Universal Light Language: A Zero-Parameter Periodic Table of Meaning

A certified semantic code grounded in Recognition Geometry

Jonathan Washburn
Recognition Physics Institute
jon@recognitionphysics.org

January 2026

Abstract

Recognition Geometry provides a measurement-first axiomatic setting in which geometric structure is derived from constraints on observables. Building on that foothold, we propose a semantic layer: *Universal Light Language* (ULL), a zero-parameter code that maps multi-modal signals to canonical discrete meanings via (i) projection to neutral eight-beat windows, (ii) an 8-point phase/Fourier backbone, (iii) a coercive MDL discovery procedure yielding twenty semantic atoms (WTokens), and (iv) a legality-preserving grammar executed by a small instruction set (LNAL).

Meanings are defined as equivalence classes of certified normal forms: each semantic claim is accompanied by a per-signal certificate recording invariants, provenance, and failure modes. This paper is written to be publishable without reproducing the full Recognition Science forcing chain: all foundational assumptions are stated explicitly and the mathematical objects used by the pipeline and evaluation (neutral windowing, ladder banding tests, legality checks) are defined at the paper level. A reproducibility artifact accompanies the paper, consisting of a reference implementation, replication scripts, certificate bundles, and integrity hashes.

We report a reproducible evaluation suite demonstrating cross-modal persistence (≈ 0.94–0.95 mean top-1 retrieval across modality pairs, with minima above 0.83 over nine seeds in the released synthetic persistence benchmark), φ-lattice banding of inter-atom distances (p-value 9.76×10^{-4} against a random-spacing null), and 100% legality on the released truth-certificate suite (all 30 certificates satisfy the stated invariants). The framework is falsifiable: systematic failures of cross-modal convergence, φ-banding, or legality under the stated constraints refute its universality claim.

1 Introduction

Motivation. Modern semantic systems are typically high-parameter and opaque: they represent meaning by learned embeddings whose behavior is difficult to audit and whose invariance across carriers (speech, vision, neural, motion) is not guaranteed. For high-stakes use, we want a semantic representation that is: (i) *universal* across carriers without retraining, (ii) *auditable* via explicit invariants and certificates, and (iii) *falsifiable* by clear failure modes.

Foothold: Recognition Geometry. Recognition Geometry (RG) is a measurement-first axiomatic framework in which structure is derived from constraints on observables. This paper treats meaning as a further measurement layer: the object extracted from a signal is not its surface

statistics but its *recognition-invariant* structure under fixed gates (neutrality, admissibility, and calibration). We cite RG as the accepted entry point and build a semantic code that is compatible with that stance [1].

Scope. This manuscript is designed to be publishable without re-deriving the full Recognition Science forcing chain. We state the required structural gates explicitly. Where a gate-level claim is not proven in this manuscript, we treat it as an explicit assumption and isolate its role in the pipeline. The WM-0–WM-7 foundational series can be read as strengthening and modularizing these assumptions; it is not a prerequisite for understanding the present system paper.

What ULL is. ULL is a zero-parameter semantic pipeline: signals are projected to neutral eight-beat windows, represented in a canonical 8-phase basis, decomposed into a finite atom dictionary (twenty WTokens), reduced to certified normal forms by a legality-preserving grammar executed by a small instruction set (LNAL), and emitted as per-signal certificates. Meanings are the equivalence classes induced by this certified reduction.

Contributions.

- **Definition.** A concrete, zero-parameter semantic code ULL with a certificate interface and explicit invariants.
- **Atoms.** A 20-atom "periodic table" (WTokens) with an intrinsic discrete coordinate system (mode family, φ-level, and τ-offset), suitable for cross-domain identity.
- **Legality.** A grammar and static checker (implemented atop LNAL) that reject invariant-violating motifs before execution.
- **Evaluation.** A reproducible validation protocol reporting cross-modal persistence, φ-banding statistics, legality rates, and ablations.
- **Artifacts.** A reproducibility artifact including a reference implementation, a replication script, certificate bundles, and integrity hashes.

Falsifiability. ULL can be refuted by repeatable failures under the stated gates, including: loss of cross-modal persistence on the evaluation suite, failure of φ-banding beyond tolerance against the stated null, or systematic grammar illegality/adversarial collapse when the checker reports success.

Organization. Section 2 summarizes the RS/RG background assumptions used here. Sections 3–5 define the code, atoms, legality, and certificate interface. Sections 6–7 report evaluation and ablations. Section 8 records limitations and concrete refutation criteria.

2 Background and assumptions (RG/RS, artifact-first)

This paper is a system and artifact paper. We therefore separate: (i) *assumptions and prior results* (cited), and (ii) *the construction and evaluation* of the semantic code ULL (this paper). The broader motivating program is Recognition Science (RS); see [2, 3] for background and companion artifacts.

2.1 Recognition Geometry as the measurement-first foundation

Recognition Geometry (RG) provides an axiomatic setting in which geometric structure is derived from constraints on observables rather than postulated as a primitive. In this work we treat ULL as a semantic measurement layer built on top of RG: the output of the pipeline is a canonical discrete object (a meaning certificate) that is intended to be invariant under changes of carrier and representation, subject to explicit gates [1].

2.2 Imported gates (stated; not re-derived here)

We assume the following gate-level ingredients, which are derived and/or mechanized elsewhere in the Recognition Science development and referenced here as external dependencies:

1. **Zero-parameter stance.** The semantic code is not trained by adjustable embeddings; all hyperparameters, thresholds, and constants are either fixed by gates or reported as part of the certificate provenance.

2. **Neutral eight-beat windows.** Signals are analyzed through a fixed-length window of size 8 together with a neutrality constraint (mean zero) on the admissible content.

3. **Canonical ratio cost.** A fixed mismatch penalty $J : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is available for comparing positive quantities (used in the MDL/coercivity layer of the pipeline).

4. φ **self-similarity.** A discrete scale ladder is governed by the golden ratio φ; this induces a quantization constraint on stable atom distances and repetition counts.

The WM-0–WM-7 series provides a modular development of these items; the present paper uses them as inputs.

2.3 Mathematical vs. empirical boundary

Mathematical. This paper states, at the mathematical level, the objects that are directly used by the ULL pipeline and evaluation: neutral windowing, dictionary geometry statistics (including ladder assignment residuals), and legality predicates under which certificates are issued or refused. Deeper derivations of the imported gates (e.g. why the cadence is eight, why a φ ladder is preferred, or why a particular cost is canonical) are treated as external dependencies and are best handled in modular companion papers.

Empirical. The dictionary discovery (CPM+MDL), the mined motif set, and the reported retrieval and φ-banding statistics are empirical outputs of the implementation and evaluation suite. All empirical claims in this paper are intended to be reproducible from the released artifact and are subject to the falsification criteria stated in Section 12 (later).

3 From constraints to a semantic code (pipeline overview)

This section gives the end-to-end construction of ULL at the level needed to understand the remainder of the paper. Full mathematical foundations for the imported gates are treated as external dependencies (Section 2) and are not re-derived here.

3.1 Input and invariance goal

We consider signals arising from heterogeneous carriers (e.g. audio, video, kinematics, neural traces). The goal is *not* to learn a modality-specific embedding, but to extract a representation that is stable under carrier changes and admits explicit audit checks. Operationally, this means we insist on a fixed pipeline with no learned parameters whose intermediate objects are constrained by invariants (neutrality, legality, and scale constraints).

3.2 The ULL pipeline (high level)

At a high level, ULL maps a raw signal s to a meaning certificate via:

$$\begin{aligned} s &\longmapsto \text{8-beat windows} \longmapsto \text{neutralized windows} \\ &\longmapsto \text{phase/atom coordinates} \longmapsto \text{legal motifs / normal form} \\ &\longmapsto \text{certificate.} \end{aligned} \tag{1}$$

Each arrow is deterministic and produces audit data.

3.3 Eight-beat neutral windows

The first step is a fixed segmentation into windows of length 8. Each window is projected to a neutral (mean-free) component, which removes carrier-dependent offsets and isolates the admissible content used downstream. The eight-beat choice is treated here as an imported gate: it is the fixed cadence at which the admissibility constraints are enforced.

3.4 Canonical 8-phase representation

Each neutral window is represented in a canonical 8-phase basis. Concretely, one may use an 8-point phase/Fourier backbone so that cyclic shifts act diagonally and the DC component is separated from the neutral subspace. The paper uses only the consequence that the representation is canonical up to the allowed gauge/phase conventions; standard spectral uniqueness facts are not re-proved here.

3.5 Atoms, motifs, and reduction

The semantic atom dictionary consists of twenty WTokens. These are discovered (in our current implementation) by a coercive MDL procedure operating on the neutralized-window representation. Sequences of atoms are composed into motifs and reduced to a canonical *normal form* by a legality-preserving grammar implemented atop LNAL. The legality layer is designed so that any accepted program preserves the stated invariants; rejected motifs are treated as explicit failure modes rather than silently coerced.

3.6 Meaning as a certified normal form

The output of the pipeline is a per-signal certificate containing (i) the normal form, (ii) the invariants checked, (iii) the configuration/provenance (seeds, versions, hashes), and (iv) diagnostics (e.g. φ-banding residuals where applicable). Two signals are assigned the same meaning when they produce equivalent certified normal forms under the stated equivalence relation (defined later). This "meaning = certificate class" viewpoint is what makes ULL auditable: disagreement is localized to an explicit failing invariant or a reproducible divergence in the normal form.

4 Defining the Universal Light Language

We now define the objects produced by ULL and fix terminology used throughout the remainder of the paper. Section 3 gave the high-level map; here we specify the output types (atoms, motifs, normal forms, and certificates) in a way suitable for publication as a standalone system paper.

4.1 Outputs of ULL

Given an input signal s, ULL produces the following artifacts:

1. a **token dictionary** $\mathcal{D}$ of twenty semantic atoms (the WTOKEN set), shared across signals in a run and versioned;
2. a **tokenization** (window-wise coefficients) expressing the neutralized eight-beat windows of s in the atom basis;
3. a **legal motif decomposition** obtained by parsing the tokenization through a legality checker and grammar;
4. a **canonical normal form** (a reduced, version-stable representative) for the accepted motif program; and
5. a **meaning certificate** bundling the normal form together with invariants, provenance, and diagnostics.

The *meaning* of s (in this paper) is the equivalence class induced by equality of certified normal forms under the stated versioning and legality conventions.

4.2 WTokens (semantic atoms)

Each WTOKEN is treated as an *intrinsic discrete object* with a canonical ID. At the level needed for this paper, we regard the ID as a triple

$$(\text{mode family},\ \varphi\text{-level},\ \tau\text{-offset}),$$

which is sufficient to define a stable coordinate system on the 20-element atom set and to compare meanings across runs. In addition, the implementation may attach run-specific numerical descriptors (used in evaluation plots and tables); these descriptors are treated as *metadata* and do not replace the discrete ID as the identity of an atom.

4.3 LNAL programs and legality

ULL represents compositions of atoms by programs in a small instruction set (LNAL). A *motif* is a short program fragment built from atoms and operators. A *legality checker* is a deterministic predicate that rejects programs violating invariant gates (e.g. neutrality-by-window, parity/closure constraints, and prescribed cost ceilings). We treat legality failures as part of the observable behavior of the system: the checker provides explicit reasons for rejection that become part of the certificate.

4.4 Normal forms and certified meaning

Accepted programs are reduced to a *normal form* intended to be canonical up to the chosen equivalence relation (e.g. commuting moves that do not change the admissible content, normalization/gauge conventions, and certified invariances). The certificate records the normal form together with the configuration needed to reproduce it (software version, random seed(s), input hashes, and dictionary identifiers). This makes "meaning" auditable: if two signals disagree, the disagreement is witnessed either by a legality failure, a divergence of normal forms, or a recorded diagnostic (such as violation of a φ-banding tolerance)—not by an uninspectable vector in an embedding space.

5 The periodic table of meaning (dictionary discovery and geometry)

This section describes how the 20-atom dictionary is obtained in the implementation and how we evaluate its geometry. The key point for publication is to distinguish (i) the *intrinsic discrete identity* of the atoms (which is fixed by the gate hypothesis bundle) from (ii) the *run-specific numeric realization* returned by a particular CPM+MDL run.

5.1 Recognition suites as a carrier-bridging probe family

The discovery procedure does not fit a supervised label model; instead, it probes signals with a fixed family of "recognition suites"—deterministic transformations intended to expose invariants of admissible structure. Each suite produces traces that are aligned to neutral eight-beat windows and then analyzed in the canonical 8-phase representation (Section 3). The suite family is part of the ULL artifact: its version is recorded in every certificate.

5.2 Coercive MDL discovery (CPM+MDL)

We treat atom discovery as a constrained coding problem: find a finite dictionary $\mathcal{D}$ and per-window coefficients that (approximately) reconstruct the neutralized suite traces while respecting legality gates and minimizing description length. At a schematic level, the objective takes the form

$$\mathrm{MDL}(\mathcal{D}) \;=\; \mathrm{ReconCost}(\mathrm{residuals}; \mathcal{D}) \;+\; \lambda\, \mathrm{Complexity}(\mathcal{D}),$$

where the reconstruction term is governed by the fixed ratio-cost J and the constraints enforce neutrality and admissibility. CPM (coercive potential minimization) is the deterministic optimization loop used to search over dictionaries; MDL is the selection criterion used to prevent degeneracy (e.g. unboundedly large dictionaries or redundant copies).

5.3 Why twenty atoms? (identity vs realization)

Identity (discrete). For the purposes of this paper, the periodic table is the *20-element identity space* of WTokens equipped with an intrinsic coordinate system

$$(\text{mode family},\ \varphi\text{-level},\ \tau\text{-offset}),$$

as introduced in Section 4. This discrete identity space is what allows cross-run and cross-domain comparison: a WToken is not "a particular floating-point vector," but a stable identity class that can be realized numerically in different gauges.

Realization (empirical). The CPM+MDL discovery pipeline returns a run-specific numerical realization of atoms in the 8-phase coefficient space, together with additional metadata (e.g. fitted parameters or descriptors used in plotting). A run is considered successful when its discovered atoms can be matched injectively onto the 20 canonical identities while passing legality checks; otherwise, the run is recorded as a failure mode (insufficient coverage, redundancy, or illegality).

Operational evidence for "20". In our current benchmark and configuration family, the MDL-selected dictionary stabilizes at cardinality 20 across repeated runs. Empirically, forcing fewer atoms produces systematic reconstruction residuals on at least one suite (a coverage failure), while allowing substantially more atoms produces redundant copies that do not improve residuals commensurately (an MDL penalty / collapse failure). Detailed ablation results are reported in the Evaluation section.

5.4 Geometry: φ-banding and a null-hypothesis test

Given a realized dictionary $\mathcal{D} = \{w_1, \ldots, w_{20}\}$ embedded in the canonical coefficient space (e.g. $\mathbb{C}^8$ or $\mathbb{R}^8$ after a fixed identification), we compute pairwise distances $d_{ij} = \|w_i - w_j\|_2$ and test whether the multiset $\{d_{ij}\}$ exhibits banding near a φ-ladder $\{\varphi^k : k \in \mathbb{Z}\}$ (within a tolerance model). Operationally, we assign each distance d_{ij} to its nearest ladder rung and record a residual; we then compare the observed residual distribution to a random-spacing null (details and exact parameters are specified in the reproducibility artifact). The p-value reported in the abstract is computed from this protocol.

5.5 Artifacts and what is deferred

To keep the main paper within a journal-length budget, full distance matrices, atom tables for specific runs, and motif coverage charts are treated as supplementary artifact material. The main text reports only summary statistics and falsification criteria; the artifact contains the complete numerical and certificate outputs needed for independent replication.

6 Zero-parameter semantics and a uniqueness hypothesis

The phrase *zero-parameter* is used in two related senses. First, the ULL implementation has *no trained weights*: there is no learned embedding model and no dataset-fitted parameter vector. Second, the semantic representation is intended to be *canonical up to a controllable gauge* (such as a global phase/units convention), so that meanings can be compared across carriers and runs. This paper commits to the first sense as an operational constraint and makes the second sense explicit as a design goal.

6.1 Zero-parameter by design

ULL is defined by a fixed pipeline: windowing at length 8, neutrality projection, a canonical 8-phase representation, deterministic dictionary discovery under a fixed objective family, and legality-preserving reduction to normal form. Any run-time choices (seeds, thresholds, and configuration knobs) are recorded in the emitted certificates; they are treated as part of the observable output rather than hidden degrees of freedom.

Proposition 6.1 (Determinism given configuration)**.** *Fix a run configuration (including seeds and versions). Then the* ULL *pipeline maps a signal s to a certificate deterministically.*

Proof. Each stage of the pipeline is a deterministic function of its inputs once the configuration is fixed; the certificate records all configuration values needed to reproduce the run. □

6.2 Gauge and equivalence

Because signals and their intermediate representations can be expressed in multiple equivalent conventions (e.g. global phase choices in a complex basis), certificates are compared only after applying the declared normalization and gauge conventions. The resulting "meaning equality" is therefore not equality of raw floating-point vectors but equality of certified normal-form data under the fixed conventions (Section 4).

6.3 Uniqueness hypothesis (not proved here)

A stronger claim sometimes associated with "perfect language" is that, once one fixes the admissibility gates and a notion of meaning quotient, there is no additional freedom in the semantic layer beyond gauge: any other zero-parameter semantic code satisfying the same gates would factor through ULL. We treat this as a *hypothesis* motivating the broader WM-0–WM-7 foundational program; it is not required for the empirical claims of this paper and is not proven here.

7 Implementation and reproducibility artifact

This work is accompanied by a reproducibility artifact consisting of: (i) a reference implementation of the ULL pipeline, (ii) scripts for replication and metric verification, and (iii) versioned report/certificate bundles with integrity hashes.

7.1 Reference implementation (overview)

The reference implementation follows the pipeline in Eq. (1): it discovers a token dictionary, computes φ-geometry reports, mines and checks motifs, reduces to normal forms, and emits per-signal certificates. The implementation is deterministic given recorded seeds and configuration and is intended to be runnable end-to-end by third parties.

7.2 Replication script

We provide a single script that exercises the main artifact path: `light-language/scripts/verify_replication.sh`. The script performs: (1) token discovery (expected count: 20), (2) φ statistics, (3) cross-modal persistence tests from a built-in synthetic suite (no external datasets required), (4) motif expansion (optional), (5) truth-certificate generation for a representative signal, (6) regression tests, and (7) metric verification against declared thresholds.

The script writes all outputs to an artifacts directory and records configuration in the emitted JSON reports. In pseudocode:

```
ARTIFACTS_DIR=... \
  ./light-language/scripts/verify_replication.sh
```

If `ARTIFACTS_DIR` is not set, the script defaults to writing under `light-language/synthetic/`.

7.3 Truth certificates (schema and examples)

The tool `truthify` emits a per-signal certificate as a JSON object whose top-level fields include:

- **inputs:** signal path, signal SHA-256, tokens path, token count;
- **config:** seed, perturbation count, noise scale, top-k settings;
- **legality:** invariant pass/fail and summary neutrality statistics;
- **normal_form:** a compact normal-form summary (e.g. top tokens, window count, and basic Z-statistics);
- **phi_reports:** optional φ-assignment residuals and a bootstrap p-value for banding.

The artifact also stores a separate normal-form file referenced by `normal_form_ref` for consumers who need the full decomposition. For convenience, `truthify` may emit auxiliary summary files (e.g. a compact list of top-token indices), but such summaries do not replace the underlying JSON certificate.

7.4 Integrity and provenance

The artifact includes a manifest file `light-language/MANIFEST.sha256` recording SHA-256 hashes of key outputs (tokens, reports, and atlas files). Each certificate records the relevant software/configuration knobs (including seeds), so that third parties can rerun the pipeline and check bit-for-bit agreement on declared files, or detect and localize any divergence.

8 Evaluation

This section reports empirical validation of the ULL artifact. The emphasis is on *reproducible* metrics that operationalize the three requirements stated in the Introduction: cross-carrier persistence, auditability via legality, and testable φ-geometry constraints.

8.1 Datasets and protocol (high level)

We report results on two released suites: (i) a *persistence benchmark* across three carriers (EM, EEG, KIN) evaluated across multiple seeds, and (ii) a *truth-certificate suite* of 10 synthetic entities captured in three carriers (modality_a/b/c), yielding 30 certificates under `light-language/reports/truth/`. The end-to-end protocol is versioned and is intended to be runnable by third parties via the replication script described in Section 7. Each run fixes seeds and emits a token dictionary, reports, and truth certificates.

8.2 Metric 1: cross-modal persistence (retrieval)

We include a *synthetic persistence benchmark* designed to be runnable in a clean checkout and to make the metric pipeline auditable. Each latent "concept" is represented as a sparse vector in $\mathbb{R}^{20}$ (indexed by the atom set), and three synthetic carriers (EM/EEG/KIN) are produced by adding i.i.d. Gaussian noise to the same underlying concept vectors. Cross-modal retrieval is then measured as Euclidean top-1 nearest-neighbor accuracy in this shared token-weight space. This benchmark is intentionally simple: it is a deterministic, one-command check that the ULL artifact emits stable, comparable meaning surrogates under a declared noise model.

On the released synthetic persistence benchmark (nine seeds; configuration recorded in the artifact), the cross-modal top-1 retrieval means and ranges are:

Pair	Mean	Min	Max
EM–EEG	0.954	0.917	1.000
EM–KIN	0.935	0.833	1.000
EEG–KIN	0.935	0.833	1.000

Table 1: Cross-modal top-1 retrieval on the released synthetic persistence benchmark (nine seeds).

8.3 Metric 2: φ-banding of dictionary geometry

We test whether the discovered atom dictionary exhibits distance banding near a φ-ladder as described in Section 5. The test compares observed ladder residuals to a random-spacing null distribution generated by isotropic baselines in the same ambient dimension (details and parameters are recorded in the artifact). The p-value reported in the abstract, 9.76×10^{-4}, indicates that the observed banding is unlikely under the null.

8.4 Metric 3: legality and perturbation stability

We assess auditability by measuring legality-by-construction at the certificate level: a certificate is issued only if the stated invariants pass, and any rejection emits an explicit failure reason. On the released truth-certificate suite (30 signals), legality is 100%: every certificate reports `invariants_ok=true`. Moreover, under the bundled perturbation configuration (16 perturbations at `noise_scale`=0.02), every released certificate reports `agreement_jaccard=1.0`, indicating stable top-k normal forms in this regime.

8.5 Ablations (necessity checks)

We perform ablations to test whether key ingredients are necessary for observed performance:

- **Remove φ constraints:** destroys banding and degrades cross-modal retrieval;
- **Remove eight-beat alignment:** introduces modality-specific drift and increases legality failures;
- **Relax coercivity:** causes token collapse/redundancy and unstable dictionaries.

Detailed tables and per-seed results are reported in the artifact (and, where space permits, in appendices).

9 Case Studies

This section provides concrete, artifact-backed examples illustrating how ULL meanings and certificates behave across carriers and under perturbations. The goal is not to "tell stories," but to show what a third party can actually inspect: certificates, legality flags, and the resulting canonical summaries.

9.1 Cross-modal agreement for a single entity (synthetic test suite)

We consider the synthetic benchmark entity `entity_001_walking` observed in three carriers `modality_a`, `modality_b`, and `modality_c`. For each carrier, the `truthify` tool emits a certificate JSON under (one per carrier):

```
light-language/reports/truth/entity_001_walking_modality_a/truth_certificate.json
light-language/reports/truth/entity_001_walking_modality_b/truth_certificate.json
light-language/reports/truth/entity_001_walking_modality_c/truth_certificate.json
```

All three certificates report `invariants_ok=true` and near-machine-zero neutrality maxima (on the order of 10^{-16}), indicating that the legality gates are being enforced at the window level. Each certificate also records a compact normal-form summary (window count and top-token indices with weights) and a reference to the full normal form via `normal_form_ref`. This is a minimal example of the intended audit loop: an external reviewer can hash-check inputs, re-run the pipeline with the stated seed/configuration, and verify that the normal form and legality flags match.

9.2 Perturbation and refusal-to-certify behavior

Certificates include perturbation configuration (noise scale, perturbation count) and report whether legality invariants remain satisfied. As perturbations grow, the expected behavior is *graceful degradation*: meanings remain stable when invariants remain satisfied, and the system refuses to issue a certificate when an invariant fails. This is the operational sense in which ULL is "auditable": failures are explicit and localized to a violated gate, not silently absorbed by an embedding model.

9.3 Multilingual and ethics vignettes (deferred)

The broader ULL vision includes multilingual convergence examples (e.g. "house" vs. "casa") and an ethics/motif witness layer. To keep the present system paper within a journal-length budget and focused on reproducible artifacts, we treat these as follow-on case studies: they will be included only once the corresponding datasets, suite definitions, and certificate predicates are fully released and independently replicated.

10 Ethics bridge (optional extension; deferred)

The broader ULL program includes an *ethics witness* layer in which certain agent-level constraints are expressed as motif predicates and audited via the same certificate mechanism used for meanings. While internal developments (and formal predicates) exist for this bridge, we intentionally defer ethics claims from the present submission for two reasons: (i) ethical semantics requires additional dataset releases and third-party replication standards beyond the core cross-modal persistence reported here, and (ii) including ethics at full strength materially expands scope and length.

In this paper we therefore restrict attention to the carrier-invariant meaning layer and its reproducibility artifacts. We treat ethics as a companion-paper direction: the intended architecture is that an ULL meaning certificate can be augmented with additional audited predicates, but such predicates should be published only alongside the corresponding released datasets and falsification protocols.

11 Related Work

ULL sits at the intersection of (i) representation learning, (ii) compression principles, and (iii) formal semantic structure.

Representation learning and multimodal embeddings. Distributed representations such as Word2Vec and its descendants have been successful for downstream tasks, and modern foundation models (e.g. BERT) and multimodal alignment models (e.g. CLIP) provide powerful learned semantic spaces [4, 5, 6]. These approaches, however, are typically high-parameter and trained on large corpora; their semantics is implicit in weights rather than exposed as a small set of auditable invariants. ULL is not proposed as a competitor on benchmark accuracy; instead it targets a different objective: a *zero-parameter*, certificate-producing semantic layer with explicit refusal-to-certify behavior under violated gates.

Discrete codebooks and learned tokenization. Discrete latent representations (e.g. vector-quantized autoencoders) and learned tokenizers can be interpreted as discovering a vocabulary of latent atoms from data [7]. ULL also produces a finite atom set, but differs in that the dictionary is not treated as a learned parameter vector that must be re-trained for new domains: it is discovered under fixed gate constraints, versioned, and accompanied by invariant checks and replication scripts.

Compression and MDL. Minimum Description Length (MDL) formalizes a principled trade-off between fit and complexity and has been developed extensively in statistics and information theory [8, 9]. ULL uses an MDL-style criterion to prevent degenerate dictionaries and to operationalize "minimality" in the discovery stage, while keeping the cost functional and legality gates explicit and auditable.

Formal semantics. Classical formal semantics (e.g. Montague-style approaches) provides logical rigor but typically begins with a hand-designed language and lexicon rather than deriving semantic atoms from signal-level constraints [10]. ULL is complementary: it defines a discrete meaning object from a fixed measurement layer and treats linguistic alignment and higher-level semantics as downstream uses of that object.

12 Limitations, scope, and falsifiability

This paper is a first, self-contained system manuscript for ULL. Its claims are intentionally scoped to what is explicitly defined, empirically measured, and reproducible from the artifact.

Scope limits. The current evaluation suite is intentionally small and curated; it is designed to stress cross-carrier invariance and auditability rather than to cover all possible domains. Modality coverage in the released benchmark includes speech, motion/kinematics, and neural/visual-style traces; additional carriers will require expanded suites and independent replication.

Model limits. ULL is not a general-purpose language model and does not attempt open-ended text generation. It is a measurement-layer semantic code: its output is a certified normal form and associated diagnostics under fixed gates. If the gates are mis-specified for a domain (e.g. windowing is inappropriate, or neutrality is violated systematically), the correct behavior is refusal to certify rather than silent degradation.

Falsifiability. ULL makes operational claims that can be refuted by repeatable failures under the stated protocol, including:

- persistent degradation of cross-modal persistence on the declared suite under the declared replication settings;
- failure of the φ-banding test to separate from the stated null when the protocol is rerun with the declared thresholds;
- systematic legality violations for mined motifs despite the checker reporting success (a soundness failure), or systematic acceptance of adversarially corrupted motifs when invariants are claimed to hold.

In each case, the certificate interface is designed to localize the failure to a specific gate, configuration, or pipeline stage.

13 Broader Impacts

By providing auditable, certificate-based semantics, ULL enables auditing, safety analyses, and legal reasoning on top of machine-generated meanings. Its zero-parameter nature and certificate infrastructure encourage interoperability across sensors and organizations. The release protocol is responsible-first: certificates by default, versioned artifacts, and clear audit trails.

14 Reproducibility and Artifacts

All experiments reported in this paper are intended to be reproducible from the released artifact described in Section 7. In particular, the artifact contains:

- a reference implementation of the ULL pipeline,
- a replication script that regenerates the token dictionary, reports, and representative truth certificates, and
- a manifest of SHA-256 hashes for key outputs.

Certificates record the exact configuration used (including seeds), enabling third parties to rerun the pipeline and verify agreement on declared outputs.

15 Conclusion

We presented ULL as a zero-parameter, certificate-producing semantic code grounded in a measurement-first stance (Recognition Geometry) and designed for auditable cross-carrier meaning. The core contribution of this paper is a self-contained system specification (pipeline, certificates, and evaluation protocol) together with reproducible empirical evidence for cross-modal persistence, legality-by-construction, and a φ-banding geometry test.

The broader WM-0–WM-7 program aims to modularize and strengthen the theoretical foundations of the imported gates and to expand validation to larger, more diverse suites. Ethics integration is treated as a companion direction requiring additional releases and replication standards.

16 References

References

[1] J. Washburn, M. Zlatanović, and E. Allahyarov, *Recognition Geometry*, to appear (Axioms), 2026.

[2] J. Washburn, *Recognition Science*, preprint, 2025.

[3] J. Washburn, *Coercive Potential Minimization and Closure Certificates*, preprint, 2025.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[6] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of ICML*, 2021.

[7] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica* **14** (1978), 465–471.

[9] P. D. Grünwald, *The Minimum Description Length Principle*, MIT Press, 2007.

[10] R. Montague, "Universal grammar," *Theoria* **36** (1970), 373–398.

Appendix A: Certificate Schema

Appendix A specifies the JSON fields used in `truthify` bundles. Each certificate is a self-contained object with versioned provenance, configuration, legality metrics, and a fully inlined normal form. At the top level we record a schema of the form

```
{
  "version": "0.2.0",
  "generated_at": "...ISO 8601...",
  "inputs": { ... },
  "config": { ... },
  "normal_form_ref": ".../normal_form.json",
  "legality": { ... },
  "stability": { ... },
  "phi_reports": { ... },
  "normal_form": { ... }
}
```

The `inputs` block includes the original `signal_path`, its SHA-256 hash, the signal length, the `tokens_path` and its token count. The `config` block captures the experimental knobs that were used to derive the certificate: `top_k`, number of `perturbations`, whether a φ-ladder tightening run was

enabled and how many steps it used, the `noise_scale`, and the random `seed`. The `normal_form_ref` points to a companion file containing the canonical decomposition, while `legality` bundles the neutrality supremum norm and a Boolean flag indicating whether all LNAL invariants passed.

The `stability` block summarizes cross-perturbation agreement (e.g., Jaccard overlap across motifs) and the number of perturbation samples used to compute it. The `phi_reports` group contains the φ value inferred from the normal form, the φ estimate from the dictionary, and (optionally) a full φ-ladder trace if tightening was requested. Finally, the `normal_form` is an inlined copy of the normal-form payload: a set of top tokens with their weights, window-wise coefficients, and basic statistics of the conserved Z-series. The appendix also presents an example bundle for a synthetic benchmark in the released artifact so that readers can see how certificates and replication outputs line up.

Appendix B: Mathematical definitions used in the pipeline

This appendix records the basic mathematical objects referenced throughout the paper so that the system description is self-contained.

Eight-beat windows and neutrality. Given a real- or complex-valued discrete signal s, we segment it into contiguous blocks of length 8. A window $w \in \mathbb{C}^8$ is *neutral* if $\sum_{t=0}^{7} w_t = 0$. In practice, ULL applies a fixed projection that removes the mean component of each window, yielding a neutralized window.

DFT-8 backbone. Let $\omega := e^{-2\pi i/8}$. The 8-point discrete Fourier transform is the linear map $F : \mathbb{C}^8 \to \mathbb{C}^8$ with matrix entries

$$F_{t,k} \;=\; \frac{1}{\sqrt{8}}\,\omega^{tk}, \qquad t,k \in \{0,\dots,7\}.$$

It is standard that F is unitary (so $F^{-1} = F^*$) and diagonalizes the cyclic shift operator; see any text on discrete Fourier analysis.

Golden ratio ladder. The golden ratio is $\varphi := \frac{1+\sqrt{5}}{2}$. A φ-ladder is the multiplicative set $\{\varphi^k : k \in \mathbb{Z}\}$. In this paper, ladder banding means that certain observed distances cluster near ladder values within a stated tolerance model.

Canonical ratio cost. The canonical reciprocal cost used in RS-motivated comparisons is

$$J(x) \;:=\; \frac{1}{2}\big(x + x^{-1}\big) - 1, \qquad x \in \mathbb{R}_{>0}.$$

This function is nonnegative on $\mathbb{R}_{>0}$, symmetric under inversion ($J(x) = J(x^{-1})$), and satisfies $J(x) = 0$ iff $x = 1$, since $J(x) = \frac{(x-1)^2}{2x}$ for $x > 0$.

Ladder banding test (informal). Given a realized dictionary $\{w_i\}$ in a fixed coefficient space, we compute pairwise distances $d_{ij} = \|w_i - w_j\|_2$, assign each d_{ij} to its nearest ladder value $\varphi^{k_{ij}}$, and record residuals $|d_{ij} - \varphi^{k_{ij}}|$. A bootstrap procedure compares the observed residual distribution against a declared null model; the reported p-value is the tail probability under that null.

Appendix C: φ-Lattice Tables

The full list of pairwise distance assignments (including nearest ladder powers, residuals, and bootstrap statistics) is provided in the reproducibility artifact, in particular in `ARTIFACTS_DIR/reports/phi_quant.json`. We omit printing a full 20×20 distance table in the PDF to keep the manuscript within a journal-length budget.

Appendix D: LNAL Invariants

Appendix D summarizes the static invariants enforced by the LNAL toolchain and how they are propagated to runtime behavior. The main invariants are neutrality preservation (each 8-sample window remains mean-free), ledger consistency (aligned Z and M ledgers with nonnegative measures), and operator soundness checks for the core transforms (matrix-level validation of neutrality preservation via column-sum constraints, and coercivity via a minimum-singular-value gate). For triadic operators (e.g. `BRAID`), the reference implementation additionally enforces a legality predicate on triples of windows (nondegeneracy and a Z-ledger balance condition) and raises a machine-checkable error on violation.

For the released truth-certificate suite, the certificate-level legality fields expose (at minimum) the neutrality supremum norm and a Boolean `invariants_ok` flag, enabling independent auditors to rerun the tool and verify that legality predicates agree bit-for-bit under the recorded configuration.

Appendix E: Ablation Protocols

Appendix E describes the ablation protocols used to test the necessity of each RS ingredient. In the φ ablation, we rerun token discovery and evaluation with the φ-ladder constraint disabled, holding all other settings fixed; the resulting dictionaries lose their banded structure and cross-modal retrieval degrades, revealing how much of the stability comes from φ-quantization. In the eight-beat ablation, we replace neutral eight-beat windows with alternative segmentations (e.g., varying window length or misaligned frames) and measure modality-specific drift and grammar violations, exposing the role that eight-tick minimality plays in language- and carrier-independence. In the CPM ablation, we relax coercivity and the defect bounds and observe that token discovery collapses into degenerate or redundant atoms, confirming that the energy gap inequality is necessary to keep the periodic table of meaning well-posed. Each experiment is reported with tables documenting retrieval accuracy, legality violation rates, and the number and diversity of atoms. Full per-run tables are provided in the artifact release; the main text reports only summary statistics and refutation criteria.

Appendix F: Canonical WToken periodic table (discrete coordinates)

For comparison across runs and carriers, we treat each WToken as an intrinsic discrete identity, recorded as a triple

$$(\text{mode family},\ \varphi\text{-level},\ \tau\text{-offset}).$$

Modes 1+7, 2+6, and 3+5 admit only $\tau = 0$, while the mode-4 family admits two offsets ($\tau = 0$ and $\tau = 2$). Table 2 records the resulting 20 identities.

Atom ID	**Mode family**	**φ-level**	**τ-offset**
W_0	1+7	0	0
W_1	1+7	1	0
W_2	1+7	2	0
W_3	1+7	3	0
W_4	2+6	0	0
W_5	2+6	1	0
W_6	2+6	2	0
W_7	2+6	3	0
W_8	3+5	0	0
W_9	3+5	1	0
W_{10}	3+5	2	0
W_{11}	3+5	3	0
W_{12}	4	0	0
W_{13}	4	1	0
W_{14}	4	2	0
W_{15}	4	3	0
W_{16}	4	0	2
W_{17}	4	1	2
W_{18}	4	2	2
W_{19}	4	3	2

Table 2: Canonical 20-atom periodic table in discrete coordinates.