Agreement without Coverage: Conditional Agreement is Non-Identifying in Variable-Support Structured Extraction

Yanru Zhou
School of Cyberspace Security
Beijing University of Technology
Beijing, China
zhouyanru@bjut.edu.cn

Dandan Song*
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
sdd@bit.edu.cn

Abstract

Conditional agreement between paired predictions is widely used as a reliability proxy in structured prediction and extraction. We show that the inference from agreement to reliability can fail in a specific and practically relevant regime: thresholded variable-support prediction, where the support on which agreement is evaluated is itself determined by the predictor. In this regime, conditional agreement is non-identifying: a predictor can strictly increase conditional agreement while simultaneously degrading the support-aware witnesses: Recall, Cardinality, and Evaluability Coverage. We formalize this failure through predictor-dependent evaluability events, which make the evaluated population an endogenous function of the predictor. A strict witness theorem gives explicit sufficient conditions under which a support-shrinking predictor ranks above a coverage-preserving one. A separate union-nonempty/Jaccard construction establishes fixed-agreement fibers with different Recall, Cardinality, and Evaluability Coverage. We instantiate the failure on Re-DocRED as a controlled high-risk stress test: a uniform consistency objective raises conditional agreement from 0.69 to 0.88 while Recall falls by 16 points, predicted support shrinks by 20%, and silent predictions increase by 68%. Contract-aligned three-seed diagnostics and directional overlap analyses confirm shrinking evaluability denominators. The same failure pattern appears across three ATLOP seeds at an intermediate consistency strength on Re-DocRED. We recommend reporting conditional agreement alongside its support-aware witnesses (Recall, Cardinality, Silence, and Evaluability Coverage) rather than treating it as a standalone reliability signal.

1 Introduction

A model can raise an agreement score by refusing to emit—and hence never being evaluated on—some of its hardest predictions. For example, retaining only four easy relations rather than ten difficult relations can increase paired agreement despite omitting potentially correct facts. A standard reliability heuristic in structured prediction and extraction is to check whether two aligned views of the same input produce the same output. Such agreement is often taken as evidence of robustness. This logic underlies counterfactual consistency in supervised extraction; related agreement and validity proxies also appear in LLM-era pipelines through schema-constrained decoding, self-consistency, verifier-style filtering, and self-refinement [18, 9, 26, 15]. We ask a more precise question: does conditional agreement identify the support-aware behavior that practitioners actually care about?

*Corresponding author: Dandan Song (sdd@bit.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This question arises whenever a set-valued system treats agreement or validity on its emitted support as reliability evidence.

Our answer is negative, but within a specific and delimited regime: thresholded, variable-support structured prediction with predictor-dependent evaluability. In this regime, the predictor does not emit a single fixed label; it produces a support set whose cardinality is itself a model choice. When the evaluability event depends on the emitted support, the predictor can improve conditional agreement by shrinking its predicted support. Fragile positives can be suppressed below threshold in both views simultaneously, making the evaluated subset easier while Recall, Cardinality, and Evaluability Coverage all deteriorate. The failure is therefore not merely noisy agreement: the population on which agreement is averaged changes with the predictor's own emitted support.

The result concerns variable-support predictors whose evaluation contracts allow low-agreement cases to leave the evaluated population: under such contracts, conditional agreement alone cannot distinguish a coverage-preserving predictor from a support-shrinking one. We study this failure on Re-DocRED, examine its behavior with ATLOP, and use biomedical and sentence-level extraction tasks to characterize its empirical boundaries.

The empirical paradox is stark. On Re-DocRED, adding a uniform counterfactual-consistency objective to a factual-only baseline (Uniform-0.5 versus No_CF; Section 5) leaves raw agreement nearly unchanged while raising J_{cond} from 0.692 to 0.877. At the same time, Silence rises by 68%, Cardinality drops by 20%, and Recall falls by 16 points. Small-budget and normalized support-recovery controls partially recover predicted support while maintaining high J_{cond}, indicating that the operationally relevant variable is not consistency in the abstract, but the support-shrinkage pressure induced by the training objective.

Fixed-threshold and matched-cardinality controls show that the observed Recall loss cannot be explained by threshold selection or output size alone; margin audits and support-state decomposition show that the support state itself shifts under consistency pressure. Unlike explicit reject-option learning, the setting has no reject head and no reported risk–coverage curve: abstention is hidden inside the thresholded structured output.

Contributions. (1) Under an explicit removable-low-agreement-state condition, we prove a strict population-level witness theorem: J_{cond} can rank a support-shrinking predictor strictly above a coverage-preserving one while Recall, Cardinality, and Evaluability Coverage all simultaneously worsen (Theorem 1). For the specific union-nonempty/Jaccard construction, a support-shrinkage fiber and a quantitative level-set corollary establish that equal or maximal J_{cond} can mask substantial variation in Recall, Cardinality, and Evaluability Coverage (Corollaries 1–2). (2) We identify a local support-shrinkage mechanism under consistency pressure that provides a sufficient account of how training can reach the failure mode (Section 3). (3) We instantiate the failure on Re-DocRED with a suite of controls—fixed-threshold replay, matched-cardinality replay, margin audits, and support-state decomposition—that rule out pure threshold-artifact explanations (Section 5). (4) We report a diagnostic witness set rather than an aggregate score, and use additional settings to delimit the regime of applicability (Section 6).

2 Setup and Non-Identifiability

Let x be an input instance and $\tilde{x}$ an aligned view of the same underlying instance. Let $\mathcal{A}(x)$ be the finite candidate atom space and let $Y(x) \subseteq \mathcal{A}(x)$ be the gold support. In document-level relation extraction, an atom is a relation triple $a = (e_h, e_t, r)$ and $Y(x)$ is the gold triple set [28, 25]. A thresholded set predictor P assigns scores $p_P(a \mid x) \in [0, 1]$ to candidate atoms and emits the predicted support

$$S_P(x) = \{a \in \mathcal{A}(x) : p_P(a \mid x) \geq \tau\}.$$

Conditional agreement as a selected-support proxy. The J_{cond} notation below is an abstraction rather than a literal restatement of any specific prior counterfactual-consistency metric. CovEReD reports a directional pairwise consistency score (CONS), computed on counterfactual triples whose factual counterparts are predicted correctly, following Paranjape et al. [22]; we use J_{cond} as a symmetric set-valued abstraction of this family of conditional agreement proxies [18]. Let $E_P(x, \tilde{x})$ denote the predictor-dependent evaluability event on which conditional agreement is conditioned. In

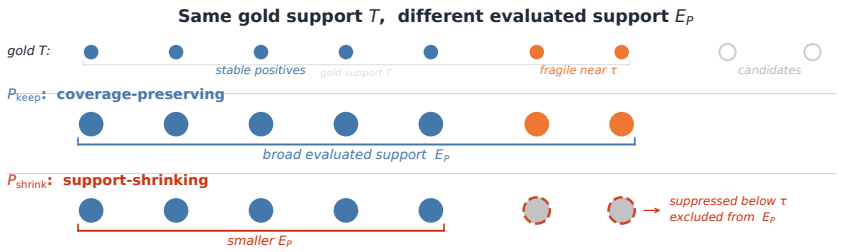


Figure 1: **Conditional agreement is computed on predictor-selected support.** The gold support T is fixed, but the evaluated support E_P depends on the predictor. P_{keep} preserves fragile positives in E_P, whereas P_{shrink} pushes them below threshold and excludes them from E_P. Consequently, J_{cond} can increase or remain tied while Recall, Cardinality K, and Evaluability Coverage C worsen (Theorem 1 and Corollaries 1–2); empirically, Silence rises as well.

CovEReD-style evaluation, E_P may correspond to factual-side correctness; in set-valued support audits, it may correspond to non-empty factual-side support or non-empty paired support. Given a per-pair agreement score $A_P(x, \tilde{x}) \in [0, 1]$, define

$$J_{\text{cond}}(P) = \mathbb{E}[A_P(x, \tilde{x}) \mid E_P(x, \tilde{x}) = 1] \,.$$

In our set-valued audits, $A_P(x, \tilde{x})$ is the Jaccard similarity between $S_P(x)$ and $S_P(\tilde{x})$ under the support-based evaluability contract. We do not claim that prior work literally optimizes J_{cond} as defined here. The theorem isolates this risk only for contracts satisfying the removable-low-agreement-state condition below; endogenous selection by itself does not imply the strict ranking.

For support-aware witnesses, define Recall, Cardinality, and Evaluability Coverage as

$$R(P) = \mathbb{E}\left[\frac{|S_P(x) \cap Y(x)|}{|Y(x)|} \,\middle|\, |Y(x)| > 0\right], \qquad K(P) = \mathbb{E}[|S_P(x)|],$$
$$C(P) = \Pr\left(E_P(x, \tilde{x}) = 1\right),$$

where $C(P)$ is the probability that a paired instance falls within the evaluability event—that is, the fraction of pairs that contribute to the J_{cond} average. Silence counts, over the gold-positive (entity-pair, relation) label slots of the aligned document pairs, the slots whose entity pair receives no emitted relation in either view; it is a slot count, not a document count, and is not the complement of Cardinality.

Support-selected contracts. A conditional-agreement contract is *support-selected* if its evaluability event $E_P(x, \tilde{x})$ is a nonconstant measurable function of the emitted paired supports $S_P(x)$ and $S_P(\tilde{x})$. We use "non-identifying" in a witness-pair sense: over this predictor class, knowing $J_{\text{cond}}(P)$ does not determine the support-aware tuple $(R(P), K(P), C(P))$. This is narrower than a proper-scoring-rule or elicitation-complexity theorem, but stronger than an empirical correlation failure. The failure mechanism is a selected-support shift: changing the predictor changes the population over which J_{cond} is averaged. The abstraction is relevant to paired structured extractors and may apply to LLM-era pipelines only when the stated sufficient conditions hold; for example, when a schema-constrained decoder, a self-consistency filter, or a verifier exposes only non-empty or accepted structured outputs. The theorem concerns the support-selected contract, not the training mechanism that produced the predictor, and it is not a monotone tradeoff between agreement and Recall: a level set of J_{cond} can have nontrivial diameter in support-aware (R, K, C) coordinates.

Theorem 1 (Strict witness under removable low-agreement mass)**.** *Consider a predictor-dependent agreement contract (E, A) on decoded paired supports with $A \in [0, 1]$. Suppose it admits three feasible paired support states z_e, z_h, z_s (with compatible fixed gold supports): (i) z_e is evaluable and $A(z_e) = a_e$; (ii) z_h is evaluable with $A(z_h) = a_h < a_e$; (iii) z_s is obtained from z_h by deleting emitted support, is unevaluable, and has strictly lower focal-view Recall and Cardinality. Suppose the retained and suppressed predictors can realize the same easy state z_e and the respective hard states z_h, z_s on a common paired evaluation distribution. Then, for every hard-branch mass $\alpha \in (0, 1)$, there exist thresholded set predictors P_{keep} and P_{shrink} such that*

$$J_{\text{cond}}(P_{\text{shrink}}) > J_{\text{cond}}(P_{\text{keep}}),$$

while simultaneously

$$R(P_{\text{shrink}}) < R(P_{\text{keep}}), \quad K(P_{\text{shrink}}) < K(P_{\text{keep}}), \quad C(P_{\text{shrink}}) < C(P_{\text{keep}}).$$

Thus a contract with removable low-agreement evaluable mass cannot use conditional agreement alone to identify these support-aware quantities. Support-dependence or nonconstancy of E alone is insufficient.

Corollary 1 (A support-shrinkage fiber at fixed agreement)**.** *For the union-nonempty evaluability event and Jaccard agreement constructed in Appendix A, there exists a one-parameter family of thresholded set predictors $\{P_t\}_{t\in[0,1]}$ such that $J_{\text{cond}}(P_t)$ is constant in t, while $R(P_t)$, $K(P_t)$, and $C(P_t)$ vary nontrivially. Exact equality of J_{cond} therefore does not identify Recall, Cardinality, or Evaluability Coverage.*

Corollary 2 (Quantitative level-set diameter)**.** *In the construction of Appendix A, fix hard mass $\alpha \in (0,1)$ and stable-support size $|C| = m \geq 1$. Let $P_0 = P_{t=0}$ and $P_1 = P_{t=1}$ be the two endpoints of the fiber in Corollary 1. With $K(P) = \mathbb{E}[|S_P(x)|]$ denoting per-view raw cardinality as above,*

$$J_{\text{cond}}(P_0) = J_{\text{cond}}(P_1) = 1,$$

but

$$R(P_1) - R(P_0) = \alpha\frac{m}{m+1}, \qquad K(P_1) - K(P_0) = \alpha m, \qquad C(P_1) - C(P_0) = \alpha.$$

The dimensionless (R, C) projection alone has diameter at least

$$|R(P_1) - R(P_0)| + |C(P_1) - C(P_0)| = \alpha\left(1 + \frac{m}{m+1}\right) \geq \frac{3}{2}\alpha.$$

Thus a maximal-J_{cond} level set can contain predictors separated by at least $\frac{3}{2}\alpha$ in dimensionless (R, C) coordinates; the raw cardinality difference αm is reported separately, since its scale depends on the candidate universe.

Proof sketch and scope. Put mass $1-\alpha$ on an easy state z_e and mass α on a lower-agreement hard state z_h. The support-shrinking predictor replaces z_h by the deleted, unevaluable state z_s. Consequently,

$$J_{\text{cond}}(P_{\text{keep}}) = (1-\alpha)a_e + \alpha a_h < a_e = J_{\text{cond}}(P_{\text{shrink}}),$$

while C changes from 1 to $1-\alpha$, and the assumed strict hard-branch gaps reduce focal-view Recall and Cardinality. Appendix A gives the proof and an explicit union-nonempty/Jaccard realization. The fixed-agreement fiber in Corollaries 1–2 holds for that specific realization, not for every contract covered by Theorem 1. A factual-correctness-conditioned directional CONS is a different estimand and is evaluated separately.

3 A Support-Shrinkage Mechanism and Witness Set

Theorem 1 is population-level: it shows that a support-shrinking predictor class exists, but not how training reaches it. We now give a sufficient local mechanism showing how a rare positive atom near the decoding threshold can be driven below it under consistency pressure. This mechanism is intentionally local. It is not a global convergence theorem for neural training; it is a threshold-crossing account of how the P_{shrink} branch can arise.

For a candidate atom a, let π_a denote its effective supervised mass, ρ_a its exposure under the consistency objective, $\mu_a > 0$ its mean positive evidence, and σ_a^2 its cross-view score variance. Let w_a be a local gain parameter controlling the strength with which the model emits this atom across views. A second-order approximation of the per-atom regularized risk is

$$\mathcal{R}_a(w_a) = \pi_a\left[(1-\mu_a w_a)^2 + \sigma_a^2 w_a^2\right] + 2\lambda\beta_a\rho_a\sigma_a^2 w_a^2,$$

where λ is the global consistency weight and β_a is the consistency budget allocated to atom a. Minimizing over w_a gives

$$w_a^*(\lambda, \beta_a) = \frac{\pi_a\mu_a}{\pi_a(\mu_a^2 + \sigma_a^2) + 2\lambda\beta_a\rho_a\sigma_a^2}.$$

Since the denominator strictly increases with the effective consistency budget $\lambda\beta_a\rho_a$, w_a^* decreases as consistency pressure grows. Solving

$$\bar{p}_a(\lambda, \beta_a) \equiv \mu_a w_a^*(\lambda, \beta_a) = \tau$$

for the transition budget gives

$$\lambda_a^* = \frac{\pi_a}{2\beta_a\rho_a\sigma_a^2}\left(\frac{\mu_a^2}{\tau} - \mu_a^2 - \sigma_a^2\right).$$

This transition budget is positive only when the atom is initially above threshold under the no-consistency local surrogate, i.e.,

$$\tau < \frac{\mu_a^2}{\mu_a^2 + \sigma_a^2};$$

otherwise the atom is already below threshold at $\lambda = 0$. Atoms with small π_a (rare labels), large σ_a^2 (high cross-view variance), or large ρ_a (broad consistency exposure) have smaller λ_a^* and are therefore the first candidates to leave the predicted support as consistency pressure grows. The surrogate illustrates how consistency pressure can suppress near-threshold positives, reaching the P_{shrink} branch of Theorem 1. The derivation is given in Appendix D.

Relation to calibration and abstention. In single-label classification, the predicted support has fixed cardinality one, so support shrinkage is impossible. In selective prediction, coverage is made explicit through a reject option and assessed through risk–coverage curves [5, 8, 3]. Structured abstention likewise introduces explicit abstention variables or abstention-aware losses [7]. Our setting differs fundamentally: there is no reject head and no explicit coverage constraint. Instead, abstention is hidden inside the thresholded structured output. The reported J_{cond} can therefore obscure the implicit abstention induced by the training objective, making support shrinkage invisible when conditional agreement is read in isolation.

Diagnostic witness set. We operationalize the non-identifiability result through a coverage-aware diagnostic witness set: raw agreement J, conditional agreement J_{cond}, Recall, Cardinality, Silence, and Evaluability Coverage C. Each quantity is either standard in the literature or directly induced by the evaluation contract; none is introduced as a standalone methodological contribution. The contribution is the theorem-grounded reporting discipline: these quantities must be read jointly, since only their combination separates the predictor classes that J_{cond} alone cannot distinguish. We therefore report the full witness tuple rather than collapsing it into a single aggregate score: aggregation would hide precisely the support-state changes that Theorem 1 identifies as the failure mode.

4 Regime of Applicability and Stress-Test Design

Regime of applicability. Theorem 1 isolates a structural risk for support-selected conditional agreement contracts admitting removable low-agreement mass. Empirically, the full support-shrinkage witness is expected to be most visible in a high-risk regime characterized by five conditions: (R1) *variable-support set-valued output*, where the predictor emits a subset whose cardinality is not fixed by the input; (R2) *thresholded decoding*, so the emitted support is monotone in a per-candidate score threshold; (R3) *sparse, long-tailed positive support*; (R4) *paired evaluation with predictor-dependent conditioning*, where the objective is computed over pairs $(x, \tilde{x})$ and conditions on an evaluability event $E_P(x, \tilde{x})$ that depends on the predictor's own output; and (R5) *optimization pressure on the conditional objective* without an explicit support- or coverage-side constraint. Fixed-support classification, single-output extraction, unconditioned agreement reporting, and objectives with explicit coverage constraints fall outside this high-risk regime.

Why Re-DocRED is the primary stress test. Re-DocRED with CovEReD-style counterfactual training is our primary controlled stress test: it sharply satisfies all five conditions and therefore sits squarely in the high-risk regime. The remaining settings are not presented as additional replications of the main witness. They delimit where the high-risk conditions weaken and where the failure mode attenuates or changes form. Throughout, the *full signature* means that J_{cond} increases while Recall, emitted support, and Evaluability Coverage all decrease relative to the baseline. ATLOP provides a

Setting	R1 support	R2 thresh.	R3 tail	R4 paired cond.	R5 pressure	Role
Re-DocRED/BERT	yes	yes	strong	yes	yes	main stress test
Re-DocRED/ATLOP	yes	yes	strong	yes	yes	cross-backbone witness/boundary
BC5CDR TitleDrop	yes	yes	task-dep.	yes	yes	seed-unstable (1/3 seeds)
CDR/BC5CDR prior probe	yes	yes	weak	partial	yes	biomedical boundary
Re-TACRED	partial	yes	weak	partial	yes	sentence-level boundary
Fixed-support tasks	no	no/limited	task-dep.	task-dep.	task-dep.	outside-regime contrast
LLM structured extraction	yes	implicit	task-dep.	task-dep.	no	motivation only

Table 1: **Regime map.** R1–R5 denote: variable-support set-valued output, thresholded decoding, sparse/long-tailed positive support, paired evaluation with predictor-dependent conditioning, and optimization pressure on the conditional objective without an explicit coverage constraint. The first row is the primary replicated witness; ATLOP is a cross-backbone witness. BC5CDR TitleDrop is a separate independent-dataset protocol whose full signature holds in one of three seeds; it is not a reinterpretation of the prior biomedical boundary probe.

Config	F1	Recall	J	J_{cond}	Silence ↓	Cardinality ↑
No_CF	0.4017 ± 0.0060	0.5378 ± 0.0248	0.4550 ± 0.0325	0.6924 ± 0.0258	5548.7 ± 401.1	13626.7 ± 571.3
Uniform-0.5	0.3434 ± 0.0093	0.3771 ± 0.0219	0.4561 ± 0.0185	0.8772 ± 0.0580	9324.7 ± 355.0	10882.3 ± 705.2
Constant-0.10	0.3774 ± 0.0033	0.4555 ± 0.0030	0.5111 ± 0.0055	0.8820 ± 0.0006	7297.7 ± 161.2	11695.7 ± 133.8
B_global_minmax	0.3886 ± 0.0097	0.4575 ± 0.0306	0.5106 ± 0.0260	0.8767 ± 0.0023	7296.3 ± 690.2	11983.7 ± 622.7

Table 2: **Three-seed Re-DocRED endpoint witness.** Under Uniform-0.5, J_{cond} rises while Silence increases and Recall and Cardinality fall. Constant-0.10 and B_global_minmax are support-recovery controls. Values are mean$\pm$std over seeds 13, 21, and 42. Cardinality and Silence are computed over the 16,576 gold-positive (entity-pair, relation) label slots of the aligned document pairs: Cardinality sums, over these slots, the number of relations emitted for the slot's entity pair in the factual view, and Silence counts the slots whose entity pair receives no emitted relation in either view.

three-seed cross-backbone witness at an intermediate consistency strength on the same Re-DocRED dataset; stronger uniform consistency still yields a boundary case where support collapses without an agreement gain. A separate BC5CDR TitleDrop paired-view experiment shows the full signature in one of three seeds under a unified rerun and is therefore a seed-unstable independent-dataset observation. CDR/BC5CDR and Re-TACRED serve as lower-support boundary checks.

5 Empirical Witness on Re-DocRED

Configurations. All Re-DocRED/BERT models share the same encoder and decoding protocol and differ only in the training objective. *No_CF* is the factual-only baseline trained with binary cross-entropy on the factual view. *Uniform-0.5* adds the counterfactual-view cross-entropy and a symmetric Bernoulli KL consistency term between the two views with a uniform per-sample weight and outer coefficient 0.5. *Constant-0.10* uses the same objective with every per-sample KL weight fixed at 0.10. *B_global_minmax* instead sets per-sample weights from the KL measured at an early (warm-up) checkpoint, winsorized at the global 1st/99th percentiles and normalized to $[0, 1]$. Exact losses and the additional allocation variants are given in Appendix B.

Main endpoint evidence. Table 2 reports the three-seed endpoint results for the primary controlled stress test. Uniform-0.5 is the primary failure witness: it sharply raises J_{cond} while simultaneously lowering Recall and shrinking the predicted support. This is the empirical counterpart of Theorem 1: conditional agreement ranks a support-shrinking predictor above a coverage-preserving baseline on the selected support on which it is evaluated. Constant-0.10 and B_global_minmax each reduce Silence and recover predicted support while maintaining high J_{cond}; we use them as support-recovery controls.

Contract-aligned coverage and directional checks. The original endpoint J_{cond} and the document-level union-nonempty Jaccard below are distinct metric implementations, so their values must not be interchanged. The additional three-seed frozen-contract analysis yields the support-aware witness in Table 3.

Setting	$J_{\mathrm{doc,union}}$	Recall	Emitted support	C_{union}	C_{both}
No_CF	0.5185 ± 0.0135	0.5618 ± 0.0307	28861 ± 2570	0.9979 ± 0.0021	0.9697 ± 0.0032
Uniform-0.5	0.7424 ± 0.0134	0.3769 ± 0.0107	21275 ± 765	0.9208 ± 0.0093	0.8836 ± 0.0093

Table 3: **Contract-aligned Re-DocRED/BERT evidence (three seeds).** Mean $\pm$ sample standard deviation across seeds 13, 21, and 42. C_{union} is the fraction of paired documents with at least one nonempty emitted support; C_{both} requires both views nonempty. Emitted support counts document-level deduplicated (h, t, r) triples summed over the 500 development documents, whereas $J_{\mathrm{doc,union}}$, C_{union}, and C_{both} are computed over the 484 aligned factual/counterfactual document pairs. These values are separate from the original endpoint table and the seed-42 support-state audit.

Audit	Config	Pred. cand.	Precision	Recall	F1
Endpoint	No_CF	28569	0.335	0.553	0.417
Endpoint	Uniform-0.5	23050	0.303	0.404	0.346
Fixed τ=0.95	No_CF	51455	0.237	0.707	0.355
Fixed τ=0.95	Uniform-0.5	566	0.792	0.026	0.050
Matched card.	No_CF	28569	0.335	0.553	0.417
Matched card.	Uniform-0.5	28569	0.268	0.443	0.334
Matched card.	Constant-0.10	28569	0.300	0.496	0.374
Matched card.	B_global_minmax	28569	0.303	0.500	0.377

Table 4: **Threshold and matched-cardinality replay on seed 42.** At matched cardinality, Uniform-0.5 still recovers fewer gold facts than No_CF.

The directional overlap $\sum_i |S_i \cap S_i'| / \sum_i |S_i|$, where S_i and S_i' are the predicted triple sets of the i-th aligned document in the factual and counterfactual view (the CovEReD-style quantity; Appendix B), rises from 0.7064 ± 0.0337 to 0.9198 ± 0.0057. A separately identified transparent CONS reimplementation also rises (0.7366 ± 0.0269 to 0.9382 ± 0.0047), although jointly correct factual–counterfactual pairs decrease (6727 ± 579 to 5876 ± 208), as the factual-correct denominator contracts (9122 ± 468 to 6262 ± 190). These directional quantities condition on factual correctness, and the symmetric proof does not transfer to them automatically. The matched-cardinality, margin left-shift, below-threshold and paired-support migration conclusions each hold qualitatively in all three seeds; the detailed seed-42 audit below provides an illustration rather than the sole evidence. In a 3×100 seed–threshold grid, $212/300$ correlated operating points exhibit the full signature under the same metric contract; they are not 300 independent replications. Unconditional variants that keep double-empty pairs in the denominator also rise on the same predictions ($J_{\mathrm{empty}=0}$: $0.5175 \pm 0.0145 \rightarrow 0.6836 \pm 0.0131$; $J_{\mathrm{empty}=1}$: $0.5195 \pm 0.0126 \rightarrow 0.7628 \pm 0.0129$), so removing the conditioning does not remove the need to report support-aware witnesses. See Appendix B.

Threshold and matched-cardinality controls. A natural alternative explanation is that the endpoint paradox is merely a consequence of threshold search or output-size differences. Table 4 therefore reports a score-trace audit designed to rule out these pure-artifact explanations. Thresholding is part of the mechanism but does not fully account for the effect: even at matched cardinality, Uniform-0.5 under-recovers gold-positive triples relative to No_CF, indicating that the support state itself has shifted. Figure 2 then localizes the mechanism: gold-positive atoms near the threshold are suppressed below τ, directly shrinking the evaluated support and shifting paired views from evaluated overlap into one-sided and silent support states.

6 Scope and Additional Settings

Cross-backbone and independent-dataset evidence. At an intermediate ATLOP consistency strength $\gamma = 0.10$ on Re-DocRED, the three-seed document-level conditional Jaccard rises from 0.521 to 0.559, 0.527 to 0.582, and 0.540 to 0.583 (seeds 42, 13, and 21), while Recall, emitted support, C_{union}, and C_{both} all decrease in each seed. This is a cross-backbone witness on the same dataset, not a cross-dataset replication. The pre-specified strength scan also identifies an important boundary: $\gamma = 0.05$ is benign and $\gamma = 0.20/0.30$ yield severe support collapse without agreement gain. Likewise, the original ATLOP Uniform-0.5 setting reduces F1 to 0.028 ± 0.012 and causes

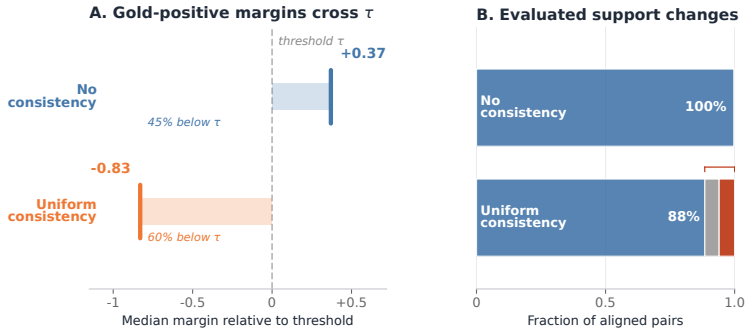


Figure 2: **Empirical mechanism of support shrinkage.** Uniform consistency raises J_{cond} while shrinking the support over which that agreement is evaluated. **(A)** Gold-positive margins cross the decoding threshold τ: the median margin moves from $+0.37$ to -0.83, and the below-τ fraction rises from 45% to 60%. **(B)** Aligned paired views decompose into evaluated-overlap, one-sided-support, and silent-pair states. No consistency yields nearly complete evaluated overlap; Uniform consistency shifts mass into one-sided and silent states, reducing both-nonempty coverage C_{both} (not the union-nonempty evaluability C_{union}). Both panels use seed 42; the three-seed endpoint changes are in Table 2. Full threshold and support-state tables are in Appendix G.

strong support collapse, but decreases its original J_{cond} from 0.750 ± 0.011 to 0.510 ± 0.063; these outcomes involve different settings/contracts and must not be conflated.

The separately precommitted *BC5CDR TitleDrop* study uses a SciBERT encoder with ATLOP, with title+abstract versus abstract-only views, and a fixed identical candidate universe. Its initial seed-42 held-out evaluation showed conditional Jaccard $0.8138 \rightarrow 0.8387$, full-view document-aware Recall $0.6923 \rightarrow 0.6332$, emitted relation memberships $1219 \rightarrow 1019$, union coverage $0.926 \rightarrow 0.922$, and silent documents $37 \rightarrow 39$ of 500, with document-aware full-view F1 $0.6459 \rightarrow 0.6475$. A subsequent unified three-seed rerun (all six arms retrained with one seed-parameterized implementation) reproduced the direction of the seed-42 result but not its exact values, and did not reproduce the full signature in seeds 13 and 21: there, conditional agreement still rose, but Recall, emitted support, and union coverage rose as well. The complete signature therefore holds in one of three seeds. We report BC5CDR TitleDrop as a possible failure case on an independent dataset, not as a seed-stable replication; it is distinct from the older benign biomedical boundary probe.

Other boundary checks. CDR/BC5CDR and Re-TACRED further delimit the scope of the claim. The earlier biomedical boundary probe has narrower output support and shows mixed or benign behavior; in the seed-42 CDR boundary probe, J_{cond} rises together with Recall while Silence falls. Sentence-level Re-TACRED shows a weaker transfer pattern in task utility: Uniform-0.5 lowers F1 relative to the baseline, while Constant-0.10 partially recovers it (Baseline 0.7699 ± 0.0043, Uniform-0.5 0.7505 ± 0.0022, Constant-0.10 0.7662 ± 0.0017). Because it does not provide the full document-level witness tuple, we use it only as lower-support transfer evidence. These outcomes sharpen the empirical regime: the observed paradox is most pronounced in long-tail, thresholded, variable-support structured extraction where strong uniform consistency can move many rare positives below the decoding threshold. Full boundary numbers are reported in Appendix H.

LLM-era structured extraction. Modern structured-output pipelines increasingly use validity-, agreement-, or verifier-style signals, including schema-constrained decoding [9, 27, 4], self-consistency [26], and self-refinement or verifier filtering [15]. These systems raise an analogous reporting question: when agreement or validity is used as evidence of reliability, is it reported together with omission, output support, and coverage? A small LLM probe is reported in Appendix H.2; it is not evidence for the supervised training mechanism studied here.

7 Related Work

Counterfactual robustness and paired-view evaluation. Counterfactual augmentation and paired-view evaluation are widely used to diagnose whether a model's predictions depend on spurious

input features. In NLP, counterfactually augmented data has been used to reveal model failures hidden by standard accuracy metrics [12]. For relation extraction, CovEReD constructs Re-DocRED counterfactual document pairs and reports a directional pairwise consistency score (CONS) following Paranjape et al. [22, 18]. Env-DocRE instead evaluates robustness to entity-name variation using F1 drops [17]. Our target is different: we do not treat conditional consistency as sufficient evidence of robustness, but rather ask whether such a proxy can identify the support-aware behavior practitioners care about.

Metric validity, property identification, and negative results. Our theorem is in the style of metric-validity and non-identification results, not metric design. Proper-scoring and property-elicitation work studies when a reported score or statistic pins down a target property of a distribution [10, 14, 6]. We use this vocabulary cautiously: our statement is a witness-pair non-separation result, not a theorem about strict scoring rules or eliciting properties from losses. High-impact negative results often identify the regime in which an appealing single criterion fails: fairness impossibility results show that natural criteria cannot jointly determine the desired behavior [13], while metric-validity work shows that BLEU can move independently of human judgment in restricted regimes [2, 23]. Our result is analogous: J_{cond} is a useful paired-view diagnostic, but cannot serve as a stand-in for support-aware evaluation in the regime of Table 1.

Selective prediction, abstention, and endogenous supports. Selective prediction makes coverage explicit through a reject option and studies the risk–coverage tradeoff [5, 8]. Structured abstention similarly introduces explicit abstention variables or abstention-aware losses [3, 7]. Our setting has no reject head: support selection is implicit in the thresholded structured output, and the conditioning event E_P depends on the predictor itself. This makes J_{cond} a selected-support quantity: the evaluated population changes with the predictor, echoing the warning from sample-selection and partial-identification work that conditional estimands need not identify their unconditional targets [16, 11].

Thresholded set prediction, coverage-aware evaluation, and proxy gaming. Multi-label and set-valued prediction work studies threshold choice, label-set utility, and calibration [19, 20], establishing that output support is a decision variable. Conformal risk control provides a constructive route to explicitly controlling coverage-sensitive losses [1]. We do not provide such a guarantee; our result is instead negative: agreement-only proxies do not imply Recall, Cardinality, Silence, or Evaluability Coverage when these quantities are left unconstrained. The failure also has a proxy-gaming interpretation [24, 21]: the proxy is conditional agreement, and the target-side witnesses are Recall, Cardinality, Silence, and Evaluability Coverage. LLM structured-output systems create an adjacent reporting risk when validity or verifier filtering selects the emitted support.

8 Limitations and Discussion

By design, the main empirical witness is Re-DocRED, the setting that most sharply satisfies the high-risk regime conditions in Table 1. ATLOP exhibits a cross-backbone witness at one strength and negative boundaries at stronger settings; on BC5CDR TitleDrop the full signature holds in one of three seeds, so the independent-dataset evidence is not seed-stable. Earlier CDR/BC5CDR and Re-TACRED probes are boundary evidence. These results do not establish universal cross-task prevalence. In practice, coverage-aware reporting can reduce overclaiming of robustness in extraction systems, whereas unreported support collapse can silently propagate omissions into downstream knowledge bases. Appendix H.1 records the final claim boundary, separating proved claims, primary empirical witnesses, and scope-setting diagnostics.

The theorem establishes non-identification for contracts admitting removable low-agreement states; it does not imply a universal tradeoff between agreement and Recall. Detailed score-trace figures remain seed-42 diagnostics, while three-seed contracted-support results and qualitative diagnostic stability are separately disclosed in Table 3 and Appendix B. Recall is a benchmark-side witness that requires labels; in unlabeled deployment, Cardinality, Silence, support-state drift, and Coverage drift remain observable alarms, but they cannot identify missing true facts and are not substitutes for Recall.

9 Conclusion

Agreement without support-state context is not a complete reliability signal when support is allowed to shrink. We proved a restricted non-identifiability result for conditional agreement under endogenous support selection and instantiated it on a controlled high-risk stress test. In this regime, agreement-style objectives can strictly increase J_{cond}—and fixed J_{cond} level sets can already vary in support-aware behavior—while Recall, Cardinality, and Evaluability Coverage deteriorate. The practical implication is to report J_{cond} alongside its support-state witnesses—Recall when labels are available, Cardinality, Silence, and contract-specified Evaluability Coverage—whenever agreement or validity proxies are used as robustness evidence, rather than collapsing the witness tuple into a single scalar.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2025YFB3003700) and the National Natural Science Foundation of China (Grant No. 62476025). The authors declare no competing interests.

References

[1] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

[2] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, 2006. URL `https://aclanthology.org/E06-1032/`.

[3] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82, 2016.

[4] X. Dong et al. Xgrammar: Flexible and efficient structured generation engine for large language models. *arXiv preprint*, 2024.

[5] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

[6] Rafael Frongillo and Ian A. Kash. Vector-valued property elicitation. In *Conference on Learning Theory*, pages 710–727, 2015.

[7] Alexandre Garcia, Slim Essid, Chloé Clavel, and Florence d'Alché Buc. Structured output learning with abstention: Application to accurate opinion prediction. In *International Conference on Machine Learning*, pages 1648–1657, 2018.

[8] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

[9] Saibo Geng et al. Generating structured outputs from language models: Benchmark and studies. *arXiv preprint arXiv:2501.10868*, 2025.

[10] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[11] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[12] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

[13] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, volume 67 of *LIPIcs*, pages 43:1–43:23, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

[14] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008. doi: 10.1145/1386790.1386813.

[15] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

[16] Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

[17] Shiao Meng, Xuming Hu, Aiwei Liu, Fukun Ma, Yawen Yang, Shuang Li, and Lijie Wen. On the robustness of document-level relation extraction models to entity name variations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16362–16374, 2024. doi: 10.18653/v1/2024.findings-acl.969.

[18] Ali Modarressi, Abdullatif Köksal, and Hinrich Schuetze. Consistent document-level relation extraction via counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11501–11507. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.672.

[19] Thomas Mortier, Marek Wydmuch, Krzysztof Dembczynski, Eyke Hüllermeier, and Willem Waegeman. Efficient set-valued prediction in multi-class classification. *Data Mining and Knowledge Discovery*, 35:1435–1469, 2021.

[20] Thomas Mortier, Eyke Hüllermeier, Krzysztof Dembczynski, and Willem Waegeman. Set-valued prediction in hierarchical classification with constrained representation complexity. In *Uncertainty in Artificial Intelligence*, 2022.

[21] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022.

[22] Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.117. URL `https://aclanthology.org/2022.acl-long.117/`.

[23] Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3): 393–401, 2018. doi: 10.1162/coli_a_00322.

[24] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, 2022.

[25] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting DocRED – addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, 2022.

[26] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[27] Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

[28] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, 2019.

A Full Proof of Theorem 1

This appendix gives the full construction and proof of Theorem 1 and Corollaries 1–2. The construction is finite and intentionally restricted: it shows that conditional agreement alone does not identify support-aware behavior when the evaluability event is selected endogenously by the predictor.

Notation. For paired views $(x, \tilde{x})$, let $S_P(x)$ and $S_P(\tilde{x})$ denote the decoded supports of predictor P. Under the support-based contract, a pair is evaluable when

$$E_P(x, \tilde{x}) = \mathbb{I}[S_P(x) \cup S_P(\tilde{x}) \neq \emptyset] .$$

When the denominator is nonzero, the per-pair set agreement is

$$J_P(x, \tilde{x}) = \frac{|S_P(x) \cap S_P(\tilde{x})|}{|S_P(x) \cup S_P(\tilde{x})|},$$

and the conditional agreement proxy is

$$J_{\mathrm{cond}}(P) = \mathbb{E}[J_P(x, \tilde{x}) \mid E_P(x, \tilde{x}) = 1] .$$

Recall $R(P)$, Cardinality $K(P)$, and Evaluability Coverage $C(P)$ are the support-aware quantities defined in Section 2. The proof below computes these quantities explicitly for the constructed predictors.

Proof of Theorem 1. Place probability $1-\alpha$ on an easy paired example with state z_e and probability α on a hard paired example. The keep predictor uses z_h on the hard example while the shrink predictor uses z_s. Both use z_e on the easy example. Evaluability is one on both easy and retained-hard examples, but zero on suppressed-hard examples. Hence

$$J_{\mathrm{cond}}(P_{\mathrm{keep}}) = (1-\alpha)A(z_e) + \alpha A(z_h) < A(z_e) = J_{\mathrm{cond}}(P_{\mathrm{shrink}}),$$

and $C(P_{\mathrm{keep}}) = 1$ while $C(P_{\mathrm{shrink}}) = 1 - \alpha$. Focal-view Recall and Cardinality remain equal on the easy mass and decrease strictly on the hard mass by hypothesis. Their population expectations therefore decrease strictly. Feasible state realizations are an explicit assumption, so this proof does not assert that nonconstancy of the event alone is sufficient. □

Explicit union-nonempty/Jaccard witness (special case). We next construct a concrete example for a support-based evaluability event

$$E_P(x, \tilde{x}) = \mathbb{I}[S_P(x) \cup S_P(\tilde{x}) \neq \emptyset] .$$

Directional factual-correctness-conditioned CONS is a distinct estimand; the construction below does not prove a strict ranking for all such directional contracts. Let $(x, \tilde{x})$ be drawn from a mixture of two subpopulations.

Easy mass. With probability $1 - \alpha$, the gold support in both views is the same non-empty stable set C with $|C| = m \geq 1$, and both predictors decode exactly C in both views. Easy examples therefore have Recall equal to one, Cardinality m, and Evaluability Coverage equal to one, and per-pair agreement one for both predictors.

Hard mass. With probability α, the gold support in both views is $C \cup \{a^\star\}$, where $a^\star$ is a rare fragile atom. We take x as the focal view used in the definitions of Recall and Cardinality, and use $\tilde{x}$ to compute paired agreement. Define P_{keep} so that on hard pairs it always decodes $C \cup \{a^\star\}$ in the focal view x, while in the aligned view $\tilde{x}$ it decodes C with probability $q > 0$ and $C \cup \{a^\star\}$ otherwise. Define P_{shrink} so that on hard pairs it suppresses $a^\star$ in both views: with probability $1 - s$ it decodes C in both views, and with probability $s \in (0, 1]$ it decodes the empty set in both views. The case $s = 1$ realizes hypothesis (iii) of Theorem 1 exactly; for $s < 1$ the evaluability of the hard branch is only partially removed, giving $C(P_{\mathrm{shrink}}) = 1 - \alpha s$, and the same strict inequalities hold.

Conditional agreement. For P_{keep}, an asymmetric hard pair (the focal view decodes $C \cup \{a^\star\}$ and the aligned view decodes C) has Jaccard overlap $m/(m+1)$, while a symmetric hard pair has overlap one. Since $q > 0$, asymmetric pairs occur with positive probability, so the hard mass strictly lowers J_{cond}:

$$J_{\text{cond}}(P_{\text{keep}}) = (1-\alpha) + \alpha\left[(1-q) + q\frac{m}{m+1}\right] = 1 - \frac{\alpha q}{m+1} < 1.$$

For P_{shrink}, all non-empty hard pairs have decoded support C in both views and thus Jaccard overlap one, while double-empty pairs are excluded by the evaluability event. Therefore

$$J_{\text{cond}}(P_{\text{shrink}}) = 1 > J_{\text{cond}}(P_{\text{keep}}).$$

This establishes the strict J_{cond} inequality in Theorem 1.

Support-aware quantities. On hard examples, P_{keep} always decodes the full gold support $C \cup \{a^\star\}$ in the focal view, so its hard-branch Recall is one and its hard-branch Cardinality is $m+1$. Thus, over the full mixture,

$$R(P_{\text{keep}}) = 1, \qquad K(P_{\text{keep}}) = m + \alpha, \qquad C(P_{\text{keep}}) = 1.$$

By contrast, P_{shrink} never recovers $a^\star$ and emits the empty set on an s-fraction of hard examples. Its hard-branch Recall is therefore $(1-s)m/(m+1)$, its hard-branch Cardinality is $(1-s)m$, and its hard-branch Evaluability Coverage is $1-s$. Over the full mixture,

$$R(P_{\text{shrink}}) = (1-\alpha) + \alpha(1-s)\frac{m}{m+1},$$

$$K(P_{\text{shrink}}) = (1-\alpha)m + \alpha(1-s)m = m(1-\alpha s),$$
$$C(P_{\text{shrink}}) = (1-\alpha) + \alpha(1-s) = 1 - \alpha s.$$

Hence

$$R(P_{\text{keep}}) - R(P_{\text{shrink}}) = \alpha\left(1 - \frac{(1-s)m}{m+1}\right) = \alpha\frac{1+sm}{m+1} > 0,$$

$$K(P_{\text{keep}}) - K(P_{\text{shrink}}) = \alpha(1+sm) > 0, \qquad C(P_{\text{keep}}) - C(P_{\text{shrink}}) = \alpha s > 0.$$

Thus P_{shrink} has strictly higher J_{cond} but strictly lower Recall, Cardinality, and Evaluability Coverage. This gives an explicit special case satisfying the support-aware inequalities in Theorem 1.

Proof of Corollary 1. Modify the hard branch of P_{shrink} as follows. For $t \in [0,1]$, define P_t to emit non-empty support C in both views on a t-fraction of hard examples, and the empty set in both views on the remaining $1-t$ fraction. All evaluable pairs have Jaccard overlap one, so

$$J_{\text{cond}}(P_t) = 1$$

for every t such that the evaluable set is non-empty. The easy branch is evaluable with probability $1-\alpha > 0$, so this condition holds for all $t \in [0,1]$. However, the support-aware coordinates vary with t:

$$R(P_t) = (1-\alpha) + \alpha t\frac{m}{m+1}, \qquad K(P_t) = (1-\alpha)m + \alpha t m, \qquad C(P_t) = (1-\alpha) + \alpha t.$$

Thus a fixed value of J_{cond} is compatible with nontrivially different (R, K, C) values, proving Corollary 1.

Proof of Corollary 2. We use the same family. On the easy branch (probability $1-\alpha$), the gold support is C and both P_0 and P_1 emit C in both views. On the hard branch (probability α), the gold support is $C \cup \{a^\star\}$: P_1 emits C in both views while P_0 emits the empty set in both views.

For every $t \in [0,1]$, every evaluable pair in the fiber has identical non-empty paired support, so $J_{\text{cond}}(P_t) = 1$. At $t = 0$ the hard branch is unevaluable, but the easy branch remains evaluable with probability $1-\alpha > 0$, so the evaluable set is non-empty and $J_{\text{cond}}(P_0) = 1$. Hence

$$J_{\text{cond}}(P_0) = J_{\text{cond}}(P_1) = 1.$$

The endpoint differences in the support-aware coordinates are

$$C(P_1) - C(P_0) = \alpha, \qquad K(P_1) - K(P_0) = \alpha m, \qquad R(P_1) - R(P_0) = \alpha \frac{m}{m+1}.$$

Because the raw cardinality coordinate depends on the size of the candidate universe, we report its αm difference separately. The dimensionless (R, C) projection has diameter at least

$$\alpha \left(1 + \frac{m}{m+1}\right) \geq \frac{3}{2}\alpha.$$

This establishes the quantitative level-set diameter bound in Corollary 2.

Selected-support interpretation. The evaluability event $E_P(x, \tilde{x})$ is a function of the predictor's emitted support. Conditional agreement is therefore computed on a predictor-selected population rather than on the full paired distribution. This is the precise sense in which J_{cond} alone cannot identify the support-aware tuple (R, K, C).

Relation to the local mechanism. The construction above is population-level and existential: it shows that the support-shrinking predictor class exists, but does not specify how training reaches it. Appendix D provides a sufficient local optimization mechanism for realizing the P_{shrink} branch: as the effective consistency budget grows, fragile positives can be driven below the decoding threshold in both views.

B Additional Evidence and Metric-Contract Clarifications

Distinct metric contracts. The original endpoint J_{cond}, document-level union-nonempty symmetric Jaccard, legacy directional overlap, and the transparent CONS reimplementation are different estimands and should not be pooled. For paired document supports S_1, S_2, the union-nonempty contract uses $E_{\mathrm{union}} = \mathbb{I}[S_1 \cup S_2 \neq \emptyset]$ and $C_{\mathrm{union}} = \Pr(E_{\mathrm{union}} = 1)$; the both-nonempty rate is $C_{\mathrm{both}} = \Pr(S_1 \neq \emptyset, S_2 \neq \emptyset)$. In the original seed-42 support-state audit, $(d_{00}, d_{10}, d_{01}, d_{11}) = (29, 25, 2, 428)$ over 484 pairs; hence $C_{\mathrm{union}} = 455/484 = 0.9401$ and $C_{\mathrm{both}} = 428/484 = 0.8843$. These values are not the means from the newly aligned three-seed pipeline.

Three-seed directional denominators. Legacy directional overlap is $0.7064 \pm 0.0337 \rightarrow 0.9198 \pm 0.0057$. Under a separate transparent reimplementation, CONS increases $0.7366 \pm 0.0269 \rightarrow 0.9382 \pm 0.0047$ even as the mean jointly correct numerator falls $6727 \pm 579 \rightarrow 5876 \pm 208$ and factual-correct denominator falls $9122 \pm 468 \rightarrow 6262 \pm 190$. The denominator covers $57.2\% \rightarrow 39.3\%$ of 15,939 eligible aligned gold pairs. These data establish denominator contraction, not a claim that the directional score inherits the symmetric theorem.

Three-seed support diagnostics and sensitivity. The matched-cardinality witness, gold-positive margin left shift, increased fraction below threshold, and migration away from the both-nonempty state hold in all three seeds. Over three seeds and 100 decoding thresholds, the complete failure signature was observed at 212 of 300 correlated operating points (70.7%), not at 212 independently repeated experiments. A document-grouped out-of-fold analysis over 51,852 gold-positive atoms found the initial No_CF margin to be the strongest measured threshold-crossing predictor ($\Delta\mathrm{AUC} = 0.181$, 95% CI $[0.162, 0.200]$). Conditional on margin, relation rarity added an association ($\Delta\mathrm{AUC} = 0.116$, 95% CI $[0.101, 0.132]$; positive in all three seeds), whereas measured cross-view instability added only 0.006. These associations do not uniquely identify a causal training mechanism.

Cross-backbone and independent-dataset extension. A frozen ATLOP strength screen on Re-DocRED found the full signature at $\gamma = 0.10$ in seeds 42, 13, and 21. Corresponding symmetric union-nonempty Jaccard values were $0.521 \rightarrow 0.559$, $0.527 \rightarrow 0.582$, and $0.540 \rightarrow 0.583$, with Recall, support size, C_{union}, and C_{both} decreasing in each seed. At $\gamma = 0.05$ behavior was benign; $\gamma = 0.20/0.30$ caused support collapse without an agreement increase. This is a within-dataset cross-backbone witness. All ATLOP arms, including the $\gamma = 0$ baseline, were decoded through the same executed path: a single global threshold τ^* on sigmoid scores selected on the development split,

rather than ATLOP's adaptive-threshold class used during ATLoss training. The witness is therefore stated for this emitted-set decoding contract.

A separate BC5CDR TitleDrop protocol used 1,500 PubMed articles in fixed 500/500/500 splits, with full title+abstract and reduced abstract-only paired views. The same candidate universe was retained across both views and arms. The control used ATLoss on the full view; the candidate used ATLoss(full)+ATLoss(reduced)+$0.5\,\mathrm{SKL}$. Both checkpoints were selected by full-view DEV F1 before one-time TEST access. The subsequent CPU-only audit of retained TEST predictions reported conditional Jaccard $0.8138 \rightarrow 0.8387$, document-aware full-view Recall $0.6923 \rightarrow 0.6332$ (against 1,066 gold document-level relation memberships), emitted memberships $1219 \rightarrow 1019$, union coverage $0.926 \rightarrow 0.922$, and silence $37 \rightarrow 39$ of 500 documents. Document-aware full-view F1 (1,066 gold memberships) was $0.6459 \rightarrow 0.6475$. A subsequent unified three-seed rerun retrained all six arms (seeds 42, 13, and 21; control and candidate) with one seed-parameterized implementation of the same protocol; all six checkpoints were fixed by full-view DEV F1 before any rerun TEST access. The rerun seed 42 reproduced the direction of the original seed-42 result but not its exact values. In seeds 13 and 21, conditional agreement again increased, but Recall, emitted support, and union coverage also increased and Silence decreased, so the full signature did not hold. Under the unified protocol the complete failure signature is therefore observed for one of three seeds. This demonstrates a possible failure case on an independent dataset but does not establish seed-stable replication; it is distinct from the original benign BC5CDR probe and does not establish cross-dataset prevalence.

Exact training objectives. No_CF uses only the factual-view binary cross-entropy (BCE). Uniform-0.5 uses factual-view BCE, counterfactual-view BCE, and one half of the mean symmetric Bernoulli KL divergence between the paired views. Constant-0.10 uses the same two BCE terms and the same outer coefficient of one half, but sets every per-sample KL allocation weight to 0.10, so its effective KL multiplier is 0.05. For B_global_minmax, each per-sample weight is obtained by winsorizing the warm-checkpoint symmetric Bernoulli KL at the global first and ninety-ninth percentiles and normalizing it to the unit interval (the degenerate case uses 0.5); this is global winsorized normalization, not batch or literal global min–max. The FASV variants share the same second-stage objective and differ only in the raw per-sample quantity used before batch min–max normalization: Real-FASV uses the warm-phase KL directly; Shuffled-FASV applies one global permutation of that vector; Random-FASV samples with replacement from the empirical Real-FASV distribution. Decoding thresholds are selected once on factual development data and reused on the counterfactual view.

Threshold sweeps are not risk–coverage curves. A decoding-threshold sweep simultaneously changes emitted support, evaluability, and predictions. We therefore refer to it as a threshold-induced support–utility curve rather than a standard selective-prediction risk–coverage curve, which would require fixed decoded outputs, a separate confidence ranking, and an explicitly defined retained-set risk.

C Additional Protocol Notes

Official task tables use the standard DocRE evaluation protocol: global τ^* search on the development split, reporting F1 and Recall. Support-aware tables use a frozen metric contract for J, J_{cond}, Silence, and Cardinality, evaluated at the endpoint threshold without re-tuning. We deliberately separate official task utility from agreement diagnostics so that a high J_{cond} value is not misread as a coverage-preserving robustness guarantee.

D Proof of the Support-Shrinkage Proposition

This appendix derives the local threshold-crossing mechanism of Section 3: as the effective consistency budget grows, fragile positive atoms can be driven below the decoding threshold, realizing the P_{shrink} branch of Theorem 1.

D.1 Setup

Let $a = (e_h, e_t, r)$ be a candidate relation triple. For a document view x, the model assigns a score $p_a(x) \in [0, 1]$ to a. The predicted support is

$$S_P(x) = \{a : p_a(x) \geq \tau\}.$$

Consider two aligned views $x^{(1)}$ and $x^{(2)}$ of the same instance, such as a factual and a counterfactual document. The consistency loss penalizes score differences between the two views. We focus on a rare positive atom a whose score is near the decoding threshold τ. Such atoms are critical because a score crossing τ changes the discrete predicted support, making them candidates for the P_{shrink} branch identified in Theorem 1.

Let π_a denote the effective supervised mass of atom a, and ρ_a its exposure under the consistency objective. Under uniform consistency, ρ_a grows large relative to π_a for many rare labels, because the regularizer is applied broadly even when supervised positive evidence is sparse–precisely the sparse, long-tailed support condition in Table 1.

D.2 Local risk

We approximate the supervised loss near the decoding threshold by a quadratic surrogate and use squared consistency for analytic clarity. Let $z_a^{(v)}$ denote the local evidence for atom a in view v, and use a one-dimensional local score model

$$p_a^{(v)} \approx w_a z_a^{(v)}.$$

For positive instances of atom a, assume

$$\mathbb{E}[z_a^{(v)} \mid y_a = 1] = \mu_a > 0, \qquad \mathbb{E}\Big[(z_a^{(v)})^2 \mid y_a = 1\Big] = \mu_a^2 + \sigma_a^2,$$

where σ_a^2 captures local cross-view instability. With squared cross-view consistency, the local disagreement term contributes $2\sigma_a^2 w_a^2$ up to the exposure weight ρ_a; other constant normalization factors are absorbed into ρ_a. The per-atom regularized risk is

$$\mathcal{R}_a(w_a) = \pi_a \left[1 - 2\mu_a w_a + (\mu_a^2 + \sigma_a^2) w_a^2\right] + 2\lambda\beta_a\rho_a\sigma_a^2 w_a^2.$$

Here λ is the global consistency weight and β_a is the allocation factor for atom a. Thus the effective consistency budget acting on this atom is $\lambda\beta_a\rho_a$.

D.3 Optimal gain and critical budget

Differentiating $\mathcal{R}_a(w_a)$ with respect to w_a and setting the derivative to zero gives the unique minimizer

$$w_a^*(\lambda, \beta_a) = \frac{\pi_a \mu_a}{\pi_a(\mu_a^2 + \sigma_a^2) + 2\lambda\beta_a\rho_a\sigma_a^2}.$$

Since the denominator is strictly increasing in the effective consistency budget $\lambda\beta_a\rho_a$, w_a^* is strictly decreasing whenever $\rho_a\sigma_a^2 > 0$. The expected positive-class score is

$$\bar{p}_a(\lambda, \beta_a) \equiv \mathbb{E}[p_a^{(v)} \mid y_a = 1] = \mu_a w_a^*(\lambda, \beta_a) = \frac{\pi_a \mu_a^2}{\pi_a(\mu_a^2 + \sigma_a^2) + 2\lambda\beta_a\rho_a\sigma_a^2}.$$

The atom crosses below the decoding threshold when $\bar{p}_a(\lambda, \beta_a) < \tau$. Solving the equality $\bar{p}_a(\lambda, \beta_a) = \tau$ gives the transition budget

$$\lambda_a^* = \frac{\pi_a}{2\beta_a\rho_a\sigma_a^2}\left(\frac{\mu_a^2}{\tau} - \mu_a^2 - \sigma_a^2\right),$$

provided $\beta_a\rho_a\sigma_a^2 > 0$. This transition budget is positive if and only if

$$\tau < \frac{\mu_a^2}{\mu_a^2 + \sigma_a^2},$$

i.e., the atom is initially above threshold in the no-consistency local surrogate. If this condition fails, the atom is already suppressed at $\lambda = 0$. When $\lambda > \lambda_a^*$, the expected score $\bar{p}_a$ falls below τ and the atom is excluded from the predicted support S_P.

Smaller π_a (rare labels) lowers λ_a^*, so rare positives collapse first; larger σ_a^2 (high cross-view variance) lowers λ_a^*, so unstable labels collapse first; larger ρ_a (broad consistency exposure) lowers λ_a^*, so uniformly regularized labels collapse earlier; smaller β_a raises λ_a^*, which explains why smaller or normalized budget controls can delay support collapse relative to Uniform-0.5.

Margin-coordinate intuition. The same threshold-crossing phenomenon can be seen directly in paired-view margin coordinates. Let

$$u = p_\theta(a \mid x) - \tau, \qquad v = p_\theta(a \mid \tilde{x}) - \tau$$

denote the factual-view and aligned-view margins for a fragile positive atom. Suppose a local quadratic approximation has task-preferred margins $\mu > 0$ and $\nu < 0$, with $\mu + \nu < 0$. The condition $\mu + \nu < 0$ captures the case where the aligned-view negative evidence is strong enough to pull the paired average below the decoding threshold. Consider

$$\mathcal{L}_{\lambda_{\text{eff}}}(u, v) = \frac{1}{2}(u - \mu)^2 + \frac{1}{2}(v - \nu)^2 + \frac{\lambda_{\text{eff}}}{2}(u - v)^2.$$

Solving the resulting 2×2 linear system gives

$$u_\lambda = \frac{\mu + \lambda_{\text{eff}}(\mu + \nu)}{1 + 2\lambda_{\text{eff}}}, \qquad v_\lambda = \frac{\nu + \lambda_{\text{eff}}(\mu + \nu)}{1 + 2\lambda_{\text{eff}}}.$$

Therefore, if

$$\lambda_{\text{eff}} > \frac{\mu}{-(\mu + \nu)},$$

then $u_\lambda < 0$ and $v_\lambda < 0$. In margin coordinates, a fragile atom that is initially emitted in the factual view can therefore be driven below threshold in both views by a sufficiently strong effective consistency budget. This paragraph is included only as a margin-level intuition for Figure 2; the preceding λ_a^* derivation remains the main mechanism analysis because it exposes the roles of rarity, cross-view variance, consistency exposure, and budget allocation.

D.4 Effect on conditional agreement

Suppose a rare positive atom $a^\star$ is unstable across views. Before suppression, a representative pair of predicted supports is

$$S_P(x^{(1)}) = C \cup \{a^\star\}, \qquad S_P(x^{(2)}) = C,$$

where C is the stable common support. The per-pair Jaccard agreement is $|C|/(|C| + 1) < 1$. After suppression of $a^\star$ in both views,

$$S_P(x^{(1)}) = S_P(x^{(2)}) = C.$$

When $C \neq \emptyset$, the agreement rises to one. When $C = \emptyset$, the pair becomes double-empty and is excluded from the evaluability event, removing it from the J_{cond} average. In both cases, suppressing the unstable atom increases or preserves J_{cond} while reducing Recall and Cardinality–the support-shrinkage failure mode of Theorem 1.

D.5 Relation to calibration and selective prediction

In ordinary single-label classification, the prediction is $\hat{y}(x) = \arg\max_k z_k(x)$, and any strictly monotone calibration map leaves the argmax unchanged. Predicted cardinality is always one and Silence is impossible. In thresholded multi-label prediction, by contrast, the model can remove atoms from the predicted support $S_P(x)$ by moving their scores below τ. The support-shrinkage failure is therefore a property of thresholded set decoding, not of probability calibration.

The mechanism can also be viewed as implicit structured abstention. Classical selective prediction gives the model an explicit reject option evaluated through a risk–coverage curve. Here, no reject head is introduced and no coverage curve is reported: the model instead implicitly abstains on parts of the structured output by suppressing difficult relation atoms below the decoding threshold. This hidden abstention is precisely what makes J_{cond} a selected-support quantity rather than an unconditional measure.

E Empirical Predictions

The λ_a^* derivation in Appendix D yields five testable empirical predictions for the support-shrinkage mechanism. These predictions are not additional theorems or universal cross-task claims; they specify what the local threshold-crossing mechanism should look like in the primary Re-DocRED stress test.

(1) Rare and long-tailed positive atoms should lose Recall before head-label atoms as the consistency budget λ grows, because their transition thresholds λ_a^* are lower. (2) Uniform-0.5 should shift gold-positive margins left and below the decoding threshold τ, particularly for mid-frequency and tail atoms. (3) Constant-0.10 should delay or attenuate this margin shift, because its smaller effective budget $\lambda\beta_a\rho_a$ raises λ_a^* for each atom. (4) J_{cond} should increase together with higher Silence and lower Evaluability Coverage C, as the evaluated support shrinks. (5) Lowering the decoding threshold τ after Uniform-0.5 should recover Recall substantially only when suppressed scores remain close to τ; if scores have been pushed far below threshold, threshold lowering provides limited recovery.

Table 6 illustrates the conditional nature of prediction (5): under strong Uniform-0.5, many scores are pushed far below common thresholds, so even aggressive threshold lowering yields near-zero Recall. Together, these five predictions guide the score-trace replay, margin audit, and support-state decomposition in Section 5 and Appendix G.

F Hard Subset-Selection Ablation (D3)

D3 denotes an ablation family that replaces the budget-controlled support-recovery variants, such as Constant-0.10 and B_global_minmax, with hard subset selection: either top-k gating or random gating over candidate groups. It asks whether hard subset selection can substitute for the budget-allocation mechanism; in this ablation, it cannot. Table 5 is retained as supporting evidence: simple top-k or random gating does not reproduce the support-aware behavior of the budget-controlled variants.

Config	F1	Recall	J	Silence
Uniform-0.5	0.3463	0.4041	0.5145	8821
Real-FASV	0.3708	0.4364	0.5202	8028
TopK-5	0.3491	0.4487	0.5162	7561
TopK-10	0.3302	0.4360	0.5280	8106
TopK-20	0.3069	0.3857	0.4910	8996
TopK-40	0.3351	0.3127	0.4157	10421
TopK-60	0.3318	0.3410	0.4547	9945
RandomGate-5	0.3456	0.3908	0.4842	8958
RandomGate-10	0.3376	0.3911	0.4862	8692
RandomGate-20	0.3222	0.3164	0.4383	10437
RandomGate-40	0.3419	0.3323	0.4456	10136
RandomGate-60	0.3333	0.3433	0.4574	9931

Table 5: **D3 diagnostic ablation family.** Hard subset selection (top-k and random gating) does not reproduce the support-aware behavior of the budget-controlled variants in this ablation. The table is retained as supporting evidence that the support-recovery behavior of Constant-0.10 and B_global_minmax is not explained by simple hard-gating strategies alone.

G Additional Threshold and Support-State Audits

The main text reports a compact subset of the score-trace replay; this appendix provides the complete audit tables. Table 6 gives the fixed-threshold grid designed to check whether the J_{cond} paradox is merely an artifact of official development threshold search. Table 7 gives matched document-coverage replay, and Table 8 gives the exact document-level support-state decomposition across 484 aligned factual/counterfactual development pairs. All three tables use the seed-42 score-trace audit and are diagnostic; they do not replace the three-seed canonical endpoint witness in Table 2.

G.1 4-bin frequency-bucket margin shift

Table 9 gives the complete 4-bin Recall breakdown across all four configurations. Bins are defined by gold-fact frequency: head ($\geq$1000), mid (300–999), tail (100–299), and extreme tail ($<$100). Mid-frequency facts collapse most severely under Uniform-0.5, consistent with the λ_a^* threshold-crossing mechanism in Appendix D; support-recovery controls partially recover mid and tail Recall.

Config	τ	Pred. cand.	Precision	Recall	F1
No_CF	0.50	115270	0.129	0.857	0.224
No_CF	0.70	91027	0.156	0.821	0.262
No_CF	0.90	63559	0.205	0.755	0.323
No_CF	0.95	51455	0.237	0.707	0.355
No_CF	0.99	28569	0.335	0.553	0.417
Uniform-0.5	0.50	50920	0.192	0.564	0.286
Uniform-0.5	0.70	31690	0.250	0.459	0.324
Uniform-0.5	0.90	8170	0.465	0.220	0.299
Uniform-0.5	0.95	566	0.792	0.026	0.050
Uniform-0.5	0.99	0	0.000	0.000	0.000
Constant-0.10	0.90	31217	0.288	0.520	0.371
Constant-0.10	0.95	19161	0.354	0.393	0.372
B_global_minmax	0.90	34126	0.276	0.545	0.367
B_global_minmax	0.95	22618	0.334	0.437	0.379

Table 6: **Fixed-threshold replay (seed 42).** At every common threshold, Uniform-0.5 shows severe Recall loss and support collapse relative to No_CF. This rules out a pure-threshold-artifact explanation of the J_{cond} paradox: the phenomenon is not merely a byproduct of official development-threshold selection.

Config	Pred. cand.	Precision	Recall	F1	J_{cond}	Silence
No_CF	28569	0.335	0.553	0.417	0.675	5258
Uniform-0.5	95736	0.132	0.733	0.224	0.929	2390
Constant-0.10	48807	0.227	0.642	0.336	0.900	4115
B_global_minmax	44871	0.238	0.617	0.343	0.893	4466

Table 7: **Matched document-coverage replay (seed 42).** When Uniform-0.5 is forced to match No_CF's document coverage, it must emit many more candidates; this sharply lowers precision and F1. The replay shows that recovering document coverage by threshold relaxation is not equivalent to recovering support-aware witness behavior.

Figure 3 visualizes the same breakdown. It shows that the largest visible Recall loss concentrates in the mid-frequency bucket, where there are enough recovered positives for the loss to be measurable, while tail buckets remain near zero under both settings.

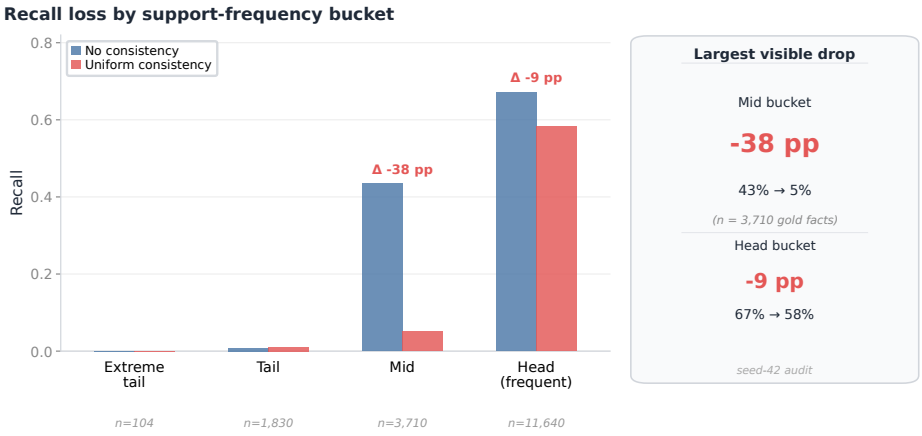


Figure 3: **Recall loss by support-frequency bucket.** Uniform consistency causes the largest visible Recall loss in the mid-frequency bucket, where Recall drops from 43% to 5% (-38 pp, $n = 3{,}710$ gold facts). The head bucket drops from 67% to 58% (-9 pp). Tail buckets have near-zero measurable Recall for both settings, so the most interpretable loss appears in buckets with enough recovered positives. This seed-42 frequency-bucket audit supports the gold-positive margin shift and evaluated-support mechanisms shown in Figure 2.

Config	Double-empty	Fact-only	CF-only	Both non-empty	Pair cov.	Factual card.
No_CF	0.0000	0.0000	0.0021	0.9979	1.0000	29368
Uniform-0.5	0.0599	0.0517	0.0041	0.8843	0.9401	22270
Constant-0.10	0.0124	0.0103	0.0083	0.9690	0.9876	23294
B_global_minmax	0.0186	0.0062	0.0124	0.9628	0.9814	24421

Table 8: **Document-level support-state decomposition (seed 42, 484 aligned pairs).** Each row reports the fraction of pairs falling into double-empty, factual-only, CF-only, and both-non-empty states, along with pair coverage and factual Cardinality. These quantities are diagnostic and use whole-document decoded triple sets; they instantiate the same predictor-dependent evaluability-coverage concern captured by $C(P)$ in Section 2.

Config	Bucket (n gold)	Recall	Below τ	Median margin	ΔRecall
No_CF	Head $\geq$1000 (11640)	0.671	32.9%	+1.088	—
	Mid 300–999 (3710)	0.434	56.6%	−0.462	—
	Tail 100–299 (1830)	0.008	91.9%	−3.085	—
	Ext. tail <100 (104)	0.000	100%	−5.344	—
Uniform-0.5	Head	0.582	41.8%	+0.289	−0.089
	Mid	0.051	94.9%	−2.000	−0.383
	Tail	0.010	99.0%	−3.586	+0.002
	Ext. tail	0.000	100%	−5.374	0.000
Constant-0.10	Head	0.580	42.0%	+0.429	−0.091
	Mid	0.275	72.5%	−1.105	−0.159
	Tail	0.032	96.8%	−3.114	+0.024
	Ext. tail	0.000	100%	−5.165	0.000
B_global_minmax	Head	0.611	38.9%	+0.704	−0.060
	Mid	0.245	75.5%	−1.184	−0.189
	Tail	0.023	97.7%	−3.261	+0.015
	Ext. tail	0.000	100%	−5.388	0.000

Table 9: **Full 4-bin frequency-bucket Recall analysis (seed 42).** Bins are defined by gold-fact count in the development set. The ΔRecall column is relative to No_CF. Mid-frequency facts show the largest collapse under Uniform-0.5 (-38 pp), consistent with the gold-positive margin-shift mechanism of Appendix D.

G.2 Budget boundary scan

Figure 4 shows the Constant-λ budget scan across four support-aware quantities as λ increases from 0.10 to 1.00. J_{cond} rises while Recall falls, Cardinality decreases, and Silence increases—consistent with the λ_a^* threshold-crossing mechanism in Appendix D. The dashed line in each panel denotes the no-consistency baseline. This scan is used only to delimit the failure mode; it is not a proposed repair method.

H Additional Cross-Backbone, Boundary, and LLM Evidence

Table 10 gives the full regime classification behind the compact Table 1 in the main text. The classification is qualitative by design: the paper does not claim a quantitative scaling law from support geometry to collapse severity.

Table 11 preserves the original cross-backbone and boundary experiments for transparency. Under its original strong Uniform-0.5 setting, ATLOP exhibits severe support collapse with a lower J_{cond} (0.510 ± 0.063 vs. 0.750 ± 0.011 for No_CF). A separate frozen intermediate-strength ATLOP analysis subsequently yielded the full witness in three seeds; see Appendix B. Likewise, the older benign CDR/BC5CDR boundary differs from the BC5CDR TitleDrop protocol. Re-TACRED remains lower-support boundary evidence. The LLM probe is reported separately in Appendix H.2 as reference-tier motivation only; it is not part of the primary evidence chain.

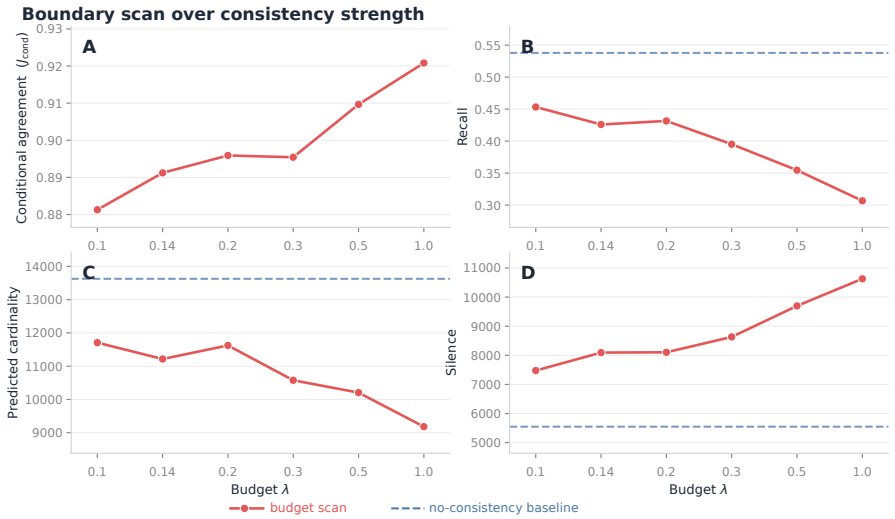


Figure 4: **Boundary scan over consistency strength.** As the consistency budget λ increases (categorical x-axis), J_{cond} rises while support-aware quantities deteriorate: Recall and Cardinality decrease, and Silence increases. The dashed line denotes the No_CF baseline. This scan delimits the failure mode and is not a proposed repair method.

Setting	R1 support	R2 thresh.	R3 tail	R4 paired cond.	R5 pressure	Role
Re-DocRED/BERT	yes	yes	strong	yes	yes	main stress test
Re-DocRED/ATLOP	yes	yes	strong	yes	yes	witness at intermediate strength
BC5CDR TitleDrop	yes	yes	task-dep.	yes	yes	seed-unstable (1/3 seeds)
CDR/BC5CDR prior probe	yes	yes	weak	partial	yes	biomedical boundary
Re-TACRED	partial	yes	weak	partial	yes	sentence-level boundary
Fixed-support tasks	no	no/limited	task-dep.	task-dep.	task-dep.	outside-regime contrast
LLM structured extraction	yes	implicit	task-dep.	task-dep.	no	motivation only

Table 10: **Full qualitative regime placement.** R1–R5 denote: variable-support set-valued output, thresholded decoding, sparse/long-tailed positive support, paired evaluation with predictor-dependent conditioning, and optimization pressure on the conditional objective without an explicit coverage constraint. The BC5CDR TitleDrop protocol (full signature in one of three seeds) is distinct from the older, benign biomedical probe. Boundary rows delimit where the full signature attenuates.

CDR biomedical boundary (seed 42). The seed-42 CDR probe serves as a benign biomedical boundary, not a second failure witness. In this setting, J_{cond} rises together with Recall while Silence falls. This is consistent with the regime characterization of Table 1: the full support-shrinkage witness requires the R1–R3 conditions to hold jointly, whereas binary biomedical settings with narrower support may behave benignly.

H.1 Claim–evidence audit

Table 12 summarizes each finding together with the evidence it rests on and the scope within which it is claimed.

H.2 LLM structured-output probes

Table 13 reports a small set of LLM structured extraction probes. These probes are included only to motivate coverage-aware reporting for LLM-era extraction systems, as described in Section 6; they do not constitute evidence for the supervised support-shrinkage mechanism studied here.

Setting	Config	F1	Recall	J	J_{cond}	Silence / Card.
ATLOP DocRE	No_CF	0.370 ± 0.010	–	0.652 ± 0.009	0.750 ± 0.011	408 ± 24 / 28435 ± 137
ATLOP DocRE	Uniform-0.5	0.028 ± 0.012	–	0.011 ± 0.005	0.510 ± 0.063	16148 ± 172 / 351 ± 143
ATLOP DocRE	Constant-0.10	0.539 ± 0.015	–	0.383 ± 0.013	0.758 ± 0.006	7948 ± 446 / 10174 ± 578
BC5CDR	No_CF	0.589–0.605	0.660–0.693	–	–	boundary
BC5CDR	Uniform-0.5	0.563–0.593	0.679–0.773	–	–	mixed
BC5CDR	Constant-0.10	0.578–0.594	0.623–0.690	–	–	mixed
CDR boundary (seed 42)	No_CF	0.5805	0.6604	–	0.9312	111 / 247
CDR boundary (seed 42)	Uniform-0.5	0.5734	0.7727	–	0.9550	65 / 289
CDR boundary (seed 42)	Constant-0.10	0.5788	0.6283	–	0.9574	117 / 235
Re-TACRED	Baseline	0.7699 ± 0.0043	–	–	–	sentence-level
Re-TACRED	Uniform-0.5	0.7505 ± 0.0022	–	–	–	sentence-level
Re-TACRED	Constant-0.10	0.7662 ± 0.0017	–	–	–	sentence-level

Table 11: **Cross-backbone and boundary evidence.** ATLOP exhibits severe support collapse under strong uniform consistency within the same DocRE regime, but J_{cond} falls rather than rises (0.510 ± 0.063 vs. 0.750 ± 0.011 for No_CF); the original setting is a cautionary boundary, whereas the distinct $\gamma = 0.10$ setting reproduces the witness in three seeds (see Appendix B). The older BC5CDR probe and Re-TACRED delimit weaker or mixed regimes; the new BC5CDR TitleDrop protocol is a separate, seed-unstable observation (one of three seeds). The CDR seed-42 probe is a benign biomedical boundary: higher J_{cond} is accompanied by higher Recall and lower Silence–the opposite of the support-shrinkage pattern.

Finding	Evidence	Scope
Conditional agreement can misrank predictors	Theorem 1	Contracts with removable low-agreement states
Fixed agreement can conceal support differences	Corollaries 1–2	Union-nonempty/Jaccard construction
Support-shrinkage witness	Re-DocRED/BERT, three seeds (Tables 2, 3)	Primary stress test
Not a threshold or output-size artifact	Fixed-threshold replay, matched-cardinality replay (Table 4)	Seed 42 in detail; qualitative in three seeds
Fragile positives cross the threshold	Margin audit, support-state decomposition (Fig. 2)	Seed 42; sufficient, non-unique mechanism
Support-recovery controls	Constant-0.10, B_global_minmax	Diagnostic, not proposed methods
Cross-backbone result	ATLOP $\gamma = 0.10$, three seeds (Appendix B)	Re-DocRED only; higher strengths collapse support
Seed sensitivity	BC5CDR TitleDrop (Appendix B)	Full signature in 1 of 3 seeds
Regime boundaries	Earlier CDR/BC5CDR and Re-TACRED probes	Weaker, mixed, or benign effects
LLM-era relevance	Appendix H.2 probe	Motivation only

Table 12: **Findings, evidence, and scope.**

I Additional Budget-Variant Ablations and Prior-Work Alignment

Table 14 gives the retained budget-variant evidence on seed 42. These ablations support the interpretation that effective budget magnitude and allocation matter more than any single fragility-specific targeting formula: small or normalized budget controls recover more support-aware witnesses than large uniform budgets, while randomized or shuffled allocations occupy an intermediate position. Table 15 reports the CovEReD-style factual/counterfactual F1 comparison from an earlier evaluation run.

J Compute Resources

The main supervised DocRE experiments used BERT-base-cased and ATLOP-style DocRE backbones on NVIDIA data-center GPUs. Each reported three-seed Re-DocRED configuration used seeds 13, 21, and 42; runtime varied with decoding and counterfactual evaluation, but was on the order of a few GPU-hours per seed for training and evaluation, with additional offline time for decoding.

Probe	Valid JSON	Mean card.	Precision	Recall	F1	Omission/empty
GPT-4o V1 closed zero-shot	–	1050	–	–	0.135	Silence 4688
GPT-4o V2 closed one-shot	–	1181	–	–	0.146	Silence 4561
GPT-4o V3 open extraction	–	1144	–	–	0.022	unsupported 91.3%
API plain JSON	0.933	18.2	0.000	0.000	0.000	1.000 omission
API schema JSON	1.000	18.6	0.298	0.161	0.197	0.839 omission
API self-consistency-3	1.000	7.8	0.426	0.095	0.152	0.905 omission
API verifier-filtered	1.000	18.7	0.270	0.154	0.193	0.846 omission

Table 13: **LLM structured-output probes.** These probes motivate coverage-aware reporting for LLM-era structured extraction systems; they are not evidence for the supervised mechanism studied here.

Config	F1	Precision	Recall	J_{cond}	Silence
Uniform-0.5	0.3463	0.3030	0.4041	0.9345	8821
Real-FASV	0.3708	0.3223	0.4364	0.9072	8028
Shuffled-FASV	0.3576	0.3111	0.4204	0.9070	8408
Random-FASV	0.3596	0.2822	0.4957	0.8852	6762
Constant-0.10	0.3761	0.3213	0.4533	0.8813	7475
Constant-0.50	0.3312	0.3107	0.3545	0.9097	9696
Constant-1.00	0.3168	0.3277	0.3066	0.9208	10627
B_global_minmax	0.3790	0.3190	0.4667	0.8760	7029

Table 14: **Budget-variant ablations (seed 42).** Randomized or shuffled allocations recover some support relative to Uniform-0.5, and small or normalized budget controls (Constant-0.10, B_global_minmax) improve Recall and reduce Silence more reliably than large budgets.

Candidate-score dumping for the threshold audit was run only for seed-42 configurations and was followed by CPU-only streaming replay to avoid loading all candidate traces into memory. Margin audits, support-state decomposition, figure generation, and table construction were CPU/offline analyses. The LLM reference probes used a small fixed document subset via API calls; they are not part of the main training pipeline and are included only as reference-tier motivation.

Config	Factual F1	CF F1	F1 drop	Factual triples
No_CF	0.424	0.359	0.065	27147
Uniform-0.5	0.352	0.350	0.002	22270
Weighted-0.1	0.344	0.335	0.009	42007
Constant-0.10	0.383	0.375	0.009	23294

Table 15: **CovEReD-style factual/counterfactual F1 comparison (seed 42, earlier evaluation run).** Consistency training narrows the factual-to-counterfactual F1 gap.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The paper is framed around a theorem-backed selected-support non-identification result, a controlled Re-DocRED empirical witness, scoped cross-backbone evidence, a seed-unstable independent-dataset observation, boundary analyses, and a theorem-grounded reporting discipline. The abstract and introduction delimit the high-risk R1–R5 regime and avoid claiming universal cross-task generalization or introducing a new repair method.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Section 8 states the Re-DocRED focus, reports that the BC5CDR TitleDrop signature holds in only one of three seeds, clarifies that the theorem is not a universal tradeoff theorem between agreement and Recall, notes that score-trace replay and matched-cardinality controls are diagnostic artifacts rather than deployment policies, and confirms that the LLM probe is reference-tier motivation only. The claim–evidence audit in Appendix H.1 further separates proved claims, primary empirical witnesses, and scope-setting diagnostics.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: Appendix A provides the complete construction and proof of Theorem 1 and Corollaries 1–2, including a self-contained notation block and explicit formulas for all support-aware quantities. Appendix D separately states the assumptions and derivation for the local support-shrinkage mechanism, including the transition budget λ_a^* and a margin-coordinate intuition paragraph aligned with the empirical margin audit.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?
   Answer: [Yes]
   Justification: The manuscript identifies the frozen metric contract, configurations, seeds (13, 21, 42), evaluation layers, threshold replay protocol, and diagnostic artifacts needed to interpret the main reported results. The manuscript documents the principal protocols, while the source-code availability is stated separately in the next checklist item.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [No]
   Justification: This manuscript does not provide a public repository URL or an executable supplemental reproducibility package. The manuscript reports the frozen metric contract, configurations, seeds, evaluation layers, and source of all main reported quantities. The main text gives metric contracts and evaluation protocols, but code and script access are not linked in the PDF. Benchmark data are obtained from the cited public releases.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details necessary to understand the results?
   Answer: [Yes]
   Justification: Sections 4–5 and Appendices D–G specify the configurations, consistency settings, thresholding contract, replay audits, metric definitions, and three-seed aggregation used to interpret the reported results. Appendix C additionally states the protocol separation between official task utility and support-aware agreement diagnostics.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate

information about the statistical significance of the experiments?
Answer: [Yes]
Justification: The main Re-DocRED endpoint table reports mean and standard deviation over seeds 13, 21, and 42. Single-seed replay and threshold audits are explicitly labeled as diagnostic rather than statistical endpoint claims.

8. **Experiments compute resources**
Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?
Answer: [Yes]
Justification: Appendix J reports the hardware class (NVIDIA data-center GPUs), seed structure (seeds 13, 21, 42), approximate compute scale (a few GPU-hours per seed), and the CPU-only offline components (threshold replay, margin audits, figure generation). LLM reference probes used API calls on a small fixed document subset and are not part of the main training pipeline.

9. **Code of ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?
Answer: [Yes]
Justification: The work is benchmark-centered robustness and evaluation research using standard, publicly available NLP datasets.

10. **Broader impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes]
Justification: The discussion notes that coverage-aware reporting can reduce overclaiming of robustness in information extraction systems, while unreported coverage collapse may mislead downstream knowledge-base construction, relation extraction deployments, and structured-output evaluation pipelines.

11. **Safeguards**
Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?
Answer: [N/A]
Justification: The paper does not release a high-risk foundation model or a high-risk scraped dataset.

12. **Licenses for existing assets**
Question: Are the creators or original owners of assets used in the paper properly credited and are the license and terms of use explicitly mentioned and respected?
Answer: [Yes]
Justification: The paper cites the benchmark, model, and software assets used in the study. No benchmark datasets or pretrained model weights are redistributed in the submission. If training code or derived artifacts are released, their package should include applicable license and attribution details.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [N/A]
Justification: The submission does not introduce a new dataset, model, or benchmark asset. Derived analysis scripts and metric artifacts are not provided as a new public benchmark asset in the current submission.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?
Answer: [N/A]
Justification: The paper does not involve crowdsourcing or human-subject experiments.

15. **Institutional review board approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such

risks were disclosed, and whether IRB approvals were obtained?
Answer: [N/A]
Justification: The paper does not involve human-subject research.

16. **Declaration of LLM usage**
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
Answer: [Yes]
Justification: The paper includes a small reference-tier LLM structured extraction probe in Appendix H.2. This probe is used only to motivate coverage-aware reporting for LLM-era structured extraction and is not part of the main supervised mechanism or proof.