# Long-Horizon Dialogue Understanding for Role Identification in the Game of Avalon with Large Language Models

Simon Stepputtis[1], Joseph Campbell[1], Yaqi Xie[1], Zhengyang Qi[1], Wenxin Sharon Zhang[1],
Ruiyi Wang[1], Sanketh Rangreji[1], Michael Lewis[2], Katia Sycara[1]
[1] *Carnegie Mellon University* [2] *University of Pittsburgh*

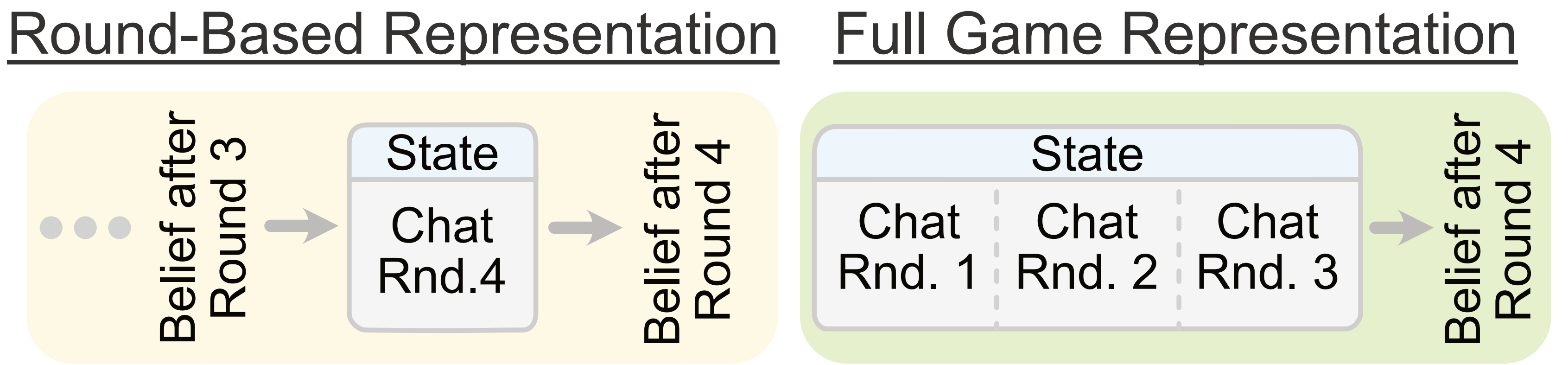
## 1. Background and Question

We address the challenging problem of understanding multi-party dialogue in a **competitive-cooperative setting** involving **persuasion and deception** amongst six humans in the game of ***Avalon: The Resistance***.

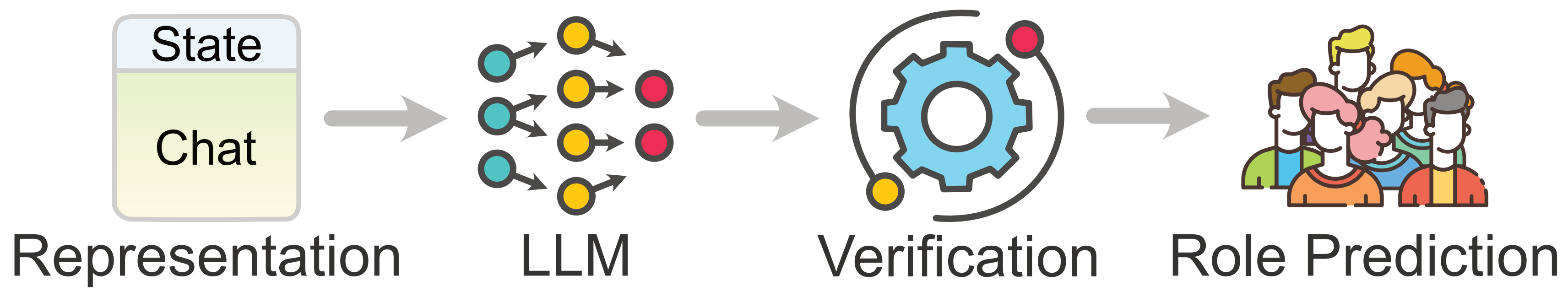
Problem Statement and Contributions:

- Large Language Models (LLM) face challenges reasoning over and identifying **persuasion and deception**
- We release a social-deduction **dataset and simulator**
- We propose **two game representations**: round-based and full game state

## 2. Representation and Inference



Large Language Model (LLM) Inference: GPT4, Llama-2

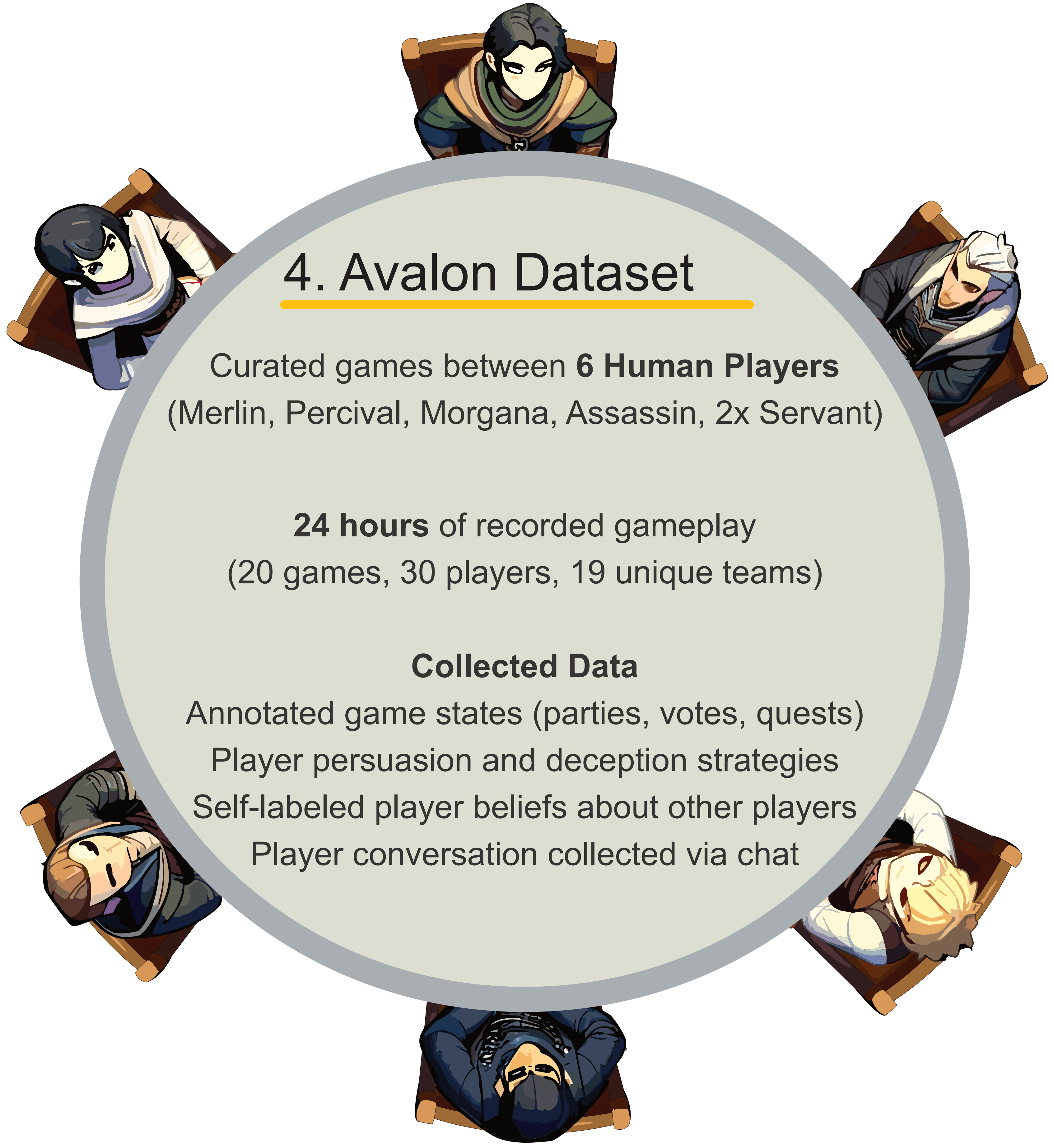


## 3. Results

We compare various LLMs including fine-tuned versions, predicting the roles (Good, Evil, Merlin) of all players

- We report F1-Scores for the round-based and full-context representation, as well as the utility of state information
- With our high-quality data, fine-tuning is successful for Llama-2, and limited success for GPT-3.5
- LLMs do not compare to human reasoning capabilities in complex social settings that require the understanding of persuasion and deception.

| Model | Modality | Role Prediction F1-Scores | | |
|---|---|---|---|---|
| | | Good | Evil | Merlin |
| GPT-4 | Chat | 0.67 / 0.67 | 0.48 / 0.55 | 0.36 / 0.20 |
| | Chat & State | 0.67 / 0.68 | 0.46 / 0.58 | 0.05 / 0.27 |
| GPT-3.5-FT | Chat & State | 0.52 / 0.59 | 0.38 / 0.41 | 0.19 / 0.15 |
| Llama-2 | Chat | 0.68 / 0.61 | 0.39 / 0.27 | 0.00 / 0.00 |
| | Chat & State | 0.61 / 0.55 | 0.33 / 0.22 | 0.00 / 0.00 |
| Llama-2-FT | Chat & State | 0.65 / 0.63 | 0.35 / 0.26 | 0.23 / 0.27 |
| Human | Chat & State | 0.76 | 0.72 | 0.33 |



## 4. Avalon Dataset

Curated games between **6 Human Players**
(Merlin, Percival, Morgana, Assassin, 2x Servant)

**24 hours** of recorded gameplay
(20 games, 30 players, 19 unique teams)

**Collected Data**
Annotated game states (parties, votes, quests)
Player persuasion and deception strategies
Self-labeled player beliefs about other players
Player conversation collected via chat

## 5. Example Round

*System* Game Started

*P4* *Persuasion: Assertion | Deception: None*
We don't have any info so just selected randomly

*System* Player-4 proposed a party: ***Player-2 Player-6***

*P5* *Persuasion: Questioning | Deception: None*
Seems like a weird choice of party in my opinion. Little bit suspicious of player-4 for not picking themself.

*P6* *Persuasion: Assertion | Deception: None*
I don't have opinions at this point

*P1* *Persuasion: Questioning | Deception: None*
Well, technically we don't know anything yet. but I agree that this choice is a little weird

*P2* *Persuasion: Suggestion | Deception: Omission*
I'm good, but I agree this choice is weird. I don't know what changes we can make since it's the first turn. I'll approve the current party unless you make some good arguments

*P3* *Persuasion: Critique/Opposition | Deception: Omission*
No opinions but a good guy will always place themselves in the team...

*System* Player-4 proposed a party: ***Player-2, Player-4***

*P4* *Persuasion: Assertion | Deception: None*
Sorry for the last turn, its still random but includes myself

*System* Party Vote Outcome: Player-1: **Yes**, Player-2: **Yes**, Player-3: **No**, Player-4: **Yes**, Player-5: **Yes**, Player-6: **Yes**

*System* Vote Succeeded! Initiating Quest Vote!

*System* Quest Succeeded!

## 6. Takeaways

- We demonstrate that current state-of-the-art LLMs struggle to understand deception and persuasion
- We provide a high-quality NLU dataset with over 20 recorded Avalon: The Resistance games
- Our dataset provides opportunities for understanding deception, agent development, and other NLU tasks

**Contact:** Simon Stepputtis (stepputtis@cmu.edu)
**Project Website:** https://sstepput.github.io/Avalon-NLU/

EMNLP 2023