STABLE REWEIGHTING FOR HETEROGENEOUS EVIDENCE FUSION UNDER SOURCE EXCLUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

How much can excluding a single evidence source change a fused prediction, even when the candidate models remain unchanged? This question is formalized as source-exclusion stability for fixed-support evidence fusion: controlling the largest shift induced in the fused predictive distribution when any one source's complete evidence contribution is removed. We propose Source-Exclusion Stable Reweighting (SESR), a reweighting objective that, to our knowledge, is the first to directly penalize the largest, rather than average, squared displacement between the all-source mixture weights and each leave-one-source-out mixture weight, jointly optimized on a shared candidate support. We prove a deterministic bound connecting this worst-case weight displacement to predictive total-variation shift, establishing weight-space coupling as a principled surrogate for prediction stability. On five weak-supervision datasets, SESR reduces worst-case prediction shift by 18.9–31.9% relative to Independent Reweighting (IR) and by 15.3–25.9% relative to Average Coupling (AC) across all dataset means, with corresponding weight-shift reductions and broadly preserved classification accuracy. These results support source-exclusion stability as an explicit, optimizable objective for heterogeneous evidence fusion.

1 INTRODUCTION

Integrating heterogeneous evidence is fundamental to scientific inference and machine learning. For example, reconstructing conformational ensembles of intrinsically disordered proteins (IDPs) requires combining complementary experimental measurements with computationally generated conformations (Janson & Feig, 2024). Recent research has integrated multiple experimental observables through ensemble reweighting and used leave-one-out validation to assess agreement with withheld measurements (Borthakur et al., 2025). More broadly, programmatic weak supervision combines information from multiple labeling functions (LFs), while predictive ensemble methods integrate outputs from multiple candidate models (Ratner et al., 2017; Zhang et al., 2021; Yao et al., 2018). Across these settings, the resulting prediction depends not only on individual evidence contributions, but also on which contributions are included.

This dependence matters when evidence is excluded after quality reassessment or deliberately withheld to examine the sensitivity of an inference. Even if the candidate models remain unchanged, excluding a single contribution may substantially alter the fused prediction. Such a change does not necessarily indicate poor evidence quality: a valid contribution may contain information that other sources cannot replace. Nevertheless, it reveals the dependence of the fused result on the choice of evidence. This motivates a general question: *how can we control the largest prediction change caused by excluding any single evidence contribution while continuing to fit the available evidence?*

We investigate this question in fixed-support evidence reweighting for discrete categorical predictions. In our weak-supervision experiments, each LF constitutes one evidence source, contributing a source-specific loss to the reweighting objective. Excluding an LF removes its complete loss contribution, while the candidate predictors, their parameters, the dataset, and their predictions remain fixed; this is neither training-sample deletion nor machine unlearning. Source importance or valuation need not equal the influence of deleting that source on a particular fusion solution. Existing source attribution, worst-group risk, and predictive stacking address related but distinct questions (Zhang et al., 2022;

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.

Guan & Koudas, 2024; Sagawa et al., 2020; Yao et al., 2018). We focus on jointly optimizing the all-source and all single-source-excluded mixture-weight vectors on a shared support.

We propose *Source-Exclusion Stable Reweighting* (SESR) to address this problem. SESR jointly optimizes the all-source and leave-one-source-out mixture weights by combining their evidence-fitting objectives with a penalty on the maximum squared distance between the all-source weights and any source-excluded weights. Unlike independent branch optimization, this formulation explicitly couples the solutions; unlike average coupling, it directly targets the largest exclusion-induced displacement. We further establish an upper bound relating mixture-weight displacement to predictive total-variation distance, connecting the optimization mechanism to prediction sensitivity.

Our contributions are threefold:

- **Source-exclusion stability.** We formulate the sensitivity of fixed-support evidence fusion to complete single-source exclusion, with the largest paired predictive-distribution change as the primary evaluation endpoint.
- **Worst-case coupled reweighting.** We develop SESR to jointly optimize the all-source and all single-source-excluded mixture weights through a maximum-displacement penalty, and establish a bound linking weight displacement to predictive sensitivity.
- **Experimental evidence and scope.** On five controlled weak-supervision testbeds, SESR reduces the primary endpoint relative to mechanism-matched independent and average-coupled controls under the training-side coupling selection procedure. Fixed-support external adaptations and source-structure controls identify complementary regimes and applicability boundaries.

2 Related Work

Weak-supervision source attribution and valuation. Data programming and Snorkel aggregate noisy labeling functions (LFs) (Ratner et al., 2016; 2017); FlyingSquid, CAGE, dependency modeling, and WRENCH extend source aggregation and its evaluation (Fu et al., 2020; Chatterjee et al., 2020; Varma et al., 2019; Zhang et al., 2021). Robust labeling-function methods address imprecise or adversarial supervision (Chatterjee et al., 2020; Sedova & Roth, 2023; An, 2025). Source-aware influence functions attribute LF-associated effects (Zhang et al., 2022); WeShap values sources by Shapley contributions to a validation-scored proxy (Guan & Koudas, 2024). Neither directly penalizes the displacement of jointly reweighted all-source and excluded mixtures. Our WeShap-based fixed-support adaptation is not its published end-to-end pipeline.

Missing evidence and robust objectives. Missing-modality methods change inputs or learned representations (Neverova et al., 2016; Li et al., 2024); SESR holds predictions fixed and removes a complete source term. Group DRO and related distributionally robust methods target worst-group or distributional risk (Sagawa et al., 2020; Duchi & Namkoong, 2021; Levy et al., 2020), not all-source versus excluded weight displacement. Our fixed-support source-group DRO adaptation uses LF losses as groups and independently re-solves after exclusion; it is not the published neural training pipeline.

Predictive mixtures and exclusion stability. Bayesian stacking uses observation-level leave-one-out scoring to combine predictive distributions (Yao et al., 2018), whose model-comparison uncertainty has also been studied (Sivula et al., 2025); algorithmic stability examines sensitivity to training-example changes (Bousquet & Elisseeff, 2002). Neither removes an entire evidence source from a fixed-support fusion objective. Worst-case consistency across augmented views also addresses a different perturbation (Jiang et al., 2023). SESR instead couples all-source and single-source-excluded solutions on a shared candidate support; IR and AC are mechanism-matched internal controls.

3 Method

We study the sensitivity of fixed-support evidence reweighting to the exclusion of an entire evidence source. Our goal is to limit the largest change in the fused prediction when any single source

is excluded, while retaining the ability to fit the available evidence. To this end, we propose Source-Exclusion Stable Reweighting (SESR), which jointly optimizes the all-source and single-source-excluded mixture weights through a worst-case displacement penalty.

Source exclusion refers to excluding one source's complete evidence-loss contribution from the fusion objective while keeping the candidate support and predictions fixed. It does not delete training data, remove candidate predictors, modify candidate predictions, or perform machine unlearning.

Figure 1 summarizes the fixed-support setup, the jointly coupled source-exclusion branches, and the primary prediction-shift comparison.

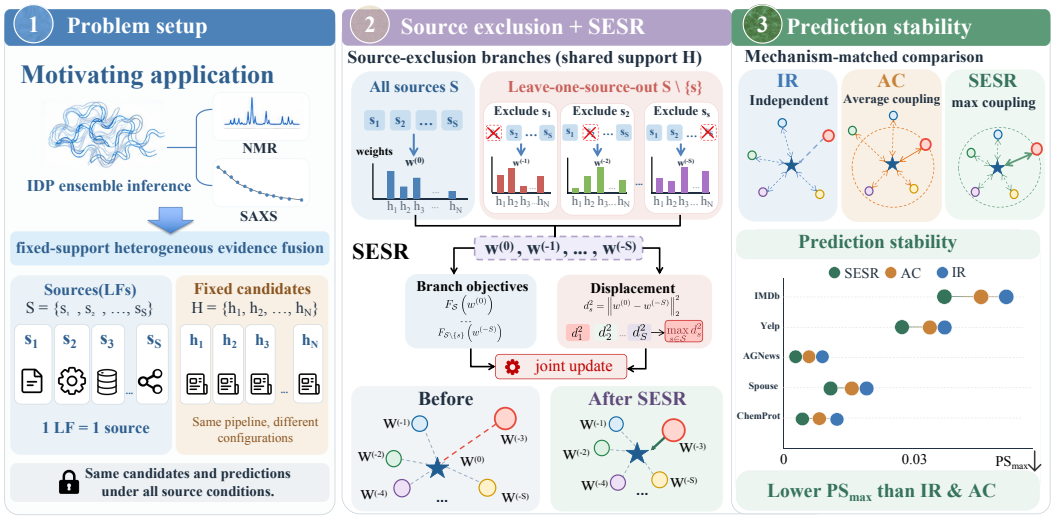

Figure 1: Overview of source-exclusion evidence fusion and SESR. (1) Heterogeneous evidence sources score a fixed candidate support; the same candidate predictors and predictions are retained across source conditions. (2) All-source and single-source-excluded branches are jointly optimized, with SESR penalizing the maximum squared weight displacement. (3) The primary endpoint, Worst-Case Prediction Shift ($\mathrm{PS}_{\max}$, lower is better), is compared across five datasets for Independent Reweighting (IR), Average Coupling (AC), and SESR.

3.1 Problem Formulation

Let $\mathcal{H} = \{h_i\}_{i=1}^{N}$ denote a fixed set of candidate predictors, where $h_i(\cdot \mid x) \in \Delta_C$ is the class-probability distribution produced by candidate i for input x. A mixture-weight vector $w \in \Delta_N$ defines the fused predictive distribution

$$p_w(k \mid x) = \sum_{i=1}^{N} w_i h_i(k \mid x), \qquad k \in \{1, \ldots, C\}, \tag{1}$$

where Δ_N denotes the probability simplex over the N candidates.

Let $\mathcal{S} = \{1, \ldots, S\}$, with $S \geq 2$, index the evidence sources. Each source supplies a fixed vector of normalized candidate losses $\widehat{\ell}_s = (\widehat{\ell}_{s,1}, \ldots, \widehat{\ell}_{s,N})$. The implemented source evidence loss is linear in the mixture weights,

$$\widehat{L}_s(w) = \langle w, \widehat{\ell}_s \rangle = \sum_{i=1}^{N} w_i \widehat{\ell}_{s,i}. \tag{2}$$

This linear form is expected candidate-level evidence loss under w, not cross entropy of the mixture predictive distribution. The candidate losses and their normalization constants are computed once before any source is excluded; Appendix B gives the construction. For a nonempty subset $A \subseteq \mathcal{S}$, we

define the evidence-fitting objective

$$F_A(w) = \frac{1}{|A|} \sum_{s \in A} \widehat{L}_s(w) + \lambda_{\mathrm{KL}} \, \mathrm{KL}(w \| \pi_0), \tag{3}$$

where π_0 is a prior over the candidate weights and $\lambda_{\mathrm{KL}} \geq 0$ controls regularization. We use a uniform prior, $\lambda_{\mathrm{KL}} = 0.1$, and $\eta = 1$ in all reported experiments.

We denote the all-source mixture weights by $w^{(0)}$ and the weights obtained when source s is excluded by $w^{(-s)}$. Source exclusion removes the complete contribution $\widehat{L}_s$ from the evidence objective, which is then averaged over the remaining $S - 1$ sources. The candidate predictors and their predictions remain fixed: source exclusion does not remove candidates or retrain them.

In our weak-supervision experiments, one labeling function (LF) constitutes one independently removable evidence source. The formulation itself is expressed in terms of source-specific evidence losses and does not require identifying the most influential source in advance.

3.2 Source-Exclusion Stable Reweighting

A straightforward source-exclusion analysis optimizes $F_{\mathcal{S}}$ and each $F_{\mathcal{S}\setminus\{s\}}$ independently. Although this produces a mixture for every evidence subset, it does not explicitly control how far a source-excluded mixture can move from the all-source mixture.

Joint optimization with an average-displacement penalty can encourage agreement across these mixtures. However, a small average displacement does not ensure that the largest displacement under a particular source exclusion is small. Because our objective is to control sensitivity to *any single* source exclusion, we target the maximum displacement directly.

SESR jointly optimizes the all-source and all single-source-excluded mixture weights:

$$\begin{aligned} \min_{\mathbf{W} \in \Delta_N^{S+1}} \quad & F_{\mathcal{S}}\big(w^{(0)}\big) + \frac{\eta}{S} \sum_{s=1}^{S} F_{\mathcal{S}\setminus\{s\}}\big(w^{(-s)}\big) \\ & + \gamma \max_{s \in \mathcal{S}} \left\| w^{(0)} - w^{(-s)} \right\|_2^2, \end{aligned} \tag{4}$$

where $\mathbf{W} = (w^{(0)}, w^{(-1)}, \ldots, w^{(-S)})$, η balances the all-source and source-excluded evidence-fitting terms, and $\gamma \geq 0$ weights the stability penalty. Here, $\| \cdot \|_2$ denotes the Euclidean norm.

The first two terms fit each mixture to its corresponding evidence set. The final term couples all branches through their largest squared weight displacement. Consequently, an exclusion branch with a large displacement directly contributes to the stability penalty, rather than being diluted by averaging its displacement with those of other branches. The method jointly balances evidence fitting and source-exclusion stability without prespecifying which source will induce the largest change.

We approximately solve Equation 4 using exponentiated-gradient mirror descent. At each iteration, the all-source and source-excluded weight vectors are updated with respect to the joint objective, incorporating the evidence gradients and a contribution from the maximum-displacement term. Exponentiated updates followed by normalization maintain the probability-simplex constraints. Numerical settings and implementation details are provided in Appendix B.

To distinguish the effect of worst-case coupling from joint optimization alone, we compare SESR with two mechanism-matched internal controls. *Independent Reweighting* (IR) independently minimizes $F_{\mathcal{S}}$ and each $F_{\mathcal{S}\setminus\{s\}}$ without cross-branch coupling. *Average Coupling* (AC) solves the complete objective

$$\begin{aligned} \min_{\mathbf{W} \in \Delta_N^{S+1}} \quad & F_{\mathcal{S}}\big(w^{(0)}\big) + \frac{\eta}{S} \sum_{s=1}^{S} F_{\mathcal{S}\setminus\{s\}}\big(w^{(-s)}\big) \\ & + \frac{\gamma}{S} \sum_{s=1}^{S} \left\| w^{(0)} - w^{(-s)} \right\|_2^2, \end{aligned} \tag{5}$$

with the same $\eta = 1$, branch objectives, initialization, optimizer, and coupling grid as SESR. All three methods use the same candidate support and source-exclusion procedure. Their comparison separates

independent reweighting, average coupling, and direct control of the largest exclusion-induced weight displacement.

3.3 From Weight-Space Coupling to Predictive Stability

SESR penalizes displacement between mixture weights, whereas our ultimate interest is the stability of the resulting predictive distributions. We therefore examine how weight-space displacement relates to prediction-level change under single-source exclusion.

Let $\mathcal{T}$ be an evaluation set. We define the *Worst-Case Prediction Shift* ($\mathrm{PS}_{\max}$) as

$$\mathrm{PS}_{\max} = \max_{s \in \mathcal{S}} \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \mathrm{TV}\left(p_{w^{(0)}}(\cdot \mid x), p_{w^{(-s)}}(\cdot \mid x)\right), \tag{6}$$

where $\mathrm{TV}(p, q) = \frac{1}{2} \sum_{k=1}^{C} |p(k) - q(k)|$ is the total-variation (TV) distance. The metric first averages predictive change over evaluation examples and then takes the maximum across excluded sources. Lower values indicate less sensitivity to single-source exclusion.

The corresponding *Worst-Case Weight Shift* ($\mathrm{WS}_{\max}$) is

$$\mathrm{WS}_{\max} = \max_{s \in \mathcal{S}} \left\| w^{(0)} - w^{(-s)} \right\|_2. \tag{7}$$

This metric uses the unsquared Euclidean distance; the stability penalty in Equation 4 uses its square.

The following proposition establishes a deterministic connection between the two quantities.

Proposition 1 (Weight-to-Prediction Stability Bound)**.** *Suppose the candidate support $\mathcal{H} = \{h_i\}_{i=1}^{N}$ is fixed and each candidate produces a valid probability distribution $h_i(\cdot \mid x) \in \Delta_C$. For any all-source and source-excluded weight vectors in Δ_N,*

$$\boxed{\mathrm{PS}_{\max} \leq \frac{\sqrt{N}}{2} \mathrm{WS}_{\max}.} \tag{8}$$

Proof sketch. For any input x and excluded source s, the fixed-support mixture representation and the triangle inequality give

$$\begin{aligned} &\mathrm{TV}\left(p_{w^{(0)}}(\cdot \mid x), p_{w^{(-s)}}(\cdot \mid x)\right) \\ &\quad \leq \frac{1}{2} \left\| w^{(0)} - w^{(-s)} \right\|_1 \leq \frac{\sqrt{N}}{2} \left\| w^{(0)} - w^{(-s)} \right\|_2. \end{aligned} \tag{9}$$

Averaging over $\mathcal{T}$ and taking the maximum over s yields Equation 8. The bound holds pointwise before averaging and requires no disjointness of candidate predictive supports. The full proof is provided in Appendix A.

Implication for SESR. Proposition 1 connects the quantity penalized by SESR to the prediction-level stability of the fixed-support mixture. Equation 4 penalizes $\mathrm{WS}_{\max}^2$, while Proposition 1 bounds $\mathrm{PS}_{\max}$ in terms of $\mathrm{WS}_{\max}$. This is an upper-bound surrogate, not a two-sided equivalence: nearly collinear candidate predictions can make the bound loose, and a smaller $\mathrm{WS}_{\max}$ need not yield a strictly smaller $\mathrm{PS}_{\max}$ in every realization.

4 Experiments

We evaluate SESR by investigating two questions: (i) Does worst-case coupling reduce prediction changes under single-source exclusion? (ii) Is the improvement accompanied by smaller changes in mixture weights while retaining comparable task performance?

4.1 Experimental Setup

Datasets and candidate support. We evaluate five weak-supervision datasets: IMDb, Yelp, AG-News, Spouse, and ChemProt, containing 5, 8, 9, 9, and 26 labeling functions (LFs), respectively. Each LF constitutes one evidence source that can be excluded independently.

For each dataset, we use a fixed support of $N = 24$ candidate classifiers and evaluate $K = 5$ support realizations. The candidates are configurations of the same text-classification pipeline rather than distinct model architectures. Their identities and predictions remain unchanged across all source-exclusion conditions.

Candidates are fitted using gold-labeled data from the local benchmark snapshots (the validation split for Spouse, whose training labels are unavailable). LF outputs define the source-specific evidence losses; excluding an LF removes its complete loss contribution without retraining candidates. These WRENCH tasks are a controlled evidence-fusion testbed, not an end-to-end weak-supervision training pipeline. Construction details are in Appendix B.

Comparison methods. IR independently optimizes the all-source and source-excluded branches; AC jointly optimizes them with average rather than maximum squared-displacement coupling. All methods share the candidate support, evidence losses, and exclusion procedure. We also evaluate fixed-support source-group DRO and WeShap-based adaptations and a source-cluster control; these are not reproductions of the external methods' end-to-end pipelines.

Evaluation metrics. The primary endpoint is Worst-Case Prediction Shift ($\mathrm{PS}_{\max}$, $\downarrow$), the largest mean predictive total-variation shift across single-source exclusions. Worst-Case Weight Shift ($\mathrm{WS}_{\max}$, $\downarrow$) measures the corresponding maximum unsquared Euclidean weight displacement. We also report all-source Accuracy and NLL, signed differences from IR, and weight-concentration diagnostics; these task-quality measures do not themselves establish performance after exclusion.

Implementation and reporting. We solve the coupled objectives with exponentiated-gradient mirror descent and select γ separately for SESR and AC using training-side weight stability and evidence risk, without test labels or test metrics. Results are dataset means over five support realizations; relative reductions use these means. Full optimization and selection details, including sensitivity to the threshold convention, are provided in Appendix B.

4.2 Prediction Stability under Source Exclusion

Table 1 presents the primary results. SESR achieves lower Worst-Case Prediction Shift than both internal controls across all five datasets.

The dataset-level reductions range from 18.9% to 31.9% relative to IR and from 15.3% to 25.9% relative to AC. Across the selected configurations, the same direction holds across all 25 matched dataset-by-support comparisons against each internal control.

Table 1: Prediction stability under single-source exclusion. Worst-Case Prediction Shift ($\mathrm{PS}_{\max}$; mean predictive total-variation distance, $\downarrow$). Values are averaged over five support realizations per dataset. Relative reductions are computed from dataset-level means.

Dataset	IR	AC	SESR	Red. vs IR	Red. vs AC
IMDb	0.049096	0.043498	**0.035497**	27.7%	18.4%
Yelp	0.033065	0.031041	**0.024068**	27.2%	22.5%
AGNews	0.004948	0.004548	**0.003371**	31.9%	25.9%
Spouse	0.015397	0.014270	**0.012093**	21.5%	15.3%
ChemProt	0.008688	0.008442	**0.007046**	18.9%	16.5%

The comparison with AC is particularly informative. Both methods jointly optimize the same branch family, but SESR explicitly penalizes the maximum rather than the average weight displacement. The observed reduction in predictive shift is consistent with the intended effect of worst-case coupling in the evaluated setting.

Fixed-γ and alternative-selector analyses preserve the dataset-level direction of the main result; detailed sensitivity results are reported in Appendix C.3.

Table 2: Weight stability and all-source diagnostics. Panel A reports Worst-Case Weight Shift ($\mathrm{WS}_{\max}$; unsquared Euclidean distance, ↓). Panel B reports SESR's all-source task-performance differences relative to IR at the selected γ, together with its weight-concentration diagnostics. Values are means over five support realizations per dataset.

Panel A: Worst-Case Weight Shift ($\mathrm{WS}_{\max}$, ↓)

Dataset	IR	AC	SESR	Red. vs IR	Red. vs AC
IMDb	0.138987	0.119421	**0.093118**	33.0%	22.0%
Yelp	0.082322	0.076694	**0.057390**	30.3%	25.2%
AGNews	0.019946	0.017887	**0.012095**	39.4%	32.4%
Spouse	0.094892	0.083924	**0.064485**	32.0%	23.2%
ChemProt	0.019413	0.018609	**0.013768**	29.1%	26.0%

Panel B: All-Source Task Performance at Selected γ and Weight Concentration

Dataset	$\Delta\mathrm{Acc}$	$\Delta\mathrm{NLL}$	Kish (%)	Max. weight
IMDb	-0.001600	$+0.005344$	60.18	0.125109
Yelp	-0.001053	$+0.004564$	75.56	0.078810
AGNews	$+0.000017$	-0.000019	22.10	0.317389
Spouse	0.000000	$+0.000290$	67.22	0.113889
ChemProt	$+0.000498$	-0.001279	37.39	0.183148

4.3 Weight Stability and Task Performance

Weight-space mechanism. Table 2, Panel A, reports Worst-Case Weight Shift. SESR achieves smaller maximum weight displacement than both controls across all five dataset means and all 25 matched support comparisons.

The dataset-level reductions range from 29.1% to 39.4% relative to IR and from 22.0% to 32.4% relative to AC. The prediction-space and weight-space results therefore show the same directional pattern, consistent with the maximum-displacement coupling and the theoretical relationship established in Section 3.3.

Task performance and weight concentration. Panel B reports all-source accuracy and NLL differences relative to IR, together with SESR's weight-concentration diagnostics. Specifically, $\Delta\mathrm{Acc} = \mathrm{Acc}_{\mathrm{SESR}} - \mathrm{Acc}_{\mathrm{IR}}$ and $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{SESR}} - \mathrm{NLL}_{\mathrm{IR}}$; Kish percent and maximum weight are computed from SESR's all-source branch $w^{(0)}$ at the selected γ.

The all-source performance differences are small in absolute magnitude, although their directions vary across datasets. Normalized Kish effective support and maximum candidate weight characterize the distribution of mixture weights.

A positive $\Delta\mathrm{Acc}$ indicates higher all-source accuracy for SESR, whereas a negative $\Delta\mathrm{NLL}$ indicates lower all-source NLL. Neither quantity measures the effect of source exclusion on task performance. Together, these results distinguish improved prediction stability from task performance and weight concentration.

LOSO task quality. Across 25 matched supports, SESR's mean minimum leave-one-source-out (LOSO) Accuracy is 0.819262, versus 0.817487 for IR and 0.818386 for AC; its mean maximum NLL increment is $+0.018227$, versus $+0.023781$ and $+0.021591$, respectively. Aggregate LOSO task quality remains comparable, while branch-level exceptions and macro-F1 results are reported in Appendix C.

4.4 External and Source-Structure Stress Tests

To probe applicability beyond the mechanism-matched controls, we use the same fixed candidates and source losses for two external-objective adaptations (Appendix D). A *fixed-support source-group DRO adaptation* minimizes worst LF evidence loss with the fixed KL coefficient, independently re-solving each deletion. It has lower $\mathrm{PS}_{\max}$ than SESR in 15/25 supports (all IMDb, Yelp, and

Spouse seeds), but higher shift in 10/25 (all AGNews and ChemProt seeds); its all-source NLL is worse than SESR in 25/25. Worst-source risk and exclusion-induced solution displacement are related but different robustness targets.

A *WeShap-based fixed-support adaptation* maps LF source valuations computed using gold validation labels to nonnegative weights on the fixed LF losses, then re-solves after deletion. These labels provide additional source-weighting supervision relative to IR, AC, SESR, and source-group DRO. Its $\mathrm{PS}_{\max}$ is lower than SESR in 10/25 supports (all Yelp and Spouse seeds) and higher in 15/25 (all IMDb, AGNews, and ChemProt seeds). Its all-source NLL is better in 20/25, an important utility distinction. Both adaptations are unified-protocol controls, not comparisons against the original neural Group DRO or end-to-end WeShap pipelines. Different objectives occupy different stability–utility operating regimes; no combined score or universal dominance is implied.

Table 3: External baselines under the common fixed-support protocol. Means over five support realizations per dataset; lower is better for both endpoints. Source-group DRO and WeShap-based are fixed-support adaptations, not the original end-to-end methods. Across 25 paired supports, DRO and WeShap-based have lower $\mathrm{PS}_{\max}$ than SESR in 15 and 10 cases, respectively.

Dataset	SESR	Source-group DRO	WeShap-based
Panel A: dataset-mean $\mathrm{PS}_{\max}$			
IMDb	0.035497	0.011537	0.060007
Yelp	0.024068	0.014272	0.017468
AGNews	0.003371	0.022511	0.005580
Spouse	0.012093	0.005959	0.003045
ChemProt	0.007046	0.013620	0.013339
Panel B: dataset-mean all-source NLL			
IMDb	0.509218	0.552302	0.491383
Yelp	0.413439	0.455862	0.403154
AGNews	0.689419	0.729161	0.692761
Spouse	0.249252	0.260278	0.243771
ChemProt	1.133157	1.364849	1.074997

Figure 2 examines three representative source structures: clustered redundancy (Spouse), dominant-source prevalence (IMDb), and source overlap (ChemProt). On Spouse, explicit clustering substantially reduces single-source prediction shift, indicating that strong source redundancy can be handled effectively by a simple cluster-aware control, although removing an entire cluster remains highly disruptive. On IMDb, the dominant source by coverage and evidence mass is distinct from the source that induces the largest deletion shift, showing that source prevalence and source valuation need not identify exclusion influence. On ChemProt, overlap and label agreement co-vary, while SESR can reduce the global worst-case shift even when some individual source-specific shifts increase. These cases illustrate that exclusion stability depends on source structure rather than on any single proxy such as coverage, valuation, or overlap. Detailed per-source and structural results are provided in Appendix D.

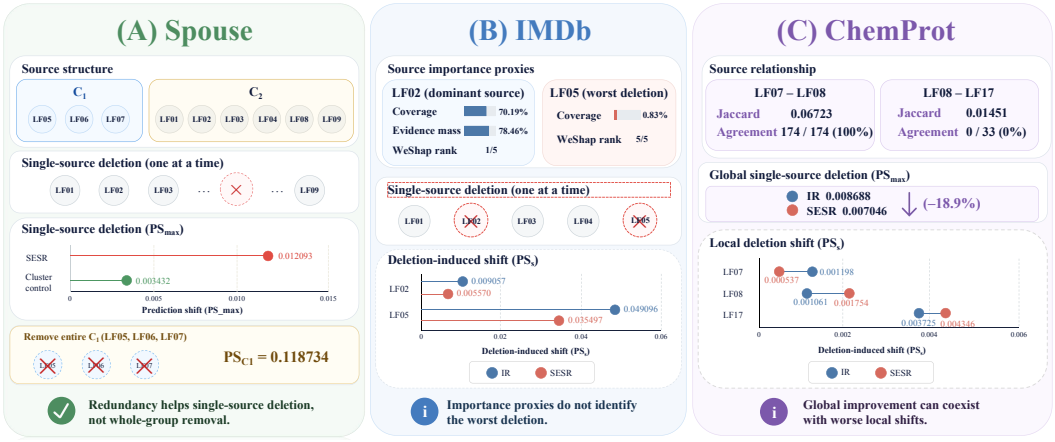

Figure 2: Source-structure case studies of exclusion-induced prediction drift. (A) On Spouse, clustered redundancy reduces single-source sensitivity, whereas whole-cluster removal is a distinct and more disruptive intervention. (B) On IMDb, the dominant source differs from the worst-deletion source, separating prevalence and valuation from deletion influence. (C) On ChemProt, overlap co-varies with agreement; global $\mathrm{PS}_{\max}$ decreases even though some local PS_s values increase. Detailed analyses are provided in Appendix D.

5 Discussion

SESR directly targets worst single-source exclusion stability, unlike IR's independent solutions or AC's average coupling. The one-sided weight-to-prediction bound motivates the mechanism but does not guarantee a strict prediction-shift reduction or higher task quality in every realization. The fixed-support external adaptations show that worst-source risk, source valuation, and exclusion stability are distinct objectives: source-group DRO sometimes lowers shift at higher all-source NLL, while the WeShap adapter achieves both lower $\mathrm{PS}_{\max}$ and lower all-source NLL on Yelp and Spouse. On IMDb, dominant source mass and WeShap value belong to LF02, but LF05 determines the worst deletion; source mass or value need not equal deletion influence.

Explicit source clustering can outperform SESR for single-source deletion, as on Spouse, without establishing robustness to removal of an entire cluster. The evidence comes from gold-trained, fixed candidate supports and single-source exclusions. Detailed task-quality and sensitivity analyses are provided in Appendices B and C. Extension to noisy candidate supports, group removal, or continuous IDP/NMR/SAXS evidence requires separate validation.

6 Conclusion

We studied source-exclusion stability in fixed-support evidence fusion, focusing on the largest change in fused predictions caused by excluding any single evidence source. We proposed Source-Exclusion Stable Reweighting (SESR), which jointly optimizes the all-source and source-excluded mixture weights through a maximum-displacement penalty. We further established a deterministic bound connecting weight displacement to predictive-distribution shift, providing a theoretical basis for the stability surrogate used in SESR.

In the evaluated fixed-support setting, experiments on five weak-supervision testbeds show lower Worst-Case Prediction Shift than independent reweighting and average coupling across all dataset means. Weight-shift reductions support the coupling mechanism, while task utility remains a separate endpoint. External adaptations and source-structure controls show that alternative robustness mechanisms can be preferable under different evidence geometries or utility criteria.

Reproducibility Statement

Appendix B specifies candidate construction, LF evidence losses, optimization, training-side γ selection, and threshold sensitivity. The fixed candidates are fitted using gold labels (using the local validation split for Spouse, whose training labels are unavailable), but test labels are used only for task-quality evaluation. Source exclusion leaves candidates unchanged; $\mathrm{PS}_{\max}$ uses test-set predictions without test labels. Supplementary numerical tables provide the matched-support and source-exclusion records.

AI Use Statement

AI tools assisted with research and method discussion, experiment and protocol design, coding, debugging, analysis support, result interpretation and consistency checking, literature search and discovery, drafting portions of the manuscript, editing, and LaTeX cleanup. Scientific decisions, experiments, numerical verification, interpretations, and final claims remain the authors' responsibility and were reviewed by the authors.

References

Steven An. Statistical analysis of an adversarial Bayesian weak supervision method. In *Advances in Neural Information Processing Systems*, volume 38, pp. 141083–141127, 2025. doi: 10.52202/085713-4239. URL https://doi.org/10.52202/085713-4239.

Kaushik Borthakur, Thomas R. Sisk, Francesco P. Panei, Massimiliano Bonomi, and Paul Robustelli. Determining accurate conformational ensembles of intrinsically disordered proteins at atomic resolution. *Nature Communications*, 16(1):9036, 2025. doi: 10.1038/s41467-025-64098-3. URL https://doi.org/10.1038/s41467-025-64098-3.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. URL https://jmlr.org/papers/v2/bousquet02a.html.

Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Robust data programming with precision-guided labeling functions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4):3397–3404, 2020. doi: 10.1609/aaai.v34i04.5742.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. doi: 10.1214/20-AOS2004.

Daniel Y. Fu, Mayee F. Chen, Frédéric Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3280–3291, 2020. URL https://proceedings.mlr.press/v119/fu20a.html.

Naiqing Guan and Nick Koudas. WeShap: Weak supervision source evaluation with Shapley values. *Proceedings of the VLDB Endowment*, 18(4):1063–1076, 2024. doi: 10.14778/3717755.3717766. URL https://doi.org/10.14778/3717755.3717766.

Giacomo Janson and Michael Feig. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLOS Computational Biology*, 20(5):e1012144, 2024. doi: 10.1371/journal.pcbi.1012144. URL https://doi.org/10.1371/journal.pcbi.1012144.

Yangbangyan Jiang, Xiaodan Li, Yuefeng Chen, Yuan He, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. MaxMatch: Semi-supervised learning with worst-case consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5970–5987, 2023. doi: 10.1109/TPAMI.2022.3208419. URL https://doi.org/10.1109/TPAMI.2022.3208419.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8847–8860, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/64986d86a17424eeac96b08a6d519059-Abstract.html.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10074–10082, 2024. doi: 10.1609/aaai.v38i9.28871. URL https://doi.org/10.1609/aaai.v38i9.28871.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Alexander J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3567–3575, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/6709e8d64a5f47269ed5cea9f625f7ab-Abstract.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Anastasiia Sedova and Benjamin Roth. ULF: Unsupervised labeling function correction using cross-validation for weak supervision. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4162–4176, 2023. doi: 10.18653/v1/2023.emnlp-main.254. URL https://arxiv.org/abs/2204.06863.

Tuomas Sivula, Måns Magnusson, Asael Alonzo Matamoros, and Aki Vehtari. Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *Bayesian Analysis*, 2025. doi: 10.1214/25-BA1569. URL https://doi.org/10.1214/25-BA1569. Advance publication.

Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, and Christopher Ré. Learning dependency structures for weak supervision models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6418–6427, 2019. URL https://proceedings.mlr.press/v97/varma19a.html.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018. doi: 10.1214/17-BA1091. URL https://doi.org/10.1214/17-BA1091.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. WRENCH: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2109.11377.

Jieyu Zhang, Haonan Wang, Cheng-Yu Hsieh, and Alexander Ratner. Understanding programmatic weak supervision via source-aware influence function. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2862–2875, 2022. doi: 10.52202/068431-0207. URL https://doi.org/10.52202/068431-0207.

A Theory and Proofs

Convexity of the SESR optimization problem. The feasible set is the product of simplices Δ_N^{S+1}. The evidence objectives in the implementation are linear functions plus KL regularizers, and the maximum of the convex squared Euclidean displacements is convex. Equivalently, one may introduce an epigraph variable t with constraints $\|w^{(0)} - w^{(-s)}\|_2^2 \le t$ for every source. Thus, the mathematical SESR objective is convex. The experiments use the fixed-budget exponentiated-gradient mirror-descent implementation described in Appendix B; we do not rely on a global-optimality or convergence claim for those finite iterates.

A.1 Proof of Proposition 1

Fix an input x and an excluded source s, and write $\delta_i = w_i^{(0)} - w_i^{(-s)}$. By the fixed-support mixture definition,

$$p_{w^{(0)}}(\cdot \mid x) - p_{w^{(-s)}}(\cdot \mid x) = \sum_{i=1}^{N} \delta_i h_i(\cdot \mid x).$$

Each $h_i(\cdot \mid x)$ is a probability-simplex vector, so $\|h_i(\cdot \mid x)\|_1 = 1$. The triangle inequality therefore yields

$$\begin{aligned}
\mathrm{TV}(p_{w^{(0)}}(\cdot \mid x), p_{w^{(-s)}}(\cdot \mid x)) &= \frac{1}{2} \left\| \sum_{i=1}^{N} \delta_i h_i(\cdot \mid x) \right\|_1 \\
&\le \frac{1}{2} \sum_{i=1}^{N} |\delta_i| \, \|h_i(\cdot \mid x)\|_1 \\
&= \frac{1}{2} \|\delta\|_1 \\
&\le \frac{\sqrt{N}}{2} \|\delta\|_2,
\end{aligned}$$

where the final inequality is Cauchy–Schwarz. The right-hand side does not depend on x, hence averaging over the evaluation set $\mathcal{T}$ gives

$$\frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \mathrm{TV}(p_{w^{(0)}}(\cdot \mid x), p_{w^{(-s)}}(\cdot \mid x)) \le \frac{\sqrt{N}}{2} \left\| w^{(0)} - w^{(-s)} \right\|_2.$$

Taking the maximum over $s \in \mathcal{S}$ proves

$$\mathrm{PS}_{\max} \le \frac{\sqrt{N}}{2} \mathrm{WS}_{\max}.$$

The bound is a geometric statement about mixtures on a fixed support. It supplies neither an accuracy guarantee nor a claim that SESR dominates every comparator.

B Experimental Protocol and Data Construction

The local IMDb snapshot contains five LFs, and the Spouse validation split contains 2,811 examples, consistent with the reported WRENCH task statistics. We could not uniquely associate these local files with an immutable public repository commit. No LF filtering was applied.

B.1 Candidate construction

For IMDb, Yelp, AGNews, and ChemProt, the candidate support is built from the gold-labeled training split before any evidence-source exclusion. Because all Spouse training labels are unavailable (encoded as -1), its candidates are fitted on the 2,811-row gold-labeled local validation split; LF evidence is still constructed on the Spouse training rows. For every dataset, the support has size $N = 24$. Weak-supervision labels are not used to fit the candidate predictors. In WRENCH, one

source means one labeling function, so LF outputs enter the source-evidence layer after the candidate bank has been generated.

The implementation constructs the 24 candidates from the Cartesian product of three word n-gram ranges $(1,1),(1,2),(1,3)$, two penalties (ℓ_1 and ℓ_2), and four regularization values $\alpha \in \{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$. Each candidate uses a `HashingVectorizer` with 2^{18} features, non-negative hashing signs, ℓ_2-normalized features, lowercase tokenization, and the token pattern `(?u)\b\w+\b`. Candidate probabilities are produced by an averaged `SGDClassifier` with logistic loss, an intercept, at most 2000 iterations, and tolerance 10^{-3}. The support seed deterministically shuffles the 24 recipes and is also used in the classifier random state (support seed plus candidate index). Thus each of the $K = 5$ pre-specified seeds defines one deterministic candidate-bank realization; all source-exclusion branches reuse that realization without changing candidate identity or ordering.

The construction recipe is: (1) start from the appropriate gold-labeled fit split described above; (2) fit the 24 candidate predictors using only gold labels and input text; (3) compute candidate predictions on the source and test rows; (4) fix the candidate set before any source-exclusion branch is computed; and (5) use weak-source labels only to form evidence losses. No candidate is pruned using SESR outcomes or reported metrics.

B.2 Dataflow and label usage

The information flow is

$$\begin{aligned}&\text{gold-labeled fit data} \to \text{candidate bank},\\ &\text{weak-source LF outputs} \to \text{evidence terms},\\ &\text{test candidate probabilities} \to \mathrm{PS}_{\max},\\ &\text{test labels and probabilities} \to \text{Accuracy/NLL diagnostics}.\end{aligned}$$

Test labels are not used for candidate fitting, evidence construction, optimization, or coupling selection. The label-free $\mathrm{PS}_{\max}$ uses test-set candidate probabilities but not test labels. Excluding source s removes only its complete evidence-loss term; candidate identity, ordering, predictions, and support space remain fixed.

B.3 Dataset statistics

Table 4: Dataset statistics for the fixed-support evaluation. $N = 24$ candidate support members are used per support realization, and $K = 5$ support realizations are nested within each dataset.

Dataset	Train	Valid	Test	Classes	Sources	N	K
IMDb	20000	2500	2500	2	5	24	5
Yelp	30400	3800	3800	2	8	24	5
AGNews	96000	12000	12000	4	9	24	5
Spouse	22254	2811	2701	2	9	24	5
ChemProt	12861	1607	1607	10	26	24	5

B.4 Coupling selection and threshold sensitivity

For either coupled method, the normalized all-source training-risk cost is

$$C_{\mathrm{train}}(\gamma) = \widehat{R}_{\mathrm{train}}\left(w_\gamma^{(0)}\right) - \widehat{R}_{\mathrm{train}}\left(w_{\mathrm{IR}}^{(0)}\right), \tag{10}$$

where $\widehat{R}_{\mathrm{train}}$ is the normalized linear evidence risk without the KL term.

The coupling grid is $\{0, 0.03, 0.1, 0.3, 1, 3, 10\}$. SESR and AC are selected separately by calls to the same selection function. For each method, it chooses the smallest γ such that the unsquared training-side $\mathrm{WS}_{\max}$ is at most 0.75 times the IR value and the normalized all-source training-risk cost in Equation 10 is at most 0.02. If no γ satisfies both conditions, it uses the cost-feasible γ with minimum unsquared training-side $\mathrm{WS}_{\max}$; if no grid value is cost-feasible, it uses $\gamma = 0$. Selection

uses training-side information only and never reads $\mathrm{PS}_{\max}$, test labels, Accuracy, NLL, or another held-out metric. AC selects $\gamma = 10$ in all 25 instances; Table 5 lists the SESR choices.

An earlier protocol specification defined the selection threshold using squared Euclidean distance, whereas the implemented selector operated on the unsquared distances returned by the evaluator. Accordingly, the objective penalty remained squared, the reported $\mathrm{WS}_{\max}$ is unsquared, and the selection threshold was applied to the unsquared quantity. Because a 25% reduction in squared distance corresponds to an L2 ratio of $\sqrt{0.75}$, rather than 0.75, we additionally evaluate the literal squared-distance convention in Appendix C.3. No test metric was used for selection under either convention.

Table 5: Selected SESR γ values.

Dataset	Selected γ values
IMDb	3, 3, 10, 10, 3
Yelp	3, 3, 10, 3, 3
AGNews	10, 10, 10, 10, 10
Spouse	3, 3, 3, 3, 3
ChemProt	3, 3, 3, 3, 3

B.5 Evidence loss and implementation details

For source s, let $\lambda_s(x)$ be its non-abstaining label on example x, let $\mathcal{T}_s$ be the training examples on which it does not abstain, and let $m_s = |\mathcal{T}_s|$. With $\epsilon = 10^{-6}$, the raw candidate-level loss is

$$\bar{\ell}_{s,i} = \frac{1}{m_s} \sum_{x \in \mathcal{T}_s} -\log\Big(\mathrm{clip}_{[\epsilon,1-\epsilon]}\left[h_i(\lambda_s(x) \mid x)\right]\Big).$$

Let $a_s = \max\{10^{-6}, \mathrm{median}_{i:\bar{\ell}_{s,i}>0}\, \bar{\ell}_{s,i}\}$ and $\widetilde{\ell}_{s,i} = \bar{\ell}_{s,i}/a_s$. The implementation then sets a block-wide scale b to the median of all positive $\widetilde{\ell}_{s,i}$ values and uses $\widehat{\ell}_{s,i} = \widetilde{\ell}_{s,i}/b$ in Equation 2. Excluding source s removes its complete normalized evidence-loss term from the active-source average and changes the denominator from S to $S-1$. The a_s, b, candidate models, and predictions are not recomputed after exclusion.

We set $\lambda_{\mathrm{KL}} = 0.1$. IR uses the exact entropic closed-form optimizer for the linear empirical loss plus the KL term. The coupled AC and SESR branches use deterministic exponentiated-gradient mirror descent initialized at the independent solutions, with 35 evaluated pre-update iterates and learning rate $0.06/\sqrt{t}$. For SESR, branches whose squared displacement is within 10^{-12} of the maximum share the active maximum subgradient equally. Exponentiated updates are normalized by softmax; the solver returns the pre-update iterate with the smallest objective (earliest iterate on exact ties). At $\gamma = 0$, both coupled methods return IR directly. The fixed 35-step budget means the coded early-stopping check (active only after step 40) cannot trigger.

The five support seeds are 2026090101, 2026090102, 2026090103, 2026090104, and 2026090105.

C Full Experimental Results and Diagnostics

C.1 Support-level results

Table 6 records the SESR results for each support realization. The dataset-level means appear in Tables 1 and 2.

Table 6: SESR results for each support realization, using the endpoint definitions in Section 3.3.

Dataset	Support realization ID	$\mathrm{PS}_{\max}$	$\mathrm{WS}_{\max}$	γ
IMDb	2026090101	0.035264	0.095828	3
IMDb	2026090102	0.041136	0.105119	3
IMDb	2026090103	0.030412	0.079587	10
IMDb	2026090104	0.031485	0.082254	10
IMDb	2026090105	0.039187	0.102804	3
Yelp	2026090101	0.024915	0.060659	3
Yelp	2026090102	0.025657	0.061607	3
Yelp	2026090103	0.018342	0.041729	10
Yelp	2026090104	0.025575	0.061157	3
Yelp	2026090105	0.025852	0.061795	3
AGNews	2026090101	0.003425	0.012338	10
AGNews	2026090102	0.003295	0.011827	10
AGNews	2026090103	0.003336	0.012003	10
AGNews	2026090104	0.003435	0.012213	10
AGNews	2026090105	0.003364	0.012096	10
Spouse	2026090101	0.011997	0.065058	3
Spouse	2026090102	0.012169	0.064549	3
Spouse	2026090103	0.012032	0.063994	3
Spouse	2026090104	0.012115	0.063964	3
Spouse	2026090105	0.012150	0.064861	3
ChemProt	2026090101	0.007217	0.013748	3
ChemProt	2026090102	0.006986	0.013723	3
ChemProt	2026090103	0.006983	0.013835	3
ChemProt	2026090104	0.007032	0.013774	3
ChemProt	2026090105	0.007012	0.013758	3

C.2 Task-quality and weight-concentration diagnostics

Panel B of Table 2 reports the all-source branch at the selected γ. The pre-specified training-side cost check uses normalized all-source training risk and does not use held-out labels or $\mathrm{PS}_{\max}$. Accuracy and NLL are descriptive all-source diagnostics; Kish percent, maximum weight, effective support size, and KL to the uniform prior characterize weight concentration. These quantities do not establish task-quality guarantees under source exclusion.

We additionally evaluated every LOSO branch with the selected fixed weights, without reselecting γ or changing candidate predictions. Table 7 reports dataset means of the minimum branch Accuracy and maximum branch NLL increment; Table 8 reports all-source and mean LOSO Macro-F1. SESR's aggregate minimum Accuracy is at least as high as both controls on these five dataset means, but across the 25 support realizations, 5/25 instances have a larger maximum NLL increment than IR and 4/25 have a lower mean LOSO Accuracy. Every branch's Accuracy, NLL and Macro-F1 is provided in the accompanying supplementary numerical results.

Table 7: Task quality under every single-source exclusion. Values are means over the five support realizations within each dataset. "Min Acc" is the minimum Accuracy over the dataset's LOSO branches; "Max NLL inc." is the largest LOSO NLL minus the corresponding all-source NLL.

	Minimum LOSO Accuracy ↑			Maximum LOSO NLL increment ↓		
Dataset	IR	AC	SESR	IR	AC	SESR
IMDb	0.790960	0.794160	0.797680	+0.056910	+0.050099	+0.040518
Yelp	0.864579	0.865211	0.866053	+0.039035	+0.036540	+0.028138
AGNews	0.832333	0.832500	0.832517	+0.005610	+0.004700	+0.004634
Spouse	0.919289	0.919289	0.919289	+0.006633	+0.006280	+0.006344
ChemProt	0.680274	0.680772	0.680772	+0.010717	+0.010334	+0.011501

Table 8: Macro-F1 for all-source and mean single-source-excluded predictions, dataset means over five support realizations. Full branch-level Accuracy, NLL and Macro-F1 are provided in the accompanying numerical data.

	All-source Macro-F1			Mean LOSO Macro-F1		
Dataset	IR	AC	SESR	IR	AC	SESR
IMDb	0.812837	0.812837	0.811235	0.810342	0.811097	0.811620
Yelp	0.871738	0.871738	0.870687	0.872007	0.871981	0.872172
AGNews	0.832891	0.832874	0.832909	0.832892	0.832902	0.832891
Spouse	0.478974	0.478974	0.478974	0.482018	0.481815	0.482018
ChemProt	0.456029	0.455923	0.456778	0.455682	0.455726	0.455673

C.3 Coupling and selection-rule sensitivity

Figure 3 traces the dataset-mean stability–risk paths on the original grid using the same fixed inputs for SESR and AC. Equal numeric γ is a controlled grid comparison, not an assertion of equal effective regularization: the objectives penalize maximum and average branch displacement, respectively. At $\gamma = 0$, the methods coincide. SESR has lower $\mathrm{PS}_{\max}$ in 24/25 instances at each of $\gamma \in \{0.03, 0.1, 0.3, 1\}$ and in 25/25 at $\gamma \in \{3, 10\}$. The fixed-γ exception is IMDb support seed 2026090104 at $\gamma = 0.03, 0.1, 0.3, 1$.

Table 9: Fixed-γ SESR versus AC on $\mathrm{PS}_{\max}$ (lower/tied/higher; 25 paired supports at each grid point).

γ	0	0.03	0.1	0.3	1	3	10
Lower/tied/higher	0/25/0	24/0/1	24/0/1	24/0/1	24/0/1	25/0/0	25/0/0

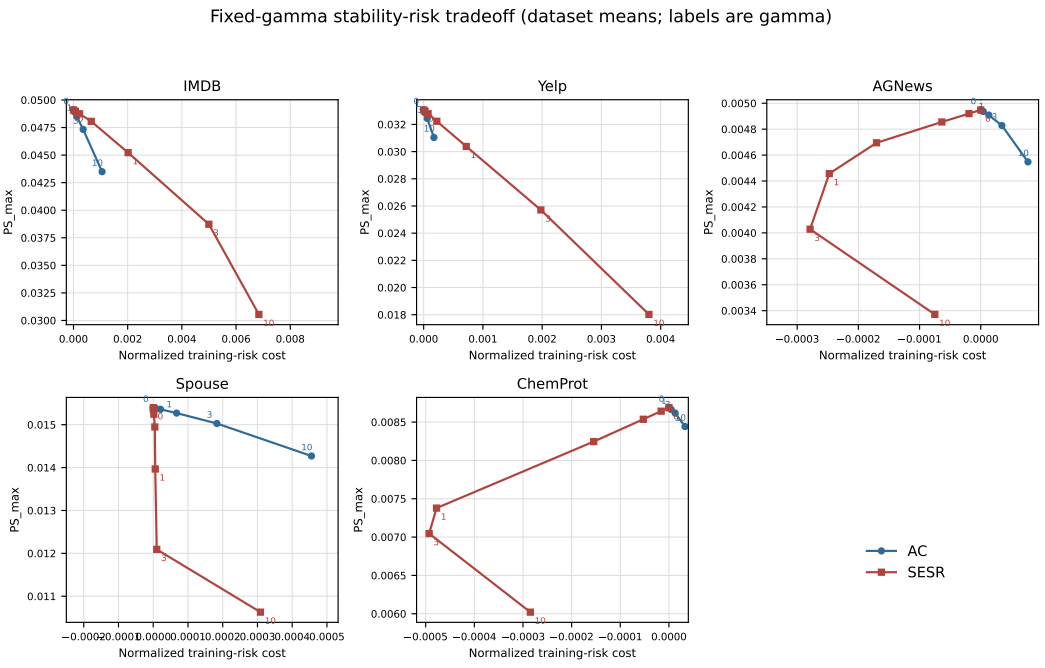

Figure 3: Fixed-γ stability–risk paths on the original seven-point grid. Points are dataset means over five support realizations and labels give γ. The horizontal axis is normalized all-source training-risk cost relative to IR; the vertical axis is $\mathrm{PS}_{\max}$. All 25 instance-level paths and task-quality diagnostics are provided in the accompanying supplementary numerical results.

Table 10 compares the implemented selector with a literal application of the squared-distance convention. AC selects the grid upper bound in every instance under both rules. The squared-distance convention changes SESR's selected γ in 19/25 instances and introduces two reversals against AC, both on IMDb: seeds 2026090101 and 2026090104. Nevertheless, SESR's dataset-mean $\mathrm{PS}_{\max}$ remains lower for all five datasets. Under the selected configurations, lower SESR $\mathrm{PS}_{\max}$ than AC coexists with worse all-source Accuracy in 11/25 instances, worse all-source NLL in 15/25, lower minimum LOSO Accuracy in 1/25, and a larger maximum LOSO NLL increment in 12/25. These cases preclude interpreting stability improvement as a task-quality guarantee.

Table 10: Sensitivity to the γ-selection rule on the 25 fixed instances. "Lower/tie/higher" compares selected SESR with selected AC on $\mathrm{PS}_{\max}$. The squared-L2 rule applies the historical written threshold literally rather than converting it to an equivalent unsquared threshold.

Selection rule	AC selected γ	SESR selected γ	Lower/tie/higher	Dataset means lower
Implemented unsquared L2	10 (25)	3 (17), 10 (8)	25/0/0	5/5
Historical squared L2	10 (25)	1 (12), 3 (13)	23/0/2	5/5

D External Adaptations and Source-Structure Controls

D.1 Fixed-support source-group DRO adaptation

Published Group DRO trains neural parameters against worst-group risk (Sagawa et al., 2020). Here an LF is a group and only the weights over the already trained 24 candidates change. For each active LF set A, we independently solve

$$\min_{w\in\Delta_{24}} \max_{s\in A} \widehat{L}_s(w) + 0.1\,\mathrm{KL}(w\|\pi_0).$$

An all-source problem and one fresh problem for each excluded LF share the fixed candidate predictions, normalized loss matrix, and test set; no joint branch coupling or test-label tuning occurs. The primary KL coefficient matches IR/AC/SESR; coefficients 0.05 and 0.2 are sensitivity checks only. Convex epigraph optimization is independently checked against the analytic source-simplex dual. This is a fixed-support source-group DRO adaptation, not the published neural training pipeline. Test labels are evaluation-only.

D.2 WeShap-based fixed-support adaptation

Published WeShap estimates LF Shapley contributions to validation probabilistic accuracy for a majority-vote/kNN proxy (Guan & Koudas, 2024); it does not learn SESR mixture weights. Our adaptation computes its vote-Shapley recurrence on fixed training LF votes and gold validation labels, with $k = 10$, Euclidean uniformly weighted neighbors, and stateless hashed unigram text features rather than the pretrained text representations used in the published evaluation. This validation supervision is extra information compared with IR/AC/SESR/DRO source weighting (Spouse validation labels also trained its candidates). Each dataset's source values ϕ_s are shared across its five supports. We freeze $q_s = \max(\phi_s, 0)$, set $\alpha_s = q_s / \sum_j q_j$ (uniform if every $q_j = 0$), and independently minimize

$$\sum_{s\in A} \alpha_s^{(A)} \widehat{L}_s(w) + 0.1\,\mathrm{KL}(w\|\pi_0)$$

on every active set A, renormalizing surviving positive mass (uniform fallback if zero). Neither source scores nor candidates are refit after deletion. This nonnegative-score-to-weight map and downstream fixed-support solve are our adaptation, not the original WeShap end-to-end pipeline. Full signed source values, including nonpositive scores clipped by this rule, accompany the source data. In IMDb the raw values for LF01–LF05 are respectively 0.00017778, 0.04133889, 0.00566222, 0.00350778, and 0.00004667; LF02 is highest and LF05 lowest. On Spouse only LF05–LF07 have positive values. No test labels were used for valuation, mapping or reweighting.

D.3 Paired external results

Table 11 records all 25 matched support-level $\mathrm{PS}_{\max}$ values for the two adaptations alongside unchanged IR/AC/SESR. All-source NLL is higher for adapted DRO than SESR in 25/25 supports

and lower for the WeShap adaptation in 20/25; hence PS alone is not a complete method comparison. The complete per-source and task-quality records are supplied in the companion numerical tables.

Table 11: Worst-Case Prediction Shift ($\mathrm{PS}_{\max}$, lower is better) by matched support. ID 01–05 abbreviates seeds 2026090101–2026090105. DRO and WeShap are fixed-support adaptations.

Dataset	ID	IR	AC	SESR	Source-group DRO	WeShap-based
IMDb	01	0.046794	0.040996	0.035264	0.007774	0.051126
IMDb	02	0.055351	0.048565	0.041136	0.009712	0.067411
IMDb	03	0.048548	0.043264	0.030412	0.006628	0.058960
IMDb	04	0.042633	0.038624	0.031485	0.020110	0.065336
IMDb	05	0.052154	0.046041	0.039187	0.013460	0.057201
Yelp	01	0.032068	0.030107	0.024915	0.017308	0.017111
Yelp	02	0.032925	0.030901	0.025657	0.006256	0.017291
Yelp	03	0.034038	0.031965	0.018342	0.012416	0.017701
Yelp	04	0.032897	0.030902	0.025575	0.019595	0.017545
Yelp	05	0.033394	0.031330	0.025852	0.015783	0.017689
AGNews	01	0.005033	0.004631	0.003425	0.021836	0.005516
AGNews	02	0.004821	0.004423	0.003295	0.014879	0.005568
AGNews	03	0.004846	0.004449	0.003336	0.026082	0.005715
AGNews	04	0.004997	0.004599	0.003435	0.026419	0.005548
AGNews	05	0.005043	0.004636	0.003364	0.023338	0.005552
Spouse	01	0.015229	0.014080	0.011997	0.004583	0.002972
Spouse	02	0.015536	0.014376	0.012169	0.005507	0.003920
Spouse	03	0.015346	0.014231	0.012032	0.007423	0.002905
Spouse	04	0.015385	0.014320	0.012115	0.006371	0.002508
Spouse	05	0.015491	0.014345	0.012150	0.005911	0.002921
ChemProt	01	0.008684	0.008438	0.007217	0.016810	0.013989
ChemProt	02	0.008734	0.008488	0.006986	0.011751	0.014626
ChemProt	03	0.008704	0.008457	0.006983	0.007412	0.012204
ChemProt	04	0.008623	0.008377	0.007032	0.008713	0.013667
ChemProt	05	0.008695	0.008447	0.007012	0.023416	0.012207

D.4 Structural interventions and per-source cases

The Spouse loss-space partition fixed from candidate-aligned source-loss profiles is C1={LF05, LF06, LF07} and C2={LF01, LF02, LF03, LF04, LF08, LF09}; a separate behavior-based partition also preserves C1. The cluster control averages source risks within each cluster, then weights clusters equally with the same KL coefficient. After deleting one LF, assignments stay fixed and the surviving members are re-averaged before independently solving the branch. Its five-seed mean single-source $\mathrm{PS}_{\max}$ is 0.003432 versus SESR's 0.012093; all-source NLL is 0.244960 versus 0.249252 and Accuracy ties at 0.919289. Removing all C1 instead gives the independently re-solved cluster control $\mathrm{PS}_{-\mathrm{C1}} = 0.118734$ and NLL increase +0.084966. We did not evaluate whole-cluster removal for SESR or AC.

IMDb LF02 provides 70.19% of non-abstaining votes and 78.46% of normalized evidence, but LF05 (0.83% of votes) sets the worst IR, AC, and SESR shift on each support. Across the five seeds, IR's LF02/LF05 shifts are 0.009057/0.049096 and SESR's are 0.005570/0.035497. SESR locally increases the LF01/LF03/LF04 shifts versus IR even while reducing the maximum. Figure 2 shows all IMDb source deletions, not just the worst one. LF02 also has unique fit-side coverage and its exclusion worsens Accuracy/NLL, so small drift alone is not evidence of preserving information.

On ChemProt, the coverage-matched LF07/LF08 pair has Jaccard overlap 0.06723 and full agreement on 174 co-covered rows, while LF08/LF17 has Jaccard 0.01451 and disagrees on all 33 co-covered rows. Their coverage is 11.59%, 9.88%, and 8.06% (LF07/LF08/LF17), respectively. IR's LF08 and LF17 shifts are 0.001061 and 0.003725; this contrast does not identify overlap separately from agreement. SESR reduces LF07 shift from 0.001198 to 0.000537 but increases LF08 from 0.001061 to 0.001754 and LF17 from 0.003725 to 0.004346. Those are local, non-worst deletions: the global ChemProt $\mathrm{PS}_{\max}$ falls from IR 0.008688 to SESR 0.007046. The even more overlapping LF16/LF17 pair conflicts in label assignments, further precluding a simple overlap law.