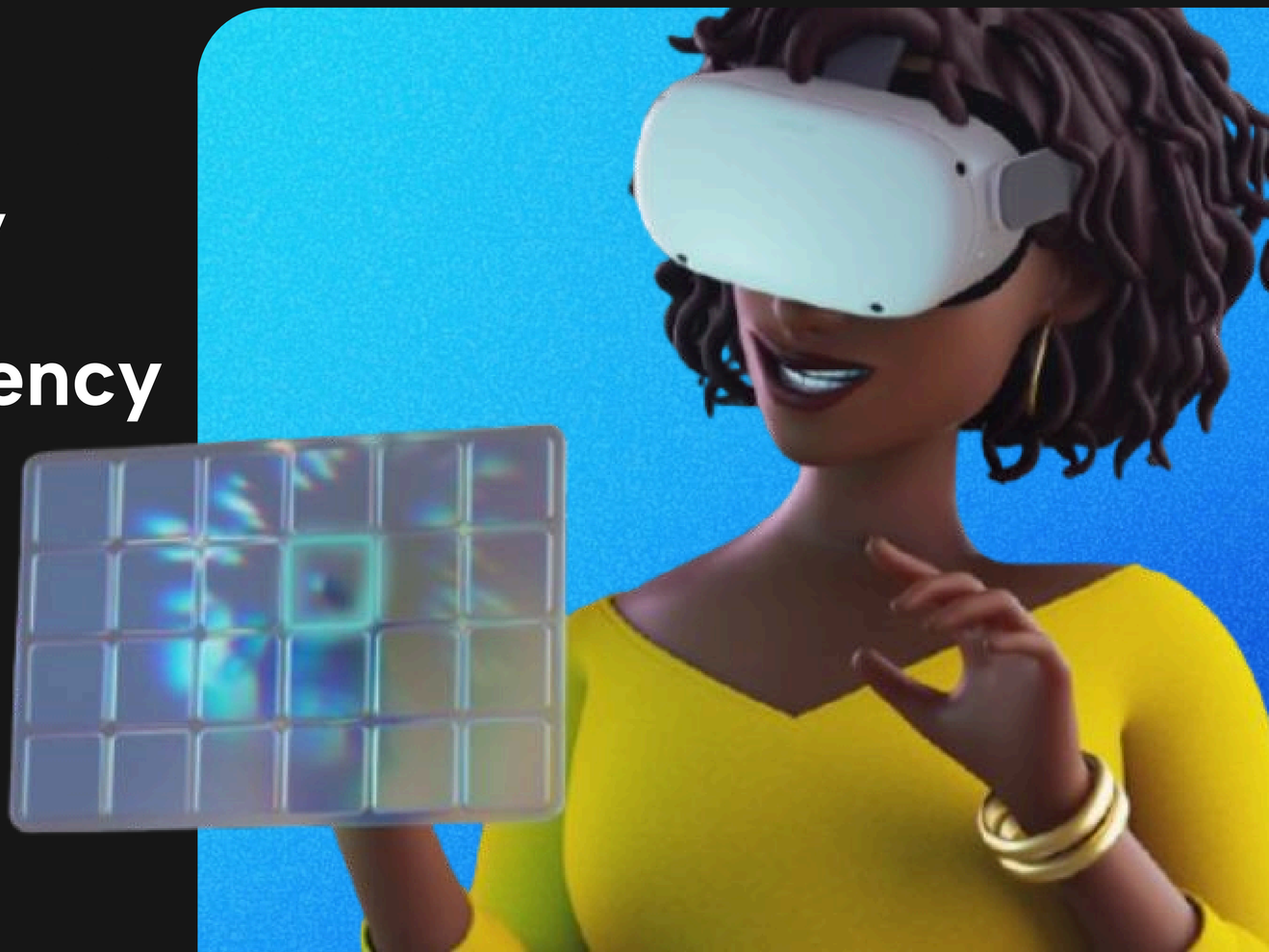
Someone: nice weather

LLM:

Maintaining Flow in VR Conversations: A Modality-Level Analysis of LLM Delay Feedback on Perceived Latency

End-sem presentation



RESEARCH QUESTION

When measured independently, which feedback modality (**verbal filler**, **gesture**, or **visual cue**) most effectively reduces perceived latency and enhances user immersion in conversations with an LLM-based 3D avatar in VR environment?

LITERATURE REVIEW

December 2024

Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality

- 18 participants; two conditions: (1) baseline with no feedback vs. (2) all feedback modalities combined (filled pauses, 2 nonverbal animations, visual spinner icon); VR interaction: 7-minute role-play conversation about workload management with LLM-powered avatar using GPT 3.5-Turbo, Whisper ASR, and OpenAI TTS.

- Provides empirical evidence that combined multimodal feedback (filled pauses + gestures + visual indicators) significantly improves presence ($p=0.03$), engagement ($p=0.05$), and response time perception ($p=0.02$) in LLM-avatar VR interactions.

Morad Elfleet and Mathieu Chollet. 2024. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. In Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA '24). Association for Computing Machinery, New York, NY, USA, Article 12, 1–9. https://doi.org/10.1145/3652988.3673965

LITERATURE REVIEW

July 2025

Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents

- Found that latency above 4 seconds degrades quality of experience, while natural conversational fillers improve perceived response time, especially in high-delay conditions.
- Participants experienced three virtual worlds, conversing with 9 virtual agents under three filler types: None, Artificial, and Natural, along with three response latency levels: Low (1.5s), Medium (4.0s), and High (6.5s)
- Created two custom questionnaires informed by prior literature, as no standardized survey exists for interface latency and its mitigation.

Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, Ryan P. McMahan, and Joseph J. LaViola Jr. 2025. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25). Association for Computing Machinery, New York, NY, USA, Article 49, 1–15. https://doi.org/10.1145/3719160.3736636

EXISTING SOLUTIONS & APPLICATIONS (where it could help!)

Games like VRCHAT

- Reduces awkward silence by making avatars (NPCs) feel more responsive.
- Increases social presence through natural filler speech/ gestures.
- Enhances immersion for roleplay, events, and live interactions.

EXISTING SOLUTIONS & APPLICATIONS (where it could help!)

- More realistic conversational flow for language practice.
- Use culture specific pauses

Language Learning Simulators like NounTown

EXISTING SOLUTIONS & APPLICATIONS (where it could help!)

- Creates smoother dialogue during mock interviews.
- Helps learners feel less anxious with natural avatar behavior.
- Makes feedback loops feel more human and timely.

Job Interview Simulators like bodyswaps

EXISTING SOLUTIONS & APPLICATIONS (where it could help!)

VR Therapy

- Builds stronger client comfort via natural conversational flow.
- Reduces the “robotic pause” that breaks therapeutic presence.
- Supports deeper immersion in exposure or guided-talk scenarios.

OUR TECH STACK

- Unity Engine with ReadyPlayerMe avatar
- Google Cloud STT, TTS
- Gemini 2.5 Flash
- Meta Quest 2

OUR METHODOLOGY

A sample experiment

Consent form & Demographic Survey

Provided insights into:

- Age & Gender
- Experience with VR
- Frequency of Chatbots/AI assistants use

Session 1

Session 2

Session 3

Session 4

Gestures

Verbal Fillers

Baseline

Visual Cue

- **Session Protocol:** Randomized sessions (5–8 mins) featuring four tasks and a unique feedback modality.
- **Post-Session Evaluation:** A survey consisting of:
  - 5 general questions (Godspeed Questionnaire)
  - 6 modality-specific questions
  - 1 open-ended feedback section

MEET ALEX


Start conversation
(Right trigger Button)


Tasks:
☐ Ask for a campus landmark
☐ Learn which dorm Alex lives in
☐ Ask if Alex has siblings
☐ Discover Alex's music taste
4

Verbal fillers (23 in total) like hmm, uh, you know, actually, and let me think

POST SESSION EVALUATION

1. Godspeed Questionnaire Series (1–5 Semantic Differential Scale):

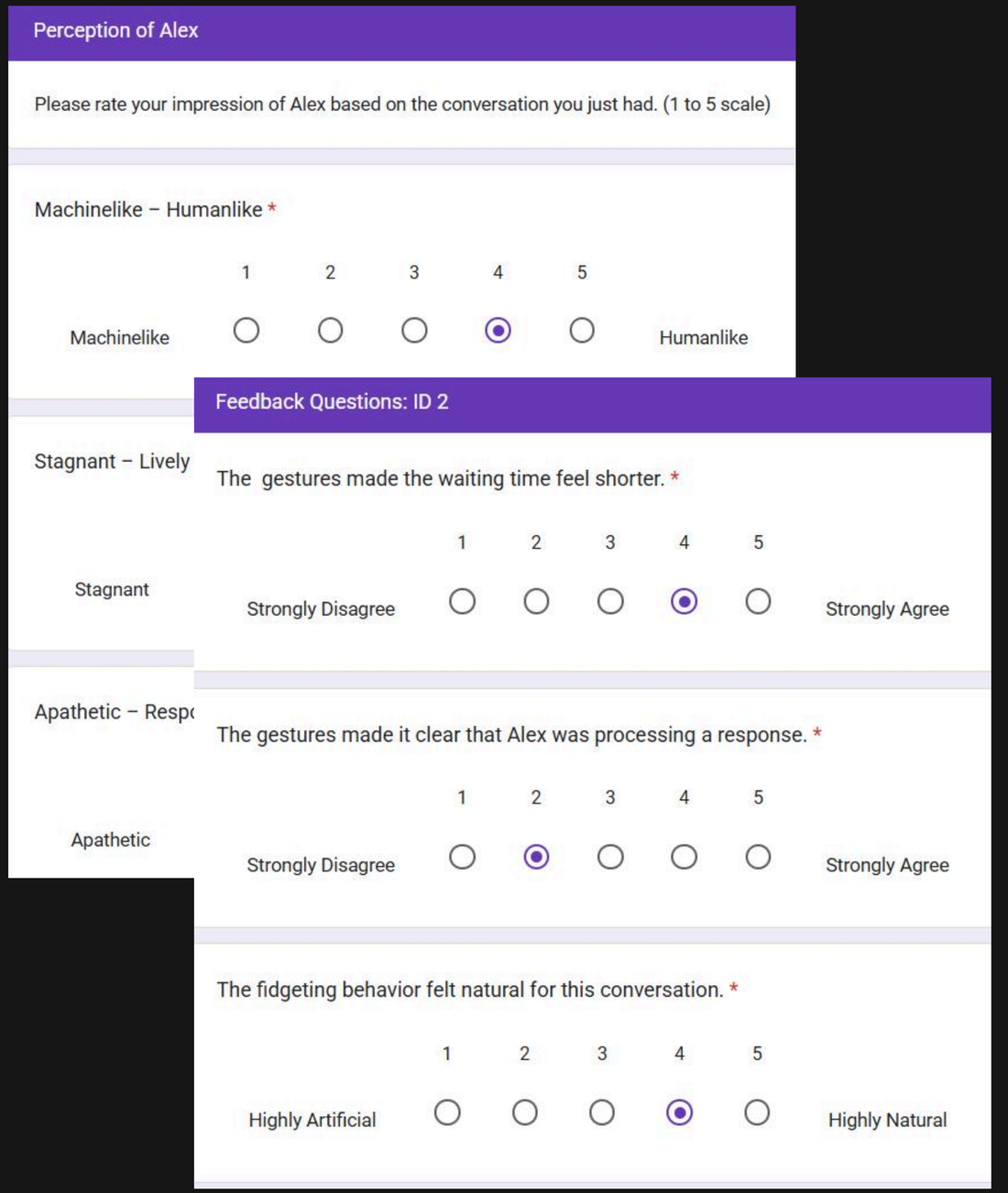
- Anthropomorphism (Humanlike vs. Machinelike)
- Animacy (Lively vs. Stagnant)
- Likeability (Friendly vs. Unfriendly)
- Perceived Intelligence (Intelligent vs. Unintelligent)
- Responsiveness (Responsive vs. Apathetic)

2. Custom Experience Metrics (Likert Scale 1–5):

- Perception of Time: "Did the feedback make the waiting time feel shorter?"
- User Experience: Naturalness, Distraction, and Co-Presence.

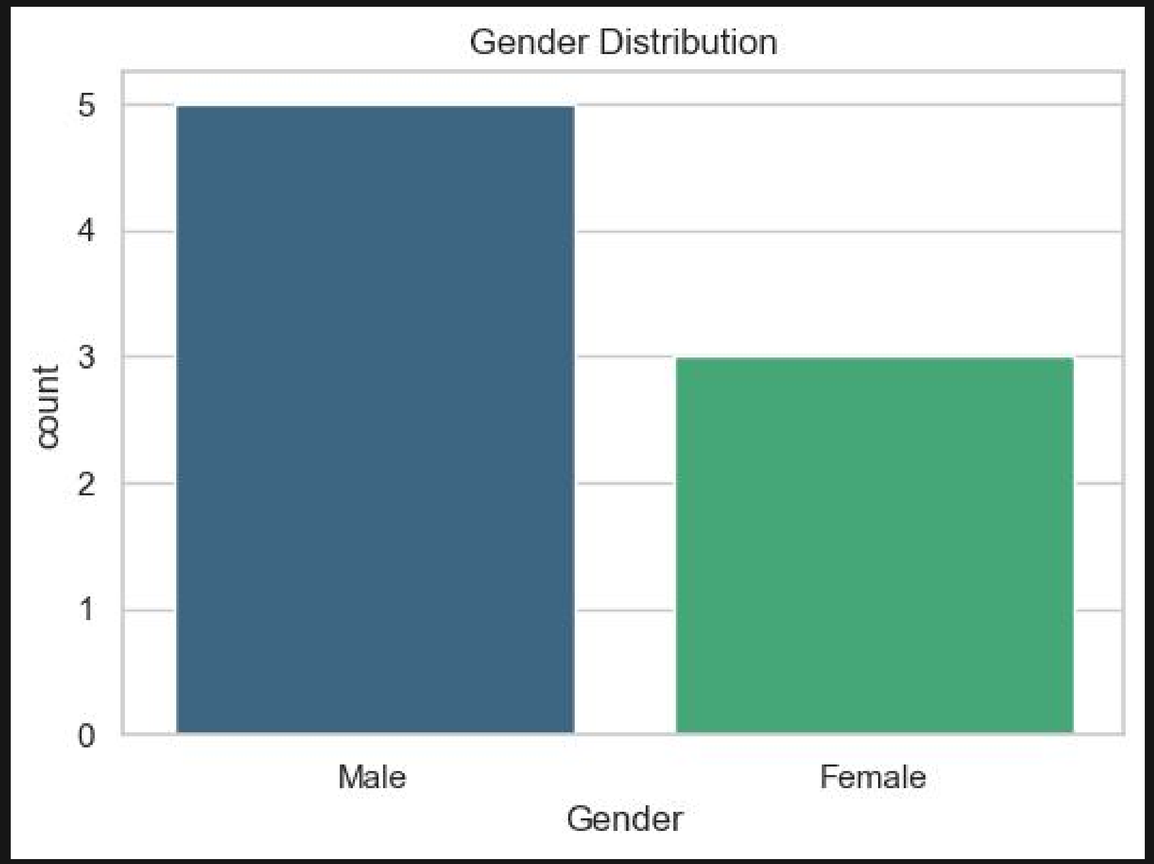
3. Interaction Logs:

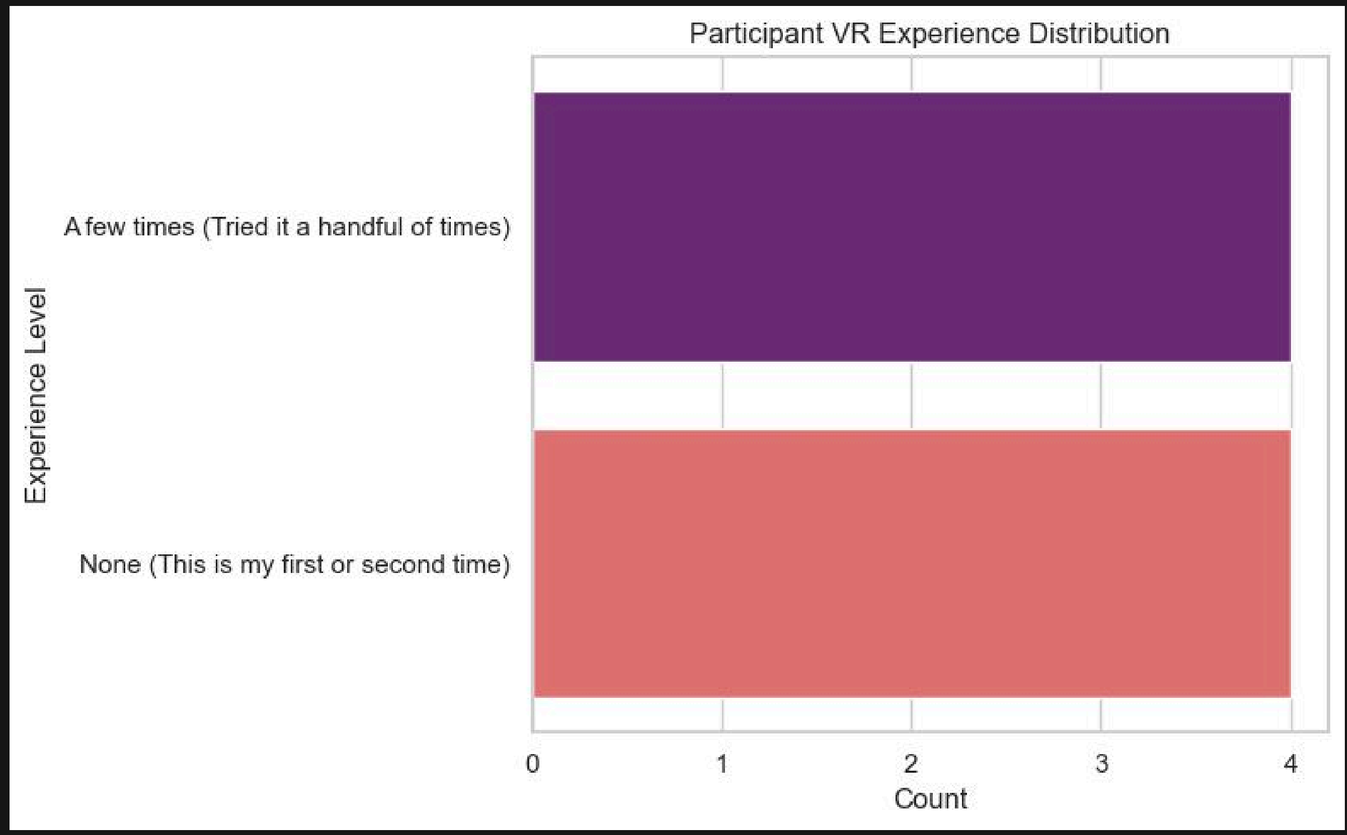
- Verbal Analysis: User Word Count (Openness) and Turn Count (Engagement).
- System Performance: Actual Response Time vs. Perceived Responsiveness.



DATA DEMOGRAPHICS

8 students (18-22 years) in a within-subjects design (each participant experienced all 4 conditions) with 2 participants in our pilot study.



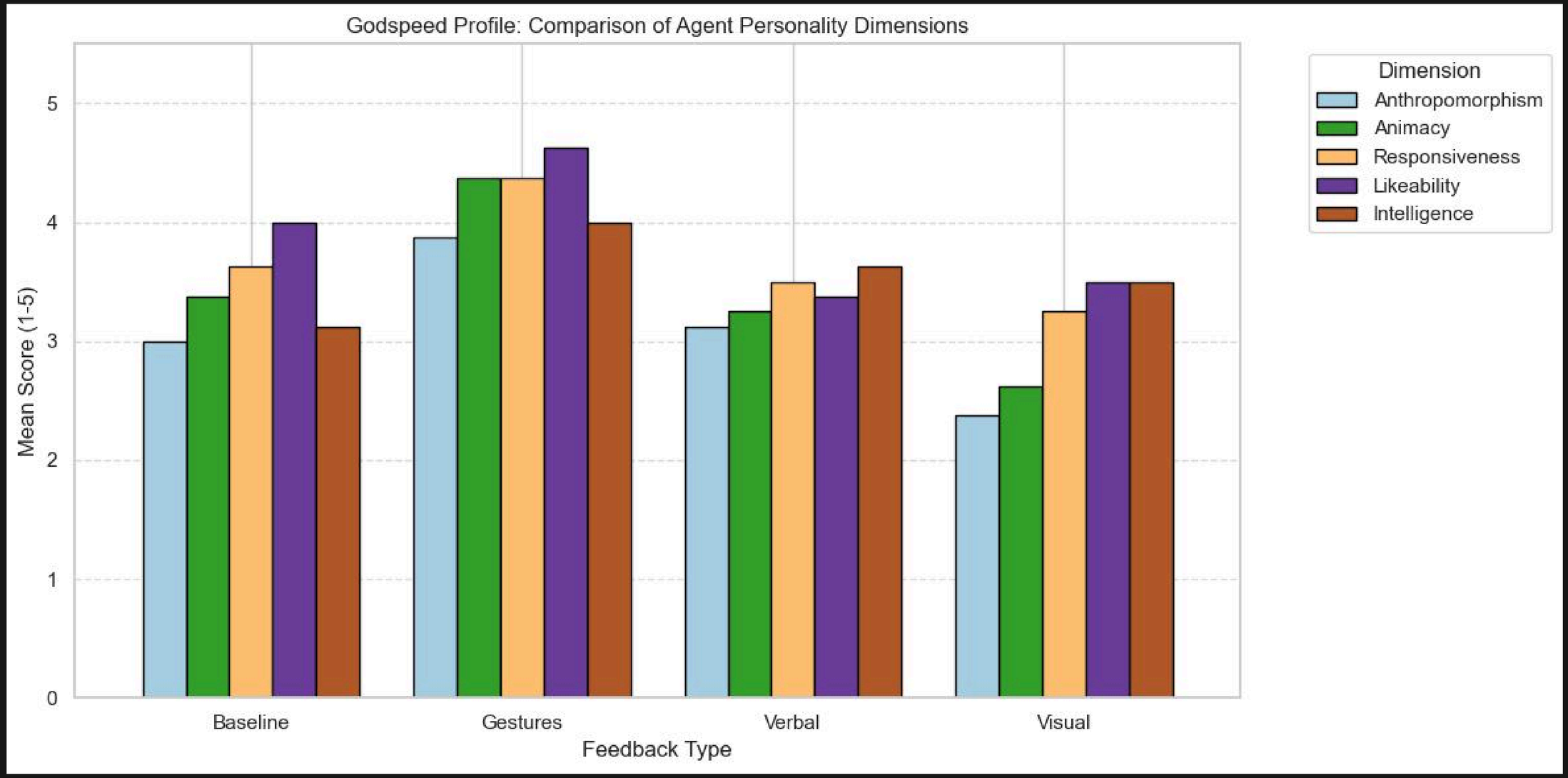


POST PILOT SESSIONS IMPROVEMENTS

1. **Problem:** The agent used the same static greeting ('Hey bro...') every time, which risked annoying users.
    - **Solution:** We randomized the opening lines to prevent fatigue and ensure a gender neutral start for each session.
2. **Problem:** The feedback mechanisms were too subtle and often missed by participants.
    - **Solution:** We adjusted the loading icon's placement and scale & increased number of gestures.
3. **Problem:** Repeating the same four conversation topics across all sessions caused user disengagement.
    - **Solution:** We implemented a dynamic system with a pool of 20 questions. Each session now randomly selects four unique topics, ensuring no participant ever answers the same question twice.

And added natural talking and blinking animations, along with context-agnostic verbal fillers that work smoothly in most situations.

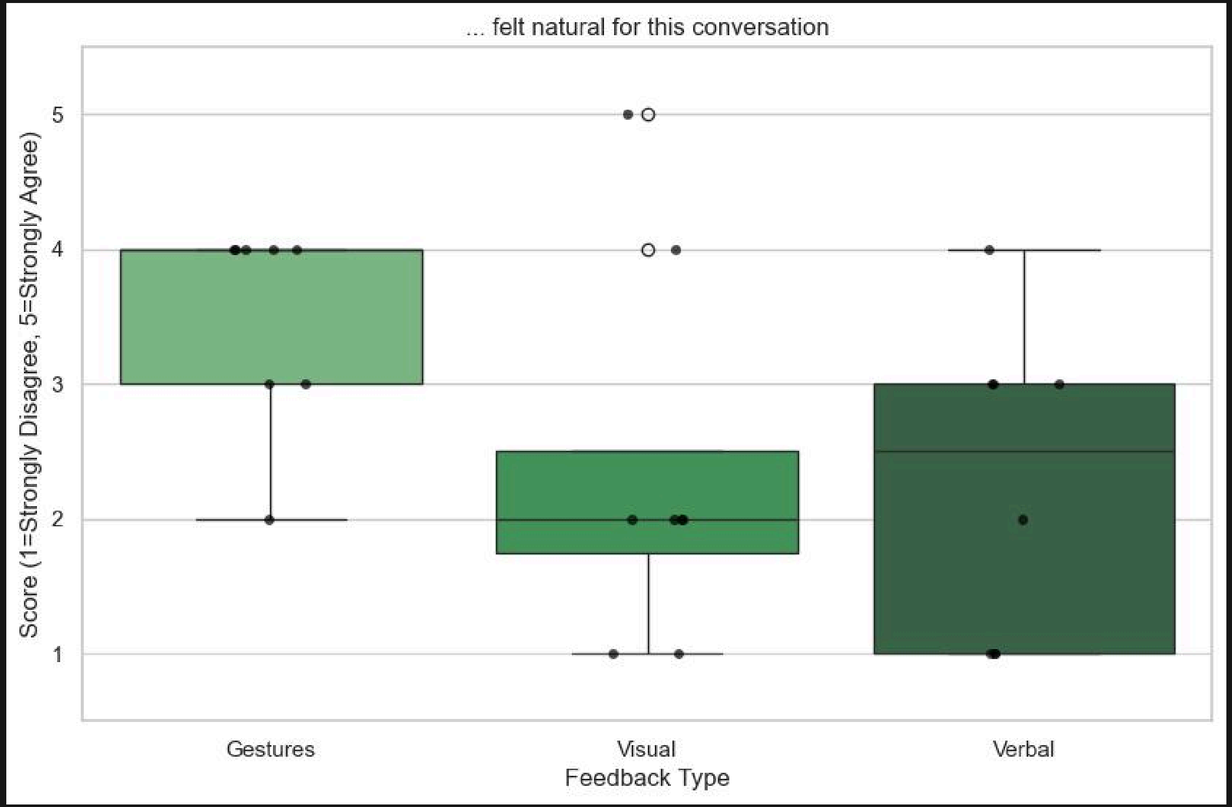
METRICS AND ANALYSIS

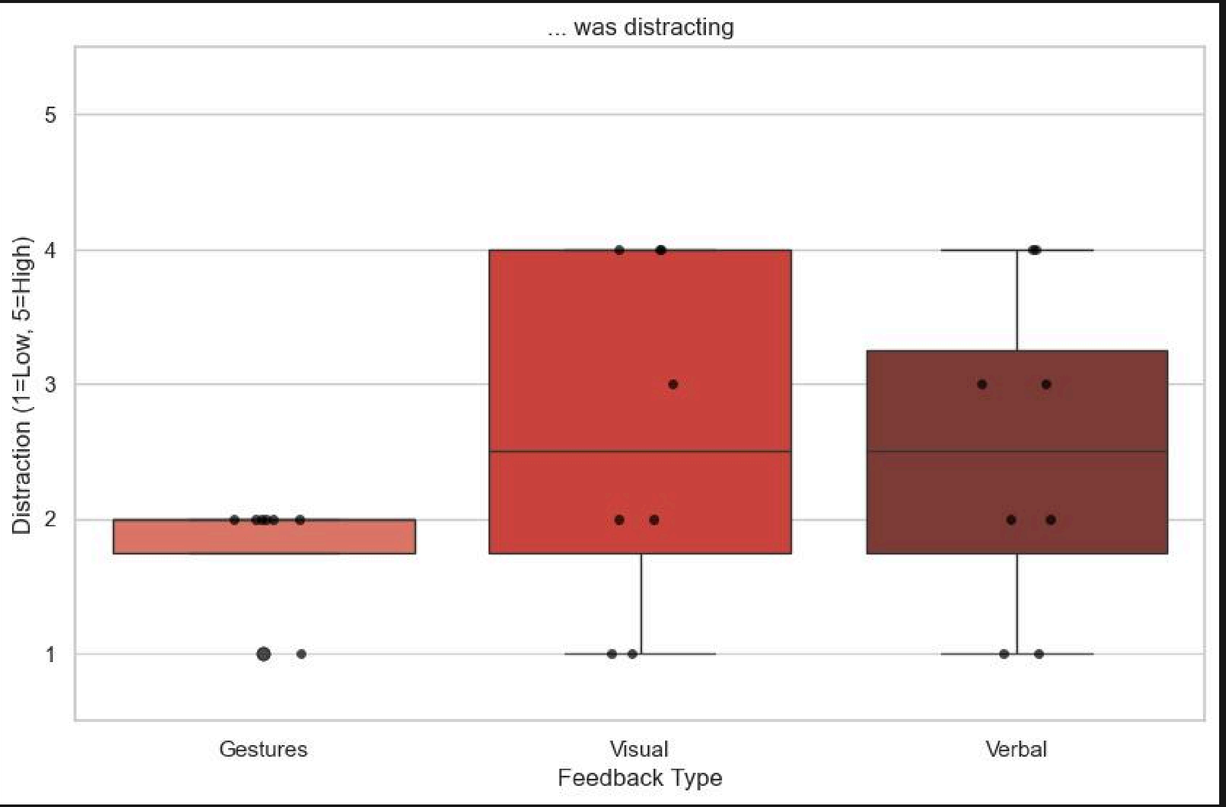


METRICS AND ANALYSIS

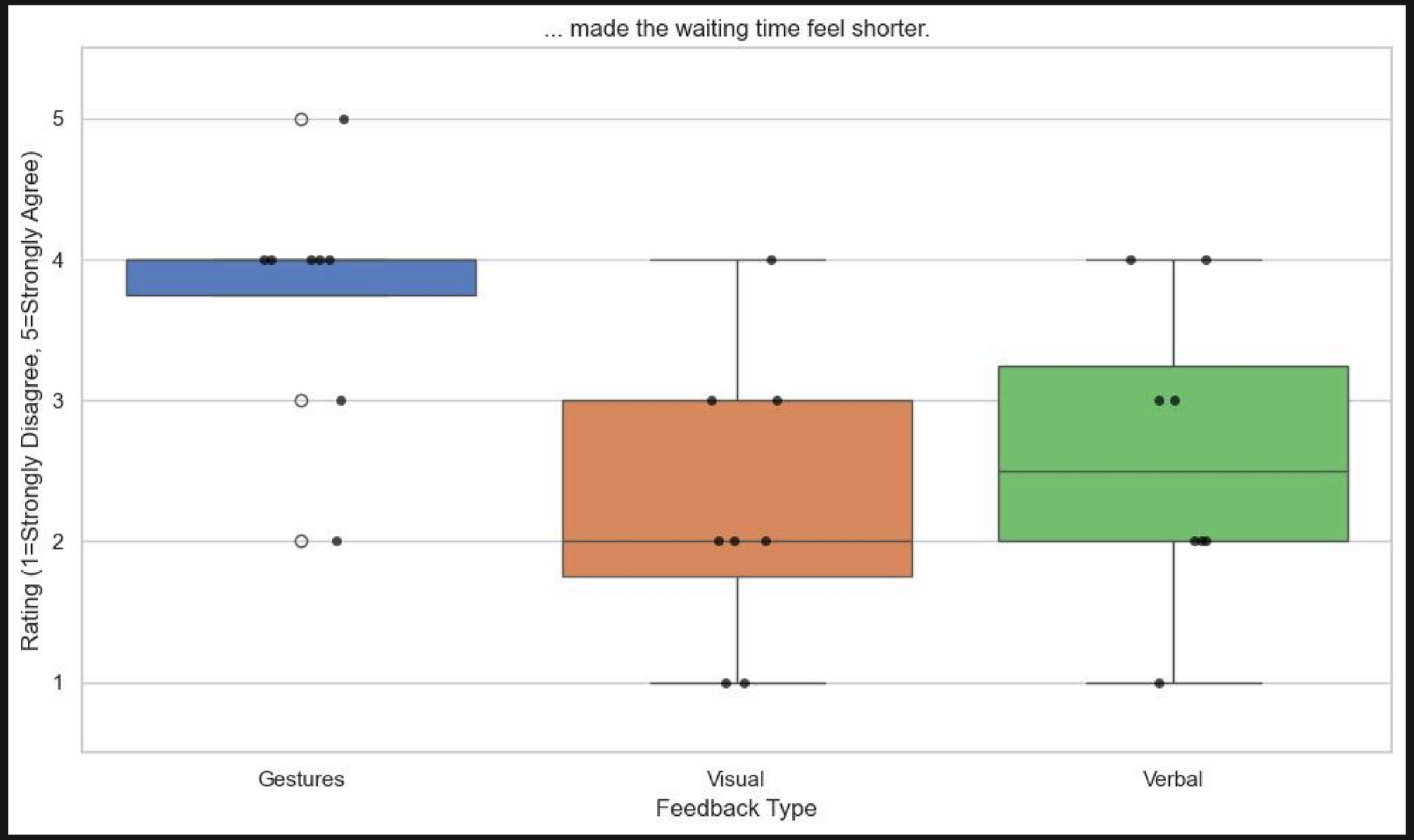
Metric	Comparison	Base Mean (SD)	Cond Mean (SD)	p-value
Anthropomorphism	Base vs Gestures	3.00 (1.07)	3.88 (0.64)	0.072
Anthropomorphism	Base vs Visual	3.00 (1.07)	2.38 (1.06)	0.26
Anthropomorphism	Base vs Verbal	3.00 (1.07)	3.12 (0.83)	0.798
Animacy	Base vs Gestures	3.38 (0.92)	4.38 (0.52)	0.021 *
Animacy	Base vs Visual	3.38 (0.92)	2.62 (0.74)	0.095
Animacy	Base vs Verbal	3.38 (0.92)	3.25 (1.04)	0.802
Responsiveness	Base vs Gestures	3.62 (0.92)	4.38 (0.74)	0.095
Responsiveness	Base vs Visual	3.62 (0.92)	3.25 (0.89)	0.419
Responsiveness	Base vs Verbal	3.62 (0.92)	3.50 (1.07)	0.805
Likeability	Base vs Gestures	4.00 (1.60)	4.62 (0.52)	0.323
Likeability	Base vs Visual	4.00 (1.60)	3.50 (0.93)	0.461
Likeability	Base vs Verbal	4.00 (1.60)	3.38 (1.19)	0.392
Intelligence	Base vs Gestures	3.12 (0.83)	4.00 (0.53)	0.028 *
Intelligence	Base vs Visual	3.12 (0.83)	3.50 (1.07)	0.448
Intelligence	Base vs Verbal	3.12 (0.83)	3.62 (0.92)	0.273

METRICS AND ANALYSIS





METRICS AND ANALYSIS



FUTURE WORK AND LIMITATIONS

Implement a compact on-device model that dynamically selects verbal fillers based on real-time conversational context

Enable more active avatar behaviour beyond idle sitting and incorporate mid-conversation streaming of LLM responses for reduced perceived latency.

Integrate more realistic facial expressions, eye-gaze, and head-movement animations, and design scenario-driven test cases to evaluate the impact of different non-verbal modalities.

Use a larger and more diverse participant sample, and experiment with controlled fixed-latency conditions (as explored in Paper 2) to better isolate modality-specific effects.

Tho it was a really fun and we learnt a lot!