LLM ROLEPLAY: SIMULATING HUMAN-CHATBOT INTERACTION

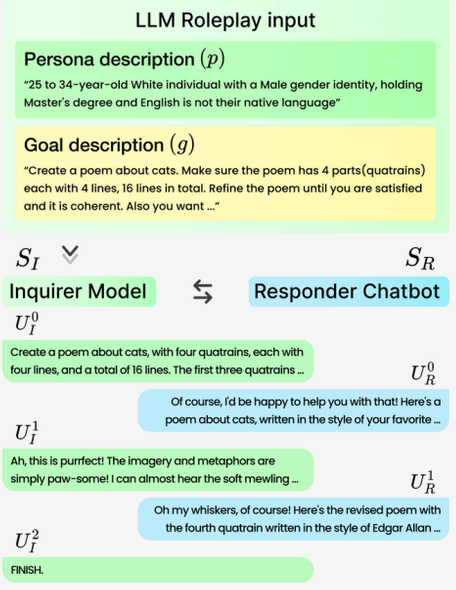
Hovhannes Tamoyan, Hendrik Schuff, and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP Lab) Technical University of Darmstadt

MOTIVATION

- Human-Chatbot Dialogue: Managing many annotators is costly.
- Synthetic Data: Chatbots offer a faster, cheaper data generation method.
- Large Language Models: Synthetic data is used for improving LLMs with methods like Alpaca and InstructGPT.
- Quality Dialogue Models: Effective dialogues depend on critical prompts, longer responses, and mimicking human styles.

RESULTS

Method

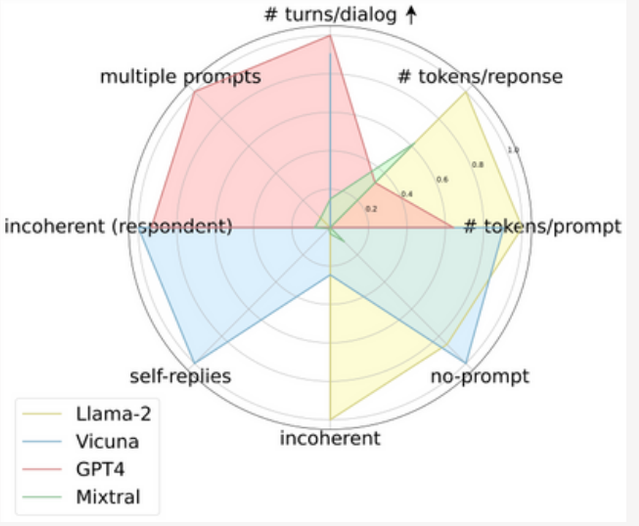


- **Role Assignments:** An **Inquirer** with a distinct persona and a goal initiates and drives the conversation, posing questions until the **Responder** accomplishes the Inquirer's task, satisfying the stopping criteria.

ABSTRACT

- **Goal-Oriented Interaction**: LLM Roleplay employs a **persona-driven**, **goal-oriented** methodology for generating dynamic, **multi-turn dialogues**.
- **Flexible Integration**: It seamlessly adapts to any LLM or chatbot, enabling realistic simulations of human-chatbot interactions for various tasks.

Persona-Specific Dialogue Collection

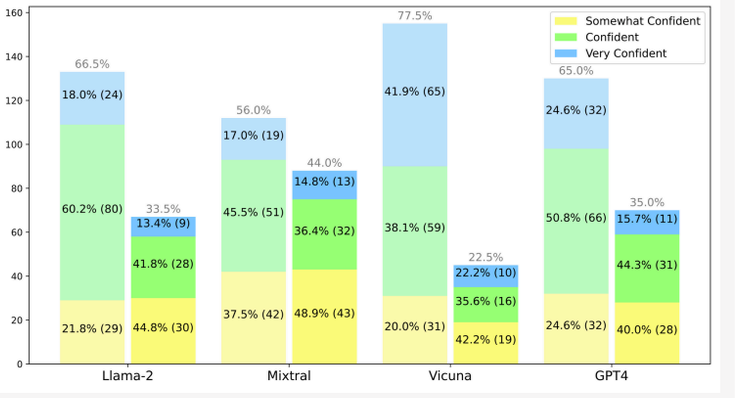


- **Mistral 8x7B**: Produces concise prompts with an average of **50.82 tokens** each.
- **GPT-4**: Maintains longer conversations with an average of **7.60 turns per dialogue**.
- Length and Detail: Dialogues average **5.30 turns** and **67.97 tokens per turn**, resulting in longer interactions than earlier models.

FINDINGS

- LLM Roleplay effectively generates multi-turn conversations that closely mimic natural human-chatbot dialogues.
- The generated dialogues demonstrate a high level of indistinguishability from real interactions.
- LLMs show strong potential in accurately embodying specified personas during simulations.
- The method opens new possibilities for generating training data for fine-tuning models.

Human-Evaluation



- Overall Indistinguishability: Among **800** samples, **33.75%** of simulated dialogues **remained undetected** (with 50% representing absolute indistinguishability).
- **Mistral 8x7B** leads with the **highest undetectability rate at 44.0%**, GPT-4: follows with a rate of 35.0%, Llama-2 and Vicuna show undetectability rates of 33.5% and 22.5%, respectively.