FACULTY OF ENGINEERING
AND ARCHITECTURE

DEPARTMENT OF ELECTRONICS AND INFORMATION SYSTEMS

IDLAB – ARTIFICIAL INTELLIGENCE AND ROBOTICS

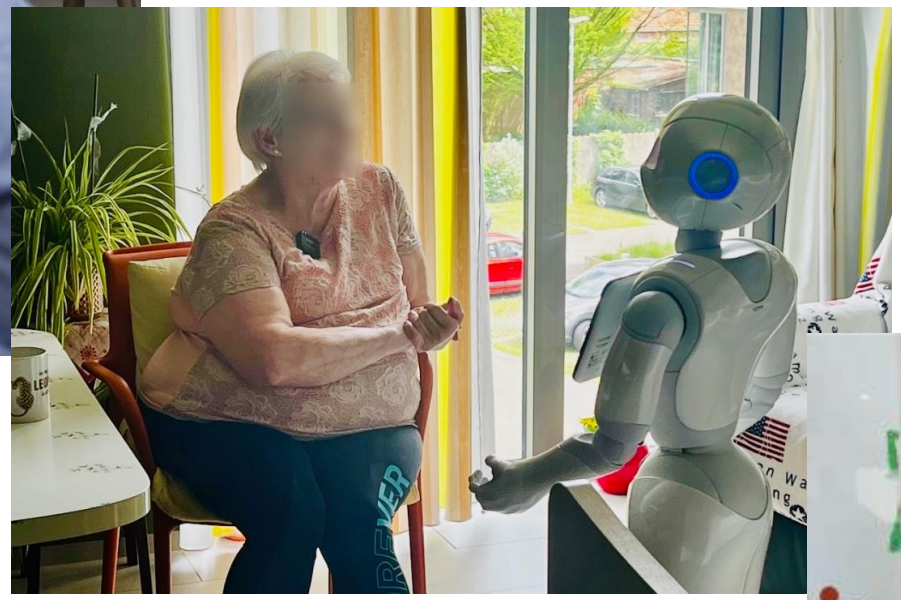
Multimodal language models for social interaction

Ruben Janssens, Highway Session Beyond Text: Multimodal LLM Integration, November 25, 2025

GHENT UNIVERSITY

imec

Robots need to be social … and autonomous



GHENT UNIVERSITY

imec

Automatic Speech Recognition (ASR)

Dialogue Manager

Text to Speech (TTS)

GHENT UNIVERSITY

imec

IDLAB AIRO

Real-time!
(< 1s)

Automatic Speech Recognition (ASR)

Dialogue Manager

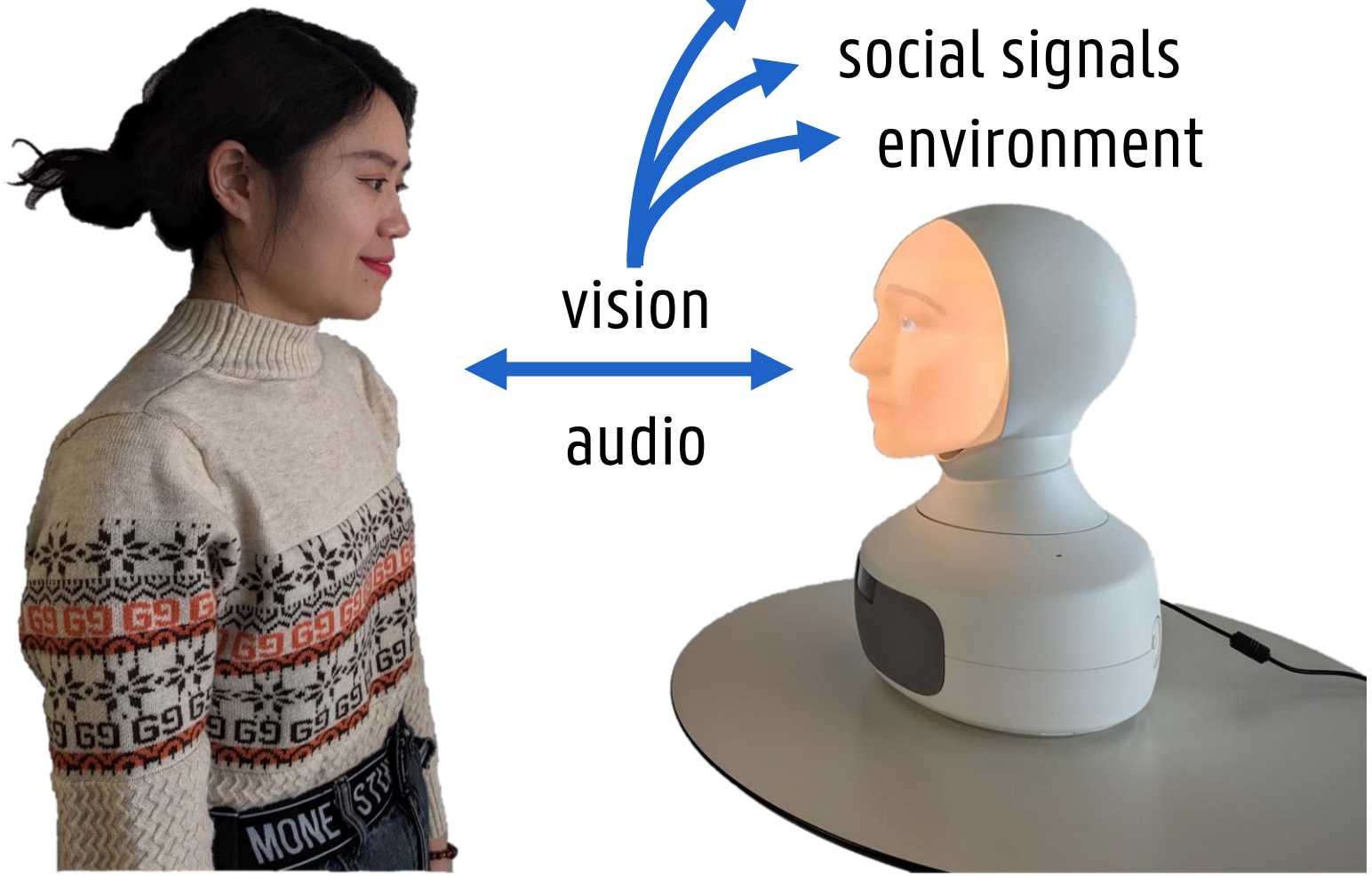
Text to Speech (TTS)

GHENT UNIVERSITY

imec

IDLAB AIRO

Robots need to be social … and multimodal



GHENT UNIVERSITY

imec

IDLAB AIRO

Computer Vision

Automatic Speech Recognition (ASR)

Dialogue Manager

Text to Speech (TTS)

GHENT UNIVERSITY

imec

IDLAB AIRO

GPT-4o

Solved?

GHENT UNIVERSITY

imec

IDLAB AIRO

GPT-4o

Solved?

No.

GHENT UNIVERSITY

imec

IDLAB AIRO

(Vision-)language models (VLMs) are not made for

embodied (first-person vision)

social (not task-oriented)

dialogue

➞ **How can we adapt them?**

GHENT UNIVERSITY

imec

IDLAB AIRO

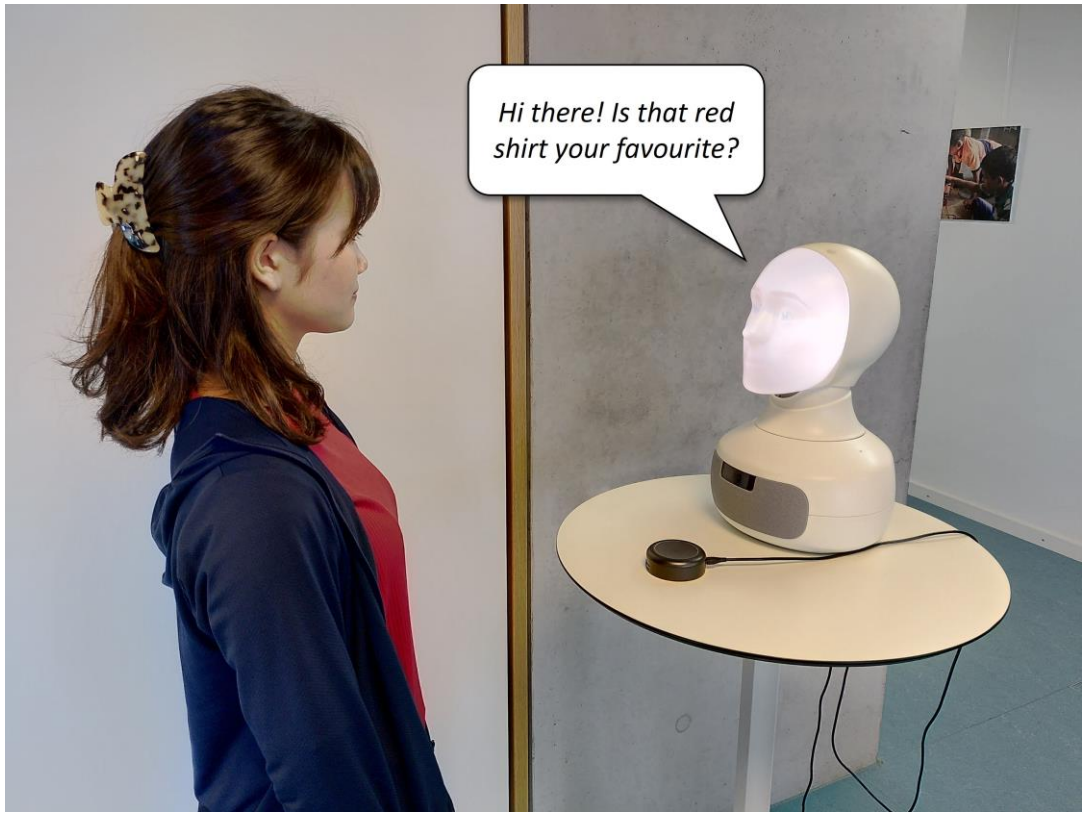
Starting small

Starting a **social** (chit-chat) conversation

with a **visual opener**

using **robot camera input**

→ **How can we integrate visual information in a language model?**



Janssens et al., 2022, IEEE/ACM HRI 2022
Janssens et al., 2024, IEEE Transactions on Affective Computing

Back to 2021

captioning model

"a man with glasses and a striped shirt."

training set

crowd-sourced questions

flickr

3000 images

3 questions per image

amazon mechanical turk

Fine-tune sequence-to-sequence language model: **BART**

question generating model

caption to question model

fine-tuned on dataset

50 test images

flickr

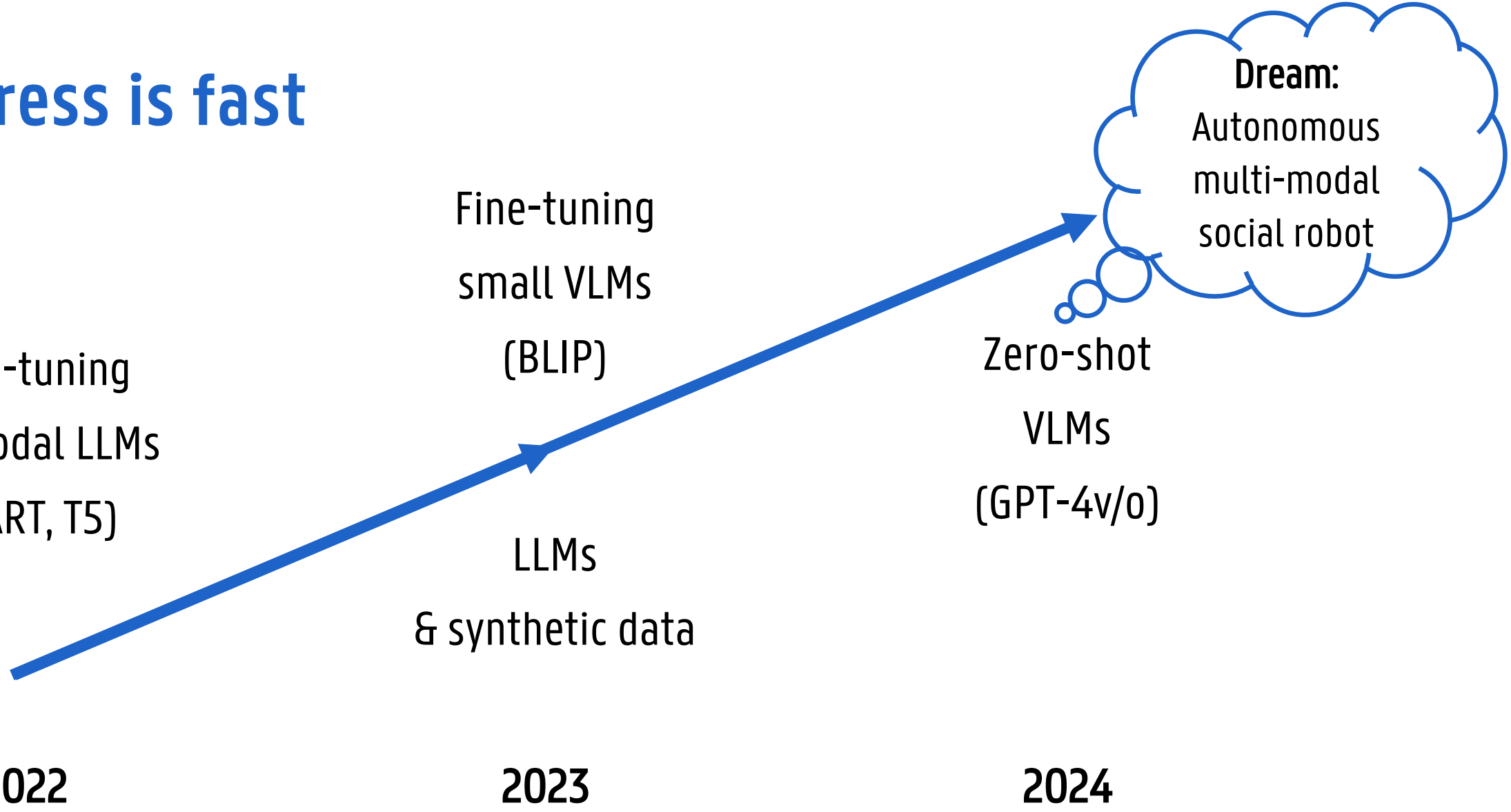
Festival

"Hi there! I couldn't help but notice your stylish striped shirt. Where did you get it from?"

GHENT UNIVERSITY

imec

Progress is fast



GHENT UNIVERSITY

imec

Janssens et al., 2022, IEEE/ACM HRI 2022
Janssens et al., 2024, IEEE Transactions on Affective Computing

IDLAB AIRO

What's still missing?

Large Language Models still struggle:

referring to the image or user in third person

extremely generic or extremely specific questions

→ LLMs don't understand embodied and social dialogue

Could we let them learn from their mistakes in interactions?

GHENT UNIVERSITY

imec

IDLAB AIRO

Can robots detect miscommunications?

GHENT UNIVERSITY
imec
language and translation technology team
+
IDLAB AIRO

Janssens et al., 2025, IEEE RO-MAN 2025

Robots make mistakes… can we detect them?

🔎 How do users express **miscommunication** in **social** human-robot dialogue?

Prior work: task-oriented dialogue e.g., assembly

🔎 Can this be detected automatically from **facial expressions**?

Prior work: facial expression recognition using over-acted emotions

(Trung et al., 2017; Kappas et al., 2020; Kontogiorgos et al., 2020 & 2021; Axelsson et al., 2022 & 2023; Stiber et al., 2023)

GHENT UNIVERSITY imec IDLAB AIRO

Experiment

Cooking dialogues

Robot explains recipe
Asks questions
Fully scripted

4 mistake types

Irrelevant
Too much info
Too little info
Interruption

GHENT UNIVERSITY

imec

IDLAB AIRO

Experiment

What do you think the next step will be?

Adding the butter?

Did you know whales are actually mammals, not fish?

??????

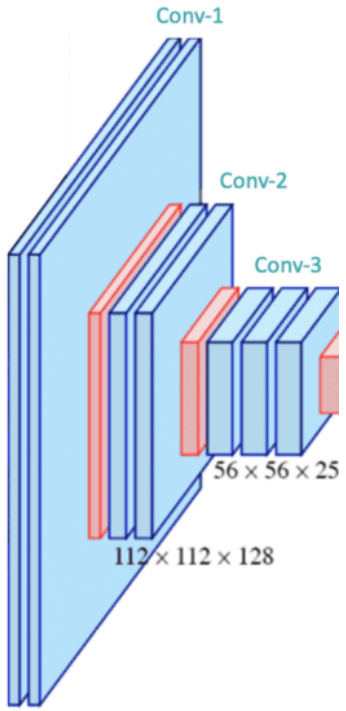
Detection pipeline

Keypoints

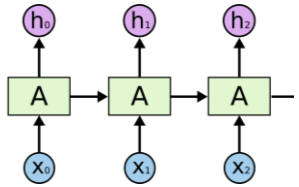
Keypoints / Blendshapes

VGG16

ResNet50



LSTM



✓ / ✗

Conversation fragment

Salient moments (3 * 1s)

Visual feature extractor

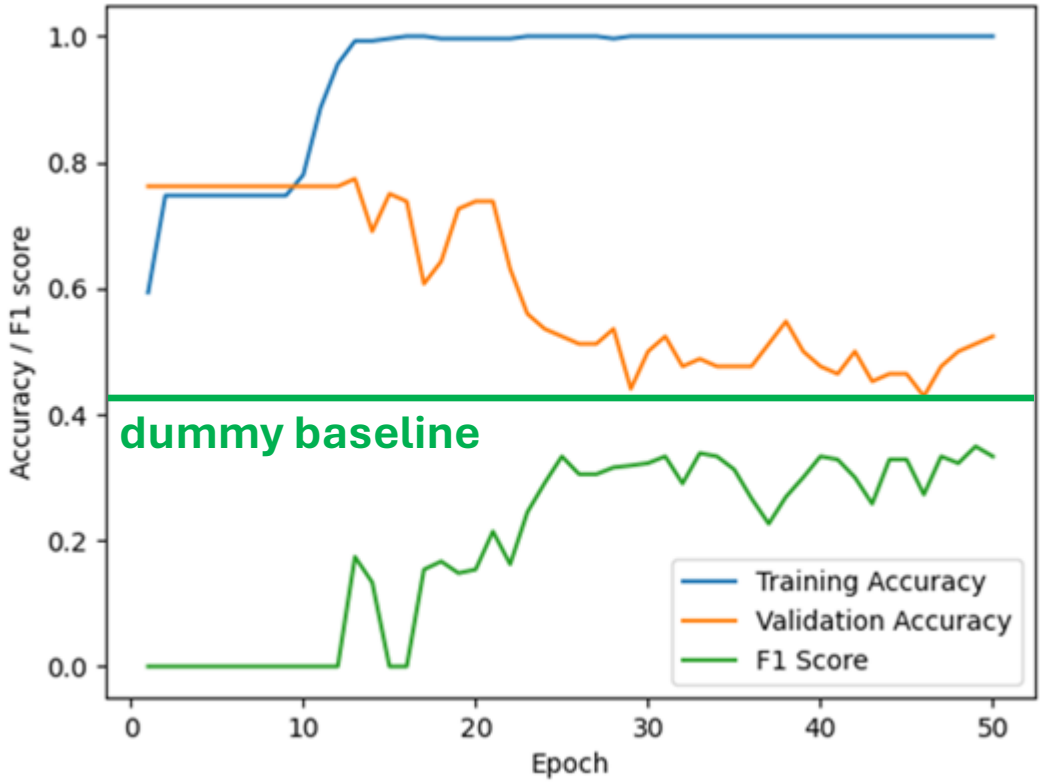
Classifier

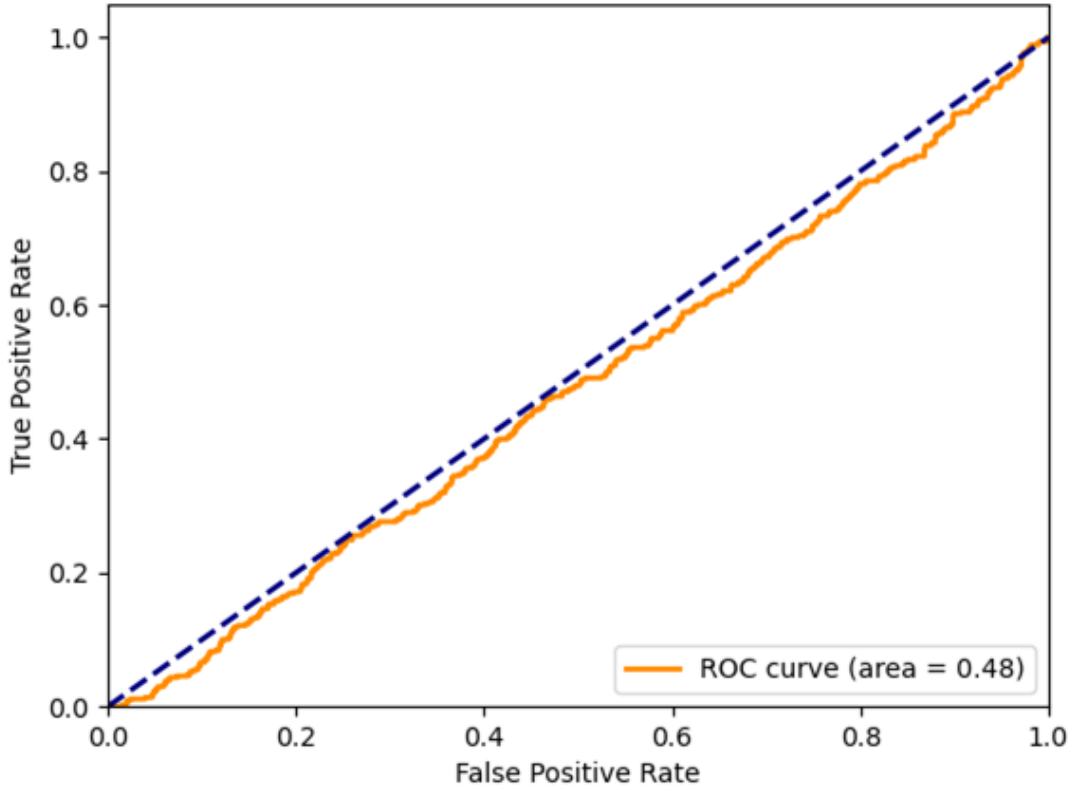
GHENT UNIVERSITY

imec

IDLAB AIRO

Classifier does not outperform random chance





GHENT UNIVERSITY

imec

IDLAB AIRO

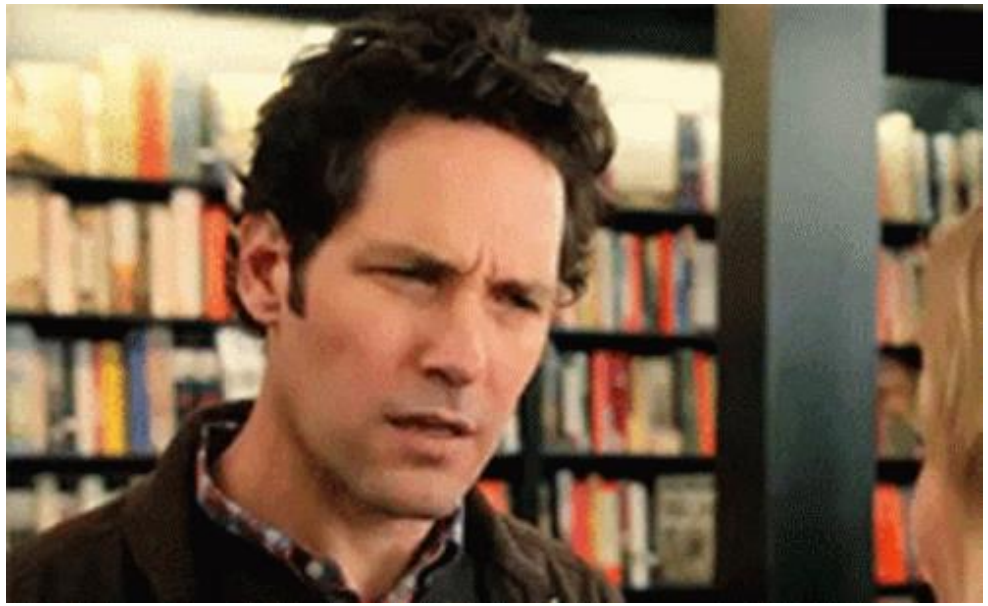
Can our model detect facial expressions?

New dataset:

expressive, over-acted displays of confusion

139 short videos (≤4s)

70% confusion, 30% neutral



Same detection model architecture

GHENT UNIVERSITY

imec

IDLAB AIRO

Can our model detect facial expressions?

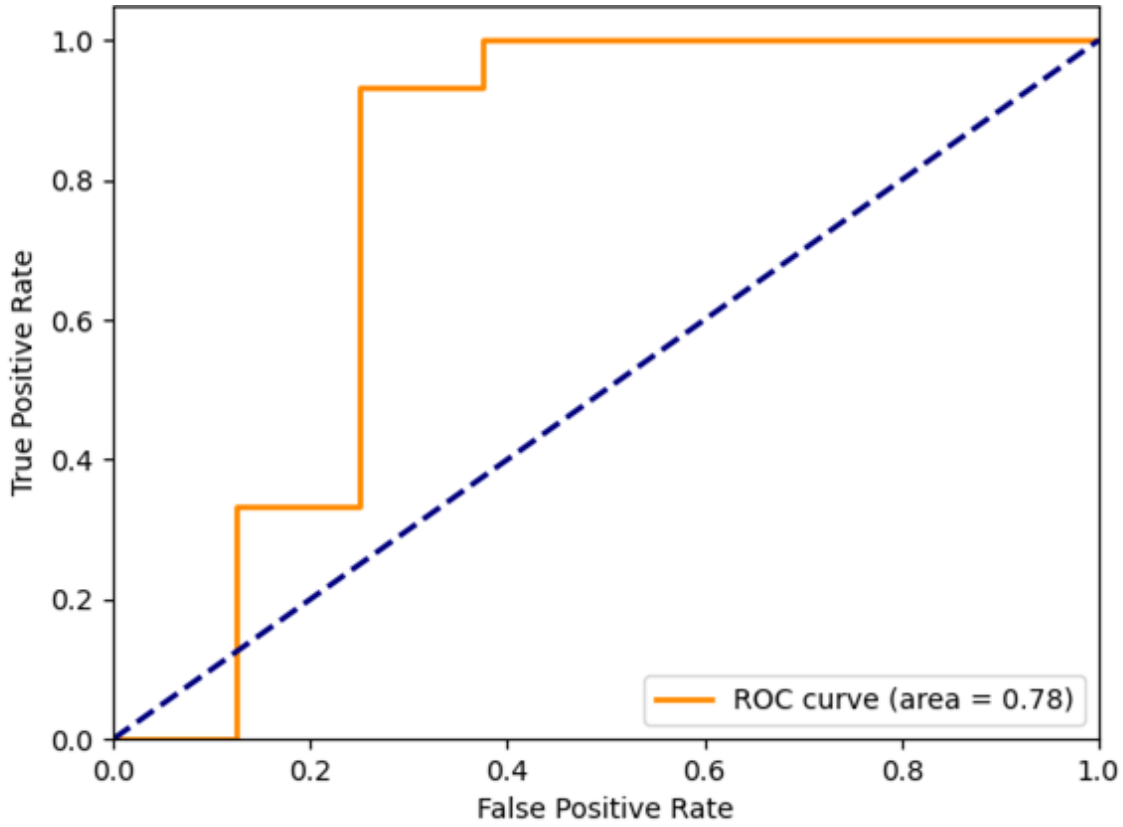
New dataset:

expressive, over-acted displays of emotion

139 short videos (≤4s)

70% confusion, 30% neutral

Same detection model architecture



GHENT UNIVERSITY

imec

IDLAB AIRO

Is it the task?

Can people detect the miscommunications?

Human evaluation:

20 video fragments (6 miscommunications) – no audio

17 participants

“Did you think the robot made a mistake?”

People do not agree.

Fleiss’ kappa = -0.012

→ no agreement

	Average accuracy per video
Miscommunication	50.00%
No miscommunication	59.24%

GHENT UNIVERSITY

imec

IDLAB AIRO

Miscommunication detection is very difficult.

It's not the robot's fault.

People often hide their emotions

even when they think the robot made a mistake.

➞ What leads users to express feedback to robots?

➞ How do different modalities connect?

➞ How can robots elicit feedback?

GHENT UNIVERSITY

imec

Why Robots are Bad at Detecting Their Mistakes:

Limitations of Miscommunication Detection in Human-Robot Dialogue

Ruben Janssens, Jens De Bock, Sofie Labat,
Eva Verhelst, Veronique Hoste, Tony Belpaeme

www.rubenjanssens.be/miscommunication-detection

IDLAB AIRO

Can robots predict the user's enjoyment?

GHENT UNIVERSITY imec

Janssens et al., 2025, IEEE/ACM HRI 2025

KTH VETENSKAP OCH KONST + IDLAB AIRO

Can robots avoid mistakes?

Large Language Models (LLMs) can drive real open-domain dialogue

Still, problems remain:

superficiality, conversation stoppers, going in circles

What if robots could predict user enjoyment?

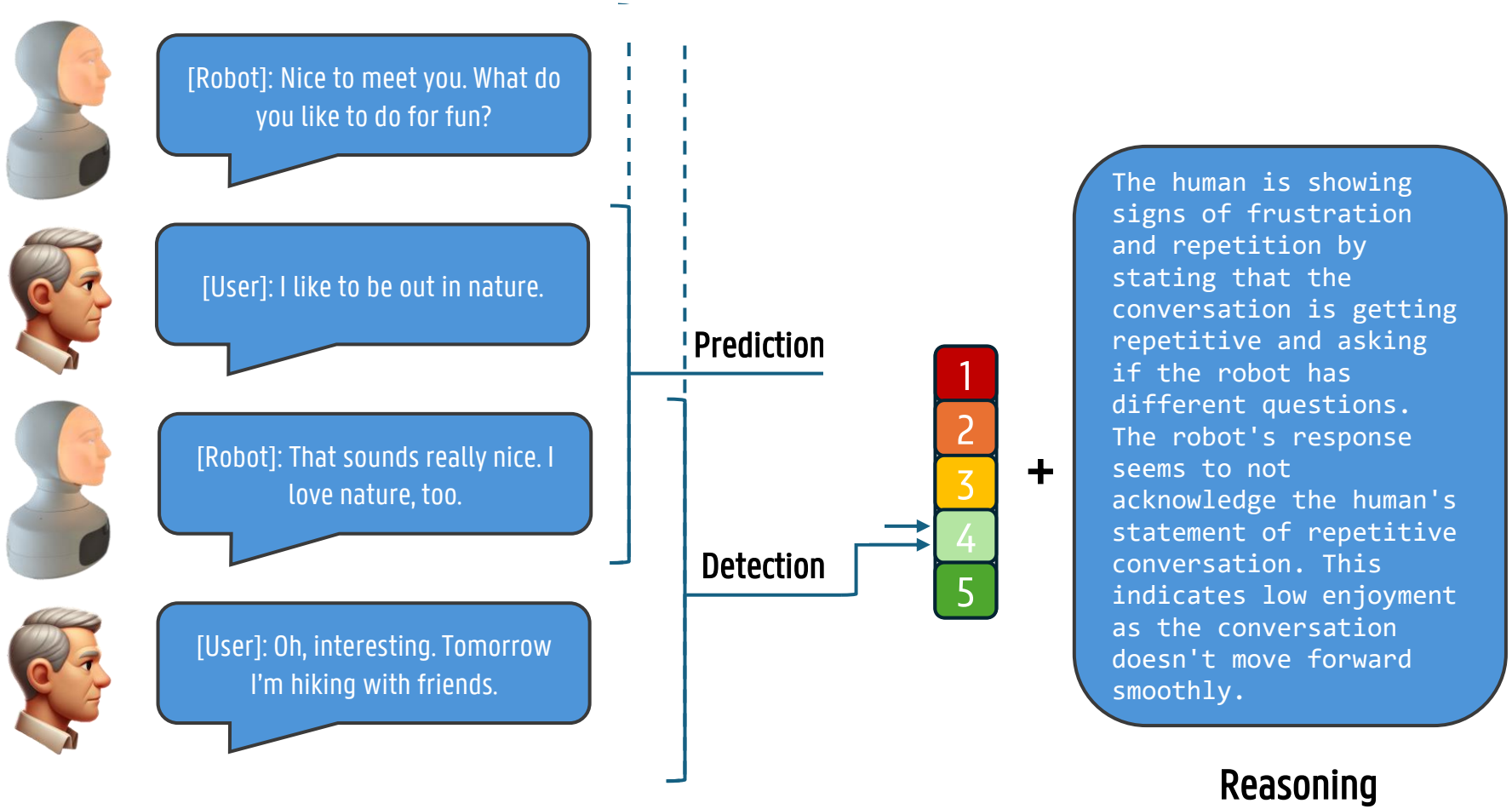
→ choose best utterance option

→ fine-tune LLM to improve enjoyability

→ adapt dialogue to improve enjoyment

Predicting user enjoyment

Large Language Models can reason about social context



Conclusions

LLMs can predict user enjoyment **without user response!**

- Performance == with user response
- Correlates **better with user perception** than human annotators

Can be transferred as-is to new context

Enables **adaptation of LLMs** to human-robot dialogue:

- online and offline

GHENT UNIVERSITY

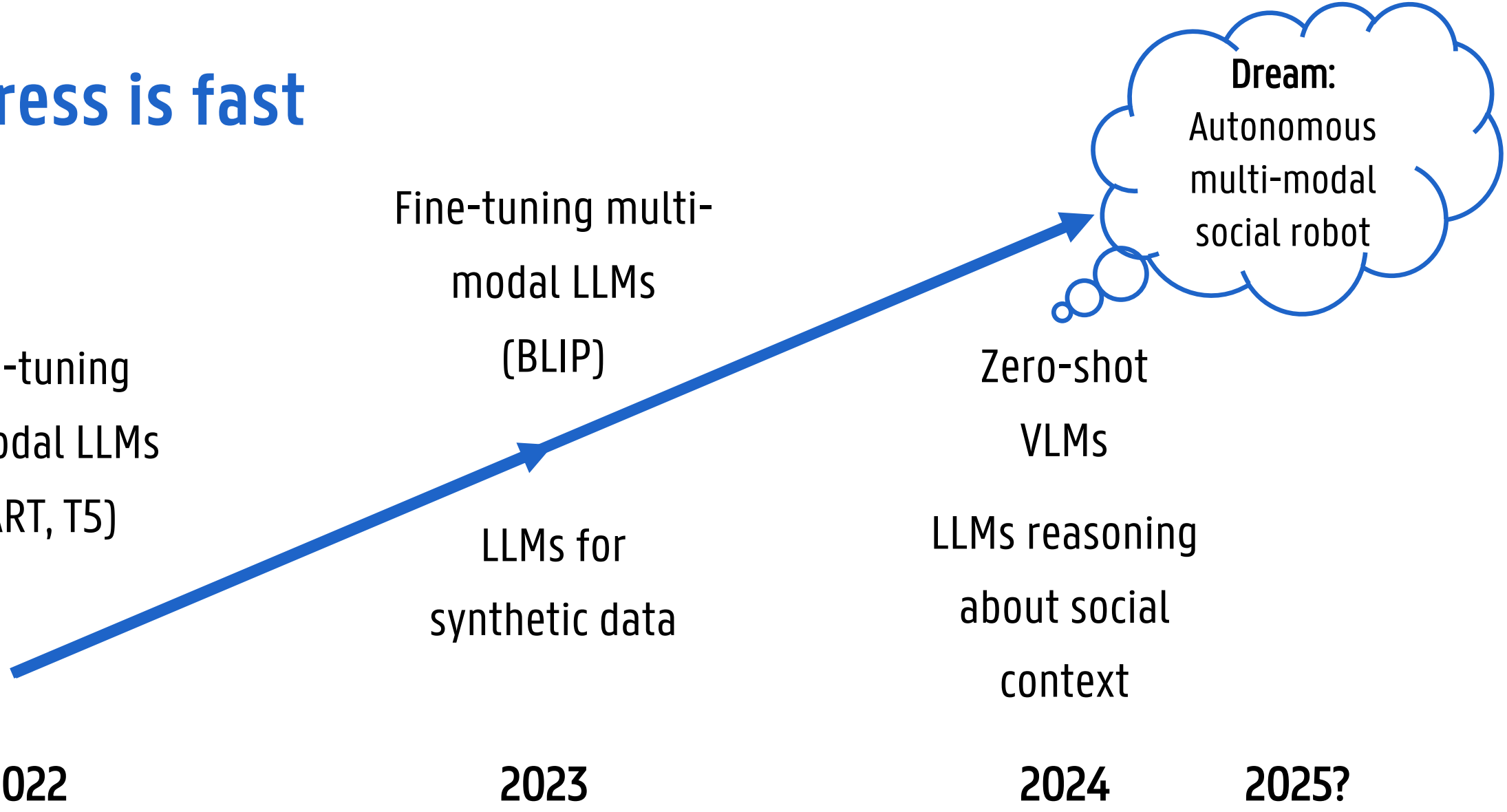
imec

Online Prediction of User Enjoyment in Human-Robot Dialogue with LLMs

Ruben Janssens, André Pereira, Gabriel Skantze, Bahar Irfan, Tony Belpaeme

www.rubenjanssens.be/enjoyment-prediction

Progress is fast



GHENT UNIVERSITY

imec

Janssens et al., 2022, IEEE/ACM HRI 2022
Janssens et al., 2024, IEEE Transactions on Affective Computing

IDLAB AIRO

Can autonomous robots have social conversations with continuous visual context?

GHENT UNIVERSITY imec

Under review

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS + KTH VETENSKAP OCH KONST + IDLAB AIRO

1

Can we build a system that drives a long social conversation with visual context?

2

Does using visual context **improve the conversation**?

GHENT UNIVERSITY

imec

IDLAB AIRO

What is a good conversation?

How to evaluate whether visual context improves conversations?

LLM dialogue: superficial, one-sided, repetitive

Prior research: **completely open-ended conversations with LLMs**

→ not comparable between conditions

Our approach: **semi-structured conversations**

= free conversations (no Q&A)

about prepared questions

with mixed initiative

Robot: If you could invite any person in the world to dinner, who would you choose?

Participant: What do you value most in a friendship?

Mixed-methods evaluation:
User preference + questionnaires + interview
External annotations

GHENT UNIVERSITY

imec

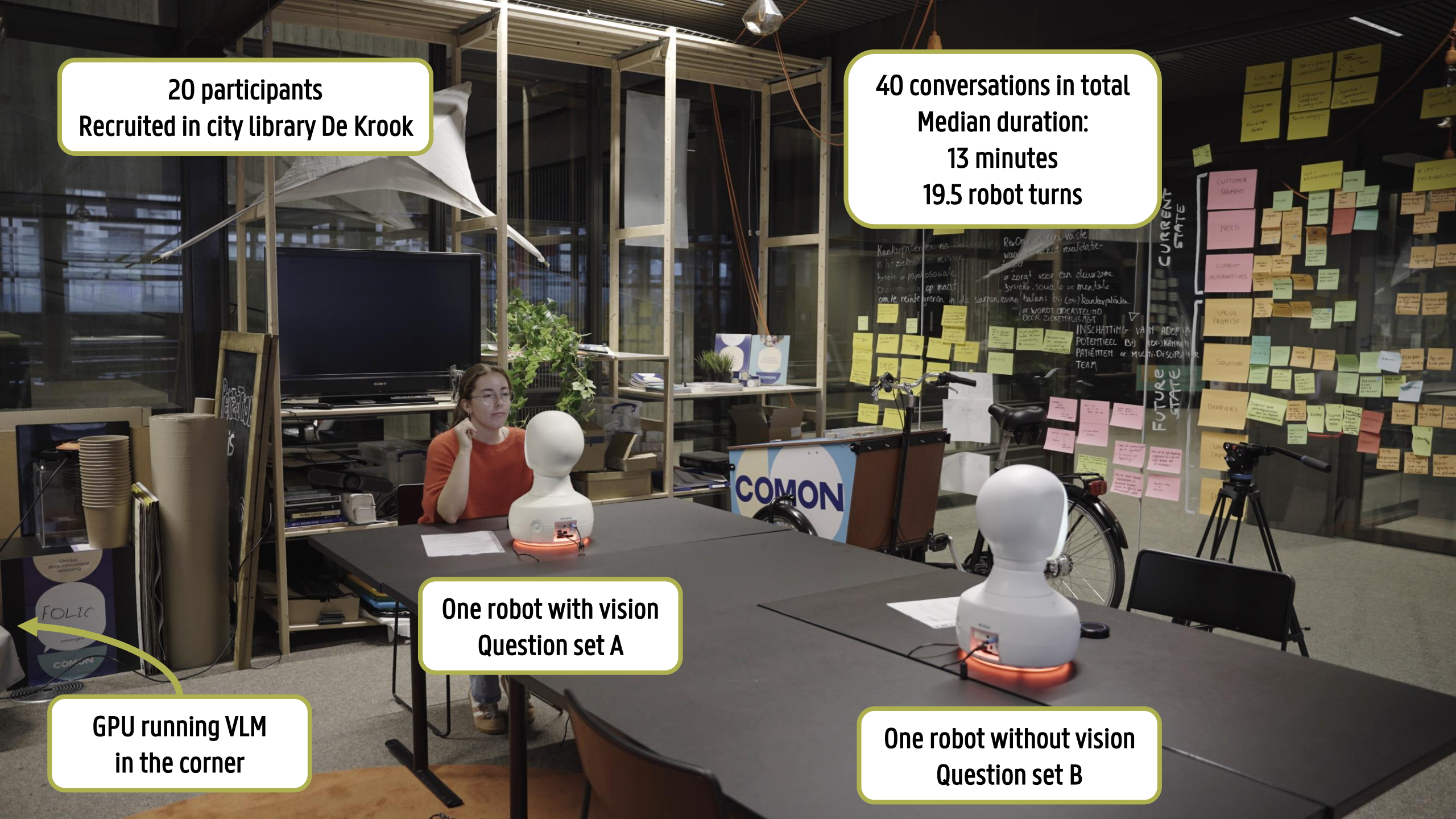
IDLAB AIRO

Multi-step, modular system

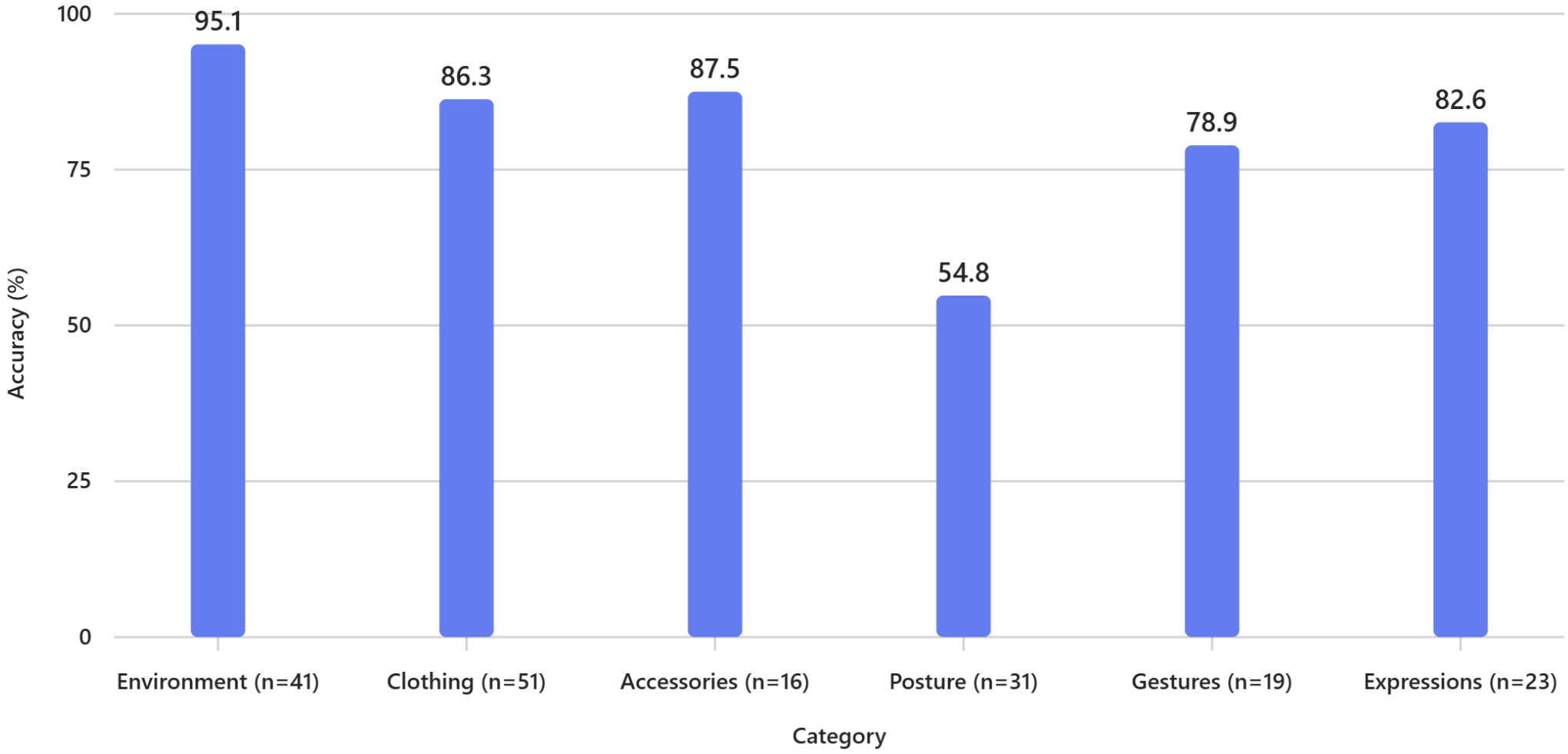
GHENT UNIVERSITY

imec

IDLAB AIRO


20 participants
Recruited in city library De Krook
40 conversations in total
Median duration:
13 minutes
19.5 robot turns
One robot with vision
Question set A
GPU running VLM
in the corner
One robot without vision
Question set B
COMON
FOLIC
CURRENT STATE
FUTURE STATE

Our local Vision-Language Model is accurate!

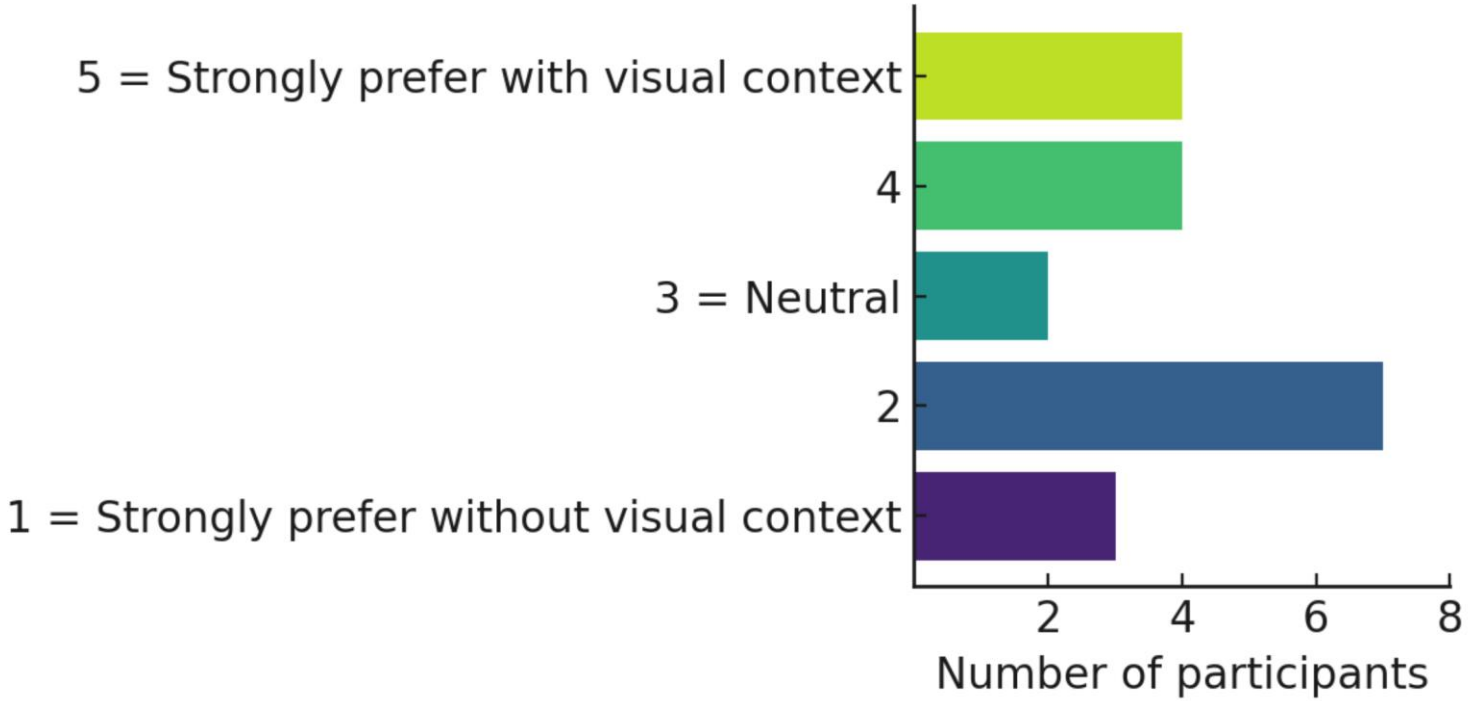


GHENT UNIVERSITY

imec

IDLAB AIRO

Yet participants don't have a clear preference…



"more personalized, interesting conversation"

"the robot was in the room with me"

"creepy"

"I felt watched"

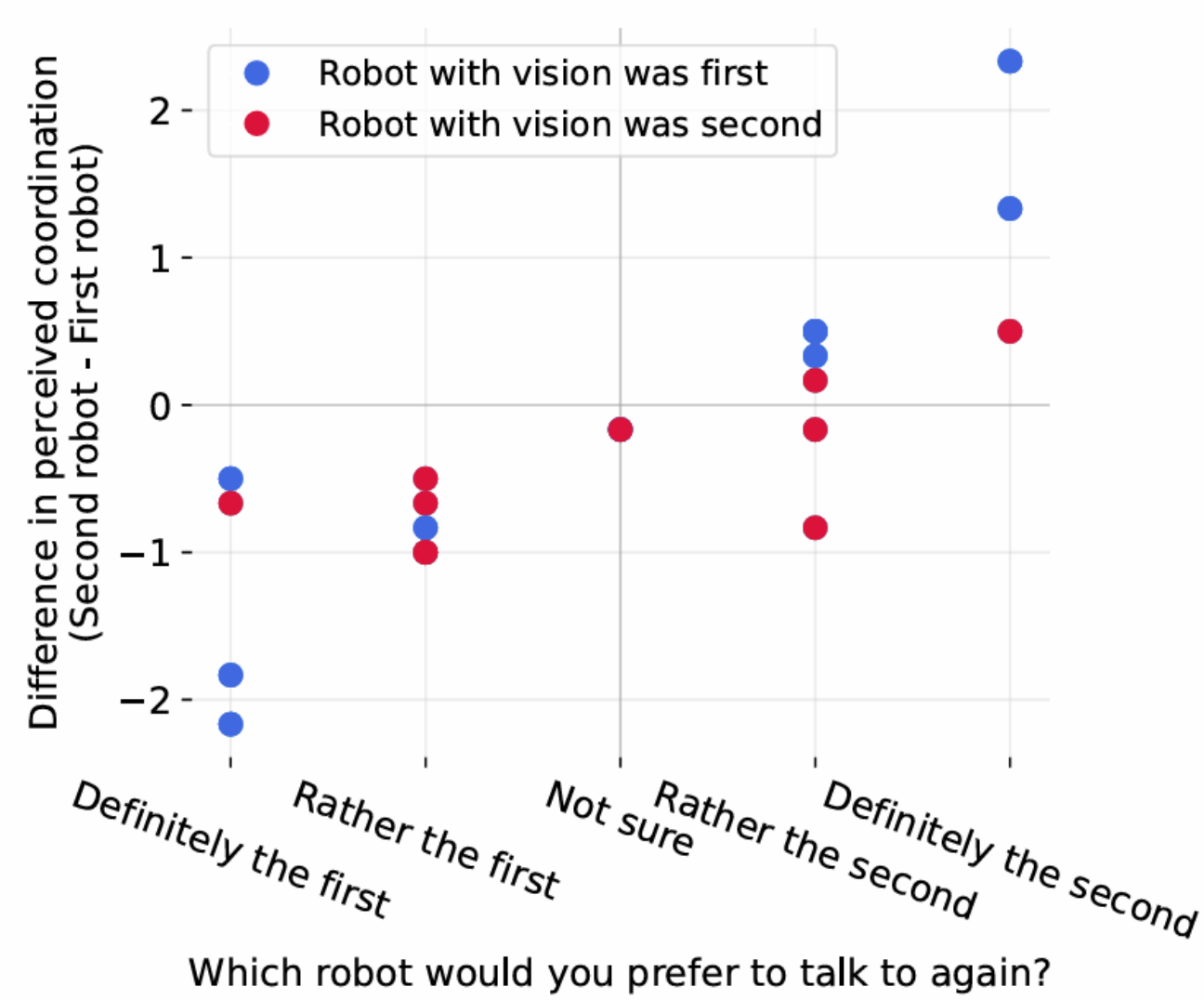
"not fitting in the conversation"

GHENT UNIVERSITY

imec

IDLAB AIRO

Participants preferred well-coordinated conversations



Strongly correlated with preference ($\rho = 0.82$)

GHENT UNIVERSITY

imec

IDLAB AIRO

What caused lack of coordination?

Recurring "conversational mistakes":

Correlation with experienced coordination

Speech recognition failures: 8% of turns

Misaligned with user on dialogue plan: 10% of turns → ρ = -0.53

Unfocused responses: 4% of turns → ρ = -0.38

Robot repeating previous comment: 10% of turns

Foundation Models lack social interaction skills

Simple “get-to-know-you” conversation is challenging

Combining multiple goals

(e.g., getting to know user, using visual context, following dialogue plan)

Planning multiple turns ahead

Switching initiative between turns

Knowing when and how to use visual context appropriately

Current Foundation Models are not trained for social, embodied dialogue!

Social interaction is not solved

Foundation Models' language and vision accuracy revolutionized robot dialogue

… yet real-world deployment quickly reaches limits

as models **lack fundamental social interaction skills**

Modelling embodied, social dialogue requires radically different training, system design, and evaluation

to benefit real users in the real world

Modular & parallel architecture

Post-training on social dialogue data

Adapting & learning from feedback in interaction

GHENT UNIVERSITY

imec

IDLAB AIRO

Lessons learned

1. Humans are not a toy problem

2. Social interaction is fast
   continuous
   and deeply multimodal

3. Evaluating in interaction is essential

GHENT UNIVERSITY

imec

IDLAB AIRO

The future

1. Modelling embodied social interaction is not a solved problem
2. The future of robotics is social and interdisciplinary
3. Engineering for social interaction can help us understand the human mind

Further work

Speech recognition for children: problem solved?

ICSR '24




Realistic mouth movements don't help pronunciation learning

HRI '23

How much do **people align to LLMs?**

ACL '25

Language learning through **personalised multimodal storytelling**

HRI '24 + '26?

GHENT UNIVERSITY

imec

www.rubenjanssens.be