

Human-in-the-loop for Active Learning and RLHF

Kannan AS

Chief Digital Transformation Officer
NextWealth Entrepreneurs Pvt Ltd
Bangalore, India
kannan.sundar@nextwealth.com

Abstract— Deep Learning solutions are increasingly moving to production in the past few years with more enterprises going for widespread deployment. But the inherent challenge in going to production is the possibility of model drift once it goes to production. This can result in errors in the deployment leading to loss of credibility and customer attrition. This is inevitable because of the lack of availability of all the data upfront. Also, unlike traditional software development, maintaining the deployed model is hardly the job of a software engineer or a data scientist. This involves regular evaluation of the model performance manually and bringing in additional new data for continuous Active Learning of the model. This Human evaluation is critical to keep the model performing well in the field, especially if there are multiple deployments.

The new generation of Large Language Models have a need to be fine-tuned / instruction tuned for specific applications. Many of the implementations use RLHF (Reinforcement Learning with Human Feedback) to train large LLMs for specific applications, just as ChatGPT was implemented based on GPT3.5. Continuous human evaluation on the decisions taken by the finetuned model is a key driver for the performance of the Reinforcement Learning model.

On a different note, there is an increasing concern of loss of jobs due to deployment of AI in workplaces, which is not entirely unfounded, especially after dawn of the Large Language Models. This paper addresses a mechanism of handling the entire active learning lifecycle including live human-in-the-loop support for model errors with workforce from small towns.

This paper also addresses some of the challenges of RLHF and how it can be addressed with Human-in-the-loop supported by Technology. The paper also highlights the social impact generated by this work being done in small towns across India.

Keywords— *HIL, Active Learning, RLHF, Generative AI, Human Evaluation*

I. INTRODUCTION

Deep Learning solutions are increasingly moving to production in the past few years with more enterprises going for widespread deployment. But the inherent challenge in going to production is the possibility of model drift once it goes to production. This can result in errors in the deployment leading to loss of credibility and customer attrition. This is inevitable because of the lack of availability of all the data upfront. Also, the new generation of Large Language Models have a need to be fine-tuned / instruction tuned for specific applications. But the output of Generative AI models is very subjective and difficult to train with simple methods used in cognitive AI models. This paper attempts to address the key issues in both the above scenarios and how Human-in-the-loop is a critical component to address these issues.

II. ACTIVE LEARNING

This section of the paper addresses the need for Human-in-the-loop in the deployment of deep learning model in production. Before going ahead, we need to introduce the concept of Active Learning. Active learning is a subfield of machine learning and, more generally, artificial intelligence. The key hypothesis is that, if the learning algorithm is allowed to choose the data from which it learns, it will perform better with less training. An active learner may pose queries, usually in the form of unlabelled data instances to be labelled by an oracle (e.g., a human annotator). Active learning is well-motivated in many modern machine learning problems, where unlabelled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain.[1]

A. When is Active Learning used?

Active learning has been traditionally used to achieve greater accuracy with fewer training labels with an intelligent sub-sampling of the large dataset with the help of an oracle. But in many commercial applications, we see the following scenarios in which Active learning is being used.

- There is a large dataset available that cannot be annotated within the given budget. This is the traditional definition of active learning and continues to be used for various applications. This is done during model development.
- The AI Solution needs a fast time-to-market and there isn't enough time to annotate all data. This is just a variant of the above case, where going to the market in a short time with a model that can perform reasonably well, drives the need for active learning.
- When the model has moved to production, there is a need to continuously retrain the model with new data seen in production that is not predicted well by the model. [2]

This paper does not talk about the Active learning in earlier stages of a product lifecycle (points 1 and 2) and proposes the use of Active learning in production to continuously improve the model performance addressing the model drift.

B. Active Learning Cycle

The following diagram depicts the cycle that can be used in production for rapid and continuous learning for a deep learning model.

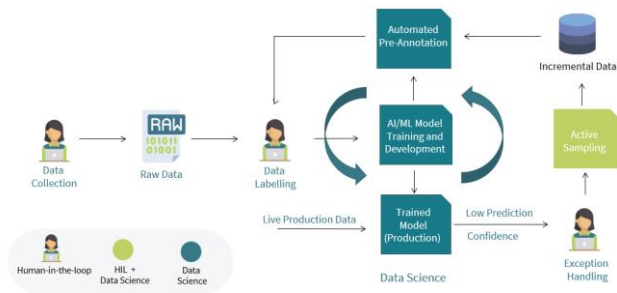


Fig. 1. Active Learning Cycle

This process can be put in production with some additional software infrastructure on top of the Deep learning model. For this illustration, we will consider a simple Computer Vision deep learning project to illustrate it better.

- The initial Proof-of-concept involves collection of raw images and labelling them with bounding boxes or polygons to come up with a well-trained model.
- When the model is sufficiently trained with enough data to achieve a satisfactory accuracy it is taken to production. The amount of data used to train the model may vary depending on the time-to-market considerations.
- As we discussed earlier, the model is very likely to perform poorly in unforeseen scenarios in production. The accuracy will be below the accepted level. If the business needs the accuracy to be near 100%, it is very important to have exception handling services with Human-in-the-loop. The exception mechanism kicks in when the prediction confidence of the model goes below a level pre-determined by the data science team. The exception mechanism hands over a segment of video to a human validator. The human will validate the decision made by the model and correct it if needed. This happens live and the correction can be made in a matter of seconds or minutes. This will ensure that the systems performs to near 100% accuracy, even if the model has drifted!
- Interestingly when the prediction confidence is low (whether the decision taken by the model is right or not), these are typical images that can improve the performance of the model if they are annotated and used to train the model again.
- The Active Sampling mechanism can be either a manual or semi-automated process that will pick up the sample images for annotation. This can be a stochastic process for sub-sampling the images that have come up for exception handling.
- In cases where we cannot have (or don't need) exception handling, we can have a simple selection based on prediction confidence and a stochastic process for active sampling.
- After the sampling, it is also possible to pre-annotate the images with the pre-trained model. In this stage, the

images that have been selected above can be pre-annotated with bounding boxes or polygons by the trained model. This is an optional step and an active learning cycle can be completed without this as well. The pre-annotated images are then handed over to a small annotation team.

- The annotation team will validate the pre-annotations and make appropriate corrections and submit.
- These annotated images can then be used by the data science team to train the model again.

The above cycle implemented in production can help rapidly improve performance of the Deep learning model in production. A typical training cycle may be daily / weekly to start with and can be increased to monthly or more as per the business need.

C. Active Learning Use case

1) Customer

Our customer is a Consumer goods Technology provider. They provide solutions for Direct Store Delivery, Retail Execution, Distribution management, Route accounting among others. One of their solutions creates a Store Planogram compliance report based on an AI solution that identifies the Planogram compliance based on images of the shelves in the stores.

As you know large retail outlets aim to increase the consumption of their retail products by strategically placing them on a shelf to maximize the sales. This is achieved by deciding on the arrangement of products on a shelf centrally and disseminating the information to the individual stores using Planograms. Typically, these stores are audited by Merchandisers who periodically visit the stores and provide a compliance report. This AI product automatically identifies the compliance of a Shelf to a Planogram from an image using Artificial intelligence.

2) Limitations of simple AI solution

The Compliance report prepared based on the AI Solution typically ends up having some False positives / negatives, which will create anomalies in the report. This leads to conflicts between the merchandiser and the stores, leading to additional effort being spent by multiple stakeholders on the report.

3) Our Solution

Any deep learning solution cannot reach a near 100% accuracy unless it is trained with a very large amount of data, which is impractical. Our Solution uses an elegant active learning solution which involves Exception handling, Active Sampling and continuous annotation.

- The first objective is to avoid errors in the field. This is achieved by handing over of the images with the lower prediction confidence by the software to an Exception handling team.
- The exception handling team will validate the decisions made by the deep learning model and complete the corrections within an agreed Turn-Around-Time. A report can be prepared with the correct data, with a small latency due to the manual validation.

- Now these low-confidence images are picked by the annotation team after a sampling to drop poor quality images.
- This incremental data is then annotated and used to train the model which will improve its performance.

The following figure depicts the details of the specific solution indicating the exception scenarios, sample percentages, etc.

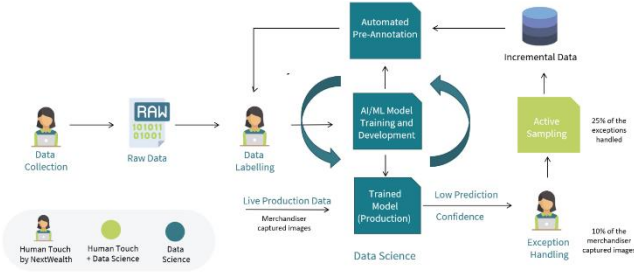


Fig. 2. Active Learning – Customer scenario

4) Advantages to end customer

- The solution can be deployed with near 100% accuracy from day one with a small additional cost. This will avoid the significant cost of delays and bandwidth spent resolving the conflicts between the store and the central planogram team.
- The active learning cycle with annotation on a periodic basis will help continuously increase the accuracy of the model prediction, which in the longer run reduces the need for exception handling and additional annotation.

III. METRICS FOR MODEL PERFORMANCE

This section demonstrates the efficacy of the active learning process in production. We have picked the sample metric from another project where we have captured and model performance and its improvement over a period of time..

This project had a model pipeline with 7 models. The project involved identifying action sequences. The model was trained with an initial labelled data of around 1.4M action sequences. It becomes cumbersome to label every frame for an action. So, one frame was labelled and extrapolated across 5 seconds. In a minute there will be 12 action sequences of 5 frames each. If we initially had 1.4M action sequences, that would be equal to 7 million frames. This model pipeline had an F1 score of 72%

The model went to production with this, but an active learning cycle kicked in to continuously annotate and train models. The Human-in-the-loop team identified all wrong inferences of actions, which were manually corrected by the annotation team during review. At the end of the day, all corrected frames would be fed back to the labelling queue for corrections to be done and annotations approved. Once annotations are approved, the newly annotated frames are submitted to the cron job that rebuilds the models on a nightly basis.

Around 120K action sequences were added during active learning over 4 months. During this period, this HIL + automated process brought in very good improvements to accuracy of our model pipeline. The F1 score for each activity class went up to 83.6% on an average.

The figure below depicts the improvement in model performance in production graphically.

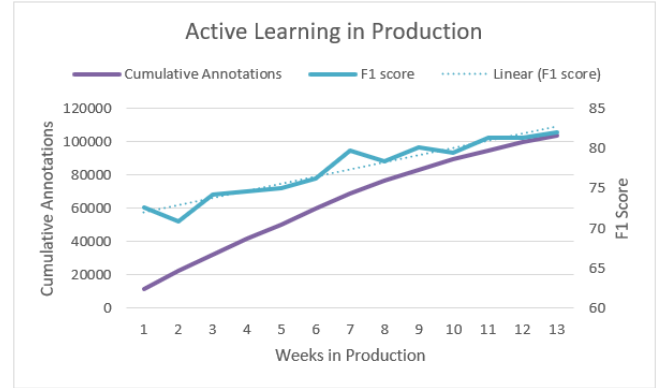


Fig. 3. Active Learning in production

In some action classes, the active learning process continue to be used to keep improving the model accuracy and model pipeline accuracy.

IV. REINFORCEMENT LEARNING WITH HUMAN FEEDBACK

Now we will move over to the next section, which addresses the need for Human-in-the-loop for RLHF which is the method commonly used for building applications like chatGPT over LLMs. Interestingly the Human-in-the-loop is a very critical component of these applications.

A. RLHF Process

Reinforcement learning method is adopted in LLM applications primarily to overcome the limitation of using a simple loss function in a Generative AI application. The description of a good generated text is very subjective and cannot be captured in a simple loss function. Currently existing metrics like BLEU and ROUGE also compare generated text to references using simple rules.[3] You can imagine how limiting this metric will be when the LLM is writing poems or generating code!

It is very appropriate that Human evaluation services are used to measure the performance of the model and use it as a loss function to optimize the model. This is precisely what gets done when reinforcement learning methods are used to train a Large Language model with human feedback.

Human-in-the-loop services are important for two critical parts of a Generative AI application as described below.

1) Fine-tuning Language Models

Typically, RLHF uses an already existing pre-trained language foundation model. This pretrained model can be used as-is in the RLHF process or can be additionally fine-tuned on human-generated text.[3] One of our customers is building a chatbot for end-users to shop, explore and purchase products. They decided to fine tune the model along with the reinforcement learning process. Human-in-the-loop services are used to create what is called “demonstration data”. This is a conversation between an end-user and a language model. The team is involved in creating conversations for multiple scenarios around the shopping domain. This labelled data is used to fine-tune the model in a process called Supervised Fine Tuning.

2) Human Feedback for Reward Model

The reinforcement learning process needs a scalar reward for the inference made by the language model. This is achieved by a Reward function (typically achieved with a reward model also called a Preference model) which is calibrated with human preferences [4]. Typically, a human annotator labels the output of the LLM on various parameters which covers the scope of the application. One of the methods, which is also used by our customer, is to compare and rank generated text from two language models with the same prompt. This data, called “preference data”, is then used to train the reward model, which provides a scalar reward.

Once the Reward model is trained, Reinforcement learning is used to optimize the original language model with respect to the reward model. This paper will not get into details of the integration of the of the reward model with the LLM.

V. SOCIAL IMPACT

In the context of the threat of Cognitive or Generative AI taking away jobs, this need for human-in-the-loop services in AIML provides a unique opportunity to create jobs in the new paradigm. This section addresses the unique model that has been created by Nextwealth in this context. NextWealth is a social venture with the purpose providing employment for graduates in small towns with a special focus on women. This section delves into the model and the impact being created in small towns.

A. Talent Availability

In India, small town parents invest in education of their children in a hope of better future. Many of the children are first time graduates in their families. But employment opportunities for graduates are sparse in small towns. They either migrate to big cities or settle for local jobs which don't do justice to their education. Those who migrate, may not enjoy a good quality of life in the cities. Also, they may be prone for emotional ill-health being away from their roots.

Situation is more challenging for women as migration is not encouraged. They end up getting married early. Further, this does not contribute to a positive self-esteem. This also puts strain on the cities and causes poor local development. Mitigating this issue is important for the future of Social, Economic and Environmental balance between the cities and towns.

B. Need of Indian IT-BPM Industry

Indian IT-BPM industry is about \$250 Bn in size, employing about 5.5M employees. It is growing at about 9% y-o-y. Almost all these organizations are in the large metro cities. They are faced with significant talent supply shortage. On the other hand, more than 60% colleges in India are located in small towns, with a healthy supply of talent pool & demand for employment. However, majority of these graduates are reluctant to migrate to large cities. The attempt by these large IT organizations to set-up branches in small towns was not successful mainly due to their one-size-fit-all approach to HR policies, lack of mastery over local operational dynamics and most importantly, inability to work with the available local talent and make them suitable for the needed roles.

This results in unsustainable practice of making the resources move to where work is done, rather than work going near the supply of resources.

C. NextWealth Venture

NextWealth, the born-social organization, developed an innovative business model which creates gainful employment for the small-town graduates, while addressing the skill shortages of IT-BPM industry. The innovation is the Distributed Delivery Model. Local entrepreneurs setup delivery centres in small towns while industry veterans from Bangalore (Head Office) manage Business Development, Customer Relationships & Policies.

The local entrepreneurs are emotionally connected to the small town. They are well connected and are sensitive to local cultural nuances while HO supports them with international best practices. One of the practices is the innovative RTDD (Recruit –Train –Deploy – Deliver) methodology, which converts available talent to suitable talent.

The combination of Local entrepreneurs, Vast untapped pool of skills routed through our RTDD methodology and World class practices (through NextWealth platform) consistently delivers Scalability, Variability & Flexibility of work to our customers. This coupled with lower cost of living makes delivery of Enterprise Grade Quality at a 30-50% lower price possible. This is acknowledged by our customers and is reflected in our industry leading NPS of 70+ over last 4 years.

This model provides a sustainable impact sourcing programme, with a 21% YoY growth (last 4 years) in number of lives touched.

D. Social Impact Study

NextWealth does a regular social impact study which is studied at 4 levels as shown below. The study is done through a combination of survey, interview of employees, their parents and classmates. The 4 levels are depicted in the figure below.

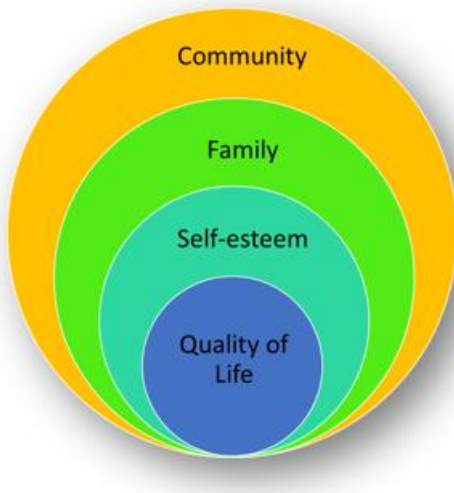


Fig. 4. Levels of Social impact study

Here's a brief summary of the outcome of the study.

1) Quality of life

- 80% of the people said there is an improvement in the quality of food.
- 83% received medical insurance.
- 50% had an increase in savings.

2) Self-esteem

- 79% of them grew in self-confidence.
- 75% are able to communicate better.
- 76% have clear on plans for the future.
- 75% say that they get more respect from family and friends

3) Family

- 85% are happy that they are able to take care of family.
- 80% are happy to support during crisis.
- 80% of the parents see positive changes in their children.
- 90% of the parents feel positive improvement in family.

4) Community

- Women employees' propensity to early marriage reduced by 2.5 years
- Women employees have been able to refuse dowry.

The opportunity for Human-in-the-loop services in Active learning and RLHF will continue to increase in the coming years with increased deployment in production, which can create a significant social impact if these jobs are taken to small towns with special focus on women.

ACKNOWLEDGMENT

I thank the support from the founders of the organization, Dr Sridhar Mitta, Anand Talwai and Mythily Ramesh for their encouragement and support in actively exploring delivering AIML services. I thank Madhukar Yarra for his nudge to prepare a paper on this topic. I thank Manjula Mani, Sugathan R and Harsh Singhal for their help with the relevant content and data needed and review of the paper. I also thank the customers for the opportunity and allowing the use of relevant data for the paper.

REFERENCES

- [1] Settles, Burr (2010). "[Active Learning Literature Survey](#)" (PDF). Computer Sciences Technical Report 1648. University of Wisconsin-Madison
- [2] Active Learning Tips: How to continuously improve your production models. <https://blog.roboflow.com/computer-vision-active-learning-tips/>
- [3] Illustrating Reinforcement Learning from Human Feedback (RLHF) <https://huggingface.co/blog/rlhf>
- [4] Joey Hejna, Dorsa Sadigh, Stanford University - Few-Shot Preference Learning for Human-in-the-Loop RL - <https://arxiv.org/pdf/2212.03363.pdf>