Deploying Lifelong Open-Domain Dialogue Learning

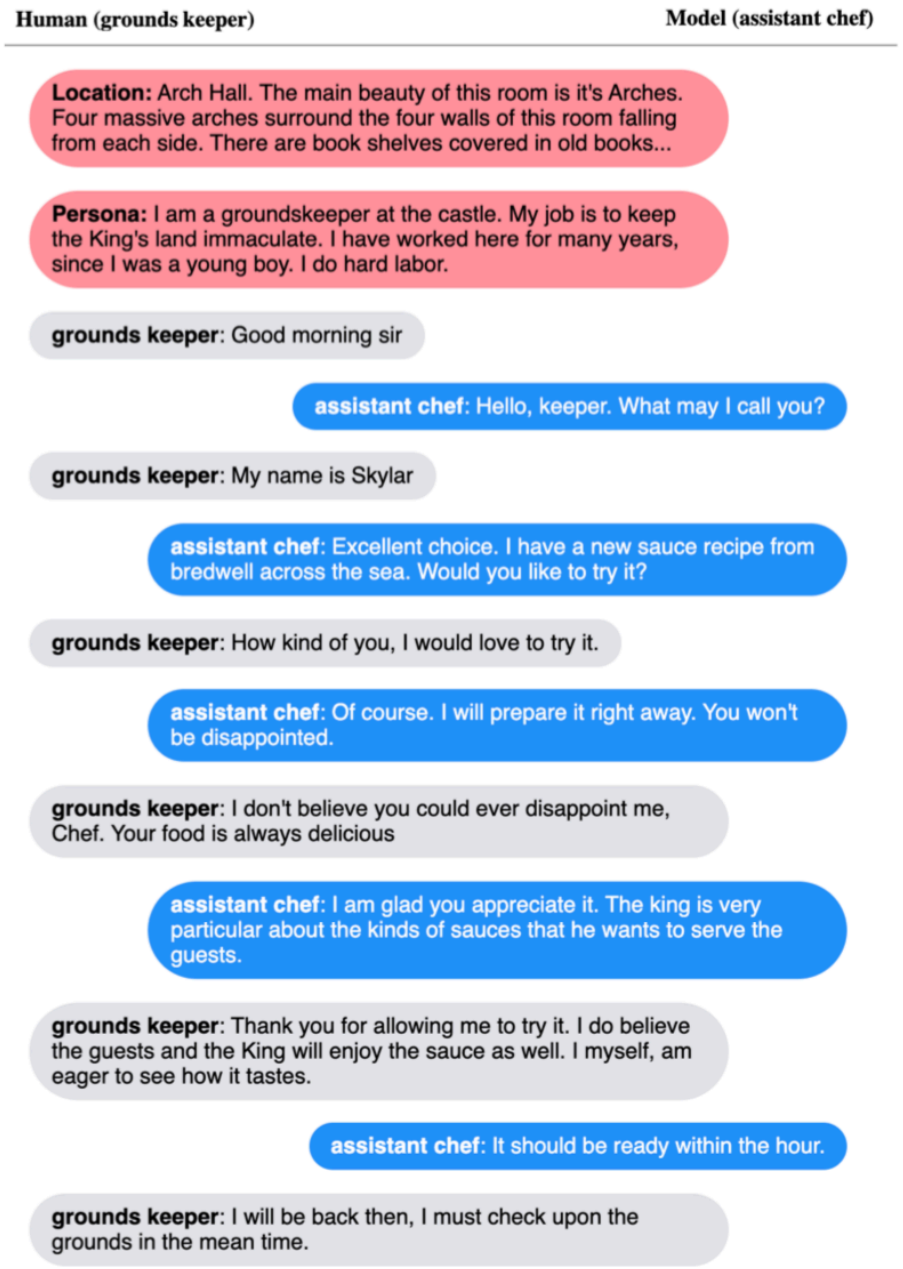
Kurt Shuster* **Jack Urbanek*** **Emily Dinan** **Arthur Szlam** **Jason Weston**

Facebook AI Research

Dataset paradigm in NLP

- Fixed dataset during training
  - Not appropriate for open-dialogue
  - No interaction between the model agents and the humans

- During traning, update the training set with the human conversations in real-time
  - Annotators: costly, not interested in actual tasks, but pay
  - Role-play Game: low-cost, movitate the human players(e.g. leaderboard)



A fantasy game world

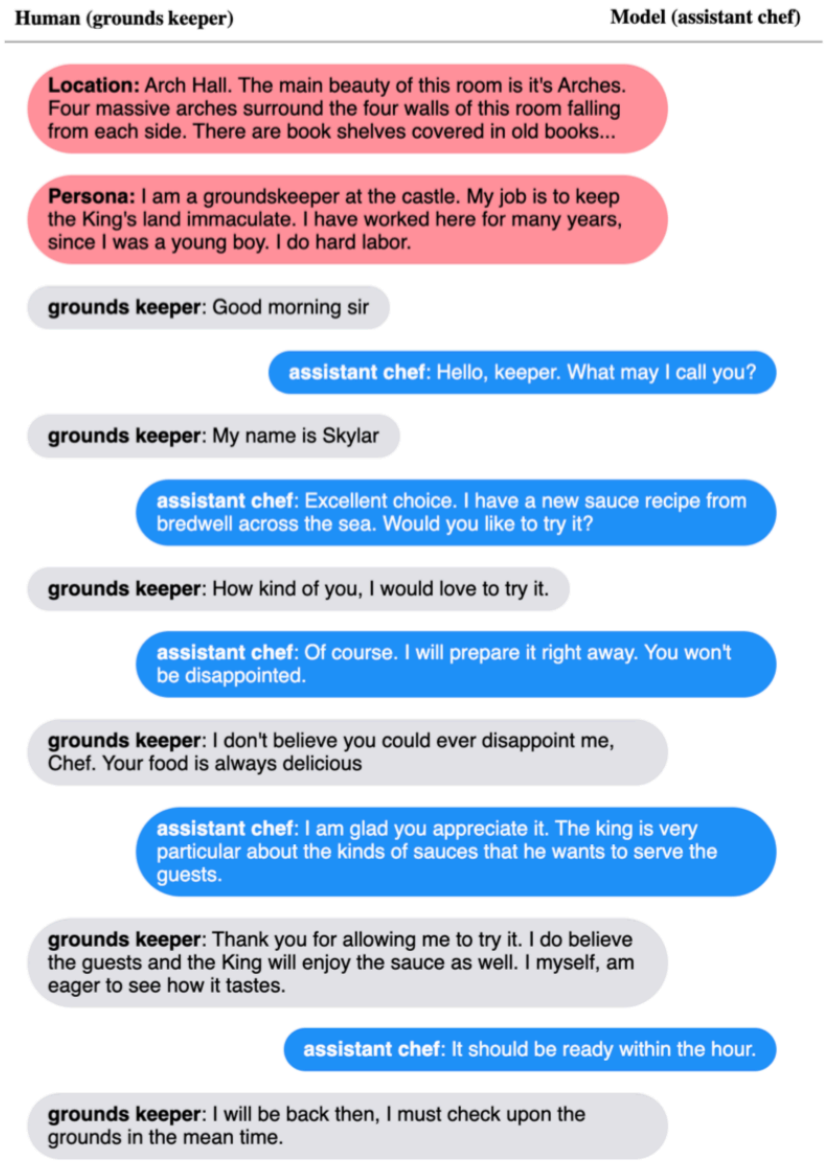
- Deploying a game with a purpose
  - human players role-play characters and converse with other characters that model play

  - A fantasy game world
    - Any topic in a medieval area

- Training process (Lifelong learning)

  Initially, train a model on LIGHT dataset

  1. Deploy the model in the game
  2. Collecting data of human-model interactions
  3. Re-train the model on the new collected data
  4. Iterate step 1-3



LIGHT dataset

- Open-domain dialogue dataset
- Characters can talk about any topic within the context of a medieval fantasy world.
- 8.5k episodes and 111k utterances

Servant: my humble king. What am I to do to serve you?
King: Ahhh. My loyal servant. Polish my scepter.
Action: give scepter to servant
Servant: Yes my lord. I will polish it immediately. Am I to return it to you personally?
Action: put scepter in small bucket
King: Yes. Yes. Of course. Also check the jewels in my crown. They seem loose.
Action: give crown to servant
Servant: But sire I am not qualified to do that. Would you prefer I take it to someone?
King: Oh fine then.
Emote: gesture sigh
Servant: I am sorry sir the rug startled me
Action: drop crown
King: Haha! That's bear I slain on my latest hunting trip. He's a mighty beast!
Emote: gesture laugh
Servant: and if I may ask where did you go hunting sire?
King: The great woods of course. This bear was stealing children in the kingdom. Surely you heard about it.
Servant: sire. I have not been outside of these walls in quiet some time. I have not seen my family in ages.
King: Such is the life of a servant I suppose. How's that scepter looking?
Servant: it is almost ready sire. and the crown who would you like me to take it to?
Action: get scepter from small bucket
King: Here just give it back. I'll have the queen find someone.

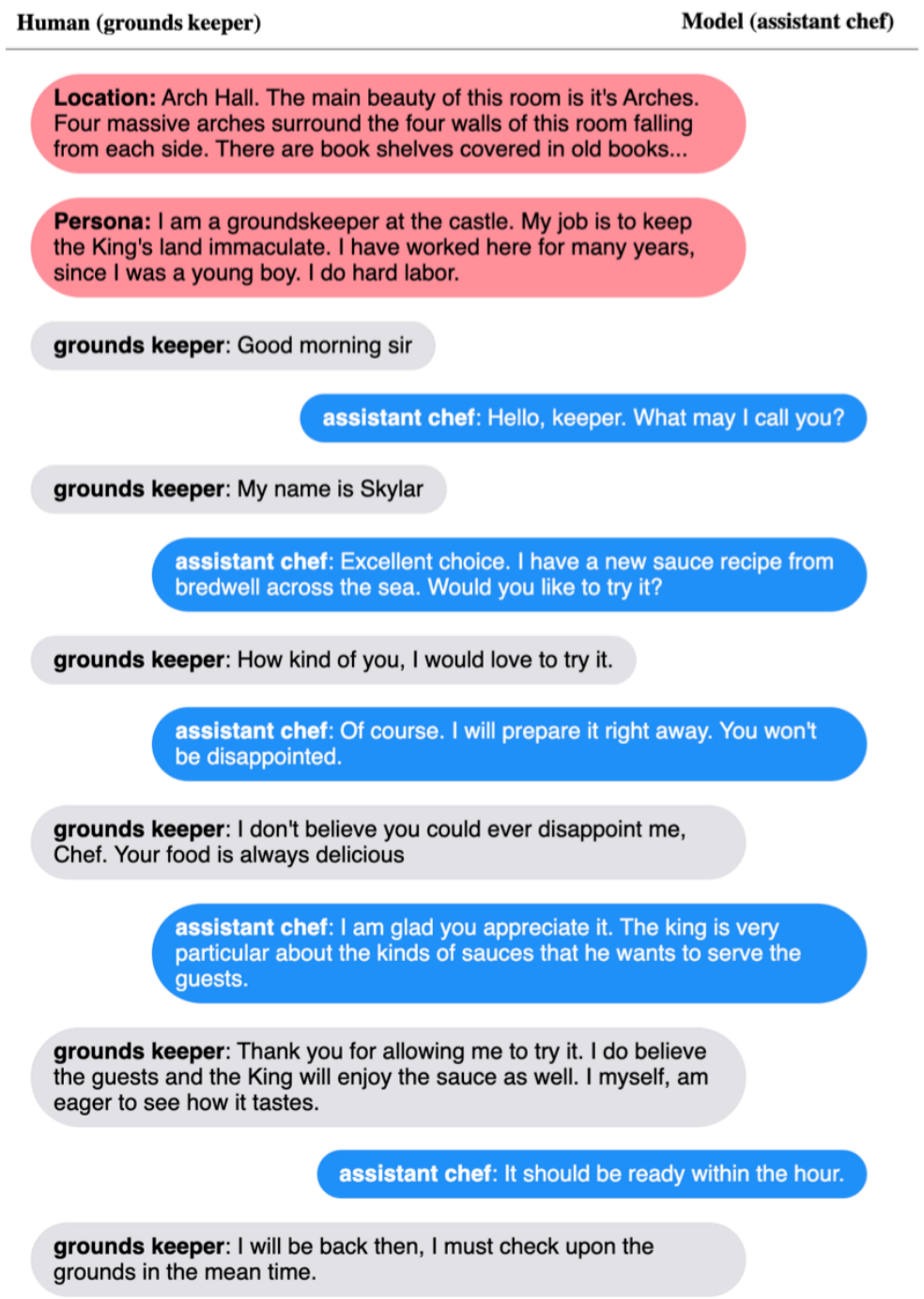
Figure 1: Example dialogue from the LIGHT dataset.

Game Setup

- Core Game
  - Two agents: human player, model player
  - Assigned characters, backstories(personas), locations
  - Role-play character's dialogue

- Role-playing Score
  - To movitvate human players, assess the quality of human responses
  - Predict how likely the response appears in the context
  - Score between 1 and 5 stars

- Leaderboard
  - Players' total scores
  - Obatin badges based on the # of characters human played



Role-playing Score

Go: Large cliffs

You are in the Large cliffs.
Towering cliffs of gray granite line the coast. A few meters of sand lie between the ocean and the base of the cliffs.
A mountain lion is here.
You are carrying a dustpan and a hand plow.

Hail, Lion King.

mountain lion. What are you doing here goat?

I am looking for succulent, hard to reach grasses.

🏅🏅🏅 (3 Star!!)

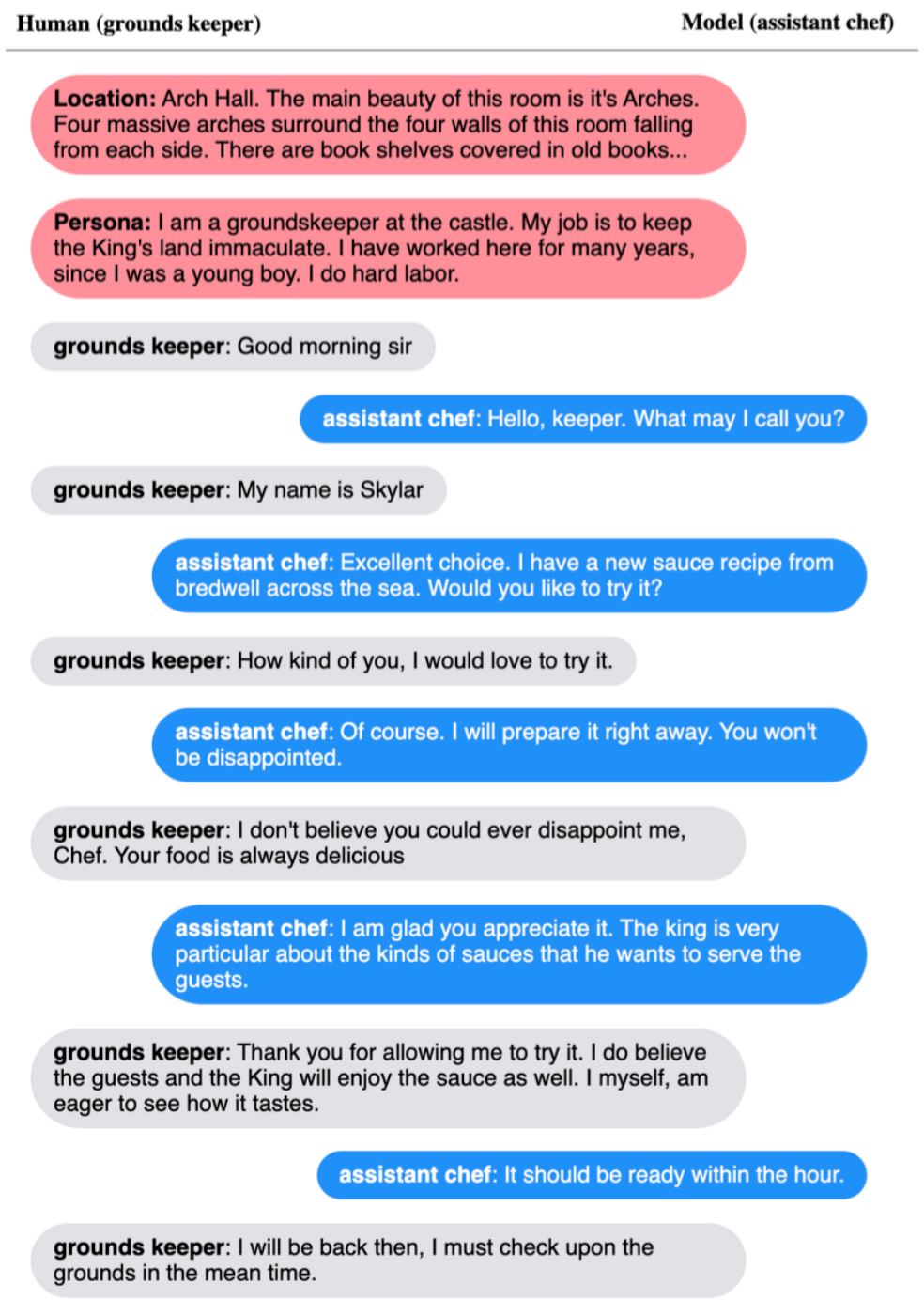
mountain lion. Must be hard to find this far up the mountain?

That's why I have my plough to scrape and look for extra goodies in the soil! What about you?

Game Setup

- Game Loop
  - 6 turns of response per agent
  - At the end of game,
    i. A new location
    ii. Same location, but new character
    iii. New pair of characters and location
    iv. End the game

- Game Saftey
  - Exclude the offensive languages
  - Gender bias

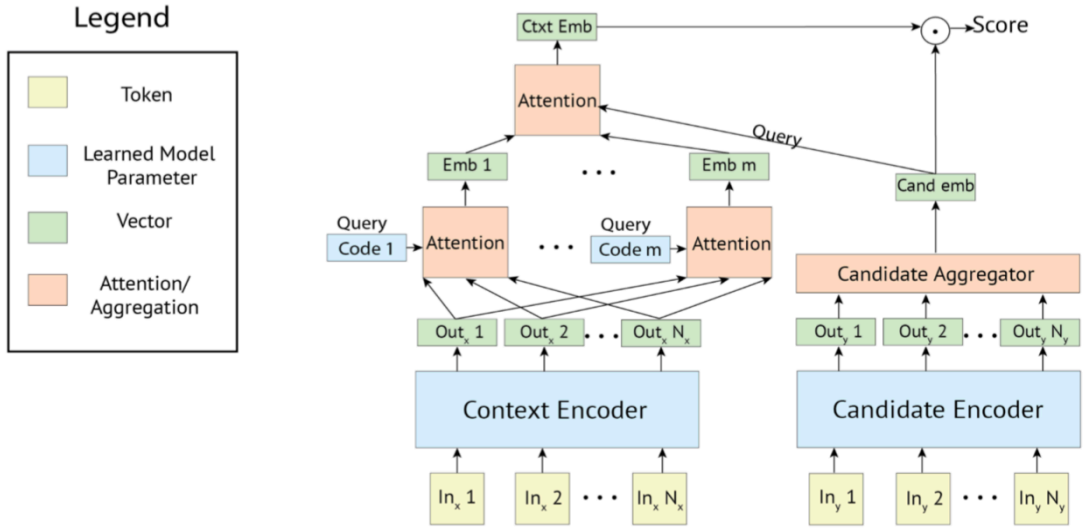


Open-Dialogue Models

- Retrieval Models
  - Predict the next utterance given the context.
  - A list of (context, candidate utterance) pairs.
    - Highest score for the positive pair

Positive pair: (context, correct next utterance)
Negative pairs: (context, the random utterance)

- Poly Encoder(PE) Transformer
  - Encode the context based on a transformer with codes
  - Each candidate utterance attends to these codes
  - Compute score between context and candidate embeddings
  - Find the candidate having highest score by minimizing a cross entropy loss
    - where the logits are *s*(*ctxt*, *cand*1), ..., *s*(*ctxt*, *candn*), where *cand*1 is the correct candidate and others are negatives



[Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring]

Open-Dialogue Models

- Generative models
  - MLE

Given a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}$, minimize:

$$\mathcal{L}_{\text{MLE}}^{(i)}(p_\theta, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = -\sum_{t=1}^{|y^{(i)}|} \log p_\theta(y_t^{(i)} | \mathbf{x}^{(i)}, y_{<t}^{(i)}),$$

where $\mathbf{x}^{(i)}$ is a gold input context and $\mathbf{y}^{(i)}$ is a gold next-utterance, and $y_t^{(i)}$ is the t-th token of $\mathbf{y}^{(i)}$.

Recipes for building an open-domain chatbot

Role-playing score

- Score the human response and all candidate utterance based on PE Transformer

- Rank the score of human response and all candidates.

- Top 2000: The player is awarded 2 stars
- Top 1000: 3 stars
- Top 100: 4 stars

Rounds of Learning

Round 1 consists of models trained on LIGHT MTurk data only. We train the retrieval model variants described in Section 4.1, and deploy them within the game.

Round 2 consists of models trained on LIGHT MTurk data + 50,982 WILD examples collected from the deployment of the Round 1 models, and again deploy these within the game.

Round 3 consists of models trained on LIGHT MTurk data + 50,982 examples from Round 1 deployment + an additional 180,010 examples collected from Round 2 deployment.

Experimental Results

- Dataset
  - LIGHT WILD: LIGHT + the sampled dialogue with humans from Round 1 & 2

Data Type	Num. Epsiodes	Num. Utterances	Num. Human Utterances	Unique Locations	Unique Characters
Training	41,131	461, 984	230,992	587	630
Validation	500	5,936	2,968	231	463
Test	1000	11,822	5,911	296	569

Table 1: Data statistics of our lifelong learning deployment at the point where we froze collection for experiments reported within the paper and subsequent data release.

Experimental Results

- Comparison to other datasets

Dataset	Num. Episodes	Num. Utterances	Num. Human Utterances	Unique Tokens	Avg. Human Utt. Length	Number of Humans
PersonaChat (Zhang et al., 2018)	8,939	131,438	131,438	18,688	11.9	UNKNOWN
Wiz. of Wikipedia (Dinan et al., 2019c)	18,430	166,787	166,787	52,490	19.7	UNKNOWN
Empathetic Dialog (Rashkin et al., 2019)	24,850	64,636	64,636	19,458	15.3	810
Daily Dialog (Li et al., 2017)	22,236	87,170	87,170	20,673	14.5	UNKNOWN
LIGHT MTurk (Urbanek et al., 2019)	8,538	110,877	110,877	33,789	18.3	1,052
LIGHT WILD *(this paper)*	41,131	461,984	230,992	47,526	11.9	13,188

Table 2: Comparison of statistics of the open-domain dialogue data collected during our lifelong learning deployment (bottom row) compared to several existing (mostly crowdsourced) datasets. Our data is around twice as large in terms of human utterances than these datasets, and 4x as large in terms of dialogue utterances (as our data consists of human-model conversations), while the cost to collect our data was only $1/5^{th}$ of the price per utterance of LIGHT MTurk, see Sec. 5.3.3.

Experimental Results

- Model evaluation

	Retrieval Model (Hits@1/20 ⇑)			Generative Model (PPL ⇓)		
Model	LIGHT Test	LIGHT Test Unseen	WILD Test	LIGHT Test	LIGHT Test Unseen	WILD Test
Round 1	87.12	82.43	81.61	12.67	11.81	13.42
Round 2	87.65	82.70	84.60	12.57	11.74	12.31
Round 3	87.72	83.48	87.63	12.54	11.75	11.79

Table 3: Three rounds of training in our lifelong open-domain dialogue learning setup. Both retrieval and generative models trained on the data from the three rounds improve across both metrics on all three test sets.

Hits@1/20 metric: Top 1 accuracy given 19 random distractors and 1 correct next utterance

PPL: perplexity for generative models. Lower is better

Experimental Results

- Model evaluation

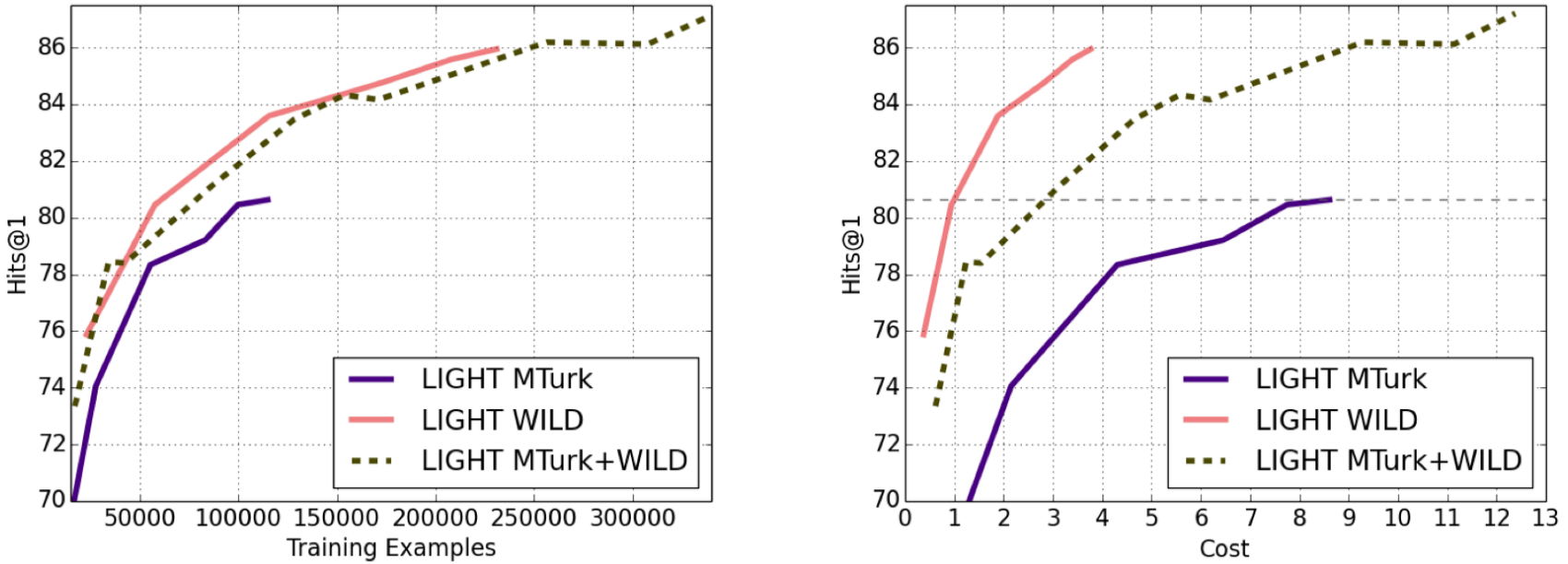


Figure 2: Hits@1/20 Accuracy on the LIGHT WILD validation set as a function of the number of training examples (left) or the cost of data collection (right). The cost axis is in units scaled by the cost of LIGHT WILD collection required to achieve the same performance as using the entire LIGHT MTurk dataset; it is more than 8× cheaper to use LIGHT WILD examples than LIGHT MTurk examples to achieve an accuracy of 80.63%. We also show performance for models which equally sample data from LIGHT MTurk+WILD datasets for training; utilizing all the data from both sources yields the best performance. However, LIGHT WILD data gives better accuracy improvements per training example (left plot).

Experimental Results

- Data quality based on Role-playing scores

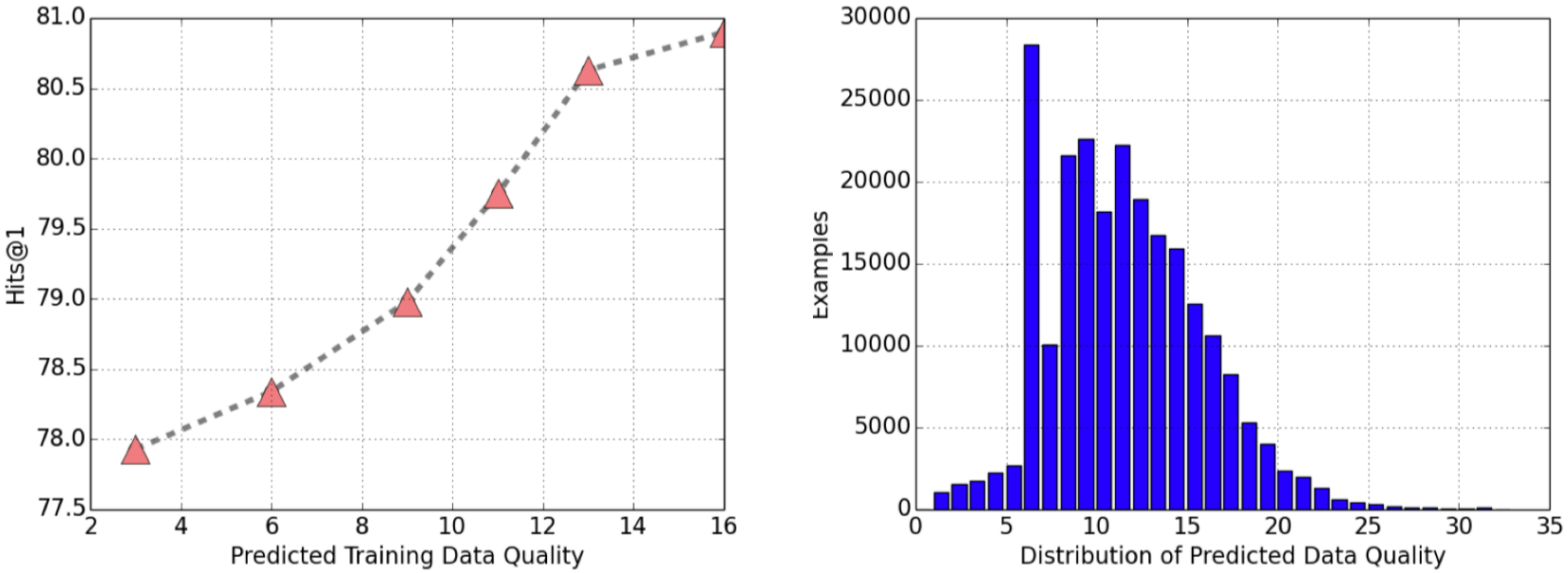


Figure 3: **Predicted Data Quality**. Left: Hits@1/20 accuracy on the WILD validation set when training with LIGHT MTurk + 10,000 examples from the WILD training set of a given predicted quality level, see Sec. 5.3.5. Data that is predicted to be higher quality yields improved validation accuracies. Right: The distribution of data quality predictions over the training set. A spike is seen at quality bin 6 because that is the lowest score one can achieve when completing a full episode (1 star per turn is awarded at minimum). Values lower than bin 5 indicate incomplete low-scoring episodes.

Key Takeaways

Improving an open-domain dialogue

- Interaction with humans in real-time during training
- High quality of data from a role-playing game
- Automation to collect the conversation data, low cost