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

Dongxu Lu, Johan Jeuring, Albert Gatt

*Utrecht University, Utrecht, The Netherlands, {**d.lu**, j.t.jeuring, a.gatt}@uu.nl*

Introduction & Motivation

- Large Language Models (LLMs) are increasingly used in **role-play dialogue systems** for professional skills training, such as simulations designed for enhancing communication skills in our study.
- These simulations require responses to be **engaging, contextually appropriate and grounded in specific knowledge**.
- **The Gap:** It is unclear **how LLM-generated responses compare against a high-quality, human-authored benchmark in knowledge-grounded role-play contexts**. Existing benchmarks often fail to assess LLM performance **over multiple turns**.

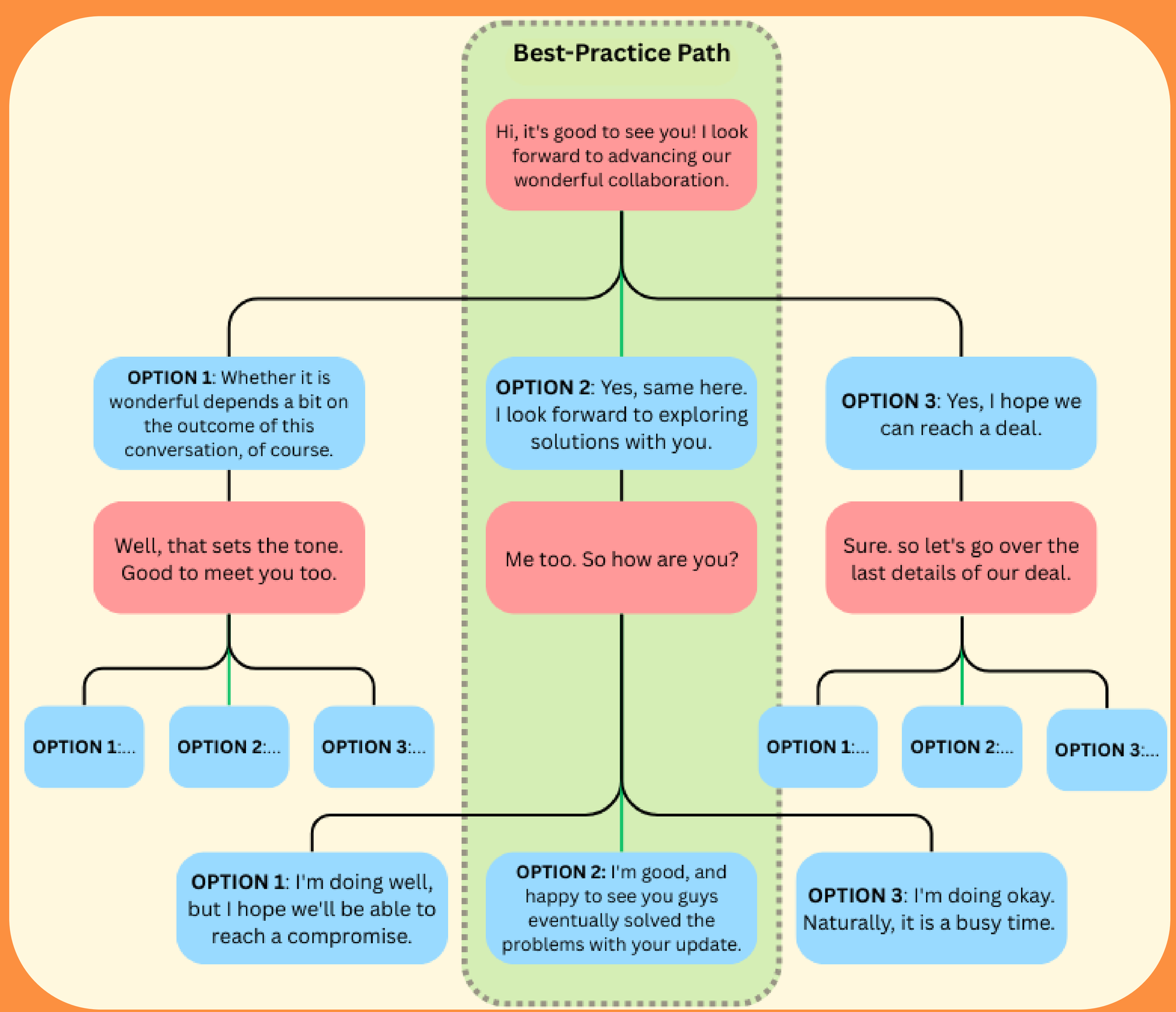
METHODS

HUMAN EVALUATION

Stimuli

We employed a negotiation scenario to obtain response pairs:

- **Human-Authored:** Pre-scripted "best-practice path" agent responses.
- **LLM-Generated:** Alternative agent responses from for the same path using a fine-tuned LLAMA 3 model.



Tasks

In-simulation Pairwise Preference: At each turn, participants chose the response that "fits best".

Post-simulation Rating: Participants rated on 6 quality dimensions (e.g., Naturalness, Maintains Context, Overall Quality).

Focus Group

Qualitative insights from two instructional designers.

AUTOMATED EVALUATION

To examine the generalizability of Exp 1's findings, we employed an LLM-as-a-judge approach on two tasks spanning three additional scenarios: motivational interviewing, selling, and consulting.

Construct Ratings

- **Validation:** We prompted **different LLMs** (LLAMA 3.1 8B, MISTRAL 7B, PHI-3 MEDIUM 14B, and GEMINI 2.0 FLASH) with **varied prompting strategies** (zero-shot, 3-shot, 6-shot, etc.) and computed their ratings' correlations against the human judgments (Exp 1).

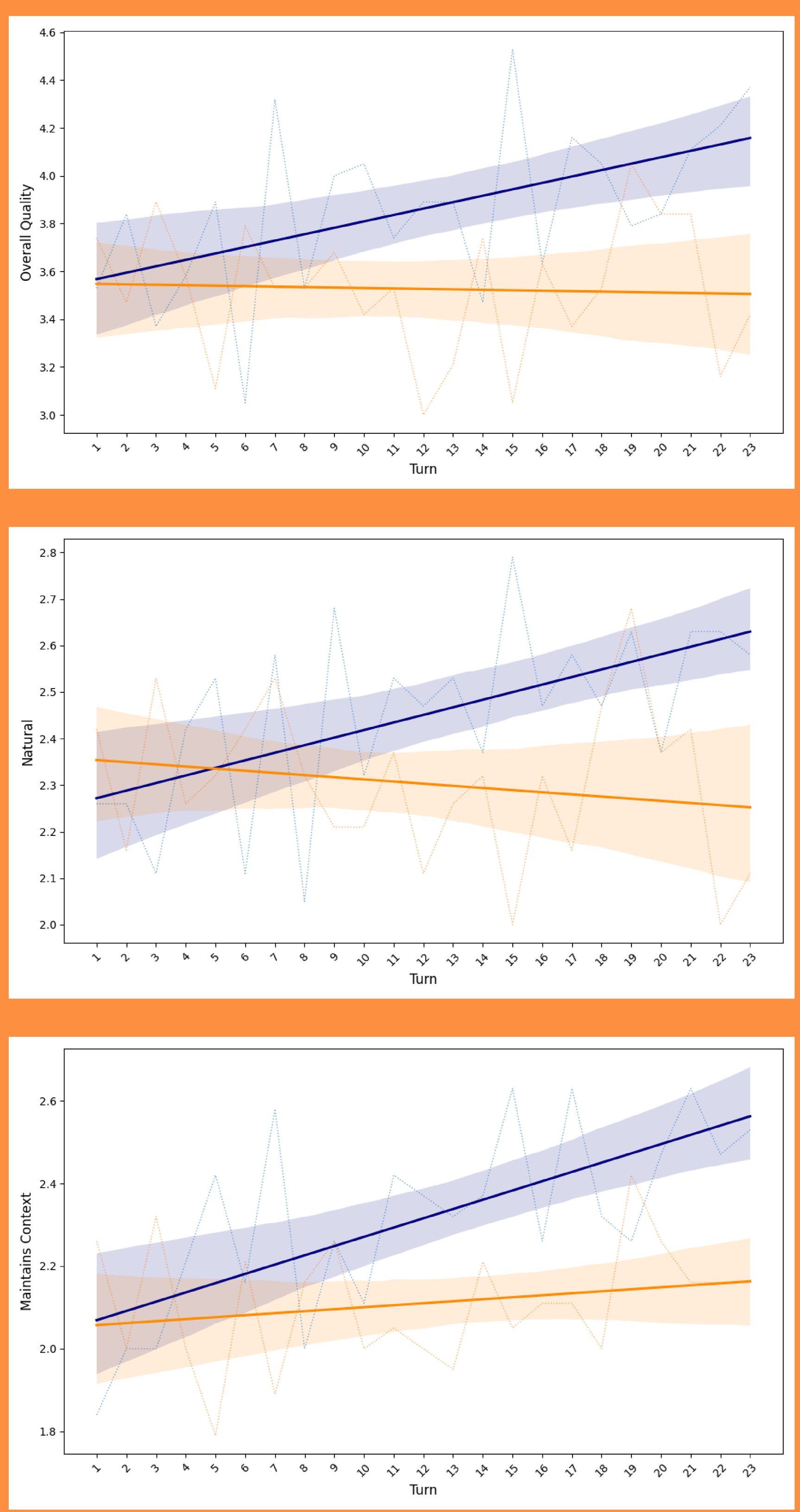
  The LLM-judge, **GEMINI 2.0 FLASH with a 6-shot random sampling strategy**, achieved highest agreement with human ratings (r_p=0.659 for Overall Quality).
- **Generalization:** The validated LLM-judge was prompted to give ratings on responses across three scenarios.

Pairwise Preference

- **Validation:** The LLM-judge, **GEMINI 2.0 FLASH under zero-shot setting**, already achieved high alignment with human preferences (r_p=0.656).
- **Generalization:** The validated LLM-judge was prompted to indicate preferences for response pairs across additional scenarios.

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

RESULTS



➢ Finding 1
LLM-generated response quality degrades over time, while human-authored response quality improves.

➢ Finding 2
Participants prefer human-authored responses in 20 out of 23 turns of the conversation.

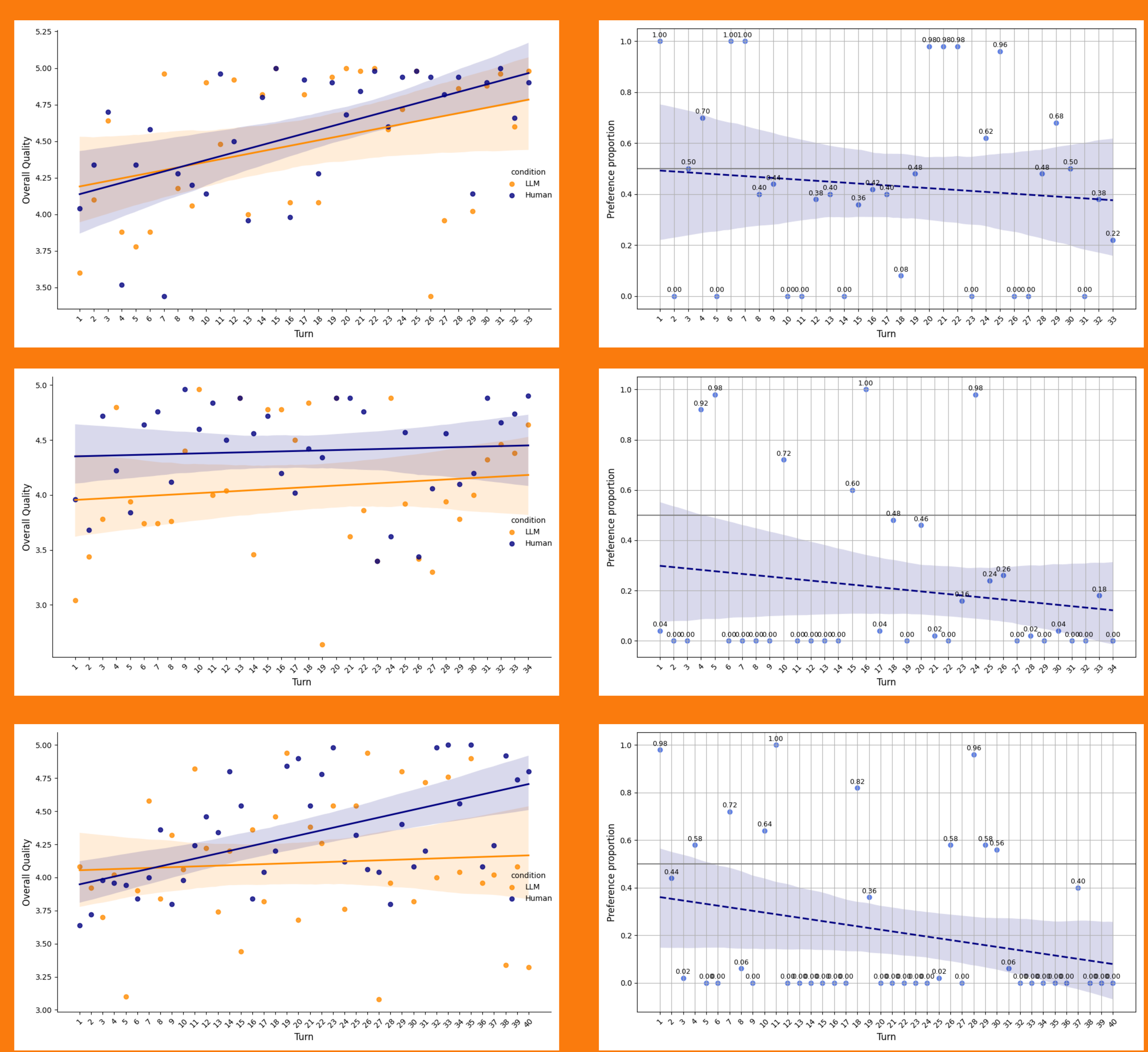
➢ Finding 3
Factors that influence quality perceptions most:

- **Interesting**
- **Naturalness**
- **Maintains Context**

More potential factors?

- Tone Appropriateness
- Pedagogical Nudging
- Sentence Length

➢ Finding 4
The quality gap between LLM-generated and human-authored responses generalizes across all three scenarios. LLM-judge also showed consistent preferences for human-authored responses.



CONCLUSIONS

- LLM-generated responses demonstrate **quality degradation** in long-form, knowledge-grounded role-play dialogues.
- Human-authored responses show the **opposite trend**, with quality improving over turns, widening the performance gap as the conversation progresses.
- We provide a **validated hybrid evaluation framework** using an LLM-as-a-judge approach.

Utrecht University

DialogueTrainer

Main Takeaway

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