REVERSIBLE HIERARCHICAL ACTIVE CORRECTION UNDER IMPERFECT HUMAN FEEDBACK FOR MLLM-HUMAN HYBRID ANNOTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

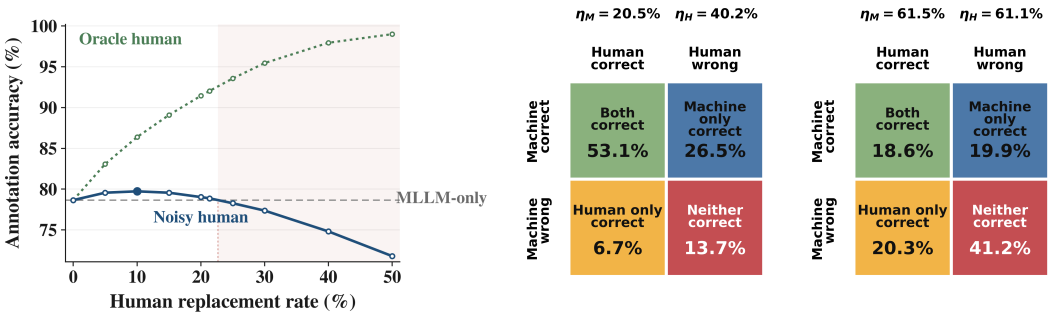
Obtaining high-quality labeled data is both expensive and time-consuming, motivating MLLM-human hybrid annotation as a scalable alternative. However, ordinary reviewers can make mistakes, while expert adjudication is too costly to apply throughout the dataset. Existing methods identify suspicious LLM-generated annotations, and selectively query human reviewers for correction. Nevertheless, they largely assume that *allocating more reviewed budget to ordinary-human review should enhance annotation accuracy*. In this paper, we demonstrate that this assumption fails when human reviewers are imperfect. We further identify two failure modes of selective human correction under imperfect feedback empirically, namely, *human-replacement failure*, where an incorrect human label overwrites a correct machine one, and *candidate insufficiency*, where neither the machine nor the human label is correct. To address these, we propose a **H**ierarchical **A**ctive **C**orrection with **R**ollback (**HAC-R**) pipeline. Specifically, HAC-R leverages a three-stage annotation pipeline to dynamically acquire reviewed samples via a training-free label-error detection score, resolving conflicts between machine-annotation and noisy-human annotation and escalating to expert-annotation. To leverage diverse annotation source, we further design a routing-aware loss tailored to HAC-R data, with theoretical guarantees on design-based unbiasedness and routing variance. Extensive experiments on six image and text classification benchmarks demonstrate that HAC-R improves annotation accuracy and downstream clean-test accuracy by up to 5% over cost-matched baselines.

1 INTRODUCTION

High-quality labels are essential for training reliable machine learning models, yet large-scale human annotation remains costly and difficult to scale (Zhang et al., 2024; Han et al., 2019; Zhou et al., 2022; Tan et al., 2024; Zhang et al., 2025a). Recent work has explored large language models (LLM) and multimodal large language models (MLLM) as scalable automatic annotators (Zhang et al., 2023; Tan et al., 2024; Xie et al., 2026). Existing methods identify suspicious LLM-generated annotations, and selectively query human reviewers for correction (Gligorić et al., 2025; Lin et al., 2025). In practice, an annotation system generally includes multiple sources (Qin et al., 2025; Wang et al., 2024) with distinct cost-quality tradeoffs, motivating hierarchical review under a shared budget (Zhang et al., 2024). In practice, machines, ordinary reviewers, and experts offer inexpensive coverage, affordable but imperfect feedback, and reliable but costly adjudication, respectively. Thus, due to expensive expert annotation, scalable annotation pipelines rely on non-expert reviewers (Snow et al., 2008; Zhang & Chaudhuri, 2015). However, this depends on an implicit monotonicity premise: *allocating more reviewed budget to ordinary-human review should produce increasingly accurate annotations.*

This assumption fails when human reviewers are imperfect (Rodrigues et al., 2014; Yan et al., 2010). In Figure 1(a), the task is to maximize the annotation accuracy given by diverse annotation sources. We progressively replace machine labels with real crowd annotations on CIFAR-100N (Wei et al., 2022), prioritizing examples ranked as potential machine errors. The blue and green curves report annotation accuracy under replacement with crowd annotations and oracle labels, respectively, for the same selected examples. The review budget is measured by the human replacement rate,

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.



(a) Budget–accuracy path. (b) All examples. (c) Reviewed examples.

Figure 1: **Failure patterns under imperfect human review on CIFAR-100N.** $\eta_{H(M)}$ refers to the human (machine) error rate. Qwen3VL generates the initial labels, and an InternVL3.5-based critic ranks examples by estimated machine-label error. **(a)** Annotation accuracy over the full dataset. **(b–c)** Joint correctness of machine and human labels in the full dataset and the selectively reviewed subset (Budget: 20.8%), respectively.

the fraction of all instances chosen for replacement. Notably, increasing the review budget does not monotonically enhance annotation quality: accuracy first increases and then declines as additional human replacements introduce new errors. Specifically, noisy-human review gives rise to two distinct failure modes. First, **human-replacement failure** occurs when an incorrect human label overwrites a correct machine label, i.e., $Y^M = Y$ and $Y^H \neq Y$. As observed in Figure 1(b), 19.9% of the reviewed examples are correctly labelled by the machine but incorrectly labelled by the human. Second, **candidate insufficiency** occurs when both machine and human labels are incorrect, so the true label is absent from the candidate set $\{Y^M, Y^H\}$, i.e., $Y^M \neq Y$ and $Y^H \neq Y$. Machine-hard examples can also be difficult for human annotators, particularly when difficulty arises from unclear class boundaries (Saha et al., 2022). As shown in Figure 1(b–c), both-wrong rates increase from 13.7% in the full dataset to 41.2% among reviewed examples. Consequently, existing methods neither alleviate the issue of noisy-replacement nor candidate insufficiency.

These findings motivate a *reversible hierarchical correction* strategy that allows human replacements to be reconsidered and reserves costly expert adjudication for annotations that remain unreliable after resolution. We propose **H**ierarchical **A**ctive **C**orrection with **R**ollback (**HAC-R**), which organizes annotation into three stages: *Acquire*, *Resolve*, and *Escalate*. Acquire prioritizes suspicious machine annotations for human review, leveraged by a label-error detection score; Resolve compares the machine and human candidates and permits rollback; and Escalate combines both candidate scores to prioritize expert adjudication. By routing examples with propensity, we construct a routing-aware objective based on hierarchical active M-estimation (Zrnic & Candes, 2024), yielding consistent downstream training performance.

Our contributions are threefold: (i) We identify two failure modes of selective human correction under imperfect feedback: human-replacement failure and candidate insufficiency. (ii) We introduce HAC-R, an *Acquire–Resolve–Escalate* framework that transforms a common reliability signal to select machine annotations for review, rollback harmful human replacements, and reserve expert adjudication for unresolved cases. (iii) We develop a hierarchical active M-estimation objective for downstream learning, establish its unbiasedness for the expert-supervised risk, and derive its variance decomposition. Empirically, we show that HAC and HAC-R lift annotation accuracy and downstream clean-test accuracy by up to 5.53 and 5.72 percentage points, respectively.

2 Problem Settings and Assumption

Data Annotation. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, \ldots, K\}$ the label space. We consider a dataset of N examples $\{X_i\}_{i=1}^N$, where each example has an unobserved label $Y_i \in \mathcal{Y}$. An MLLM annotator produces an initial machine annotation Y_i^M for every example at negligible marginal cost, while ordinary-human annotations Y_i^H and expert annotations Y_i^E are acquired selectively under a

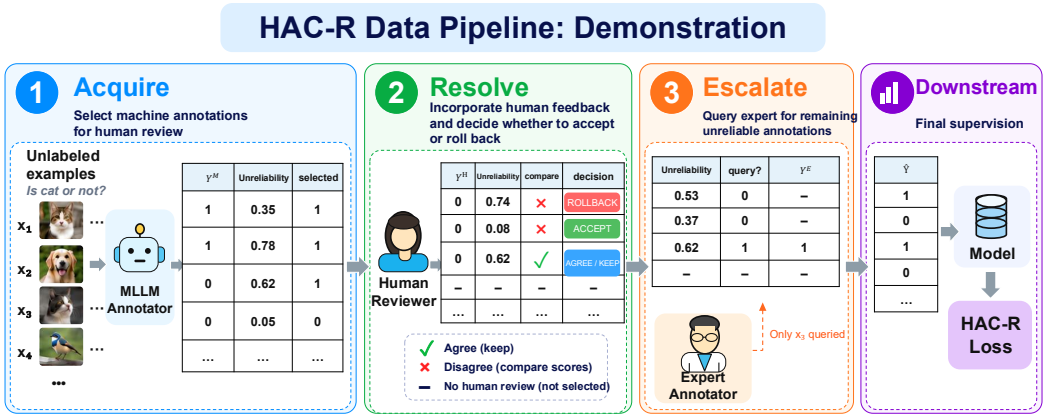


Figure 2: HAC-R Annotation Pipeline.

shared annotation budget. We assume that expert adjudication recovers the ground-truth label, i.e., $Y_i^E = Y_i$.

Hierarchical expertise assumption. For an annotation source $S \in \{M, H, E\}$, define its source noise rate as $\eta_S = 1/N \sum_{i=1}^{N} \mathbb{1}\{Y_i^S \neq Y_i\}$. We assume that paid reviews become more reliable and more costly with increasing expertise. In the main two-tier setting, $\eta_H \geq \eta_E = 0$. The machine is an initial automatic source, we impose no ordering between η_M and η_H. For a generic annotation $\widetilde{Y}_i$, let $s(X_i, \widetilde{Y}_i)$ denote its unreliability score, where a larger value indicates lower reliability. The acquisition decision under an expected query budget b is $\delta_i(b) \sim \text{Bernoulli}\left(\pi_b\left(s(X_i, \widetilde{Y}_i)\right)\right)$, where $\pi_b(\cdot) \in [0, 1]$ is a budget-dependent transformation map satisfying $\sum_{i=1}^{N} \pi_b\left(s(X_i, \widetilde{Y}_i)\right) \leq b$.

Transformation. We consider two acquisition maps: threshold sampling $\pi_b^{\text{thres}}(s_i)$ and exponential sampling $\pi_b^{\text{exp}}(s_i)$ as (Lin et al., 2025):

$$\pi_b^{\text{thres}}(s_i) = \begin{cases} 1 - \varepsilon_b, & s_i \geq t_b, \\ \varepsilon_b, & s_i < t_b, \end{cases} \qquad \pi_b^{\text{exp}}(s_i) = \frac{1}{1 + \exp[-\kappa(s_i - \mu_b)]}, \tag{1}$$

where $\varepsilon_b > 0$ is a probability floor. Thresholding selects the highest-scoring examples, where t_b is the threshold to satisfy the query budget. For exponential weighting, $\kappa > 0$ controls sharpness and μ_b is chosen such that $\sum_{i \in \mathcal{S}} \pi_b^{\text{exp}}(s_i) \leq b$. These rules control the realized and expected query counts, respectively.

Budget Settings. Let $c_h > 0$ denote the unit cost of an ordinary-human annotation and $c_e > c_h$ the unit cost of an expert adjudication. We reserve B for the total annotation-cost budget and allocate stage-wise budgets n_h and n_e to human review and expert adjudication, respectively: $c_h n_h + c_e n_e \leq B$. The corresponding target expected query counts are $n_h := \frac{B_h}{c_h}, n_e := \frac{B_e}{c_e}$.

For the two stages, let $\delta_i^H(n_h)$ and $\delta_i^E(n_e)$ denote the human-review indicator and the conditional expert-query indicator, respectively. We abbreviate them as δ_i^H and δ_i^E when the query budgets are fixed. Their conditional sampling probabilities for the acquisition stage and escalation stage are $p_i := \pi_{n_h}(\hat{s}_i^M), q_i := \pi_{n_e}(\hat{s}_i^E)$, where $\hat{s}_i^M = s(X_i, Y_i^M)$ is the machine-label unreliability score and $\hat{s}_i^E$ is the post-review escalation score defined in Section 3.4. The expert acquisition map operates conditionally on the reviewed set. The transformation parameters are calibrated such that $\mathbb{E}\left[\sum_{i=1}^{N} \delta_i^H\right] = \sum_{i=1}^{N} p_i = n_h, \mathbb{E}\left[\sum_{i=1}^{N} \delta_i^H \delta_i^E\right] = \sum_{i=1}^{N} p_i q_i = n_e$. Equivalently, the expected annotation cost satisfies $C(p, q) = c_h n_h + c_e n_e = B_h + B_e \leq B$.

3 HIERARCHICAL ACTIVE CORRECTION WITH ROLLBACK

3.1 PIPELINE OVERVIEW

Figure 2 details pipeline for HAC(-R). HAC-R contains three decisions. Hierarchical correction organizes annotation sources into multiple levels of supervision, with each level refining the annotation from the previous stage. *Acquire* selects machine annotations for ordinary-human review; *Resolve* chooses between the machine and human candidates; and *Escalate* reserves expert adjudication for reviewed examples that remain unreliable.

Data annotation objective. A hierarchical policy first decides whether to acquire Y_i^H, represented by $\delta_i^H \sim \text{Bernoulli}(p_i)$. After human review, the acceptance indicator $\alpha_i \in \{0, 1\}$ sets $Y_i^R = Y_i^H$ when $\alpha_i = 1$ and $Y_i^R = Y_i^M$ otherwise. Conditional on review, δ_i^E follows a Bernoulli distribution with probability q_i, conditional on the information available after human review, and determines whether Y_i^R should be replaced by Y_i^E. The resulting annotation is $\widehat{Y}_i = (1 - \delta_i^H)Y_i^M + \delta_i^H(1 - \delta_i^E)Y_i^R + \delta_i^H\delta_i^E Y_i^E$, $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_N)$. Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$, and let $Q(\mathbf{Y}, \widehat{\mathbf{Y}})$ denote a dataset-level annotation-quality functional, such as accuracy. The cost-aware HAC-R annotation problem is

$$\max_{\pi_{n_h}, \pi_{n_e}} \mathbb{E}\Big[Q\Big(\mathbf{Y}, \widehat{\mathbf{Y}}\Big)\Big] \quad \text{s.t.} \quad C(p, q) \leq B. \tag{2}$$

3.2 ACQUIRE

The MLLM annotator M first assigns a machine label $Y_i^M = M(X_i)$ to each instance. To better rank potentially unreliable annotations for human review, we adapt a feature- and neighbourhood-based label-error score (Zhang et al., 2025b) to the frozen machine-labeled pool $\mathcal{D}_M = \{(X_j, Y_j^M)\}_{j=1}^N$. The intuition behind it is to leverage the labels of neighbouring samples to identify and correct potential label errors. This score also provides a training-free acquisition signal without relying exclusively on an additional MLLM criticizer, whose judgements may be affected by evaluator bias and self-preference (Chen et al., 2024b; Wataoka et al., 2024). The score s combines input–label compatibility with neighbourhood support:

$$s_i(y) = d_C(i, y) + \beta r_i^X(y) + \gamma r_i^Y(y), \tag{3}$$

where $r_i^X(y)$ measures agreement with labels assigned to nearby examples and $r_i^Y(y)$ measures whether examples associated with nearby labels are also close to X_i, and β, γ weight the two neighbourhood terms. The complete score definition and fixed configuration are provided in Appendix A.4. For text classification, we compute neighbourhood support using frozen text representations and omit the cross-modal term. We evaluate the machine label as $\hat{s}_i^M = s_i(Y_i^M)$ and acquire an ordinary-human annotation with probability $p_i = \pi_{n_h}(\hat{s}_i^M)$, where the budget-dependent map π_{n_h} prioritizes higher-scoring annotations under the human-review budget n_h. This acquisition signal requires neither task-specific model training nor access to the annotator's internal logits, making it applicable to black-box MLLMs. The same candidate-conditioned score is subsequently used to compare machine and human annotations in RESOLVE.

3.3 RESOLVE

Human review provides an alternative annotation instead of a guaranteed correction. The purpose of RESOLVE is to determine which available candidate should retain. Ideally, an oracle resolver would prefer whichever of Y_i^M and Y_i^H has lower annotation error. To alleviate the *human-replacement failure* under zero-shot settings, we approximate this decision by evaluating both candidates with the same label-conditional unreliability score from Eq. 3. This paired comparison requires only the within-example ordering of the two scores. Specifically, we compute $\hat{s}_i^H = s_i(Y_i^H)$ and define human-acceptance indicator α_i, then the estimated human-acceptance indicator $\hat{\alpha}_i$ and resolved label Y^R are:

$$\hat{\alpha}_i = \mathbb{1}\{\hat{s}_i^H < \hat{s}_i^M\}, \qquad Y_i^R = \begin{cases} Y_i^H, & \hat{\alpha}_i = 1, \\ Y_i^M, & \hat{\alpha}_i = 0. \end{cases} \tag{4}$$

Otherwise, HAC-R accepts human feedback only when it is better supported than the machine annotation under the shared scoring function; otherwise, it rolls back to Y_i^M. This comparison only

determines which available candidate is better supported. It cannot recover the correct label when both candidates are unreliable, which motivates the expert-escalation stage.

3.4 ESCALATE

Expert adjudication addresses *candidate insufficiency*, where neither the machine nor the human candidate is correct. To prioritize such cases, we combine $\hat{s}_i^M$ and $\hat{s}_i^H$ into a residual-unreliability score $\hat{s}_i^E$ for expert routing. Since lower scores indicate greater reliability, $\min(\hat{s}_i^M, \hat{s}_i^H)$ provides a natural starting point: expert review is warranted when even the candidate with stronger evidential support remains unreliable. To smooth the transition between the two candidate scores, we replace the hard minimum with a differentiable approximation:

$$\mathrm{SoftMin}_\tau(a,b) := -\tau \log\left(\frac{e^{-a/\tau}+e^{-b/\tau}}{2}\right), \qquad \hat{s}_i^E = \mathrm{SoftMin}_\tau(\hat{s}_i^M, \hat{s}_i^H), \qquad \tau = 1. \quad (5)$$

We set $\tau = 1$ in all experiments; the fixed-budget sensitivity analysis in Appendix D.5 shows that annotation accuracy is stable across moderate and large temperatures The normalization gives $\mathrm{SoftMin}_\tau(a,a) = a$, with $\min(\hat{s}_i^M, \hat{s}_i^H) \le \hat{s}_i^E \le (\hat{s}_i^M + \hat{s}_i^H)/2$ and $\mathrm{SoftMin}_\tau(a,b) \to \min(a,b)$ as $\tau \to 0^+$. Expert routing takes place only among reviewed examples. Conditional on the information available after human review and before expert adjudication, the sampling rule is $\delta_i^E(n_e) \sim \mathrm{Bernoulli}\big(\pi_{n_e}(\hat{s}_i^E)\big), \delta_i^H(n_h) = 1$, where π_{n_e} allocates the expert-query budget over the reviewed set. If the example is escalated, the expert annotation Y_i^E replaces Y_i^R; otherwise, Y_i^R is retained. Examples not acquired for human review retain Y_i^M. Expert adjudication can therefore correct both *candidate insufficiency* and errors in source selection, regardless of whether RESOLVE accepted the human label or rolled back to the machine label.

4 DOWNSTREAM LEARNING WITH HAC-R DATA

The annotation pipeline in Sec. 3 produces a dataset containing retained machine labels, accepted human annotations, and selectively acquired expert labels. Learning from this dataset requires accounting for both annotation errors and selective supervision. Training exclusively on expert-labeled examples discards substantial supervision from the remaining data and confines learning to the selectively adjudicated subset. Treating all retained annotations as ground truth makes use of the full dataset but propagates residual labeling errors into downstream training. We therefore formulate a hierarchical active M-estimation objective that integrates machine, resolved, and expert supervision through sequential loss corrections weighted by the corresponding inverse query propensities.

Hierarchical active M-estimation. Active M-estimation integrates machine predictions and selectively acquired human annotations through inverse-propensity correction (Lin et al., 2025; Zrnic & Candes, 2024). HAC-R introduces an additional decision between human review and expert escalation: after observing Y_i^H, Resolve determines whether the machine or human annotation should be retained. We therefore apply the correction twice, following the two levels of the annotation hierarchy. Let $\ell_{\theta,i}^M$, $\ell_{\theta,i}^H$, and $\ell_{\theta,i}$ denote the losses computed using Y_i^M, Y_i^H, and Y_i^E. The resolved loss is

$$\ell_{\theta,i}^R = \ell(f_\theta(X_i), Y_i^R) = \alpha_i \ell_{\theta,i}^H + (1-\alpha_i)\ell_{\theta,i}^M, \quad (6)$$

where $\alpha_i = 1$ accepts the human annotation and $\alpha_i = 0$ rolls back to the machine annotation. We retain the notation p_i and q_i from Sec. 2: q_i is conditional on the information available after human review, and both q_i and α_i are chosen before observing the expert annotation. The HAC-R loss is

$$\mathcal{L}_\theta^{\text{HAC-R}} = \frac{1}{N}\sum_{i=1}^N \Big[\ell_{\theta,i}^M + \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)}\big(\ell_{\theta,i}^R - \ell_{\theta,i}^M\big) + \frac{\delta_i^H(n_h)\delta_i^E(n_e)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}\big(\ell_{\theta,i} - \ell_{\theta,i}^R\big)\Big]. \quad (7)$$

where the three terms represent machine supervision, resolution-induced correction, and expert correction of the residual loss discrepancy. The standard HAC is recovered when $\alpha_i \equiv 1$, in which case every human annotation is accepted and $\ell_{\theta,i}^R = \ell_{\theta,i}^H$. The resolution correction satisfies $\ell_{\theta,i}^R - \ell_{\theta,i}^M = \alpha_i(\ell_{\theta,i}^H - \ell_{\theta,i}^M)$. When human feedback is accepted, the objective corrects the machine loss through the human loss. Under rollback, the intermediate correction vanishes, and the expert term directly corrects the machine loss while retaining the two-stage propensity weight.

Proposition 1 (Statistical properties of the HAC-R loss) *For fixed θ, let $\delta_i^H(n_h) \sim \mathrm{Bernoulli}(\pi_{n_h}(\hat{s}_i^M))$, with $\pi_{n_h}(\hat{s}_i^M) \geq p_{\min} > 0$. Conditional on the human-review history, each reviewed example is queried for expert adjudication with probability $\pi_{n_e}(\hat{s}_i^E) \geq q_{\min} > 0$. Both α_i and the expert-query propensity are determined before observing $Y_i^E = Y_i$. Then $\mathbb{E}\big[\mathcal{L}_\theta^{\text{HAC-R}}\big] = \mathcal{L}_\theta$ Under independent first-stage sampling and conditionally independent expert sampling given the human-review history, the routing variance is*

$$\mathrm{Var}\big(\mathcal{L}_\theta^{\text{HAC-R}}\big) = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Bigg[\Big(\frac{1}{\pi_{n_h}(\hat{s}_i^M)}-1\Big)(\ell_{\theta,i}-\ell_{\theta,i}^M)^2 + \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)^2}\Big(\frac{1}{\pi_{n_e}(\hat{s}_i^E)}-1\Big)(\ell_{\theta,i}-\ell_{\theta,i}^R)^2\Bigg]. \tag{8}$$

By Proposition 1 the HAC(-R) objective is unbiased under positive-support stochastic routing greater than a lower bound $p_{\min}, q_{\min}$, while its variance depends on both acquisition and resolution (See proof in Appendix B.5). In practice, we set $p_{\min}$ and $q_{\min}$ as 10^{-4}. The first arises from discrepancies between machine and expert losses, motivating human acquisition that prioritizes unreliable machine annotations. The second reveals two complementary ways to improve estimation precision: *(1) perfect resolver*, which retains the candidate whose loss is closer to the expert loss; and *(2) astute escalation*, which assigns higher expert-query probabilities to examples with large post-resolution residuals. For a reviewed example, this term vanishes when $\ell_{\theta,i}^R = \ell_{\theta,i}$ or $\pi_{n_h}(\hat{s}_i^M) = 1$. The decomposition reveals complementary roles for resolution and escalation: accurate resolution reduces the residual loss discrepancy, while expert escalation attenuates its contribution to variance.

5 Experiments

In this section, we empirically evaluate the performance of HAC and (HAC-R) from both of the *annotation* and *downstream learning* perspectives, aiming to answer the following questions: (1) Does ACQUIRE identify machine errors that ordinary-human review can correct? (2) When does RESOLVE prevent more harmful replacements than beneficial corrections it rejects? (3) Does ESCALATE address *candidate insufficiency* by targeting cases where neither the machine nor human annotation is correct? (4) How do improvements in annotation quality translate into downstream?

5.1 Experimental Setup

Datasets. We empirically evaluate our pipeline on six image- and text-classification benchmarks, including Image Classification: CIFAR-100N (Wei et al., 2022), Fashion-MNIST (Xiao et al., 2017), Stanford Cars (Krause et al., 2013), and Mini-ImageNet (Vinyals et al., 2016); Text Classification: TREC (Li & Roth, 2002), and 20 Newsgroups (Lang, 1995).

Imperfect human feedback and metrics. Machine predictions and human annotations are generated once and reused across all methods. Reference annotations are reserved for expert adjudication, with each query charged to the annotation budget, and offline evaluation. For CIFAR-100N, Y^H is the original crowd annotation. Stanford Cars, Mini-ImageNet, Fashion-MNIST, TREC, Irony, and 20 Newsgroups use fixed controlled human sources contaminated by Instance-Dependent Noise (IDN) (Xia et al., 2020). The main controlled setting has error rate $\eta_H = \{0.1, 0.4\}$. The resulting machine-human error structure is summarised in Appendix D.4. The generated annotations are fixed across compared methods; Appendix A.3 and Algorithm 1 provide the generation details. We report *final annotation accuracy* as the metric for evaluating annotation quality; and report *clean-test accuracy* for evaluating downstream learning.

MLLMs. For image datasets, we use Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) as the machine annotator, and InternVL3.5-30B-A3B (Wang et al., 2025) as the criticizer for ACT Lin et al. (2025). For text datasets, we use Qwen3-30B-A3B (Yang et al., 2025) as the corresponding text-only backbone.

Downstream Models. We use an ImageNet-pretrained ResNet-18 (He et al., 2016) for image classification and RoBERTa-base (Liu et al., 2019) for text classification. For tasks shared with ACT (Lin et al., 2025), we follow its downstream architecture and training hyperparameters. Each annotation method uses the same model, data split, and optimization protocol. Complete configurations are reported in Appendix A.2.

Comparison methods. We compare HAC and HAC-R with machine-only and noisy-human-only training, as well as cost-matched CDI (Gligorić et al., 2025) and ACT (Lin et al., 2025) baselines using ordinary-human or expert corrections.

Annotation cost and budget settings. We set $c_h = 1$ and $c_e = 10$ and compare all methods under the nominal budget $B = \widehat{\eta}_M N$, where $\widehat{\eta}_M$ is the empirical machine error rate (Lin et al., 2025). This budget corresponds to one ordinary review per estimated machine error. HAC and HAC-R reserve a fraction $\rho_E = \widehat{\eta}_H$ for expert adjudication, yielding $n_e = \lfloor \rho_E B / c_e \rfloor$ expert queries and $n_h = \lfloor (1 - \rho_E) B / c_h \rfloor$ human reviews. For example, on a dataset of 10,000 instances with 80% machine accuracy and a human error rate of 40%, the total budget is 2,000 cost units: 1,200 fund ordinary reviews, while the remaining 800 fund 80 expert queries. Complete budget settings and sensitivity analysis are provided in Appendix Table 6 and D.1.

5.2 Experiment Results

Table 1: Annotation accuracy (%) under a common nominal budget $B = N\widehat{\eta}_M$. For CIFAR-100N with real crowd annotation. $+H$ and $+E$ denote correction using imperfect ordinary-human labels or oracle-expert labels. $\text{HAC}_{\text{Thres}}$ and HAC_{Exp} refers to HAC with threshold sampling and exponential sampling. Bold marks the highest accuracy among budget-constrained methods for each dataset and human-noise setting.

Method	Stanford Cars	Mini-ImageNet	Fashion-MNIST	TREC	20 Newsgroups	CIFAR-100N
Machine only	70.887	86.234	82.918	64.602	72.299	79.766
Noisy human only	90.005/59.995	90.000/60.000	90.000/60.000	89.994/59.996	90.003/60.002	59.800
ACT $+H$	76.117/67.350	88.766/84.156	86.147/80.442	74.221/63.460	**82.982**/73.035	76.260
ACT $+E$	72.065	86.796	83.503	65.967	**73.994**	80.776
CDI $+H$	78.200/68.610	87.060/82.743	84.353/79.060	73.324/63.012	77.771/69.073	75.911
CDI $+E$	71.933	86.462	83.240	65.910	73.163	80.245
$\text{HAC}_{\text{thres}}$	**84.443**/75.258	**92.384**/87.814	**86.420**/82.570	**77.603**/67.129	80.853/73.237	80.374
$\text{HAC-R}_{\text{thres}}$	83.730/**75.626**	91.844/**88.066**	85.958/**83.848**	76.605/**67.781**	77.908/73.832	**81.530**
HAC_{exp}	79.786/71.719	88.693/85.422	85.109/81.485	75.313/65.731	78.689/71.746	78.837
$\text{HAC-R}_{\text{exp}}$	79.484/73.310	88.690/86.973	84.609/83.336	73.654/66.860	76.233/73.181	80.870

Experiments on annotation quality. Table 1 compares the annotation quality of HAC and HAC-R with the baselines under the same budget. The gains are more pronounced on image datasets, with improvements of 5–8% on Cars and Mini-ImageNet. HAC performs well when η_h is low and human feedback is reliable, while HAC-R becomes more advantageous as η_h increases, highlighting the value of rollback under noisy human supervision. Overall, the pipeline improves annotation quality across the evaluated levels of human reliability without increasing the annotation budget.

5.3 Discussion

Table 2: Summary of ablation experiments. All metrics are reported as percentages.

Stage	Variant	CIFAR-100N	Cars	Mini-ImageNet	TREC
Acquire	Random	12.1	17.6	8.3	21.3
	Entropy	39.0	38.9	38.1	30.5
	Criticizer	**41.1**	17.7	22.4	20.5
	Ours	36.0	**46.9**	**42.8**	**33.1**
Resolve	Random	50.0	50.0	50.0	50.0
	HAC (Always H)	50.3	80.9	71.4	66.1
	HAC-R	**79.0**	**89.0**	**82.4**	**78.2**
Escalate	Min	**49.8**	67.0	59.3	34.8
	Mean	48.8	**71.3**	64.4	**42.0**
	Ambiguity	24.1	17.0	29.1	31.9
	SoftMin	48.8	**71.3**	**65.1**	**42.0**

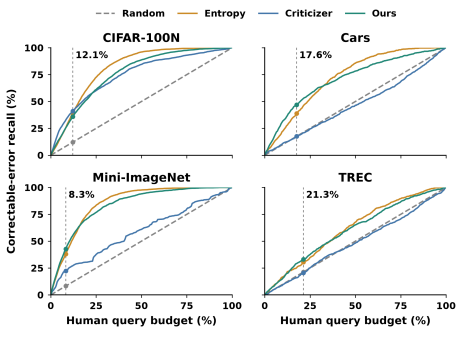


Figure 3: Acquisition recall with fixed annotations.

Ablation Study. Table 2 evaluates each stage across image and text datasets. *Correctable-error recall* measures the proportion of machine errors with correct human alternatives ($Z_i^M = 1, Z_i^H =$

0) selected for review ($\delta_i^H = 1$). *Correct-source accuracy* measures how often resolution selects the correct label ($Z_i^R = 0$) among reviewed examples with exactly one correct candidate ($Z_i^M + Z_i^H = 1$). *Expert-query correction rate* measures the proportion of expert queries ($\delta_i^H \delta_i^E = 1$) that correct an erroneous resolved label ($Z_i^R = 1$). Label-error scoring improves correctable-error recall over random acquisition. Among reviewed examples with exactly one correct candidate, RESOLVE enhances source-selection accuracy over always accepting human labels by 8.1–28.7 percentage points. SoftMin outperforms ambiguity-based escalation on all four datasets, although the best scoring rule varies by dataset.

Answering RQ1: Analysis on the acquisition strategy. Figure 3 compares logit-based predictive entropy (Li et al., 2023), an MLLM-based criticizer (Wang et al., 2024), and our label-error detection score under the same human-review budget. Our score achieves the highest correctable-error recall on Cars, Mini-ImageNet, and TREC, outperforming entropy by 8.0, 4.7, and 2.6 percentage points, respectively. On CIFAR-100N, the criticizer performs best, while our score remains competitive with entropy. Nevertheless, the criticizer does not transfer consistently across datasets, falling close to random acquisition on Cars, Mini-ImageNet, and TREC.

Answering RQ2: Analysis on the resolve module. Figure 4(a–b) shows the trade-off between prevented harm ($Y^M = Y$ and $Y^H \neq Y$) and rejected corrections ($Y^M \neq Y$ and $Y^H = Y$). At $\eta_H = 0.10$, rejected corrections outweigh prevented harm across all five controlled-noise datasets, reducing annotation accuracy by 0.49-2.99 percentage points relative to HAC before expert intervention. At $\eta_H = 0.40$, prevented harm outweighs rejected corrections, yielding net gains of 0.30–1.32 percentage points. Rollback improves annotation accuracy by 1.35 percentage points on CIFAR-100N. **Resolve therefore improves annotation quality in the high-noise setting evaluated, but rejects beneficial corrections in the low-noise setting**, when human annotations are more reliable. Further details are provided in Appendix D.2.

Answering RQ3: Analysis on the expert adjudication. Figure 4(c) shows how expert escalation addresses *candidate insufficiency* by correcting cases where both machine and human annotations are wrong. Under high-noise setting ($\eta_h = 0.40$), 40.4-68.1% of expert queries correct such cases across six datasets. These corrections recover labels that RESOLVE module cannot obtain by selecting between the two candidates, demonstrating the complementary role of expert supervision.

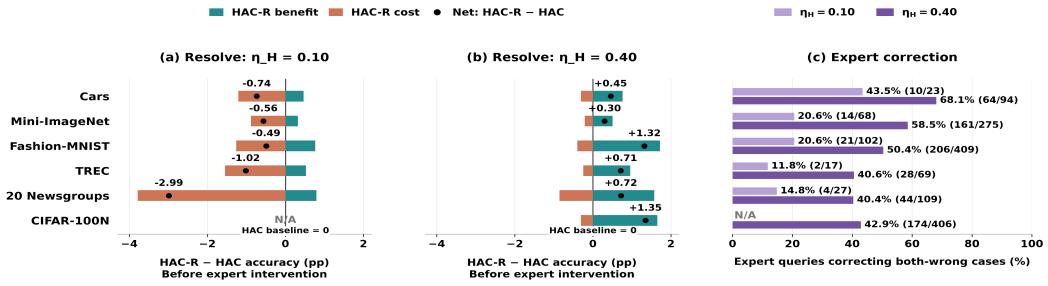


Figure 4: (a–b) Rollback benefits and costs. Black dots showing the net annotation accuracy gain of HAC-R over HAC. (c) Expert queries correcting cases where both annotations are wrong.

Answering RQ4: Experiments on downstream learning. We adopt the Cross-entropy loss and apply the power-tuning hyperparameter λ for all HAC losses following (Lin et al., 2025). Table 3 reports clean-test accuracy under matched annotation costs across controlled and real-crowd settings. HAC performs consistently well across both human-noise levels, showing that hierarchical correction effectively combines machine, human, and expert supervision for downstream learning. HAC-R further improves performance in selected high-noise settings, including Stanford Cars and real-crowd CIFAR-100N. This result aligns with our analysis: rollback preserves the target risk, but improves downstream learning only when it reduces the residual error after resolution.

Table 3: Matched-cost downstream clean-test accuracy (%; mean and standard deviation over five seeds). Bold indicates the best result among cost-matched methods.

Dataset	Machine only	Human only	ACT+H	ACT+E	CDI+H	CDI+E	HAC Exp	HAC-R Exp	HAC Thres	HAC-R Thres
Controlled human error $\eta_H = 0.10$										
Stanford Cars	54.65(1.17)	79.95(0.40)	67.17(0.24)	63.69(0.26)	69.18(0.69)	63.09(0.37)	75.60(0.48)	74.83(0.35)	**75.80(0.24)**	74.90(0.27)
Mini-ImageNet	70.89(1.19)	72.25(1.53)	71.93(0.73)	71.38(1.33)	71.39(1.11)	70.77(0.56)	73.73(1.13)	73.02(1.08)	**73.75(1.04)**	73.72(0.52)
Fashion-MNIST	83.77(1.28)	92.20(0.34)	85.56(0.92)	84.15(1.19)	85.09(1.34)	83.95(0.69)	**86.66(1.21)**	84.68(1.29)	85.39(1.67)	84.48(1.17)
TREC	76.32(1.16)	97.04(0.17)	85.12(1.08)	77.56(1.07)	86.60(1.48)	77.60(1.08)	89.08(0.52)	**89.20(1.08)**	88.36(1.09)	88.52(1.30)
20 Newsgroups	66.68(0.23)	71.75(0.28)	**70.06(0.24)**	66.90(0.26)	68.23(0.48)	66.72(0.39)	69.42(0.23)	68.66(0.23)	69.32(0.13)	68.55(0.40)
Controlled human error $\eta_H = 0.40$										
Stanford Cars	54.65(1.17)	55.72(0.85)	61.67(0.27)	63.69(0.26)	62.56(0.16)	63.09(0.37)	64.58(0.48)	65.35(0.66)	67.38(0.42)	**67.67(0.37)**
Mini-ImageNet	70.89(1.19)	48.48(1.33)	69.93(1.45)	71.38(1.33)	68.77(0.72)	70.77(0.56)	71.92(0.82)	71.54(1.51)	71.67(1.07)	**72.01(1.12)**
Fashion-MNIST	83.77(1.28)	79.42(3.38)	83.16(1.52)	**84.15(1.19)**	82.60(2.27)	83.95(0.69)	84.05(1.00)	82.77(1.53)	83.40(1.58)	83.57(1.21)
TREC	76.32(1.16)	92.60(0.95)	80.48(1.33)	77.56(1.07)	80.48(0.89)	77.60(1.08)	81.24(1.47)	79.44(1.31)	**82.64(0.93)**	81.96(0.26)
20 Newsgroups	66.68(0.23)	65.66(0.25)	**67.73(0.47)**	66.90(0.26)	66.78(0.48)	66.72(0.39)	67.69(0.28)	67.56(0.29)	67.62(0.42)	67.67(0.27)
Real crowd feedback on CIFAR-100N ($\eta_H = 0.402$)										
CIFAR-100N	66.43(0.63)	57.64(0.98)	65.42(0.89)	67.35(0.88)	64.44(0.57)	66.44(0.62)	66.47(1.29)	**67.64(0.80)**	67.57(0.76)	67.51(0.92)

6 Related Work

Human-LLM Collaborative Annotation. Annotation from LLM has emerged as a promising alternative for human labelling in the area of NLP (Tan et al., 2024; Zhang et al., 2023; Pangakis & Wolken, 2024; Tseng et al., 2024; Casanova et al., 2026). Despite these advances, fully automated annotation remains susceptible to systematic and instance-dependent errors, motivating selective human involvement for unreliable machine-generated labels. Additionally, the potential of annotation by MLLMs tailored to image datasets has not been fully explored. Hybrid annotation leverages two annotation sources to optimize the hybrid annotation accuracy (Xiao et al., 2023). Several works leverage Human-LLM collaboration for annotation via uncertainty-based rules to query unreliable samples for human correction (Li et al., 2023; Kim et al., 2024), or by introducing another MLLM verifier to route low-confidence samples (Wang et al., 2024). CDI (Gligorić et al., 2025) and ACT (Lin et al., 2025) balance between annotation accuracy and cost via the LLM confidence. The most relevant work is ACT (Lin et al., 2025), which leverages a cross-family MLLM criticizer to collaborate with a human annotator. Nevertheless, HAC-R treats human feedback as an imperfect intermediate source. We establish an *Acquire–Resolve–Escalate* pipeline that can roll back harmful human corrections and selectively invoke expert adjudication when both annotations remain suspicious, while a hierarchical active M-estimation objective is tailored to downstream inference.

Cost-Aware Hierarchical and Multi-Source Annotation. Prior work has considered annotation sources with heterogeneous costs and reliability, ranging from weak and strong labelers to modern multi-source annotation systems (Zhang & Chaudhuri, 2015; Zhang et al., 2025a; Qin et al., 2025). Recent methods further optimize annotation routing across multiple sources under explicit quality-cost constraints with a labeled calibration set, as in HyPAC (Zeng et al., 2026), whereas QUORUM dynamically allocates examples across human and LLM annotators and supports repeated annotation under a fixed budget (Purificato et al., 2026). These methods primarily address which source should annotate an instance. In contrast, HAC-R considers *reversible sequential correction* after an initial machine annotation, where an acquired human review is imperfect and requires being rolled back before unresolved cases are selectively escalated to an expert sequentially.

7 Conclusions

In conclusion, we identify two failure patterns when human feedback is imperfect, which are *human-replacement failure* and *candidate insufficiency*. We address them with Hierarchical Active Correction (HAC) and its rollback extension, HAC-R. HAC-R introduces an *Acquire–Resolve–Escalate* pipeline that treats human review as an imperfect, reversible intermediate stage. We further develop a hierarchical active M-estimation objective for downstream learning from the resulting multi-source supervision. Extensive experiments across diverse image and text datasets validate its robust empirical performance on both annotation quality and downstream learning. Future work will explore reliability-adaptive rollback and extend HAC-R to imperfect and heterogeneous real experts.

AI USE STATEMENT

In this work, we used generative AI tools to help develop and refine the conceptual framework, formulate mathematical claims, assist with proof derivations, refine hypotheses, design experiments, implement and debug methods, and interpret experimental results. We did not use generative AI to generate datasets, translate research material, clean or reformat datasets, or perform qualitative or thematic data analysis; the remaining required-disclosure tasks are not applicable to this work. We additionally used generative AI tools for brainstorming, literature search and summarization, suggesting experimental parameters, creating and editing code and scientific figures or tables, structuring and drafting portions of the manuscript, and improving its readability. We reviewed all AI-assisted work. The mathematical arguments were checked by the authors, code was executed and tested, and reported measurements were verified against the corresponding experimental artifacts. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Gabriel Afriat, Ryan Lucas, Xiang Meng, Yufang Hou, Yada Zhu, and Rahul Mazumder. Seeing what's wrong: A trajectory-guided approach to caption error detection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 66918–66948, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6db3ea527f53682657b3d6b02a841340-Paper-Conference.pdf.

Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-NN for noisy labels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 540–550. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bahri20a.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Etienne Casanova, Rafal Kocielnik, and R. Michael Alvarez. On the limits of LLM adaptability: Impact of model-internalized priors on annotation task performance. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=oTv2bKG5Qg. Spotlight, Oral Presentation.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024a. URL https://arxiv.org/abs/2402.04788.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/chen24h.html.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. Can unconfident LLM annotations be used for confident conclusions? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3514–3533, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.179. URL https://aclanthology.org/2025.naacl-long.179/.

Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5137–5146, 2019. doi: 10.1109/ICCV.2019.00524.

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 (CVPR)*, June 2016.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. MEGAnno+: A human-LLM collaborative annotation system. In Nikolaos Aletras and Orphee De Clercq (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 168–176, St. Julians, Malta, March

2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.18. URL https://aclanthology.org/2024.eacl-demo.18/.

Shuitsu Koyama, Yuiga Wada, Daichi Yashima, and Komei Sugiura. Mllm-as-a-judge exhibits model preference bias, 2026. URL https://arxiv.org/abs/2604.11589.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Ken Lang. Newsweeder: learning to filter netnews. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pp. 331–339, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603778.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1487–1505, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.92. URL https://aclanthology.org/2023.emnlp-main.92/.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150/.

Lequan Lin, Dai Shi, Andi Han, Feng Chen, Qiuzheng Chen, Jiawen Li, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, and Junbin Gao. Act as human: Multimodal large language model data annotation with critical thinking. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 24107–24138. Curran Associates, Inc., 2025. doi: 10.52202/085713-0814. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/22bbfc36cb32d0baf7cd1c832da6f5e9-Paper-Conference.pdf.

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.

Nicholas Pangakis and Sam Wolken. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In Dallas Card, Anjalie Field, Dirk Hovy, and Katherine Keith (eds.), *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pp. 113–131, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlpcss-1.9. URL https://aclanthology.org/2024.nlpcss-1.9/.

Antonio Purificato, Maria Sofia Bucarelli, Andrea Bacciu, Amin Mantrach, and Fabrizio Silvestri. Quorum: Quality-optimized routing using multiple annotators, 2026. URL https://arxiv.org/abs/2608.27974.

Maosheng Qin, Renyu Zhu, Mingxuan Xia, Chenchenkai, Zhen Zhu, Minmin Lin, Junbo Zhao, Lu Xu, Changjie Fan, Runze Wu, and Haobo Wang. CrowdAgent: Multi-agent managed multi-source annotation system. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 925–942, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.72. URL https://aclanthology.org/2025.emnlp-demos.72/.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Gaussian process classification and active learning with multiple annotators. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 433–441, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/rodrigues14.html.

Swarnadeep Saha, Peter Hase, Nazneen Rajani, and Mohit Bansal. Are hard examples also harder to explain? a study with human and model-generated explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2121–2131, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.137. URL https://aclanthology.org/2022.emnlp-main.137/.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In Mirella Lapata and Hwee Tou Ng (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1027/.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 930–957, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.54. URL https://aclanthology.org/2024.emnlp-main.54/.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. Are expert-level language models expert-level annotators?, 2024. URL https://arxiv.org/abs/2410.03254.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641960. URL https://doi.org/10.1145/3613904.3641960.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations, 2022. URL https://arxiv.org/abs/2110.12088.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7597–7610. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5607fe8879e4fd269e88387e8cb30b7e-Paper.pdf.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. FreeAL: Towards human-free active learning in the era of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14520–14535, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.896. URL https://aclanthology.org/2023.emnlp-main.896/.

Ming-Kun Xie, Jia-Hao Xiao, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Are multimodal large language models good annotators for image tagging?, 2026. URL https://arxiv.org/abs/2602.20972.

Yan Yan, Romer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 932–939, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/yan10a.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hao Zeng, Huipeng Huang, Xinhao Qu, Jianguo Huang, Bingyi Jing, and Hongxin Wei. Hypac: Cost-efficient llms-human hybrid annotation with pac error guarantees, 2026. URL https://arxiv.org/abs/2602.02550.

Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/eba0dc302bcd9a273f8bbb72be3a687b-Paper.pdf.

Haodi Zhang, Junyu Yang, Wenxi Huang, Min Cai, Jiahong Li, Chen Zhang, and Kaishun Wu. Recognizing textual entailment by hierarchical crowdsourcing with diverse labor costs. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 453–458, 2024. doi: 10.1109/CSCWD61410.2024.10580016.

Haodi Zhang, Junyu Yang, Jinyin Nie, Peirou Liang, Kaishun Wu, Defu Lian, Rui Mao, and Yuanfeng Song. Efficient data labeling by hierarchical crowdsourcing with large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11290–11303, Abu Dhabi, UAE, January 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.748/.

Haoran Zhang, Aparna Balagopalan, Nassim Oufattole, Hyewon Jeong, Yan Wu, Jiacheng Zhu, and Marzyeh Ghassemi. LEMon: Label error detection using multimodal neighbors. In *Forty-second International Conference on Machine Learning*, 2025b.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. LLMaAA: Making large language models as active annotators. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13088–13103, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.872. URL https://aclanthology.org/2023.findings-emnlp.872/.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Prototype-anchored learning for learning with imperfect annotations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27245–27267. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22f.html.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27412–27427. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhu22a.html.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

APPENDIX CONTENTS

A Datasets and Experiment Details **17**

A.1 Dataset Details. 17

A.2 Backbone 17

A.3 Experiment Setting 17

A.4 Training-free LEMoN Score. 19

A.5 Additional Method Overview 19

B Theoretical Proofs **21**

B.1 Notation 21

B.2 HAC and HAC-R Objectives with Different Sampling Rules 22

B.3 Additional Statistical Guarantees 22

B.4 Proof of HAC 23

B.5 Proof of HAC-R 28

B.6 Corollary 1 29

C Extension to Multiple Imperfect Human Annotators **29**

D Supplementary Experiment Details **31**

D.1 Analysis of Budget Allocation. 31

D.2 Net Value of Rollback 33

D.3 Rollback Retention and Estimator Variance 33

D.4 Joint-Error Analysis 33

D.5 Sensitivity Analysis of τ 34

E Prompts **35**

E.1 Annotator Prompts 35

E.2 Criticizer Prompt 36

F Additional Discussions and Related Works **37**

F.1 Additional Related Works. 37

A Datasets and Experiment Details

A.1 Dataset Details.

CIFAR-100N provides real crowd annotations and serves as our primary real-world imperfect-human benchmark. Stanford Cars evaluates fine-grained recognition with a large label space, while Mini-ImageNet provides a broader 100-class recognition setting. Fashion-MNIST provides a lower-complexity, low-class-count regime for evaluating whether hierarchical correction remains stable when the initial annotation problem is relatively easy. TREC and 20 Newsgroups extend the study to short- and long-form text classification, respectively.

Table 4 distinguishes real crowd feedback from controlled human noise and summarizes the role of each benchmark. The controlled settings isolate changes in reviewer reliability, while CIFAR-100N tests the pipeline with naturally occurring annotation errors.

Table 4: Dataset annotation sources and roles. Controlled human sources are generated once and shared by all compared methods.

Dataset	Modality	Human source	Primary role
CIFAR-100N	Image	Real crowd	Main benchmark
Stanford Cars	Image	Controlled IDN	Fine-grained
Mini-ImageNet	Image	Controlled IDN	General recognition
Fashion-MNIST	Image	Controlled IDN	Low-complexity regime
TREC	Text	Controlled IDN	Short-text transfer
20 Newsgroups	Text	Controlled IDN	Long-form transfer

A.2 Backbone

Annotation pipeline backbones. All annotation decisions are produced in a zero-shot manner, without fine-tuning any of the multimodal or language models. The machine annotator and criticizer use different visual encoders and multimodal alignment pipelines.

For image annotation, the machine annotator is `Qwen/Qwen3-VL-30B-A3B-Instruct`, a 30B-parameter mixture-of-experts model with approximately 3B active parameters. The criticizer is `OpenGVLab/InternVL3_5-30B-A3B`, which combines an InternViT-6B visual encoder with a Qwen3-MoE language backbone. Although the annotator and criticizer share related language-model components, they use distinct visual encoders and multimodal alignment pipelines.

For text datasets, we use the text-only counterparts from the same model families whenever available. In particular, the default machine annotator is the 30B-A3B Qwen3 instruction model, while the criticizer retains the InternVL/Qwen3 backbone. All competing annotation pipelines share the same machine predictions and human labels; hence, differences between methods cannot be attributed to different annotator outputs.

A.3 Experiment Setting

Table 6 compares how the methods spend the nominal budget on human and expert queries. ACT uses a single paid source, whereas HAC and HAC-R split queries across the two stages. CDI's fractional counts are averages over routing seeds rather than deterministic query targets, so equal nominal budgets need not imply equal realized costs.

Implementation Details. All experiments are conducted on high-performance computing infrastructure using PyTorch 1.11.0 and an NVIDIA A100 GPU. All multi-run experiments use five random seeds, listed as `0,1,2,3,4`.

Controlled instance-dependent human noise. We use a fixed-count variant of feature-dependent label corruption. For each dataset, we uniformly select $m = \mathrm{round}(\eta_H N)$ instances without replacement, where $\eta_H \in \{0.10, 0.40\}$. We sample a shared projection matrix $W \in \mathbb{R}^{d \times K}$ with

Table 5: MLLM configuration for machine annotation.

Parameter	Setting
Backbone	Qwen3-VL-30B-A3B-Instruct
Inference backend	Hugging Face Transformers
Numerical precision	BF16
Input image size	224×224
Image interpolation	Bicubic
Inference batch size	1
Annotation scheme	Direct K-way classification
Prediction rule	Argmax over concatenated option logits
Confidence	Maximum joint softmax probability
Softmax temperature	1 (unscaled logits)
Chain-of-thought	Disabled
Self-consistency	Disabled
Autoregressive generation	Not used
Fine-tuning	None (frozen backbone)

Table 6: Per-method query allocations under the matched nominal cost budget $B = \widehat{\eta}_M N$, with human and expert query costs $c_H = 1$ and $c_E = 10$. Mean query counts are rounded to the nearest integer.

Modality	Dataset	η_H	B	ACT+H n_h	ACT+E n_e	CDI+H n_h	CDI+E n_e	HAC / HAC-R	
								n_h	n_e
Image	Stanford Cars	0.1	2,371	2,371	237	2,354	236	2,133	23
		0.4	2,371	2,371	237	2,355	236	1,422	94
Image	Mini-ImageNet	0.1	6,883	6,883	688	6,802	668	6,194	68
		0.4	6,883	6,883	688	6,815	668	4,129	275
Image	Fashion-MNIST	0.1	10,249	10,249	1,024	10,218	1,021	9,224	102
		0.4	10,249	10,249	1,024	10,212	1,021	6,149	409
Text	TREC	0.1	1,737	1,737	173	1,720	168	1,563	17
		0.4	1,737	1,737	173	1,721	168	1,042	69
Text	20 Newsgroups	0.1	2,746	2,746	274	2,734	274	2,471	27
		0.4	2,746	2,746	274	2,733	274	1,647	109

independent standard-normal entries. For each selected instance, we draw an incorrect label according to

$$\Pr(\hat{y}_i^H = c \mid i \text{ is selected}) = \frac{\exp(x_i^\top w_c)}{\sum_{k \neq y_i} \exp(x_i^\top w_k)}, \qquad c \neq y_i.$$

Unselected labels remain unchanged, yielding exactly m annotation errors. Thus, corruption locations are sampled uniformly, while the distribution over incorrect classes depends on instance features. The generated annotations are reused across compared methods, as shown in Algorithm 1

Downstream Training. Following ACT (Lin et al., 2025) For image classification, we fine-tune an ImageNet-pretrained ResNet-18; for text classification, we fine-tune RoBERTa-base. We select the optimization settings separately for each dataset and keep the backbone, training duration, and batch size identical across competing methods within each dataset.

Learning rates are selected from $\{10^{-2}, 10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$ for image tasks and $\{10^{-4}, 5 \times 10^{-5}, 2 \times 10^{-5}, 10^{-5}\}$ for text tasks. For HAC and HAC-R, we tie the power-tuning parameters, $\lambda_M = \lambda_H = \lambda$, and search over $\lambda \in \{0.5, 0.75, 1.25, 1.5\}$. Hyperparameters are selected separately for each method and dataset using validation accuracy, with the same validation split and search budget across methods. The selected configuration is then evaluated over five random seeds without further tuning on the test set.

Algorithm 1 Fixed-count instance-dependent human-noise generation

Require: Features $\{z_i \in \mathbb{R}^d\}_{i=1}^N$, reference labels $\{y_i\}_{i=1}^N$, number of classes K, noise level η_H, seed $s = 0$
Ensure: Simulated human labels $\{\hat{y}_i^H\}_{i=1}^N$
$m \leftarrow \mathrm{round}(\eta_H N)$
Draw $u_i \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0,1)$
$C \leftarrow$ indices of the m largest u_i
$\hat{y}^H \leftarrow y$
Sample $W \in \mathbb{R}^{d\times K}$ with independent $\mathcal{N}(0,1)$ entries
for $i \in C$ in ascending index order **do**
    $a_i \leftarrow W^\top z_i$
    $a_{i,y_i} \leftarrow -\infty$
    $p_i \leftarrow \mathrm{softmax}(a_i)$
    Sample $\hat{y}_i^H \sim \mathrm{Categorical}(p_i)$
end for
return $\{\hat{y}_i^H\}_{i=1}^N$

Stabilized HAC-Exp training. The exploratory stabilized HAC-Exp configuration uses CE warm-up on the frozen final annotations before switching to HAC-Exp with $(\lambda_M, \lambda_H) = (1, 1)$ and bounded logits. The warm-up provides a CE-trained initialization for the hierarchical objective and is intended to mitigate early optimization instability. This configuration applies specifically to the stabilized HAC-Exp runs, currently available for seed 0, and not to the Threshold or final-label CE runs. Since these Exp runs also use a different routing history, they compare complete training configurations rather than isolate the effect of the loss alone.

Budget Settings. Table 6 relates dataset size and class count to the review workload. For controlled datasets, paired entries show how the human–expert allocation changes between the two noise settings; CIFAR-100N is reported separately with its real crowd labels.

A.4 Training-free LEMoN Score.

Let $v_i = g_X(X_i)$ and $t_y = g_Y(y)$ denote the frozen CLIP embeddings of image X_i and label y. We use cosine distance to capture cross-modal spaces, $d_X(i,j) = 1 - \cos(v_i, v_j), d_C(i,y) = 1 - \cos(v_i, t_y)$, and the discrete distance $d_Y(y,y') = \mathbb{1}\{y \neq y'\}$ in the label space. Let $\mathcal{N}_X(i)$ be the indices of the k reference images closest to X_i under d_X, and let $\mathcal{N}_Y(y)$ be the indices of the k reference labels closest to candidate y under d_Y. We score an arbitrary candidate label y as

$$
\begin{aligned}
r_i^X(y) &= \frac{1}{k} \sum_{j \in \mathcal{N}_X(i)} d_Y(y, Y_j^M) \exp\left[-\tau_{1X} d_X(i,j) - \tau_{2X} d_C(j, Y_j^M)\right], \\
r_i^Y(y) &= \frac{1}{k} \sum_{j \in \mathcal{N}_Y(y)} d_X(i,j) \exp\left[-\tau_{1Y} d_Y(y, Y_j^M) - \tau_{2Y} d_C(j, Y_j^M)\right], \\
s_i(y) &= d_C(i,y) + \beta r_i^X(y) + \gamma r_i^Y(y).
\end{aligned}
\tag{9}
$$

The first neighbourhood term $r_i^X(y)$ measures the support for y among images similar to X_i. The second $r_i^Y(y)$ measures whether images carrying labels close to y are also close to X_i. The direct term measures the compatibility between X_i and y. We use the fixed LEMoN configuration $k = 30$, $\beta = \gamma = 5$, $\tau_{1X} = \tau_{1Y} = 0.1$, and $\tau_{2X} = \tau_{2Y} = 5$ (Zhang et al., 2025b), without task-specific training or tuning. The machine-label unreliability score is $\hat{s}_i^M = s_i(Y_i^M)$, with larger values indicating less reliable labels. For text classification, we compute neighbourhoods using frozen text representations and omit the cross-modal term.

A.5 Additional Method Overview

Figure 5 contrasts the routing structures of expert-only ACT, HAC, and HAC-R. HAC inserts ordinary-human review before expert adjudication, while HAC-R additionally compares the human

candidate with the machine label before accepting a replacement. The rollback arrow denotes retaining the original machine label, not an additional annotation query. The paths show the available decisions; expert adjudication is selective rather than mandatory for every reviewed example.

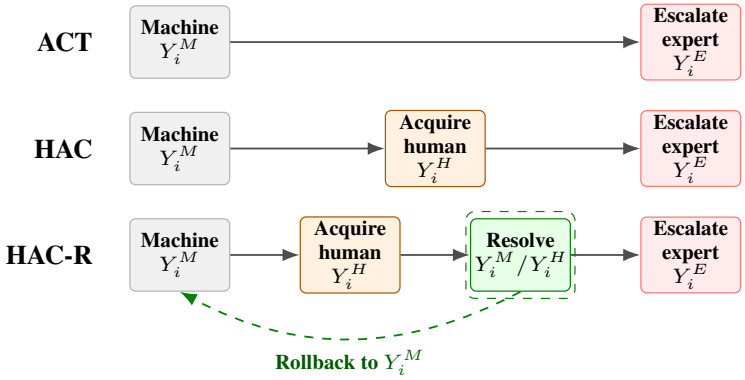


Figure 5: Routing comparison. ACT routes machine labels directly to experts, HAC adds human review, and HAC-R resolves human feedback with optional rollback before escalation.

B THEORETICAL PROOFS

B.1 NOTATION

Table 7: Summary of notation.

Symbol	Definition
X_i	Input example i
Y_i	Ground-truth annotation
Y_i^E	Expert annotation
Y_i^M	Initial machine annotation
Y_i^H	Annotation provided by the imperfect human reviewer
Y_i^R	Resolved annotation after machine–human comparison
$\widehat{Y}_i$	Final annotation produced by the routing pipeline
$\mathcal{F}_i^M$	Information available before human review
$\mathcal{F}_i^H$	Information available after human review and before expert adjudication
Z_i^M	Machine-annotation error indicator, $\mathbb{1}\{Y_i^M \neq Y_i\}$
Z_i^H	Human-annotation error indicator, $\mathbb{1}\{Y_i^H \neq Y_i\}$
Z_i^R	Resolved-annotation error indicator, $\mathbb{1}\{Y_i^R \neq Y_i\}$
η_i^M	Conditional machine error risk, $\Pr(Z_i^M = 1 \mid \mathcal{F}_i^M)$
η_i^H	Conditional human error risk, $\Pr(Z_i^H = 1 \mid \mathcal{F}_i^H)$
$\widehat{\eta}_H$	Human-source error rate specified or estimated before routing
d_X, d_Y, d_C	Image, label, and cross-modal distances used by Full LEMoN
$\mathcal{N}_X(i), \mathcal{N}_Y(y)$	The k nearest reference indices in image and label space
$r_i^X(y), r_i^Y(y), s_i(y)$	Image-neighbor term, label-neighbor term, and Full LEMoN candidate score in Eq. 3
$\hat{s}_i^M$	Machine unreliability score, $s_i(Y_i^M)$
$\hat{s}_i^H$	Human unreliability score, $s_i(Y_i^H)$
$\hat{s}_i^E$	Normalized SoftMin escalation score defined in Eq. 5
$g \in \{\mathrm{A}, \mathrm{R}\}$	Acquisition or within-example resolution score-comparison regime
$\widetilde{s}_i^M, \widetilde{s}_i^E$	Stable rank-normalized Acquire and Escalate scores
$\pi_b(\cdot)$	Mapping from a routing score to an acquisition probability under expected query budget b
n_h	Expected number of human-review queries, $\sum_i p_i$
n_e	Expected number of expert-adjudication queries, $\mathbb{E}[\sum_i \delta_i^H \delta_i^E]$
ρ_E	Fraction of the total annotation cost allocated to expert adjudication
ρ_H	Fraction of the total annotation cost allocated to ordinary-human review
$p_{\min}, q_{\min}$	Positive lower bounds on the two routing propensities
γ_H, γ_E	Normalizers chosen to satisfy the human and expert query budgets
δ_i^H	Human-review indicator, shorthand for $\delta_i^H(n_h)$
δ_i^E	Conditional expert-query indicator, shorthand for $\delta_i^E(n_e)$
δ_i^{ACT}	Indicator that ACT acquires example i directly for expert correction
$\delta_i(b)$	Generic acquisition indicator under expected query budget b, with $\delta_i(b) \sim \mathrm{Bernoulli}\Big(\pi_b(s_i(\widetilde{Y}_i))\Big)$ for a generic annotation $\widetilde{Y}_i$
p_i	Human-review propensity, $p_i = \pi_{n_h}(\hat{s}_i^M)$
q_i	Conditional expert-adjudication propensity, $q_i = \pi_{n_e}(\hat{s}_i^E)$
α_i	Human-acceptance indicator, $\mathbb{1}\{\hat{s}_i^H < \hat{s}_i^M\}$
c_h	Cost of one ordinary-human annotation
c_e	Cost of one expert adjudication
B	Total annotation-cost budget
$\ell_{\theta,i}^M$	$\ell(f_\theta(X_i), Y_i^M)$
$\ell_{\theta,i}^H$	$\ell(f_\theta(X_i), Y_i^H)$

Continued on the next page

Table 7 continued

Symbol	Definition
$\ell^R_{\theta,i}$	$\ell(f_\theta(X_i), Y_i^R)$
$\ell_{\theta,i}$	$\ell(f_\theta(X_i), Y_i)$
λ_M	Machine-annotation power-tuning parameter controlling the contribution of ℓ_i^M
λ_H	Human-annotation power-tuning parameter controlling the contribution of ℓ_i^H when the human annotation is accepted
$L^\star_\theta$	Fully reference-supervised empirical objective

B.2 HAC and HAC-R Objectives with Different Sampling Rules

For all rules, define the power-weighted resolved loss as

$$\ell^{R,\lambda}_{\theta,i} = \alpha_i \lambda_H \ell^H_{\theta,i} + (1-\alpha_i)\lambda_M \ell^M_{\theta,i}. \tag{10}$$

Exponential weighting. Let

$$p_i^{\exp} = \left[1+\exp\{-\beta_H(\widetilde{s}_i^M - a_H)\}\right]^{-1}, \qquad q_i^{\exp} = \left[1+\exp\{-\beta_E(\widetilde{s}_i^E - a_E)\}\right]^{-1}, \tag{11}$$

where a_H, a_E and $\beta_H, \beta_E > 0$ control the centers and sharpness of the two allocation maps and are set to meet the corresponding budgets. For the Bernoulli decisions induced by these propensities, the objective is

$$\mathcal{L}_\theta^{\text{HAC-R-exp}} = \frac{1}{N}\sum_{i=1}^N \left[\lambda_M \ell^M_{\theta,i} + \frac{\delta_i^{H,\exp}}{p_i^{\exp}}\left(\ell^{R,\lambda}_{\theta,i} - \lambda_M \ell^M_{\theta,i}\right) + \frac{\delta_i^{H,\exp}\delta_i^{E,\exp}}{p_i^{\exp} q_i^{\exp}}\left(\ell_{\theta,i} - \ell^{R,\lambda}_{\theta,i}\right)\right]. \tag{12}$$

Threshold. Let t_H and t_E be thresholds selected such that the resulting propensities satisfy the expected human-review and expert-query budgets. We define the positive-support threshold propensities as

$$p_i^{\text{thre}} = \begin{cases} 1-p_{\min}, & \hat{s}_i^M \ge t_H, \\ p_{\min}, & \hat{s}_i^M < t_H, \end{cases} \qquad q_i^{\text{thre}} = \begin{cases} 1-q_{\min}, & \hat{s}_i^E \ge t_E, \\ q_{\min}, & \hat{s}_i^E < t_E. \end{cases} \tag{13}$$

Human review and expert adjudication are then sampled according to

$$\delta_i^{H,\text{thre}} \sim \text{Bernoulli}(p_i^{\text{thre}}), \qquad \delta_i^{E,\text{thre}} \mid \delta_i^{H,\text{thre}} = 1 \sim \text{Bernoulli}(q_i^{\text{thre}}). \tag{14}$$

The corresponding positive-support stochastic objective is

$$\begin{aligned}\mathcal{L}_\theta^{\text{HAC-R-thre}} = \frac{1}{N}\sum_{i=1}^N \Bigg[&\lambda_M \ell^M_{\theta,i} + \frac{\delta_i^{H,\text{thre}}}{p_i^{\text{thre}}}\left(\ell^{R,\lambda}_{\theta,i} - \lambda_M \ell^M_{\theta,i}\right) \\ &+ \frac{\delta_i^{H,\text{thre}}\delta_i^{E,\text{thre}}}{p_i^{\text{thre}} q_i^{\text{thre}}}\left(\ell_{\theta,i} - \ell^{R,\lambda}_{\theta,i}\right)\Bigg].\end{aligned} \tag{15}$$

We set $p_{\min} = q_{\min} = 10^{-4}$ and use the realized propensities in the inverse-propensity weights. Setting $\alpha_i \equiv 1$ recovers the corresponding HAC objective.

In practice, we impose probability floors of $p_{\min} = q_{\min} = 10^{-4}$ before stochastic sampling and use the resulting probabilities in the inverse-propensity weights.

B.3 Additional Statistical Guarantees

B.3.1 Probabilistic upper bound of the parameter gap

Theorem 1 (Parameter gap under hierarchical active correction) *Suppose that $\mathcal{L}_\theta^{\text{HAC}}$ is μ-strongly convex. For each example i, let $\delta_i^H(B_h) \sim \text{Bernoulli}(\pi_{B_h}(\hat{s}_i^M))$ and, conditional on*

$\delta_i^H(B_h) = 1$, *let* $\delta_i^E(B_e) \sim \text{Bernoulli}(\pi_{B_e}(\hat{s}_i^E))$, *where* $\pi_{B_h}(\hat{s}_i^M), \pi_{B_e}(\hat{s}_i^E) \in (0, 1]$. *Assume that the sequential sampling decisions are independent across examples.*

Define the machine and human gradient discrepancies at $\theta^\star$ *as*

$$D_i^M = \nabla \ell_{\theta^\star,i}^M - \nabla \ell_{\theta^\star,i}, \qquad D_i^H = \nabla \ell_{\theta^\star,i}^H - \nabla \ell_{\theta^\star,i}. \tag{16}$$

Let

$$V_N = \sum_{i=1}^N \left[\frac{1 - \pi_{B_h}(\hat{s}_i^M)}{\pi_{B_h}(\hat{s}_i^M)} \left\| D_i^M \right\|_2^2 + \frac{1 - \pi_{B_e}(\hat{s}_i^E)}{\pi_{B_h}(\hat{s}_i^M)\pi_{B_e}(\hat{s}_i^E)} \left\| D_i^H \right\|_2^2 \right], \tag{17}$$

and

$$M_N = \max_{1 \le i \le N} \left[\max\left\{1, \frac{1 - \pi_{B_h}(\hat{s}_i^M)}{\pi_{B_h}(\hat{s}_i^M)}\right\} \left\| D_i^M \right\|_2 + \frac{1}{\pi_{B_h}(\hat{s}_i^M)} \max\left\{1, \frac{1 - \pi_{B_e}(\hat{s}_i^E)}{\pi_{B_e}(\hat{s}_i^E)}\right\} \left\| D_i^H \right\|_2 \right]. \tag{18}$$

Then, for any confidence parameter $\eta \in (0, 1)$, *with probability at least* $1 - \eta$ *over the hierarchical sampling decisions,*

$$\left\| \widehat{\theta}_{\text{HAC}} - \theta^\star \right\|_2 \le \frac{1}{\mu N} \left[\sqrt{2 V_N \log\left(\frac{2}{\eta}\right)} + \frac{2 M_N}{3} \log\left(\frac{2}{\eta}\right) \right]. \tag{19}$$

B.4 Proof of HAC

B.4.1 Proof of Proposition 1

Proof 1 *Recall that the HAC loss is defined as*

$$\mathcal{L}_\theta^{(\text{HAC})} = \frac{1}{N} \sum_{i=1}^N \left[\ell_{\theta,i}^M + \left(\ell_{\theta,i}^H - \ell_{\theta,i}^M\right) \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)} + \left(\ell_{\theta,i} - \ell_{\theta,i}^H\right) \frac{\delta_i^H(n_h)\delta_i^E(n_e)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)} \right], \tag{20}$$

where

$$\delta_i^H(n_h) \sim \text{Bernoulli}(\pi_{n_h}(\hat{s}_i^M))$$

and

$$\delta_i^E(n_e) \mid \delta_i^H(n_h) = 1 \sim \text{Bernoulli}(\pi_{n_e}(\hat{s}_i^E)).$$

We define

$$Z_i = \ell_{\theta,i}^M + \left(\ell_{\theta,i}^H - \ell_{\theta,i}^M\right) \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)} + \left(\ell_{\theta,i} - \ell_{\theta,i}^H\right) \frac{\delta_i^H(n_h)\delta_i^E(n_e)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}, \tag{21}$$

such that

$$\mathcal{L}_\theta^{(\text{HAC})} = \frac{1}{N} \sum_{i=1}^N Z_i.$$

Assuming that the examples and their sampling decisions are independent across i, *the variance of the HAC loss is*

$$\text{Var}\left(\mathcal{L}_\theta^{(\text{HAC})}\right) = \text{Var}\left(\frac{1}{N} \sum_{i=1}^N Z_i\right) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(Z_i). \tag{22}$$

Hence, it suffices to calculate the variance of Z_i, *which can be written as*

$$\text{Var}(Z_i) = \mathbb{E}[Z_i^2] - \left(\mathbb{E}[Z_i]\right)^2.$$

For notational simplicity, let

$$A_i = \ell_{\theta,i}^H - \ell_{\theta,i}^M$$

and

$$C_i = \ell_{\theta,i} - \ell_{\theta,i}^H.$$

Then

$$Z_i = \ell^M_{\theta,i} + A_i \frac{\delta^H_i(n_h)}{\pi_{n_h}(\hat{s}^M_i)} + C_i \frac{\delta^H_i(n_h)\delta^E_i(n_e)}{\pi_{n_h}(\hat{s}^M_i)\pi_{n_e}(\hat{s}^E_i)}.$$

Expanding its square gives

$$\begin{aligned} Z_i^2 = \ & \left(\ell^M_{\theta,i}\right)^2 + A_i^2 \frac{\delta^H_i(n_h)}{\pi_{n_h}(\hat{s}^M_i)^2} + C_i^2 \frac{\delta^H_i(n_h)\delta^E_i(n_e)}{\pi_{n_h}(\hat{s}^M_i)^2\pi_{n_e}(\hat{s}^E_i)^2} \\ & + 2\ell^M_{\theta,i} A_i \frac{\delta^H_i(n_h)}{\pi_{n_h}(\hat{s}^M_i)} \\ & + 2\ell^M_{\theta,i} C_i \frac{\delta^H_i(n_h)\delta^E_i(n_e)}{\pi_{n_h}(\hat{s}^M_i)\pi_{n_e}(\hat{s}^E_i)} \\ & + 2A_i C_i \frac{\delta^H_i(n_h)\delta^E_i(n_e)}{\pi_{n_h}(\hat{s}^M_i)^2\pi_{n_e}(\hat{s}^E_i)}, \end{aligned} \tag{23}$$

where we use

$$\left(\delta^H_i(n_h)\right)^2 = \delta^H_i(n_h)$$

and

$$\left(\delta^H_i(n_h)\delta^E_i(n_e)\right)^2 = \delta^H_i(n_h)\delta^E_i(n_e).$$

Since

$$\mathbb{E}\left[\delta^H_i(n_h)\right] = \pi_{n_h}(\hat{s}^M_i)$$

and

$$\mathbb{E}\left[\delta^H_i(n_h)\delta^E_i(n_e)\right] = \pi_{n_h}(\hat{s}^M_i)\pi_{n_e}(\hat{s}^E_i),$$

taking the expectation of Equation 23 yields

$$\begin{aligned} \mathbb{E}[Z_i^2] = \ & \left(\ell^M_{\theta,i}\right)^2 + \frac{A_i^2}{\pi_{n_h}(\hat{s}^M_i)} + \frac{C_i^2}{\pi_{n_h}(\hat{s}^M_i)\pi_{n_e}(\hat{s}^E_i)} \\ & + 2\ell^M_{\theta,i} A_i + 2\ell^M_{\theta,i} C_i + \frac{2A_i C_i}{\pi_{n_h}(\hat{s}^M_i)}. \end{aligned} \tag{24}$$

Using

$$A_i + C_i = \ell_{\theta,i} - \ell^M_{\theta,i},$$

Equation 24 can be rearranged as

$$\begin{aligned} \mathbb{E}[Z_i^2] = \ & \ell^2_{\theta,i} \\ & + \left(\ell_{\theta,i} - \ell^M_{\theta,i}\right)^2 \left(\frac{1}{\pi_{n_h}(\hat{s}^M_i)} - 1\right) \\ & + \left(\ell_{\theta,i} - \ell^H_{\theta,i}\right)^2 \left(\frac{1}{\pi_{n_h}(\hat{s}^M_i)\pi_{n_e}(\hat{s}^E_i)} - \frac{1}{\pi_{n_h}(\hat{s}^M_i)}\right). \end{aligned} \tag{25}$$

Equivalently,

$$\begin{aligned} \mathbb{E}[Z_i^2] = \ell^2_{\theta,i} & + \left(\ell_{\theta,i} - \ell^M_{\theta,i}\right)^2 \left(\frac{1}{\pi_{n_h}(\hat{s}^M_i)} - 1\right) \\ & + \left(\ell_{\theta,i} - \ell^H_{\theta,i}\right)^2 \frac{1}{\pi_{n_h}(\hat{s}^M_i)} \left(\frac{1}{\pi_{n_e}(\hat{s}^E_i)} - 1\right). \end{aligned} \tag{26}$$

Moreover, the HAC correction is unbiased:

$$\begin{aligned} \mathbb{E}[Z_i] &= \ell^M_{\theta,i} + \left(\ell^H_{\theta,i} - \ell^M_{\theta,i}\right) + \left(\ell_{\theta,i} - \ell^H_{\theta,i}\right) \\ &= \ell_{\theta,i}. \end{aligned} \tag{27}$$

Applying the law of total variance, we obtain

$$\begin{aligned}\mathrm{Var}(Z_i) =\ & \mathrm{Var}\left(\mathbb{E}[Z_i]\right) + \mathbb{E}\left[\mathrm{Var}(Z_i)\right] \\ =\ & \mathrm{Var}(\ell_{\theta,i}) \\ & + \mathbb{E}\left[\left(\ell_{\theta,i} - \ell_{\theta,i}^{M}\right)^2 \left(\frac{1}{\pi_{n_h}(\hat{s}_i^M)} - 1\right)\right] \\ & + \mathbb{E}\left[\left(\ell_{\theta,i} - \ell_{\theta,i}^{H}\right)^2 \frac{1}{\pi_{n_h}(\hat{s}_i^M)} \left(\frac{1}{\pi_{n_e}(\hat{s}_i^E)} - 1\right)\right].\end{aligned} \tag{28}$$

Finally, substituting Equation 28 into Equation 22 gives

$$\begin{aligned}\mathrm{Var}\left(\mathcal{L}_\theta^{(\mathrm{HAC})}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\Bigg\{ & \mathrm{Var}(\ell_{\theta,i}) \\ & + \mathbb{E}\left[\left(\ell_{\theta,i} - \ell_{\theta,i}^{M}\right)^2 \left(\frac{1}{\pi_{n_h}(\hat{s}_i^M)} - 1\right)\right] \\ & + \mathbb{E}\left[\left(\ell_{\theta,i} - \ell_{\theta,i}^{H}\right)^2 \frac{1}{\pi_{n_h}(\hat{s}_i^M)} \left(\frac{1}{\pi_{n_e}(\hat{s}_i^E)} - 1\right)\right]\Bigg\}.\end{aligned} \tag{29}$$

This completes the proof.

B.4.2 Proof of Theorem 1

Proof 2

$$\begin{aligned}\nabla\mathcal{L}_\theta^{(\mathrm{HAC})} &= \frac{1}{N}\sum_{i=1}^{N}\pi_i\nabla\ell_{\theta,i} + \frac{1}{N}\sum_{i=1}^{N}\left(\pi_i^H - \pi_i\right)\nabla\ell_{\theta,i}^H \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left(1-\pi_i^H\right)\nabla\ell_{\theta,i}^M \\ &= \nabla\mathcal{L}_\theta + \left(\frac{1}{N}\sum_{i=1}^{N}\pi_i\nabla\ell_{\theta,i} - \frac{1}{N}\sum_{i=1}^{N}\nabla\ell_{\theta,i}\right) \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left(\pi_i^H - \pi_i\right)\nabla\ell_{\theta,i}^H + \frac{1}{N}\sum_{i=1}^{N}\left(1-\pi_i^H\right)\nabla\ell_{\theta,i}^M \\ &= \nabla\mathcal{L}_\theta + \frac{1}{N}\sum_{i=1}^{N}\left(1-\pi_i^H\right)\left(\nabla\ell_{\theta,i}^M - \nabla\ell_{\theta,i}\right) \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left(\pi_i^H - \pi_i\right)\left(\nabla\ell_{\theta,i}^H - \nabla\ell_{\theta,i}\right).\end{aligned}$$

where $\mathcal{L}_\theta = \frac{1}{N}\sum_{i=1}^{N}\ell_{\theta,i}.$, $\pi_i^H = \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)}$, $\pi_i = \frac{\delta_i^H(n_h)\delta_i^E(n_e)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}$.

It is easy to see that

$$\mathbb{E}[Z_i] = 0,$$

where

$$Z_i = \left(1-\pi_i^H\right)\left(\nabla\ell_{\theta,i}^M - \nabla\ell_{\theta,i}\right) + \left(\pi_i^H - \pi_i\right)\left(\nabla\ell_{\theta,i}^H - \nabla\ell_{\theta,i}\right).$$

Recall that $\pi_{n_h}(\hat{s}_i^M)$ *is the probability of querying the human reviewer, while* $\pi_{n_e}(\hat{s}_i^E)$ *is the conditional probability of querying the expert after human review. Assume that there exist constants* $p_{\min}, q_{\min} > 0$ *such that*

$$\pi_{n_h}(\hat{s}_i^M) \geq p_{\min}, \qquad \pi_{n_e}(\hat{s}_i^E) \geq q_{\min}$$

We further assume that there exist constants $C_m, C_h > 0$ such that

$$\left\|\nabla \ell^M_{\theta,i} - \nabla \ell_{\theta,i}\right\|_2 \le C_m, \qquad \left\|\nabla \ell^H_{\theta,i} - \nabla \ell_{\theta,i}\right\|_2 \le C_h$$

for all $i \in \{1, \dots, N\}$.

We first derive an almost-sure bound on $\|Z_i\|_2$. Since

$$\pi_i^H = \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)},$$

we have

$$\left|1 - \pi_i^H\right| = \begin{cases} 1, & \delta_i^H(n_h) = 0, \\ \dfrac{1 - \pi_{n_h}(\hat{s}_i^M)}{\pi_{n_h}(\hat{s}_i^M)}, & \delta_i^H(n_h) = 1. \end{cases}$$

Therefore,

$$\left|1 - \pi_i^H\right| \le \max\left\{1, \frac{1 - p_{\min}}{p_{\min}}\right\}.$$

Similarly, since

$$\pi_i^H - \pi_i = \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)} \left(1 - \frac{\delta_i^E(n_e)}{\pi_{n_e}(\hat{s}_i^E)}\right),$$

we have

$$\left|\pi_i^H - \pi_i\right| = \begin{cases} 0, & \delta_i^H(n_h) = 0, \\ \dfrac{1}{\pi_{n_h}(\hat{s}_i^M)}, & \delta_i^H(n_h) = 1, \delta_i^E(n_e) = 0, \\ \dfrac{1 - \pi_{n_e}(\hat{s}_i^E)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}, & \delta_i^H(n_h) = 1, \delta_i^E(n_e) = 1. \end{cases}$$

It follows that

$$\left|\pi_i^H - \pi_i\right| \le \frac{1}{p_{\min}} \max\left\{1, \frac{1 - q_{\min}}{q_{\min}}\right\}.$$

Applying the triangle inequality gives

$$\begin{aligned} \|Z_i\|_2 \le\ & \left|1 - \pi_i^H\right| \left\|\nabla \ell^M_{\theta,i} - \nabla \ell_{\theta,i}\right\|_2 \\ & + \left|\pi_i^H - \pi_i\right| \left\|\nabla \ell^H_{\theta,i} - \nabla \ell_{\theta,i}\right\|_2 \\ \le\ & \max\left\{1, \frac{1 - p_{\min}}{p_{\min}}\right\} C_m + \frac{1}{p_{\min}} \max\left\{1, \frac{1 - q_{\min}}{q_{\min}}\right\} C_h =: c_0. \end{aligned} \tag{30}$$

We next bound the second moment of Z_i. Under sequential Bernoulli sampling,

$$\mathbb{E}\left[\left(1 - \pi_i^H\right)^2\right] = \frac{1 - \pi_{n_h}(\hat{s}_i^M)}{\pi_{n_h}(\hat{s}_i^M)}.$$

Moreover,

$$\mathbb{E}\left[\left(\pi_i^H - \pi_i\right)^2\right] = \frac{1 - \pi_{n_e}(\hat{s}_i^E)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}.$$

The cross term vanishes under iterated expectation:

$$\begin{aligned} \mathbb{E}\Big[& \left(1 - \pi_i^H\right)\left(\pi_i^H - \pi_i\right) \\ & \times \left\langle \nabla \ell^M_{\theta,i} - \nabla \ell_{\theta,i}, \nabla \ell^H_{\theta,i} - \nabla \ell_{\theta,i} \right\rangle \Big] = 0. \end{aligned} \tag{31}$$

Consequently,

$$\begin{aligned} \mathbb{E}\left[\|Z_i\|_2^2\right] &= \frac{1 - \pi_{n_h}(\hat{s}_i^M)}{\pi_{n_h}(\hat{s}_i^M)} \left\|\nabla \ell^M_{\theta,i} - \nabla \ell_{\theta,i}\right\|_2^2 + \frac{1 - \pi_{n_e}(\hat{s}_i^E)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)} \left\|\nabla \ell^H_{\theta,i} - \nabla \ell_{\theta,i}\right\|_2^2 \\ &\le \frac{1 - p_{\min}}{p_{\min}} C_m^2 + \frac{1 - q_{\min}}{p_{\min} q_{\min}} C_h^2 =: c_1. \end{aligned} \tag{32}$$

To retain the instance-dependent sampling information, define

$$v_i := \frac{1-\pi_{n_h}(\hat{s}_i^M)}{\pi_{n_h}(\hat{s}_i^M)} \left\|\nabla \ell^M_{\theta^\star,i} - \nabla \ell_{\theta^\star,i}\right\|_2^2 + \frac{1-\pi_{n_e}(\hat{s}_i^E)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)} \left\|\nabla \ell^H_{\theta^\star,i} - \nabla \ell_{\theta^\star,i}\right\|_2^2,$$

and let

$$V_N = \sum_{i=1}^N v_i. \tag{33}$$

By the preceding second-moment decomposition, $V_N = \sum_{i=1}^N \mathbb{E}\left[\|Z_{\theta^\star,i}\|_2^2\right]$. *In particular, the worst-case variance bound in Eq. 32 implies*

$$V_N \le N c_1,$$

whereas V_N *retains the individual sampling probabilities and gradient gaps.*

We also define an instance-dependent almost-sure bound

$$\begin{aligned} M_i := \max\left\{1, \frac{1-\pi_{n_h}(\hat{s}_i^M)}{\pi_{n_h}(\hat{s}_i^M)}\right\} \left\|\nabla \ell^M_{\theta^\star,i} - \nabla \ell_{\theta^\star,i}\right\|_2 \\ + \frac{1}{\pi_{n_h}(\hat{s}_i^M)} \max\left\{1, \frac{1-\pi_{n_e}(\hat{s}_i^E)}{\pi_{n_e}(\hat{s}_i^E)}\right\} \left\|\nabla \ell^H_{\theta^\star,i} - \nabla \ell_{\theta^\star,i}\right\|_2, \end{aligned} \tag{34}$$

and let

$$M_N = \max_{1\le i\le N} M_i. \tag{35}$$

It follows that $\|Z_{\theta^\star,i}\|_2 \le M_N$ *almost surely for every* i.

We now apply the vector Bernstein inequality to the independent, zero-mean random vectors $Z_{\theta^\star,1}, \dots, Z_{\theta^\star,N}$. *For every* $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^N Z_{\theta^\star,i}\right\|_2 \ge \sqrt{2V_N t} + \frac{2M_N t}{3}\right) \le 2e^{-t}. \tag{36}$$

Setting

$$t = \log(2/\eta), \qquad \eta \in (0,1),$$

shows that, with probability at least $1-\eta$,

$$\left\|\frac{1}{N}\sum_{i=1}^N Z_{\theta^\star,i}\right\|_2 \le \frac{\sqrt{2V_N \log(2/\eta)}}{N} + \frac{2M_N \log(2/\eta)}{3N}. \tag{37}$$

Finally, let

$$\widehat{\theta}_{\mathrm{HAC}} \in \arg\min_\theta \mathcal{L}_\theta^{(\mathrm{HAC})}, \qquad \theta^\star \in \arg\min_\theta \mathcal{L}_\theta.$$

Since $\mathcal{L}_\theta^{(\mathrm{HAC})}$ *is* μ*-strongly convex,*

$$\begin{aligned} \left\|\widehat{\theta}_{\mathrm{HAC}} - \theta^\star\right\|_2 &\le \frac{1}{\mu}\left\|\nabla \mathcal{L}^{(\mathrm{HAC})}_{\widehat{\theta}_{\mathrm{HAC}}} - \nabla \mathcal{L}^{(\mathrm{HAC})}_{\theta^\star}\right\|_2 \\ &= \frac{1}{\mu}\left\|\nabla \mathcal{L}^{(\mathrm{HAC})}_{\theta^\star}\right\|_2, \end{aligned} \tag{38}$$

where

$$\nabla \mathcal{L}^{(\mathrm{HAC})}_{\widehat{\theta}_{\mathrm{HAC}}} = 0.$$

Moreover, since

$$\nabla \mathcal{L}_{\theta^\star} = 0,$$

the HAC gradient decomposition gives

$$\nabla \mathcal{L}^{(\mathrm{HAC})}_{\theta^\star} = \frac{1}{N}\sum_{i=1}^N Z_{\theta^\star,i}. \tag{39}$$

Combining Eqs. 37, 38, and 39, we obtain, with probability at least $1-\eta$,

$$\left\|\widehat{\theta}_{\mathrm{HAC}} - \theta^\star\right\|_2 \le \frac{1}{\mu N}\left[\sqrt{2V_N \log(2/\eta)} + \frac{2M_N \log(2/\eta)}{3}\right]. \tag{40}$$

This completes the proof.

B.5 PROOF OF HAC-R

For sample i, let $\mathcal{F}_i^M \subseteq \mathcal{F}_i^H$ denote the information available before and after human review, respectively. We assume

$$\begin{aligned} \Pr(\delta_i^H(n_h) = 1 \mid \mathcal{F}_i^M) &= \pi_{n_h}(\hat{s}_i^M) > 0, \\ \Pr(\delta_i^E(n_e) = 1 \mid \mathcal{F}_i^H, \delta_i^H(n_h) = 1) &= \pi_{n_e}(\hat{s}_i^E) > 0. \end{aligned} \tag{41}$$

The hard resolution decision $\alpha_i \in \{0, 1\}$ is $\mathcal{F}_i^H$-measurable and is chosen before observing the expert label. It is fixed with respect to the downstream parameter θ. Define the resolved loss by

$$\ell_{\theta,i}^R = (1 - \alpha_i)\ell_{\theta,i}^M + \alpha_i \ell_{\theta,i}^H, \tag{42}$$

The HAC-R estimator is

$$\widehat{\ell}_{\theta,i}^{\text{HAC-R}} = \ell_{\theta,i}^M + \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)}\big(\ell_{\theta,i}^R - \ell_{\theta,i}^M\big) + \frac{\delta_i^H(n_h)\delta_i^E(n_e)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}\big(\ell_{\theta,i} - \ell_{\theta,i}^R\big). \tag{43}$$

Proposition 2 (Unbiasedness and exact routing variance) *For fixed data and machine annotations, taking expectations over the scoring, review, and routing randomness gives*

$$\mathbb{E}\big[\widehat{\ell}_{\theta,i}^{\text{HAC-R}}\big] = \ell_{\theta,i}. \tag{44}$$

Hence, for $\mathcal{L}_\theta^{\text{HAC-R}} = N^{-1}\sum_i \widehat{\ell}_{\theta,i}^{\text{HAC-R}}$ *and* $\mathcal{L}_\theta = N^{-1}\sum_i \ell_{\theta,i}$, $\mathbb{E}[\mathcal{L}_\theta^{\text{HAC-R}}] = \mathcal{L}_\theta$. *Moreover,*

$$\begin{aligned} \operatorname{Var}\big(\widehat{\ell}_{\theta,i}^{\text{HAC-R}}\big) = \mathbb{E}\left[\left(\frac{1}{\pi_{n_h}(\hat{s}_i^M)} - 1\right)\big(\ell_{\theta,i} - \ell_{\theta,i}^M\big)^2\right] \\ + \mathbb{E}\left[\frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)^2}\left(\frac{1}{\pi_{n_e}(\hat{s}_i^E)} - 1\right)\big(\ell_{\theta,i} - \ell_{\theta,i}^R\big)^2\right]. \end{aligned} \tag{45}$$

Proof 3 *Taking expectation over expert acquisition first, the human and resolved corrections telescope. Taking expectation over human acquisition then gives*

$$\mathbb{E}\big[\widehat{\ell}_{\theta,i}^{\text{HAC-R}}\big] = \ell_{\theta,i}^M + \mathbb{E}\left[\frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)}\big(\ell_{\theta,i} - \ell_{\theta,i}^M\big)\right] = \ell_{\theta,i}.$$

For the variance, write

$$\begin{aligned} \widehat{\ell}_{\theta,i}^{\text{HAC-R}} - \ell_{\theta,i} = \left(\frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)} - 1\right)\big(\ell_{\theta,i} - \ell_{\theta,i}^M\big) \\ + \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)}\left(\frac{\delta_i^E(n_e)}{\pi_{n_e}(\hat{s}_i^E)} - 1\right)\big(\ell_{\theta,i} - \ell_{\theta,i}^R\big). \end{aligned} \tag{46}$$

The second term has conditional mean zero given the information available after human review, whereas the first is already determined at that point. Their cross moment is therefore zero. The conditional Bernoulli second moments are $1/\pi_{n_h}(\hat{s}_i^M) - 1$ *for the first stage and* $1/\pi_{n_e}(\hat{s}_i^E) - 1$ *for the expert stage. Taking expectations gives Eq. 45. Under the conditional independence assumptions of Proposition 1, cross-example moments also vanish. Summing these variances and dividing by* N^2 *yields Eq. 8. HAC follows by setting* $\alpha_i = 1$.

The corresponding HAC estimator replaces $\ell_{\theta,i}^R$ with $\ell_{\theta,i}^H$. Therefore, Proposition 2 immediately yields the following comparison.

Parameter error. Let $\theta^\star$ minimize the expert-supervised empirical objective and define

$$\Delta_i^S = \nabla\ell_{\theta^\star,i}^S - \nabla\ell_{\theta^\star,i}, \qquad S \in \{M, H, R\}. \tag{47}$$

Set

$$V_N^R = \sum_{i=1}^N \left[\frac{1 - \pi_{n_h}(\hat{s}_i^M)}{\pi_{n_h}(\hat{s}_i^M)}\|\Delta_i^M\|_2^2 + \frac{1 - \pi_{n_e}(\hat{s}_i^E)}{\pi_{n_h}(\hat{s}_i^M)\pi_{n_e}(\hat{s}_i^E)}\|\Delta_i^R\|_2^2\right], \tag{48}$$

and let M_N^R uniformly bound the norm of the corresponding centered single-sample gradient increment in equation 46.

Proof 4 *Replace the human residual gradient in the proof of the HAC parameter bound by the predictable resolved residual Δ_i^R, and apply the same vector Bernstein and strong-convexity argument.*

In particular, if $\|\Delta_i^R\|_2 \le \|\Delta_i^H\|_2$ for every i and $M_N^R \le M_N^H$, then the HAC-R parameter-gap bound is no larger than the corresponding HAC bound.

B.6 Corollary 1

How HAC(-R) reduces variance Equation 8 shows that Resolve and Escalate act on the same residual. Resolve determines $\ell_{\theta,i} - \ell_{\theta,i}^R$, while expert routing determines how strongly that residual contributes to variance through $(1 - \pi_{n_e}(\hat{s}_i^E))/\pi_{n_e}(\hat{s}_i^E)$. A resolved annotation has no residual contribution when it is always escalated, whereas its quality becomes increasingly important when expert supervision is sparse. This relationship can be stated by comparing HAC-R with HAC under the same human and expert routing propensities.

Corollary 1 (Effect of rollback) *Under the conditions of Proposition 1, suppose HAC and HAC-R use the same $\pi_{n_h}(\hat{s}_i^M)$ and conditional $\pi_{n_e}(\hat{s}_i^E)$. Their variance difference is*

$$\begin{aligned} &\mathrm{Var}\left(\mathcal{L}_\theta^{\text{HAC-R}}\right) - \mathrm{Var}\left(\mathcal{L}_\theta^{\text{HAC}}\right) \\ &= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N} \frac{\delta_i^H(n_h)}{\pi_{n_h}(\hat{s}_i^M)^2} \frac{1-\pi_{n_e}(\hat{s}_i^E)}{\pi_{n_e}(\hat{s}_i^E)} \left\{\left(\ell_{\theta,i} - \ell_{\theta,i}^R\right)^2 - \left(\ell_{\theta,i} - \ell_{\theta,i}^H\right)^2\right\}\right]. \end{aligned} \tag{49}$$

Hence, rollback reduces variance when the machine annotation retained by Resolve is closer in downstream loss to the expert annotation than the rejected human annotation, after weighting by the expert routing propensity.

The corollary also clarifies why rollback and escalation should not be treated as independent operations. Rollback changes the intermediate annotation that remains when expert supervision is unavailable, while escalation determines which of these intermediate decisions are ultimately corrected. Their joint effect is governed by the post-resolution residual in Eq. 8.

C Extension to Multiple Imperfect Human Annotators

Although we focus on a single imperfect human annotator in the main paper, HAC-R naturally extends to multiple sequential human-review stages. The key idea is to maintain a *resolved intermediate state* rather than irrevocably replacing the current annotation whenever new human feedback is acquired.

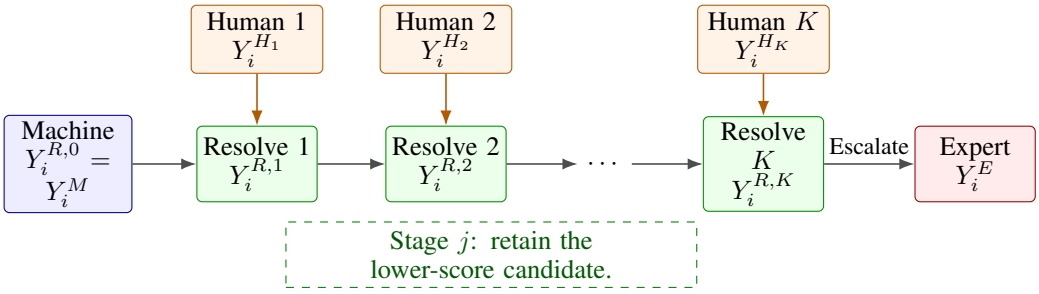


Figure 6: Multi-human HAC-R. Each new human annotation is resolved against the current state before optional expert escalation.

Let $Y_i^{R,0} = Y_i^M$ denote the initial machine annotation and $\ell_{\theta,i}^{R,0} = \ell_{\theta,i}^M$ its corresponding loss. Suppose that up to K imperfect human annotations $Y_i^{H_1}, \ldots, Y_i^{H_K}$ may be acquired sequentially. For each potential human source, define $\eta_{H_j} = N^{-1}\sum_i \mathbb{1}\{Y_i^{H_j} \neq Y_i\}$ and let c_{H_j} denote its

annotation cost. We assume that review cost increases and average noise decreases with expertise,

$$\begin{aligned} &c_{H_1} \le c_{H_2} \le \cdots \le c_{H_K} \le c_e, \\ &\eta_{H_1} \ge \eta_{H_2} \ge \cdots \ge \eta_{H_K} \ge \eta_E, \qquad \eta_E = 0. \end{aligned} \tag{50}$$

This average ordering does not require the correctness indicators to be nested for every example; a more expert reviewer may still be wrong when an earlier source is correct. The ordering motivates cost-aware escalation but is not required for design unbiasedness, which depends only on predictable resolution, positive acquisition propensities, and the information ordering of the stages. After acquiring the j-th human annotation, we compare its reliability with that of the current resolved annotation and define an instance-wise resolution decision

$$\alpha_{ij} = \mathbb{1}\left\{ s_i^{H_j} < s_i^{R,j-1} \right\}, \tag{51}$$

where a smaller reliability score indicates a more reliable annotation. The resolved loss is then updated recursively as

$$\ell_{\theta,i}^{R,j} = \alpha_{ij} \ell_{\theta,i}^{H_j} + (1 - \alpha_{ij}) \ell_{\theta,i}^{R,j-1}, \qquad j = 1, \ldots, K. \tag{52}$$

Hence, newly acquired human feedback replaces the current supervision only when it is assessed as more reliable; otherwise, the previous resolved state is retained.

Let $\delta_{ij}^H \in \{0, 1\}$ denote the acquisition decision at human stage j, with conditional propensity $p_{ij} \in (0, 1]$ given the information available before that acquisition. Define the cumulative acquisition indicator and propensity as

$$A_{ij} = \prod_{t=1}^{j} \delta_{it}^H, \qquad P_{ij} = \prod_{t=1}^{j} p_{it}, \tag{53}$$

with $A_{i0} = P_{i0} = 1$. If expert adjudication is subsequently acquired with conditional propensity $q_i \in (0, 1]$, the multi-human extension of HAC-R can be written as

$$\widehat{\ell}_{\theta,i}^{K\text{-HAC-R}} = \ell_{\theta,i}^{R,0} + \sum_{j=1}^{K} \frac{A_{ij}}{P_{ij}} \left(\ell_{\theta,i}^{R,j} - \ell_{\theta,i}^{R,j-1} \right) + \frac{A_{iK} \delta_i^E}{P_{iK} q_i} \left(\ell_{\theta,i} - \ell_{\theta,i}^{R,K} \right). \tag{54}$$

The same telescoping argument underlying HAC-R immediately extends to this setting. Under sequential positivity and correct conditional propensities,

$$\mathbb{E}\left[\frac{A_{ij}}{P_{ij}} \right] = 1, \qquad \mathbb{E}\left[\frac{A_{iK} \delta_i^E}{P_{iK} q_i} \right] = 1. \tag{55}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\widehat{\ell}_{\theta,i}^{K\text{-HAC-R}} \right] &= \ell_{\theta,i}^{R,0} + \sum_{j=1}^{K} \left(\ell_{\theta,i}^{R,j} - \ell_{\theta,i}^{R,j-1} \right) + \left(\ell_{\theta,i} - \ell_{\theta,i}^{R,K} \right) \\ &= \ell_{\theta,i}. \end{aligned} \tag{56}$$

Thus, the fully expert-supervised target is invariant to the number and form of intermediate resolution stages. In particular, the single-human HAC-R formulation studied in the main paper is recovered by setting $K = 1$. This extension also permits heterogeneous and correlated human annotators: such dependencies may affect the quality of the resolved state and the variance of the estimator, but do not alter its design-based unbiasedness as long as each acquisition propensity is defined conditionally on the information available at that stage.

Proposition 3 (Variance of multi-stage HAC-R) *Consider the K-stage HAC-R estimator in Eq. 54. Let*

$$P_{ij} = \prod_{t=1}^{j} p_{it}, \qquad P_{i0} = 1,$$

and suppose that the acquisition decisions are randomized sequentially with the stated conditional propensities. Conditional on the potential annotations, pre-computed reliability scores, and the resulting resolved states, the per-example sampling variance is

$$\mathrm{Var}\left(\widehat{\ell}_{\theta,i}^{\text{K-HAC-R}}\right) = \sum_{j=1}^{K} \frac{1-p_{ij}}{P_{ij}} \left(\ell_{\theta,i} - \ell_{\theta,i}^{R,j-1}\right)^2 + \frac{1-q_i}{P_{iK}q_i} \left(\ell_{\theta,i} - \ell_{\theta,i}^{R,K}\right)^2. \tag{57}$$

If the acquisition randomizations are independent across examples, then

$$\mathrm{Var}\left(\mathcal{L}_{\theta}^{\text{K-HAC-R}}\right) = \frac{1}{N^2} \sum_{i=1}^{N} \left[\sum_{j=1}^{K} \frac{1-p_{ij}}{P_{ij}} \left(\ell_{\theta,i} - \ell_{\theta,i}^{R,j-1}\right)^2 + \frac{1-q_i}{P_{iK}q_i} \left(\ell_{\theta,i} - \ell_{\theta,i}^{R,K}\right)^2 \right]. \tag{58}$$

Proof 5 *For notational simplicity, define*

$$A_{ij} = \prod_{t=1}^{j} \delta_{it}^{H}, \qquad A_{i0} = 1,$$

and let

$$r_{ij} = \ell_{\theta,i} - \ell_{\theta,i}^{R,j}.$$

The estimation error can be rewritten as the sum of sequential martingale differences

$$\begin{aligned} \widehat{\ell}_{\theta,i}^{\text{K-HAC-R}} - \ell_{\theta,i} = &\sum_{j=1}^{K} \frac{A_{i,j-1}}{P_{i,j-1}} \left(\frac{\delta_{ij}^{H}}{p_{ij}} - 1\right) r_{i,j-1} \\ &+ \frac{A_{iK}}{P_{iK}} \left(\frac{\delta_{i}^{E}}{q_i} - 1\right) r_{iK}. \end{aligned} \tag{59}$$

Each term has conditional mean zero given the information available before its corresponding acquisition decision. Hence, the cross terms vanish by sequential orthogonality.

For human stage j,

$$\mathbb{E}\left[\left(\frac{A_{i,j-1}}{P_{i,j-1}}\right)^2 \left(\frac{\delta_{ij}^{H}}{p_{ij}} - 1\right)^2\right] = \frac{1-p_{ij}}{P_{ij}},$$

where we use $P_{ij} = P_{i,j-1}p_{ij}$. Similarly, for expert adjudication,

$$\mathbb{E}\left[\left(\frac{A_{iK}}{P_{iK}}\right)^2 \left(\frac{\delta_{i}^{E}}{q_i} - 1\right)^2\right] = \frac{1-q_i}{P_{iK}q_i}.$$

Summing the variances of the orthogonal increments yields Eq. 57. The dataset-level expression follows by independence of the acquisition randomizations across examples.

D Supplementary Experiment Details

D.1 Analysis of Budget Allocation.

Figure 7 shows that the benefit of additional human review depends on annotation reliability. At $\eta_H = 0.40$, increasing the human-review budget causes HAC accuracy to decline on CIFAR-10, mini-ImageNet, and Fashion-MNIST, whereas HAC-R maintains relatively stable accuracy. The same pattern appears on CIFAR-10N, where HAC-R improves with greater review coverage while HAC deteriorates. These results highlight HAC-R's ability to prevent additional noisy feedback from degrading annotation quality. Increasing the expert budget improves both methods across all evaluated datasets, although their relative performance varies. HAC remains stronger on Clothing1M and in several low-noise settings, indicating that HAC-R's main benefit lies in handling unreliable human feedback rather than uniformly improving accuracy across all budget allocations.

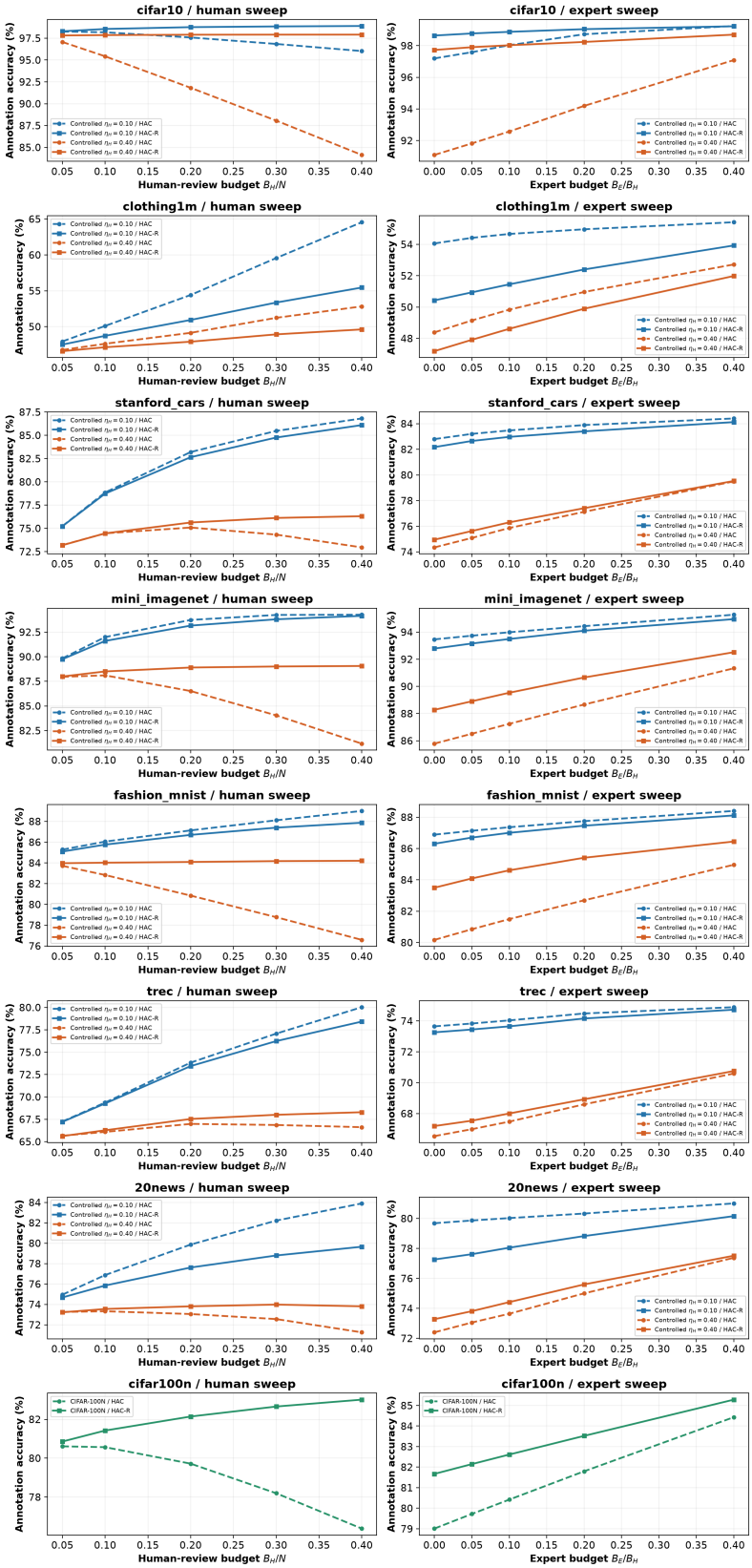


Figure 7: Sensitivity to annotation budget allocation under a fixed total annotation cost.

D.2 Net Value of Rollback

Rollback has two opposing effects: retaining a correct machine label can prevent a harmful human replacement, but retaining an incorrect machine label can reject a useful human correction. Before expert adjudication, the net accuracy change relative to Always-H is the number of the former minus the latter, divided by the full routing-pool size. Using this common denominator permits comparison across review budgets; it differs from the reviewed-set rates in Figure 4.

Table 9 compares HAC-R with HAC on Cars and TREC at $\eta_H = 0.40$ and on CIFAR-100N with real crowd feedback. Under both threshold and exponential routing, prevented harmful replacements outnumber rejected useful corrections. For example, threshold routing on CIFAR-100N prevents 827 harmful replacements while rejecting 154 useful corrections. Rollback therefore provides a net annotation benefit in these settings rather than merely reducing human-label acceptance. Final annotation accuracy increases in all six comparisons, by 0.37–2.03 percentage points; the gain is larger under exponential routing for each of the three datasets.

The downstream benefit is less consistent. Cars improves by 0.29 and 0.77 points under threshold and exponential routing, respectively, but both paired 95% intervals include zero. Under exponential routing, CIFAR-100N improves by 1.17 points (95% interval: $[0.14, 2.21]$), whereas TREC declines by 1.80 points ($[-2.64, -0.96]$), despite improved annotation accuracy in both cases. Neither dataset has a downstream interval excluding zero under threshold routing. These results separate the net value of source selection from its effect on learning: preventing more label errors than rollback introduces does not by itself guarantee improved clean-test accuracy.

Table 9: Net rollback effects under threshold and exponential routing. Cars and TREC use $\eta_H = 0.40$; CIFAR-100N uses real crowd labels. P counts prevented harmful replacements and L counts rejected useful corrections. Accuracy changes are HAC-R minus HAC (pp); brackets give paired 95% intervals for downstream changes.

Dataset	Routing	P	L	Final annotation Δ acc.	Downstream Δ acc. [paired 95% interval]
Stanford Cars	Thres	60	25	$+0.37$	$+0.29$ $[-0.26, +0.84]$
Stanford Cars	Exp	171.4	30.6	$+1.57$	$+0.77$ $[-0.14, +1.67]$
TREC	Thres	47	11	$+0.65$	-0.68 $[-1.67, +0.31]$
TREC	Exp	86.2	26.0	$+1.12$	$\mathbf{-1.80}$ $[\mathbf{-2.64}, \mathbf{-0.96}]$
CIFAR-100N	Thres	827	154	$+1.16$	-0.07 $[-0.42, +0.29]$
CIFAR-100N	Exp	1229.8	119.4	$+2.03$	$\mathbf{+1.17}$ $[\mathbf{+0.14}, \mathbf{+2.21}]$

D.3 Rollback Retention and Estimator Variance

Figure 8 examines rollback retention on Irony using a shared fixed model, paired human-review draws, and expert-query probabilities based on the Min score. Retention ranges from always accepting human feedback (0%) to the original Resolve rule (100%). At $\eta_H = 0.10$, increasing retention raises the loss-estimator variance and reduces annotation accuracy both before and after expert queries. At $\eta_H = 0.40$, rollback improves annotation accuracy, whereas variance initially decreases slightly and then increases. Thus, improved label accuracy does not by itself imply reduced loss-estimator variance at a fixed model.

D.4 Joint-Error Analysis

Figure 9 shows that active acquisition substantially changes the composition of the reviewed population. Under low human noise, reviewed examples contain a large proportion of machine-wrong/human-correct cases (28.2–53.5%), indicating that ordinary-human review provides useful corrections for many acquired machine errors. This explains why unconditional human acceptance is competitive and why rollback may reject beneficial corrections in this regime.

Under high human noise, however, the both-wrong rate in the reviewed subset rises to 27.3–51.2%, while harmful machine-correct/human-wrong cases also become more frequent. The former cannot be addressed by selecting between the two available candidates and therefore motivates expert escalation; the latter motivates reversible resolution. These results show that the overall human-noise

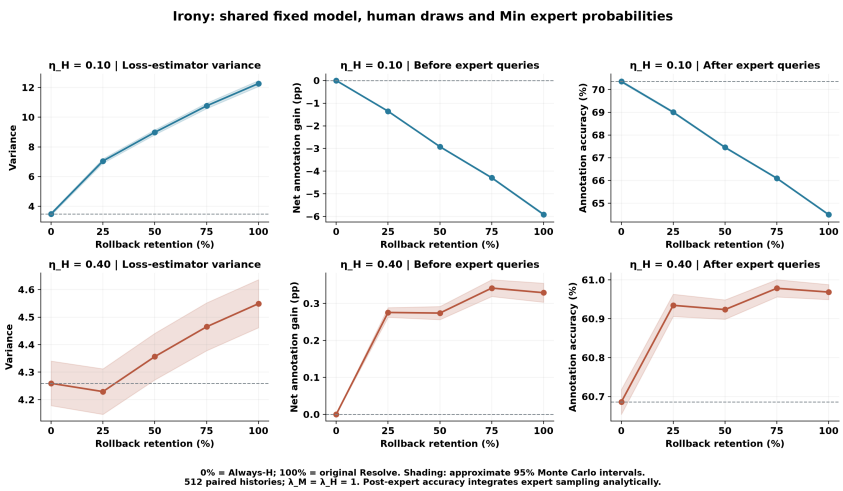


Figure 8: Rollback-retention diagnostic on Irony with $\lambda_M = \lambda_H = 1$. Shading denotes approximate 95% Monte Carlo intervals over 512 paired histories. Post-expert annotation accuracy integrates expert sampling analytically; dashed lines mark no rollback.

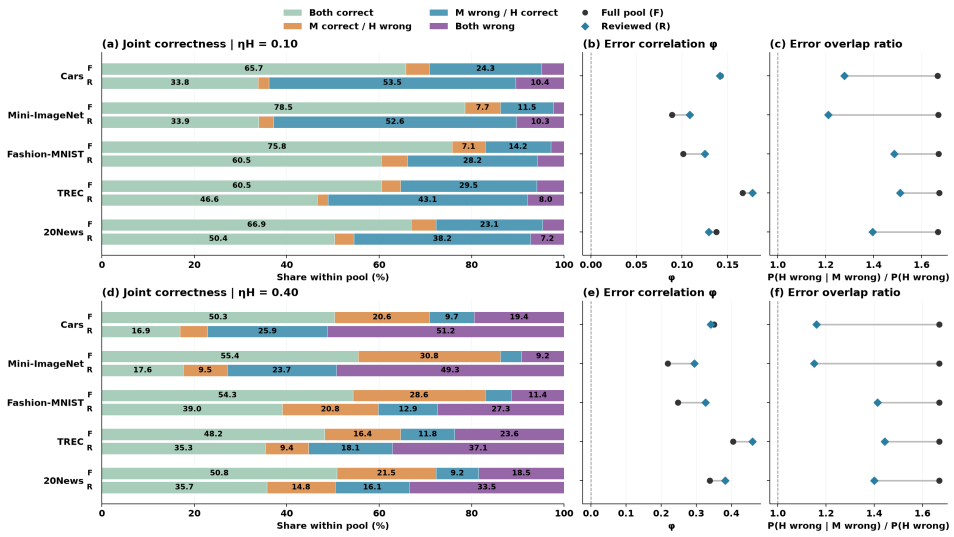


Figure 9: Joint machine–human correctness in the full (F) and reviewed (R) pools under controlled human noise.

rate alone does not characterize the supervision encountered after selective acquisition: Resolve and Escalate address distinct failure modes induced by the joint machine–human error structure.

D.5 Sensitivity Analysis of τ

Sensitivity to the SoftMin temperature. Figure 10 shows that HAC and HAC-R are largely insensitive to τ under fixed query budgets. Across all datasets and noise settings, the variation in annotation accuracy remains below 0.15 percentage points. Moderate and large temperatures produce nearly identical results, whereas very small values occasionally reduce accuracy, particularly on TREC and CIFAR-100N. Intermediate temperatures can yield small gains on Cars and Fashion-MNIST, but these improvements are not consistent across datasets. We therefore use $\tau = 1$ as a stable default.

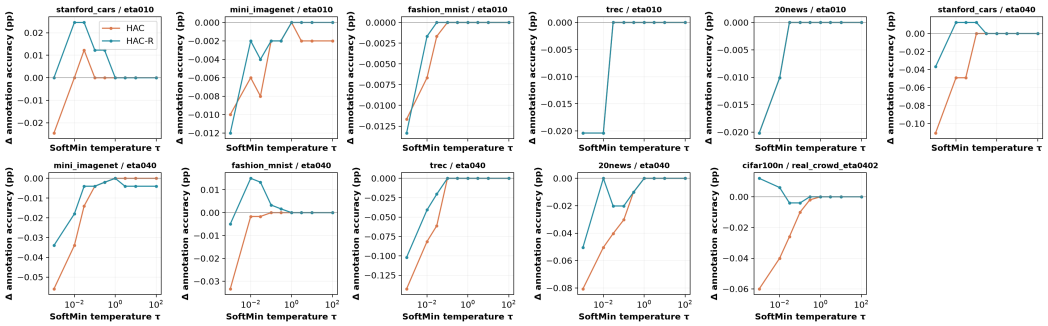


Figure 10: Sensitivity of HAC and HAC-R annotation accuracy to the SoftMin temperature τ under fixed query budgets. Each panel reports the change in annotation accuracy relative to the corresponding reference configuration, in percentage points, across controlled human-noise settings $\eta_H \in \{0.10, 0.40\}$ and real crowd feedback on CIFAR-100N ($\eta_H = 0.402$). Performance is generally stable once τ is sufficiently large, while very small temperatures can produce dataset-dependent changes.

E PROMPTS

Here, we present the prompt templates used by the annotator and criticizer. We show one representative template for each module rather than repeating dataset-specific instances. In the following examples, red denotes input placeholders, while blue denotes the expected output format. Dataset-specific class names and task hints are substituted into the corresponding placeholders at inference time.

Illustrations of Inputs and Outputs

- {image}: the input image.
- {text}: the input text.
- {sample_text}: the passage to be classified.
- {label_list_with_index}: the complete indexed label set.
- {options}: the candidate labels rendered as lettered options.
- {task_hint}: a short dataset-specific task description.
- {predicted_label}: the annotation produced by the base annotator.
- {label_index}: the index of the candidate label being evaluated.
- {slot_a} and {slot_b}: two candidate labels displayed in randomized and source-anonymous order.
- [label_index]: the selected class index.
- [option]: the selected option letter.
- [error_probability]: the estimated probability that a label is wrong.

E.1 ANNOTATOR PROMPTS

[Text Classification] Annotation Prompt — Direct

Prompt
You are a careful text classifier. Read the passage and pick the single most likely category from the options below.
Passage:
{sample_text}

Options:
{options}
Answer:
Output
[option]

[Image Classification] Annotation Prompt — Direct

Prompt
{image}
Determine the label of the image classification task. The list of labels is: {label_list_with_index}. The required output format is: [label_index]. For example, if the label is the first class, you should output [0]. Do not return other texts.
Output
[label_index]

[Image Classification] Annotation Prompt — CoT

Prompt
{image}
Determine the label of the image classification task. The list of labels is: {label_list_with_index}. The required output format is: [label_index]. Think step-by-step and provide your reasoning. Example of required output format is: [reasoning][label_index]. The first brackets contain your step-by-step reasoning, and the second brackets contain the label index such as [0] for {first_label}. Do not return other texts.
Output
[CoT_A][label_index]

For TREC, the options are ABBR, ENTY, DESC, HUM, LOC, and NUM. For 20 Newsgroups, the options are the 20 canonical newsgroup names. The text annotator uses direct prompting without chain-of-thought generation.

E.2 Criticizer Prompt

[Classification] Criticizer Prompt

Prompt
{input}
Your task is to decide whether the predicted label is wrong. The list of labels is: {label_list_with_index}. The predicted label is: {predicted_label}. The required output is either Yes or No, where Yes means that the prediction is incorrect and No means otherwise. Do not return other texts.
Options:
A. Yes
B. No
Answer:
Output
[option]

The normalized probability assigned to option A is used as the criticizer score. The criticizer only estimates whether the annotation is incorrect and does not generate a replacement label.

F Additional Discussions and Related Works

F.1 Additional Related Works.

Label-Error Detection. Label-Error detection aims to identify mislabeled examples in the data (Bahri et al., 2020; Zhu et al., 2022; Zhang et al., 2025b; Afriat et al., 2026). Deep KNN (Bahri et al., 2020) employs k-NN entropy on embedding for noise identification,and SimiFeat (Zhu et al., 2022) designs a local-neighboorhood voting for noise detection. LEMoN (Zhang et al., 2025b) designs a label-noise detector based on the local-neighborhood across multiple data modalities. Inspired by LEMoN, we proposes a training-free label-error detection score for detecting labelling errors from both of human-annotators and machine-annotators, mitigating bias against the machine-annotator labels (Chen et al., 2024a; Koyama et al., 2026).