

Supplementary Information

GPT4 model primer

Here we present the full model primer that was used to prompt GPT4. Note that while we focused on expressive communication in this study, we also asked GPT4 to predict the receptive communication scores.

“””

You will be provided with multiple conversations between two people: a person with autism, and their study partner.

BACKGROUND ON THE FORMAT OF THE CONVERSATIONS:

The person with autism is labelled as [participant], and the study partner is labelled as [study_partner]. However, the study_partner might sometimes be labelled as [F1] or [M1].

There may or may not be other people in the conversation, labelled as [F2] or [M2] etc...

Any names in the conversation have been redacted for privacy concerns, and replaced with <redacted>.

YOUR JOB:

Your job is to provide a score between 0 and 100 of both the expressive and receptive communication abilities of the participant based on the Vineland Adaptive Behaviour Scales clinical scale.

OUTPUT FORMAT:

This is VERY IMPORTANT, the format of the final score should be as follows:

Final Expressive Communication Score: X

Final Receptive Communication Score: Y

“””

Where example conversations were used, the primer also included lines like the following:

“””

Here is an example of a conversation that should score very low on both scales:

[study_partner] Hi, how was school today? [participant] Good.

...

And here is an example of a conversation that should score very high on both scales:

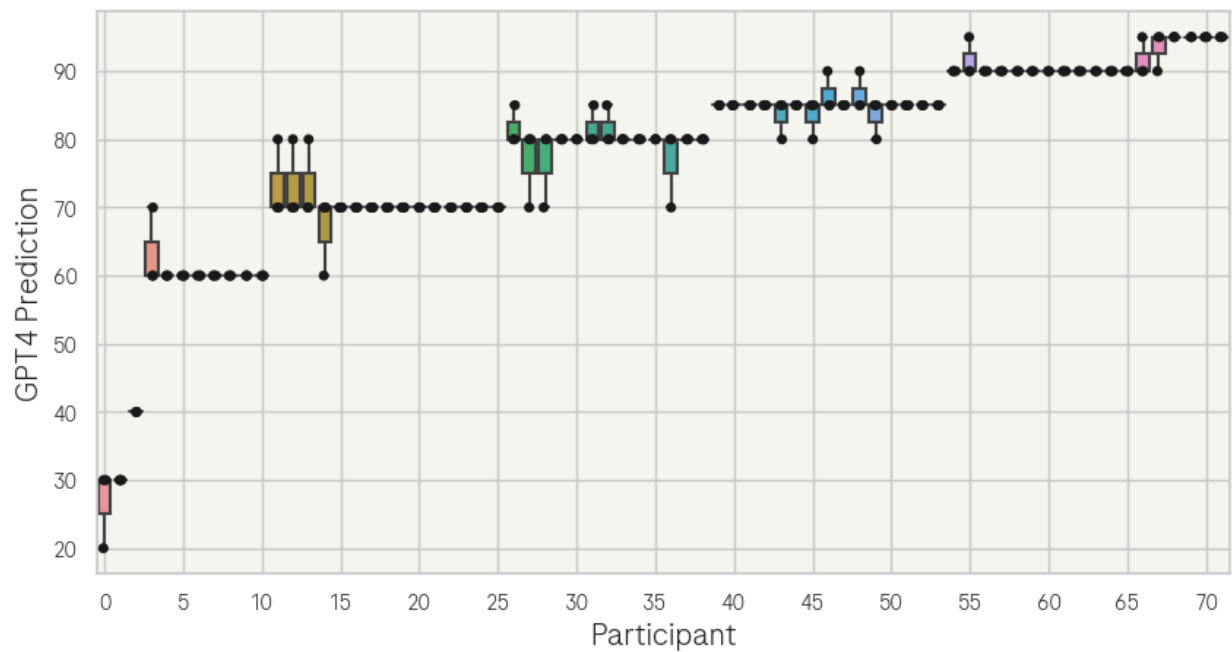
[study_partner] Ok so what should we talk about today? [participant] Oh I don't know, how about the weather?

...

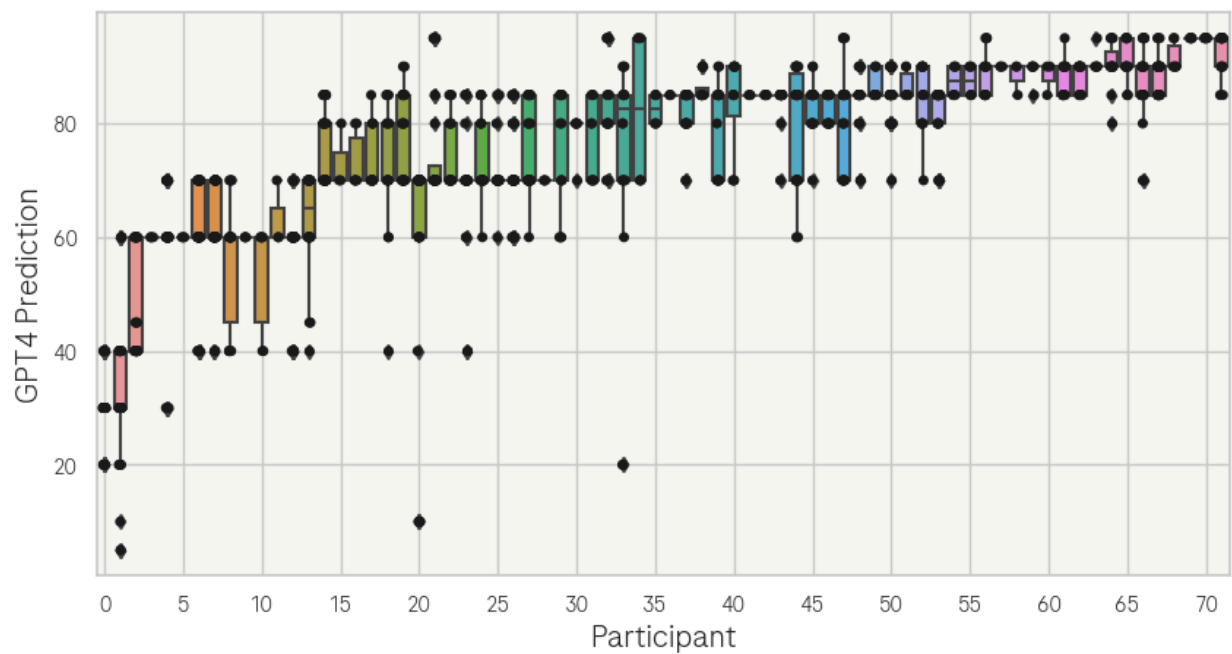
“””

However, please note that unfortunately we are unable to provide full transcripts. This is to ensure participant privacy, because the informed consent agreements limit the individuals that are allowed to read the transcripts.

Intraclass Correlation Visualization



Supplementary Figure 1: A boxplot illustrating the variability across iterations of GPT4 ratings when using the concatenated conversations of each participant. The concatenated conversations were each submitted to GPT4 on 3 separate iterations, revealing a remarkable consistency ($ICC(2,1) = 0.97$, [95% CI: 0.96, 0.98]).



Supplementary Figure 2: A boxplot illustrating the variability of GPT4 predictions within each subject, across all individual conversations of each participant and GPT4 ratings iterations. This therefore assesses not just the consistency of GPT4, but also the consistency of each conversation. $ICC(2,1) = 0.79$ [95% CI: 0.72, 0.84]

When asking GPT4 for symptom descriptions

Below are some example outputs of GPT4 when asking it to summarise the conversation of a participant and describe their autism symptoms if evident:

Participant 1

Summary of Conversations:

The conversations involve a participant with autism engaging with their study_partner. The participant often expresses a desire to eat, mentioning various foods like pizza, fish, McDonald's, and ice cream. They also discuss activities such as going to the zoo, getting haircuts, and attending school. The participant frequently mentions wanting to visit places like Chuck E. Cheese's and Six Flags, and expresses interest in toys, particularly from the Power Rangers Dino Charge series. The participant has difficulty staying on topic, often repeating phrases or fixating on certain subjects, such as food or specific toys. The study_partner prompts the participant to use their words, slow down, and stay focused during the conversation. The participant also shows some confusion about days and dates, and there are instances where they need to be reminded of their schedule and activities.

Assessment of Autism Symptoms:

Expressive Communication: The participant demonstrates repetitive speech, limited variety in language use, and a strong focus on specific interests (e.g., food, toys). They have difficulty with topic maintenance and require prompts to expand on their thoughts. The participant also exhibits echolalia, repeating phrases or words they have heard.

Receptive Communication: The participant shows some understanding of questions and prompts but often provides unrelated or repetitive responses. They seem to have difficulty processing and following multi-step conversations, and the caregiver frequently has to redirect them or simplify questions.

Based on the Vineland Adaptive Behavior Scales clinical scale, the participant's scores are as follows:

Final Expressive Communication Score: 40

Final Receptive Communication Score: 40

Participant 2

Summary of Conversations:

The conversations between the participant and the study partner revolve around various topics, including the participant's interest in buses, subways, and their experiences at school. The participant expresses a fondness for New York City, where they were born, and a preference for the Bronx Division of buses. They also discuss their behaviour at school, specifically an incident on Tuesday where they were reported for being mean, which led to them being grounded from electronics. The participant acknowledges their mistake and indicates they have learned their lesson. They also talk about their day playing WiiSports and their visit to their grandmother's house. Additionally, the participant shows a keen interest in trains, discussing different models and their construction. They also touch upon historical figures like Martin Luther King Jr. and the significance of Mount Rushmore. The participant seems to have a good understanding of the differences between school systems in the city and their current location. They also express a desire to visit places like Washington D.C. and the Dominican Republic, provided they learn to eat a wider variety of foods.

Assessment of Autism Symptoms:

Expressive Communication: The participant demonstrates a strong ability to express themselves, sharing detailed information about their interests, experiences, and feelings. They engage in back-and-forth conversation, ask and answer questions, and express complex ideas, such as their understanding of the differences in school systems and their thoughts on historical events.

Receptive Communication: The participant shows good comprehension of the study partner's questions and comments. They respond appropriately to the context of the conversation and are able to follow the flow of dialogue, indicating a strong receptive communication ability.

Final Expressive Communication Score: 85

Final Receptive Communication Score: 85

Participant 3

Summary of Conversations:

The conversations between the participant and the study partner revolve around daily activities, upcoming trips, school experiences, meals, and holiday plans. The participant discusses going shopping with Uncle <redacted>, eating at Olive Garden, and watching movies at school. They also talk about a trip to Philadelphia and New York City, where the participant is excited to visit Rocky-related sites and the Nintendo store. The participant expresses a desire for specific Christmas gifts, including movies and video games, and discusses preferences for meals like Popeye's chicken. There is also mention of getting a haircut and packing for a trip. The participant occasionally uses filler words like "uh" and "um" and sometimes gives brief or one-word answers but is also capable of providing more detailed responses when prompted.

Assessment of Autism Symptoms:

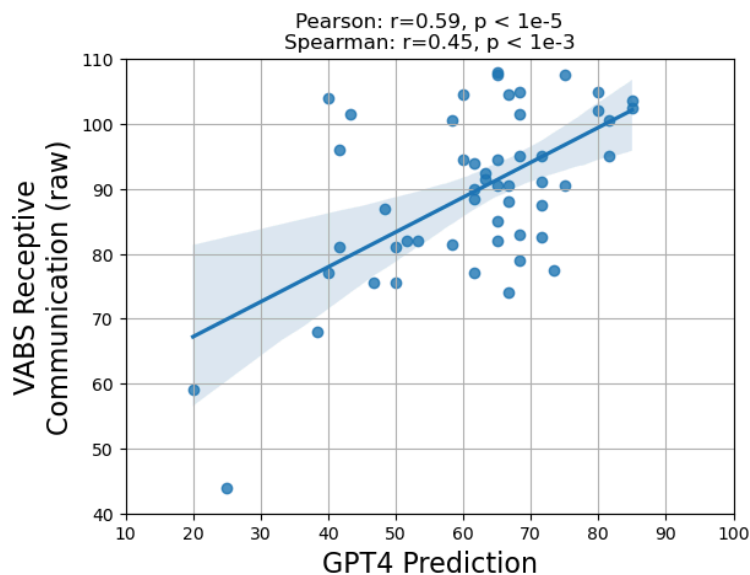
Expressive Communication: The participant is able to communicate their thoughts and desires, such as their excitement about the trip and specific Christmas gifts they want. They use complete sentences and can engage in back-and-forth conversation, although they sometimes give minimal responses or need prompting to elaborate.

Receptive Communication: The participant demonstrates understanding of the study partner's questions and comments, responding appropriately to the context of the conversation. They follow the conversation's flow and can answer questions related to their activities and preferences.

Final Expressive Communication Score: 70

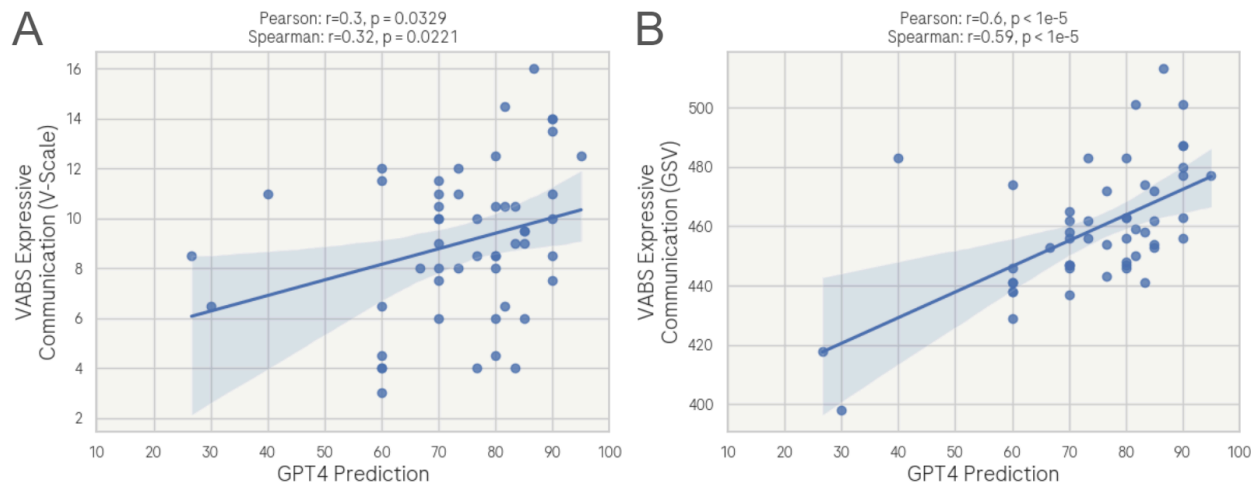
Final Receptive Communication Score: 75

Correlation with the actual VABS scores when also asking for additional information



Supplementary Figure 3: The Correlation between the actual VABS expressive communication scores (y-axis) and the predicted scores from GPT4 (x-axis) when asking GPT4 for additional information, such as a description of the conversations and autism symptoms. Each dot is a participant (n=52).

Correlation with V-Scale and GSV scores



Supplementary Figure 4: The Correlation between the predicted scores from GPT4 (x-axis) and the V-Scale (A) and Growth Scale Value (GSV; B) scores (y-axes). Each dot is a participant (n=52).

Inclusion and Exclusion Criteria

The inclusion criteria for autistic participants were as follows:

- Diagnosis of ASD based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), and the Autism Diagnostic Observation Schedule (ADOS-2).
- Children's Yale-Brown Obsessive Compulsive Scale modified for ASD (CY-BOCS-ASD) total score of at least 12.
- Clinical Global Impression-Severity (CGI-S) score of at least 4 about participant's current autism severity.
- Intelligence quotient (IQ) score of 50 or above as assessed by the Abbreviated Intelligence Quotient (ABIQ) SB5 scale.

- English proficiency compatible with the study measurements as judged by the investigator.
- Hearing, vision, and speech compatible with the study measurements as judged by the Investigator.
- All medications and treatments were expected to be stable for the duration of the study.

The diagnostic evaluations were completed at the study site by research staff and supervised by a licensed psychologist.

The exclusion criteria for all participants were as follows:

- Participation in an investigational drug or device study within 4 weeks, or five times the half-life (if it is a drug study) of the investigational molecule (whichever is longer), prior to screening and the participant is expected not to enrol in any other trial during the study.
- Co-occurring disease or condition that could interfere with, or treatment which might interfere with, the conduct of the study, or that would, in the opinion of the Investigator, pose an unacceptable risk to the participant in this study.
- Unstable or uncontrolled clinically significant psychiatric and/ or neurological disorder that may interfere with the objectives of the study.
- Participants with known "syndromic" ASD (e.g., Fragile-X syndrome, Angelman Syndrome, Prader-Willi, Rett's syndrome, tuberous sclerosis, Dup15q syndrome).
- History of alcohol misuse and/ or illicit drug use during the last 12 months prior to screening.

The inclusion criteria for the study partners were as follows:

- A staff member of the residential home can be the caregiver if this person spends sufficient time with the participant. In the opinion of the Investigator, the caregiver must be able to reliably assess the participant's mental status, activities, and behaviour, and report on the participant's adherence and health. This would normally be possible when the caregiver spends a few hours each day with the participant.
- A family member living at the participant's home can be the caregiver if the participant returns home every night. When the participant returns home only over the weekend, a family member can only be the caregiver if they have intensive interaction with the participant during the week e.g., via phone calls, calls via Skype, SMS messages, etc. The quality of these interactions between caregiver and participant needs to be assessed for each participant to determine whether they are sufficient.

Transcription Guidelines

- Turns were coded as utterances, not C-units or T-units or intonational phrases.
- Each line should contain one utterance from one speaker in the following format:
[speaker] [[timestamp start end]] [transcript][end of line]
E.g.:
P [timestamp 2:04.3 2:05.4] well that's a good one.
P [timestamp 2:07.8 2:09.1] my last name is Charlene.
- If an utterance contains a silence longer than 1 second it should be separated over 2 lines, each with their own timestamps.
- Only the last line of the utterance should end with a full stop, question, or exclamation mark.
- There should be a separate transcription of each individual speaker, also in cases when two or more people speak at the same time.

With regards pauses and interruptions etc., we gave detailed instructions to the transcribers as follows:

Filled pauses. We are interested in hesitations, and negative and positive backchannel responses.

Whenever possible, use the following spellings:

■ hesitation:

uh, um, ah, er, eee, eh, ew, ha, hee, hmm, huh, mm, oof

■ negative responses/backchannelling:

huh-uh, uh-uh, nuh-uh, mm-mm, tsk-tsk

■ positive responses/backchannelling:

mm-hmm, uh-huh, mm-huh, uh-hmm, um-hmm, um-hum

- Provide a consistent annotation of any other sounds that the speaker makes to fill silence, if none of the above can be used to describe them.
- False starts and interruptions involving word fragments: Write word fragments with a trailing asterisk, e.g. would you like some oran* apple juice?
- If it's clear what the full word would be, please provide the unspoken part of the word fragment in parentheses, e.g.: would you like some oran(ge) apple juice?
- Treat stutter the same way as other word fragments: s* s* s* stutter
- False starts and interruptions involving full words: end the false start with a double hyphen, e.g.: I went to the mall -- I mean, to the store and bought groceries.
- Whenever an interruption is continued (e.g. after the other speaker says something), begin it with double hyphen, e.g.:

C did you do that or --

P oops. no, I forgot.

C -- did you forget again?