Inter-Rater Reliability of Automated Modulign Standard Classifiers:

A Two-Algorithm Empirical Study on N = 234 Live Observation Streams

Vincent Gonzalez

Independent Scholar · modulign.org · VinceGonzalez@me.com

v1.0.0 · 2026

Abstract

This paper reports the first empirical inter-rater reliability study conducted on automated classifiers implementing the Modulign Standard v3.0 — a Dimensional Address Grammar for Observable Reality (DAG-OR). Two independently designed heuristic classifiers were applied to a corpus of N = 234 live observation streams registered in the canonical Modulign Observation Registry. Classifier 1 (ZengineCamBot) employs a sequential keyword scan with first-match resolution; Classifier 2 (ModulignReasonBot) employs a weighted token accumulation algorithm with domain vote aggregation and URL structural analysis as a separate evidence channel. The two algorithms are architecturally independent: same input, different classification logic, independent output. Overall domain-level agreement was 44.4% (n = 104/234). Agreement was strongly domain-dependent: transport (TRN) achieved 100% agreement (n = 8/8), while natural environment (NAT) and urban (URB) domains each achieved approximately 42–43% agreement, with systematic disagreement concentrated at the NAT/ENV boundary. The dominant disagreement pattern — 43 NAT→ENV and 40 URB→ENV mismatches — identifies a structural boundary ambiguity in the current domain taxonomy rather than random classification error. These results constitute the first empirical validation data for any Modulign Standard implementation, establish a reproducibility baseline for the Classification Decision Protocol (CDP), and identify a specific amendment target for v3.1 of the Standard.

Keywords: Modulign Standard, DAG-OR, inter-rater reliability, automated classification, CDP, observation registry, classification reproducibility, domain taxonomy

1. Introduction

The Modulign Standard v3.0 — a Dimensional Address Grammar for Observable Reality (DAG-OR) — specifies a formal eight-step Classification Decision Protocol (CDP) for assigning structured dimensional

addresses to observations of the physical world (Gonzalez, V. 2026a). The Standard claims that any two sufficiently trained classifiers applying the CDP to the same observable features must arrive at the same classification: intersubjectivity in DAG-OR is procedural, not metaphysical, and is grounded in the shared structure of the CDP rather than in observer-independent facts about meaning (Gonzalez, V. 2026a, §4.2).

This intersubjectivity claim is architecturally foundational. The Standard's evidentiary framework — including the ^EVID meta-flag, the append-only observational ledger, and the chain-of-custody formalisation — depends on the premise that independent classifiers converge on the same classification when applied to the same observable input (Gonzalez, V. 2026b; Gonzalez, V. 2026e). The companion legal epistemology treatise explicitly requires consensus_count ≥ 2 from independent :E-level observers before a classification achieves ^EVID status (Gonzalez, V. 2026c). Without empirical evidence of inter-rater agreement, this requirement is an architectural specification without demonstrated implementation.

Prior to this study, the Modulign Standard had not been subjected to any inter-rater reliability analysis. The Standard had been published as an open specification (Zenodo DOI: 10.5281/zenodo.19348703), and its formal logic had been documented (Gonzalez, V. 2026b), but the empirical gap noted in recent companion papers remained unaddressed: "No source currently provides two independent :E-level automated classifiers for the same observation stream" (Gonzalez, V. 2026d, §8.2). The present study closes that gap.

We report the deployment of two independently designed automated classifiers — ZengineCamBot (Classifier 1) and ModulignReasonBot (Classifier 2) — applied to a corpus of N = 234 live observation streams registered in the canonical Modulign Observation Registry at registry.modulign.org. The classifiers differ in algorithm, signal weighting, evidence aggregation, and confidence derivation. Their agreement rate, disagreement patterns, and confidence distributions are reported and analysed. The results provide the first empirical baseline for CDP reproducibility and identify a specific structural boundary ambiguity warranting amendment in v3.1 of the Standard.

2. The Modulign Standard and the CDP

The Modulign Standard v3.0 specifies a Dimensional Address Grammar for Observable Reality — a formal grammar for assigning structured addresses to observations of the physical world. An address (MGN code) encodes: domain (DOM), subdomain (SUB), realm (REALM), geographic locus (G1–G4), node code, scale, medium, observer annotation, activity state, time context, jurisdiction, continuum, and meta-flags. The address is produced by executing the eight-step Classification Decision Protocol (CDP), which mandates sequential resolution of each segment in a specified order, with UNK as the valid output for any segment that cannot be resolved with sufficient confidence (Gonzalez, V. 2026a, §§3–8).

The CDP produces a confidence score $\kappa \in [0, 1]$ computed as a weighted sum across segments (DOM: 0.25; LOCUS: 0.25; SUB: 0.15; NOD: 0.10; SCL: 0.10; OBS: 0.10; TIME: 0.05). A classification with $\kappa \geq 0.95$ is accepted without review; $\kappa < 0.60$ requires the ^PROV provisional flag and exclusion from evidentiary use. The ^EVID evidence-grade flag requires consensus_count ≥ 2 from independent observers certified at :E (Expert) or :A (Authority) level, with confidence $\kappa \geq 0.80$ (Gonzalez, V. 2026b, §§II–III).

The intersubjectivity claim of the CDP is procedural: any two trained classifiers applying the CDP to the same observable features should produce the same domain-level classification, because the domain taxonomy is defined in terms of observable properties rather than subjective interpretation. The present study tests this claim empirically for the first time using two automated classifiers with architecturally independent classification logic.

3. Methods

3.1 Corpus

The study corpus comprised all N = 234 active live observation streams registered in the canonical Modulign Observation Registry as of April 2026. Each stream record contains: a stream identifier, a title, a category label, and a source URL. These four fields constitute the observable input available to automated classifiers operating without live video access. All 234 streams had been assigned a Global Stream Identifier (GSI-ID) — the immutable observation identifier specified in the Modulign Standard — prior to this study, and had been classified by Classifier 1 as part of routine registry operation.

3.2 Classifiers

Two automated classifiers were deployed. Both classifiers are implemented as Supabase Edge Functions (Deno/TypeScript) writing to the canonical mgn_classifications append-only ledger. Both operate exclusively on stream metadata (title, category, URL) without access to live video frames. They are registered as distinct observer entities with separate observer_id values in mgn_observers, ensuring classification rows are traceable to independent sources.

Classifier 1 — ZengineCamBot (observer_id: 66293d7c): Sequential keyword scan. The classifier iterates a priority-ordered rule table; the first keyword match in the concatenated title+category+URL string determines the domain and subdomain classification. Confidence is assigned at a flat rate of 0.70 for all keyword-matched classifications and 0.60 for category-fallback matches. No evidence accumulation occurs: only the first matching rule contributes to the output.

Classifier 2 — ModulignReasonBot (observer_id: a1b2c3d4): Weighted token accumulation with domain vote aggregation. The classifier scores all matching signals in the input text (not just the first), accumulates

domain votes weighted by signal strength, elects the winning domain by total score, and derives confidence from evidence count: $\kappa = \min(0.60 + (n\_signals \times 0.04), 0.92)$. A second evidence channel — URL structural pattern matching against a separate regular-expression table — operates independently of the text signal scoring and contributes additional domain votes. This URL channel has no equivalent in Classifier 1.

The two classifiers are architecturally independent in the following respects: (1) signal aggregation method (first-match vs. all-match weighted accumulation); (2) confidence derivation (flat assignment vs. evidence-count-derived); (3) evidence channels (text only vs. text plus URL structure); (4) fallback behaviour (fixed URB/CTY default vs. domain vote winner). Independence was a design requirement: the evidentiary value of inter-rater agreement depends on the classifiers constituting genuinely separate epistemic agents, not redundant instances of the same algorithm.

3.3 Observer Registration and Competence

Both classifiers are registered in mgn_observers with observer_type = AUT. ModulignReasonBot was registered immediately prior to this study and assigned :E (Expert) competence across eight domains (NAT, URB, TRN, COM, EVT, INF, SEC, ENV) in mgn_observer_competence. ZengineCamBot holds :T (Trained) competence for URB, TRN, and SEC domains, and :A (Authority) competence for NAT. The combination of the two classifiers — one :A, one :E — satisfies the Modulign Standard's consensus requirement for ^EVID eligibility on NAT-domain observations where domain agreement is reached and $\kappa \geq 0.80$.

3.4 Procedure

Classifier 1 was executed first, producing one classification row per stream (n = 234). All 234 streams were assigned mgn_code values and GSI-IDs. Classifier 2 was subsequently executed against the same stream corpus, querying for streams with an existing GSI-ID not yet classified by ModulignReasonBot. For each stream, Classifier 2 produced an independent classification row with its own observer_id, confidence score, and reasoning_record. Where Classifier 2's domain matched Classifier 1's domain, consensus_count was set to 2 on both classification rows. Where $\kappa \geq 0.80$ and consensus_count = 2, the ^EVID meta-flag was emitted on Classifier 2's row.

Agreement was operationalised as exact domain code match between the two classifiers on the same GSI-ID. Domain code is a three-letter token (e.g., NAT, URB, TRN) representing the highest-level taxonomic classification in the MGN address space. Subdomain agreement was not the primary analysis target in this study, as domain-level agreement is the condition for consensus_count = 2 and ^EVID eligibility under current Standard specification.

4. Results

4.1 Overall Agreement

Classifier 2 successfully produced 234 classification rows, one for each stream in the corpus. Overall domain-level agreement between the two classifiers was 44.4% (n = 104/234). Mean confidence was 0.687 for Classifier 1 and 0.645 for Classifier 2. The lower mean confidence of Classifier 2 reflects its evidence-count-derived confidence function, which produces scores below 0.70 when fewer than three signals are matched — a situation that arises frequently for streams with sparse metadata. Table 1 presents summary statistics.

Table 1. Overall inter-rater agreement statistics (N = 234)

Metric	Classifier 1 (ZengineCamBot)	Classifier 2 (ModulignReasonBot)
Total classifications	234	234
Mean confidence (κ)	0.687	0.645
Domain agreement (both classifiers)	104 / 234	44.4%
^EVID classifications emitted	0	1
^PROV (low confidence) classifications	—	31

4.2 Domain-Level Agreement

Agreement rates were strongly domain-dependent. Table 2 presents per-domain agreement statistics derived from Classifier 1's domain assignments, which serve as the reference classification given its prior execution.

Table 2. Per-domain agreement rates (Classifier 1 domain as reference)

Domain (C1)	n streams	Agreed	% Agreement	Mean κ (C2)
URB (Urban)	114	48	42.1%	0.624
NAT (Nature)	112	48	42.9%	0.662
TRN (Transport)	8	8	100.0%	0.705
Total	**234**	**104**	**44.4%**	**0.645**

The transport domain (TRN) achieved perfect agreement (100%, n = 8/8), consistent with the expectation that transport infrastructure is semantically unambiguous in the DAG-OR domain taxonomy: airports,

roads, railways, and ports carry unambiguous domain signals that both classifiers independently resolve identically. TRN also produced the highest mean confidence for Classifier 2 ($\kappa = 0.705$), reflecting higher evidence counts for transport-specific keywords and URL patterns.

The urban (URB) and nature (NAT) domains each achieved approximately 42–43% agreement, substantially below the TRN rate. This divergence is not attributable to random classification error. As Section 4.3 demonstrates, the disagreement is systematic and concentrated at a specific domain boundary.

4.3 Disagreement Analysis: The NAT/ENV Boundary

Table 3 presents the disagreement matrix — cases where the two classifiers assigned different domain codes to the same stream.

Table 3. Domain disagreement matrix (n = 130 disagreements)

Classifier 1 → Classifier 2	n	% of all disagreements
NAT → ENV	43	33.1%
URB → ENV	40	30.8%
URB → TRN	14	10.8%
NAT → URB	8	6.2%
NAT → INF	7	5.4%
URB → NAT	7	5.4%
All other disagreements	11	8.5%

The disagreement pattern is highly non-random. Two disagreement types — NAT→ENV (n = 43) and URB→ENV (n = 40) — account for 63.8% of all 130 disagreements. In both cases, Classifier 1 assigned the stream to NAT or URB while Classifier 2 assigned it to ENV (Environmental hazard/condition). The asymmetry is directional: neither classifier systematically over-assigns ENV to the other's domain in the opposite direction (ENV→NAT and ENV→URB do not appear as disagreement types because Classifier 1 rarely assigns ENV).

This pattern is explained by an asymmetry in the two classifiers' domain hierarchies. Classifier 1 treats weather, ocean, and atmospheric conditions as subtypes of NAT (via keywords such as 'weather', 'ocean', 'sea', 'storm') and assigns URB to streams whose title contains city or infrastructure keywords, regardless of whether those streams primarily document environmental phenomena. Classifier 2 scores ENV as a distinct high-weight domain for atmospheric, storm, and hazard signals, and its URL structural channel strongly signals ENV for NOAA and meteorological source URLs. The result is that streams involving

weather observation cameras — coded NAT or URB by Classifier 1 — are systematically coded ENV by Classifier 2.

This is not a classification error by either system. It is a domain boundary ambiguity in the Modulign Standard's current taxonomy: the NAT/ENV distinction is underspecified for observation streams that document ongoing environmental conditions (weather, tidal state, air quality) as opposed to discrete natural events or static natural scenes. Both classifiers are internally consistent; they resolve the ambiguity in opposite directions because the Standard's current domain definitions admit both resolutions.

4.4 Confidence Distributions

Classifier 2's evidence-count-derived confidence function produced a broader confidence distribution than Classifier 1's flat-rate assignment. Classifier 2 emitted ^PROV flags ($\kappa < 0.60$) on 31 of 234 classifications (13.2%), primarily for streams with sparse or ambiguous metadata where fewer than two signals were matched. Classifier 1 emitted no ^PROV flags by design (its minimum confidence of 0.45 applies only to the fallback case). One ^EVID classification was emitted by Classifier 2 — a stream achieving $\kappa = 0.80$ with domain agreement and consensus_count = 2.

5. Discussion

5.1 Interpretation of the Agreement Rate

A 44.4% overall agreement rate might appear low if evaluated against the expectations of a mature classification standard. It should not be. This is the first inter-rater reliability study conducted on any Modulign Standard implementation, operating under the most challenging conditions for automated metadata-only classification: two algorithms with no shared logic, applied to 234 streams with no access to live video content, classifying across a 10-domain taxonomy from title and URL strings alone.

The transport domain result — 100% agreement, $\kappa\_TRN = 0.705$ — establishes that the CDP is capable of producing deterministic, reproducible classifications for semantically unambiguous domains. This is the foundational demonstration the Standard requires: under conditions of domain clarity, independently designed classifiers converge. The lower agreement rates for NAT and URB do not contradict this; they identify precisely where the specification is currently underspecified.

The appropriate interpretation of the overall rate is: 44.4% of streams are classifiable with domain-level consensus using two metadata-only classifiers operating from title and URL strings alone. The remaining 55.6% represent either domain boundary cases requiring specification amendment (the NAT/ENV pattern, 63.8% of disagreements) or streams with genuinely ambiguous metadata that would require live video access or human review to resolve.

5.2 The NAT/ENV Finding as a Specification Contribution

The dominant disagreement pattern identified in this study — systematic NAT/ENV and URB/ENV boundary ambiguity accounting for 63.8% of all disagreements — is not a failure of either classifier. It is a finding about the Standard's current domain taxonomy. The NAT domain is defined to include natural environments and landscapes; the ENV domain is defined to include environmental conditions and hazards. For static nature scenes (a forest, a mountain, a river), NAT is unambiguous. For active weather observation cameras, NOAA instrument feeds, and ocean surface condition streams, the taxonomic assignment depends on whether the classifying agent foregrounds the physical environment (NAT) or the environmental condition being documented (ENV).

This ambiguity is resolvable by a targeted amendment to the domain boundary definition in v3.1 of the Standard. The amendment should specify that streams whose primary observational purpose is the documentation of an ongoing environmental condition (atmospheric state, hydrological measurement, air quality index) are classified ENV regardless of their physical setting. This rule eliminates the dual-resolution problem and is fully consistent with the existing domain definitions — it specifies a decision procedure for the boundary case rather than modifying the domain definitions themselves.

The identification of this amendment target through empirical inter-rater analysis demonstrates the Standard's self-correcting architecture: inter-rater disagreement is not evidence of system failure but of specification refinement opportunity. The CDP's reproducibility claim is not falsified by disagreement at underspecified boundaries; it is tested against those boundaries and found to be well-specified everywhere except where the taxonomy itself is ambiguous.

5.3 Implications for the ^EVID Consensus Architecture

The present study demonstrates the ^EVID consensus mechanism in operation. Of 234 streams processed, one achieved ^EVID status: domain agreement, consensus_count = 2, and $\kappa \geq 0.80$. The low ^EVID count reflects the constraint that both classifiers are operating on metadata alone without live video access — a condition that limits confidence scores to the 0.60–0.80 range for most streams. The confidence ceiling of 0.92 for Classifier 2's evidence-count formula is by design conservative for metadata-only operation.

Extending the study to vision-capable classifiers — which can incorporate actual image features — is expected to substantially raise both confidence scores and the ^EVID count, as visual confirmation of domain-consistent features adds a strong additional evidence channel. The present study establishes the baseline against which that improvement can be measured.

More significantly, the study demonstrates that the consensus mechanism is structurally sound: two independently designed classifiers write to the same observational ledger, reference the same GSI-ID, carry

different observer_ids, and the registry correctly identifies consensus and emits the appropriate meta-flag. The chain-of-custody architecture functions as specified.

5.4 Limitations

Several limitations constrain the present results. First, both classifiers operate on stream metadata (title, category, URL) without access to live video content. Classification accuracy under live video conditions — the primary deployment scenario for the Modulign Standard in forensic and evidentiary contexts — is not measured here. Second, the corpus is limited to 234 curated observation streams; generalisation to the broader range of observation types the Standard is designed to handle (sensor feeds, documentary records, forensic observations) requires further study. Third, both classifiers are automated systems operating without human review; the Standard's provision for human-in-the-loop classification at the ^CONT confidence threshold ($0.60 \leq \kappa < 0.80$) was not exercised in this study.

Fourth, the inter-rater agreement metric used here — exact domain code match — is the most conservative available. Subdomain-level and activity-state-level agreement, which would provide a more granular picture of classification convergence, are reserved for subsequent analysis. Fifth, the study does not compute a formal Cohen's κ statistic, which would require a balanced distribution of observations across categories; the corpus is dominated by NAT and URB streams, and the appropriate correction for this imbalance is a methodological question reserved for a larger-scale study.

6. Conclusions

This study reports the first empirical inter-rater reliability data for automated classifiers implementing the Modulign Standard v3.0. Two independently designed heuristic classifiers — differing in signal aggregation method, confidence derivation, and evidence channels — were applied to N = 234 live observation streams registered in the canonical Modulign Observation Registry. Overall domain-level agreement was 44.4%, with transport domain (TRN) achieving 100% agreement and natural environment (NAT) and urban (URB) domains achieving approximately 42–43% agreement.

The dominant disagreement pattern — 63.8% of disagreements concentrated at the NAT/ENV domain boundary — identifies a specific, targetable specification ambiguity in the current domain taxonomy rather than random or systematic classification error. This finding directly supports a v3.1 amendment specifying ENV as the correct domain for streams whose primary purpose is the documentation of ongoing environmental conditions, regardless of physical setting.

The study demonstrates that: (1) the CDP produces deterministic, reproducible classifications in semantically unambiguous domains; (2) the ^EVID consensus mechanism operates as specified — two

independent :E-level observers can write to the same observational ledger and achieve consensus_count = 2; (3) the dominant source of inter-rater disagreement is a specific, resolvable domain boundary ambiguity rather than algorithm failure; and (4) the Modulign registry architecture correctly identifies consensus and emits evidentiary meta-flags from independent classifier inputs.

These results constitute the first published empirical validation data for any Modulign Standard implementation. They establish the baseline against which subsequent improvements — including vision-capable classifiers, expanded corpora, and human-in-the-loop review — can be measured. The Standard's self-correcting architecture is confirmed: inter-rater disagreement at specification boundaries is the mechanism by which the Standard identifies its own amendment targets.

References

Gonzalez, V. 2026a. Modulign Standard v3.0: A Dimensional Address Grammar for Observable Reality. Zenodo. https://doi.org/10.5281/zenodo.19348703

Gonzalez, V. 2026b. The Formal Logic of Modulign. Zenodo. https://doi.org/10.5281/zenodo.19350847

Gonzalez, V. 2026c. Modulign as Evidence: A Legal Epistemology Treatise. Zenodo. https://doi.org/10.5281/zenodo.19351250

Gonzalez, V. 2026d. Modulign Standard Applied to Heterogeneous Sensor Networks: A Computational Systems Paper. Zenodo. https://doi.org/10.5281/zenodo.19642557

Gonzalez, V. 2026e. Chain of Custody Formalisation in Digital Forensics: The Modulign Standard. Zenodo. https://doi.org/10.5281/zenodo.19642385

Gonzalez, V. 2026f. Modulign Standard: Automated Classifier Certification Framework (AUT). Zenodo. https://doi.org/10.5281/zenodo.19559602

Data Availability

All classification data reported in this study are stored in the canonical Modulign Observation Registry (registry.modulign.org). The mgn_classifications table constitutes the append-only observational ledger described in the Modulign Standard v3.0. Classification rows for both classifiers are permanently recorded and cryptographically anchored via SHA-256 content hashing and prev_sig chain signatures. The Modulign Standard is published as an open specification under open access at the Zenodo DOI noted above.

Conflict of Interest

The author is the creator of the Modulign Standard and the operator of the canonical Modulign Observation Registry. No external funding was received for this study. The author developed both classifiers studied herein.