NLP LAB 2021 - VDA

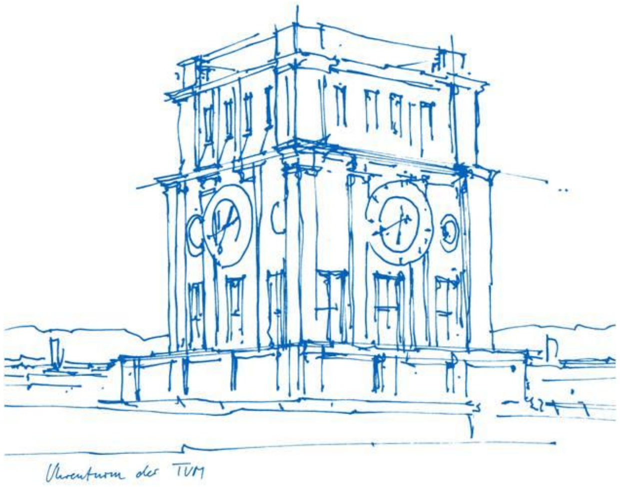
Ananta Bhattarai, Sophie Schoen, Viviane Rehor

Technische Universität München

Department of Informatics

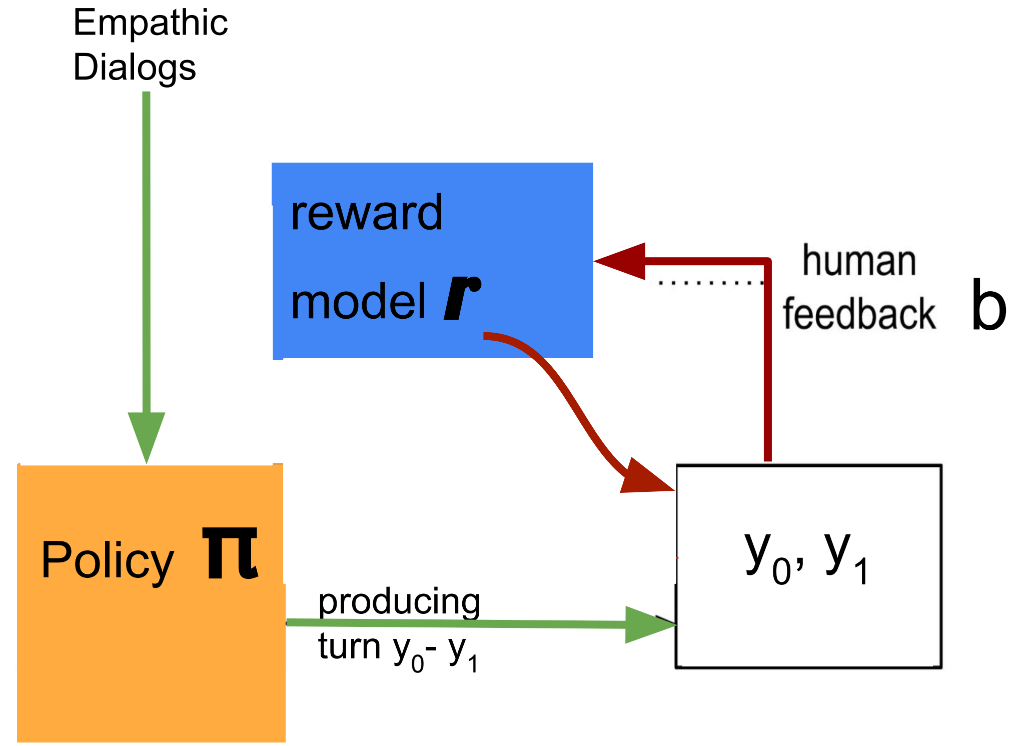
Research Group Social Computing

Munich, June 22 2021



Our Vision - Part 1

1. Creating a Baseline: Fine-Tuning GPT-2 with Empathic Dialogues datasets
   a. Data Preprocessing (Encoding, Decoding & Importing Data)
   b. Fine-Tuning to get π
2. Producing turns y_0- y_1 to give human feedback on
   a. *With test/validation set of Empathic Dialogs dataset*
3. Train reward model r with those turns and human feedback b

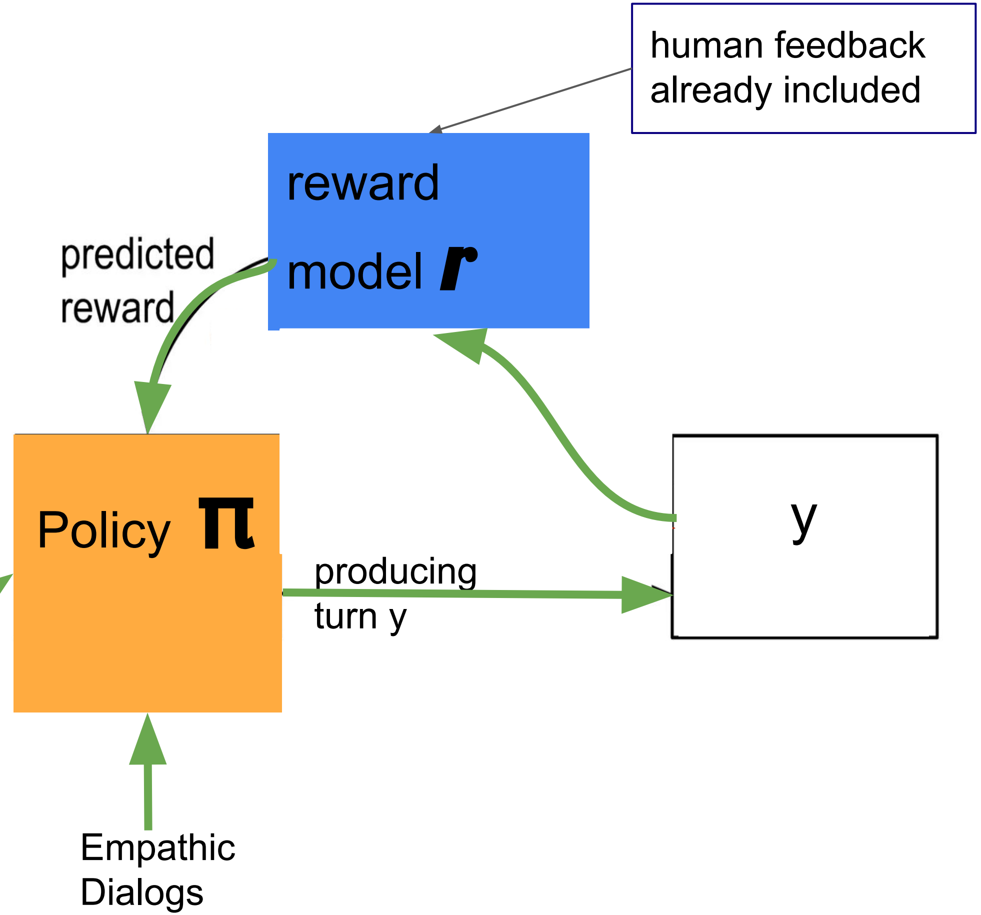


Our Vision - Part 2

4. Train policy π using reward model *r* with those turns and human feedback b

+ Evaluate Model performance with metrics

5. Improve the idea with input from other papers



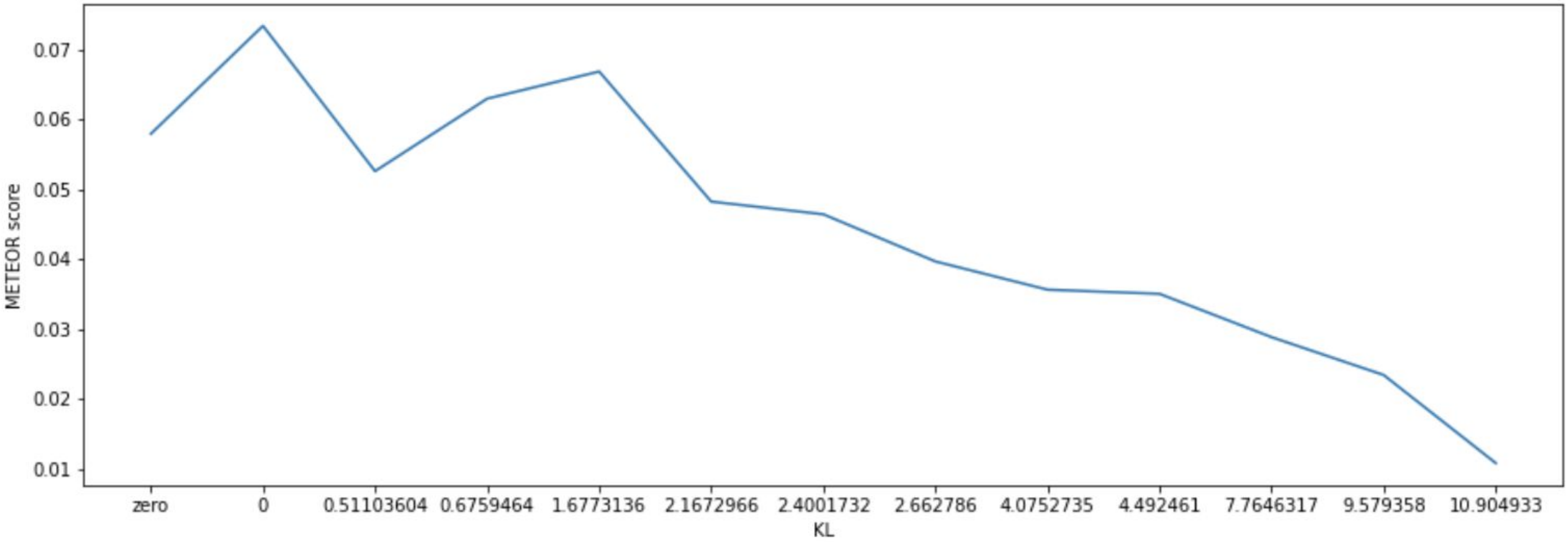
5. Maybe add Knowledge base + Dialog History

Annotation Statistics

	Ananta		Vivi		Sophie	
	Agreement	Cohen Kappa	Agreement	Cohen Kappa	Agreement	Cohen Kappa
Ananta	-	-	68%	0.35	73%	0.45
Vivi	68%	0.35	-	-	73%	0.46
Sophie	73%	0.45	73%	0.46	-	-

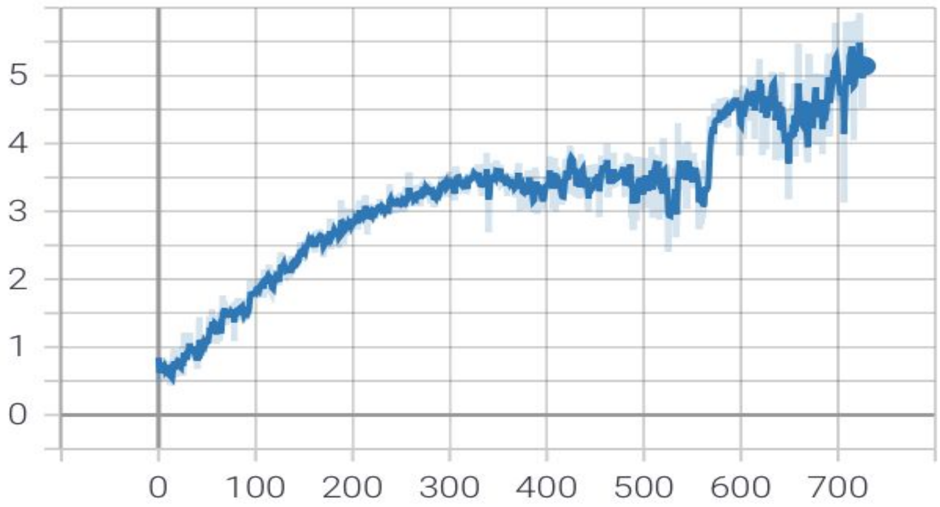
All agreement: 57%
Average Cohen Kappa: 0.41

KL VS METEOR



METEOR score
0.07
0.06
0.05
0.04
0.03
0.02
0.01
zero
0
0.51103604
0.6759464
1.6773136
2.1672966
2.4001732
2.662786
4.0752735
4.492461
7.7646317
9.579358
10.904933
KL

KL vs Returns



Two possible reasons for high rewards even after over optimization:

1. Distributional shift of the samples that reward model hasn’t seen while training.
2. Not enough training data to train reward model (2200 samples)

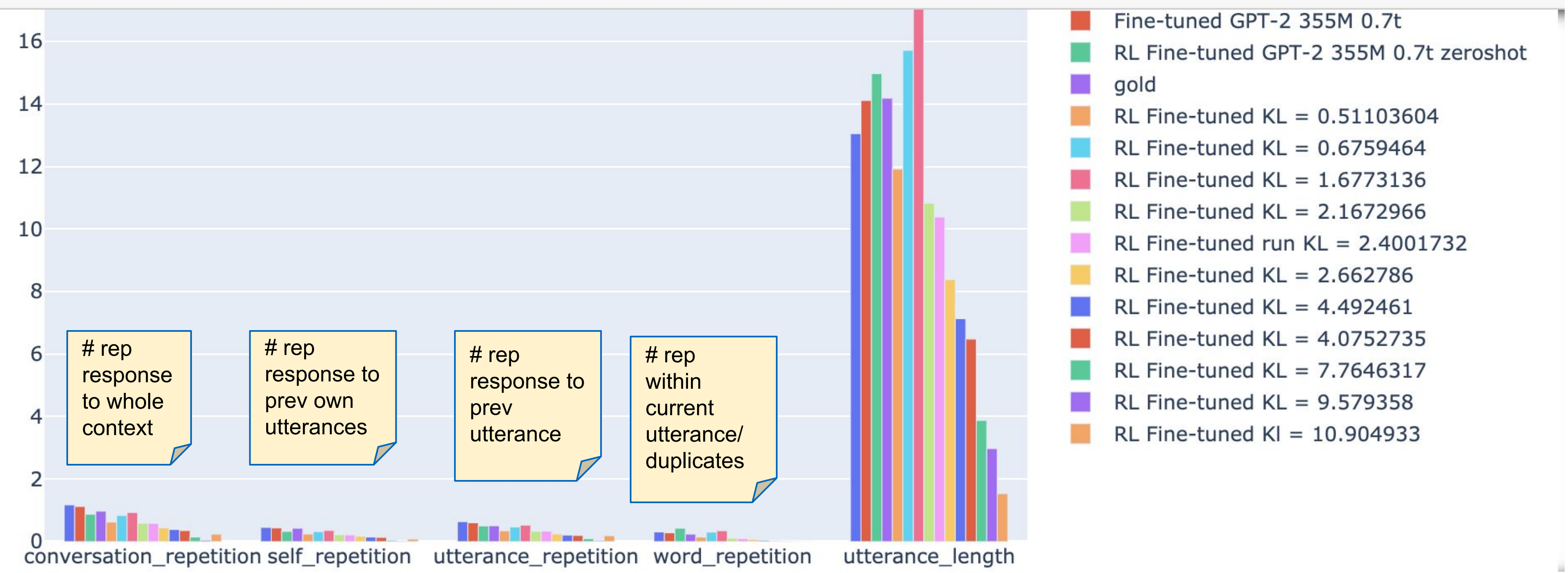
Evaluation Metrics - Florian

- **Utterance length**
- **Self repetition**
- **Utterance repetition**
- **Word repetition**
- **Question**
- **Conversation repetition**
- **Emotional reaction level**
- **Interpretation level**
- **Exploration level**
- **QuestionVsGold (Monika's idea from last meeting)**

- Deepmoji sentiment pos
- Deepmoji sentiment neg
- Deepmoji coherence
- Infersent coherence
- USE similarity
- Word2Vec coherence

TODO

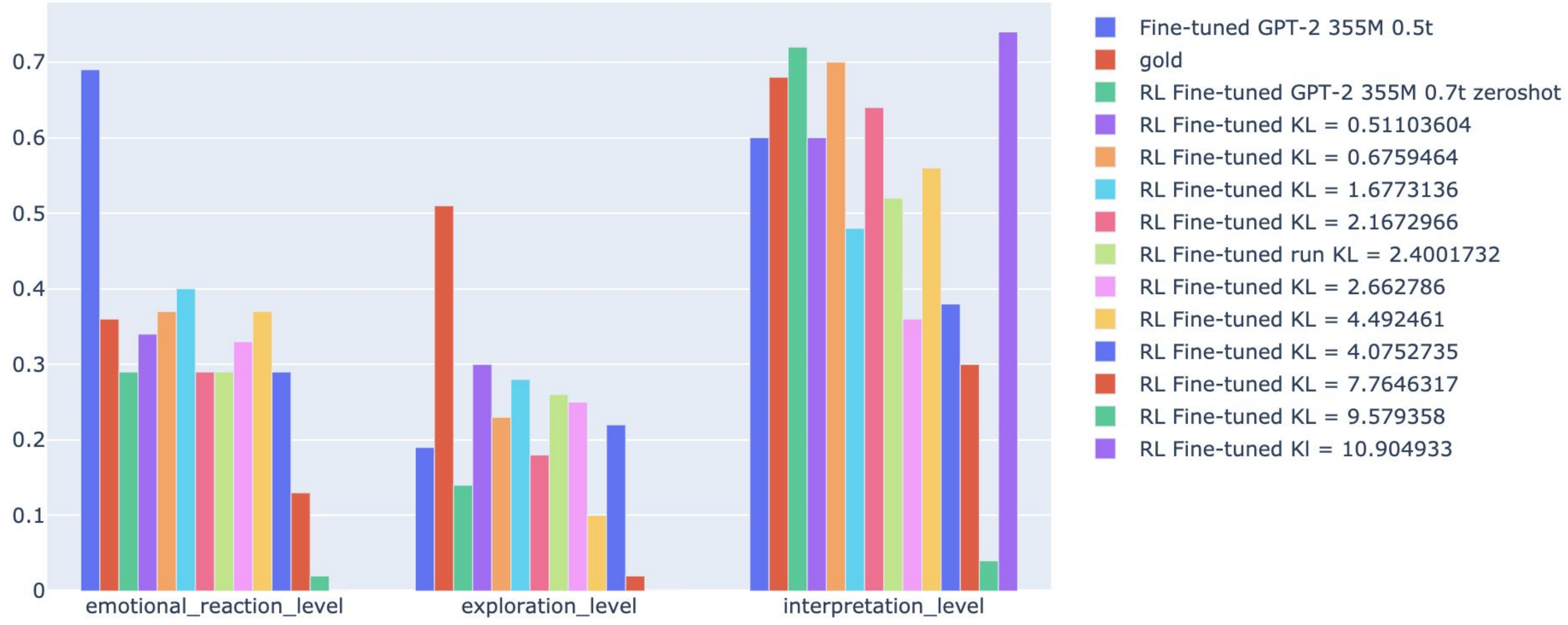
Word count metrics



Question metric: Generated vs Gold utterance

Empathy metrics

=> Weird values => Ask Florian



Automated Evaluation Metrics scores

1. BLEU Score = Basis
    - judges translations on a per-word basis
    - measures MT **adequacy** by looking at word precision
    - measures MT **fluency** by calculating n-gram precisions
    - n-gram matching requires exact word matches
      → better: METEOR Score

Automated Evaluation Metrics scores

2. METEOR Score

- allows multiple reference translations
  → addresses the problem of variability with flexibility in word matching
- Extra features to BLEU:
  - stemming
  - synonymy matching

Automated Evaluation Metrics scores

3. NIST Score

- weights n-gram matches by their information gain & indirectly penalizes uninformative n-grams

  → BLEU calculates n-gram precision by adding equal weight to each n-gram
  → NIST also calculates how relevant a particular n-gram is
  → **More weight** is given to n-grams that are considered **less likely to occur**

"Yes I made an interesting calculation"

Automated Evaluation Metrics scores

4. TER Score

- measures the number of edits required to change a system output into one of the references
  → evaluating the quality

Evaluating chatbots

Combine sensibleness and specificity in one metric: SSA (sensibleness and specificity average) ≈ human likeness

Human Evaluation setups:

a) *Static*: benchmark models on a fixed set of multi-turn contexts to generate responses
b) *Interactive*: allow humans to chat freely with chatbots

A: "I love tennis,"	A: "I love tennis,"
B: "That's nice,"	B: "Me too, I can't get enough of Roger Federer!"
→ not specific	→ specific

Automatic Evaluation

From Google Research:Towards a Human-like Open-Domain Chatbot

Automatic Perplexity metric

- correlates with human judgement of sensibleness and specificity (SSA metrics)
- seq2seq model outputs a probability distribution over possible next response tokens
- Correlation static sensibleness & specificity vs perplexity: $R2=0.93$
  → perplexity= good automatic metric for measuring sensibleness and specificity

Perplexity Score

- good evaluation metric for chatbots
- With perplexity you are trying to evaluate the similarity between the token (in your case probably sentences) distribution generated by the model and the one in the test data.
- For instance, assuming you have
-

What are our best next steps?