Universität Stuttgart
Institute for Natural Language Processing

LREC-COLING 2024

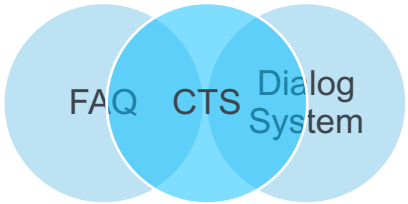
Towards a Zero-Data, Controllable, Adaptive Dialog System

Dirk Väth, Lindsey Vanderlyn, Ngoc Thang Vu

Motivation

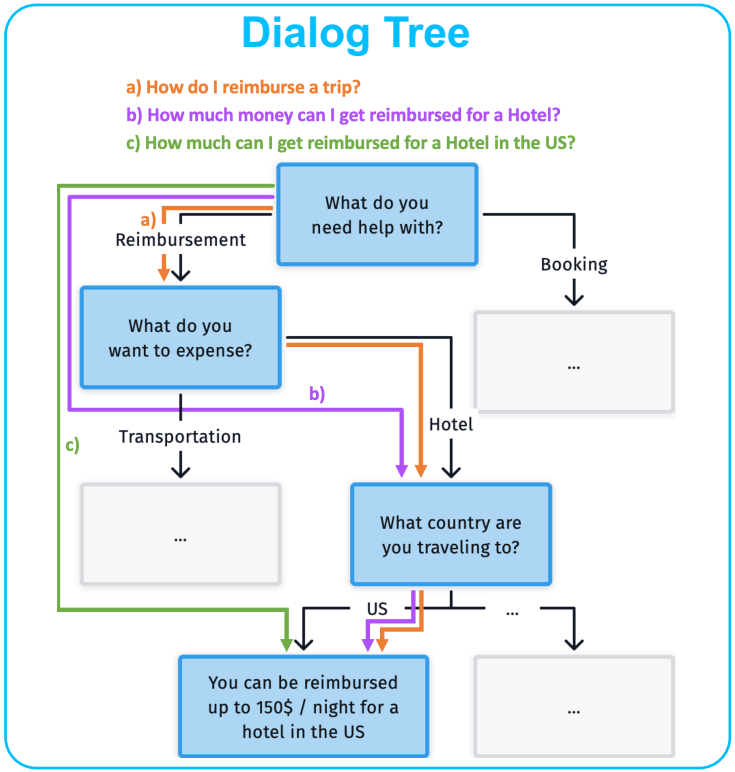
- Research & Industry transition towards Large Language Models (LLM) for dialog systems
- BUT: Controlling output remains an open challenge (e.g. Hallucination)

  ➔ LLMs are currently unsuitable for sensitive domains (e.g. legal, medical)
- Main text-based user interfaces in such domains are
  - FAQ-style Retrieval Systems
  - Multi-Turn Dialog Systems
- **Conversational Tree Search (CTS)** [2] **:**
  - “Interpolates” between FAQ-style Information Retrieval and multi-turn chatbot-style interaction
  - But: CTS needs data from both categories:
    - FAQ: retrieval model (potentially finetinung)
    - Dialog System: Dialog graph, user questions, user responses
  - **How can we minimize data requirements**?



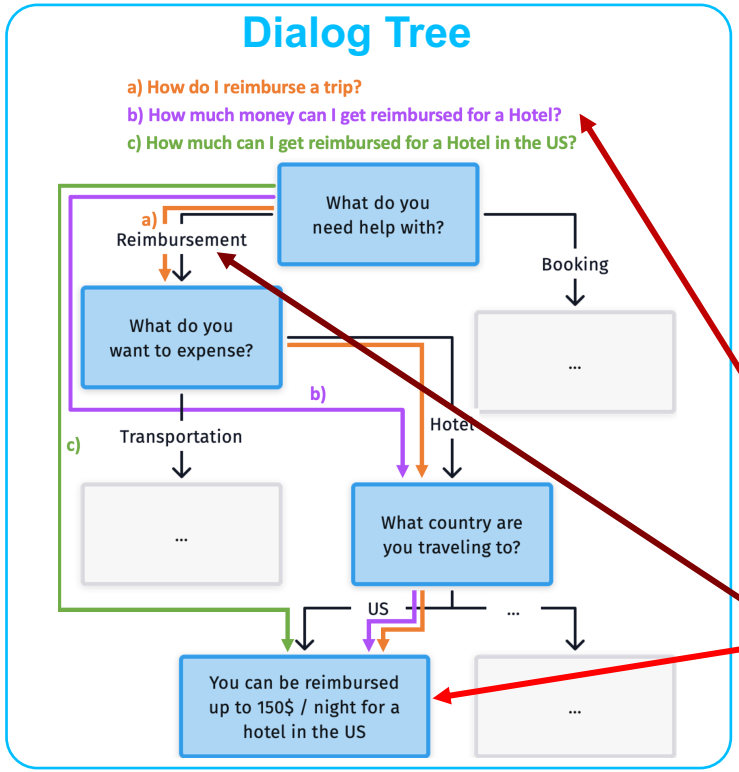
[1] Thakur et al. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.
[2] Väth et al. 2023. Conversational tree search: A new hybrid dialog task. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 12641280, Dubrovnik, Croatia. Association for Computational Linguistics.

Conversational Tree Search
Task & Training



- Reinforcement Learning (RL) agent learns to traverse dialog tree node by node
  - Dialog flow pre-defined by domain-experts → Controllability
- Agent learns to skip unnecessary dialog nodes on the way to answering the user question
  - Agent can not “hop” between branches → Consistent dialog context

Conversational Tree Search
Task & Training



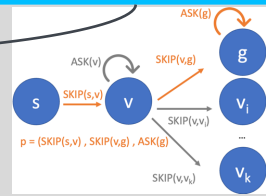
User utterance (How do I reimburse a trip?),
Turn reward (-1),
Node text (What do you need help with?),
Current responses (Reimbursement, Booking),
…

Simulator

1. Draw goal node
2. Draw user utterance
…
N. Draw responses

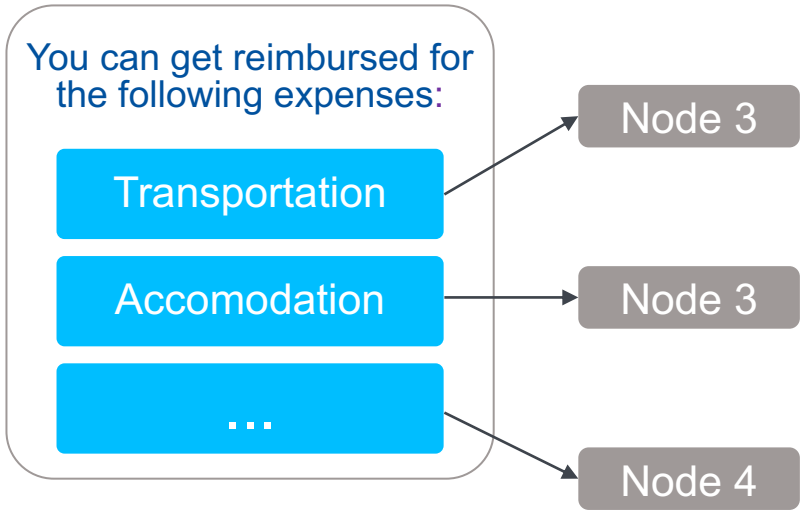
Select response (Reimbursement),
Interaction style prediction (Question)

RL Agent



Conversational Tree Search
Task & Training

- Required training data for RL agent:
  - User questions (per node, node contains the answer)
  - User responses (per node response option, node contains list of prototypical user response options)



- **Question**: What costs can I get reimbursed?
- **Responses**:
  - **Transportation**: Train ticket
  - **Accomodation**: My hotel

Datasets

New datasets

Dataset	Split	#Nodes	Tree Depth	Max. Node Degree	#User Questions	Avg. User Questions	#Answer Paraphrases	Avg. Answer Paraphrases
REIMBURSE	Train Test	123	32	14	279 173	3.5 2.2	246 162	3.4 2.2
REIMBURSE-En	Train Test	123	32	14	279 173	3.5 2.2	246 162	3.4 2.2
DIAGNOSE **DIAGNOSE**	Train Test	98	10	6	219 150	2.9 2.0	298 298	3.0 3.0
ONBOARD **ONBOARD**	Train Test	88	15	9	141 117	2.4 2.0	175 152	3.1 2.7

Table 1: Overview of original *REIMBURSE*, translated *REIMBURSE-En*, and newly created *ONBOARD* and *DIAGNOSE* datasets (numbers rounded to one decimal).

Research Questions

- **RQ1 How can we effectively generate data for a zero data approach to training CTS agents?**
  - **RQ1.1** How can we analyze the quality of generated data?
  - **RQ1.2** How do agents trained on generated data perform in simulation, compared to agents trained on human data?
  - **RQ1.3** How well do the data generation techniques transfer to new domains?
- **RQ2** How does a CTS agent trained on generated data perform with real users compared to an agent trained on human data?

Research Questions

- **RQ1 How can we effectively generate data for a zero data approach to training CTS agents?**
  - **RQ1.1** How can we analyze the quality of generated data?
  - **RQ1.2** How do agents trained on generated data perform in simulation, compared to agents trained on human data?
  - **RQ1.3** How well do the data generation techniques transfer to new domains?
- **RQ2** How does a CTS agent trained on generated data perform with real users compared to an agent trained on human data?

Effectively generating data for a zero data approach to training CTS agents Approach

- We generate data via prompting using
  - ChatGPT (gpt-3.5-turbo)[1]
  - Quantized Llama (fits on single NVIDIA GeForce RTX 3090)[2]
- **Goals**: Generated data should be
  - **Similar to human data**
    - Syntactically
    - Semantically
  - **Diverse**
  - **Answerable**

[1] https://platform.openai.com/docs/ models/gpt-3-5
[2] https://huggingface.co/TheBloke/ upstage-llama-30b-instruct-2048-GPTQ

Effectively generating data for a zero data approach to training CTS agents
Question Generation & Analysis

Gen_{V1}

- **System Instruction:** You are a truthful assistant, generating diverse FAQ-style questions given some facts. The generated questions should be answerable using the given facts only, without additional knowledge. The questions should also be human-like. Try to vary the amount of information between questions. Present the results in a numbered list
- **User prompt:** Generate {#questions} FAQ-style questions about the given facts: "{NODE TEXT}"

- **Naïve approach**: ask LLM to generate truthful, answerable questions given dialog node context
- **Observation**: Questions are much longer than human questions, also less natural

? Can a change in the system instruction fix that?

Effectively generating data for a zero data approach to training CTS agents
Question Generation & Analysis

Gen_{V1}

- **System Instruction:** You are a truthful assistant, generating diverse FAQ-style questions given some facts. The generated questions should be answerable using the given facts only, without additional knowledge. The questions should also be human-like. Try to vary the amount of information between questions. Present the results in a numbered list
- **User prompt:** Generate {#questions} FAQ-style questions about the given facts: "{NODE TEXT}"

Gen_{V2}

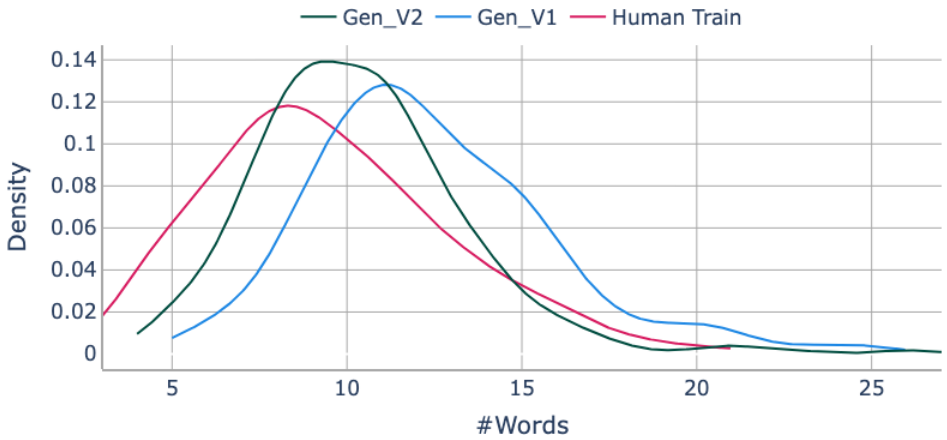
- **System Instruction:** You are a truthful assistant, generating diverse FAQ-style questions given some facts. The generated questions should be answerable using the given facts only, without additional knowledge. The questions should also be short and human-like. Try to vary the amount of information between questions. Present the results in a numbered list

Effectively generating data for a zero data approach to training CTS agents

Question Generation & Analysis

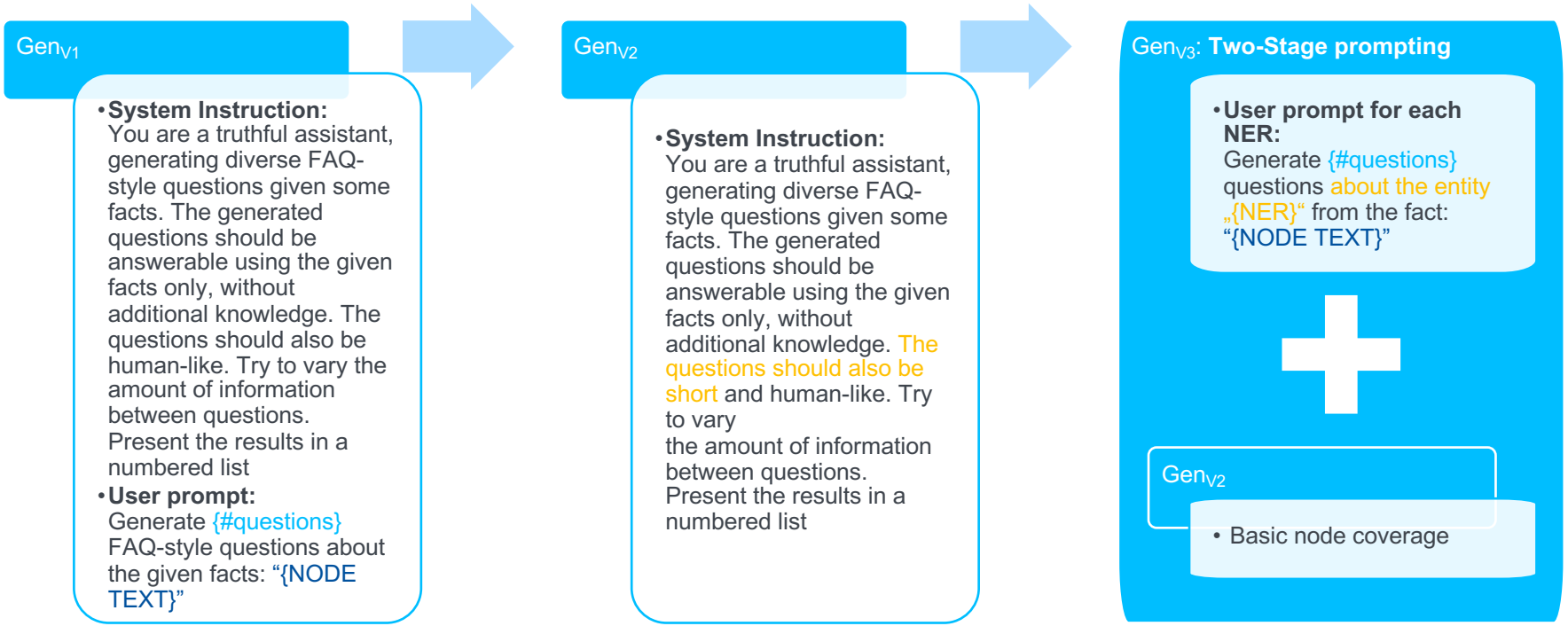
- **Observations**:
  - ✔ Gen_{V2} shifts the distribution of question lengths more towards the human data
  - **But**: Generated questions tend to focus only on one part of the node text
    - Lack diversity
    - Omit topics

? Can explicitly asking about different topics per node help?



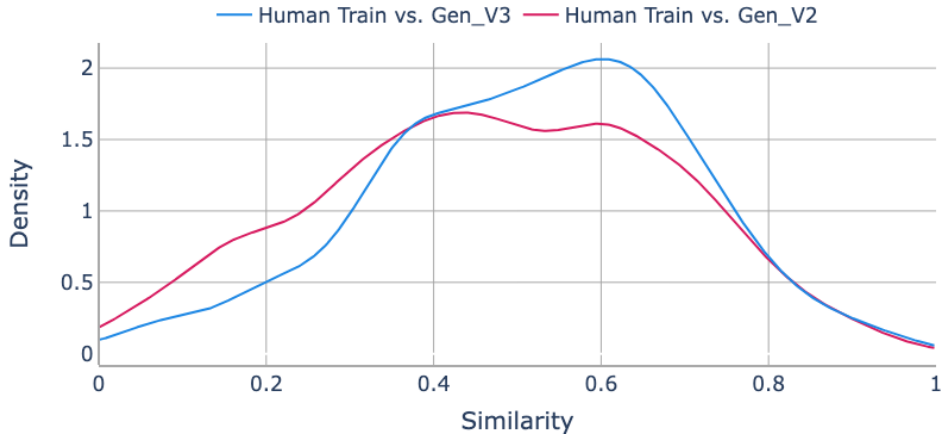
Comparison between human data and generated data question length

Effectively generating data for a zero data approach to training CTS agents
Question Generation & Analysis



Effectively generating data for a zero data approach to training CTS agents
Question Generation & Analysis

- Investigate semantic similarity with human data:
  - Calculate pair-wise cosine similarities between all human and generated questions for each node from the dialog graph
    - Average similarities across all nodes
- **Observations**:
  - ✔ Significantly[1] increases similarity of generated (avg.: 0.52) vs. human training data (avg.: 0.47),



Semantic similarity: artificial vs with human data

[1] ($p < 1.82e-11$)

Effectively generating data for a zero data approach to training CTS agents
Question Generation & Analysis

- Investigate diversity:
  - Calculate self-BLEU scores (lower is better)
- **Observations**:

✔ Gen_{V3} data is most diverse

Training Data	n-1	n-2	n-3	n-4	n-5
Human	**0.78**	**0.68**	**0.60**	**0.54**	**0.49**
V1	0.95	0.92	0.87	0.83	0.80
V2	0.95	0.90	0.85	0.80	0.76
V3	**0.85**	**0.78**	**0.71**	**0.66**	**0.62**

Self-BLEU scores for different n-gram sizes on human and generated data

Effectively generating data for a zero data approach to training CTS agents
Question Generation & Analysis

- Investigate answerability:
  - Calculate QA confidence scores of the generated questions, given the node text as document containing the answer
- **Observation:** Gen_{V3} data is significantly[1] more answerable than Gen_{V2}

✔ We have now improved towards all goals (similarity to human data, diversity & answerability)

? Will the improvements translate to the downstream task (higher dialog success rate)?

[1] ($p < 0.0003$)

Research Questions

- **RQ1** How can we effectively generate data for a zero data approach to training CTS agents?
  - **RQ1.1** How can we analyze the quality of generated data?
  - **RQ1.2 How do agents trained on generated data perform in simulation, compared to agents trained on human data?**
  - **RQ1.3** How well do the data generation techniques transfer to new domains?
- **RQ2** How does a CTS agent trained on generated data perform with real users compared to an agent trained on human data?

Agents trained on generated data in simulation vs. agents trained on human data

- ✔ Best performing agent trained on artificial data (Gen_{V3} : 69.44% success) performs comparably to the best performing agent trained on human data (CTS ours : 73.86% success)
- ✔ Using a standard T-Test, we find no statistically significant difference

Model	Training Data	Avg. Perceived Length (guided)	Avg. Perceived Length (free)	Success (combined)	Dialog Mode Prediction F1	Dialog Mode Prediction Consistency
Original	human (GER)	n/a	n/a	62.58%	0.85	**1.0**
Original	human (EN)	n/a	n/a	55.28%	0.86	0.87
Ours	human (EN)	13.56	2.95	**73.86%**	**0.94**	0.96
Ours	V1 (LLAMA)	13.53	3.41	64.17%	**0.98**	**0.97**
Ours	V2 (LLAMA)	11.71	3.65	65.02%	**0.98**	0.95
Ours	V3 (LLAMA)	12.89	3.45	**69.44%**	0.96	0.95
Ours	V1 (ChatGPT)	13.02	3.65	64.35%	**0.98**	**0.97**
Ours	V2 (ChatGPT)	14.55	3.71	66.67%	0.95	**0.97**
Ours	V3 (ChatGPT)	12.87	3.59	**68.41%**	**0.98**	**0.97**

Simulation results on **REIMBURSE(-En)** test splits of original CTS agent (German), our improved agent (English), and our CTS agent trained on generated data only (English).

Research Questions

- **RQ1** How can we effectively generate data for a zero data approach to training CTS agents?
  - **RQ1.1** How can we analyze the quality of generated data?
  - **RQ1.2** How do agents trained on generated data perform in simulation, compared to agents trained on human data?
  - **RQ1.3 How well do the data generation techniques transfer to new domains?**
- **RQ2** How does a CTS agent trained on generated data perform with real users compared to an agent trained on human data?

Transfer to new domains

- ✔ Comparable results between human data and artificially generated data
- ✔ T-Tests show no statistically significant differences between the best synthetically trained agents (ChatGPT) and the agents trained on human data

Domain	Training Data	Avg. Perceived Length (guided)	Avg. Perceived Length (free)	Success (combined)
DIAGNOSE	human	6.42	2.29	76.31%
DIAGNOSE	V3 (LLAMA)	6.62	2.95	71.08%
DIAGNOSE	V3 (ChatGPT)	5.65	2.46	**85.12%**
ONBOARD	human	7.88	2.98	**73.61%**
ONBOARD	V3 (LLAMA)	7.91	3.52	63.38%
ONBOARD	V3 (ChatGPT)	7.60	3.58	70.72%

Performance of CTS agents trained on human and generated data on the new domains **DIAGNOSE and ONBOARD** in simulation.

Research Questions

- **RQ1** How can we effectively generate data for a zero data approach to training CTS agents?
  - **RQ1.1** How can we analyze the quality of generated data?
  - **RQ1.2** How do agents trained on generated data perform in simulation, compared to agents trained on human data?
  - **RQ1.3** How well do the data generation techniques transfer to new domains?
- **RQ2 How does a CTS agent trained on generated data perform with real users compared to an agent trained on human data?**

Evaluation of CTS agents trained on generated data vs human data by real users

Performance of CTS agents trained on human and generated data on the new domains **DIAGNOSE and ONBOARD** in simulation.

Training Data	# Turns	Success	Perceived Length	Answer Satisfaction
Human	6.14	77.59	2.88	2.93
V3	5.27	72.73	2.65	2.73

✔ No statistically significant differences (standard T-Test) between either subjective or objective measures:

- Success
- dialog length

✔ No significant difference in the reported trust, reliability, or usability scores between either group

➔ **No perceived loss in performance when using generated data compared to real data, either in objective or subjective metrics**

➔ **Simulator can be a good proxy for real user evaluation**: Using Welch's t-test (to account for the difference in sample size), we find no significant difference in dialog success between the simulated and human dialogs

Conclusion

- We present two new and publicly available datasets (DIANGOSE & ONBOARD)
- We explore several zero-data prompting-based methods for generating data to train a CTS agent
  - We develop a novel two-stage prompting approach to increase question diversity
  - We find that automatic scores for diversity and answerability can be indicative of downstream dialog task performance
- We show that there is no statistically significant difference in objective or subjective metrics between agents trained on human data or on generated data
  - Verified through simulation & user study
  - We tested agent performance on 3 datasets to verify generalizability

➔ **We can effectively generate training data from a dialog tree, such that CTS agents can be trained in zero data settings with negligible performance impact**

Universität Stuttgart
Institute for Natural Language Processing

Thank you!

Dirk Väth, Lindsey Vanderlyn, Ngoc Thang Vu

Dirk.vaeth@ims.uni-Stuttgart.de

ims.uni-stuttgart.de/en/institute/team/Vaeth-00001/

University of Stuttgart
Institute for Natural Language Processing