

Kurt Lewin Institute Conference 2026
May 11-12, 2026, Conference Centre Woudschoten, Zeist

Trustworthy Conversational AI Agents for Mental Health Support

Speaker: Jiahuan Pei

Assistant Professor
Vrije Universiteit Amsterdam

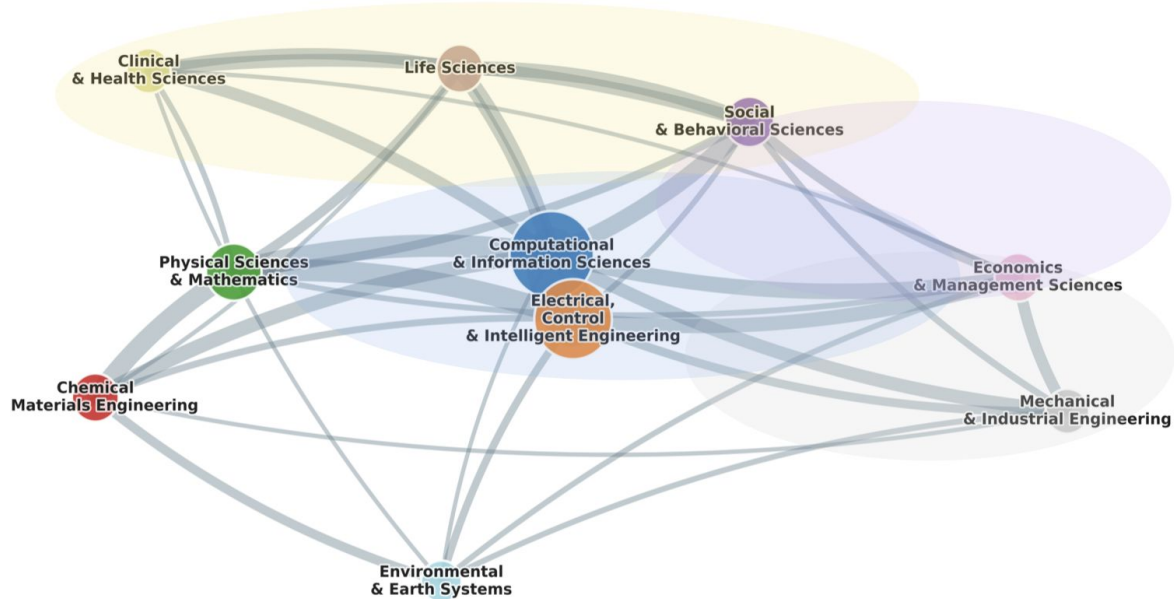


Agenda

- Introduction
- Agent Simulation for Counselling & Coaching
- Psycholinguistic Bonding in Conversational Interactions
- Gaze Perception and Understanding Beyond Conversations
- Challenges & Opportunities

Research across Multiple Disciplines

Discipline network of 1,706 surveyed papers (2021-2025) related to trust spans multiple disciplines but remains fragmented. Node size indicates literature volume; edge thickness reflects connection strength.



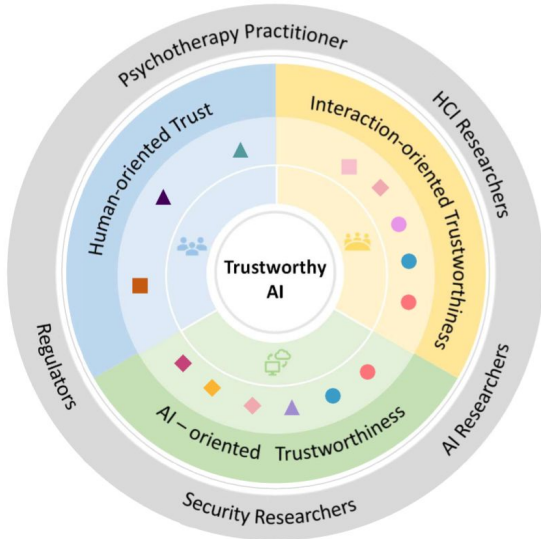
Conversations shape how we learn, create, and connect.

What measurable criteria should determine whether a conversational AI is trustworthy enough for mental health support?



Trustworthy Criteria

- Transparency ■ Empathy ● Controllability ● Safety
- User Characteristics ▲ Explainability ◆ Reliability ◆ Fairness
- ▲ Anthropomorphism ◆ Accuracy ▲ Robustness



1 TRANSPARENCY

LOGS
ALGORITHMS
DECISION PATHS

USERS CAN UNDERSTAND AI PROCESSES

2 EMPATHY

AI RECOGNIZES & RESPONDS TO EMOTION

3 CONTROLLABILITY

PAUSE ALGORITHM ADJUST SENSITIVITY CHANGE TONE

USER HAS OVERRIDE & SETTING POWER

4 SAFETY

SECURE DATA PRIVATE CLOUD

DATA PRIVACY & RISK MITIGATION

5 USER CHARACTERISTICS

CULTURAL CONTEXT LANGUAGE (SPANISH/ ENGLISH)
AGE (ADULT/TEEN)

AI ADAPTS TO INDIVIDUAL DIVERSITY

6 EXPLORABILITY

CBT APPROACH DBT EXPLORATION
ACT OPTIONS ACT OPTIONS

ABILITY TO DISCOVER & CHOOSE APPROACHES

7 RELIABILITY

SESSION COMPLETION

50
40
30
20
10

3/7/18 2/7/18 5/1/18 3/1/17 2/2/17

DEPENDABLE RESULTS OVER TIME

8 FAIRNESS

FAIR ADVICE NO BIAS

EQUITABLE TREATMENT FOR ALL USERS

9 ANTHROPOMORPHISM

Is it real?
Feels too connected.

AVOID OVERLY HUMAN & MISLEADING CONNECTIONS

10 ACCURACY

MEDICAL DATABASE CLINICAL DATA

MEDICAL KNOWLEDGE & CLINICAL VALIDITY

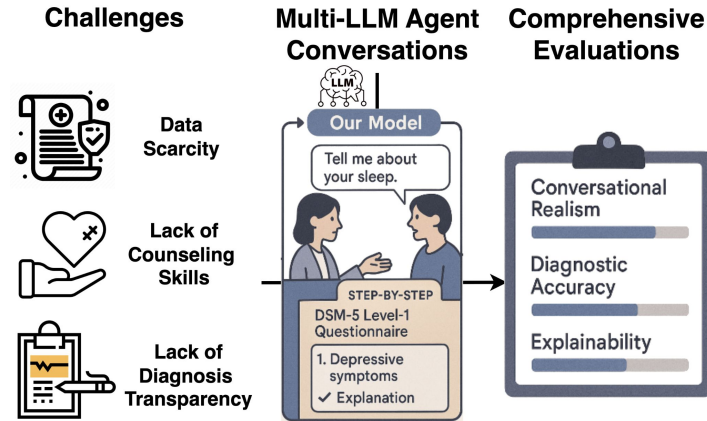
11 ROBUSTNESS

garbled strange queries
garbl so... strange queries

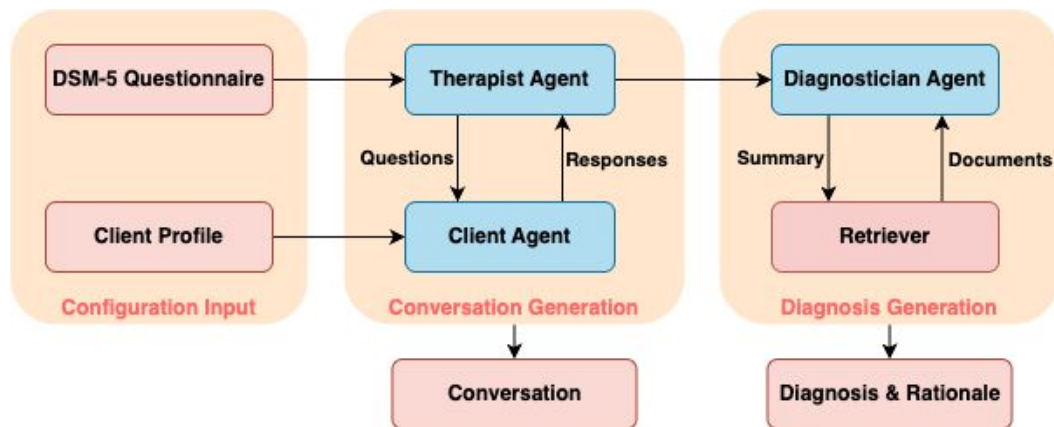
AI successfully converts these into a coherent, helpful, and conversational flow.

RESILIENT IN FACE OF UNPREDICTABLE INPUTS

Agent Simulation for Counselling & Coaching



DSM5AgentFlow



- Can LLMs simulate therapist-client conversations, to effectively complete the diagnostic DSM-5 questionnaire?
- Can disorder type predictions be made by linking questionnaire responses to specific disorder descriptions?
- Is the LLM able to make explainable and transparent diagnoses to enhance the perceived trustworthiness?

DSM5AgentFlow Demo

The screenshot displays the 'Mental Health Multi-Agent' interface. On the left, there is a sidebar with 'Individual Logs' and 'Batch Logs' tabs. The 'Individual Logs' tab shows a list of log entries with timestamps and file names, such as '2025-04-19 18:32:04 - ptsd - cleaned_questionnaire.txt'. The main area is titled 'Conversation: cleaned_questionnaire.txt' and shows a chat interface. The chat history includes an Assistant message: 'Hello, my name is Dr. Ramirez, and I'm a mental health clinician. I'm using the DSM-5 questionnaire today to help us understand your experiences and explore any potential mental health concerns. The purpose of this assessment is to gather information to inform a comprehensive understanding of your well-being. It's important to know that everything you share here is completely confidential, and we're creating a safe space for you to be open and honest. This assessment will involve answering 23 questions. Are you ready to begin with the first question?' and a Patient response: 'Yes, let's start.'. Below the chat, the 'Diagnosis' section provides a 'Summary:' of the patient's presentation, a 'Diagnosis:' of 'Major Depressive Disorder' with a brief explanation, and a 'Reasoning:' section detailing the supporting evidence from the patient's responses.

Mental Health Multi-Agent [Logs] [Generate] [Chat]

Individual Logs [Batch Logs]

Patient Profile:
All [v]
Date Range:
All time [v]
Search:
Search...

2025-04-19 18:32:04 - ptsd - cleaned_questionnaire.txt
2025-04-19 18:31:52 - anxiety_moderate - cleaned_questionnaire.txt
2025-04-19 18:31:41 - anxiety_moderate - cleaned_questionnaire.txt
2025-04-19 18:31:36 - ptsd - cleaned_questionnaire.txt
2025-04-19 18:31:31 - anxiety - cleaned_questionnaire.txt
2025-04-19 18:31:25 - depression - cleaned_questionnaire.txt
2025-04-19 18:31:20 - bipolar - cleaned_questionnaire.txt
2025-04-19 18:31:15 - anxiety_moderate - cleaned_questionnaire.txt
2025-04-19 18:31:09 - ptsd - cleaned_questionnaire.txt
2025-04-19 18:31:04 - ptsd - cleaned_questionnaire.txt
2025-04-19 17:57:08 - ptsd - cleaned_questionnaire.txt

Conversation: cleaned_questionnaire.txt

Date: 4/9/2025, 6:08:27 PM Questionnaire: cleaned_questionnaire.txt Patient Profile: anxiety Assistant Model: gemma3:4b Patient Model: qwen2.5:3b

Question Count: 23

Conversation

Assistant

Hello, my name is Dr. Ramirez, and I'm a mental health clinician. I'm using the DSM-5 questionnaire today to help us understand your experiences and explore any potential mental health concerns. The purpose of this assessment is to gather information to inform a comprehensive understanding of your well-being. It's important to know that everything you share here is completely confidential, and we're creating a safe space for you to be open and honest. This assessment will involve answering 23 questions. Are you ready to begin with the first question?

Patient

Yes, let's start.

Assistant

Okay, great. Let's begin with question 1: Over the past two weeks, how often have you felt down, depressed, or hopeless?

Diagnosis

Summary:

Patient presents with a deeply concerning presentation of persistent emotional distress. The individual expresses profound feelings of hopelessness and low self-worth, coupled with significant anxiety and difficulty managing their emotional responses. The reported avoidance of social interaction and fleeting thoughts about death or suicide highlight a serious level of suffering and underscore the urgent need for intervention. The patient's vulnerability is evident in their repeated expressions of negative self-perception and the associated emotional turmoil. It's clear this individual is struggling immensely, and immediate support is critical to prevent further deterioration.

Diagnosis:

Major Depressive Disorder - likely meeting criteria for single episode or recurrent episodes, given the persistent **dysphoria**, **difficulty concentrating**, and **negative self-perception**. Further assessment is needed to rule out other potential contributing factors and to determine the severity and chronicity of the symptoms.

Reasoning:

The patient's responses strongly suggest a diagnosis of **Major Depressive Disorder**. Several key indicators support this conclusion. Firstly, the repeated expression of **hopelessness** and **low self-worth** is a hallmark symptom of **Major Depressive Disorder**. Secondly, the reported **difficulty concentrating** aligns with the cognitive impairment often associated with

Three Gaps

1. Situational grounding

Likert scores (“depression = 3/4”) don’t tell the client’s *story*. MI is situational.

2. Dynamic strategy control

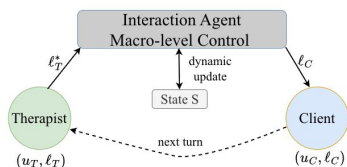
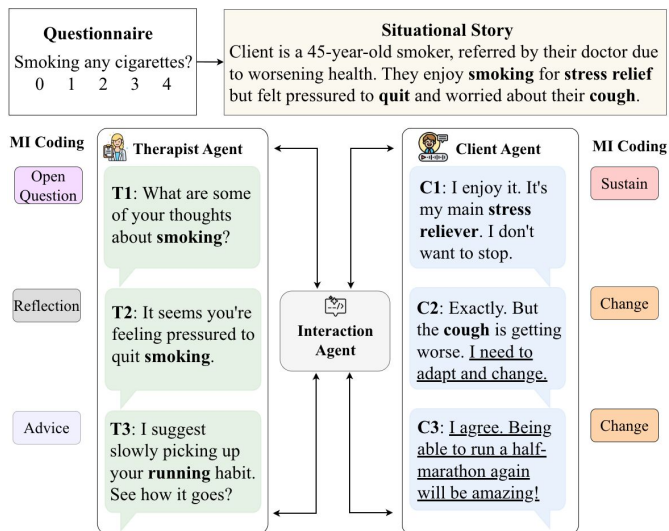
Fluent text \neq MI-faithful text. Each turn must be the right kind of move.

3. Evaluation fidelity

BLEU / ROUGE miss the therapy. Even LLM-as-judge tends to inflate.

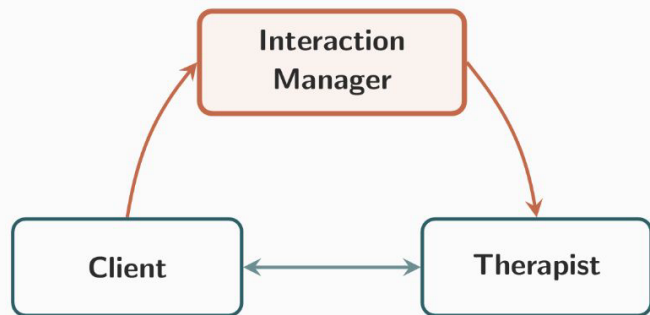
↓ **StoryMI** addresses all three.

StoryMI: Steerable Therapeutic Strategies



- Can we simulate client-therapist conversations adhering to MI strategies with multiple LLM-based agents?
- How does situational story and MI codes influence the mental health conversation quality?
- How well does LLM-as-a-Judge align with human annotators with respect to metric and model evaluation levels?

StoryMI: Three Agents Interaction Loop



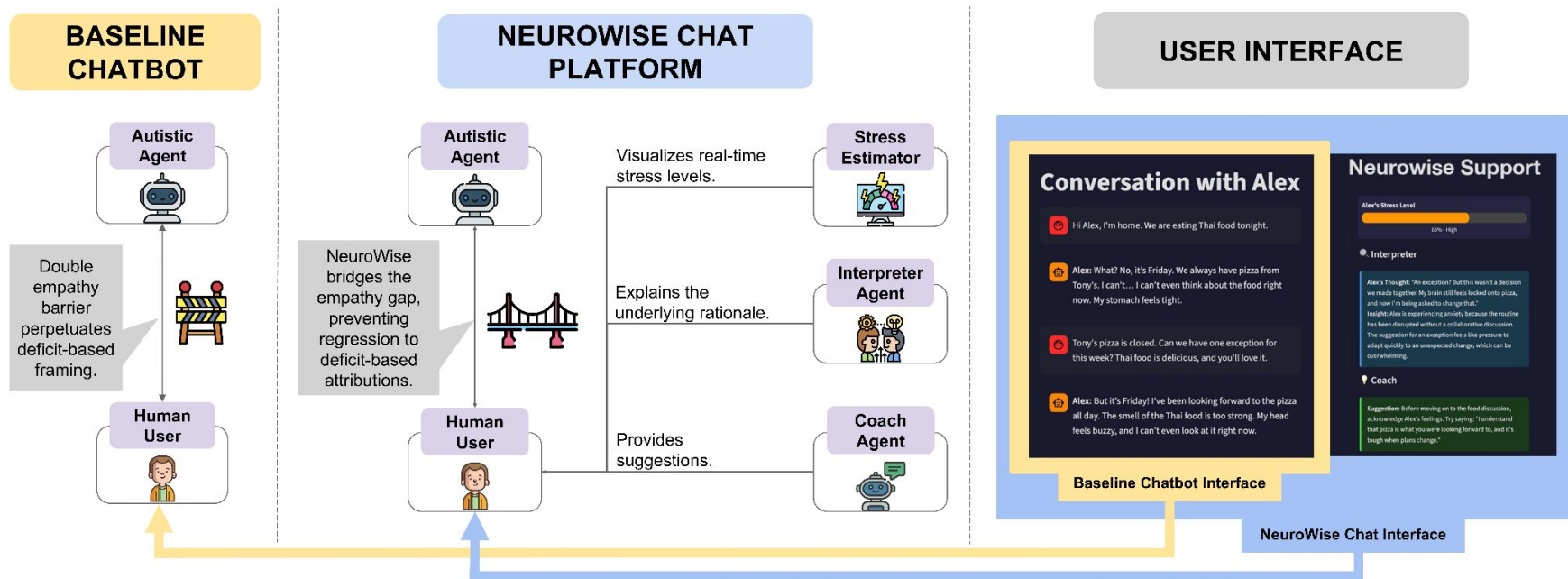
Two role-players; one always-on coordinator.

Each turn, the **Interaction Manager** adds one agent call that:

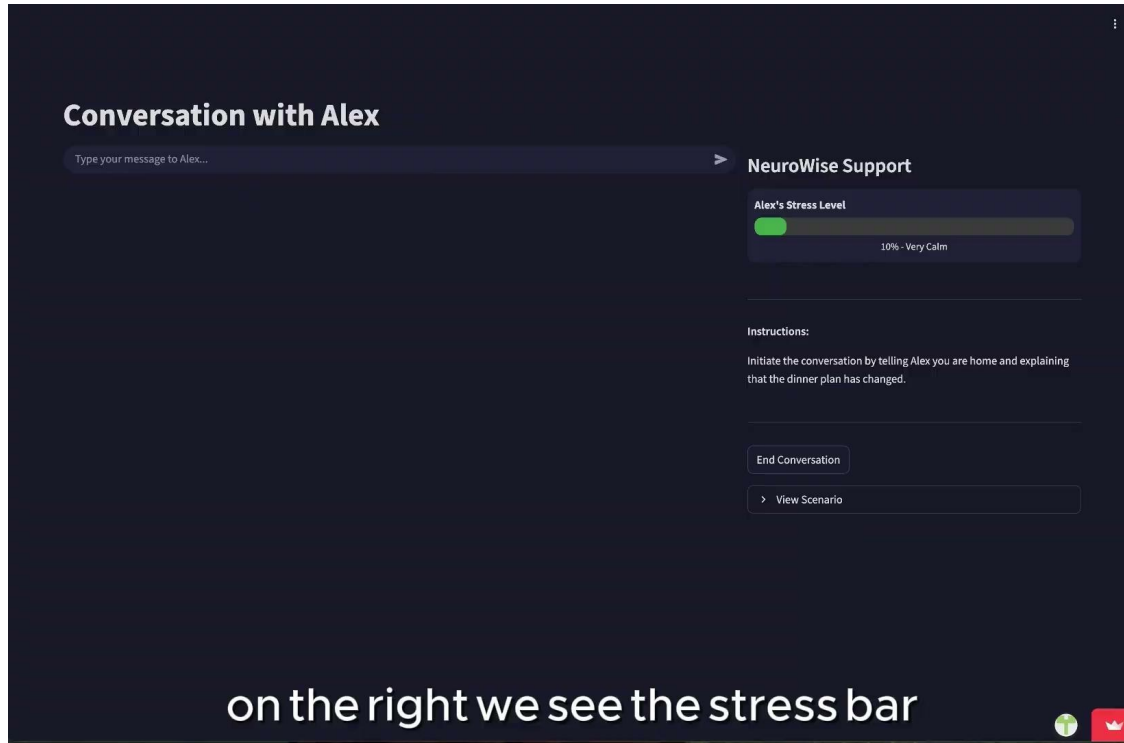
1. **Identifies** the client's MI code from the latest utterance (change / sustain / neutral talk).
2. **Selects** the therapist's next MI move ℓ_T^* from the recent code window $\mathcal{C}[-k :]$ — the macro-level control.
3. **Updates** dialogue state and **checks termination** after $T_{\min}=10$ turns.

Heuristic: change-talk → reflection; sustain-talk → open question.

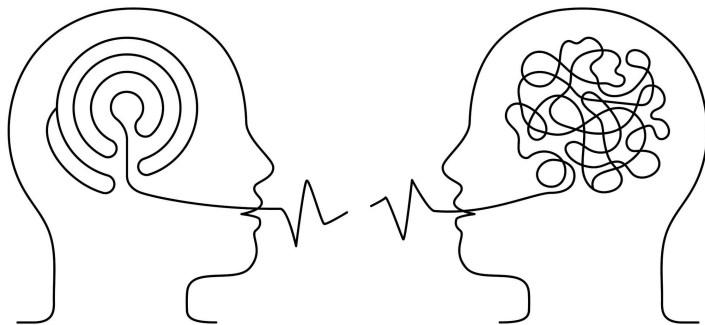
NeuroWise: Double-Empathy Communication /w Autistic Partners



NeuroWise: Double-Empathy Communication /w Autistic Partners



Psycholinguistic Bonding in Conversational Interactions



Why does language shape bonding?

Conversational AI is everywhere — yet why some agents feel close and others do not is poorly understood.

01

Ubiquitous AI partners

Therapy bots, tutors, companions, customer agents — daily conversations now span humans and AI.

02

Engagement is fragile

Trust and sustained use depend on how the agent talks, not just what it says.

03

Language carries the signal

Word choice, style, rhythm reveal the cognitive and social processes that build (or break) bonding.

Core gap: which psycholinguistic mechanisms drive bonding in human–AI interaction, and how can we engineer them?

Research Questions

From mechanism to transfer to design — bridging psychology and NLP.

RQ1

Mechanisms

Which psycholinguistic mechanisms — synchrony, alignment, dependency — actually contribute to interpersonal bonding?

Psycholinguistics

RQ2

Transfer

Do these mechanisms operate in human–AI interaction the way they do in human–human interaction?

Psychology × NLP

RQ3

Implementation

How can these mechanisms be implemented in LLMs to design more engaging, adaptive conversational agents?

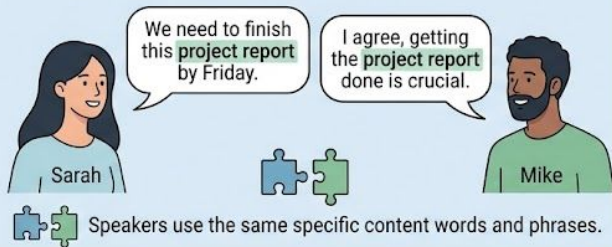
NLP · System design

We are exploring the first mechanism: linguistic synchrony.

What is Linguistic Synchrony?

FOUR PATTERNS OF LINGUISTIC SYNCHRONY

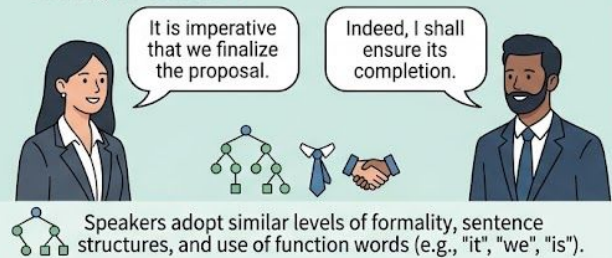
1. LEXICAL SYNCHRONY: WORD & PHRASE OVERLAP



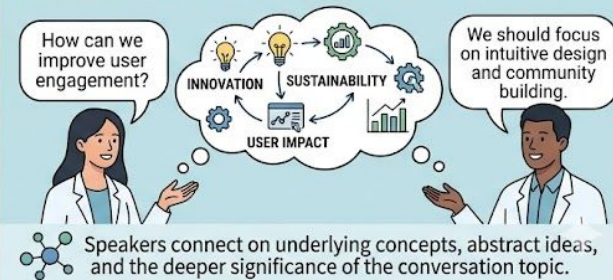
2. EMOTIONAL CONVERGENCE: MIRRORING AFFECTIVE TONE



3. STYLISTIC ALIGNMENT: MATCHING FUNCTION WORDS, SYNTAX, & FORMALITY



4. SEMANTIC ALIGNMENT: SHARED THEMES & DEEPER MEANING



Goals

Theory → Method → Causal test

GOAL 1

Theoretical

Develop the framework

- Systematic review of synchrony in human–human interaction
- Theoretical framework for synchrony in human–AI settings

GOAL 2

Computational

Measure synchrony at scale

- Quantify lexical, stylistic & dependency-based synchrony
- Compare operationalizations on conversational data

GOAL 3

Experimental

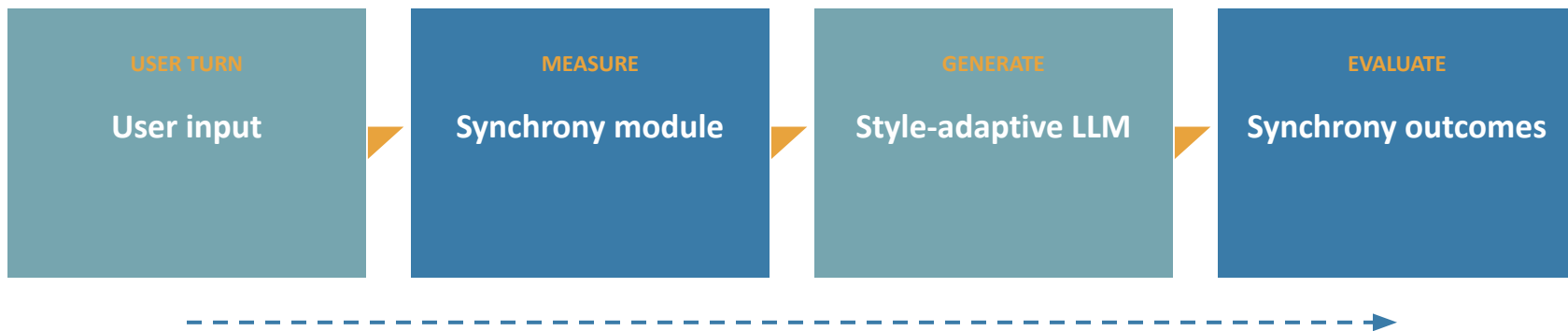
Test causal effects

- LLM agents with manipulated high vs. low synchrony
- Outcomes: bonding, perceived empathy, trust, engagement

OUR APPROACH

An agent that measures synchrony — and adapts to it

Current LLMs do not adapt to a user's linguistic style in a principled, measurable way. We build that loop.



DESIGN PRINCIPLES

Measurable *synchrony quantified turn-by-turn*

Causal *manipulable for controlled experiments*

Causal test of synchrony on bonding

Manipulate synchrony, randomize, control confounders — isolate the causal effect.

INDEPENDENT VARIABLE

Synchrony level

manipulated in an LLM-based agent

HIGH synchrony

LOW synchrony

DEPENDENT VARIABLES

● **Bonding / closeness**

IOS scale, rapport items

● **Trust**

trust-in-agent scale

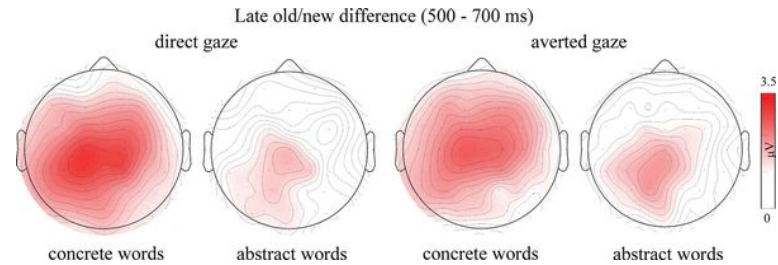
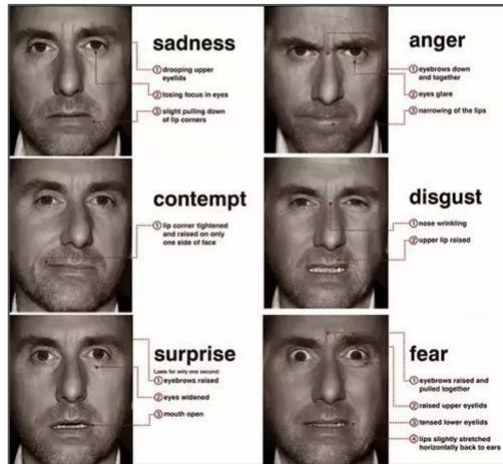
● **Perceived empathy**

validated empathy scale

● **Willingness to engage**

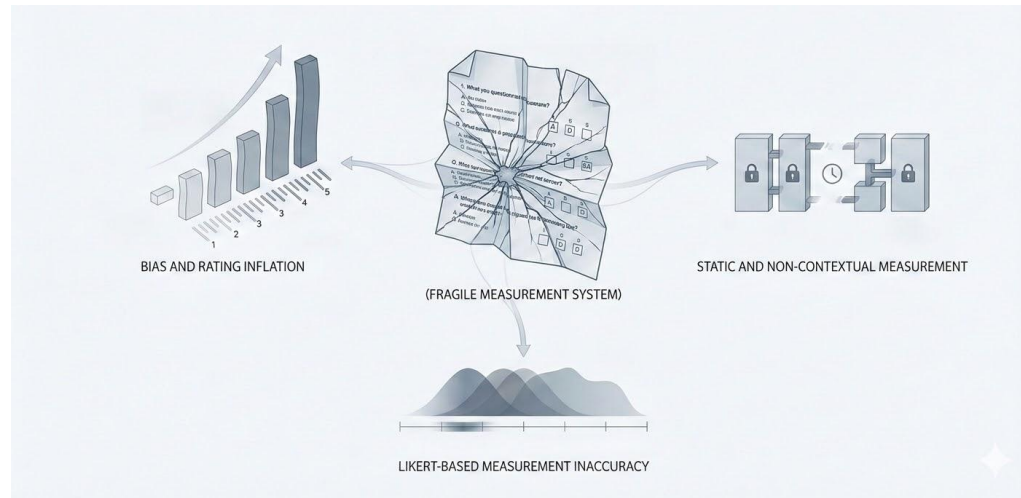
intent + behavioral measure

Gaze Perception & Understanding Beyond Conversations



Challenges in Questionnaire-base Measurement

- Bias and Rating Inflation(Anvari et al., 2022; McGrath et al., 2010)
- Cannot be taken in real-time during a task (Jyotsna et al., 2023)
- Measurement Inaccuracy (Likert Limitations)



The Role of Gaze

- Multiple non-verbal signals provide **objective context** beyond self-report (Bhatara et al., 2014; de Gelder et al., 2015).
- Gaze behavior is difficult to consciously control and **emerges early** in mental disorders (Jagla, 2016; Navalón et al., 2021).
- Gaze features achieve high predictive **accuracy and scalability**, outperforming EEG- and heart-rate–based models in mental state recognition (Jyotsna et al., 2023; Bozkir et al., 2023; Chlasta et al., 2024).

Motivation

The main goal is to infer potential mental-health risks, **enhance self-awareness**, and ultimately improve well-being.

- Higher self-awareness is linked to better well-being (Richards et al., 2010).
- Emotional awareness is a key component of mental health (Weissman et al., 2020).

Gaze signals → behavioral indicators of mental health risk

A high-level view of the pipeline. The modeling layer is treated as a black box for this overview.



DESIGN PRINCIPLES

Behavior-grounded

Built on patterns established in clinical eye-tracking research.

Task-flexible

Designed to work across different stimuli and screening contexts.

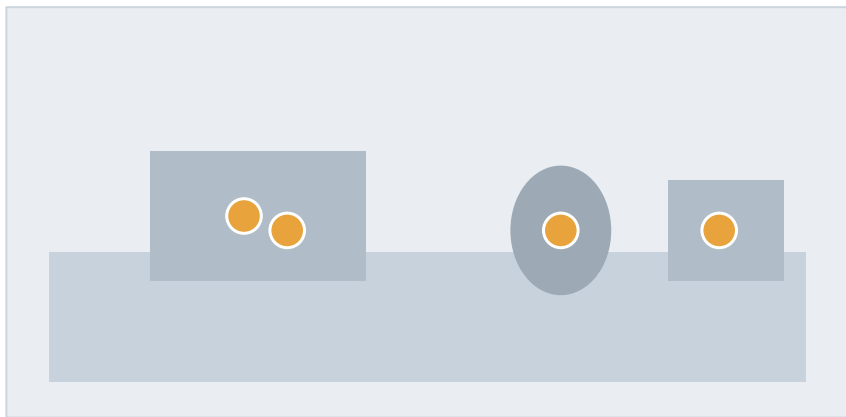
Interpretable output

Produces structured indicators clinicians can examine — not opaque scores.

Vision-language models bridge the gap

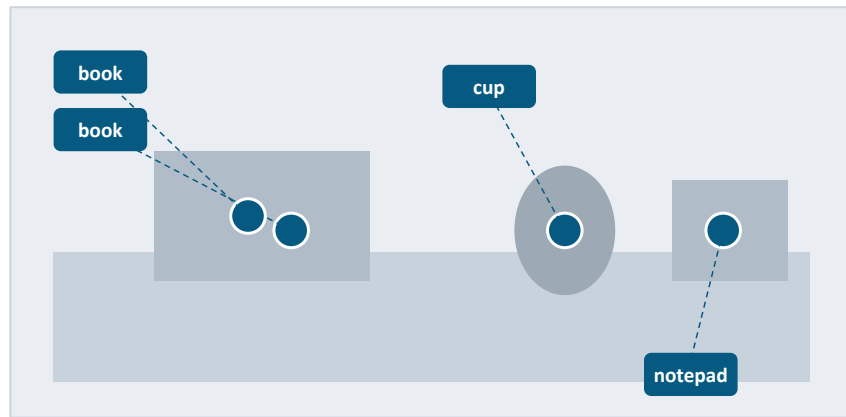
We develop **vision-language models** (VLMs) to `look` at the same image a participant saw, and identify what was the semantic meanings at each **Gaze Signals** for decision-making.

BEFORE



Gaze Signals *on a scene* — *coordinates only*

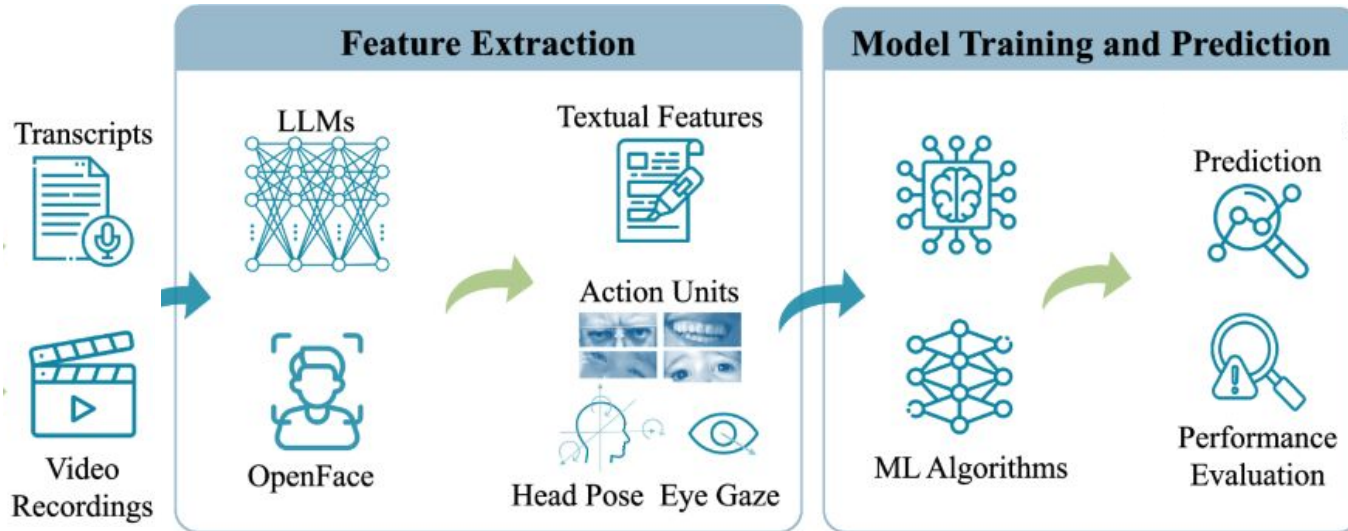
AFTER



Each signal now carries an object label

Vision-language models bridge the gap

We develop **vision-language models** (VLMs) to `look` at the same image a participant saw, and identify what was the semantic meanings at each **Gaze Signals** for decision-making.



Challenges & Opportunities

The pipeline is technical infrastructure. To turn it into something clinically meaningful, we need expertise.

01 Clinical grounding

Help us define which mental-health indicators are most worth pursuing first, and which are scientifically defensible to claim.

02 Study design

Co-design participant recruitment, screening protocols, and stimuli selection so our results meet clinical-research standards.

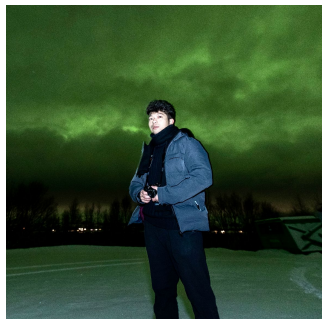
03 Data access

Access to existing eye-tracking datasets with clinical annotations, or collaboration on new data collection in clinical settings.

04 Joint publication

Co-authorship on methodology and validation papers — bridging the technical and clinical communities is the point.

Team



Qingyu Meng

PhD student,

Trustworthy Multi-Agents and Safety Alignment



Yifan Mo

PhD student, *Psycholinguistic Bonding in Human-AI Interaction for Conversational Agents*



Yue Su

PhD student,

Inferring Mental Health Risks from Gaze-Based Signals



Mithat Can Ozgun

MSc student & Software Engineer@Booking.com



Albert Tang

High school student,

Human-Centered AI for Neurodiverse Communication



Koen Hindriks

Full Prof. of AI,
*Socially AI, Social Robots,
Human-Robot Interaction*



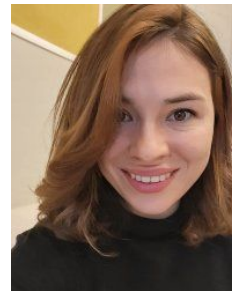
Sander Koole

Full Prof. of Psychology,
*Emotion regulation,
Personality, AI*



Artem Belopolsky

Assistant Prof. of Neurocontrol,
Cognitive Science, Human-AI Interaction, Gaze estimation



Aneta Lisowska

Assistant Prof. of AI and Health,
*Machine Learning,
Health Informatics*



Lucia Donatelli

Assistant Prof. of Computational Linguistics,
*Multimodal natural language understanding &
human-robot dialogue*

Reference

- Ozgun, M. C., Pei, J., Hindriks, K., Donatelli, L., Liu, Q., & Wang, J. (2025). Trustworthy AI Psychotherapy: Multi-agent LLM Workflow for Counseling and Explainable Mental Disorder Diagnosis. In Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM'25).
- Qingyu Meng, Min Chen, Dingming Liu, Yifan Mo, Yue Su, Xin Sun, Koen Hindriks, Jiahuan Pei, StoryMI: Steerable Multi-Agent Therapeutic Dialogue Generation. In: Findings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL'26)
- Tang, A., Mo, Y., Li, J., Su, Y., Zhang, M., Koole, S. L., ... & Pei, J. (2026, April). NeuroWise: A Multi-Agent LLM "Glass-Box" System for Practicing Double-Empathy Communication with Autistic Partners. In Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI'26).
- Sun, X., Su, Y., Mo, Y., Meng, Q., Li, Y., Sugawara, S., ... & Pei, J. (2026). Aligning Human-AI-Interaction Trust for Mental Health Support: Survey and Position for Multi-Stakeholders. arXiv preprint arXiv:2604.20166.

Thank you for your attention!
Q & A

Jiahuan Pei
j.pei2@vu.nl
Assistant Professor
Vrije Universiteit Amsterdam