

Supplementary Data 1: Data Schema and Questionnaire Mappings

Legend: This document provides the complete data schema and concept mappings required to understand the structure of the input data used for the machine learning models. It is composed of two parts:

- **Part A: Questionnaire to Concept Mappings.** These tables detail the mapping between the raw questions in the MCHAT questionnaires and the corresponding medical concepts used as features in our text-based models.
- **Part B: Final Data Schema.** This table provides a dictionary for the minimal dataset used in the analysis. It lists all variables programmatically generated in the preprocessing scripts and used as direct input for the models. The scripts can be found in our GitHub repository:
 - **Stage 1 Preprocessing:** https://github.com/skwgbobf/Multimodal-AI-ASD-Risk-ScreeningF/blob/main/Stage1_model/notebooks/1.Preprocess_git.ipynb
 - **Stage 2 Preprocessing:** https://github.com/skwgbobf/Multimodal-AI-ASD-Risk-ScreeningF/blob/main/stage2_model/notebooks/1_Data_Preprocessing/model2_data%20preprocess_Final.ipynb

Part A: MCHAT Questionnaire to Concept Mapping

DSM5 Criteria	Symptom of DSM5 Criteria	MCHAT Question_eng	mapping ASD terms
A1	social-emotional reciprocity	1. Does your child enjoy being swung, bounced on your knee, etc.? Yes No	problem parent-child
A1	failure to initiate social interactions	2. Does your child take an interest in other children? Yes No	does not play with other children
A1	failure to respond to social interactions	2. Does your child take an interest in other children? Yes No	does not interact with other children
A1	failure to respond to social interactions	2. Does your child take an interest in other children? Yes No	does not interact much with other children

B2	difficulties with transitions	3. Does your child like climbing on things, such as up stairs? Yes No	difficulty walking up stairs
A1	failure to initiate social interactions	4. Does your child enjoy playing peek-a-boo/hide-and-seek? Yes No	does not seek out contact with other children
A3	difficulties in sharing imaginative play	5. Does your child ever pretend, for example, to talk on the phone or take care of dolls, or pretend other things? Yes No	learning through pretend play
A1	failure to initiate social interactions	6. Does your child ever use his/her index finger to point, to ask for something? Yes No	does not use gestures to communicate
A1	failure to initiate social interactions	7. Does your child ever use his/her index finger to point, to indicate interest in something? Yes No	does not use gestures to communicate
A1	failure to initiate social interactions	8. Can your child play properly with small toys (e.g. cars or bricks) without just mouthing, fiddling, or dropping them? Yes No	inspected toys but did not play with toys
A1	failure to initiate social interactions	8. Can your child play properly with small toys (e.g. cars or bricks) without just mouthing, fiddling, or dropping them? Yes No	play with toys is not very functional
B3	highly restricted interests	8. Can your child play properly with small toys (e.g. cars or bricks) without just mouthing, fiddling, or dropping them? Yes No	no play
A1	social-emotional reciprocity	9. Does your child ever bring objects over to you (parent) to	parent-child relationships

		show you something? Yes No	
A1	failure to initiate social interactions	9. Does your child ever bring objects over to you (parent) to show you something? Yes No	does not play with others
B2	need to take same route	10. Does your child look you in the eye for more than a second or two? Yes No	does not establish eye contact
A2	abnormalities in eye contact	10. Does your child look you in the eye for more than a second or two? Yes No	does not make good eye contact
A2	abnormalities in eye contact	10. Does your child look you in the eye for more than a second or two? Yes No	impairment in eye-to-eye gaze
B4	adverse response to specific sounds	11. Does your child ever seem over sensitive to noise? (e.g., plugging ears) Yes No	ringing in ears sounds like
B4	adverse response to specific sounds	11. Does your child ever seem over sensitive to noise? (e.g., plugging ears) Yes No	slight sensitivity to noise
B4	adverse response to specific sounds	11. Does your child ever seem over sensitive to noise? (e.g., plugging ears) Yes No	definite sensitivity to noises that are not distressing to most other people
B4	adverse response to specific sounds	11. Does your child ever seem over sensitive to noise? (e.g., plugging ears) Yes No	definite sensitivity to noises
A2	lack of facial expressions	12. Does your child smile in response to your face or your smile? Yes No	milestone: has a social smile
A1	failure to respond to social interactions	12. Does your child smile in response to your face or your smile? Yes No	response delay
A3	absence of interest in peers	13. Does your child imitate you? (e.g., you make a face-will your child imitate it?) Yes	atypical child with whom to interact

		No	
A1	social-emotional reciprocity	13. Does your child imitate you? (e.g., you make a face- will your child imitate it?) Yes No	interacts positively with child
A1	failure to respond to social interactions	14. Does your child respond to his/her name when you call? Yes No	does not appropriately respond verbally
A1	failure to respond to social interactions	14. Does your child respond to his/her name when you call? Yes No	not initiate or not respond to social overtures
A3	absence of interest in peers	15. If you point at a toy across the room, does your child look at it? Yes No	does not interact with peers
B1	stereotyped motor movements	16. Does your child walk? Yes No	unable to walk (finding)
A3	absence of interest in peers	17. Does your child look at things you are looking at? Yes No	atypical child with whom to interact
B1	repetitive motor movements	18. Does your child make unusual finger movements near his/her face? Yes No	repetitive finger movements
A1	failure to initiate social interactions	19. Does your child try to attract your attention to his/her own activity? Yes No	does not play with others
E	intellectual developmental disorder	20. Have you ever wondered if your child is deaf? Yes No	disorder, blind-deaf
A1	social-emotional reciprocity	20. Have you ever wondered if your child is deaf? Yes No	social/emotional reciprocity appears deaf
A2	poorly integrated verbal communication	21. Does your child understand what people say? Yes No	speech difficult to understand
B4	hyporeactivity to	22. Does your child sometimes	loss of postural

	sensory input	stare at nothing or wander with no purpose? Yes No	tone
A2	lack of nonverbal communication	23. Does your child look at your face to check your reaction when faced with something	non-communicating child

Part B: Final Data Schema

SubjectId	Gender	SubjectMonthAge	language_label	path	label	Stage1 text	Stage2 text
1523050225	FEMALE	41	no	/data1/workspace/skbae/ASDDATA/original/IF2007_1_1_1523050225_1.wav	TD	inappropriate facial expressions behavior is peer group influenced lack of close friend is seeking social interactions some indication of spontaneous imitation that goes beyond copying a frequent use of an object spontaneously pointing head-nodding head shaking some evidence of reciprocal social smiling sharing social behavior body language in children range of facial expressions used to communicate response to approaches of other children usually looks up and pays attention when spoken to imaginative cooperative play with other children in spontaneous pretend activities subject does engage in interactive play with other children easily nan	Task Results: Success of Responded to name ,Success of Mimicked actions2 ,Success of Played catch ,Success of Fed baby doll ,Success of Reacted to snack Combined: restless behavior , , ,demonstrated behavior multiple patterns , , ,unusual behavior ,dependent on others , , ,difficulty taking turns in conversation , ,abnormal movements ,speech nonverbal communication skills facial expressions , ,avoided eye contact , , ,poor self-esteem (finding) , ,difficulty communicating thoughts ,unusual sensory interests , , ,seeks no play that involves participation in groups of other children , ,very focused and intense on single objects to the point of tuning everyone out , , ,not initiate social interactions ,repetitive stereotypical behavior ,child always different ,emotional outbursts ,perseverative thoughts , , , ,not

							attending to social approaches from adults ,lack of peer relationships , ,restricted range of interests , , ,sensory discomfort ,self-care behaviour ,not spontaneously engaging in other children , , , ,inappropriate laughing, joking, or punning , , , ,difficulty answering questions , ,speech tone ,uses people as objects , , ,unusual social interaction ,does not usually play with other children ,limited attention and focus , ,suspicious of others ,little or no emotions indication facial expression range , , ,unusual social interaction ,discomfort in social situations ,
--	--	--	--	--	--	--	---

Supplementary Tables and Figures

Supplementary Table 1. 3-class Classification Performance Metrics

Legend: Performance metrics (AUROC, mean, and 95% confidence intervals) for 3-class classification across different modalities. AUROC = Area Under the Receiver Operating Characteristic.

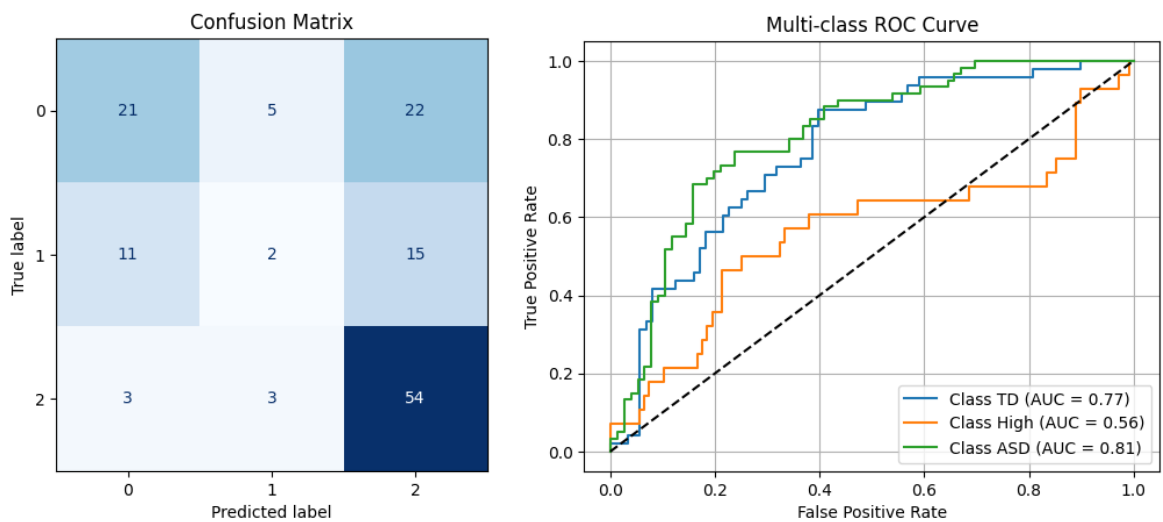
3-class classification

Model	Metric	Mean	CI Lower	CI Upper
Stage 1	AUROC	0.6875	0.62743	0.75196
Audio only				
Stage 1	AUROC	0.8733	0.8166	0.9206
Text only				
Stage 1	AUROC	0.8825	0.83365	0.9270
Audio+ Text				
Stage2	AUROC	0.8549	0.7680	0.9284
SRS+5 Tasks				

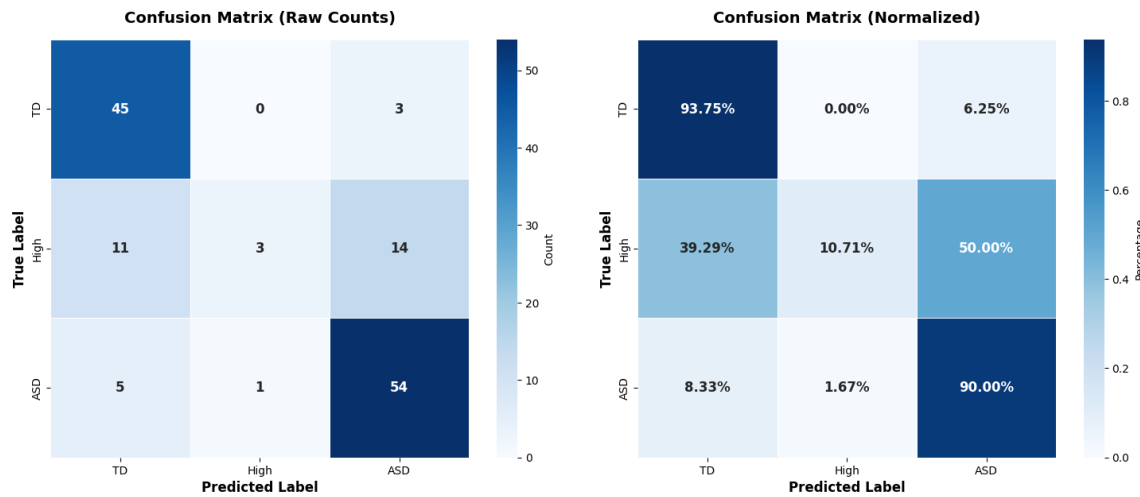
Supplementary Figure 1. Confusion Matrix & AUROC for 3-Class Classification

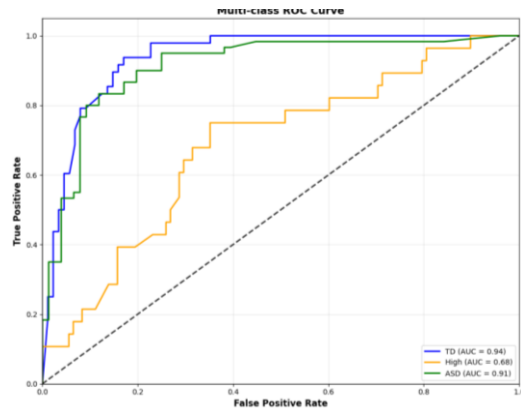
Confusion Matrix & AUROC

1) Stage 1 Audio only

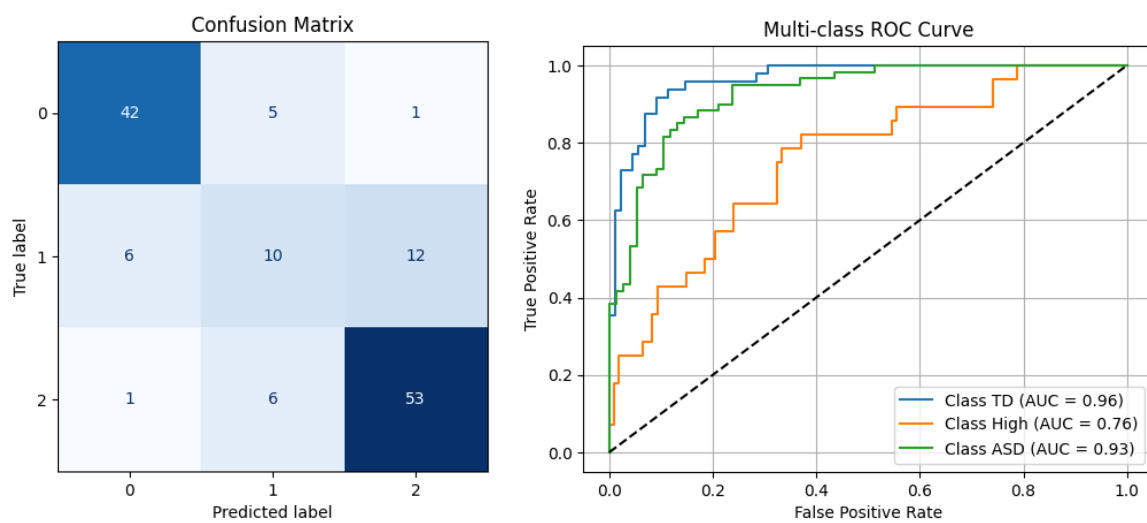


2) Stage 1 Text only

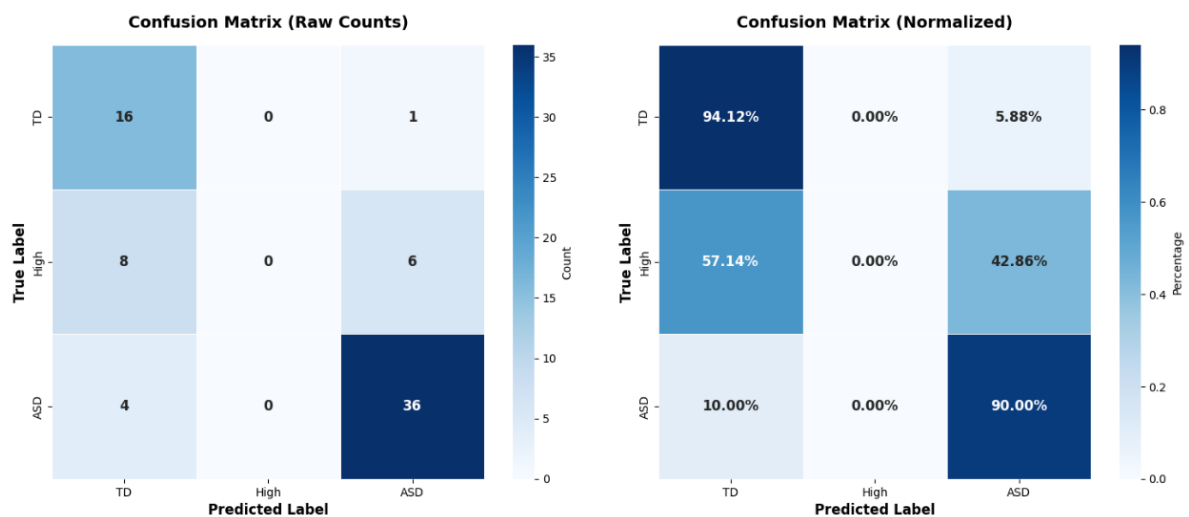


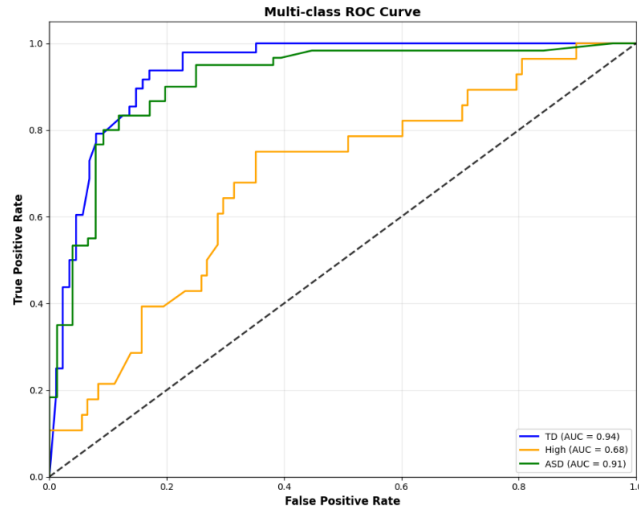


3) Audio text all label stage1



4) Stage2 SRS+5 Tasks





Supplementary Table 2. Stage 1 Classification Performance Metrics

Legend: Performance metrics (AUROC, mean, and 95% confidence intervals) for Stage 1 classification across different modalities. AUROC = Area Under the Receiver Operating Characteristic.

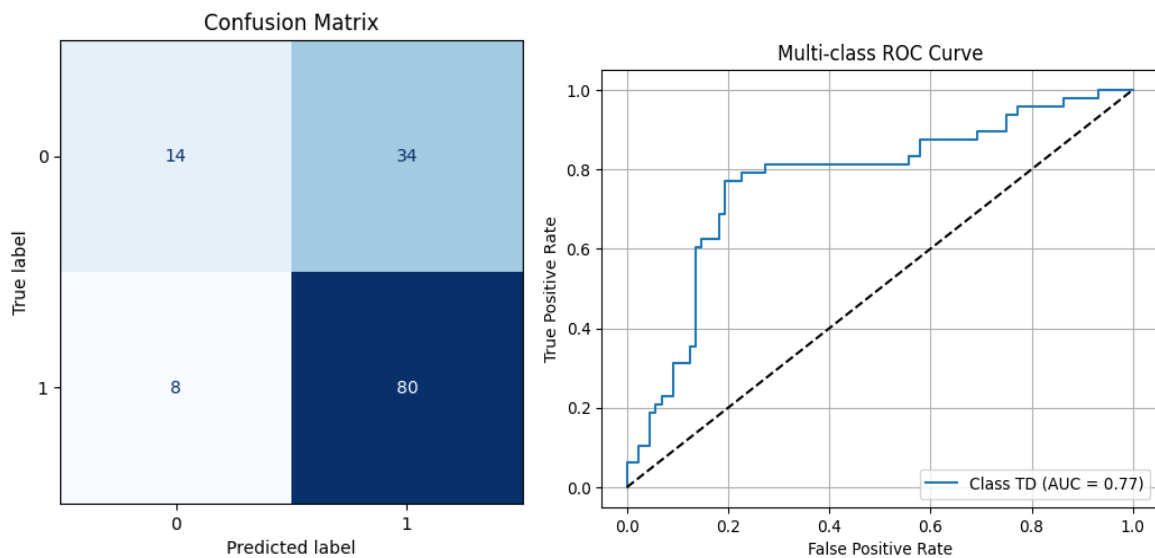
Stage 1 Classification

Model	Metric	Mean	CI Lower	CI Upper
Audio only	AUROC	0.7691	0.6758	0.8521
Text only	AUROC	0.9418	0.9020	0.9726
Audio+ Text with language delay (Stage1 model)	AUROC	0.9551	0.8247	0.9812
Audio+ Text w/o language delay	AUROC	0.9496	0.9143	0.9777

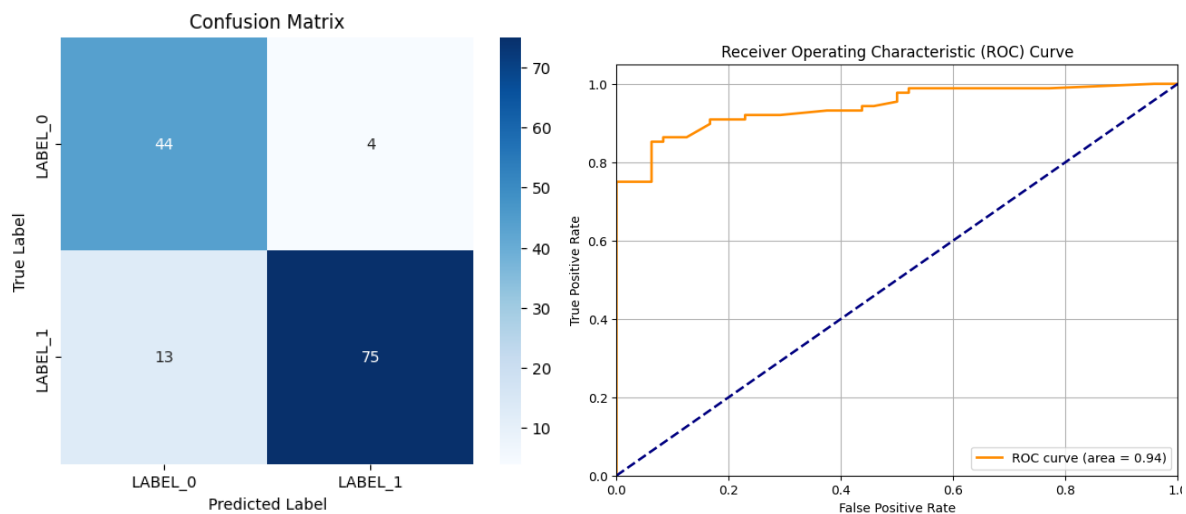
Supplementary Figure 2. Confusion Matrix & AUROC for Stage 1 Classification

Legend: Confusion matrix and AUROC for Stage 1 classification using audio and text modalities.

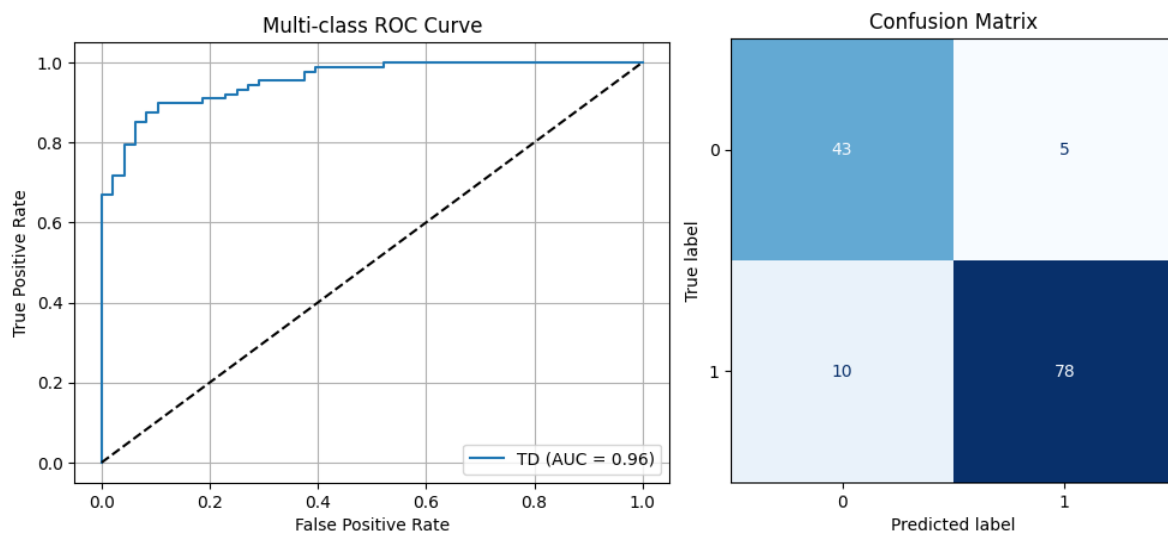
1) Stage1 Audio(parent-child interaction) only



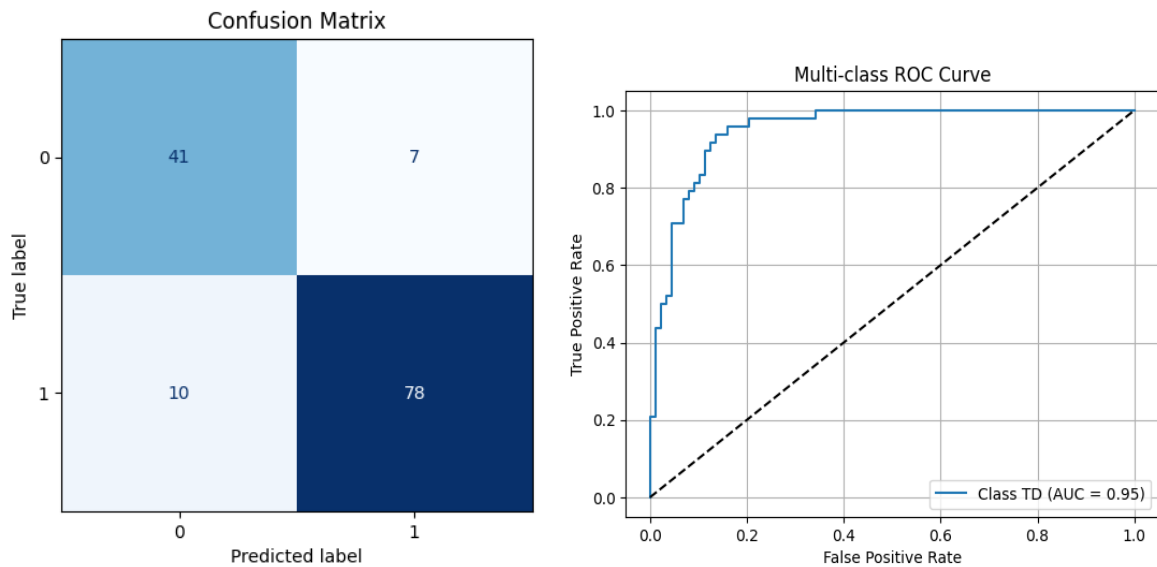
2) Stage1 Text(MCHAT+SCQ-L) only



3) Stage 1 Audio+ Text with language delay



4) Stage 1 Audio+ Text w/o language delay



Supplementary Table 3. Stage 2 Classification Performance Metrics

Legend: Performance metrics (AUROC, mean, and 95% confidence intervals) for Stage 2 classification models. AUROC = Area Under the Receiver Operating Characteristic; SRS = Social Responsiveness Scale; S/F = Success/Failure.

Stage 2 Classification

Model	Metric	Mean	CI Lower	CI Upper
Base SVM	AUROC	0.8263	0.6482	0.9696
SRS Only	AUROC	0.8831	0.7634	0.9697
5 Tasks S/F Only	AUROC	0.7607	0.6330	0.8750
SRS+5 Tasks	AUROC	0.9317	0.8585	0.9854

Supplementary Figure 3. Confusion Matrix & AUROC for Stage 2 Classification

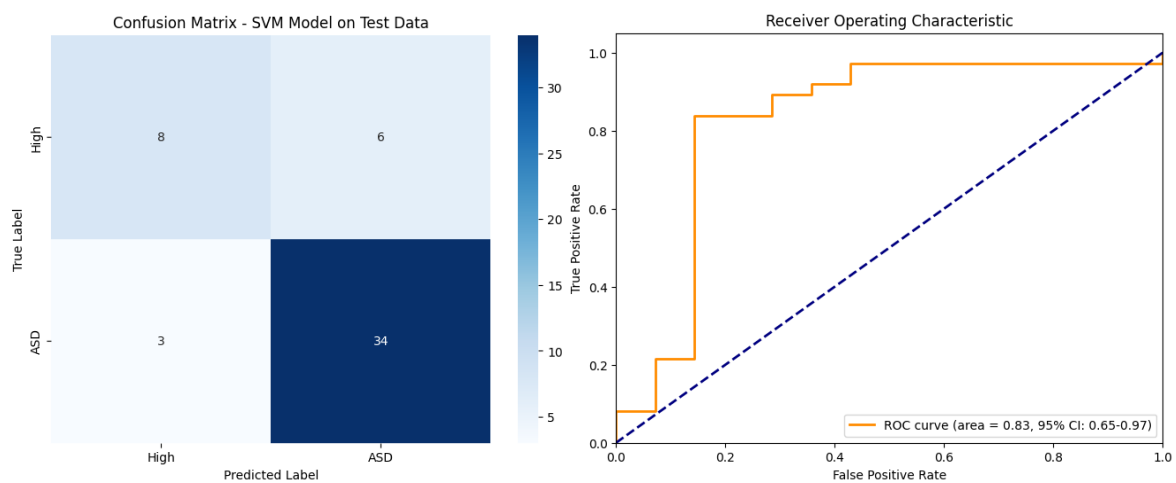
Abbreviations

AUROC: Area Under the Receiver Operating Characteristic

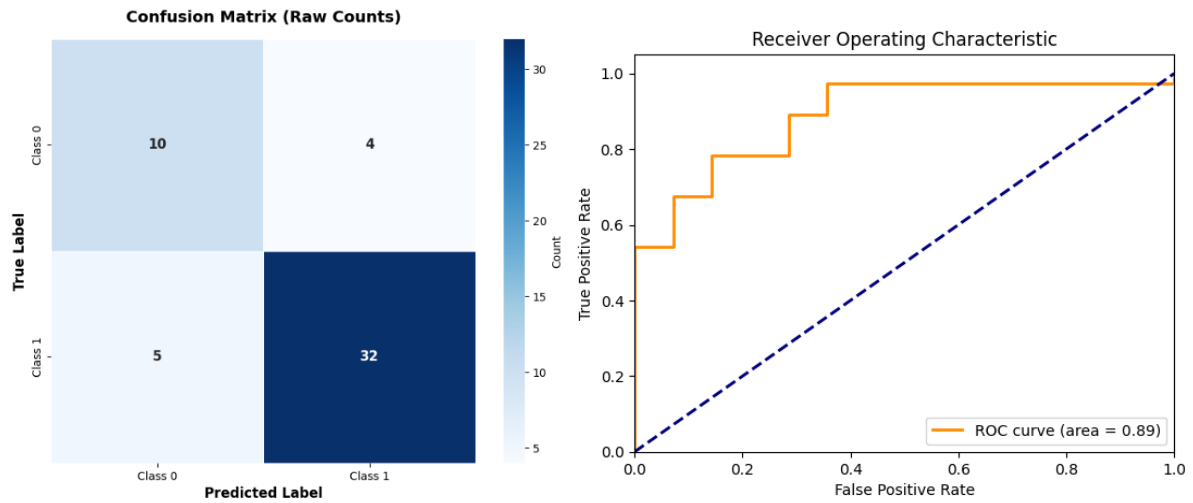
SRS: Social Responsiveness Scale

S/F: Success/Failure

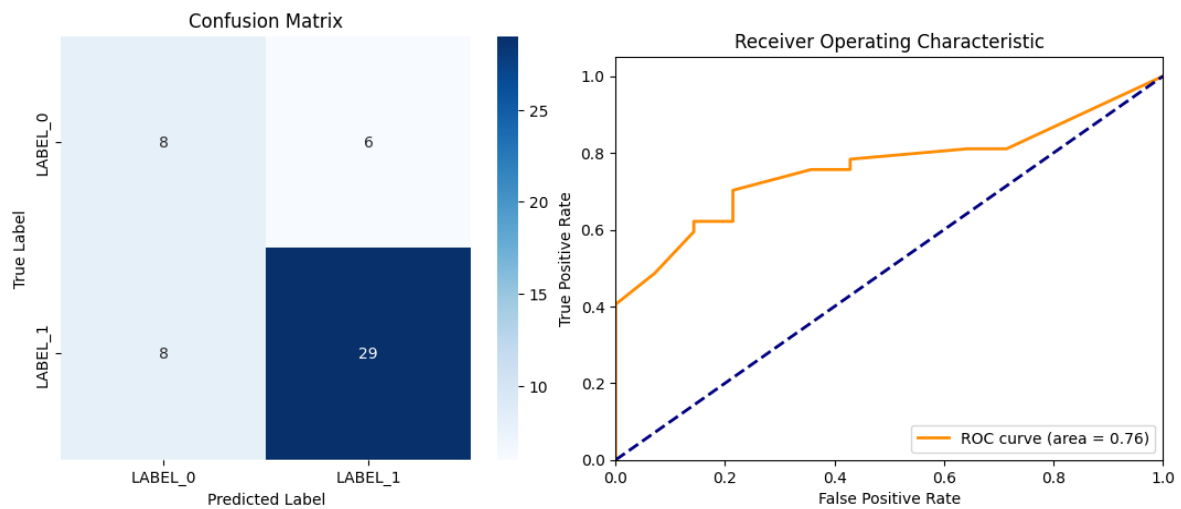
1) Stage2 Base SVM (Stage2 SRS+5 Tasks Success/Failure)



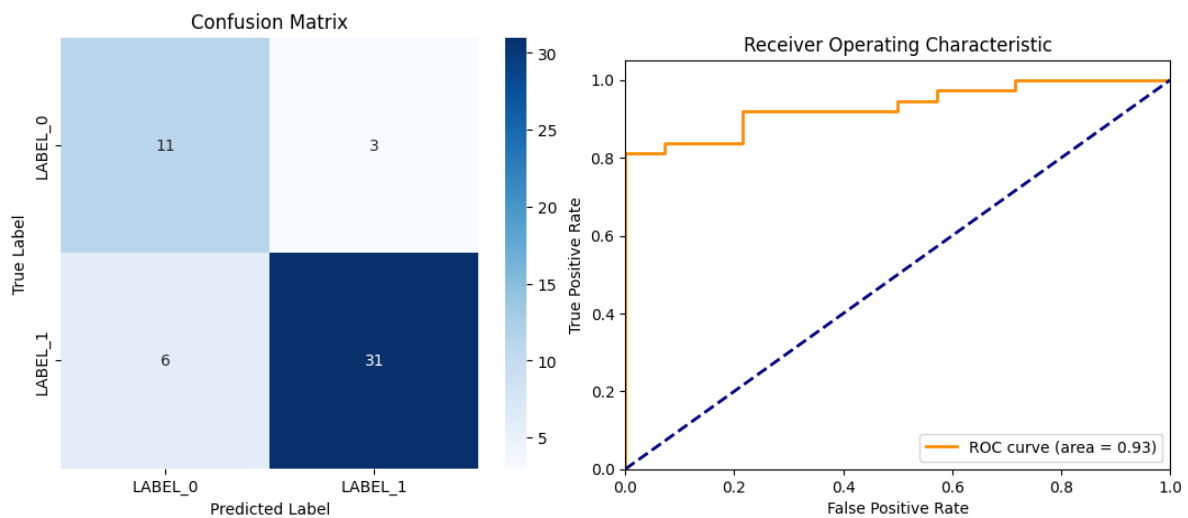
2) Stage2 SRS only



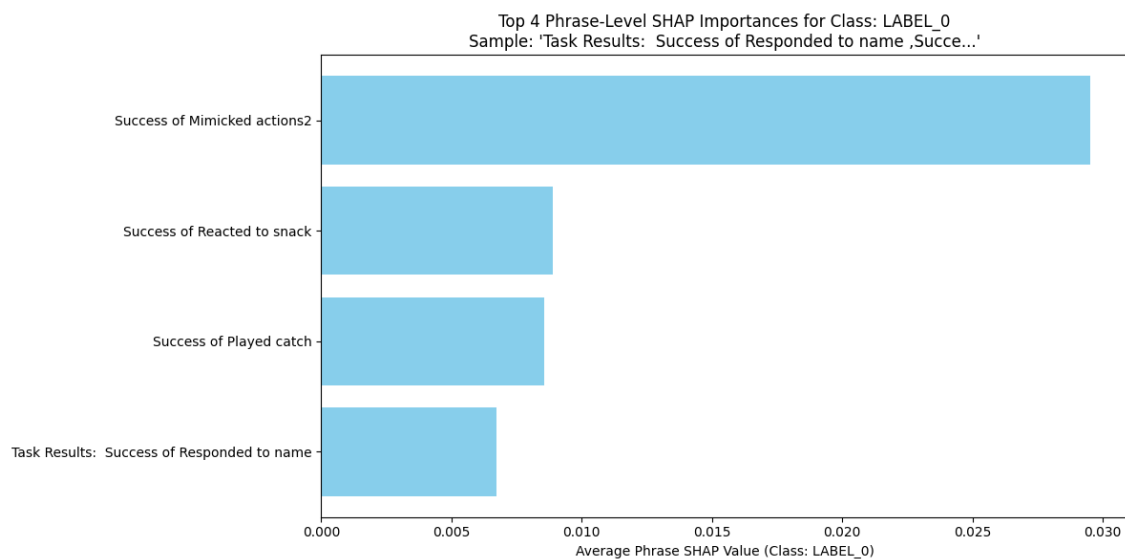
3) Stage2 5 Tasks S/F Only



4) Stage2 SRS+5 Tasks S/F



Supplementary Phrase-Level SHAP Importance and Phrase-Level CLS Attention

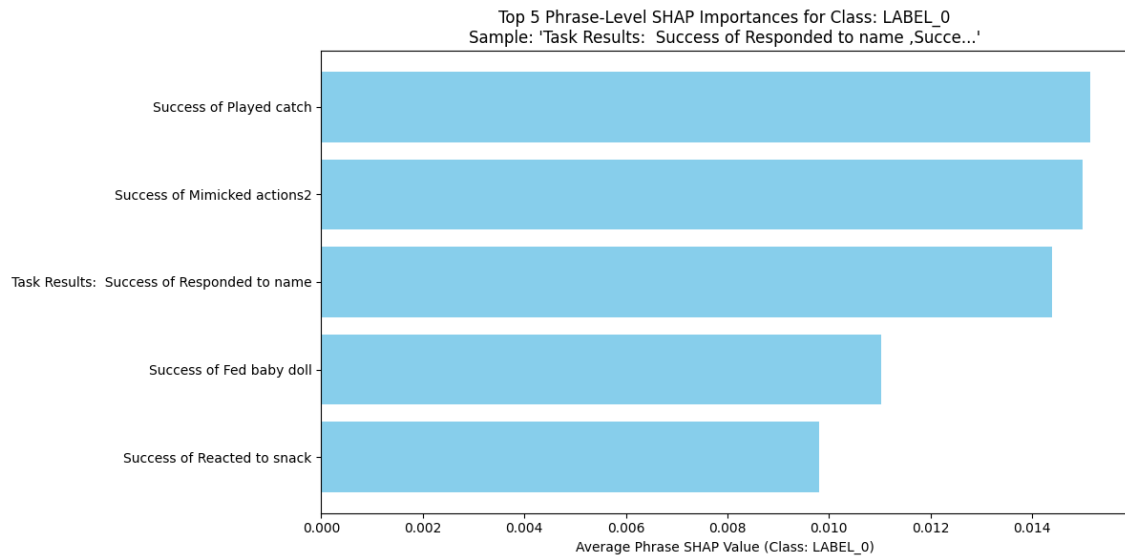


Supplementary Figure 4. Top 4 Phrase-Level SHAP Importances for CLASS_0 in the Sage2 Model with 5-Tasks Success/Failure only (Label 0: High)

Legend: This figure displays the aggregated SHAP (SHapley Additive exPlanations) values for CLASS_0 (High-Risk) in the Stage 2 model trained on five task success/failure outcomes. Positive SHAP values indicate features that push the model toward predicting High-Risk. The four most influential phrases are:

- Success of Mimicked actions2 (+0.030)
- Success of Reacted to snack (+0.009)
- Success of Played catch (+0.009)
- Task Results: Success of Responded to name (+0.007)

These task success indicators exhibit the largest positive contributions to the High-Risk prediction, demonstrating that correct performance on the "Mimicked actions2" task in particular is the strongest driver of CLASS_0 (High-Risk) assignments.

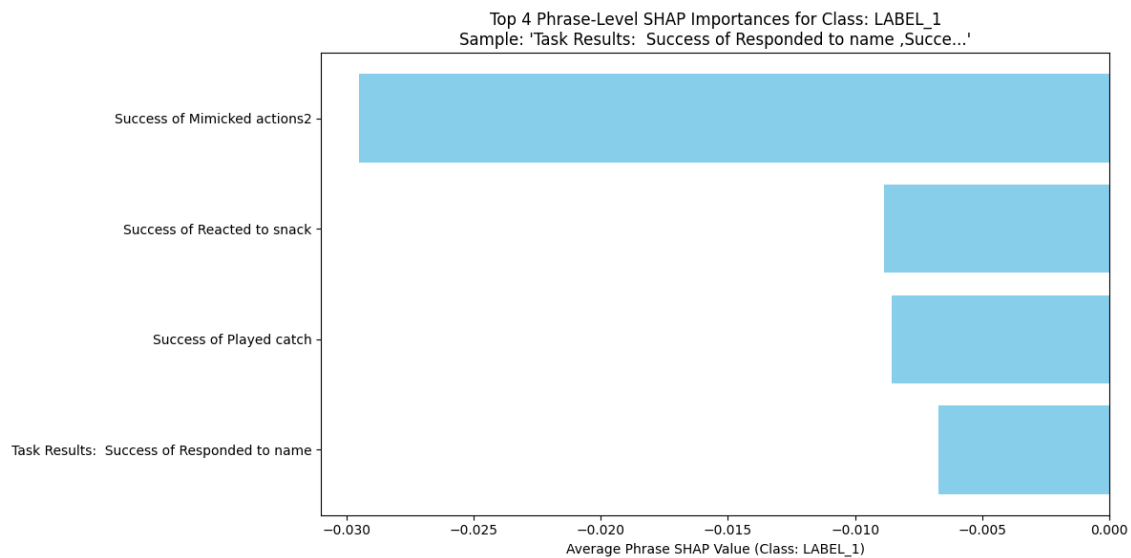


Supplementary Figure 5. Top 4 Phrase-Level SHAP Importances for CLASS_0 in the Sage2 Model with 5-Tasks Success/Failure only (Label 0: High)

Legend: This plot shows the four phrases with the highest average SHAP values for CLASS_0 (High-Risk) when the Stage 2 model is trained solely on the five task success/failure indicators. Positive SHAP values indicate features that push the model toward predicting High-Risk. The top phrases are:

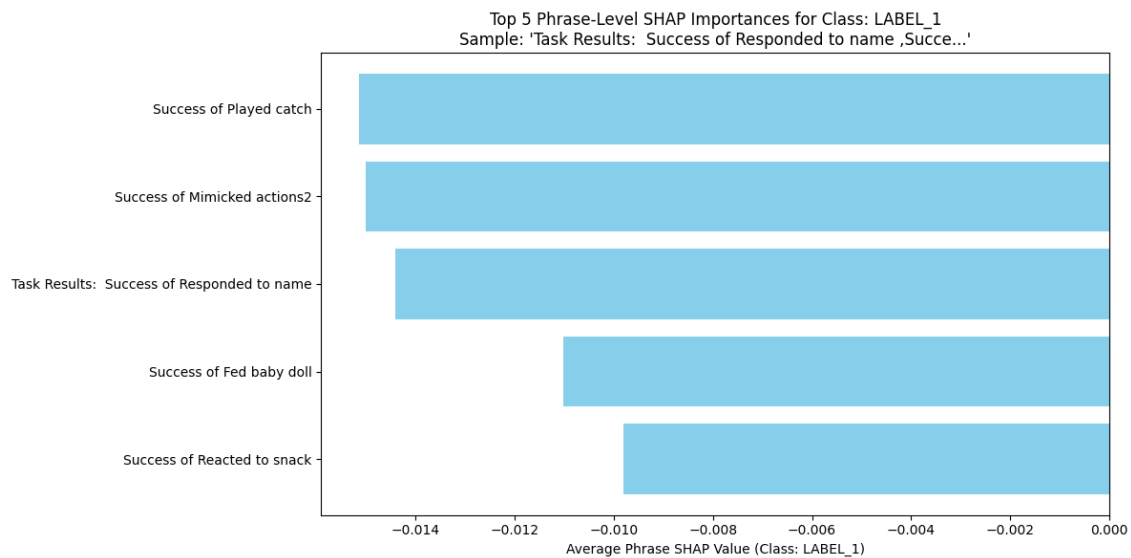
- "Success of Played catch" (≈ 0.015)
- "Success of Mimicked actions2" (≈ 0.015)
- "Task Results: Success of Responded to name" (≈ 0.014)
- "Success of Fed baby doll" (≈ 0.011)

Among these, "Success of Played catch" and "Success of Mimicked actions2" contribute most strongly to a High-Risk prediction, followed closely by "Success of Responded to name" and "Success of Fed baby doll."



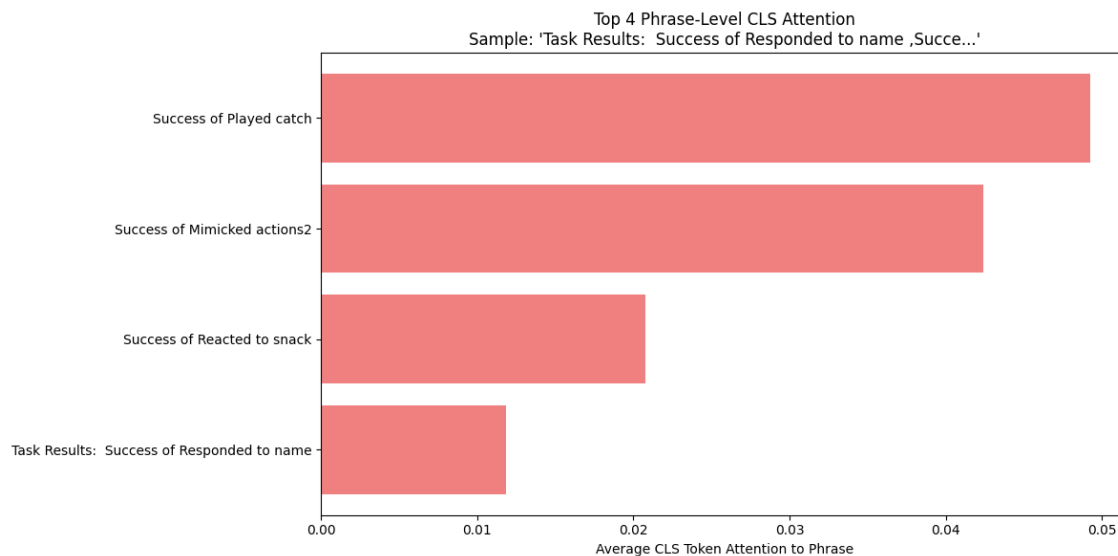
Supplementary Figure 6. Top 4 Phrase-Level SHAP Importances for CLASS_0 in the Sage2 Model with 5-Tasks Success/Failure only (Label 1: ASD)

Legend: This figure shows the aggregated SHAP values for CLASS_1 (ASD) in the Stage 2 model trained on five task success/failure results. Negative SHAP values for phrases like "Success of Mimicked actions2" (−0.030), "Success of Reacted to snack" (−0.010), "Success of Played catch" (−0.009), and "Task Results: Success of Responded to name" (−0.007) indicate that these task outcomes strongly decrease the likelihood of an ASD prediction. Such patterns highlight the inverse relationships between task success indicators and ASD classification.



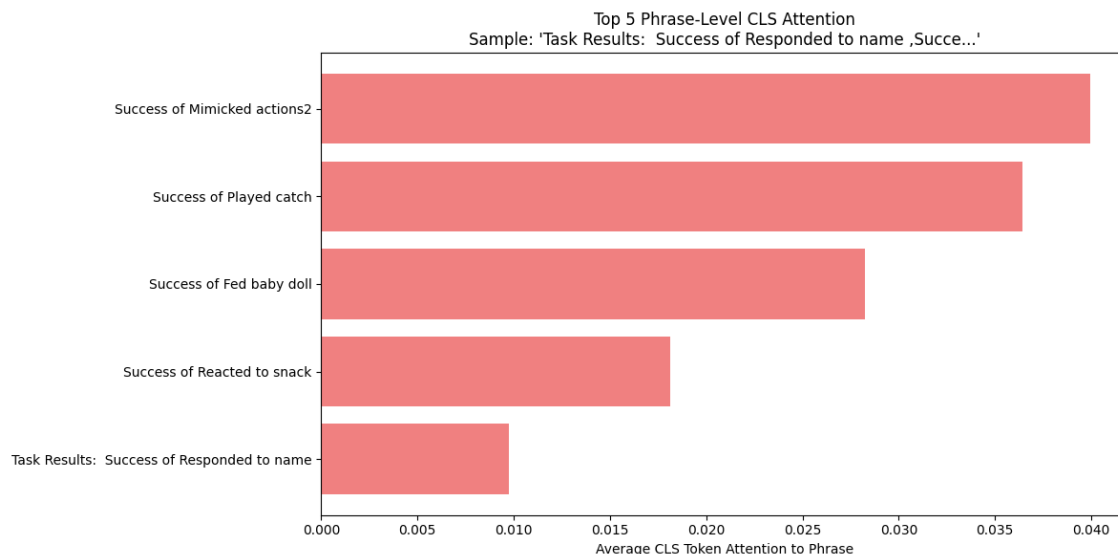
Supplementary Figure 7. Top 5 Phrase-Level SHAP Importances for CLASS_0 in the Sage2 Model with 5-Tasks Success/Failure only (Label 1: ASD)

Legend: This figure shows the aggregated SHAP values for CLASS_1 (ASD) in the Stage 2 model trained on five task success/failure results. Negative SHAP values for phrases such as "Success of Played catch" (-0.015), "Success of Mimicked actions2" (-0.015), "Task Results: Success of Responded to name" (-0.012), "Success of Fed baby doll" (-0.010), and "Success of Reacted to snack" (-0.009) indicate that successful performance on these tasks strongly decreases the likelihood of an ASD prediction. These patterns underscore the inverse relationship between task success indicators and ASD classification.



Supplementary Figure 8. Top 4 Phrase-Level CLS Attention for in the Sage2 Model with 5-Tasks Success/Failure only

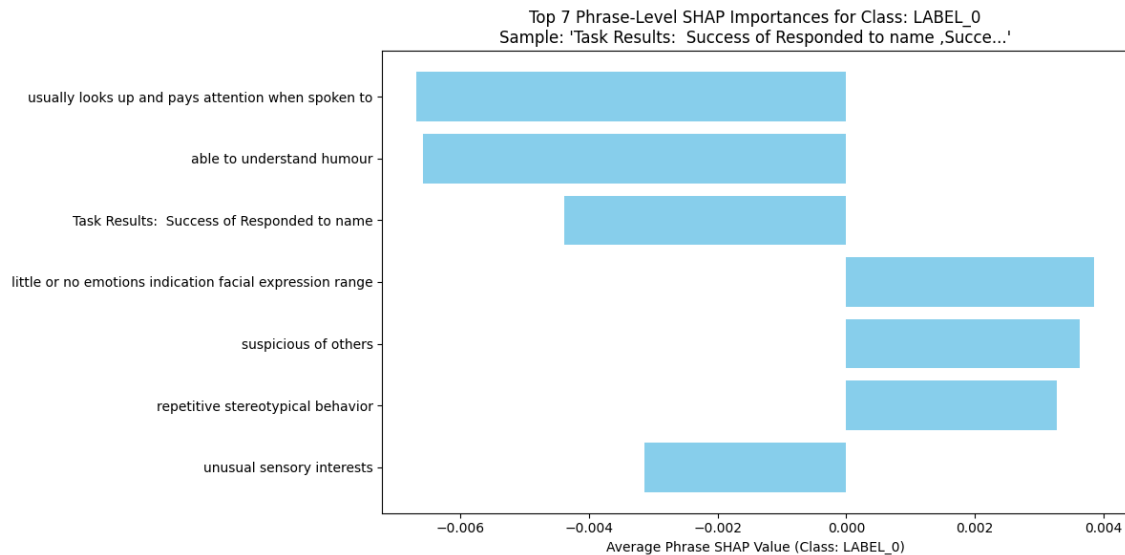
Legend: This figure shows the phrase-level [CLS] token attention weights in the Stage 2 model trained on five task success/failure results. Phrases reflecting direct task outcomes—such as “Success of Played catch” (≈ 0.050), “Success of Mimicked actions2” (≈ 0.045), “Success of Reacted to snack” (≈ 0.020), and “Task Results: Success of Responded to name” (≈ 0.012)—receive the highest attention scores. These high attention values confirm that the model’s global representation is driven primarily by explicit success indicators when distinguishing between high-risk and ASD outcomes.



Supplementary Figure 9. Top 5 Phrase-Level CLS Attention for in the Sage2 Model with 5-Tasks Success/Failure only

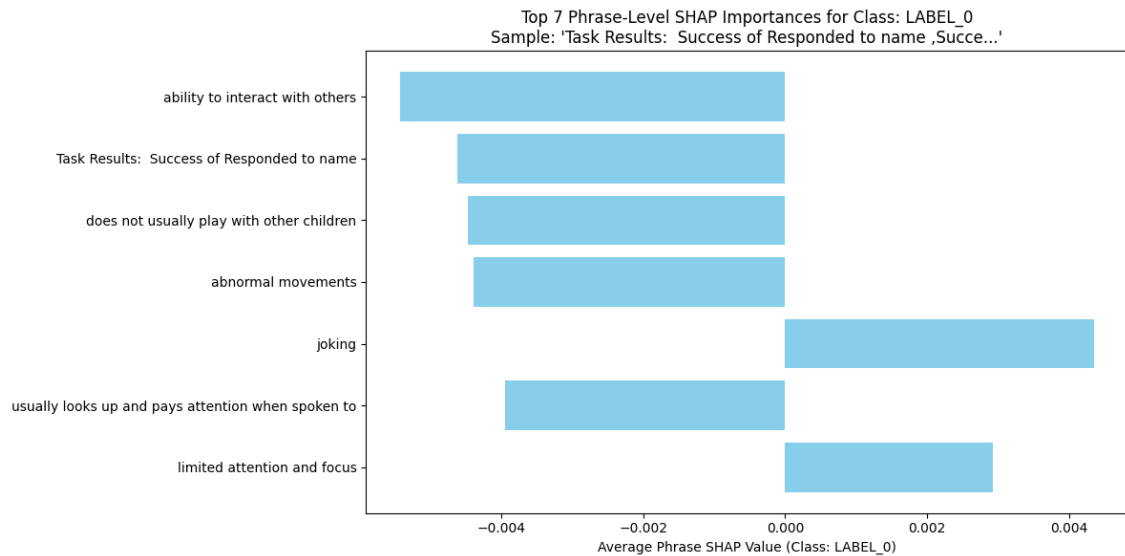
Legend: This figure shows the phrase-level [CLS] token attention weights in the Stage 2 model trained on five task success/failure results. High attention scores for phrases such as "Success of Mimicked actions2" (≈ 0.040), "Success of Played catch" (≈ 0.036), "Success of Fed baby doll" (≈ 0.030), "Success of Reacted to snack" (≈ 0.018), and "Task Results: Success of Responded to name" (≈ 0.010) indicate that the model's global representation is driven primarily by these explicit task-success indicators.

Sage2 Model with SRS + 5-Tasks Success/Failure



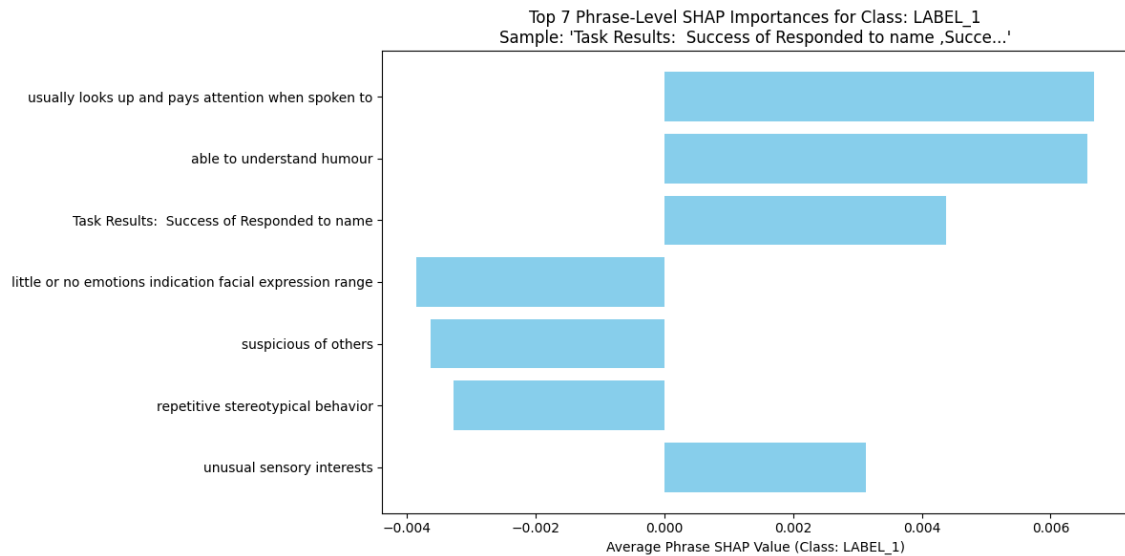
Supplementary Figure 10. Top 7 Phrase-Level SHAP Importances for High in the Sage2 Model with SRS + 5-Tasks Success/Failure (Label0: High)

Legend: This figure shows the aggregated SHAP values for CLASS_0 (High-Risk) in the Stage 2 model trained on both SRS items and five task success/failure results. Negative SHAP values for phrases such as "usually looks up and pays attention when spoken to" (-0.006), "able to understand humour" (-0.006), and "Task Results: Success of Responded to name" (-0.003) indicate that these features decrease the likelihood of a High-Risk classification. In contrast, positive SHAP values for phrases such as "little or no emotions indication facial expression range" ($+0.0035$), "suspicious of others" ($+0.0033$), "repetitive stereotypical behavior" ($+0.0030$), and "unusual sensory interests" (≈ 0.000) indicate that these behaviors increase the probability of predicting High-Risk. These patterns highlight that reduced social-emotional engagement and the presence of core ASD-related behaviors drive the model toward a High-Risk classification.



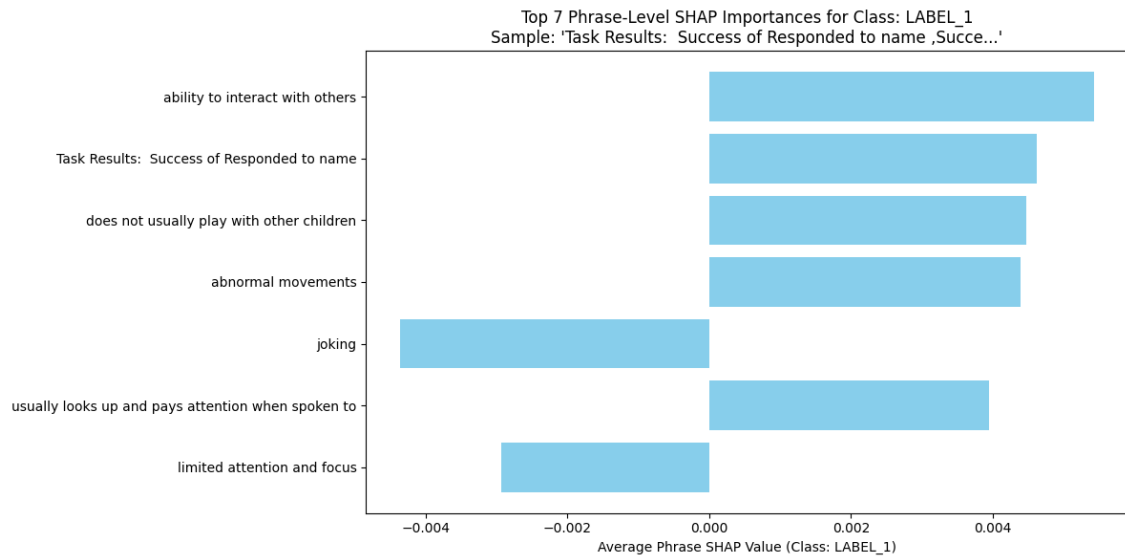
Supplementary Figure 11. Top 7 Phrase-Level SHAP Importances for High in the Sage2 Model with SRS + 5-Tasks Success/Failure (Label0: High)

Legend: This figure shows the aggregated SHAP values for CLASS_0 (High-Risk) in the Stage 2 model trained on both SRS items and five task success/failure results. Negative SHAP values for phrases such as "ability to interact with others" (-0.005), "Task Results: Success of Responded to name" (-0.003), "does not usually play with other children" (-0.003), "abnormal movements" (-0.003), and "usually looks up and pays attention when spoken to" (≈ -0.001) indicate that these behaviors decrease the likelihood of a High-Risk classification. In contrast, positive SHAP values for phrases like "joking" ($+0.004$) and "limited attention and focus" ($+0.003$) indicate that these features increase the probability of predicting High-Risk. These patterns highlight that reduced social engagement and the presence of certain atypical behaviors drive the model toward a High-Risk classification, whereas stronger interactive abilities and task successes pull the prediction away from High-Risk.



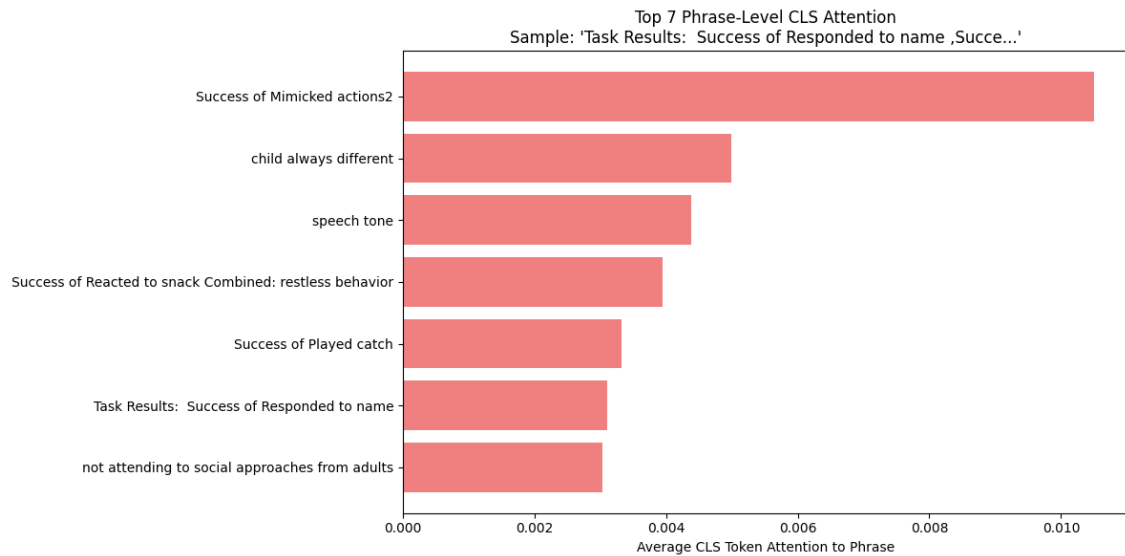
Supplementary Figure 12. Top 7 Phrase-Level SHAP Importances for High in the Sage2 Model with SRS + 5-Tasks Success/Failure (Label1: ASD)

Legend: This figure shows the aggregated SHAP values for CLASS_1 (ASD) in the Stage 2 model trained on SRS items combined with five task success/failure results. Positive SHAP values for phrases such as "usually looks up and pays attention when spoken to" (+0.007), "able to understand humour" (+0.007), "Task Results: Success of Responded to name" (+0.005), and "unusual sensory interests" (+0.003) indicate that these features increase the probability of an ASD prediction. Conversely, negative SHAP values for "little or no emotions indication facial expression range" (-0.003), "suspicious of others" (-0.003), and "repetitive stereotypical behavior" (-0.002) indicate that these behaviors decrease the likelihood of classifying a child as ASD. These patterns highlight how certain social-communication strengths and sensory atypicalities drive the model toward ASD, while emotional flatness, suspiciousness, and stereotyped behaviors push it away.



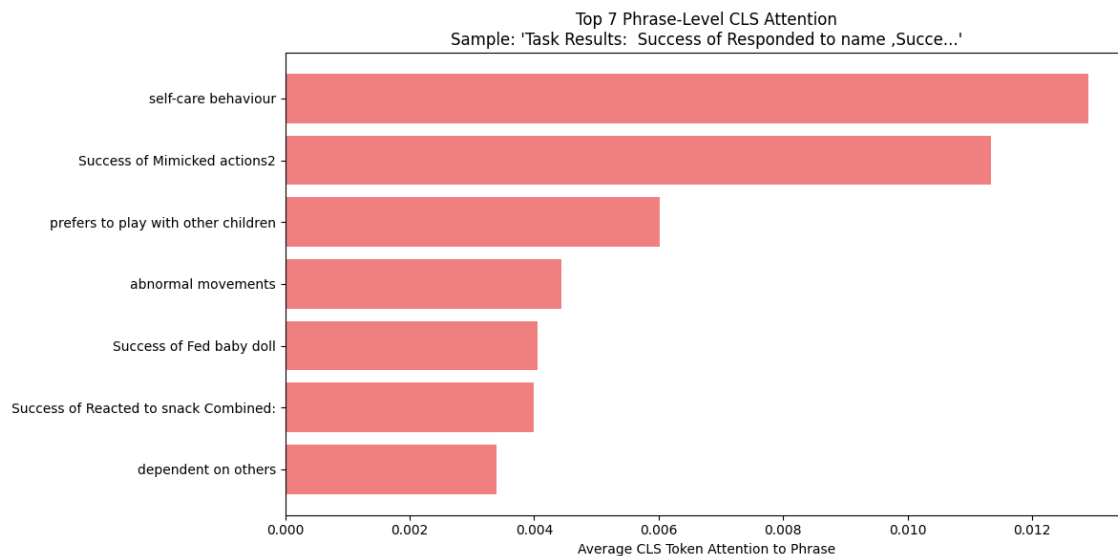
Supplementary Figure 13. Top 7 Phrase-Level SHAP Importances for High in the Sage2 Model with SRS + 5-Tasks Success/Failure (Label1: ASD)

Legend: This figure shows the aggregated SHAP values for CLASS_1 (ASD) in the Stage 2 model trained on SRS items combined with five task success/failure results. Positive SHAP values for phrases such as "ability to interact with others" ($\approx +0.005$), "Task Results: Success of Responded to name" ($\approx +0.005$), "does not usually play with other children" ($\approx +0.005$), "abnormal movements" ($\approx +0.005$), and "usually looks up and pays attention when spoken to" ($\approx +0.004$) indicate that these features increase the likelihood of an ASD prediction. In contrast, negative SHAP values for "limited attention and focus" (≈ -0.002) and "joking" (≈ -0.004) show that these traits decrease the probability of classifying a child as ASD.



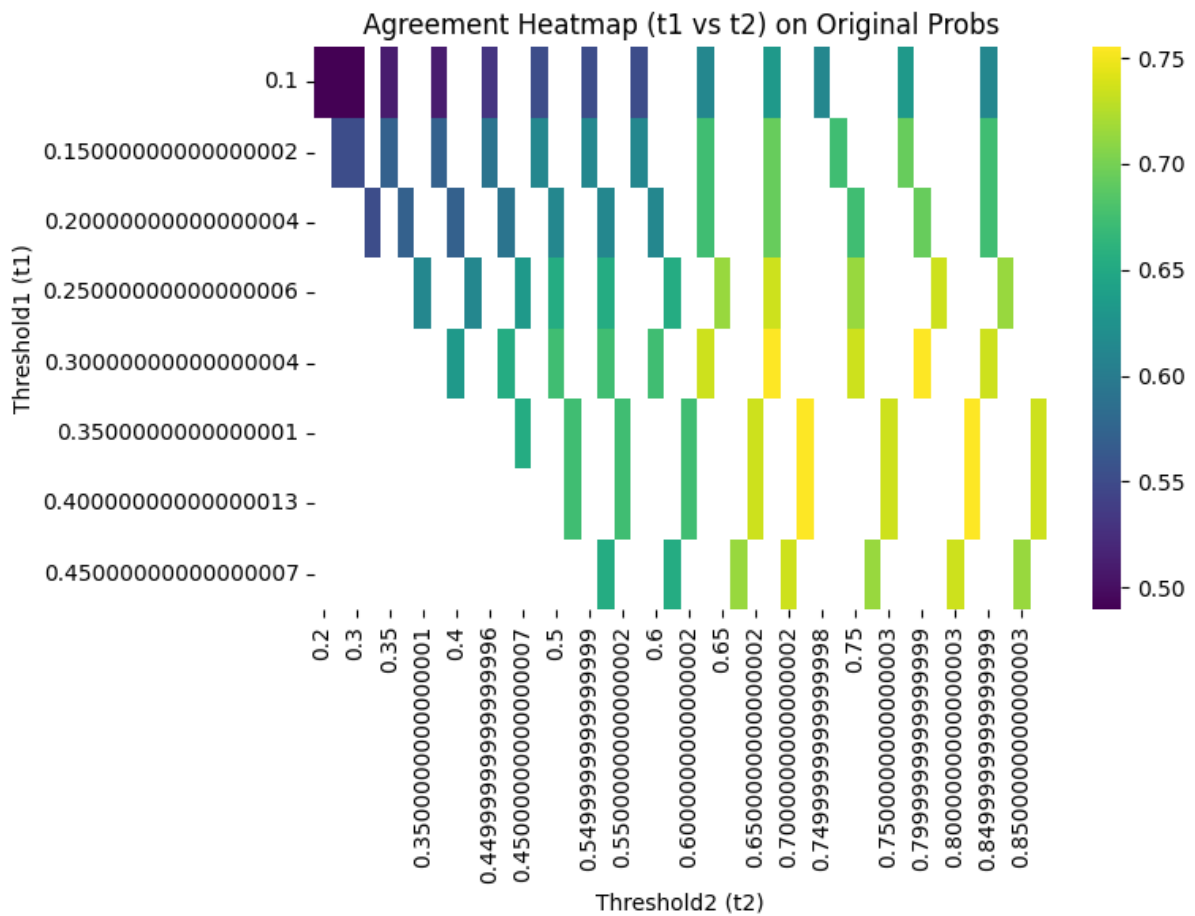
Supplementary Figure 14. Top 7 Phrase-Level CLS Attention for in the Sage2 Model with SRS + 5-Tasks Success/Failure

Legend: This figure shows the phrase-level [CLS] token attention weights in the Stage 2 model trained on both SRS items and five task success/failure results. High attention scores for phrases such as "Success of Mimicked actions2" (≈ 0.010), "child always different" (≈ 0.005), "speech tone" (≈ 0.0045), "Success of Reacted to snack Combined: restless behavior" (≈ 0.004), "Success of Played catch" (≈ 0.0035), "Task Results: Success of Responded to name" (≈ 0.003), and "not attending to social approaches from adults" (≈ 0.003) indicate that the model's global representation places greatest emphasis on these specific behavioral and task-outcome phrases. These attention patterns confirm that the model integrates both core ASD-related features (e.g., "child always different," "speech tone") and explicit task successes (e.g., "Mimicked actions2," "Played catch") when forming its overall encoding.

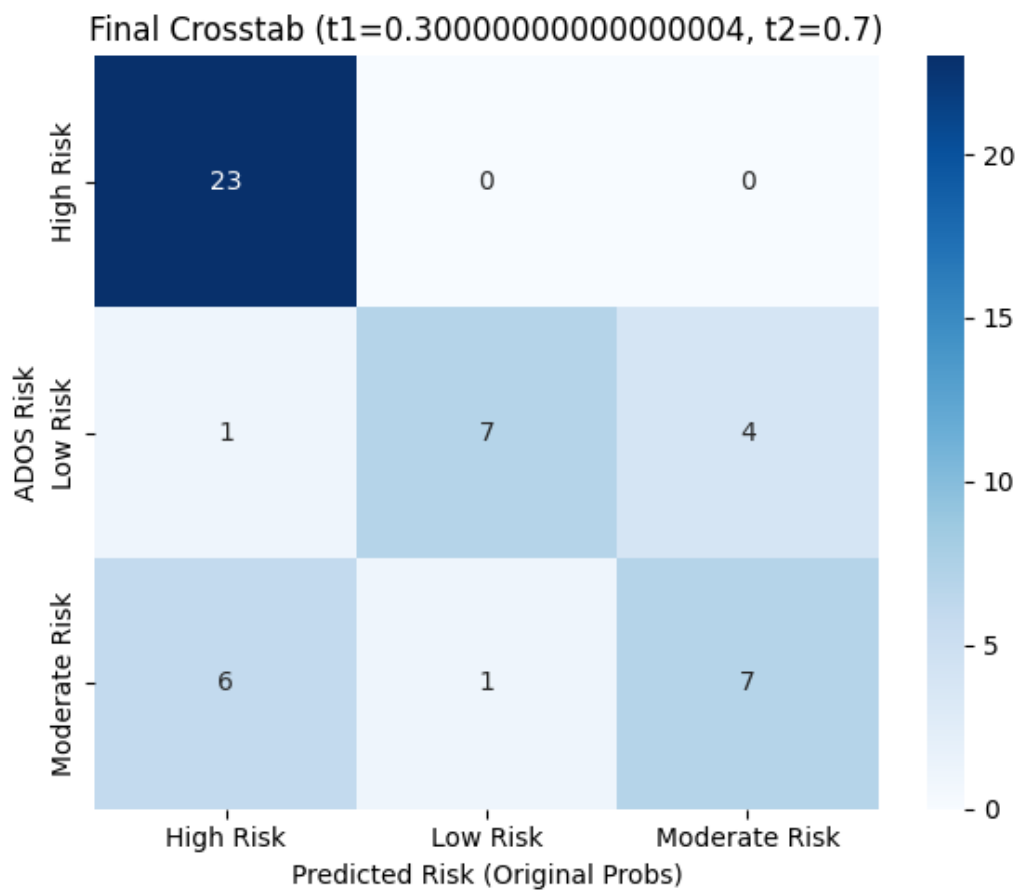


Supplementary Figure 15. Top 7 Phrase-Level CLS Attention for in the Sage2 Model with SRS + 5-Tasks Success/Failure

Legend: This figure shows the phrase-level [CLS] token attention weights in the Stage 2 model trained on SRS items combined with five task success/failure results. High attention scores for phrases such as "self-care behaviour" (≈ 0.013), "Success of Mimicked actions2" (≈ 0.011), "prefers to play with other children" (≈ 0.006), "abnormal movements" (≈ 0.0045), "Success of Fed baby doll" (≈ 0.004), "Success of Reacted to snack Combined" (≈ 0.004), and "dependent on others" (≈ 0.0035) indicate that the model's global representation places greatest emphasis on these specific behavioral and task-outcome features. These attention patterns confirm that the model integrates both core social-adaptive behaviors (e.g., "self-care behaviour," "prefers to play with other children," "dependent on others") and explicit task outcomes (e.g., "Mimicked actions2," "Fed baby doll," "Reacted to snack") when encoding the combined SRS and five-task inputs.

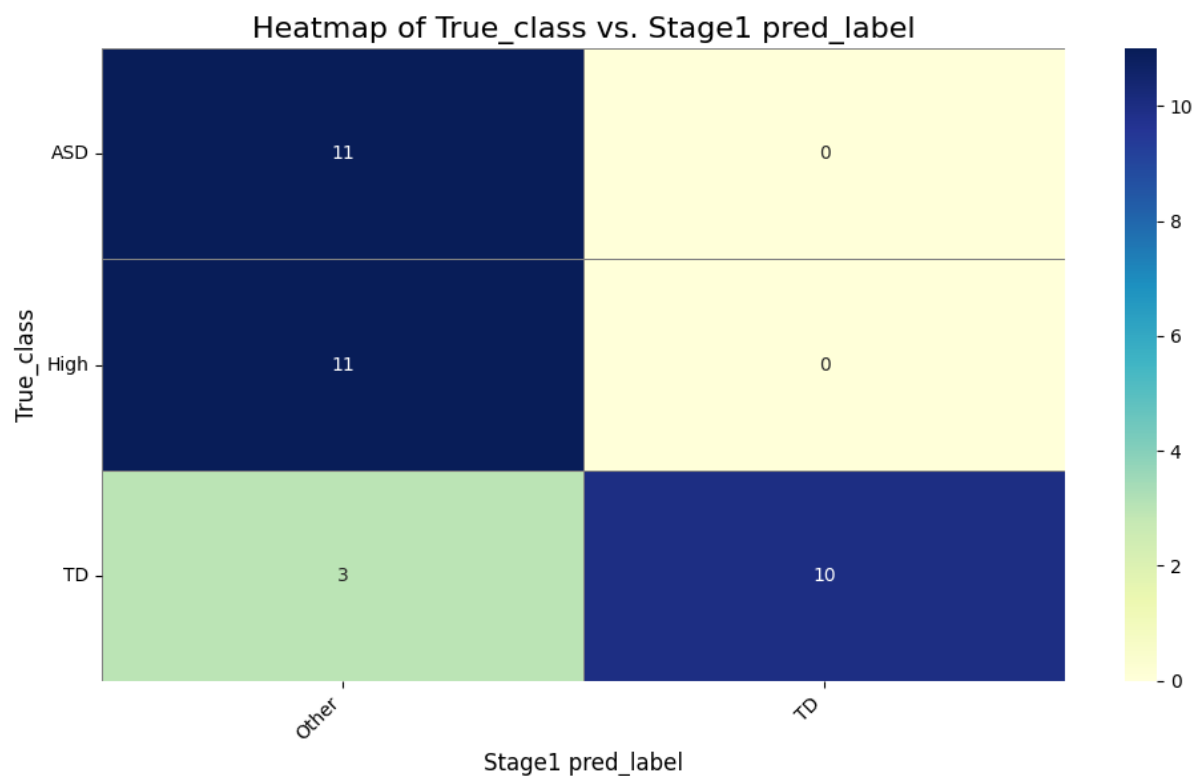


Supplementary Figure 16. Agreement between Model-Predicted and ADOS-Based Risk Categories (Uncalibrated Probabilities). This figure displays a cross-tabulation heatmap illustrating the agreement between the model's predicted risk categories (High Risk, Low Risk, Moderate Risk) and the ADOS-based risk categories (High Risk, ADOS Risk Low Risk, Moderate Risk) when using uncalibrated probabilities. The overall agreement demonstrated by the uncalibrated model is 0.6552, highlighting its initial performance before calibration



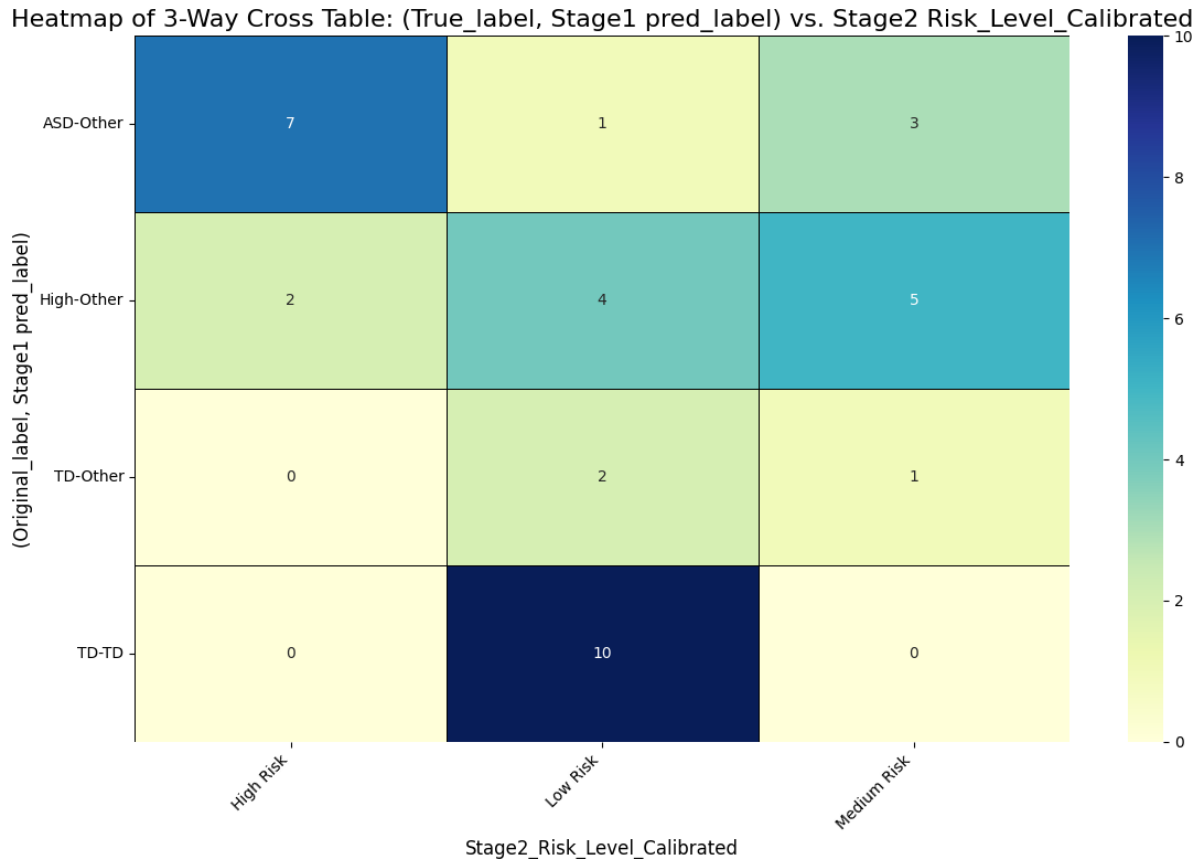
Supplementary Figure 17. Optimized Cross-tabulation Heatmap of ADOS Risk vs. Predicted Risk (Calibrated Probabilities). This heatmap displays the agreement between ADOS-2 defined risk categories (High Risk, ADOS Risk Low Risk, Moderate Risk) and the model's predicted risk categories (High Risk, Low Risk, Moderate Risk) at the optimized thresholds (**0.30** and **0.70**). The overall agreement at these thresholds is 0.7959, demonstrating the model's ability to stratify individuals into clinically relevant risk groups with calibrated probabilities

Independent, newly collected external cohort



Supplementary Figure 18. Heatmap of Ground-Truth Class versus Predicted Label in Stage 1 Classification

Description: This heatmap cross-tabulates each subject's true diagnostic class (True_class: ASD, High-risk, TD) against the binary model prediction (stage1 pred_label: Other = non-TD, TD). Cell color intensity and overlaid integer annotations denote the count of subjects in each true-vs-predicted pairing. In this Stage 1 classifier, all ASD (n = 11) and High-risk (n = 11) children were correctly identified as "Other," while among TD children (n = 13), 10 were correctly classified as TD and 3 were misclassified as Other. This illustrates the model's perfect sensitivity for non-TD cases with a small number of false positives in the TD group.



Supplementary Figure 19. Heatmap of Calibrated Risk Levels by True Class and Stage-1 Prediction

Description:

This heatmap shows how subjects' stage2 calibrated risk categories ("High Risk", "Low Risk", "Medium Risk") distribute across each combination of their true diagnostic class (True_class) and the Stage 1 binary prediction (stage1 pred_label: Other = non-TD vs. TD). Each row corresponds to a specific (True Class – Predicted Label) pairing:

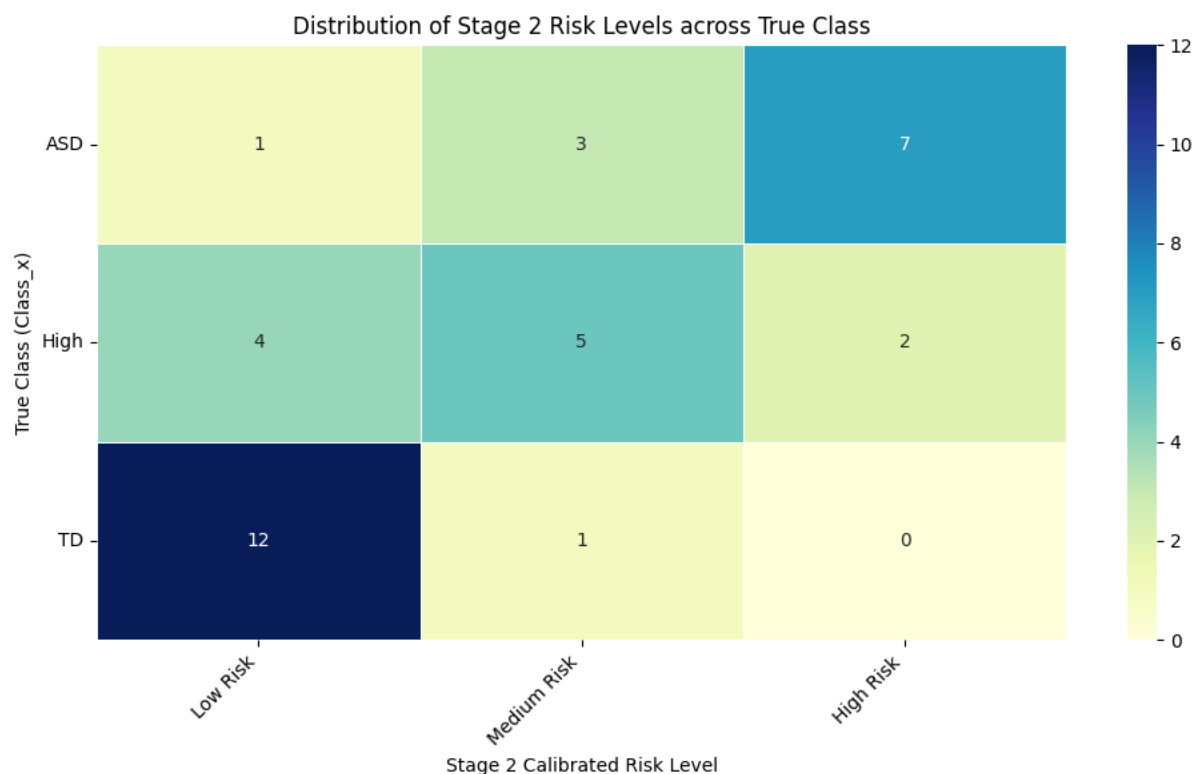
- **ASD – Other** (true ASD cases flagged as non-TD)
- **High – Other** (true High-risk cases flagged as non-TD)
- **TD – Other** (true TD cases misclassified as non-TD)
- **TD – TD** (true TD cases correctly classified as TD)

Cell color intensity and overlaid counts indicate the number of subjects falling into each calibrated risk bucket. Notable patterns:

- Among **ASD–Other**, the majority (7/11) were assigned High Risk, with smaller numbers in Low (1) and Medium (3) Risk.

- **High–Other** subjects show a more even spread across High (2), Low (4), and Medium (5) Risk.
- Misclassified TD cases (**TD–Other**) tend toward lower risk (2 Low, 1 Medium), while correctly classified TDs (**TD–TD**) are exclusively in the Low Risk category (10/10).

This visualization highlights that the calibration step generally places non-TD predictions into higher risk levels, whereas correctly identified TDs are consistently low risk, with only a few false-positive TDs receiving elevated risk scores

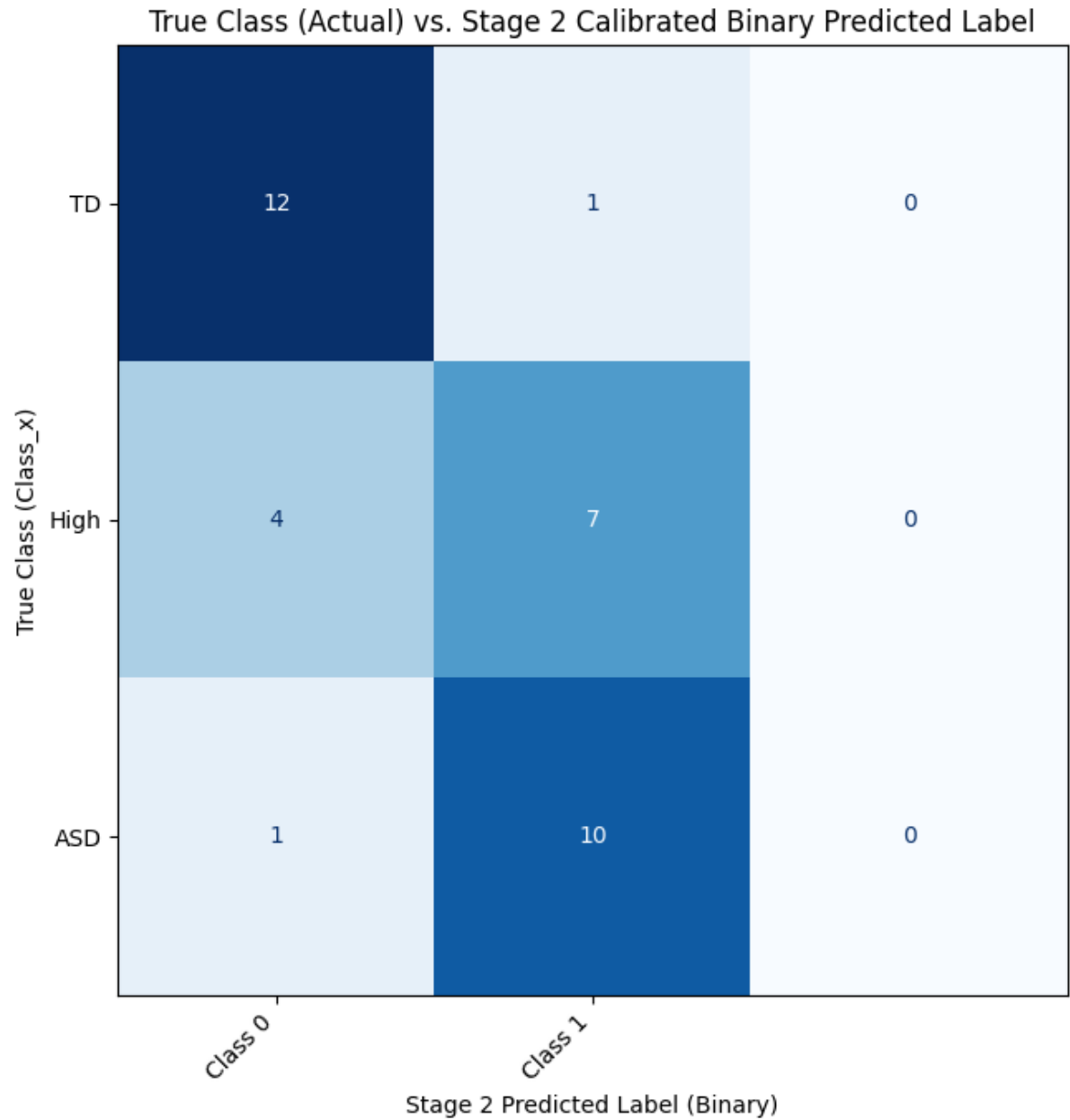


Supplementary Figure 20. Heatmap of Stage-2 Calibrated Risk Levels by True Diagnostic Class

Description:

This heatmap illustrates how the Stage 2 classifier's three calibrated risk levels ("Low Risk", "Medium Risk", "High Risk") distribute across the true diagnostic categories (True_class: ASD, High-risk, TD). Each row corresponds to one true class, and each

column to a calibrated risk bucket, with cell color intensity and overlaid counts indicating the number of subjects. Notably, most ASD cases (7/11) are assigned High Risk, with the remainder split between Medium (3) and Low (1) Risk. The High-risk group shows a more even spread—4 Low, 5 Medium, and 2 High Risk—reflecting residual uncertainty in this intermediate cohort. True TD subjects are almost exclusively classified as Low Risk (12/13), with only one Medium-Risk misclassification and no High-Risk flags. This pattern confirms that Stage 2 calibration robustly elevates ASD cases to higher risk, maintains a predominantly low-risk score for TD, and appropriately stratifies the High-risk group.



Supplementary Figure 21. Confusion Matrix of True Diagnostic Class versus Stage-2 Calibrated Binary Prediction

Description:

This heatmap displays the counts of subjects by their true diagnostic class (True_class: TD, High-risk, ASD) against the binary Stage 2 prediction label (Class 0 = non-ASD, Class 1 = ASD), evaluated on the independent external cohort. Cell color intensity and overlaid integers indicate frequency. True TD children (n = 13) were almost entirely recognized as non-ASD (12/13), with one false-positive ASD call. Among true High-risk subjects (n = 11), 7 were predicted ASD and 4 non-ASD. True ASD cases (n = 11) were correctly flagged in 10 instances, with a single miss. This matrix highlights the Stage 2 model's strong sensitivity for ASD and its high specificity against TD, while showing residual misclassification within the intermediate high-risk group.