Leveraging Home Video Analysis with Large Language Models to Identify Developmental Performance
Insights Beyond Standardized Assessments in Rare Disease Trials

Anzalee Khan[1,2]; Sheraz Hussain[3]; James Lefkowitz[4]; Stacey Eckert[3]; Mary Seddo[3]; Christian Yavorsky[3]

Affiliations: 1. Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY; 2. Manhattan Psychiatric Center, NY, NY; 3. Valis Biosciences, Inc., New York, NY; 4. Wandering Monk Productions

METHODOLOGICAL QUESTION

To what extent can large language model–based analysis of ecologically valid home video data extract quantifiable developmental indicators, and how do these LLM-derived behavioral features statistically correspond with, diverge from, or augment standardized norm-referenced developmental assessment scores? Additionally, can integration of these multimodal data sources improve sensitivity to real-world functional abilities over time?

BACKGROUND

- Standardized developmental assessments (e.g., Vineland-3, Bayley-4, CGI) remain the gold standard for measuring developmental change, yet they are conducted in structured clinical settings and may underrepresent context-dependent, emerging, or environmentally scaffolded skills.
- Ecologically valid behavioral sampling through home video captures real-world functioning, including spontaneous social engagement, naturalistic communication, adaptive problem-solving, and motor behaviors that may not be elicited during clinic-based testing.
- Large Language Models (LLMs) provide scalable tools for structured extraction of behavioral features from unstructured narrative data, enabling quantification of subtle developmental signals across domains (motor, language, cognitive, social-emotional, sensory).
- Integrating clinician-rated standardized measures with multimodal, naturalistic data streams represents a next-generation measurement approach, potentially improving sensitivity to functional change and enhancing longitudinal developmental monitoring in clinical trials.

STUDY DESIGN

Study Design: Retrospective multimodal secondary analysis of participants enrolled in two pediatric interventional clinical trials in rare and orphan indications that included standardized developmental assessments, Caregiver-recorded home videos

Analyses focused on baseline and post-intervention timepoints

Participants: N = 41 participants with:

Inclusion for this analysis required:

- Complete Vineland-3 and BSID-4 Growth Scale Values (GSV)
- Sufficient video quality for behavioral interpretation

Home Video Data Acquisition

Naturalistic, caregiver-recorded videos

Unstructured home environments. Activities included:

- Play
- Caregiver-child interaction
- Task completion
- Communication attempts
- Motor exploration

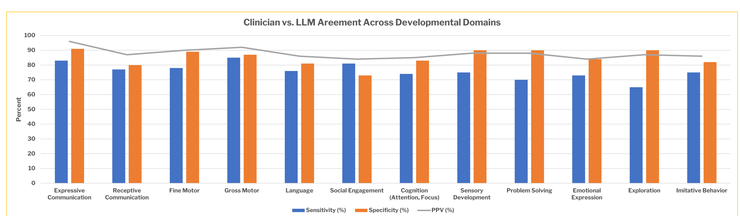
Videos ranged in length and context, representing ecologically valid behavioral sampling.

BEHAVIORAL EXTRACTION AND MAPPING

Step	Process	Description	Output
Step 1	Narrative Generation	Home videos were reviewed and summarized into structured narrative descriptions capturing: Observable behaviors • Contextual cues • Functional performance indicators • Social reciprocity • Motor coordination• Language attempts • Sensory responses	Structured behavioral narrative summaries
Step 2	Clinician (Human) Review	Independent clinician review of home videos to identify and document observable developmental skills across domains, including Expressive and receptive communication behaviors • Gross and fine motor skills • Social engagement and reciprocity • Task persistence and cognitive engagement • Sensory responses • Clinician notes were used for comparison with LLM-derived outputs to assess concordance and interpretability.	Clinician-coded behavioral observations
Step 3	LLM Behavioral Structuring	A large language model was used to: ▪ Extract domain-specific behavioral indicators from the home videos Identify developmental constructs (e.g., imitation, cause–effect reasoning, joint attention) ▪ Categorize behaviors into standardized developmental domains: Motor, Language (Expressive/Receptive), Cognitive, Social-Emotional, Sensory	Domain-structured behavioral feature set
Step 4	Domain Mapping	Qualitative video-derived observations (LLM + clinician review) were mapped to corresponding standardized domains:Vineland-3, BSID-4. This enabled directional comparison (improvement / decline / mixed) across modalities.	Cross-modal domain alignment dataset for concordance analysis

We employed an integrated, multimodal analytic framework to evaluate developmental trajectories across communication, motor, cognitive, and social domains for home videos collected within a secure video-audio integrated LLM developmental analysis system. Concordance between standardized measures (Vineland-3 and BSID-4 GSV change scores) was assessed using Pearson correlation coefficients ($\alpha = 0.05$). Domain-level agreement between clinician developmental notes and LLM-derived behavioral indicators was analyzed using frequency-based agreement coding, reporting percent agreement and divergence. For domain-level pattern analysis, descriptive statistics summarized mean GSV change, direction of change, and alignment with video-derived signals across expressive communication, receptive communication, fine and gross motor skills, and socialization behaviors.

CLINICIAN VS. LLM AGREEMENT ACROSS DOMAINS



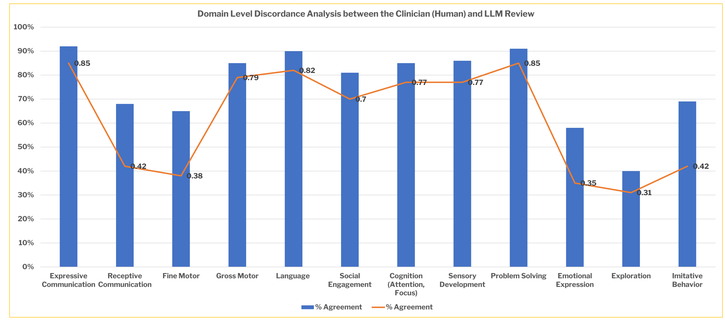
Domain	LLM: Present / Clinician: Present (True Positive)	LLM: Present / Clinician: Absent (False Positive)	LLM: Absent / Clinician: Present (False Negative)	LLM: Absent / Clinician: Absent (True Negative)	Total Clinician Present	Total Clinician Absent	Total Observations	Sensitivity (%)	Specificity (%)	PPV (%)
Expressive Communication	25	1	5	10	30	11	41	83	91	96
Receptive Communication	20	3	6	12	26	15	41	77	80	87
Fine Motor	18	2	5	16	23	18	41	78	89	90
Gross Motor	22	2	4	13	26	15	41	85	87	92
Language	19	3	6	13	25	16	41	76	81	86
Social Engagement	21	4	5	11	26	15	41	81	73	84
Cognition (Attention, Focus)	17	3	6	15	23	18	41	74	83	85
Sensory Development	15	2	5	19	20	21	41	75	90	88
Problem Solving	14	2	6	19	20	21	41	70	90	88
Emotional Expression	16	3	6	16	22	19	41	73	84	84
Exploration	13	2	7	19	20	21	41	65	90	87
Imitative Behavior	18	3	6	14	24	17	41	75	82	86

CORRELATION AND DOMAIN DISCORDANCE

Correlation Between Structured Developmental Measures

Comparison	N	Pearson r	p-value	Interpretation
Vineland-3 GSV Change vs. BSID-4 GSV Change	41	0.95	< 0.05	Strong cross-measure concordance

Domain Level Discordance Analysis Between the Clinician (Human) and LLM



Overall, clinician-LLM concordance was strongest in observable, behaviorally concrete domains such as expressive communication, language, problem solving, gross motor, and attention, all demonstrating substantial to near-perfect agreement. Lower agreement emerged in more interpretive or context-dependent domains, particularly emotional expression, exploration, receptive communication, and fine motor, suggesting greater variability in how nuanced socio-emotional and emerging skills are identified across modalities.

RESULTS: RPI CORRELATIONS

Domain	Vineland-3 GSV Change	BSID-4 GSV Change	Direction (Standardized)	Video-Derived Observations of Improvement	Alignment between video observations and clinician-assessment
Expressive Communication	17	1	Improvement	Vocalizations, word approximations, social engagement	Concordant
Receptive Communication	-20	-22	Decline	Following instructions, emerging receptive behaviors	Discordant
Gross Motor	5	1	No change	Walking, climbing, coordinated play	Discordant
Fine Motor	-5	-6	Decline	Grasping, stacking, object manipulation	Discordant
Social Interaction	10	N/A	Improvement	Social smiling, engagement	Concordant

Across developmental domains, standardized measures and video-derived observations demonstrated variable alignment. Expressive communication and social interaction showed improvements on standardized scales that were **concordant** with video-observed behaviors. In contrast, receptive communication, fine motor and Gross Motor domains showed **discordant** video observations, indicating that subtle motor behaviors and the participant's expressive language may not be fully captured by standard instruments.

CONCLUSIONS

- **Domain-Specific Concordance:** LLM-based analysis of home videos demonstrated strongest agreement with clinicians in concrete, observable domains (i.e., expressive communication, problem solving, gross motor, attention), while more nuanced or context-dependent domains (e.g., emotional expression, exploration, receptive communication, fine motor) showed lower concordance.
- **Complementary Insights:** Video-derived behavioral features captured subtle, emerging, or context-specific skills that were sometimes discordant with standardized measures, highlighting the added ecological value of naturalistic home video assessment.
- **Implications for Developmental Monitoring:** Integrating automated video analysis with standardized assessment data can enhance developmental monitoring by improving sensitivity to real-world performance, strengthening ecological validity, and supporting objective, multimodal evaluation of functional abilities.

The Nathan S. Kline Institute for Psychiatric Research

VALIS BIOSCIENCE