Original Paper

The Spectrum & Clinical Tools for Reliable Assessment (SPECTRA) Evidence Stack for Adult Neurodivergence Assessment: Framework Development and Synthetic Data Evaluation

Lucas Fortaleza[1,2], MD, MSc; Lenifer Siqueira Landim[1]; Giovana Melo Bezerra Teixeira[3]; Charlene Coleman[4]; Ana Maria Luna Neri Benevides[1], MD, MSc

[1]Undergraduate Medical Program, Universidade Christus (Unichristus), Fortaleza, Brazil

[2]NeurAtip, Fortaleza, Brazil

[3]Undergraduate Medical Program, Universidade de Fortaleza (Unifor), Fortaleza, Brazil

[4]NeurAtip, Dublin, Ireland

Corresponding Author:
Lucas Fortaleza, MD, MSc
Undergraduate Medical Program
Universidade Christus (Unichristus)
133 João Adolfo Gurgel Street
Cocó
Fortaleza, 60192-345
Brazil
Phone: 55 (85) 8767 0252
Email: lucas@neuratip.com

Abstract

Background: Psychological assessment of adult neurodivergence relies on multi-informant data that frequently disagree; discrepancies may reflect masking, context-dependent expression, or limited self-insight rather than measurement error.

Objective: This paper specifies the SPECTRA (Spectrum & Clinical Tools for Reliable Assessment) Evidence Stack, a rule-based, neuroaffirmative architecture that preserves cross-informant discrepancies as interpretable evidence rather than collapsing them into opaque scores, producing quality flags, convergence profiles, evidence-strength bands, and reason codes.

Methods: This is a design specification, not a validation study: no clinical participants were involved. Framework behavior was evaluated on 250 author-generated synthetic records through a 4-condition, prespecified rule-configuration comparison and a threshold-perturbation sensitivity analysis. Stochastic generation and score jitter were controlled by a recorded random seed.

Results: With the disclosed code, version, and seed, the reference run was reproducible: mean evidence strength was 0.60 for the full-stack configuration, 0.53 for the no-quality comparator, 0.54 for the no-convergence comparator, and 0.50 for the baseline comparator, with the ordering A > C > B > D preserved in all 12 threshold-perturbation scenarios. These descriptive differences verify the behavior of the 4 specified configurations; they are not causal estimates of the isolated effects of individual layers. An extended sensitivity analysis covering all 22 reference-configuration parameters (one-at-a-time ±10%, a 1000-sample Latin hypercube joint perturbation of ±20%, and tipping-point sweeps of the band cutoffs) showed that the ordering was preserved in 38 of 44 single-parameter scenarios but not under joint perturbation (278/1000, 27.8% of samples); the caution cap and the high-band cutoff were the dominant drivers of band assignment. The reference configuration is fully disclosed and independently runnable.

Conclusions: Within the disclosed reference configuration, the SPECTRA Evidence Stack produced reproducible, conditionally assigned evidence-strength outputs that preserved cross-informant discrepancy as interpretable evidence and made uncertainty visible through bands, flags, and reason codes; the ordering of the 4 rule configurations was preserved in most single-parameter scenarios but not under joint perturbation, and the parameters most influential for band assignment (the caution cap and the high-band cutoff) were identified for priority calibration. Diagnostic accuracy, criterion validity, reliability, and clinical utility

XSL•FO
RenderX

are not evaluated here and are reserved for a subsequent clinical validation study, for which a prespecified calibration protocol is now provided.

(JMIR Form Res 2026;10:e106266) doi: 10.2196/106266

KEYWORDS

adult autism; measurement framework; multi-informant assessment; neurodivergence; psychological assessment

Introduction

The Assessment Problem: Fragmentation, Masking, and Opaque Integration

Autism, attention-deficit/hyperactivity disorder (ADHD), and related neurodivergent presentations are increasingly being identified and assessed in adulthood. However, validated, accessible, and bias-conscious assessment tools remain limited [1,2]. Most existing instruments were originally developed for children or rely primarily on single-source self-report, reducing their usefulness for adults who have developed sophisticated compensatory strategies, particularly women and gender-diverse individuals [3,4]. For example, the Ritvo Autism Asperger Diagnostic Scale–14 Screen (RAADS-14) showed adequate sensitivity (95%) but poor specificity (70%) when validated outside its original population, suggesting that single-source screening instruments may lack the discriminative precision needed for culturally and linguistically diverse adult populations [5]. Social camouflaging, or masking, is associated with increased anxiety, depression, exhaustion, burnout, and delayed identification [6-8]. Rather than representing merely a source of measurement error, masking is itself a clinically meaningful phenomenon that assessment systems should detect, preserve, and interpret.

Contemporary digital assessment platforms increasingly integrate multiple sources of evidence, including self-report questionnaires, informant reports, structured performance tasks, clinician-observed behavioral tasks, semistructured interview observations, response-time telemetry, and open-ended responses. Although this multimodal approach has the potential to improve assessment, these sources frequently yield discrepant findings. Self- and informant reports may diverge because informants observe individuals across different contexts [9]. Likewise, performance- or observation-based measures may underestimate traits that are successfully masked in structured settings, whereas unusual response patterns may reflect fatigue, alexithymia, or demand characteristics rather than the absence of neurodivergent traits. When these heterogeneous signals are collapsed into opaque composite scores, clinically meaningful discrepancies may be obscured, leading to overconfident interpretations and reduced transparency.

This creates a recurring clinical challenge. In adult neurodivergence consultations—particularly among late-diagnosed adults, women, and gender-diverse individuals—the primary clinical bottleneck is rarely a scarcity of questionnaires or metrics, but rather the interpretation of legitimate discrepancies between information sources. For instance, an adult patient may report severe internal masking effort, cognitive exhaustion, and sensory overload, while a workplace colleague or parent informant describes seemingly adaptive, functional outward behavior. Crucially, conventional psychometric practice frequently resolves cross-source discrepancy by calculating an arithmetic mean or composite score across informants. From a neuroaffirmative clinical standpoint, arithmetic averaging across sources is fundamentally flawed: it incorrectly assumes that self- and informant reports measure the identical construct under identical environmental conditions; it collapses the vital distinction between subjective internal cost and observable external appearance; and it treats discrepancy as mere “measurement error” to be averaged away, thereby invalidating the lived experience of the evaluated individual.

In the absence of transparent integration logic, clinicians are often forced either to favor a single source of evidence or to disregard discrepancies altogether. The SPECTRA (Spectrum & Clinical Tools for Reliable Assessment) Evidence Stack was developed to address this gap by treating discrepancies not as noise to be resolved, but as interpretable, traceable, and clinically meaningful signals. Under the SPECTRA framework, cross-source discordance opens a structured investigative hypothesis—prompting targeted clinical inquiry into compensation history, environmental scaffolding, or contextual variance—rather than closing inquiry with a premature, automated diagnostic verdict. It is intended as a clinician-facing decision-support module that structures multisource assessment data within a clinical workflow, not as a stand-alone diagnostic system. In the intended workflow, a qualified clinician reviews the SPECTRA output (evidence-strength band, reason codes, and next-step recommendations) during a multidisciplinary assessment or clinical formulation meeting; the clinician decides whether additional history, informant clarification, or reassessment is needed and can override or defer the output when it is inconsistent with the broader clinical context, following a standardized override, deferral, and audit protocol (Methods section). Training requirements for users will be evaluated in future usability studies.

Discrepancy as Signal Rather Than Noise

A neuroaffirmative assessment framework—one that respects neurodivergent identity and experience rather than treating difference as deficit—should regard discrepancies between sources as potentially informative rather than automatically invalidating. Differences across self-report, informant-report, and performance-based evidence may reflect masking (self-report lower than informant or performance measures), hypervigilance or heightened self-monitoring (self-report higher than informant measures), context-dependent trait expression, limited self-insight, or state-dependent variability in functioning. Each pattern has distinct clinical implications and therefore warrants explicit interpretation rather than forced convergence.

XSL•FO
RenderX

This perspective is consistent with growing calls for transparency and explainability in clinical decision-support systems [10-12]. It also responds to longstanding critiques from the autistic community regarding deficit-oriented frameworks and opaque scoring methods [13-16], as well as participatory research approaches that emphasize the involvement of autistic individuals in the design and evaluation of assessment tools [17,18]. Within this framework, identity-first, nonpathologizing language and explicit decision logic are not merely stylistic preferences; they are methodological requirements for transparent and trustworthy assessment.

Existing approaches to multisource evidence in digital assessment typically fall into 2 categories: opaque composite scores that collapse discrepancies into a single metric, and machine learning models that optimize prediction at the cost of interpretability [10,11]. Neither approach is well suited to neurodivergence assessment, where disagreement across sources is often clinically meaningful and where autistic communities have emphasized the importance of transparent, explainable decision-making [13-16]. The SPECTRA Evidence Stack differs from these approaches in 3 ways. First, unlike interpretable early-warning scores or sepsis-alert systems that combine multiple data sources into a single risk score, SPECTRA preserves discrepancy as a structured convergence profile rather than resolving it into one number. Second, unlike multi-informant autism instruments that report source discrepancies descriptively, SPECTRA feeds discrepancy profiles directly into an evidence-strength estimate with mandatory reason codes and next-step recommendations. Third, unlike black-box predictive models, SPECTRA makes every output traceable to explicit data conditions and predefined rules, with a hard nondiagnostic boundary enforced by output contracts.

Why a Rule-Based Architecture?

At this formative stage, a rule-based architecture was intentionally prioritized over machine learning because interpretability, auditability, and regulatory transparency were considered more important than predictive optimization. This prioritization is consistent with emerging regulatory expectations for AI-enabled clinical decision support, which emphasize transparent documentation, human oversight, and postmarket monitoring [19]. Rule-based systems have well-known limitations: they cannot easily capture complex, nonlinear interactions unless those interactions are explicitly encoded, and their performance depends heavily on the completeness and accuracy of the specified rules. However, in neurodivergence assessment, disagreement across information sources is often clinically meaningful rather than undesirable, and autistic communities have consistently emphasized the importance of transparent decision-making. Under these conditions, an entirely black-box model would limit interpretability at precisely the stage where transparency is most critical. A rule-based approach therefore trades some predictive flexibility for accountability and explainability.

Accordingly, the SPECTRA Evidence Stack adopts a rule-based architecture because it provides 3 properties that are essential for an early-stage assessment framework: (1) interpretability—every output can be traced to explicit data conditions and predefined decision rules, enabling clinicians and respondents to understand why a particular confidence estimate, convergence profile, or reason code was generated; (2) auditability—the integration logic of the reference configuration can be independently inspected, challenged, and refined without retraining opaque model parameters, thereby supporting iterative improvement through clinical, community, and regulatory review; and (3) regulatory appropriateness—rule-based clinical decision-support systems are generally easier to document, audit, and evaluate than autonomous diagnostic algorithms, making them well suited to emerging expectations for transparency in AI-enabled health technologies [10,20].

The current reference configuration was refined during revision in response to editorial and reviewer comments. The parameter values used for the final reported execution were frozen before that execution and recorded in the accompanying implementation. They were selected through an expert-informed design process led by a psychiatrist with expertise in adult neurodevelopmental assessment and reviewed against the framework's design principles—transparency, uncertainty preservation, and conservative evidence integration—by the coauthors. They were not preregistered or empirically optimized against clinical outcomes and are provisional rather than optimal; empirical calibration is reserved for future validation studies. Accordingly, the principal contribution of this paper lies in the framework architecture and explicit decision logic rather than in the numerical values assigned to individual parameters.

Importantly, this rule-first strategy should not be interpreted as a rejection of machine learning. If future empirical studies demonstrate that learned components improve calibration while maintaining interpretability and auditability, they could be incorporated within the existing architectural framework without altering its layer contracts or transparency principles.

Study Aims and Scope

This multistudy report presents 3 interrelated methodological evaluations of the SPECTRA Evidence Stack. First, in study 1 (architecture specification), we describe the SPECTRA Evidence Stack, a 3-layer framework for integrating heterogeneous evidence in adult neurodivergence assessment, specifying transparent and traceable decision pathways linking raw assessment inputs to quality flags, convergence profiles, confidence estimates, reason codes, and next-step recommendations. Second, in study 2 (synthetic rule-configuration comparison), we compare 4 deliberately specified rule configurations on 250 author-generated synthetic records (seed=42) to verify their documented behavior under varying configurations of evidence quality, missingness, and cross-source discrepancy. Configurations B and C are comparators, not controlled knockouts, and their differences from A are not causal estimates of the isolated effects of a layer. Third, in study 3 (sensitivity analyses and clinical vignettes), we evaluate (1) threshold-perturbation sensitivity across 12 scenarios (±10% perturbation on 6 quality thresholds); (2) an extended parameter-sensitivity analysis covering all 22

XSL•FO
RenderX

reference-configuration parameters one at a time, a global joint perturbation (Latin hypercube, 1000 samples, ±20%) with PRCC indices, and tipping-point sweeps of the 3 confidence-band cutoffs; and (3) a set of 3 worked clinical demonstration vignettes (masking-compatible divergence, informant-context mismatch, and sparse evidence) with full decision traces.

These studies collectively support the development of neuroaffirmative, transparent, and clinically accountable assessment technologies. The proprietary implementation, calibration assets, and software components of the SPECTRA platform remain protected intellectual property of the developing organization.

The present study represents an early-stage clinical decision-support framework development phase. A DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Using Artificial Intelligence) reporting checklist [20] is provided as a prospective alignment—a roadmap for the framework's future clinical evaluation—rather than as a record of a clinical evaluation that has already been performed. No clinical validation or effectiveness study is reported here. No formal preregistration was performed because the work is an early-stage methodological specification and synthetic stress test rather than a confirmatory hypothesis-testing study.

Equally important is defining the boundaries of this work. This paper does not evaluate diagnostic validity against independent clinical adjudication, report instrument-specific psychometric properties, or make claims regarding clinical effectiveness or real-world utility. Those questions will be addressed in subsequent publications within the SPECTRA research program, including validation studies of the Autism Understanding by Relative Assessment (AURA), Neurodivergent Expression Assessment (NEXA), Veiling Effort & Identity Load (VEIL), and Textual Recognition of Pragmatic Expressions (TROPE) instruments; a test-retest reliability study of Autism Spectrum Test–10 Items (TEA-10); and a capstone clinical validation study comparing Evidence Stack outputs with independent expert adjudication. Establishing these scope boundaries is essential to prevent premature claims of validity and to emphasize the formative nature of the present work.

Methods

Study 1: Architecture Specification

Design Principles and Scope

The SPECTRA Evidence Stack was designed around 6 principles that distinguish it from conventional scoring pipelines.

Neuroaffirmative Framing

All outputs use identity-first language ("autistic adult," "ADHDer") and avoid deficit-only framing. Traits are described as differences that may create challenges in specific contexts, not as intrinsic impairments. The framework does not label respondents as "invalid" when data quality is low; instead, it flags data conditions and recommends remediation. Neuroaffirmative language and principles are presented throughout as design intentions—commitments guiding the architecture—rather than as empirically validated characteristics because no formal evaluation with autistic or otherwise neurodivergent users has yet been conducted.

Community Involvement and Positionality

The framework's design principles were shaped through the first author's ongoing engagement with the neurodivergent community. The first author is an autistic psychiatrist who founded 2 Brazilian online communities for neurodivergent adults—1 focused on autistic adults (more than 54,000 followers) and 1 focused on adults with ADHD (more than 37,000 followers). Members of these communities contributed informally and predominantly indirectly to the framework's development: their dialogue, lived-experience perspectives, and feedback validated or contextualized design choices and shaped the first author's understanding of neuroaffirmative, assertive, and contemporary assessment practice, rather than formally co-designing the architecture. No contributor received financial compensation. This engagement is reported here as design context, not as a formal participatory or usability evaluation; a dedicated participatory and usability evaluation with autistic and otherwise neurodivergent users is planned (see Future Directions section).

Discrepancy Preservation

The stack does not force a single latent score when sources disagree. It preserves divergence as structured convergence profiles with explicit interpretation constraints. This principle is central to detecting and interpreting masking.

Auditability and Explainability

Every output must be traceable to inputs and rules. Reason codes are mandatory, human-readable, and linked to the specific quality flags and convergence profiles that produced them. The reference configuration described in this paper—including its thresholds, tolerances, convergence rules, and code—is fully disclosed and deposited at Open Science Framework (OSF; code component [21]), enabling independent verification of the reference configuration. Internal traceability within this disclosed reference configuration is not equivalent to independent auditability of the deployed production system. Production calibration values, which are tuned against platform operating data and may differ from the published reference configuration, remain proprietary to the developing organization and are outside the claims of this paper.

Conservative Confidence

Confidence, in this context, refers to evidence strength—the completeness, consistency, and cross-source traceability of the available evidence—not to diagnostic confidence or probability of condition. Confidence is capped when evidence is sparse, low-quality, or highly discrepant. The system is designed to be transparently uncertain rather than falsely precise. Low confidence is a useful output, not a failure.

Nondiagnostic Boundary

The SPECTRA Evidence Stack is assessment support, not diagnostic automation. It organizes evidence and supports clinical formulation; it does not produce diagnostic labels or

XSL•FO
RenderX

replace clinical judgment. This boundary is enforced by output contracts and reason-code wording.

Privacy by Design

Multisource data are collected with independent consent. Informant responses are not shared with participants and vice versa. Telemetry is used exclusively to characterize response quality and is never used to infer traits outside the scope of the assessment. All processing follows applicable data protection legislation (including Lei Geral de Proteção de Dados Pessoais [LGPD] and General Data Protection Regulation [GDPR]).

Study Design: The SPECTRA Evidence Stack

Overview

The framework comprises three sequential layers (Figure 1): (1) layer 1 (response quality) evaluates data completeness, response patterns, and telemetry to produce quality flags and gate profiles; (2) layer 2 (multisource convergence) compares evidence across sources (self, informant, and performance or observational) and assigns a convergence profile; and (3) layer 3 (confidence synthesis) combines quality and convergence artifacts into a confidence band, reason codes, and next-step recommendations.

Figure 1. The SPECTRA (Spectrum & Clinical Tools for Reliable Assessment) Evidence Stack as a linear evidence pipeline. Raw inputs (self-report, informant-report, and performance or observational data) are normalized through a canonical ResponseRecord schema and flow sequentially through layer 1 (response quality), layer 2 (multisource convergence), and layer 3 (confidence synthesis). Each layer emits structured artifacts—quality flags and gates, a convergence profile, and a confidence band with mandatory reason codes and next-step clinical recommendations. The vertical arrows emphasize the unidirectional, traceable path from input to actionable output; every output of the reference configuration can be reconstructed from the input records and the published rule set.

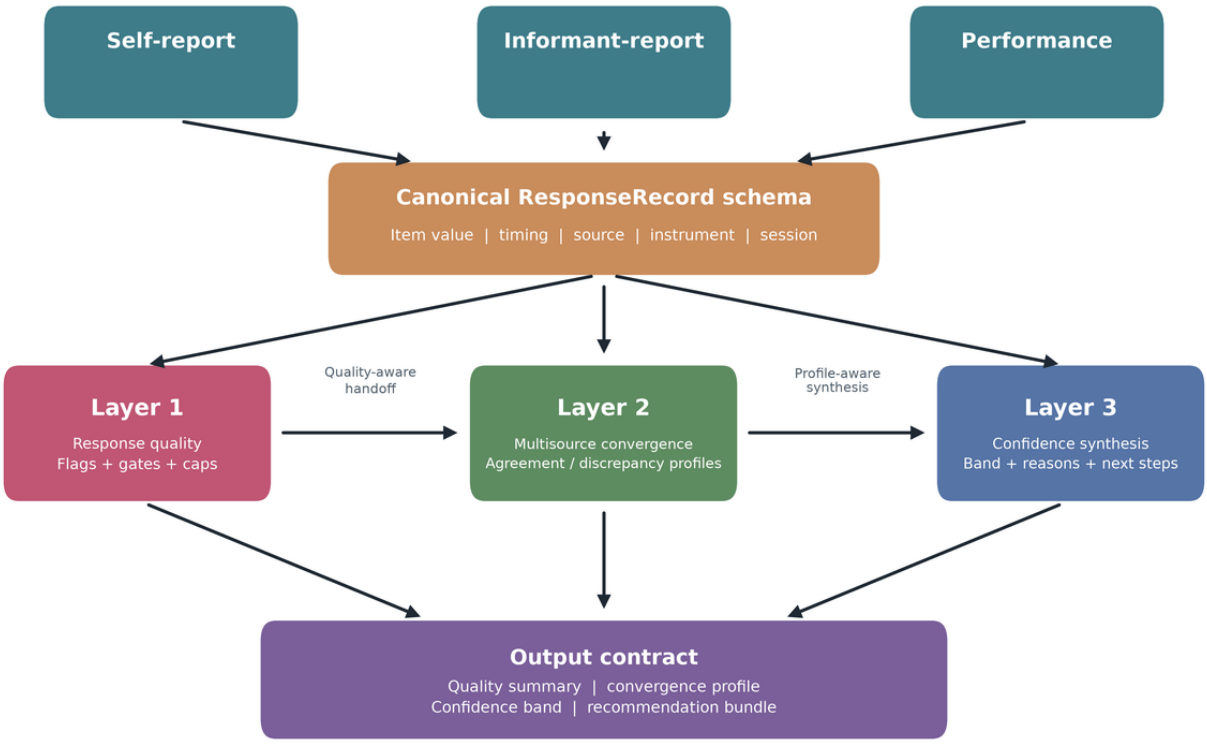


Figure 1 shows the complete data flow as a linear evidence pipeline: input → quality → convergence → confidence → reason codes → clinical recommendation. This visual layout is intended to make the architecture memorable and to highlight that each downstream output is fully determined by the preceding layers. The pipeline is deliberately modular so that each layer can be tested, inspected, and updated independently. The canonical ResponseRecord schema normalizes heterogeneous inputs, allowing the same quality and convergence rules to operate across instruments. Each layer emits structured artifacts consumed by the next, creating a fully traceable path from input to output. This traceability within the disclosed reference configuration—not independent auditability of any deployed system—is the property demonstrated here: any output of the reference configuration can be reconstructed from the input records and the published rule set. The SPECTRA Evidence Stack described in this paper is methodologically specified at the architectural level. The SPECTRA production platform is developed and operated by the developing organization; production implementation details, calibration assets, and software components remain proprietary. This manuscript provides a complete methodological specification of the Evidence Stack sufficient to understand its design, decision logic, and intended behavior, but does not disclose commercially sensitive implementation details.

Input Schema

All assessment responses pass through a canonical ResponseRecord schema with fields for item identifier, response value, response type, per-item timing, perspective (self, informant, or clinician), instrument identifier, participant identifier, and session identifier. This schema enables

XSL•FO
RenderX

cross-instrument timing analytics and convergence analysis without instrument-specific data pipelines.

Output Contract

Every case produces the following four artifacts: (1) a quality summary consisting of a set of flags and gate status; (2) a convergence profile describing the labeled pattern of agreement or divergence across sources; (3) a confidence estimate presented as a bounded score with an explicit interpretation band; and (4) a next-step bundle containing reason codes and recommended actions. This contract prevents silent failures and supports reproducible reporting.

Layer 1: Response Quality

Overview

Response quality evaluates whether the available data are sufficient and interpretable. A quality flag is a property of data conditions, not a judgment about the person. This distinction is essential for neuroaffirmative implementation.

Quality Flag Definitions

Table 1 summarizes the core quality flags. Each flag is computed from response behavior, completion patterns, or telemetry. Flags are grouped into 3 severity levels: information (noted but does not gate interpretation), caution (triggers a confidence cap), and critical (prevents automated interpretation and recommends reassessment).

Table 1. Response quality flags, thresholds, and gate behavior.

Flag	Definition	Input features	Severity	Gate effect
high_missingness	≥20% of items in a domain unanswered	Missingness rate	Caution	Retain technical cap at moderate; suppress reader-facing score and band
straightlining	≥80% of consecutive items with identical responses	Response sequence	Caution	Cap confidence at moderate
rapid_responding	≥30% of items with response time <,2 s	Response time	Caution	Flag for review; cap confidence
slow_responding	≥20% of items with response time >,60 s	Response time	Information	Note possible distraction
fatigue_slope	Response time increases ≥50% from first to last third of assessment	Response time trend	Caution	Cap confidence at moderate
acquiescence	≥70% of nonreverse items endorsed in same direction	Response distribution	Caution	Flag for review
incomplete_session	Session terminated before mandatory modules	Completion status	Critical	Retain capped technical score for QA[a]; suppress reader-facing band and require human review
single_source_only	Only one evidence source available	Source count	Information	Document in reason codes
informant_mismatch	Informant reports knowing respondent <6 months	Informant metadata	Caution	Cap confidence at 0.74 (reference value; Reference Configuration Parameters subsection)

[a]QA: quality assurance.

Flag thresholds for the reference configuration are listed in Table 1 and deposited with the accompanying code at OSF [21]. Production calibration values, tuned against platform operating data, may differ and remain proprietary. The architecture itself is fully disclosed herein. The revision-stage parameter history and final-execution freeze are described in the Introduction; no claim of global preregistration or prereview prespecification is made.

Gate Logic

Gates translate flag patterns into downstream constraints. A gate can cap evidence strength, require a specific reason code, or suppress reader-facing interpretation. For a critical record or a record with high missingness, the implementation retains the technical score and band for software quality assurance (QA) and aggregate verification, but it does not display either value to the user. The reader-facing output is "Insufficient — human review required," and human review is an action recommendation rather than a reason-code identifier. Multiple flags can activate simultaneously; the most conservative rule prevails.

Clinician Override, Deferral, and Audit Protocol

Overview

Because the framework is decision support rather than decision automation, the intended workflow defines a standardized protocol for what happens when outputs and clinical judgment diverge. The protocol has 4 parts.

Decision Types

For every reviewed output, the reviewing clinician records one of four decisions: (1) accept (output incorporated into the formulation as an interpretive draft), (2) override (the clinician's judgment replaces the output's recommendation, with rationale), (3) defer (interpretation is postponed pending additional data, with the specific gap named), or (4) escalate (the case is referred

XSL•FO
RenderX

for multidisciplinary review, used when outputs conflict with each other or with the clinical picture in ways the clinician judges too complex for a single-reviewer decision).

Mandatory Deferral Triggers

Independent of the clinician's judgment, 3 output states require deferral rather than acceptance: a reader-facing output of "Insufficient — human review required" (critical records or high missingness), activation of a critical quality flag, and a single-source record whose convergence profile is discordance-sensitive (masking-compatible, hypervigilance-compatible, or context-dependent). These triggers exist because the underlying evidence is structurally incomplete, not because the clinician's judgment is in question.

Minimum Documentation Fields

Every override or deferral is logged with the contested output element (band, profile, or specific reason code), the clinician's decision and rationale (free text), any additional evidence relied on, the timestamp, and the clinician's role. The log entry is stored in the clinical record alongside the framework output, so that the framework's contribution and the clinician's decision remain separable and auditable.

Automation-Bias Safeguards

Three safeguards address automation bias explicitly: (1) where workflow permits, the clinician records an initial structured impression before viewing the evidence-strength band, so that the band cannot anchor the formulation; (2) all reason codes must be read before a decision is recorded—the reason codes, not the band, carry the interpretive content; and (3) override rates are monitored as a safety signal in both directions: a near-zero override rate is treated as seriously as a near-total one, since the former suggests rubber-stamping (automation bias) and the latter suggests that the framework is not supporting decisions.

Audit Rubric

Deployments are audited on a quarterly cadence against five indicators: (1) override rate overall and stratified by band and by reason code; (2) clinician-output agreement on deferral triggers (deferral is mandatory, so agreement should be 100%); (3) documentation completeness (proportion of overrides with all minimum fields); (4) appropriateness review, in which a sample of overrides is rated by an independent senior clinician as clinically appropriate or not; and (5) outcome linkage, where available. Prespecified action thresholds—for example, an override rate above 40% or below 5%, documentation completeness below 90%, or a calibration slope outside 0.8 to 1.25 on monitoring data—trigger recalibration under the prespecified empirical calibration protocol or suspension of deployment pending review. The rubric will be validated in the planned usability and clinical validation studies before any deployment.

Neuroaffirmative Interpretation

Quality flags are reported as "data conditions" rather than "validity failures." For example, rapid responding might reflect anxiety, executive-function overload, or a mismatch between interface pacing and processing speed, not necessarily careless responding. The reason code links the flag to a recommended action (eg, offer a break and readminister with accommodations) rather than a pejorative label.

Layer 2: Multisource Convergence

Overview

Multisource convergence evaluates agreement and divergence across evidence sources. Instead of computing a single latent score, the layer assigns one of several convergence profiles. Each profile carries interpretation constraints that feed into confidence synthesis.

Evidence Sources

The framework currently integrates three source families: (1) self-report, which includes questionnaires completed by the respondent (eg, NEXA and VEIL); (2) informant-report, which includes questionnaires completed by someone who knows the respondent well (eg, AURA); and (3) performance or observational evidence, which includes structured performance tasks, clinician-observed behavioral tasks, and semistructured interview observations (eg, TROPE and SPECTRA Clinical Interview [SPECTRA-CI]). Additional sources (eg, ecological momentary assessment and clinical interview records) can be incorporated through the same schema.

Convergence Profiles

Table 2 defines the core convergence profiles. Profiles are determined by comparing standardized scores across sources within each domain (eg, social communication, restricted or repetitive behaviors, attention, and emotional regulation).

XSL•FO
RenderX

Table 2. Convergence profile definitions and interpretation constraints.

Profile	Pattern	Clinical interpretation	Confidence implication
convergent	All sources agree within a predefined tolerance	Strong cross-source evidence	Permits higher confidence
masking_compatible	Self-report lower than informant or performance	Respondent may underreport due to masking or limited insight	Cap confidence; flag masking
hypervigilance_compatible	Self-report higher than informant or performance	Respondent may overreport due to anxiety, hypervigilance, or recent stress	Cap confidence; flag hypervigilance
context_dependent	Informant and performance diverge without clear self-report alignment	Trait expression varies by setting	Moderate confidence; recommend multisetting observation
limited_evidence	Only 1 source available or data too sparse	Insufficient cross-source evidence	Low confidence; recommend additional data
mixed_discrepancy	Multiple contradictory patterns across domains	Complex presentation requiring clinical integration	Low confidence; recommend clinical review

Profile assignment uses domain-specific rules rather than a single global discrepancy threshold. The rules and tolerances for the reference configuration are fully disclosed in the deposited code (reference_implementation.py at OSF [21]); the extended-sensitivity script and its captured output are also available in the OSF code component [21]. The numerical parameters are listed in the Reference Configuration Parameters subsection below.

Discrepancy as Signal

Each profile maps to a set of candidate reason codes. For example, masking_compatible divergence in the social domain may trigger the reason code "self-report lower than informant or performance-consider social camouflaging or context-dependent expression." This wording validates masking as a real phenomenon rather than treating the respondent as unreliable.

Layer 3: Confidence Synthesis

Overview

Confidence synthesis combines layer 1 and layer 2 artifacts into a final output. Confidence here means evidence confidence—the completeness, consistency, and cross-source traceability of the evidence—not diagnostic confidence.

Confidence Components

Table 3 lists the components that contribute to the confidence estimate. Each component is bounded and combined through a gated rule set. The reference configuration uses no free component weights: each component acts through explicit gates and caps (Table 3, right columns), and the only additive parameters are the profile base scores, the clean-record bonus (+0.06), the no-quality penalty (−0.10), and the jitter SDs, all listed in the Reference Configuration Parameters subsection. These values are expert-derived and provisional; empirical calibration is planned (see Prespecified Empirical Calibration Protocol subsection).

Table 3. Confidence components, gates, and reason-code mapping.

Component	Description	Gate or cap	Reason-code family
data_quality	Inverse of quality-flag severity	Critical flags cap confidence at low	QUALITY_*
source_coverage	Number and balance of evidence sources	Single source caps at moderate	COVERAGE_*
convergence_strength	Agreement across sources	Discrepant profiles cap at moderate or low	CONVERGENCE_*
domain_completeness	Coverage of relevant assessment domains	Missing domains cap confidence	DOMAIN_*
temporal_stability	Consistency with prior assessments if available	Large change triggers review	STABILITY_*

Confidence Bands

The continuous confidence score is mapped to four bands for reporting: (1) high (≥0.75), indicating that evidence is complete, consistent, and multisource, and that interpretation is well-supported; (2) moderate (0.50-0.74), indicating that evidence is adequate but has notable limitations (eg, single source, mild discrepancy, or quality caution); (3) low (0.25-0.49), indicating that evidence is sparse, discrepant, or affected by quality issues, and that interpretation requires substantial caution; and (4) insufficient (<0.25), indicating that evidence is inadequate for any automated interpretation and that reassessment or clinical integration is required.

XSL•FO
RenderX

Reference Configuration Parameters

For transparency, the numerical parameters of the reference configuration (the object evaluated in this paper) are listed here, including the confidence band cutoffs (high≥0.75, moderate 0.50-0.74, low 0.25-0.49; and insufficient<0.25). These values are implemented verbatim in the deposited reference_implementation.py and are the same values that produced the reported results. The profile base evidence-strength scores are listed in Table 4. They are the reference configuration only; production calibration values differ and remain proprietary.

Table 4. Profile base evidence-strength scores (before gates), assigned to respect the "confidence implication" ordering in Table 2.

Profile	Base score
convergent	0.88
context_dependent	0.62
masking_compatible	0.60
hypervigilance_compatible	0.60
limited_evidence	0.45
mixed_discrepancy	0.30

Gate behavior (reference configuration) includes the following: a critical flag caps confidence at 0.20, a caution flag caps confidence at 0.74, single-source coverage applies a hard cap of 0.74 (below the high band), and a clean (flag-free) record earns a quality-confirmation bonus of +0.06 in the full-stack configuration (configuration A). These caps and the bonus are explicit reference-configuration choices; the manuscript describes the gates qualitatively (eg, "caps confidence at moderate"), and these values make them quantitative.

The rule-configuration arithmetic (reference configuration) for the 4 verification configurations computes evidence-strength scores using deliberately specified configurations. Configurations B and C alter more than 1 operational feature and therefore are comparison configurations, not controlled knockouts or causal ablations. The full arithmetic is stated here so that every reported numerical result can be reconstructed from the disclosed code, version, and seed:

- Configuration A (full stack): score = base[profile] + bonus + ε, where bonus = +0.06 if the record is clean (no flags) and multisource, otherwise bonus = 0; ε ~ N(0, 0.03). The score is then capped: min(base, 0.20) for critical records (with ε ~ N(0, 0.02)), and min(base, 0.74) for caution-flagged or single-source records. The final score is bounded between 0.05 and 0.98. For critical or high-missingness records, this value is retained only as an internal technical output; the reader-facing output suppresses the score and band and states "Insufficient — human review required."
- Configuration B (no quality): the response quality layer is removed, so the system cannot distinguish clean from noisy records. It therefore forgoes the confirmation bonus and applies a flat reliability penalty of −0.10 to every record, with additional convergence uncertainty ε ~ N(0, 0.11): score = base[profile] − 0.10 + ε. The final score is bounded between 0.05 and 0.98.
- Configuration C (no convergence): only single-source (self-report) evidence is used, with quality gating retained: score = 0.55 + ε, with ε ~ N(0, 0.04), capped at 0.20 for critical records, 0.70 for caution records, and a hard ceiling of 0.74 for all records (the source-coverage gate keeps single-source evidence below the high band).
- Configuration D (baseline): uniform scoring with no quality or convergence logic: score ~ N(0.50, 0.20), clipped between 0.05 and 0.98.

These formulas make the record-level differences reported in the Results section directly legible. Clean records have mean 0.628 in configuration A and 0.521 in configuration B, a descriptive difference of 0.106. This difference reflects both the configuration choices and the recorded random stream: 59 clean multisource records are eligible for the +0.06 bonus in configuration A, whereas configuration B applies a −0.10 penalty and larger jitter; 68 clean single-source records receive neither the configuration A bonus nor identical treatment in configuration B. The weighted descriptive decomposition, (59 × 0.1546 + 68 × 0.0646) / 127, is 0.1064. The larger dispersion in configuration B (SD 0.21 vs 0.17) is also specified by its larger jitter (SD 0.11 vs 0.03). These are reproducible configuration contrasts conditional on the recorded seed and implementation version, not causal effects attributable to removing one layer.

Reason Codes and Next-Step Recommendations

Reason codes are mandatory and human-readable. Each code has a fixed identifier (used internally to trace which rule fired) and a plain-language explanation shown to the clinician. Examples of code identifiers and their explanations include the following:

- QUALITY_HIGH_MISSINGNESS: "Some domains had many unanswered items. Consider readministering with breaks or simplified instructions."
- CONVERGENCE_MASKING_COMPATIBLE: "Self-report was lower than other sources. Consider whether social camouflaging or context-dependent expression may be present."
- COVERAGE_SINGLE_SOURCE: "Only self-report was available. Informant or performance data would strengthen interpretation."

Next-step recommendations are framed as suggestions for the clinician and respondent, not as deterministic directives. They

XSL•FO
RenderX

are restricted to evidence-gathering and assessment-process guidance (eg, obtain additional informant information, review discrepant evidence, consider further assessment when evidence is incomplete) and do not include diagnostic labels, treatment recommendations, medication advice, or clinical management decisions.

Reporting Standards

The methods reported in this paper reflect principles of transparent reporting for measurement frameworks and decision-support tools [20,22,23]. We provide explicit architectural documentation, decision rules, parameter values, and rule-configuration logic to enable conceptual replication and critical evaluation. Although the present work involves a synthetic stress test rather than a clinical validation study, we address key transparency elements: (1) description of the framework and its components (Study Design: The SPECTRA Evidence Stack; Layer 1: Response Quality; Layer 2: Multisource Convergence; and Layer 3: Confidence Synthesis sections); (2) user population and setting (Introduction and Design Principles and Scope sections); (3) data inputs, quality, and governance (Input Schema and Layer 1: Response Quality sections); (4) logic, algorithms, and rules (Layer 1: Response Quality, Layer 2: Multisource Convergence, and Layer 3: Confidence Synthesis sections); (5) outputs and their presentation to users (Layer 3: Confidence Synthesis and Study 3 Results: Clinical Demonstration Vignettes sections); (6) human oversight (Design Principles and Scope; Clinician Override, Deferral, and Audit Protocol; and Ethical Considerations sections); (7) ethical safeguards (Ethical Considerations section); and (8) limitations and next steps (Discussion section). Future clinical validation studies will follow the COSMIN (Consensus-based Standards for the selection of health Measurement Instruments) guidelines for evaluating measurement properties [23] and will report psychometric performance (criterion validity, test-retest reliability, and measurement invariance) alongside clinical utility end points.

Study 2: Synthetic Rule-Configuration Comparison

Design and Scope of Inference

We conducted a formative synthetic stress test using 4 prespecified rule-behavior verification configurations on a synthetic dataset (N=250 records; seed=42). The purpose was computational verification under varying evidence configurations, not inferential hypothesis testing or demonstration of empirical superiority. Configurations B and C were deliberately specified comparator configurations: they change scoring penalties, jitter, source coverage, or caps in addition to omitting named layer behavior. They are therefore not controlled knockouts, and contrasts with configuration A cannot identify the isolated causal contribution of a layer. Because the records contain the patterns the architecture is designed to detect, the comparison is an internal rule-behavior check rather than a clinical evaluation. Fixed rules are combined with seeded stochastic generation and jitter; reported outputs are reproducible conditional on the code version, arguments, and seed, not deterministic mappings from input records to final scores irrespective of seed.

The synthetic dataset was not generated using a probabilistic population model; rather, it was constructed as a controlled rule-based stress test designed to evaluate whether the evidence stack produced the expected behavior under predefined evidence configurations.

The 4 configurations were as follows:

- Configuration A (full stack): quality + convergence + confidence (reference configuration)
- Configuration B (no quality): convergence + confidence, with quality gating removed
- Configuration C (no convergence): quality + confidence, using only self-report (single source)
- Configuration D (baseline): no quality + no convergence; uniform scoring without stack logic.

Dataset and Generation Parameters

Records were generated to include realistic patterns: complete and incomplete responses, variable response times, single-source and multisource cases, convergent and discrepant profiles, and masking-compatible divergence. The synthetic generator and analysis script are described in the Data Availability section. The synthetic data generation parameters are summarized in Table 5.

The profile targets in Table 5 are percentages of 250 records; the realized sample sizes reflect rounding to whole records and truncation of the final record so that the cohort totals exactly 250 records. The discordant-profile subset (masking-compatible + hypervigilance-compatible + context-dependent) therefore totals 113 (45.2%) records, which is the value reported throughout the Results and Discussion sections. Note that “limited evidence” (a convergence profile) and “single source” (a source-coverage property) are distinct quantities: the limited-evidence profile is single-source by definition (Table 2), whereas the 60/40 multisource/single source-coverage target applies to records independently of their profile, so the realized proportion of single-source records (128/250, 51.2%) exceeds the 40% target because every limited-evidence record is forced single-source.

Labels were assigned before scores were generated, so the comparison evaluates stipulated rule behavior rather than discovery of unknown structure. The final revision-stage configuration was frozen before the reported execution. The generation algorithm is summarized in Table 5 and deposited in full with the accompanying code at OSF (code component [21]; the earlier preprint is also available there [24]).

XSL•FO
RenderX

Table 5. Synthetic data generation parameters (N=250 records).

Parameter	Value
Random seed, n	42
Bootstrap replications (B), n	2000
Convergent profiles, n (%)	75 (30)
Masking-compatible divergence, n (%)	50 (20)
Hypervigilance-compatible divergence, n (%)	25 (10)
Context-dependent divergence, n (%)	38 (15)
Limited evidence (single source), n (%)	38 (15)
Mixed discrepancy, n (%)	24 (10)
Missingness rate (where induced)	15%-40% per domain
Response time distribution	Lognormal (μ=8 seconds, σ=0.8), with 10% fast (<2 seconds) and 5% slow (>60 seconds) outliers
Source coverage	60% multisource, 40% single-source (target)
Programming language	Python
Libraries	Python standard library only (no third-party packages)

Study 3: Sensitivity Analyses

Overview

To examine whether the rule-configuration findings depend on the specific illustrative parameters of the reference configuration, we conducted a 3-part sensitivity analysis. Part 1 involved threshold perturbation (Table 6), in which each quality threshold (missingness, straight-lining, rapid responding, fatigue slope, acquiescence, and informant-mismatch minimum acquaintance) was varied independently by ±10% around its nominal value, and the full-stack configuration was rerun under each perturbation.

Table 6. Threshold-perturbation sensitivity analysis across 12 scenarios (all scenarios use N=250 records). Recomputed August 22, 2026, using the deposited reference implementation (OSF [21]); values are reproducible by rerunning the command as follows: “reference_implementation.py --seed 42 --n 250.”

Scenario	Perturbation	Mean (SD)	Insufficient, n (%)	Low, n (%)	Moderate, n (%)	High, n (%)	Band changed, n
1	Missingness −10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
2	Missingness +10%	0.602 (0.170)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
3	Straightlining −10%	0.603 (0.172)	8 (3.2)	59 (23.6)	153 (61.2)	30 (12)	1
4	Straightlining +10%	0.602 (0.170)	8 (3.2)	59 (23.6)	155 (62)	28 (11.2)	1
5	Rapid responding −10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
6	Rapid responding +10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
7	Fatigue slope −10%	0.603 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
8	Fatigue slope +10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
9	Acquiescence −10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
10	Acquiescence +10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0
11	Informant mismatch −10%	0.603 (0.172)	8 (3.2)	59 (23.6)	153 (61.2)	30 (12)	1
12	Informant mismatch +10%	0.602 (0.171)	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)	0

Full-stack mean evidence strength ranged from 0.602 (SD 0.170) to 0.603 (SD 0.172), and the ordering A > C > B > D was retained in every scenario. The 10% perturbations were relative changes to each nominal detection rate, varied one at a time while all others remained nominal; the same latent-draw boundary rule was used. One of 250 records changed band in each of the 2 straight-lining scenarios and in the −10% informant-mismatch scenario; no record changed band in the other 9 scenarios. The complete 4-band distributions appear in Table 6. This seeded synthetic analysis describes sensitivity of the specified implementation; it does not validate the thresholds or identify causal layer effects.

Part 2 involved extended parameter sensitivity (Table 7). Because the editor’s calibration concern extends beyond the 6 quality thresholds, we extended the analysis to all 22 numerical parameters of the reference configuration, including the 6 profile base scores, the 3 gate caps (critical 0.20, caution 0.74, and single-source 0.74), the clean-record bonus (+0.06), the no-quality penalty (−0.10), the 3 band cutoffs, the 7 jitter SDs, and the configuration C constants. Three complementary probes

XSL•FO
RenderX

were run with the same cohort (seed 42; Latin hypercube seed 7), using a parameterized scorer that reproduces the deposited reference implementation exactly under default parameters (verified identity check; the previously reported results are unchanged).

Table 7. Extended parameter-sensitivity summary (seed=42; LHSa seed=7; full output: reference_run_output_extended.txt, OSF code component [21]).

Probe	Scenarios/samples	Full-stack mean range	Ordering A > C > B > D preserved, n/N (%)	Max band changes
OAT, all 22 parameters, ±10%	44	0.591-0.614	38/44 (86.4)	30/250 (12%; caution cap)
Global joint (LHS), ±20%	1000	0.530-0.671	278/1000 (27.8)	Up to 51.2% of records
Tipping point, high cutoff 0.60-0.90	13	Mean unchanged	Always (labels only)	Up to 46.4 pp[b] distribution shift

[a]LHS: Latin hypercube sampling.

[b]pp: percentage points.

One-at-a-time (OAT) perturbation analysis (±10%; 44 scenarios) showed that, across the 44 OAT scenarios (±10%), the mean confidence score for configuration A ranged from 0.591 to 0.614; the baseline mean was 0.602 (SD 0.171). The ordering A > C > B > D was preserved in 38 of 44 scenarios. The 6 rank-breaking scenarios were four ±10% increases in profile base scores (convergent, context-dependent, masking-compatible, and limited-evidence) plus ±10% decreases in the no-quality penalty and the configuration C base constant; they shifted configuration means by ≤0.011 and changed the band of at most 3.6% (9/250) of records. The largest band-impact parameter was the caution cap (±10% changed the band of up to 30/250, 12% records).

Global joint perturbation analysis (Latin hypercube, 1000 samples, all parameters varied simultaneously by ±20%) showed that, across 1000 Latin hypercube parameter samples (joint ±20% perturbation), the mean confidence score for configuration A, calculated across the 250 synthetic records in each sample, ranged from 0.530 to 0.671 (median 0.596; 5th-95th percentiles 0.558-0.639); the ordering A > C > B > D was preserved in only 27.8% (278/1000) of the joint-perturbation samples; the mean proportion of records changing band was 18.9%. Partial rank correlation (PRCC) identified the caution cap (ρ=0.884 with the full-stack mean) and the high-band cutoff (ρ=−0.573 with the proportion of band changes) as the dominant drivers of output behavior.

Tipping-point sweeps of the 3 band cutoffs showed that band cutoffs changed labels only, never scores or the configuration ranking. The high cutoff was the most consequential: moving it from 0.75 to 0.60 shifted the band distribution by up to 46 percentage points (pp); a 5 pp move down (to 0.70) shifted it by 17.6 pp, and a 5 pp move up (to 0.80) shifted it by 5.2 pp. The moderate and low cutoffs were comparatively stable (≤15.2 pp and ≤6.8 pp across their swept ranges).

Part 3 involved a prespecified empirical calibration protocol. The sensitivity analyses above characterize how outputs respond to parameter changes; they do not select parameter values. Parameter selection against real-world data will follow a prespecified protocol, described next.

Prespecified Empirical Calibration Protocol

The reference configuration's parameters are expert-derived and provisional. Before any clinical use, they will be calibrated against real-world clinical data under the following prespecified protocol, which will be reported in the clinical validation study:

The data will be split into a calibration set and a held-out validation set drawn from the clinical validation cohort, stratified by convergence profile and quality state, with the validation set sealed until calibration is frozen.

The reference standard will be blinded expert clinical adjudication of evidence strength, with adjudicators blinded to all battery scores and evidence stack outputs (preventing incorporation bias).

The objective function will guide parameter selection to maximize agreement between evidence-strength bands and blinded adjudication, operationalized using (1) calibration slope and intercept of score vs adjudicated category, (2) Brier score for band membership, and (3) a calibration curve with visual inspection, following established calibration-assessment practice.

Parameter selection will use constrained grid search (with Bayesian optimization as fallback) while respecting the framework's ordinal constraints: profile base scores must preserve the Table 2 confidence-implication ordering; caps must remain below the band they guard; the bonus must remain nonnegative and smaller than the penalty's magnitude; and band cutoffs must remain ordered (insufficient < low < moderate < high). Any parameter set violating these constraints is inadmissible regardless of fit.

The calibrated configuration will be frozen, versioned, and deposited before validation-set evaluation, with the validation set evaluated exactly once. Recalibration will be triggered only by prespecified triggers (eg, sustained override rates above the audit thresholds defined in the Clinician Override, Deferral, and Audit Protocol section, or a calibration slope outside 0.8-1.25 on monitoring data).

The recalibrated configuration will be accepted only if the validation-set calibration slope falls within 0.85 to 1.15, the Brier score improves on the reference configuration by a

XSL•FO
RenderX

prespecified margin, and the configuration ordering remains interpretable. Otherwise, the reference configuration will remain in force, and the discrepancy will be reported.

This protocol concerns future clinical data; the present paper reports no clinical calibration, and the synthetic analyses above remain configuration-sensitivity descriptions only.

Metrics

The primary metrics included the mean and SD of confidence scores per configuration; 95% bootstrap intervals for mean evidence strength (B=2000), which describe variability of the statistic under empirical resampling of the fixed 250-record synthetic cohort but are not population CIs and do not measure Monte Carlo numerical stability; the proportion of cases in each confidence band; the frequency of reason codes; and threshold-perturbation sensitivity of the full-stack configuration to ±10% changes in quality thresholds.

Ethical Considerations

This manuscript reports a methods description and a synthetic computational evaluation. No human participants were involved in the empirical component, and therefore no research ethics committee (Comitê de Ética em Pesquisa/Comissão Nacional de Ética em Pesquisa [CEP/CONEP]) approval was required for the work reported here, consistent with the scope of Brazilian National Health Council (Conselho Nacional de Saúde) Resolution 510/16, which exempts research that does not involve human participants or identifiable human data. The SPECTRA platform operates under an institutional umbrella research ethics approval held by the NeurAtip/SPECTRA program (Research Ethics Committee of Universidade Christus [Unichristus], CAAE 95653025.2.0000.5049, Parecer 8.397.639, approved April 30, 2026), which governs the platform-wide assessment studies—including the instrument-validation papers for NEXA, VEIL, TEIA, and others—that will provide the real clinical data referenced by this architecture. The present paper reports no participant data under that approval and cites it only as institutional context. The community and clinician input described above was informal, non-systematic consultation—individuals were approached directly for design feedback and lived-experience perspectives, and this activity did not involve systematic collection or analysis of identifiable opinions; it therefore did not constitute human-subjects research and required no ethics determination. Subsequent validation studies that involve human data collection will be submitted for review by a research ethics committee via Plataforma Brasil (CEP/CONEP) before initiation, in accordance with National Health Council (Conselho Nacional de Saúde) Resolutions 466/12 and 510/16 and the Declaration of Helsinki.

The framework incorporates the following safeguards, which govern any future deployment. Informed consent requires independent electronic consent for all data sources, with respondents and informants permitted to withdraw at any time. Data minimization ensures that only data necessary for the specified assessment purpose are collected. Pseudonymization requires identifiers to be separated from responses, with linkage restricted to authorized access. Quality flags trigger support actions rather than negative labels, ensuring that no punitive flags are generated. Human-in-the-loop oversight requires all outputs to be reviewed by a qualified clinician before incorporation into clinical interpretation. The clinician examines confidence bands, reason codes, quality flags, and next-step recommendations; determines whether additional history, informant clarification, or reassessment is needed; and can override or defer the output when it is inconsistent with the broader clinical context. LGPD and GDPR compliance is supported by consent as the legal basis for data processing and encryption of data at rest and in transit.

Foreseeable harms and inappropriate-use scenarios include the possibility that, because the convergence profiles carry clinically suggestive labels (eg, “masking-compatible,” “hypervigilance-compatible”), the framework could, if used without qualified clinical oversight or as a stand-alone decision tool, anchor interpretation, pathologize a respondent, or be mistaken for a diagnostic result. The framework is therefore explicitly designed as decision support within a clinical workflow and is not authorized for stand-alone diagnostic use, for high-stakes determinations (eg, legal, occupational, or custody decisions), or for any use without review by a qualified clinician. The nondiagnostic boundary is enforced by output contracts and reason-code wording, and the next-step recommendations are restricted to evidence-gathering and assessment-process guidance.

Override documentation will follow the intended workflow, in which the reviewing clinician may override or defer the framework’s output when it is inconsistent with the broader clinical context. Overrides and deferrals follow the standardized Clinician Override, Deferral, and Audit Protocol subsection (Methods), which specifies decision types, mandatory deferral triggers, minimum documentation fields, automation-bias safeguards, and a quarterly audit rubric with prespecified action thresholds; the full rubric and log schema are provided in Multimedia Appendix 1. Safety performance, including the rate and appropriateness of overrides, the frequency of clinician disagreement with outputs, and any adverse consequences of use, will be formally evaluated in the planned usability and clinical validation studies before any deployment.

The formative evaluation used synthetic data and therefore cannot speak to demographic fairness. Production calibration will include differential item functioning and measurement invariance analyses across gender, age, education, and neurotype. These analyses are planned for the instrument-specific validation papers in the SPECTRA program.

Results

Study 2 Results: Synthetic Rule-Configuration Comparison

The rule-configuration results are summarized in Table 8. Under the recorded implementation version and seed, configuration A had the highest mean evidence strength. This ordering is a descriptive consequence of the complete configuration definitions and is not a causal estimate of any isolated layer.

XSL•FO
RenderX

Table 8. Synthetic rule-configuration results (N=250 records). Recomputed during revision using the reference implementation; values are reproducible conditional on the implementation version and running the following command "--seed 42 --n 250."

Configuration	Evidence strength, mean (SD)	95% bootstrap interval	Insufficient (<0.25), n (%)	Low (0.25-0.49), n (%)	Moderate (0.50-0.74), n (%)	High (≥0.75), n (%)
A (full stack)	0.60 (0.17)	0.58-0.62	8 (3.2)	59 (23.6)	154 (61.6)	29 (11.6)
B (no quality)	0.53 (0.21)	0.51-0.56	25 (10)	91 (36.4)	87 (34.8)	47 (18.8)
C (no convergence)	0.54 (0.07)	0.53-0.55	8 (3.2)	23 (9.2)	219 (87.6)	0 (0)
D (baseline)	0.50 (0.20)	0.47-0.52	24 (9.6)	115 (46)	85 (34)	26 (10.4)

The low dispersion in configuration C reflects the source_coverage gate's hard ceiling on single-source evidence (Table 3: "single source caps at moderate"), rather than higher measurement precision; consistent with this rule, no configuration C record reaches the high band.

Paired comparisons with configuration A (descriptive contrasts within the fixed synthetic dataset) were constructed as follows. Because the same 250 records were evaluated under each complete configuration, these differences describe configuration-level separation and do not isolate a causal layer effect. The bootstrap intervals describe empirical resampling variability of each statistic within the fixed synthetic cohort, not population uncertainty or numerical stability. The mean difference between configuration A and configuration B (no quality) was 0.07 (paired-difference SD 0.13; 95% bootstrap interval 0.06-0.09). The mean difference between configuration A and configuration C (no convergence) was 0.06 (paired-difference SD 0.15; 95% bootstrap interval 0.05-0.08). The mean difference between configuration A and configuration D (baseline) was 0.11 (paired-difference SD 0.23; 95% bootstrap interval 0.08-0.14).

The bootstrap intervals were computed by resampling the 250 synthetic records with replacement (B=2000), separately for each paired difference, and taking the 2.5th and 97.5th percentiles. We report raw mean differences and paired-difference SDs because this is a seeded comparison of rule-specified outputs on the same fixed records, not an inferential test between independent clinical samples.

Discordance-Sensitive Cases (Illustrative Post Hoc Analysis)

We identified the subset of cases with a discordant convergence profile (masking-compatible, hypervigilance-compatible, or context-dependent; n=113, 45.2%). This subset was defined post hoc by the same convergence profiles the framework detects, so the analysis is illustrative rather than confirmatory. In this subset, the full stack produced higher evidence-strength scores than the no-convergence configuration (full-stack mean 0.61, SD 0.10; no-convergence mean 0.54, SD 0.08; raw mean difference 0.07). This pattern is a direct arithmetic consequence of the framework's design: when sources disagree, the stack's explicit handling of discrepancy preserves evidence strength rather than collapsing it to the single-source baseline or default cap. The effect is directional and method-level; it demonstrates rule execution rather than an empirical clinical effect size.

Interpretation of Rule-Configuration Behavior

The purpose of the full stack is not to maximize evidence strength but to assign it conditionally. In configuration C, no record reached high because the configuration includes a single-source ceiling. Configuration B produced 10% insufficient, 36.4% low, 34.8% moderate, and 18.8% high, with SD 0.21. Configuration D had SD 0.20 and no interpretive layer structure. These patterns describe the four complete configurations. Because configuration B and configuration C also change penalties, jitter, source coverage, or caps, the contrasts do not identify the causal effect of removing a single layer.

These findings are formative and synthetic. They demonstrate intended rule behavior, structural consistency (ie, outputs change monotonically and traceably with changes in inputs and layers), and traceability within the disclosed reference configuration, not clinical validity.

Record-Level Transitions Across Quality States

To make the aggregate rule-configuration results traceable at the record level, Table 9 reports mean evidence strength for the full-stack (A) and no-quality (B) configurations, stratified by the record's quality state (clean, caution, or critical). The quality state reflects the count and severity of activated quality flags in the full-stack configuration.

Table 9. Record-level mean evidence strength by quality state and band (full-stack vs no-quality configurations).

Quality state	Full-stack (A), mean (SD)	No-quality (B), mean (SD)	Records, n
Clean	0.63 (0.17)	0.52 (0.21)	127
Caution	0.60 (0.14)	0.55 (0.22)	115
Critical	0.19 (0.02)	0.50 (0.17)	8

Across the 250 records, the full-stack configuration produced higher mean confidence than the no-quality configuration within the clean (0.63 vs 0.52) and caution (0.60 vs 0.55) states. In the critical state (n=8), the full stack constrained confidence substantially (0.19), whereas comparison configuration B did not discount these degraded records (0.50) because that configuration does not use the response quality layer. This record-level pattern is the mechanism behind the aggregate

XSL•FO
RenderX

finding that configuration B (no quality) produced the largest combined insufficient + low proportion (46.4%) and the highest dispersion (SD 0.21) of any configuration, and it illustrates the intended conditional behavior: degraded records are discounted only when quality gating is active.

Study 3 Results: Clinical Demonstration Vignettes

Overview

Clinical interpretation is best understood through concrete clinical archetypes encountered in adult assessment. Figure 2 presents 3 synthetic vignettes that illustrate how the framework handles common but challenging patterns in adult neurodivergence assessment: masking-compatible divergence, informant-context mismatch, and sparse evidence. These vignettes are not empirical cases; they are constructed demonstrations of the rule paths described in the Methods section, grounded in the formative clinical experience that motivated the architecture.

Figure 2. Three visual clinical demonstration vignettes, each presented as a self-contained case box flowing from top to bottom: input pattern → quality flags → convergence profile → confidence band → reason codes → clinical recommendation. (A) Masking-compatible divergence, in which self-report is lower than informant or performance. (B) Informant-context mismatch, in which the informant observes limited contexts. (C) Sparse evidence, in which only a single source is available and missingness is present. Color coding distinguishes input sources, quality cautions, and confidence levels. All vignettes are synthetic and illustrative.

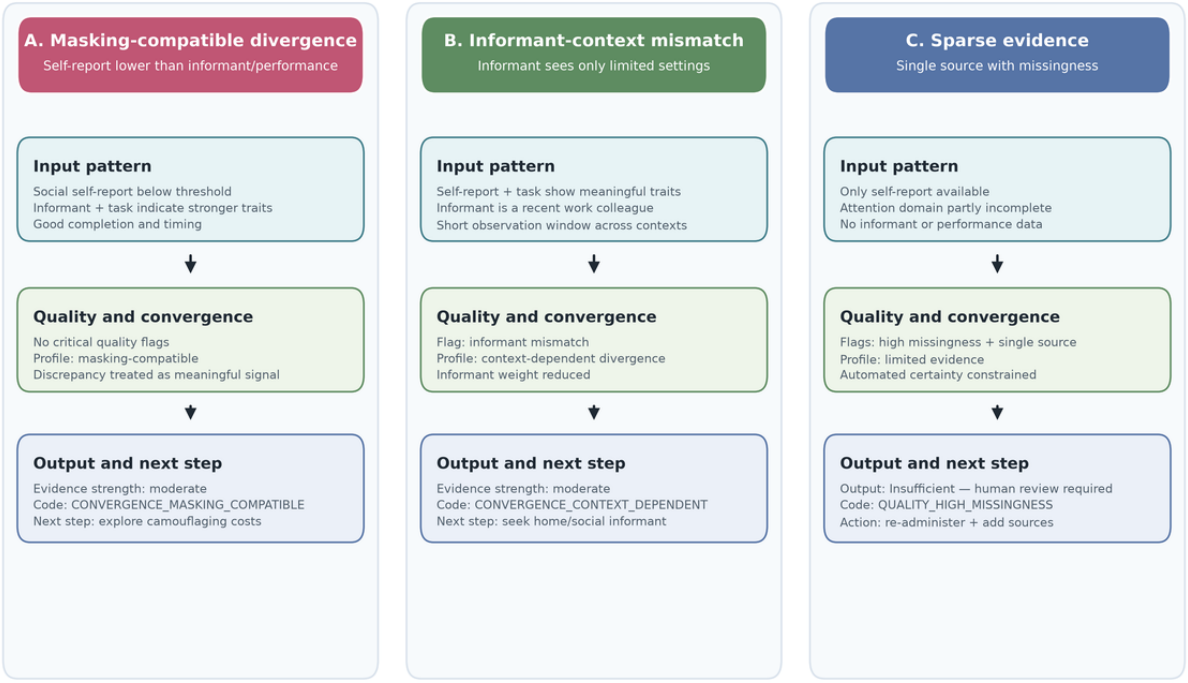


Figure 2 demonstrates how the same architecture produces distinct confidence bands, convergence profiles, and reason codes depending on the input pattern.

Vignette A (masking-compatible divergence) reflects the archetypal clinical presentation in which an individual demonstrates substantial divergence between subjective internal cost and observable external behavior. For instance, self-report indicates lower explicit traits or elevated camouflaging effort alongside high cognitive strain, whereas informant or structured performance metrics suggest preserved functional competence. Rather than hastily concluding that the informant is mistaken or that self-report is inflated or distorted, the framework fires CONVERGENCE_MASKING_COMPATIBLE and constrains confidence to moderate (0.50-0.74). This directs the clinician toward targeted inquiry into the patient's developmental compensation history, postsocial collapse, and internalized masking strategies, without privileging any single source as an absolute diagnostic truth.

Vignette B (informant-context mismatch/context-dependent functioning) illustrates a discrepancy between structured and unstructured settings. In this scenario, an informant who observes the person only in a professional context reports fewer observable difficulties, whereas other sources differ. The framework assigns the context_dependent profile and emits CONVERGENCE_CONTEXT_DEPENDENT, recommending collateral data collection across settings rather than forcing cross-contextual consensus.

Vignette C (sparse evidence) illustrates how incomplete domain coverage and substantial item nonresponse activate QUALITY_HIGH_MISSINGNESS. If the record is also critical because the session ended before mandatory modules, the implementation retains a capped technical score for QA but suppresses the reader-facing score and interpretive band. The displayed output is "Insufficient — human review required"; human review is the recommended action, not a reason code.

Reproducibility Illustration

To show how the rule path is traced, Table 10 presents one worked example using vignette A. Fixed identifiers and transformations follow the published rules; the numerical score

XSL•FO
RenderX

is reproducible conditional on the recorded implementation version and seed. The example traces the synthetic input pattern through quality flags, convergence profile, band, reason codes, and recommended next step. This worked example is synthetic and illustrative; it documents traceability, not clinical validity.

Table 10. Worked example: vignette A (masking-compatible divergence).

Step	Component	Value in vignette A
1	Input pattern	Low self-report; moderate informant and performance evidence
2	Quality flags	No missingness, no straight-lining, no rapid responding, mild fatigue slope
3	Convergence profile	Self-report lower than informant/performance (masking-compatible divergence)
4	Confidence band	Moderate (0.50-0.74)
5	Reason codes	CONVERGENCE_MASKING_COMPATIBLE; QUALITY_FATIGUE_SLOPE
6	Next-step recommendation	Explore camouflaging and context-dependent expression

Sensitivity Analyses

The sensitivity analyses described in the Methods section show the sensitivity of this seeded rule-configuration run to its parameters. Across all twelve ±10% threshold perturbations, the ordering A > C > B > D was preserved; three scenarios changed the band of 1 of 250 records, and 9 changed none (Table 6). In the extended analysis (Table 7), single-parameter ±10% moves preserved the ordering in 38 of 44 scenarios and changed at most 12% of records' bands (caution cap); under joint ±20% perturbation, the ordering was preserved in 27.8% of 1000 samples, with the caution cap and the high-band cutoff identified as the dominant drivers (PRCC ρ=0.884 and −0.573). Tipping-point sweeps showed that the high cutoff is the most consequential label boundary (up to 46 pp distribution shift when moved to 0.60); a 5 pp move down (to 0.70) shifted the distribution by 17.6 pp, and a 5 pp move up (to 0.80) by 5.2 pp. These results characterize parameter sensitivity of the specified implementation; they do not validate the parameters against real-world data, which requires the prespecified calibration protocol described in the Methods section.

Scope Firewall

Textbox 1 summarizes what this paper evaluates and what it does not evaluate. This firewall protects against premature validity claims and aligns with the SPECTRA publication program.

Textbox 1. Scope firewall: assessed and unassessed domains.

Evaluated in this study

- Computational behavior of the framework
- Distribution and stability of confidence outputs
- Traceability of reason codes
- Discrepancy-handling patterns
- Synthetic stress-test behavior
- Transparency of integration logic

Not evaluated in this study

- Diagnostic validity or accuracy
- Clinical sensitivity or specificity
- Instrument-level psychometric properties
- Agreement with clinical gold standards
- Real-world clinical utility end points
- Longitudinal stability or test-retest reliability

Discussion

Principal Results

The main contribution of this paper is methodological: it presents a formally specified and neuroaffirmative framework for integrating heterogeneous assessment signals in adult neurodivergence assessment; traceability is demonstrated within the disclosed reference configuration. By separating response quality, multisource convergence, and confidence synthesis into explicit layers, the SPECTRA Evidence Stack makes integration logic inspectable and disputable. By treating discrepancy as structured signal rather than noise, it aligns assessment technology with contemporary clinical and community understanding of masking and context-dependent expression.

XSL•FO
RenderX

The synthetic stress test was consistent with the intended behavior: the framework assigns evidence strength conditionally rather than maximizing it. In the rule-configuration comparison, configuration A had a mean of 0.60 (SD 0.17), configuration C had a mean of 0.54 (SD 0.07), and configuration D had a mean of 0.50 (SD 0.20). These are descriptive outputs of complete, deliberately specified configurations rather than causal layer effects or findings of clinical superiority. The separation observed among discordance-sensitive cases (n=113, 45.2%) is likewise descriptive. Because the synthetic records were constructed to contain the patterns the framework detects, the findings show rule traceability and programmatic fidelity rather than independent empirical validation. They demonstrate internal consistency and seeded reproducibility within the disclosed reference configuration; they do not establish diagnostic validity or comparative effectiveness.

By providing reason codes and next-step recommendations, the framework is designed to support shared decision-making and reduce the risk that a single low-quality or single-source score drives clinical conclusions. These are design affordances demonstrated at the architectural level; whether they actually facilitate shared decision-making in clinical practice is a question for future usability and implementation studies.

Comparison With Prior Work

Digital phenotyping and mobile health studies in psychiatry often prioritize signal capture and predictive modeling rather than transparent handling of quality or discrepancy across multiple sources of evidence [25]. Some clinical assessment systems include validity flags, but these are often framed as respondent-level judgments and may not feed into confidence estimates. The SPECTRA Evidence Stack extends this work by (1) making quality flags data conditions rather than person judgments, (2) preserving multisource discrepancy as interpretable profiles, and (3) generating mandatory reason codes and next-step recommendations. The novel contribution, as a design specification, lies in combining these 3 elements within a single neuroaffirmative, nondiagnostic architecture for adult neurodivergence assessment, where discrepancy is treated as clinically meaningful signal rather than noise to be averaged away. This design is closer to explainable clinical decision-support systems—such as rule-based sepsis alerts or interpretable early-warning scores that combine multiple data sources into structured, auditable outputs—than to traditional test-scoring pipelines. It also aligns with the broader regulatory principle that transparent documentation and human oversight should be established before claiming clinical utility for AI-enabled clinical decision support [19]. Notably, multi-informant assessment methods have gained traction even in affect-regulation domains: the Alexithymia Questionnaire for Children–Parent version (AQC-P) demonstrated that parent-reported and self-reported alexithymia measures yield meaningful but imperfect concordance, with measurement invariance supported across gender [26]. This finding resonates with the SPECTRA framework's treatment of discrepancy as interpretable signal, suggesting that cross-informant divergence in alexithymia assessment—a feature relevant to autistic populations—should be preserved rather than averaged away.

Formative Clinical Relevance and Human-in-the-Loop Safeguards

The primary value of the SPECTRA Evidence Stack for digital health assessment lies not in providing an automated, closed diagnostic classifier, but in rendering clinical reasoning and assessment uncertainty explicit, structured, and auditable. In conventional digital assessment, the synthesis of conflicting clinical signals typically occurs either inside opaque machine-learning models or through unstandardized clinician heuristics. By contrast, the formative contribution of this architecture is to make visible the precise conditions under which evidence converges or diverges, providing an objective scaffolding that supports rather than supplants clinical expertise.

To maintain ethical and clinical integrity, the framework incorporates 5 foundational human-in-the-loop safeguards. First, uncertainty preservation over forced certainty is maintained. When input data are degraded, missing, or contradictory, the system systematically discounts confidence rather than manufacturing false precision. Second, rule-specified suspension of reader-facing interpretation is applied. Under a critical incomplete session or high missingness, the stack retains a technical score and band for QA but suppresses both values in the reader-facing output. It emits the applicable reason code and displays "Insufficient — human review required"; human review is an action, not a reason code. Third, traceable, nonblack-box scoring is ensured. Fixed transformations and reason-code selection follow published rules, while synthetic generation and score jitter are stochastic and reproducible conditional on the recorded seed and implementation version. Fourth, interpretive draft status is maintained. All generated outputs—including convergence profiles, confidence bands, reason codes, and next-step suggestions—constitute a structured interpretive draft. Sovereign diagnostic judgment and clinical formulation remain exclusively with the licensed health care professional who reviews, contextualizes, and signs the final evaluation. Fifth, standardized override, deferral, and audit are implemented. Divergence between output and clinical judgment is handled by a prespecified protocol—4 decision types, mandatory deferral triggers, minimum documentation fields, explicit automation-bias safeguards (including blind initial impression and mandatory reason-code reading), and a quarterly audit rubric with action thresholds that trigger recalibration—rather than left to unstandardized clinician heuristics (Methods section; Multimedia Appendix 1).

Importantly, the formative clinical experience and operational insights that informed this architecture highlight recurring challenges in adult neurodevelopmental assessment that the framework was engineered to make structured and observable. The vignettes, archetypes, and decision pathways presented here are formative and illustrative; they do not constitute a clinical trial, nor do they establish empirical diagnostic efficacy. To translate this architecture into robust clinical workflows, future prospective research must systematically evaluate clinician usability, cognitive load, user comprehension, and informational safety across real-world practice settings.

XSL•FO
RenderX

Limitations

This study has important limitations. The evaluation used synthetic data constructed to contain the patterns the architecture detects, so the rule-configuration comparison is an internal consistency check rather than an independent test. Because the synthetic cases were intentionally constructed to represent framework-relevant scenarios, they should not be interpreted as approximations of clinical prevalence distributions or real-world patient populations. The synthetic dataset, generation algorithm, and analysis script for the reference configuration are deposited at OSF (code component [21]) to enable independent replication of the numerical results. The architectural rules and reference-configuration parameters are fully specified in this paper and the accompanying code. Flag thresholds, convergence tolerances, and evidence-strength weights were refined during revision through a structured expert-informed process. The final values were frozen before the reported execution but were not globally preregistered or established before peer review. This process was led by a psychiatrist with expertise in adult neurodevelopmental assessment and reviewed by the coauthors, with additional consultative input acknowledged below. The band cutoffs and quality thresholds require empirical calibration; a prespecified calibration protocol (data split, blinded reference standard, calibration metrics, constrained optimization, and freeze and acceptance criteria) is now provided in the Methods section and will be executed in the clinical validation study, following established measurement-property standards such as COSMIN [23]. The extended sensitivity analysis identified the caution cap and the high-band cutoff as the parameters with the largest influence on outputs; these will be prioritized in calibration. The joint-perturbation results also show that the configuration ordering is not invariant under large simultaneous parameter moves (preserved in 27.8% of ±20% samples), reinforcing that the reported ordering is a property of the reference configuration rather than a universal law. Relatedly, the response quality metrics used in the present framework rely on multiitem composites; brief or single-item measures—which have been identified as a pervasive but underscrutinized practice in psychological assessment [27]—were deliberately avoided to preserve measurement precision across the response quality, multisource convergence, and confidence synthesis layers. The framework has not been tested beyond its source language, and the neuroaffirmative framing has not yet been evaluated with autistic users. Furthermore, the multisource architecture presupposes the availability of informant reports and performance-based measures, which may be challenging in resource-constrained public health settings where informant availability is limited and structured performance tasks are not routinely administered. The feasibility and cost-effectiveness of deploying the full evidence stack in such contexts remain unexamined. Diagnostic accuracy, clinical utility, and subgroup fairness are reserved for subsequent studies. No formal preregistration was performed for this formative framework-development study; the study was designed as an early-stage methodological specification and synthetic stress test rather than a confirmatory hypothesis-testing study.

Future Directions

The two most urgent next steps are: (1) a clinical validation study comparing evidence stack outputs with independent gold-standard diagnostic adjudication performed by clinicians blinded to all battery and evidence stack outputs, reported in accordance with DECIDE-AI and the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines; and (2) usability studies with clinicians and autistic adults to evaluate the interpretability of convergence profiles, reason codes, and next-step recommendations, as well as the training requirements for intended users. The blinded adjudication procedure is essential to prevent incorporation bias: if the adjudicating clinician has access to the instrument scores or evidence stack output, the reference standard is not independent and accuracy estimates become uninterpretable. Clinical validation is prioritized because the framework's intended value depends on whether its outputs correspond with expert clinical judgment obtained under blinded conditions; usability studies will then determine whether the outputs are interpretable and actionable in practice. After these steps, the program will proceed with instrument-specific validation studies of AURA, NEXA, VEIL, and TROPE; longitudinal stability analyses using TEA-10; and fairness and measurement invariance analyses across gender, age, education, and neurotype. Because the stack's layer contracts separate decision logic from parameter values, future iterations could substitute learned components (for example, machine learning–based item scoring or ensemble adjudication) wherever they demonstrably improve calibration without sacrificing traceability—a path toward explainable, traceable AI-mediated assessment that is the subject of ongoing development and will be reported separately.

Open Science

The rule logic, parameter values, pseudocode, synthetic dataset, and analysis script that support the findings of this study are deposited at OSF (code component [21]). The Methods section describes how the records were constructed to enable conceptual replication. Online supplementary material is provided with this submission: Multimedia Appendix 2 (DECIDE-AI prospective-alignment checklist) and Multimedia Appendix 1 (clinician override, deferral, and audit protocol).

An earlier version of this manuscript was deposited as a preprint at OSF [24]. The reference configuration and outputs reported here were refined and recomputed during peer review in response to editorial and reviewer comments. The historical preprint record may retain an earlier title or earlier results and should not be read as already containing every revision reported here.

This research was not preregistered. The present work is an early-stage methodological specification and synthetic stress test rather than a confirmatory hypothesis-testing study.

Conclusions

The SPECTRA Evidence Stack offers a transparent, traceable within the disclosed reference configuration, and neuroaffirmative conceptual architecture for multisource adult neurodivergence assessment. It makes uncertainty visible, preserves discrepancy as interpretable evidence, and is designed

XSL•FO
RenderX

to support transparent and collaborative clinical decision-making, pending future clinical evaluation. This formative methods paper establishes the architecture and demonstrates conservative, traceable behavior under varying evidence conditions, as summarized in the scope firewall (Textbox 1). It does not establish diagnostic accuracy; that evaluation is reserved for a subsequent clinical validation study. We offer the framework as a contribution to the growing literature on explainable, ethical, and community-informed digital health assessment.

Acknowledgments

The authors thank autistic community members and clinicians whose informal design feedback and lived-experience perspectives informed the framing of the framework.

Generative AI Use

From July through November 2026, the authors used GitHub Copilot, Anthropic Claude (including Claude Sonnet in recoverable interactions), and ChatGPT (OpenAI) during manuscript preparation and revision (Multimedia Appendix 3). Exact historical model snapshots could not be recovered for every interaction. The tools were used for language editing, structural organization, formatting, drafting and critique of reviewer responses, consistency auditing, and critical review. They did not execute the analyses reported in this article or make autonomous scientific or editorial decisions. The authors reviewed, modified, accepted, or rejected all suggestions; checked suggested citations against primary sources; and independently verified AI-assisted scientific statements and reported numbers. Calculations were generated with the versioned reference implementation described in the Methods section. No participant data were used in this paper. AI tools are not listed as authors. A categorized log and representative paraphrased examples are provided as Multimedia Appendix 3; this material is not represented as a complete historical record, and the examples are not verbatim quotations or reconstructed historical transcripts. Redactions are limited to personal, security-sensitive, third-party, internal-workflow, and proprietary information unrelated to evaluating the AI contribution. Complete raw records are not published because they contain irrelevant or confidential material, file paths, internal governance information, and potentially proprietary or third-party content. Retained manuscript-relevant records can be supplied confidentially to the editor after appropriate sanitization. During copyediting, the tools were also used for consistency checking and drafting responses to copyeditor queries.

Funding

The authors declared no financial support was received for this work.

Data Availability

The rule logic, parameter values, pseudocode, synthetic dataset, and analysis script for the reference configuration are deposited at OSF (code component [21]; with the associated preprint also available at OSF [24]). The reference implementation and its inputs are openly downloadable without a request. The extended-sensitivity analysis script and its captured output are available in the OSF code component [21]. Production calibration values are proprietary to NeurAtip and are outside the scope of this paper; they are not required to reproduce any result reported here, and no claim of access to them is made. The present study analyzed synthetic records only; no participant data were collected or analyzed.

Authors' Contributions

LF: Conceptualization; Methodology; Software; Formal Analysis; Investigation; Data Curation; Writing – Original Draft; Writing – Review & Editing; Project Administration; Resources; Visualization.
LSL: Writing – Review & Editing.
GMBT: Formal Analysis; Writing – Review & Editing.
CC: Writing – Review & Editing.
AMLNB: Supervision; Writing – Review & Editing.

Conflicts of Interest

LF is the founder and lead developer of the NeurAtip/SPECTRA platform and holds equity in NeurAtip. He developed the instruments described in this manuscript as part of the NeurAtip proprietary assessment ecosystem. The SPECTRA platform is proprietary technology of NeurAtip. CC is a partner and equity holder in NeurAtip and may benefit financially from the results of this research. LSL, GMBT, and AMLNB declare no conflicts of interests.

Multimedia Appendix 1

Clinician override, deferral, and audit protocol: full audit rubric, override log schema, and prespecified action thresholds.
[DOCX File , 24 KB-Multimedia Appendix 1]

XSL•FO
RenderX

Multimedia Appendix 2

DECIDE-AI prospective-alignment checklist (Vasey et al [20]; 2022), provided as a roadmap for the framework's future clinical evaluation.
[DOCX File , 25 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Generative AI use during manuscript preparation, revision, and copyediting: categorized activities and representative paraphrased examples. The examples are not verbatim quotations or reconstructed historical transcripts.
[DOCX File , 39 KB-Multimedia Appendix 3]

References

1. Lai M, Baron-Cohen S. Identifying the lost generation of adults with autism spectrum conditions. Lancet Psychiatry. 2015;2(11):1013-1027. [doi: 10.1016/S2215-0366(15)00277-1] [Medline: 26544750]
2. Lewis LF. Realizing a diagnosis of autism spectrum disorder as an adult. Int J Ment Health Nurs. 2016;25(4):346-354. [doi: 10.1111/inm.12200] [Medline: 26940281]
3. Hull L, Petrides KV, Allison C, Smith P, Baron-Cohen S, Lai M, et al. "Putting on my best normal": social camouflaging in adults with autism spectrum conditions. J Autism Dev Disord. 2017;47(8):2519-2534. [FREE Full text] [doi: 10.1007/s10803-017-3166-5] [Medline: 28527095]
4. Loomes R, Hull L, Mandy WPL. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry. 2017;56(6):466-474. [doi: 10.1016/j.jaac.2017.03.013] [Medline: 28545751]
5. Kember SM, Williams MN. Autism in Aotearoa: is the RAADS-14 a valid tool for a New Zealand population? European Journal of Psychological Assessment. 2021;37(3):247-257. [doi: 10.1027/1015-5759/a000598]
6. Cage E, Troxell-Whitman Z. Understanding the reasons, contexts and costs of camouflaging for autistic adults. J Autism Dev Disord. 2019;49(5):1899-1911. [FREE Full text] [doi: 10.1007/s10803-018-03878-x] [Medline: 30627892]
7. Hull L, Mandy W, Lai M, Baron-Cohen S, Allison C, Smith P, et al. Development and validation of the camouflaging autistic traits questionnaire (CAT-Q). J Autism Dev Disord. 2019;49(3):819-833. [FREE Full text] [doi: 10.1007/s10803-018-3792-6] [Medline: 30361940]
8. Hull L, Levy L, Lai M, Petrides KV, Baron-Cohen S, Allison C, et al. Is social camouflaging associated with anxiety and depression in autistic adults? Mol Autism. 2021;12(1):13. [FREE Full text] [doi: 10.1186/s13229-021-00421-1] [Medline: 33593423]
9. Kanne SM, Abbacchi AM, Constantino JN. Multi-informant ratings of psychiatric symptom severity in children with autism spectrum disorders: the importance of environmental context. J Autism Dev Disord. 2009;39(6):856-864. [FREE Full text] [doi: 10.1007/s10803-009-0694-7] [Medline: 19191016]
10. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323. [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
11. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
12. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform. 2020;8(7):e15182. [FREE Full text] [doi: 10.2196/15182] [Medline: 32673244]
13. Botha M, Frost DM. Extending the minority stress model to understand mental health problems experienced by the autistic population. Society and Mental Health. 2020;10(1):20-34. [doi: 10.1177/2156869318804297]
14. Cascio MA, Weiss JA, Racine E, the Autism Research Ethics Task Force. Person-oriented ethics for autism research: creating best practices through engagement with autism and autistic communities. Autism. 2020;24(7):1676-1690. [doi: 10.1177/1362361320918763] [Medline: 32551887]
15. Kapp SK, Gillespie-Lynch K, Sherman LE, Hutman T. Deficit, difference, or both? Autism and neurodiversity. Dev Psychol. 2013;49(1):59-71. [doi: 10.1037/a0028353] [Medline: 22545843]
16. Kenny L, Hattersley C, Molins B, Buckley C, Povey C, Pellicano E. Which terms should be used to describe autism? Perspectives from the UK autism community. Autism. 2016;20(4):442-462. [doi: 10.1177/1362361315588200] [Medline: 26134030]
17. Fletcher-Watson S, Adams J, Brook K, Charman T, Crane L, Cusack J, et al. Making the future together: shaping autism research through meaningful participation. Autism. 2019;23(4):943-953. [FREE Full text] [doi: 10.1177/1362361318786721] [Medline: 30095277]
18. Pellicano E, den Houting J. Annual research review: shifting from 'normal science' to neurodiversity in autism science. J Child Psychol Psychiatry. 2022;63(4):381-396. [FREE Full text] [doi: 10.1111/jcpp.13534] [Medline: 34730840]

XSL•FO
RenderX

19. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. URL: https://www.fda.gov/media/153486/download [accessed 2026-11-09]
20. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
21. The SPECTRA Evidence Stack: a rule-specified neuroaffirmative framework for integrating multi-source evidence in adult neurodivergence assessment. Open Science Framework. URL: https://osf.io/e2rhj/ [accessed 2026-10-30]
22. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res. 2011;13(4):e126. [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
23. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1171-1179. [FREE Full text] [doi: 10.1007/s11136-017-1765-4] [Medline: 29260445]
24. Fortaleza L, Landim LS, Coleman C. The SPECTRA Evidence Stack: a rule-specified neuroaffirmative framework for integrating multi-source evidence in adult neurodivergence assessment. PsyArXiv. Preprint posted online on June 24, 2026. [FREE Full text] [doi: 10.31234/osf.io/ebn98_v1]
25. Birk RH, Samuel G. Digital phenotyping for mental health: reviewing the challenges of using data to monitor and predict mental health problems. Curr Psychiatry Rep. 2022;24(10):523-528. [doi: 10.1007/s11920-022-01358-9] [Medline: 36001220]
26. Brown RH, Murray AL, Stewart ME, Auyeung B. Psychometric validation of a parent-reported measure of childhood alexithymia: the Alexithymia Questionnaire for Children–Parent (AQC-P). Eur J Psychol Assess. 2022;38(1):24-31. [doi: 10.1027/1015-5759/a000640]
27. Allen MS, Iliescu D, Greiff S. Single item measures in psychological science: a call to action. Eur J Psychol Assess. 2022;38(1):1-5. [doi: 10.1027/1015-5759/a000699]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
AQC-P: Alexithymia Questionnaire for Children–Parent version
AURA: Autism Understanding by Relative Assessment
CEP/CONEP: Comitê de Ética em Pesquisa/Comissão Nacional de Ética em Pesquisa
COSMIN: Consensus-based Standards for the selection of health Measurement Instruments
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Using Artificial Intelligence
GDPR: General Data Protection Regulation
LGPD: Lei Geral de Proteção de Dados Pessoais
NEXA: Neurodivergent Expression Assessment
OAT: one-at-a-time
OSF: Open Science Framework
pp: percentage points
PRCC: partial rank correlation
QA: quality assurance
RAADS-14: Ritvo Autism Asperger Diagnostic Scale–14 Screen
SPECTRA: Spectrum & Clinical Tools for Reliable Assessment
SPECTRA-CI: SPECTRA Clinical Interview
STARD: Standards for Reporting of Diagnostic Accuracy Studies
TEA-10: Autism Spectrum Test–10 Items
TROPE: Textual Recognition of Pragmatic Expressions
VEIL: Veiling Effort & Identity Load

XSL•FO
RenderX

Edited by J Sarvestan; submitted 12.Jul.2026; peer-reviewed by A Shateri; comments to author 01.Sep.2026; revised version received 07.Oct.2026; accepted 23.Oct.2026; published 12.Nov.2026

Please cite as:
Fortaleza L, Landim LS, Teixeira GMB, Coleman C, Luna Neri Benevides AM
The Spectrum & Clinical Tools for Reliable Assessment (SPECTRA) Evidence Stack for Adult Neurodivergence Assessment: Framework Development and Synthetic Data Evaluation
JMIR Form Res 2026;10:e106266
URL: https://formative.jmir.org/2026/1/e106266
doi: 10.2196/106266
PMID: 42925227

©Lucas Fortaleza, Lenifer Siqueira Landim, Giovana Melo Bezerra Teixeira, Charlene Coleman, Ana Maria Luna Neri Benevides. Originally published in JMIR Formative Research (https://formative.jmir.org), 12.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX