WHAT DO LABELS LET US CONCLUDE?

Anonymous authors
Paper under double-blind review

ABSTRACT

The same human responses can be evaluated against a single answer or a response distribution. Choosing between these targets can change what counts as a good prediction, and making sense of label disagreement requires evidence about how the data was labelled. We propose a framework for making these evaluation judgements explicit. A measurement contract states the target, and a Mediated-Label Reliability Graph records the production evidence. Analysts use shared rules to determine the supported conclusions. A compiler then combines this record with the requested performance claim to construct an evaluation protocol. Controlled evidence profiles show how removing target or process information changes rule matches and eligible actions while responses stay fixed. In real-data comparisons, changing the target and score changes two model rankings. Applying the complete requirements of twelve performance claims withdraws support for six and narrows one. Matched image and language experiments show a further consequence of target choice: majority reduction increases divergence from the response-distribution target. Together, these studies show how explicit targets and evidence guide model selection, the interpretation of performance gains, and label treatment.

1 INTRODUCTION

Suppose eight of ten annotators assign an image to class A and two to class B. For a single-answer task, the B responses may be errors. For a task predicting human responses, both frequencies belong to the target. This choice changes scoring and training (Figure 1) and depends on the scientific question (Jacobs & Wallach, 2021; Peterson et al., 2019).

The target does not explain why responses vary. An independent reference can establish error relative to a point target, while source identities permit comparisons across annotators or protocols. These distinctions guide action: correcting error can improve a point-target dataset, whereas collapsing responses can remove variation a distributional predictor should learn.

Existing approaches adopt particular label interpretations. Noisy-label learning uses corruption assumptions to estimate errors (Northcutt et al., 2021; Li et al., 2020), while distributional learning retains response frequencies (Peterson et al., 2019; Barsoum et al., 2016). Dataset documentation provides context for this choice (Gebru et al., 2021). When several explanations remain possible, the evaluator must determine which conclusions are supported. We ask how a declared target and incomplete production evidence guide this judgement.

A *measurement contract* states the target and population. Production evidence restricts the compatible explanations, and agreement across them supports a conclusion. The Mediated-Label Reliability Graph (MLRG) records this evidence. Analysts apply shared rules to form a source and action record, which a compiler connects to an evaluation protocol for the requested claim.

The protocol connects improvements to supported judgements. A model can rank first by majority-label accuracy yet fit the response distribution poorly. A component gain may support a narrower claim than one requiring several properties together. Our studies examine these decisions through controlled profiles, fixed predictions, and matched training.

Our contributions are:

- A criterion for supported conclusions based on agreement across compatible measurement processes. It distinguishes source interpretation from action eligibility. For observation-

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.

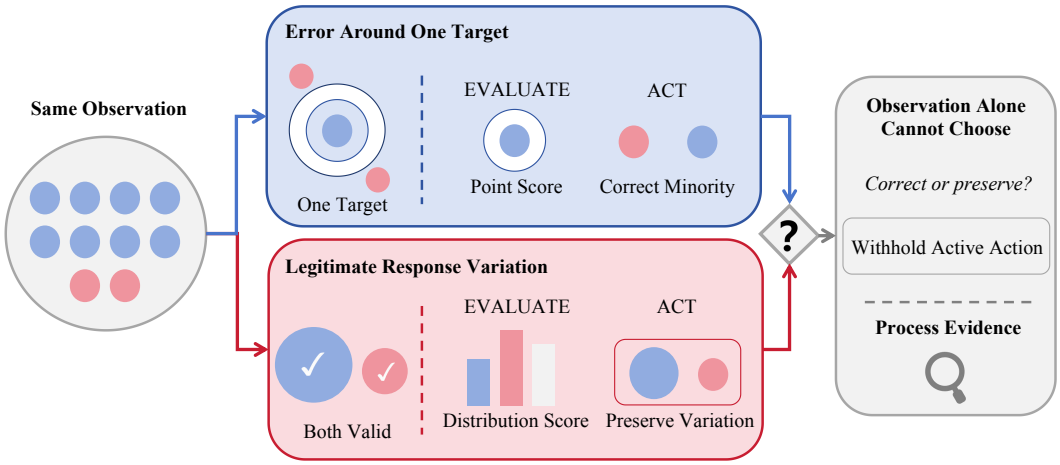


Figure 1: **Two targets for the same responses.** A point score evaluates a single answer, while a distributional score evaluates response frequencies. Correction requires evidence that a recorded response is an error for the declared target.

ally equivalent processes with conflicting valid actions, a bound relates active coverage to invalid-action risk.

- A typed MLRG representation that connects production evidence to source and action judgements through shared rules. A compiler checks the resulting record and constructs the evaluation protocol for a declared performance claim.
- Studies showing how these distinctions affect model comparison and label treatment. Controlled profiles isolate the information used by the rules, and real comparisons reveal changes in model order and claim support. Matched training measures the cost of majority reduction, while a coverage study represents 28 dataset contracts.

2 Related Work

Point-target methods explain disagreement relative to a latent answer. Crowd models estimate it alongside annotator reliability (Dawid & Skene, 1979; Raykar et al., 2010). Noisy-label methods use corruption assumptions to identify errors or reduce their training influence (Northcutt et al., 2021; Li et al., 2020). Correction thus depends on both the target and the label-production model.

Distributional methods predict variation. Label-distribution learning estimates how strongly each label describes an instance (Geng, 2016), while human-response methods use empirical frequencies (Peterson et al., 2019; Barsoum et al., 2016). Style models describe systematic segmentation-label differences (Zepf et al., 2023), and hidden context motivates distributions over preference utilities (Siththaranjan et al., 2024). We use established distributional treatments to examine evaluation-target choice.

Within a specified task, calibration relates probabilities to outcomes (Guo et al., 2017), while selective prediction studies acceptance and rejection (Geifman & El-Yaniv, 2019). Application requirements determine which quantities matter (Kahl et al., 2024). DomainBed makes model selection explicit (Gulrajani & Lopez-Paz, 2021), and WILDS organises comparisons around distribution shifts (Koh et al., 2021). Both show how evaluation conditions shape conclusions.

Measurement research relates constructs to their operationalisation (Jacobs & Wallach, 2021), and datasheets document production context (Gebru et al., 2021). Underspecification shows that similar in-distribution scores can coexist with different deployment behaviour (D'Amour et al., 2022). We connect targets and production evidence to conclusions by specifying eligible comparisons and actions, then applying the performance requirements.

3 METHOD

The method first specifies the prediction target, then uses production evidence to determine the supported conclusions. These conclusions guide the evaluation protocol in Figure 2.

3.1 SPECIFYING THE TARGET

Label production begins with acquisition, which turns a task-relevant state S into input $X \in \mathcal{X}$. An annotator records response $\widetilde{Y}$ under a context and protocol. The prediction target selects the quantity of interest from this process.

A measurement contract τ states the population and context to which the prediction refers. Its functional T_τ selects the target quantity from a measurement process P:

$$\theta_\tau(x) = T_\tau(P \mid X = x) \in \mathcal{Q}_\tau. \tag{1}$$

For a point target, T_τ selects the single answer declared by the contract. This can be an answer about the task state or the modal response in a population. A majority vote estimates the modal response. Using that vote as a reference for a task state requires a response model or independent validation linking the responses to that state. The contract records this reference procedure so that the meaning of accuracy remains explicit.

For a distributional response target, T_τ selects the distribution of recorded responses under the declared annotation protocol. The contract specifies how repeated responses are aggregated and how missing observations are treated (Appendix B.1).

When responses are pooled, each respondent group is weighted according to its proportion in the declared population. A group-conditioned target instead describes responses within a group. Changing group composition can therefore change the pooled target despite constant within-group behaviour.

A predictor $f_\tau : \mathcal{X} \to \mathcal{Q}_\tau$ is evaluated in the target space. For a distributional target, a proper scoring rule is optimised in expectation by that distribution (Gneiting & Raftery, 2007). The protocol selects the score and further requirements, such as improvement across seeds.

In the 80/20 example, predictions $(0.9, 0.1)$ and $(0.8, 0.2)$ have log losses 0.105 and 0.223 against point target A, favouring the first. Against the response distribution, their Kullback–Leibler (KL) divergences are 0.044 and 0.000, favouring the second. Both select A but assign different probabilities to B (Appendix A.5).

3.2 DETERMINING THE SUPPORTED CONCLUSION

With the target fixed, we ask which conclusions hold across the processes compatible with the evidence. Let $\mathcal{O}_{d,\tau}$ contain the observations and process evidence for dataset d, and let $\mathcal{A}_{d,\tau}$ contain the assumptions. At the population level, the compatible processes are

$$\mathfrak{M}_{d,\tau} = \{P : P_{\mathcal{O}_{d,\tau}} = p_{\text{obs}},\ P \models \mathcal{A}_{d,\tau},\ T_\tau(P) \text{ is well-defined}\}. \tag{2}$$

For a requested conclusion j, consider its values across this set:

$$\mathcal{H}_j(d, \tau) = \{h_j(P, \tau) : P \in \mathfrak{M}_{d,\tau}\}. \tag{3}$$

Let $\mathcal{R}_j(\tau)$ be the answer range allowed by the declared hypothesis class before applying the dataset's production evidence. For a non-empty compatible set, a singleton $\mathcal{H}_j$ identifies the conclusion (I). A non-singleton proper subset of $\mathcal{R}_j(\tau)$ gives partial identification (P). If the full non-singleton range remains possible, the conclusion is unidentifiable (U).

A source judgement records both the comparison scope and the mechanism, $h_{\text{source}}(P, \tau) = (b, z)$. Here, b specifies the source comparisons supported by the recorded links, and z describes how variation arises within that scope. Evidence can fix b while leaving several mechanisms possible, giving a partially identified joint answer. For actions, $h_{\text{action}}(P, \tau)$ is the set of eligible treatments or comparisons under the contract. Its status concerns whether that set is determined. The performance of a particular eligible treatment is a further empirical question.

Distinguishing evidence narrows these alternatives by restricting the compatible set. If $\mathcal{O}'$ extends $\mathcal{O}$ under the same contract and assumptions, then

$$\mathfrak{M}(\mathcal{O}', \tau) \subseteq \mathfrak{M}(\mathcal{O}, \tau), \qquad \mathcal{H}_j(\mathcal{O}', \tau) \subseteq \mathcal{H}_j(\mathcal{O}, \tau). \tag{4}$$

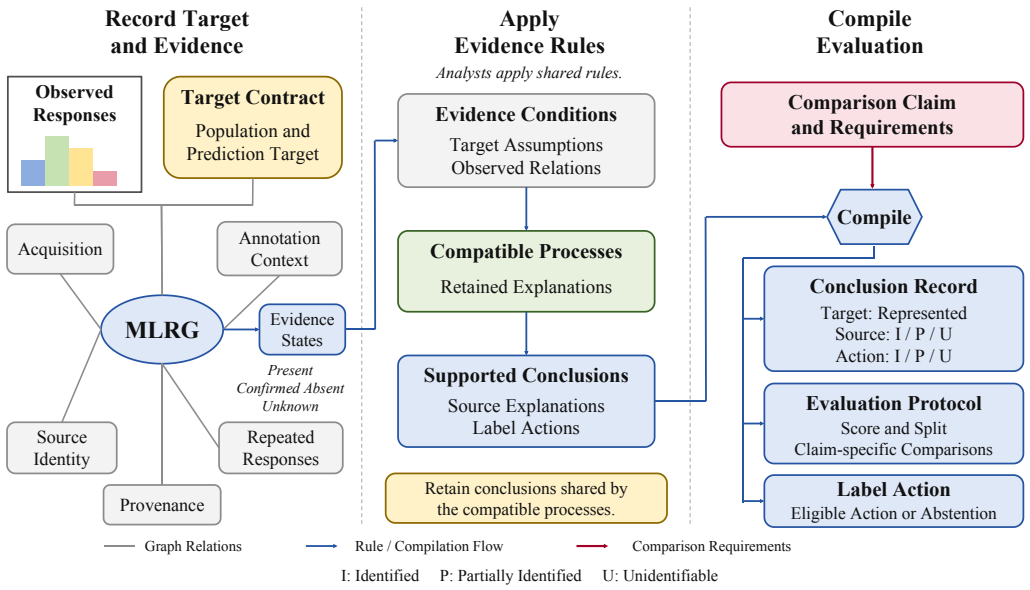


Figure 2: **From target and evidence to evaluation.** Analysts apply shared rules to the MLRG. The compiler combines the resulting source and action record with the comparison claim to specify the evaluation protocol.

An independent reference, for example, can exclude a response-variation explanation for a recorded error relative to a point target.

3.3 From Process Evidence to Evaluation

The MLRG links each observation to evidence about its production. Repeated-item links connect responses to the same input. Source links identify the annotator, protocol or domain associated with them. Each relation is present, confirmed absent, or unknown (Appendix B.2).

Even without an identifier column, a protocol specifying that one person labels all items establishes their common annotator. If the protocol leaves identity unspecified and identifiers are unavailable, the relation remains unknown. Analysts therefore interpret available fields together with the production protocol before assigning an evidence state.

Analysts apply six evidence-rule templates (Table 1, Appendix B.3) and record source and action judgements in $\mathcal{J}_{d,\tau}$. This record states the supported comparison scope and the evidence for any mechanism interpretation.

The compiler checks that this record refers to the declared contract and constructs the evaluation:

$$(\mathcal{C}_{d,\tau}, \mathcal{P}_{d,\tau}) = \mathrm{Compile}_{\mathcal{R}}(\sigma_d, \tau, \mathcal{G}_{d,\tau}, \mathcal{A}_{d,\tau}, \mathcal{J}_{d,\tau}, \kappa_{d,\tau}), \tag{5}$$

where σ_d supplies the dataset interface and $\kappa_{d,\tau}$ states the performance requirements. The compiler receives the analyst's statuses and eligible actions, retaining the scope of I and P records. Source links permit conditional reporting, and complete group links permit group-disjoint splits. The protocol uses the declared score and fits transforms on training data. Joint metric improvement or consistent modality contribution is required when specified by $\kappa_{d,\tau}$.

Researchers using the framework specify these requirements according to the intended use of the predictions. Average distributional fit and reliable confidence concern different properties of the same predictor. A claim about both calls for both assessments, whereas a claim restricted to distributional fit can use its proper score alone. Each numerical tolerance then qualifies a stated property, such as an acceptable loss of task accuracy. Appendix E.2 gives the requirements used in the real-data comparisons, including their aggregation across items and seeds.

Table 1: **Examples of evidence rules.** Each row connects evidence interpreted under a target to a supported conclusion and its evaluation consequence. Appendix B.3 gives all six templates.

Contract and evidence	Supported distinction	Action or comparison
Point target with a validated reference or distinguishing corruption model	A recorded deviation is error relative to the reference target	Correct or relabel
Response target with repeats and valid response mass	Recorded response variation is included in the target	Preserve the distribution
Source identity with linked repeated items and a conditional comparison	Variation can be compared within the same item across sources	Condition on source

The controlled study evaluates Boolean rule conditions. A reference and known corruption channel completely match the error rule, permitting correction. Hiding both leaves responses unchanged but the rule unresolved, retaining only abstention. In the shortcut profile, context identity and a fixed split retain a comparison scope after hiding the matched control, giving a partial match. These are finite rule-evaluation states (Section 4.2).

ChaosNLI collects repeated judgements for natural language inference (NLI) (Nie et al., 2020). The subsets used here label premise–hypothesis pairs as entailment, neutral or contradiction. Our contract targets the response distribution, estimated from 100 judgements per item.

Item links distinguish the Stanford NLI (SNLI) and Multi-Genre NLI (MNLI) subsets, fixing comparison scope b. Without an independent reference, error and valid variation remain possible mechanisms z, giving a P source record. The response target supports distribution preservation and scoring.

To score that prediction, let n_{ic} count responses in class c for item i, so that $q_{ic} = n_{ic}/100$. For predicted probabilities p_i on the same classes, the mean KL divergence over N items is

$$L_{\mathrm{KL}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c} q_{ic} \log \frac{q_{ic}}{p_{ic}}. \tag{6}$$

Separate SNLI and MNLI scores average to pooled KL when weighted by item counts. Hiding source links preserves pooled KL but removes source comparisons. The target supports the score, while production evidence determines its reporting scope (Appendix B.4).

Some compatible processes require conflicting actions. Under a fixed, known contract, let δ_n use n observed records to select an active action or abstain. Let R_k be its probability of selecting an invalid active action in setting k. If two processes induce the same law on these records and have disjoint valid active-action sets, their common active coverage C satisfies

$$R_0(\delta_n) + R_1(\delta_n) \geq C, \qquad \max\{R_0(\delta_n), R_1(\delta_n)\} \geq C/2. \tag{7}$$

Every active output is invalid in at least one setting, so always acting has worst-setting risk of at least $1/2$. Abstention reduces exposure to this conflict, while distinguishing evidence can resolve it. Appendix A gives the proof and a separate example of how two target contracts assign different actions to the same response law.

4 Experiments

4.1 Study Design

Shared response arrays test how rules use target and process information. Existing fixed-system evaluations then test changes in model order and claim support. Matched training measures the consequence of majority reduction, and a coverage study applies the representation to further contracts.

4.2 How Available Information Changes a Rule Output

Six evidence profiles have visible and hidden states under the rules in Section 3.3. Five pairs hide production information, while the response-variation pair hides the distributional-target declaration. Responses stay identical within each pair, isolating the information supplied for interpretation.

The 36 configurations cross these twelve profiles with three seeds. At each seed, all profiles share one array of 20,000 responses. Class balance and response entropy are therefore identical across profiles. Appendix D specifies the construction and removed conditions.

Visible information gives a complete match in all 18 configurations (Figure 3). Hiding the designated conditions leaves five profiles unresolved. The shortcut profile retains a partial match because its context-defined comparison survives the removal of the matched control. Every hidden profile permits only abstention. The rules therefore distinguish an eligible action from a comparison whose missing evidence still needs to be supplied.

4.3 How the Target and Claim Affect Model Comparisons

Seven of 13 candidate datasets provide paired evaluations on the same validation split, yielding two ranking comparisons and 12 conclusion contrasts (Appendix E). We compare their recorded metrics under the declared target and performance requirements.

Target-dependent model order. FERPlus records ten emotion judgements per face image (Barsoum et al., 2016), forming an empirical response distribution. Majority-label accuracy scores the most frequent response, while KL assesses the whole distribution.

The five-system comparison uses a residual network with 18 layers (ResNet-18) (He et al., 2016). An entropy-calibration term aligns predicted and response entropy, while an edge objective learns relations from response-distribution similarity. We test each objective, their combination, and random and within-class relation controls. The three-system comparison tests residual edge against global and within-class controls. Residual edge compares minority responses within a majority class, weighting similarities by non-majority mass.

The controls permute the response distributions used by the relation objective while retaining the original task targets. The random control permutes across training items. Permutation preserves the collection of response distributions and their entropies, so each control retains the amount of annotation variation available for relational training. The within-class control restricts this permutation to items with the same majority label, retaining class membership while disrupting the assignment of a particular response distribution to its image. This comparison tests whether the relations carry useful information beyond majority-class structure. The residual comparison's global control uses the same unrestricted permutation under its residual objective.

We rank each system by its mean improvement over the baseline across three training seeds. For fixed response targets, cross-entropy and KL differ only by target entropy. Since this quantity is constant across systems, cross-entropy gives the same distributional ordering.

Changing from accuracy to KL changes both orders (Figure 4a,b). In the five-system comparison, the within-class control falls from first to fourth, while the edge model rises from second to first. Four of ten pairwise orders change. In the three-system comparison, the global control and residual-edge model exchange places, changing one of three pairwise orders.

	Information visible		Information hidden	
Evidence template	Rule match	Eligible action	Rule match	Eligible action
Label error	**Complete**	Correct or relabel	**Unresolved**	**Abstain**
Valid response variation	**Complete**	Preserve distribution	**Unresolved**	**Abstain**
Protocol or annotator style	**Complete**	Condition on source	**Unresolved**	**Abstain**
Group or domain shift	**Complete**	Group or domain evaluation	**Unresolved**	**Abstain**
Missing information	**Complete**	Acquire modality or abstain	**Unresolved**	**Abstain**
Shortcut contamination	**Complete**	Group or domain evaluation	**Partial**	**Abstain**

Figure 3: **Rule matches and eligible actions under visible and hidden information.** Complete and partial refer to matches of the conditions in Appendix B.3. Unresolved profiles retain only abstention. All three seeds give the displayed results, and complete matches also permit abstention.

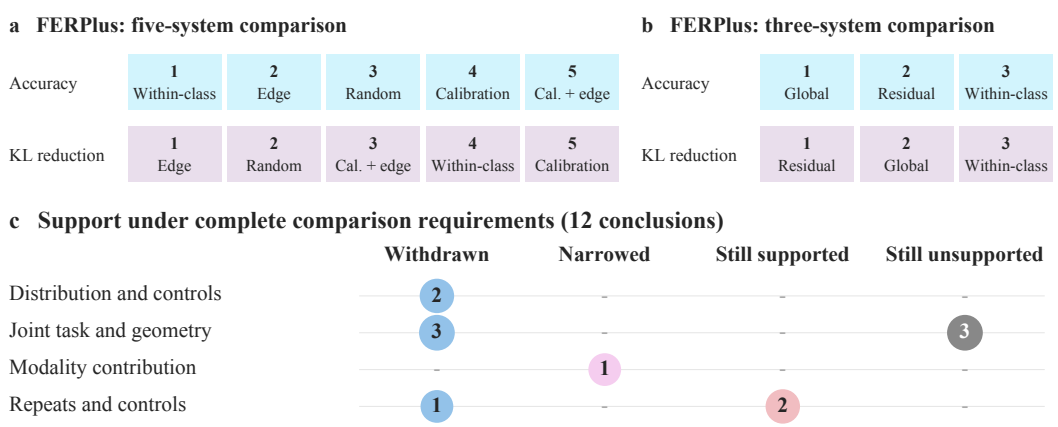


Figure 4: **Target-dependent rankings and claim-dependent support. (a,b)** FERPlus rankings use the mean accuracy improvement or KL reduction over three seeds, relative to each comparison's shared baseline. Larger improvements rank first, using unrounded values. Cal. denotes calibration. **(c)** Changes in support across 12 claims, grouped by comparison requirement.

The within-class control illustrates why the scores disagree. It improves accuracy by 0.007 relative to the shared baseline but increases KL by 0.007 nat. The edge model improves accuracy by 0.004 and reduces KL by 0.014 nat. Selecting the majority response more often can therefore coincide with a poorer fit to the response distribution.

Support for the complete performance claim. The 12 conclusion contrasts test whether a favourable component result supports a broader claim. Its requirements specify which results must hold jointly within the evidence-supported scope.

For FERPlus, the full utility claim combines distributional improvement with calibration, repeat consistency, and matched-control comparisons. Both tested methods improve mean accuracy and mean KL. Their calibration and repeat results fall short of the full requirement, so the broader utility claim loses support.

Geometry comparisons ask whether similar response targets have similar representations globally and locally. In ChaosNLI, response-distribution similarity defines target affinity, while source identities select cross-source pairs. Global Spearman correlation compares target affinity with representation cosine similarity, measuring agreement in pair ordering. Local affinity measures the mean target affinity among nearest cross-source neighbours above a random-pair reference. Improvements among distant pairs can raise global correlation without improving the nearest neighbours.

On ChaosNLI, global correlation improves by 0.024 but local affinity decreases by 0.001. Pair ordering improves while the neighbourhood statistic declines, so the joint geometry requirement fails. Appendix E.2 gives the affinity definition and settings.

This pattern recurs in GoEmotions, with fine-grained text-emotion annotations (Demszky et al., 2020), and OK-VQA, with multiple answers to visual questions requiring external knowledge (Marino et al., 2019). Global statistics improve, but the joint geometry criteria fail. Separate task-utility comparisons for both datasets and ChaosNLI remain unsupported.

The Multimodal EmotionLines Dataset (MELD) labels conversational emotion from text, audio, and video (Poria et al., 2019). Its full model exceeds text alone in weighted-F1 at all three validation seeds. Consistent contribution from both added modalities also requires a loss when either is removed. These differences change direction across seeds, retaining the overall gain but narrowing the modality claim.

Claims can also require consistency across training runs. COCO-Search18 records human fixations during object search (Yang et al., 2020). Mean attention similarity improves by 0.001, but one of three seeds fails the complete task criterion. The average gain remains supported, while the every-seed claim loses support. The geometry comparison remains supported.

Full claims can also retain support. The Continuously Annotated Signals of Emotion (CASE) dataset pairs continuous valence and arousal annotations with physiological signals (Sharma et al., 2019). Its fusion comparison meets both task-performance requirements and subject-related controls.

Across 12 contrasts, six full claims lose support and one narrows. Two supported and three unsupported claims retain their status (Figure 4c). Complete requirements preserve favourable component results while specifying which broader conclusions follow.

4.4 Consequences of Majority Reduction

The preceding comparisons hold predictions fixed. We now test how majority reduction during training affects response-distribution prediction, using repeated judgements for CIFAR10H images (Peterson et al., 2019) and three-class ChaosNLI sentence pairs (Nie et al., 2020).

We use three treatments from FERPlus (Barsoum et al., 2016). Preservation (A) trains on response frequencies. Majority reduction (I) uses the modal class, while response sampling (N) draws one response per item and epoch. Model and optimisation settings are matched within each dataset and seed.

Response sampling provides a control that retains the distributional target in expectation. For a fixed prediction p_i and a sampled response $Y_i \sim q_i$,

$$\mathbb{E}_{Y_i \sim q_i}[-\log p_{iY_i}] = -\sum_c q_{ic} \log p_{ic}. \tag{8}$$

This equals the full-distribution loss. At a fixed model state, preservation averages losses and gradients over responses with their empirical probabilities. Sampling selects one response gradient with the same expectation, adding variation to updates under a finite training budget.

Sampling tests the effect of averaging over responses. Majority reduction instead assigns all mass to the modal class. Three matched seeds use ResNet-18 for images and RoBERTa for NLI (Liu et al., 2019).

KL from the empirical response distribution is the primary endpoint. A positive lower 95% confidence bound for $I - A$ indicates increased divergence. For preservation relative to sampling, non-inferiority requires the upper bound for $A - N$ to be at most 0.010 nat per item, the pre-specified margin for negligible excess loss. Checkpoint selection uses a separate validation partition.

Contrasts pair the same held-out items and average over three seeds. The pre-specified bootstrap resamples seeds, then independently resamples paired items within each selected seed. Because seeds share evaluation items, a post-hoc sensitivity analysis resamples common item indices across selected seeds (Appendix F.6).

An item can favour the same treatment at several seeds. With common item weights, that item enters or leaves all selected seed means together, preserving its recurring contribution to the estimated difference. Independently sampling seed indices still captures variation among training runs. The two resampling dimensions therefore correspond to distinct sources of uncertainty.

Majority reduction increases KL by 0.254 on CIFAR10H (95% CI $[0.195, 0.317]$) and 0.095 on ChaosNLI ($[0.061, 0.133]$), as shown in Figure 5. Both intervals remain above zero under shared-item resampling. Preservation also gives lower mean KL than response sampling, with paired differences of -0.063 and -0.014. For ChaosNLI, the upper bound is 0.008 nat in the pre-specified analysis and 0.016 nat under shared-item resampling, compared with the 0.010-nat non-inferiority margin. CIFAR10H exceeds the margin in both analyses. The non-inferiority result thus depends on the resampling scheme, while the increase under majority reduction persists.

4.5 Coverage Across Dataset Contracts

Core-24 applies the representation to 24 datasets spanning 14 families and 18 target specifications. Prospective-4 adds four contracts held out from the Core analysis (Appendix C).

All 28 targets are represented (Table 2). Statuses reflect available evidence in the analyst records. The four prospective contracts retain a source-comparison scope, giving P for the joint judgement, while active label actions remain unresolved.

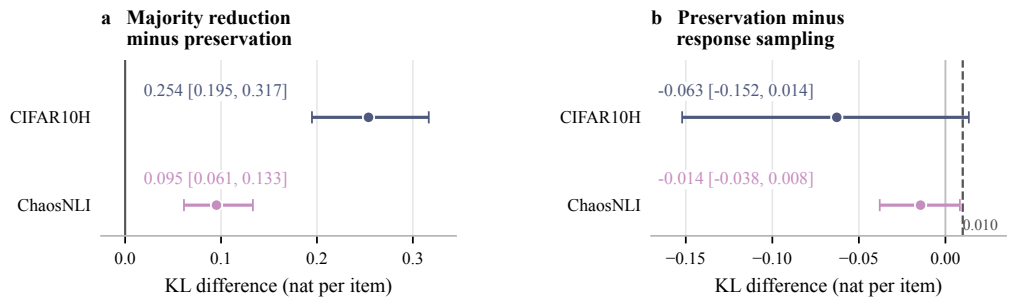


Figure 5: **Paired KL differences between label treatments. (a)** Majority reduction minus preservation. **(b)** Preservation minus response sampling, with the 0.010-nat non-inferiority margin dashed. Points and bars show means and 95% intervals from the pre-specified hierarchical bootstrap in nat per item. Appendix F.6 reports sensitivity to resampling shared evaluation items.

Table 2: **Target representation and conclusion status.** Target entries count represented contracts. Source and action entries count identified (I), partially identified (P), and unidentifiable (U) records, as defined in Section 3.2.

Cohort	Targets represented	Source	Action
Core-24	24/24	1 I, 20 P, 3 U	9 I, 15 P
Prospective-4	4/4	4 P	4 U

5 Discussion

The framework connects model selection and performance claims to explicit targets and evidence. FERPlus shows why higher accuracy can mislead selection for a response target. Complete requirements in the other comparisons retain component gains while delimiting the conclusions they support.

Unresolved mechanisms still permit useful evaluation. ChaosNLI source links support conditional reporting, and its response target supports preservation. Repeats improve target estimation, whereas mechanism attribution needs distinguishing evidence (Appendix A.4). Real-data judgements rely on analyst interpretation, with inter-analyst agreement still to be studied. Controlled profiles establish rule behaviour under supplied conditions.

Action eligibility and training performance answer different questions. Preservation and sampling share an expected objective, yet, finite optimisation yields different predictors. Their non-inferiority comparison depends on uncertainty estimation from three seeds and shared items. Majority reduction increases KL under both analyses, supporting preservation for response targets.

6 Conclusion

This study connected label targets and production evidence to evaluation judgements. The framework specified supported comparisons and label actions. Real-data studies showed how target choice changed model order and complete requirements refined performance claims. Matched training quantified the cost of majority reduction for response prediction. Together, these results supported model selection and label treatment consistent with the research question.

AI use statement

Generative AI tools assisted with finding related work, drafting sections of the manuscript, polishing the writing, and proving mathematical claims. The authors verified the resulting claims, citations and proofs and take responsibility for the final content.

REFERENCES

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 437–442, 2013. URL https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf.

Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES dataset: Diversity in conversational AI evaluation for safety. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53330–53342, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/a74b697bce4cac6c91896372abaa8863-Abstract-Datasets_and_Benchmarks.html.

Oresti Banos, Rafael Garcia, and Alejandro Saez. MHEALTH. UCI Machine Learning Repository, 2014. URL https://archive.ics.uci.edu/dataset/319/mhealth+dataset.

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 279–283, 2016. doi: 10.1145/2993148.2993165. URL https://doi.org/10.1145/2993148.2993165.

Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, 2017. doi: 10.18653/v1/E17-2092. URL https://aclanthology.org/E17-2092/.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979. doi: 10.2307/2346806. URL https://doi.org/10.2307/2346806.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

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

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. doi: 10.1109/TKDE.2016.2545658. URL https://doi.org/10.1109/TKDE.2016.2545658.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

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

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Kim-Celine Kahl, Carsten T. Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F. Jaeger. ValUES: A framework for systematic validation of uncertainty estimation in semantic segmentation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yV6fD7LYkF.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344, 2024. doi: 10.52202/079017-3342. URL https://doi.org/10.52202/079017-3342.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011. doi: 10.1145/1964897.1964918. URL https://doi.org/10.1145/1964897.1964918.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, 2021. doi: 10.1609/aaai.v35i17.17745. URL https://doi.org/10.1609/aaai.v35i17.17745.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

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*, 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://www.jair.org/index.php/jair/article/view/12125.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122. URL https://doi.org/10.1214/07-AOAS122.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Peterson_Human_Uncertainty_Makes_Classification_More_Robust_ICCV_2019_paper.html.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390. URL https://doi.org/10.1145/2733373.2806390.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

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/.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. URL https://jmlr.org/papers/v11/raykar10a.html.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 400–408, 2018. doi: 10.1145/3242969.3242985. URL https://doi.org/10.1145/3242969.3242985.

Karan Sharma, Claudio Castellini, Egon L. van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific Data*, 6(1):196, 2019. doi: 10.1038/s41597-019-0209-0. URL https://doi.org/10.1038/s41597-019-0209-0.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010. doi: 10.1109/TBME.2009.2036000. URL https://doi.org/10.1109/TBME.2009.2036000.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6. URL https://doi.org/10.1038/s41597-020-0495-6.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018. URL https://arxiv.org/abs/1804.03209.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1391–1399, 2017. doi: 10.1145/3038912.3052591. URL https://doi.org/10.1145/3038912.3052591.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8. URL https://doi.org/10.1038/s41597-022-01721-8.

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 193–202, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yang_Predicting_Goal-Directed_Human_Attention_Using_Inverse_Reinforcement_Learning_CVPR_2020_paper.html.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. doi: 10.1109/MIS.2016.94. URL https://doi.org/10.1109/MIS.2016.94.

Kilian Zepf, Eike Petersen, Jes Frellsen, and Aasa Feragen. That label's got style: Handling label style bias for uncertain image segmentation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wZ2SVhOTzBX.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 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. 3277–3287, 2021. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

A Formal Basis of Supported Conclusions

Equation 7 concerns processes under a fixed, known contract. It follows when the rule observes the same law in both settings and the valid active-action sets are disjoint. We state these conditions explicitly, then illustrate the separate role of target choice through two contracts with a shared response law.

A.1 Notation and assumptions

An evaluation setting consists of a measurement process P and a contract τ. Let $O = (X, \widetilde{Y}, M_{\text{obs}})$ be the observed record, including recorded process metadata. For the fixed-contract result, $A^*(P_k)$ denotes the active actions valid under the common contract τ. The rule can use this known contract, which is held constant and suppressed in the notation. The symbol $\perp$ denotes abstention. A possibly randomised rule at sample size n is

$$\delta_n : (O_1, \ldots, O_n, U) \longrightarrow \mathcal{A} \cup \{\perp\}, \tag{9}$$

where U is independent of the observed records. Table 3 collects the objects used below.

Table 3: **Notation for the formal result.** Each setting specifies its target and valid action set.

Symbol	Meaning
P	Complete measurement process under a declared contract.
$O_{1:n}$	Observed records available to the decision rule.
$P^{O_{1:n}}$	Finite-sample law induced by P on those records.
$A^*(P)$	Contract-valid active actions under P.
δ_n	Rule using $O_{1:n}$ and independent randomness.
$C_P(\delta_n)$	Active coverage, $\Pr_P\{\delta_n \neq \perp\}$.
$R_P(\delta_n)$	Probability of selecting an invalid active action.

For a fixed n, the lower bound requires two processes P_0 and P_1 such that

$$P_0^{O_{1:n}} = P_1^{O_{1:n}}, \qquad A^*(P_0) \cap A^*(P_1) = \varnothing. \tag{10}$$

The first condition gives every rule in Equation 9 the same output law in both settings. The result applies to any fixed contract for which these two conditions hold. Extending it to every finite n requires equality of all finite-dimensional observed laws. The example below changes the contract and illustrates what can be inferred when the decision rule receives only the shared response records.

A.2 A shared response law under two target contracts

Let $X \sim \text{Bernoulli}(1/2)$ and

$$\Pr(\widetilde{Y} = X \mid X) = \frac{4}{5}, \qquad \Pr(\widetilde{Y} \neq X \mid X) = \frac{1}{5}. \tag{11}$$

Consider the choice between correcting responses to X and preserving their distribution. Under contract τ_0, process P_0 has point target $S = X$, and the off-diagonal mass is corruption to be corrected. Under contract τ_1, process P_1 has the conditional response distribution as its target, so preservation is the valid choice. Both produce the joint law in Table 4.

Table 4: **Exact shared law for the two measurement processes.**

$(X, \widetilde{Y})$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
Probability	$2/5$	$1/10$	$1/10$	$2/5$

Records are sampled independently from this law under both processes. Consequently, $P_0^{O_{1:n}} = P_1^{O_{1:n}}$ for every finite n. A rule receiving only these records faces different valid choices under the two contracts. The same inequality applies to that restricted information setting. Providing the contract resolves this particular choice, illustrating why target specification precedes evidence interpretation.

A.3 FORCED ACTION AND THE COVERAGE–RISK BOUND

Proposition A.1 (forced action). Under Equation 10, every rule that always selects an active action satisfies

$$R_{P_0}(\delta_n) + R_{P_1}(\delta_n) \geq 1, \qquad \max_{k\in\{0,1\}} R_{P_k}(\delta_n) \geq \frac{1}{2}. \tag{12}$$

Proof. The two processes induce the same distribution over the rule output. Let $p(B)$ denote the common probability of selecting an action in B. Because the rule is active with probability one,

$$\begin{aligned} R_{P_0} + R_{P_1} &= 2 - p(A^*(P_0)) - p(A^*(P_1)) && (13)\\ &= 1 + [1 - p(A^*(P_0) \cup A^*(P_1))] && (14)\\ &\geq 1. && (15) \end{aligned}$$

The second inequality in Equation 12 follows because at least one of two non-negative risks is no smaller than half their sum. □

Proposition A.2 (coverage–risk trade-off). Allow $\delta_n = \perp$ and let C be its common active coverage under P_0 and P_1. Then,

$$R_{P_0}(\delta_n) + R_{P_1}(\delta_n) \geq C, \qquad \max_{k\in\{0,1\}} R_{P_k}(\delta_n) \geq \frac{C}{2}. \tag{16}$$

Proof. Observed-law equality again gives a common output distribution. Since the valid active-action sets are disjoint and their union lies within the active outputs,

$$\begin{aligned} R_{P_0} + R_{P_1} &= 2C - p(A^*(P_0)) - p(A^*(P_1)) && (17)\\ &= C + [C - p(A^*(P_0) \cup A^*(P_1))] && (18)\\ &\geq C. && (19) \end{aligned}$$

This proves Equation 16. In particular, zero wrong-action risk for both processes requires zero active coverage on this equivalence class. □

A.4 REPEATED RESPONSES

Suppose k repeated responses are conditionally independent given X and the two processes retain the same conditional response law. Every response array then has probability

$$\prod_{r=1}^{k} P(\widetilde{Y}_r \mid X), \tag{20}$$

which is identical under P_0 and P_1. Repeats estimate the shared response law. Distinguishing its process interpretations requires evidence with a different law under the competing processes, such as an independent reference or informative joint response structure. The declared contract specifies which quantity that evidence is used to interpret.

A.5 CONTRACT-DEPENDENT RANKING

The two contracts also reverse a model ranking. Let q_h assign probability $9/10$ to X, and let q_d assign $4/5$. The point contract scores log loss against $S = X$ and ranks q_h first because $-\log(9/10) < -\log(4/5)$. Under the response-distribution contract, the expected-log-loss difference is

$$L(q_h) - L(q_d) = \frac{1}{5}\log\left[(8/9)^4\, 2\right] = \frac{1}{5}\log\left(\frac{8192}{6561}\right) > 0. \tag{21}$$

Thus, q_d ranks first under the distributional estimand. The observed law is unchanged, and the contract determines which ranking answers the evaluation question.

B CONTRACT REPRESENTATION AND COMPILATION

The evaluation record in Section 3.3 connects the declared target to evidence interpreted by an analyst. The compiler checks this record and constructs its protocol. The definitions below specify the contract fields, graph relations, and rules used in that process.

B.1 MEASUREMENT MODEL AND CONTRACT

Let S denote a task-relevant latent state, M_X the acquisition process, and $X^{1:M}$ the observed modalities. Annotator A records response $\widetilde{Y}$ under context C and protocol Π:

$$(S, M_X) \sim p(s, m_x), \qquad X^{1:M} \sim p_X(\cdot \mid S, M_X), \qquad \widetilde{Y} \sim p_Y(\cdot \mid S, X^{1:M}, A, C, \Pi). \tag{22}$$

The target functional selects the quantity of interest from this process. A point contract can select a latent answer associated with S or the modal response in the declared population. A distributional response contract can select $P(\widetilde{Y} \mid X)$ for the declared respondent population. The contract records the context required to interpret that quantity.

Table 5: **Contract fields by scientific role.**

Field group	Meaning
Target and population	Quantity to be predicted and the population to which it refers
Prediction and aggregation	Output space and the use of repeated responses to form an empirical target
Protocol and context	Conditions under which the response is recorded
Dependence units	Repeated items, respondent groups, and source domains used in splits or conditional evaluation
Missing observations	Meaning of unavailable measurements and their treatment in the comparison
Scoring specification	A score compatible with the target, accompanied by the requirements of the requested performance claim

B.2 GRAPH AND EVIDENCE STATES

The typed graph is

$$\mathcal{G}_{d,\tau} = (V, E, \phi_V, \phi_E, \Omega), \qquad \Omega_k = (o_k, \mathcal{R}_k, \iota_k), \tag{23}$$

where ϕ_V and ϕ_E assign scientific roles to nodes and relations. For evidence relation $\mathcal{R}_k$, o_k is present, confirmed absent or unknown, and ι_k records its interpretation under the contract. Present or confirmed-absent evidence constrains the relevant process explanation. If the relation is unknown, the alternatives it could distinguish remain compatible.

Table 6: **Evidence dimensions in the graph.**

Relation	Observed distinction
Repeated responses	Variation within an item
Annotator or protocol identity	Variation conditioned on the response source
Group or domain identity	Population-specific comparisons and dependence units
Modality or context	Information available during response production
Provenance	Connection to acquisition and annotation procedures
Reference or counterfactual evidence	Evidence distinguishing competing production explanations

B.3 PROCESS RULES AND PROTOCOL CONSTRUCTION

Table 7 organises evidence conditions into six templates for interpreting variation. Each condition is interpreted under the declared target. For example, a reference must distinguish an error relative to

that target, while source links establish which responses can be compared. Analysts record the conclusions supported by these conditions. The controlled study evaluates their Boolean representations directly.

Table 7: **Evidence templates used to construct the analysis record.** Conditions concern the declared target and the available production evidence. Source and group links determine comparison scope. Mechanism attribution depends on the distinguishing evidence within that scope.

Template	Required information	Supported action or comparison
Label error	A validated independent reference, known distinguishing corruption channel, protocol-based adjudication, or a logical constraint uniquely rejecting the recorded label	Correct or relabel relative to the point target
Valid response variation	A response-distribution or answer-set target, repeated responses or an explicit set, and positive valid response mass	Preserve the response distribution or answer set
Protocol or annotator style	Source identity, repeated-item links across sources, and a source-conditioned comparison	Condition on response source
Group or domain shift	Group identity, complete item–group linkage, and a constructible group-disjoint split	Evaluate across the declared groups or domains
Missing information	Availability indicator and declared missingness semantics, with controlled removal, paired complete/incomplete measurements, or a protocol-defined unanswerable state	Acquire the relevant measurement or abstain
Shortcut contamination	Context identity, a label-matched or counterfactual control, and a fixed context split	Evaluate the context-related explanation within that split

Rule evaluation in the controlled profiles. A complete match satisfies every jointly required condition in Table 7, including at least one member of each alternative condition set. A unique complete match selects the corresponding template and makes its action eligible alongside abstention. Multiple complete matches leave the selection unresolved and retain only abstention.

When no complete rule matches, two rules can retain a comparison scope. An availability indicator with declared missingness semantics gives a partial match to the missing-information rule. Context identity with a fixed split gives a partial match to the shortcut rule. A unique partial match retains that template with abstention. Otherwise, the result is unresolved and retains all six templates. Complete, partial and unresolved describe this finite matching operation. The I/P/U records in Section 3.2 concern the values of a requested conclusion over compatible processes.

A reproducible visible/hidden example. The point-target profile supplies a validated independent reference and a known corruption channel. Either satisfies the alternative condition in the error rule. The other five rules lack required conditions, so the match is complete and correction or relabelling is eligible. Removing both predicates leaves the same responses but no rule match. The profile then retains only abstention.

For the shortcut pair, hiding the matched-control predicate leaves context identity and the fixed split. These satisfy the partial rule, preserving the comparison scope with abstention as the eligible action.

For real data, the compiler receives the analyst's source and action statuses together with the corresponding protocol fields. It checks that the record binds to the declared contract and that the statuses and actions are consistent. It then carries the supported scope into the output protocol, using metadata to construct the required splits and controls. A partial source record can therefore support conditional evaluation while its mechanism attribution remains open.

The evidence rules also connect the record to controls. Present metadata permits a hidden-metadata comparison, and unknown evidence retains an abstention control. A non-point target calls for a target-preserving control. Numerical thresholds and joint performance requirements are supplied by the comparison claim.

Target representation records the quantity being evaluated. Source status describes the joint scope-and-mechanism judgement in Section 3.2, and action status describes the eligible-action set. These are the three fields summarised by the coverage tables.

B.4 CHAOSNLI EVALUATION RECORD

The ChaosNLI example uses the Stanford Natural Language Inference (SNLI) and Multi-Genre NLI (MNLI) subsets. Its target is the distribution of recorded judgements under the annotation protocol, with $q_{ic} = n_{ic} / \sum_{c'} n_{ic'}$ and predictions p_i on the three-class simplex. Repeated responses estimate within-item variation, and source-dataset identity allows this variation to be examined through conditional reporting or a source-disjoint comparison.

In the source record, b is the comparison within each source dataset. The recorded item-to-source links fix this component. Candidate mechanisms z include error, valid variation, annotation style, domain differences and mixed sources. With no independent reference to distinguish them, mechanism attribution remains U. The joint source record is P because its scope is fixed. Preservation is supported by the response target, giving an identified action record and a multiclass proper score (Table 8).

Table 8: **ChaosNLI evaluation record.** The source record combines a defined comparison by source dataset with unresolved mechanism attribution.

Component	Status	Content
Target	Represented	Response distribution under the NLI annotation protocol
Source	Partial	Defined comparison by source dataset with unresolved mechanism attribution
Action	Identified	Preserve the distribution and condition comparisons on source
Split	Specified	Item split with source-domain reporting, or a source-disjoint comparison when required
Score	Specified	Multiclass proper score, with $D_{\mathrm{KL}}(q_i \| p_i)$ used in the action study
Controls	Specified	Majority reduction and response sampling in the action study

C CONTRACT COVERAGE

The coverage tables give the dataset-level records summarised in Section 4.5. Core-24 contains the main analysis contracts. Prospective-4 tests the same interface on four additional contracts held out from that analysis.

C.1 CORE-24 CONTRACTS

Tables 9 and 10 list each Core contract and its analyst-assigned conclusion scope. A represented target has an explicit prediction quantity and aggregation rule. VAD denotes valence, arousal, and dominance. ECG denotes electrocardiography. I, P and U denote identified, partially identified and unidentifiable source or action conclusions.

C.2 PROSPECTIVE-4 COMPILATION

Prospective-4 contains four dataset contracts held out from the Core analysis. Table 11 reports their target representation and conclusion scope. EEG denotes electroencephalography.

Table 9: **Core-24 contracts and conclusion status (1/2).**

Dataset	Family	Target contract	Represented	Source	Action
CIFAR10H (Peterson et al., 2019)	Visual disagreement	Soft multiclass	Yes	P	I
CMU-MOSI (Zadeh et al., 2016)	Multimodal affect	Continuous	Yes	U	P
UCI HAR (Anguita et al., 2013)	Wearable activity	Hard multiclass	Yes	P	I
CASE (Sharma et al., 2019)	Physiological affect	Continuous	Yes	P	P
WESAD (Schmidt et al., 2018)	Physiological affect	Hard multiclass	Yes	P	I
HANS (McCoy et al., 2019)	Language understanding	Hard binary	Yes	I	I
ChaosNLI (Nie et al., 2020)	Language understanding	Soft multiclass	Yes	P	I
GoEmotions (Demszky et al., 2020)	Language understanding	Soft multilabel	Yes	P	I
Speech Commands (Warden, 2018)	Speech audio	Hard multiclass	Yes	P	I
MELD (Poria et al., 2019)	Multimodal affect	Hard multiclass	Yes	P	P
OK-VQA (Marino et al., 2019)	Vision–language	Soft answer distribution	Yes	P	I
COCO-Search18 (Yang et al., 2020)	Visual search behaviour	Spatial distribution	Yes	P	P

Table 10: **Core-24 contracts and conclusion status (2/2).**

Dataset	Family	Target contract	Represented	Source	Action
DICES (Aroyo et al., 2023)	Annotator-conditioned judgement	Soft multilabel	Yes	P	I
BBQ (Parrish et al., 2022)	Structured QA context	Contextual hard multiclass	Yes	U	P
MovieLens 25M (Harper & Konstan, 2015)	Recommender preferences	Ordinal rating	Yes	P	P
MHEALTH (Banos et al., 2014)	Wearable activity	Grouped time-series class	Yes	P	P
ESC-50 (Piczak, 2015)	Environmental audio	Audio multiclass	Yes	P	P
MedMNIST (Yang et al., 2023)	Biomedical imaging	Image task suite	Yes	U	P
HateXplain (Mathew et al., 2021)	Annotator-conditioned judgement	Soft class with rationales	Yes	P	P
DynaSent (Potts et al., 2021)	Annotator-conditioned judgement	Soft multiclass	Yes	P	P
EmoBank (Buechel & Hahn, 2017)	Annotator-conditioned judgement	Continuous VAD	Yes	P	P
Wikipedia Talk Toxicity (Wulczyn et al., 2017)	Annotator-conditioned judgement	Repeated binary/continuous	Yes	P	P
WISDM (Kwapisz et al., 2011)	Wearable activity	Single-stream time series	Yes	P	P
PTB-XL (Wagner et al., 2020)	Clinical ECG	Soft multilabel	Yes	P	P

Table 11: **Prospective-4 compilation.** Four additional contracts are evaluated through the same interface.

Slot	Dataset	Domain	Target form	Represented	Source	Action
P1	Anonymised EEG cohort	Physiological measurement	Hard class	Yes	P	U
P2	PRISM Alignment (Kirk et al., 2024)	Pluralistic language-model feedback	Soft distribution	Yes	P	U
P3	Parkinsons Telemonitoring (Tsanas et al., 2010)	Clinical voice telemonitoring	Continuous	Yes	P	U
P4	TAT-QA (Zhu et al., 2021)	Financial hybrid reasoning QA	Answer set	Yes	P	U

D Controlled Evidence Profiles

The controlled study applies six evidence templates to shared response arrays. Each template has a visible and a hidden profile at three seeds, giving 36 configurations. One array of 20,000 responses is generated per seed and reused across its twelve profiles. The profiles vary the target or production information supplied to the rules.

D.1 Visible and hidden information

Table 12 gives the information supplied and hidden for each template. Five pairs remove production evidence. The response-variation pair removes the declaration of a distributional target while retaining repeated responses and positive valid response mass. All pairs retain their generated observations.

Table 12: **Information changed in each controlled pair.** The hidden profile retains all visible conditions except those in the final column.

Template	Information in the visible profile	Information hidden
Label error	Known corruption channel and independent reference	Both the channel and reference
Valid response variation	Distributional target, repeated responses, and positive valid response mass	Declaration of the distributional target
Protocol or annotator style	Source identity, repeated-item linkage, and a source-conditioned comparison	Source identity
Group or domain shift	Group identity, complete joins, and a group-disjoint split	Group identity
Missing information	Availability indicator with declared semantics and controlled modality removal	Indicator, semantics and removal evidence
Shortcut contamination	Context identity, matched control, and fixed context split	Matched control

D.2 Matched construction

For each seed, the array contains equal numbers of $X = 0$ and $X = 1$. Within each class, exactly one fifth of the recorded responses differ from X. The seed changes their ordering. A baseline predicts X with confidence 0.800, giving agreement with the generated responses of 0.800. These summaries are identical across profiles and seeds (Table 13). The six templates supply different interpreted evidence conditions for the same arrays.

Table 13: **Observable summaries shared by the controlled profiles.** Conditional entropy uses natural logarithms. The disagreement rate describes the generated responses relative to X, and baseline agreement is measured against those responses.

n	Class balance	$\Pr(\widetilde{Y} \neq X)$	Conditional entropy	Confidence	Agreement
20,000	0.500	0.200	0.500	0.800	0.800

D.3 Controlled results

Table 14 reports the rule matches and eligible actions. All visible profiles have a unique complete match. Hiding the designated information leaves five profiles unresolved and the shortcut profile partially matched. Each row gives the same result at all three seeds.

Table 14: **Rule outputs for each controlled pair.** Complete and partial denote the matching states defined in Appendix B.3. The visible action sets also include abstention.

Template	Visible match	Visible action	Hidden match	Hidden action
Label error	Complete	Correct or relabel	Unresolved	Abstain
Valid response variation	Complete	Preserve distribution	Unresolved	Abstain
Protocol or annotator style	Complete	Condition on source	Unresolved	Abstain
Group or domain shift	Complete	Evaluate by group/domain	Unresolved	Abstain
Missing information	Complete	Acquire modality or abstain	Unresolved	Abstain
Shortcut contamination	Complete	Evaluate by group/domain	Partial	Abstain

D.4 Aggregate results

The matching states and eligible-action sets agree with their specified reference outputs in all 36 configurations (Table 15). The denominator counts rule evaluations over the supplied profiles.

Table 15: **Agreement with specified rule outputs.** The denominators follow from six templates, two information states and three seeds.

Measure	Result	Rate
Visible-profile match state	18/18	1.000
Hidden-profile match state	18/18	1.000
Evidence-dependent action agreement	36/36	1.000

E Real-Data Contrasts

The real-data contrasts use existing metrics for the same systems on the same validation split. For a performance claim, we compare a component result with the complete requirements attached to that claim. For a ranking, we change the score used to order the systems while their predictions remain fixed.

E.1 Comparison set

Of 13 candidate datasets with existing model comparisons, seven provide the paired validation evaluations needed for this analysis. Each contrast requires a fixed system comparison with the compo-

nent metric and the results needed to assess the full claim. Table 16 records why six candidates fall outside that comparison set. The seven included datasets yield all twelve outcomes in Table 19.

Table 16: **Candidates outside the real-data comparison set.** Selection concerns the availability of comparable validation results. The CIFAR10H entry refers to an existing method comparison, separate from the matched training study in Appendix F.

Dataset	Comparison unavailable for this analysis
CIFAR10H	The existing paired method summary reports test results.
CMU-MOSI	Geometry evidence has no standalone joint validation utility comparison for the same system.
UCI HAR	No applicable system comparison provides both metrics on the same split.
WESAD	No applicable intervention comparison is available.
HANS	The failure control has no comparable set of validation intervention systems.
Speech Commands	No applicable intervention comparison is available.

E.2 Scores and comparison conditions

The distributional comparisons use the KL score in Equation 6, with empirical response probabilities q_i and predictions p_i in the corresponding target space. For a fixed target, KL differs from cross-entropy only by the target entropy, so both scores give the same ranking. The FERPlus distributional comparison ranks reductions in L_{KL} relative to a common baseline. Its point-target comparison ranks improvements in accuracy against the modal response. Each system's improvement is averaged over three seeds. Larger means rank first, using unrounded values.

The geometry comparisons test whether similar response targets have similar learnt representations. Spearman correlation between target affinity and representation similarity measures this relation globally across sampled cross-group pairs. To examine local structure, we measure mean target affinity among nearest cross-group neighbours above a random-pair reference. Both changes are measured against the baseline representation. The response target defines affinity, and the available group links define which pairs can be compared. Requiring global and local improvement together is part of the declared geometry claim.

For ChaosNLI, target affinity is $a(q_i, q_j) = 1 - \mathrm{JS}(q_i, q_j)/\log 2$, where $\mathrm{JS}(q_i, q_j) = \frac{1}{2}D_{\mathrm{KL}}(q_i\|m) + \frac{1}{2}D_{\mathrm{KL}}(q_j\|m)$ and $m = (q_i + q_j)/2$. Representation similarity is cosine similarity. Global correlation uses sampled cross-source pairs. Local affinity averages up to five cross-source neighbours per query, selected from 96 candidates, then subtracts the random-pair mean.

The ChaosNLI protocol requires positive changes in both geometry statistics while limiting the maximum absolute change in task predictions to 10^{-6}. The global improvement must also exceed the best of five random projections, and the local improvement must exceed their mean. The learnt representation passes the random-projection controls and preserves task predictions. However, its global change of $+0.024$ accompanies a local change of -0.001. This supports improved global alignment without meeting the joint requirement for local improvement.

Entropy alignment measures the mean absolute gap between prediction and target entropy. High-ambiguity or high-entropy comparisons use thresholds fitted on training responses. For COCO-Search18, this threshold is the upper quartile of training target entropy. Its geometry comparison requires global improvement above each of five random projections at every seed, together with positive global and local mean changes. The preserved task predictions bind this comparison to the same task model.

CASE uses subject-related probes to test the information carried by fusion weights and predictions. The weight-only probe may exceed chance accuracy by at most 0.010. Prediction and joint prediction–weight probes may exceed their equal-weight-fusion references by at most 0.010. The task and probe conditions hold at all three seeds, which yield identical global weights and predictions.

The complete FERPlus criterion requires positive mean accuracy improvement, including improvement at two or more seeds. Mean KL and ECE must remain non-degraded. For high-ambiguity

Table 17: **Requirements attached to the real-data performance claims.** Improvements are method minus baseline for higher-is-better metrics and baseline minus method for losses. Each row requires its conditions jointly. ECE denotes 15-bin expected calibration error.

Comparison	Complete requirement
CASE fusion	At every seed, mean absolute error improves over equal-weight fusion, with non-negative Pearson and RMSE improvements. Subject-probe accuracies satisfy the tolerances described below.
ChaosNLI utility	Accuracy and ECE improvements are at least -0.010. Overall KL, high-ambiguity KL and entropy alignment improve strictly.
GoEmotions utility	Micro-F1, macro average precision and ECE improvements are at least -0.010, -0.005 and -0.002, respectively. Soft binary cross-entropy, high-ambiguity KL and entropy alignment improve strictly.
OK-VQA utility	VQA and modal-answer accuracy improvements are at least -0.010. ECE, overall KL, high-ambiguity KL and entropy alignment improve strictly.
COCO-Search18 utility	At every seed, normalised JS divergence and attention similarity improvements are at least -0.005, while high-entropy divergence and entropy alignment improve strictly. All four mean improvements are positive.
Response geometry	Both global and local changes are positive with task predictions preserved within 10^{-6}. For ChaosNLI and GoEmotions, global and local gains must exceed the best and mean of five random projections, respectively. GoEmotions and OK-VQA require lower affinity prediction loss. COCO-Search18 applies its conditions at every seed.
MELD modality contribution	Weighted-F1 is higher for the full model than for text alone, without audio and without video at each of three seeds.
FERPlus utility	Mean distributional and calibration results, repeat consistency and comparisons with matched controls meet the conditions stated below.

items, KL must improve both on average and at two or more seeds. The ambiguity threshold is fitted from response entropy on the training partition.

The five-system claim tests calibration plus edge against the random and within-class controls. Mean accuracy must be at least as high as both controls, and high-ambiguity KL must be lower. High-ambiguity KL must also be lower than for calibration alone. The three-system claim applies the corresponding comparisons to residual edge against global and within-class controls. Both tested methods improve mean accuracy and mean KL, but their calibration, repeat, and some control results fall short of the full criterion.

For MELD, let F_s, T_s, A_s and V_s be validation weighted-F1 at seed s for the full model, text alone, without audio and without video. The component result requires $F_s > T_s$ at every seed. Consistent contribution from both added modalities also requires $F_s > A_s$ and $F_s > V_s$ at every seed. We evaluate these conditions directly on the existing validation results in Table 18. The full model exceeds text alone throughout, while each modality's leave-out comparison changes direction across seeds.

Table 18: **MELD validation differences in weighted-F1.** A positive leave-out difference indicates a lower score after removing that modality. Conditions use unrounded values.

Seed	Full − text	Full − without audio	Full − without video
20260802	0.004	-0.005	0.002
20260803	0.006	$0 < \Delta < 0.001$	-0.004
20260804	0.006	0.009	0.002

E.3 ALL CONCLUSION CONTRASTS

Table 19 distinguishes withdrawal of support for a full claim from a narrower interpretation that retains an overall gain. Stable outcomes keep the same support status under both comparisons.

Table 19: **All 12 conclusion contrasts.** R and F indicate support under the component and full comparison, respectively. The reduced value is the component statistic or decision.

Dataset	Conclusion	Reduced value	Full requirement	R	F	Outcome
CASE	Fusion utility	Task conditions met	Task and subject-control criteria	Yes	Yes	Stable support
ChaosNLI	Calibrated utility	Accuracy -0.154	Task and calibration criterion	No	No	Stable unsupported
ChaosNLI	Response geometry	Correlation $+0.024$	Joint geometry criterion	Yes	No	Support withdrawn
COCO-Search18	Calibrated utility	Mean $+0.001$	Complete criteria at every seed	Yes	No	Support withdrawn
COCO-Search18	Response geometry	Mean $+0.066$	Multi-seed geometry criterion	Yes	Yes	Stable support
FERPlus (five systems)	Distributional utility	$0 < \Delta_{\mathrm{acc}} < 0.001$	Distribution, calibration, repeats and controls	Yes	No	Support withdrawn
FERPlus (three systems)	Distributional utility	Accuracy $+0.006$	Distribution, calibration, repeats, and controls	Yes	No	Support withdrawn
GoEmotions	Calibrated utility	Micro-F1 -0.004	Task and calibration criterion	No	No	Stable unsupported
GoEmotions	Response geometry	Correlation $+0.050$	Joint geometry criterion	Yes	No	Support withdrawn
MELD	Modality contribution	Full $>$ text at all seeds	Audio and video leave-out support	Yes	No	Narrowing
OK-VQA	Calibrated utility	Accuracy -0.006	Task and calibration criterion	No	No	Stable unsupported
OK-VQA	Response geometry	Correlation $+0.033$	Joint geometry criterion	Yes	No	Support withdrawn

Across the 12 rows, six conclusions lose support and one narrows. Two supported and three unsupported conclusions retain their status.

E.4 FERPLUS SYSTEMS AND RANKINGS

The FERPlus systems use residual networks with 18 layers (ResNet-18) (He et al., 2016). The calibration variant adds the absolute difference between predicted entropy and response-target entropy to the training objective, with weight 0.1. The edge variant trains pairwise relations from label-distribution affinity while protecting the task objective. Combining the two gives the calibration-plus-edge model. To test the role of the relations, the random and within-class controls replace the relation targets with their respective control relations.

The three-system comparison tests a residual-edge objective against global and within-class controls. This objective compares the alternative-response distributions of images sharing a majority class, weighted by their non-majority response mass. Tables 20 and 21 report the three-seed mean accuracy improvement and KL reduction against the shared baseline of each comparison.

Table 20: **FERPlus five-system comparison.** Values are means over three seeds. Ranks use unrounded values, giving four changes among ten pairwise orders.

System	Accuracy change	Rank	KL reduction (nat)	Rank
Calibration + edge	$0 < \Delta < 0.001$	5	0.007	3
Calibration only	0.001	4	-0.019	5
Edge only	0.004	2	0.014	1
Random control	0.001	3	0.012	2
Within-class control	0.007	1	-0.007	4

Table 21: **FERPlus three-system comparison.** Values are means over three seeds. Ranks use unrounded values, giving one change among three pairwise orders.

System	Accuracy change	Rank	KL reduction (nat)	Rank
Global control	0.007	1	0.012	2
Residual edge	0.006	2	0.013	1
Within-class control	0.002	3	0.003	3

In both comparisons, the score changes while each system's predictions and evaluation split remain fixed. The displayed orders describe the observed means for these systems.

F Training with Repeated-Response Targets

The action study compares full response distributions, sampled responses and majority labels on CIFAR10H (Peterson et al., 2019) and ChaosNLI (Nie et al., 2020). Separate validation partitions are used for model selection and the final comparison. All reported results come from the held-out decision partition.

F.1 Procedures and evaluation criteria

Let q_i denote the empirical response distribution for item i. The three procedures share the model and optimisation protocol and differ in the target presented during training:

- preservation (A) minimises soft cross-entropy against q_i.
- response sampling (N) draws one categorical response from q_i for each item and epoch.
- majority reduction (I) replaces q_i with a one-hot target at its lowest-index mode.

Within each dataset and seed, the procedures share their initialisation, data order and training settings. Each selects its checkpoint on the selection partition, then undergoes one evaluation on the held-out decision partition.

The primary item-level endpoint is

$$D_{i,m} = D_{\mathrm{KL}}(q_i \,\|\, p_{i,m}), \tag{24}$$

where $p_{i,m}$ is the prediction from procedure m. The majority-reduction contrast is $I - A$, with a positive lower 95% confidence limit indicating increased KL. The non-inferiority contrast is $A - N$, with an upper confidence limit at most 0.010 nat per item required for non-inferiority. The margin defines the largest negligible average excess log loss.

F.2 Models, splits, and optimisation

Table 22 gives the dataset-level configuration, using ResNet-18 for CIFAR10H (He et al., 2016) and RoBERTa for ChaosNLI (Liu et al., 2019).

Table 22: **Action-study configuration.** Optimisation settings are shared by the three procedures within each dataset.

Dataset	Model	Epochs	Batch	LR	WD
CIFAR10H	ResNet-18	60	256	10^{-3}	10^{-4}
ChaosNLI	RoBERTa	12	32	2×10^{-5}	10^{-2}

The image-level split contains 7,000/1,500/1,500 CIFAR10H records for training, validation, and test. The three-class SNLI and MNLI subsets of ChaosNLI contain 3,113 items, split into 2,169/486/458 records. Both splits are item-disjoint. Within each hard-label stratum, validation records are divided into checkpoint-selection and held-out decision partitions, giving 753/747 for CIFAR10H and 244/242 for ChaosNLI. These partitions and their item order are shared across all treatments and seeds.

F.3 Paired hierarchical bootstrap

The pre-specified analysis uses bootstrap resampling (Efron, 1979) of paired item-level KL differences. Each of 10,000 draws samples the three seeds with replacement, then independently resamples paired items within each selected seed. Averaging the resulting seed means gives one draw of the contrast. The 2.5th and 97.5th percentiles give the reported interval. The two contrasts use independent random streams. Appendix F.6 examines an alternative resampling scheme that preserves common item weights across seeds.

F.4 SEED-LEVEL RESULTS

Table 23 reports the primary KL endpoint for each seed. A, N, and I denote full-distribution training, response sampling, and majority reduction.

Table 23: **Seed-level KL on the held-out decision partition.**

Dataset	Seed	A	N	I
CIFAR10H	1	0.626	0.626	0.854
CIFAR10H	2	0.514	0.670	0.818
CIFAR10H	3	0.555	0.587	0.784
ChaosNLI	1	0.261	0.285	0.331
ChaosNLI	2	0.278	0.287	0.405
ChaosNLI	3	0.266	0.276	0.354

F.5 AGGREGATE COMPARISONS

Table 24 reports the four paired contrasts under the pre-specified analysis. Majority reduction increases KL on both datasets, with both confidence intervals above zero. Preservation has lower mean KL than sampling on both datasets. In this analysis, the upper bound is below the 0.010 margin on ChaosNLI and above it on CIFAR10H.

Table 24: **Paired KL contrasts under the pre-specified analysis.** Intervals are paired 95% hierarchical bootstrap intervals in nat per item. Outcomes refer to this analysis. Table 25 reports the shared-item sensitivity analysis.

Dataset	Contrast	Estimate	Lower	Upper	Outcome
CIFAR10H	$I - A$	0.254	0.195	0.317	Increased KL
CIFAR10H	$A - N$	-0.063	-0.152	0.014	Upper bound > 0.010
ChaosNLI	$I - A$	0.095	0.061	0.133	Increased KL
ChaosNLI	$A - N$	-0.014	-0.038	0.008	Non-inferior

F.6 SENSITIVITY TO SHARED EVALUATION ITEMS

The same evaluation items appear at every seed, giving a crossed seed–item structure. We examine sensitivity to that structure with a post-hoc row-and-column bootstrap (Owen, 2007). For either treatment contrast, let d_{si} be the paired KL difference at seed s and item i. Each draw samples three seed indices with replacement and independently samples N item indices with replacement. The sampled item indices are shared across the selected seeds. The bootstrap statistic is

$$\overline{d}^{\,*} = \frac{1}{3N} \sum_{a=1}^{3} \sum_{b=1}^{N} d_{s_a^*, i_b^*}. \tag{25}$$

We use 10,000 draws and the same percentile limits as in the pre-specified analysis. Table 25 compares the resulting intervals. This resampling scheme retains both the pairing between treatments and the shared items across seeds.

Shared-item resampling keeps both $I - A$ lower bounds above zero and places both $A - N$ upper bounds above the 0.010 margin. The negative $A - N$ means are unchanged.

Table 25: **Sensitivity of paired KL intervals to shared-item resampling.** Means and 95% intervals are in nat per item. Both schemes use the same three trained seeds and evaluation items.

Dataset	Contrast	Mean	Pre-specified	Shared-item
CIFAR10H	$I - A$	0.254	$[0.195, 0.317]$	$[0.176, 0.333]$
CIFAR10H	$A - N$	-0.063	$[-0.152, 0.014]$	$[-0.161, 0.024]$
ChaosNLI	$I - A$	0.095	$[0.061, 0.133]$	$[0.053, 0.144]$
ChaosNLI	$A - N$	-0.014	$[-0.038, 0.008]$	$[-0.045, 0.016]$