



**GESIS** Leibniz-Institute  
for the Social Sciences



# Integrating ChatGPT into cognitive pretesting procedures

Timo Lenzner & Patricia Hadler

CS3 Meeting, 12 June 2024



# ChatGPT (chat.openai.com)

ChatGPT 3.5 ▾

**You**  
What do the acronyms DGOF and GOR stand for?

**ChatGPT**  
DGOF stands for "Deutsche Gesellschaft für Online-Forschung," which translates to "German Society for Online Research." It's an organization in Germany dedicated to promoting and advancing online research methods.

GOR stands for "General Online Research," which is an annual conference organized by the DGOF. The GOR conference is a major event in the field of online research, bringing together researchers, practitioners, and experts to discuss the latest developments, methodologies, and applications in online research.

📄 👍 👎 ↻

Is this conversation helpful so far? 👍 👎 ×

Message ChatGPT... ⬆

ChatGPT can make mistakes. Consider checking important information.

# Cognitive Pretesting

The aim of cognitive pretesting is to get insights into the cognitive processes underlying survey responding:



- How do respondents interpret questions?
- How do they retrieve relevant information and events from memory?
- How do they arrive at a judgment about what to answer?
- How do they map their internal answer to the answer categories provided?

(Tourangeau, 1984; Tourangeau et al., 2000)

# Example

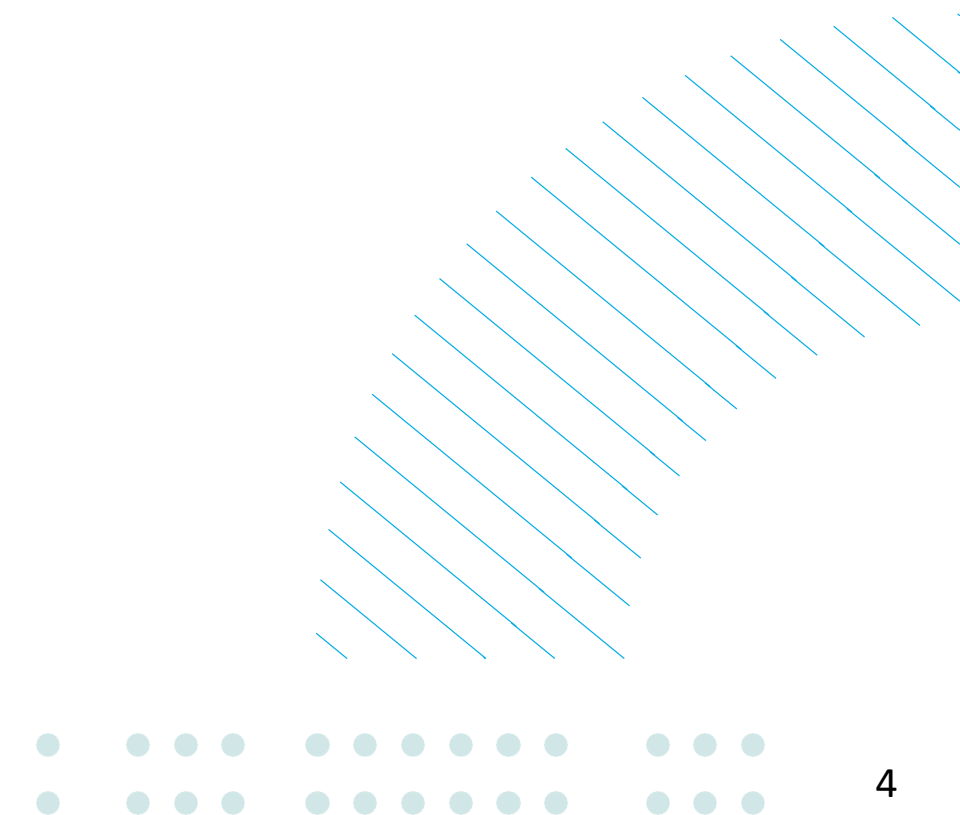
Question:

“I feel more like a citizen of the world than of any other country.”

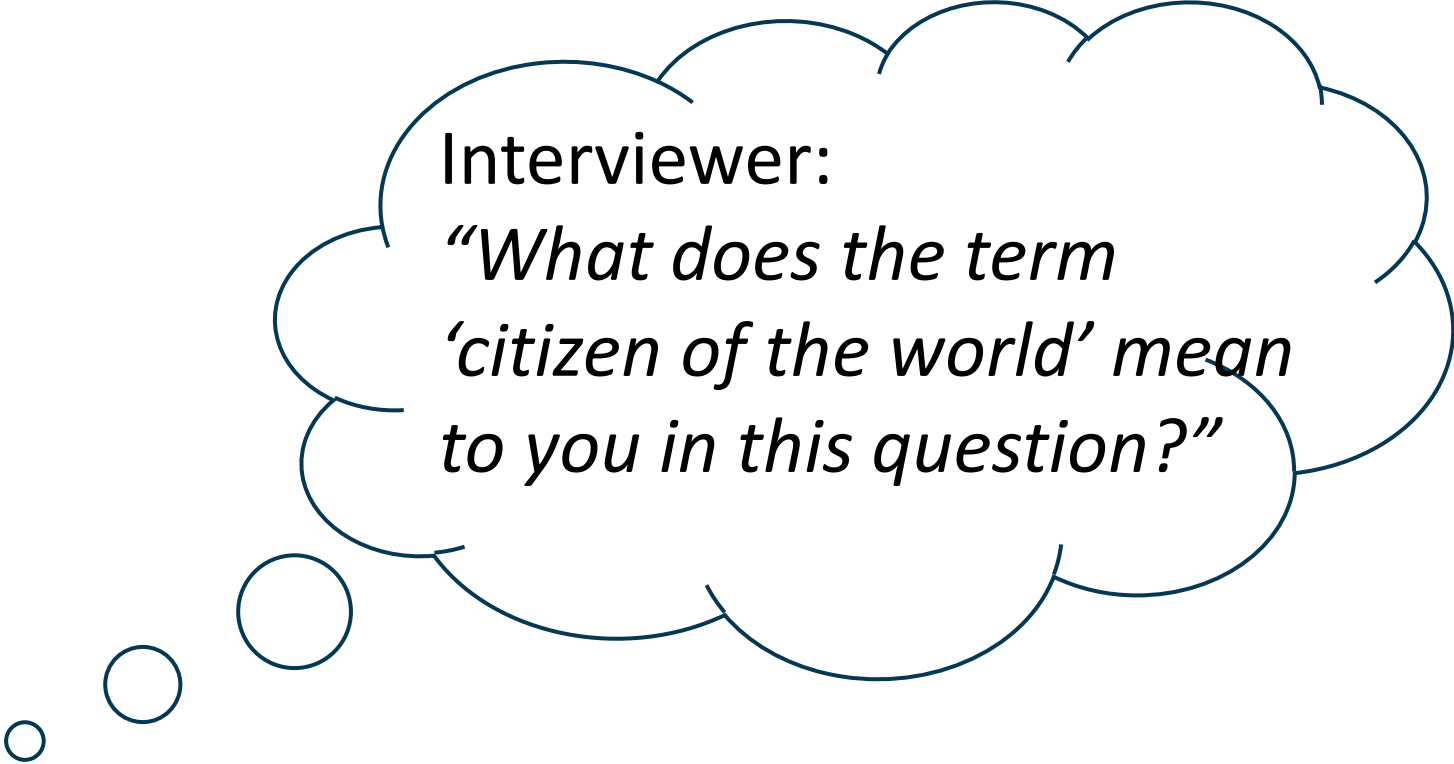
Strongly agree / Strongly disagree (5-point scale)

Results:

- Only 2 out of 20 test persons said that they don't know (or that they are unsure of) what the phrase “citizen of the world” refers to
- 18 test persons selected an answer category (without hesitation)



# Example



Interviewer:  
*“What does the term  
‘citizen of the world’ mean  
to you in this question?”*

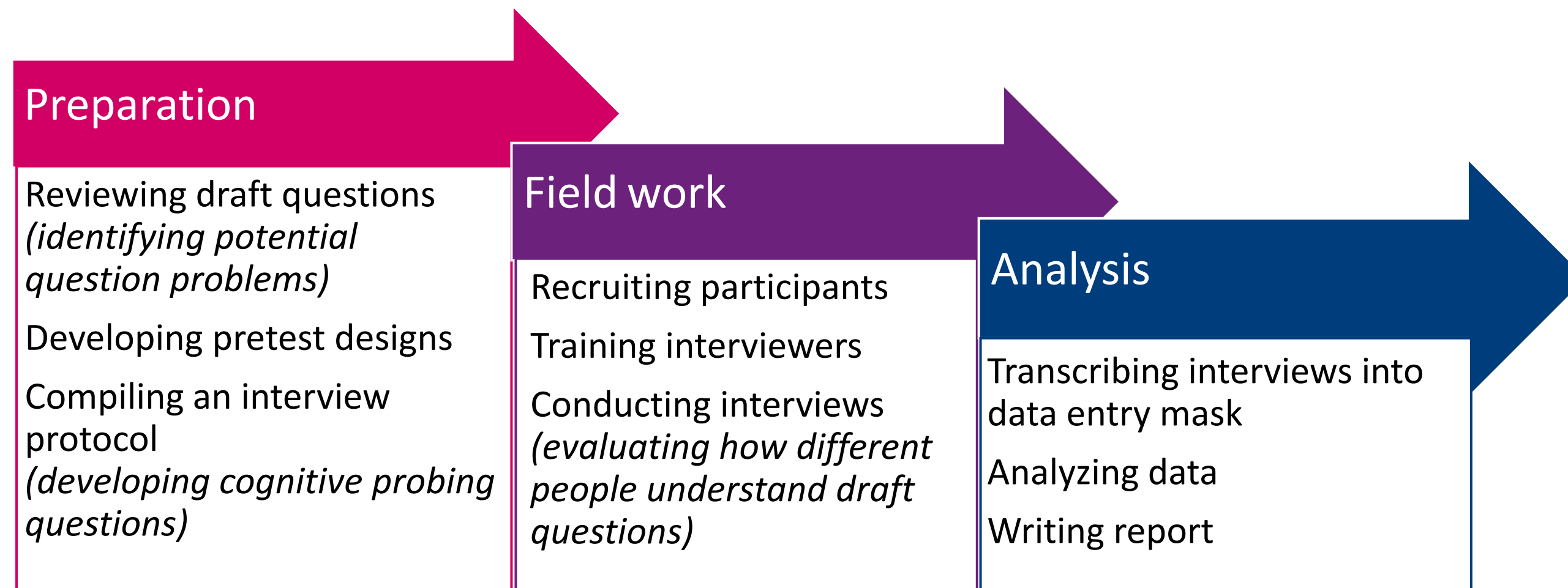
## Results of probing:

- 5 out of 18 test persons interpret the phrase wrongly in the sense of “human being“, for example:
  - *“For me, citizen of the world refers to all human beings living on earth.”*
  - *“Everyone who lives on earth is a citizen of the world.”*
- All 5 respondents “strongly/rather agree“ with this statement
  - Proportion of people feeling like “citizens of the world“ in a population would probably be overestimated (considerably) with this item

# Cognitive Pretesting

**Aim:** Evaluating the content-related validity of (draft) survey questions

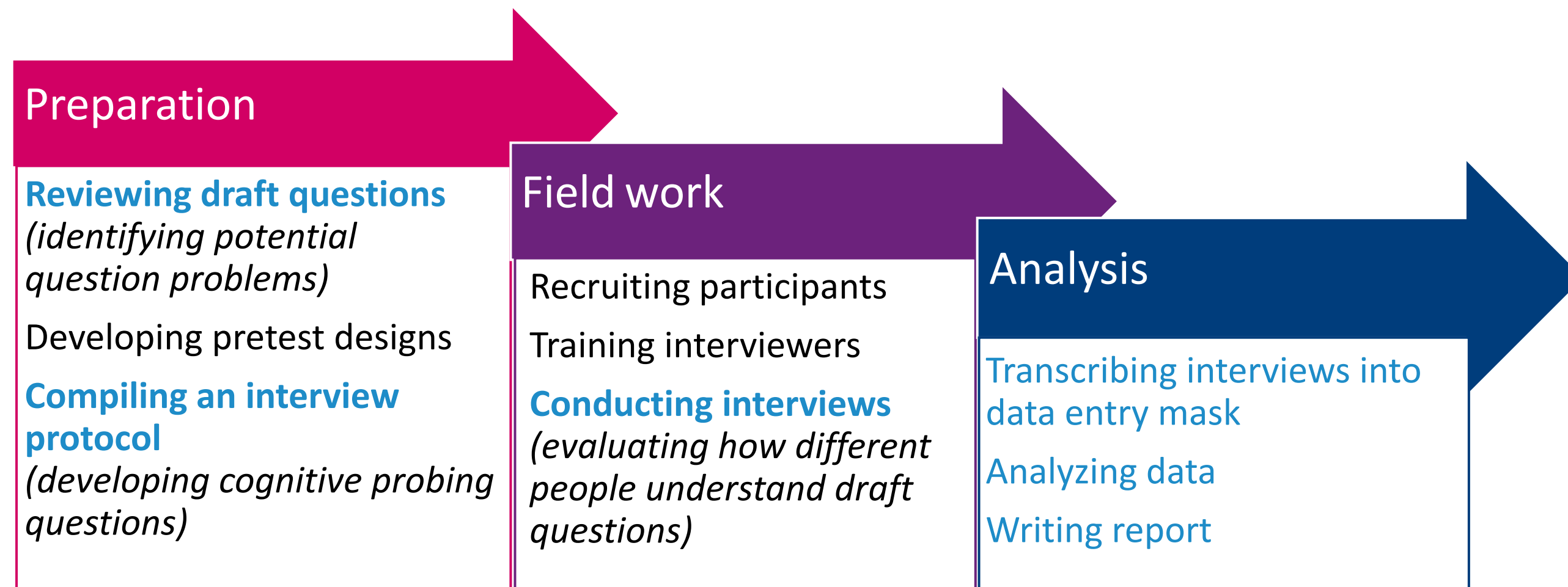
**Project stages:**



# Integration of LLMs

**Aim:** Increase efficiency and objectivity

**Project stages:**



# Research Questions

How good is ChatGPT at

- ...identifying potential question problems prior to cognitive testing?  
(RQ1)
- ...suggesting cognitive probing questions to test draft survey questions?  
(RQ2)
- ...simulating or predicting respondents' answers to probing questions?  
(RQ3)



# Methods & Data

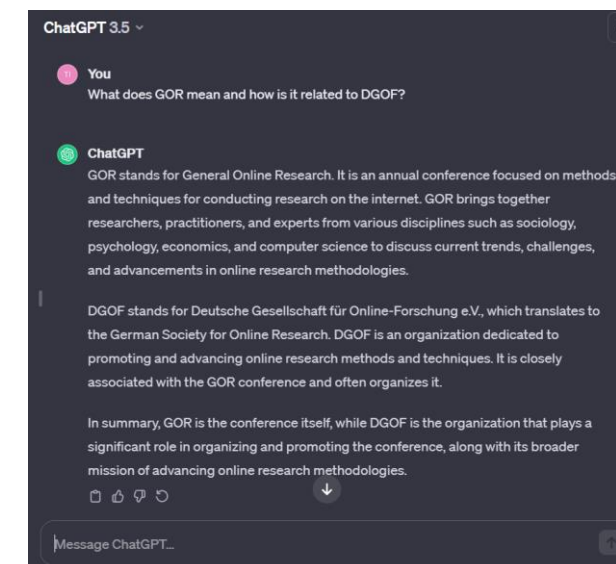
## Cognitive interviews (CI)

- 4 versions of a question on children's activity levels
- N = 24 (parents, children)
- Mode: CAPI, Remote-CAPI



## ChatGPT

- Version 3.5
- chat.openai.com
- Prompted in parallel to CI study (June/July 2023)

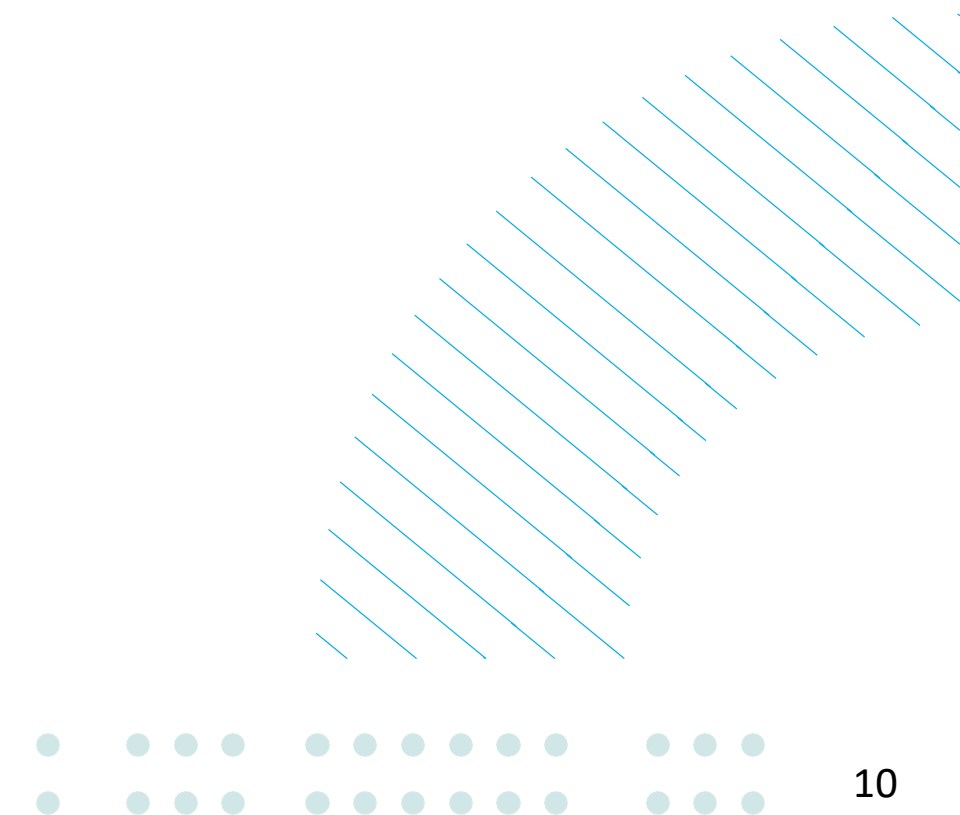


# Tested Question (Version 1, Parents)

The following question is about your child's movement. Please think of lively movements such as romping, brisk running, cycling and sports in daycare, school, education/work, and leisure time.

**How much lively movement has your child done in the last week?**

- less than 1 hour
- 1-2 hours
- 3-4 hours
- 5-6 hours
- at least 7 hours
- no answer*



# RQ1 – Identifying potential question problems

Prompt pattern



You

**Task:** What problems can arise when answering the following survey item?

**Survey item/question text:** The following question is about your child's movement.

Please think of lively movements such as romping, brisk running, cycling and sports in daycare, school, education/work and leisure time. How much lively movement has your child done in the last week?

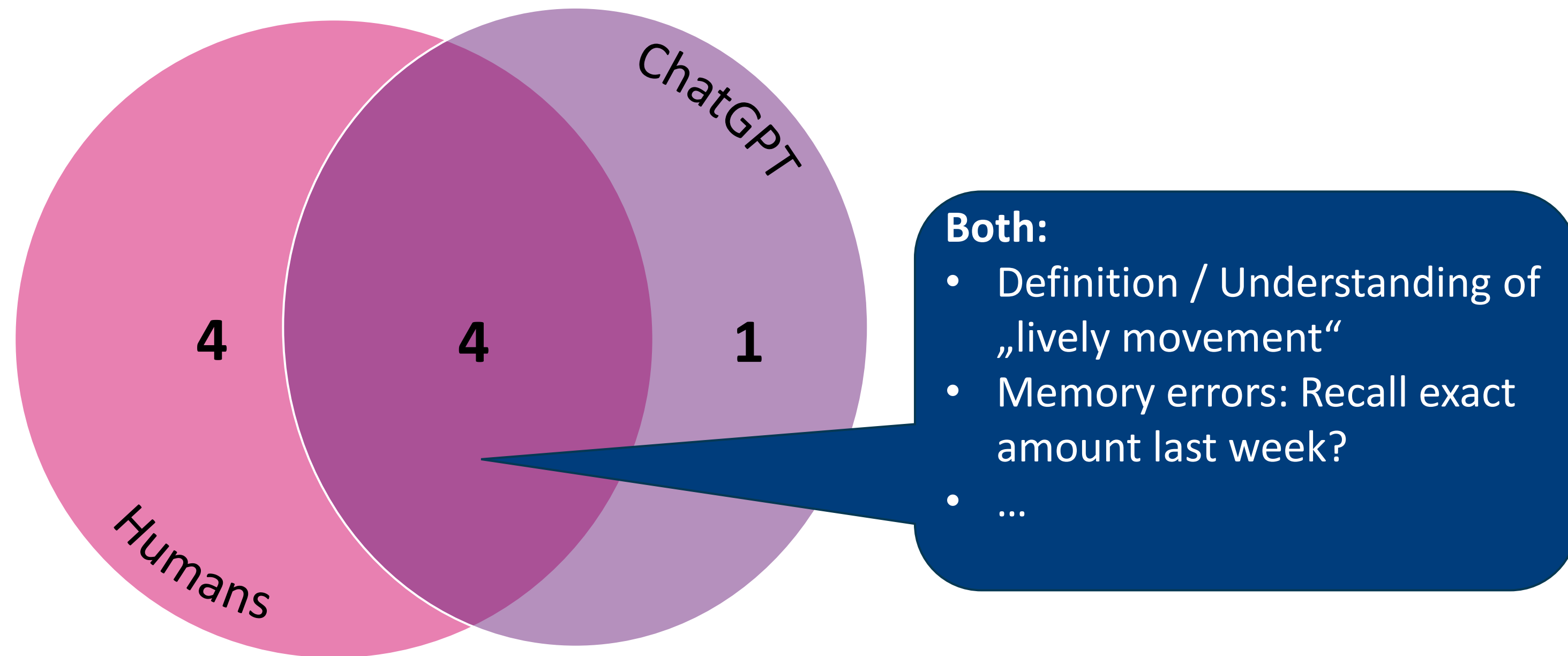
**Survey item/answer options:** less than 1 hour, 1-2 hours, 3-4 hours, 5-6 hours, at least 7 hours, no answer

**Target group:** Parents of children between the ages of 3 and 15

**Interview mode:** face-to-face

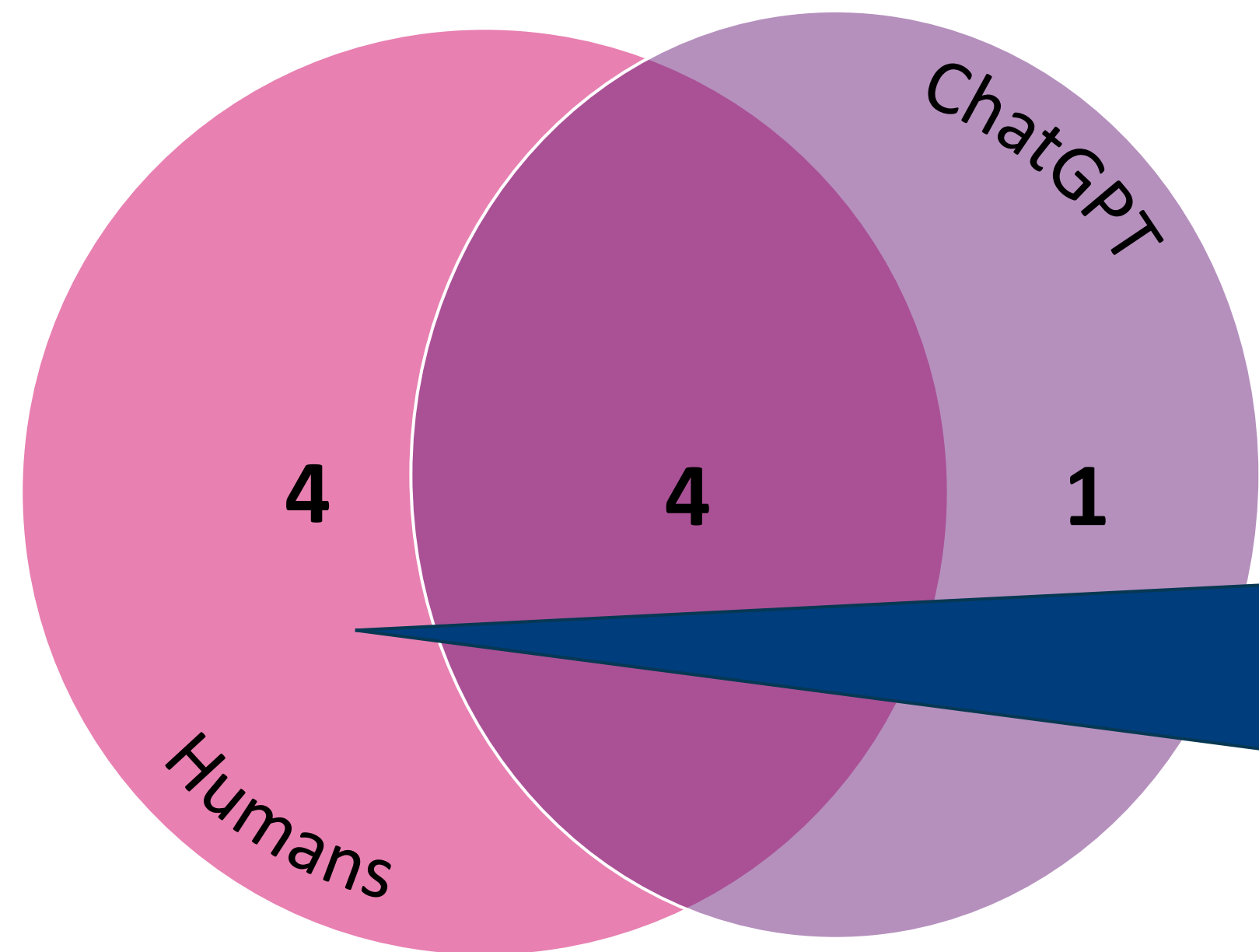
# RQ1 – Identifying potential question problems

Results



# RQ1 – Identifying potential question problems

Results

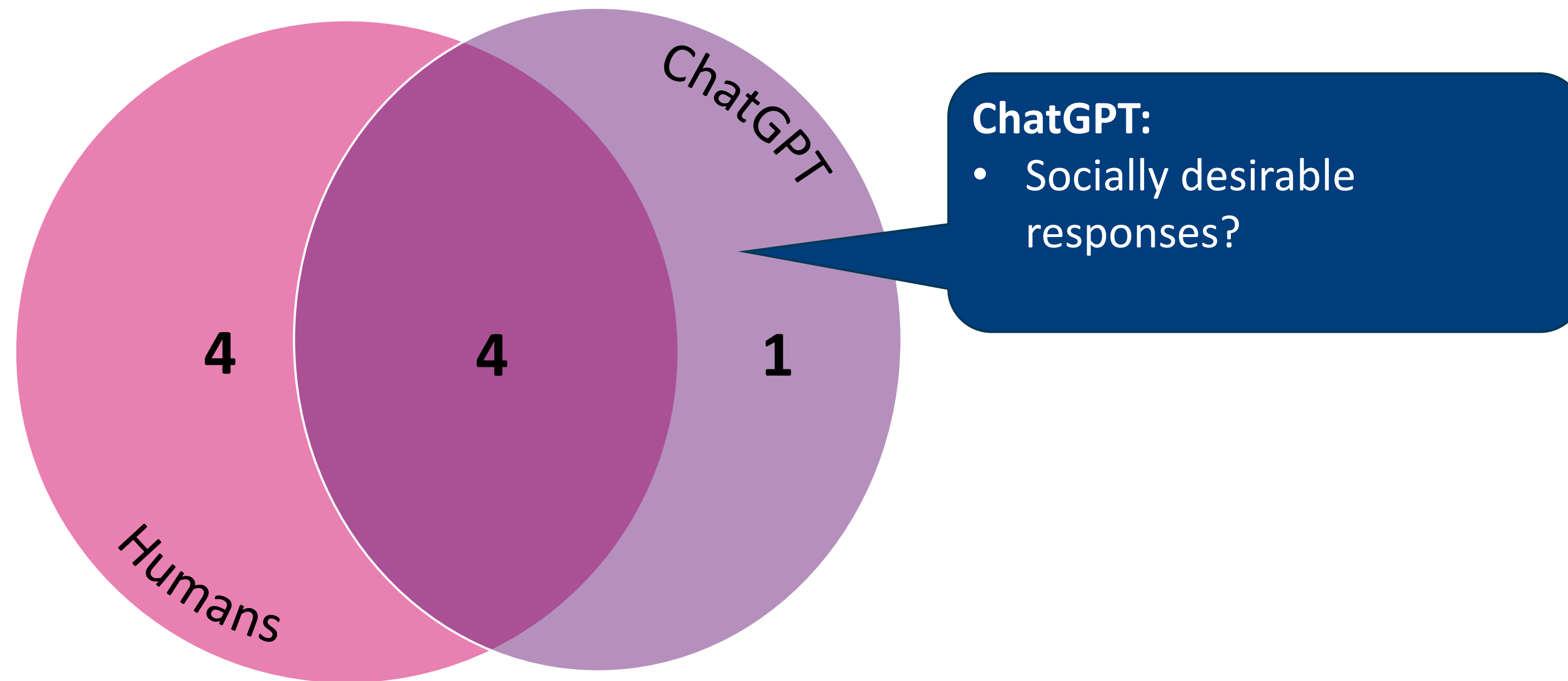


## Humans:

- Possible for parents to quantify children's behavior at all?
- „Last week“ vs. „normal week“: What if last week was untypical (illness, holidays)?
- ...

# RQ1 – Identifying potential question problems

Results



# RQ2 – Suggesting cognitive probes

## Prompt pattern

TI You

**Survey item/question text:** The following question is about your child's movement.

Please think of lively movements such as romping, brisk running, cycling, and sports in daycare, school, education/work, and leisure time. How much lively movement has your child done in the last week?

**Survey item/answer options:** less than 1 hour, 1-2 hours, 3-4 hours, 5-6 hours, at least 7 hours, no answer

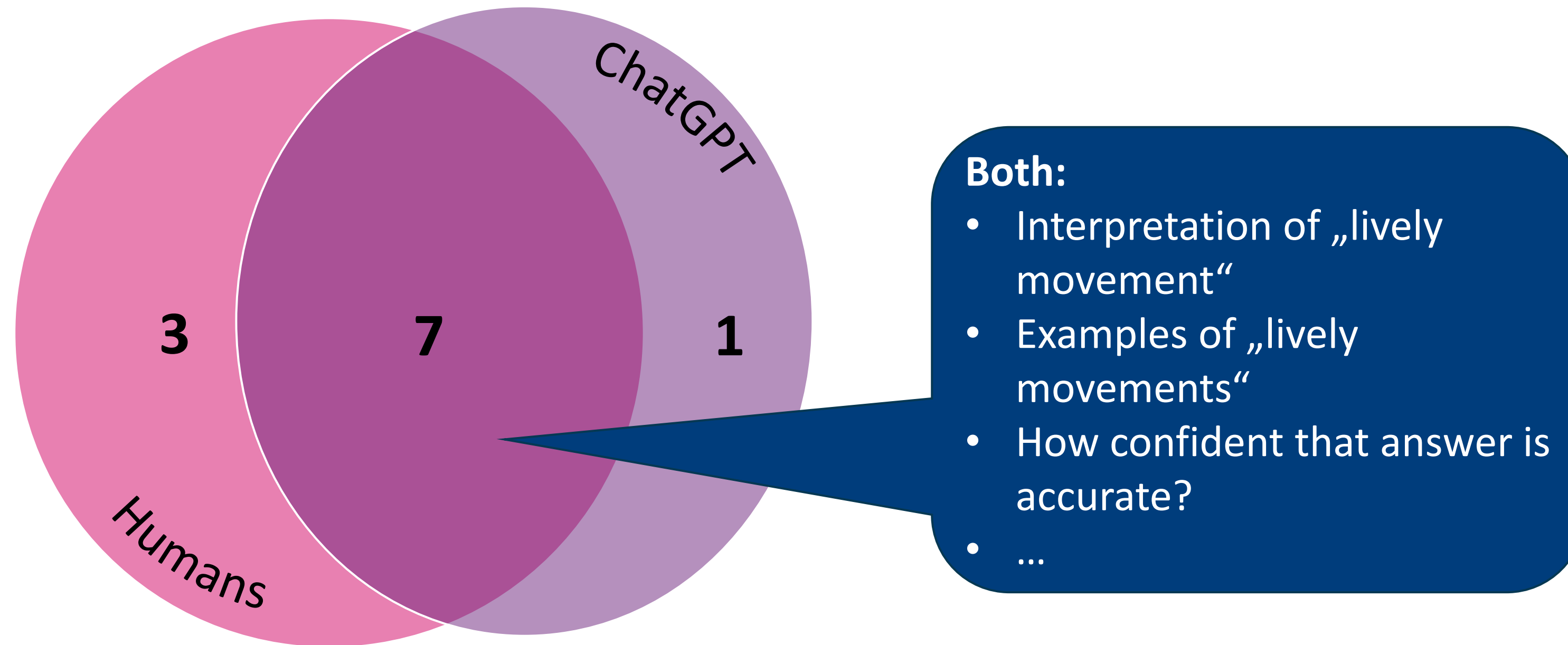
**Task:** Create an interview protocol with cognitive probing questions that can be used to answer the following questions.

### Questions:

1. How is the term "lively" understood and does it match the intended understanding?  
(Intended interpretation: "The term "lively" refers to movements that increase the child's breathing rate and may cause them to sweat. This can be targeted sports activities, lively movement games, but also everyday activities, such as cycling from place to place or brisk walking/running or romping.")
2. Is it understood that it can be both "lively movement in everyday life", such as cycling

# RQ2 – Suggesting cognitive probes

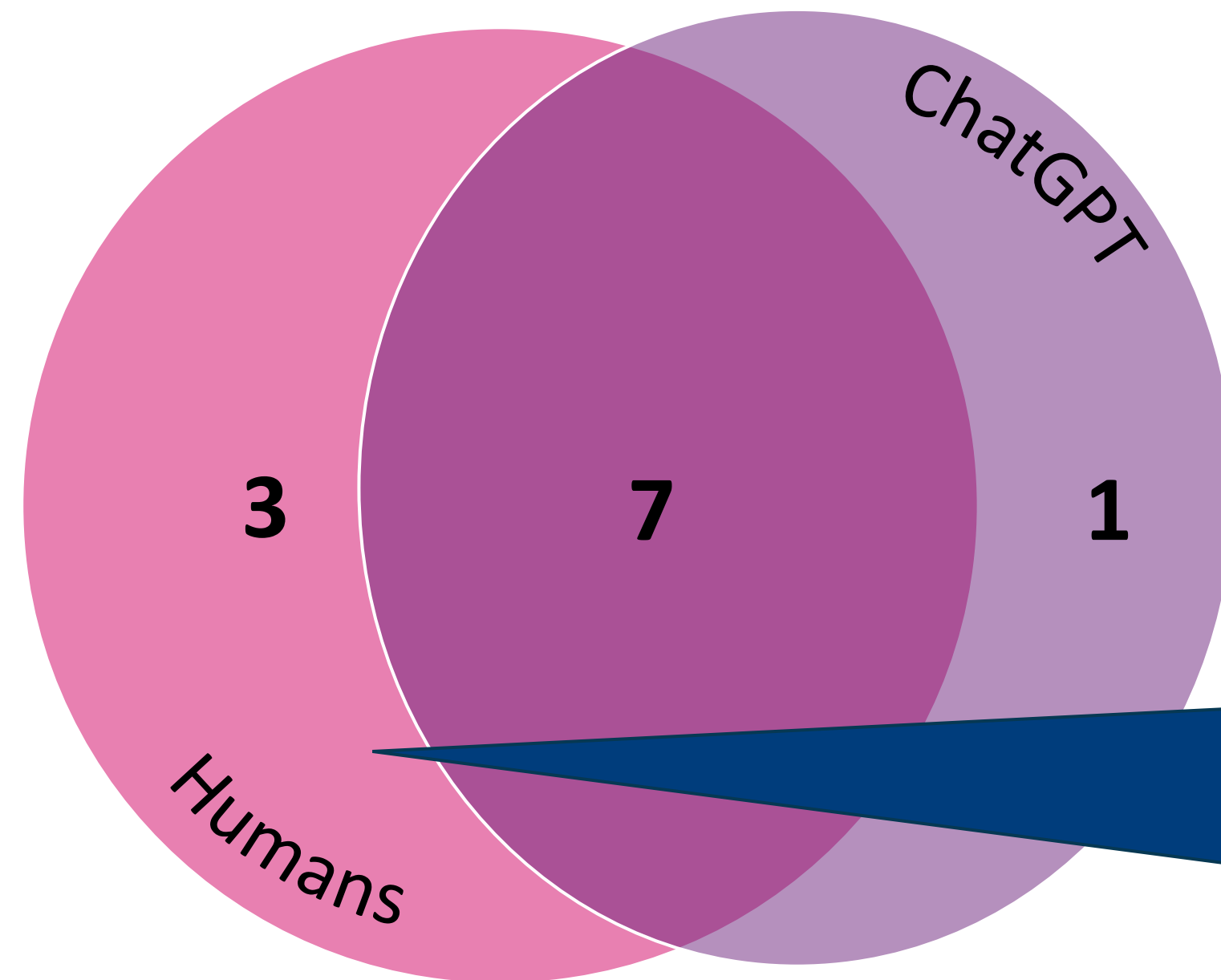
Results





# RQ2 – Suggesting cognitive probes

Results

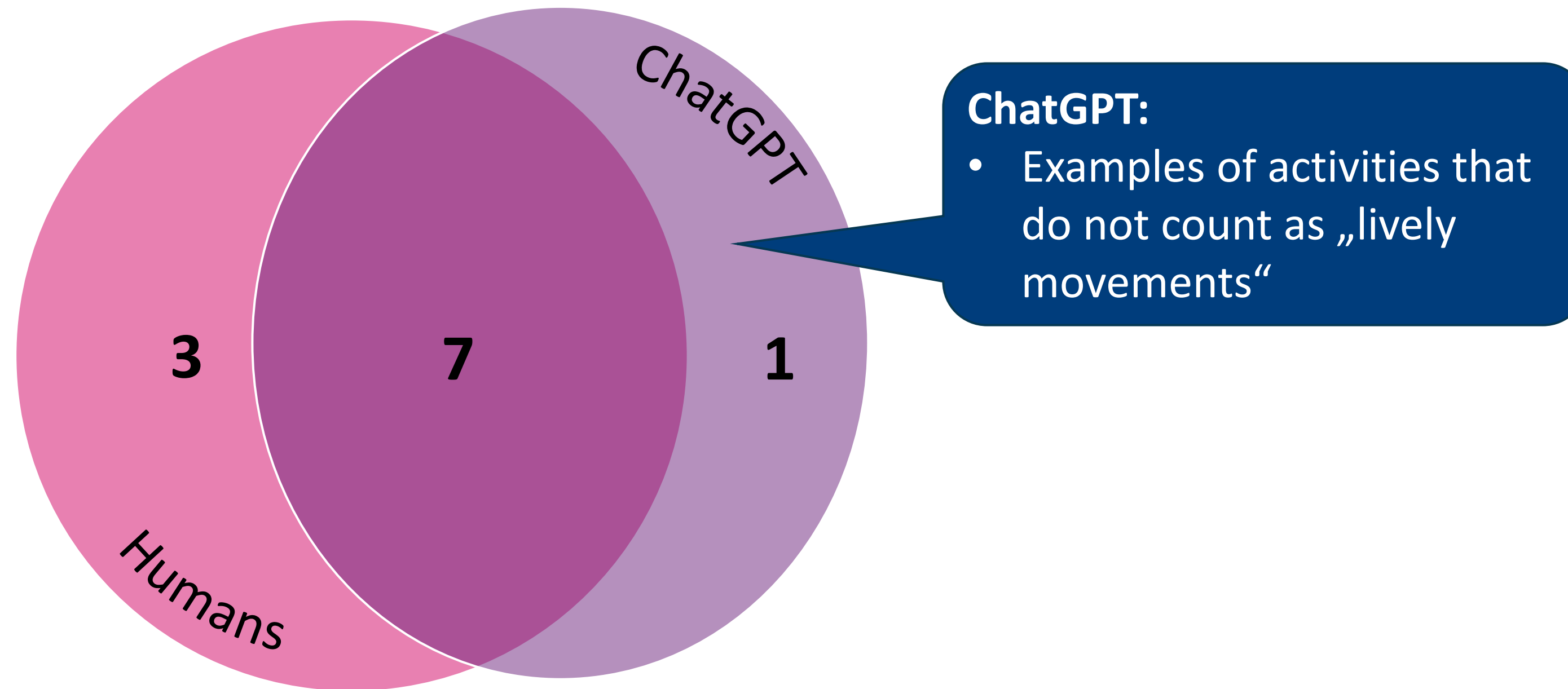


## Humans:

- How did you go about answering the questions?
- Spontaneous answer (if no answer categories were provided)?
- Counting vs. estimating no. of hours?

# RQ2 – Suggesting cognitive probes

Results



# RQ3 – Predicting probe answers

Prompt pattern („persona pattern“)

TI You

**Task:** Act as a 16-year-old girl attending a grammar school. Answer the following survey question and the cognitive probing questions.

**Mode of interview:** Face-to-face

**Survey question:** The following question is about your movement. Please think of lively movements such as running around, brisk walking, cycling and sports at school, training/work and leisure.

How much lively movement have you done in the last week?

**Answer options:** less than 1 hour, 1-2 hours, 3-4 hours, 5-6 hours, at least 7 hours, no answer

**Cognitive probing questions:**

1. How did you go about answering the question? Please explain in more detail what you were thinking about when you answered the question.
2. What would you have spontaneously given as an answer if no answer options had been given?
3. Did you think of all the activities during the individual days of the week last week and add them up or did you rather estimate the number of hours over the week?

## RQ3 – Predicting probe answers

- Overall, ChatGPT's answers hardly match the answers of real respondents
- Deviations:
  - Very little variation in answers (of people with different characteristics)
    - *“By ‘lively movement’ I mean physical activities where I move intensively, my heart rate is increased and I may work up a sweat.”* (female, 16 years)
    - *“By ‘lively movement’ I mean physical activities in which I move more intensively, increase my breathing rate and possibly work up a sweat.”* (male, 17 years)
    - *“For me, ‘lively movement’ means being physically active, that your breathing and heartbeat are increased.”* (female, 17 years)
    - *“For me, ‘lively movement’ means being active and moving around a lot, so you might get a bit out of breath and sweat.”* (male, 12 years)

## RQ3 – Predicting probe answers

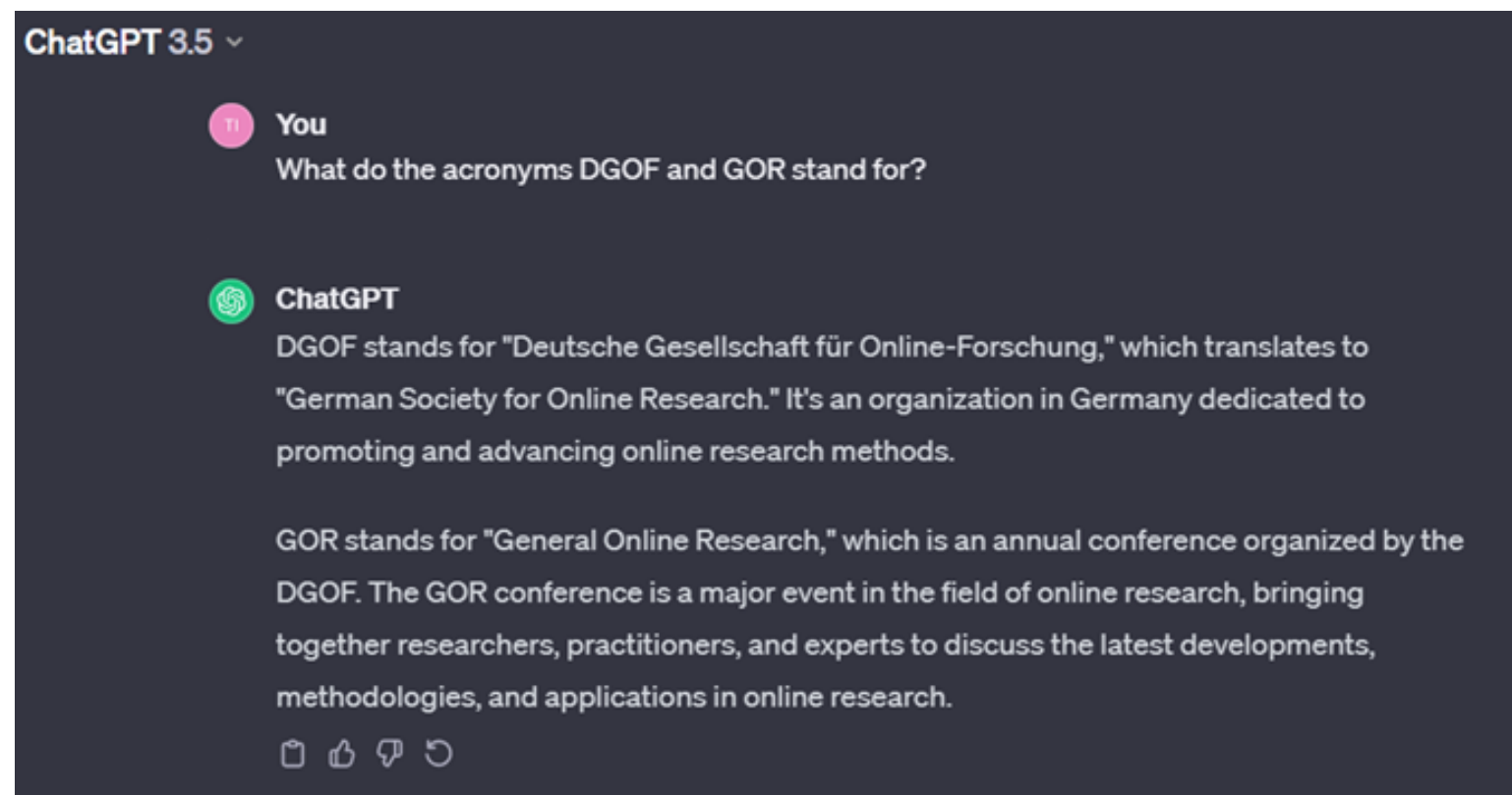
- Overall, ChatGPT's answers hardly match the answers of real respondents
- Deviations:
  - Very little variation in answers (of people with different characteristics)
  - Strange choice of words (for the age of the respondents)
    - Male, 16 years old: *"I often just **play outside** [Ich spiele oft einfach draußen]";  
"I thought of ... running, and **romping in the playground** [...Toben auf dem Spielplatz]"*
  - Reproduction of gender stereotypes (to some extent): Girls do running or dancing, boys play soccer or lift weights at the gym
  - ChatGPT's answers do not address any problems with the response scale, while the real respondents criticize the scale several times

# Summary & Conclusion

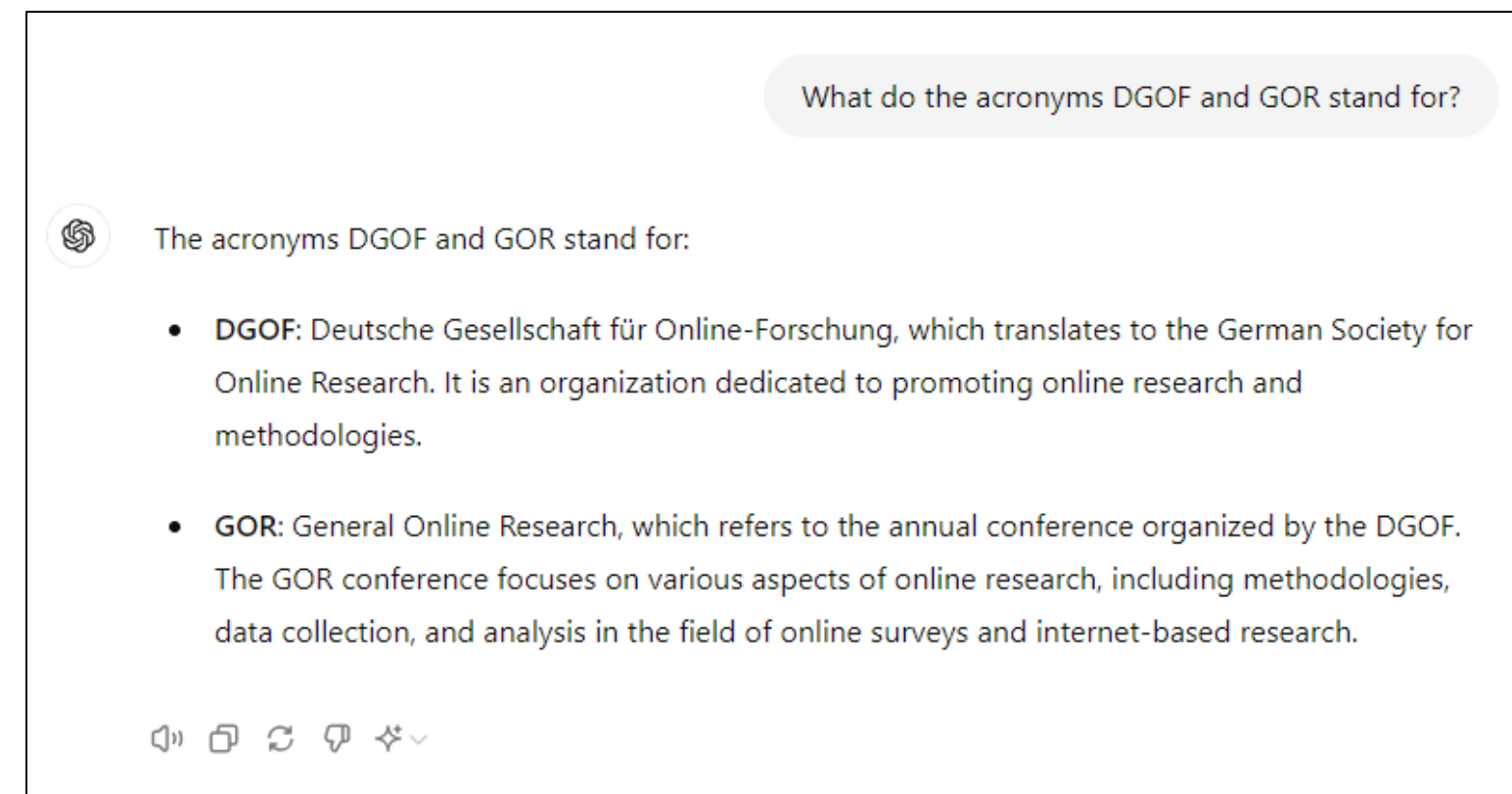
- ChatGPT
  - ...is helpful in identifying potential question problems and suggesting cognitive probes to test draft questions
  - ...does not (yet?) predict individual responses to cognitive probes accurately
- Potentially increases objectivity of pretesting procedures, but not (yet?) the efficiency
- For now: LLMs add to the researchers' toolkit; they neither replace researchers/interviewers nor respondents

# Limitations

- Limited data basis (few pretest projects and questions to date)
- Effects of prompt design on output
- GPT 3.5 (already outdated?)
- Replicability issues



Feb 2024



June 2024

**Thank you for your attention.**

Contact:  
[timo.lenzner@gesis.org](mailto:timo.lenzner@gesis.org)