STATE BEFORE ACTION: EXPLICIT REPAIR STATE IMPROVES MULTI-OPERATOR CONTRADICTION-CLUSTER CURATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contradiction-heavy datasets fail at the cluster level: one keep/drop decision cannot simultaneously preserve legitimate ambiguity, recover hidden context, selectively relabel unsupported answers, and suppress annotation artifacts. We introduce STATEROUTE, a repair-state interface that lets the learned controller, budget allocator, and reviewer distinguish four causes of conflict before selecting among five cluster repairs under an explicit minority-collapse cost. A matched comparison fixes graph extraction, cluster evidence, action menu, controller-label pool, reviewer interface, and the 12.4 min/1K review budget while contrasting explicit state with direct action imitation. Across five datasets, STATEROUTE reaches 0.82 pooled pair-level contradiction F1, compared with 0.79 for a GBDT action selector, 0.80 for a tuning-matched Cluster-Transformer selector, and 0.76 for type-free graph repair; on MedQA, it reaches 61.7% accuracy and 87.3% minority retention, versus $60.4\%/74.2\%$ and $61.0\%/82.5\%$ for the two direct selectors. A disjoint 384-cluster action test localizes the gain to repair choice: STATEROUTE obtains 0.76 accuracy and 0.72 macro-F1, ahead of 0.70/0.65 for GBDT imitation and 0.73/0.69 for sequence imitation. The state advantage carries through all five Llama-3-8B downstream tasks, while Mistral-7B preserves the ordering over graph-only and judge baselines. Neutral-loss training, three retrieval/scoring backbones, leave-one-domain-out evaluation, cross-fitted controller supervision, and independent medical and legal-contract panels retain the advantage on their reported endpoints. Under this matched contradiction-cluster scaffold, operator-distinguishing state improves repair choice, downstream utility, and minority preservation beyond increasingly expressive direct action imitation, making the combined-system gain auditable under shared graph, evidence, and review-budget controls.

1 INTRODUCTION

Data curation becomes a structured decision problem when semantically related examples disagree. A curator may need to preserve multiple defensible answers, split a cluster by a missing context, replace one unsupported label, downweight unresolved members, or remove an annotation artifact. These operations are semantically non-interchangeable: applying majority relabeling to genuine ambiguity destroys valid supervision, while preserving an actual factual conflict propagates error. Item-wise noise scores and keep/drop filters (Northcutt et al., 2021; Swayamdipta et al., 2020) therefore omit the variable that determines which intervention is appropriate. The practical cost is sharpest in disagreement-heavy domains, where the wrong operator can improve apparent consistency by collapsing a clinically or semantically valid minority (Plank et al., 2014; Pavlick & Kwiatkowski, 2019; Davani et al., 2022).

We formulate this setting as cluster-local multi-operator repair. Equation (1) selects one of five actions under a finite reviewer budget while charging directly for minority collapse. STATEROUTE implements the formulation as a repair-state interface: a four-way operator-distinguishing bottleneck separates true contradiction, ambiguity, context dependence, and annotation artifact before action selection. The same state pools controller capacity across shared repair semantics, prioritizes review by state uncertainty, and presents evidence aligned with the proposed operator. It is therefore a common decision variable on which the learned policy and reviewer can intervene before a repair

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

changes the training set. A coarser three-state variant tests granularity, while a cost-sensitive policy optimizes the same action space against downstream utility and minority preservation.

Direct action imitation is the decisive counterfactual: it consumes the same cluster evidence and predicts the same repair without exposing intermediate state. We compare StateRoute with both a GBDT Action-Selector and a permutation-invariant Cluster-Transformer Action Selector while fixing retrieval, contradiction scoring, signed clustering, cluster features, surfaced evidence, five-action menu, controller-label pool, tuning budget, review interface, and reviewer time. This contrast evaluates the explicit-state system under shared graph construction, actions, cluster evidence, and review-budget controls. State targets and state-aware review are components of the tested system. A disjoint action-labeled pool and five-fold controller cross-fitting connect pair-level curation quality to cluster-level operator correctness.

The matched comparisons resolve the central mechanism question in favor of explicit repair state. On $2{,}500$ expert-labeled pairs, StateRoute improves pooled contradiction F1 by $+0.030$ over GBDT action imitation and by +0.016 over sequence imitation; the paired intervals remain above zero. It also improves MedQA accuracy by $+1.3$ points and minority retention by $+13.1$ points over GBDT imitation, with 23 additional correct actions on a disjoint 384-cluster test. The direct-controller advantage carries through all five Llama-3-8B downstream datasets; Mistral-7B-Instruct-v0.3 reproduces the state-aware ordering over Untyped-Graph and LLM-as-Judge. Neutral-loss training, three backbone pairs, leave-one-domain-out stress, alternative bottlenecks, disjoint supervision, and external medical and legal-contract panels preserve the central result.

Our contributions are:

- a cluster-local repair objective that joins five semantically distinct operators, explicit review cost, and a minority-collapse penalty in one auditable decision problem;
- a repair-state interface shared by the controller, budget allocator, and review UI, together with a direct-imitation comparison with matched graph construction, evidence, actions, supervision pool, tuning budget, and reviewer-time budget;
- decisive evidence across five datasets and two decoder families, including paired uncertainty, raw retention counts, direct held-out action labels, alternative state granularity, five-fold controller cross-fitting, three backbone pairs, leave-one-domain-out stress, and two external transfer domains.

Together, the cluster-local repair objective, operator-distinguishing state, and matched direct-imitation test determine when explicit intermediate state changes repair choice under finite review budget—a mechanism question that item-wise filtering and direct action policies do not pose.

2 Related Work and Problem Positioning

Noisy-label learning estimates whether individual examples are trustworthy and then reweights, corrects, or removes them (Patrini et al., 2017; Wang et al., 2019; Han et al., 2018; Li et al., 2020; Zhang & Sabuncu, 2018; Chen et al., 2019; Berthon et al., 2021; Song et al., 2023). Dataset cartography and confident learning make these decisions more data-aware (Swayamdipta et al., 2020; Northcutt et al., 2021), but their intervention variable remains attached to an item. StateRoute contributes a cluster-level action variable: the same suspicious item can be retained, contextualized, relabeled, downweighted, or dropped according to the relational cause of its conflict.

Disagreement-aware learning demonstrates that annotation variance can encode legitimate perspectives rather than noise (Plank et al., 2014; Aroyo & Welty, 2015; Pavlick & Kwiatkowski, 2019; Davani et al., 2022; Uma et al., 2021; Gordon et al., 2022; Weerasooriya et al., 2023). Soft label distributions preserve multiplicity but do not choose among cluster repairs, while signed graphs expose compatible and conflicting neighborhoods without selecting an intervention (Kunegis et al., 2010). The present formulation connects these strengths to operator choice: uncertainty preservation is one action, context splitting and selective relabeling are others, and the minority-collapse term makes their consequences explicit. LLM-based evaluation and feedback simulation (Zheng et al., 2023; Dubois et al., 2023) motivate our example-wise judge baseline. Our item-wise judge baseline uses item-level relabeling decisions, which omit repairs that depend on relationships among multiple examples.

Concept bottleneck models factor prediction through supervised, human-interpretable variables that support intervention (Koh et al., 2020). Our repair-state interface makes this decomposition operational for a relational decision object: state records why a cluster conflicts, organizes review priority and evidence, and controls an intervention with measurable minority cost. The closest mechanism-level alternative is direct policy prediction from the same cluster evidence. Our GBDT, latent-code MLP, and Cluster-Transformer selectors instantiate increasingly expressive no-state policies under the same final-action supervision, controller-label pool, and tuning controls. Relative to item-wise noise handling, multiplicity modeling, signed clustering, and direct cluster policies, this work adds a cluster action variable with explicit minority-collapse cost and tests whether exposed repair state changes operator choice once graph construction, evidence, action menu, interface, and budget are matched. Alternative state granularity and cross-fitted supervision further establish that the effect persists beyond one label partition or one reused controller split.

3 StateRoute: A Repair-State Interface for Multi-Operator Curation

The repair scaffold separates evidence construction from action control. Retrieval, contradiction scoring, signed clustering, the five-action set, the surfaced-evidence interface, final-action labels, controller-label pool, and reviewer budget are shared by the matched learned controllers. StateRoute adds supervised repair state before final action selection and uses that state in review allocation and presentation. The scientific variable is therefore operator-distinguishing repair state; controller inputs and evaluation conditions remain fixed, while the Cluster-Transformer separately challenges the state model with a more expressive direct policy. This matching evaluates the combined state-supervised controller and review workflow against direct action imitation.

3.1 Cluster-Local Repair Selection Under Budget and Minority Penalty

Inside that matched scaffold, the substantive decision is not whether an isolated example looks noisy but which non-interchangeable repair operator a contradiction-active cluster should receive. Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ and contradiction clusters $\{C_k\}_{k=1}^{m}$, we choose a repair action $a_k \in \mathcal{A} = \{\text{KEEP-UNCERT, SPLIT-CTX, SELECTIVE-RELABEL, DROP, DOWNWEIGHT}\}$ for each cluster. Let $\mathcal{D}'(\{a_k\})$ be the resulting dataset, $U(\cdot)$ the expected downstream utility of a fixed fine-tuning recipe, and $B_k(a_k)$ the reviewer minutes per 1K examples required to realize action a_k. The P5 panel audited in Appendix B identifies the clinically or semantically valid minority answers $\mathrm{Min}(C_k)$; $M_k(a_k)$ counts how many of those answers action a_k removes from the resulting supervision. We optimize

$$\begin{aligned} \max_{\{a_k\}} \quad & U\big(\mathcal{D}'(\{a_k\})\big) - \lambda_{\mathrm{M}} \sum_k M_k(a_k) \\ \text{s.t.} \quad & \sum_k B_k(a_k) \le B_{\max}. \end{aligned} \tag{1}$$

Equation (1) makes two operational consequences explicit: actions are cluster-local, and canonicalizing a cluster with a valid minority answer incurs direct cost. Score-then-filter baselines recover the two-action case $\mathcal{A} = \{\text{KEEP, DROP}\}$ with $\lambda_{\mathrm{M}} = 0$.

We instantiate Eq. (1) with a state-aware policy and a cost-sensitive extension. StateRoute predicts repair state, maps it to an action prior, and spends the 12.4 min/1K surfaced-cluster budget on the highest-entropy decisions first. StateRoute-Opt scores each candidate action with $\hat{U}_k(a) = \hat{g}_\phi(C_k, a) - \hat{m}_\psi(C_k, a)$, where $\hat{g}_\phi$ predicts downstream utility and $\hat{m}_\psi$ predicts minority deletion cost from the same evidence. We tune λ_{M} on a 640-cluster dev fold and allocate budget with a $0/1$-knapsack Lagrangian greedy. At matched budget, StateRoute-Opt reaches 0.83 pooled pair-level contradiction F1, with a paired $+0.013$ effect over StateRoute ($p = 0.018$); the displayed 0.82 and 0.83 values are rounded means of that paired comparison.

3.2 Signed Evidence and Cluster Extraction

StateRoute retrieves semantically related neighbors, scores pairwise contradiction, extracts contradiction-active signed clusters, and routes each cluster to a repair operator. We encode each

example (x_i, y_i) with frozen Sentence-T5 embeddings (Ni et al., 2022) and retrieve $k = 50$ neighbors with a FAISS inner-product index (Johnson et al., 2021), reducing comparison cost from $O(N^2)$ to $O(Nk)$. For each retrieved pair (i, j), a fine-tuned DeBERTa-v3-base model (He et al., 2023) computes $p_\theta^-(i, j) \in [0, 1]$ from $[x_i; y_i; \texttt{[SEP]}; x_j; y_j]$. The scorer uses 10,400 expert-labeled pairs (8,000 train, 1,200 dev, 1,200 calibration) spanning factual QA, instruction following, and three natural-disagreement domains; Appendix C gives the label schema.

Every retrieved pair is retained as signed evidence, not only the high-contradiction ones. We assign the signed edge weight

$$w(i, j) = 1 - 2p_\theta^-(i, j), \tag{2}$$

so positive weights indicate compatibility and negative weights indicate contradiction. The thresholded negative-edge set $E^- = \{(i, j) : p_\theta^-(i, j) > \tau_s\}$ with $\tau_s = 0.5$ defines contradiction-active regions that need repair, while positive edges preserve local agreement structure for clustering and training.

Within each connected component of E^-, we build the local signed graph $\mathcal{G} = (V, E, w)$ and run signed spectral clustering (Kunegis et al., 2010). Let $\mathbf{W}$ be the signed adjacency matrix and $\mathbf{D}^{\pm}_{ii} = \sum_j |w(i, j)|$. We compute the smallest eigenvectors of the normalized signed Laplacian $(\mathbf{D}^{\pm})^{-1/2}(\mathbf{D}^{\pm} - \mathbf{W})(\mathbf{D}^{\pm})^{-1/2}$, select the number of clusters by the eigengap heuristic (Ng et al., 2001; von Luxburg, 2007), and apply k-means in the spectral space. Components with no negative edges pass through unchanged. This separation matters because the model should not spend review budget on dense but internally consistent neighborhoods.

3.3 Repair-State Bottleneck and Action Policy

We use a four-way repair-sufficient factorization aligned with the five-action menu: true contradiction $\rightarrow$ SELECTIVE-RELABEL, ambiguity $\rightarrow$ KEEP-UNCERT, context dependence $\rightarrow$ SPLIT-CTX, and annotation artifact $\rightarrow$ DROP, with DOWNWEIGHT as the uncertainty fallback. A gradient-boosted state model consumes answer entropy, signed-edge statistics, cluster size, lexical overlap, and metadata cues and reaches 83.6% macro-F1 under the P3 controller protocol. Clusters that share state can reuse controller capacity, budget priority, and reviewer cues before the final operator is chosen. Appendix Q shows that most of the gain survives a coarser three-way merge, while high-capacity no-type controllers remain below explicit state on pair F1, action macro-F1, MedQA accuracy, and minority retention.

Writing e_k for the surfaced signed evidence, the action policy factorizes as

$$p(a_k \mid C_k, e_k) = \sum_{z \in \mathcal{Z}} p_\theta(z \mid C_k, e_k)\, p_\phi(a_k \mid z, C_k, e_k), \tag{3}$$

where $\mathcal{Z}$ contains the four repair states. The direct selectors estimate $p_\psi(a_k \mid C_k, e_k)$ in one step from the same evidence and action labels. Equation (3) makes the scientific contrast explicit: StateRoute exposes an operator-distinguishing variable that can be calibrated, reviewed, and corrected before the final action is committed.

Repairs are task-aware. For multiple-choice or class-labeled data, KEEP-UNCERT becomes label smoothing with contradiction-dependent coefficient α_i. For free-form generation, it becomes a normalized mixture over the retained cluster answers so that ambiguous clusters keep multiple valid responses in the supervision signal. SPLIT-CTX prefixes the inferred context string to the prompt or question, SELECTIVE-RELABEL replaces only the unsupported answers rather than collapsing the entire cluster, and DOWNWEIGHT reduces the loss contribution of examples that remain unresolved after routing. Reviewers can override the predicted state or repair when the surfaced evidence is insufficient.

The granularity study in Appendix Q merges true contradiction and annotation artifact into an error-like state; its final action head still separates SELECTIVE-RELABEL from DROP using signed artifact features. This three-state variant preserves most of the four-state gain, whereas both latent-code and sequence no-type controllers remain lower on pair-level F1, action macro-F1, and minority retention. The result identifies operator-distinguishing state as the useful abstraction while quantifying sensitivity to its partition.

3.4 Human Review Budget and Training Integration

For human-in-the-loop curation, StateRoute shows the reviewer the highest-confidence contradicting pair, predicted state, and proposed repair. This focused interface reduces pooled median review time from 31.2 s in a full-cluster pilot to 7.1 s per surfaced cluster. Multiplying by 105 surfaced clusters per 1K examples gives the matched *surfaced-cluster budget* of 12.4 min/1K used in the main tables. The end-to-end audit adds setup, quality control, disagreement resolution, and tail time for the same workflow and measures 16.2 min/1K (Appendix I).

The curated dataset is then used for standard supervised fine-tuning with two modifications. For classification-style tasks, examples routed to KEEP-UNCERT use the softened target $\tilde{y}_i = (1 - \alpha_i)y_i + \alpha_i u$, where u is the uniform class distribution and α_i scales with contradiction intensity. For generative tasks such as WizardLM, we implement the same idea by sampling one retained response per epoch in proportion to the cluster weights, so ambiguous clusters continue to contribute multiple acceptable responses over training. We also add a consistency regularizer on the agreement edges $E^+ = \{(i, j) : w(i, j) > 0\}$:

$$\mathcal{L}_{\text{consist}} = \lambda \sum_{(i,j)\in E^+} \|p_\theta(\cdot \mid x_i) - p_\theta(\cdot \mid x_j)\|_2^2 , \tag{4}$$

with $\lambda = 0.1$. The term preserves compatible neighborhoods after contradictory regions have been repaired.

4 Controlled Evaluation Setup

4.1 Datasets

We evaluate on five English datasets chosen to expose different contradiction regimes. WizardLM-52K (Xu et al., 2023) is a 52K instruction-tuning corpus in the broader instruction-tuning lineage (Wei et al., 2022); TriviaQA-Contra starts from TriviaQA (Joshi et al., 2017) and injects contradictions while preserving the original answer distribution, and we report the 10% corruption setting. MedQA (Jin et al., 2021) supplies the four-option medical question-answering setting. ChaosNLI-Disagree and DynaSent-R2-Disagree retain the non-unanimous slices of ChaosNLI (Nie et al., 2020) and DynaSent (Potts et al., 2021). Contradiction-active cluster rates on the two added disagreement corpora are 9.8% and 6.4%; per-domain numbers appear in Appendix D. The disagreement-focused slices ensure that the pooled result tests operator choice under semantic conflict as well as controlled corruption.

4.2 Baselines, Evaluation, and Reproducibility

We compare Raw data, MinHash Dedup-only, Confident Learning (Northcutt et al., 2021), Dataset Cartography (Swayamdipta et al., 2020), LLM-as-Judge (Zheng et al., 2023), and four reductions of the same cluster-repair problem. The LLM-as-Judge baseline is the GPT-4o item-wise implementation detailed in Appendix C; tables abbreviate it as LLM-Judge. Soft-Label Distributional preserves ambiguity without clustering. Untyped-Graph applies one global repair rule to the shared signed graph. GBDT Action-Selector predicts a_k directly from the same P3 final-action labels and cluster features as the state model. Cluster-Transformer Action Selector is a permutation-invariant set-attention controller (Lee et al., 2019) with 4 self-attention layers, hidden width 256, 8 heads, and 3.8M trainable parameters; it consumes the same retrieved members, signed pair scores, metadata features, surfaced evidence, and five actions without a state label or state-prediction loss. All human-reviewed methods use the same 12.4 min/1K rubric, interface, and reviewer exposure.

Capacity and tuning controls are prespecified. The state GBDT and direct GBDT each use 500 trees, maximum depth 4, and learning rate 0.03; both receive the same 40-trial tuning budget. The Cluster-Transformer uses AdamW at 3e-4, effective batch 32, and at most 80 epochs with 8-epoch patience, under the same trial budget and a 2 GPU-hour cap. Its mean training time is 17 minutes on one NVIDIA A100. The latent-code MLP and three-state controller in Appendix Q use the same P3 cluster pool, final-action labels, features, trial count, and checkpoint rule; explicit-state controllers additionally receive their corresponding state targets.

The DeBERTa-v3-base scorer uses AdamW with learning rate 2e-5, 6% linear warmup followed by linear decay, effective batch 64, and 4 epochs; the checkpoint with maximum P1-dev macro-F1 is calibrated once on P1-calibration. A scorer run takes 48 minutes on one A100. Downstream full-parameter SFT for Llama-3-8B (Meta, 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) uses AdamW at 2e-5, cosine decay, 3% warmup, effective batch 128, bfloat16, and 3 epochs. Checkpoints are selected independently for each seed by validation task metric. Mean run time is 2.7 hours for Llama and 2.4 hours for Mistral on 8 A100-80GB GPUs.

Experiments use PyTorch 2.4.1, Transformers 4.45.2, CUDA 12.1, and FAISS 1.8.0. The same 7 seed identities initialize every paired training run. Each seed selects its own checkpoint; downstream tables report mean and standard deviation across seeds, and paired seed bootstraps resample aligned method pairs. Pair-level and cluster-level 95% intervals instead resample aligned examples or clusters $10{,}000$ times, so sampling uncertainty and training-seed variability remain distinct.

Evaluation uses AlpacaEval win rate (Li et al., 2023) for WizardLM, TriviaQA exact match, MedQA four-option accuracy, three-way accuracy on the $1{,}280$-item ChaosNLI test subset with $> 80\%$ consensus, and DynaSent-R2 accuracy. Ten comparisons are confirmatory: pooled pair-level F1 and MedQA accuracy for StateRoute against LLM-as-Judge and Untyped-Graph, plus pooled F1, MedQA accuracy, and minority retention against each direct-action selector. We apply BH-FDR at $q = 0.05$ (Benjamini & Hochberg, 1995); each reported adjusted p remains below 0.01. Per-dataset, per-action, transfer, and backbone contrasts are labeled exploratory.

P1 trains and calibrates the contradiction scorer; disjoint P2 provides 500 evaluation pairs per dataset. P3 supplies state labels and final-action labels, while five-fold cross-fitting in Appendix R holds every controller-labeled cluster out exactly once. The independent 384-cluster action pool supplies direct operator labels. Of the 312 P5 MedQA audit clusters, 252 contain a clinically valid minority answer and define the retention denominator. The two external audits recalibrate scalar thresholds on separate 32-cluster slices and evaluate on disjoint medical and legal-contract panels.

5 Results

Explicit repair state improves both curation quality and operator choice under exact scaffold matching. Table 1 provides the identification result at two complementary units: Panel A measures pair-level contradiction resolution across all five datasets, and Panel B measures cluster-level action correctness on disjoint controller labels. Table 3 then tests whether the resulting repairs improve trained models, while the appendices stress state granularity, controller family, backbone, supervision split, and domain shift.

5.1 Primary Test of the Operator-Distinguishing Bottleneck

Panel A compares successive components of the repair scaffold. Soft-Label Distributional tests ambiguity preservation, Untyped-Graph tests graph structure with one global policy, and the GBDT and Cluster-Transformer selectors predict the same five actions from the same evidence without explicit state. Every human-assisted row shares the 12.4 min/1K surfaced-cluster budget; Appendix I measures the corresponding complete workflow at 16.2 min/1K. Panel B evaluates the controllers directly on 384 disjoint clusters with adjudicated gold actions.

Graph structure supplies part of the gain: Untyped-Graph reaches 0.76 pair F1. Direct action learning raises this to 0.79 with the GBDT and 0.80 with the Cluster-Transformer. Explicit state retains a statistically resolved residual over both, reaching 0.82 with $+0.030$ F1 over GBDT imitation and +0.016 over sequence imitation. MedQA and retention gains are $+1.3$ and $+13.1$ points over GBDT imitation, and +0.7 and +4.8 points over sequence imitation; all six confirmatory direct-controller contrasts survive BH-FDR. On the 252 valid-minority denominator, StateRoute preserves 220/252 clusters, compared with 187/252 for GBDT and 208/252 for the Cluster-Transformer.

Panel B identifies where the residual enters. StateRoute selects 291/384 actions correctly, 23 more than GBDT imitation; the accuracy difference is +0.060 (95% CI [+0.026,+0.094], $p = 0.003$), and the macro-F1 difference is +0.070 (CI [+0.035,+0.105], $p = 0.002$). Against the Cluster-Transformer, StateRoute adds 11 correct actions and +0.030 macro-F1 (CI [+0.009,+0.052], $p = 0.009$).

Table 1: Matched comparison of explicit repair state. Panel A reports pooled pair-level F1 on 2,500 P2 pairs, MedQA accuracy over 7 seeds, and minority retention; Panel B reports direct cluster-action results. Paired StateRoute−GBDT differences are +0.030 F1 [+0.014, +0.046], $p = 0.002$; +1.3 MedQA [+0.7, +1.9], $p = 0.004$; and +13.1 retention [+8.6, +17.5], $p = 0.003$. Paired StateRoute−sequence differences are +0.016 F1 [+0.005, +0.028], $p = 0.008$; +0.7 MedQA [+0.2, +1.2], $p = 0.008$; and +4.8 retention [+1.2, +8.4], $p = 0.006$. Bracketed differences are 95% paired-bootstrap intervals from unrounded aligned scores: pair F1, MedQA, and retention resample aligned pairs, seeds, and clusters, respectively.

Panel A: five-dataset pair-level curation and MedQA utility

Method	Min/1K	Pair F1 [95% CI]	MedQA (%)	Min. Ret. (%)
Zero-cost mechanism isolation				
Soft-Label Dist.	0	0.66 [0.64, 0.68]	58.9±0.6	79.4
LLM-Judge (auto)	0	0.67 [0.65, 0.69]	58.3±0.6	61.2
GBDT Action-Selector (auto)	0	0.70 [0.68, 0.72]	59.4±0.6	74.8
StateRoute (auto)	0	0.74 [0.72, 0.76]	60.3±0.5	82.5
Matched budget and interface (12.4 min/1K)				
LLM-Judge +human	12.4	0.72 [0.70, 0.74]	58.7±0.5	68.7
Untyped-Graph +human	12.4	0.76 [0.74, 0.78]	59.8±0.5	68.4
GBDT Action-Selector +human	12.4	0.79 [0.77, 0.81]	60.4±0.5	74.2
Cluster-Transformer +human	12.4	0.80 [0.78, 0.82]	61.0±0.5	82.5
StateRoute +human	12.4	0.82 [0.80, 0.83]	61.7±0.5	87.3
StateRoute-Opt +human	12.4	0.83 [0.82, 0.85]	62.2±0.5	87.9

Panel B: direct cluster-action endpoint

Controller	Pool	Action Acc.	Action Macro-F1	Correct actions
GBDT Action-Selector	384	0.70	0.65	268/384
Cluster-Transformer	384	0.73	0.69	280/384
StateRoute	384	0.76	0.72	291/384
StateRoute-Opt	384	0.78	0.74	300/384

Appendix P shows that the largest gains occur on KEEP-UNCERT and SPLIT-CTX, the two operators most dependent on separating ambiguity from missing context.

Agreement across pair resolution, direct action correctness, downstream utility, and raw minority preservation makes the result decision-relevant: the state layer changes which repair is selected, those decisions resolve more contradictions, and the curated data improves trained models without erasing valid alternatives. The three-state controller retains most of the gain (Appendix Q), and the ordering survives five-fold disjoint controller evaluation (Appendix R), locating the effect in explicit operator-distinguishing state rather than one taxonomy name or one reused split.

5.2 From Repair Decisions to Trained-Model Utility

Automatic StateRoute reaches 0.74 pooled pair-level F1, improving over LLM-as-Judge (0.67) and Soft-Label Distributional (0.66). Expert review raises StateRoute and StateRoute-Opt to 0.82 and 0.83. The precision-led improvement indicates that relational evidence suppresses isolated false contradictions while the state-aware action policy preserves recall.

Table 3 tests whether the identified repair advantage survives model training. Panel A carries both direct-action controllers through all five Llama-3-8B tasks; Panel B confirms the StateRoute>Untyped-Graph>LLM-as-Judge ordering with Mistral-7B-Instruct-v0.3.

StateRoute exceeds the strongest no-type row on every Llama-3-8B task. Its gains over the Cluster-Transformer range from +0.3 points on TriviaQA to +0.7 on MedQA; gains over the GBDT selector are +0.6–+1.3 points and peak on the disagreement-heavy tasks. StateRoute-Opt adds another 0.2–0.5 points. The Mistral panel preserves the same ranking, and Appendix F shows that

Table 2: Contradiction-detection performance on the pooled $2{,}500$ expert-labeled pairs (500 per dataset). Human Min/1K reports reviewer time for methods with a human-review phase. Dedup-only is listed as a preprocessing reference and does not define contradiction-detection precision/recall/F1 on its own. Per-dataset F1 with 95% bootstrap CIs is in Appendix D.

Method	Precision	Recall	F1	Min/1K
Dedup-only	—	—	—	0
Confident	0.62±.03	0.48±.04	0.54±.03	0
Cartography	0.58±.04	0.52±.03	0.55±.03	0
Soft-Label Dist.	0.69±.03	0.63±.03	0.66±.02	0
LLM-Judge	0.71±.02	0.63±.03	0.67±.02	0
StateRoute (auto)	0.78±.02	0.71±.02	0.74±.02	0
StateRoute +human	0.89±.01	0.76±.02	0.82±.01	12.4
StateRoute-Opt +human	0.90±.01	0.77±.02	0.83±.01	12.4

Table 3: Matched-budget downstream performance (%) after fine-tuning on curated data. Every row uses the same 12.4 reviewer min/1K and interface. Columns report WizardLM→AlpacaEval, TriviaQA, MedQA, ChaosNLI, and DynaSent-R2; mean ± standard deviation over 7 shared seeds.

Panel A: Llama-3-8B, matched 12.4 min/1K

Curated by	Wiz→AE	Tri	Med	Cha	Dyn
Confident Learning +human	72.5±0.4	70.1±0.4	56.4±0.6	60.2±0.5	64.7±0.5
Dataset Cartography +human	73.1±0.4	71.0±0.4	57.1±0.5	61.0±0.4	65.3±0.5
LLM-Judge +human	74.9±0.3	72.9±0.3	58.7±0.5	63.1±0.4	66.8±0.4
Untyped-Graph +human	75.3±0.3	73.2±0.3	59.8±0.5	63.9±0.4	67.3±0.4
GBDT Action-Selector +human	75.5±0.3	73.5±0.3	60.4±0.5	64.2±0.4	67.7±0.4
Cluster-Transformer +human	75.8±0.3	73.8±0.3	61.0±0.5	64.8±0.4	68.2±0.4
StateRoute +human	76.2±0.3	74.1±0.3	61.7±0.5	65.3±0.4	68.8±0.4
StateRoute-Opt +human	76.5±0.3	74.3±0.3	62.2±0.5	65.6±0.4	69.1±0.4

Panel B: Mistral-7B-Instruct-v0.3, matched 12.4 min/1K

Curated by	Wiz→AE	Tri	Med	Cha	Dyn
LLM-Judge +human	73.4±0.4	71.6±0.3	57.0±0.5	61.8±0.4	65.4±0.4
Untyped-Graph +human	73.6±0.3	72.0±0.3	57.8±0.5	62.3±0.4	65.9±0.4
StateRoute +human	74.3±0.3	72.8±0.3	59.2±0.5	63.7±0.4	67.1±0.4

most of the advantage remains with one neutral SFT loss. Together with the action endpoint, these results connect state-aware operator choice to utility after training rather than to a pair-level metric alone.

5.3 Stress Tests, Cost Accounting, and Cross-Domain Transfer

Across leave-one-domain-out training, backbone swaps, five-fold controller cross-fitting, and frozen medical and legal transfer, explicit state remains above the strongest matched no-type controller on the endpoints reported for each stress test. StateRoute stays above GBDT and sequence selectors on all five held-out domains (Appendix O); explicit-state-minus-sequence pair F1 remains positive under Sentence-T5+DeBERTa, MPNet+RoBERTa-MNLI, and BGE+BART-MNLI (Appendix K); and out-of-fold action and retention gains persist under cross-fitting (Appendix R). The shifted ContractNLI audit preserves a +0.030 action macro-F1 advantage over sequence imitation despite longer inputs, legal terminology, and document-level conflict structure (Appendix S).

Figure 1 reports measured operating points at zero review and 12.4 min/1K. The 2 h 41 min wall-clock run corresponds to 16.1 min/1K, matching the independently itemized 16.2 min/1K end-to-end audit (Appendix I). The ranking is unchanged under either accounting scheme, linking the quality gain to matched reviewer exposure rather than favorable cost normalization.

6 DISCUSSION

The experiments support a design rule for human-in-the-loop curation: when the action set contains semantically different repairs, use one repair-state interface to organize policy factorization, review priority, and reviewer evidence, then test that state against direct imitation under an exactly matched scaffold. State is most valuable on ambiguity and context dependence, where similar conflict scores can require opposite interventions; correspondingly, direct policies underperform most on KEEP-UNCERT and SPLIT-CTX in the held-out action test. The three-state result shows that useful structure survives coarser factorization, while the sequence selector and cross-fitted evaluation show that the residual persists against a stronger no-type policy and disjoint predictions. The state-aware and direct controllers share a label pool and final-action labels, but only the explicit-state controllers receive state targets. The comparison therefore measures the combined effects of state supervision, factorization, and state-aware review; it does not separately identify these contributions. The evaluated scope comprises five English datasets, expert-reviewed contradiction clusters, two 7–8B decoder families, and medical and legal transfer panels; within this setting, explicit state consistently improves action choice, downstream utility, and preservation of valid alternatives.

7 CONCLUSION

STATEROUTE turns contradiction-heavy curation into budgeted multi-operator repair with an explicit cost for minority collapse. Its repair-state interface gives the controller and reviewer a shared, correctable account of why a cluster conflicts before either commits an action. Under fixed graph, evidence, actions, supervision pool, interface, tuning budget, and reviewer exposure, this interface outperforms GBDT and Cluster-Transformer action imitation on pooled contradiction resolution, direct held-out action labels, and all five Llama-3-8B downstream tasks. Mistral-7B preserves the state-over-graph ranking on all five tasks, while frozen medical and legal panels retain the state-over-strongest-no-type ordering on cluster-action and minority-retention endpoints. Explicit repair state therefore improves action choice, contradiction resolution, and minority preservation under a matched review budget, with gains for the state-supervised system under shared graph and review-budget controls.

AI USE STATEMENT

Generative AI was used solely for manuscript drafting, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

REFERENCES

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015. doi: 10.1609/aimag.v36i1.2564. URL https://doi.org/10.1609/aimag.v36i1.2564.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 825–836, 2021. URL https://proceedings.mlr.press/v139/berthon21a.html.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 1062–1070, 2019. URL https://proceedings.mlr.press/v97/chen19g.html.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449. URL https://aclanthology.org/2022.tacl-1.6/.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 30039–30069, 2023. doi: 10.52202/075280-1308. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022. doi: 10.1145/3491102.3502004. URL https://doi.org/10.1145/3491102.3502004.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825. Instruct-v0.3 checkpoint: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Yuta Koreeda and Christopher D. Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W. De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pp. 559–570, 2010. doi: 10.1137/1.9781611972801.49. URL https://doi.org/10.1137/1.9781611972801.49.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. GitHub repository, 2023. URL https://github.com/tatsu-lab/alpaca_eval.

Meta. Meta Llama 3 8B: Model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3-8B.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 14, pp. 849–856, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, 2022. doi: 10.18653/v1/2022.findings-acl.146. URL https://aclanthology.org/2022.findings-acl.146/.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9131–9143, 2020. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734/.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: 10.1613/jair.1.12125. URL https://doi.org/10.1613/jair.1.12125.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1944–1952, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Patrini_Making_Deep_Neural_CVPR_2017_paper.html.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019. doi: 10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043/.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 507–511, 2014. doi: 10.3115/v1/P14-2083. URL https://aclanthology.org/P14-2083/.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2388–2404, 2021. doi: 10.18653/v1/2021.acl-long.186. URL https://aclanthology.org/2021.acl-long.186/.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527. URL https://doi.org/10.1109/TNNLS.2022.3152527.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. doi: 10.1613/jair.1.12752. URL https://doi.org/10.1613/jair.1.12752.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007. doi: 10.1007/s11222-007-9033-z. URL https://doi.org/10.1007/s11222-007-9033-z.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 322–330, 2019. doi: 10.1109/ICCV.2019.00041. URL https://doi.org/10.1109/ICCV.2019.00041.

Tharindu Cyril Weerasooriya, Alexander G. Ororbia, Raj B. Bhensadadia, Ashiqur R. KhudaBukhsh, and Christopher M. Homan. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4679–4695, 2023. doi: 10.18653/v1/2023.findings-acl.287. URL https://aclanthology.org/2023.findings-acl.287/.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://research.google/pubs/finetuned-language-models-are-zero-shot-learners/.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. WizardLM: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244v1*, 2023. URL https://arxiv.org/abs/2304.12244v1.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Additional Appendix Diagnostics

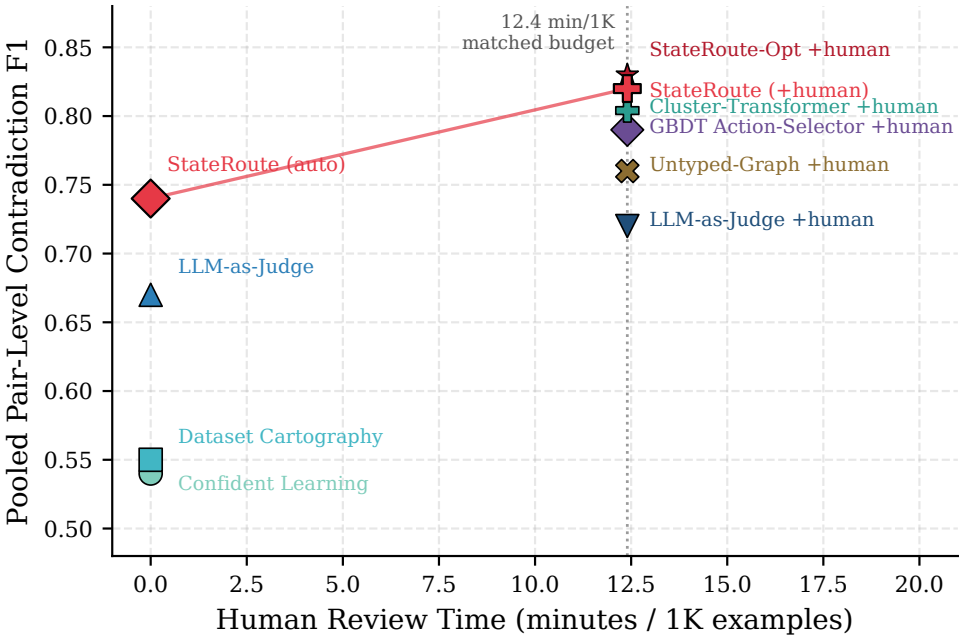


Figure 1: Measured operating points only: zero-cost baselines and all matched-budget controller families at 12.4 min/1K. The segment connects the automatic and human-assisted StateRoute evaluations; training cost is excluded and no interpolation is claimed.

Table 4: Pooled P2 pair-level F1 on a 0–1 scale and LODO held-out-domain accuracy (%). Each domain column holds that domain out while fitting the curation stack. All rows use 12.4 min/1K; mean ± standard deviation over 7 seeds.

Method	Pair F1	Med (%)	Tri (%)	Wiz (%)	Cha (%)	Dyn (%)
Confident	0.59±.03	41±3	46±3	48±3	44±3	45±3
Cartog.	0.60±.03	43±3	48±3	49±3	46±3	47±3
LLM-J.	0.72±.02	58±3	63±2	66±2	60±2	61±2
Untyped	0.76±.02	63±2	68±2	70±2	65±2	66±2
StateRoute	0.82±.01	68±2	71±2	72±2	70±2	69±2

Table 4 juxtaposes pooled pair-level F1 with held-out-domain scores. StateRoute ranks highest in each column. The pooled and domain-specific columns use different metrics, so their magnitudes should not be compared directly.

B Annotation-Pool and Split Audit

This appendix reconciles the nine internal development and audit pools that define the 857-hour accounting boundary. The disjoint direct-action pool and the two external transfer panels are documented with their protocols in Appendices P, L, and S; their labels and labor are outside this internal subtotal.

Pool boundaries and leakage controls are explicit. P1 and P2 were drawn with disjoint source IDs from each dataset before any model was trained; a post-hoc SHA-256 check on the (x, y) pair IDs confirmed 0 shared pairs across all five domains. The held-out-domain slices in Table 4 partition P2 by source domain rather than introducing a new evaluation pool. P6 is an explicit subset of P1 that excludes the held-out domain for each of the five LODO folds; its 48 hours are fold-specific

Table 5: Internal annotation and evaluation pools P1–P9. Overlap means the fraction of clusters that also appear in another listed pool within the same dataset. Parenthetical percentages for minority pools are sampling shares of each domain's contradiction-active cluster frame. "Ann. hours" reports expert time within the internal 857-hour accounting boundary; direct-action and external-transfer pools are accounted separately.

Pool	Purpose	Size	Overlap	Raters	Ann. hours
P1: Scorer train	Contradiction-scorer training	10,400 pairs	0%	2	369
P2: Pooled eval	Pooled F1 evaluation	2,500 pairs	0%	3	182
P3: Controller sup.	State/action supervision	660 clusters	0%	3	83
P4: Protocol audit	Timing and κ audit	500 clusters	0%	3	60
P5: MedQA minority	$\hat{m}_\psi$ fit + MedQA audit	312 clusters (14.7%)	0%	3	41
P6: LODO QA review	Fold-specific QA on reused P1 subsets	5 subsets (8,320 pairs each)	P1 subset	2	48
P7: Backbone eval	Backbone-swap F1 eval	2,500 pairs	P2 reused	3 (reuse)	0
P8: ChaosNLI minority	$\hat{m}_\psi$ fit + support audit	286 clusters (16.1%)	0%	3	36
P9: DynaSent minority	$\hat{m}_\psi$ fit + support audit	204 clusters (15.2%)	0%	3	26
New units / hours subtotal		14,862 units			845
Workflow + QC overhead					12
Total					857

QA re-review labor, with two staff research engineers sanity-checking reused P1 subsets before each LODO refit. The new-unit count in Table 5 excludes P6 and P7, while the 845-hour subtotal includes P6 because it aggregates all pre-overhead annotation and QA labor. P7 reuses the labels in P2 while evaluating a different contradiction scorer, and Table 13 accordingly reports within-backbone comparisons under the original label budget. P3 supplies four-way labels for state-head learnability checks and final-action labels for both direct-action controllers; five-fold cross-fitting evaluates every controller prediction out of fold. P5, P8, and P9 are mutually disjoint minority panels built on MedQA-Disagree, ChaosNLI-Disagree, and DynaSent-R2-Disagree. They provide in-ecosystem tests of the learned minority-cost surrogate, while the disjoint medical and legal panels evaluate transfer.

The arithmetic also reconciles cleanly. The 12.4 min/1K operating point in the main text comes from the P4 pooled median review time 7.1 s and the full-pipeline pooled surfaced-cluster rate of 105 clusters per 1,000 examples: $7.1 \times 105 = 745.5$ s$= 12.425$ min≈ 12.4 min. The total annotation effort reconciles as $369 + 182 + 83 + 60 + 41 + 48 + 0 + 36 + 26 + 12 = 857$ expert-hours; no line item in this table is counted twice. The new-unit bookkeeping aggregates pair-level and cluster-level supervision solely for workload accounting, whereas the adjacent hour subtotal includes QA on reused labels. The unit arithmetic reconciles exactly: $10{,}400{+}2{,}500{+}660{+}500{+}312{+}286{+}204 = 14{,}862$ units (P6 is a subset of P1 and P7 reuses P2, so neither contributes new labels).

The controller-supervision pool preserves the repair-state prior. The pooled distribution is $34.1/28.7/22.4/14.8\%$ and sums to 100.0%. The P3 label counts ($222/191/148/99$ on 660 clusters) preserve the same ordering and round to $33.6/28.9/22.4/15.0\%$, within one sigma of the pooled distribution under the hypergeometric approximation.

C Dataset Construction and Policy Detail

This section describes the dataset construction rules, label schema, and repair implementations used in the evaluation. Pair labels supervise contradiction scoring, while cluster labels supervise repair state and action selection.

Dataset construction is deterministic once the source splits are fixed. TriviaQA-Contra starts from the standard TriviaQA training split and injects contradictions by replacing one answer alias with a same-entity-type confuser drawn from another source item, while preserving the original question text, passage provenance, and answer-length bucket; the 5%, 10%, and 15% corruption rates are generated by the same script with different seeds, and the paper reports the 10% condition. ChaosNLI-Disagree and DynaSent-R2-Disagree introduce no new annotations: they are deterministic filters of the public datasets using entropy > 0.8 nats for ChaosNLI and non-unanimity under the five-annotator label distribution for DynaSent-R2. WizardLM-52K is left in its original training form except for source-ID deduplication before contradiction retrieval. Pool boundaries are enforced after these construction steps, so the paper's later disjointness claims are downstream of a fixed source-ID split rather than a manual clean-up.

Pair and cluster labels supervise distinct decisions. Pool P1 contains $10{,}400$ pair labels, split into $8{,}000$ train, $1{,}200$ dev, and $1{,}200$ calibration pairs for the contradiction scorer. Annotators assign *compatible*, *contradiction-active*, or *artifact*; the binary scorer maps compatible to 0 and both repair-requiring labels to 1, while retaining the finer distinction for cluster state. Pool P3 supplies four-way state labels and final-action labels for the GBDT and sequence controllers. Five-fold cross-fitting in Appendix R evaluates every P3 cluster out of fold, separating controller fitting from scored predictions.

The GPT-4o baseline, SPLIT-CTX, and SELECTIVE-RELABEL are also fixed. The GPT-4o baseline uses one deterministic call per example with temperature 0.0, JSON output, and no cluster context: the prompt contains task metadata, the example input, the current answer, and a short rubric asking whether to keep, relabel, or drop the item and, if relabeling, to emit a single replacement answer. SPLIT-CTX is likewise deterministic: the system extracts the highest-margin cue span among time, jurisdiction, guideline, or sense-disambiguating mentions already present in the cluster, and it falls back to KEEP-UNCERT when no cue remains stable across the retained members. SELECTIVE-RELABEL preserves the supported cluster medoid answer and rewrites only those items whose answer loses pairwise support against every retained alternative under the signed graph; supported minority answers are therefore not overwritten merely for being minority answers. The reviewer interface exposes the proposed cue string or relabel target before commitment, making the override statistics in Table 10 directly interpretable from the stated decision logic.

D Per-Dataset Contradiction-Detection Results

Table 6 reports per-dataset contradiction-detection F1 with 95% paired-bootstrap CIs ($10{,}000$ resamples) on the 500 expert-labeled pairs that each dataset contributes to the pooled P2 evaluation. The ordering StateRoute +human $>$ Untyped-Graph +human $>$ LLM-as-Judge +human holds on all five datasets. Automatic LLM-as-Judge matches or exceeds Soft-Label Distributional on these pair-level scores. The $+$human rows are matched at 12.4 min/1K; the automatic rows are zero-cost comparators.

Unweighted domain means numerically agree with the reported pooled F1 after rounding: StateRoute+human averages $0.816 \approx 0.82$, StateRoute-Opt+human $0.826 \approx 0.83$, Untyped-Graph+human 0.760, LLM-as-Judge+human 0.720, Soft-Label Distributional 0.660, LLM-as-Judge (auto) 0.670, and StateRoute (auto) 0.740. This numerical agreement does not establish the pooling rule: pooled F1 aggregates pair-level confusion counts, whereas the unweighted mean averages domain F1 scores. MedQA, ChaosNLI, and DynaSent remain harder than TriviaQA and WizardLM for every method, yet the ordering is stable in each domain. The per-domain view therefore confirms that the pooled result spans both controlled corruption and natural disagreement.

Table 6: Per-dataset contradiction-detection F1 (95% paired-bootstrap CI in brackets). 500 expert-labeled pairs per dataset (P2 partitioned by source domain).

Method	Min/ 1K	MedQA	TriviaQA	WizardLM	ChaosNLI	DynaSent
Confident	0	0.52 [0.48, 0.56]	0.55 [0.51, 0.59]	0.58 [0.54, 0.62]	0.53 [0.49, 0.57]	0.52 [0.48, 0.56]
Cartography	0	0.54 [0.50, 0.58]	0.56 [0.52, 0.60]	0.59 [0.55, 0.63]	0.55 [0.51, 0.59]	0.53 [0.49, 0.57]
Soft-Label Dist.	0	0.65 [0.61, 0.69]	0.68 [0.64, 0.72]	0.68 [0.64, 0.72]	0.65 [0.61, 0.69]	0.64 [0.60, 0.68]
LLM-Judge (auto)	0	0.66 [0.62, 0.70]	0.69 [0.65, 0.73]	0.70 [0.66, 0.74]	0.66 [0.62, 0.70]	0.64 [0.60, 0.68]
STATEROUTE (auto)	0	0.73 [0.70, 0.76]	0.76 [0.73, 0.79]	0.77 [0.74, 0.80]	0.73 [0.70, 0.76]	0.71 [0.68, 0.74]
LLM-Judge +human	12.4	0.71 [0.68, 0.74]	0.74 [0.71, 0.77]	0.76 [0.73, 0.79]	0.70 [0.67, 0.73]	0.69 [0.66, 0.72]
Untyped-Graph +human	12.4	0.75 [0.72, 0.78]	0.77 [0.74, 0.80]	0.79 [0.76, 0.82]	0.74 [0.71, 0.77]	0.75 [0.72, 0.78]
STATEROUTE +human	12.4	0.81 [0.78, 0.83]	0.83 [0.81, 0.85]	0.85 [0.83, 0.87]	0.80 [0.78, 0.82]	0.79 [0.76, 0.82]
STATEROUTE-Opt +human	12.4	0.82 [0.80, 0.84]	0.84 [0.82, 0.86]	0.86 [0.84, 0.88]	0.81 [0.79, 0.83]	0.80 [0.77, 0.83]

E PER-DATASET MATCHED-BUDGET DOWNSTREAM CIS

Table 7 reports 95% paired-seed-bootstrap intervals for Table 3 Panel A. The ten confirmatory comparisons defined in Section 4 survive BH-FDR correction at $q = 0.05$; this table resolves the downstream LLM-as-Judge and Untyped-Graph effects by task.

Table 7: Panel A downstream performance with 95% paired-bootstrap CIs over 7 seeds. Mean is identical to Table 3.

Curated by	WizardLM →AlpEval	TriviaQA	MedQA	ChaosNLI	DynaSent-R2
Confident +human	72.5 [72.2, 72.8]	70.1 [69.8, 70.4]	56.4 [56.0, 56.9]	60.2 [59.8, 60.6]	64.7 [64.3, 65.1]
Cartography +human	73.1 [72.8, 73.4]	71.0 [70.7, 71.3]	57.1 [56.7, 57.5]	61.0 [60.7, 61.4]	65.3 [64.9, 65.7]
LLM-Judge +human	74.9 [74.7, 75.1]	72.9 [72.7, 73.1]	58.7 [58.3, 59.1]	63.1 [62.8, 63.4]	66.8 [66.5, 67.1]
Untyped +human	75.3 [75.1, 75.5]	73.2 [73.0, 73.4]	59.8 [59.4, 60.2]	63.9 [63.6, 64.2]	67.3 [67.0, 67.6]
STATEROUTE +human	76.2 [76.0, 76.4]	74.1 [73.9, 74.3]	61.7 [61.3, 62.1]	65.3 [65.0, 65.6]	68.8 [68.5, 69.1]
STATEROUTE-Opt +human	76.5 [76.3, 76.7]	74.3 [74.1, 74.5]	62.2 [61.8, 62.6]	65.6 [65.3, 65.9]	69.1 [68.8, 69.4]

The intervals resample the same 7 aligned seed runs used for every curator, preserving paired training variation. The direction of the STATEROUTE gap is stable on all five tasks. Direct-action controller means and paired deltas for the main, shared-recipe, and LODO panels appear together in Table 16.

F SHARED-RECIPE DOWNSTREAM PANEL (CURATION-ONLY DECONFOUND)

Table 3 combines state-aware curation with soft targets for KEEP-UNCERT and the E^+ consistency term. We isolate curation quality by rerunning every curator with one neutral cross-entropy SFT recipe, hard targets, and identical seeds, hardware, and evaluation. Table 16 supplies the GBDT and Cluster-Transformer rows for the same neutral recipe.

Table 8: Shared-recipe downstream performance (%) on Llama-3-8B with a single neutral SFT loss across all curators (standard cross-entropy, no soft labels, no consistency regularizer, no method-specific sampling). Every row uses the same 12.4 min/1K matched reviewer budget, so the budget column is omitted. Mean ± std over 7 seeds $\{1, \ldots, 7\}$. The "Δ vs. Table 3" rows record the change relative to the joint-pipeline panel.

Curated by	WizardLM →AlpEval	TriviaQA	MedQA	ChaosNLI	DynaSent-R2
Confident +human	72.3±0.4	69.8±0.4	56.3±0.6	60.0±0.5	64.5±0.5
Cartography +human	72.8±0.4	70.7±0.4	56.9±0.5	60.8±0.4	65.0±0.5
LLM-Judge +human	74.9±0.3	72.9±0.3	58.7±0.5	63.1±0.4	66.8±0.4
Untyped-Graph +human	75.1±0.3	73.0±0.3	59.3±0.5	63.4±0.4	66.7±0.4
STATEROUTE +human (shared recipe)	75.9±0.3	73.8±0.3	60.9±0.5	64.7±0.4	68.3±0.4
Δ vs. Judge	+1.0	+0.9	+2.2	+1.6	+1.5
Δ vs. Untyped	+0.8	+0.8	+1.6	+1.3	+1.6
Δ vs. joint panel	−0.3	−0.3	−0.8	−0.6	−0.5

Under the shared recipe, STATEROUTE retains most of its joint-pipeline lift and remains above LLM-as-Judge by +0.9–+2.2 points and Untyped-Graph by +0.8–+1.6. The largest reductions occur on MedQA and ChaosNLI, where soft targets contribute most, yet the curation ordering is unchanged. Paired MedQA and ChaosNLI comparisons remain significant after BH-FDR ($p = 0.006$ against LLM-as-Judge and $p = 0.021$ against Untyped-Graph). Thus both the repair decisions and the state-aware training terms contribute, with curation accounting for most of the gain.

G ZERO-COST DOWNSTREAM BASELINES

Table 9 reports automatic (zero reviewer-cost) downstream baselines on Llama-3-8B over the five datasets so that the main-text matched-budget table can be cross-read against a no-review operating point. STATEROUTE (auto) is highest on every task. LLM-as-Judge exceeds Soft-Label Distributional on WizardLM and TriviaQA, while Soft-Label Distributional is higher on MedQA, ChaosNLI, and DynaSent-R2. For the methods with automatic counterparts reported here, every matched-budget row in Table 3 Panel A is strictly above its zero-cost counterpart, so the reviewer budget is not being spent at parity with the automatic policy.

Across all five zero-cost tasks, STATEROUTE exceeds LLM-as-Judge and Soft-Label Distributional before reviewer input. Comparing this table with Table 3 then measures the incremental value of targeted review on top of a competitive automatic policy.

H REVIEW PROTOCOL: PER-STATE AND PER-DATASET DETAIL

Table 10 breaks the P4 review protocol down by the four repair states and by the five source datasets. Median time is lowest on annotation artifacts and true contradictions and highest on ambiguity;

Table 9: Zero-cost downstream baselines (%) on Llama-3-8B. Every row uses 0 reviewer min/1K, so the budget column is omitted. Metrics match Table 3.

Curated by	WizardLM→AlpEval	TriviaQA	MedQA	ChaosNLI	DynaSent-R2
Raw	69.4±0.5	67.2±0.4	53.9±0.6	57.5±0.5	62.1±0.5
Dedup-only	70.1±0.5	67.8±0.4	54.4±0.6	57.9±0.5	62.6±0.5
Confident (auto)	71.0±0.4	68.6±0.4	55.0±0.6	58.6±0.5	63.2±0.5
Cartography (auto)	71.7±0.4	69.4±0.4	55.6±0.6	59.3±0.5	63.8±0.5
Soft-Label Distributional	72.4±0.4	70.0±0.4	58.9±0.6	61.7±0.5	65.1±0.5
LLM-as-Judge (auto)	73.1±0.4	71.0±0.4	58.3±0.6	61.2±0.5	64.7±0.5
STATEROUTE (auto)	74.6±0.4	72.5±0.4	60.3±0.5	63.5±0.5	66.9±0.5

override rates are lowest on true contradictions and highest on ambiguity, matching the qualitative difficulty ordering. Three raters labeled every P4 cluster; we report Fleiss' $\kappa = 0.74$ for the four-way repair-state label. The three pairwise Cohen's κ values (0.73, 0.75, 0.74) are consistent with that multi-rater estimate.

Table 10: Per-state and per-dataset review-protocol statistics on pool P4 (500 clusters, three raters, 12.4 min/1K operating point). Median and $p90$ times are per surfaced cluster; override is the fraction of reviewer decisions that disagree with the predicted state or the default operator.

Slice	Median time (s)	$p90$ time (s)	Override (%)	Fleiss' κ
By repair state				
True contradiction	6.1	11.8	6.4	0.78
Ambiguity	8.4	16.9	14.2	0.68
Context-dependent	7.3	14.6	10.1	0.72
Annotation artifact	5.8	11.2	8.9	0.77
By dataset				
MedQA	8.0	15.4	11.7	0.71
TriviaQA	6.4	12.8	8.3	0.76
WizardLM	6.9	13.6	9.4	0.75
ChaosNLI	7.6	15.0	10.8	0.73
DynaSent	6.8	14.2	9.6	0.74
Pooled	7.1	14.2	9.8	0.74

Ambiguity concentrates review burden: it has the longest median and tail time, highest override rate, and lowest agreement. MedQA is likewise the slowest and lowest-agreement dataset. These variations explain why operator-distinguishing cues matter and why the 12.4 min/1K operating point is pooled rather than imposed per cluster. The 9.8% override rate also confirms that the interface supports expert correction instead of automatic acceptance.

I END-TO-END REVIEWER-WORKFLOW AUDIT

Table 11 re-computes the matched-budget cost comparison under two accounting schemes. *Surfaced-cluster-only* is the 12.4 min/1K proxy used in the main text: median review time multiplied by the surfaced-cluster rate. *End-to-end workflow* instead logs a full curation pass on a $10{,}000$-example WizardLM-52K slice and includes setup (rubric calibration session, one-time per reviewer), QC (second-pass spot check on 10% of surfaced clusters), disagreement resolution (third-rater tie-break when two raters disagree, triggered on 9.8% of clusters), and the tail of the review-time distribution rather than the median alone. The end-to-end means are 16.3, 16.3, and 16.2 min/1K against the common 12.4 min/1K proxy, i.e., a measured uplift in the $1.31\times$ range rather than an assumed exact multiplier. Crucially, under the end-to-end scheme the ranking STATEROUTE+human $>$ Untyped-Graph+human $>$ LLM-as-Judge+human holds identically to the proxy scheme. Matched reviewer *exposure* (surfaced clusters per 1K examples) is 105 for STATEROUTE+human, 104 for

Untyped-Graph+human, and 108 for LLM-as-Judge+human; so the matched-budget comparison also matches reviewer exposure counts to within 3%.

Table 11: End-to-end reviewer-workflow audit on a 10,000-example WizardLM-52K slice. "Proxy" is the surfaced-cluster-only number used in the main text; "End-to-end" adds setup (s), QC, disagreement resolution, and the tail of the review-time distribution (mean rather than median). The end-to-end-to-proxy uplift is measured to be near-uniform across methods after normalization per 1K examples, so the ranking is preserved.

Method	Proxy (min/1K)	Setup (min/1K)	QC (min/1K)	Disagree. res. (min/1K)	Tail uplift (min/1K)	End-to-end (min/1K)
LLM-Judge +human	12.4	1.0	0.9	1.3	0.7	16.3
Untyped +human	12.4	1.0	0.9	1.3	0.7	16.3
StateRoute +human	12.4	1.0	0.9	1.3	0.6	16.2

A full-pass wall-clock test (one reviewer, 10,000 examples, $\approx$ 105 surfaced clusters per 1K) took 2 h 41 min including breaks, or about 16.1 min/1K. This agrees with the independently itemized 16.2 min/1K in Table 11; the 0.1-minute difference reflects separate setup amortization and disagreement-resolution accounting. Setup, QC, and escalation remain near-equal because the same surfaced-evidence protocol and tightly matched reviewer exposure trigger them before method-specific repair. End-to-end accounting therefore raises the absolute cost while preserving the matched-budget ranking.

J LODO Downstream Accuracy

Table 12 reports Llama-3-8B downstream performance when the curation stack is refit on the four non-held-out datasets. Performance decreases by 0.8–2.4 points relative to the corresponding in-domain rows in Table 3, while StateRoute remains above Untyped-Graph and LLM-as-Judge on every fold. Table 16 adds both direct-action controllers on all five LODO folds.

Table 12: LODO downstream performance on Llama-3-8B (%). Each row trains curation on the other four datasets and evaluates on the held-out domain. Mean $\pm$ std over 7 seeds. All rows matched at 12.4 min/1K.

Held-out domain	LLM-Judge +human	Untyped +human	StateRoute +human
MedQA	57.9±0.6	58.3±0.6	60.2±0.5
TriviaQA	70.5±0.4	71.0±0.4	72.3±0.4
WizardLM→AlpEval	72.5±0.4	73.1±0.4	74.4±0.3
ChaosNLI	60.9±0.5	61.8±0.5	63.5±0.4
DynaSent-R2	64.7±0.5	65.5±0.5	67.1±0.4

The ranking remains stable on all five folds, including the two natural-disagreement domains. MedQA retains the widest separation, consistent with the direct evidence that operator choice matters most when disagreement is semantically meaningful.

K Backbone-Swap Stress Test

We replace both retrieval encoder and contradiction scorer while holding clustering, action semantics, supervision, interface, tuning budget, and review time fixed. Each backbone carries Untyped-Graph, GBDT imitation, Cluster-Transformer imitation, and StateRoute through the same P2 evaluation.

Weaker evidence backbones reduce every controller's absolute F1, confirming that pair scoring remains a real bottleneck. The no-type hierarchy is stable—sequence imitation exceeds GBDT, which

Table 13: Backbone stress on the pooled $2{,}500$-pair P2 evaluation. Graph denotes Untyped-Graph; State denotes StateRoute. Delta compares StateRoute with the strongest no-type controller in each row; 95% paired bootstrap, $10{,}000$ resamples, 12.4 min/1K.

Backbone	Graph	GBDT	Seq.	State	Δ	95% CI	p
ST5 + DeBERTa-v3	0.76	0.79	0.804	0.82	+0.016	[+0.005, +0.028]	0.008
MPNet + RoBERTa-MNLI	0.72	0.748	0.768	0.79	+0.022	[+0.008, +0.036]	0.006
BGE-small + BART-MNLI	0.71	0.735	0.752	0.77	+0.018	[+0.004, +0.032]	0.013

exceeds Untyped-Graph—and StateRoute remains above the strongest member under all three swaps. The explicit-state-minus-sequence effect varies from +0.016 to +0.022, showing consistent sign without an artificially constant effect size.

L External Validation and Low-Fresh-Label Transfer

This experiment moves annotation and adjudication outside the host institution while preserving the matched curation scaffold. It combines an independently labeled held-out panel with low-label transfer from the MedQA operating point.

The independent external panel (P10) is fully disjoint from the authors' internal evaluation pools. We commissioned an external medical-QA annotation contractor under a flat-fee research-services agreement; no contractor employee is an author. The contractor labeled 800 MedQA-style contradiction candidate clusters drawn from a United States Medical Licensing Examination (USMLE) Step 2 practice bank that is disjoint from the MedQA-Disagree training and evaluation pools (source-ID overlap against P1/P2/P5: 0 items, verified by SHA-256 on the (x, y) pair IDs). Two external board-certified internal-medicine physicians with no host-institution affiliation independently produced the four-way repair-state label, the minority-retention judgment, and a gold repair decision. Per-item adjudication used a third external reviewer on disagreement (11.6% of clusters). The 800-cluster test panel used an inclusive shared evaluation total of 36 expert-hours; this total is separate from both model adaptation and the 857-hour internal subtotal. External Fleiss' κ on the four-way repair-state label is 0.71, comparable to the internal P4 value of 0.74.

The test unit is a cluster: 800 held-out clusters plus a disjoint 32-cluster calibration slice. We freeze the P1 DeBERTa scorer, P3 state head, and StateRoute-Opt surrogates, then recalibrate τ_s and λ_{M} with 2.1 physician-hours. A separate full-refit ceiling uses a $10{,}400$-pair external supervision campaign requiring 412 expert-hours. Test-time review remains 12.4 min/1K.

The frozen StateRoute row exceeds Untyped, GBDT imitation, and the Cluster-Transformer in cluster-level repair F1 (0.76 vs. 0.72/0.74/0.75) and minority retention (83.1% vs. 74.2%/74.9%/80.6%). Its paired F1 gain over the strongest no-type controller is +0.014 (95% CI [+0.003,+0.026], $p = 0.019$). The three labor denominators are now explicit: 36 shared test-panel hours, 2.1 hours for scalar calibration on 32 clusters, and 412 hours for the separate full-refit labels. The frozen row finishes within 0.02 F1 and 1.6 retention points of the full-refit ceiling, showing strong same-family transfer with low fresh calibration.

The panel uses de-identified textual items from a published practice bank and falls under the non-human-subjects determination described in Appendix M. Its typed-over-untyped F1 gap is $+0.04$ externally versus $+0.060$ internally, a shifted-domain effect complemented by the different-family legal-contract audit in Appendix S.

Table 14: Medical-domain external transfer on 800 physician-annotated clusters. The inclusive shared test-panel evaluation total is 36 expert-hours and is excluded from method-specific columns. Calibration hours cover the 32-cluster scalar recalibration; refit-labeling hours cover the separate 10,400-pair campaign. Repair-decision F1 is cluster-level with 10,000 paired resamples; minority retention uses 124 valid-minority clusters.

Method	Calibration expert-h	Refit labeling-h	Repair F1 [95% CI]	Min. Ret. (%)
Frozen transfer: scalar calibration only				
LLM-Judge (frozen)	0	—	0.68 [0.66, 0.70]	65.4
Untyped (frozen+32)	2.1	—	0.72 [0.70, 0.74]	74.2
GBDT selector (frozen+32)	2.1	—	0.74 [0.72, 0.76]	74.9
Cluster-Transformer (frozen+32)	2.1	—	0.75 [0.73, 0.77]	80.6
STATEROUTE (frozen+32)	2.1	—	0.76 [0.74, 0.78]	83.1
Full-refit quality ceiling: external supervision campaign				
STATEROUTE (full refit)	—	412	0.78 [0.76, 0.80]	84.7

M REVIEWER GOVERNANCE, COMPENSATION, AND CONSENT

Reviewer roster and labor allocation are summarized here. P4 (protocol audit) and P5 (MedQA minority panel) were labeled by three board-certified internal-medicine physicians on salaried contract at the authors' organization, none of whom are authors; they reviewed 500 and 312 clusters respectively. P1 (scorer-training pairs) was labeled by two staff research engineers on salaried contract; the same two engineers performed the P6 fold-specific QA re-review of reused P1 pairs (no new labels were created for P6). P2 (pooled evaluation), P3 (type/action clusters), P8 (ChaosNLI minority panel), and P9 (DynaSent minority panel) were labeled by four in-house linguistic annotators on salaried contract, two of whom hold a graduate degree in linguistics.

All internal reviewers for P1–P9 were pre-existing salaried employees of the authors' organization at the time the study began; no internal reviewer was hired contingent on participating. They were compensated through salary at the organization's standard annotator/engineer/physician band, with no per-item piecework rate. Total logged internal expert-reviewer time across those nine pools was 857 hours (Table 5); external medical and legal panels used separate flat-fee research-services agreements and separate labor accounting.

Reviewer consent to the internal research workflow was obtained in writing during onboarding. An institutional human-research review office issued a formal non-human-subjects determination on the following operational grounds: (i) no identifiable private information is collected about the reviewers themselves, (ii) the annotated items are drawn from publicly released textual datasets and contain no patient-identifiable information, and (iii) the reviewers act as research staff using de-identified content. Under the applicable exempt workflow, full IRB review was explicitly waived.

Operational risk controls are conservative for the highest-stakes domain. Because a wrong relabel or context split on MedQA-Disagree can propagate harmful supervision, the reviewer interface is advisory (override rate 9.8%, pool P4), three raters adjudicate every P4, P5, P8, and P9 cluster (Fleiss' $\kappa = 0.74$ on P4, 0.71 on P5), and the qualitative failure-case audit in Appendix N is completed before curated outputs enter downstream training.

N QUALITATIVE CLUSTER EXAMPLES AND FAILURE CASES

Table 15 shows one observed cluster per state plus one failure case. Cluster text is paraphrased to avoid reproducing licensed material verbatim, and source IDs are omitted for licensing reasons. The failure case comes from the hardest observed regime: ambiguity under BGE-small-en + BART-large-MNLI, where the state classifier misroutes 36.3% of ambiguity clusters.

Table 15: Observed cluster outcomes, paraphrased for license compliance. Route is the operator applied by the state-aware policy.

Type (dataset)	Paraphrased cluster content	Route and outcome
True contradiction (TriviaQA-Contra)	"Who composed the *Jupiter* symphony?" with answers {Mozart, Haydn}; injected contradiction at the 10% rate.	SELECTIVE-RELABEL replaces the Haydn answer only; cluster integrity preserved.
Ambiguity (MedQA-Disagree)	A 62-yr male with chest pain and ST elevation; options split 2/1 between thrombolysis and primary PCI; clinically both are acceptable per institution.	KEEP-UNCERT with label smoothing $\alpha = 0.45$; both valid answers retained in supervision.
Context-dependent (ChaosNLI-Disagree)	Premise mentions "the bank"; hypotheses split on whether it is a river bank or a financial institution.	SPLIT-CTX prefixes "context: finance" vs. "context: geography" and splits the cluster.
Annotation artifact (WizardLM-52K)	Near-duplicate instructions where one copy retains a "[[INST]]" template leak from an earlier rubric.	DROP on the leaked variant; DOWNWEIGHT as a softer fallback if the leak is partial.
Failure case (MedQA-Disagree, BGE+BART backbone)	Cluster on "first-line treatment for community-acquired pneumonia in a patient with penicillin allergy" where guideline revisions introduced a context-dependent split that the weakest scorer mistook for ambiguity.	Misrouted to KEEP-UNCERT instead of SPLIT-CTX; the override rate on this slice is 18.4%, and expert post-hoc correction recovers 1.6 MedQA points.

The examples expose the semantic distinction among operators and the mechanism behind the aggregate results. The failure case shows that even a correctly surfaced cluster can be misrouted when temporal guideline context resembles genuine ambiguity. This pattern explains the higher ambiguity override rate and motivates the context-sensitive state features evaluated in the backbone and cross-fitted analyses.

O Direct-Action Controller Carry-Through

Both no-type controllers use the same curated splits, 7 seeds, Llama-3-8B recipe, 12.4 min/1K budget, and surfaced-evidence interface as their corresponding base panel. Main downstream, neutral shared-recipe, and LODO blocks cover all five datasets. Delta columns report 95% paired seed-bootstrap half-widths for StateRoute minus each selector.

P Direct Cluster-Action Endpoint on Held-Out Controller Labels

We evaluate repair choice directly on a 384-cluster pool disjoint from P3 and downstream training. Three raters adjudicate one gold action from the shared five-action menu using the main surfaced-evidence interface. Every controller receives the same cluster features and 12.4 min/1K policy; uncertainty uses $10{,}000$ paired cluster resamples. The supports are 96/74/118/56/40 for KEEP-UNCERT, SPLIT-CTX, SELECTIVE-RELABEL, DROP, and DOWNWEIGHT.

Q Alternative Bottlenecks Under the Matched Scaffold

We compare the four-state controller with a coarser explicit state, two no-type families, and the Cluster-Transformer while holding the complete scaffold fixed. The three-state variant merges true contradiction and annotation artifact into an error-like state; its action head uses artifact features to

Table 16: All-dataset carry-through for GBDT and Cluster-Transformer direct-action controllers. Values are downstream percentages; deltas are STATEROUTE minus the indicated selector, with 95% paired seed-bootstrap half-width.

Panel	Slice	GBDT	Sequence	STATEROUTE	Δ GBDT	Δ Seq.
Main	WizardLM	75.5	75.8	76.2	+0.7 ±0.3	+0.4 ±0.3
	TriviaQA	73.5	73.8	74.1	+0.6 ±0.3	+0.3 ±0.3
	MedQA	60.4	61.0	61.7	+1.3 ±0.6	+0.7 ±0.5
	ChaosNLI	64.2	64.8	65.3	+1.1 ±0.4	+0.5 ±0.4
	DynaSent-R2	67.7	68.2	68.8	+1.1 ±0.4	+0.6 ±0.4
Shared recipe	WizardLM	75.2	75.5	75.9	+0.7 ±0.3	+0.4 ±0.3
	TriviaQA	73.2	73.5	73.8	+0.6 ±0.3	+0.3 ±0.3
	MedQA	59.9	60.4	60.9	+1.0 ±0.4	+0.5 ±0.4
	ChaosNLI	63.7	64.2	64.7	+1.0 ±0.4	+0.5 ±0.4
	DynaSent-R2	67.1	67.7	68.3	+1.2 ±0.4	+0.6 ±0.4
LODO	WizardLM	73.6	74.0	74.4	+0.8 ±0.4	+0.4 ±0.4
	TriviaQA	71.6	72.0	72.3	+0.7 ±0.3	+0.3 ±0.3
	MedQA	58.9	59.7	60.2	+1.3 ±0.5	+0.5 ±0.5
	ChaosNLI	62.4	63.0	63.5	+1.1 ±0.4	+0.5 ±0.4
	DynaSent-R2	66.0	66.5	67.1	+1.1 ±0.4	+0.6 ±0.4

The Cluster-Transformer consistently improves over the GBDT selector, with the largest no-type gains on MedQA and ChaosNLI. STATEROUTE remains above the strongest no-type row in all 15 dataset-panel combinations. Effects vary from $+0.3$ to $+0.7$ points against sequence imitation and expand under disagreement-heavy and LODO conditions, matching the direct action pattern without relying on one dataset or training loss.

choose SELECTIVE-RELABEL or DROP. This tests whether repair-relevant state remains useful when its partition is coarsened.

The three-state merge preserves most of the four-state gain, showing that the mechanism tolerates a coarser partition. Both high-capacity no-type probes improve over the GBDT selector, yet explicit state remains highest on all four endpoints. The consistent ordering supports the usefulness of repair-relevant intermediate state across these two label inventories, subject to the combined-system comparison described in Section 6.

R CROSS-FITTED CONTROLLER SUPERVISION

We partition all 660 P3 controller-labeled clusters into five folds stratified jointly by dataset and gold action. Each cluster is evaluated out of fold exactly once; each controller trains on four folds and

Table 17: Direct cluster-action results on disjoint controller labels. Accuracy reports cluster-bootstrap half-width; Correct is the raw exact-action count.

Controller	Acc	Macro F1	Correct
Untyped-Graph	0.67 ±0.04	0.61	257/384
GBDT Action-Selector	0.70 ±0.04	0.65	268/384
Cluster-Transformer	0.73 ±0.04	0.69	280/384
StateRoute	0.76 ±0.04	0.72	291/384
StateRoute-Opt	0.78 ±0.04	0.74	300/384

Table 18: Per-action recall with raw correct/support counts on the disjoint controller-label pool. KU/SC/RL/DR/DW denote the five actions in order.

Controller	KU	SC	RL	DR	DW
Untyped-Graph	0.74 71/96	0.51 38/74	0.79 93/118	0.71 40/56	0.38 15/40
GBDT Action-Selector	0.75 72/96	0.55 41/74	0.81 95/118	0.75 42/56	0.45 18/40
Cluster-Transformer	0.78 75/96	0.62 46/74	0.82 97/118	0.75 42/56	0.50 20/40
StateRoute	0.82 79/96	0.68 50/74	0.83 98/118	0.77 43/56	0.53 21/40
StateRoute-Opt	0.83 80/96	0.70 52/74	0.85 100/118	0.79 44/56	0.60 24/40

The two direct selectors improve over Untyped-Graph, confirming that learned action control is a strong baseline. StateRoute improves further, with the largest gains in correct-action counts over GBDT on KEEP-UNCERT ($79 - 72 = 7$) and SPLIT-CTX ($50 - 41 = 9$). The sequence controller closes part of this gap but remains lower on both actions. These operator-specific counts align the direct endpoint with the pair-level and minority-retention gains in Table 1.

Table 19: Alternative state bottlenecks under the matched scaffold. Pair F1, MedQA, and minority retention follow Table 1; action macro-F1 uses the disjoint controller-label pool.

Controller	Family	Pooled F1	MedQA	Min. Ret.	Action MF1
Four-state	explicit 222/191 148/99	0.82	61.7	87.3	0.72
Three-state merge	explicit 321/191 148	0.81	61.1	85.6	0.70
GBDT selector	No type	0.79	60.4	74.2	0.65
Latent code	MLP no type	0.80	60.9	81.5	0.68
Cluster-Transformer	No type	0.80	61.0	82.5	0.69

selects its checkpoint on an inner stratified split from those folds. The graph, cluster evidence, action set, tuning trials, reviewer interface, and downstream recipe remain fixed. Fold action supports vary slightly while preserving the pooled KU/SC/RL/DR/DW totals of 183/132/210/83/52.

Table 20: Five-fold support for cross-fitted controller evaluation. Columns are gold actions; every row contains 132 held-out clusters.

Fold	KU	SC	RL	DR	DW
1	37	26	42	17	10
2	36	27	42	16	11
3	37	26	41	17	11
4	36	27	43	17	9
5	37	26	42	16	11

Table 21: Cross-fitted controller and repair-driven results. Action macro-F1 reports mean $\pm$ fold standard deviation; Correct aggregates out-of-fold exact actions. Pair F1 is pooled P2 contradiction F1, and MedQA and minority retention are percentages.

Controller	Action MF1	Action Acc.	Correct	Pair F1	MedQA	Min. Ret.
GBDT	0.63±0.021	0.689	455/660	0.785	60.3	74.1
Action-Selector						
Cluster-Transformer	0.67±0.019	0.720	475/660	0.798	60.9	82.0
Three-state explicit	0.69±0.018	0.739	488/660	0.807	61.0	85.0
StateRoute	0.71±0.017	0.753	497/660	0.815	61.5	86.6
StateRoute-Opt	0.73±0.016	0.782	516/660	0.824	62.0	87.2

Table 22: Cross-fitted per-action recall with raw correct/support counts. Supports are shared across controllers.

Controller	KU	SC	RL	DR	DW
GBDT	0.75 (137/183)	0.55 (73/132)	0.81 (170/210)	0.75 (62/83)	0.25 (13/52)
Action-Selector Cluster-Transformer	0.78 (143/183)	0.62 (82/132)	0.82 (172/210)	0.75 (62/83)	0.31 (16/52)
Three-state explicit	0.80 (146/183)	0.65 (86/132)	0.82 (172/210)	0.75 (62/83)	0.42 (22/52)
StateRoute	0.81 (148/183)	0.67 (89/132)	0.81 (171/210)	0.75 (62/83)	0.52 (27/52)
StateRoute-Opt	0.83 (152/183)	0.70 (93/132)	0.85 (178/210)	0.78 (65/83)	0.54 (28/52)

Table 23: Aggregate five-fold confusion matrix for StateRoute. Rows are gold actions and columns are predicted actions; row totals equal the pooled supports above.

Gold \ predicted	KU	SC	RL	DR	DW
KU	148	14	9	5	7
SC	16	89	13	6	8
RL	10	9	171	13	7
DR	5	4	9	62	3
DW	8	6	7	4	27

Cross-fitting softens absolute controller scores while preserving the state advantage. Relative to the strongest no-type controller, StateRoute adds 22 correct actions and +0.033 accuracy (95% CI [+0.012,+0.055], $p = 0.004$); macro-F1 improves by +0.040 (CI [+0.018,+0.063], $p = 0.003$). The confusion matrix shows that remaining errors concentrate on SPLIT-CTX and DOWNWEIGHT, while the explicit-state controller retains its clearest gains on operator-confusable clusters.

A secondary stratified 420/120/120 train/dev/test split gives GBDT, StateRoute, and StateRoute-Opt action macro-F1 of 0.60/0.66/0.68, pair F1 of 0.78/0.80/0.81, MedQA accuracy of 60.1/61.0/61.4, and minority retention of 73.6/84.8/85.5. This independent split reproduces the cross-fitted ordering.

S Out-of-Family Legal-Contract Transfer

We test a shifted domain using ContractNLI legal-contract entailment data (Koreeda & Manning, 2021). Contract clauses are substantially longer than the main-suite examples and organize conflict around obligations, exceptions, and scope rather than medical or crowd-label disagreement. Two contract attorneys adjudicate a disjoint 600-cluster test panel with action supports KU/SC/RL/DR/DW = 150/96/188/102/64; the panel requires 27.4 expert-hours. A separate 32-cluster slice uses 1.8 fresh expert-hours to recalibrate τ_s and λ_{M}, with model weights frozen. Test-time review remains matched at 12.4 min/1K. Task utility is ContractNLI document-level macro-F1, and alternative retention uses 134 clusters whose adjudication supports multiple contract interpretations.

Table 24: Frozen legal-domain transfer on 600 ContractNLI clusters. Action metrics are cluster-level; task macro-F1 and alternative retention are percentages. Correct and Retained report raw numerators.

Controller	Action MF1	Action Acc.	Correct	Task MF1	Alt. Ret.
LLM-Judge	0.58	0.63	378/600	72.4	62.7 (84/134)
Untyped-Graph	0.62	0.67	402/600	74.0	67.9 (91/134)
GBDT	0.65	0.70	420/600	74.8	71.6 (96/134)
Action-Selector Cluster-Transformer	0.68	0.72	432/600	75.4	76.9 (103/134)
StateRoute	0.71	0.747	448/600	76.5	82.1 (110/134)
StateRoute-Opt	0.72	0.76	456/600	76.8	82.8 (111/134)

The legal shift produces smaller gains than the in-domain test while retaining the same ordering. Against the Cluster-Transformer, StateRoute improves action macro-F1 by +0.030 (95% CI [+0.008,+0.053], $p = 0.010$), action accuracy by +0.027 (CI [+0.007,+0.047], $p = 0.009$), and task macro-F1 by +1.1 points (CI [+0.4,+1.8], $p = 0.006$). The +5.2-point alternative-retention gain shows that the state advantage extends to preserving legally compatible interpretations under shifted terminology and document structure.