Evaluating LLM-Generated Versus Human-Authored Responses in Role-Play Dialogues

Dongxu Lu, Johan Jeuring, Albert Gatt

Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

Table of contents

I Introduction & Motivation

II Exp 1. Human Evaluation

38 **human participants** to evaluate **1** scenario & focus group with two designers

III Exp 2. Automated Evaluation

Using **LLM-as-a-judge** to evaluate **3** additional scenarios

IV Discussion & Conclusions

I

Introduction & Motivation

LLMs for Professional Role-Play

Virtual Patient[1]

AI Tutor[2]

Communication Trainer[3]

[1] Patrick G. Kenny and Thomas D. Parsons. 2024. Virtual Standardized LLM-AI Patients for Clinical Practice. Annual Review of CyberTherapy and Telemedicine, 22:177–182.
[2] Sixu An, Yicong Li, Yunsi Ma, Gary Cheng, and Guan-dong Xu. 2024. Developing an LLM-Empowered Agent to Enhance Student Collaborative Learning Through Group Discussion. In International Conference on Computers in Education.
[3] Lu, D., Jeuring, J., & Gatt, A. (2025). Evaluating LLM-Generated Versus Human-Authored Responses in Role-Play Dialogues. arXiv preprint arXiv:2509.17694.

LLMs for Professional Role-Play

- **High accessibility** as an AI-based solution
- Targeted scenarios for **communication skill** training
- **Scenario design:**
  time-consuming & requires domain-specific knowledge

Francisco
DialogueTrainer

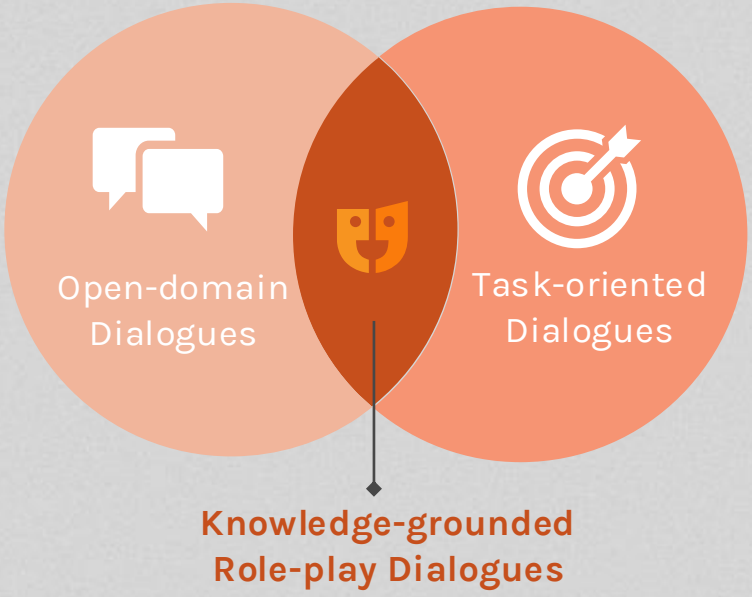
LLMs for Professional Role-Play

- **High accessibility** as an AI-based solution
- Targeted scenarios for **communication skill** training
- **Scenario design:**
  time-consuming & requires domain-specific knowledge

LLMs for scenario design of training simulations

Francisco
DialogueTrainer

Knowledge Gap: Dialogue Evaluation



Knowledge Gap: Dialogue Evaluation

Knowledge Gap: Dialogue Evaluation

Research Question

- Lack of benchmarks for knowledge-grounded role-play dialogues
- Limited interactions instead of extended interactions

*"How do **LLM-generated and human-authored** responses compare in **knowledge-grounded role-play dialogues over multiple turns**?""*

II

Human Evaluation

Study Design

SubRQ1:

“How do the quality perceptions of LLM-generated and human-authored responses change as a dialogue progresses over turns?”

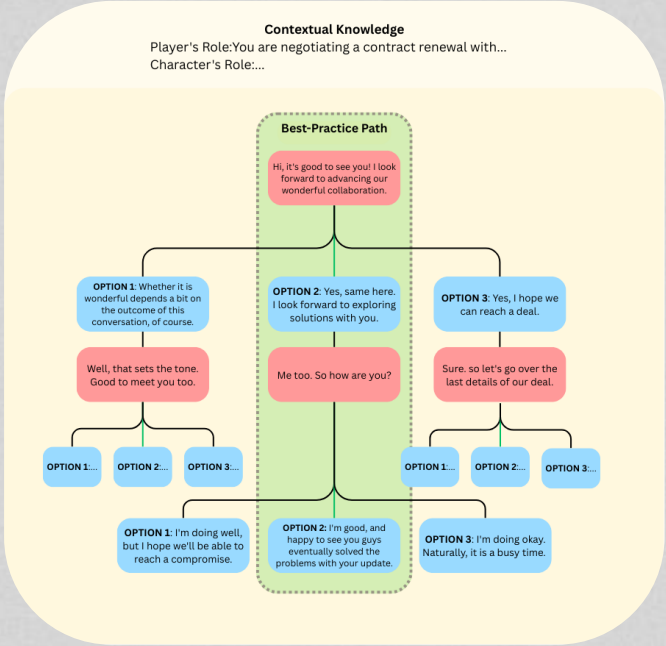
SubRQ2:

“What key factors most strongly influence participants’ perceptions of response quality?”

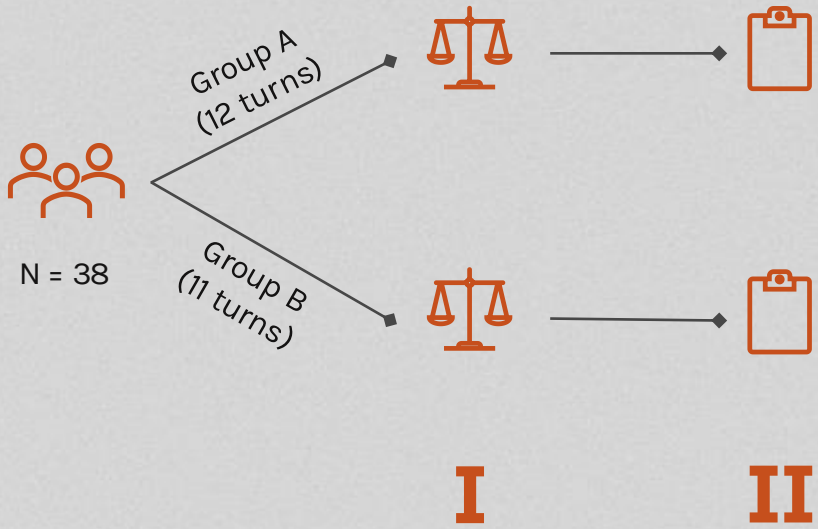
Study Design

Negotiation skill (23 exchanges)

- ◊ **Human-authored Responses**
  Agents' responses in the best-practice path, written by instructional designers.
- ◊ **LLM-generated Responses**
  An alternative response generated at every agent's turn along the same path using a fine-tuned LLAMA 3 model.



Study Design



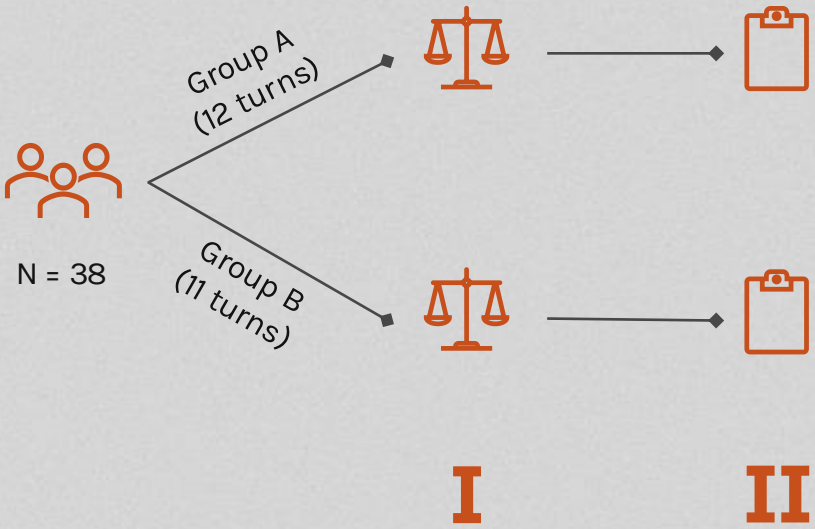
I: In-simulation Preference

"Which response do you think fits best within the conversation?"

II: Post-simulation Rating

Understandable, Natural, Maintains Context, Interesting, Uses Knowledge, Overall Quality

Study Design



Focus Group

2 instructional designers

- Discussed examples of extreme preferences
- Explored factors influencing perceived response quality

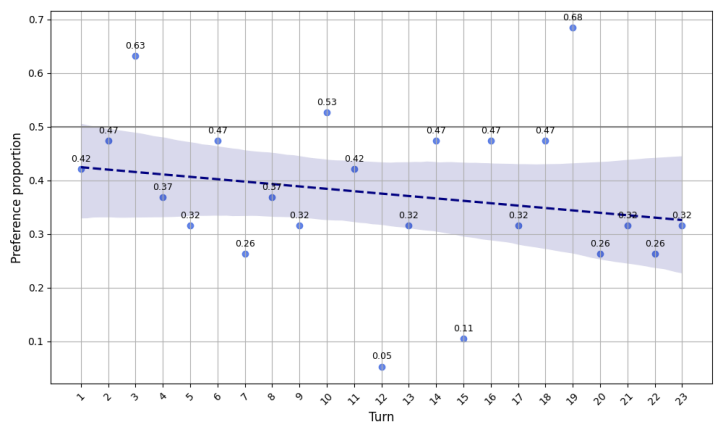
Data Analysis

- Proportional Analysis
- Trend Visualisation using OLS

- Linear Mixed-Effects Models
  - Fixed Effects: Condition (LLM vs. Human), Turn, and their interaction
  - Random Effects: Random intercepts
- Trend Visualisation using OLS
- Correlation Analysis (Spearman)

- Transcription & Data Cleaning
- Thematic Analysis
  - Labelled key arguments and concepts
  - Grouped and refined themes

Results

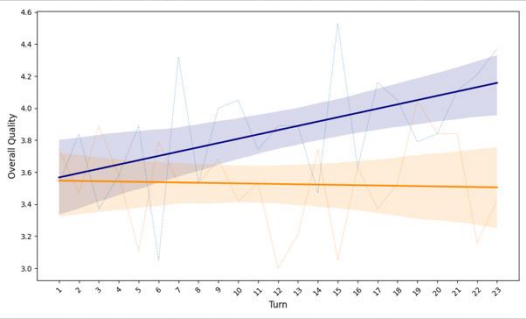


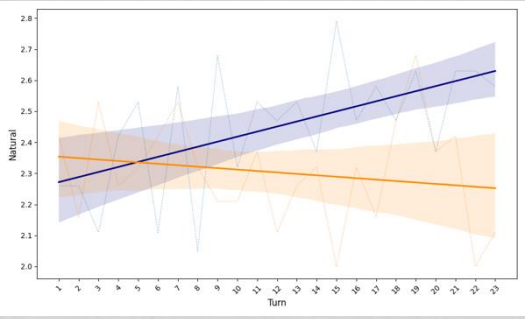
- **Finding 1:**
  Participants prefer human-authored responses in 20 out of 23 turns of the conversation

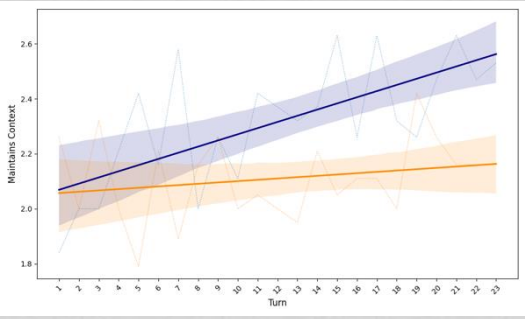
Results

- **Finding 2:**
  LLM-generated response quality degrades over time, while human-authored response quality improves.
  - Overall Quality (β = –0.029, p = .001)
  - Natural (β = –0.021, p < .001)
  - Maintains Context (β = –0.020, p < .001).

“How do the quality perceptions of LLM-generated and human-authored responses change as a dialogue progresses over turns?”



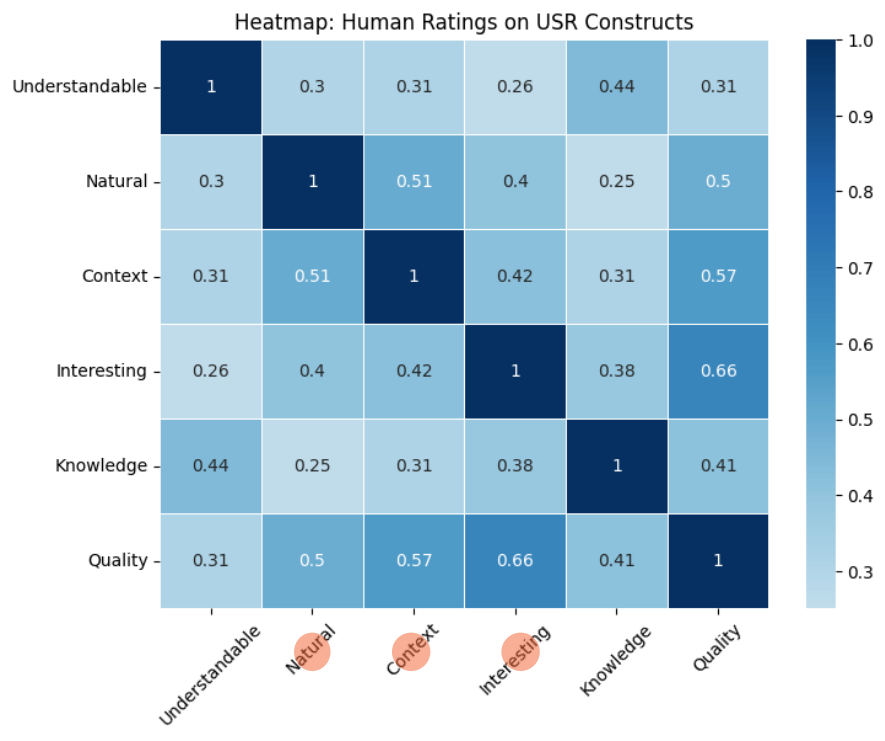




Results

- **Finding 3:**
  Factors that influence quality perceptions most:
  - Interesting
  - Maintains Context
  - Naturalness

““What key factors most strongly influence participants’ perceptions of response quality?”



Results

- **Finding 3:**
  Factors that influence quality perceptions most:
  - Interesting
  - Maintains Context
  - Naturalness

""What key factors most strongly influence participants' perceptions of response quality?"

- **More potential factors?**
  - Tone Appropriateness:
    the tone of the response must align with the scenario dynamics
  - Pedagogical Nudging
    a response can nudge towards intended learning objectives without feeling artificial
  - Sentence Length
    a response's primary intent should be obvious within first few words

III

Automated Evaluation

Study Design

To examine the **generalizability** of Human Evaluation's findings, we employed an **LLM-as-a-judge** approach on two tasks spanning **three additional scenarios**: motivational interviewing, selling, and consulting.

I

Pairwise Preference

"Which response do you think fits best within the conversation?"

II

Construct Rating

Understandable, Natural, Maintains Context, Interesting, Uses Knowledge, Overall Quality

Study Design

I

Pairwise Preference

“Which response do you think fits best within the conversation?”

- **Validation**
  GEMINI 2.0 FLASH under zero-shot setting ($r_P = 0.656$).
- **Generalization**

Study Design

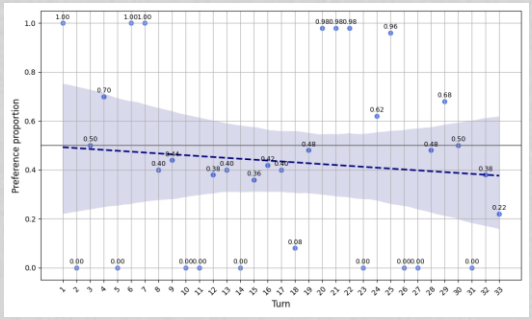
II

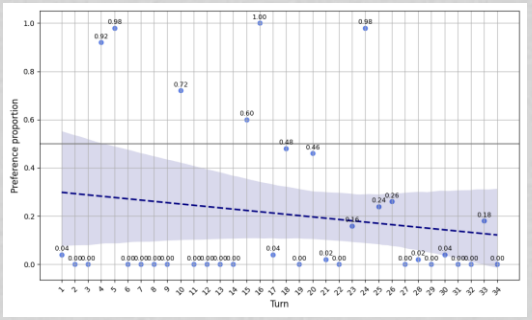
Construct Rating

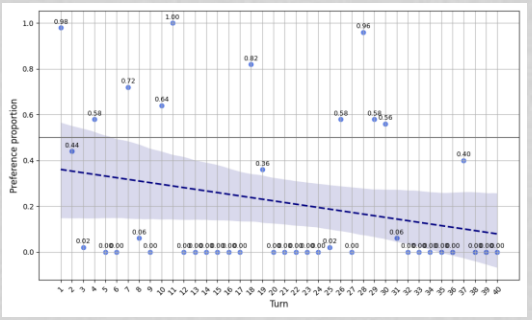
Understandable, Natural, Maintains Context, Interesting, Uses Knowledge, Overall Quality

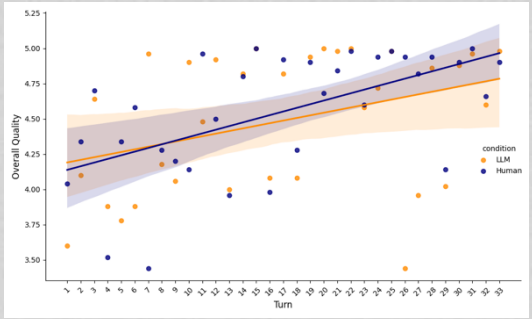
- **Validation**
  - **LLMs:** LLAMA 3.1 8B, MISTRAL 7B, PHI-3 MEDIUM 14B, and GEMINI 2.0 FLASH
  - **Prompting Strategies:** Zero-shot/ 3-shot/ 6-shot, First-k/Random
  - GEMINI 2.0 FLASH with 6-shot random sampling strategy (r_p =0.659 for Overall Quality).
- **Generalization**

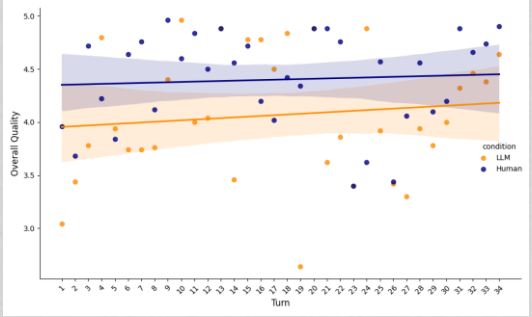
Results

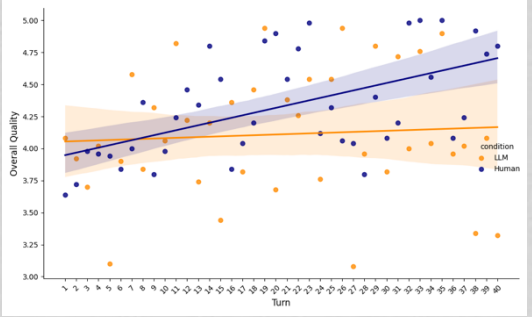

Preference proportion
Turn

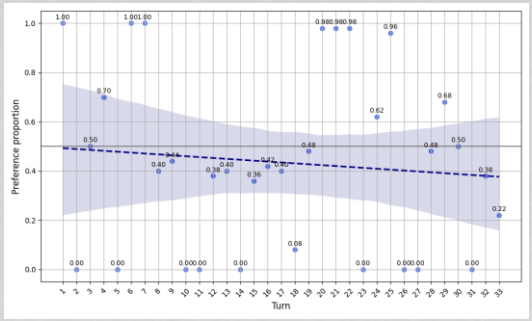

Preference proportion
Turn

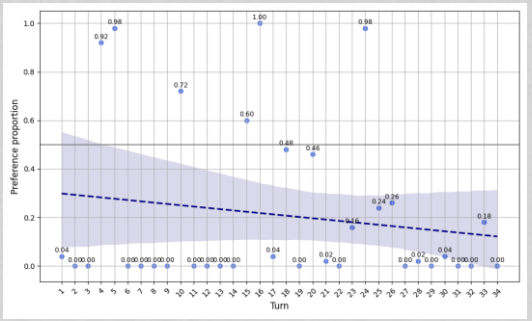

Preference proportion
Turn

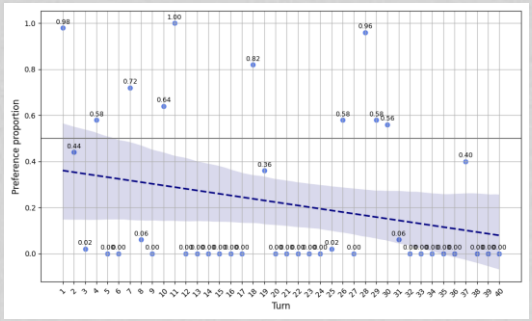

Overall Quality
Turn
condition
LLM
Human


Overall Quality
Turn
condition
LLM
Human


Overall Quality
Turn
condition
LLM
Human

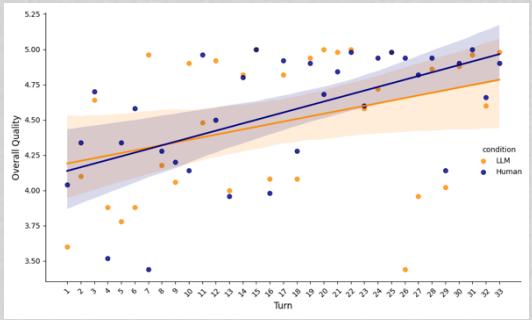


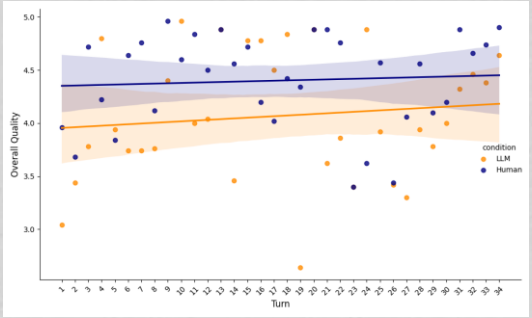


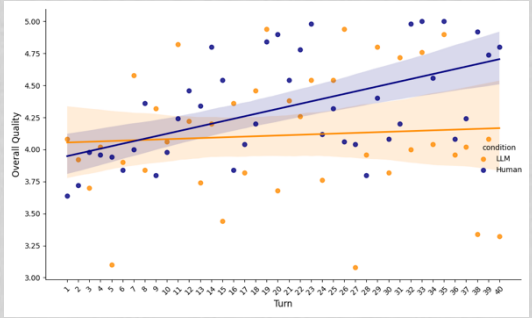


- ◊ **Finding 4:**

  The preferences towards human-authored responses and quality gap between LLM-generated/human-authored responses generaliszes across additional scenarios.







IV

Discussion & Conclusions

Conclusions

I:

LLM-generated responses demonstrate **quality degradation** in long-form, knowledge-grounded role-play dialogues.

II:

Human-authored responses show the **opposite trend**, with quality improving over turns.

.

III:

We provide a **validated hybrid evaluation framework** using an **LLM-as-a-judge** approach.

.

Limitations

I:

Human evaluation was limited to a **single scenario**.

II:

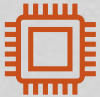
The LLM-as-a-judge showed **potential biases** and **calibration needs**.

III:

The study relied on one generation model (LLAMA 3) , which showed significant **performance decay**.

Future Work

Extend to more scenarios and topics



Explore models of varied capabilities

Explore strategies to mitigate performance decay

Look into linguistic dialogue structures

Main Takeaway

Due to the degradation of LLM performance over extended interactions, **human authors remain the "gold standard"** for crafting role-play training simulations.

Thank You!

Dongxu Lu (Dawn)
contact: d.lu@uu.nl