

Human as a Knowledge Augmenter in Human–AI Interaction

Yujin Cha

chayj@kaist.ac.kr

Korea Advanced Institute of Science and Technology

Daejeon, Republic of Korea

Abstract

Humans exhibit greater flexibility in learning than machine learning models and are known to achieve robust generalization even from a small number of training examples. This ability underscores their efficiency under data-scarce conditions. In the context of mutual learning between humans and machine learning models, we hypothesize that humans can generalize from a limited set of samples provided by the model and, without external information, transfer their acquired knowledge back to the model, thereby further improving model performance. We define this as the knowledge augmentation effect, which, despite its theoretical plausibility, has not been empirically validated. We design a human–AI interaction framework grounded in a canonical active learning paradigm to empirically examine the hypothesis. Specifically, we incorporate a feedback mechanism into the active learning loop in which the model’s knowledge is selectively provided to human oracles. Experimental results demonstrate that such feedback improves the labeling accuracy of the oracle and, in turn, substantially improves model performance. Furthermore, through human–AI interaction experiments involving 127 clinicians using real-world medical imaging datasets, we provide empirical evidence of the knowledge augmentation effect and demonstrate the practical significance of integrating human–AI co-learning within expert-in-the-loop systems.

Keywords

Active learning, Human-in-the-loop, Noisy oracle, Uncertainty

1 Introduction

Recent advances in machine learning have surpassed human-level performance across many tasks, but such achievements largely rely on massive training datasets. In contrast, cognitive science research [5] shows that humans can generalize from sparse data and generate novel inferences, demonstrating a flexible learning system that achieves robust generalization even from only a few samples. This suggests that, under the same data constraints, humans may achieve higher levels of generalization than machine learning models. We refer to this phenomenon—where human learning provides additional generalization beyond that of models trained on the same data scale—as the knowledge augmentation effect. While this property of human learning is well recognized, its quantitative impact has not been rigorously studied. Under the assumption of the knowledge augmentation effect, if humans and models iteratively exchange knowledge in a learning loop, a human trained on limited model-provided data may recursively feed back generalized knowledge to the model, thereby progressively improving model performance without external information. This can be tested by extending the standard active learning loop [2, 7] with a reverse

learning phase, where the model’s training data are provided back to the oracle for learning. We define this extended loop as bidirectional active learning (BAL). Beyond verifying the knowledge augmentation effect, BAL also provides an opportunity to improve noisy oracles [1, 6], thereby offering a pathway to enhance active learning performance in challenging domains such as medical image interpretation. In this work, we quantify the knowledge augmentation effect in BAL and investigate its utility for mitigating the limitations of noisy oracles. To this end, we designed a BAL framework and conducted a large-scale study with 127 physicians performing medical image labeling tasks. Models retrained on annotations produced by trained oracles significantly outperformed those trained on labels from untrained oracles. In summary, this work reframes the oracle as an adaptive learner within the active learning loop and provides empirical evidence that human–model co-learning can enhance system performance through a recursive knowledge augmentation effect.

2 Materials and methods

This study builds upon the conventional active learning framework by introducing a **bidirectional structure** in which the model and the oracle engage in mutual learning. The dataset is defined as $D = \{X, Y\}$, where X denotes the set of input images and Y represents corresponding binary labels. The primary model W selects a subset of unlabeled samples $x_U \subset X_U$ that are considered informative from the model’s perspective and queries the oracle O for their labels y_T , which are then used for training. In each round r , the newly labeled pairs $\{x_U, y_T\}$ are added to the training dataset D_T , and the model is retrained from randomly initialized weights using the updated dataset. The performance of both the model W and the oracle O is evaluated using a separate, fixed evaluation dataset D_E . A key feature of this framework is that the training data D_T serve as a source not only for model training but also for *oracle learning*. To manage this dual usage, D_T is partitioned into two subsets:

- D_T^O : Samples newly labeled by the current oracle
- D_T^* : Samples either present initially or labeled by prior oracles

To isolate the effect of *model-to-oracle knowledge transfer* and exclude any confounding self-reinforcement effects, the oracle in this study is trained **exclusively on D_T^*** . This design ensures that improvements in the oracle’s performance are attributable solely to information previously acquired by the model.

2.1 The hypothesis of knowledge augmentation effect

The knowledge augmentation effect is evidenced not only by improvements in the oracle’s own learning but by measurable gains in model performance arising from interaction with the oracle. We

define augmentation as a statistically significant increase in model accuracy when the model receives feedback from an oracle that has been trained on its prior outputs. The baseline condition is the performance gain obtained when interacting with an untrained oracle under otherwise identical active learning settings. Throughout, we assume that the oracle is noisy yet generally reliable, adapts based on queries, and does not behave adversarially. In this section, we formalize the following assumptions. Let the baseline performance of the model be defined as

$$s_W^{\text{base}} = \text{Perf}(W \mid D_T \cup Q(W, O; X_U)). \quad (1)$$

Here, $Q(W, O; X_U)$ denotes the model W 's selection of informative samples from the unlabeled pool X_U and the corresponding queries to oracle O . Consider an alternative scenario in which oracle O is explicitly trained on the X_T data utilized by model W . This updated oracle is denoted as

$$O' = \text{Learn}(O \mid X_T),$$

and the resulting model performance as

$$s_W^{\text{amp}} = \text{Perf}(W \mid D_T \cup Q(W, O'; X_U)). \quad (2)$$

Our null hypothesis is

$$H_0 : s_W^{\text{amp}} \leq s_W^{\text{base}}. \quad (3)$$

This hypothesis asserts that oracle learning solely from X_T does not yield additional performance gains for model W , as no new knowledge is introduced beyond the model's existing capabilities. To support this, we present a set-theoretic framework. Let

K_W : knowledge acquired by model W from X_T ,

K_O : prior knowledge of oracle O ,

$g : K_W \rightarrow K'_O$ a mapping from K_W
to the subset learnable by the oracle

Thus, the updated oracle knowledge is

$$K'_O = K_O \cup g(K_W), \quad g(K_W) \subseteq K_W.$$

In other words, the oracle can only assimilate a projection of the model's knowledge that resides entirely within the model's knowledge space. Consequently, the labels provided by the trained oracle O' satisfy

$$\forall x \in X_U, \quad Q(W, O'; x) \in \text{span}(K_W).$$

Therefore, no new information is incorporated into the model beyond its current knowledge boundaries, and

$$\Delta s = s_W^{\text{amp}} - s_W^{\text{base}} \approx 0 \quad (4)$$

is expected to hold true. However, experimental results indicated that

$$s_W^{\text{amp}} > s_W^{\text{base}} + \epsilon, \quad \text{for some noise } \epsilon > 0. \quad (5)$$

In the case of a statistically significant difference, the oracle has not simply replicated the knowledge of the model, but instead has internalized and generalized it in a manner that enhances its subsequent labeling performance. These findings provide empirical evidence of the knowledge-augmentation effect resulting from human learning.

3 Experiments

3.1 Dataset and Participants

This study employed chest X-ray (CXR) images randomly sampled from the CheXpert 1.0 [3] and MIMIC-CXR 1.0 [4] datasets as experimental data. Two distinct datasets reflecting different pathological classifications were constructed: CXR-A (Normal vs. Abnormal) and CXR-B (Edema vs. Pneumonia). Participants were recruited from licensed medical doctors without a history of psychiatric illness, with radiology specialists excluded. A total of 127 participants were included in the final analysis (Table 1). Each participant was assigned to either the CXR-A or CXR-B experimental group.

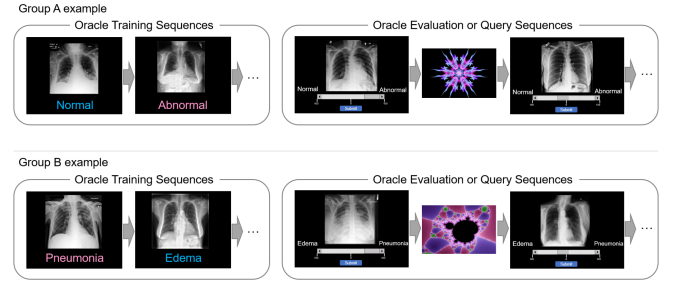


Figure 1: Illustration of the human experiment interface across different datasets

Table 1: Characteristics of participants by experimental group and subgroup

	Characteristics	Learning Group	Control Group
Group A (n=62)	Total participants (Female)	33 (19)	29 (16)
	Age (Years)	29.1±3.0	28.3±2.8
	Clinical experience (Years)	3.0±2.0	2.6±1.4
Group B (n=65)	Total participants (Female)	29 (15)	36 (23)
	Age (Years)	28.7±3.3	29.8±3.7
	Clinical experience (Years)	2.7±2.0	3.2±2.0

3.2 Experimental Design and Procedure

Participants engaged in a BAL task via a custom computer interface (Fig. 1). All participants completed four consecutive main experimental rounds. The model underwent five rounds in total, including Round 0, which served as a pretraining phase without oracle intervention. During all rounds, access to external information and communication between participants were strictly prohibited. In each round, both the model and the human oracle (participant) performed their respective learning tasks sequentially. Across both CXR-A and CXR-B groups, the datasets used for training history (D_T^*), evaluation (D_E), and unlabeled pool (X_U) consisted of 90, 20, and 290 samples, respectively. The model was trained by minimizing a cross-entropy loss function on the current labeled dataset D_T , while the oracle was trained by observing 7 selected samples

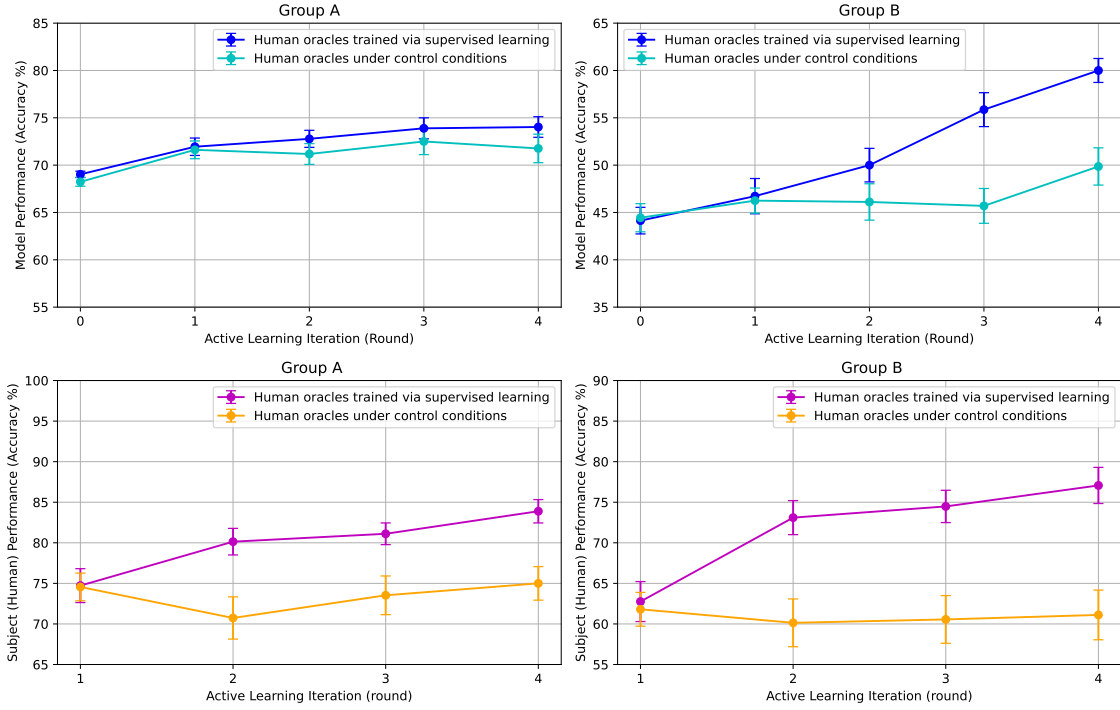


Figure 2: Round-wise performance trajectories of the main models and human participants across experimental subgroups.

Table 2: Statistical comparison of performance across experimental groups in human-ML bidirectional active learning.

Main Group	Learner	Round	t(df)	p-value	Cohen's d	95% CI
Group A	Oracle	Round 1	t(60) = -0.09	0.9276	-0.02	[-0.49, 0.48]
		Round 4	t(60) = 3.03	0.0036	0.77	[0.30, 1.27]
	Model	Round 0	t(60) = 1.34	0.1842	0.34	[-0.08, 0.85]
		Round 4	t(60) = 2.16	0.0349	0.55	[0.11, 0.98]
Group B	Oracle	Round 1	t(63) = 0.30	0.7671	0.07	[-0.41, 0.60]
		Round 4	t(63) = 4.03	0.0002	1.01	[0.58, 1.54]
	Model	Round 0	t(63) = -0.15	0.8835	-0.04	[-0.52, 0.47]
		Round 4	t(63) = 4.10	0.0001	1.02	[0.59, 1.55]

from D_T^* , each displayed for 12 seconds. Participants were randomly assigned to one of two experimental conditions:

- (1) **Experimental condition:** Samples were presented with ground-truth labels for observation.
- (2) **Control condition:** Samples were presented without labels.

Each round consisted of the following three phases:

- **Oracle Evaluation:** Participants provided probabilistic class predictions via a slider interface, allowing evaluation of their accuracy.
- **Oracle Training Phase:** Seven training samples were displayed for observational learning, with or without labels depending on the group.

- **Model Query and Update:** The model randomly selected 30 unlabeled samples from X_U , which were labeled by the participant via the probability slider. These labeled samples were then used to retrain the model.

The classification model $W_{(r)}$ was implemented as a simple CNN architecture accepting 64×64 grayscale input, composed of two convolutional layers followed by two fully connected layers.

3.3 Bidirectional Learning Performance with Human Oracles

In both CXR-A and CXR-B groups, participants in the experimental condition showed significant improvements in classification accuracy across rounds, suggesting that exposure to labeled feedback

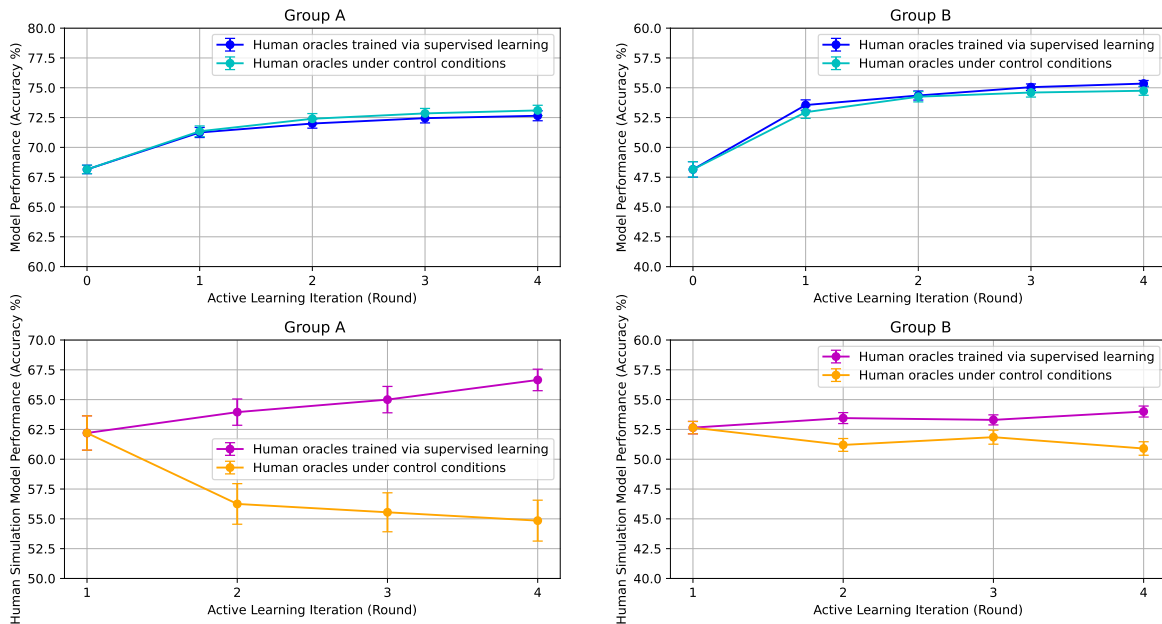


Figure 3: Round-wise performance trajectories of simulated oracle models (and their corresponding main models)

facilitated learning (Table 2). The model’s performance also improved proportionally with the oracle’s learning (Fig. 2). In contrast, participants in the control group showed minimal improvement. At the final round (Round 4), the performance of models trained in the experimental group significantly exceeded that of the control group, which serves as a baseline for expected model gains without oracle learning. These findings provide empirical support for the emergence of a knowledge amplification effect, where recursive oracle learning contributes to enhanced model performance.

3.4 Simulation with Machine Learning-Based Oracles

To isolate the contribution of human learning, we repeated the same experiment using simulated CNN-based oracles. Unlike the human experiment, learning by these artificial oracles did not lead to meaningful improvements in the model’s performance. Although the simulated oracles were capable of learning, this did not translate into amplified model gains across rounds (Fig. 3). No knowledge amplification effect was observed, indicating that the human oracle’s intrinsic learning capability was the critical factor behind the observed performance improvements.

4 Discussion and Conclusion

This work introduced Bidirectional Active Learning (BAL) as a framework to empirically test the knowledge augmentation effect in the absence of external information. We evaluated whether backward feedback within the active learning loop could enhance oracle labeling performance and, in turn, improve model accuracy. If human oracles, after learning from model-provided signals, generate labels surpassing their baseline capacity, the effect constitutes genuine knowledge augmentation rather than mere information

recycling. Beyond serving as an experimental testbed, BAL also provides a practical solution to the oracle noise problem inherent in standard active learning. By simply adding a backward feedback channel, the method achieves scalability and ease of integration. Controlled studies with practicing clinicians confirmed that oracle accuracy improved through model-based learning without external knowledge access, which subsequently yielded significant gains in model performance—clearly outperforming the unidirectional condition. Moreover, the large-scale dataset collected under these constraints holds substantial academic value given the rarity of extended clinical experiments. To further examine whether knowledge augmentation is uniquely human, we conducted a simulation study where the oracle was replaced with a machine-learning model. Although the simulated oracle improved its own accuracy, these gains did not translate into meaningful downstream improvements, underscoring the distinctive contribution of human learning. In sum, this work presents a practical and extensible approach to enhancing oracles through human learning, demonstrating concurrent improvements in both model performance and label quality. While oracle learning is not without limits, our findings suggest that BAL is particularly effective in realistic settings where oracle uncertainty is unavoidable. More broadly, this study establishes the foundation for human-centered active learning frameworks that leverage cognitive learning mechanisms and offers a strategy readily integrable into existing systems.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, RS-2023-00210106) and a Medical Scientist Training Program from the Ministry of Science & ICT of Korea.

References

- [1] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 259–268.
- [2] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*. PMLR, 1183–1192.
- [3] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilicus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.
- [4] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).
- [5] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350, 6266 (2015), 1332–1338.
- [6] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).
- [7] Burr Settles. 2009. Active learning literature survey. (2009).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009