Empathetic Dialogue Agent Trained with Reinforcement Learning

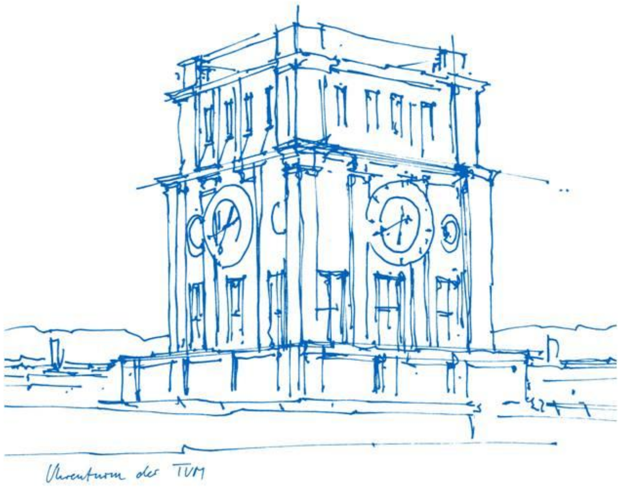
Ananta Bhattarai, Sophie Schoen, Viviane Rehor

Technische Universität München

Department of Informatics

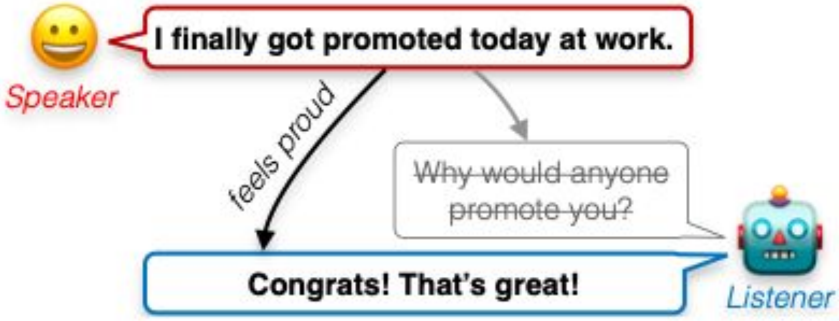
Research Group Social Computing

Munich, July 13 2021

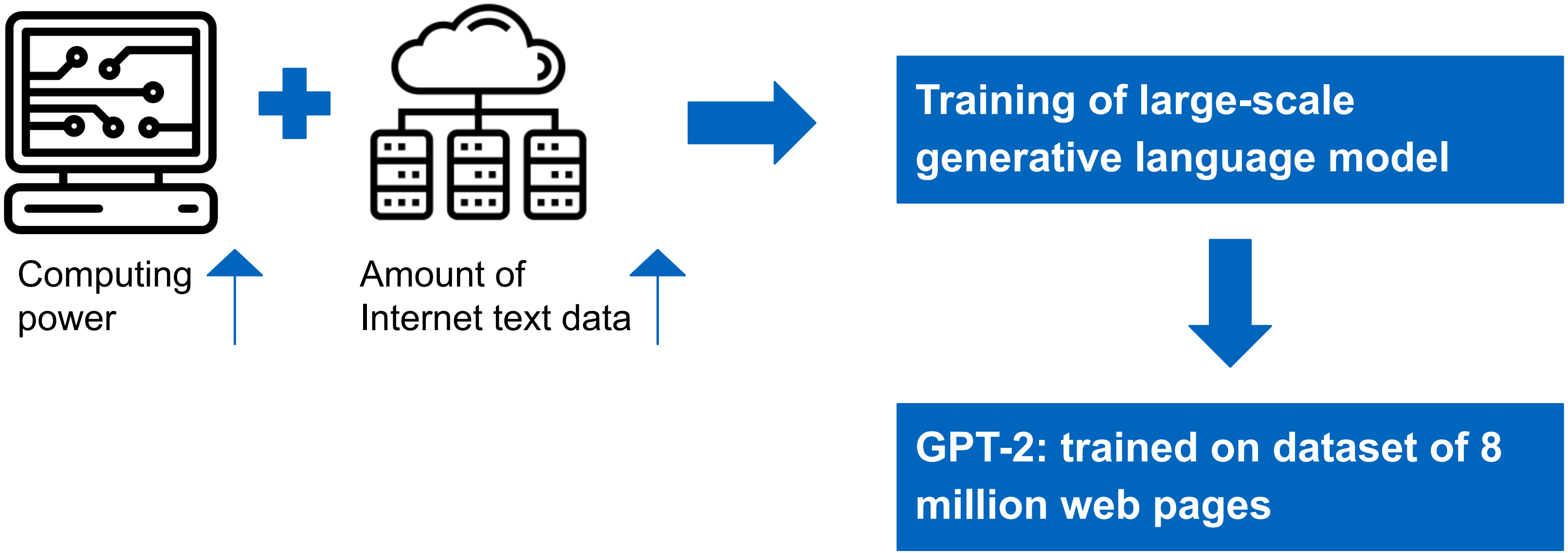


Motivation

Why focusing on the
Empathy
of the dialog model?


Speaker
I finally got promoted today at work.
feels proud
~~Why would anyone promote you?~~
Congrats! That's great!
Listener

Related Work



Computing power

Amount of Internet text data

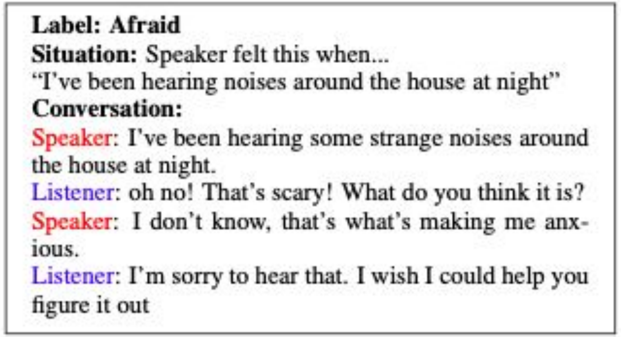
Training of large-scale generative language model

GPT-2: trained on dataset of 8 million web pages

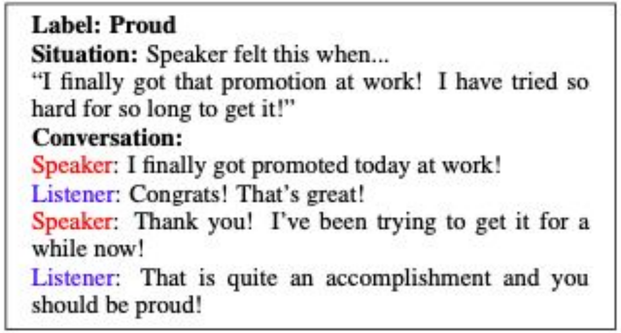
Empathetic Dialogue Dataset

= dataset of 25k conversations
based on emotional situations
→ produces more empathetic responses

Representation of empathetic conversations:

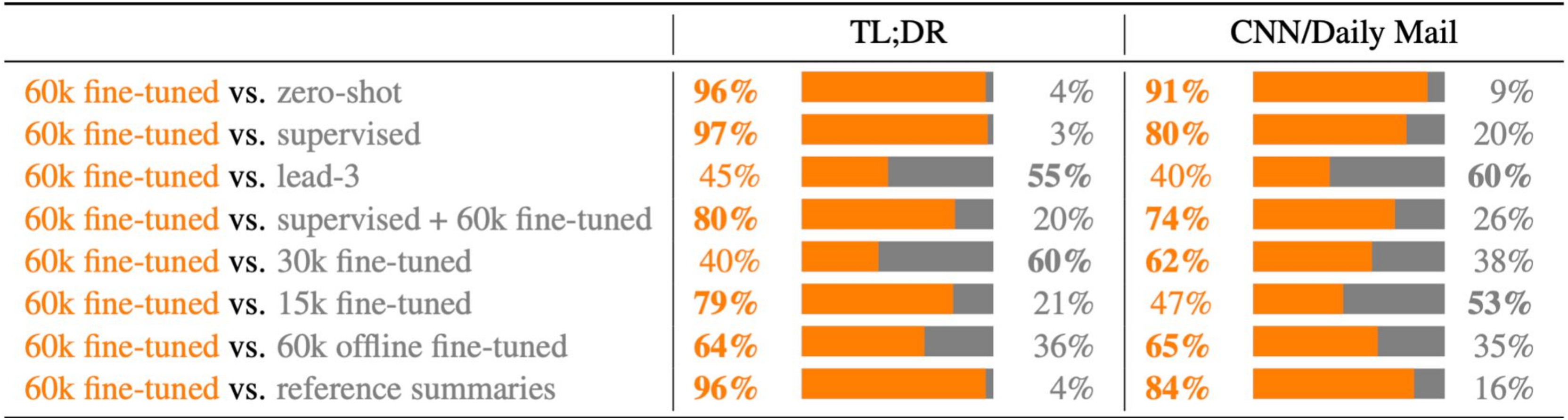


Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out



Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Training with human feedback

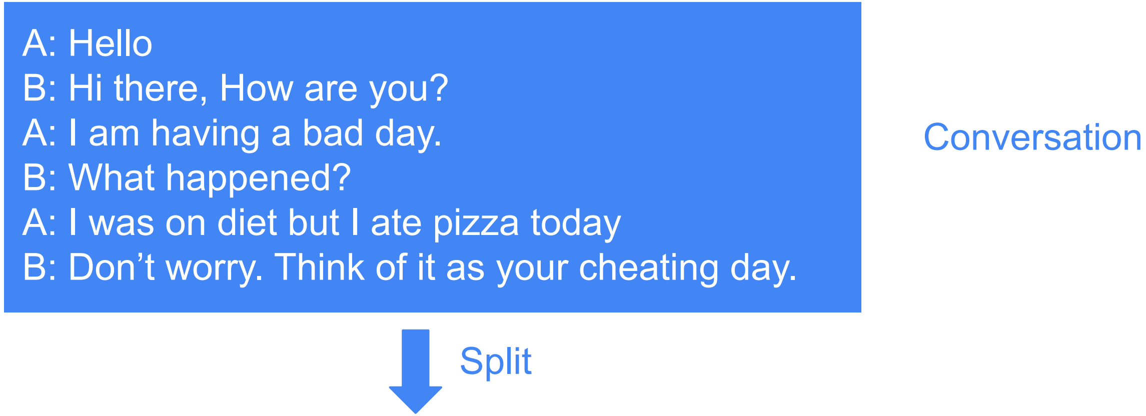


	TL;DR		CNN/Daily Mail	
60k fine-tuned vs. zero-shot	**96%**	4%	**91%**	9%
60k fine-tuned vs. supervised	**97%**	3%	**80%**	20%
60k fine-tuned vs. lead-3	45%	**55%**	40%	**60%**
60k fine-tuned vs. supervised + 60k fine-tuned	**80%**	20%	**74%**	26%
60k fine-tuned vs. 30k fine-tuned	40%	**60%**	**62%**	38%
60k fine-tuned vs. 15k fine-tuned	**79%**	21%	47%	**53%**
60k fine-tuned vs. 60k offline fine-tuned	**64%**	36%	**65%**	35%
60k fine-tuned vs. reference summaries	**96%**	4%	**84%**	16%

RL fine-tuned models with human feedback (60k fine-tuned) are preferred more than the other models.

Data Pre-Processing

- Training Examples: Use training set of Empathetic dialogues dataset to create a context-response pair



A: Hello
B: Hi there, How are you?
A: I am having a bad day.
B: What happened?
A: I was on diet but I ate pizza today
B: Don’t worry. Think of it as your cheating day.

Conversation

Split

Training 1: Hello </s> (Context) → response: Hi there, How are you?
Training 2: Hello <SOC> Hi there, How are you? </s> (Context) → response:I am having a bad day
Training 3: Hello <SOC> Hi there, How are you? <SOC> I am having a bad day </s> (Context)
...

- Inference Step: Create contexts from the validation set and sample two responses to each context.

Data Annotation - Annotation Criteria

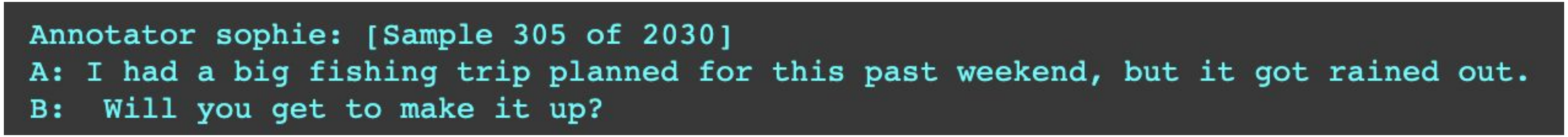
Empathy/Sympathy

Relevance

Fluency

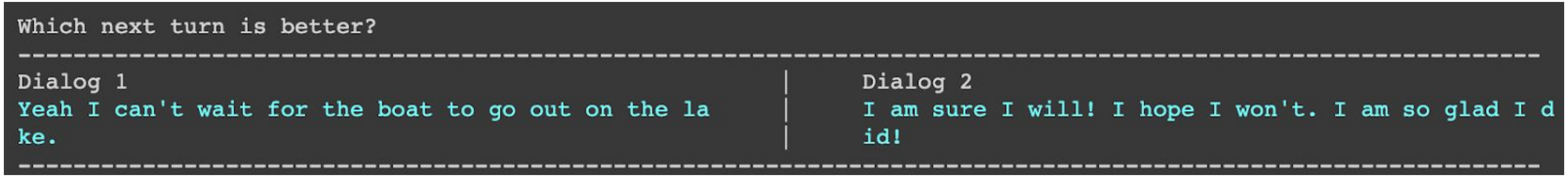
Data Annotation - Labeling Script

Input

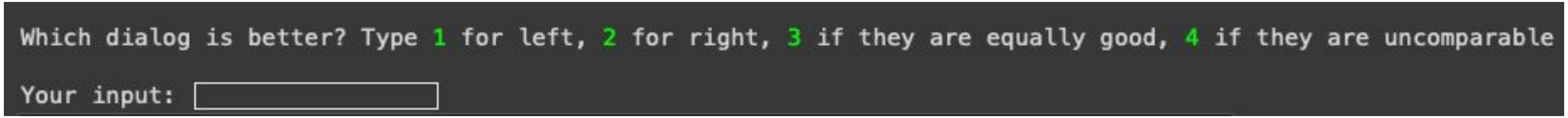


```
Annotator sophie: [Sample 305 of 2030]
A: I had a big fishing trip planned for this past weekend, but it got rained out.
B:  Will you get to make it up?
```

Output



```
Which next turn is better?
-------------------------------------------------------------------------------------
Dialog 1                                     |      Dialog 2
Yeah I can't wait for the boat to go out on the la  |      I am sure I will! I hope I won't. I am so glad I d
ke.                                          |      id!
-------------------------------------------------------------------------------------
```



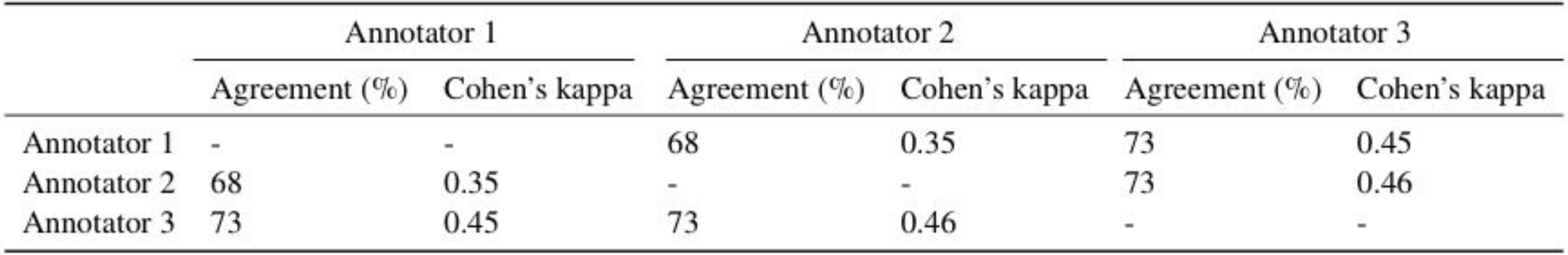
```
Which dialog is better? Type 1 for left, 2 for right, 3 if they are equally good, 4 if they are uncomparable

Your input: [                    ]
```

Options for Interaction:

- 1 / 2 : Left or right sample is better
- 3 : Both are good, so random choice
- 4 : Both samples do not fit → kick them out
- 5 / 6 : Copy part of left or right sample to console

Data Annotation - Agreement

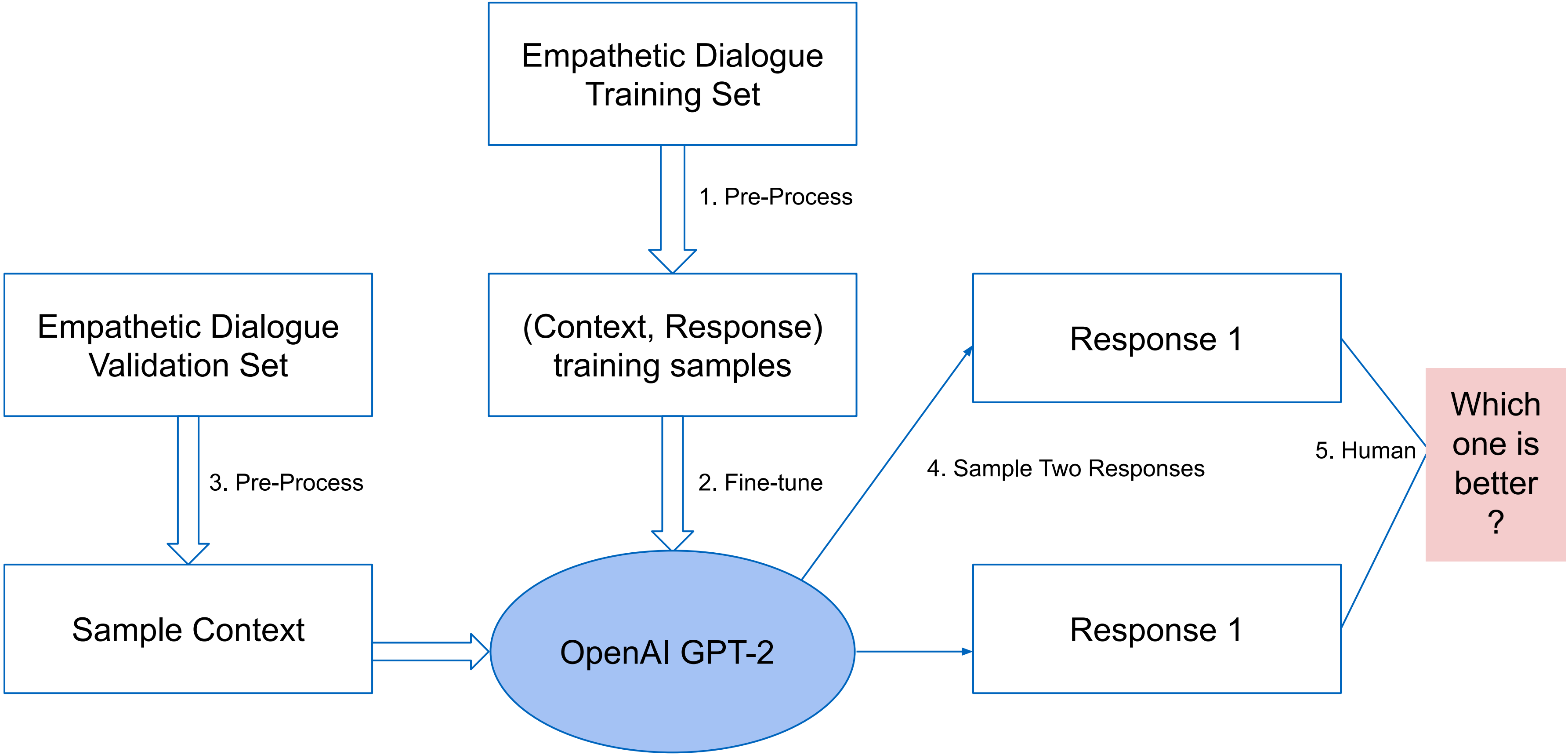


	Annotator 1		Annotator 2		Annotator 3	
	Agreement (%)	Cohen's kappa	Agreement (%)	Cohen's kappa	Agreement (%)	Cohen's kappa
Annotator 1	-	-	68	0.35	73	0.45
Annotator 2	68	0.35	-	-	73	0.46
Annotator 3	73	0.45	73	0.46	-	-

Table 1 Confusion matrix showing agreement percentage and Cohen's kappa score between the annotators

→ reasonable agreement result

Method: Supervised Training and Inference



Empathetic Dialogue Training Set

1. Pre-Process

(Context, Response) training samples

2. Fine-tune

Empathetic Dialogue Validation Set

3. Pre-Process

Sample Context

OpenAI GPT-2

4. Sample Two Responses

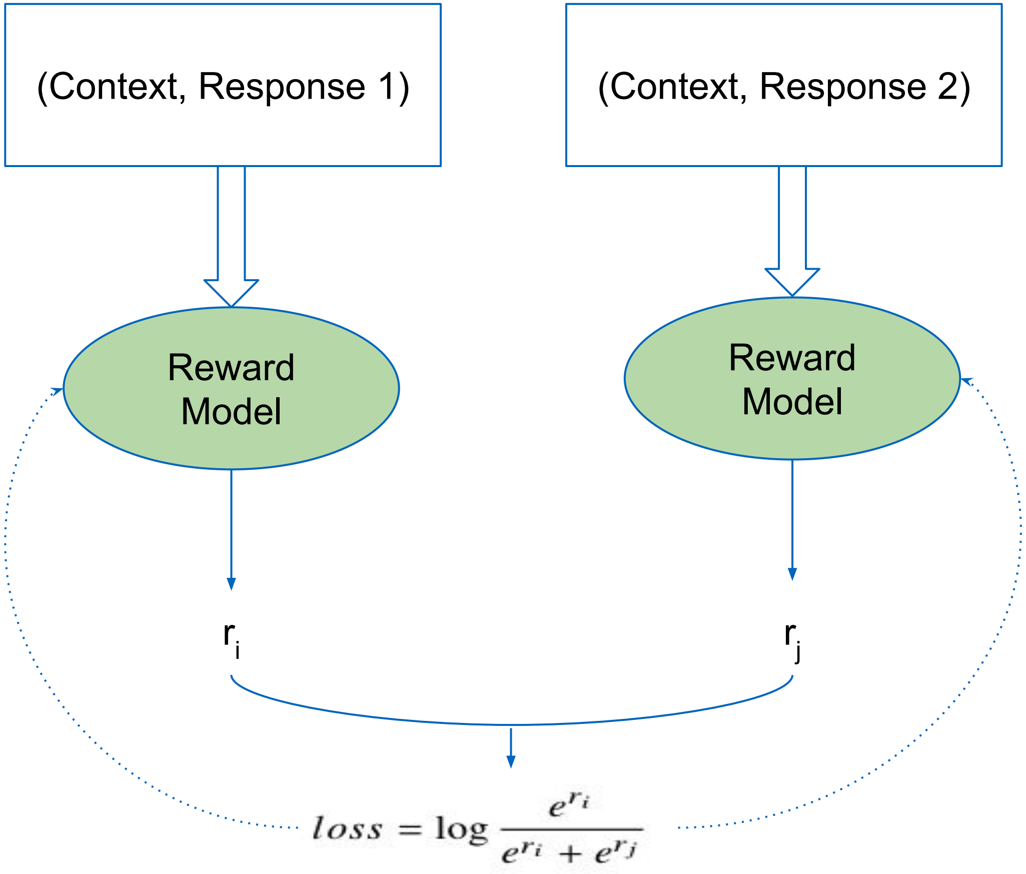
Response 1

Response 1

5. Human

Which one is better ?

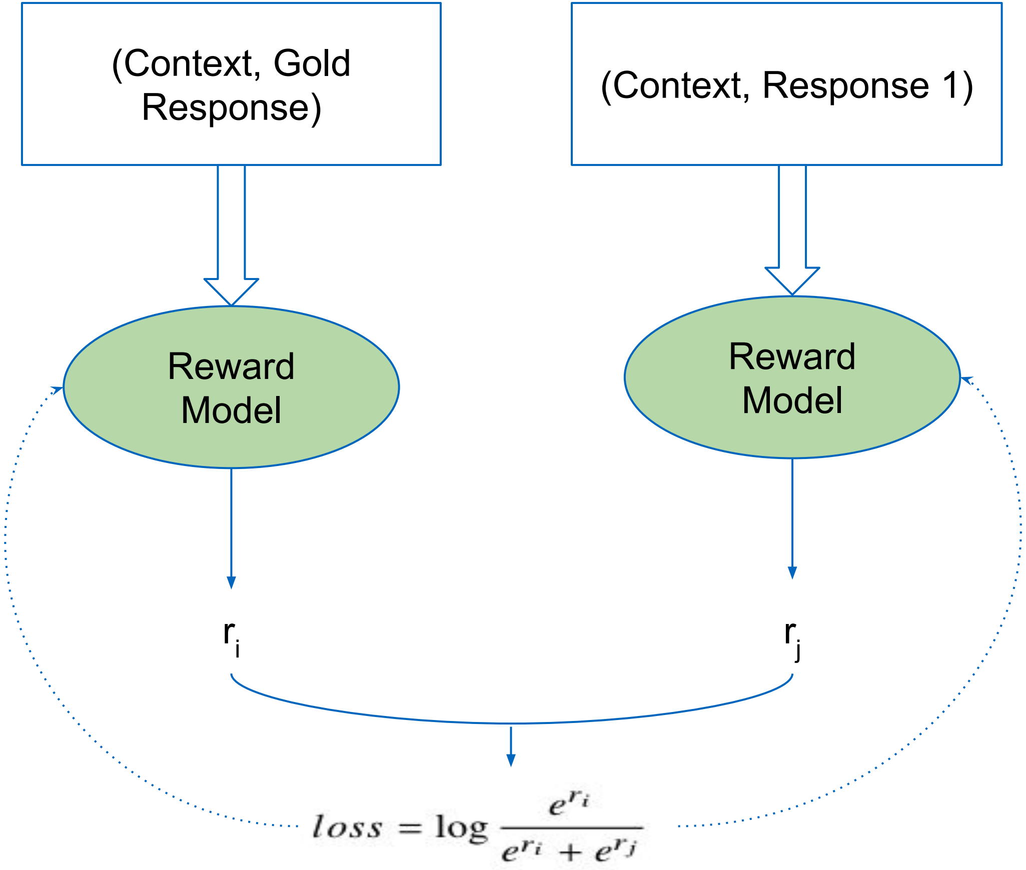
Method: Reward-1 Training



$$loss = \log \frac{e^{r_i}}{e^{r_i} + e^{r_j}}$$

r_i is better than r_j

Method: Reward-2 Training

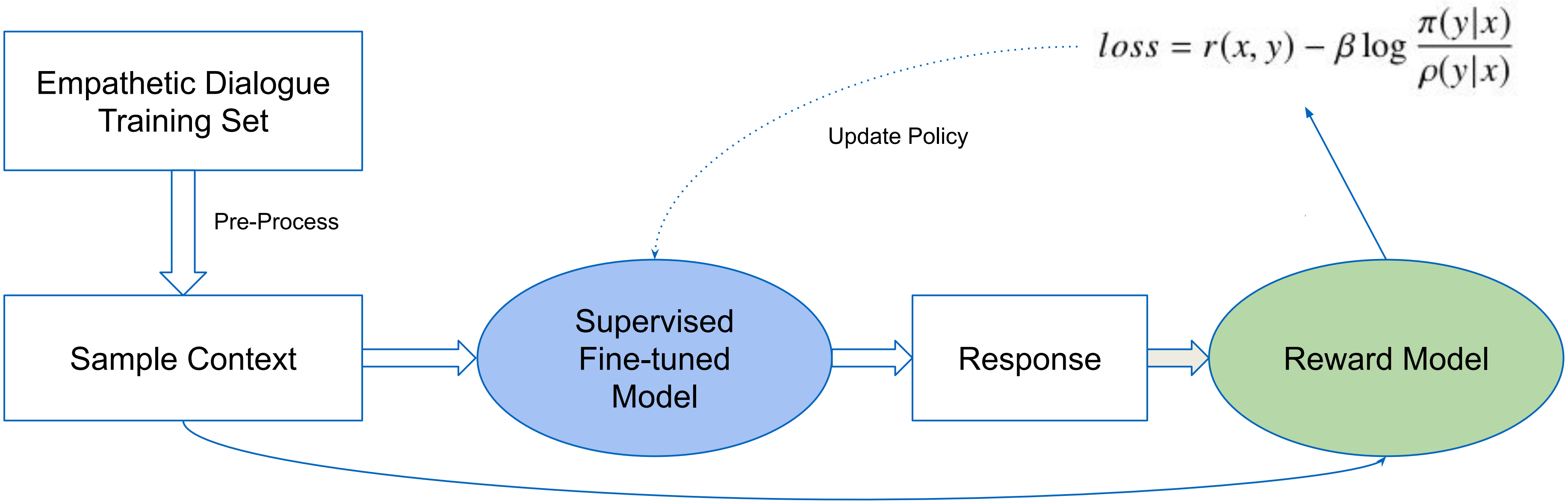


Validation Set
Gold Response

r_i is better than r_j

$$loss = \log \frac{e^{r_i}}{e^{r_i} + e^{r_j}}$$

Method: Policy Training



$$loss = r(x, y) - \beta \log \frac{\pi(y|x)}{\rho(y|x)}$$

Empathetic Dialogue Training Set

Pre-Process

Sample Context

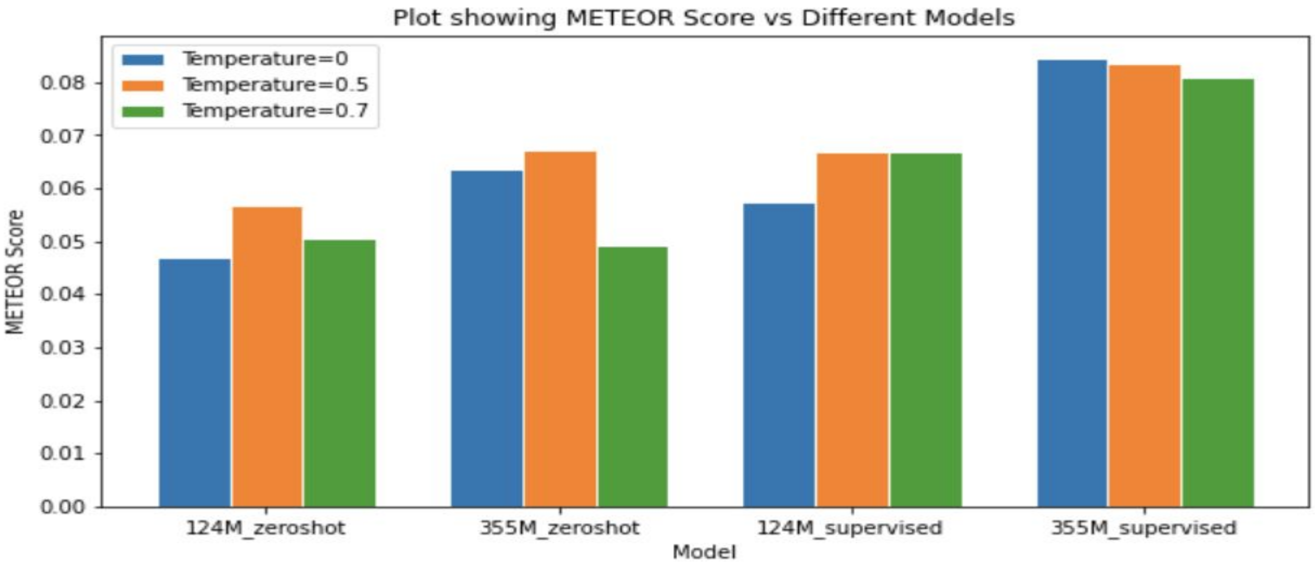
Supervised Fine-tuned Model

Response

Reward Model

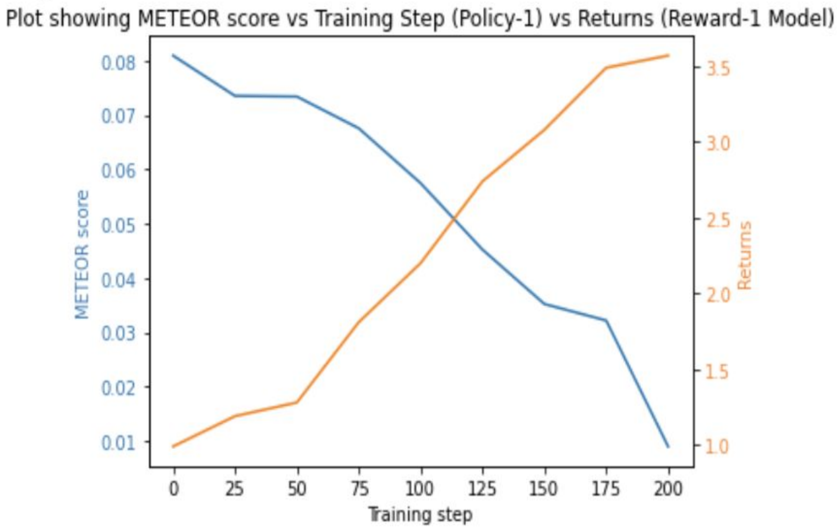
Update Policy

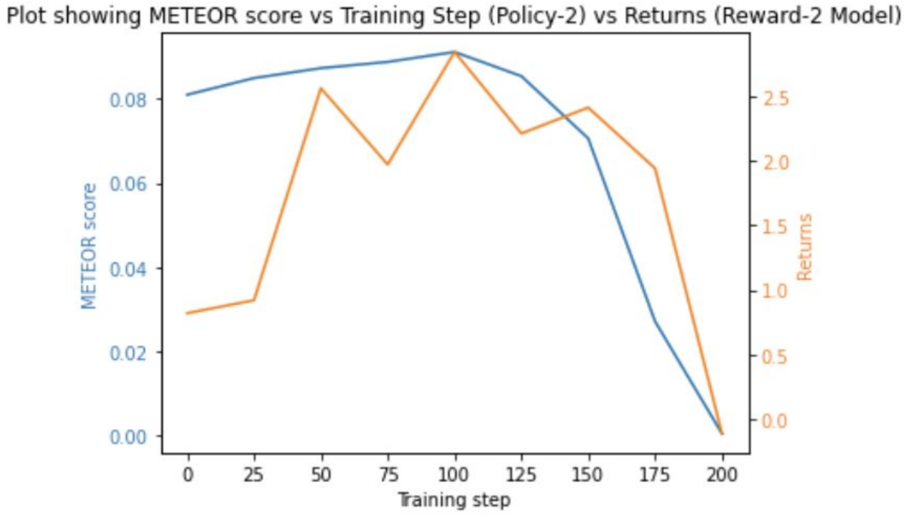
Supervised Baselines



- Comparing sampled responses for given context in the validation set with gold responses in the validation set using METEOR score for different models.
- Fine-tuned 355M OpenAI GPT-2 achieves a higher METEOR score. Therefore, we continue from fine-tuned 355M for the reward and policy training.

Policy-1 vs Policy-2 Evaluation





- For policy-1, returns of the model increases whereas the METEOR score decreases, as we optimize the policy. This shows reward-1 model didn't learn anything useful from less annotated data (2212).
- For policy-2, returns of the model increase in the first few steps of policy training and then start to decrease. The same trend is observed with the METEOR score. This shows as we diverge too far from the initial policy, the model is over-optimized and indicates overfitting.

Evaluation Metrics

- Meteor

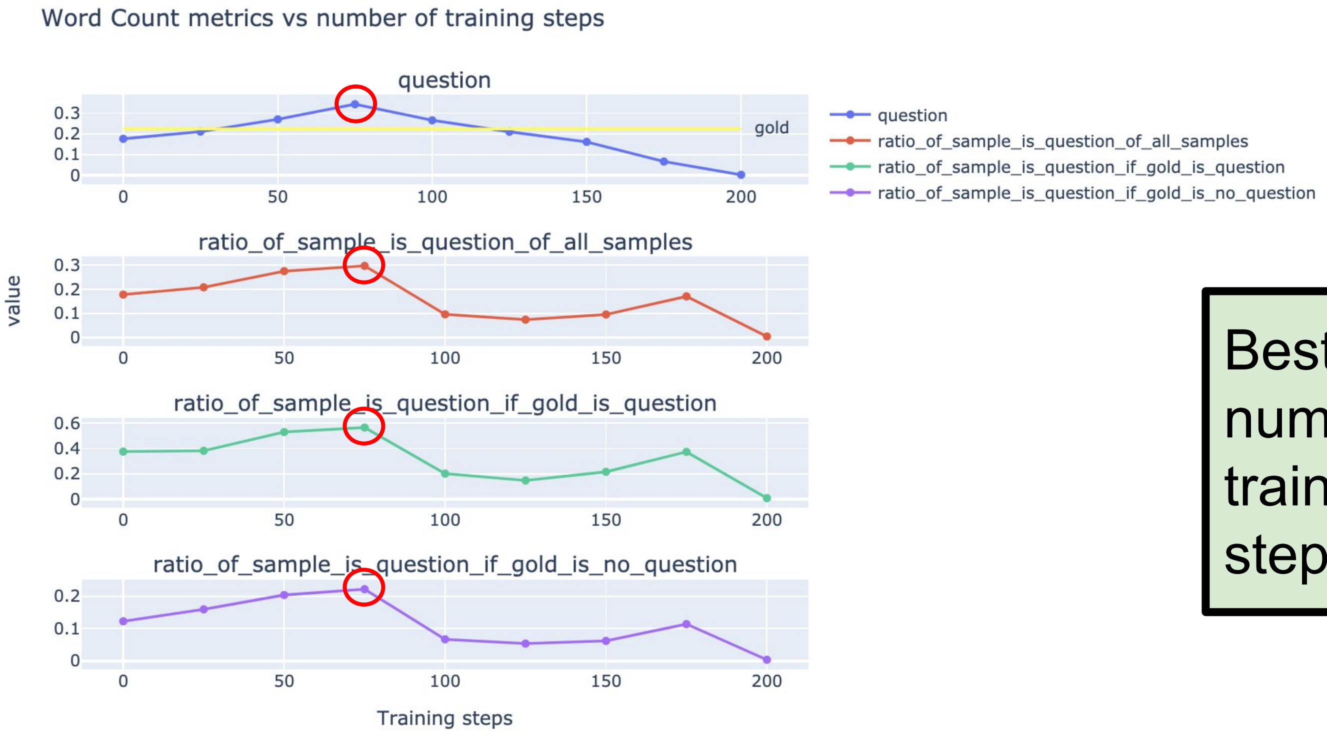
Metric’s from Florian’s Thesis:

- Utterance length
- Self repetition
- Utterance repetition
- Word repetition
- Conversation repetition
- Emotional reaction level
- Interpretation level
- Exploration level

} **Empathy metrics**

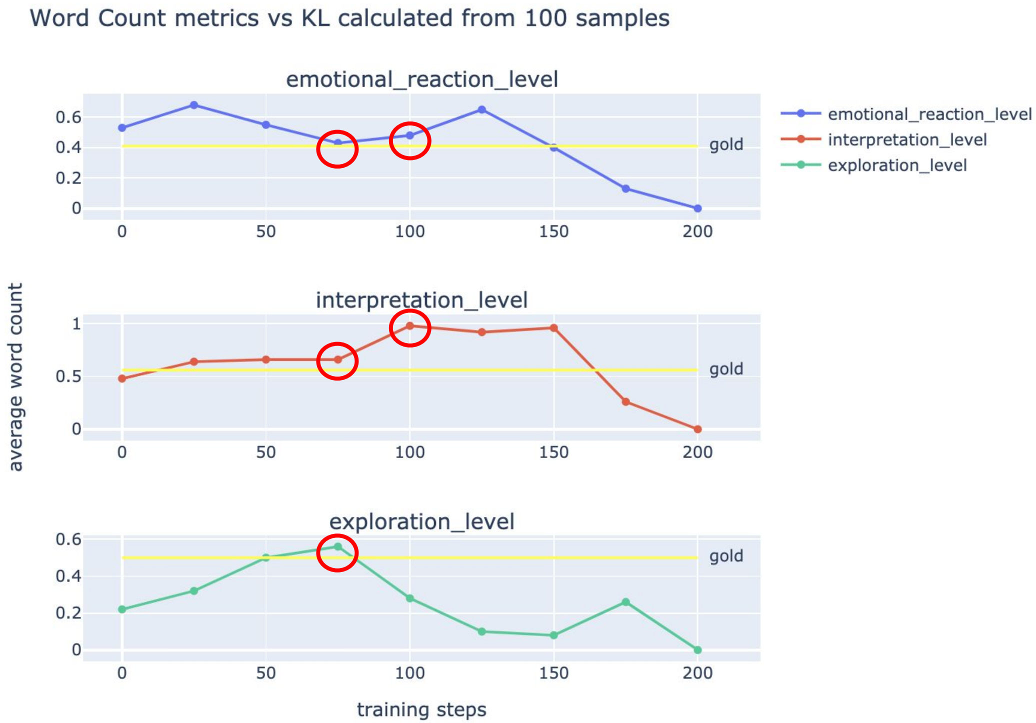
- Question (1 else 0 -> avg)
- Question ration of all samples
- If gold is question ratio of sample questions
- If gold is no question ration of sample questions

Question metrics (Policy 2)



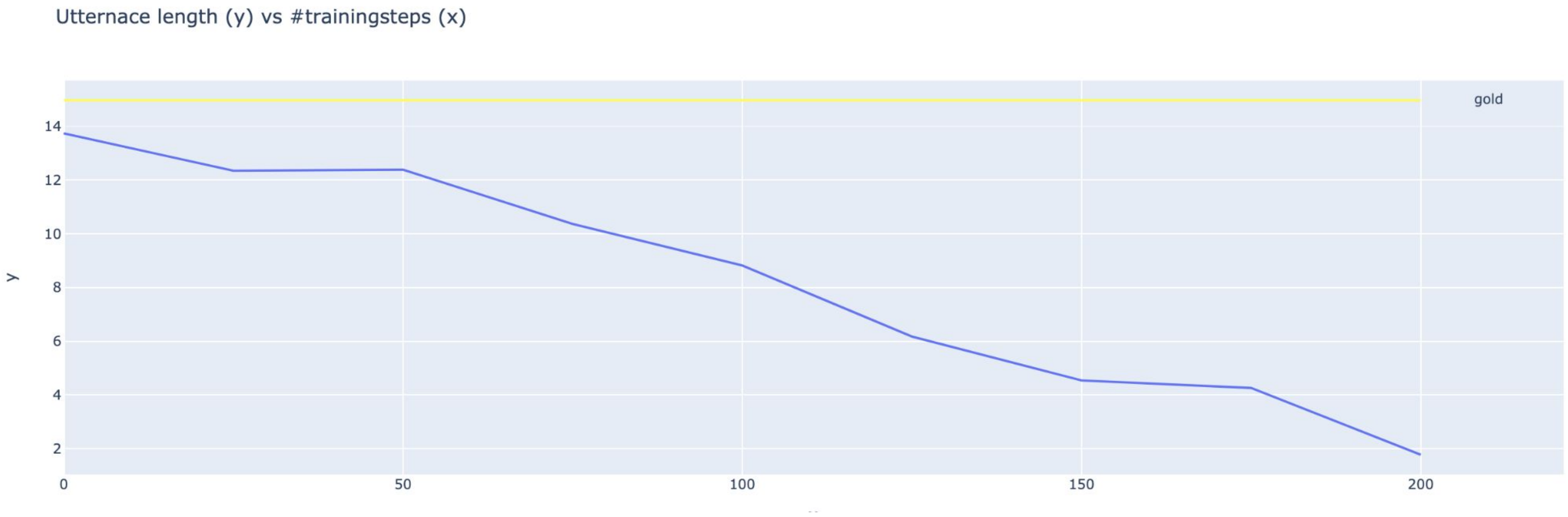
Best number of training steps: 75

Empathy metrics (Policy 2)

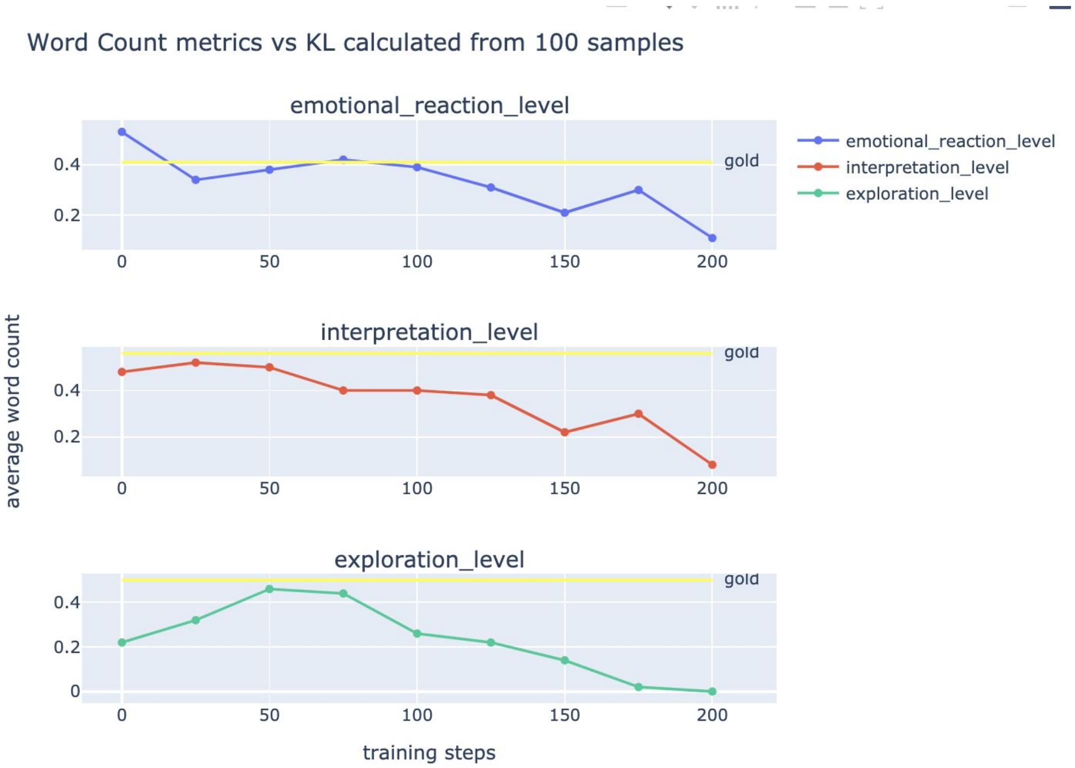


All metrics over the generated response from policy-2‘s model show that after training step **75/100** the model reached it’s best values!

Utterance length (Policy 1)



Empathy metrics (Policy 1)



All metrics over the generated response from policy-1‘s model show that the model is getting worse while training!

Which model is the best? - Human Evaluation

- Policy 2’s model after 75 training steps
  - Largest number of questions
  - Emotional reaction level closest to gold
  - Interpretation and exploration level very close to gold
  - => very human like

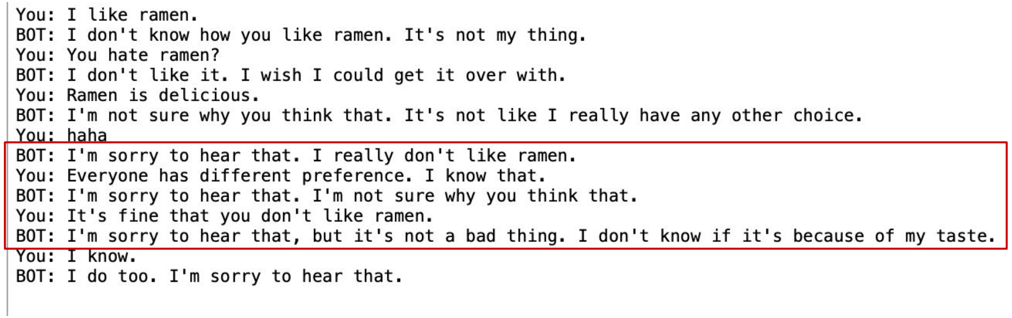
- Policy 2’s model after 100 training steps
  - Number of questions close to gold
  - Highest interpretation level (shows understanding of feelings)
  - Emotional reaction level still close to gold

Human Evaluation (50 samples each)

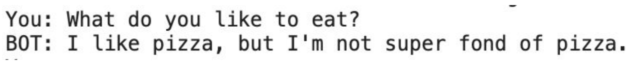
Empathy: Which model generates answer that show more understanding of feelings?		
	P75 Model	**P100 Model**
% Preference over supervised baseline	52.66	**57.34**
% Preference over state-of the art DialoGPT	**60.0**	-
Fluency: Which model generates answers that seem fluent and accurate?		
	P75 Model	**P100 Model**
% Preference over supervised baseline	**57.34**	46.66
% Preference over state-of the art DialoGPT	**76.67**	-
Relevance: Which model generates answers that seem appropriate and on-topic?		
	P75 Model	**P100 Model**
% Preference over supervised baseline	**55.34**	43.34
% Preference over state-of the art DialoGPT	**67.33**	-

Conclusion & Future Ideas

- Policy 2’s model shows good results after 75 training steps
- Improvement ideas:
  - Repetition in generated answers
  - Contradiction in gen. answers
  - Safety issues, work against toxic responses
- Policy 1’s with a lot more annotated data
- Start fine-tuning directly with DialoGPT (already trained on reddit Dialogue dataset)



```
You: I like ramen.
BOT: I don't know how you like ramen. It's not my thing.
You: You hate ramen?
BOT: I don't like it. I wish I could get it over with.
You: Ramen is delicious.
BOT: I'm not sure why you think that. It's not like I really have any other choice.
You: haha
BOT: I'm sorry to hear that. I really don't like ramen.
You: Everyone has different preference. I know that.
BOT: I'm sorry to hear that. I'm not sure why you think that.
You: It's fine that you don't like ramen.
BOT: I'm sorry to hear that, but it's not a bad thing. I don't know if it's because of my taste.
You: I know.
BOT: I do too. I'm sorry to hear that.
```



```
You: What do you like to eat?
BOT: I like pizza, but I'm not super fond of pizza.
```

Thank You for your attention

Do you have any questions?

References

- Hannah Rashkin et al.Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. 2019. arXiv: 1811.00207[cs.CL].
- Nisan Stiennon et al.Learning to summarize from human feedback. 2020. arXiv: 2009.01325[cs.CL].
- Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2019).
- Yizhe Zhang et al.DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. 2020. arXiv: 1911.00536[cs.CL].
- Daniel M. Ziegler et al.Fine-Tuning Language Models from Human Preferences. 2020. arXiv: 1909.08593[cs.CL].
- Florian von Unhold. “State-Based Dialog Modeling Using Human Preferences”. 2021.

Backup Slides

(Context, Response) samples of our model

Context	Response
A: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road.	That's terrible. I'm sorry to hear that. Was it your fault?
A: I there, dont know what to do, jst broke up with my girlfirned, we were 8 years together B: sorry to hear! do you have any idea about the break up? did you think about it ?	I did, but i wasnt sure what to do, i was really upset about it.
A: my husband lost a job but i'm hoping he can find a full time job soon B: He will, I have faith. A: thank you so much! B: No problem. What kind of work does he do?	I think he does a lot of work in the field of IT.
A: i'm so excited because i'm finally going to visit my parents next month! I didn't see them for 3 years B: 3 years is a long time. How come? A: I moved to the US and they stayed in my homecountry	Oh my gosh, I hope you're really excited about it. I'm sure it will be a great experience for you.

Note: Contexts are taken from the test set.

Conversation sample of our model with human

YOU: Hi! How are you?! :)

BOT: I'm doing fine, I just need some time to recover from my injury.

YOU: What happened to you?

BOT: I broke my ankle and it looked like it was going to be a long recovery.

YOU: I am so sorry. How long do you need to rest?

BOT: I need a lot of time to recover from the injury. I'm not sure if I will be able to return to the field for the rest of the season.

YOU: You will. Don't worry. Don't lose hope.

BOT: I hope so too. I'm just thankful for the doctors and everything else I have to work with.

YOU: Yeah. Doctors are great.

BOT: I know. I'm just thankful that I'm able to keep going. I know I will be fine.

YOU: You will be. Do you have any plans for today?

BOT: I have a long day tomorrow. I'm going to be working on my leg, too. I'm really feeling it now.

YOU: I see. I am sad though.

BOT: I know. I just want to be able to get back to work.

YOU: You will.

BOT: I hope so. I think I am going to be okay.

YOU: What's your favourite movie?

BOT: I just like the ones that are about the frontiersman.

YOU: Do you have any specific one that you like?

BOT: I do. I think I like the one where the main character is a mountain man.

YOU: I haven't seen that movie. What's the name of the movie?

BOT: I think it's called "Mountain Man, The Biggest of the Big."

YOU: I will definitely watch it.

BOT: I'm sure you'll enjoy it. I'm not sure if I agree with the movie's ending, but I'm sure I'll like it.

Reward-2 Evaluation

Context	Response	Reward
A: I've been hearing some strange sounds around the house at night lately.	What are they, and how bad are they, is it a ghost or something?	+14.17
A: We went to our local county fair a couple of days ago. B: I love the fair! Did you have a good time? C: We did. We listened to music, rode some rides.	That's cool. I love fairs.	+11.48
A: I am so ready for this semester B: do you go to college? A: Yes I am a senior this yar B: That's good. I am sure your parents are proud	I think so.	+2.05
A: I lent my friend 500$ and she never paid me back. Ugh. I am too nice sometimes.	I hope she didn't go to hell.	-4.96
A: Last week we had a fire start in our bathroom.	-	-5.27

Reward-1 Evaluation

Context	Response	Reward
A: YI still can't believe my parents for divorced.	What happened?	+9.93
A: My friend is coming in from out of town today.	Oh wow, I'm very excited. Who is he?	+9.32
A: A few days ago, I was looking through a bunch of old photo albums. B: Were you looking for something in particular or just for fun? A: Just for fun. It brought back a lot of good memories!	Did you get anything?	+3.11
A: It was a tough day at work B: Awe, I am sorry. What made it tough?	I am sorry	+6.41
A: So my dog passed away a couple of weeks ago. Its been really hard not having him around. B: Oh no! How old was he?	-	+11.88

Hyperparameters (Supervised Model Training)

GPT-2 355M Training Details (Until 146 Steps):

- Batch Size: 4 * 5
- Patience = 10
- Steps = 146
- Lr = 0.00001
- History_max = 4
- Context_max = 100
- Response_max = 100

GPT-2 124M Training Details:

- Batch Size: 16 * 5
- Patience = 10
- Steps = 282
- Lr = 0.00001
- History_max = 4
- Context_max = 100
- Response_max = 100

Hyperparameters (Supervised Model Training)

GPT-2 355M Training Details (After 146 Steps):

- Batch Size: 8 * 5
- Patience = 50
- Steps = 594
- Lr = 0.000007
- History_max = 4
- Context_max = 100
- Response_max = 100

Hyperparameters (Reward Model Training)

Reward-1 Training Details:

- Batch Size: 4
- Training Samples = 2212
- Lr = 1.5e-5
- Label Type = Best of 2
- Normalize After = True
- Normalize Before = True
- Normalize Samples = 256

Reward-2 Training Details:

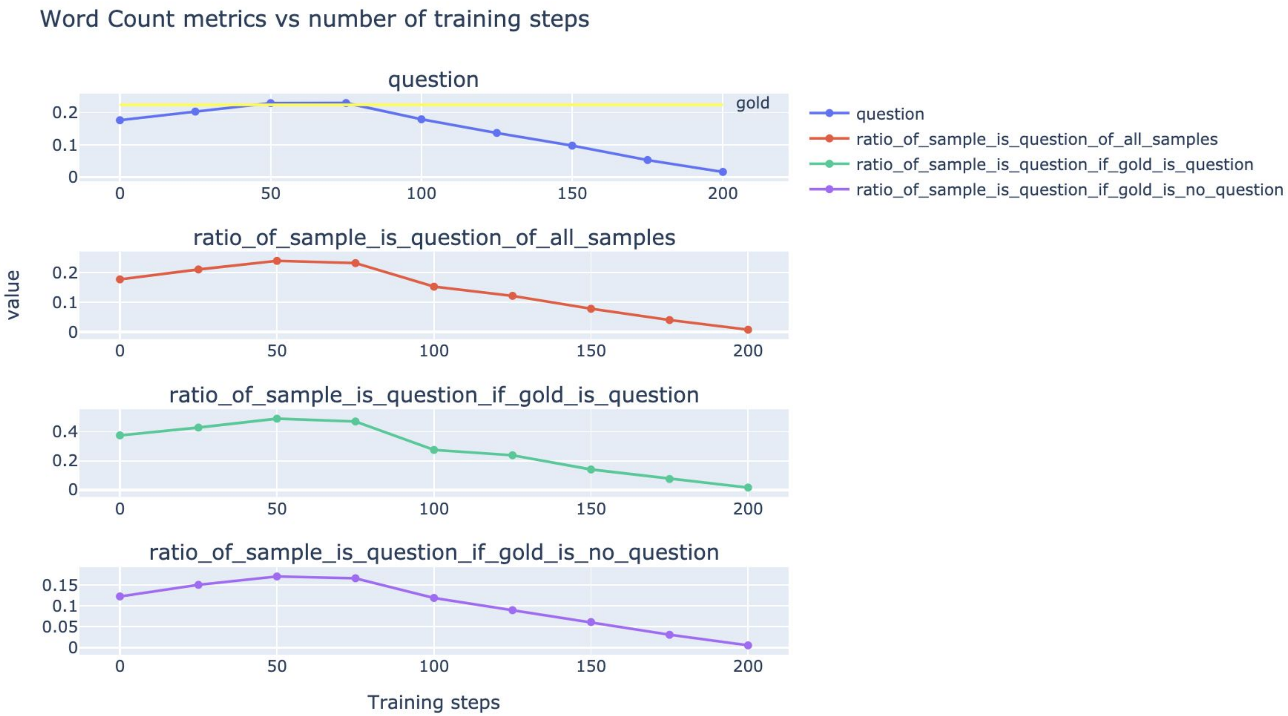
- Batch Size: 4
- Training Samples = 9280
- Lr = 1.5e-5
- Label Type = Best of 2
- Normalize After = True
- Normalize Before = True
- Normalize Samples = 256

Hyperparameters (Policy Training)

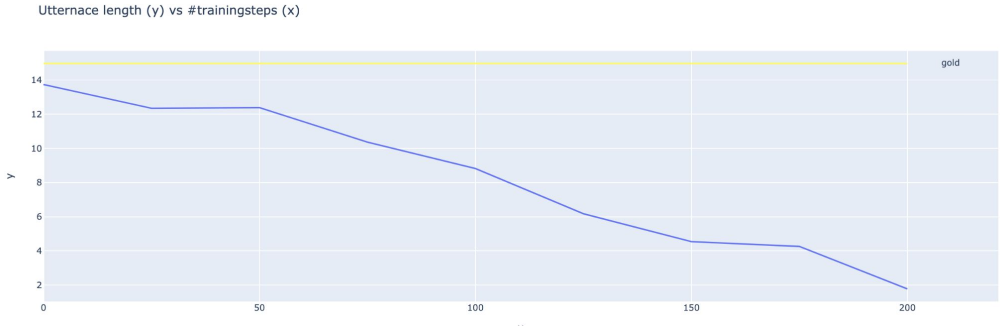
Policy Training Details:

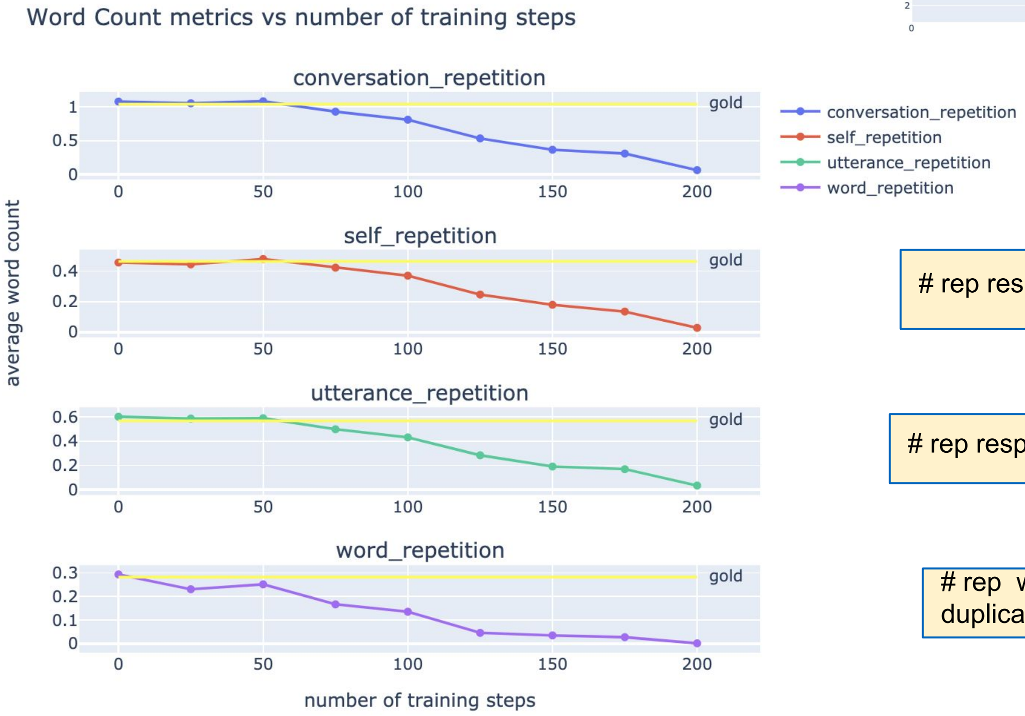
- Batch Size: 5
- KL coefficient = 0.03
- Lr = 2e-6
- Total episodes = 1000
- Training Step = 200

Question metrics (Policy 1)



Word count metrics (Policy 1)


Utternace length (y) vs #trainingsteps (x)


Word Count metrics vs number of training steps

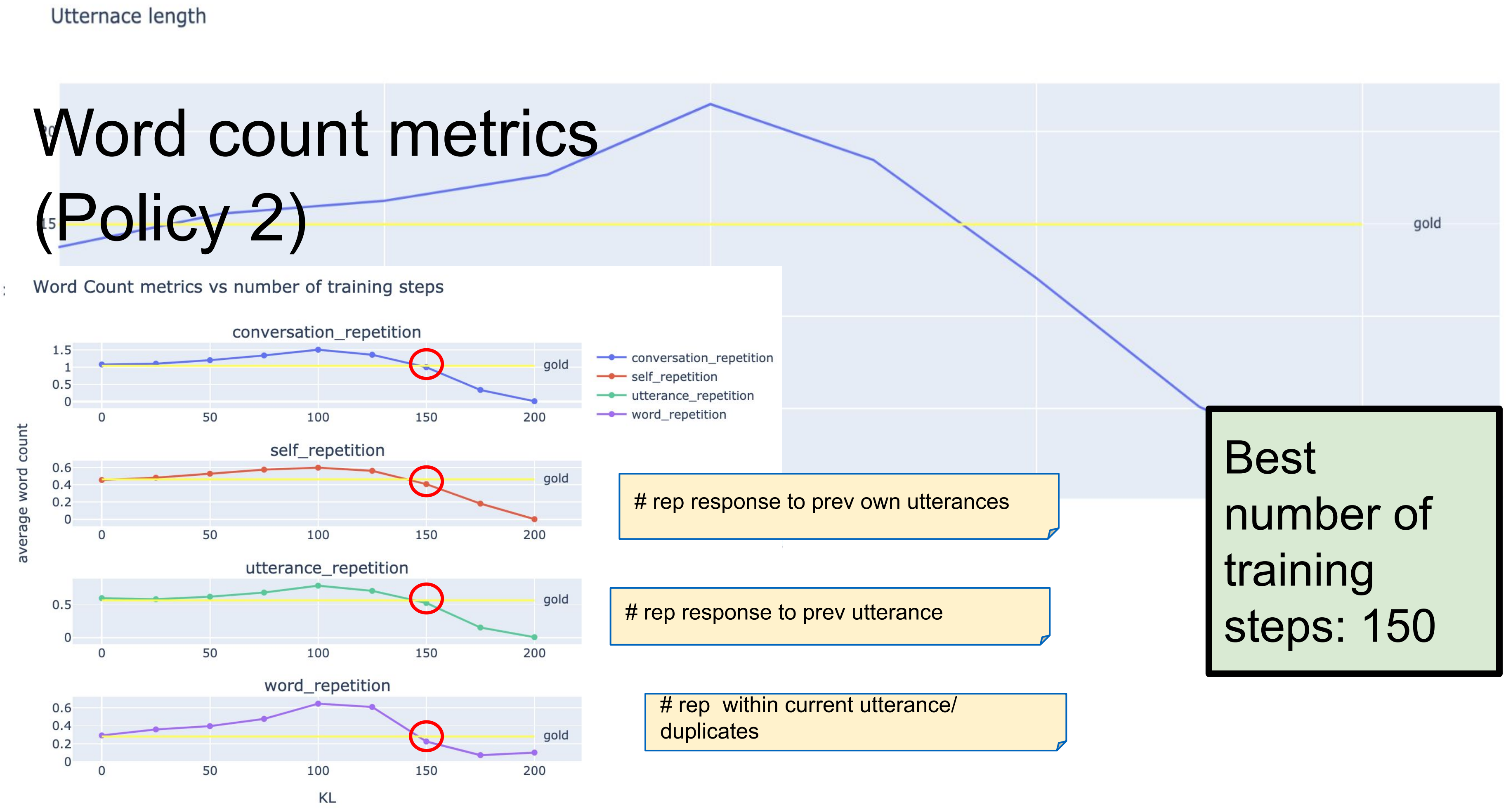
# rep response to whole context

# rep response to prev own utterances

# rep response to prev utterance

# rep within current utterance/
duplicates

Word count metrics (Policy 2)



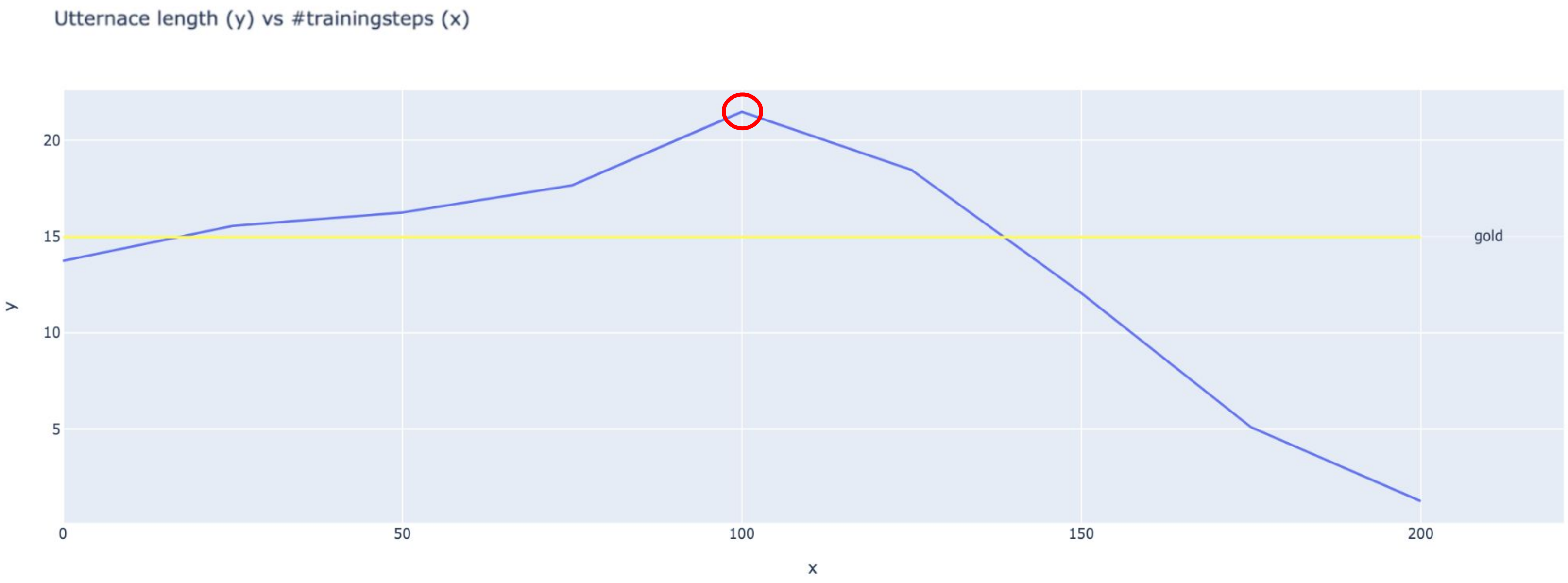
# rep response to prev own utterances

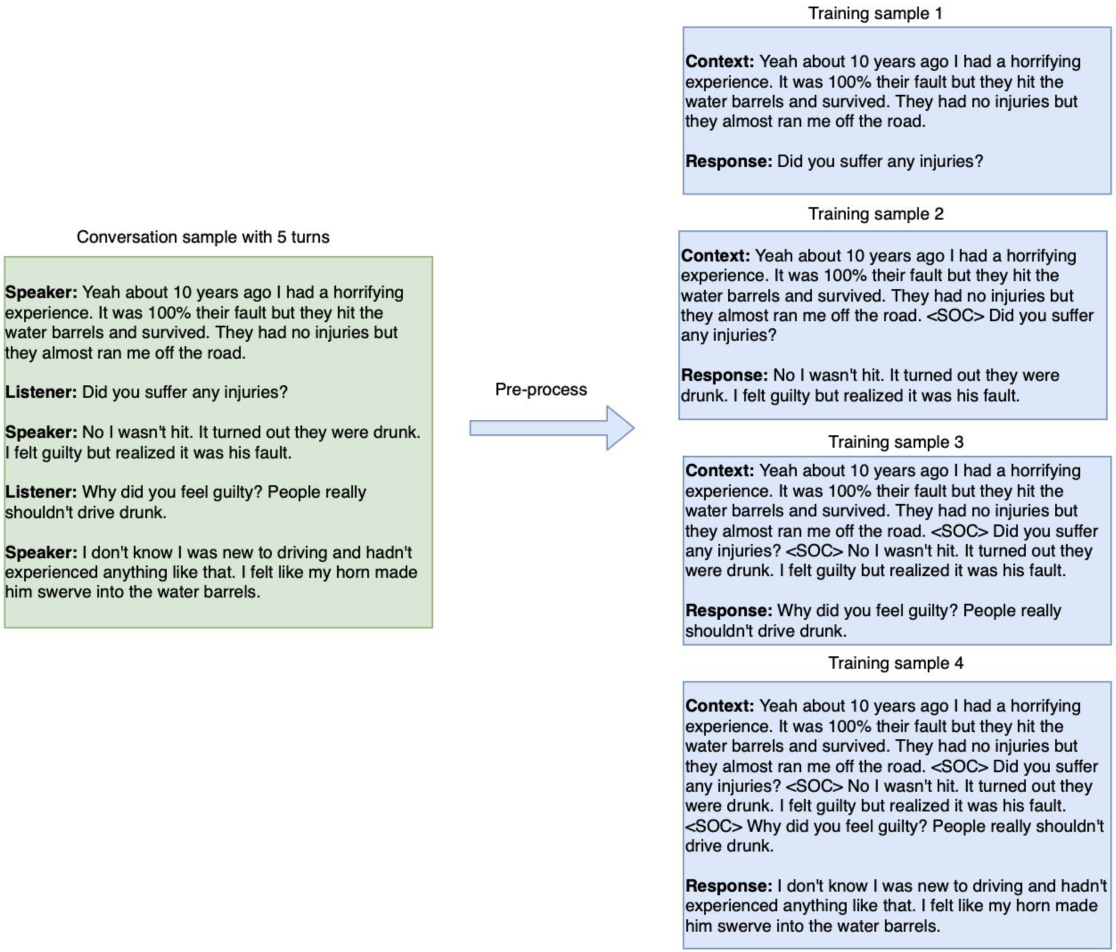
# rep response to prev utterance

# rep within current utterance/ duplicates

Best number of training steps: 150

Word count metrics (Policy 2)





Conversation sample with 5 turns

Speaker: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road.

Listener: Did you suffer any injuries?

Speaker: No I wasn't hit. It turned out they were drunk. I felt guilty but realized it was his fault.

Listener: Why did you feel guilty? People really shouldn't drive drunk.

Speaker: I don't know I was new to driving and hadn't experienced anything like that. I felt like my horn made him swerve into the water barrels.

Pre-process

Training sample 1

Context: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road.

Response: Did you suffer any injuries?

Training sample 2

Context: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road. <SOC> Did you suffer any injuries?

Response: No I wasn't hit. It turned out they were drunk. I felt guilty but realized it was his fault.

Training sample 3

Context: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road. <SOC> Did you suffer any injuries? <SOC> No I wasn't hit. It turned out they were drunk. I felt guilty but realized it was his fault.

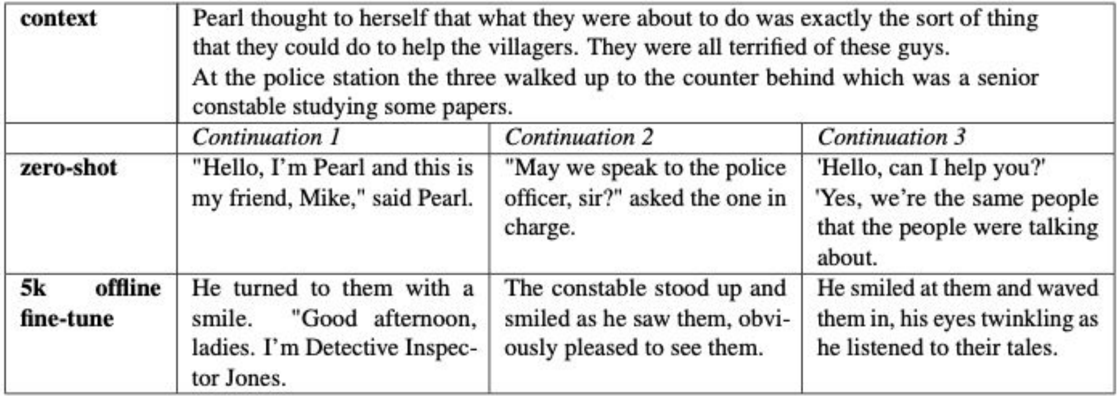
Response: Why did you feel guilty? People really shouldn't drive drunk.

Training sample 4

Context: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road. <SOC> Did you suffer any injuries? <SOC> No I wasn't hit. It turned out they were drunk. I felt guilty but realized it was his fault. <SOC> Why did you feel guilty? People really shouldn't drive drunk.

Response: I don't know I was new to driving and hadn't experienced anything like that. I felt like my horn made him swerve into the water barrels.

HUMAN EVALUATIONS OF CONTINUATIONS “SENTIMENT”



context	Pearl thought to herself that what they were about to do was exactly the sort of thing that they could do to help the villagers. They were all terrified of these guys. At the police station the three walked up to the counter behind which was a senior constable studying some papers.		
	Continuation 1	*Continuation 2*	*Continuation 3*
zero-shot	"Hello, I’m Pearl and this is my friend, Mike," said Pearl.	"May we speak to the police officer, sir?" asked the one in charge.	'Hello, can I help you?' 'Yes, we’re the same people that the people were talking about.
5k offline fine-tune	He turned to them with a smile. "Good afternoon, ladies. I’m Detective Inspector Jones.	The constable stood up and smiled as he saw them, obviously pleased to see them.	He smiled at them and waved them in, his eyes twinkling as he listened to their tales.

Human annotates best continuation