ACT as Human: Multimodal Large Language Model Data Annotation with Critical Thinking

Lequan Lin, Dai Shi, Andi Han, Feng Chen, Qiuzheng Chen,
Jiawen Li, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, & Junbin Gao

Accepted at NeurIPS 2025

26 Sep 2025

*The work is done during Lequan (Lena) Lin’s internship at TnS-Algo-Live Safety, TikTok.

*All images in this presentation are from public sources. No internal user data is involved.

Background

Industrial Background - Content Moderation

Content moderation is the process of reviewing, filtering, and removing content that violates Community Guidelines, such as nudity, violence, or misinformation.

Background

Industrial Background - Content Moderation

Content moderation is the process of reviewing, filtering, and removing content that violates Community Guidelines, such as nudity, violence, or misinformation.

Data Challenges - Limited Human Resources & Ethical Issues

- AI content moderation relies on large-scale labeled data for supervised training, yet human annotation resources remain limited and costly.
- Sensitive content such as nudity or graphic violence can harm the mental and physical well-being of human annotators.

Background

Industrial Background - Content Moderation

Content moderation is the process of reviewing, filtering, and removing content that violates Community Guidelines, such as nudity, violence, or misinformation.

Data Challenges - Limited Human Resources & Ethical Issues

- AI content moderation relies on large-scale labeled data for supervised training, yet human annotation resources remain limited and costly.
- Sensitive content such as nudity or graphic violence can harm the mental and physical well-being of human annotators.

Technical Challenges - LLM Data Annotation

- While large language models (LLMs) have recently emerged as an alternative for data annotation, the labels they generate often lack the accuracy required for training reliable downstream models.

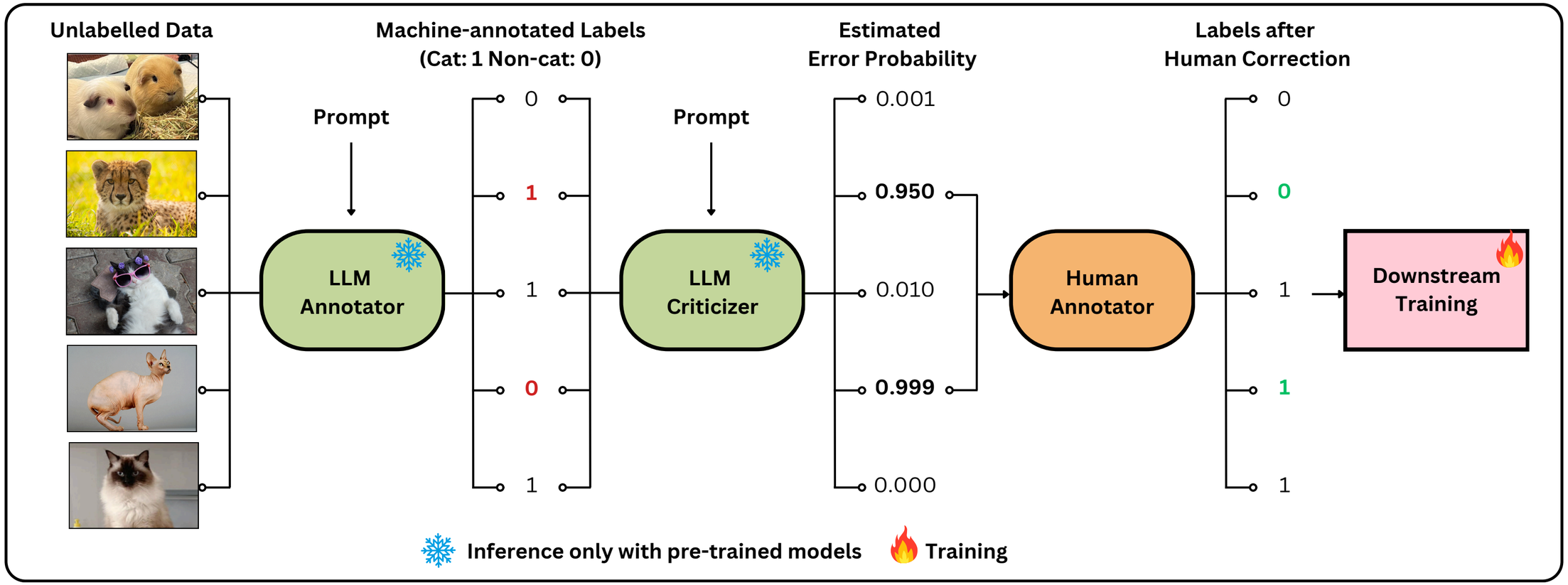
Our Research Question

“How can we incorporate LLMs into the data pipeline to efficiently reduce human cost without compromising downstream training performance?”

The Proposed Method

In this paper, we propose the **Annotation with Critical Thinking (ACT)** data pipeline.

With this approach, an LLM handles the majority of annotation workloads, while a limited human budget is strategically allocated to review samples flagged as potentially erroneous by another LLM-based error detector (a.k.a, the criticizer).



Mathematical Setups

Annotation

Suppose we have a regular human annotator $f^{(h)}$, LLM annotator $f^{(m)}$, and LLM criticizer g.

We then denote data and its ground truth label as (x_i, y_i) for $i = 1, 2, \ldots, N$.

The labels provided by human and LLM are denoted as $\hat{y}_i^{(h)}$, and $\hat{y}_i^{(m)}$, respectively.

In this work, we assume that y_i is approximated by $\hat{y}_i^{(h)}$.

Criticism & Human Review

Let $\epsilon_i = P(y_i \neq \hat{y}_i^{(m)} | x_i)$ denote the true error probability, which is then estimated by $\hat{\epsilon}_i = g(x_i, \hat{y}_i^{(m)})$.

Given a human budget $B \leq N$, we define a budget-aware sampling function $\delta(B)$ where $\delta_i(B) \sim \text{Ber}(\pi_B(\hat{\epsilon}_i))$.

$\pi_B(\cdot)$ is a transformation that adjusts the error probability $\hat{\epsilon}_i$ based on the budget B, ensuring that $\sum_{i=1}^{N} \delta_i(B) \leq B$.

The samples with $\delta_i(B) = 1$ will be reviewed by human annotators.

Problems to be Solved...

ACT may appear to be a straightforward application of “LLM-as-a-Judge” and “Human-in-the-Loop,”

but...

- **[Pipeline Implementation]** How should we select the LLM annotator and LLM criticizer? Which models are most appropriate for specific tasks? How to design the prompts?
- **[Downstream Training]** Since the ACT data pipeline cannot guarantee 100% correct annotation (i.e., the criticizer may fail to identify some errors), how can we ensure reliable downstream performance?

Problems to be Solved...

ACT may appear to be a straightforward application of "LLM-as-a-Judge" and "Human-in-the-Loop,"

but...

- **[Pipeline Implementation]** How should we select the LLM annotator and LLM criticizer? Which models are most appropriate for specific tasks? How to design the prompts?
- **[Downstream Training]** Since the ACT data pipeline cannot guarantee 100% correct annotation (i.e., the criticizer may fail to identify some errors), how can we ensure reliable downstream performance?

We aim to address these questions by:

- **[Pipeline Implementation]** Systematically exploring each component of ACT, including the LLM annotator, the criticizer, and the associated prompt strategies;
- **[Downstream Training]** Theoretically showing how downstream training can be secured with a simple modification of the loss function.

Experimental Setups

Datasets

We consider a comprehensive variety of annotation tasks, including **image classification**: CIFAR10, Fashion-MNIST (Fashion), and Stanford Cars (Cars); **text classification**: TweetEval Emotion and Irony; and **vision question-answering**: VQA-RAD.

Multimodal LLMs

In this paper, we mainly focus on multimodal LLMs (MLLMs) because they can handle diverse data types.

We conduct experiments using models from six prominent MLLM families: ChatGPT 4o 2024-08-06 (GPT4o), Gemini-1.5-Pro-002 (Gemini1.5P), Claude 3.5 Sonnet (Claude3.5S), LLaVA OneVision 72B (LLaVA-OV), Qwen 2.5 VL 72B (Qwen2.5VL), and InternVL 2.5 78B (InternVL2.5).

Downstream Models

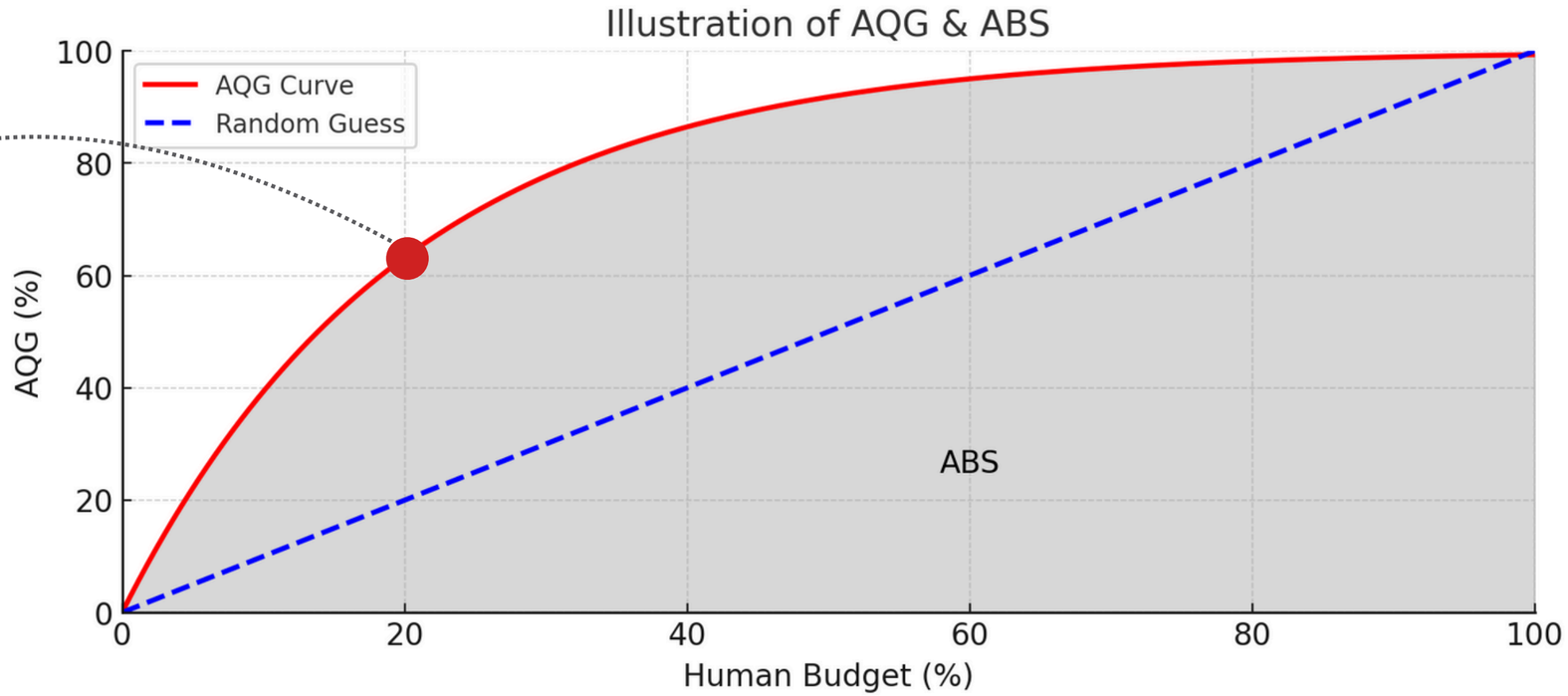
For image classification tasks, we use ResNet18, initialized with default weights pretrained on ImageNet-1k. For text classification, we utilize the RoBERTa-base model. For the VQA task, we adopt the BLIP-VQA model.

What makes ACT work better?

“Work better” means less errors with lower human budget. We design two metrics to evaluate the performance of ACT data pipeline:

- **Annotation Quality Gain (AQG)** measures the annotation quality (e.g., classification accuracy) improved by ACT from naive LLM annotation given a fixed human budget B;
- **Area under Budget Sensitivity (ABS)** measures the overall efficiency of human budget, for which a higher value implies better budget utilization.
- **We use ABS for our explorations because we do not assume a specific human budget.**

When 20% of the data are reviewed by human, ~60% of the LLM annotation errors are fixed.



What makes ACT work better?

Via comprehensive experiments (ABS as the key evaluation metric):

6 Annotation Tasks x 6 MLLMs x 2 Annotation Prompts x 7 Criticizer Prompts

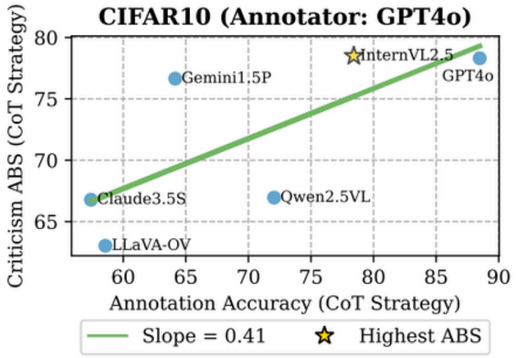
We have gained 7 insights of ACT implementation:

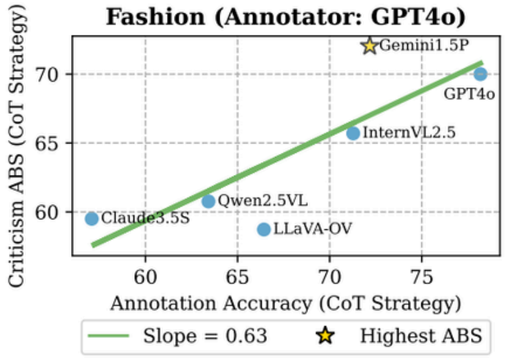
- Insight 1: **Models from ChatGPT family are generally promising annotators.**
- **Insight 2**: CoT is not consistently helpful with annotation.
- Insight 3: **Cross-criticism outperforms self-criticism.**
- **Insight 4**: Black-box models are better criticizers with black-box strategies.
- **Insight 5**: CoT is more helpful with criticism than annotation.
- **Insight 6**: White-box strategies can occasionally enhance criticism performance.
- Insight 7: **White-box strategies do not consistently outperform black-box strategies.**

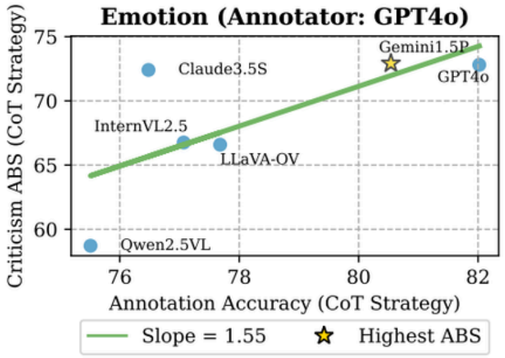
From Insights to Practice

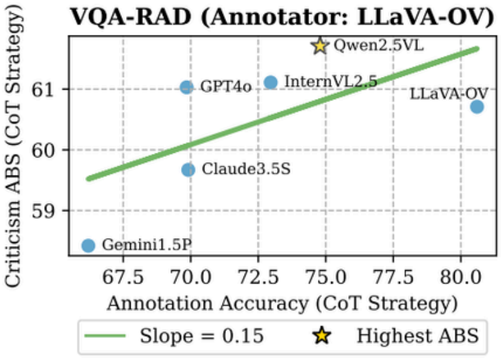
For a given dataset, how do we choose the annotator and criticizer to maximize pipeline efficiency?

- **Default**:
  - [CoT Prompt Strategy] GPT annotation + GPT criticism is a generally good option.
- **“Pro Max”**:
  - Prepare the candidate MLLMs and a small set of human-annotated data for testing;
  - Test the annotation accuracy of candidate MLLMs;
  - Use the MLLM with the best accuracy as the annotator, and the second best as the criticizer.









ACT as Human in Downstream Training

Up to this point, we have only considered data annotation, focusing on how to correct most LLM annotation errors with minimal human effort. However, errors can rarely be fully eliminated.

So, how can we ensure reliable downstream training performance when ACT labels are not entirely accurate?

Our goal is **“ACT as Human”**, which means downstream models trained with ACT data should match the performance of models trained with fully human-annotated data.

ACT as Human in Downstream Training

Up to this point, we have only considered data annotation, focusing on how to correct most LLM annotation errors with minimal human effort. However, errors can rarely be fully eliminated.

So, how can we ensure reliable downstream training performance when ACT labels are not entirely accurate?

Our goal is **"ACT as Human"**, which means downstream models trained with ACT data should match the performance of models trained with fully human-annotated data.

The Solution: Active M-Estimation

A potential solution to the downstream challenge is active M-estimation (Zrnic and Candès, ICML2024). In this work, we adapt the modified loss to ACT data and design the ACT loss as follows:

$$\mathcal{L}_{\theta}^{(ACT)} = \frac{1}{N}\sum_{i=1}^{N}\left(\ell_{\theta,i}^{(m)} + \left(\ell_{\theta,i} - \ell_{\theta,i}^{(m)}\right)\frac{\delta_i(B)}{\pi_B(\hat{\epsilon}_i)}\right)$$

N : dataset size; $\ell_{\theta,i}^{(m)}$: loss computed with LLM label; $\ell_{\theta,i}$: loss computed with human label;

$\pi_B(\hat{\epsilon}_i)$: transformed $\mathrm{P}(\text{error})$ to satisfy human budget B; $\delta_i(B)$: 0-1 indicator of human review, $\delta_i(B) \sim \mathbb{B}(\pi_B(\hat{\epsilon}_i))$

ACT as Human in Downstream Training

$$\mathcal{L}_\theta^{(ACT)} = \frac{1}{N}\sum_{i=1}^{N}\left(\ell_{\theta,i}^{(m)} + \left(\ell_{\theta,i} - \ell_{\theta,i}^{(m)}\right)\frac{\delta_i(B)}{\pi_B(\hat{\epsilon}_i)}\right)$$

Why the ACT loss works?

- **Statistical Properties of ACT Loss**
  - Assume that the loss is strongly convex (e.g. CE Loss).
  - **Unbiased:** $\mathbb{E}\left(\mathcal{L}_\theta^{(ACT)}\right) = \mathcal{L}_\theta$
  - **Variance:** $\mathrm{Var}\left(\mathcal{L}_\theta^{(ACT)}\right) = \frac{1}{N}\left(\mathrm{Var}\left(\ell_\theta\right) + \mathbb{E}\left[\left(\ell_\theta - \ell_\theta^{(m)}\right)^2\left(\frac{1}{\pi_B(\hat{\epsilon})} - 1\right)\right]\right)$
  - **Variance is lower for:**
    - More accurate LLM annotator;
    - OR More astute LLM criticizer.
    - This demonstrates the importance of our previous exploration.

ACT as Human in Downstream Training

$$\mathcal{L}_{\theta}^{(ACT)} = \frac{1}{N}\sum_{i=1}^{N}\left(\ell_{\theta,i}^{(m)} + \left(\ell_{\theta,i} - \ell_{\theta,i}^{(m)}\right)\frac{\delta_i(B)}{\pi_B(\hat{\epsilon}_i)}\right)$$

Why the ACT loss works?

- **Probabilistic Upper Bound of the Parameter Gap**
  - For an arbitrary $p \in (0,1)$, with probability at least $1-p$, we can bound:

$$\|\theta_*^{(\mathrm{ACT})} - \theta_*\| \leq \sqrt{\frac{8c_1\log(2/p)}{\mu^2 N}} \text{ where } c_1 = (1-q)C^2/q$$

  - The upper bound of the gap between parameters learned with ACT loss and ground truth loss is decided by q, the lower bound of the transformed errors of samples reviewed by human, which means $\pi_B(\hat{\epsilon}_i) \geq q$ for all samples with $\delta_i(B) = 1$.
  - The probabilistic parameter gap can be lower with larger q.
  - Thus, we propose “exponential weighting” and “thresholding” transformation rules , which map errors of selected samples close to 1 ($q \to 1$).

ACT as Human in Downstream Training

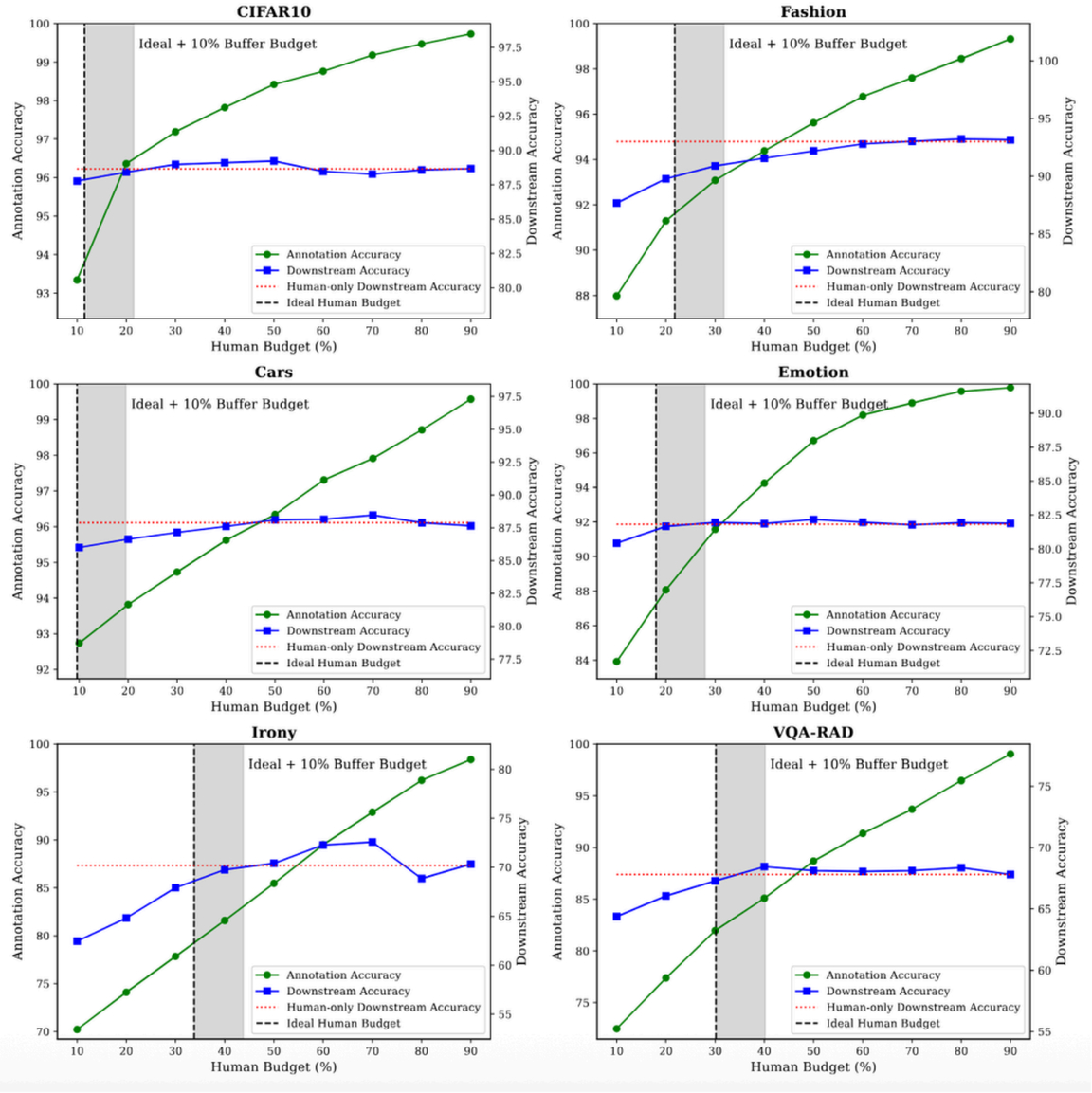
Key downstream experiment results (human budget is set as estimated annotator error proportion):

- Training solely on LLM-annotated data leads up to 10.15% lower accuracy compared to human-annotated data, whereas ACT reduces the performance gap to **less than 2%** for most datasets while **saving up to 90% human costs**.

- Exponential Weighting (exp.) and Thresholding (thre.) outperform Normalization (norm.), especially under limited human budgets. Note that norm. is the method proposed originally in active M-estimation (Zrnic and Candès, ICML2024).

Training Data - Loss	CIFAR10 (ResNet18)	Fashion (ResNet18)	Cars (ResNet18)	Emotion (RoBERTa)	Irony (RoBERTa)	VQA-RAD (BLIP-VQA)
Human only - Cross-entropy Loss	88.66 ± 0.97	93.01 ± 0.63	87.88 ± 0.36	81.82 ± 0.57	70.18 ± 3.23	67.81 ± 1.47
Machine only - Cross-entropy Loss	81.55 ± 1.93	82.86 ± 0.84	83.68 ± 0.17	78.96 ± 2.40	60.71 ± 5.43	61.03 ± 2.05
ACT data - Cross-entropy loss	85.59 ± 0.52	87.50 ± 0.86	85.88 ± 0.26	80.82 ± 1.08	67.83 ± 2.82	61.83 ± 3.27
ACT data - ACT norm. loss	64.70 ± 5.46	69.27 ± 7.25	11.54 ± 0.96	79.87 ± 0.88	65.66 ± 2.00	62.55 ± 3.01
ACT data - ACT exp. loss (Ours)	87.73 ± 0.36	89.73 ± 0.35	86.19 ± 0.14	81.44 ± 0.51	68.49 ± 3.20	67.73 ± 1.33
ACT data - ACT thre. loss (Ours)	87.95 ± 0.35	89.16 ± 0.89	86.00 ± 0.26	81.41 ± 0.64	68.21 ± 1.94	67.02 ± 1.32
Human-Machine performance gap (%)	7.11%	10.15%	4.20%	2.86%	9.47%	6.87%
Human-ACT performance gap (%)	0.71%	3.28%	1.69%	0.38%	1.69%	0.08%
ACT human budget (%)	11.52%	21.81%	9.56%	17.98%	33.79%	30.15%

Set Human Budget to Close the Gap



How to set the human budget to close the Human-ACT performance gap? Of course we don't want human budget = dataset size!

- When using the ideal budget (LLM annotator error proportion), a performance gap can be observed across all datasets. This is because the criticizer is not perfectly accurate, leading to some overlooked mislabeled data, which slightly degrades the final training performance.
- However, for most datasets, the performance gap can be closed with 10% buffer adding up to the ideal human budget.
- **So, we recommend first evaluating the annotator's accuracy with a small test set, and then adding a reasonable buffer to the ideal budget based on the observed error proportion.**

Discussions & Future Works

Discussions

Although our study is based on current MLLMs, and the efficiency of the proposed pipeline is constrained by their capabilities, our approach can be readily adapted to more advanced models in the future.

The insights gained from our explorations are likely to generalize and provide valuable guidance for practical applications and future research.

Some future directions

- More complex tasks: e.g., image captioning and open questions
- LLM self/cross-improvement: LLM downstream models

Thank You
Q&A