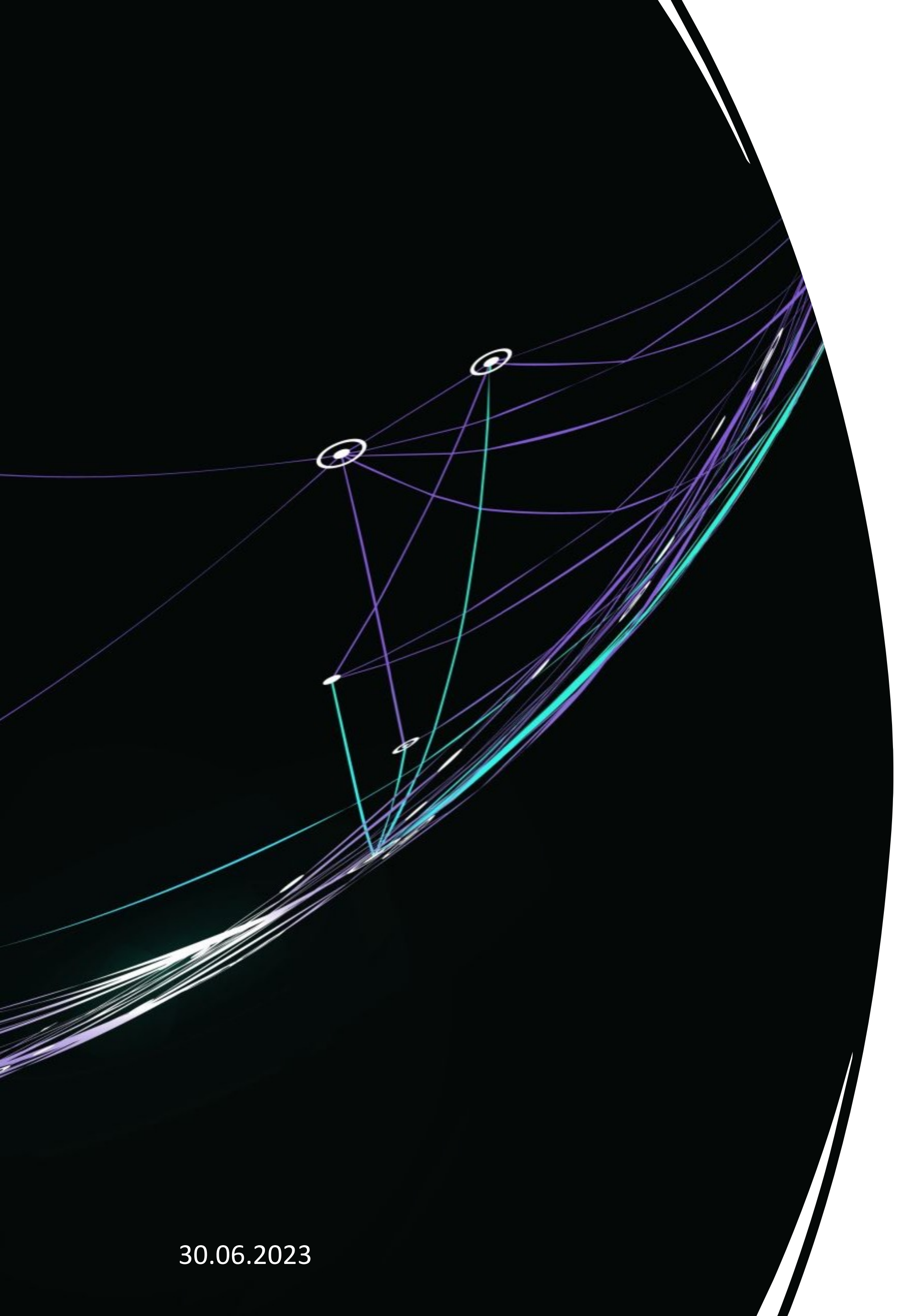


Dialogue Evaluation via Offline Reinforcement Learning and Emotion Prediction

Dr. Nurul Lubis
Dialog Systems and Machine Learning
Heinrich-Heine University Düsseldorf

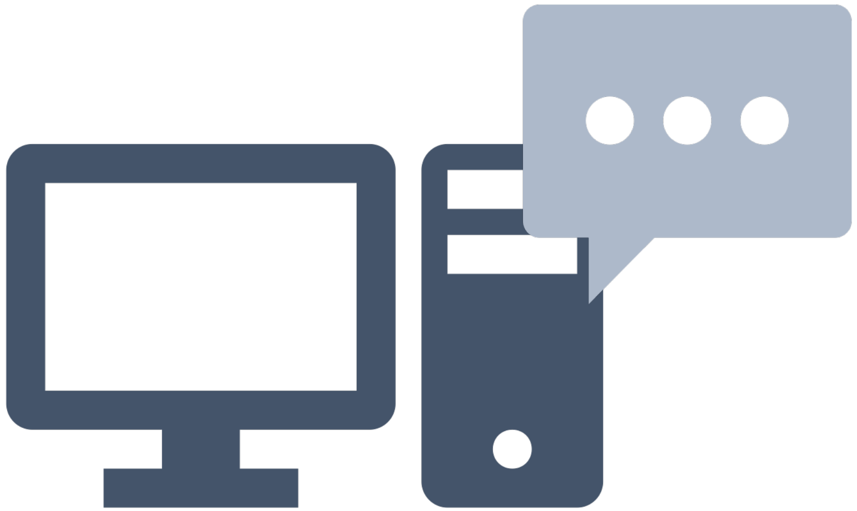
Why dialogue?

- Natural language has evolved to facilitate communication
- Dialogue is a prime interactive NLP task
    - Turing poses dialogue as a core AI problem (Turing, 1950)

https://prowritingaid.com/

What makes it challenging?

- Infinite possibilities of how a dialogue can go
  - We can always think of a dialogue that was never produced before
  - Can not be solved with handcrafted rules
- Dialogue can be viewed as an AI-complete problem (Shapiro, 1992)
  - Recognition, reasoning, and generation



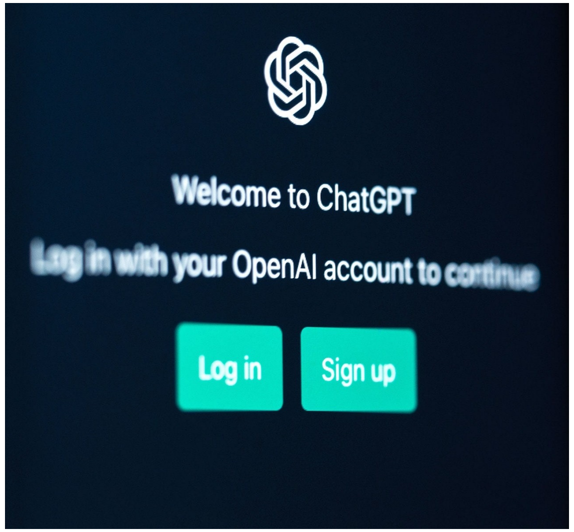
What are good dialogue properties?

- Understanding the user
- Handling different (new) topics in a dynamic world
- Understanding emotions and sentiment
- Responding in a human-like manner
- Responding sensibly, intelligently and fluently
- Providing personalised outputs
- …

Dialogue systems are becoming more ubiquitous

reuters.com

bloomberg.com



mdr.de

What makes one system better than the other?

Dialogue systems

Task oriented dialogues (ToD)

- Centered around fulfilling user goals
- Domain specific

I'm looking for a nice restaurant in the center of town

What type of food would you like to have?

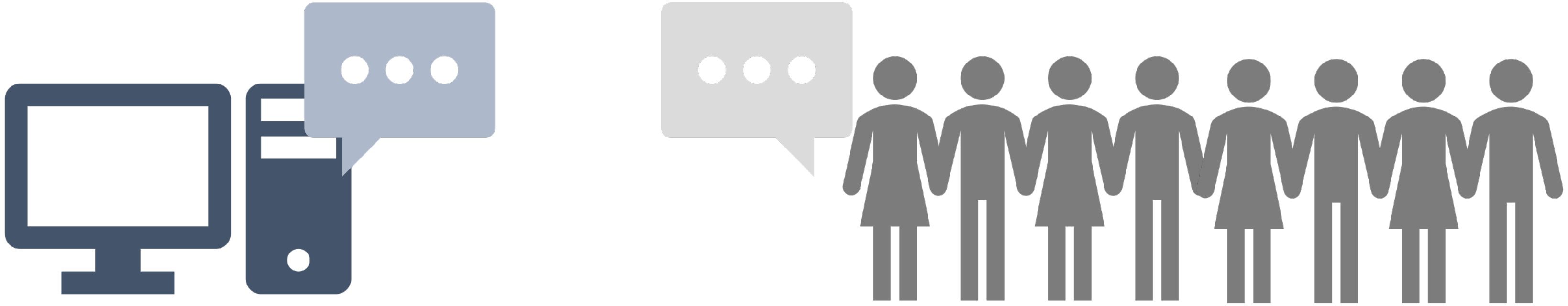
Chit-chat dialogues

- Typical aims are user engagement or entertainment
- Open-ended

How many pets do you have?

I have two dogs and a cat. I love animals.

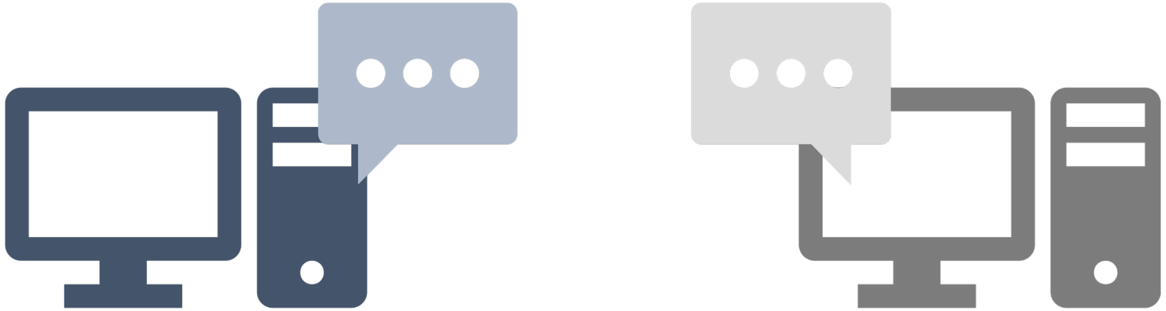
Subjective Human Evaluation



User-centered criteria (Walker et al., 1997; Lee and Eskenazi, 2012; Ultes et al., 2017)

- Time- and cost-intensive
- Hard to compare

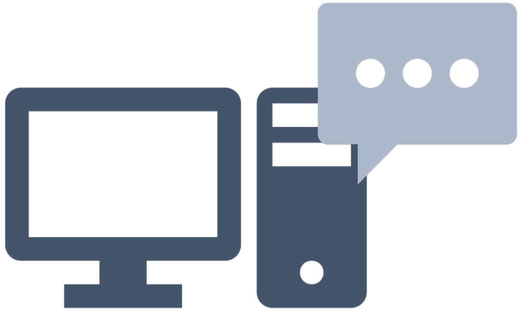
Interactive User Simulator



Interactive user **simulator** (US) (Schatzmann, 2008; Lin et al., 2021)

- Not straightforward to build

Automatic evaluation with static corpora

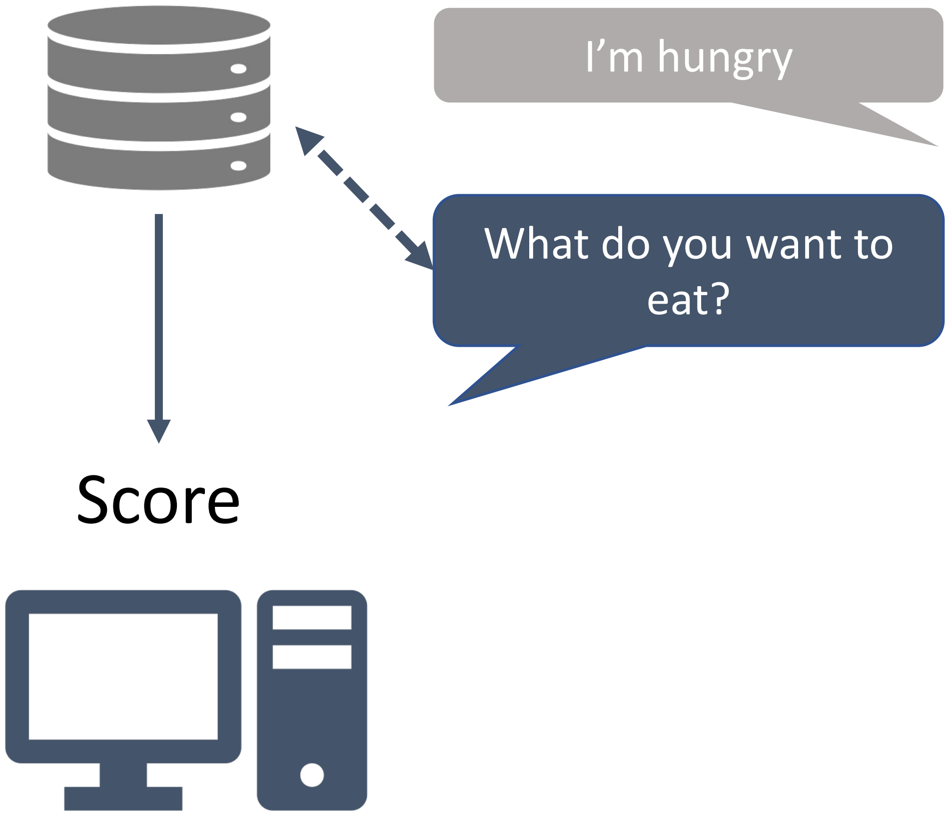


Can we use a test set for dialogue evaluation?

Practical	Easily reproducible
• Easy and fast to compute	• Suitable for benchmarking

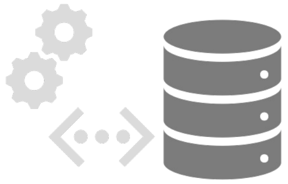
Corpus-based evaluation

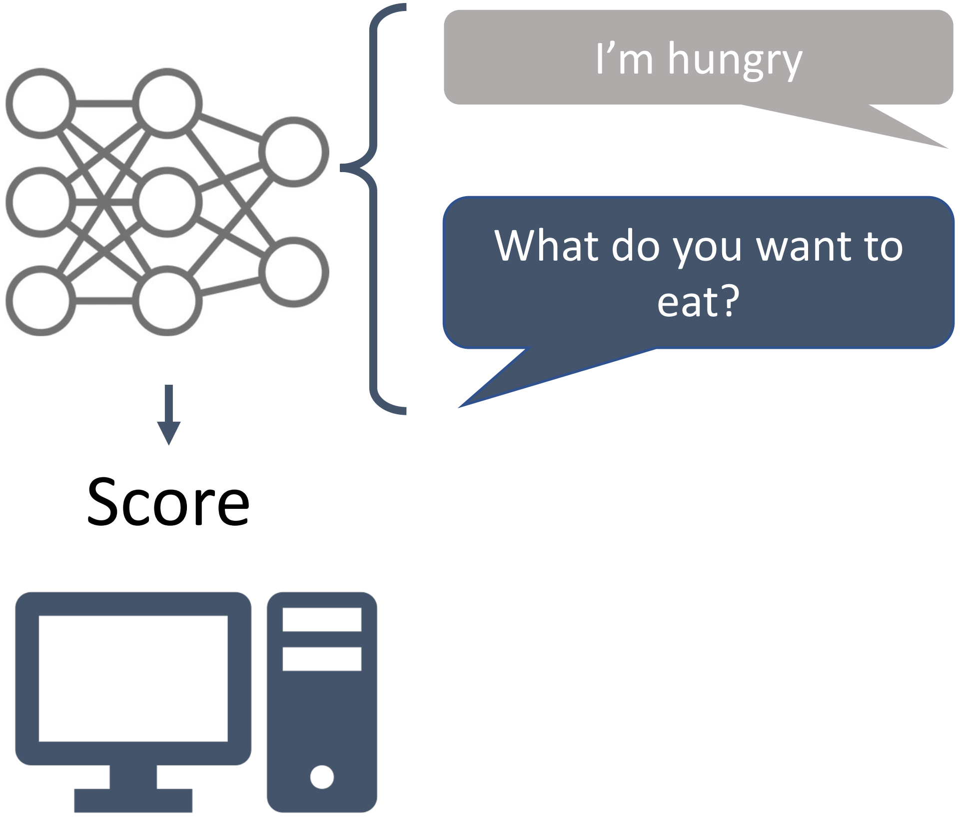
Method 1: Response matching



- Match system’s outputs to gold responses from the corpus
  - E.g. N-gram based BLEU (Papineni et al., 2002)
- Poorly correlates with human judgement (Liu et al., 2016)
  - Dialogue is a one-to-many problem
- Turn-based, ignores the dialogue as a whole

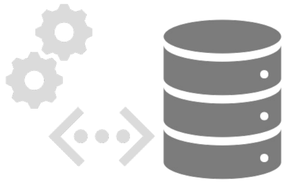
Method 2: Predict a score

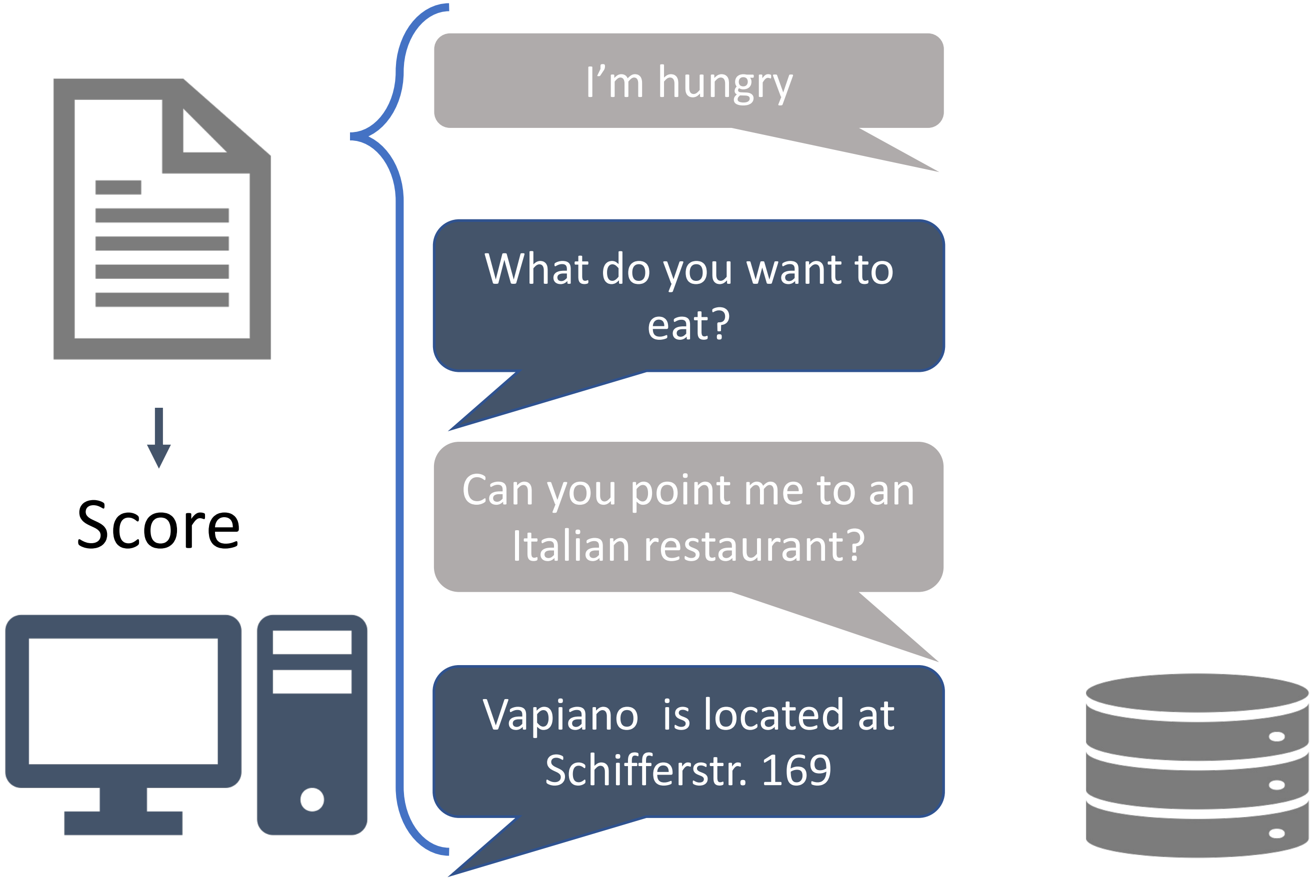




- Train a model to output a score given a response/dialogue
- Considering dialogue context
- Focus on subjective quality
    - E.g FED (Mehri and Eskanazi, 2020), USR (Mehri and Eskanazi, 2020)
- *Did the user fulfill their goal?*

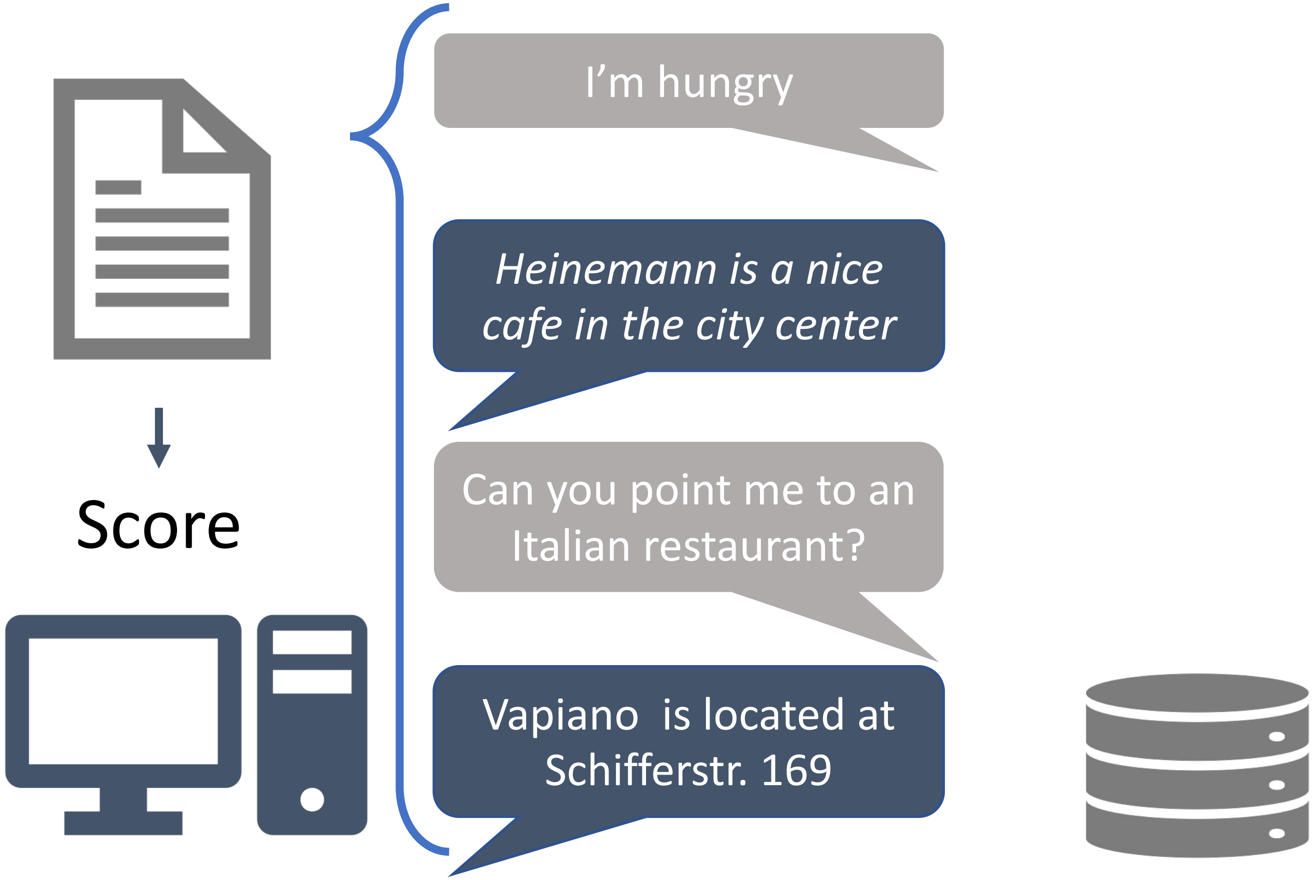
Method 3: Construct pseudo-dialogue





- Replace golden system turns with generated ones
- Evaluate a dialogue as a whole
- Rules to check whether user goal is fulfilled
  - Objective measure
  - Corpus-specific
- *Does pseudo-dialogue still make sense?*

Method 3: Construct pseudo-dialogue



- Context mismatch between user and system turns
  - Pseudo-dialogue is not the result of an interaction
- Overlooks specific types of mistakes
  - May overestimate dialogue policy performance

Corpus-based evaluation: current challenges

Not yet strongly correlated with human judgements	Focus on limited, subjective qualities	Lack of generalization across datasets and models

- A recent National Science Foundation (NSF) report (Mehri et al., 2022)

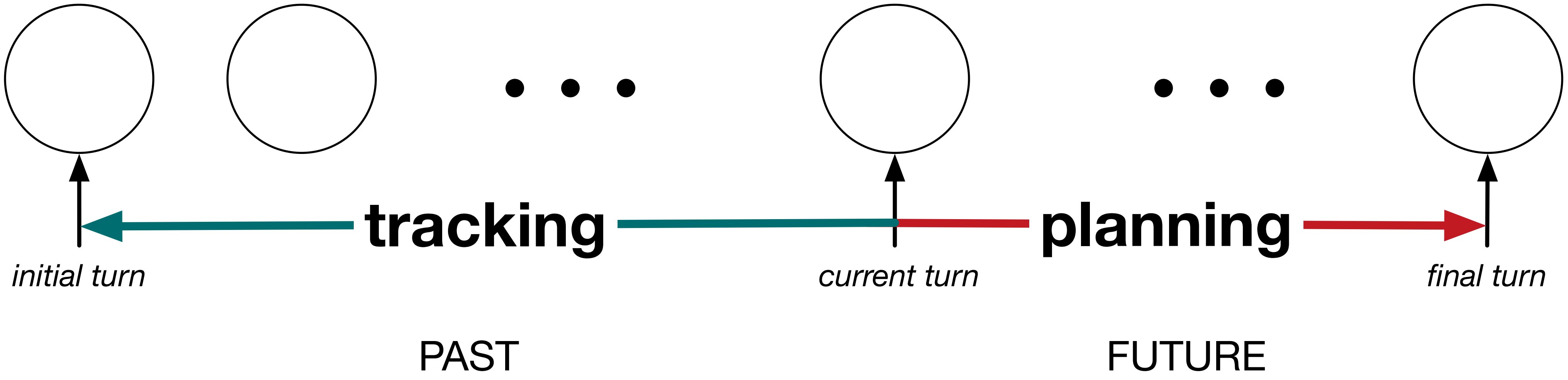
Tackling the issues

How can we solve current challenges with an efficient and reliable method to evaluate dialogue systems?

We propose to use offline reinforcement learning (RL) critic as dialogue evaluator

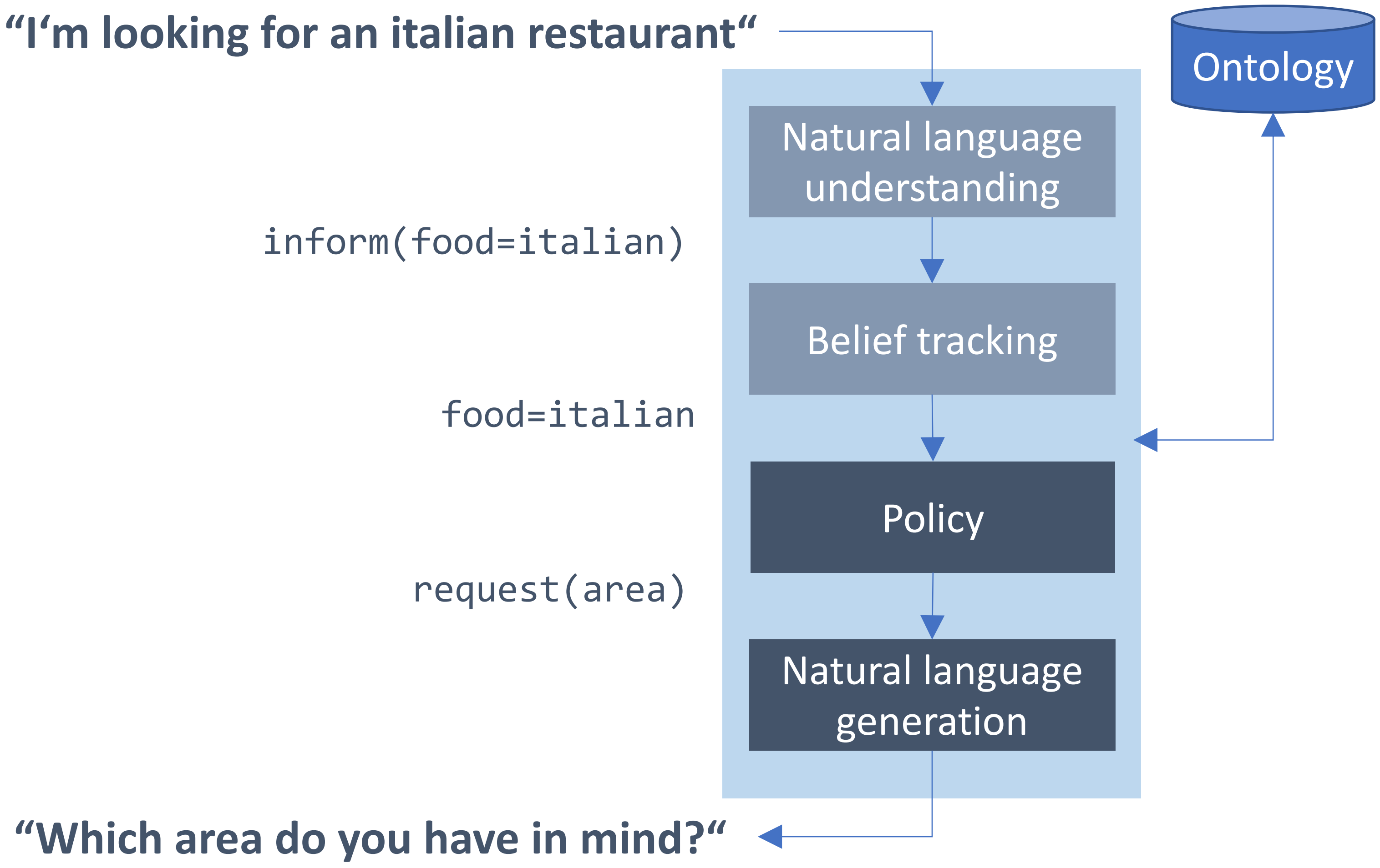
Hello, dialogue systems

Fundamental tasks in dialogue



Credit: Prof. Milica Gašić

Modular ToD systems

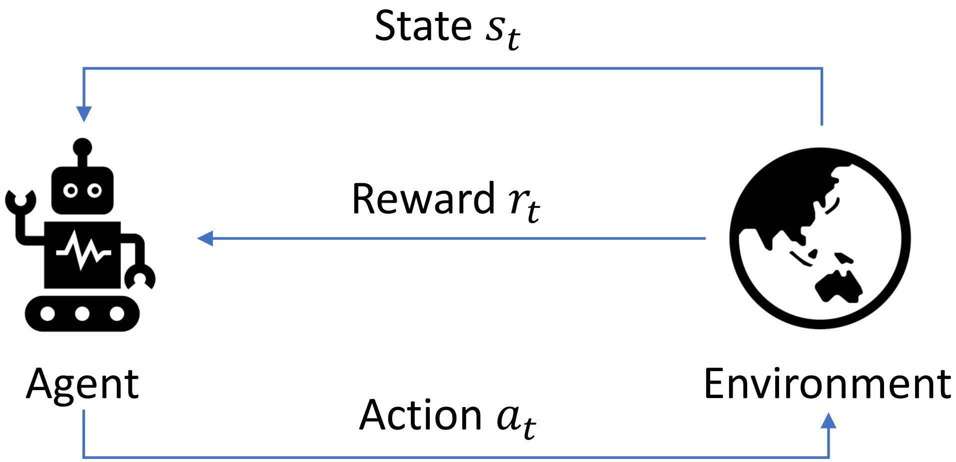


- Modular approach: Dialogue systems are divided into modules (Williams, 2006; Thomson and Young, 2010)
- Policies can be trained with supervised or reinforcement learning

Reinforcement Learning (RL)

A gentle primer

RL Primer: Notations



- Through interactions with the environment, the agent tries to find the best policy based on some measure of reward.
- Huge amount of interactions are needed

Trajectory $\tau = \{(s_1, a_1, s_2, r_1), \dots, (s_T, a_T, s_{T+1}, r_T)\}$
Sequence of state, action, reward tuples from a sequence of time steps, e.g. 1 to T. (we assume a finite horizon case)

Return R_t
Discounted cumulative reward: *How much reward have I collected from timestep t until the end?*

$$R_1 = \sum_{n \geq 1} \gamma^{n-1} r_n$$

Policy $\pi(a|s)$
Probability distribution over actions in a given state
In a given state, which action should I take?

Value functions

- $Q^\pi(s, a)$ expected return of being in state s, taking action a, and following policy π afterwards
- *How good is it to take a particular action in a given state?*

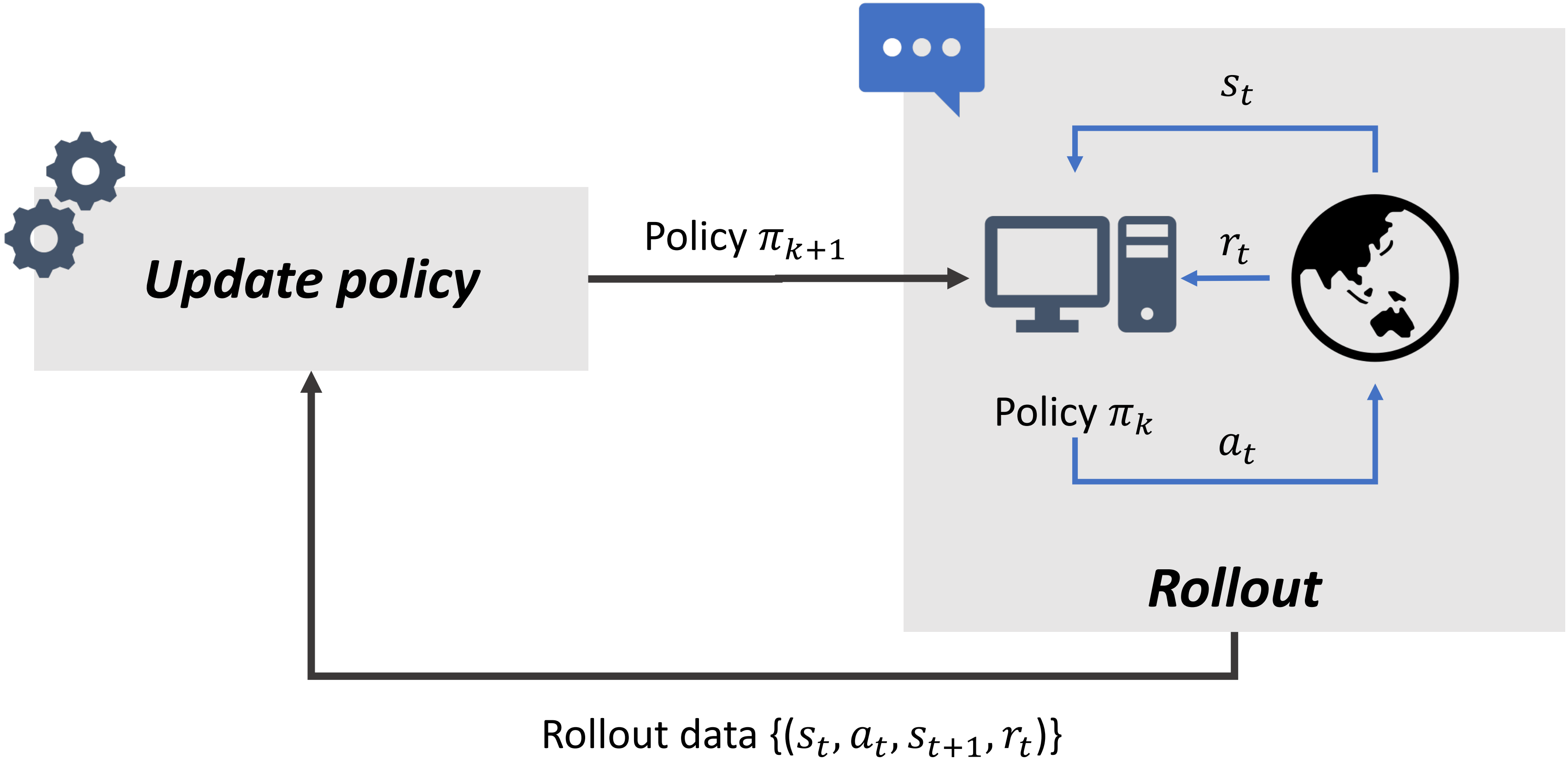
Learning methods

- **Policy-based:** We learn the policy $\pi_\theta(a|s)$...
  - Using parameters θ to map state to action
- **Value-based:** We learn the value function $Q^\pi(s, a)$...
  - Bellman Equation: the value of any state can be calculated with one-step look ahead, as opposed to having to inspect every future state

$$\mathcal{T}Q(s_t, a_t) = \mathbb{E}_{s_{t+1}}[r_t + \gamma Q(s_{t+1}, a_{t+1})].$$

  - Act greedily: choose action with the highest value estimate
- **Actor-critic:** We learn both!
  - Learn a policy that maximizes value estimate

Optimizing policies with online RL



The need for offline RL

Learning from online interaction can be expensive and time consuming

- Even more than evaluation!

Some environments are high-risk

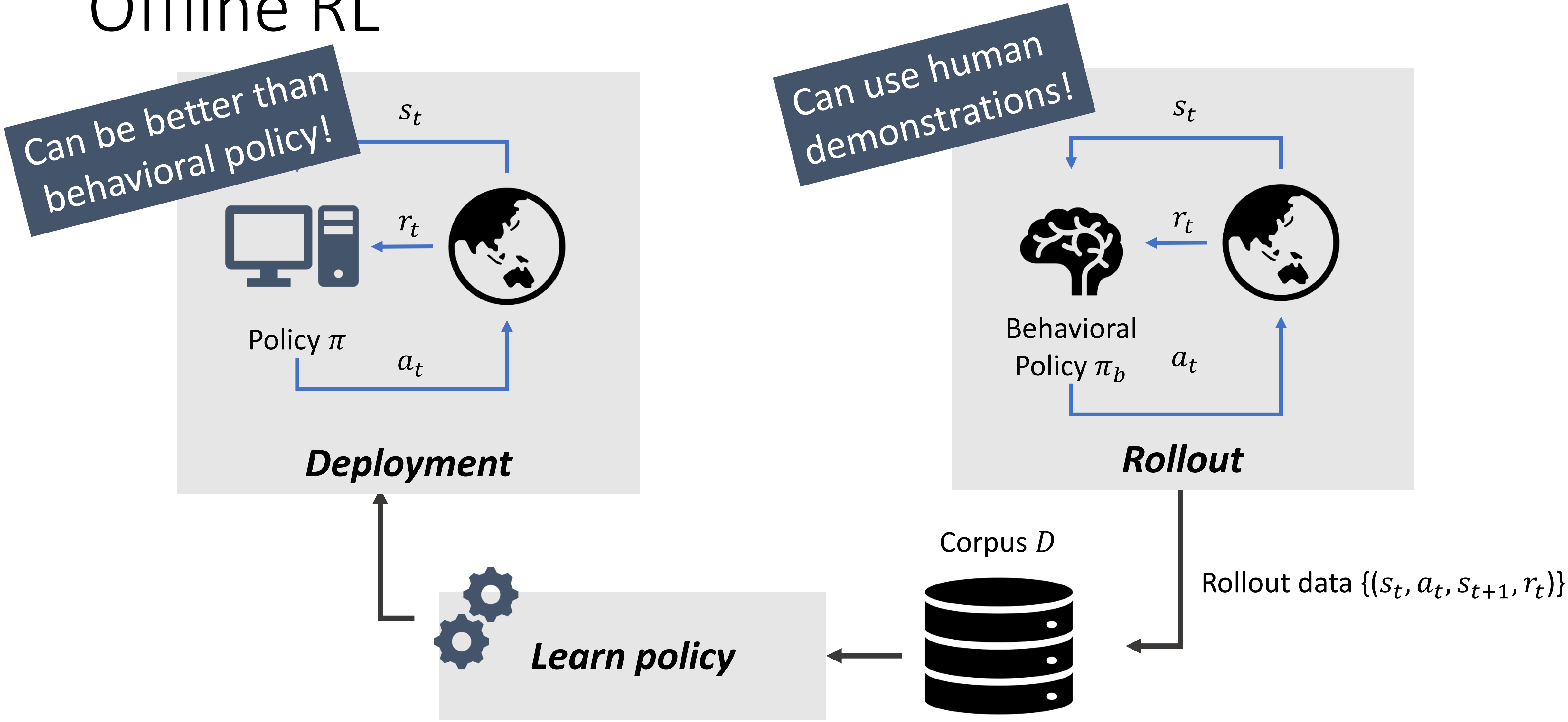
- Dialogue systems for emotional distress?

Some behavior we want to learn are highly complex

- How do we model the environment?

Can we leverage datasets to learn a policy?

Offline RL

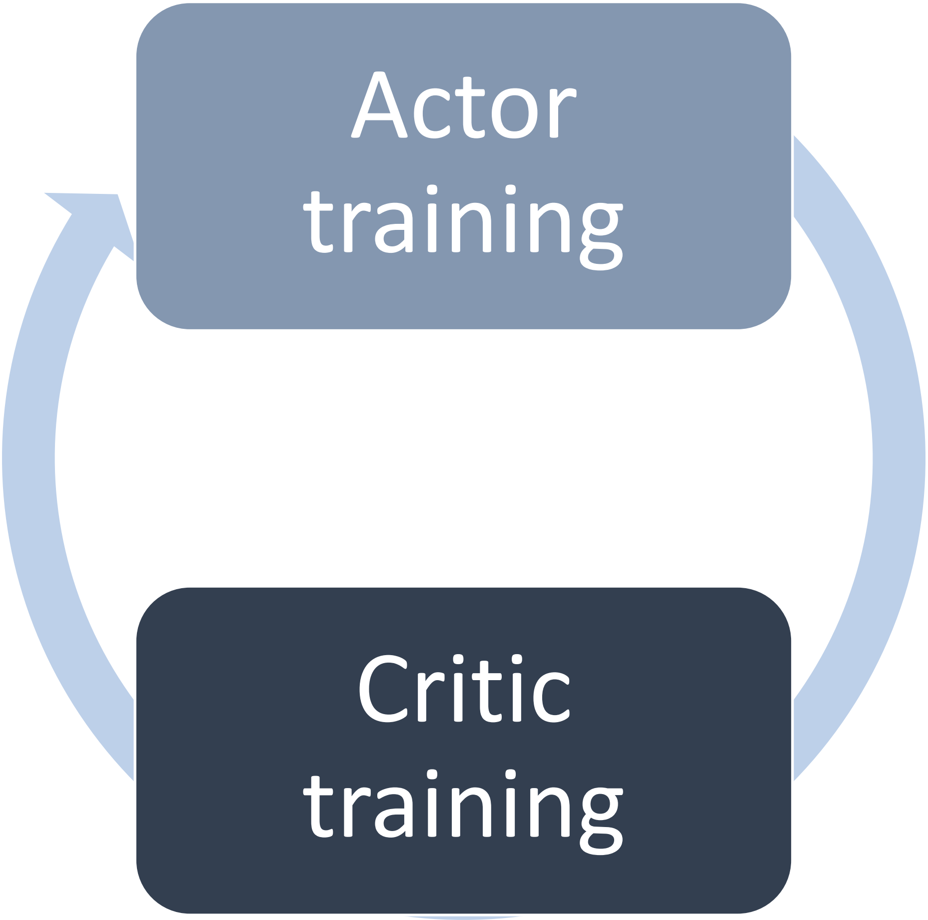


Can be better than behavioral policy!
s_t
r_t
Policy π
a_t
Deployment
Can use human demonstrations!
s_t
r_t
Behavioral Policy π_b
a_t
Rollout
Corpus D
Rollout data $\{(s_t, a_t, s_{t+1}, r_t)\}$
Learn policy

Dialogue Evaluation with Offline Reinforcement Learning

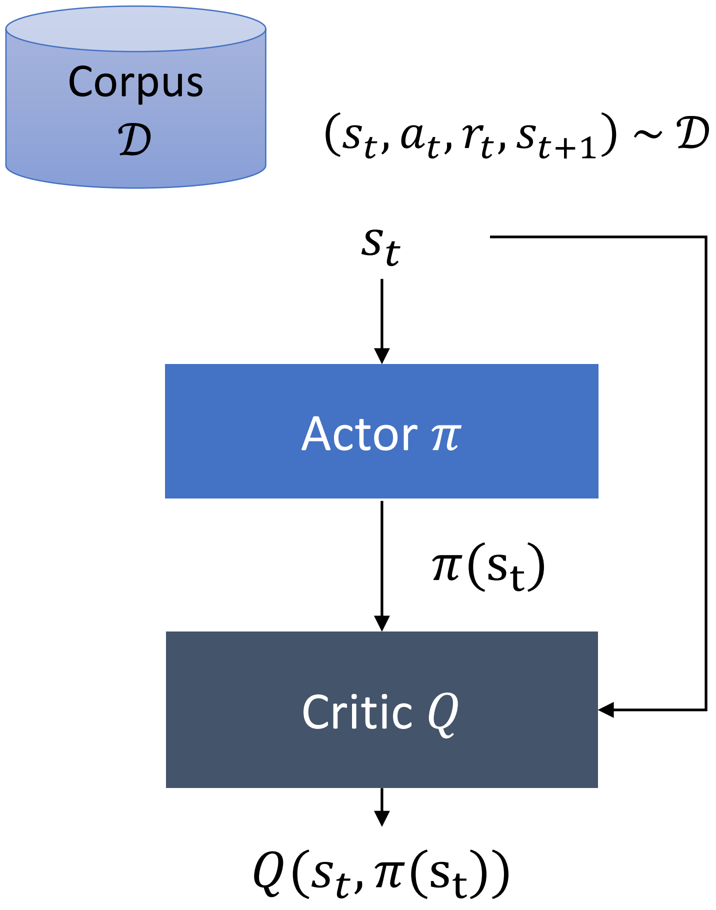
Lubis, Nurul, et al. "Dialogue Evaluation with Offline Reinforcement Learning." *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2022.

Learning with critic



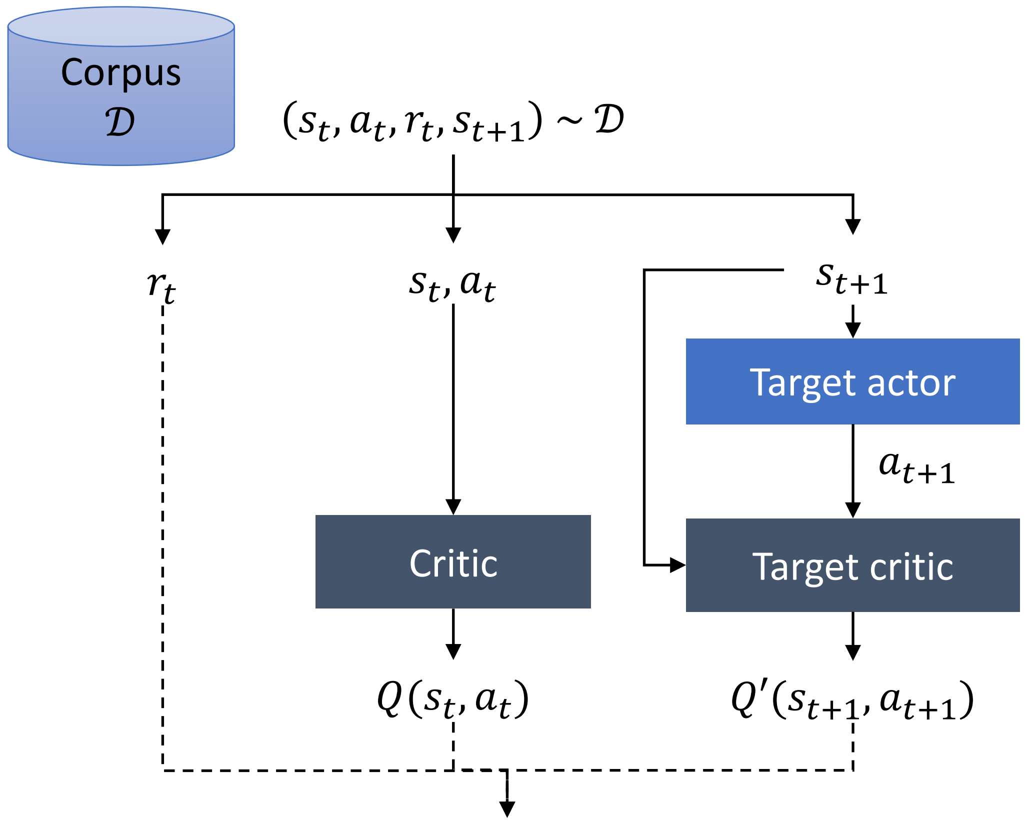
- Actor and critic are optimized alternatingly throughout training

Actor training



- Start with supervised learning (SL) pre-training to initialize the actor
- Continue training with offline RL
  - For each state, actor predicts the action
  - Critic estimate the value function
  - Actor tries to maximize critic's estimate

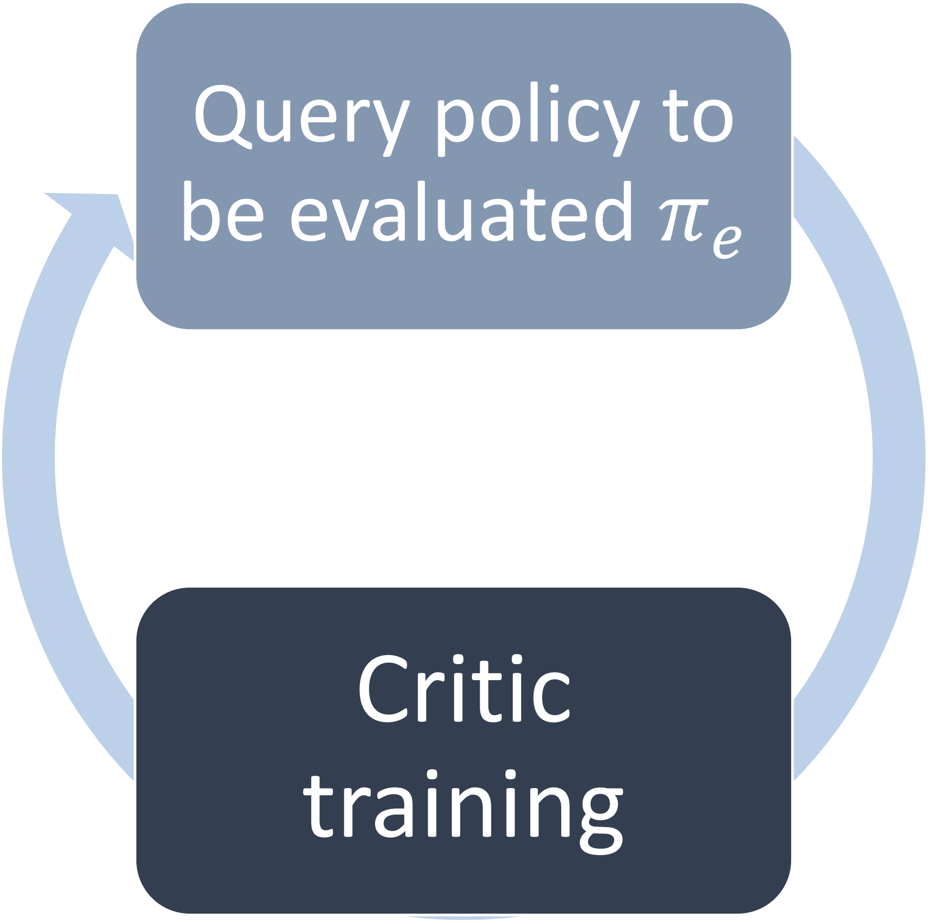
Critic training



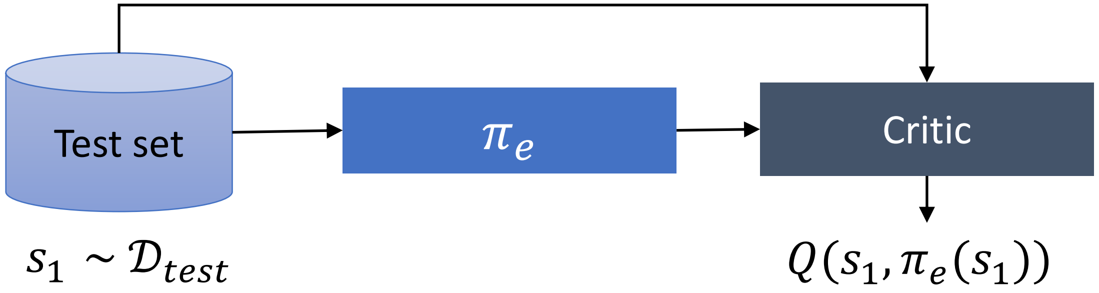
$$\mathcal{L}_{\text{critic}} = (Q(s_t, a_t) - (r_t + \gamma Q'(s_{t+1}, \pi'(s_{t+1}))))^2$$

- Critic produce value estimates
  - a_t comes from data
  - a_{t+1} comes from actor
- Estimate is refined by minimizing the error of Bellman equation

Evaluation with critic



- For any policy, we can train a critic independently after-the-fact
- Use policy to be evaluated to estimate $Q(s_{t+1}, a_{t+1})$
  - Used to compute critic loss
- Use the final critic to estimate Q-values over a test set
  - Average Q-value on the initial states



Advantages

Theoretically grounded
solves context mismatch

Ontology, data, and model independent
model-based, no handcoded rules

State, action, and reward can take any form
adjust the architecture of critic and actor

Experiments

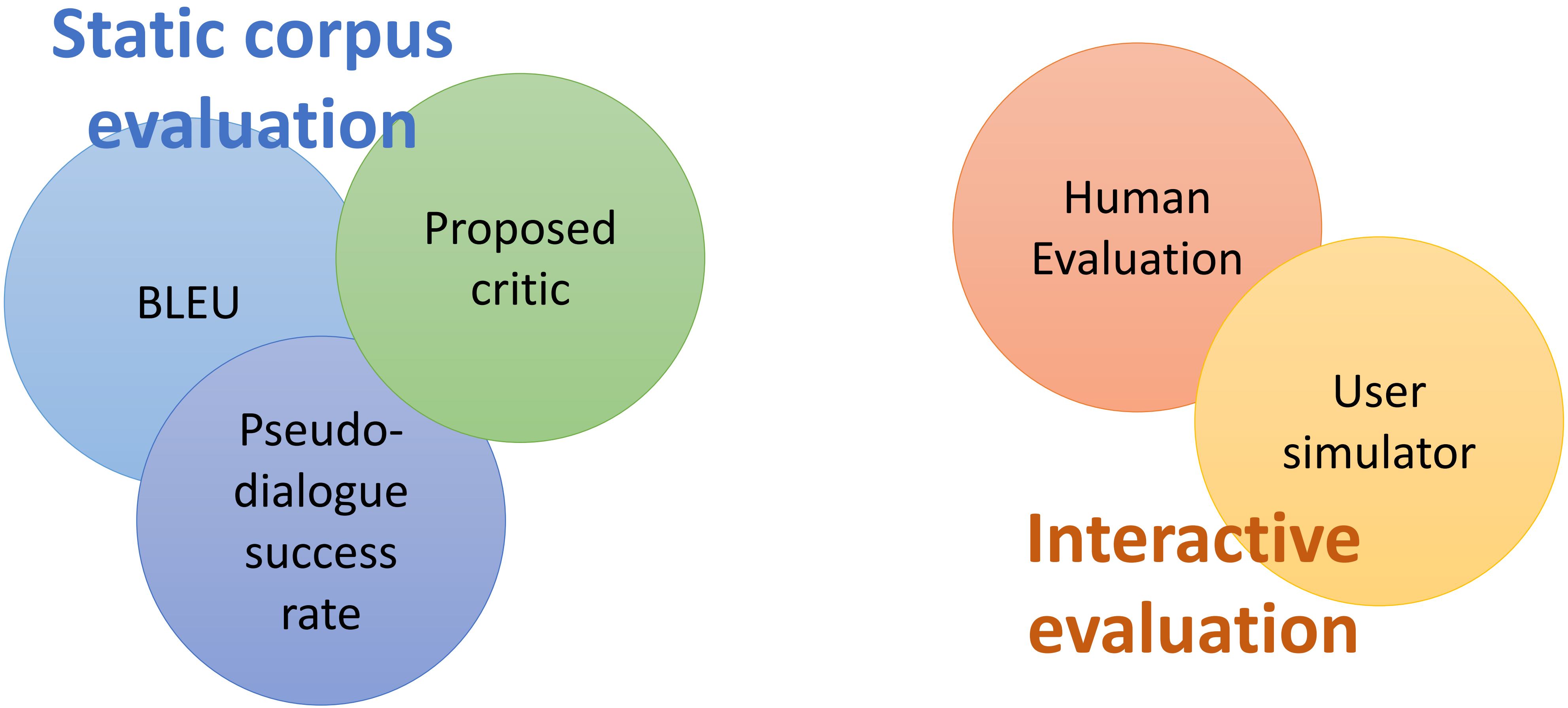
Task-oriented dialogue benchmark

- MultiWoZ corpus (Budzianowski et al., 2018)
  - Information seeking and reservation making
- Multi-domain dialogues
  - Restaurants, hotels, attractions, taxi, train, hospital, police
  - Multiple domain can occur in one dialogue

Policy optimization

(INFORM + SUCCESS)*0.5 + BLEU	MultiWOZ 2.0			MultiWOZ 2.1		
Model	INFORM	SUCCESS	BLEU	INFORM	SUCCESS	BLEU
TokenMoE* (Pei et al. 2019)	75.30	59.70	16.81			
Baseline* (Budzianowski et al. 2018)	71.29	60.96	18.8			
Structured Fusion* (Mehri et al. 2019)	82.70	72.10	16.34			
LaRL* (Zhao et al. 2019)	82.8	79.2	12.8			
SimpleTOD (Hosseini-Asl et al. 2020)	88.9	67.1	16.9	85.1	73.5	16.22
MoGNet (Pei et al. 2019)	85.3	73.30	20.13			
HDSA* (Chen et al. 2019)	82.9	68.9	23.6			
ARDM (Wu et al. 2019)	87.4	72.8	20.6			
DAMD (Zhang et al. 2019)	89.2	77.9	18.6			
SOLOIST (Peng et al. 2020)	89.60	79.30	18.3			
MarCo (Wang et al. 2020)	92.30	78.60	20.02	92.50	77.80	19.54
UBAR (Yang et al. 2020)	94.00	83.60	17.20	92.70	81.00	16.70
HDNO (Wang et al. 2020)	96.40	84.70	18.85	92.80	83.00	18.97
LAVA (Lubis et al. 2020)	97.50	94.80	12.10	96.39	83.57	14.02
JOUST (Tseng et al. 2021)	94.70	86.70	18.70			
CASPI (Ramachandran et al. 2021)	96.80	87.30	19.10			
GALAXY (He et al. 2021)	94.8	85.7	19.93	94.8	86.2	20.29

Metrics to be compared



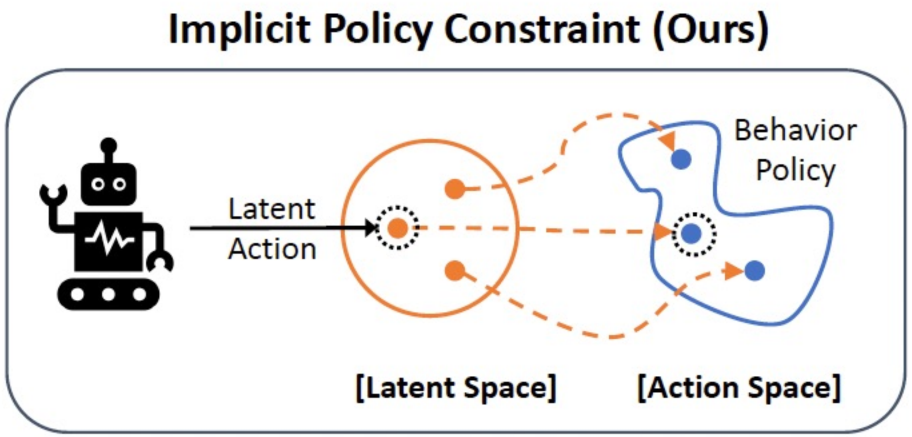
Policy Optimization

Experiments

LAVA + PLAS

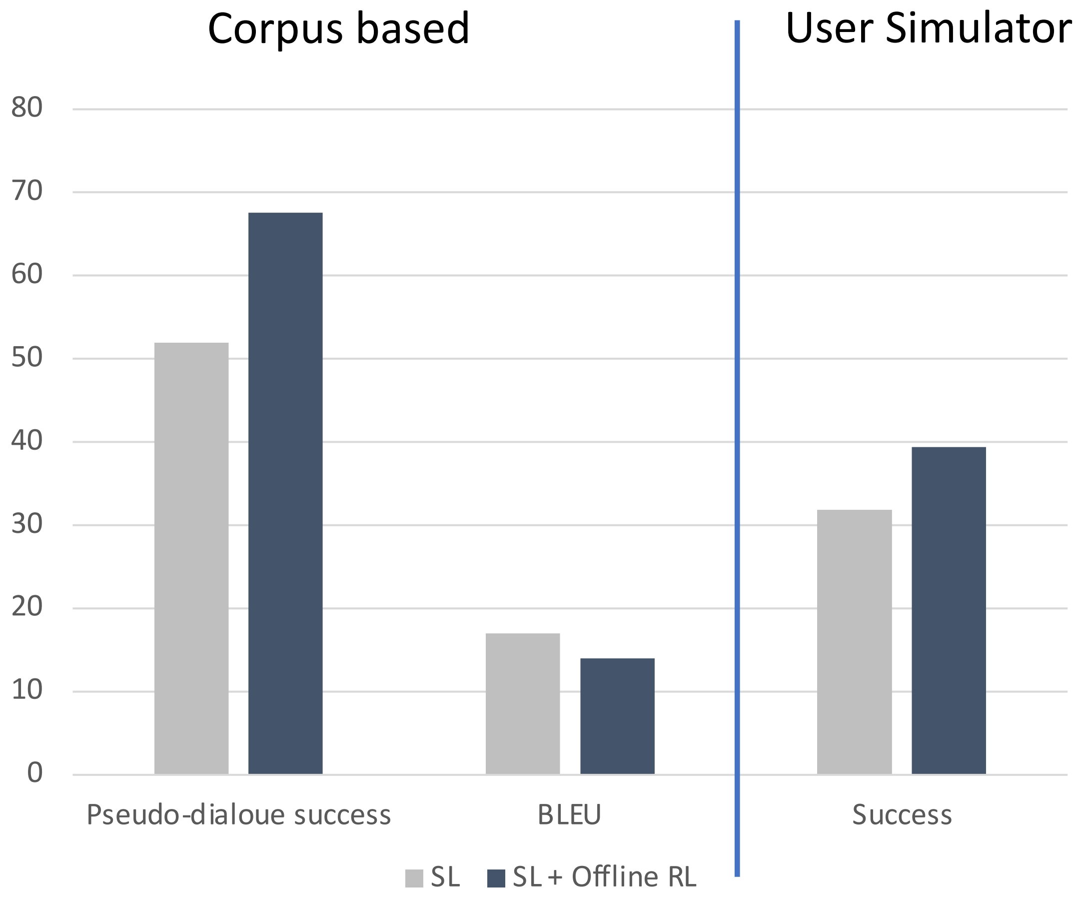
- Goal: To use critic's signal to optimize a dialogue policy via offline RL
- Hypothesis: optimizing critic's signal will also improve policy performance as measured by existing metrics

- PLAS (Zhou et al., 2020) offline RL algorithm on latent space



- Train a VAE to reconstruct actions found in corpus

Can the critic optimize the policy as measured by established metrics?



- Task-related metrics are consistently improved via offline RL on critic's signal
- Slight decrease on BLEU
  - Trade off between BLEU and success has been observed before (Zhou et al., 2020; Lubis et al., 2021)

Policy Evaluation

Experiments

Dialogue evaluation with offline RL

Goal: Investigate critic's value estimate as an evaluation metric compared to existing ones

Hypothesis: Critic can serve as a corpus-based evaluation metric that is better correlated with human judgements

Policies to be evaluated

SL

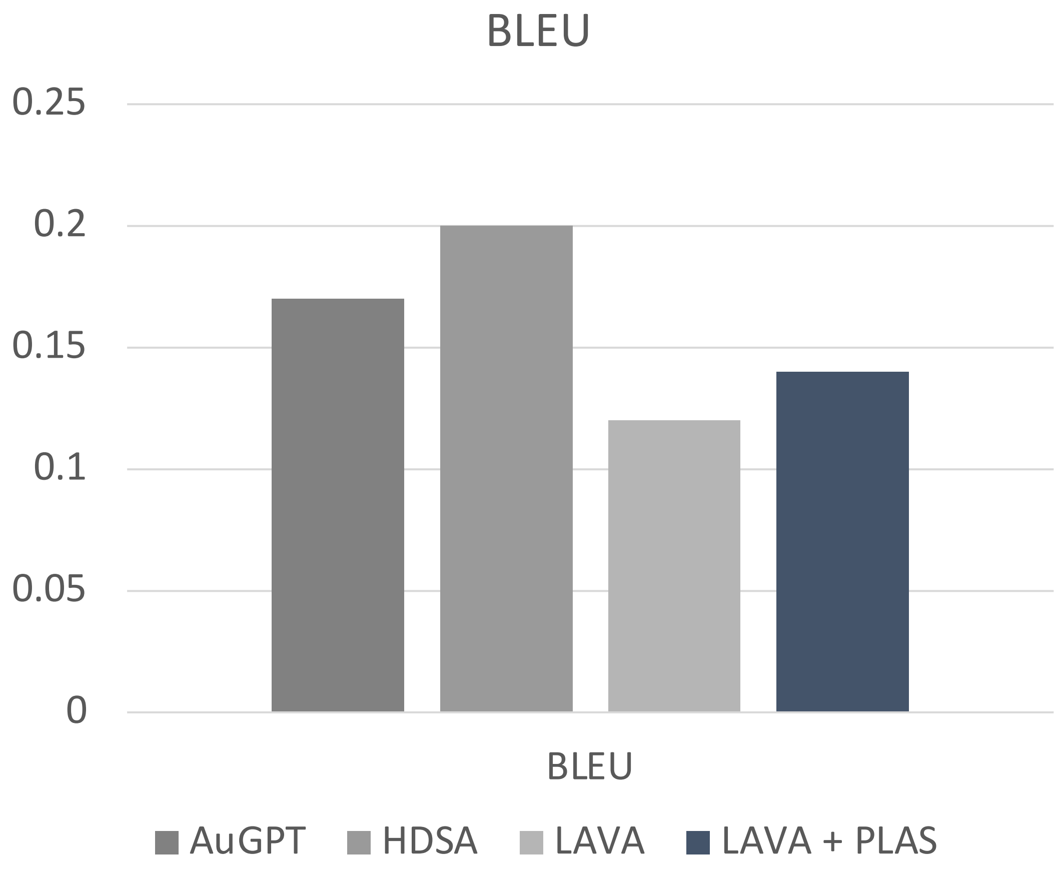
- **AuGPT** (Kulhanek et al., 2021) end-to-end **transformer-based** dialogue system, **large** amounts of data and labels
- **HDSA** (Chen et al., 2019) Policy operates on semantic-level action with **a dedicated NLG module**

SL + RL

- **LAVA** (Lubis et al., 2020) Policy with latent action, optimized on corpus-based **success rate** using RL
- **LAVA + PLAS** (proposed) Policy with latent action, optimized on **critic's signal** using offline RL

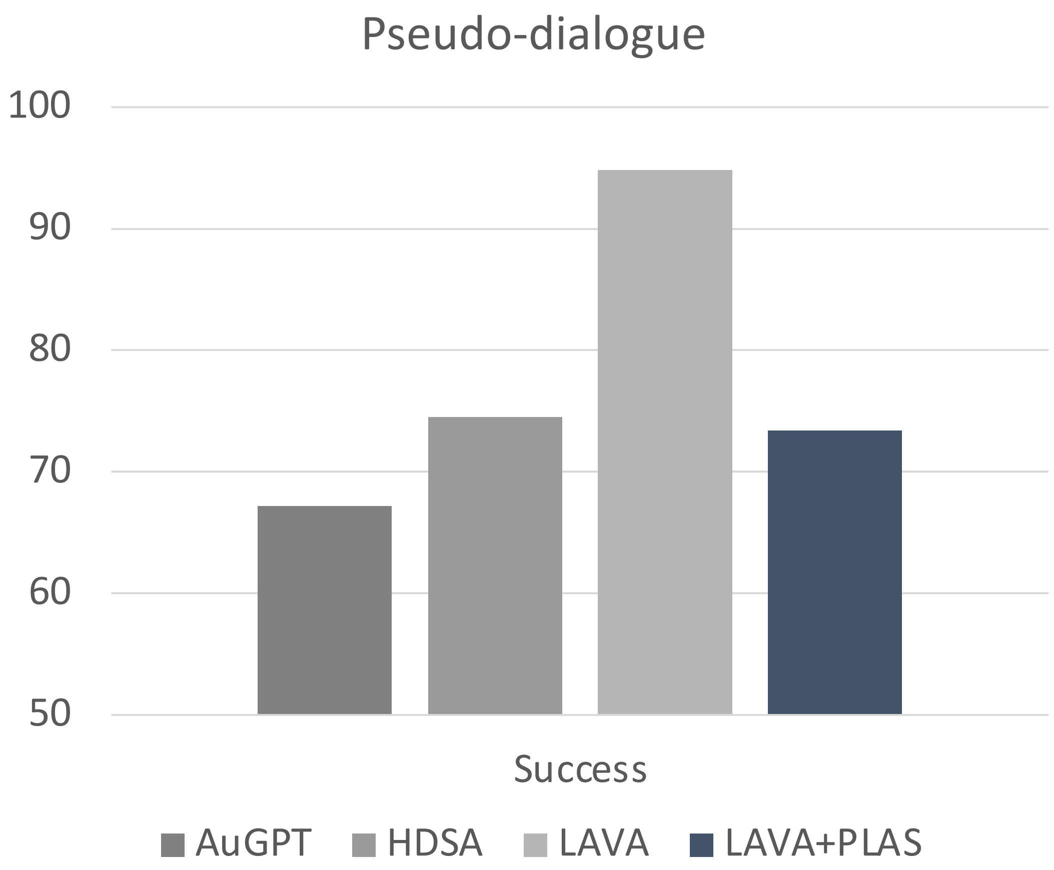
Corpus-based evaluation

- HDSA has highest BLEU score
  - Trained emphasis on generation



Corpus-based evaluation

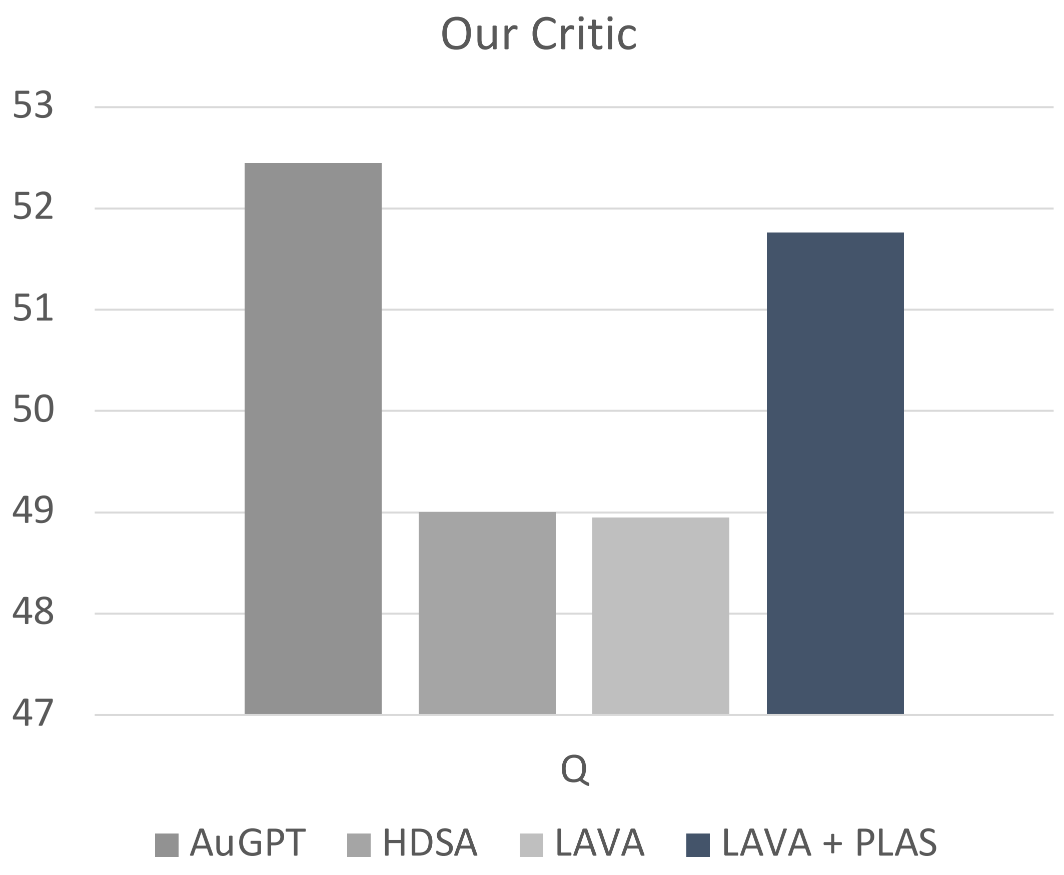
- HDSA has highest BLEU score
  - Trained emphasis on generation
- LAVA has highest corpus success rate
  - Optimized with RL on this metric



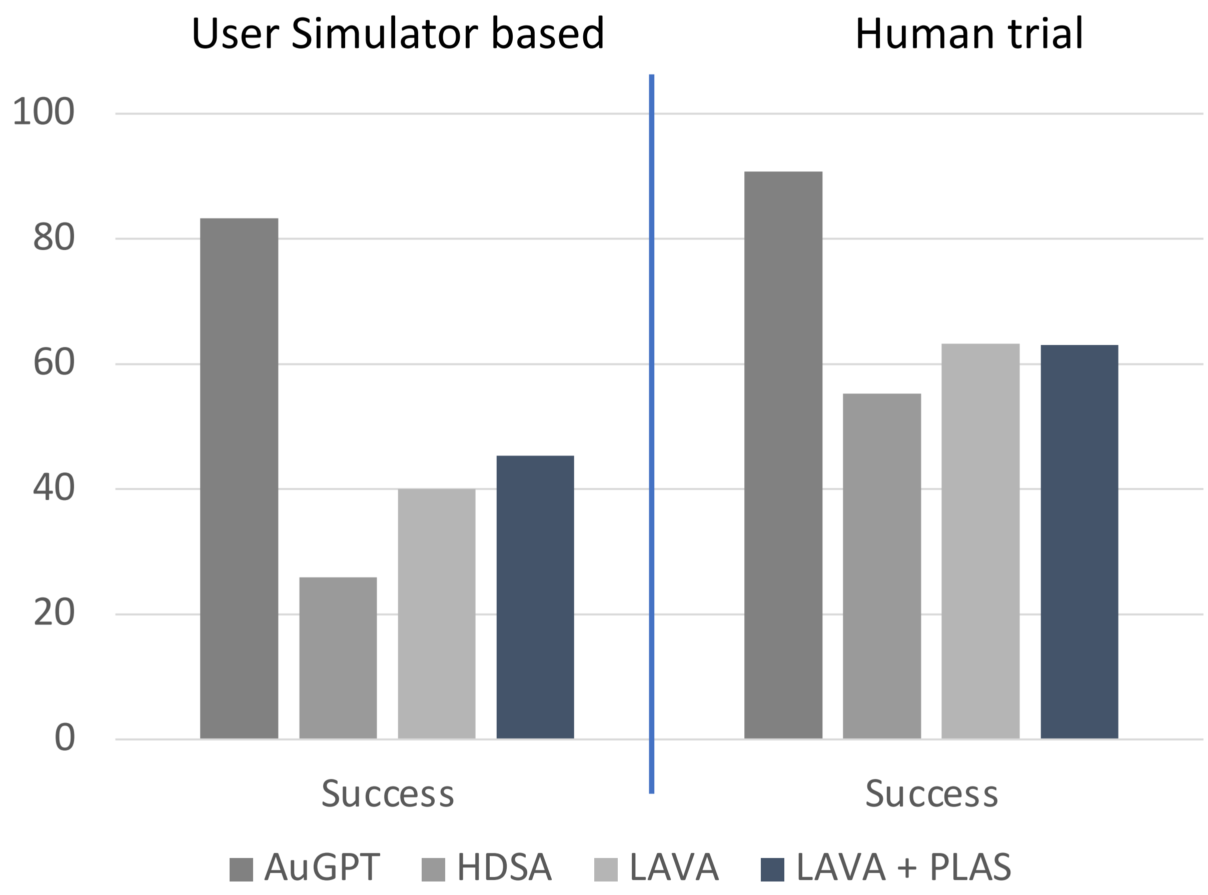
Corpus-based evaluation

- HDSA has highest BLEU score
  - Trained emphasis on generation
- LAVA has highest corpus success rate
  - Optimized with RL on this metric
- AuGPT has highest Q-value, followed by LAVA + PLAS
  - LAVA + PLAS is optimized on this metric

"Best model" on each corpus-based metric differs!



Interactive Evaluation



- Different trend compared to corpus evaluations
- AuGPT does very well on interactive evaluations
  - Large pre-trained model with data augmentation

Does the critic correlate with human judgement?

Fleiss' Kappa			Human Evaluation	
			Success	Rating
Corpus-based	Corpus	Match	-0.623	-0.571
		Success	-0.460	-0.397
		BLEU	0.343	0.299
	Critic		**0.755**	**0.713**
Interactive	US	Complete	0.992	0.984
		Success	0.991	0.984
		Book	0.789	0.802
		F1	0.990	0.978
		Turn	-0.967	-0.956

- Corpus-based metrics
  - Standard corpus-based metrics are negatively correlated with human evaluation
  - Our experiment confirm that BLEU has poor correlation
  - Our critic has strong correlation with human judgement
- Interactive metrics
  - User simulator is a good proxy to estimate system performance in human trial
- Critic training has the advantage of being corpus- and model-independent

Corpus- and model-independent evaluation

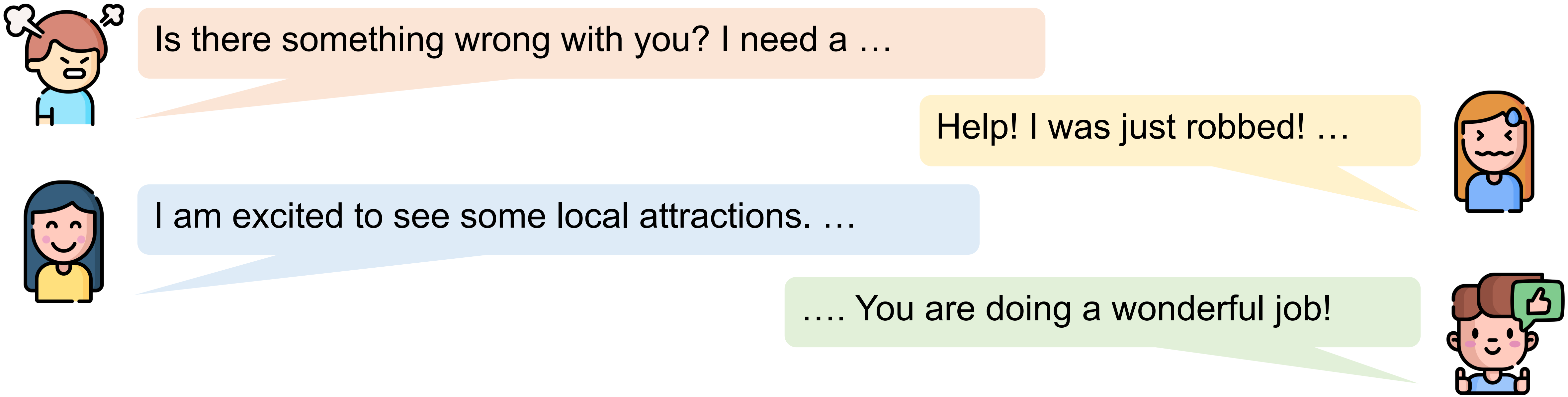
- Can we infer dialogue success from other signals?
- How does a successful dialogue look like?
- How do users behave in a successful dialogue?
- How do users react to a failed dialogue?
- ...

Emotional signal for task-oriented dialogue evaluation

Emotion in task-oriented dialogues

Feng, Shutong, et al. "EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems." *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022.

- Emotions are part of a natural human-like dialogue
- However, emotions are mainly studied in **chit-chat** dialogues
- User also expresses emotion as it relates to their goal



Emotion in task-oriented dialogues

Feng, Shutong, et al. "EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems." *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022.

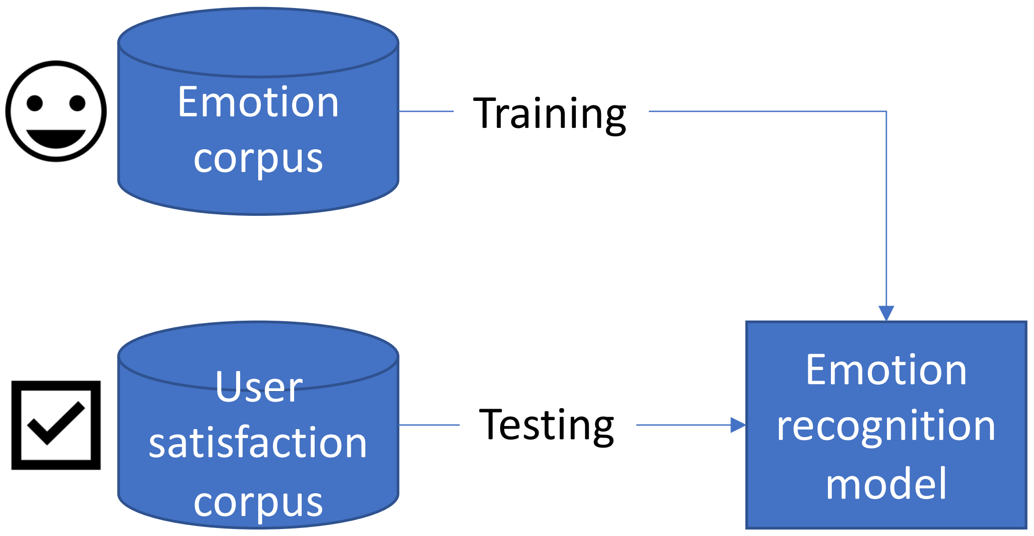
- We identified 7 classes to model user emotion

Valence	Elicitor	Conduct	Emotion Tokens
Neutral	-	-	**Neutral**
Negative	Event/Fact	Neutral/Polite	**Fearful**, sad, disappointed
Negative	System	Neutral/Polite	**Dissatisfied**, disliking
Negative	User	Neutral/Polite	**Apologetic**
Negative	System	Impolite	**Abusive**
Positive	Event/Fact	Neutral/Polite	**Excited**, happy, anticipating
Positive	System	Neutral/Polite	**Satisfied**, liking, appreciative

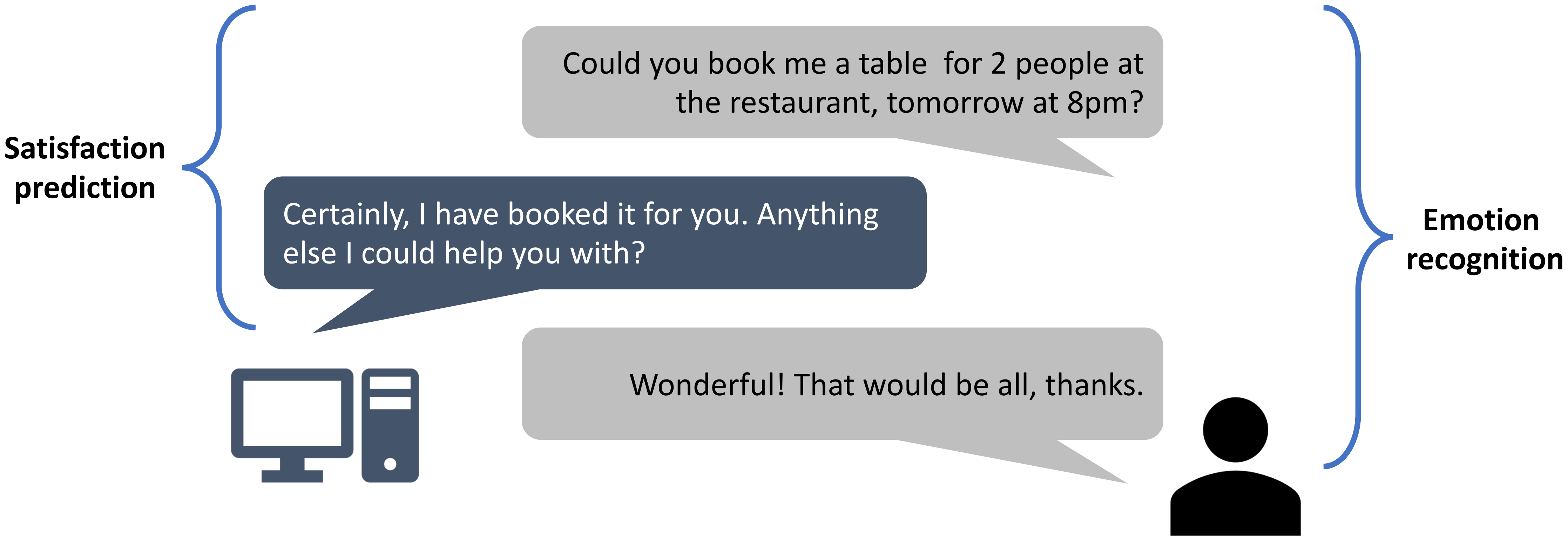
Emotion recognizer for dialogue evaluation
(Feng et al., under review)

- Goal: Leverage emotion recognition model to infer dialogue system performance
- Hypothesis: A model that can recognize emotions can also infer task success via user satisfaction

- Zero-shot prediction of user satisfaction



What’s the difference?



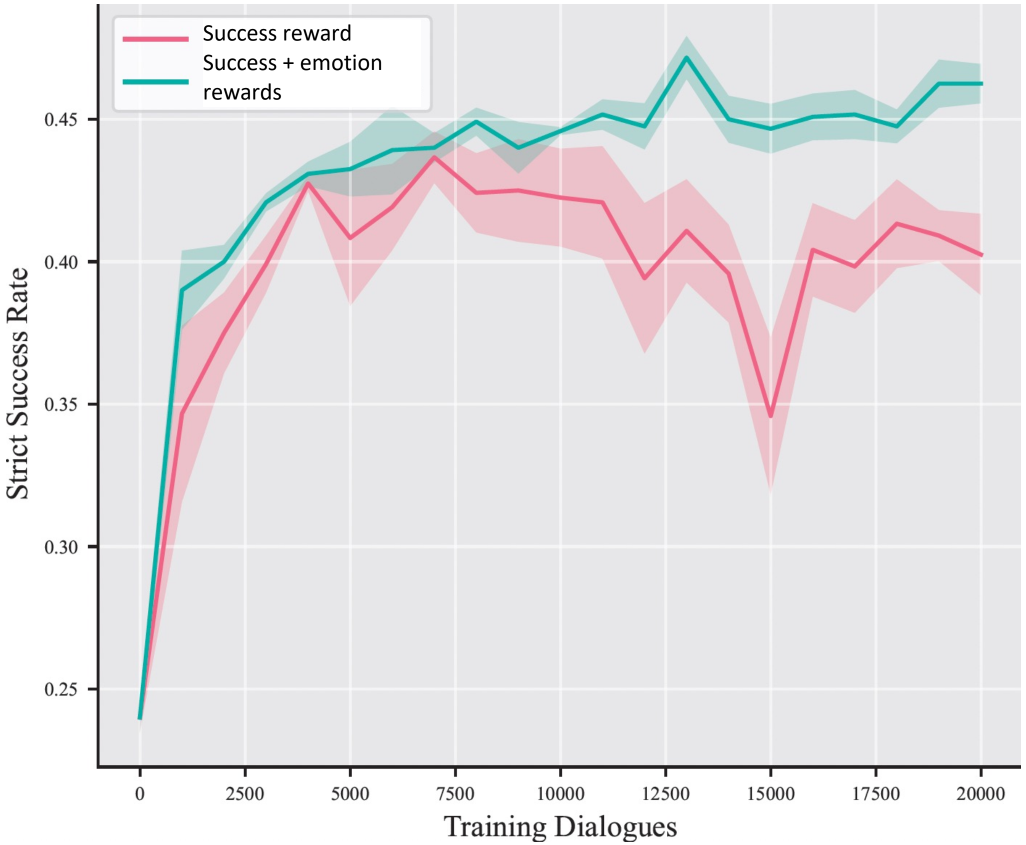
Zero-shot satisfaction prediction

(Feng et al., under review)

	JDDC	SGD	ReDial	CPPE
HiGRU (Sun et al., 2021)	17.1	8.6	8.3	27.4
BERT (Sun et al., 2021)	18.5	4.8	12.5	24.5
SatAct (Kim and Lipani, 2022)		71.3		16.5
SatActUtt (Kim and Lipani, 2022)		**84.7**		73.4
Zero-shot ERToD	**50.1**	78.8	**78.1**	**77.6**

ERToD achieves state-of-the-art performance on satisfaction prediction

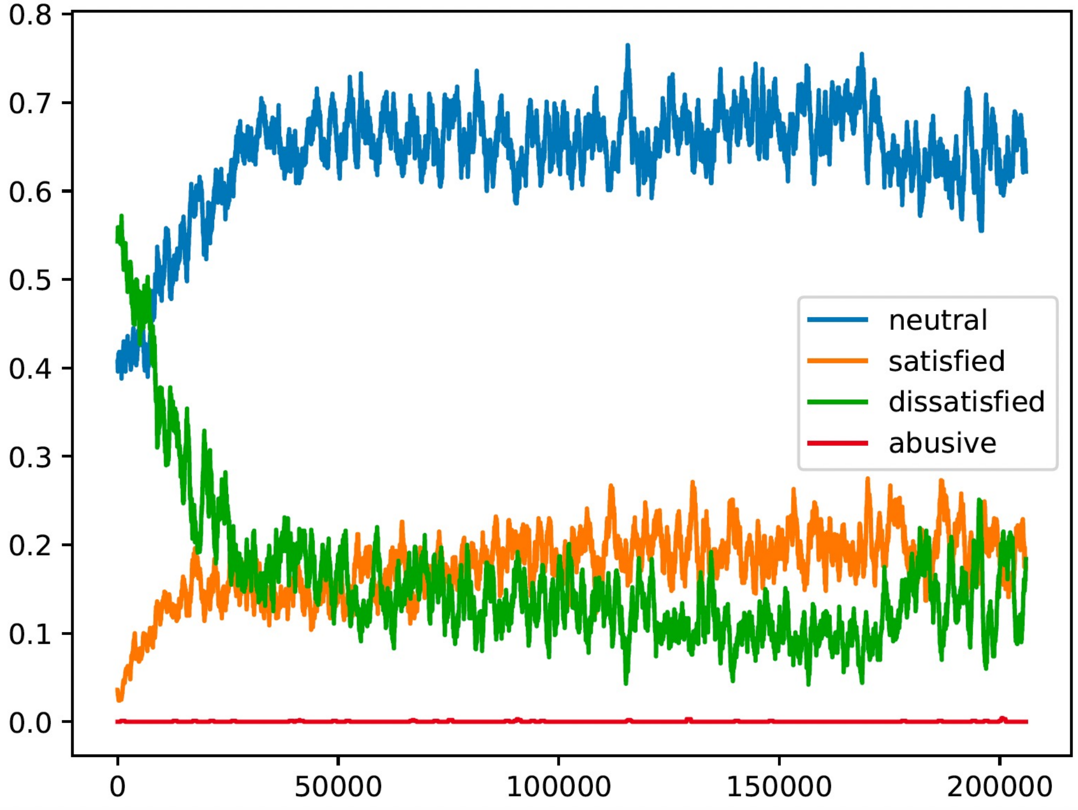
What happens if we use emotion as reward signal?
(Geishauser et al., work in progress)



Emotions provide useful signal to help improve task success

Positive emotion correlates with task success

(Geishauser et al., work in progress)



As task success improves, “satisfied” emotion becomes more probable and “dissatisfied” less

Conclusion

Conclusion

- We propose utilizing **offline RL for dialogue evaluation**
  - Towards solving current challenges of dialogue evaluation
  - Efficient and reliable metric
  - Strong correlation with human judgments
- **Emotional signal** can serve as a proxy to user satisfaction
  - Evaluate user satisfaction with zero-shot satisfaction prediction
  - Emotion as reward signal improves dialogue success
  - A universal, ontology-independent signal
- **Wide-range** of applications
  - Chat-oriented system
  - Human annotations as reward signal

hhu
Heinrich Heine
Universität Düsseldorf

Alexander von Humboldt
Stiftung / Foundation

Google Cloud

erc
European Research Council
Established by the European Commission

Thank you