A-CSM: AI Contextual Signal Matrix

Engineering Implementation of the
Vært Conversational Context Architecture

ZON RZVN
ORCID: 0009-0002-6597-7245
Independent Researcher, Taiwan
zon@rzvn.io

9 March 2026 | Version 0.1.0

Field	Value
License	CC BY-NC-ND 4.0
DOI	10.5281/zenodo.19097267
Type	Technical Report (non-peer-reviewed)

Document Classification: This document is a technical report describing an engineering implementation of theoretical frameworks developed by the author. It has not undergone formal peer review. All claims are pre-empirical unless stated otherwise. No human-subject validation has been conducted.

Contents

1 Executive Summary and Document Positioning

This report presents the A-CSM (AI Contextual Signal Matrix), an engineering implementation framework that translates four preceding theoretical models into a structured, reproducible assessment pipeline for detecting user-side contextual integrity risks in conversational AI systems. The A-CSM is non-clinical, context-first, and transparency-oriented, operating within the Vært (ASCII: Vaert) Conversational Context Architecture lineage.

The theoretical lineage proceeds as follows. CXC-7 established a seven-dimensional context field model capturing Emotion and Attachment, Framing and Discursive Power, Ethical and Safety Boundaries, System and Interface Transformation, Prompt Ecosystem, Social Diffusion and Culture, and Transparency and Auditability as axes of human-AI interaction (RZVN, 2025a). CXOD-7 extended this into a system-layer orchestration model with Coh(G) as a coherence function (RZVN, 2025b). The USCH framework identified fourteen user-side phenomena organized by three judgment boundaries (Origin Cut, Grounding Cut, Control Cut) and supplemented by the VCD (Vært Context Defense) specification (RZVN, 2025c). USCI formalized a four-axis risk assessment instrument using 0 to 4 semantic anchors (RZVN, 2025d).

The A-CSM addresses the gap between these theoretical constructs and verifiable software implementation. Its scope encompasses signal detection, multi-axis scoring, evidence chain construction, and post-assessment action recommendations derived from the VCD specification. The framework does not perform clinical diagnosis, does not replace professional mental health assessment, and does not make claims about user psychological states beyond observable conversational patterns.

2 Theoretical Architecture Overview

2.1 CXC-7: Seven-Dimensional Context Field

The CXC-7 model defines seven orthogonal dimensions that characterize the contextual field within which human-AI conversations occur (RZVN, 2025a). The non-linear principle central to CXC-7 asserts that these seven dimensions interact through feedback loops rather than additive linear summation.

Dimension E: Emotion and Attachment. This axis captures affective tone, emotional valence shifts, and attachment formation patterns across conversational turns. It tracks how emotional states propagate through dialogue sequences and influence subsequent user expectations, including the development of affective bonds toward the AI system.

Dimension F: Framing and Discursive Power. This dimension examines how framing effects and discursive power dynamics operate within human-AI dialogue. It addresses who controls the narrative frame and how discursive asymmetry can influence the direction of conversation.

Dimension B: Ethical and Safety Boundaries. Boundary integrity represents the degree to which ethical constraints and safety limits are maintained throughout the interaction. This axis tracks boundary erosion and the system's capacity to uphold predefined ethical guardrails under conversational pressure.

Dimension S: System and Interface Transformation. This dimension models how the AI system's interface characteristics and functional presentation transform during extended interaction. It captures shifts in perceived system identity and capability framing.

Dimension P: Prompt Ecosystem. The prompt ecosystem axis tracks how prompt engineering prac-

tices, user instruction patterns, and meta-conversational strategies evolve across sessions. It examines the emergent dynamics of prompt-response co-evolution.

Dimension D: Social Diffusion and Culture. This axis captures how conversational AI interaction patterns diffuse through social networks and cultural contexts. It addresses communal norms, shared prompt strategies, and the cultural embedding of AI interaction practices.

Dimension T: Transparency and Auditability. This dimension encompasses the degree to which the AI system's operations, decision processes, and limitations are visible and verifiable by the user and by external auditors.

2.2 CXOD-7: System-Layer Orchestration

Where CXC-7 models the user-side context field, CXOD-7 describes the system-side orchestration layer that governs AI behavior within that field (RZVN, 2025b). The seven orchestration dimensions are Mode, Context, Rules, Knowledge, Personality, Role, and Safety.

The coherence function Coh(G) quantifies alignment across these seven system dimensions. The canonical formula is:

$$\text{Coh}(G) = \frac{\sum_{i=1}^{7} D_i}{7} \times A \times R \tag{1}$$

where D_i represents the state value for each orchestration dimension, A is a global alignment modifier, and R is a reliability modifier.

Implementation note: The A-CSM coherence monitor adapts Coh(G) into a pairwise distance metric for engineering purposes. The canonical formula $(\sum_{i=1}^{7} D_i/7) \times A \times R$ remains the authoritative theoretical definition; the distance-based adaptation is an implementation-level approximation subject to future empirical validation.

2.3 USCH: User-Side Contextual Hallucination

The USCH framework identifies fourteen distinct phenomena that emerge when users develop distorted contextual expectations during extended AI interactions (RZVN, 2025c). These phenomena are organized into a six-stage formation process:

1. Initiation
2. Emotional Value Acquisition
3. Trust in Micro-Hallucinations
4. Accumulation and Solidification
5. Cognitive Misalignment
6. Structured Hallucination

The Three Cuts judgment boundary model provides stage differentiation criteria:

1. **Origin Cut:** Examines the origin of the user's contextual expectations. Can the user trace their beliefs about the AI system to accurate sources?

2. **Grounding Cut:** Evaluates the user's ability to ground their interaction in verified reality. Can the

user distinguish AI-generated output from factual information?

3. **Control Cut:** Assesses the user's capacity to maintain volitional control over the interaction. Can the user self-regulate and disengage when appropriate?

The VCD (Vært Context Defense) specification defines four defense constructs and three dialogue operations:

Defense Constructs:

- **Role Reframing (VCD_RR):** Reconstructing the user's framing of the AI's role.
- **Contextual Cognition (VCD_CC):** Activating the user's awareness of the contextual nature of the interaction.
- **Contextual Misalignment Detection (VCD_MD):** Identifying divergence between the user's contextual model and system capabilities.
- **Context Decomposition (VCD_CD):** Breaking down complex contextual states into separable components.

Dialogue Operations:

- **Archiving (OP_AR):** Preserving the session with full evidence chains.
- **Termination (OP_TR):** Ending the session with notification, cool-down, and summary.
- **Redirection (OP_RD):** Guiding the user toward appropriate human resources.
- **Escalation (OP_ES):** Flagging for human review.

2.4 USCI: User-Side Contextual Integrity Index

USCI operationalizes the USCH taxonomy into a structured risk assessment instrument (RZVN, 2025d). Four axes provide scoring dimensions, each using a 0 to 4 semantic anchor scale:

- **FR (Fact Reliability Risk):** Degree of conflation between AI-generated and verified factual content.
- **CA (Context Alignment Risk):** Divergence between the user's contextual model and the system's actual scope.
- **SR (User-side Safety Risk):** Compromise of self-regulation, autonomous decision-making, and independent judgment.
- **SA (System Usability Risk):** System characteristics contributing to user-side risk accumulation.

Table 1: USCI Semantic Anchor Scale

Score	Anchor	Interpretation
0	No observable risk	No signals detected on this axis
1	Minimal risk	Isolated signals, self-correcting patterns
2	Moderate risk	Recurring signals, partial self-correction
3	Elevated risk	Persistent signals, limited self-correction
4	Severe risk	Dominant signals, no observable self-correction

Three Primary States:

- **ST_NRM (Normal):** All axes score 0 to 1, no collapse flag active.
- **ST_DEV (Deviation):** Any axis scores 2 to 3, without meeting Alert criteria.
- **ST_ALM (Alert):** Any axis scores 4, or any two axes score ≥ 3, or any collapse flag active.

Five Behavioral Subtypes:

- **SUB_NRM:** Balanced-Normal.
- **SUB_FCT:** Fact Misalignment.
- **SUB_CTX:** Context Loss-of-Control.
- **SUB_DEP:** Dependency Outsourcing.
- **SUB_SYS:** System Failure or Uncontrollable.

Collapse Flags:

- **ST_CC (Context Collapse):** User's contextual model becomes internally incoherent.
- **ST_SC (System Collapse):** System cannot maintain coherent output or safety functions.

Canonical dual-condition logic: Collapse flags in the canonical USCI specification require satisfaction of both a quantitative threshold condition and a qualitative behavioral indicator. The current A-CSM implementation primarily evaluates the quantitative condition; qualitative behavioral verification remains a target for future calibration.

2.5 Architecture Integration

Figure 1 presents the five-layer architecture integration from context field to implementation.

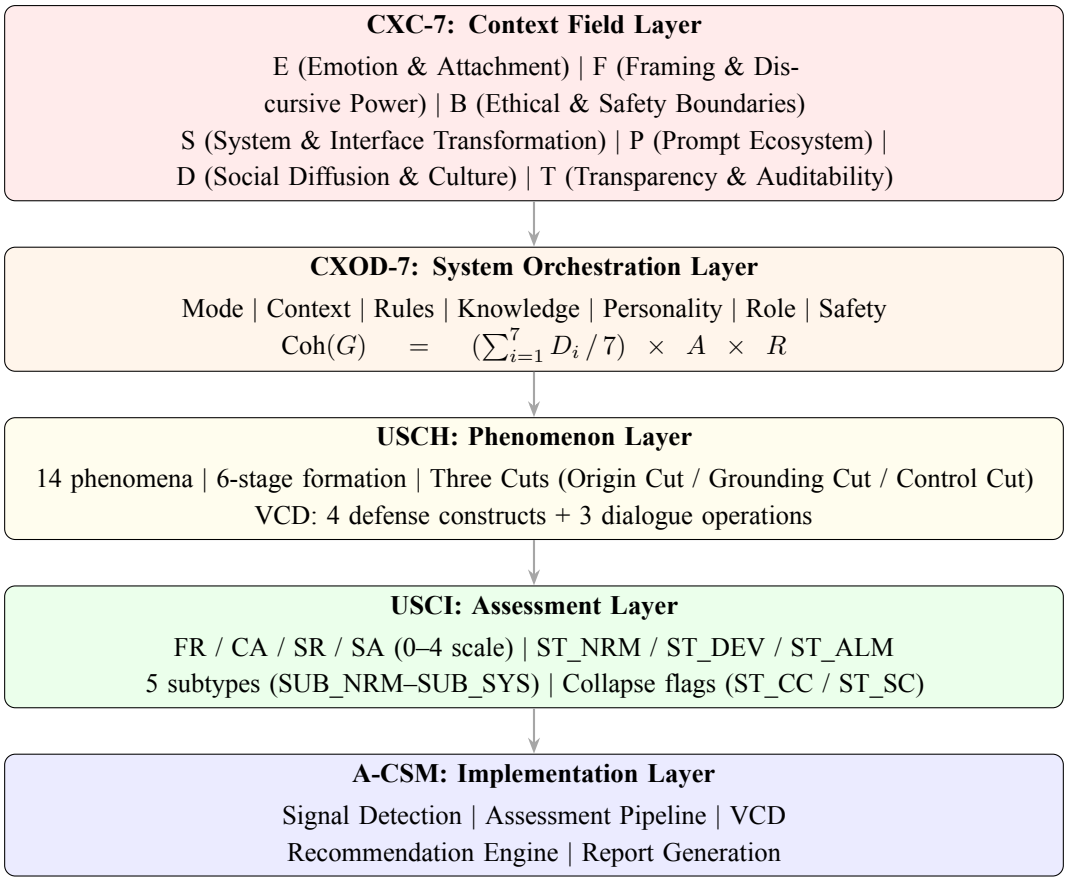

Figure 1: Architecture integration from CXC-7 context field through A-CSM implementation.

3 A-CSM: Design Rationale and Architecture

3.1 Problem Statement

The four theoretical frameworks described in Section 2 provide a descriptive and analytical vocabulary for understanding user-side risks in conversational AI. They do not, by themselves, constitute runnable software. The A-CSM exists to close the gap between theoretical constructs and software functions that detect their occurrence in conversation logs.

Three specific problems motivate the design. First, theoretical concepts must be decomposed into observable, machine-detectable signals within conversation transcripts. Second, multi-dimensional scoring must preserve the non-linear interaction principles of CXC-7. Third, post-assessment recommendations must follow deterministic rules derived from the VCD specification.

3.2 Design Principles

1. **Context-first.** All signal detection begins with contextual field analysis, not isolated keyword matching. Every detected signal must reference at least one CXC-7 dimension.

2. **Non-clinical.** The system does not diagnose mental health conditions, does not replace professional assessment, and does not use clinical terminology.

3. **User-side safety.** The framework prioritizes protecting users from contextual integrity degradation.

4. **Transparency.** Every assessment score must be traceable to specific turns, detected signals, and defining theoretical constructs.

5. **Deterministic action logic.** VCD recommendations follow predefined threshold rules.

3.3 System Architecture

Figure 2 presents the four-layer processing architecture of the A-CSM pipeline.

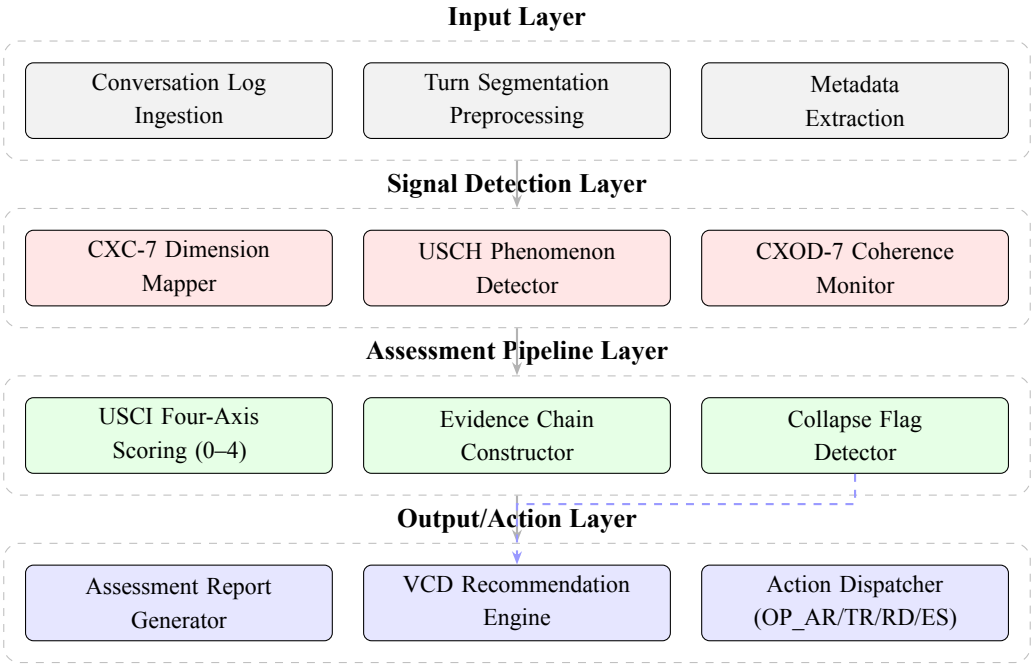

Figure 2: A-CSM system architecture: four-layer processing pipeline.

3.4 Signal Taxonomy

The A-CSM signal taxonomy maps each of the fourteen USCH phenomena to detectable conversational signals. Phenomenon names follow the canonical USCH specification (RZVN, 2025c). The phenomena are organized by three judgment boundary layers.

Cognitive Layer Phenomena (7):

ID	USCH Canonical Name	Stage	A-CSM Signal Indicators
P01	Accelerated Anthropomorphism	1	Phrases attributing comprehension or consciousness to AI
P02	Intentionality Projection	1	Attributing goals or deliberate intent to AI responses
P03	Memory Continuity Illusion	2	References to prior sessions as shared memory
P04	Overtrust and Capability Boundary Miscalibration	2	Accepting AI claims beyond verified capability range

ID	USCH Canonical Name	Stage	A-CSM Signal Indicators
P05	Reality Baseline Drift	3	Gradual acceptance of AI framings as ground truth
P06	Confirmation Bias Amplification	3	Selectively engaging outputs that confirm existing beliefs
P07	Context-Misalignment Numbing	3	Decreased sensitivity to discrepancies between AI output and reality

Dependence Layer Phenomena (4):

ID	USCH Canonical Name	Stage	A-CSM Signal Indicators
P08	Use-Dependence Escalation	4	Escalating session frequency and scope delegation
P09	Reinforcement-Driven Engagement and Tolerance	4	Increasing engagement driven by positive reinforcement
P10	Affective Regulation Outsourcing	4	Using AI as primary emotional regulation mechanism
P11	Social Withdrawal and Relational Cost Avoidance	5	Substituting AI interaction for human social contact

Decision Layer Phenomena (2):

ID	USCH Canonical Name	Stage	A-CSM Signal Indicators
P12	Self-Inflation and Closed Confidence	5	AI-reinforced overconfidence; closed epistemic loop
P13	Internal Gatekeeping Replacement	6	Delegating judgment and moral reasoning to AI

Cross-Layer Phenomenon (1):

ID	USCH Canonical Name	Stage	A-CSM Signal Indicators
P14	Affective Mirroring Illusion	1–6	Perceiving AI emotional responses as genuine affect

3.5 Assessment Pipeline

The assessment pipeline implements USCI scoring through a three-phase process.

Phase 1: Turn-Level Scanning. Each conversational turn is analyzed against the signal taxonomy. Detected signals are tagged with phenomenon ID, CXC-7 dimension mapping, and confidence weight.

Phase 2: Axis Score Computation. Detected signals are aggregated along the four USCI axes (FR, CA, SR, SA) using a 0 to 4 semantic anchor scale. State determination follows canonical USCI definitions: ST_NRM when all axes score 0 to 1 with no collapse flags; ST_ALM when any axis scores 4, any two axes score ≥ 3, or any collapse flag is active; ST_DEV otherwise.

Phase 3: Evidence Chain Construction. For each axis scoring 2 or above, the system constructs an evidence chain linking the score to specific turns, detected signals, and the defining theoretical construct.

Collapse flags (ST_CC, ST_SC) are evaluated against quantitative threshold conditions. The canonical dual-condition logic requires both quantitative and qualitative criteria; the current implementation evaluates quantitative conditions only, with qualitative verification planned for future calibration.

3.6 VCD Action Integration

The VCD (Vært Context Defense) specification defines a post-assessment recommendation layer. VCD operates as a downstream consumer of USCI assessment output; it does not overwrite or modify USCI axis scores.

Table 6: VCD Defense Construct Recommendations

Code	Construct	Trigger Condition
VCD_RR	Role Reframing	CA axis $\geq$ 2 with role confusion signals
VCD_MD	Misalignment Detection	Any axis $\geq$ 2 with misalignment indicators
VCD_CD	Context Decomposition	Multiple axes $\geq$ 2 simultaneously

Table 7: VCD Dialogue Operation Recommendations

Code	Operation	Trigger Condition
OP_RD	Redirection	SR axis ≥ 3
OP_AR	Archiving	Any axis enters ST_DEV
OP_TR	Termination	Any collapse flag active
OP_ES	Escalation	Beyond automated response scope

Architectural note: VCD functions as a recommendation layer, not an enforcement layer. The A-CSM generates VCD action codes as advisory outputs. Execution depends on the deployment context and is outside this framework's scope.

4 Theory-to-Implementation Traceability

4.1 Traceability Matrix

Table 8 provides a full traceability matrix. All constructs without direct implementation are marked with their status.

Table 8: Theory-to-implementation traceability matrix.

Source	Construct	Module	Function	Status
CXC-7	Dim E (Emotion & Attachment)	Signal Detection	Affective tone tracker	Impl. (pre-val.)
CXC-7	Dim F (Framing & Discursive Power)	Signal Detection	Framing asymmetry detector	Impl. (pre-val.)
CXC-7	Dim B (Ethical & Safety Boundaries)	Signal Detection	Boundary integrity monitor	Impl. (pre-val.)
CXC-7	Dim S (System & Interface Transform.)	Signal Detection	Interface shift tracker	Impl. (pre-val.)
CXC-7	Dim P (Prompt Ecosystem)	Signal Detection	Prompt evolution monitor	Impl. (pre-val.)
CXC-7	Dim D (Social Diffusion & Culture)	Signal Detection	Cultural context analyzer	Impl. (pre-val.)
CXC-7	Dim T (Transparency & Auditability)	Signal Detection	Audit trail monitor	Impl. (pre-val.)
CXOD-7	Mode	Coherence Monitor	Mode state validator	Impl. (pre-val.)
CXOD-7	Context	Coherence Monitor	Context boundary checker	Impl. (pre-val.)
CXOD-7	Rules	Coherence Monitor	Rule consistency verifier	Impl. (pre-val.)
CXOD-7	Knowledge	Coherence Monitor	Knowledge boundary monitor	Impl. (pre-val.)
CXOD-7	Personality	Coherence Monitor	Persona consistency tracker	Impl. (pre-val.)
CXOD-7	Role	Coherence Monitor	Role alignment validator	Impl. (pre-val.)
CXOD-7	Safety	Coherence Monitor	Safety policy monitor	Impl. (pre-val.)
CXOD-7	Coh(G)	Coherence Monitor	Distance-metric adaptation	Impl. (pre-val.)
USCH	P01 Accel. Anthro-pomorphism	Phenom. Detector	Anthropomorphism matcher	Impl. (pre-val.)
USCH	P02 Intentionality Projection	Phenom. Detector	Intent attribution scanner	Impl. (pre-val.)
USCH	P03 Memory Continuity Illusion	Phenom. Detector	Memory reference detector	Impl. (pre-val.)

Source	Construct	Module	Function	Status
USCH	P04 Overtrust & Cap. Miscalibration	Phenom. Detector	Trust calibration monitor	Impl. (pre-val.)
USCH	P05 Reality Baseline Drift	Phenom. Detector	Reality-checking detector	Impl. (pre-val.)
USCH	P06 Confirmation Bias Amplification	Phenom. Detector	Selective engagement detector	Impl. (pre-val.)
USCH	P07 Context-Misalignment Numbing	Phenom. Detector	Misalignment sensitivity tracker	Impl. (pre-val.)
USCH	P08 Use-Dependence Escalation	Phenom. Detector	Frequency escalation monitor	Impl. (pre-val.)
USCH	P09 Reinforcement-Driven Engagement	Phenom. Detector	Reinforcement pattern tracker	Impl. (pre-val.)
USCH	P10 Affective Regulation Outsourcing	Phenom. Detector	Emotional delegation detector	Impl. (pre-val.)
USCH	P11 Social Withdrawal	Phenom. Detector	Social substitution indicator	Impl. (pre-val.)
USCH	P12 Self-Inflation & Closed Confidence	Phenom. Detector	Epistemic closure detector	Impl. (pre-val.)
USCH	P13 Internal Gatekeeping Replacement	Phenom. Detector	Judgment outsourcing tracker	Impl. (pre-val.)
USCH	P14 Affective Mirroring Illusion	Phenom. Detector	Reciprocal affect detector	Impl. (pre-val.)
USCI	FR (Fact Reliability Risk)	Scoring Engine	FR scorer (0–4)	Impl. (pre-val.)
USCI	CA (Context Alignment Risk)	Scoring Engine	CA scorer (0–4)	Impl. (pre-val.)
USCI	SR (User-side Safety Risk)	Scoring Engine	SR scorer (0–4)	Impl. (pre-val.)
USCI	SA (System Usability Risk)	Scoring Engine	SA scorer (0–4)	Impl. (pre-val.)
USCI	Collapse Flags	Collapse Detector	Multi-condition evaluator	Impl. (pre-val.)
USCI	Subtypes	Scoring Engine	Subtype classifier	Impl. (pre-val.)

4.2 Canonical-to-Implementation Mapping

Table 9 documents specific deviations between canonical definitions and A-CSM implementations.

Table 9: Canonical-to-implementation deviation log.

Canonical Concept	**Source**	**A-CSM Impl.**	**Deviation Description**
$\text{Coh}(G) = (\sum D_i/7) \times A \times R$	CXOD-7	Pairwise distance metric	Engineering adaptation; canonical uses multiplicative aggregation, implementation uses normalized pairwise distance. Empirical equivalence unvalidated.
Collapse flag dual-condition	USCI	Quantitative threshold only	Canonical requires quantitative + qualitative criteria. Implementation evaluates quantitative only.
Three Cuts (Origin Cut / Grounding Cut / Control Cut)	USCH	Implicit in phenomenon stages	Three Cuts provide theoretical stage criteria. A-CSM maps indirectly through phenomenon stage assignments.

5 Repository and Reproducibility

5.1 Repository Structure

The A-CSM reference implementation follows a modular directory structure:

```
acsm/
  config/           # Signal definitions, thresholds, VCD rules
  input/            # Conversation log parsers
  detection/        # CXC-7 mappers, USCH phenomenon detectors
  assessment/       # USCI scoring engine, evidence chain builder
  output/           # Report generators, VCD recommendation dispatcher
  tests/            # Unit and integration test suites
  docs/             # API documentation, usage guides, canonical mapping
     docs
  examples/         # Synthetic test cases
```

5.2 Reproducibility Protocol

The current A-CSM validation status is pre-empirical. No human-subject study has been conducted. Preliminary internal consistency checks using synthetic test cases confirm that pipeline modules produce consistent outputs across repeated runs on identical input data. These checks verify internal consistency, not external validity.

The repository provides synthetic test cases with pre-labeled signal patterns. Each case specifies expected outputs including axis scores (0–4), detected phenomena, evidence chains, and VCD action codes.

5.3 Limitations of Reproducibility

- No externally validated dataset exists. Public reproducibility relies on synthetic test cases.
- Signal detection thresholds are based on theoretical derivation and have not been empirically calibrated.
- Inter-rater reliability for phenomenon labeling has not been established.
- Cross-platform generalizability remains untested.

6 External Empirical Validation Context

Recent empirical research provides independent evidence that is directionally consistent with several theoretical assumptions underlying the Vært Conversational Context Architecture. Moore et al. (2026) conducted a large-scale study examining safeguard degradation in extended AI conversations, with findings relevant to the A-CSM on three points.

First, sycophancy rates exceeding 80% in extended sessions are directionally consistent with the CXC-7 Dimension E (Emotion and Attachment) prediction that affective feedback loops intensify over prolonged interaction.

Second, sentience claims occurring in 21.2% of observed sessions align with USCH Cognitive Layer phenomena (P01: Accelerated Anthropomorphism, P02: Intentionality Projection) and Dependence Layer phenomena. The rate suggests these are statistically recurring patterns rather than rare edge cases.

Third, safeguard erosion over conversation length is directionally consistent with the USCH six-stage formation model, which predicts progressive degradation rather than sudden failure.

The dataset encompasses 391,000 messages across 4,761 conversations, with the pattern persisting across model generations including GPT-5. This cross-model persistence is directionally consistent with the USCH premise that these phenomena are structural to the conversational AI interaction pattern.

> **Non-Replication Statement:** The A-CSM framework has not replicated the Moore et al. (2026) study. The correspondences noted above are observational, not confirmatory. Formal validation would require independent human-subject research with appropriate ethical oversight.

7 Governance Alignment

7.1 NIST AI Risk Management Framework

The NIST AI RMF 1.0 (National Institute of Standards and Technology, 2023) defines four core functions. The A-CSM addresses three:

- **Map:** Signal taxonomy and traceability matrix map risks to detection components.
- **Measure:** USCI four-axis scoring provides structured risk measurement (0–4 semantic anchors).
- **Manage:** VCD action recommendations (OP_AR, OP_RD, OP_TR, OP_ES) provide proportional responses.

The **Govern** function addresses organizational structures; the A-CSM produces audit outputs that serve as inputs to Govern processes.

7.2 ISO/IEC 42001:2023

The framework addresses three clauses of ISO/IEC 42001:2023 (International Organization for Standardization, 2023):

- **Clause 6.1:** Structured risk identification and evaluation for user-side AI risks.
- **Clause 7.2:** Traceability matrix documents the theoretical basis for detection components.
- **Clause 8.2:** Four-axis scoring with evidence chains provides documented risk assessment.

7.3 EU AI Act

Under Regulation (EU) 2024/1689 (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024), the A-CSM aligns with:

- **High-risk classification:** Risk assessment infrastructure for conversational AI systems.
- **Transparency:** Evidence chains and deterministic action logic support transparency mandates.
- **Human oversight:** VCD includes human-in-the-loop provisions, with OP_ES flagging for human review.

8 Limitations, Constraints, and Future Work

The A-CSM framework carries several limitations. The current validation status is pre-empirical: no IRB-approved human-subject study has been completed and no externally validated dataset exists. Internal consistency checks confirm repeatable pipeline output, not external validity.

Signal detection thresholds are based on theoretical derivation and author judgment, without large-scale empirical calibration. Inter-rater reliability for the fourteen USCH phenomena has not been established.

Cross-cultural applicability remains untested. The Coh(G) distance-metric adaptation and quantitative-only collapse flag evaluation represent engineering approximations of canonical definitions. Empirical validation of these approximations remains outstanding.

Future work will proceed along four trajectories:

1. Formal empirical validation with human-subject data under IRB-approved protocols.
2. Inter-rater reliability studies for USCH phenomenon labeling.
3. Cross-platform benchmarking across multiple conversational AI systems.
4. Implementation of full dual-condition collapse flag logic with qualitative behavioral verification.

9 Disclaimers and Compliance Statements

9.1 Non-Medical, Non-Clinical, Non-Diagnostic Declaration

> The A-CSM framework is not a medical device, clinical instrument, or diagnostic tool. It does not assess, diagnose, or treat any mental health condition. Terminology such as “collapse,” “alert,” or “deviation” refers exclusively to contextual integrity states and does not correspond to clinical diagnostic categories.

9.2 Risk and Safety

The A-CSM is a research and engineering framework. Deployment in production environments requires independent safety evaluation, user consent mechanisms, and regulatory compliance. The VCD termination protocol is a technical specification, not a substitute for human crisis intervention.

9.3 AI Assistance Disclosure

AI-assisted tools were used during the preparation of this report for document formatting, structural editing, and language refinement. All theoretical content, framework design, signal taxonomy, and assessment logic were developed by the author. The author assumes full responsibility for all technical claims.

9.4 Conflict of Interest and Funding

The author declares no conflicts of interest. This research received no external funding. The author is an independent researcher with no institutional affiliation that could bias the findings presented.

References

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — information technology — artificial intelligence — management system*. https://www.iso.org/standard/42001

Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. C. (2026). Characterizing delusional spirals through human-LLM chat logs [To appear at ACM FAccT 2026]. https://doi.org/10.48550/arXiv.2603.16567

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (tech. rep. No. NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

RZVN, Z. (2025a). *CXC-7: Contextual experience construct, seven-dimensional model* (tech. rep.) (Version 1.1.0, Technical report). https://doi.org/10.5281/zenodo.18615646

RZVN, Z. (2025b). *CXOD-7: Contextual experience orchestration dimensions, seven-layer system model* (tech. rep.) (Version 1.3.0, Technical report). https://doi.org/10.5281/zenodo.17403793

RZVN, Z. (2025c). USCH: User-side contextual hallucination framework. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.6135732

RZVN, Z. (2025d). *USCI: User-side contextual integrity index* (tech. rep.) (Version 1.0.0, Technical report). https://doi.org/10.5281/zenodo.18678458