When Global Grades Fail Locally: Learning Local Supervision from Aggregate Feedback and Hindsight

Anonymous authors
Paper under double-blind review

Abstract

Agentic systems often receive one grade for an entire trajectory, even though learning requires feedback on individual steps. Copying the final grade to every step can therefore create wrong local labels: a failed trajectory may contain correct steps, and a successful trajectory may contain mistakes. We study how to recover local correctness without densely labeling every step. Our method combines aggregate feedback with a small number of local labels to train a local critic. Hindsight then allows the critic to reassess earlier claims using evidence that becomes available later in the trajectory. Our theory explains what aggregate feedback leaves unresolved and when hindsight provides additional information. It also shows how local correctness relates to errors in inherited labels. We evaluate the method in a controlled multimodal forecast-and-recount setting with independent simulator ground truth. On a frozen final set, the retrospective critic reduces local-correctness log loss by 32% relative to a matched static critic, from 0.5781 to 0.3943. A separate frozen validation set shows the same pattern. These results show that aggregate feedback, sparse local supervision, and hindsight can recover useful local supervision without dense step-level labels.

1 Introduction

Long-horizon models and agents are often judged by a final outcome (Cobbe et al., 2021; Zhou et al., 2024; Liu et al., 2024). Did the task succeed? Did the response receive a positive grade (Ouyang et al., 2022)? But training may require feedback on individual steps (Lu et al., 2026; Li et al., 2026; Yao et al., 2026). A trajectory can fail even when some steps are correct, and it can succeed even when some steps are wrong. Copying the final grade to every step can therefore create wrong local labels.

Figure 1 shows the basic problem. An early forecast is correct, but the final recount is wrong. If the final failure grade is copied to every forecast, the correct forecast receives the wrong label. This is the global-to-local supervision problem that we study.

One solution is to label every intermediate step directly. Process supervision follows this idea and has proved useful for reasoning models (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024). But dense local labels can be expensive (Luo et al., 2024). We ask whether useful local supervision can instead be learned from coarser feedback and a small number of local labels (Zhou, 2018; Ratner et al., 2017; Chen et al., 2024; Schroeter et al., 2019).

Our first source of supervision is aggregate feedback (Zhang et al., 2020; Quadrianto et al., 2009; Yu et al., 2014). In our setting, an exact subtotal constrains several local outcomes without revealing each one. A small number of local labels reduces the remaining ambiguity (Shi et al., 2020). We combine these signals to train a critic for local correctness.

Hindsight provides another source of information (Andrychowicz et al., 2017; Harutyunyan et al., 2019). A claim may be made before its outcome is known. Later in the trajectory, the relevant evidence may become available (Turtel et al., 2026; Karlsson et al., 2022). Our critic uses this later evidence to reassess the earlier claim (Feng et al., 2026; Zheng et al., 2026). Only after estimating

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

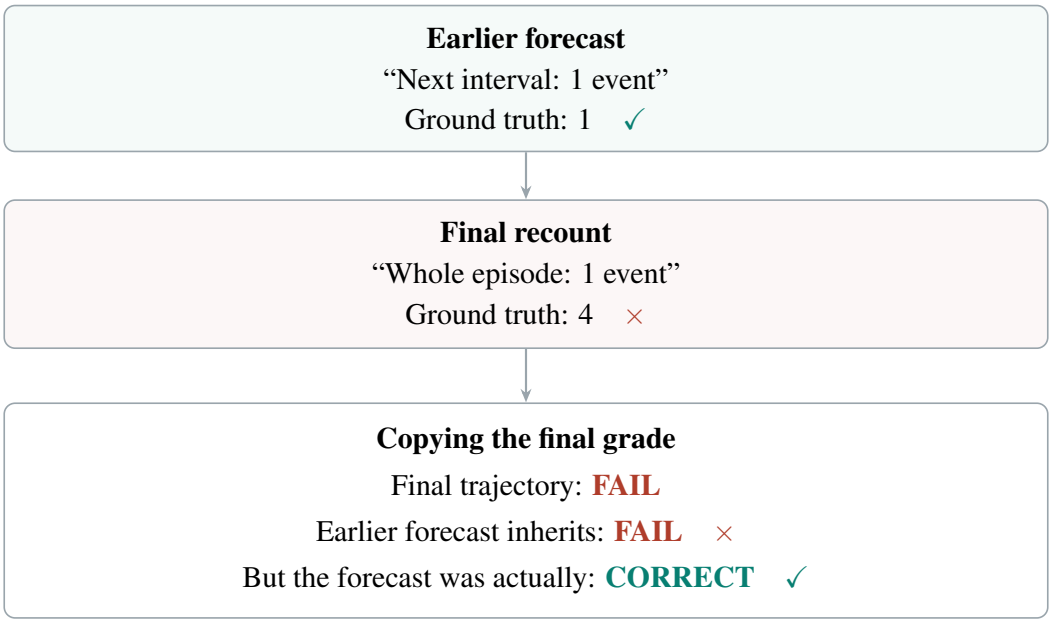

Figure 1: A final grade can be wrong for an intermediate claim. The first forecast correctly predicts one event in the next interval, but the later whole-episode recount is wrong. Copying the final failure grade to every forecast therefore gives the first forecast the wrong label.

local correctness do we compare it with the final grade and ask whether the inherited local label is wrong (Northcutt et al., 2021; Bernhardt et al., 2022).

To study this problem cleanly, we use a controlled multimodal forecast-and-recount setting. A fixed language model makes three event-count forecasts over successive video intervals and later recounts the complete episode (Qwen Team, 2026). Simulator records provide independent ground truth for both the local forecasts and the final recount (Yi et al., 2020). During training, the critic receives the exact subtotal across the three target intervals together with a small number of local count labels. At evaluation, the retrospective critic sees the target interval after it has occurred, while a matched static critic only sees evidence available before that interval. The final binary grade is a separate signal.

The results show a large benefit from hindsight. On a frozen final set, the retrospective critic reaches a local-correctness log loss of 0.3943, compared with 0.5781 for the matched static critic, a reduction of about 32%. With the same hindsight but without the sparse local labels, log loss is 0.5440. A separate frozen validation set shows the same pattern. The critic also improves the prioritization of wrong inherited labels under a fixed review budget.

Our theory explains why these ingredients fit together. Aggregate feedback constrains local outcomes but does not generally determine them (Li et al., 2024). Sparse local labels reduce this ambiguity. Hindsight can help when it contains information that was not available at prediction time. Finally, a probability of local correctness conditional on the final grade can be converted directly into the probability that an inherited label is wrong. The theory follows the same sequence as the method: coarse supervision, local information, hindsight, and local label correction.

Our contributions are threefold:

1. We formulate a global-to-local supervision problem for agentic trajectories: a final grade can be valid for the full trajectory but wrong for an individual step.

2. We show how aggregate feedback, sparse local labels, and hindsight can be combined to learn local correctness without dense step-level supervision.

3. We provide statistical results that explain the information available from these signals, together with frozen validation and final experiments in a controlled multimodal setting.

2 From Global Grades to Local Supervision

Consider a trajectory with several local claims and one final grade. Let $Y_j \in \{0, 1\}$ denote whether local claim j is correct, and let $L \in \{0, 1\}$ be the grade assigned to the trajectory as a whole. If L is copied to every local claim, then claim j receives the wrong label whenever

$$M_j = \mathbf{1}_{\{Y_j \neq L\}}.$$

This is the supervision error we want to identify.

In our multimodal setting, each local claim is an event-count forecast. Let a_j be the model's forecast and N_j the reference count for that interval. Its local correctness is

$$Y_j = \mathbf{1}_{\{a_j = N_j\}}.$$

The final grade is defined separately. After making three forecasts, the model recounts the entire episode. The binary grade L says whether this final recount is correct. A local forecast can therefore be correct even when $L = 0$, or wrong even when $L = 1$.

Learning from coarse feedback. Dense local supervision would reveal every N_j. We instead train with a coarser signal. For each training trajectory, we observe the subtotal

$$T = N_1 + N_2 + N_3,$$

together with a small number of individual local counts.

The subtotal constrains the local counts but does not usually determine them. For example, $T = 3$ is consistent with $(0, 1, 2)$, $(1, 1, 1)$, and many other allocations. A local label removes part of this ambiguity. This is why aggregate feedback and sparse local supervision provide complementary information.

Notice that the subtotal T and the final grade L play different roles. The subtotal is used to train the local critic. The final grade is used only after the critic has estimated local correctness.

From local correctness to inherited-label error. Let p_j be the critic's estimated probability that local claim j is correct. Once the final grade L is known, the corresponding probability that the inherited label is wrong is

$$q_j = L(1 - p_j) + (1 - L)p_j.$$

The formula is simple. If the final grade is positive, the inherited label is wrong when the local claim is wrong, so $q_j = 1 - p_j$. If the final grade is negative, the inherited label is wrong when the local claim is correct, so $q_j = p_j$.

This separation is central to our approach. We first learn whether a local claim is correct. We only then use the final grade to determine whether its inherited label should be trusted.

3 Learning from Aggregate Feedback and Hindsight

Figure 2 summarizes the method.

For each local claim j, the critic predicts a distribution over its event count,

$$f_\theta(n \mid X_j),$$

where X_j is the visual evidence available to the critic. If the model forecast is a_j, we use

$$p_j = f_\theta(a_j \mid X_j)$$

as the critic's estimate that the forecast is correct.

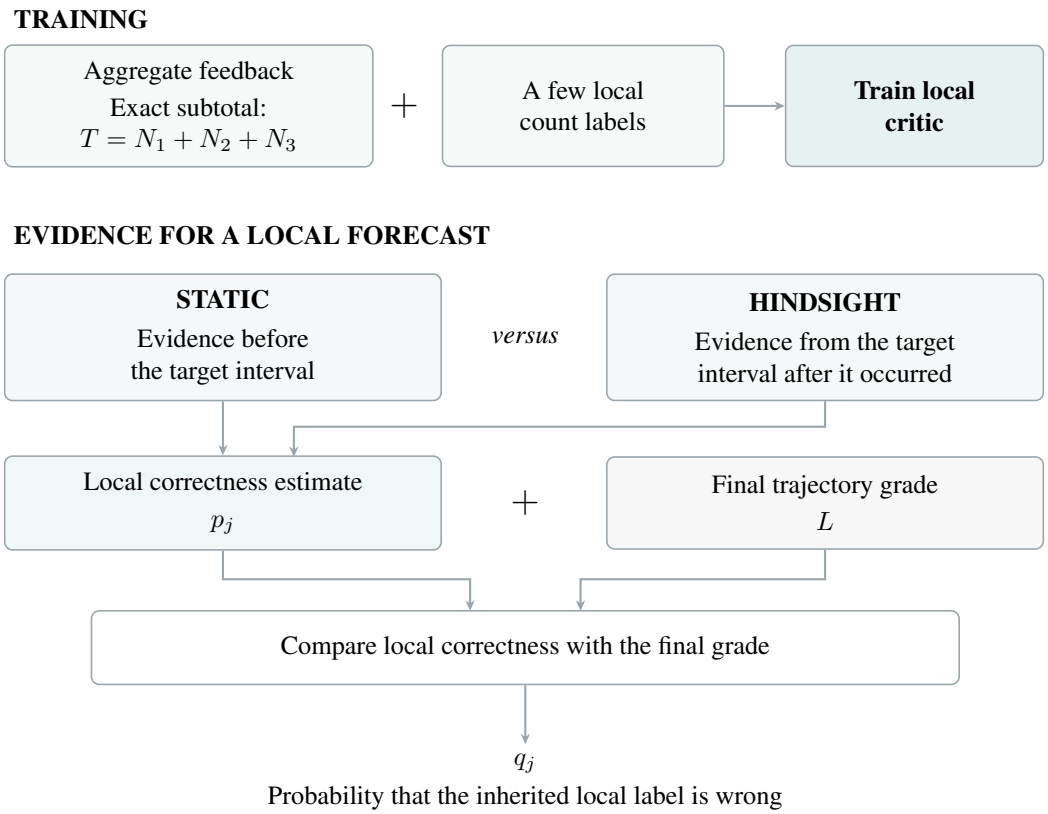

Figure 2: Learning local supervision from aggregate feedback and hindsight. Exact subtotals and a small number of local counts train the local critic. The static critic uses evidence available before the target interval, while the hindsight critic uses the target interval after it has occurred. The final grade is used afterward to identify inherited local labels that are likely to be wrong.

Learning from aggregate feedback. Most training trajectories do not reveal the three local counts separately. Instead, we observe only their subtotal,

$$T = N_1 + N_2 + N_3.$$

If $T = t$, the three local predictions are trained through

$$P_\theta(T = t \mid X) = \sum_{n_1+n_2+n_3=t} \prod_{j=1}^{3} f_\theta(n_j \mid X_j).$$

The model therefore receives credit when its local predictions are jointly consistent with the observed total.

A small number of training trajectories also reveal one local count. When this happens, we fix that count and consider only the remaining count combinations that agree with the same subtotal. These local labels help determine how the aggregate total should be distributed across the individual claims.

We first train on 256 exact subtotals and then continue training with 77 local count labels. The matched static and hindsight critics use the same supervision and training schedule.

Using hindsight. The two critics differ only in the evidence used to judge each forecast. The **static critic** sees the video immediately before the target interval. The **hindsight critic** sees the target interval after it has occurred.

The distinction is simple. The static critic must judge a forecast using only earlier evidence. The hindsight critic can also use what actually happened afterward. Neither critic sees the reference count at evaluation.

Both critics use the same frozen video representation and the same prediction network. We fit two count models and average their probabilities with a fixed $1/2$–$1/2$ mixture. The mixture weight is fixed before validation and final evaluation. Architectural and training details are given in Appendix B.

The local critic does not use the final trajectory grade. It first estimates whether each forecast is correct. We then compare this estimate with the final grade. If they disagree, the label inherited from the full trajectory is likely to be wrong. Section 2 gives the corresponding probability q_j.

Our main comparisons test these two ideas directly: sparse local labels can reduce the ambiguity left by aggregate feedback, and hindsight can add evidence that is unavailable to a static critic.

4 Controlled Multimodal Evaluation

We evaluate the method on CLEVRER videos (Yi et al., 2020), where simulator records provide independent ground truth for individual events. This lets us measure local forecast correctness and the final trajectory grade separately.

Each episode contains 128 video frames, divided into four 32-frame intervals. A fixed Qwen3.6-35B-A3B model (Qwen Team, 2026) makes three forecasts. After each of the first three intervals, it predicts the number of events in the next interval. After seeing the full episode, it separately recounts the total number of events.

The three forecasts define the local claims. A forecast is correct when its predicted count matches the simulator count in its target interval. The final grade is determined separately by whether the whole-episode recount is correct. The final recount can therefore be right even when an earlier forecast is wrong, or wrong even when an earlier forecast is right.

Frozen validation and final sets. We use two separate evaluation sets, each containing 256 scene groups and 768 local forecasts. Before examining their outcome labels, we fixed the models, baselines, parsers, evidence available to each critic, review budgets, and evaluation code.

We first evaluated the frozen validation set. The final set was evaluated only after the prespecified validation criteria were satisfied. The two sets are reported separately and are never pooled. Full cohort construction and validation criteria are given in Appendix C.

Comparisons. Our main comparison isolates hindsight. The hindsight and static critics use the same training data, architecture, and optimization schedule. They differ only in the video evidence used for each forecast: the static critic sees the preceding interval, while the hindsight critic sees the realized target interval.

We also compare with a hindsight critic trained only from aggregate subtotals. This measures the value of adding sparse local labels while keeping the visual evidence fixed. Additional controls, including local-label-only, fully supervised, and alternative training objectives, are described in Appendix D.

Evaluation. Our primary metric is log loss for local correctness; lower is better. We also report the Brier score, the mean squared error of the predicted probability, as a second measure of probabilistic accuracy (Gneiting & Raftery, 2007).

As a secondary evaluation, we rank local claims by the probability that their inherited label is wrong. We then measure how many wrong labels are found when reviewing a fixed fraction of the claims. We use review budgets of 5%, 10%, and 20%.

Confidence intervals use 2,000 paired whole-group bootstrap samples. All reported validation and final comparisons use the models and evaluation rules fixed before their corresponding outcome labels were examined.

5 Results

Global grades are often wrong locally. On the frozen final set, 401 of 768 local claims (52.2%) disagree with the final grade they inherit. The validation set shows almost the same rate: 399 of

768 (52.0%). The errors occur in both directions. A correct local claim can inherit failure, and an incorrect claim can inherit success.

Sparse local labels improve learning from aggregate feedback. We first isolate the value of sparse local supervision while keeping hindsight fixed. With aggregate feedback alone, the final log loss is 0.5440. Adding 77 local count labels reduces it to 0.3943. Brier score improves from 0.1850 to 0.1235.

The validation set shows the same pattern: log loss improves from 0.5704 to 0.4061. Sparse local labels therefore improve the local critic substantially beyond aggregate-only training.

Hindsight improves local correctness. We next compare the hindsight critic with a matched static critic. The two models use the same supervision, architecture, training schedule, and number of input tokens. They differ only in the evidence used for each forecast.

On the final set, hindsight reduces log loss from 0.5781 to 0.3943, a 32% reduction. Brier score improves from 0.1987 to 0.1235. The validation set again shows the same pattern, with log loss improving from 0.6409 to 0.4061.

These two comparisons support the two main components of our approach: sparse local labels help resolve the ambiguity left by aggregate feedback, and hindsight provides better evidence for judging earlier claims.

Table 1 summarizes the main comparisons.

Table 1: Main local-correctness results. Lower is better. The static and hindsight critics in the first two rows use the same supervision, architecture, and training schedule. The aggregate-only comparison keeps hindsight fixed but removes the sparse local labels. The fully supervised model uses all 768 local count labels.

	Validation		Final	
Method	Log loss	Brier	Log loss	Brier
Aggregate + sparse labels, hindsight	0.4061	0.1278	0.3943	0.1235
Aggregate + sparse labels, static	0.6409	0.2252	0.5781	0.1987
Aggregate only, hindsight	0.5704	0.1946	0.5440	0.1850
Fully supervised hindsight	0.2978	0.0891	0.3111	0.0958

Sparse supervision does not match dense local supervision: the fully supervised hindsight model reaches a final log loss of 0.3111. Our result is therefore about learning useful local supervision from much coarser feedback, not replacing dense labels altogether.

Finding wrong inherited labels. We also rank local claims by the probability that their inherited label is wrong. At a 20% review budget, the hindsight critic finds 149 wrong labels on the final set, compared with 136 for the matched static critic, a gain of 13 [4.98, 22]. The validation gain is larger: 149 versus 116, or +33 [23, 43].

6 Discussion and Limitations

Our results show that a final trajectory grade need not be a reliable label for every intermediate claim. In our controlled setting, aggregate feedback and a small number of local labels are enough to learn useful local correctness estimates, and hindsight improves those estimates substantially over matched static evidence. This gives a simple alternative to copying the final grade to every step.

The theory is broader than the particular video-counting experiment. It explains why an aggregate signal can leave local outcomes ambiguous, how sparse local information can reduce that ambiguity, and when hindsight can add information that was unavailable earlier. It also connects local correctness to errors in inherited labels. These results describe the statistical structure of the supervision problem; they do not guarantee that hindsight will improve every learned critic.

Our experiments have a deliberately narrow scope. They use one model family, one synthetic video domain, and a counting task with exact simulator ground truth. The aggregate subtotal is richer than a binary final grade, and the hindsight critic observes the realized target interval directly.

A natural next step is to study the same supervision problem in broader agentic trajectories, where local outcomes can be verified after a task unfolds. It would also be useful to learn which local labels to request and to test whether better local supervision improves downstream training.

Reproducibility Statement

We provide complete proofs of the theoretical results and additional technical details in the appendix. The experimental setup, implementation details, hyperparameters, and additional numerical results are also documented in the appendix.

AI use statement

Generative AI tools were used to assist with software implementation and debugging, as well as with drafting and editing portions of the manuscript for clarity and presentation. The core research questions, methodology, theoretical results, and scientific conclusions were developed by the authors. All AI-assisted code and text were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/453fadbd8a1a3af50a9df4df899537b5-Abstract.html.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2103.15691.

Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13:1161, 2022. doi: 10.1038/s41467-022-28818-3. URL https://www.nature.com/articles/s41467-022-28818-3.

Hao Chen, Jindong Wang, Lei Feng, Xiang Li, Yidong Wang, Xing Xie, Masashi Sugiyama, Rita Singh, and Bhiksha Raj. A general framework for learning from weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7462–7485, 2024. URL https://proceedings.mlr.press/v235/chen24ar.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006. doi: 10.1002/047174882X. URL https://doi.org/10.1002/047174882X.

Yangyang Feng, Zhuoyan Feng, and Junlan Chen. PAST: Privileged adaptation from complete student trajectories for on-policy self-distillation. *arXiv preprint arXiv:2608.08726*, 2026. URL https://arxiv.org/abs/2608.08726.

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

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4651–4664, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Rickard K. A. Karlsson, Martin Willbo, Zeshan M. Hussain, Rahul G. Krishnan, David Sontag, and Fredrik Johansson. Using time-series privileged information for provably efficient learning of prediction models. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 5459–5484, 2022. URL https://proceedings.mlr.press/v151/k-a-karlsson22a.html.

Gene Li, Lin Chen, Adel Javanmard, and Vahab Mirrokni. Optimistic rates for learning from label proportions. In *Proceedings of Thirty Seventh Conference on Learning Theory*, 2024. URL https://proceedings.mlr.press/v247/li24b.html.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents. *arXiv preprint arXiv:2605.19447*, 2026. URL https://arxiv.org/abs/2605.19447.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Zhicong Lu, Zichuan Lin, Wei Jia, Changyuan Tian, Deheng Ye, Peiguang Li, Li Jin, Nayu Liu, Guangluan Xu, and Wei Feng. HISR: Hindsight information modulated segmental process rewards for multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2603.18683*, 2026. URL https://arxiv.org/abs/2603.18683.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024. URL https://arxiv.org/abs/2406.06592.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026. URL https://arxiv.org/abs/2603.14482.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021. URL https://jair.org/index.php/jair/article/view/12125.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances*

in Neural Information Processing Systems, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Novi Quadrianto, Alex J. Smola, Tibério S. Caetano, and Quoc V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 2009. URL https://www.jmlr.org/papers/v10/quadrianto09a.html.

Qwen Team. Qwen3.6-35B-A3B model card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Accessed 25 September 2026; experimental checkpoint and quantization pinned separately.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797. URL https://vldb.org/pvldb/vol11/p269-ratner.pdf.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010. URL https://www.jmlr.org/papers/v11/reid10a.html.

S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977. doi: 10.1108/eb026647. URL https://doi.org/10.1108/eb026647.

Julien Schroeter, Kirill Sidorov, and David Marshall. Weakly-supervised temporal localization via occurrence count learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5649–5659, 2019. URL https://proceedings.mlr.press/v97/schroeter19a.html.

Yong Shi, Jiabin Liu, Bo Wang, Zhiquan Qi, and YingJie Tian. Deep learning from label proportions with labeled samples. *Neural Networks*, 128:73–81, 2020. doi: 10.1016/j.neunet.2020.04.026. URL https://www.sciencedirect.com/science/article/pii/S0893608020301544.

Benjamin Turtel, Paul Wilczewski, Danny Franklin, and Kris Skothiem. Future-as-label: Scalable supervision from real-world outcomes. *arXiv preprint arXiv:2601.06336*, 2026. URL https://arxiv.org/abs/2601.06336.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2211.14275. URL https://arxiv.org/abs/2211.14275.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://aclanthology.org/2024.acl-long.510/.

Jiashu Yao, Heyan Huang, Zeming Liu, and Yuhang Guo. Utilizing and calibrating hindsight process rewards via reinforcement with mutual information self-evaluation. *arXiv preprint arXiv:2604.11611*, 2026. URL https://arxiv.org/abs/2604.11611.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.01442.

Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014. URL https://arxiv.org/abs/1402.5902.

Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5b0fa0e4c041548bb6289e15d865a696-Paper.pdf.

Haoyu Zheng, Yun Zhu, Qing Wang, and Wenqiao Zhang. Trajectory-relative hindsight distillation for agentic reinforcement learning. *arXiv preprint arXiv:2608.07371*, 2026. URL https://arxiv.org/abs/2608.07371.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 2018. URL https://academic.oup.com/nsr/article/5/1/44/4093912.

A Statistical Structure of Global-to-Local Supervision

The results in this appendix explain the statistical structure behind the method in the main paper. We first ask what an aggregate observation can identify about its components. We then relate local correctness to errors in inherited labels, characterize the information gained from hindsight, and connect probability estimation to fixed-budget review. Only the first result uses the Poisson count model; the remaining results apply more generally to binary local correctness.

A.1 What Aggregate Feedback Leaves Unresolved

Let $N_1, \dots, N_m$ denote local nonnegative counts and let

$$T = \sum_{j=1}^{m} N_j$$

be their aggregate total. Observing T constrains the local counts, but in general it does not determine how the total is distributed across them (Zhang et al., 2020). The following simple Poisson model makes this ambiguity explicit.

Proposition 1 (Totals and local labels). Suppose the local counts are independent with

$$N_j \sim \mathrm{Pois}(\lambda_j), \qquad \lambda_j > 0,$$

and let $S \subsetneq \{1, \dots, m\}$ be the set of components whose local counts are also observed. Two rate vectors λ and λ' induce the same distribution of

$$(T, (N_j)_{j \in S})$$

if and only if

$$\lambda_j = \lambda'_j \quad \text{for every } j \in S,$$

and

$$\sum_{j \notin S} \lambda_j = \sum_{j \notin S} \lambda'_j.$$

Thus the aggregate total identifies the sum of the unobserved rates, but in general not their individual allocation.

Proof. Let $R = S^c$. The observation

$$(T, (N_j)_{j \in S})$$

is equivalent to observing

$$\left(T - \sum_{j \in S} N_j, (N_j)_{j \in S} \right).$$

The first coordinate is

$$\sum_{j \in R} N_j \sim \mathrm{Pois}\left(\sum_{j \in R} \lambda_j \right),$$

and it is independent of the anchored counts $(N_j)_{j \in S}$. The observed distribution is therefore determined exactly by the anchored rates and the sum of the remaining rates. Equality of these quantities is sufficient for equality of the observed distributions. Conversely, equality of the observed distributions implies equality of the corresponding Poisson means, which gives the stated conditions. □

Without any local labels, all positive rate vectors with the same total rate are observationally equivalent under this model. With one local label, one component rate is fixed and the remaining rates must share the residual total. More local information therefore narrows the ambiguity, although it need not remove it completely.

Shared structure across trajectories can provide additional information. For example, if the rates are constrained by a common parameterization, aggregate observations from many trajectories may identify parameters that a single unrestricted total cannot identify.

There is also a simpler distinction at the level of realized counts. Suppose $d \geq 1$ local counts remain unknown and their residual subtotal is $s \geq 0$. The number of nonnegative integer allocations is

$$\binom{s+d-1}{d-1}.$$

The allocation is unique when $s = 0$ or $d = 1$. In particular, an observed total $T = 0$ immediately implies that every local count is zero. This is a statement about realized counts, not identification of their underlying rates.

For $r = |S| < m$, the remaining positive rates range over a simplex of dimension $m - r - 1$. With $m - 1$ anchored rates the last rate is determined; with all m rates anchored nothing remains unobserved. Dividing the joint Poisson mass by the subtotal mass also gives

$$(N_1, \ldots, N_m) \mid T = t \sim \text{Multinomial}\left(t; \lambda_1/\Lambda, \ldots, \lambda_m/\Lambda\right), \qquad \Lambda = \sum_{j=1}^{m} \lambda_j.$$

Identification from an observed law does not equate a single count with its unknown Poisson mean. With $d = 0$, consistency requires $s = 0$. An anchor can reveal complementary counts, but an already implied answer adds no constraint.

A.2 From Local Correctness to an Inherited-Label Error

Let $Y \in \{0, 1\}$ denote local correctness and let $L \in \{0, 1\}$ be the final grade inherited by the local claim. Define

$$M = \mathbf{1}_{\{Y \neq L\}},$$

so that $M = 1$ exactly when the inherited label is wrong.

The relation between local correctness and inherited-label error is deterministic once L is known.

Proposition 2 (Probability reflection). Let $\mathcal{G}$ be any information set that contains L. If

$$p = \mathbb{P}(Y = 1 \mid \mathcal{G}),$$

then

$$\mathbb{P}(M = 1 \mid \mathcal{G}) = L(1 - p) + (1 - L)p.$$

Proof. Because L is binary and belongs to $\mathcal{G}$,

$$M = L(1 - Y) + (1 - L)Y.$$

Taking conditional expectations with respect to $\mathcal{G}$ gives the result. □

For $L = 0$, the probability of an inherited-label error is p: the negative final grade is wrong exactly when the local claim is correct. For $L = 1$, the probability is $1 - p$: the positive final grade is wrong exactly when the local claim is incorrect.

The same reflection also preserves any label-symmetric loss (Reid & Williamson, 2010).

Corollary 1 (Loss preservation under reflection). Let

$$q = L(1 - p) + (1 - L)p.$$

If a binary loss satisfies

$$\ell(1 - p, 1 - y) = \ell(p, y),$$

then

$$\ell(q, M) = \ell(p, Y).$$

In particular, this identity holds for log loss and Brier score.

Proof. If $L = 0$, then $(q, M) = (p, Y)$. If $L = 1$, then

$$(q, M) = (1 - p, 1 - Y),$$

and the result follows from label symmetry. □

The algebraic reflection is exact, but its posterior interpretation requires care. Proposition 2 uses the probability of local correctness conditional on information that already contains the final grade. Replacing

$$\mathbb{P}(Y = 1 \mid X, a, L)$$

with

$$\mathbb{P}(Y = 1 \mid X, a)$$

requires these two conditional probabilities to agree. The critic used in the main experiments produces a model-based local correctness score and does not assume this equality. The reflection in the main method should therefore be understood as the specified readout from that score and the observed final grade, rather than as an assumption that the critic is an oracle posterior.

The loss identity holds for every (p, Y, L) if and only if the loss is label-symmetric: taking $L = 1$ proves necessity. Equality includes infinite endpoint losses; no infinite quantities are subtracted. This also applies to spherical scores. It does not require properness: absolute loss is symmetric but not proper for probability estimation. Conversely, $(p - y)^2 + y$ is strictly proper and not symmetric.

Equality of the two conditional probabilities is equivalent to $Y \perp L \mid (X, a)$. For balanced Y, the calibrated score $p = 1/2$ with $L = Y$ violates this condition. Even evaluating a true count law at an informative random claim requires care about conditioning on that claim.

A.3 When Hindsight Adds Information

We next separate two questions: whether later evidence contains useful information in principle, and whether a fitted model can exploit that information.

Let W denote the evidence available before the target interval, let H denote hindsight evidence, and let

$$C = (a, L)$$

collect the information shared by both comparisons. For $U \in \{W, H, (W, H)\}$, define

$$q_U = \mathbb{P}(M = 1 \mid U, C).$$

Let

$$R_U = \mathbb{E}[\ell_{\log}(q_U, M)]$$

be the Bayes log-loss risk, and let

$$B_U = \mathbb{E}\big[(M - q_U)^2\big]$$

be the corresponding Bayes Brier risk. Use natural logarithms, $0 \log 0 = 0$ in entropy, and infinite loss for an incompatible deterministic prediction.

Proposition 3 (Value of additional evidence). Adding hindsight to static evidence gives

$$R_W - R_{W,H} = I(M; H \mid W, C) \geq 0,$$

and

$$B_W - B_{W,H} = \mathbb{E}\left[(q_{W,H} - q_W)^2\right] \geq 0.$$

Both differences are zero exactly when

$$M \perp H \mid (W, C).$$

Proof. Under log loss, the Bayes risk is the conditional entropy of M (Cover & Thomas, 2006). Therefore

$$\begin{aligned} R_W - R_{W,H} &= H(M \mid W, C) - H(M \mid W, H, C) \\ &= I(M; H \mid W, C). \end{aligned}$$

For Brier score, write

$$M - q_W = (M - q_{W,H}) + (q_{W,H} - q_W).$$

The cross term has zero expectation after conditioning on (W, H, C), which gives the stated decomposition. Each difference vanishes exactly when the corresponding conditional distributions agree. □

This proposition concerns adding hindsight to information that is already available. Our implemented hindsight critic instead uses H in place of W. For this replacement,

$$R_W - R_H = I(M; H \mid W, C) - I(M; W \mid H, C),$$

and

$$B_W - B_H = \mathbb{E}\left[(q_{W,H} - q_W)^2\right] - \mathbb{E}\left[(q_{W,H} - q_H)^2\right].$$

Hindsight can therefore add information while also discarding information contained in W. There is no general guarantee that replacing W by H must improve prediction. The matched comparison in the main experiments measures which effect dominates in our setting.

The distinction is immediate in simple examples. With C constant and M nondegenerate, if W determines M exactly and H is independent of M, replacing W with H loses information. Reversing the roles of W and H reverses the conclusion.

Bayes information versus fitted performance. The identities above concern Bayes-optimal predictions. A fitted critic also has approximation and estimation error. For any score $s_U \in [0, 1]$ measurable from (U, C),

$$\mathbb{E}(M - s_U)^2 = B_U + \mathbb{E}(s_U - q_U)^2,$$

and

$$\mathbb{E}[\ell_{\log}(s_U, M)] = R_U + \mathbb{E}\left[D_{\mathrm{KL}}(\mathrm{Bern}(q_U) \| \mathrm{Bern}(s_U))\right].$$

Thus a fitted-model comparison reflects both the information in the evidence and how well the critic learns from it. The information identities alone do not guarantee a finite-sample improvement for a particular trained model.

The log-loss reduction in Proposition 3 also equals

$$\mathbb{E} D_{\mathrm{KL}}(\mathrm{Bern}(q_{W,H}) \| \mathrm{Bern}(q_W)).$$

Since L is common information, the map between Y and M is a conditional bijection. The Brier difference and the mutual information are identical with Y in place of M. One may therefore state the result for either local truth or inherited-label error, provided the conditioning includes L.

A.4 From Probability Accuracy to Fixed-Budget Review

Finally, consider reviewing exactly K of n local claims. Let

$$q_i^\star = \mathbb{P}(M_i = 1 \mid \mathcal{F})$$

be the true conditional probability that inherited label i is wrong, and let $\widehat{q}_i$ be an estimated probability based on the same available information. Let $A^\star$ contain the K largest values of $q_i^\star$, and let $\widehat{A}$ contain the K largest values of $\widehat{q}_i$, with fixed tie breaking (Robertson, 1977).

The conditional expected number of wrong labels missed by using $\widehat{A}$ instead of $A^\star$ is

$$\Delta = \sum_{i \in A^\star} q_i^\star - \sum_{i \in \widehat{A}} q_i^\star.$$

Proposition 4 (Fixed-budget review bound). For $0 < K < n$, let

$$c_K = \min(K, n - K).$$

Then

$$\Delta \le \min\left\{ c_K,\ \sqrt{2c_K \sum_{i=1}^{n} (q_i^\star - \widehat{q}_i)^2},\ \sqrt{c_K \sum_{i=1}^{n} D_{\mathrm{KL}}(\mathrm{Bern}(q_i^\star) \| \mathrm{Bern}(\widehat{q}_i))} \right\}.$$

For $K = 0$ or $K = n$, $\Delta = 0$.

Proof. Write

$$e_i = q_i^\star - \widehat{q}_i.$$

Because $\widehat{A}$ maximizes the sum of $\widehat{q}_i$ over all sets of size K,

$$\Delta \le \sum_{i \in A^\star \setminus \widehat{A}} e_i - \sum_{i \in \widehat{A} \setminus A^\star} e_i.$$

Applying Cauchy–Schwarz gives

$$\Delta \le \sqrt{|A^\star \triangle \widehat{A}| \sum_{i \in A^\star \triangle \widehat{A}} e_i^2}.$$

At most $2c_K$ entries belong to the symmetric difference, hence

$$\Delta \le \sqrt{2c_K \sum_{i=1}^{n} e_i^2}.$$

At most c_K selected entries can differ and every probability lies in $[0, 1]$, which also gives

$$\Delta \le c_K.$$

Finally, Bernoulli Pinsker's inequality gives

$$D_{\mathrm{KL}}(\mathrm{Bern}(q_i^\star) \| \mathrm{Bern}(\widehat{q}_i)) \ge 2(q_i^\star - \widehat{q}_i)^2.$$

Substituting this inequality into the squared-error bound yields the KL bound. The endpoint cases $K = 0$ and $K = n$ select the same set under every ranking. □

When the final grade L_i is fixed, reflection from local correctness to inherited-label error preserves the absolute probability error:

$$|q_i^\star - \widehat{q}_i| = |p_i^\star - \widehat{p}_i|.$$

It also preserves the corresponding Bernoulli KL divergence. The bound therefore connects probability accuracy for local correctness to the expected quality of a fixed-budget review.

The result is one-way. Better probability estimates can reduce the worst-case ranking loss, but a strict improvement in log loss or Brier score need not change the top-K ordering in a particular dataset.

The two sums are n times conditional mean excess Brier and log loss over the posterior (Gneiting & Raftery, 2007). Expected discovery counts use linearity and require no independence among components. Incompatible probability endpoints make the KL bound infinite, while the Brier bound remains valid.

Additional information and the optimal review. For nested common batch information fields $\mathcal{F}_0 \subseteq \mathcal{F}_1$, let $q_{ui} = \mathbb{E}[M_i \mid \mathcal{F}_u]$, $V_u = \max_{|A|=K} \sum_{i \in A} q_{ui}$, and let A_0 be a fixed, measurable optimizer under $\mathcal{F}_0$. The tower property gives

$$\mathbb{E}V_1 - \mathbb{E}V_0 = \mathbb{E}\left[V_1 - \sum_{i \in A_0} q_{1i} \right] \ge 0.$$

Equality holds precisely when A_0 remains optimal under the refined risks almost surely. Strict proper-loss improvement is insufficient: equiprobable refined risk vectors $(.7, .1)$ and $(.9, .3)$ average to $(.8, .2)$; both coordinates improve in Bayes Brier loss, but the top-one choice and expected discovery count remain unchanged.

B Model and Training Details

This appendix gives the implementation details omitted from the main paper. We first describe the evidence seen by each critic, then the video representation and count models, followed by the training objective and optimization schedule.

B.1 Evidence Seen by the Critic

Each video is divided into four 32-frame intervals. For the three local forecasts, the target intervals are frames 32–63, 64–95, and 96–127.

The **static critic** uses the complete 32-frame interval immediately before the target interval. The **hindsight critic** uses the complete target interval after it has occurred. A full-prefix static control uses all video intervals available before the forecast.

No critic receives the simulator event records, the reference count for the target interval, or held-out aggregate counts at evaluation. The primary local critic also does not receive the model's final recount or its final grade as a neural input. The target index determines which video interval is used, but is not itself provided as a learned feature. The model's forecast is not a neural input.

B.2 Video Representation

We use frozen V-JEPA 2.1 features for the video input (Mur-Labadia et al., 2026). Each 32-frame interval is encoded independently. A fixed channel projection produces a spatiotemporal representation with 16 temporal positions, a 20×30 spatial grid, and 128 channels. The shared label-free scaler uses only the nested 80-family source subset.

The prediction network first summarizes each temporal position using four learned spatial queries. Temporal self-attention and a feed-forward block then combine information across time before pooling to a single representation (Arnab et al., 2021; Jaegle et al., 2021).

Relative time and spatial coordinates are included. For inputs containing multiple preceding intervals, relative time is measured with respect to the most recent interval.

The video encoder remains frozen throughout training. The static and hindsight critics use the same prediction architecture and have the same number of input tokens in the primary matched comparison.

B.3 Count Distributions

We fit two models for the local event count.

The first uses a Poisson distribution with rate

$$\lambda(X) = \mathrm{softplus}(z(X)) + 10^{-6}.$$

The second is a more flexible count distribution. It predicts separate probabilities for counts $0, \ldots, 8$, together with a geometric tail for counts $n \geq 9$. If $\pi_0, \ldots, \pi_8, \pi_{\mathrm{tail}}$ are the normalized masses and $r \in (0, 1)$ is the tail parameter, then

$$f_F(n \mid X) = \begin{cases} \pi_n, & 0 \leq n \leq 8, \\ \pi_{\mathrm{tail}}(1-r)r^{n-9}, & n \geq 9. \end{cases}$$

Both distributions therefore have support on every nonnegative integer.

The two models are trained separately. The primary critic averages their probabilities with a fixed mixture,

$$f_{\mathrm{mix}}(n \mid X) = \frac{1}{2} f_P(n \mid X) + \frac{1}{2} f_F(n \mid X).$$

The mixture weight is fixed and is never selected using validation or final data. For a forecast a_j, the local correctness score used in the main paper is

$$p_j = f_{\mathrm{mix}}(a_j \mid X_j).$$

The Poisson and flexible prediction networks contain 59,009 and 59,659 learned parameters, respectively, in addition to the frozen video encoder.

B.4 Training from Aggregate and Local Counts

For one trajectory, let

$$N = (N_1, N_2, N_3)$$

denote the three target counts. We model their conditional distribution as

$$f_\theta(N \mid X) = \prod_{j=1}^{3} f_\theta(N_j \mid X_j).$$

This factorization is a modeling approximation; it does not assume that the physical events themselves are independent.

When only the subtotal

$$T = N_1 + N_2 + N_3 = t$$

is observed, the likelihood is

$$F_\theta(t \mid X) = \sum_{n_1+n_2+n_3=t} \prod_{j=1}^{3} f_\theta(n_j \mid X_j).$$

When one local count is also observed, say $N_j = n$, the likelihood becomes

$$A_\theta(n, t \mid X) = f_\theta(n \mid X_j) \sum_{n_k+n_\ell=t-n} f_\theta(n_k \mid X_k) f_\theta(n_\ell \mid X_\ell).$$

For trajectories with a local count label, this second likelihood replaces the subtotal-only likelihood; it does not add a second copy of the subtotal loss.

We minimize negative log likelihood (Zhang et al., 2020). Each trajectory-level contribution is divided by three so that the scale is comparable to a per-local-claim loss. Finite subtotal probabilities can be computed exactly even though the component count distributions have unbounded support.

B.5 Training Schedule

Training has two stages.

First, each critic is trained for 720 updates using 256 exact subtotals. We then restart the optimizer and continue for another 720 updates while incorporating 77 local count labels.

The 77 source trajectories are chosen from positive-subtotal families using a fixed hash ordering. An independently salted hash chooses which of the three local counts is revealed. These choices are fixed before the local count values are accessed.

Each update uses four trajectories, with three target intervals per trajectory. Optimization uses AdamW with learning rate 3×10^{-4}, weight decay 0.01, and gradient clipping at 1.

The same supervision, architecture, optimizer settings, and update schedule are used for the matched static and hindsight critics. Their primary difference is therefore the visual evidence described in Appendix B.1.

C Experimental Protocol

This appendix describes how the actor outputs, reference labels, validation set, and final set were constructed. The goal is to keep model selection and evaluation separate from the ground-truth labels used to measure performance.

C.1 Forecast-and-Recount Workflow

We use a fixed Qwen3.6-35B-A3B actor throughout the study. The exact checkpoint, quantization, revision, and decoding settings are fixed for all reported trajectories.

Each video contains 128 frames, divided into four 32-frame intervals. The actor makes three forecasts:

- after frames 0–31, it predicts the event count in frames 32–63;
- after frames 0–63, it predicts the event count in frames 64–95;
- after frames 0–95, it predicts the event count in frames 96–127.

After seeing the full 128-frame episode, the actor makes a separate whole-episode recount. The recount sees the earlier forecasts but is instructed to recount the complete episode rather than add or revise them.

Let N_0, N_1, N_2, N_3 denote the reference counts in the four intervals, and let c denote the actor's final recount. The three local forecasts concern N_1, N_2, N_3. Their training subtotal is

$$T = N_1 + N_2 + N_3.$$

The final grade is defined separately as

$$L = \mathbf{1}_{\{c=N_0+T\}}.$$

Thus the aggregate training signal T and the binary final grade L are different quantities.

Frozen actor configuration. The actor conversion is `mlx-community/Qwen3.6-35B-A3B-4bit`, revision `38740b847e4cb78f352aba30aa41c76e08e6eb46`. Decoding uses temperature 0, top-p 1, top-k 0, and a maximum of 512 generated tokens per call, with thinking disabled. Repetition, presence and frequency penalties, KV quantization, cross-request caches and draft models are disabled; prefill is unchunked.

C.2 Simulator Ground Truth

Ground truth comes from the released CLEVRER simulator records rather than from the actor or the local critic. We count deduplicated event records indexed by integer frame and unordered object pair, retaining events whose two participants are within camera view.

Adjacent records are not merged. The prompt asks for contact onsets, so continued contact is counted once rather than repeatedly. CLEVRER also excludes repeated collisions between the same pair within an episode (Yi et al., 2020). In the original 128-scene development inventory, the 321 recorded events contain no repeated object pairs and no discrepancy caused by adjacency merging.

We use the released event times and camera-visibility information directly. We do not perform a separate manual certification of event onset or occlusion. The reference counts are constructed independently of model scores, actor answers, and the critic.

C.3 Figure 1 Example

Figure 1 is an illustration of the global-to-local label problem, not a case selected by the critic.

The example is the first development trajectory in fixed source order for which the sum of the three parsed forecasts exceeds the parsed final recount. This rule was applied before the simulator truth for the example was inspected.

Its four reference counts are

$$(0, 1, 2, 1),$$

the three forecasts are

$$(1, 1, 0),$$

and the final recount is 1. The first forecast is therefore correct, while the final recount is wrong. Copying the final failure grade to the first forecast gives it the wrong local label.

C.4 Validation and Final Cohorts

The reported validation and final results use two separate cohorts of 256 scene groups each, giving 768 local forecasts per cohort.

Before examining their outcome labels, we started from a fixed public pool of 528 candidate scenes. These candidates were screened jointly against 1,060 scenes that had already been exposed during development. A frozen visual-similarity procedure formed connected components over the combined set. Any component touching a previous exposure was excluded.

This left 527 eligible components. We took the first 512 representatives in the original nomination order: 256 for validation and 256 for final evaluation. The screen examined all 1,260,078 pairs among the 1,588 combined scenes.

The similarity screen is intentionally conservative. It reduces reuse of visually similar scenes but does not prove that the selected simulator worlds are statistically independent or uniformly sampled.

Frozen similarity rule. Two scenes are connected if their video SHA-256 hashes match, their ordered RGB-frame hashes match, they share any RGB-frame hash, or the mean Hamming distance between their 128 aligned 64-bit difference hashes is at most 4. Each difference hash uses a grayscale frame resized to 9×8 with bilinear interpolation and the 64 comparisons of each pixel with its left neighbor. All connected components are formed before selection. The earliest candidate in the original nomination order represents each component that contains no previous exposure.

C.5 Freeze-Before-Truth Protocol

All candidate models, prediction rules, parsers, evidence choices, cohort rules, review budgets, and evaluation code were fixed before the validation ground truth was joined.

The validation set could determine only whether the already frozen candidate passed the prespecified criteria. It could not be used to fit a new model, change a seed, select a new mixture weight, or redefine the final evaluation.

The final ground truth was released only after the complete validation criterion passed. Validation and final results are reported separately and are never pooled.

The primary validation criterion required improvement over every required control by at least 0.05 log loss and 0.02 Brier score, with the paired 95% upper difference limit below zero. The primary also had to satisfy

$$\text{log loss} \leq 0.65, \qquad \text{Brier} \leq 0.23, \qquad \text{incorrectness AUC} \geq 0.75,$$

with at least 730 jointly evaluable local forecasts.

Together these formed 88 prespecified conditions. Four additional comparisons with the two individual count models were recorded separately. Performance under a fixed review budget could not rescue a model that failed the validation criterion.

C.6 Metrics and Uncertainty

Each validation or final cohort contains 768 local forecasts. The primary predictive metrics are log loss and Brier score for local correctness. We also retain incorrectness AUC and count-level diagnostics.

For the fixed-budget review, we evaluate

$$K \in \{39, 77, 154\},$$

corresponding to 5%, 10%, and 20% of the 768 local forecasts.

Uncertainty intervals use 2,000 paired whole-group bootstrap samples with equal group weighting. Entire scene groups, rather than individual forecasts, are resampled together. For each bootstrap sample, claims are ranked again before computing fixed-budget results.

All reported intervals condition on the frozen models and earlier development choices; they do not integrate over model selection or training-seed uncertainty.

Bootstrap execution and unknown predictions. The fixed resampling seed is 2026092261. Repeated draws of a group create separate occurrences that are ranked again. Unknown predictions consume review budget rather than being dropped; selected unknown labels retain lower and upper discovery bounds. Relative gain and oracle-headroom capture can be undefined when their denominator vanishes; undefined replicates are not assigned zero. Passing a required validation comparison requires all 2,000 paired loss differences to be finite and defined. Parsed public integers above 2048 or an unsupported grade-conditioned branch fail the phase rather than being clipped or omitted. Unknown final claims or grades use the prespecified unconditioned fallback.

D Additional Comparisons and Diagnostics

The main paper focuses on two comparisons: adding sparse local labels while holding hindsight fixed, and replacing static evidence with hindsight under matched training. This appendix reports the additional controls used to interpret those results. These comparisons answer different questions and should not be read as interchangeable tests of a single effect.

D.1 What Each Comparison Tests

The primary hindsight critic is trained first from aggregate subtotals and then from 77 local count labels. We use several controls to separate the role of supervision, evidence, and training objective.

Aggregate-only hindsight. This control uses the same hindsight evidence and the same initial aggregate training, but does not add the 77 local labels. It isolates the value of sparse local supervision under the tested continuation schedule.

Matched static critic. This control uses the same architecture, supervision, and training schedule as the primary critic, but replaces the realized target interval with the preceding interval. It isolates the empirical value of hindsight in this setting.

Local-label-only critics. These models use the same 77 local labels without aggregate pretraining. They test whether the sparse labels alone are sufficient, but they also differ in training coverage and curriculum from the primary critic. They should therefore be interpreted as whole-recipe comparisons rather than a pure estimate of the aggregate signal alone.

Fully supervised hindsight. These models use all 768 local counts. They provide a stronger-supervision reference, not an equal-cost comparison. As reported in the main paper, dense supervision remains better in point predictive loss.

Alternative aggregate objective. The residual-MSE control uses the same hindsight evidence, local labels, model capacities, and number of updates, but replaces the aggregate likelihood with a loss on the subtotal mean. This tests one alternative way to use aggregate information; it does not establish that every mean-based objective must perform worse.

Table D.1 summarizes these comparisons.

Table D.1: What the main controlled comparisons change. Not every row isolates a single causal effect; the text states the interpretation of each comparison.

Comparison	Held fixed	Changed
Hindsight vs. aggregate-only	hindsight evidence, architecture, aggregate pretraining, continuation budget	add 77 local labels
Hindsight vs. static	architecture, supervision, training schedule, input-token count	target interval vs. preceding interval
Hindsight vs. residual MSE	evidence, model capacity, local labels, update budget	aggregate likelihood vs. mean-based objective
Hindsight vs. local-label-only	hindsight evidence, architecture, same 77 local labels	aggregate pretraining, coverage, and curriculum
Sparse vs. fully supervised	hindsight evidence	77 local labels vs. 768 local labels

D.2 Alternative Training Objectives

The primary method models the full distribution of the aggregate count. One natural alternative is to fit only its mean.

On the final cohort, the residual-MSE mixture reaches log loss 0.4904 and Brier score 0.1669, compared with 0.3943 and 0.1235 for the primary likelihood model. The corresponding validation values are $0.5129/0.1758$ and $0.4061/0.1278$. Thus the full likelihood performs better for the primary local-correctness metric in this matched comparison.

This difference should not be interpreted as a universal advantage of likelihood over mean fitting. The comparison evaluates one specified residual-MSE objective under the same architecture and supervision. In fact, some fixed subgroups favor the MSE model, as reported below.

D.2.1 A Same-Information Development Comparison

After the frozen study was complete, we also compared the joint likelihood with a simpler aggregate-plus-singleton objective on an already exposed development set. This diagnostic used the same 256 subtotals and the same 26 local count answers for both methods.

The joint objective reaches log loss 0.44682 and Brier score 0.14401, while the aggregate-plus-singleton objective reaches 0.45133 and 0.14609. The paired differences are

$$-0.00451\ [-0.01045,\ 0.00146]$$

for log loss and

$$-0.00208\ [-0.00464,\ 0.00051]$$

for Brier score.

The difference is small and uncertain. This development comparison therefore does not show that the joint likelihood is necessary or generally superior. It does not alter the frozen validation or final comparisons, which test the value of sparse local information and hindsight rather than universal objective optimality.

Both methods used target-interval hindsight, the original aggregate-only checkpoints, 118,668 effective parameters, a fixed-half Poisson/flexible mixture, and one 720-update continuation per model with matched optimizer, batch stream and clipping. The 26 review-selected answers use a different label budget, mask and continuation stream from the frozen 77-label primary. They imply another 46 counts: 24 zeros from eight zero-subtotal groups and 22 from exhausted positive-subtotal groups. These are consequences of the same information, not extra purchased labels. The auxiliary retains subtotal log loss and adds singleton count log loss for purchased and logically implied counts; this changes weighting, doubling zero-subtotal loss. Both methods evaluate count probabilities at the forecast and use the same final-grade reflection. No binary-correctness loss is used; an unused binary head adds no effective predictive capacity. Evaluation uses 384 claims in 128 already exposed development groups and 2,000 paired whole-group resamples.

D.3 Count Models and Final-Grade Conditioning

The primary critic averages two separately trained count models: a Poisson model and a more flexible count distribution.

On the final cohort, the Poisson hindsight model reaches log loss 0.4700, the flexible model 0.4724, and their fixed mixture 0.3943. On validation, the corresponding values are 0.4878, 0.5003, and 0.4061. The mixture therefore gives the best proper-loss values among these three fixed readouts on both frozen cohorts.

This does not imply that the mixture is uniformly better for every downstream ranking. For example, the Poisson model finds 151 wrong inherited labels at the 20% final review budget, compared with 149 for the mixture. Proper-loss accuracy and top-K ranking are different objectives.

We also evaluated readouts that condition the count prediction on the observed final grade. These comparisons keep the underlying count models fixed and apply grade information only at prediction time.

The effects are mixed. On the final cohort, conditioning improves the Poisson hindsight log loss by 0.01035, but the change for the primary mixture is only -0.00162, with an interval that includes zero. Conditioning worsens the flexible matched-static model by 0.01523. We therefore retain the simpler grade-blind local critic as the primary method rather than claiming that final-grade conditioning is uniformly helpful.

Calibration shows the same distinction between different notions of predictive quality. On the final cohort, the primary critic has lower log loss and Brier score than the matched static critic, but its pooled local expected calibration error is higher. Better proper loss therefore does not imply better calibration under every calibration summary.

Tables D.2 and D.3 report the conditioning results.

Conditioning rule. The 18 supporting readouts are Poisson/flexible/half, each at H/latest/prefix, each with normalized blind and grade-conditioned variants. Both grade branches are formed from public claims and fixed count distributions before observed grades or references are accessed. For a target law, the other three terminal components have a fixed sum prior $\mathrm{Pois}(3\mu)$ with $\mu = 2/3$, obtained from the original source subtotals (sum 512 across 768 targets). This extrapolates to the unobserved first block N_0, which is not supervised by the target subtotal. It is an assumption of this diagnostic, not a learned simulator law.

The entire half-mixture target law is conditioned before posterior normalization. Averaging two separately conditioned heads would be a different readout. Flexible normalization includes the full geometric tail. Both grade events have positive modeled mass; the execution cap for parsed public integers is 2048, with an unsupported branch failing the phase rather than being clipped or silently excluded. Unknown terminal claims or grades use the prespecified blind fallback. The six grade-conditioned static readouts are required controls; H companions remain supporting comparisons.

Table D.2: Validation final-grade conditioning: conditioned minus normalized blind loss. Negative values favor conditioning. Intervals are paired pointwise 95%.

Law/view	Δ log loss[interval]	Δ Brier[interval]
Poisson H	$-.0199[-.0255, -.0141]$	$-.0075[-.0099, -.0050]$
Poisson latest	$-.0086[-.0160, -.0013]$	$-.0034[-.0068, -.0001]$
Poisson prefix	$-.0225[-.0303, -.0150]$	$-.0072[-.0106, -.0042]$
Flexible H	$.0056[-.0006, .0112]$	$.0014[-.0008, .0034]$
Flexible latest	$.0150[.0067, .0230]$	$.0039[.0011, .0069]$
Flexible prefix	$-.0003[-.0076, .0067]$	$.0009[-.0012, .0029]$
Half H	$-.0055[-.0105, -.0005]$	$-.0023[-.0045, -.0003]$
Half latest	$.0031[-.0042, .0103]$	$.0017[-.0014, .0047]$
Half prefix	$-.0094[-.0160, -.0030]$	$-.0018[-.0044, .0007]$

Table D.3: Final-cohort grade conditioning: conditioned minus normalized blind loss. Negative values favor conditioning. Intervals are paired pointwise 95%.

Target law/view	Δ log loss [95% interval]	Δ Brier [95% interval]
poisson / H256	-0.01035 [-0.01605, -0.00434]	-0.00356 [-0.00592, -0.00109]
poisson / latest_S256	0.00094 [-0.00526, 0.00712]	0.00054 [-0.00228, 0.00328]
poisson / prefix_S256	-0.00983 [-0.01661, -0.00329]	-0.00186 [-0.00430, 0.00059]
flexible / H256	0.00635 [0.00095, 0.01135]	0.00215 [0.00005, 0.00405]
flexible / latest_S256	0.01523 [0.00793, 0.02211]	0.00412 [0.00139, 0.00674]
flexible / prefix_S256	0.00668 [0.00129, 0.01258]	0.00160 [-0.00007, 0.00321]
half / H256	-0.00162 [-0.00609, 0.00292]	-0.00029 [-0.00218, 0.00166]
half / latest_S256	0.00796 [0.00178, 0.01383]	0.00341 [0.00072, 0.00595]
half / prefix_S256	-0.00091 [-0.00647, 0.00468]	0.00110 [-0.00106, 0.00335]

D.4 Subgroups and Training Seeds

The main results average over all local forecasts. We also retain prespecified descriptive results across target position, reference count, and final-grade outcome.

The most important reversal occurs between zero-count and positive-count intervals. On the final cohort, the primary critic is much better than residual MSE on positive-count intervals:

$$0.5275/0.1744 \quad \text{vs.} \quad 0.7890/0.2869$$

in log loss/Brier score.

For zero-count intervals, however, residual MSE is better:

$$0.1734/0.0386 \quad \text{vs.} \quad 0.2685/0.0759.$$

These subgroup results are descriptive and do not define a new model-selection rule.

We also trained the frozen development recipe with two additional training seeds before validation. Across all three seeds, the hindsight model with aggregate and sparse supervision consistently improves over its own aggregate-only and matched static controls. The mixture's improvement over the flexible constituent is not statistically conclusive for every seed.

These checks support the two main comparisons—sparse local supervision and hindsight—without establishing that one particular constituent or mixture is universally optimal. Tables D.4 and D.5 report the subgroups; Table D.6 reports the seed results.

Table D.4: Validation fixed subgroups: log loss/Brier score. Errors are inherited-label disagreements. Static uses the preceding interval. Rows overlap.

Scope	Slots/groups	Errors	Primary	Static	MSE
All	768/256	399	.4061/.1278	.6409/.2252	.5129/.1758
Count zero	367/236	263	.2665/.0741	.6666/.2410	.1795/.0404
Count positive	401/245	136	.5354/.1778	.6148/.2096	.7961/.2899
Terminal failed	639/213	350	.4013/.1256	.6355/.2233	.5036/.1713
Terminal successful	129/43	49	.4298/.1386	.6675/.2346	.5591/.1979
Target 1	256/256	98	.4647/.1494	.6290/.2195	.5961/.2066
Target 2	256/256	139	.4073/.1285	.6786/.2431	.5532/.1937
Target 3	256/256	162	.3462/.1054	.6151/.2130	.3894/.1270

Table D.5: Final fixed subgroups: log loss/Brier score. Errors are inherited-label disagreements. Rows overlap.

Scope	Slots	Groups	Errors	Primary NLL/Brier	Latest static	MSE
all768	768	256	401	0.3943/0.1235	0.5781/0.1987	0.4904/0.1669
target1	256	256	110	0.4211/0.1320	0.5583/0.1894	0.5739/0.1988
target2	256	256	130	0.4468/0.1458	0.6711/0.2388	0.5693/0.2007
target3	256	256	161	0.3151/0.0928	0.5048/0.1679	0.3281/0.1014
count_zero	372	239	251	0.2685/0.0759	0.6063/0.2131	0.1734/0.0386
count_positive	396	244	150	0.5275/0.1744	0.5595/0.1886	0.7890/0.2869
L_negative	651	217	338	0.4019/0.1263	0.5805/0.1999	0.4922/0.1671
L_positive	117	39	63	0.3520/0.1080	0.5647/0.1922	0.4807/0.1663

Table D.6: Existing development proper losses across fixed seeds. Values are log loss/Brier; seed labels abbreviate 2026092202–2026092204. No seed selection or pooling is performed.

Readout	2202	2203	2204
Half H +77	.3996/.1244	.3641/.1113	.3440/.1068
Half H aggregate-only	.5650/.1938	.4756/.1558	.5427/.1855
Half latest +77	.6175/.2153	.6077/.2095	.5815/.1985
Half prefix +77	.6531/.2181	.6069/.2060	.5967/.2013
Poisson H +77	.4655/.1518	.4716/.1576	.4269/.1424
Flexible H +77	.5089/.1693	.3750/.1142	.3670/.1159

E Fixed-Budget Review Results

The main paper reports the 20% review result as an additional use of the local critic. This appendix gives the complete results at all three prespecified review budgets.

For each local claim i, we rank the probability q_i that its inherited label is wrong. At budget K, the K highest-ranked claims are reviewed. The budgets are fixed at

$$K \in \{39, 77, 154\},$$

corresponding to 5%, 10%, and 20% of the 768 local claims in each cohort.

The **oracle** ranks claims using their known inherited-label errors and therefore gives the maximum possible number of discoveries at each budget. We also report a fixed hash-random ordering as a reference. These baselines do not affect model selection.

E.1 Validation Results

The validation set contains 399 wrong inherited labels among 768 local claims. Table 7 reports the review results.

Table 7: Complete fixed-budget results on the frozen validation cohort.

Method	5% (39)	10% (77)	20% (154)
Hindsight critic	39	77	149
Latest static	32	63	116
Prefix static	32	62	123
Poisson hindsight	39	77	152
Flexible hindsight	38	72	138
Fixed hash-random	22	43	87
Oracle	39	77	154

The hindsight critic reaches the oracle count at the 5% and 10% budgets. At 20%, its paired gain over the matched latest-static critic is

$$33\ [23, 43].$$

The corresponding gains at 5% and 10% are

$$7\ [2, 12] \qquad \text{and} \qquad 14\ [7, 21].$$

These results were part of the frozen validation evaluation. They were not used to modify the ranking rule before final evaluation.

E.2 Final Results

The final set contains 401 wrong inherited labels among 768 local claims. Figure 3 reports the review results.

At the 5% and 10% budgets, the hindsight critic finds 39 and 77 wrong labels, while the matched latest-static critic finds 38 and 73. The paired gains are

$$1\ [0, 3] \qquad \text{and} \qquad 4\ [-1, 7.03].$$

At the 20% budget, the hindsight critic finds 149 wrong labels, compared with 136 for the matched latest-static critic. The paired gain is

$$13\ [4.98, 22].$$

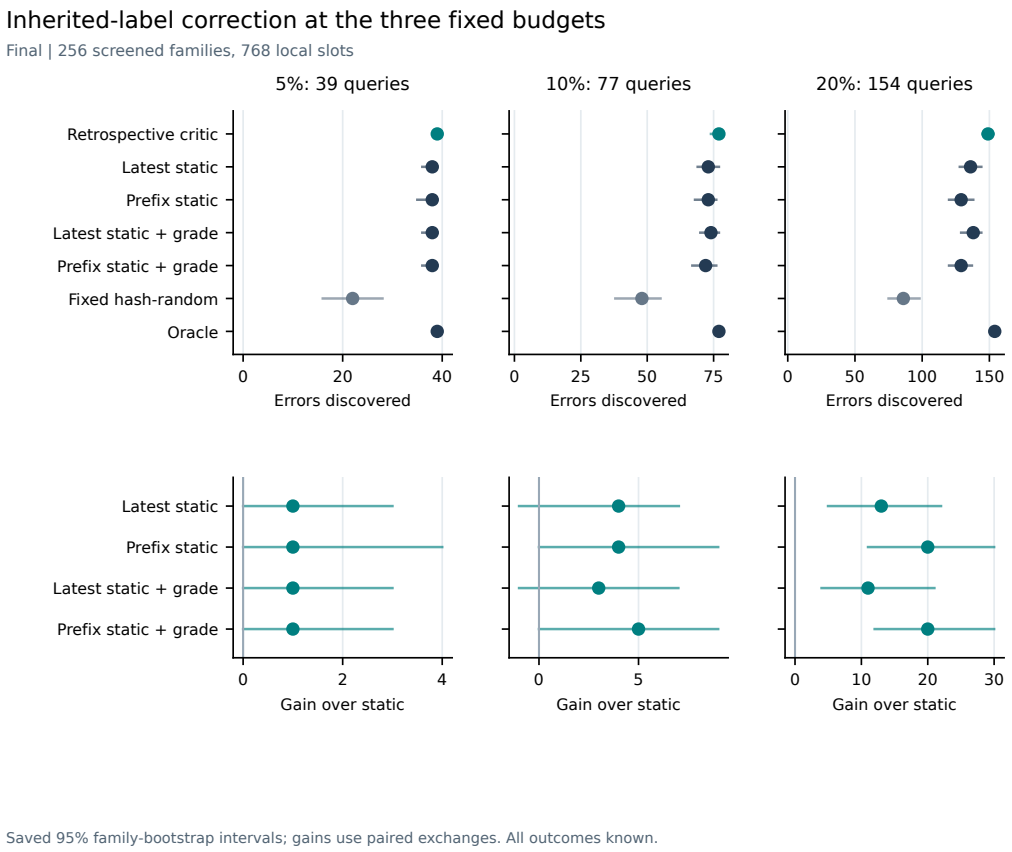

Figure 3: Fixed-budget review on the final cohort. The top panels show the number of wrong inherited labels found at the three prespecified review budgets. The bottom panels show paired gains of the hindsight critic over the static controls. All labels are known when evaluating this figure; the ranking rules were fixed before final outcomes were examined.

E.3 Prediction Accuracy and Ranking Are Different Objectives

Better probability estimates do not necessarily produce a better ordering at every review budget. This distinction is reflected both in the theory of Appendix A.4 and in the empirical results.

For example, the Poisson hindsight model finds 151 wrong labels at the 20% final budget, while the primary mixture finds slightly fewer, even though the mixture has substantially lower log loss. At the smaller budgets, the primary ranking is already saturated at the oracle count.

The validation and final cohorts also differ in the strength of the static ranking. The validation gain at 20% is $+33$, whereas the final gain is $+13$. We report the two cohorts separately rather than combining them.

F Additional Prospective Confirmation

The original validation and final experiments use the frozen training schedule reported in the main paper. A later development experiment suggested that another round of training could further improve the hindsight critic. Rather than replacing the frozen primary model, we evaluated this longer schedule in a separate prospective confirmation.

F.1 Confirmation Design

The confirmation adds 720 training updates after the original two-stage schedule. The optimizer is restarted and the data stream begins again. The matched static and aggregate-only controls receive the same additional training.

All model states, readouts, mixture weights, random seeds, review budgets, and evaluation code were fixed before the new cohort was selected. The confirmation cohort contains 256 scene groups and 768 local forecasts. Its reference labels were joined only after the frozen evaluation pipeline had passed the required checks.

This experiment is reported separately. It does not replace or pool the original validation and final cohorts.

A fixed 272-candidate pool was screened jointly with all 1,588 prior public exposures, including unused bridge candidates, using the unchanged transitive exact/RGB/mean-dHash≤ 4 rule. The 1,860-node graph checked all 1,728,870 pairs. All 272 candidates were eligible; the first 256 representatives in public order were selected without refill. This is heuristic family screening, not certified physical independence or uniform sampling. The ten saved states, 13 readouts, mixture weight, seed, budgets and evaluator preceded membership. References were joined only after public-program checks.

The 12 prespecified conditions comprise eight material comparisons against the original endpoint and three matched controls, plus NLL≤ 0.65, Brier≤ 0.23, incorrectness AUC≥ 0.75, and known support≥ 730. Material conditions require point improvements of 0.05 NLL/0.02 Brier and strictly negative paired upper limits; they do not require the intervals to exclude those margins. There is no review gate.

F.2 Predictive Results

The additional training improves the hindsight critic on the new cohort. Table 8 summarizes the main results; Table 9 retains all frozen readouts.

Table 8: Main predictive results on the separate prospective confirmation cohort.

Method	Log loss	Brier
Original hindsight model	0.4186	0.1313
Continued hindsight model	0.3519	0.1043
Continued latest-static	0.5846	0.1981
Continued prefix-static	0.6661	0.2126
Continued aggregate-only hindsight	0.4899	0.1610

Relative to the original hindsight model, the continued model improves log loss by

$$0.0667\ [0.0456, 0.0885]$$

and Brier score by

$$0.0270\ [0.0194, 0.0350].$$

It also outperforms each equally continued static and aggregate-only control. The confirmation passes all 12 prespecified conditions.

The flexible count model has better point loss than the Poisson model, and the fixed mixture is close to the flexible constituent. The paired difference between the mixture and the flexible model is

uncertain. This confirmation therefore supports the supervision and hindsight comparisons rather than superiority of a particular constituent.

Table 9: All 13 fixed readouts on the separate schedule cohort. Brackets are saved 95% fixed-model, paired family-bootstrap intervals (2,000 draws). "Matched" denotes the same added schedule; all mixtures use fixed weight 0.5. The original endpoint is evaluated on this cohort, not copied from final.

Readout	Log loss [95% interval]	Brier [95% interval]
Original H mixture +77	0.4186 [0.3898, 0.4491]	0.1313 [0.1199, 0.1433]
Continued Poisson H +77	0.4921 [0.4375, 0.5501]	0.1538 [0.1381, 0.1713]
Continued flexible H +77	0.3632 [0.3067, 0.4226]	0.1065 [0.0886, 0.1245]
Matched Poisson latest +77	0.5965 [0.5592, 0.6383]	0.2035 [0.1888, 0.2194]
Matched flexible latest +77	0.8108 [0.7397, 0.8891]	0.2599 [0.2373, 0.2826]
Matched Poisson prefix +77	0.7031 [0.6397, 0.7735]	0.2216 [0.2020, 0.2425]
Matched flexible prefix +77	0.9570 [0.8675, 1.0502]	0.2885 [0.2633, 0.3150]
Matched Poisson aggregate-only H	0.5333 [0.5044, 0.5662]	0.1813 [0.1687, 0.1954]
Matched flexible aggregate-only H	0.7293 [0.6424, 0.8220]	0.2132 [0.1885, 0.2400]
Continued H mixture +77	0.3519 [0.3142, 0.3930]	0.1043 [0.0909, 0.1184]
Matched latest mixture +77	0.5846 [0.5462, 0.6257]	0.1981 [0.1841, 0.2124]
Matched prefix mixture +77	0.6661 [0.6059, 0.7272]	0.2126 [0.1939, 0.2316]
Matched aggregate-only H mixture	0.4899 [0.4590, 0.5232]	0.1610 [0.1484, 0.1744]

F.3 Fixed-Budget Review

Table 10 reports all confirmation review results. The continued model also retains the fixed-budget ranking improvement over the matched latest-static critic. At the 20% budget, the paired gain is

$$11\ [4, 22].$$

At the 5% and 10% budgets, the gains are

$$3\ [-2, 7] \qquad \text{and} \qquad 4\ [-1, 10].$$

The additional training clearly improves predictive loss, but it does not establish a corresponding ranking improvement over the original hindsight endpoint or over either individual count model. This is consistent with the distinction between probability accuracy and top-K ranking discussed in Appendix A.4.

Table 10: All schedule-confirmation discovery points. Budgets are fixed at 39/77/154 (5/10/20%); no selection rule was changed after evaluation.

Readout	39	77	154
Original H mixture +77	39	75	144
Continued Poisson H +77	38	74	147
Continued flexible H +77	39	74	146
Matched Poisson latest +77	36	70	135
Matched flexible latest +77	38	72	122
Matched Poisson prefix +77	35	68	134
Matched flexible prefix +77	31	65	122
Matched Poisson aggregate-only H	38	76	146
Matched flexible aggregate-only H	32	57	117
Continued H mixture +77	38	74	147
Matched latest mixture +77	35	70	136
Matched prefix mixture +77	33	67	127
Matched aggregate-only H mixture	38	73	140
Fixed random hash	13	30	70
Oracle	39	77	154
Uniform-random expectation	19.5	38.5	77.0