighteen inference arms cover three folds, two views, and three WiCE-trained encoder families: DeBERTa-v3-base, RoBERTa-large, and ModernBERT-base. Every arm, composed prediction, PIAF certificate, selector score, and 5/10/20% row mask was complete before the key was opened once. No TREC label fits a checkpoint, composer,

calibrator, selector, or cutoff. All 4,162 rows, all three backbones, every registered selector, and every discordant action are retained in the result.

The replacement rule required a positive PIAF topic-macro point effect in at least two families, a positive lower bound for the equal-family shared-topic bootstrap, and no family with more harms than corrections. All conditions pass. The equal-family effect is +4.05 points with a 95% interval [2.89, 5.27]. PIAF keeps 444/631, 383/532, and 86/356 RA-DVRC actions, respectively.

Table 40: Protected TREC assessor-visible replacement sensitivity. Accuracy and deltas are topic-macro percentage points.

Family	Anchor	PIAF Δ [95% CI]	C / H	Harm (%)	Δ recall$_P$
DeBERTa-v3-base	54.12	+5.26 [3.61, 6.92]	312 / 89	20.05	-5.95
RoBERTa-large	51.33	+5.33 [3.66, 7.01]	278 / 59	15.40	-4.09
ModernBERT-base	51.34	+1.55 [0.94, 2.24]	71 / 6	6.98	0.00

The boundary decomposition matters. DeBERTa and RoBERTa trade partial-support recall for large insufficient-recall gains; 73/89 and 44/59 harms have gold partial support. ModernBERT makes no changed prediction on a gold-partial row, but its anchor already has only 10.87% partial recall. Thus the positive replacement result establishes cross-family utility, not resolution of the graded severity boundary.

At the registered 10% row budget, WCSD must beat the best scalar uncertainty control in at least two families. It does so in all three. Paired intervals resample complete topics and compare exactly matched masks. The final column is the scalar control's retained risk minus WCSD's, so positive values favor WCSD.

Table 41: Protected TREC retained risk at the registered 10% row budget. The final column is best scalar risk minus WCSD risk, in percentage points.

Family	WCSD	Best scalar	Learned	S-MMCE	Advantage [95% CI]
DeBERTa-v3-base	42.42	44.07	43.30	44.02	+1.66 [0.82, 2.51]
RoBERTa-large	44.69	46.80	44.63	46.13	+2.11 [1.46, 2.81]
ModernBERT-base	45.41	46.80	44.55	46.77	+1.39 [0.64, 2.08]

WCSD also beats restricted S-MMCE in all three point comparisons. It does not dominate learned deferral: WCSD is better for DeBERTa, statistically tied for RoBERTa, and worse for ModernBERT. This is the intended control. Warning priority isolates structured local–global disagreement from scalar uncertainty, while the transferred learned selector shows what source labels can buy when its ranking ports.

This experiment closes two breadth gaps without erasing the negative results. PIAF now has a label-sealed positive cross-family replication, and WCSD beats scalar uncertainty under the same protected budgets. ExpertQA still shows that positive aggregate accuracy can conceal a worse support/partial boundary. XRAG still shows that learned deferral and target-risk attainment are architecture- and shift-dependent. The combined evidence supports an action-conditioned claim rather than universal portability.

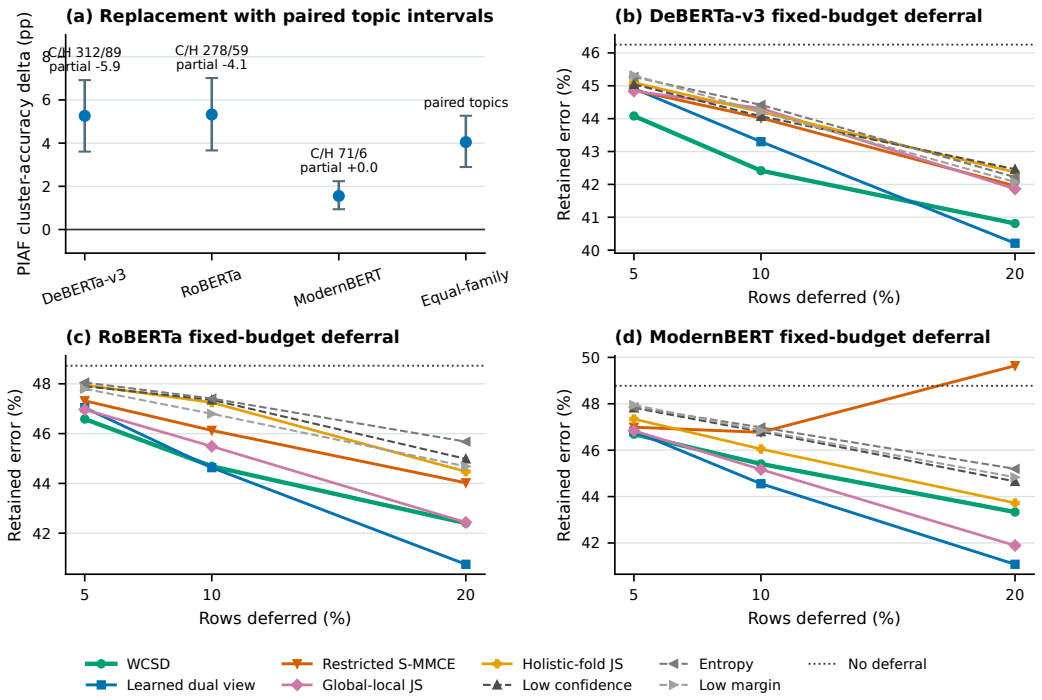


Figure 10: Protected TREC assessor-visible architecture sensitivity. Panel (a) reports PIAF topic-macro effects; the equal-family interval jointly resamples the same 14 topic IDs across backbones. Panels (b–d) show fixed-row-budget retained risk for every pre-registered selector. The dotted line is the no-deferral anchor risk. These curves are complete sensitivity results, but the assessment condition prevents interpreting this cohort as prediction-blind ground truth.

S POST-HOC CONTROLLED HOSTED-JUDGE AUTHORITY AUDIT

This reviewer-motivated audit asks whether stronger black-box judges remove the need for an action layer. It is post-hoc: WiCE labels were already open from the primary experiment. Before comparing any v54 prediction with those labels, we fixed one zero-shot prompt, two model aliases, all policies, all metrics, and a 358-row label-free batch. Every authority sees the exact claim and evidence used by the frozen WiCE systems. The unit of inference remains the title cluster.

The relay reported `gpt-5.6-sol` and `gemini-3-flash-preview`. Both were called through the same third-party relay, so this is a controlled two-model-family comparison, not an official-provider replication. GPT used low reasoning effort and Gemini used minimal effort. A content-independent format rule retried 14 Gemini rows that either did not end with `stop` or did not contain parseable canonical JSON. The retry rule was fixed before label comparison, inspected neither predicted relations nor gold labels, and allowed one recovery round. Both final files contain 358 unique, normally stopped, parseable responses.

Table 42: Post-hoc controlled hosted-judge authority audit on the same 358 WiCE test rows and 326 title clusters. Accuracy and deltas are title-cluster macro percentages; intervals use 10,000 paired title-cluster bootstraps. C/H/N denotes corrections, harms, and neutral changes relative to the holistic anchor. Harm is the percentage of changed rows that were correct under the anchor and wrong after the change. Every policy fixed before label comparison is shown.

Policy	Acc.	Δ [95% CI]	Changes	C/H/N	Harm	Recall I/P/S
Anchor	45.97	0.00 [0.00, 0.00]	0	0/0/0	–	71.9/21.9/89.2
RA-DVRC	54.48	+8.48 [5.52, 11.73]	33	30/1/2	3.0	78.1/32.1/93.7
PIAF	51.03	+5.06 [2.76, 7.52]	19	17/0/2	0.0	71.9/28.8/91.0
GPT direct	71.98	+25.97 [17.81, 33.99]	232	156/68/8	29.3	65.6/89.3/39.6
Gemini direct	75.62	+29.66 [22.55, 36.63]	201	143/45/13	22.4	68.8/84.2/57.7
GPT ∩ Gemini direct	74.24	+28.28 [21.16, 35.25]	180	133/40/7	22.2	62.5/82.3/58.6
GPT-confirmed RA-DVRC	52.95	+6.98 [4.22, 9.97]	26	25/1/0	3.8	75.0/31.6/91.0
Gemini-confirmed RA-DVRC	52.87	+6.88 [4.14, 9.82]	25	24/1/0	4.0	75.0/30.7/91.9
GPT-confirmed PIAF	50.42	+4.44 [2.45, 6.75]	15	15/0/0	0.0	71.9/28.4/90.1
Gemini-confirmed PIAF	49.81	+3.84 [1.84, 5.98]	13	13/0/0	0.0	71.9/27.9/89.2
Both-confirmed PIAF	49.50	+3.51 [1.69, 5.52]	12	12/0/0	0.0	71.9/27.4/89.2

The hosted judges are much stronger global classifiers than the anchor. GPT and Gemini gain 25.97 and 29.66 cluster-accuracy points. Their authority is also much broader: they alter 232 and 201 of 358 rows. This breadth carries a different cost. GPT harms 68 previously correct decisions and Gemini harms 45, for action-harm rates of 29.3% (Wilson 95% interval [23.8, 35.5]) and 22.4% ([17.2, 28.6]). Both over-demote support: support recall falls from 89.2% to 39.6% and 57.7%. Unanimous direct authority still has 40 harms among 180 changes, or 22.2% ([16.8, 28.8]). Agreement is not a safety certificate.

Confirmation does not rescue the trade-off. GPT- and Gemini-confirmed RA-DVRC retain 25 and 24 corrections, but both retain RA-DVRC's single harm. They give up five or six changes without improving observed safety. PIAF has 17 corrections, zero harms, and two neutral moves. Judge-confirmed PIAF policies retain only 13–15 corrections at the same zero observed harms; requiring both judges retains 12. Relative to PIAF, these confirmation rules lose 0.61–1.54 cluster-accuracy points. Thus the hosted judges improve global classification, whereas PIAF supplies a narrower low-harm intervention. Neither result subsumes the other.

The correction/harm counts make the cost dependence concrete. If one correction is worth one unit and one harm costs λ units, the descriptive utility is $C - \lambda H$. PIAF overtakes direct GPT once $\lambda > 2.04$ and direct Gemini once $\lambda > 2.80$. These crossovers are opened-result descriptions, not fitted costs or prospective guarantees. Recorded API usage was 369,820 total tokens for GPT and 665,563 for Gemini, including 306,602 Gemini reasoning tokens. The relay provenance, post-hoc status, and English-only scope limit the comparison. Its role is narrower but useful: a high-accuracy judge does not automatically satisfy the loss attached to an automatic action.

T PROSPECTIVE SUFFICIENTFACTS EVIDENCE-OMISSION CONFIRMATION

SufficientFacts holds a claim fixed and removes fluent constituents or full sentences from its evidence, then records whether enough evidence remains for fact checking (Atanasova et al., 2022). We pin the official repository at commit `1fbc36364f5639d2942997849585 99fda0978472`. The primary cohort includes only rows whose original relation is `SUPPORTS`; contradiction is outside the paper's $I/P/S$ space. Original endpoints map to S. Official reduced labels `ENOUGH` and `NOT ENOUGH` map to S and I. Exact normalized claim text is the cluster, so all omission variants of one claim stay together.

The protocol was written before row download. The fixed files contained 1,200 FEVER and 1,200 HoVer rows rather than the 1,000/1,000 counts stated on the data card. Preparation stopped before writing a cohort. We amended only those two expected file counts, then materialized 1,696 pairs, 1,693 claim clusters, and 3,392 endpoints: 698 FEVER, 600 HoVer, and 398 VitaminC pairs. Models, mappings, actions, metrics, and budgets did not change. All 18 fold predictions, three backbone action files, matched-count rankings, and 953 PIAF certificates were complete before the separate key was opened.

Table 43: Prospective SufficientFacts evidence-omission confirmation. Endpoint and pair deltas are percentage points relative to each same-family anchor; intervals resample exact-claim clusters. C/H denotes corrections/harms.

Family	Anchor endpoint/pair	RA endpoint Δ [95% CI]	RA pair Δ [95% CI]	RA C/H	PIAF endpoint Δ [95% CI]	PIAF pair Δ [95% CI]	PIAF C/H
DeBERTa-v3	65.36/33.79	+5.49 [4.55,6.44]	+9.79 [8.09,11.52]	230/44	+3.22 [2.51,3.96]	+5.84 [4.49,7.21]	137/28
RoBERTa	60.94/24.47	+8.20 [7.27,9.14]	+15.51 [13.73,17.34]	289/11	+6.02 [5.23,6.85]	+11.38 [9.86,12.99]	212/8
ModernBERT	58.90/21.46	+4.34 [3.63,5.05]	+6.96 [5.79,8.15]	151/4	+0.65 [0.38,0.95]	+1.18 [0.71,1.71]	23/1

The composer result is the cleanest cross-family finding: all six endpoint and pair intervals are positive. PIAF retains positive effects but reduces authority sharply, especially for ModernBERT. It does not uniformly improve the conditional harm rate. Relative to RA-DVRC, PIAF's harm-rate difference is +3.88 points ([0.82, 7.26]) for DeBERTa and has an interval containing zero for RoBERTa and ModernBERT. At each PIAF action count, candidate confidence, raw advantage, and generic fold disagreement are close in aggregate accuracy. The certificate therefore controls the declared coupling/loss sensitivity; it is not an empirical claim of universal ranking or harm-rate dominance.

Source stratification locates the largest failure. DeBERTa PIAF makes 30 corrections and 24 harms on HoVer, compared with 83/3 on FEVER and 24/1 on VitaminC. Of its 28 total harms, 25 occur under full-sentence removal. We report these strata because the protocol requires every source and removal type, not because they revise the frozen policy.

Deferral is also architecture- and budget-dependent. RoBERTa WCSD beats matched low confidence in retained risk and precision at every 5/10/20/30% review budget, with all paired intervals positive. ModernBERT does so at 5% and 10% but ties at larger budgets. DeBERTa improves precision at 20%, loses to entropy at 30%, and otherwise has intervals spanning zero. This new confirmation thus answers the single-family concern without asserting universal WCSD dominance.

U PROTECTED BINGCHECK NATIVE-PARTIAL CONFIRMATION

BingCheck contains fact-checking records for LLM-generated responses, including human-reviewed evidence and verdicts (Li et al., 2024a). We pin the public repository at commit `d33157db37c4475ae1b9a1df4907e7368b35e995`. Native `supported`, `partially supported`, and `not supported` labels map to S, P, and I. We exclude `refuted` instead of conflating contradiction with missing support. Every claim from one original response remains in the same cluster.

Before inference, we excluded the paper's worked response and its separately quoted Brain-virus example. Public-schema inspection revealed one claim string that differed from the shared evidence/verdict dictionary key. The two dictionaries had identical order in all 396 response files, so we used their shared human-reviewed key. This rule was fixed before predictions or key access. The

final blind cohort contains 3,791 claims in 395 response clusters: 2,676 S, 899 P, and 216 I after the key is opened. The blind file contains no label field.

All 18 fold outputs, three composed policies, 1,032 PIAF certificates, three matched-authority rankings, and four deferral budgets were complete before opening the separate key. A family passes only if PIAF improves cluster-macro accuracy, has more corrections than harms, keeps action harm at or below 15%, preserves partial recall within two points, and removes exact harms to correct gold-P anchors relative to RA-DVRC. Confirmation requires two passing families and a positive pooled interval. Every endpoint is reported regardless of sign.

Table 44: Protected BingCheck replacement result. Deltas and intervals are response-cluster macro percentage points relative to each same-family anchor. C/H/N denotes corrections, harms, and neutral changes. The final column reports exact harms to previously correct Partial anchors before and after PIAF.

Family	Anchor	PIAF Δ [95% CI]	C/H/N	Harm (%)	ΔR_P	Correct-P anchor harm RA→PIAF	Gate
DeBERTa-v3	67.55	+0.80 [0.06,1.56]	102/74/37	34.74	-7.23	151→71	fail
RoBERTa	70.86	+0.47 [-0.32,1.27]	103/90/49	37.19	-8.57	94→78	fail
ModernBERT	73.23	+0.21 [-0.03,0.46]	14/5/6	20.00	-0.33	46→3	fail

The equal-family pooled effect is +0.50 points with a 95% interval [0.05, 0.95]. All three point effects are positive, and every family makes more corrections than harms. Nevertheless, 0/3 families pass the full rule. DeBERTa and RoBERTa violate both action safety and partial non-inferiority. ModernBERT violates action safety. The protected conclusion is therefore failure, not weak success.

After that decision was fixed, we stratified by a model-facing property that was already present in the blind file. The decomposition does not alter any policy or endpoint.

Table 45: Opened-result evidence-presence decomposition. Effects are PIAF-minus-anchor response-cluster macro accuracy in percentage points. This post-hoc table explains the registered result; it does not revise it.

Family	Empty rows Δ	Present rows Δ	Empty C / all C	Present C/H
DeBERTa-v3	+46.04	-1.55	81/102	21/74
RoBERTa	+48.14	-2.05	85/103	18/90
ModernBERT	+8.04	-0.13	13/14	1/5

The aggregate gain comes from 178 empty-evidence rows. On the 3,613 evidence-present rows, every family has a nonpositive PIAF effect. Most retained harms are demotions of genuinely partial support. The registered harm and partial-recall conditions therefore catch a real metric reversal: better aggregate accuracy does not justify the intended severity action.

The lower-authority result is narrower. At the registered 5% budget, WCSD's deferral precision is 69.84%, 70.90%, and 73.54% for DeBERTa, RoBERTa, and ModernBERT. Entropy obtains 57.67%, 62.96%, and 57.67%; paired precision advantages have intervals [3.78,20.83], [0.15,15.98], and [7.79,24.17]. WCSD has the highest point estimate among all registered controls at this budget. The advantage does not persist uniformly at 10%–20%. BingCheck thus supports low-budget review enrichment, not a universal selector or portable replacement policy.

V Protected SciTail Cross-Task Partial-Support Confirmation

SciTail constructs entailment hypotheses from science questions and answer choices, and premises from retrieved web sentences (Khot et al., 2018). The published benchmark collapses the task to binary entailment, but the official raw archive retains five-worker Support, Partial, and None vote counts. We use those native votes rather than the released binary label.

The protocol is fixed after inspecting the paper, repository documentation, archive paths, checksum, TSV header, and aggregate line count, but before any annotation row or vote value. It keeps every raw row whose normalized question–answer hypothesis occurs in the official test split and has a unique

winner with at least 80% of at least five votes. Paper and README examples are excluded. The result is 2,593 pairs in 171 official test-question clusters: 853 S, 454 P, and 1,286 I. All 18 frozen outputs, three action files, matched-authority rankings, four review budgets, and 1,585 certificates are complete before opening the separate key. No SciTail label fits any model, composer, calibrator, threshold, ranking, or certificate.

A family passes only if PIAF improves question-cluster macro accuracy, makes more corrections than harms, keeps action harm at or below 15%, preserves partial recall within two points, and removes gold-P harms relative to unfiltered RA-DVRC. Confirmation requires two passing families and a positive equal-family cluster-bootstrap interval.

Table 46: Protected SciTail replacement result. Deltas and intervals are question-cluster macro percentage points relative to each same-family anchor. C/H/N denotes corrections, harms, and neutral changes.

Family	Anchor	PIAF Δ [95% CI]	C/H/N	Harm (%)	ΔR_P	Gold-P harm RA→PIAF	Gate
DeBERTa-v3	55.14	+1.90 [0.63,3.15]	90/31/25	21.23	-3.08	102→14	fail
RoBERTa	43.83	+14.53 [11.33,17.82]	451/39/25	7.57	-7.05	46→33	fail
ModernBERT	53.86	+0.50 [0.22,0.85]	16/1/0	5.88	-0.22	24→1	pass

The equal-family PIAF effect is +5.64 points with a 95% interval [4.36, 6.94]. Every family improves aggregate accuracy and makes more corrections than harms. PIAF also removes many partial-support harms relative to the unfiltered composer. The registered confirmation still fails. DeBERTa exceeds the action harm ceiling. DeBERTa and RoBERTa exceed the partial-recall non-inferiority margin. Only ModernBERT passes all five checks, below the required two families.

The registered gate above intentionally remains unchanged, but it used point estimates rather than a confidence-bound safety criterion. ModernBERT makes only 17 actions with one harm. Its exact one-sided 95% binomial upper bound is 25.0%, above the declared 15% target. A post-hoc uncertainty-aware audit therefore finds that no family establishes the declared boundary-safety target. This is a power and gate-design limitation, not a change to the protected outcome. Future action gates require both a minimum number of independent action clusters and the larger of action-level exact and cluster-bootstrap harm upper bounds.

The same warning is more useful at lower authority. At the locked 15% row budget, WCSD has the best deferral-precision and retained-risk point estimate among all registered scalar-uncertainty and generic-disagreement controls in all three families.

Table 47: Protected SciTail deferral at the locked 15% row budget. Best control is selected by retained-risk point estimate from the four registered controls.

Family	Reviewed	WCSD precision	WCSD risk	Control precision	Control risk
DeBERTa-v3	388	77.84	37.01	72.68	37.91
RoBERTa	388	95.88	48.16	71.39	52.47
ModernBERT	388	91.24	36.69	59.28	42.31

For RoBERTa and ModernBERT, paired question-cluster intervals favor WCSD over every registered control for both precision and retained risk. DeBERTa has the best point estimates, but comparisons with generic fold disagreement and some scalar controls include zero. SciTail therefore supplies a protected cross-task replication of the action split: aggregate correction is not boundary-safe replacement, while the unresolved warning remains useful for review.

W Protected SALAD Native-Partial Multi-Backbone Study

SALAD contains human sentence-level attribution judgments for long-form QA under six generator–document settings (Chen et al., 2024). We use the official repository at commit `4e905be93fbe919082d62faa77ee32c68e90237c`. Each sentence has three labels from Supported, Partially Supported, and Not Supported. We retain the unique majority label and map these outcomes to $S/P/I$. One punctuation-only sentence is removed by a label-free modelability

rule fixed before inference. The final cohort has 3,856 rows in 100 question clusters: 2,290 S, 234 P, and 1,332 I.

The six settings vary WebGPT, GPT-3.5, and Alpaca answers and whether evidence comes from WebGPT documents, human documents, or no generation-time documents. For the human-document condition, evaluation uses the released `docs-human.json`. This matches the paper's study design and corrects a stale automatic-input binding in the public repository. All three folds of the global and local arms for DeBERTa, RoBERTa, and ModernBERT finish before key access. The same is true of 1,403 proposals and certificates, four review budgets, and all registered controls. No SALAD label fits a model, composer, calibrator, threshold, ranking, or certificate.

The label-free activation gate requires at least 14 boundary-separable PIAF actions in 14 question clusters for at least two families. DeBERTa has 145 actions in 66 clusters, and RoBERTa has 134 in 58. ModernBERT has only four in four clusters. The first two families therefore authorize one-time key access; ModernBERT remains a reported non-activation.

A family passes replacement only if PIAF improves question-cluster accuracy, makes more corrections than harms, has row harm at most 15%, has a one-sided exact cluster-harm upper bound at most 20%, preserves Partial recall within two points, and reduces gold-P harms relative to unfiltered RA-DVRC. Strict confirmation requires two family passes and a positive pooled equal-family question-cluster interval.

Table 48: Protected SALAD replacement. C/H/N denotes corrections, harms, and neutral changes. The harm upper bound is a one-sided exact 95% interval over action-bearing question clusters. ModernBERT did not pass the label-free activation gate.

Family	Proposals	PIAF actions/clusters	C/H/N	Δ acc. [95% CI]	ΔR_P	Cluster-harm U95	Gate
DeBERTa-v3	646	145/66	132/8/5	+3.13 [2.38,3.94]	+1.28	20.81	fail
RoBERTa	475	134/58	131/3/0	+3.20 [2.42,4.05]	-0.43	12.83	pass
ModernBERT	282	4/4	3/0/1	+0.08 [0.00,0.19]	0.00	52.71	not activated

The pooled PIAF effect is +2.14 question-cluster accuracy points with a 95% interval [1.72, 2.58]. This efficacy estimate is positive, but the registered confirmation fails: only RoBERTa passes every family gate, below the required two. DeBERTa fails only the cluster-action safety bound. ModernBERT supplies too few action clusters to estimate the declared safety target. PIAF reduces unfiltered gold-P harms from 19 to 3 for DeBERTa, from 17 to 3 for RoBERTa, and from 12 to 0 for ModernBERT, but those reductions do not override the strict gate.

The lower-authority result is more stable. The table below reports the locked 10% budget. The best control is the highest-precision member of anchor low confidence, entropy, margin, and generic fold disagreement.

Table 49: Protected SALAD deferral at the locked 10% row budget. Precision and retained risk are percentages.

Family	Reviewed	WCSD precision	Best control	WCSD risk	Error capture
DeBERTa-v3	385	94.29	80.78	26.74	28.12
RoBERTa	385	93.77	72.21	26.16	28.45
ModernBERT	385	87.53	71.95	24.86	28.08

WCSD has higher point deferral precision than every registered control at each 5%, 10%, 15%, and 20% budget for every family. More strongly, all 12 family-budget comparisons have paired question-cluster precision intervals above zero against every control. Retained-risk intervals are less uniform, so we do not claim simultaneous dominance on both metrics. SALAD therefore reproduces the authority split under a native Partial label and three encoder families: automatic replacement remains population- and boundary-sensitive, whereas warning-conditioned review transfers cleanly.

X Fresh GaRAGe Conditional Action-Risk Confirmation

X.1 Population, split, and frozen action

GaRAGe contains 2,366 questions with human-curated long-form answers and passage-level grounding annotations (Sorodoc et al., 2025). We use the public release at commit `c3604d0fa2a169c2d30f893fc0ae701b58604898` and pair each material generated-answer sentence with its explicitly cited grounding passage. Construction accesses only the question, generated answer, citation markers, and grounding text. It does not use answer validation, comments, evidence correctness, evidence relevance, or citation-quality fields.

After deterministic parsing and an exact-text overlap exclusion against the earlier GenSearch calibration source, the cohort contains 15,967 pairs in 1,939 question clusters. A seeded permutation assigns complete questions to 970 risk-calibration and 969 protected-confirmation clusters. No question crosses the split. The WiCE checkpoints, RA-DVRC composers, thresholds, point-PIAF certificates, and representative rule are unchanged from the earlier study. Within each question the rule chooses at most one action, preferring a proposal, then larger frozen advantage, then lexical unit ID. It selects 71 risk-half and 74 confirmation-half question clusters.

No GaRAGe $S/P/I$ relation label existed when these choices were made. A label-free packet check removed one malformed residual citation control token from annotator display only. It changed no semantic text, model input, score, selection, or action. All earlier artifacts containing the unclean display were quarantined before annotation.

X.2 Outer risk gate and protected decision

The estimand is conditional action-cluster harm among questions selected by this fixed policy. A harm occurs when the anchor matches the adjudicated relation and the candidate does not; a correction reverses those roles. Every other selected action is neutral. Because the policy selects at most one action per question, selected cluster outcomes are Bernoulli observations under the declared independent-exchangeable-cluster assumption.

The risk half tests the single fixed policy against a 20% harm target with an exact lower-tail binomial test at level .05. At least 40 selected clusters are required. Passing releases all 74 already-frozen confirmation actions. The confirmation rule is separate and fixed: all 74 actions must be adjudicated, corrections must exceed harms, and the one-sided 95% Clopper–Pearson harm upper bound must be at most 20%. No confirmation p-value is required, and no action or threshold is revised after either label set opens.

Table 50: Fresh GaRAGe conditional action-risk study for the unchanged point-PIAF policy. C/H/N denotes corrections, harms, and neutral selected actions. U95 is the exact one-sided 95% Clopper–Pearson harm upper bound.

Stage	Clusters	Actions	C/H/N	Harm	U95	Coverage	Decision
Risk calibration	970	71	56/5/10	7.04%	14.24%	7.32%	release
Protected confirmation	969	74	53/7/14	9.46%	17.03%	7.64%	pass

The risk-stage exact p-value is .00231. On confirmation, 40 demotions yield 32 corrections, 2 harms, and 6 neutral actions; 34 promotions yield 21, 5, and 8. The two independent annotators agree on 69/74 labels (93.24%), with Cohen's $\kappa = .840$ and quadratic-weighted $\kappa = .871$. A third person, blinded to the two labels and all model/action fields, adjudicates only the five disagreements. An independent recomputation from the final labels reproduces 53/7/14 and U95=.1703413875 exactly.

The conditional anchor and candidate accuracies on selected confirmation actions are 9.46% and 71.62%, respectively, a 62.16-point difference. This is not an overall GaRAGe accuracy result. The other 895 confirmation clusters do not receive new relation labels. The defensible conclusion is that an outer target-domain risk gate grants population authority to the fixed selected policy for this GaRAGe population under exchangeability. It does not provide arbitrary-shift or distribution-free safety.

What the harm target can change. GaRAGe tests one fixed policy, not a family of coverage thresholds. Its risk-stage upper bound is 14.24%; the independent confirmation upper bound is 17.03%. The already computed bounds therefore rule out a 10% target at risk calibration. A 15% target is compatible with calibration but fails the separate confirmation criterion; 20% and 25% are compatible with both. These are retrospective readings of two fixed intervals, not additional confirmatory tests. Only 20% was locked.

Changing the acceptable target does not generate new candidate actions. The fixed policy still selects 74/969 confirmation questions, so the evidence does not identify an optimized coverage–risk curve. Such a curve requires a prespecified policy family and valid selection across that family on independent calibration clusters. Arbitrarily subsampling the 74 known outcomes or choosing a new target cannot establish improved deployment coverage. The current result demonstrates nonzero authority at the declared target, not the largest achievable action set.

X.3 Negative marginal-band sensitivity

The original GaRAGe protocol also tested a stricter inner filter that enlarged the ambiguity set from point gate marginals to source-calibrated Venn–Abers multiprobability bands. This policy authorizes only 19 risk-half and 25 confirmation-half question representatives, below the locked minimum of 40 in each half. It therefore fails closed before relation labels and remains a negative activation result. The successful outer study does not rescue or retune that filter. It evaluates the already-frozen point-PIAF policy and moves the unresolved population uncertainty to an explicit target-domain LTT layer.

Y REPRESENTATION-FAMILY STRESS TEST ON MEDFACT

MEDFACT is a Chinese evidence-based medical fact-checking dataset with graded claim labels (Chen et al., 2025). We use only its public training release at repository commit `503a7aab21603a86520a41bbe774b6e3c71d7f2f`; this is not an official hidden-test result. Question IDs define disjoint clusters. The three-way map keeps Supported, Partially Supported, and Uncertain as $S/P/I$ and excludes Refuted and Not Applicable.

The protocol was frozen before MEDFACT model inference. A multilingual MiniLM checkpoint trained before dataset screening supplies one global head and one token head. We change only the deterministic token-to-unit memberships: `cjk_atomic`, `tokenizer_material`, and `shallow_clause`. For every proposal, the representation-robust filter requires nonnegative advantage for every unitization, every Frechet-consistent coupling, and three frozen loss vertices (zero–one, partial-exit cost 3, and false-support cost 3). The outer LTT layer can run only if the inner filter retains eligible action clusters.

Risk calibration contains 1,568 rows in 296 question clusters. The composer proposes 1,154 changes, but the representation-family filter authorizes none. The worst-representation counts are 994 CJK-atomic, 120 shallow-clause, and 40 tokenizer-material. Confirmation contains 1,523 rows in 270 clusters and 1,135 proposals. It again yields zero authorized actions, with corresponding counts of 988, 108, and 39.

Both partitions fail closed before their action labels are used. The result does not rank the three unitizations: `worst` identifies the view supplying a counterexample for a particular frozen proposal. It establishes a narrower point. Extending PIAF from one lexical representation to a declared finite family is computationally direct, but the full intersection can have zero coverage under cross-domain and cross-language shift. LTT cannot repair this activation failure because no action reaches the population-control layer. Subsequent outcome-opened MEDFACT development is excluded from the paper's confirmation claims.

Y.1 OUTCOME-OPENED PARSER-INFORMED SENSITIVITY

We separately test whether a parser-informed unitization repairs the lexical representation concern on the already opened SemanticCite cohort. This is a descriptive sensitivity analysis, not a new confirmation. The action builder loads no labels and changes no backbone, composer, calibrator, threshold, loss vertex, or nominal proposal. It adds spaCy dependency frames to the two frozen deterministic unitizations and requires the certificate to survive all three. Exact ordered-ID checks align 1,109 rows in 1,107 source clusters with the three-stage human audit before evaluation.

The parser view disagrees with the nominal candidate on 63 of 529 proposals (11.9%). The three-view intersection retains 47 of the 57 previously certified actions. It blocks ten actions, but nine were corrections and only one was a harm. Consequently, corrections/harms move from 39/17 to 30/16 and net corrections fall from 22 to 14. This analysis rules out a convenient story: adding a syntactic view does not automatically make the local representation more semantically faithful. A finite-family contract can expose representation sensitivity, but it cannot identify which unitization is correct after outcomes have opened. We therefore leave the frozen policy unchanged and treat semantic unit construction as an unresolved modeling problem.